ANYTHING GOES?!

FORGE23

Forschungsdaten in den Geisteswissenschaften

04. - 06. OKTOBER | TÜBINGEN

EBERHARD KARLS UNIVERSITÄT TÜBINGEN

AG Datenzentren

DAASI International

Gefördert und unterstützt durch:

DHd digital humanities im deutschsprachigen raum

GEFÖRDERT VOM Bundesministerium für Bildung und Forschung

Baden-Württemberg MINISTERIUM FÜR WISSENSCHAFT, FORSCHUNG UND KUNST

FORGE 2023
Forschungsdaten in den Geisteswissenschaften

ANYTHING GOES?!
Forschungsdaten in den Geisteswissenschaften – kritisch betrachtet

– Konferenzabstracts –

Eberhard Karls Universität Tübingen
04. bis 06. Oktober 2023

Das Book of Abstracts zur FORGE 2023 ist online verfügbar unter:
https://doi.org/10.5281/zenodo.8341605

Herausgeber

Dr. Michael Derntl
Digital Humanities Center
Eberhard Karls Universität Tübingen
Keplerstr. 2
72074 Tübingen

Peter Gietz
DAASI International GmbH
Europaplatz 3
72072 Tübingen

Patrick Helling
Data Center for the Humanities (DCH)
Universität zu Köln
Albertus-Magnus-Platz
50923 Köln

Layout-Redaktion

Nicole Majka

Design

Jennifer Vosseler

Programmkomitee

Dr.in Sabine Bartsch

Fabian Cremer

Alexander Czmiel

Prof.in Dr. Aline Deicke

Dr. Michael Derntl

Dr.in Lisa Dieckmann

Swantje Dogunke

Peter Gietz

Patrick Helling

Dr. Matthias Lang

Marina Lemaire

Dr.in Katrin Moeller

Dr. Jonas Müller Laackman

Janna Neumann

Dr.in Carolin Odebrecht

Prof. Dr. Nils Reiter

Dr. Felix Schäfer

Dr. Claus-Michael Schlesinger

Dr.in Birgit Schmidt

Dr. Fabian Schwabe

Sibylle Söring

Dr. Dirk von Suchodoletz

Prof. Dr. Thomas Walter

Dr. Peter Wittenburg

Dr. Kai Joachim Wörner

Prof.in Dr. Ulrike Wuttke

Grußwort

Auch für die Geisteswissenschaften werden Daten für den Forschungsprozess immer wichtiger. Sie werden im Fall digitaler Editionen während des Forschungsprozesses erzeugt, bereits vorhandene Daten werden analysiert, es werden Geodaten mit Zeitangaben verknüpft, es werden Objekte vermessen, mit Hilfe von 3D-Scannern digitalisiert, statistische Verfahren verwendet, letztendlich auch mit KI-Methoden analysiert, Metadaten werden von Fachleuten oder via Crowdsourcing manuell erstellt oder automatisch berechnet. Heutzutage wird von jedem Forschungsprojekt bereits bei der Antragstellung ein durchdachtes Konzept für den professionellen Umgang mit Forschungsdaten erwartet, und von vielen Geldgebern sogar ein detaillierter Datenmanagementplan. Digitale Daten sind mittlerweile Teil der meisten geisteswissenschaftlichen Forschung. Es haben sich Datenzentren etabliert und mit ihnen auch ein ganzes Ökosystem.

Alle diese Trends haben sich in den letzten Jahrzehnten entwickelt und sind zu Selbstverständlichkeiten geworden. Oft werden solche Prozesse nicht mehr hinterfragt und Datenkritik, also die kritische Auseinandersetzung, warum und wie Daten erstellt oder verarbeitet werden, wird immer wichtiger. Daten müssen zugänglicher gemacht werden, es muss gesichert werden, dass Daten auch über zehn Jahre hinaus zugänglich bleiben, auch um Intersubjektivität zu gewährleisten.

Es ist deshalb ein gutes Zeichen, dass sich ein Konferenzformat zunehmend etabliert, das sich ganz den geisteswissenschaftlichen Forschungsdaten widmet. So findet die FORGE nunmehr zum vierten Mal (nach 2015 und 2016 an der Universität Hamburg und 2021 an der Universität zu Köln) als Veranstaltungsreihe der AG Datenzentren des Verbands Digital Humanities im deutschsprachigen Raum e.V. statt, um sich gezielt mit dieser Thematik zu beschäftigen und den kritischen Umgang mit geisteswissenschaftlichen Forschungsdaten zu fördern. Das Motto der FORGE 2023 ist *Forschungsdaten in den Geisteswissenschaften – kritisch betrachtet* und es wird die Frage gestellt, ob hier alles möglich ist („Anything goes?"). Was ist noch Teil des wissenschaftlichen Prozesses, was ist Spielerei, wo liegt der Erkenntnisgewinn? Der Call-for-Papers wollte genau hierzu für Beiträge werben:

„Wollen die Geisteswissenschaften ihrem gesellschaftlichen Auftrag gerecht werden und die Digitalisierung in Wissenschaft und Gesellschaft kritisch begleiten und unterstützen, bedarf es vor diesem Hintergrund einer kritischen Auseinandersetzung mit Forschungsdaten, bspw. in Bezug auf ihre Entstehung und Generierung, ihre Epistemologie sowie Provenienz oder Qualität. Hierbei stellen die zunehmende Menge und der Trend zur automatischen Verarbeitung eine Herausforderung dar."

„Dies gilt sowohl aus der Sicht der Datenerzeugung als auch der Daten(nach)nutzung für die diversen Forschungsdatentypen und ihre Entstehungskontexte. Die Entstehung literaturwissenschaftlicher Textkorpora etwa auf der Grundlage heute fragwürdiger Kanonbildung kann hier ebenso als Beispiel dienen wie die Verfügbarmachung von Nachlässen bei unklarer Rechtslage oder die Aufbereitung von Sammlungen aus kolonialen Kontexten. Wiederum ganz eigene Herausforderungen in Bezug auf die Beachtung von Gesetzeslagen (z. B. DSGVO) ergeben sich bei der (Nach-)Nutzung von z. B. Audio-visuellen oder Social Media Daten. Auch in Bezug auf die in den Geisteswissenschaften zunehmende Verwendung von Machine-Learning-Systemen und die Wichtigkeit der Beachtung von Bias, Sampling und Passgenauigkeit der Trainingsdaten spielt eine kritische Betrachtung der Forschungsdaten eine immer größere

Rolle. Viele dieser Aspekte gelten Disziplin-unabhängig und international, und gerade die Geisteswissenschaften könnten bei der kritischen Betrachtung und Nutzung von Forschungsdaten wertvolle Impulse geben. Im internationalen Kontext spielen die RDA-Resultate und die FAIR-Prinzipien eine wichtige Rolle, an die angeknüpft werden sollte.“

Und wieder sind viele Forschungsdatenexpert*innen im Bereich Digital Humanities diesem Ruf gefolgt und haben ihr Interesse an diesen Fragen mit zahlreichen Einreichungen bewiesen. Das Programmkomitee hat im Rahmen eines maximal transparenten Begutachtungsprozesses ein vielfältiges Programm zusammengestellt, das im hier vorliegenden Book of Abstracts dokumentiert ist und den Diskurs um die kritische Betrachtung von Forschungsdaten in den Geisteswissenschaften voranbringen wird.

Das Digital Humanities Center der Universität Tübingen und die DAASI International Tübingen haben sich der Organisation der diesjährigen FORGE mit Freude gewidmet und danken der AG Datenzentren für das gesetzte Vertrauen, sowie dem Programmkomitee und den Gutachter*innen für die Gestaltung des Programms.

Tübingen, Oktober 2023

Peter Gietz, Geschäftsführer der DAASI International Tübingen

Dr. Michael Derntl, Leiter des Digital Humanities Center der Universität Tübingen

Inhaltsverzeichnis

Workshops

Vorträge

Poster

Workshops

Die Ermittlung von Data Literacy Bedarfen in Studium, Lehre und Qualifikation in den historisch arbeitenden Wissenschaften. Ein kritischer Blick auf die Zwischenergebnisse der Erhebung

Döring, Laura
doeringl[at]uni-trier.de
Universität Trier, Deutschland
ORCID-iD: 0009-0001-7129-2018

Lemaire, Marina
marina.lemaire[at]uni-trier.de
Universität Trier, Deutschland
ORCID-iD: 0000-0003-4726-2481

Zusammenfassung. In Forschung und Lehre ist die frühe und zielgruppengerechte Integration von FDM- und DL-Inhalten elementar, um Datenkompetenzen effizient zu vermitteln sowie nachhaltig in die Forschungspraxis zu integrieren. Um den aktuellen Bedarf an Data Literacy Lehr- und Lernangeboten für die verschiedenen Qualifikationsstufen in den historisch orientierten Disziplinen zu ermitteln, wird in NFDI4Memory im Laufe des Jahres eine Bedarfserhebung durchgeführt. In dem Workshop sollen erste Zwischenergebnisse präsentiert und auffällige Beobachtungen in den Auswertungen mit den Workshopteilnehmenden diskutiert werden.

1 Abstract

Im Rahmen des NFDI4Memory Konsortiums beschäftigt sich die Task Area 4 „Data Literacy“ mit der Entwicklung von Services, Handreichungen und innovativen Lehr- und Lernformaten, um die Datenkompetenz (DL) und das aktive Forschungsdatenmanagement (FDM) auf allen Qualifikationsstufen in den historisch arbeitenden Disziplinen zu verbessern. Um zielgruppenspezifische und dem Niveau entsprechende DL-Angebote zu entwickeln, wird als ein erster Schritt eine Bedarfserhebung durchgeführt. Für die Befragung kann kaum auf vorhandene Fragebogendesigns zurückgegriffen werden, da bislang nur wenige Erhebungen dieser Art durchgeführt wurden. Die auffindbaren Studien ermitteln eher Digital Literacy. DL-Kompetenzen wurden dabei eher nur am

Rande erhoben.[1] Für den Bereich des FDM gibt es zwar diverse Bedarfserhebungen, die aber weniger auf den Kenntnisstand der Befragten ausgerichtet sind, sondern vielmehr auf benötigte Infrastrukturen und Services abzielen. Eine systematische Erhebung mit einem geistes- oder gar geschichtswissenschaftlichen Fokus ist bislang eine Leerstelle. Dementsprechend ist das Fragebogendesign eine besondere Herausforderung: Für dessen Entwicklung führen wir derzeitig bereits vorhandene FDM- und DL-Kompetenzmatrizen zu einem DL-Kompetenzkatalog für die historisch arbeitenden Wissenschaften zusammen. Als Grundlage werden die „Lernzielmatrix zum Themenbereich Forschungsdatenmanagement (FDM) für die Zielgruppen Studierende, PhDs und Data Stewards"[2], die „FDM-Kompetenzmatrix" des DIAMANT-Modells[3] und die „Taxonomy of Digital Research Activities in the Humanities (TaDiRAH)"[4] verwendet und anschließend an die historisch arbeitenden Disziplinen angepasst. Der so entstehende Kompetenzkatalog ist die Ausgangsbasis zur Entwicklung des Erhebungsdesigns. Es wird eine breite online-Befragung in den historisch arbeitenden Disziplinen erfolgen, die eventuell noch durch Leitfadeninterviews ergänzt wird. Die Erhebungen werden im August, September und Oktober 2023 durchgeführt, sodass zum Zeitpunkt der FORGE23 erste Auswertungen der bis dahin erhobenen Daten vorliegen werden.
Ziel des Workshops ist es, nach der Präsentation der vorläufigen Analyseergebnisse in einem intensiven Austausch mit den Teilnehmenden, besondere Auffälligkeiten in den Daten mit Expert*innen zu diskutieren, ihre Einschätzung zu Auffälligkeiten in den Daten einzuholen sowie Anregungen/Fragen für die weitere Auswertung der Daten zu erhalten.

---

1 Vgl. Roland A. Stürz u. a., *Das bidt-SZ Digitalbarometer*, pdf, hg. von Bayerisches Forschungsinstitut für Digitale Transformation bidt (München: bidt - Bayerisches Forschungsinstitut für Digitale Transformation, 2022), https://doi.org/10.35067/XYPQ-KN66; OECD, *Ländernotiz. Erhebung über die Fähigkeiten und Fertigkeiten Erwachsener. Deutschland*, 2013, http://www.oecd.org/skills/piaac/Country%20note%20-%20Germany%20(DEU).pdf.

2 Britta Petersen u. a., „Lernzielmatrix zum Themenbereich Forschungsdatenmanagement (FDM) für die Zielgruppen Studierende, PhDs und Data Stewards" (Zenodo, 5. September 2022), https://doi.org/10.5281/zenodo.7034478.

3 Marina Lemaire u. a., *Das DIAMANT-Modell 2.0. Modellierung des FDM-Referenzprozesses und Empfehlungen für die Implementierung einer institutionellen FDM-Servicelandschaft*, eSciences Working Papers 05 (Trier, 2020), 30–40, https://doi.org/10.25353/ubtr-xxxx-f5d2-fffb.

4 „TaDiRAH - Taxonomy of Digital Research Activities in the Humanities (v. 0.5.1, 05/2014)" (2014; repr., Digital Humanities Taxonomy Group, 28. Juli 2022), https://github.com/dhtaxonomy/TaDiRAH.

Der Workshop gliedert sich somit in drei Teile: Nach einem einführenden Vortrag zum Forschungskonzept und den Stand der bisherigen Erhebungen, werden Ergebnisse der vorläufigen Auswertung präsentiert und mit den Teilnehmer*innen kritisch reflektiert. Ein besonderes Augenmerk soll dabei auf Besonderheiten und Ausreiser in den Zwischenergebnissen gelegt werden.
Die Diskussionen und Anregungen werden gemeinsam mit den Workshopteilnehmenden in einem Pad dokumentiert, sodass die Ideen in die Abschlussauswertung der Daten noch mit einfließen können.
Um einen effektiven Austausch zu ermöglichen, soll die Anzahl der Workshopteilnehmenden auf 15 begrenzt sein und die Dauer des Workshops 120 Minuten betragen. Die Teilnehmenden werden im Vorfeld des Workshops gebeten, sich mit dem Fragebogen vertraut zu machen. Die dafür notwendigen Materialien werden rechtzeitig zur Verfügung gestellt.

Wissenslücken, sensible Daten, größtmögliche Transparenz?

Provenienzforschung zu Sammlungsgut aus kolonialen Kontexten ermöglichen und nach den FAIR- und CARE-Prinzipien zugänglich machen

Fründt, Sarah
sarah.fruendt[at]kulturgutverluste.de
Deutsches Zentrum Kulturgutverluste, Deutschland

Köhler, Romy
r.koehler[at]hv.spk-berlin.de
Deutsche Digitale Bibliothek, Deutschland

Werner, Sabrina
sabrina.werner[at]kulturgutverluste.de
Deutsches Zentrum Kulturgutverluste, Deutschland

Zusammenfassung. Seit einigen Jahren ist der angemessene Umgang mit Sammlungsgut aus kolonialen Kontexten in deutschen Einrichtungen ein wichtiges kulturpolitisches Thema. In diesem Zusammenhang wurden zwei Portale entwickelt, die Zugang zu den Sammlungen und ihrer jeweiligen Provenienz gewährleisten sollen: das Portal „Sammlungsgut aus kolonialen Kontexten" („Collections from Colonial Contexts"= CCC-Portal) innerhalb der Deutschen Digitalen Bibliothek und die Datenbank „Proveana" am Deutschen Zentrum Kulturgutverluste. Im Workshop können verschiedene Formen der Provenienzausweisung in beiden Portalen erprobt und diskutiert werden. Dabei geht es um das spezifische Zusammenwirken von FAIR- und CARE-Prinzipien und das Spannungsfeld zwischen erwünschter Transparenz und gebotener Vorsicht im Hinblick auf sensible Informationen.

Seit 2019 existiert durch den „Fachbereich Kultur- und Sammlungsgut aus kolonialen Kontexten" des Deutschen Zentrums Kulturgutverluste erstmals eine gezielte Förderung für die Erforschung der Provenienz von Sammlungsgut im Bereich der kolonialen Kontexte.[1] Die Ergebnisse

[1] Zuvor hat eine Arbeitsgruppe beim Deutschen Museumsbund (DMB) einen „Leitfaden zum Umgang mit Sammlungsgut aus kolonialen Kontexten in Museen und Sammlungen" herausgegeben, der seit 2018 zum dritten Mal aktualisiert wurde, siehe DMB, dritte Fassung 2021.

werden in der Forschungsdatenbank Proveana[2] dokumentiert. Seit 2021 wird in der Deutschen Digitalen Bibliothek das CCC-Portal[3] (Collections from Colonial Contexts), als ein zentrales Online-Portal aufgebaut, das bereits erschlossenes und digitalisiertes Sammlungsgut aus kolonialen Kontexten in deutschen Museen und Wissenseinrichtungen mit einem besonderen Fokus auf die Ergebnisse jüngerer Provenienzforschung weltweit zugänglich macht.[4] Die kontextbedingte Sensibilität der Daten erfordert ein konzeptuelles Zusammenwirken der FAIR[5]- und CARE[6]-Prinzipien im CCC-Datenfeldkatalog.
Ziel ist nun mittelfristig, die Synergieeffekte zwischen den aus der Provenienzforschungsförderung resultierenden Forschungsdaten zwischen beiden Portalen zu eruieren - und damit nicht nur den politisch vielfach geforderten transparenten Zugang zu den Beständen, sondern auch die Forschung zu kolonialhistorischen Zusammenhängen aus musealen Perspektiven zu ermöglichen.

Die Entwicklung von Onlineportalen wie Proveana oder Collections from Colonial Contexts wirft eine Reihe von Fragen zu den Möglichkeiten und Grenzen auf, (weltweite) digitale Transparenz sensibler Daten herzustellen – neben dem Umfang und der Qualität der zur Verfügung gestellten Daten geht es dabei insbesondere um Fragen der

---

[2] https://www.proveana.de/de/start
[3] https://ccc.deutsche-digitale-bibliothek.de/
[4] Für eine Beschreibung des politischen Kontextes des CCC-Portals und die bisher in der DDB umgesetzten Arbeitsschritte siehe Köhler/Spohr 2023.
[5] Die vier Grundprinzipien des fairen Datenmanagements lauten, Daten auffindbar (**f**indable), zugänglich (**a**ccessible), **i**nteroperabel und nachnutzbar (**r**eusable) zu machen – Für eine Erläuterung der einzelnen Prinzipien und Anwendungseempfehlungen für Akteure der NFDI4Culture-Communities siehe Kailus 2023.
[6] Die 2019 gegründete Global Indigenous Data Alliance (GIDA) hat im Zusammenhang mit der Internatonal Indigenous Data Sovereignty Interest Group innerhalb der Research Data Alliance (RDA) die CARE-Prinzipien für die Handhabung indigener Daten dezidiert als notwendige Ergänzung der FAIR-Prinzipien formuliert, um die Daten(nach)nutzung im Sinne der Interessen indigener Völker und in Übereinstimmung mit indigenen Werten zu lenken sowie deren Beteiligung an Entscheidungsprozessen zu stärken: CARE steht für **C**ollectve Benefit (Kollektiver Nutzen), **A**uthority to Control (Kontrolle über die Daten), **R**esponsibility (Verantwortungsbewusstsein) und **E**thics (Ethik), siehe Imeri/Rizolli 2022: 3, vgl. auch Carroll et al. 2022.

kolonialzeitlichen Agency von Wissenschaftler*innen und anderen Akteur*innen beim „Erwerb“ von Sammlungsgut, die einen Würdeverlust kolonisierter Akteur*innen mit und nach sich zogen, der nicht digital reproduziert werden kann und darf. Das schließt den Umgang mit diskriminierenden oder auch verharmlosenden Begriffen oder Abbildungen im historischen Material und den jeweiligen institutionellen Objekterschließungssystemen und -terminologien mit ein.
Der Workshop geht der Frage nach, inwiefern eine digitale Ausweisung der Ergebnisse jüngerer Provenienzforschung nach internationalen Standards unter Berücksichtigung der spezifischen Historizität und Sensibilität der Daten gelingen kann.[7] Während Proveana vor allem detaillierte Informationen zu Akteur*innen, historischen Ereignissen oder verschiedene Quellen zur weiteren Forschung, verzeichnet, verknüpft und visualisiert[8], stellt das CCC-Portal neue Provenienzforschungsdaten einzelobjektbezogen und ereignisbasiert (Ort, Zeit, beteiligte Akteur*innen) in einem zentralen Onlineportal zu den Beständen an Sammlungsgut aus kolonialen Kontexten in den verschiedenen deutschen Museen und Wissenseinrichtungen zur Verfügung. Die Datensätze zu Sammlungsgut aus ethnologischen, naturkundlichen, archäologischen, kunst- und kulturhistorischen Museen und Universitätssammlungen wie auch kleineren Regional- und Lokalmuseen stellt – über den ethisch angemessenen Umgang mit in verschiedener Hinsicht sensiblen Informationen hinaus – auch erhöhte Anforderungen an die Standardisierung von Daten und die Entwicklung zentraler Datenfeldkataloge.

Im Workshop können die Teilnehmenden die unterschiedlichen Formen der Provenienzausweisung in beiden Datenbanken kennenlernen. Anhand der konkreten Arbeit an einzelnen Objektdatensätzen werden die Möglichkeiten und Grenzen der objekt- und ereignisbasierten Provenienzausweisung erfahrbar und das spezifische Zusammenwirken von CARE- und FAIR-Prinzipien im CCC-Datenfeldkatalog erläutert. Der Umgang mit ethisch und rechtlichen sensiblen Informationen wird erweiternd anhand von Akteurs- und

[7] Vgl. auch Hahn et al. 2021.
[8] Werner 2020, 2023.

Ereignisdatensätzen sowie hinterlegten Volltexten in Proveana vorgestellt.

In der auf diesen praktischen Einblicken in die ereignisbasierte und netzwerkzentrierte Provenienzausweisung aufbauenden Diskussion soll es zum einen um die Schnittstellen beider Datenbanken sowie mit anderen nationalen und internationalen Initiativen (Normdaten, Daten-Standardisierungsverfahren, Verwendung nationaler und internationaler Thesauri) gehen. Zum anderen werden inhaltliche und programmatische Fragen postkolonialer Ausrichtung erörtert: Wie könnte eine ereignisbasierte Provenienzkette aussehen, die nicht auf europäische „Sammler*innen“ fokussiert ist? Wie lässt sich in beiden Portalen die Agency der nicht historisch überlieferten Hersteller*innen und Bewahrer*innen von Kulturgut herauslesen und sichtbar machen? Oder etwas allgemeiner formuliert: wie kann verhindert werden, dass Wissenslücken in der Dokumentation der Provenienz von Sammlungsgut aus kolonialen Kontexten zu einseitigen inhaltlichen Aussagen oder fortgeführten Asymmetrien führen? Wie kann der Umgang mit sensiblen Daten im Einklang mit den Vorgaben der verschiedenen Akteur*innen gestaltet werden, ohne dabei das Ziel der größtmöglichen Transparenz aus den Augen zu verlieren?

Der Workshop soll Teilnehmende zusammenbringen, die – auch aus inhaltlich anderen Themengebieten – mit ähnlichen Möglichkeiten und Herausforderungen konfrontiert sind und/oder sich für die angemessene ‚Übersetzung‘ ethischer Fragen in digitale Infrastrukturen interessieren. Als ein wesentliches Ergebnis werden verschiedene Modelle des konzeptuellen Zusammenwirkens von CARE und FAIR-Prinzipien in der Ausweisung von Provenienzdaten vorgestellt, weiterentwickelt und in ihren Möglichkeiten und Grenzen diskutiert. Außerdem sollen Weiterentwicklungspotenziale in der Datenstruktur und der Normierung von Akteurs- und Ereignisdatendatensätzen diskutiert werden, um diese für die Forschung und die Vernetzung zu optimieren. (Weltweite) Transparenz über die Provenienz von Sammlungsgut aus kolonialen Kontexten wird so als die kollaborative Entwicklung dekolonialer digitaler Infrastrukturangebote in zwei verschiedenen Provenienzdatenbanken erfahrbar.

Zeitbedarf: 180 Minuten

Koordination und Durchführung:

Sarah Fründt M.A., wissenschaftliche Referentin im Fachbereich „Kultur- und Sammlungsgut aus kolonialen Kontexten“ des Deutschen Zentrum Kulturgutverluste

Romy Köhler M.A., wissenschaftliche Mitarbeiterin im Projekt „Sammlungsgut aus kolonialen Kontexten“ in der Geschäftsstelle der Deutschen Digitalen Bibliothek

Sabrina Werner M.A., Referentin „Proveana“ im Fachbereich „Dokumentation und Forschungsdatenmanagement“ des Deutschen Zentrums Kulturgutverluste

Bibliografie

Sammlungsgut aus kolonialen Kontexten - Sammlungsgut aus kolonialen Kontexten (deutsche-digitale-bibliothek.de)

Proveana

Carroll, Stephanie Russo, Ibrahim Garba, Rebecca Plevel, Desi Small-Rodriguez, Vanessa Y Hiratsuka., Maui Hudson, and Nanibaa' A. Garrison "Using Indigenous Standards to Implement the CARE Principles: Setting Expectations through Tribal Research Codes". *Frontiers in Genetics* (Sec. ELSI in Science and Genetics) 13, (March 2022). https://doi.org/10.3389/fgene.2022.823309.

Deutscher Museumsbund e.V. *Leitfaden Umgang mit Sammlungsgut aus kolonialen Kontexten*. 3. Fassung. Berlin, 2021. Verfügbar unter: https://www.museumsbund.de/wp-content/ uploads/2021/03/mb-leitfanden-web-210228-02.pdf [Zugriff am: 28.4.2023].

Hahn, Peter, Oliver Lueb, Katja Müller, und Karoline Noack. *Digitalisierung ethnologischer Sammlungen: Perspektiven aus*

Theorie und Praxis. Bielefeld: transcript Verlag, 2021. https://doi.org/10.1515/9783839457900-008

Sabine Imeri, Michaela Rizzolli (2022): *CARE Principles for Indigenous Data Governance: Eine Leitlinie für ethische Fragen im Umgang mit Forschungsdaten?.* O-Bib. Das Offene Bibliotheksjournal 9(2), 1-14. https://doi.org/10.5282/o-bib/5815

Kailus, Angela. *Handreichung für ein FAIRes Management kulturwissenschaftlicher Forschungsdaten*. Zenodo (2023) https://doi.org/10.5281/zenodo.7716941

Köhler, Romy; Spohr, Julia. „Sammlungsgut aus kolonialen Kontexten – ein Subportal der Deutschen Digitalen Bibliothek." *Zeitschrift für Bibliothekswesen und Bibliografie* 1, (2023): 19-25. http://dx.doi.org/10.3196/186429502070130

Werner, Sabrina: „Proveana – Eine Datenbank für die Provenienzforschung am Deutschen Zentrum Kulturgutverluste," *Bibliothekdienst* 57, No. 1 (2023): 19-27. https://doi.org/10.1515/bd-2023-0006

Werner, Sabrina. "Proveana. Zur Entwicklung der Forschungsdatenbank des Deutschen Zentrums Kulturgutverluste," *Provenienz und Forschung* 1 (2020): 26-33.

Faires FDM für digitale Editionen

Hegel, Philipp
philip.hegel[at]tu-darmstadt.de
Technische Universität Darmstadt, Deutschland
ORCID-iD: 0000-0001-6867-1511

Hensen, Kilian
kilian.hensen[at]uni-koeln.de
CCeH, Universität zu Köln, Deutschland
ORCID-iD: 0000-0001-6708-1237

König, Sandra
sandra.koenig[at]leopoldina.org
Leopoldina - Nationale Akademie der Wissenschaften, Deutschland
ORCID-iD: 0000-0002-0615-0523

Kudella, Christoph
kudella[at]sub.uni-goettingen.de
Niedersächsische Staats- und Universitätsbibliothek Göttingen, Deutschland
ORCID-iD: 0000-0002-9645-7122

Lemke, Karoline
karoline.lemke[at]bbaw.de
Berlin-Brandenburgische Akademie der Wissenschaften, Deutschland
ORCID-iD: 0000-0002-1604-672X

Schulz, Daniela
schulz[at]hab.de
Herzog August Bibliothek Wolfenbüttel, Deutschland
ORCID-iD: 0000-0003-3167-5089

Seltmann, Melanie
melanie.seltmann[at]tu-darmstadt.de
ULB Darmstadt, Deutschland
ORCID-iD: 0000-0002-7588-4395

Zusammenfassung. Die Berücksichtigung der FAIR-Prinzipien während des editorischen Prozesses bildet die Grundlage für ein ganzheitliches und nachhaltiges Forschungsdatenmanagement, welches nur gemeinschaftlich umgesetzt werden kann. Ziel des Workshops ist es, unter den Teilnehmer: innen das Problembewusstsein dafür zu schärfen, vor welchen Herausforderungen sie in diesem Prozess stehen. Anhand kurzer Impulse werden die FAIR-Prinzipien, weitere Leitlinien sowie deren Niederschlag in der Forschungsförderung skizziert, bevor gemeinsam

Maßnahmen zu deren Umsetzung erarbeitet werden. Der Workshop richtet sich an alle am Prozess von Datenerstellung, Datenmanagement und -nutzung beteiligten Personen. Der Workshop wird von der Datendomäne Editionen des NFDI-Konsortiums Text+ organisiert und findet im World Café-Format statt.

1 Problemstellung und Ziel

1.1 Die FAIR-Prinzipien in der editionswissenschaftlichen Praxis

Die FAIR[1]-Prinzipien haben sich seit ihrer Veröffentlichung als Richtlinien für den Umgang mit Forschungsdaten etabliert und sind fester Bestandteil von Förderprogrammen. In der Praxis der digitalen Editionen zeigt sich allerdings, dass der genaue Inhalt der als Leitlinien gedachten Prinzipien – Findable, Accessible, Interoperable, Reusable – in geringerem Maße bekannt ist als gemeinhin angenommen.[2] Sie werden häufig mit anderen Begrifflichkeiten und Aspekten wie Open Access, offenen Lizenzen oder Barrierefreiheit gleichgesetzt oder vermengt.[3] Zugleich sind die FAIR-Prinzipien bewusst allgemein formuliert, sodass ihre Anwendung im jeweiligen disziplinären Kontext analysiert und adaptiert werden müssen.[4] Bislang fand im Bereich der digitalen Editorik noch wenig kritische Reflexion zur Umsetzbarkeit statt.[5] Hieraus ergibt sich der Bedarf, die FAIR-Prinzipien domänen- bzw. disziplinspezifisch auszudeuten und mit Maßnahmen zu konkretisieren.

Digitale Editionsprojekte zeichnen sich auf mehreren Ebenen durch große Heterogenität aus: Die Vielzahl der mit der Erarbeitung von Editionen befassten Fachwissenschaften mit ihren spezifischen Forschungsfragen und -gegenständen (z. B. Briefe, Manuskripte, Filme, Bilder, Münzen) und ganz eigenen editorischen Traditionen, resultiert unweigerlich in einem Nebeneinander verschiedenster editorischer Modelle, Methoden und Datenformaten.[6] Neben fachlichen Standards und der Zielgruppe determiniert auch die Ressourcenverfügbarkeit die Umsetzung von Editionsprojekten und damit die Möglichkeiten ihrer

[1] Vgl. „FAIR Principles“; Wilkinson et al. 2016.
[2] Vgl. dagegen die breiten Vorstellungen in Bezug auf Accessibilty: Martinez et al. 2019.
[3] Vgl. Gengnagel, Neuber, Schulz 2023, 6.
[4] Vgl. Gengnagel, Neuber, Schulz 2023, 1; vgl. z.B. die Umsetzung der FAIR-Prinzipien für den spezifischen Fall der Forschungssoftware, Chue Hong et al. 2022.
[5] Vgl. unveröffentlichte Masterarbeit Windeck 2019.
[6] Vgl. Sahle 2013, S. 12.

nachhaltigen Verfügbarmachung. Zudem sind digitale Editionen oft weniger abgeschlossen als Printausgaben. Diese Prozesshaftigkeit bringt zusätzliche Herausforderungen für das Forschungsdatenmanagement mit sich.[7]

Digitale Editionen sind in eine Prinzipien-Landschaft eingebettet:[8] Während die FAIR-Prinzipien den Fokus primär auf technisch-formale Aspekte legen, zielen die CARE[9]-Prinzipien der Global Indigenous Data Alliance (GIDA) darauf ab, Rechte und Interessen indigener Völker adäquat zu berücksichtigen. Für Aspekte der Barrierearmut entwickelte die W3C-Web Accessibility Initiative die POUR[10]-Prinzipien. Den gewachsenen Anforderungen an einen verantwortungsvollen Umgang mit Daten insbesondere bei automatisierten Vorgängen begegnen die vom niederländischen Konsortium Responsible Data Science (RDS) entwickelten FACT[11]-Prinzipien.

Die Organisator:innen sind Mitarbeiter:innen der Datendomäne Editionen im NFDI-Konsortium Text+ und setzen sich dort intensiv mit den Fragen der praktischen Umsetzung der FAIR- und weiterer Prinzipien auseinander. Eine erste öffentliche Annäherung an das Thema erfolgte mit den Meetups des FAIR February.[12] Die Überlegungen in einem Workshop fortzuführen und zu festigen, ergab sich als logische Konsequenz mit dem Ziel, den Austausch mit der Community zu stärken und Maßnahmen aus der Praxis heraus zu entwickeln.

1.2 Ziel des Workshops

Ziel des Workshop ist es, unter den Teilnehmer:innen die Kenntnis der Prinzipien zu stärken und ein vertieftes Verständnis für Herausforderungen, aber auch Chancen herauszubilden, wenn die genannten Leitlinien in den Editionsprozess eingebunden werden. Darauf aufbauend wird gemeinsam explorativ erarbeitet, in welcher Phase der Editionsarbeit sie jeweils zum Tragen kommen.

---

[7] Vgl. Dängeli, Stuber 2020, 37.
[8] Vgl. Deppe 2020.
[9] Vgl. „CARE Principles“ (CARE: Collective Benefit, Authority to Control, Responsibility, Ethics).
[10] Vgl. „Accessibility Principles“ (POUR: Perceivable, Operable, Understandable, Robust).
[11] Vgl. „Mission“ (FACT: Fairness, Accuracy, Confidentiality, Transparency).
[12] Vgl. „FAIR February Meetup“.

Die Veranstaltung richtet sich an alle am editorischen Prozess sowie der Datenerstellung Beteiligten, wie Editor:innen, weitere Fachwissenschaftler:innen, Forschungsdatenmanager:innen, Datenkurator:innen, Datenmodellierer:innen, Research Software Engineers, Vertreter:innen von Träger- und Förderinstitutionen sowie Nutzer:innen digitaler Editionen. Der Fokus des Workshops wird auf den Typ der Textedition beschränkt. Als Leitfaden dient ein generischer Datenmanagementplan. Die Ergebnisse werden im Text+ Blog publiziert. Ebenso werden sie in Handreichungen von Text+ eingehen.

2 Ablauf

Um die Teilnehmer:innen in einen guten Austausch zu bringen, wird der Workshop im Format eines World Cafés durchgeführt. Nach Impulsen zu den genannten Prinzipien und den Anforderungen aus Förderperspektive (inkl. Datenmanagementplanung) folgen drei durch Leitfragen orientierte World Café-Runden:

1. Die Gruppen machen sich jeweils mit einer Rolle im Prozess der digitalen Editionspraxis vertraut.
2. Die Gruppen kommen in verschiedenen Rollen zusammen und beschäftigen sich gemeinsam mit den Anforderungen eines Datenmanagementplans.
3. Diese Datenmanagementpläne werden mit der Praxis in digitalen Editionsprojekten abgeglichen.
4.

In jeder Runde werden die Gruppen neu zusammengesetzt, so dass jeweils unterschiedliche Perspektiven aufeinandertreffen. So werden unvermutete bzw. aus nur einer Perspektive leicht übersehbare Herausforderungen aufgedeckt. Pat:innen protokollieren die Diskussionen der einzelnen Tische.

2.1 Zeitplan

Der Workshop ist auf eine Dauer von 180 min angelegt.

(15 min) Begrüßung und Vorstellungsrunde
(10 min) Einführung
(20 min) Impuls I + II
(30 min) World Café-Runde 1
(05 min) Pause
(30 min) World Café-Runde 2
(30 min) World Café-Runde 3

(05 min) Pause
(20 min) Vorstellung der Tisch-Ergebnisse
(15 min) Diskussion und Wrap-Up

2.2 Technische Angaben zum Workshop

Der Workshop findet auf Deutsch statt und ist auf 12–40 Teilnehmer ausgerichtet, Vorwissen wird nicht vorausgesetzt. Für die Einführung, Impulsvorträge und Abschluss werden Computer und Beamer benötigt; für die World Café-Runden 4-5 Flipcharts, entsprechend Papier, Klebezettel in verschiedene Farben, Stifte.

Bibliografie

„Accessibilty Principles“: W3C Web Accessibility Initiative WAI, „Accessibilty Principles“. Accessed April 27, 2023. https://www.w3.org/WAI/fundamentals/accessibility-principles/.

„CARE Principles“: GIDA Global Indigenous Data Alliance, „CARE Principles for Indigenous Data Governance“. Accessed April 27, 2023. https://www.gida-global.org/care.

Carroll et al. 2021: Carroll, Stephanie Russo, Edit Herczog, Maui Hudson, Keith Russell, and Shelley Stall. „Operationalizing the CARE and FAIR Principles for Indigenous Data Futures.“ *Scientific Data* 8, no. 1 (April 16, 2021): 108. https://doi.org/10.1038/s41597-021-00892-0.

Chue Hong et al. 2022: Chue Hong, Neil P., Daniel S. Katz, Michelle Barker, Anna-Lena Lamprecht, Carlos Martinez, Fotis E. Psomopoulos, Jen Harrow, Leyla Jael Castro, Morane Gruenpeter, Paula Andrea Martinez, et al. „FAIR Principles for Research Software (FAIR4RS Principles)“, May 24, 2022. https://doi.org/10.15497/RDA00068.

Dängeli, Stuber 2020: Dängeli, Peter, and Martin Stuber. „Nachhaltigkeit in langjährigen Erschliessungsprojekten.“ *xviii.Ch. Schweizerische Zeitschrift für die Erforschung des 18. Jahrhunderts* 11 (2020): 34–51. https://doi.org/10.24894/2673-4419.00004.

Deppe 2020: Arvid Deppe. „FAIR, CARE und mehr Prinzipien für einen verantwortungsvollen Umgang mit Forschungsdaten.“ In Historisches Erbe und zeitgemäße Informationsinfrastrukturen:

Bibliotheken am Anfang des 21. Jahrhunderts. Festschrift für Axel Halle, edited by Matthias Schulze, 299–312. Kassel: kup – kassel university press, 2020. https://doi.org/10.17170/kobra-202010131934.

„FAIR February Meetup“: „Erstes FAIR February Meetup – Let’s talk FAIR Digital Editions“. Accessed April 27, 2023, https://events.gwdg.de/event/413/.

„FAIR Principles“: GO FAIR Initiative, „FAIR Principles“. Accessed April 27, 2023. https://www.go-fair.org/fair-principles/.

Gengnagel, Neuber, Schulz 2023: Gengnagel, Tessa, Frederike Neuber, and Daniela Schulz. „EDITORIAL: FAIR Enough? Evaluating Digital Scholarly Editions and the Application of the FAIR Data Principles.“ *RIDE* 16 (2023). https://doi.org/10.18716/ride.a.16.0.

Martinez et al. 2019: Martinez, Merisa, Wout Dillen, Elli Bleeker, Anna-Maria Sichani, and Aodhán Kelly. „Refining Our Conceptions of ‘Access’ in Digital Scholarly Editing: Reflections on a Qualitative Survey on Inclusive Design and Dissemination.“ *Variants. The Journal of the European Society for Textual Scholarship*, no. 14 (March 20, 2019): 41–74. https://doi.org/10.4000/variants.1070.

„Mission“: Responsible Data Science, „Mission“. Accessed April 27, 2023. https://redasci.org.

Sahle 2013: Sahle, Patrick. Digitale Editionsformen. Zum Umgang mit der Überlieferung unter den Bedingungen des Medienwandels. Teil 1: Das typografische Erbe (Schriften des Instituts für Dokumentologie und Editorik 7). Norderstedt: BoD, 2013. http://kups.ub.uni-koeln.de/id/eprint/5351.

Wilkinson et al. 2016: Wilkinson, Mark D., Michel Dumontier, IJsbrand Jan Aalbersberg, Gabrielle Appleton, Myles Axton, Arie Baak, Niklas Blomberg, et al. „The FAIR Guiding Principles for Scientific Data Management and Stewardship.“ *Scientific Data* 3, no. 1 (March 15, 2016): 160018. https://doi.org/10.1038/sdata.2016.18.

Windeck 2019: Jürgen Windeck, „FAIRe digitale Editionen. Anforderungen für nutzbare wissenschaftliche Editionen“ (Masterarbeit, Technische Hochschule Köln, 2019; Manuskript).

Einführung in ExPresS XR

Einfache Erstellung von öffentlichkeitswirksamen XR-Ausstellungen im Kunst- & Kulturbereich

Körner, Kevin
kevin.koerner[at]uni-tuebingen.de
Universität Tübingen, Deutschland
ORCID-iD: 0000-0001-7607-9876

Dreiling, Luca
luca.dreiling[at]student.uni-tuebingen.de
Universität Tübingen, Deutschland

Zusammenfassung. Um (junge) Menschen zu motivieren, mehr ins Museum zu gehen, ist es wichtig die medialen Interessen der Zielgruppe zu kennen – beispielsweise die Erweiterte Realität (XR) – und diese in Kunst- und Kulturausstellungen zu integrieren. Wir möchten auf der Forge23-Tagung einen Workshop zum ExPresS XR-Projekt anbieten, dass es ermöglicht Applikationen der Erweiterten Realität ohne Programmiererfahrung zu entwickeln. Zudem stellt es über den OpenXR-Standard sicher, dass erstellte Anwendungen auf den verbreitetsten XR-Geräten ohne großen Änderungsaufwand verwendet werden können. Der Workshop soll eine XR-Livesession umfassen, auf der die Teilnehmenden Erfahrungen mit der Technologie sammeln dürfen sowie eine Hands-on-Session zur Arbeit mit ExPresS XR.

In den vergangenen Jahren ist die Anzahl von Kunst- und Kultureinrichtungsbesuchenden immer wieder starken Schwankungen unterlegen (Institut für Museumsforschung, 2022). Der wachsende Anspruch an Digitalisierung sowie Barrierefreiheit bei der Besucherschaft sowie in der Forschung stellt hier insbesondere das Personal aus den Geisteswissenschaften vor große Herausforderungen: Wie können eine zeitgemäße Bewahrung, Aufarbeitung und insbesondere Präsentation von Kulturgütern, zur Festigung des Zusammenhalts der Gesellschaft, erreicht werden?

Ein Ansatz, um diesem Trend entgegenzuwirken und einen Vorteil im Werben um (junge) Besuchende zu erhalten, ist der Einsatz moderner Technologien in Ausstellungen, um die medialen Interessen der Zielgruppe zu erreichen (Pop & Borza, 2016); z.B. die sich langsam etablierende Erweiterte Realität (XR). Einige geisteswissenschaftliche Fachbereiche, wie beispielsweise die Archäologie oder die

Geschichtswissenschaften, forschen bereits, indem sie Objekte in dreidimensionalen Digitalisaten erfassen und diese analysieren. Diese somit ohnehin digital verfügbaren Kulturgüter können zudem mit beschreibenden Forschungsdaten erweitert auf dem modernen Medium XR präsentiert werden, was den Nutzenden beispielsweise ermöglicht, immersiv an historischen Ereignissen teilzunehmen, Exponate zu betrachten, die aufgrund von Platzrestriktionen im Museum nicht ausgestellt werden könnten oder mit diesen zu interagieren (Shehade, 2020).

Jedoch ist diese strategische Ausrichtung für Einrichtungen auch mit Problemen und Risiken verbunden: Einerseits fehlt häufig technikerfahrenes Personal in der kuratorischen Leitung, andererseits entwickelt sich der Technologiemarkt derart schnell weiter, dass sich die Einarbeitung in eine Technologie oder der Erwerb von Hardware bereits nach minimaler Zeit als nicht mehr zeitgemäß und somit finanzieller Misserfolg erweist.

Diese Problemstellung geht das Open-Source-Projekt ExPresS XR[1] an. Als Plugin der etablierten Spielentwicklungsumgebung Unity[2] ermöglicht sie die Erstellung immersiver XR-Erlebnisse, ohne dafür Programmierkenntnisse zu benötigen (Abb. 1). Neben einer dialogbasierten Erstellung der nötigen Basiselemente (Abb. 2) bietet ExPresS XR eine grafische Oberfläche zur Bearbeitung der individuellen Elemente in der XR-Anwendung während des Entwicklungsprozesses. Beispielsweise können Nutzende somit per Klick virtuelle Räume anlegen und in diesen 3D-Modelle platzieren. Darüber hinaus enthält ExPresS XR eine Vielzahl an bereits fertig implementierten dynamischen und interaktiven Komponenten, die per Drag&Drop integriert werden können; beispielsweise Animationen für Exponate oder auch frei konfigurierbare Quiz.

Weiterhin umfasst das Projekt optionale Elemente für die Erfassung des Verhaltens von Nutzenden (z.B. Tracking von Bewegungen oder Reaktionszeit) zur Unterstützung XR-basierter Studien. Auch erlaubt der modulare Aufbau die Einbettung eigener Unity-Projekte, wodurch bei Bedarf extern erstellte Projekte in die Ausstellungen integriert werden können; beispielsweise fachlich passende Serious Games.

[1] https://github.com/eisclimber/ExPresS-XR
[2] https://unity.com/

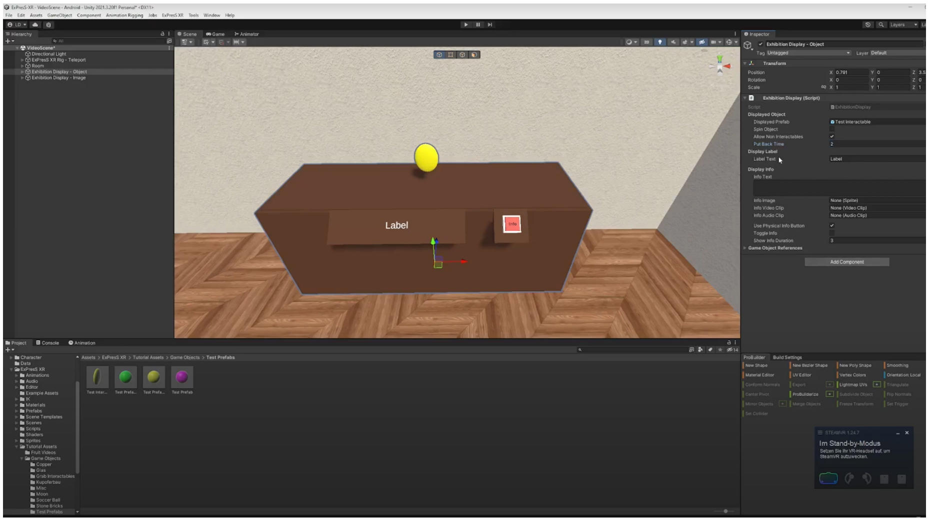


Abb. 1. Oberfläche von Unity mit ExPresS XR zur Bearbeitung von Ausstellungsräumen

ExPresS XR basiert auf dem OpenXR-Standard[3], welcher die Kommunikation zwischen verschiedenen Plattformen zur Entwicklung von Anwendungen der Erweiterten Realität und deren Hardware vereinheitlicht. Dies stellt sicher, dass erstellte Ausstellungen, ohne zusätzlichen Entwicklungsaufwand, auf den verbreitetsten XR-Geräten funktionieren; beispielsweise auf den Geräten der Firmen Oculus (Quest), HTC (Vive), Valve (Index), Microsoft (HoloLens 2) oder auch als Augmented Reality Anwendung auf dem Smartphone.

Für die Forge23 Tagung möchten wir mit dieser Einreichung einen Workshop anbieten, der in 120 Minuten anhand von konkreten Fallbeispielen aus dem Museumsbereich den grundlegenden Funktionsumfang von ExPresS XR präsentiert und den Teilnehmenden einen Einstieg in dessen Nutzung vermittelt.

[3] https://www.khronos.org/openxr/

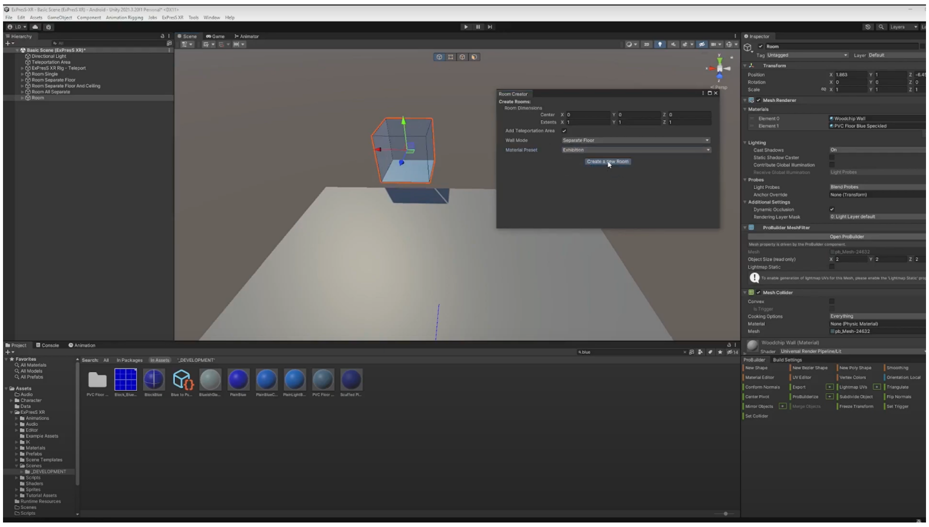


Abb. 1. Raumerstellung in ExPresS XR.

Wir planen, den Workshop in zwei Blöcke aufzuteilen. Der erste, 60 Minuten umfassende, Block erörtert den Teilnehmenden zunächst als Impulsvortrag die didaktischen Konzepte von zwei Ausstellungen, die wir in Kooperation mit Museen unter Einbeziehung von ExPresS XR entwickelt haben. Um die Teilnehmenden direkt in die immersiven Erfahrungen eintauchen zu lassen, planen wir, dass sie vor Ort die erstellten XR-Anwendungen auf von uns bereitgestellter Hardware begutachten und somit Erfahrung mit der Technologie sammeln können.

Der zweite Block führt die Teilnehmenden in einer geführten Hands-on-Session an die Entwicklung von XR-Ausstellungen mittels ExPresS XR heran. Basierend auf einem von uns bereits mehrfach verwendeten Ablaufplan, erlernen sie die Einrichtung eines virtuellen Ausstellungsraums, die Integration von Exponaten in diesen und dessen Erweiterung um spielerische Elemente. Ihre somit erstellte Ausstellung können die Teilnehmenden direkt vor Ort auf der vorhandenen Hardware testen und ihre Projektdateien für weitere Experimente mit nach Hause nehmen. Um Probleme mit der Installation von Unity und ExPresS XR während des Blocks zu vermeiden, stellen wir eine ausreichend große Menge an vorinstallierten Laptops bereit. Je nach Teilnehmerzahl müssen die Teilnehmenden diesen Block in Gruppe bearbeiten.

Den Workshop beenden wir in einer Plenumsdiskussion, welche den Teilnehmenden nochmals die Möglichkeit bietet, Fragen zu stellen sowie Kritik und Anmerkungen anzubringen und in der Gruppe zu diskutieren.

Bibliografie

Pop, Izabela Luiza & Borza, Anca. (2016). “Technological innovations in museums as a source of competitive advantage”, 2016, Munich Personal RePEc Archive, https://mpra.ub.uni-muenchen.de/76811/1/MPRA_paper_76811.pdf, Last access: 14.05.2023

Institut für Museumsforschung, “Materialien aus dem Institut für Museumsforschung, Heft 76: Statistische Gesamterhebung an den Museen der Bundesrepublik Deutschland für das Jahr 2020.” Including an English Summary, Berlin 2022 (179 S.), Publiziert bei arthistoricum.net, Universitätsbibliothek Heidelberg 2022, ISSN-Internet: 2747-9382; ISSN-Print: 2747-9366, DOI: https://doi.org/10.11588/ifmzm.2022.1

Shehade, Maria, and Theopisti Stylianou-Lambert. 2020. "Virtual Reality in Museums: Exploring the Experiences of Museum Professionals" *Applied Sciences* 10, no. 11: 4031. https://doi.org/10.3390/app10114031

Semantic Annotation of Heterogeneous, Multimedia Cultural Research Data

A FOSS Toolchain for the Digital Humanities

Rossenova, Lorenza
lozana.rossenova[at]tib.eu
TIB – Leibniz Information Centre for Science and Technology, Deutschland
ORCID-iD: 0000-0002-5190-1867

Sohmen, Lucia
lucia.sohmen[at]tib.eu
TIB – Leibniz Information Centre for Science and Technology, Deutschland

Duchesne, Paul
paul.duchesne[at]tib.eu
TIB – Leibniz Information Centre for Science and Technology, Deutschland

Günther, Lukas
lukas.guenther[at]tib.eu
TIB – Leibniz Information Centre for Science and Technology, Deutschland

Schubert, Zoe
zoe.schubert[at]tib.eu
TIB – Leibniz Information Centre for Science and Technology, Deutschland

Blumel, Ina
ina.blumel[at]tib.eu
TIB – Leibniz Information Centre for Science and Technology, Deutschland
ORCID-iD: 0000-0002-3075-7640

Summary. This workshop will present a workflow for structuring and annotating multimedia datasets within a collaborative, linked open data environment. Participants will be able to take part in practical demonstrations and provide feedback on the Semantic Kompakkt toolchain that connects three existing open source software tools: 1) OpenRefine – for data reconciliation and batch upload; 2) Wikibase – for linked open data storage; and 3) Kompakkt – for rendering and annotating 3D models, and other 2D and AV media files. This toolchain was developed in the context of NFDI4Culture with a particular focus on increasing interoperability and data reuse across different domains of cultural research.

1 Extended Abstract

Traditionally, the development of descriptive metadata standards and collection management systems has focused on the management of the separations between analog objects, their descriptive surrogates and the creation of finding aids.[1] But in the context of digital humanities research, born-digital artifacts, such as various media files, software applications and their associated metadata records, are created, managed and used in the same environment. Describing the ever-expanding range of significant properties of these artifacts,[2] tracking data provenance,[3] and allowing collaborative research and annotation[4] can prove particularly challenging to traditional relational database systems and their information architecture models.[5,6] The latter tend to separate schema from content, depend on fixed vocabularies and categorizations, and remain siloed and closed off to external audiences. In the field of cultural heritage and the digital humanities, 2D- and 3D-digital representations of cultural assets are particularly heterogeneous in formats and structure[7], hence standardized access and visualisation tools fail to meet the objectives of critical digital humanities research as well as related research funding requirements (e.g. for FAIR data). To bridge the gaps across traditional research data management tools and media-rendering environments, at TIB's Open Science Lab we have developed a suite of tools as part of a larger national effort which

[1] Chris Hurley, "Parallel Provenance," 2005, https://www.descriptionguy.com/images/WEBSITE/parallel-provenance.pdf

[2] Pip Laurenson, "Old Media, New Media? Significant difference and the conservation of software based art," In *New Collecting: Exhibiting and Audiences after New Media Art*, ed. by Beryl Graham (Farnham, Surrey, England; Burlington, Vermont: Ashgate, 2014) 73–96.

[3] Lozana Rossenova, Karin de Wild, and Dragan Espenschied, "Provenance for Internet Art: Using the W3C PROV data model," In *Proceedings of 16th International Conference on Digital Preservation iPRES 2019,* Sept 16–20, 2019 Amsterdam, The Netherlands (2019).

[4] Øyvind Eide, Zoe Schubert, Enes Türkoğlu, Jan G. Wieners, and Kai Niebes, "The intangibility of tangible objects: re-telling artefact stories through spatial multimedia annotations and 3D objects," in *ICOM Kyoto 2019, 25th ICOM General Conference: Museums as Cultural Hubs: The Future of Tradition, Kyoto*. (2019). DOI: 10.5281/zenodo.3878966.

[5] Hurley, "Parallel Provenance."

[6] Gregory Wiedeman, "The Historical Hazards of Finding Aids." *University Libraries Faculty Scholarship* 124 (2019). https://scholarsarchive.library.albany.edu/ulib_fac_scholar/124

[7] Ina Blümel, and Raoul Wessel, "DDB goes 3D," *Zenodo*, July 2, 2019. https://doi.org/10.5281/zenodo.5579159.

involves the partnership between research, library and cultural institutions, namely the NFDI4Culture consortium.[8]

In this workshop, researchers, digital curators and data managers will learn how to make datasets including 3D models and other media files available as linked open data within Semantic Kompakkt, the integrated FOSS (Free and Open Source Software) toolchain developed at TIB.[9] The toolchain consists of three main components (Fig. 1): 1) OpenRefine – for data reconciliation and batch upload;[10] 2) Wikibase – for linked open data storage;[11] and 3) Kompakkt – for rendering and annotating 3D models, and other 2D and AV media files.[12] All components of the toolchain feature graphical user interfaces aiming to lower the barrier of participation in the semantic web for a wide range of practitioners and researchers (Fig. 2). The workshop will feature practical demonstrations of the collaborative environment with different levels of read/write access, wherein researchers can try out the data upload and annotation functionalities for themselves.

---

[8] Altenhöner, Reinhard, Ina Blümel, Franziska Boehm, et al., "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.

[9] Lozana Rossenova, Zoe Schubert, Richard Vock, and Ina Blümel, "Beyond the render silo - Semantically annotating 3D data within an integrated knowledge graph and 3D-rendering toolchain," in *DHd 2022 Kulturen des digitalen Gedächtnisses. 8. Tagung des Verbands "Digital Humanities im deutschsprachigen Raum", Potsdam* (2022). DOI: 10.5281/zenodo.6328155.

[10] Elizabeth Sterner, "Cleaning Collections Data Using OpenRefine," *Issues in Science and Technology Librarianship* 92 (2019). DOI: 10.29173/istl30

[11] Samantha Alípio, Mohammed S. Abdulai, Georgina Burnett, and Dan Shick, "Wikibase: the Software for Open Data projects," *Wikimedia Tech News*, April 14, 2021. https://tech-news.wikimedia.de/en/2021/04/14/wikibase-the-software-for-open-data-projects/

[12] Eide, et al, "The intangibility of tangible objects."

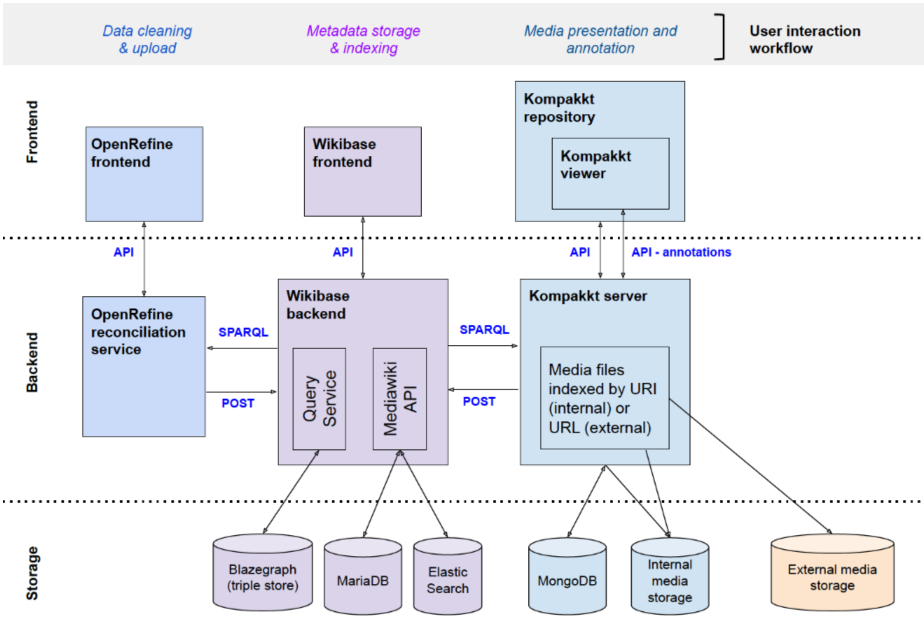


Fig. 1. Diagrammatic representation of the toolchain architecture.

Furthermore, the workshop will highlight how the toolchain follows FAIR principles and adopts a common data model with mappings to standard ontology terms (including CIDOC-CRM, FRBR, CRMdig, and DC terms) resulting in increased interoperability and data reuse across datasets from different research domains. In this way, media objects and annotations, as well as their cultural context (including historical people and places, geo-location and digital-capture-technology metadata), can be linked to the broader semantic web and various national and international authority records (GND, Getty's AAT, VIAF and more). The data model also takes into account the need for clear data provenance across heterogeneous data formats and data sources, and is compliant with the W3C Web Annotation Standard for the annotation of digital media artefacts.

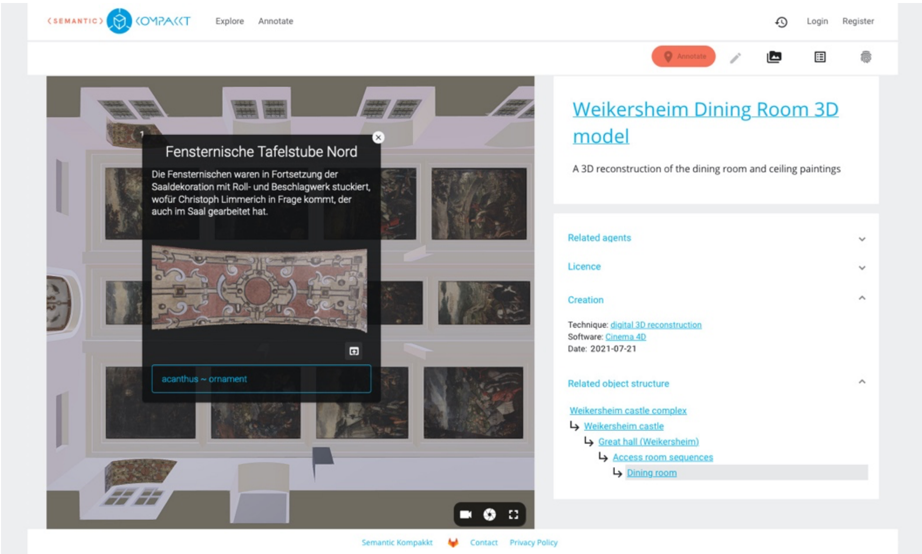


Fig. 2. Image of the Semantic Kompakkt interface showing items from the Corpus der barocken Deckenmalerei in Deutschland (CbDD) project.[13]

The toolchain has so far been developed iteratively thanks to continuous testing with real-world use cases from the NFDI4Culture community.[14] Following this agile approach, after the hands-on demonstrations, the workshop will also include requirements gathering exercises. These will seek to gather feedback and critical perspectives from participants concerning additional features or further areas of development that could benefit their specific research domains. The workshop will be of interest to researchers, digital curators and information science professionals who work with datasets containing 3D media, and want to explore the possibilities of linked open data, open source software and collaborative annotation workflows.

---

[13] "Corpus der barocken Deckenmalerei in Deutschland (CbDD)," Bayerische Akademie der Wissenschaften, accessed January 18 2022, https://deckenmalerei.badw.de/.

[14] Lozana Rossenova, Zoe Schubert, Richard Vock, Lucia Sohmen, Lukas Günther, Paul Duchesne, and Ina Blümel, "Collaborative annotation and semantic enrichment of 3D media: a FOSS toolchain,." in *Proceedings of the 22nd ACM/IEEE Joint Conference on Digital Libraries (JCDL '22).* Association for Computing Machinery, New York, NY, USA, Article 40, 1–5. (2022). DOI: 10.1145/3529372.3533289.

Bibliography

Alípio, Samantha, Mohammed S. Abdulai, Georgina Burnett, and Dan Shick. “Wikibase: the Software for Open Data projects,” *Wikimedia Tech News,* April 14, 2021. https://tech-news.wikimedia.de/en/2021/04/14/wikibase-the-software-for-open-data-projects/.

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.

Bayerische Akademie der Wissenschaften. “Corpus der barocken Deckenmalerei in Deutschland (CbDD).” Accessed January 18, 2022. https://deckenmalerei.badw.de/.

Blümel, Ina, and Raoul Wessel. “DDB goes 3D”. Zenodo, July 2, 2019. https://doi.org/10.5281/zenodo.5579159.

Eide, Øyvind, Zoe Schubert, Enes Türkoğlu, Jan G. Wieners, and Kai Niebes. “The intangibility of tangible objects: re-telling artefact stories through spatial multimedia annotations and 3D objects.” In *ICOM Kyoto 2019, 25th ICOM General Conference: Museums as Cultural Hubs: The Future of Tradition, Kyoto*. (2019). DOI: 10.5281/zenodo.3878966.

Hurley, Chris. “Parallel Provenance.” (2005). https://www.descriptionguy.com/images/WEBSITE/parallel-provenance.pdf.

Laurenson, Pip. “Old Media, New Media? Significant difference and the conservation of software based art.” In *New Collecting: Exhibiting and Audiences after New Media Art*, edited by Beryl Graham, 73–96. Farnham, Surrey, England; Burlington, Vermont: Ashgate, 2014.

Rossenova, Lozana, Karin de Wild, and Dragan Espenschied. "Provenance for Internet Art: Using the W3C PROV data model." In *Proceedings of 16th International Conference on Digital Preservation iPRES 2019,* Sept 16–20, 2019 Amsterdam, The Netherlands (2019).

Rossenova, Lozana, Zoe Schubert, Richard Vock, and Ina Blümel. "Beyond the render silo - Semantically annotating 3D data within an integrated knowledge graph and 3D-rendering toolchain." In *DHd 2022 Kulturen des digitalen Gedächtnisses. 8. Tagung des Verbands "Digital Humanities im deutschsprachigen Raum",* Potsdam (2022). DOI: 10.5281/zenodo.6328155.

Rossenova, Lozana, Zoe Schubert, Richard Vock, Lucia Sohmen, Lukas Günther, Paul Duchesne, and Ina Blümel. "Collaborative annotation and semantic enrichment of 3D media: a FOSS toolchain." In *Proceedings of the 22nd ACM/IEEE Joint Conference on Digital Libraries (JCDL '22).* Association for Computing Machinery, New York, NY, USA, Article 40, 1–5. (2022). DOI: 10.1145/3529372.3533289

Sterner, Elizabeth. "Cleaning Collections Data Using OpenRefine." *Issues in Science and Technology Librarianship* 92 (2019). DOI: 10.29173/istl30.

Wiedeman, Gregory. "The Historical Hazards of Finding Aids." *University Libraries Faculty Scholarship* 124 (2019). https://scholarsarchive.library.albany.edu/ulib_fac_scholar/124

KI statt Paläographie: Automatische Transkription von Handschriften und Drucken
Einführung in Transkribus und eScriptorium

Will, Larissa
larissa.will[at]uni-mannheim.de
Universitätsbibliothek Mannheim, Deutschland
ORCID-iD: 0009-0004-6220-8939

Huff, Dorothee
dorothee.huff[at]uni-tuebingen.de
Universitätsbibliothek Tübingen, Deutschland
ORCID-iD: 0000-0003-0866-9967

Zusammenfassung: Texterkennungs- und Transkriptionsplattformen wie Transkribus und eScriptorium können bei der Erschließung historischer Quellen unterstützen und bieten gegenüber kommandozeilenbasierter Texterkennungssoftware den Vorteil einer einfachen Anwendung über eine graphische Oberfläche. Dabei spielt es keine Rolle, ob es sich um eine Postkarte, eine historische Zeitung oder eine mittelalterliche Handschrift handelt. Von der Layouterkennung und -korrektur über die automatisierte Texterkennung bis hin zur manuellen Korrektur der OCR-Ergebnisse und dem Training eigener Modelle werden die grundlegenden Funktionen von Transkribus und eScriptorium vorgestellt und können von den Teilnehmenden selbst an Beispielen getestet werden.

Abstract

Egal, ob es sich um Editionsvorhaben, ein Seminar zur Handschriftenkunde oder Citizen Science-Projekte handelt: Schriftstücke aus vergangenen Zeiten lassen sich ohne paläographische Kenntnisse oft nur mit hohem Aufwand entziffern und maschinell weiterverarbeiten. Texterkennungs- und Transkriptionsplattformen wie Transkribus und eScriptorium können hier bei der Erschließung von historischen Quellen unterstützen bzw. diese bei großen Korpora überhaupt erst ermöglichen. Sie bieten gegenüber kommandozeilenbasierter Texterkennungssoftware den Vorteil einer einfachen Anwendung über eine graphische Oberfläche. Dabei spielt es keine Rolle, ob es sich um eine Postkarte, eine historische Zeitung oder eine mittelalterliche Handschrift handelt.

Transkribus ist eine Plattform zur Transkription handschriftlicher und gedruckter Dokumente, die verschiedene Tools wie die automatische

Layout- und Texterkennung auf Grundlage von neuronalen Netzen vereint und daneben unter anderem auch das Tagging von Entitäten und Strukturelementen sowie das Training von eigenen Modellen ermöglicht. Die erzeugten Daten können in vielen unterschiedlichen Formaten für die Weiterverarbeitung exportiert werden. Transkribus war 2016 bis 2019 Teil des EU geförderten Projekts READ (Recognition and Enrichment of Archival Documents) und wird seitdem von der READ-COOP SCE weiterbetrieben.[1] Aktuell kann das Programm in einer Desktop-Version (Expert Client) und Browser-Version (Transkribus Lite) nach einer einfachen Registrierung ohne Lizenz genutzt werden. Abgesehen von der automatischen Texterkennung sind alle Funktionen gebührenfrei nutzbar. Aufgrund der großen Usercommunity stehen viele öffentlich zugängliche Texterkennungsmodelle für Drucke und Handschriften in unterschiedlichen Sprachen und Schriftarten zur Verfügung, die oftmals schon bei geringem Eigenaufwand mit gutem Ergebnis auf eigene Texte angewandt werden können. Während eine automatische Transkription auf Knopfdruck sehr bequem ist, so muss dabei jedoch beachtet werden, dass der Output auf diese Weise unter Umständen Richtlinien entspricht, die sich nicht mit den eigenen decken, uneinheitlich und/oder nicht nachvollziehbar sind. Sollen diese Probleme mit einem eigenen Modelltraining umgangen werden, kann jedoch auch ein nichtstandardisierter Output als Ausgangspunkt für selbst aufbereitete Ground-Truth-Daten dienen und deren Erzeugung erleichtern.

eScriptorium ist eine freie Open-Source-Transkriptionsplattform, die als Teil des scripta-Projektes[2] an der Université Paris Sciences et Lettres entwickelt wird.[3] Für die Texterkennung und die Segmentierung verwendet eScriptorium die OCR-Engine Kraken.[4] Die Quelloffenheit von eScriptorium ermöglicht es der Anwenderschaft einen Beitrag zur Entwicklung zu leisten oder an eigene Anforderungen zu adaptieren. eScriptorium lässt sich auf dem eigenen PC oder Server betreiben. Handschriftliche und gedruckte Texte können automatisiert oder manuell segmentiert und transkribiert werden. Da die Daten hierfür nicht auf Fremd-Server hochgeladen werden müssen, ist es damit auch möglich, Digitalisate mit restriktiven Auflagen für die Weitergabe zu bearbeiten.

---

[1] READ-COOP, 2023.
[2] PSL, 2023.
[3] Scripta/escriptorium, 2023.
[4] Kiessling, 2015 und Kiessling, 2021, S. 87. (Bei Kraken handelt es sich um einen Quelltext-Fork der OCR-Lösung OCRopus, siehe dazu: OCRopus, 2023).

Ein Vorteil von eScriptorium ist, dass nicht nur die Daten, die für das Modelltraining genutzt wurden gemäß der FAIR-Prinzipien bereitgestellt werden können, sondern auch die Segmentierungs- und Texterkennungsmodelle selbst. Dadurch wird die Interoperabilität und Wiederverwendbarkeit der Modelle gefördert. Die Auffindbarkeit wird durch die Bereitstellung von mit einer DOI versehenen Modellen z. B. auf Zenodo gewährleistet. Außerdem können die Modelle außerhalb des Systems für Massen-OCR mit Kraken verwendet werden. Dieser offene Umgang mit den erarbeiteten Daten fördert den Wissenstransfer und ermöglicht es auch kleineren lokalhistorischen Projekten, Studierenden und Forschungsprojekten mit wenig finanziellen Mitteln eine kooperative und komfortable Texterkennungs- und Transkriptionsumgebung zu nutzen. Die Möglichkeit, Modelle auszutauschen und zu teilen, fördert die Zusammenarbeit in der Community und ermöglicht es den Benutzern, auf bereits existierende Ressourcen zurückzugreifen. Dies trägt dazu bei, die Effizienz und den Fortschritt bei der Transkription von historischen Dokumenten zu steigern.
Die Universitätsbibliothek Mannheim betreibt seit Oktober 2021 eine eigene Instanz von eScriptorium[5] und hat mittlerweile schon einige Projekte u. a. in Lehrveranstaltungen damit umgesetzt sowie einen Transcribathon.[6] Mittlerweile hat die eScriptorium-Community ein breites Portfolio an Texterkennungsmodellen trainiert.[7] Die UB Mannheim hat dazu einen wichtigen Teil beigetragen, dazu gehören diverse Frakturmodelle, ein generisches Handschriftenmodell sowie ein Modell zur Erkennung von Behördenschriftgut, welches mit Schreibmaschine erstellt wurde.[8]

Für die Teilnahme am Workshop sollten alle Beteiligten Zugriff auf einen eigenen PC oder Laptop mit Internetanschluss haben. Anhand eigener oder von uns bereitgestellter Dokumente kann der gesamte OCR-Workflow durchlaufen werden. Von der Layouterkennung und -korrektur über die automatisierte Texterkennung bis hin zur manuellen Korrektur der OCR-Ergebnisse und dem Training eigener Modelle bietet der Workshop einen spannenden Einstieg in die Welt der modernen Transkriptionsarbeit. Die Teilnehmenden können sich beim Workshop einen Einblick in die Vor- und Nachteile von Transkribus und eScriptorium verschaffen und deren Nutzen für ihre Forschungsarbeit prüfen.

[5] Universitätsbibliothek Mannheim/eScriptorium, 2023.
[6] z. B. Kamlah et. al, 2022.
[7] OCR/HTR model repository, 2023.
[8] Weil, 2023.

Bibliografie

„Home - READ-COOP". READ-COOP. Abgerufen am 27.04.2023. https://readcoop.eu/.

Kamlah, Jan, Thomas Schmidt und Renat Shigapov. 2022. "Extracting Research Data from Historical Documents with EScriptorium and Python.", präsentiert bei Focused Tutorial on Capturing, Enriching, Disseminating Research Data Objects, Use Cases from Text+, NFDI4Culture and BERD@NFDI, Mannheim, Deutschland. https://doi.org/10.5281/zenodo.7373134.

Kiessling, B. (2021). Advances in Optical Character Recognition for Historical Arabic. Université Paris Sciences et Lettres.

Kiessling, Benjamin. „Kraken Home". Kraken, 2015. http://kraken.re/.

„OCR/HTR model repository | Zenodo". Zenodo - Research. Shared., 21.03.2023. https://zenodo.org/communities/ocr_mod-els?page=1&size=20.

„OCRopus". GitHub. Abgerufen am 27.04.2023. https://github.com/ocropus.

„Scripta | PSL". Université PSL (Paris Sciences & Lettres) | PSL. Abgerufen am 27.04.2023. https://www.psl.eu/en/scripta.

„Scripta / escriptorium - GitLab". GitLab, 27.04.2023. https://gitlab.com/scripta/escriptorium/.

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. „Index of /~stweil/tesstrain/kraken". 11.05.2023. https://ub-backup.bib.uni-mannheim.de/~stweil/tesstrain/kraken/.

Vorträge

Inhaltserschließung für Forschungsdaten

TextGrid Repository, Normdaten und Basisklassifikation

Calvo Tello, José
calvotello[at]sub.uni-goettingen.de
Niedersächsische Staats- und Universitätsbibliothek Göttingen, Deutschland
ORCID-iD: 0000-0002-1129-5604

Funk, Stefan
funk[at]sub.uni-goettingen.de
Niedersächsische Staats- und Universitätsbibliothek Göttingen, Deutschland
ORCID-iD: 0000-0003-1259-2288

Kurzawe, Daniel
kurzawe[at]sub.uni-goettingen.de
Niedersächsische Staats- und Universitätsbibliothek Göttingen, Deutschland
ORCID-iD: 0000-0001-5027-7313

Ventjeer, Ubbo
veentjer[at]sub.uni-goettingen.de
Niedersächsische Staats- und Universitätsbibliothek Göttingen, Deutschland
ORCID-iD: 0000-0002-9726-3135

Zusammenfassung. Bibliotheken verwenden Klassifikationssysteme für die Sach- und Inhaltserschließung von Primär- und Sekundärliteratur. Für andere Arten von Publikationen, wie z.B. Forschungsdaten, werden sie jedoch nicht eingesetzt. In diesem Beitrag wird die Einführung eines solchen Klassifikationssystems in das TextGrid Repository vorgestellt, das für textuelle Daten (Editionen und Korpora) in XML-TEI geeignet ist. Konkret wird die Basisklassifikation verwendet, ein offenes und im deutschsprachigen Raum weit verbreitetes Klassifikationssystem mit einer mittleren Anzahl von Klassen. In diesem Beitrag möchten wir dafür plädieren, dass Bibliotheken eine aktivere Rolle bei der Beschreibung von Forschungsdaten übernehmen sollten, weil dies zu einer Verbesserung des FAIR-Status der Forschungsdaten führt.

In den letzten Jahren haben sich die FAIR-Kriterien zu einer der wichtigsten Richtlinien für Forschungsdaten entwickelt (Wilkinson u. a. 2016; Kraft 2017; Burger, Cordts, und Habermann 2021; Helling, Jung, und Pielström 2022). In zwei dieser Kriterien (Findability und Reusability) spielt die Beschreibung durch Metadaten eine wesentliche Rolle.

Bibliotheken, die häufig Forschungsdatenrepositorien betreiben (Bertelmann und Pfeiffenberger 2014), haben eine lange Tradition in der Erstellung und Anwendung von Normdaten und Klassifikationssystemen. Bislang werden solche Systeme jedoch für die Erschließung von Primärliteratur (z.B. literarische Editionen) oder Sekundärliteratur (Monographien, Sammelbände, Zeitschriften etc.) sowohl in digitaler als auch in gedruckter Form verwendet (Gantert 2016). Neben Primär- und Sekundärliteratur sind Forschungsdaten als eine weitere Variante von Forschungspublikationen zu verstehen (Bertelmann und Pfeiffenberger 2014). Es wäre von Vorteil, wenn die in Bibliothekskatalogen verwendeten Mittel zur Erschließung von Primär- und Sekundärliteratur auch in Forschungsdatenrepositorien zum Einsatz kämen.

Das TextGrid Repository (TGR; Neuroth 2015; Funk 2018) ist spezialisiert auf text- und sprachbasierte Daten (Kett u. a. 2022) und ein Beitrag zum NFDI-Konsortiums Text+. Als eines der ersten fachspezifischen Repositorien der Digital Humanities im deutschsprachigen Raum ist das TGR mit der "digitalen Bibliothek" eine der wichtigsten Quellen für digitale literarische Texte in deutscher Sprache in XML-TEI (Betz 2015). Seit dem Start von Text+ wurde das Repositorium in verschiedene Richtungen weiterentwickelt, u.a. um die Veröffentlichung bereits vorhandener Korpora zu erleichtern. In diesem Zusammenhang wurden neue Korpora wie CoNSSA (Calvo Tello 2021) oder die ELTeC-Korpora der COST Action Distant Reading (Schöch u. a. 2021; Burnard, Schöch, und Odebrecht 2021) veröffentlicht. Damit stehen ca. 1500 neue literarische Texte aus verschiedenen europäischen Sprachen im TGR zur Verfügung.

Normdaten spielen schon im TGR eine wichtige Rolle. Das TG Metadatenschema ermöglicht es, diese Identifikatoren sowohl auf der Ebene des Autors als auch auf der Ebene des Werkes einzugeben, was bisher in der Regel über die GND erfolgte (de la Iglesia, Moretto, und Brodhun 2015). Seit Beginn des neuen Konsortiums sind weitere Optionen in Bezug auf die Normdaten möglich, wie z.B. die Suche nach Texten bestimmter Personen über die GND-ID oder die Eintragung anderer Identifikatoren, wie z.B. aus Wikidata oder VIAF (Calvo Tello, Rißler-Pipka, und Barth 2023).

Wie in vielen anderen Forschungsdatenrepositorien fehlte jedoch auch in TGR die Möglichkeit, Daten für die eigene Teildisziplin abzufragen: Welche Daten sind für die Germanistik, für die Romanistik oder für die Slawistik in diesem Repositorium interessant? Viele Forschungsdatenrepositorien verwenden Ad-hoc-Hierarchien mit einigen Dutzend Klassen, in denen sich die einzelnen Philologien in der Regel nicht wiederfinden. Andere Plattformen verwenden Taxonomien, die bisher nicht für Publikationen verwendet wurden (wie die DFG-Fachsystematik, Kindling 2023) oder das Dewey Decimal Classification System (Kennel 2019), ein proprietäres Klassifikationssystem.

Um diese Abfrage zu ermöglichen und ein offenes und bereits etabliertes Klassifikationssystem zu verwenden, wurde in TGR die Basisklassifikation (BK) eingeführt (Schulz 1991; Zimmermann 1994). Dieses Klassifikationssystem ist eines der am weitesten verbreiteten Klassifikationssysteme im deutschsprachigen Raum und umfasst ca. 2000 Klassen. Diese Anzahl an Klassen ist deutlich höher als in den üblichen einfachen Ad-hoc-Hierarchien von Forschungsdatenrepositorien, wird aber im Bibliothekswesen als "grob" oder "klein" bezeichnet, da die Anzahl der Klassen deutlich geringer ist als in anderen gängigen Klassifikationssystemen. Allerdings können Kombinationen von BK-Klassen den gleichen Informationsgehalt wie die Klassen der großen Klassifikationssysteme haben.

Da TGR bisher nur für digitale Editionen und literarische Korpora vorgesehen ist, kann es als Brücke zwischen Forschungsdatenrepositorien und klassischen bibliothekarischen Katalogen gesehen werden. Deswegen sind prinzipiell nur die Klassen der Hauptklassen 17 und 18 von Bedeutung. Dies bedeutet, dass die Forschungsdaten (bzw. digitale Editionen) im TGR mit den gleichen Klassen beschrieben werden, wie die gedruckten oder E-Book-Ausgaben im Bibliothekskatalog, was zugleich auch einen Schritt in Richtung Interoperabilität der Ressourcen ermöglicht. Die Hierarchie der BK-Klassen ermöglicht die Suche nach bestimmten Sprachen oder Sprachgruppen. Beispielsweise kann ein Benutzender nun nach Texten suchen, die für die Romanistik oder die Slawistik relevant sind. Es bleibt daher offen, inwieweit die gleiche Strategie für andere Repositorien von anderen Forschungsdaten sowohl in den Geisteswissenschaften als auch in den Naturwissenschaften verwendet werden kann. Für andere Forschungsdatenrepositorien könnte es sinnvoller sein, die Ad-hoc-Taxonomien durch die BK-Hauptklassen zu ersetzen, um die Homogenität und Wiederverwendbarkeit der Ressourcen zu verbessern.

Wenn ein Forschender seine Daten in einem Repositorium oder einer ähnlichen Plattform ablegt, muss die Person die Metadaten dazu beitragen. Mit anderen Worten: Die Beschreibung der Forschungsdaten ist derzeit eine Aufgabe der Forschenden und nicht der Betreibenden des Repositoriums. Wie gut und eindeutig diese Beschreibung ist, liegt in den Händen der Forschenden. Ist es aber realistisch zu erwarten, dass Personen, die in der Regel wenig oder nichts mit Normdaten oder Klassifikationssystemen zu tun haben, diese gut erfassen können? Der Vergleich mit anderen klassischen Forschungspublikationen zeigt einen deutlichen Unterschied: Bibliotheken spielen eine viel aktivere Rolle bei der Beschreibung bzw. Katalogisierung und Erschließung von Primär- und Sekundärliteratur. Ihre Rolle bei der Beschreibung von Forschungsdaten ist dagegen grundsätzlich passiv (Bertelmann und Pfeiffenberger 2014).

Dieser Beitrag zeigt ein Beispiel für die Integration von Normdaten und Klassifikationssystemen in Forschungsdatenrepositorien. Wir plädieren für die Einführung und Nutzung der BK in weiteren Repositorien und Quellenverzeichnissen (bzw. Registries), einschließlich der neuen NFDI-Bestrebungen. Diese Einführung muss jedoch mit einer Diskussion über die Rollen der Forschenden einerseits und der Repositorienbetreiber andererseits sowie über mögliche Änderungen in den Prozessen einhergehen.

Bibliografie

Bertelmann, Roland, und Hans Pfeiffenberger. 2014. „Forschungsdaten und Bibliotheken“. In *Praxishandbuch Bibliotheksmanagement*, herausgegeben von Rolf Griebel, Hildegard Schäffler, und Konstanze Söllner, 639–51. Reference. Berlin ; Boston: de Gruyter. https://discovery.sub.uni-goettingen.de/id{colon}729019497.

Betz, Katrin. 2015. „Ein virtuelles Bücherregal: Die Digitale Bibliothek im TextGrid Repository“. In *TextGrid: Von der Community - für die Community*, herausgegeben von Heike Neuroth, Andrea Rapp, und Sibylle Söring, 229–39. Glückstadt: Verlag Werner Hülsbusch. https://discovery.sub.uni-goettingen.de/id{colon}1748350463.

Burger, Marleen, Anette Cordts, und Ted Habermann. 2021. „Bausteine Forschungsdatenmanagement : 2021, Wie FAIR sind unsere Metadaten?“ Application/pdf. *BausteineForschungsdatenmanagement. Empfehlungen und Erfahrungsberichte für die Praxis vonForschungsdatenmanagerinnen und -managern* 3 (Oktober). https://doi.org/10.17192/BFDM.2021.3.8351.

Burnard, Lou, Christof Schöch, und Carolin Odebrecht. 2021. „In search of comity: TEI for distant reading“. *Journal of the Text Encoding Initiative*, Nr. Issue 14 (März). https://doi.org/10.4000/jtei.3500.

Calvo Tello, José. 2021. *The Novel in the Spanish Silver Age: A Digital Analysis of Genre Using Machine Learning*. Digital Humanities Research 4. Bielefeld: transcript. https://www.transcript-verlag.de/978-3-8376-5925-2/the-novel-in-the-spanish-silver-age/?c=331025282.

Calvo Tello, José, Nanette Rißler-Pipka, und Florian Barth. 2023. „GND und Normdaten für europäische Literatur? Personen und Werke in den multilingualen Korpora von ELTeC". In *Open Humanities, Open Culture*, 160–65. Luxemburg, Trier: Digital Humanities im deutschsprachigen Raum e.V. https://doi.org/10.5281/zenodo.7688631.

Funk, Stefan E. 2018. „Elektronisches Publizieren von Digitalen Forschungsdaten am Beispiel des TextGrid Repositorys – Umsetzung von Digitalen Publikationsworkflows für die eHumanities". Köln. http://dx.doi.org/10.20375/0000-000B-D269-2.

Gantert, Klaus. 2016. *Bibliothekarisches Grundwissen*. *Bibliothekarisches Grundwissen*. Berlin, Boston: De Gruyter Saur. https://www.degruyter.com/view/title/302969.

Helling, Patrick, Kerstin Jung, und Steffen Pielström. 2022. „Making Research Data FAIR. Seriously? - Reflections on Research Data Management in the Computational Literary Studies". In *Responding to Asian Diversity*, 230–33. Tokyo: ADHO. https://dh2022.dhii.asia/abstracts/247.

Iglesia, Martin de la, Nicolas Moretto, und Maximilian Brodhun. 2015. „Metadaten, LOD und der Mehrwert standardisierter und vernetzter Daten". In *TextGrid: Von der Community - für die Community*, herausgegeben von Heike Neuroth, Andrea Rapp, und Sibylle Söring, 91–102. Glückstadt: Verlag Werner Hülsbusch. https://discovery.sub.uni-goettingen.de/id{colon}1748350463.

Kennel, Patrik. 2019. „Sacherschließung von Forschungsdaten". https://doi.org/10.25651/1.2019.0017.

Kett, Jürgen, Christoph Kudella, Andrea Rapp, Regine Stein, und Thorsten Trippel. 2022. „Text+ Und Die GND – Community-Hub Und Wissensgraph". *Zeitschrift Für Bibliothekswesen Und Bibliographie* 69 (1–2): 37–47. https://doi.org/10.3196/1864295020691262.

Kindling, Maxi. 2023. „Qualitätssicherung von Datenpublikationen bei Data Journals und Forschungsdatenrepositorien“. DoctoralThesis, Humboldt-Universität zu Berlin. https://doi.org/10.18452/26023.

Kraft, Angelina. 2017. „The FAIR Data Principles for Research Data“. *TIB-Blog* (blog). 12. September 2017. https://blogs.tib.eu/wp/tib/2017/09/12/the-fair-data-principles-for-research-data/.

Neuroth, Heike, Hrsg. 2015. *TextGrid: Von der Community - für die Community: eine virtuelle Forschungsumgebung für die Geisteswissenschaften*. Glückstadt: Hülsbusch.

Schöch, Christof, Tomaz Erjavec, Roxana Patras, und Diana Santos. 2021. „Creating the European Literary Text Collection (ELTeC): Challenges and Perspectives“. *Modern Languages Open*, Nr. 1 (Dezember): 25. https://doi.org/10.3828/mlo.v0i0.364.

Schulz, Ursula. 1991. „Die niederländische Basisklassifikation : eine Alternative für die ‚Sachgruppen‘ im Fremddatenangebot der Deutschen Bibliothek“. *Bibliotheksdienst* 25: 1196–1219.

Wilkinson, Mark D., Michel Dumontier, IJsbrand Jan Aalbersberg, Gabrielle Appleton, Myles Axton, Arie Baak, Niklas Blomberg, u. a. 2016. „The FAIR Guiding Principles for scientific data management and stewardship“. *Scientific Data* 3 (März). https://doi.org/10.1038/sdata.2016.18.

Zimmermann, Harald H. 1994. „Zur Struktur und Nutzung von Klassifikationen im Bibliothekswesen (am Beispiel der Klassifikation der Deutschen Bibliothek und der sog. Niederländischen Basisklassifikation)“. In *Mehrwert von Information - Professionalisierung der Informationsarbeit: Proceedings des 4. Internationalen Symposiums für Informationswissenschaft, ISI 1994, Graz, Austria, November 2-4, 1994*, herausgegeben von Wolf D. Rauch, Franz Strohmeier, Harald Hiller, und Christian Schlögl, 16:187–200. Schriften zur Informationswissenschaft. Hochschulverband für Informationswissenschaft.

Total Error Sheets for Datasets (TES-D) zur Dokumentation Digitaler Verhaltensdaten

Eine Kritische Reflektion des Datensammlungsprozesses

Fröhling, Leon
leon.froehling[at]gesis.org
GESIS, Deutschland
ORCID-iD: 0000-0002-5339-7019

Sen, Indira
indira.sen[at]gesis.org
GESIS, Deutschland
ORCID-iD: 0000-0003-3475-0371

Soldner, Felix
felix.soldner[at]gesis.org
GESIS, Deutschland
ORCID-iD: 0000-0001-5324-3581

Steinbrinker, Leonie
leonie.steinbrinker[at]uni-leipzig.de
Universität Leipzig, Deutschland
ORCID-iD: 0009-0003-7078-4111

Zens, Maria
maria.zens[at]gesis.org
GESIS, Deutschland
ORCID-iD: 0000-0001-7461-8231

Weller Katrin
katrin.weller[at]gesis.org
GESIS / CAIS, Deutschland
ORCID-iD: 0000-0003-3799-1146

Zusammenfassung. Die in den Sozialwissenschaften zunehmend zur Untersuchung bekannter Verhaltensmuster und neuartiger Kommunikationsphänomene verwendeten digitalen Verhaltensdaten sind häufig das Resultat eines mehrstufigen Prozesses der Datensammlung. Die bereits zur Sammlung der Daten zu treffenden Designentscheidungen können sich auf häufig unerwartete Weise in der Zusammensetzung und Qualität des resultierenden Datensatzes niederschlagen. Zur besseren Erkennung, Dokumentation und Kommunikation systematischer

Verzerrungen, Eigenheiten und potenzieller Fehler in Datensätzen digitaler Verhaltensdaten präsentieren wir die *Total Error Sheets for Datasets (TES-D)*. Das TES-D leitet Forschende durch die kritische Reflektion des Prozesses der Datensammlung und unterstützt bei der Erstellung einer umfänglichen Dokumentation des resultierenden Datensatzes.

Zur Untersuchung sozialer Phänomene werden zunehmend auch neuartige Arten digitaler Verhaltensdaten genutzt, insbesondere solche, die aus verschiedenen Online-Plattformen gesammelt werden können. Während Plattformen teilweise ermöglichen, Daten mit Hilfe spezieller Schnittstellen (APIs) abzurufen, so sind diese in den wenigsten Fällen auf die Bedürfnisse wissenschaftlicher Forschung zugeschnitten. Um auf Basis dieser Daten also valide Aussagen unter Einhaltung wissenschaftlicher und ethischer Standards treffen zu können muss sichergestellt sein, dass die Charakteristika, Limitationen und potenziellen systematischen Verzerrungen (Biases) dieser Daten erkannt, erfasst und dokumentiert werden. Während diese Art der Auseinandersetzung mit der Qualität von Forschungsdaten in den traditionelleren Sozialwissenschaften weitestgehend etabliert ist und Forschenden eine Reihe von Hilfsmitteln wie etwa Error Frameworks zur Qualitätsbeurteilung und Metadatenschemata zur Dokumentation zur Verfügung stehen, fehlen diese Werkzeuge für das junge Feld der Computational Social Science zumeist.

Wir präsentieren unsere *Total Error Sheets for Datasets (TES-D)* als ein Hilfsmittel für die standardisierte, ganzheitliche und kritische Dokumentation von Datensätzen im Bereich Computational Social Science, mit einem Fokus auf der Sammlung von Daten aus Online-Plattformen. TES-D baut einerseits in der Konzeption und der Struktur auf den checklistenartigen Ansätzen der Datendokumentation für Datensätze im Bereich Machine Learning auf (Bender and Friedman 2018, Gebru et al. 2021). Andererseits bedient sich TES-D inhaltlich der kritischen, fehlerfokussierten Reflektion des idealisierten Forschungsprozess im Sinne von Error Frameworks. Konkret baut es auf der Auseinandersetzung mit Datenqualität von digitalen Verhaltensdaten gesammelt aus Online-Plattformen auf, wie sie im *TED-On: A Total Error Framework for Digital Traces of Human Behavior on Online Platforms (TED-On)* (Sen et al. 2021) strukturiert wird. TES-D trägt somit auf zwei Arten zu einem verbesserten Umgang mit Forschungsdaten bei. Erstens werden die Forschenden, die an der Erstellung von Datensätzen arbeiten, durch die Auseinandersetzung der Dokumentationsfragen im TES-D dazu angeregt und darin angeleitet, den Prozess der Datensammlung und -analyse kritisch zu reflektieren. Wenn das TES-D Formular begleitend zur Sammlung eines Datensatzes ausgefüllt wird, so hat es das Potenzial unmittelbar das Bewusstsein für den Einfluss von Designentscheidungen (z.B. Auswahl der Plattform, Verzerrung der Repräsentativität durch Auswahl von Suchkriterien) auf die Datensammlung zu erhöhen und so direkt zu besseren Entscheidungen in der Planung und Umsetzung des Forschungsdesigns zu führen. Bei einer nachträglichen Konsultation des TES-D ermöglicht es die Identifizierung und transparente Kommunikation von möglichen Biases des Datensatzes. Zweitens erhöht das TES-D durch eben diese systematische Dokumentation und transparente Kommunikation von Verzerrungen, Eigenheiten und potenziellen Fehlern die Nachnutzbarkeit der Daten durch andere Forschende. Nachnutzenden wird eine informierte(re) Entscheidung über die Eignung des Datensatzes für die eigene Forschungsfrage ermöglicht. Dies ist eine wesentliche Voraussetzung um die Nachnutzung von Datensätzen aus Online-Plattformen gewinnbringend zu ermöglichen.

Unser Vortrag stellt die theoretischen Grundlagen der Total Error Sheets for Datasets sowie den Entwicklungsprozess vor, ordnet den Ansatz in den Kontext etablierter Methoden der Reflektion zu Datenqualität in den Sozialwissenschaften und zur Frage der Reproduzierbarkeit ein, und zeigt schließlich unsere Vision der Einbindung von TES-D in den Forschungsprozess auf. Dazu wird anhand eines Prototyps eine Möglichkeit der technischen Implementierung und Integration von TES-D in den typischen Prozess der automatischen Datensammlung in der Programmierumgebung *Jupyter Notebooks* aufgezeigt.

Bibliografie

Bender, Emily M., and Batya Friedman. "Data statements for natural language processing: Toward mitigating system bias and enabling better science." *Transactions of the Association for Computational Linguistics* 6 (2018): 587-604.

Gebru, Timnit, Jamie Morgenstern, Briana Vecchione, Jennifer Wortman Vaughan, Hanna Wallach, Hal Daumé III, and Kate Crawford. "Datasheets for datasets." *Communications of the ACM* 64, no. 12 (2021): 86-92.

Sen, Indira, Fabian Flöck, Katrin Weller, Bernd Weiß, and Claudia Wagner. "A total error framework for digital traces of human behavior on online platforms." *Public Opinion Quarterly* 85, no. S1 (2021): 399-422.

Ein weiteres Toolverzeichnis für die Digital Humanities?!

Aber diesmal offen und mit Wikidata

Grallert, Till
till.grallert[at]hu-berlin.de
Humboldt-Universität zu Berlin, Deutschland
ORCID-iD: 0000-0002-5739-8094

Eckenstaler, Sophie
sophie.eckenstaler.1[at]ub.hu-berlin.de
Humboldt-Universität zu Berlin, Deutschland

Tirtohusodo, Samatha
samantha.tirtohusodo.1[at]hu-berlin.de
Humboldt-Universität zu Berlin, Deutschland

Schlesinger, Claus-Michael
claus-michael.schlesinger[at]hu-berlin.de
Humboldt-Universität zu Berlin, Deutschland
ORCID-iD: 0000-0001-6718-5773

Zusammenfassung. Der eingereichte Beitrag skizziert die Landschaft der bestehenden Toolverzeichnisse in den Digital Humanities und ihrer Schwächen und stellt vor diesem Hintergrund unseren Vorschlag eines Wikidata-basierten offenen Ansatzes als einen von minimal computing, making und Open Science informierten Beitrag zu den Digital Commons vor.

Toolverzeichnisse sind Legion und ein etabliertes Genre in den Digital Humanities: 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 Forschungsdaten Infrastruktur (NFDI), den Fachinformationsdiensten (FID) oder individuellen Bibliotheken und Instituten. Ihnen allen ist gemeinsam, dass sie ein offensichtliches und reales Bedürfnis der Forschungscommunities nach einem Überblick über computergestützte Werkzeuge mit kuratorischen Ansätzen der Wissensorganisation bedienen. Ihnen ist außerdem gemeinsam, dass sie über keine

[1] https://tapor.ca/
[2] https://marketplace.sshopencloud.eu/

dauerhafte Finanzierung verfügen, dass sie primär auf die Kuratierung durch (unbezahlte) Expert_innen-Gremien 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. Im Ergebnis sind diese Toolverzeichnisse in dem Anspruch eines umfassenden, representativen und je aktuellen Abbildes der verfügbaren Möglichkeiten computationeller Forschung und digitaler Wissenschaft als gescheitert zu verstehen (vgl. Dombrowski 2021). Vielmehr handelt es sich im besten Fall um Momentaufnahmen bzw. die Dokumentation historischer Praktiken. Der Fokus auf eine volatile, kontinuierliche Wartung erfordernde Präesentationsschicht führt dazu, dass auch die Daten als Ergebnis und wissenschaftlicher Mehrwert der Kuratierung nicht dauerhaft, mit permanenten URIs, und in einem stabilen, maschinenlesbaren Format als Basis für verlinkte Normdatensätze zur Verfügung stehen.

Für den DFG-geförderten prototypischen *Scholarly Makerspace* als Lernort zur Förderung der Werkzeugkompetenz (*tool literacy*) in den Geistes- und Kulturwissenschaften an der Universitätsbibliothek der Humboldt-Universität zu Berlin verfolgen wir daher einen von *minimal computing*, *making* und Open Science inspirierten Ansatz. *Minimal computing* erfordert dabei die Frage “was brauchen *wir*?” mit den Antworten auf die Frage “was steht *uns* je zur Verfügung?” in Einklang zu bringen (vgl. Gil und Ortega 2016; Risam und Gil 2022). Für unsere Arbeit in der Beratung benötigen wir eine Infrastruktur, die es uns erlaubt unser sich entwickelnes Wissen über computationelle Werkzeuge für die Wissenschaft nachhaltig festzuhalten und teilen zu können: was gibt es, wofür kann es eingesetzt werden, wer hat es eingesetzt und dabei welche Erfahrungen gemacht, und wo und wie kann ich den Einsatz erlernen? Den Projektzyklen der Wissenschaftsförderung unterliegend, sind unsere Möglichkeiten dabei beschränkt: wir sind darauf angewiesen auf bestehenden Datensätzen aufzusetzen und offene, kostenfreie und etablierte Software und Platformen zu nutzen. Der wissenschaftliche Mehrwert liegt dann darin gut dokumentierte Workflows, Datenmodelle und Beispielimplementationen für den Aufbau von offenen Toolverzeichnissen zu entwickeln und den Forschungscommunities zur Verfügung zu stellen.

Der Beitrag stellt unseren Vorschlag einer gemeinsamen offenen Basisinfrastruktur für Toolverzeichnisse für die Fachcommunities vor. Dabei steht Wikidata[3] als eine verteilte, community-kuratierte Normdatei und offene Softwareplatform im Zentrum unseres Vorschlages, die mehrere Schwächen bestehender Toolverzeichnisse adressiert. Wikidata erlaubt es minimale Datenmodelle iterativ zu entwickeln, Datensätze zu pflegen und diese in Wikiprojekten zu kuratierten Sammlungen zusammenzustellen. Auf der Datenebene erlaubt Wikidata die unmittelbare Nutzung sämtlicher Informationen als Linked Open Data (LOD) 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 viele 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. Die Wikimedia-Software und -Plattform bietet die Möglichkeit, dies auch direkt in Wikidata zu tun. Unser Vorschlag erlaubt 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

[3] https://wikidata.org/
[4] https://viaf.org/
[5] https://vocabs.dariah.eu/tadirah/
[6] https://scholia.toolforge.org/

Projektlaufzeit. Damit ist unser Vorschlag Teil einer Bewegung, Wikidata in der Wissenschaft und GLAM-Institution nicht mehr nur als Anbieter von Inhalten wahrzunehmen (vgl. Zhao 2022; Fischer und Ohlig 2019).

Unserer Kenntnis nach, gibt es mit der Research Software Encyclopaedia[7] nur eine weitere community-kuratierte, den FAIR-Prinzipien (vgl. Barker u. a. 2022) und nachhaltigen Technologie-Stacks verpflichtete und damit unserem Vorschlag verwandte Infrastruktur. Allerdings verfolgte diese keinen LOD-Ansatz, fokussiert ausschließlich auf Forschungssoftware, die auf GitHub oder GitLab gehostet ist, und konzentriert sich nicht auf die (Digital) Humanities (Sochat u. a. 2022).

Bibliografie

Barker, Michelle, Neil P. Chue Hong, Daniel S. Katz, Anna-Lena Lamprecht, Carlos Martinez-Ortiz, Fotis Psomopoulos, Jennifer Harrow, u. a. 2022. „Introducing the FAIR Principles for Research Software“. *Scientific Data* 9 (1): 622. https://doi.org/10.1038/s41597-022-01710-x.

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.

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

[7] https://rseng.github.io/projects/research-software-encyclopedia/

2019.
https://wiki.dnb.de/pages/viewpage.action?pageId=147754828.

Gil, Alex, und Élika Ortega. 2016. „Global Outlooks in Digital Humanities: Multilingual Practices and Minimal Computing". In *Doing digital humanities: practice, training, research*, herausgegeben von Constance Crompton, Richard J Lane, und Ray Siemens, 22–34. Abingdon: Routledge.

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.

Risam, Roopika, und Alex Gil. 2022. „Introduction: The Questions of Minimal Computing". Herausgegeben von Alex Gil und Roopika Risam. *Digital Humanities Quarterly* 16 (2, "Minimal Computing"). http://digitalhumanities.org/dhq/vol/16/2/000646/000646.html.

Sochat, Vanessa, Nicholas May, Ian Cosden, Carlos Martinez-Ortiz, und Sadie Bartholomew. 2022. „The Research Software Encyclopedia: A Community Framework to Define Research Software". *Journal of Open Research Software*, März. https://doi.org/10.5334/jors.359.

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.

Data affairs
Ein Portal zum Forschungsdatenmanagement in der Sozial- und Kulturanthropologie

Heldt, Camilla
camilla.heldt[at]fu-berlin.de
Freie Universität Berlin, Deutschland

Voigt, Anne
anne.voigt[at]fu-berlin.de
Freie Universität Berlin, Deutschland

Röttger-Rössler, Birgitt
birgitt.roettger-roessler[at]fu-berlin.de
Freie Universität Berlin, Deutschland

Grote, Brigitte
brigitte.grote[at]fu-berlin.de
Freie Universität Berlin, Deutschland

Zusammenfassung. Im Vortrag wird ein Webangebot des Sonderforschungsbereichs Affective Societies vorgestellt, das qualitativ arbeitende Wissenschaftler*innen und Lehrende aus der Sozial- und Kulturanthropologie (SKA) beim Forschungsdatenmanagement (FDM) unterstützen soll. Bisher stehen aus diesen Fächergruppen aufgrund der Besonderheit von Forschung und Methodik kaum geteilte Forschungsdaten gemäß den FAIR-Prinzipien zur Verfügung.
Das frei zugängliche Angebot gibt einen Überblick über Themen des FDMs, diskutiert den aktuellen Stand anthropologischer Debatten und lädt zum interaktiven Selbststudium ein. Erfahrungsberichte und Praxisbeispiele aus ethnographischen Forschungsfeldern regen zur (kritischen) Auseinandersetzung mit den Forderungen nach FAIR geteilten Forschungsdaten an. Inhalte und Quellcode sind unter einer freien Lizenz veröffentlicht.

1 Ausgangslage

Anthropologische Forschungsdaten, die *in situ*, d.h. durch teilnehmende Beobachtung in einem Forschungsfeld, erhoben werden, existieren nicht jenseits der persönlichen und affektiven Interaktionen zwischen Forschenden und Forschungsteilnehmer*innen. In der reflexiven Filterung ihrer im Feld getätigten Beobachtungen und Befragungen produzieren Anthropolog*innen erst ihre Forschungsdaten, die somit immer subjektiv und affektiv geprägt sind. Daher kann

innerhalb der ethnographischen Feldforschung auch nicht von der Sammlung bereits existierender Roh- oder Primärdaten gesprochen werden: Vielmehr werden Daten während der Forschung in sozialen Interaktionen erstellt und müssen somit als konstruiert aufgefasst werden. Anthropologische Forschung schließt also „nicht einer Datengewinnung deren Analyse an, sondern stellt in einer theorie-orientierten Analyse den Wert bestimmter Informantenäußerungen als Datum erst her“[1].
Hier zeichnet sich ab, dass eine offene und transparente Einsicht in diese Art von Daten im Sinne der FAIR-Prinzipien mit persönlichen, rechtlichen und ethischen Herausforderungen – sowohl für Ethnograph*innen als auch für Informant*innen – einhergeht. In der ethnographischen Feldforschung werden sowohl qualitative als auch quantitative Daten erhoben, deren Unterscheidung nicht immer trennscharf zu ziehen ist.[2] Während sich quantitative Daten einfacher depersonalisieren lassen, sind qualitative Daten durch ihre semantische Dichte und Mehrdeutigkeit geprägt und ohne präzise Kontextualisierung kaum zu verstehen.[3] Ein Teilen von quantitativen Daten (die während der teilnehmenden Beobachtung auch erhoben werden) ist somit unkomplizierter, während das Teilen von qualitativen Daten herausfordernder ist. Insofern ist stets eine sorgfältige (Vor-) Auswahl und Kuration von zu veröffentlichenden Daten erforderlich, nicht zuletzt auch, da „die Datenmassen jeder […] Feldforschung so umfangreich sind, dass der Feldforscher selbst sie bis an sein Lebensende meist nicht veröffentlichen kann“[4]. So begegnen ethnographisch Forschende der Forderung und Erwartung nach offenen und geteilten Daten berechtigterweise mit Skepsis und Kritik.

2 Das Webangebot Data affairs

Aus diesen Überlegungen ergibt sich der Bedarf nach einem Unterstützungsangebot, welches die fachspezifischen Besonderheiten des FDM in der SKA adressiert. Zwar liegen mit Angeboten wie forschungsdaten.info, dem VerbundFDB und Cessda uvm.[5] bereits Informations- und Lernangebote zum FDM vor, doch bieten diese einen allgemeineren Einstieg in das FDM. Das im Rahmen des SFB Affective

---

1 Hirschhauer 2014, 303
2 Beer 2003, 11
3 Hirschhauer 2014, 303; Kretzer 2013, 153
4 Fischer 2003, 285
5 https://forschungsdaten.info/, https.www.forschungsdaten-bildung.de, https://dmeg.cessda.eu/Data-Management-Expert-Guide

Societies entwickelte Portal Data affairs versteht sich als ein ergänzendes Angebot, welches Maßnahmen und Methoden des Forschungsdatenmanagements, u.a. unter Nachnutzung bereits vorhandener Ressourcen, präsentiert und gleichzeitig auf die Problematiken und Grenzen in den anthropologischen Fächern eingeht, indem der aktuelle Stand anthropologischer Debatten zu den Themen des FDM diskutiert wird. Obwohl viele Fragen noch nicht abschließend beantwortet werden können, soll dieses Webangebot mit praxis-orientierten Hinweisen, Beispielen und Übungen Hilfestellungen beim Umgang mit (eigenen) Forschungsdaten geben.
Das Portal richtet sich an interessierte Nachwuchswissenschaftler*innen, Studierende und Lehrende mit jeweils zielgruppenspezifischen Zugängen.[6] Das frei zugängliche unter einer offenen Lizenz veröffentlichte Angebot[7] gibt einen Überblick über Themen des FDMs (Wissenschaftler*innen und Lehrende) und lädt mit vielfältigen Aufgaben zum interaktiven Selbststudium ein (Studierende). Erfahrungsberichte und Praxisbeispiele aus ethnographischen Forschungsfeldern untermauern die Informationen beispielhaft und regen zur (kritischen) Auseinandersetzung mit den Forderungen der Open Science an. Eine inhaltliche Weiterentwicklung und Pflege der Inhalte über die 3. Laufzeit des SFB wird derzeit eruiert.

3 Umsetzung und Nachnutzbarkeit

Zentral für die technische Umsetzung waren Überlegungen zum Publizieren in offenen Formaten:[8] Dieses beinhaltet u.a. die Publikation von zitierfähigen Inhalten im Open Access sowie die Nachnutzbarkeit der Daten, Versionierung, Verknüpfung der Inhalte mit externen Ressourcen und ggf. Anreicherung durch Normdaten (u. a. GND, GeoNames) sowie die freie Verfügbarkeit des Quellcodes. Die redaktionelle Arbeitsumgebung soll weiterhin die kollaborative Erstellung multimedialer Inhalte und deren Verknüpfungen, die Erstellung und Verwaltung von bibliographischen Daten und Glossareinträgen sowie eine Verschlagwortung der Inhalte unterstützen.
Die technische Implementation des Webangebots erfolgte auf Basis des freien Content Management Systems WordPress[9] unter Verwendung

---

[6] Bei der Umsetzung spezifischer Zugänge zu den Portalinhalten haben wir uns von „forTEXT – Literatur digital erforschen“ (https://fortext.net/) inspirieren lassen.
[7] Geplanter Launch Ende September 2023 unter: https://fdm.sfb-affective-societies.de/
[8] U.a. Kleineberg, Kaden. 2017; Wissenschaftsrat (Hg.) 2022
[9] https://wordpress.com/de/

des an der FU entwickelten OES-Plugins[10]. WordPress bietet bereits eine Redaktionsumgebung, die kollaboratives Arbeiten sowie Vernetzung und Verschlagwortung der Inhalte unterstützt; mit der WordPress-Erweiterung H5P[11] können interaktive Aufgaben und Übungen erstellt werden.
Das OES-Plugin erweitert diese Redaktionsumgebung um Funktionen zum wissenschaftlichen Publizieren im Open Access wie Verwaltung bibliographischer Einträge, Anbindung von Normdateien, Vergabe einer CC-Lizenz und eines persistenten Identifiers, und ermöglicht die Verwendung eines projektspezifischen Datenmodells. Die elementaren Inhaltsbausteine (*micro content*) im Portal entsprechen „Beiträgen" im WordPress-Redaktionssystem, eine projektspezifische Erweiterung erlaubt deren Aggregation zu komplexen Datenstrukturen (Artikeln, Lerneinheiten). Die als Wordpress-Theme realisierte barrierefreie und responsive Webseite garantiert die o.g. flexible und zielgruppenspezifische Anzeige der multimedialen Inhalte, die strukturierte Ausgabe der Metadaten und einen intuitiven Zugriff auf die Inhalte über Suche, Filter und Verschlagwortung. Der OES-Code ist unter einer freien Lizenz (GPLv2) auf Github veröffentlicht,[12] die projektspezifischen Erweiterungen stehen für eine Nachnutzung durch Vorhaben mit ähnlichen Anforderungen an zielgruppenspezifische Ausgabeformate bereit.

Bibliografie

Beer, Bettina ed. Methoden und Techniken der Feldforschung. Ethnologische Paperbacks. Berlin: Reimer. 2003.

Fischer, Hans. Dokumentation. In: Beer, B., ed. Methoden und Techniken der Feldforschung. Ethnologische Paperbacks. Berlin: Reimer. Pp. 265-295. 2003.

Hirschauer, Stefan. Sinn im Archiv? Zum Verhältnis von Nutzen, Kosten und Risiken der Datenarchivierung. In: Georg Vobruba ed. Soziologie Jg. 43 (2014) 3 (2014), Campus Frankfurt / New York.

[10] https://www.open-encyclopedia-system.org/ OES (Open Encyclopedia System) ist eine am CeDiS/Freie Universität Berlin im Rahmen einer DFG-Förderung (LIS, 2016-2020) entwickelte Plattform zur Erstellung, Pflege und Publikation von Online-Referenzwerken. Der langfristige Betrieb der an der FU betriebenen OES-Instanzen und die Pflege des OES-Codes sind durch die FU gesichert.

[11] https://h5p.org/wordpress%20

[12] https://github.com/open-encyclopedia-system

pp. 300-3012. 2014.

Kleineberg, Michael and B. Kaden, Ben. „Open Humanities? Expertenmeinungen über Open Access in den Geisteswissenschaften“. In LIBREAS: Library Ideas, 32. 2017.

Kretzer, Susanne. Arbeitspapier zur Konzeptentwicklung der Anonymisierungs-/Pseudonymisierung in Qualiservice. Mannheim. 2013.

Wissenschaftsrat (Hg. Empfehlungen zur Transformation des wissenschaftlichen Publizierens zu Open Access. 2022. https://doi.org/10.57674/fyrc-vb61

Automatische Texterkennung von Handschriften und historischen Drucken

Qualität und Normierung von Ground-Truth-Daten in der Praxis

Huff, Dorothee
dorothee.huff[at]uni-tuebingen.de
Universitätsbibliothek Tübingen, Deutschland
ORCID-iD: 0000-0003-0866-9967

Will, Larissa
larissa.will[at]uni-mannheim.de
Universitätsbibliothek Mannheim, Deutschland
ORCID-iD: 0009-0004-6220-8939

Stöbener, Kristina
kristina.stoebener[at]uni-tuebingen.de
Universitätsbibliothek Tübingen, Deutschland
ORCID-iD: 0000-0002-3299-974X

Zusammenfassung: Automatische Texterkennung (OCR) übersetzt textliche Bildinhalte in digitale Textformate. Auf diese Weise werden der Zugang zu historischen Drucken und Handschriften erhöht und neue Forschungsfragen an das Material ermöglicht. Vor der wissenschaftlichen Auswertung der Daten gilt es jedoch, sich über Aspekte wie Qualität und Normierung der Ground-Truth-Daten und des erzeugten Outputs bewusst zu werden, diese zu hinterfragen und bei der Nachnutzung der Daten in Betracht zu ziehen. Anhand von Beispielen sollen unterschiedliche Vorgehensweisen bei der Erzeugung von Ground-Truth-Daten sowie Ergebnisse der jeweiligen Modelltrainings vorgestellt und problematisiert werden.

Automatische Texterkennung (OCR) übersetzt textliche Bildinhalte in digitale Textformate. Auf diese Weise werden der Zugang zu historischen Drucken und Handschriften erhöht und neue Forschungsfragen an das Material ermöglicht. Erkannte Texte können durchsucht, kopiert, bearbeitet und für eine Extraktion von Forschungsdaten verwendet werden. Große Mengen an Text können bei geringem Ressourceneinsatz – im Gegensatz zur manuellen Transkription – erzeugt und verarbeitet werden. Auch paläographische Kenntnisse sind auf diesem Weg keine zwingende Voraussetzung für die Arbeit mit primärem Quellenmaterial.

Während die automatische Texterkennung von historischen Dokumenten mit klassischer OCR kaum möglich war, wurden mithilfe von Machine Learning in den letzten Jahren große Fortschritte gemacht.[1] Moderne Texterkennungsmodelle sind neuronale Netze, die anhand von korrigierten Transkriptionen, sogenannten Ground-Truth-Daten, trainiert werden. Während der Begriff „Ground-Truth" auf den Ausschließlichkeitsanspruch nur einer möglichen richtigen Lösung hinzudeuten scheint, können die zugrunde liegenden historischen Schriftzeichen bei der Transkription jedoch oftmals unterschiedlich interpretiert und wiedergegeben sowie sogar je nach Zielsetzung bewusst an individuelle Standards angepasst werden. So kann zwar die Weiterverarbeitung der Daten in spezifischen Kontexten wie z. B. als Editionsgrundlage, für linguistische Untersuchungen oder die Einbringung in Datenbanken erleichtert werden, allerdings entsteht auf diese Weise ein Spannungsverhältnis eines solchermaßen passgenau personalisierten Datensatzes zu einer allgemeinen Nachnutzbarkeit der Daten. Zudem fehlt gerade bei historischen Dokumenten oftmals eine Normierung bezüglich der Wiedergabe von historischen Schriftzeichen mit einem Codepoint, damit diese einheitlich von Maschinen verarbeitet werden können. Eine Lösung zumindest für den Bereich historischer Druckwerke sind die Ground-Truth-Richtlinien der koordinierten Förderinitiative OCR-D, die normierte Handlungsempfehlungen für drei Transkriptionslevel bieten.[2]

Die Erzeugung von Ground-Truth-Daten für das Training von Texterkennungsmodellen erfolgt an den Universitätsbibliotheken Tübingen und Mannheim zeichennah nach Level 2 der OCR-D-Transkriptionsrichtlinien, um eine möglichst breite Nachnutzbarkeit der Daten zu gewährleisten und Transformationsmöglichkeiten offenzuhalten. Da der Standardisierungsprozess noch nicht abgeschlossen ist und nicht alle Fälle abgedeckt werden, ist oftmals zwangsweise doch wieder ein zumindest in Teilen individualisiertes Vorgehen notwendig. In der Praxis ist daher zu überlegen, inwieweit dies die Weiterverarbeitung der Daten beeinflusst, wie bei Veröffentlichung der Daten mit Abweichungen umzugehen ist und wie konvertibel die Daten sind. Ist der diplomatische Ansatz jedoch überhaupt und wenn ja, in welchen Kontexten sinnvoll? So gewährleistet die alternative Herangehensweise einer Normalisierung der Daten in der Regel eine bessere Lesbarkeit und Durchsuchbarkeit. Es soll beispielhaft analysiert werden, wie sich unterschiedliche Transkriptionsrichtlinien auf die Erzeugung standardisierter Daten und die Zeichenfehlerrate der Texterkennungsmodelle auswirken.

---

[1] Hodel 2023, 154–157.
[2] OCR-D, n.d.

Die für die Erstellung der Ground-Truth-Daten getroffenen Entscheidungen haben jedoch nicht nur einen Einfluss auf die Nachnutzung dieser Daten in weiteren Kontexten, wie z. B. bei der Zusammenstellung verschiedener Ground-Truth-Datensätze für Modelltrainings, wo uneinheitliche Transkriptionsrichtlinien zu Problemen führen können, sondern bestimmen im nächsten Schritt auch den Output der auf Grundlage dieser Daten trainierten Texterkennungsmodelle. Unter Umständen können zwar regelhaft verwendete Zeichen sowohl in den Ground-Truth-Daten wie auch in der automatisch erzeugten Transkription im Nachhinein ersetzt und die Daten so für andere Nutzungskontexte angepasst werden, jedoch hängt das Ergebnis der automatischen Transkription grundsätzlich von den bei der Erzeugung des Trainingsmaterials getroffenen Entscheidungen und der Qualität desselben ab. Egal wie gut letztere ist, muss beachtet werden, dass mit den aktuellen technischen Möglichkeiten in der Regel kein hundertprozentig korrektes Ergebnis erzielt werden kann, und überlegt werden, wie damit umzugehen ist.[3]

Beim Einsatz von Texterkennungssoftware auf historische Dokumente gilt es somit, sich verschiedener Fragestellungen bei der Datenerzeugung und -nachnutzung bewusst zu werden. Das 2019 im Rahmen des Projekts *OCR-BW* als Service der Universitätsbibliotheken Tübingen und Mannheim eingerichtete *Kompetenzzentrum „Volltexterkennung von handschriftlichen und gedruckten Werken“* betreut Wissenschaftlerinnen und Wissenschaftler sowie Bibliotheken und Archive in Baden-Württemberg bei der Anwendung von automatischer Texterkennungs- und Transkriptionssoftware.[4] Anhand eigener Textkorpora aus Beständen der UB Tübingen, wie z. B. Expeditionstagebüchern, juristischen Konsilien und mittelalterlichen Handschriften, wie auch bei der Unterstützung von wissenschaftlichen Projekten aus verschiedenen Fachdisziplinen werden die Transkriptionsplattformen *Transkribus*[5] und *eScriptorium*[6] für die Erzeugung von automatischen Volltexten für Handschriften und Drucke systematisch getestet und eingesetzt.[7] Anhand von Beispielen sollen unterschiedliche Vorgehensweisen bei der Erzeugung von Ground-Truth-Daten sowie Ergebnisse der jeweiligen Modelltrainings vorgestellt und problematisiert werden.

---

[3] Neudecker et al. 2021, 153–157.
[4] OCR-BW, n.d.
[5] Transkribus, n.d.
[6] Scripta/eScriptorium, n.d.
[7] Huff and Stöbener 2022.

Bibliografie

Hodel, Tobias. "Konsequenzen der Handschriftenerkennung und des maschinellen Lernens für die Geschichtswissenschaft. Anwendung, Einordnung und Methodenkritik." *Historische Zeitschrift* 316, no. 1 (February 2023): 151–180. https://doi.org/10.1515/hzhz-2023-0006.

Huff, Dorothee, and Kristina Stöbener. "Projekt OCR-BW: Automatische Texterkennung von Handschriften." *O-Bib. Das Offene Bibliotheksjournal / Herausgeber VDB* 9, no. 4 (2022): 1–19. https://doi.org/10.5282/o-bib/5885.

Neudecker, Clemens, Karolina Zaczynska, Konstantin Baierer, Georg Rehm, Mike Gerber and Julián Moreno Schneider. "Methoden und Metriken zur Messung von OCR-Qualität für die Kuratierung von Daten und Metadaten." In *Qualität in der Inhaltserschließung*, edited by Michael Franke-Maier, Anna Kasprzik, Andreas Ledl and Hans Schürmann, 137–166. Berlin, Boston: De Gruyter Saur, 2021. https://doi.org/10.1515/9783110691597-009.

OCR-BW. "OCR-BW. Kompetenzzentrum OCR der Universitätsbibliotheken Mannheim und Tübingen." Accessed July 21, 2023. https://ocr-bw.bib.uni-mannheim.de/.

OCR-D. "Die Ground Truth Richtlinien." Accessed July 21, 2023. https://ocr-d.de/de/gt-guidelines/trans/.

Scripta/eScriptorium. "scriptorium." Accessed July 21, 2023. https://gitlab.com/scripta/escriptorium/.

Transkribus. "READ-COOP – Transkribus." Accessed July 21, 2023. https://readcoop.eu/transkribus/.

Feministische Forschungsdaten FAIR gestalten? Kritische Reflexionen zur Modellierung feministischer Filmgeschichte als Linked Open Data

Jungiger, Pauline
pauline.junginger[at]uni-marburg.de
Philipps-Universität Marburg, Deutschland
ORCID-iD: 0000-0003-3899-1674

Zusammenfassung. Das *Women Film Pioneers Project* (WFPP) ist eine etablierte Online-Ressource zum Frühen Kino, die individuelle Geschichten von Filmpionierinnen erzählt, um die vielfältigen Tätigkeiten von Frauen in der Frühen Filmindustrie sichtbarer zu machen. Mein Projekt zielt darauf ab, strukturierte Metadaten für das WFPP zu generieren und diese als Linked Open Data aufzubereiten. Im Spannungsfeld zwischen praktischer Umsetzung und kritischer Reflexion untersucht mein Projekt dabei, wie die Prinzipien des Forschungsdatenmanagements in der Medienwissenschaft angewendet werden können und welche kritischen Fragen feministische Theorien diesbezüglich ermöglichen. Der Vortrag präsentiert die methodische Gestaltung des Projekts mit einem Schwerpunkt auf deren Reflexion aus feministischer Perspektive.

1 Feministische Filmgeschichtsschreibung FAIR gestalten

1.1 Das *Women Film Pioneers Project* (WFPP)

Das *Women Film Pioneers Project* (WFPP) ist eine etablierte Online-Ressource für die Forschung zu Frauen im Frühen Kino (Gaines, Vatsal und Dall'Asta, o. J.). Der Fokus des WFPP liegt auf dem Erzählen individueller Geschichten von Filmpionierinnen, um die vielfältigen Tätigkeiten von Frauen in der frühen Filmindustrie sichtbarer zu machen. Das Kernstück der Plattform bilden knapp 310 Profile über Filmpionierinnen mit Informationen zu Karrieren und Filmografien sowie den Archiven, die Werke der Pionierinnen und Sekundärmaterial besitzen. Obwohl strukturierten Metadaten eine zentrale Bedeutung bei der Sichtbarmachung, Zugänglichkeit und Nachnutzung von Forschungsdaten und digitalen Publikationen zukommt (Baca 2016;

Flanders und Jannidis 2018), arbeitet das WFPP bisher nur sehr eingeschränkt mit Metadaten. Zudem verfolgt das editorische Team einen Ansatz, der eher die Vielfalt feministischer Filmgeschichte in den Vordergrund hebt, als eine möglichst große Einheitlichkeit der Profile anzustreben (Dang 2023).

1.2 Feministische Filmgeschichtsschreibung als LOD aufbereiten

Um die Sichtbarkeit der feministischen Filmgeschichtsschreibung zu erhöhen, bestehende Lücken in Wissensdatenbanken zu schließen und disparate Datenbestände zu Frauen im Frühen Kino nachhaltig zu verlinken, zielt mein Projekt darauf ab, strukturierte Metadaten für das WFPP zu generieren, diese mit Normdaten anzureichern und perspektivisch auf Wikidata einzuspeisen. Um Metadaten im Sinne der FAIR-Prinzipien (Wilkinson et al. 2016) interoperabel zu gestalten und mit anderen Daten verlinkbar zu machen, bietet sich ihre Bereitstellung als Linked Open Data (LOD) an. In Filmarchiven gibt es bereits seit einigen Jahren Bestrebungen, Filmdaten als LOD aufzubereiten (Heftbeger 2019). Diese Institutionen verfolgen mit ihrer Orientierung an dem Standard EN15907 jedoch ein weitaus strukturierteres Datenmanagement, als dies beim WFPP der Fall ist. Deshalb ist es nur in Teilen möglich, sich an diesen Vorarbeiten aus dem Filmkulturerbebereich zu orientieren.

Da der Großteil der Informationen, die das WFPP zu Filmpionierinnen bereithält, bisher nur in Form von Fließtexten vorhanden ist[1], ist es zunächst erforderlich, relevante Informationen zu extrahieren, um diese strukuturieren, mit Normdaten anreichern und als LOD aufbereiten zu können.

2 Feministisches Forschungsdatenmanagement in Theorie und Praxis

2.1 Kritische Reflexionen zur Modellierung als Linked Open Data

Aus einer feministischen Perspektive ist bei der Arbeit mit Linked Open Data zu fragen, welche Vorannahmen und Machtverhältnisse in

[1] Das WFPP hat 2020 einen Datensatz mit ausgewählten biografischen Daten in tabellarischer Form veröffentlicht.

Technologien des Semantic Web, in Normdaten wie die GND und in nachnutzbare Vokabulare wie der Art & Architecture Thesaurus (AAT) oder VIAF eingeschrieben sind. Auch die Prinzipien des Forschungsdatenmanagements müssen daraufhin befragt werden, inwiefern sie es erlauben, feministische Filmgeschichte adäquat zu erfassen und nachhaltig verfügbar zu machen. Laut Deb Verhoeven verschleiern die FAIR-Prinzipien die epistemologischen und ontologischen Annahmen der Fachcommunity aus der sie stammen (2022). Imeri und Rizzolli stellen zudem fest, dass „ethische Aspekte sowie historische Kontexte der Datenproduktion (…) weitgehend unberücksichtigt“ (2022: 3) bleiben.

Was bedeutet dies folglich für ein feministisches Forschungsdatenmanagement? Inwiefern ist es möglich, der Vielfalt und Komplexität feministischer Filmgeschichte gerecht zu werden, wenn die Aufbereitung als LOD eine Strukturierung und Standardisierung der Daten erfordert? Wie kann der spezifische Kontext der Wissensproduktion des WFPP in strukturierten Daten abgebildet werden?

Im Spannungsfeld zwischen praktischer Umsetzung und kritischer Reflexion untersuche ich in meinem Projekt, wie die Prinzipien des Forschungsdatenmanagements in der Medienwissenschaft angewendet werden können und welche kritischen Fragen mit Hilfe feministischer Ansätze (D’Ignazio und Klein 2020) und kritischer Klassifikationstheorie (Bowker und Star 1999) gestellt werden können.

Der Vortrag thematisiert die methodische Gestaltung des Projekts und reflektiert über die spezifischen Fragen, die sich diesbezüglich aus einer feministischen Perspektive stellen.

Bibliografie

Baca, Murtha. “Introduction.” *Introduction to Metadata*. Ed. Murtha Baca. 3rd ed. Los Angeles: Getty P, 2016. <http://www.getty.edu/publications/intrometadata/introduction/> [11.05.2023]

Bowker, Geoffrey C., and Susan Leigh Star. *Sorting Things Out: Classification and its Consequences.* Cambridge: MIT Press, 1999.

Dang, Sarah-Mai, „Forschung explorieren. Zu den Möglichkeiten digitaler Datenvisualisierungen für die feministische Filmgeschichtsschreibung“, in: Laura Niebling, Felix Raczkowski, Sven Stollfuß (Hrsg.): *Handbuch Digitale Medien und Methoden*, Wiesbaden: Springer VS, 2023.

D'Ignazio, Catherine und Lauren F. Klein. *Data Feminism*. Cambridge: MIT Press, 2020.

Flanders, Julia und Fotis Jannidis (Hrsg.). *The Shape of Data in the Digital Humanities: Modeling Texts and Text-based Resources*. London, 2018: Routledge. https://doi.org/10.4324/9781315552941.

Gaines, Jane, Radha Vatsal und Monica Dall'Asta, "Women Film Pioneers Project", New York: Columbia University Libraries. <https://wfpp.columbia.edu/> [11.05.2023]

Heftberger, Adelheid, "Building Resources Together – Linked Open Data for Film Archives", in: *Journal of Film Preservation* 101 (2019): 65-73 <https://www.proquest.com/docview/2317838940/abstract/B700AB2C594C4879PQ/1> [01.05.2023]

Imeri, Sabine und Michaela Rizzoli, "CARE Principles for Indigenous Data Governance. Eine Leitlinie für ethische Fragen im Umgang mit Forschungsdaten?", in: O-Bib. Das Offene Bibliotheksjournal / Herausgeber VDB, 9(2) (2022): 1–14. https://doi.org/10.5282/o-bib/5815

Verhoeven, Deb, "Scholarship in a Clopen World ." Pop! Public. Open. Participatory, 4 (2022). https://doi.org/10.54590/pop.2022.002.

Wilkinson, Mark D., Michel Dumontier, IJsbrand Jan Aalbersberg, Gabrielle Appleton, Myles Axton, Arie Baak, Niklas Blomberg, et al., "The FAIR Guiding Principles for Scientific Data Management and Stewardship," *Scientific Data* 3 (2016): 160018. https://doi.org/10.1038/sdata.2016.18.

Digital Humanities in Discuss Data

Aufbau eines Communityspace

Kahlert, Torsten
kahlert[at]hab.de
Herzog-August-Bibliothek, Wolfenbüttel

Kurzawe, Daniel
kurzawe[at]sub.uni-goettingen.de
Staats- und Universitätsbibliothek Göttingen
ORCID-iD: 0000-0001-5027-7313

Zusammenfassung. In diesem Beitrag beschreiben wir, wie die Forschungsdatenplattform Discuss Data um einen Bereich (“Communityspace”) für die Digital Humanities (DH) erweitert wird. Dazu ergründen wir die Spezifika dieses Forschungsbereichs für einen entsprechenden Communityspace in Discuss Data und hinterfragen auch kritisch, wie erfolgreich bisherige Ansätze der Plattform Discuss Data im Aufbau der Community des Space für die “Osteuropa, Südkaukasus und Zentralasien” Forschung verlaufen sind und wie diese Erfahrungen bei dem Aufbau eines neuen Communityspace einbezogen werden können. Dies betrifft auch Kernbestandteile, wie die Vernetzungskomponenten und die Möglichkeit für Diskussionen über Daten auf der Plattform.

1 Einleitung

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.[1] Aktuelle Standards und die Digitalisierung bestehender Prozesse und Angebote benötigen jedoch Strukturen, um nachhaltige Entwicklungsmodelle zu schaffen und zu ermöglichen.[2] Dies betrifft auch die Betrachtung der Datenqualität, welche in der offenen akademischen Diskussion von immer zentralerer Bedeutung

[1] Rapp (2021).
[2] Bingert, Buddenbohm, Engelhardt, Kurzawe (2017).

wird.[3] Discuss Data[4] bietet hierfür eine Plattform, die dem digitalen Forschungsdatenmangement (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.[5] 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, im Fall von Discuss Data sind das bisher die sozialwissenschaftlichen Regionalstudien zum postsowjetischen Raum. Discuss Data setzt auf die FAIR Prinzipien, ermöglicht aber auch den Einsatz restriktiverer Standards bei besonders schützenswürdigen Datensets.

2 Erfahrungen im Aufbau von Discuss Data

Discuss Data startete 2016[6] als 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 102 Datensätze publiziert[7] und es haben sich 118 Nutzende registriert. 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.[8] Diese auf Konferenzen und Reviews

[3] Hierzu ist insbesondere auch der Bericht des Rat für Informationsinfrastrukturen (2019) zu nennen, welcher die Problemstellung und damit verbundenen Herausforderungen ausführlich beschreibt.
[4] https://www.discuss-data.net.
[5] Herrmann und Kurzawe (2020).
[6] Erste Projektphase 2016 bis 2021, gefördert durch die DFG (Projektnummer 323616639).
[7] Stand: 20.04.2023.
[8] Barlösius (2023).

durchaus übliche und fachlich äußerst wichtige Diskussionskultur hat sich, trotz der positiven Haltung dazu, bisher nicht etabliert.

3 Aufbau eines Discuss Data Communityspaces für die Digital Humanities

Die DH-Community eignet sich aus mehreren Gründen gut für den Aufbau eines weiteren Discuss Data Communityspaces. Die DH-Community ist als Community of Practice mit dem gegenseitigen Kommentieren und Begutachten im digitalen öffentlichen Raum vertrauter, als traditionsreichere Fachcommunities. Methodisch vereinen sich agile und vernetzte Ansätze aus der Softwareentwicklung mit geisteswissenschaftlichen Forschungsmethoden und Fragestellungen.[9] Mit dedizierten Lehrstühlen und einem eigenen Fachverband sowie zahlreichen Forschungsprojekten ist die DH-Forschung fest im Wissenschaftssystem verankert. Sie differenziert sich intern zunehmend aus, wofür auch die Gründung neuer Fachzeitschriften ein guter Indikator ist.[10]

Forschung in den DH ist durch intensive Nutzung von digitalen Methoden gekennzeichnet.[11] 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[12] 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.[13]

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

[9] Balzer, Eleftheriadis und Kurzawe (2018).
[10] Beispiele sind das Journal of Digital History (seit 2021) oder das Journal of Computational Literary Studies (seit 2022).
[11] Hughes, Constantopoulos und Dallas (2015).
[12] Kurzawe (2023).
[13] Asef et al. (2022).

Jahren eine der zentralen Herausforderungen der DH geworden.[14] 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. Bei ebenfalls häufig genutzten kommerziellen Angeboten (z.B. Github) kommt zusätzlich das Risiko hinzu, dass unklar ist, wie sich die Unternehmen entwickeln werden (z.B. ob sie zukünftig Gebühren verlangen oder Angebote einstellen) und ob die Forschungsdaten langfristig verfügbar bleiben.

Die erste Förderphase zeigte, dass es essentiell ist Datenkurator:innen und Redaktionsmitglieder zu gewinnen, um die Communityspaces langfristig von der Community tragen zu lassen. Dafür sind Positivbeispiele notwendig, die Mehrwerte aufzeigen und den Zeitaufwand rechtfertigen. Die stärkere Einbindung von Diskussionen als Micropublikationen könnte hierzu beitragen. Auch die Begutachtung und das Review von Forschungsdatensätzen wird als wichtiges Instrument der Qualitätssicherung im Forschungs- und Publikationsprozess an Bedeutung gewinnen.

Bibliografie

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.

Balzer, Wolfgang, A. Eleftheriadis und Daniel Kurzawe. „Digital Humanities and Hermeneutics“. Philosophical Inquiry 42, Nr. 3 (2018): 103–19. https://doi.org/10/gddms9.

[14] Fickers (2020).

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.

Kurzawe, Daniel. 2023. Die Dynamik von Forschung und Gesellschaft: Simulationen von Wissenschaftsprozessen. Hildesheim, München: Olms Universitätsbibliothek Ludwig-Maximilians-Universität München. https://doi.org/10.5282/edoc.29687.

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.

Wilkinson, Mark D., Michel Dumontier, IJsbrand Jan Aalbersberg, Gabrielle Appleton, Myles Axton, Arie Baak, Niklas Blomberg, et al., “The FAIR Guiding Principles for Scientific Data Management and

Stewardship," *Scientific Data* 3 (2016): 160018. https://doi.org/10.1038/sdata.2016.18.

Kontingente Beobachtungen: Forschungsdaten unter konstruktivistischem Paradigma

Kleymann, Rabea
rabea.kleymann[at]phil.tu-chemnitz.de
Leibniz-Zentrum für Literatur- und Kulturforschung, Deutschland
ORCID-iD: 0000-0003-3856-2685

Zusammenfassung: Geisteswissenschaftliche Daten können als Voraussetzung und Ergebnis kontingenter Forschungskontexte und infrastruktureller Settings gelesen werden. In den DH wird häufig der von Johanna Drucker eingeführte Terminus *capta* verwendet, um den konstruktivistischen Charakter von Forschungsdaten zu betonen. Der Konstruktivismus stellt ein latentes Forschungsparadigma der DH dar, das sich unter anderem in Datenpraktiken manifestiert. Der Vortrag widmet sich den Voraussetzungen, Implikationen und Problemfeldern des konstruktivistischen Paradigmas und fragt mithilfe von alternativen Theorieentwürfen nach den (noch) nicht ausgeschöpften transformativen Potenzialen. Drei Fallbeispiele zu Datensammlungen, -bereinigungen und -analysen werden dazu vorgestellt.

Dass Forschungsdaten kontextabhängig gesammelt, analysiert und visualisiert werden, das heißt mithin konstruiert werden, ist ein Allgemeinplatz in den Digital Humanities (DH).[1] Über den von Drucker eingeführten Begriff *capta* wird seither der damit einhergehenden Situiertheit und Lokalität der Forschungsdaten Ausdruck verliehen.[2] Gegenstand der DH sind eben nicht Daten, sondern *capta*. Druckers terminologische Setzung beruht auf einer wissenschaftstheoretischen Unterscheidung zwischen Realismus und Konstruktivismus. Mit *capta* wird folglich ein konstruktivistisches Paradigma für geisteswissenschaftliche Forschungsdaten vorausgesetzt.[3] Implizit ist

[1] Der Ausdruck *raw data* wurde im Kontext der DH zurückgewiesen, vgl. Gitelmann 2013; Owens 2011; Kemman 2022, 288. Interessanterweise findet sich der Ausdruck an prominenter Stelle im ersten Satz der Einleitung von *R for Data Science* (Wickham, Cetinkaya-Rundel, und Grolemund 2023).

[2] Vgl. Drucker 2011. Vgl. Kritik am Begriff u. a. von Lavin 2021.

[3] Weitere Fundstellen: Vgl. Jannidis 2017, 107. Im Rahmen seiner Überlegungen zur Datenmodellierung merkt Jannidis an, dass „in den Geisteswissenschaften erkenntnistheoretische Positionen, die mehr oder weniger klar als Varianten eines nicht allzu radikalen Konstruktivismus zu erkennen sind, besonders häufig sind.“ Vgl. Clemont 2016.

damit auch eine Kontingenz datenbasierter Wissensansprüche verbunden, die nicht nur mit Blick auf Zeitdiagnosen von der Postfaktizität an Relevanz gewinnen. Vielmehr stellen auch Verfahren des maschinellen Lernens die Situiertheit geisteswissenschaftlicher Forschungsdaten zur Disposition.

In wissenschaftstheoretischer Perspektive widmet sich der Vortrag Voraussetzungen, Implikationen und Problemfeldern eines konstruktivistischen Paradigmas für geisteswissenschaftliche Forschungsdaten. Ich argumentiere, dass konstruktivistische Paradigmen in Anlehnung an Kuhn einen Einfluss auf die in den DH emergierenden Datenkulturen haben. Zugleich ist die Verhandlung latenter Paradigmen für eine kritische Datenpraxis in den DH unerlässlich. Zu fragen ist daher, inwiefern ein konstruktivistisches (Daten-)Paradigma für die DH weiterhin vertretbar und wünschenswert ist. Mit anderen Worten: Hält der Konstruktivismus als Alleinstellungsmerkmal geisteswissenschaftlicher Forschungsdaten immer noch ein transformatives Potenzial bereit, wie Drucker es vor mehr als zehn Jahren postulierte? Im Vortrag diskutiere ich alternative Theorieentwürfe (Akteur-Netzwerk-Theorie, Postfundamentalismus, Standpunkttheorie, spekulativer Konstruktivismus), um mögliche Grenzen des konstruktivistischen Paradigmas aufzuzeigen.[4] Methodisch gehe ich wie folgt vor: In einem ersten Schritt wird die diskursive Formation des Konstruktivismus im Forschungsdatenmanagement der DH in den Blick genommen. In einem zweiten Schritt untersuche ich anhand von drei Fallbeispielen, wie sich konstruktivistische Annahmen in etablierten Datenpraktiken manifestieren.

In der Wissenschaftsgeschichte bezeichnet der Begriff *Konstruktivismus* eine Reihe unterschiedlicher biologischer, psychologischer, kybernetischer und erkenntnistheoretischer Positionen, die sich im Allgemeinen mit Fragen der Wirklichkeitskonstruktion und den Erkenntnisbedingungen durch Beobachter*innen beschäftigen.[5] Merkmale des Konstruktivismus sind

[4] Vgl. Latour 2007; Smithies 2014; Marchart 2010; Harding 2003; Stengers 2008.
[5] Vgl. Keller und Zierold 2011, 421; Collin 2008, 24f.

unter anderem die Kritik an „realistische[n], ontologische[n] sowie korrespondenztheoretische[n] Auffassungen von Wahrheit und Wissen“[6], eine Ablehnung von Objektivitätsidealen sowie Letztbegründungen, eine Verschiebung vom *Was* der Erkenntnis zum *Wie* des Erkenntnisvorgangs sowie eine Einführung von Intersubjektivität als Erkenntnismaßstab.[7] Von Kritiker*innen werden konstruktivistische Ansätze mit Blick auf einen (Werte-)Relativismus abgelehnt.[8] Im Rahmen der feministischen Wissenschaftstheorie wurde vor allem der fehlende Diskurs ontologischer und politischer Konstellationen problematisiert.[9]

Im Vortrag werden drei Fallbeispiele aus den DH diskutiert, die auf etablierte Datenpraktiken der Sammlung, Bereinigung bzw. Manipulation und Analyse Bezug nehmen. Ein erstes Fallbeispiel widmet sich der Sammlung von Daten aus dem GLAM-Bereich anhand von Application Programming Interfaces (API). Die mit Loukissas gesprochenen „data settings“[10] werden mit Blick auf die Dokumentation von Kontingenzen ins Verhältnis gesetzt. Wie lässt sich die radikale Vielfalt lokaler Datensettings eigentlich bei einer API-Abfrage nachvollziehen? Reicht es aus, Kontingenzen geisteswissenschaftlicher Forschungsdaten durch Reflexivität zu adressieren? Am zweiten Fallbeispiel des *R*-Pakets *tidyverse* wird das Verhältnis der Datenbereinigung und -manipulation zur Beobachtung diskutiert.[11] Von Interesse ist unter anderem die in *R* formulierte Regel „[E]ach column is a variable, and each row is an observation“[12] für bereinigte Datensätze. Rawson und Muñoz sowie Klein und D’Ignazio weisen bereits darauf hin, dass Verfahren der Datenbereinigung und -manipulation Differenzen verschwinden lassen.[13] Ein drittes Fallbeispiel rückt

[6] Prechtl und Burkard 2008, 310.
[7] Vgl. Keller und Zierold 2011, 426; Hacking 2003, 23f.
[8] Vgl. Latour 2003; Schmidt 2003, 128.
[9] Vgl. Haraway 1995, 78. Haraway hält fest: „Es reicht nicht aus, auf die grundlegende historische Kontingenz zu verweisen und zu zeigen, wie alles konstruiert ist. […] In traditionellen philosophischen Kategorien formuliert, heißt das, dass es möglicherweise stärker um Ethik und Politik geht als um Epistemologie.“
[10] Loukissas 2019, 2.
[11] Vgl. Wickham 2014.
[12] Wickham, Cetinkaya-Rundel, und Grolemund 2023.
[13] Vgl. Rawson und Muñoz 2019; D’Ignazio und Klein 2020, 132.

statistischen Inferenzbildungen bei Datenanalysen in den Fokus. Anhand von Analysen aus dem Bereich der Computational Literary Studies diskutiere ich mögliche Auswirkungen von statistischen Messungen, z. B. dem Bayes-Theorem oder dem p-Wert, auf ein konstruktivistisches (Daten-)Paradigma.[14]

Bibliographie

Alvarado, Rafael C. „Datawork and the Future of Digital Humanities." In *The Bloomsbury Handbook to the Digital Humanities*, herausgegeben von James O'Sullivan, 361–72. London: Bloomsbury Academic, 2022. https://doi.org/10.5040/9781350232143.

Clemont, Tanya E. „Where is Methodology in Digital Humanities?" In *Debates in the Digital Humanities 2016*, herausgegeben von Matthew K. Gold und Lauren F. Klein, 153–75. Minneapolis: University of Minnesota Press, 2016. https://doi.org/10.5749/j.ctt1cn6thb.

Collin, Finn. *Konstruktivismus*. Paderborn: Fink, 2008.

D'Ignazio, Catherine, und Lauren F. Klein. *Data feminism*. Cambridge, Massachusetts: The MIT Press, 2020.

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.

Hacking, Ian. „Soziale Konstruktion beim Wort genommen." In *Entdeckung und Konstruktion. Erkenntnistheoretische Kontroversen*, herausgegeben von Matthias Vogel und Lutz Wingert, 23–54. Frankfurt/Main: Suhrkamp, 2003.

Haraway, Donna Jeanne. „Situiertes Wissen. Die Wissenschaftsfrage im Feminismus und das Privileg einer partialen Perspektive." In *Die Neuerfindung der Natur: Primaten, Cyborgs und Frauen*, herausgegeben von Carmen Hammer und Immanuel Stieß, 73–98. Frankfurt/Main: Campus-Verl., 1995.

[14] Vgl. Alvarado 2022, 366; Karsdorp et al. 2021.

Harding, Sandra. „Starke Objektivität.“ In *Entdeckung und Konstruktion. Erkenntnistheoretische Kontroversen*, herausgegeben von Matthias Vogel und Lutz Wingert, 162–90. Frankfurt/Main: Suhrkamp, 2003.

Jannidis, Fotis. „Grundlagen der Datenmodellierung.“ In *Digital Humanities: eine Einführung*, herausgegeben von Fotis Jannidis, Hubertus Kohle, und Malte Rehbein, 99–107. Stuttgart: J.B. Metzler Verlag, 2017.

Gitelman, Lisa, Hrsg. *Raw data is an oxymoron*. Cambridge, Massachusetts: The MIT Press, 2013. https://doi.org/10.7551/mitpress/9302.001.0001?locatt=mode:legacy.

Karsdorp, Folgert, Mike Kestemont, und Allen Ridell. *Humanities Data Analysis: Case Studies with Python*. Princeton: Princeton University Press, 2021.

Keller, Katrin, und Martin Zierold. „Konstruktivismus.“ In *Lexikon der Geisteswissenschaften: Sachbegriffe-Disziplinen-Personen*, herausgegeben von Helmut Reinalter und Peter J. Brenner, 421–27. Wien: Böhlau, 2011.

Kemman, Max. „Tool Criticism through Playful Digital Humanities Pedagogy.“ In *The Bloomsbury Handbook to the Digital Humanities*, herausgegeben von James O’Sullivan, 287–94. London: Bloomsbury Academic, 2022. https://doi.org/10.5040/9781350232143.

Latour, Bruno. „The promises of constructivism.“ In *Chasing technoscience: matrix for materiality*, herausgegeben von Don Ihde und Evan Selinger, 27–46. Bloomington: Indiana University Press, 2003.

Latour, Bruno. *Eine neue Soziologie für eine neue Gesellschaft: Einführung in die Akteur-Netzwerk-Theorie*. 1. Aufl. Frankfurt/Main: Suhrkamp, 2007.

Lavin, Matthew. „Why Digital Humanists Should Emphasize Situated Data over Capta.“ *Digital Humanities Quarterly* 15, Nr. 2 (2021). http://www.digitalhumanities.org/dhq/vol/15/2/000556/000556.html.

Loukissas, Yanni Alexander. *All data are local: Thinking critically in a data-driven society*. Cambridge Massachusetts and London England: The MIT Press, 2019.

Marchart, Oliver. *Die politische Differenz: zum Denken des Politischen bei Nancy, Lefort, Badiou, Laclau und Agamben*. Berlin: Suhrkamp, 2010.

Prechtl, Peter, und Franz-Peter Burkard, Hrsg. *Metzler Lexikon Philosophie: Begriffe und Definitionen*. 3., Erweiterte und Aktualisierte Auflage. Stuttgart Weimar: Verlag J.B. Metzler, 2008.

Rawson, Katie, und Trevor Muñoz. „Against cleaning.“ In *Debates in the Digital Humanities 2019*, herausgegeben von Matthew K. Gold und Lauren F. Klein, Bd. 5. Minneapolis: University of Minnesota Press, 2019. https://doi.org/10.5749/9781452963785.

Schmidt, Siegfried J*. Geschichten & Diskurse. Abschied vom Konstruktivismus*. Reinbek bei Hamburg: Rowohlt, 2003.

Smithies, James. „Digital Humanities, Postfoundationalism, Postindustrial Culture“. *Digital Humanities Quarterly* 8, Nr. 1 (2014). http://www.digitalhumanities.org/dhq/vol/8/1/000172/000172.html

Stengers, Isabelle. *Spekulativer Konstruktivismus*. Internationaler Merve-Diskurs 312. Berlin: Merve-Verl, 2008.

Wickham, Hadley. „Tidy Data.“ *Journal of Statistical Software* 59, Nr. 10 (2014). https://doi.org/10.18637/jss.v059.i10.

Wickham, Hadley, Mine Cetinkaya-Rundel, und Garrett Grolemund. „R for Data Science. Import, Tidy, Transform, Visualize and Model Data,“ 2023. Letzter Aufruf am 25.04.2023. https://r4ds.hadley.nz/.

Immer FAIR?!

Problematische Inhalte in den Datenbeständen der Provenienzforschung

Lang, Sabine
sab.lang[at]fau.de
Friedrich-Alexander-Universität Erlangen-Nürnberg, Deutschland

Zusammenfassung. Die Datenbestände der Provienzforschung weisen problematische Inhalte auf. Im Kontext der FAIR-Prinzipien für Forschungsdaten, die die Auffindbarkeit, Zugänglichkeit, Interoperabilität und Wiederverwendbarkeit von Daten fordern, stellen sich sodann folgende Fragen: Wie soll man mit problematischen Inhalten im Digitalen umgehen? Welche Strategien gibt es? Wie offen können Inhalte der Provenienzforschung sein? Sind die FAIR-Prinzipien für die Provenienzforschung überhaupt umsetzbar? Und wie legt das Fach den Inhalt der Prinzipien aus? Der Beitrag widmet sich diesen Fragen anhand verschiedener Datenbankbeispiele aus der Provenienzforschung zu NS-verfolgungsbedingt entzogenem Kulturgut und schlägt mögliche Strategien für den Umgang mit problematischen Inhalten vor.

1 Immer FAIR oder wie FAIR?

1.1 Einführung

Die Datenbestände der Provenienzforschung können problematische Inhalte aufweisen, die im Kontext der FAIR-Prinzipien[1] für Forschungsdaten Fragen aufwerfen: Wie soll man mit problematischen Inhalten im Digitalen umgehen? Welche Strategien gibt es? Wie offen können die Inhalte der Provenienzforschung sein und sind die FAIR-

[1] Vgl. Wilkinson et al. 2016.
FAIR steht für **F**indability, **A**ccessibility, **I**nteroperability und **R**eusability (vgl. Wilkinson et al. 2016, 1). Die Prinzipien gestatten eine Einschränkung des Zugriffs, wenn es sich um sensible Inhalte handelt und der Zugangsweg nachvollziehbar ist. “Data should be as open as possible and as closed as necessary[.]” (zitiert nach: Landi et al. 2020, 50). Im Kontext der Provenienzforschung kann eine Einschränkung des Zugangs allerdings problematisch sein, da es der Forderung nach Transparenz widerspricht und z.B. die Suchen der Nachkommen der NS-Opfer behindert. Eine Auseinandersetzung mit der inhaltlichen Auslegung und Umsetzung der FAIR-Prinzipien, vor allem mit der Forderung nach Zugänglichkeit, hat noch nicht in ausreichendem Maß stattgefunden. Konsens darüber ist aber Voraussetzung für die Entwicklung brauchbarer Strategien (vgl. Forschungsdaten und Forschungsdatenmanagement, n.d.; Landi et al. 2020).

Prinzipien überhaupt umsetzbar? Der Vortrag widmet sich diesen Fragen anhand verschiedener Datenbankbeispiele aus der Provenienzforschung zu NS-Raubgut. Außerdem werden Strategien für den Umgang mit problematischen Inhalten benannt. Diskriminierende und rassistische Begriffe, personenbezogene Daten, die persönliche und deshalb sensible Informationen enthalten, oder Inhalte politischer Propaganda – all dies sind Beispiele von möglichen problematischen Inhalten in der Provenienzforschung zu NS-Raubgut. Hinzukommen problematische Abbildungen, die hier aus Platzgründen ausgeklammert werden müssen. Strategien sind deshalb unbedingt notwendig, um zu verhindern, dass Diskriminierungen weitergetragen werden, bestimmte Personengruppen verletzt oder Inhalte für politische Zwecke missbraucht werden.[2]

Der Beitrag ist eingebettet in aktuelle Debatten über sensible Sprache in Museen[3], den Umgang mit sensiblen Objekten wie menschlichen Überresten oder Sammlungsgut aus kolonialen Kontexten[4] oder mit rassistischen und diskriminierenden Begriffen in Werktiteln und Werkbeschreibungen.[5] Andere Beiträge haben sich diskriminierenden oder rassistischen Abbildungen gewidmet.[6] Auch die Rolle der Digitalisierung wird in diesem Kontext besprochen; so werden Potentiale und Herausforderungen der Digitalisierung ethnologischer Sammlungen und die Reproduktion problematischer Inhalte thematisiert.[7] Auch eine Diskussion über die FAIR-Prinzipien hat in verschiedenen Bereichen stattgefunden wie z.B. den Bibliothekswissenschaften[8] oder in Bezug auf den Zugang zu Gesundheitsdaten[9]. Da sich erst kürzlich gegründete Initiativen wie die Nationale Forschungsdateninfrastruktur (NFDI)[10] mit Datenbeständen und im Besonderen mit Standards für historische Forschungsdaten[11] befassen, muss eine Auseinandersetzung mit problematischen Inhalten

---

[2] Vgl. Tayiana, n.d.
[3] Vgl. Retour 2022; Lenbachhaus, n.d.
[4] Vgl. Deutscher Museumsbund e.V. 2021a; Deutscher Museumsbund e.V. 2021b.
[5] Vgl. Soltau 2021; Staatliche Kunstsammlungen Dresden, Online Collection, n.d.
[6] Vgl. Harbeck und Strickert 2020a; Harbeck und Strickert 2020b.
Der Umgang mit problematischen Abbildungen kann im Kontext des Beitrags nicht thematisiert werden, da es den Umfang übersteigen würde. Auch eine Besprechung anderer Kontexte der Provenienzforschung wie koloniale Kontexte muss an dieser Stelle entfallen.
[7] Vgl. Hahn et al. 2021.
[8] Vgl. Deppe 2020.
[9] Vgl. Landi et al. 2020.
[10] Vgl. Nationale Forschungsdateninfrastruktur (NFDI), n.d.
[11] Vgl. NFDI4Memory, n.d.

und Umgangsmöglichkeiten jetzt stattfinden. Nur dann können Bedürfnisse und Anforderungen bei der Entwicklung von Standards berücksichtigt werden.

1.2 Strategien für den Umgang mit problematischen Inhalten

Um mehr über die Herkunft eines Objektes herauszufinden, stehen der Provenienzforschung zu NS-Raubgut verschiedene Datenbanken zur Verfügung, die problematische Inhalte abbilden können. Eine wichtige Quelle ist die Datenbank German Sales[12], die über 11.500 Auktions- und Verkaufskataloge enthält. Die öffentlich zugänglichen Kataloge sind vorwiegend aus dem deutschsprachigen Raum und im Zeitraum zwischen 1901 und 1945 entstanden.[13] Aufgrund ihres zeitlichen und geographischen Schwerpunkts können die Kataloge Begrifflichkeiten enthalten, die die nationalsozialistische Ideologie widerspiegeln. Ein Beispiel hierfür ist die Namensliste der Besitzer*innen, die im Katalog des Berliner Auktionshauses von Max Perl aus dem Jahr 1939 abgedruckt ist.[14] Neben diskriminierenden oder rassistischen Begriffen sind es auch private und deshalb sensible Informationen, die in Datenbeständen abgebildet sein können. Diese enthalten z.B. Namen von Geschädigten, Adressen, Angaben zu Vermögen oder Verfolgungsschicksalen. Die WGA Datenbank[15] ist ein Beispiel hierfür; der auf der Webseite aufrufbare Bereich Erläuterungen zur Datenbank zeigt Abbildungen mit den vorhandenen Datenbankfeldern wie Verfahren (Antragsteller), geschädigt oder Gegenstand sowie Beispieldatensätze, die z.B. Informationen zu Adressen oder

[12] Vgl. arthistoricum.net, n.d.
[13] Vgl. arthistoricum.net, n.d.
[14] Vgl. Perl 1939, 4.
Die Autorin hat sich an dieser Stelle dazu entschieden, den Begriff nicht zu reproduzieren, sondern ‚nur' darauf zu verweisen.
[15] Vgl. Landesarchiv Berlin, n.d. "WGA Datenbank".
Die WGA Datenbank (WGA = Wiedergutmachungsämter) ermöglicht eine umfassende Recherche zu den Verfahrensakten der Berliner Wiedergutmachungsämter, die im Landesarchiv Berlin aufbewahrt werden. Inhaltliche Grundlage sind ungefähr 440.000 Karteikarten, die von Mitarbeiter*innen der Wiedergutmachungsämter auf Basis der Verfahrensakten angelegt wurden. Die Karten benennen z.B. die Namen der Antragsteller*innen, Geschädigten oder Adressen zum Zeitpunkt der Antragstellung und Schädigung. Die Karten wurden zunächst transkribiert, die Informationen in die Datenbank eingepflegt und schließlich mit weiteren Angaben zu Kulturgütern angereichert (vgl. Landesarchiv Berlin, n.d. "WGA Datenbank. Bestandsinformation. Erläuterungen zu den Akten"; Landesarchiv Berlin, n.d. "WGA Datenbank. Bestandsinformation. Erläuterungen zur Datenbank").

Geburtsdaten der Antragsteller*innen enthalten.[16] Die Tatsache, dass zahlreiche Datenbestände der Provenienzforschung problematische Inhalte aufweisen, lässt die Frage aufkommen, ob die FAIR-Prinzipien zulässig sind. Vor allem die Forderungen nach Zugänglichkeit und Wiederverwendbarkeit müssen kritisch reflektiert werden: Sollen alle Daten für jeden/jede uneingeschränkt einsehbar und nutzbar sein? Oder nur bestimmte Daten und nur für autorisierte Nutzer*innen? Sollte die Datennachnutzung an Bedingungen gebunden sein? Aufgrund dieser zahlreichen Fragen, Unsicherheitsfaktoren und der Gefahren einer Veröffentlichung, benötigt es im Umgang mit problematischen Inhalten standardisierte Strategien, nur dann sind die tatsächlich zugänglichen Inhalte nachvollziehbar und bewertbar. Damit Strategien in der Praxis anwendbar sind und den Anforderungen der betroffenen und beteiligten Interessengruppen genügen, müssen sie zudem von der Forschungsgemeinschaft gemeinsam entwickelt und etabliert werden.

Indem historische Werktitel mit Anführungszeichen gekennzeichnet und problematische Begriffe in Werktiteln oder Werkbeschreibungen durch Sternchen (****) ersetzt und erst durch die Aktivierung des Nutzenden eingeblendet werden, haben die Staatlichen Kunstsammlungen Dresden (SKD) eine Strategie aufgezeigt, wie man mit problematischen Inhalten im Digitalen umgehen kann (Abb.1).[17] Eine weitere Möglichkeit besteht in der Kontextualisierung der Begriffe durch z.B. Erläuterungen zur Historie und Problematik der Begriffe. Im Zuge aktueller Debatten über Biases in Trainingsdaten informatischer Modelle hat man in der Machine Learning-Community bereits vor einigen Jahren Datasheets for Datasets[18] vorgeschlagen, die „[…] creation, composition, intended uses, maintenance, and other properties […]"[19] dokumentieren. In ähnlicher Weise könnte man auch für Datenbestände der Provenienzforschung Datenblätter anlegen, die Angaben zu problematischen Inhalten machen. Möglich wäre zudem der Austausch durch neutrale Begriffe oder die Löschung problematischer Inhalte. Da Quellen aber immer auch Zeugnisse historischer Tatsachen sind, könnten die Änderung oder Löschung der Inhalte die damit verbundene Problematik mindern oder sogar Geschichte verfälschen. Und schließlich: „The erasure of problematic and offensive terms does

[16] Vgl. Landesarchiv Berlin, n.d. "WGA Datenbank. Bestandsinformation. Erläuterungen zur Datenbank".
[17] Siehe beispielhaft den Datensatz zum Objekt von Johann Gotthelf Studer (Münzmeister) in der online Sammlung der SKD.
[18] Vgl. Gebru et al. 2018.
[19] Gebru et al. 2018, 1.

equate to the removal of problematic context and histories […].“[20] Der Beitrag wird die Anwendbarkeit dieser Strategien deshalb kritisch reflektieren.

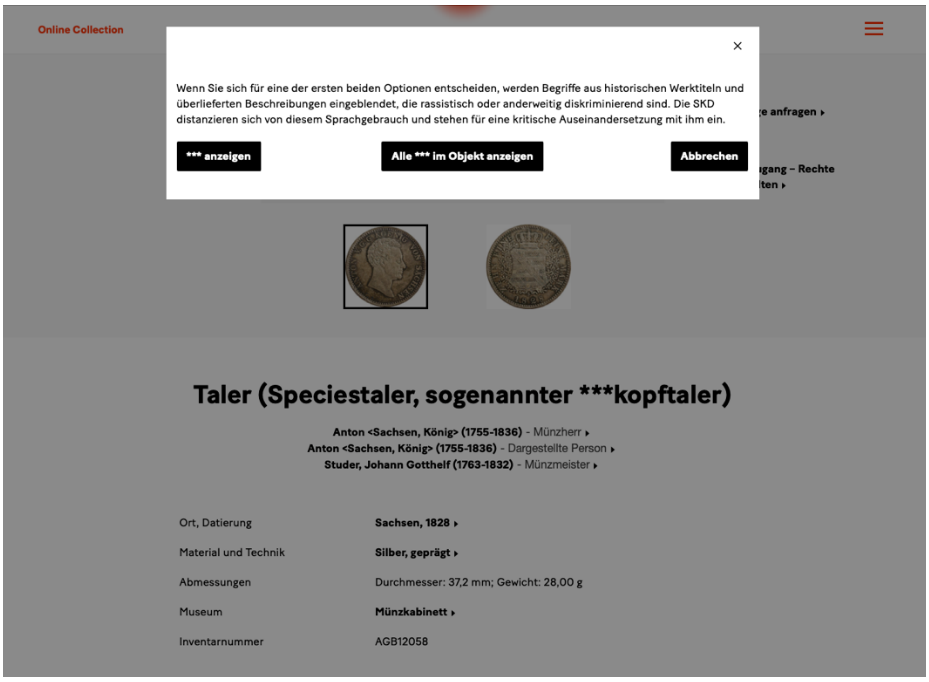


Abb. 1. Die SKD haben sich entschieden, problematische Begriffe in Werktiteln oder Werkbeschreibungen durch (****) zu ersetzen. Erst durch den Klick auf die Asterisken kann der Titel vollständig angezeigt werden. Quelle: Datensatz zu Objekt von Johann Gotthelf Studer (Münzmeister), Online Collection der SKD. Bildnachweis: Münzkabinett, Staatliche Kunstsammlungen Dresden, Foto: SKD.

Bibliografie

arthistoricum.net. n.d. “German Sales.“ Abgerufen am 3. April 2023. https://www.arthistoricum.net/themen/portale/german-sales.

Deppe, Arvid. 2020. "FAIR, CARE und mehr. Prinzipien für einen verantwortungsvollen Umgang mit Forschungsdaten." In *Historisches Erbe und zeitgemäße Informationsinfrastrukturen: Bibliotheken am Anfang des 21. Jahrhunderts,* herausgegeben von Matthias Schulze, 299-312. Kassel: Kassel University Press.

[20] Tayiana, n.d.

Deutscher Museumsbund e.V., ed. 2021a. *Leitfaden Umgang mit Sammlungsgut aus kolonialen Kontexten.* 3. Fassung. Berlin. https://www.museumsbund.de/wp-content/uploads/2021/02/leitfaden-zum-umgang-mit-sammlungsgut-aus-kolonialen-kontexten-web.pdf.

Deutscher Museumsbund e.V., ed. 2021b. *Leitfaden Umgang mit menschlichen Überresten in Museen und Sammlungen.* Überarbeitete Fassung. Berlin. https://www.museumsbund.de/wp-content/uploads/2021/06/dmb-leitfaden-umgang-menschl-ueberr-de-web-20210623.pdf.

Forschungsdaten und Forschungsdatenmanagement. n.d. "FAIRe Daten." Letzte Änderung am 5. Mai 2023, https://forschungsdaten.info/themen/veroeffentlichen-und-archivieren/faire-daten/.

Gebru, Timnit, Jamie Morgenstern, Briana Vecchione, Jennifer Wortman Vaughan, Hannah Wallach, Hal Daumé III, und Kate Crawford. 2018. "Datasheets for Datasets." *Proceedings of the 5th Workshop on Fairness, Accountability, and Transparency in Machine Learning*. Stockholm, Schweden. https://www.microsoft.com/en-us/research/uploads/prod/2019/01/1803.09010.pdf.

Hahn, Hans Peter, Oliver Lueb, Katja Müller, und Karoline Noack, eds. 2021. *Digitalisierung ethnologischer Sammlungen. Perspektiven aus Theorie und Praxis.* Bielefeld: transcript Verlag. https://doi.org/10.1515/9783839457900.

Harbeck, Matthias und Moritz Strickert. 2020a. "Freiwilligkeit und Zwang. Eine Diskussion im Kontext der frühen ethnologischen Fotografie." Potsdam: ZZF – Centre for Contemporary History: Visual History. https://doi.org/10.14765/zzf.dok-1928.

Harbeck, Matthias und Moritz Strickert. 2020b. "Zeigen / Nichtzeigen." Potsdam: ZZF – Centre for Contemporary History: Visual History. https://doi.org/10.14765/zzf.dok-1927.

Landesarchiv Berlin. n.d. "WGA Datenbank." Accessed April 3, 2023. http://www.wga-datenbank.de/starten.php?s=1.

Landesarchiv Berlin. n.d. "WGA Datenbank. Bestandsinformation. Erläuterungen zu den Akten." Abgerufen am 3. April 2023. http://www.wga-datenbank.de/erlaeuterungen-zu-den-akten.php?s=2&sub=3.

Landesarchiv Berlin. n.d. "WGA Datenbank. Bestandsinformation. Erläuterungen zur Datenbank." Abgerufen am 11. April 2023. http://www.wga-datenbank.de/erlaeuterungen-zur-datenbank.php?s=2&sub=4.

Landi, Annalisa, Mark Thompson, Viviana Giannuzzi, et al., "The "A" of FAIR – As Open as Possible, as Closed as Necessary." *Data Intelligence* 2, issue 1-2 (2020): 47-55. https://doi.org/10.1162/dint_a_00027.

Lenbachhaus. n.d. "Worte finden – Sensible Sprache in Provenienzforschung und im musealen Kontext." Abgerufen am 3. April 2023. https://www.lenbachhaus.de/besuchen/kalender/termin/digitale-vortrags-und-gespraechsreihe-16940.

Nationale Forschungsdateninfrastruktur (NFDI). n.d. Abgerufen am 3. April 2023. https://www.nfdi.de.

NFDI4Memory. n.d. Abgerufen am 3. April 2023. https://4memory.de.

Perl, Max, ed. 22. Mai 1939. *Alte und moderne Bücher, Graphik, Handzeichnungen.* Ausstellungskatalog. Berlin. https://doi.org/10.11588/diglit.6293#0004.

Retour – Freier Blog für Provenienzforschende. 2022. "Worte finden – Sensible Sprache in Provenienzforschung und im musealen Kontext." Veröffentlicht am 5. Januar 2022. https://retour.hypotheses.org/1695.

Soltau, Hannes. 2021. "Debatte um Umbenennung von Kunstwerken. In Dresdner Museen findet weder ein „Bildersturm" noch „Zensur" statt." *Tagesspiegel*, 15. September 2021. https://www.tagesspiegel.de/kultur/in-dresdner-museen-findet-weder-ein-bildersturm-noch-zensur-statt-4276738.html.

Staatliche Kunstsammlungen Dresden, Online Collection. n.d. "Ethische Leitlinien der Online Collection (Arbeitsprozess, diskriminierungsfreie Sprache, sensible Objekte).“ Abgerufen am 16. April 2023. https://skd-online-collection.skd.museum.

Studer, Johann Gotthelf (Münzmeister), *Taler (Speciestaler, sogenannter ***kopftaler)*, 1828, Sachsen. Staatliche Kunstsammlungen Dresden, Online Collection. Abgerufen am 8. Juli 2023. https://skd-online-collection.skd.museum/Details/Index/1466101.

Tayiana, Chao. n.d. “Use of derogatory, racist and harmful language.“ Abgerufen am 10. Juli 2023. https://digitalbenin.org/documentation/use-of-derogatory-racist-and-harmful-language.

Wilkinson, Mark D., Michel Dumontier, IJsbrand Jan Aalbersberg, Gabrielle Appleton, Myles Axton, Arie Baak, Niklas Blomberg, et al., “The FAIR Guiding Principles for Scientific Data Management and Stewardship.” *Scientific Data* 3 (2016): 160018. https://doi.org/10.1038/sdata.2016.18.

Von der Herkunft zur Zukunft

Interdisziplinäre Ansätze zur Erforschung von Provenienzen in Museen

Ludwig, Elisa
elisa.ludwig[at]kunstgeschichte.uni-muenchen.de
Ludwig-Maximilians-Universität München, Deutschland
ORCID-iD: 0009-0006-5647-3474

Maget Dominicé, Antoinette
antoinette.maget[at]kunstgeschichte.uni-muenchen.de
Ludwig-Maximilians-Universität München, Deutschland
ORCID-iD: 0000-0001-9056-4544

Schneider, Stefanie
stefanie.schneider[at]itg.uni-muenchen.de
Ludwig-Maximilians-Universität München, Deutschland
ORCID-iD: 0000-0003-4915-6949

Vollmer, Ricarda
ricarda.vollmer[at]campus.lmu.de
Ludwig-Maximilians-Universität München, Deutschland
ORCID-iD: 0000-0002-4105-1045

Zusammenfassung. Infolge des gesteigerten öffentlichen Interesses und des wachsenden Bewusstseins bezüglich der Herkunft von Kulturgütern, gewinnt die Provenienzforschung als Methode und Forschungsgebiet zunehmend an Bedeutung. Die Einreichung untersucht die Relevanz von Forschungsdaten in Bezug auf Provenienzangaben in Museumsdatenbanken. Es wird aufgezeigt, dass Forschungsdaten dort eine essenzielle Rolle bei der Gewährleistung der Transparenz, Vernetzung und Zugänglichkeit einnehmen. Dabei werden Online-Sammlungen internationaler Kunstmuseen mittels aus relevanten Leitfäden abgeleiteten Kriterien vergleichend analysiert, wobei quantitative und qualitative Methoden komplementär eingesetzt werden. Die Einreichung betont die Bedeutung (nach)nutzbarer Forschungsdaten, um provenienzwissenschaftliche Belange zu unterstützen und zu einem verantwortungsbewussten Umgang mit dem kulturellen Erbe beizutragen.

1 Hintergrund

Eine der zentralen Aufgaben eines Museums ist es, materielles und immaterielles Erbe zu erforschen und zu dokumentieren – wie 2022 in der neu verabschiedeten Definition des Internationalen Museumsverbands (International Council of Museums; ICOM) bestätigt wurde.[1] Durch diesen Schritt hat die letzte Generalversammlung des ICOM die in den ethischen Richtlinien des Verbands verankerte Verpflichtung als festen Bestandteil der Museumsarbeit anerkannt. Betont wird, dass „die vollständige Dokumentation eines Gegenstandes und seiner Besitzverhältnisse vom Zeitpunkt seiner Entdeckung oder Schöpfung bis in die Gegenwart“[2] von grundlegender Bedeutung sei; die Forschenden hätten eine Sorgfaltspflicht zur bestmöglichen, reflektierten und lückenlosen Angabe dieser Daten inne.[3] In der Provenienzforschung werden derartige Herkunftsnachweise ermittelt und dokumentiert. Sie dienen gegenwärtig Museen, dem Kunsthandel und der privaten Sammeltätigkeit als Instrument, um die Authentizität von Kulturgütern zu überprüfen und Eigentumsansprüche zu klären. Dies trägt maßgeblich zu einem verantwortungsvollen und ethisch fundierten Umgang mit dem kulturellen Erbe bei und fördert eine transparente und nachvollziehbare Darstellung der Objektgeschichten.

Ereignisse wie der Schwabinger Kunstfund von Cornelius Gurlitt im Jahr 2012, politische Stellungnahmen und fachliche Impulse haben im vergangenen Jahrzehnt die Erforschung von (Museums-)Sammlungen und deren Dokumentation nicht nur in Deutschland wesentlich an öffentlicher Bedeutung gewinnen lassen. Infolgedessen werden sowohl für das breite Publikum nachvollziehbar als auch innerhalb der Fachwelt (unrechtmäßige) Sammlungskontexte neu verhandelt. Gegenwärtige Untersuchungen richten ihren Fokus auf die systematische Aufarbeitung umfangreicher Archivkonvolute und das globalgeschichtliche Zusammenspiel von Akteur:innen, Objekten und Institutionen in den jeweiligen Kontexten. Die Provenienzforschung sieht sich mit Herausforderungen konfrontiert, die sowohl in der Erfassung von Objektinformationen als auch in der Heterogenität der Verwaltung und Sicherung dieser liegen.

---

[1] ICOM, „Neufassung der ICOM-Museumsdefinition beschlossen“.
[2] ICOM (2010, 28).
[3] ICOM (2010, 12).

2 Methodik

Ein Beitrag zur langfristigen und nachhaltigen Bewältigung jener Herausforderungen soll durch eine digital motivierte Provenienzforschung geleistet werden. Hier setzt die vorgeschlagene Einreichung an: Verfolgt wird ein interdisziplinärer Ansatz, bei dem Forschungsstände in deutschen und US-amerikanischen Museen erfasst, Defizite aufgedeckt und Vorschläge für zukünftige Entwicklungen präsentiert werden.

Seit der Verabschiedung der Washingtoner Prinzipien im Jahr 1998 – und angeregt durch Fälle wie die Causa Gurlitt – konzentrierten sich Untersuchungen in deutschen Museen lange Zeit vorwiegend auf die Aufklärung von verfolgungsbedingt entzogenem Kulturgut während der nationalsozialistischen Herrschaft.[4] In diesem Zusammenhang vermittelt der Leitfaden des im deutschsprachigen Raum tätigen Arbeitskreises Provenienzforschung Empfehlungen zur Standardisierung von Provenienzangaben.[5] Museumsdatenbanken nehmen hierbei idealerweise eine zentrale Rolle ein, um einheitliche, qualitativ und quantitativ hochwertige Daten zu Provenienzen zugänglich zu machen.[6] Festgestellt wurde jedoch immer wieder, dass Provenienzangaben in Museumsdatenbanken nicht einheitlich erfasst werden.[7]

In Anlehnung daran analysieren wir systematisch öffentlich zugängliche (Auszüge von) Datenbanken deutscher und US-amerikanischer Museen, die sich in Bezug auf ihre Organisationsstruktur sowie die Qualität und Quantität der bereitgestellten Daten unterscheiden. Die Untersuchung fokussiert dezidiert auf Kunstmuseen, die in beiden Leitfäden als Musterbeispiele gelten.[8] Folgende drei Untersuchungsperspektiven dienen als Ausgangspunkte: (1) Die erste Kategorie betrifft die Veröffentlichung digitaler Herkunftsinformationen, die für eine bessere Sichtbarkeit und Zugänglichkeit für ein breiteres Publikum entscheidend sind. (2) Die zweite Kategorie thematisiert die Interaktion mit digitalen Provenienzangaben, die es Nutzenden erlaubt, tiefere Einblicke in Informationen zu gewinnen und Zusammenhänge aufzudecken. (3) Die dritte Kategorie bezieht sich auf die technische Dimension, die bei der

[4] Deutsches Zentrum Kulturgutverluste, „Grundsätze der Washingtoner Konferenz in Bezug auf Kunstwerke, die von den Nationalsozialisten beschlagnahmt wurden (Washington Principles)".

[5] Andratschke et al. (2018).

[6] Haffner (2019, 2020); Rother, Koss und Mariani (2022).

[7] Hopp (2018); Fuhrmeister und Hopp (2019); Lang (2023).

[8] Andratschke et al. (2018): Galerie des 20. Jahrhunderts in West-Berlin – Stiftung Preußischer Kulturbesitz; Yeide, Akinsha und Walsh (2001): National Gallery of Art, Washington, D.C., Art Gallery of Ontario, Los Angeles County Museum of Art, und Museum of Fine Arts, Boston.

Entstehung und Implementierung digitaler Provenienzangaben eine bedeutende Rolle einnimmt.

Indem wir uns auf national einschlägige Leitfäden[9] stützen und sie komplementär betrachten, erforschen wir die genannten Perspektiven in neun sorgfältig ausgearbeiteten Unterkategorien wie Chronologie und Vollständigkeit. Mittels Web-Scraping extrahieren wir Daten von jeweils 20 repräsentativen deutschen und US-amerikanischen Museen unterschiedlicher Größe und thematischer Schwerpunkte. Die Erfassung struktureller Unterschiede in der Darstellung von Provenienzangaben wird mittels einer semi-automatisierten Mustererkennung realisiert, um systematische Unterschiede zwischen den Institutionen aufzudecken. Zudem unterziehen wir die ermittelten Daten einer qualitativen Analyse: Sowohl die inhaltliche Tiefe als auch der Informationsgehalt der Angaben wird auf diese Weise evaluiert.

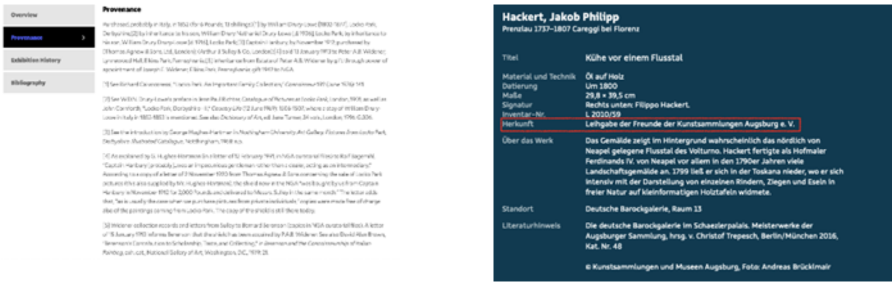
3 Ergebnisse

In zahlreichen Datenbanken werden Provenienzinformationen gemäß den konsultierten Leitfäden in chronologischer Reihenfolge präsentiert und Angaben online dokumentiert, beispielsweise zur Objektidentität, Datierungen und zu Gebrauchsspuren (Abb. 1a und 2a). Trotz dieser Bemühungen wird der aktuelle Stand der Forschung vielerorts (online) nicht transparent dargestellt; Quellenangaben und Provenienzlücken bleiben unerwähnt (Abb. 1b und 2b). Es mangelt an der Verwendung standardisierter Vokabulare wie der Gemeinsamen Normdatei (GND), dem Virtual International Authority File (VIAF) oder Allgemeinen Künstlerlexikon (AKL). Dies hängt unter anderem mit den unterschiedlichen Organisationsstrukturen der Museen zusammen, die sich ebenfalls auf die Erfassung und Dokumentation der Provenienzangaben auswirken. Um die Transparenz, Vernetzung und die Zugänglichkeit der Daten entsprechend unseren Untersuchungsperspektiven zu gewährleisten, sollten daher zukünftig Kriterienkataloge entwickelt werden, die diese Unterschiede berücksichtigen und eine intuitiv-praktische Implementierung in museale Datenbankmanagementsysteme (DBMS) erlauben. Auch eine nur partielle (Online-)Publikation der Daten könnte für die erwähnten Mängel ursächlich sein: Wir gehen davon aus, dass intern abgelegte Informationen vielfach noch nicht für eine breitere Öffentlichkeit freigegeben wurden.

Die konsequente Erfassung und Veröffentlichung von Informationen zur Herkunft von Kunstwerken und Sammlungsobjekten ist für eine

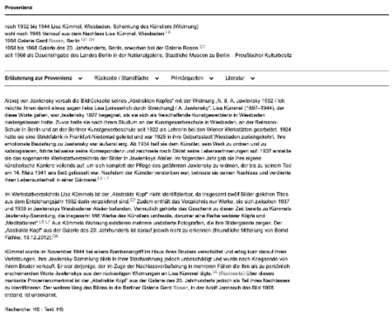
[9] Andratschke et al. (2018); Yeide, Akinsha und Walsh (2001).

kritische Auseinandersetzung mit der Geschichte von Kunstwerken und deren Sammlungen jedoch essenziell: Insbesondere die Aufarbeitung von NS-Raubkunst hat verdeutlicht, dass Provenienzforschung eine moralische Verantwortung innehat und die systematische Untersuchung von Gewaltkontexten ermöglicht. Entscheidend ist in diesem Zusammenhang die Qualität der Forschungsdaten. Sie bildet die Grundlage für eine transparente, nachvollziehbare und präzise Dokumentation von Provenienzen.

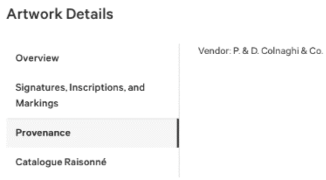


(a) (b)

Abb. 1. Screenshot der Provenienzangaben für (a) Andrea del Castagno, *David with the Head of Goliath*, 1450/1455 (National Gallery of Art, Washington, D.C., https://www.nga.gov/collection/art-object-page.1145.html); (b) Jakob Philipp Hackert, *Kühe vor einem Flusstal*, um 1800 (Kunstsammlungen & Museen Augsburg, https://kunstsammlungen-museen.augsburg.de/sammlung-online).



(a)



(b)

Abb. 2. Screenshot der Provenienzangaben für (a) Alexej von Jawlensky, *Abstrakter Kopf*, 1932 (Die Galerie des 20. Jahrhunderts in West-Berlin. Ein Provenienzforschungsprojekt, https://www.galerie20.smb.museum/werke/958555.html); (b) Albrecht Altdorfer, *Covered Goblet with Three Pomegranates*, o.J. (The Metropolitan Museum of Art, https://www.metmuseum.org/art/collection/search/430277?dep-tids=9&ao=on&ft=*&offset=120&rpp=40&pos=149).

Bibliografie

Andratschke, Claudia, Jasmin Hartmann, Johanna Poltermann, Brigitte Reuter, Iris Schmeisser und Wolfgang Schöddert. *Leitfaden zur Standardisierung von Provenienzangaben*. Hamburg, Arbeitskreis Provenienzforschung e. V., 2018. Zugriff 26.4.2023. https://wissen-schaftliche-sammlungen.de/files/4515/2585/6130/Leitfaden_AP-FeV_online.pdf.

Deutsches Zentrum Kulturgutverluste. „Grundsätze der Washingtoner Konferenz in Bezug auf Kunstwerke, die von den Nationalsozialisten beschlagnahmt wurden (Washington Principles).“ Zugriff: 27.4.2023. https://www.kulturgutverluste.de/Webs/DE/Stiftung/Grundla-gen/Washingtoner-Prinzipien/Index.html.

Fuhrmeister, Christian und Meike Hopp. „Rethinking Provenance Research.“ *Getty Research Journal* 11 (2019): 213–231. 10.1086/702755.

Haffner, Dorothee. „Provenienzforschung digital vernetzt. Ergebnisse sichtbar machen.“ Museumskunde 84 (2019): 90–97. Zugriff: 11.4.2023. https://www.museumsbund.de/wp-content/up-loads/2022/07/museumskunde-2019-1-online.pdf.

Haffner, Dorothee. „Provenienzen in Sammlungsdatenbanken. Digitale und virtuelle Chancen für die Vermittlung.“ *Provenienz & Forschung. Digitale Provenienzforschung* (2020): 36–42.

Hopp, Meike. „Provenienzrecherche und digitale Forschungsinfrastrukturen in Deutschland: Tendenzen, Desiderate, Bedürfnisse.“ In *…(k)ein Ende in Sicht. 20 Jahre Kunstrückgabegesetz in Österreich*, Schriftenreihe der Kommission für Provenienzforschung Band 8 (2018), hg. Eva Blimlinger und Heinz Schödl, 35–59. Wien, Köln: Böhlau Verlag. 10.7767/9783205201274.37.

ICOM – Conseil international des musées, Hg. *Ethische Richtlinien für Museen von ICOM*. Zürich: ICOM, 2010. https://icom-deutschland.de/images/Publikationen_Buch/Publikation_5_Ethische_Richtlinien_dt_2010_komplett.pdf.

ICOM – Conseil international des musées. „Neufassung der ICOM-Museumsdefinition beschlossen.“ Letzte Bearbeitung: 8.9.2022. https://icom-deutschland.de/de/component/content/category/31-museumsdefinition.html?Itemid=114.

Lang, Sabine. „Mind the Gap: Von Lücken in der Provenienzforschung und ihrer Präsenz im digitalen Raum.“ In *Abstracts zur 9. Jahrestagung des Verbands Digital Humanities im deutschsprachigen Raum e.V. „DHd2023: Open Humanities, Open Culture“ an den Universitäten Luxemburg und Trier, 13.3.2023–17.3.2023*, hg. Anna Busch und Peer Trilcke, 212–217. 10.5281/zenodo.7688632.

Rother, Lynn, Max Koss und Fabio Mariani. „Taking Care of History: Toward a Politics of Provenance Linked Open Data in Museums.“ In *Perspectives on Data* (2022), hg. Emily Lew Fry und Erin Canning, Art Institute of Chicago. 10.53269/9780865593152/06.

Yeide, Nancy H., Konstantin Akinsha und Amy L. Walsh. *The AAM Guide to Provenance Research.* Washington, D.C.: American Association of Museums, 2001.

Das TOSCA Modelling Tool
Nachhaltige Dokumentation von Forschungssoftware

Neuefeind, Claes
c.neuefeind[at]uni-koeln.de
Universität zu Köln, Deutschland
ORCID-iD: 0000-0002-9377-9492

Schaeben, Marcel
m.schaeben[at]uni-koeln.de
Universität zu Köln, Deutschland
ORCID-iD: 0000-0001-9672-9856

Schildkamp, Philip
philip.schildkamp[at]uni-koeln.de
Universität zu Köln, Deutschland
ORCID-iD: 0000-0003-0209-2837

Zusammenfassung. Für die Nachnutzbarkeit von Forschungsanwendungen, welche einen zunehmenden Anteil wissenschaftlicher Forschungsergebnisse ausmachen, ist Dokumentation unerlässlich. Dies betrifft neben nutzungsorientierten Bedienungsanleitungen auch die technische Dokumentation der Funktionalität und Betriebsbedingungen solcher Applikationen. Wir schlagen daher vor, den TOSCA-Standard für die Beschreibung von Forschungsanwendungen, deren Bereitstellung und Laufzeitumgebungen einzusetzen und stellen dazu in unserem Beitrag das durch NFDI4Culture geförderte “TOSCA Modelling Tool” vor. Dabei handelt es sich um einen Desktop-basierten, visuellen Editor zur TOSCA-konformen Modellierung von Anwendungen und deren Laufzeitumgebung. In unserem Beitrag präsentieren wir die zentralen Konzepte des TOSCA-Standards und Anwendungsbeispiele für das “TOSCA Modelling Tool”.

1 Einleitung

Der Fließtext FDM-Initiativen in den digitalen Geisteswissenschaften umfassen schon lange nicht mehr nur die Archivierung von Artikeln und anderen Publikationen, sondern betreffen zunehmend auch die in Forschungsprojekten erhobenen und verarbeiteten Datenbestände in mannigfaltigen Formaten. Neben solchen statischen Daten, die bspw. als CSV, SQL oder XML vorgehalten werden können, fallen auch immer mehr sog. “lebende Systeme” (vgl. Sahle/Kronenwett 2013) als Forschungsergebnisse an, die neue Herausforderungen hinsichtlich der etablierten Nachhaltigkeitsstrategien mit sich bringen.

Erhalt und Nachnutzung solcher “lebenden Systeme” sind in hohem Maße abhängig von guter Dokumentation, sowohl hinsichtlich ihrer Funktionalität als auch der technischen Betriebsbedingungen. So bezeichnen z.B. Henny und Jettka (2022) in ihrem ”Leitfaden für die nachhaltige Entwicklung von Forschungssoftware” Dokumentation als einen der wichtigsten Aspekte zur Erhöhung der Nachnutzungsperspektive von Forschungsanwendungen. Dabei sollte allerdings nicht nur der Quelltext dokumentiert werden (etwa in Form sog. “docstrings”), sondern es sollten darüber hinaus niedrigschwellige Anleitungen der Arbeitsabläufe und Betriebsbedingungen erzeugt und bereitgestellt werden. Darüber hinaus erleichtert solch eine umfassende Dokumentation auch die Referenzierbarkeit und Übertragbarkeit des Dokumentierten, was letztendlich auch die Nachhaltigkeit fördert.

2 Das “TOSCA Modelling Tool”

In unserem Beitrag stellen wir das “TOSCA Modelling Tool” (TMT) vor, das am Cologne Center for eHumanities (CCeH) im Rahmen der Flex-Fund Förderung von NFDI4Culture entwickelt wurde[1]. Das TMT ist ein niederschwelliges Tool zur Modellierung der Struktur von Forschungssoftware und deren Paketierung gemäß des TOSCA-Standards (“Topology and Orchestration Specification for Cloud Applications”, OASIS: 2013, 2020). Grundlage für die Entwicklung des TMT war das OpenTOSCA Ökosystem[2] und das konkrete Entwicklungsziel war es, die grafische Modellierungskomponente des OpenTOSCA Ökosystems (die sog. „Winery“[3], s. Kopp 2013) dahingehend weiterzuentwickeln, dass sie als eigenständige Desktop-Anwendung unabhängig von OpenTOSCA installiert und auf allen gängigen Betriebssystemen (Linux, macOS, Windows) genutzt werden kann.

Der zugrunde liegende TOSCA-Standard definiert eine in XML bzw. YAML implementierte, maschinenlesbare Beschreibung von Anwendungskomponenten und deren Beziehungen untereinander und

[1] https://nfdi4culture.de/de/nachrichten/flexfunds-2022-tosca.html
[2] https://www.opentosca.org
[3] https://projects.eclipse.org/projects/soa.winery

ermöglicht somit eine standardisierte und anbieterunabhängige Modellierung, Provisionierung und Bereitstellung von Forschungsanwendungen. Darüber hinaus bietet TOSCA die Möglichkeit diese maschinenlesbare Beschreibung einer Applikation mit Metadaten wie Bedienungsanleitungen, Lizenzen etc. anzureichern und eignet sich daher sehr gut zur langfristigen Dokumentation, Archivierung und Bereitstellung von in den DH erzeugter Forschungssoftware (vgl. Neuefeind et al.: 2018, 2019, 2020). Weiterhin zielt das Forschungsprojekt ReSUS (Reusable Software University Stuttgart)[4] explizit auf die Entwicklung einer Erweiterung des OpenTOSCA Ökosystems ab, mit der sich Forschungssoftware und -daten eindeutig referenzierbar und lizenzrechtlich korrekt ablegen lassen und somit auffindbar und zitierfähig gemacht werden.

3 Von visueller Modellierung zur Community-weiten Nachhaltigkeitsstrategie?

Das "TOSCA Modelling Tool" ermöglicht es Entwickler:innen sowie Betreuer:innen von DH-Anwendungen, die Struktur ihrer Anwendungen in einem grafischen Editor auf Basis von und konform zu dem TOSCA-Standard zu modellieren und diese anschließend in Form eines standardisierten Paketierungsformats (sog. CSAR, für "Cloud Service Archive") zu exportieren. Die so erzeugten CSARs enthalten, ähnlich wie ein Docker-Image, alle für den Betrieb der paketierten Anwendung nötigen Komponenten (z. B. als fertiges Kompilat oder unter Angabe von Paketquellen). Darüber hinaus enthalten CSARs auch eine exakte Spezifikation der Abhängigkeiten zwischen den einzelnen Komponenten einer Anwendungsstruktur – eine Art "Beipackzettel", vergleichbar zu Docker-Compose-Stacks.

Die exportierten CSARs können bspw. mit einer DOI versehen (z. B. auf Zenodo) publiziert werden, was die Zitation und Referenzierbarkeit der enthaltenen Anwendungen jenseits einfacher Verweise auf den Quelltext ermöglicht. Perspektivisch können die CSARs zudem – eine korrekte und vollständige Modellierung vorausgesetzt – von jeder TOSCA-fähigen Laufzeitumgebung interpretiert und als lauffähige und damit direkt (nach-)nutzbare Instanz bereitgestellt werden.

Mit dem "TOSCA Modelling Tool" als sehr einfach nutzbare Desktop-Anwendung hoffen wir, die technischen Anforderungen an (DH-)

[4] https://www.iaas.uni-stuttgart.de/en/projects/resus/

Forscher:innen für die TOSCA-konforme Modellierung von Forschungssoftware erheblich gesenkt zu haben. Weiterhin erachten wir das TMT als einen weiteren Schritt hin zu einem nachhaltigeren Umgang mit Forschungssoftware. Unsere Dissemination des TMT soll in erster Linie dazu beitragen, TOSCA als Dokumentationsstandard in den DH einzuführen bzw. zu festigen. Perspektivisch lässt sich das Vorhaben mit der Bereitstellung dezentraler, TOSCA-fähiger Runtimes als Community-weite Strategie zur Förderung der Nachhaltigkeit und Referenzierbarkeit von Forschungsanwendungen fortführen.

4 Förderung und Ressourcen

Die Entwicklung des “TOSCA Modelling Tools” erfolgte quelloffen erfolgte quelloffen im Rahmen der Flex-Funds-Förderung des NFDI-Konsortiums NFDI4Culture und ist unter folgendem Github-Repositorium nachvollziehbar: https://github.com/olvidalo/desktop-winery. Für die Bereitstellung der jeweils aktuellsten Version des "TOSCA Modelling Tools" wurde zudem eine eigene Projektwebsite eingerichtet: https://olvidalo.github.io/desktop-winery.

Bibliografie

Henny, Ulrike; Jettka, Daniel (2022): ”Leitfaden für die nachhaltige Entwicklung von Forschungssoftware”. NFDI4Culture Handreichung. Online verfügbar unter https://nfdi4culture.de/go/E3332. Zuletzt abgerufen am 15.05.2023

Kopp, Oliver; Binz, Tobias; Breitenbücher, Uwe and Leymann, Frank (2013). "Winery – A Modeling Tool for TOSCA-Based Cloud Applications." In: Service-Oriented Computing. ICSOC 2013. Lecture Notes in Computer Science, vol 8274. Springer, Berlin, Heidelberg. DOI: 10.1007/978-3-642-45005-1_64.

Neuefeind, Claes; Mathiak, Brigitte; Schildkamp, Philip; Karadkar, Unmil; Stigler, Johannes; Steiner, Elisabeth; Vasold, Gunter; Tosques, Fabio; Ciula, Arianna; Maher, Brian; Newton, Greg; Arneil, Stewart; Holmes, Martin (2020). “Sustainability Strategies for Digital Humanities Systems,” in: Book of Abstracts of the Digital Humanities Conference 2020, University of Ottawa und Carleton University: Alliance of Digital Humanities Organizations (ADHO).

Neuefeind, Claes; Schildkamp, Philip; Mathiak, Brigitte; Marčić, Aleksander; Hentschel, Frank; Harzenetter, Lukas; Breitenbücher, Uwe; Barzen, Johanna; Leymann, Frank. (2019). "Sustaining the Musical Competitions Database: a TOSCA-based Approach to Application Preservation in the Digital Humanities." In: Book of Abstracts of the Digital Humanities Conference 2019. Utrecht University: Alliance of Digital Humanities Organizations (ADHO).

Neuefeind, Claes; Harzenetter, Lukas; Schildkamp, Philip; Breitenbücher, Uwe; Mathiak, Brigitte; Barzen, Johanna; Leymann, Frank (2018). "The SustainLife Project – Living Systems in Digital Humanities". In: Proceedings of the 12th Advanced Summer School on Service-Oriented Computing (IBM Research Report RC25681), 101-112.

OASIS (2013). “Topology and Orchestration Specification for Cloud Applications Version 1.0”. URL: http://docs.oasis-open.org/tosca/TOSCA/v1.0/TOSCA-v1.0.html.

OASIS (2020). “TOSCA Simple Profile in YAML”, Version 1.3. URL: https://docs.oasis-open.org/tosca/TOSCA-Simple-Profile-YAML/v1.3/TOSCA-Simple-Profile-YAML-v1.3.html.

Sahle, P. und Kronenwett, S. (2013): ”Jenseits der Daten: Überlegungen zu Datenzentren für die Geisteswissenschaften am Beispiel des Kölner ’Data Center for the Humanities’“. In: LIBREAS. Library Ideas 23, S. 76–96.

New Ways of Creating Research Data

Conversion of Unstructured Text to TEI XML using GPT on the Correspondence of Hugo Schuchard with a Web Prototype for Prompt Engineering

Pollin, Christopher
christopher.pollin[at]uni-graz.at
Zentrum für Informationsmodellierung, Österreich
ORCID-iD: 0000-0002-4879-129X

Steiner, Christian
christian.steiner[at]dhcraft.org
Digital Humanities Craft OG, Österreich
ORCID-iD: 0000-0002-6658-4622

Zach, Constantin
constantin.zach[at]proton.me
Independent Software Developer

Summary. This paper explores the use of prompt engineering to streamline the creation of humanities research data by converting unstructured correspondence texts into the TEI XML format. The approach optimizes language models such as GPT to produce accurate structured data while preserving context. The paper discusses the iterative refinement of the conversion process, challenges, and potential solutions, and presents a user-friendly web prototype. Overall, prompt engineering shows potential for improving the efficiency of research data creation in the humanities.

1 The Role of Text Conversion in Humanities Research

A crucial aspect of creating research data in the humanities is the efficient conversion of unstructured text into structured formats such as Text Encoding Initiative (TEI) XML, which facilitates the analysis, preservation, and dissemination of historical and scholarly documents. Correspondence is a valuable resource for humanities research, and its conversion to TEI XML is an essential step in the creation of digital editions and research data. This paper focuses on the use of prompt engineering to streamline the process of converting unstructured correspondence texts into TEI XML, thereby increasing the efficiency of research data creation in the humanities.

2 Prompt Engineering and Text Conversion in Humanities Research

Prompt engineering, an approach to fine-tuning and optimizing large-scale language models such as GPT, has shown promise in various natural language processing tasks. By designing effective prompts, it is possible to guide these models to perform specific tasks with improved accuracy and precision. In the context of converting unstructured correspondence text to TEI XML, prompt engineering can be used to generate highly accurate structured data that retains the semantic and contextual information of the original text.

This paper describes the workflow for converting unstructured correspondence text to TEI XML using prompt engineering techniques. We will begin by analyzing the structure of the correspondence and identifying various textual and semantic elements such as page beginnings, footnotes, paragraphs and named entities. This analysis will inform the development of a set of prompts that will guide the GPT model through the conversion process, transforming the unstructured text into a structured TEI XML format.

As part of our investigation, we will explore the iterative nature of prompt engineering and its role in refining the conversion process. By examining the TEI XML generated by GPT, we will identify discrepancies and areas for improvement, leading to the modification and fine-tuning of prompts. This iterative approach will ensure the generation of more accurate TEI XML data, contributing to the overall efficiency of research data creation in the humanities, particularly for digital editions.

The paper also discusses the challenges and limitations of using GPT models for this task, including the need for domain-specific training and potential problems with consistency and correctness. We will also explore potential solutions and troubleshooting techniques to overcome these limitations and improve the conversion process. One possibility is the integration of vector databases.

In addition, we will present a web prototype that supports prompt engineering, providing researchers with a user-friendly interface for iteratively generating and refining prompts. This web-based tool will enable humanities researchers to effectively harness the power of GPT models to enable more streamlined and efficient conversion of unstructured text into TEI XML.

3 Practical Application: An Example of Prompt-Engineered TEI XML Conversion

To illustrate the practical application of our approach, we will include an example of unstructured text, its query, and the resulting TEI XML data. This example will demonstrate the potential of prompt engineering and the web prototype to facilitate the creation of high-quality research data in the humanities.

|1|
[1]Wien 14/2 79
IV Victorg. 5[2]

Auf Ihre freundlichen Zeilen vom 8. d. M., die mir erst am 11. zukamen, verehrter Herr College, würde ich gern umgehend geantwortet haben, wenn sich nicht gerade in diesen Tagen sehr Viel gehäuft hätte. Leider bin ich ausser Stande das befremdliche Verhalten aufzuklären, das Ihnen einen so unwillkommenen Eindruck mit allem Recht macht. Ich kenne E. Hübner[3] fast gar nicht, bin überhaupt nur zwei Mal mit ihm persönlich zusammengekommen, sonst nur aber auch das selten in brieflichem Verkehr gewesen aus geschäftlichem Anlass.[4] Aber ich sollte meinen dass ein objectives offenes Wort den Zwischenfall von Ihrer Seite ohne Weiteres begleichen müsste.

Ein Missbehagen wie Sie es schildern |2| ist mir leider aus eigener Erfahrung nur zu bekannt. Möchte es sich Ihnen rasch beheben. Munterkeit brauchen wir zu unserm Beruf so gar mehr als viele andere.

Mit hochachtungsvollem Gruss

Ihr ergebener

Bdf

Abb. 1. Example of an input

```xml
<div subtype="original" type="letter" xml:id="L.7711" xml:lang="de">
    <pb facs="#IMG.1" n="1"/>
    <dateline> <note n="1" place="foot" type="editorial">Antwortbrief auf
        01-HSFB.</note>Wien 14/2 79<lb/> IV Victorg. 5<note n="2" place="foot"
        type="editorial">Benndorf wohnte seit seinem Wechsel von Prag nach Wien in der
        Victorgasse 5 in Wien-Wieden, einem Stadtteil, die bei Universitätsprofessoren
        sehr beliebt war. In dieser Gegend hatte schon sein Vorgänger am Wiener
        Lehrstuhl für Klassische Archäologie, <persName>Alexander Conze</persName>,
        gewohnt, und auch sein Kollege am Archaeologisch-epigraphischen Seminar,
        <persName>Otto Hirschfeld</persName>, hatte in Wieden sein Zuhause
        gefunden.</note> </dateline>
    <p>Auf Ihre freundlichen Zeilen vom 8. d. M., die mir erst am 11. zukamen, verehrter
        Herr College, würde ich gern umgehend geantwortet haben, wenn sich nicht gerade
        in diesen Tagen sehr Viel gehäuft hätte. Leider bin ich ausser Stande das
        befremdliche Verhalten aufzuklären, das Ihnen einen so unwillkommenen Eindruck
        mit allem Recht macht. Ich kenne <persName>E. Hübner</persName> <note n="3"
        place="foot" type="editorial"> <persName>Emil Hübner</persName> (1834-1901),
        klassischer Philologe und Epigraphiker, seit 1863 Professor für klassische
        Philologie an der Friedrich-Wilhelms-Universität Berlin </note> fast gar nicht,
        bin überhaupt nur zwei Mal mit ihm persönlich zusammengekommen, sonst nur aber
        auch das selten in brieflichem Verkehr gewesen aus geschäftlichem Anlass.<note
        n="4" place="foot" type="editorial">Mit <persName>Emil Hübner</persName> stand
        Benndorf von 1868 bis 1895 in brieflicher Korrespondenz (16 Schreiben), vgl.
        Österreichische Nationalbibliothek, Sammlung von Handschriften und alten
        Drucken, Autogr. 646/36</note> Aber ich sollte meinen dass ein objectives
        offenes Wort den Zwischenfall von Ihrer Seite ohne Weiteres begleichen
        müsste.</p>
    <p>Ein Missbehagen wie Sie es schildern <pb facs="#IMG.2" n="2"/> ist mir leider aus
        eigener Erfahrung nur zu bekannt. Möchte es sich Ihnen rasch beheben. Munterkeit
        brauchen wir zu unserm Beruf so gar mehr als viele andere.</p>
    <closer> <salute>Mit hochachtungsvollem Gruss</salute> <lb/> <salute>Ihr
        ergebener</salute> <lb/> <signed> <persName>Bdf</persName> </signed> </closer>
</div>
```
Abb. 2. Example of the corresponding result after iterative prompt engineering

In conclusion, this paper demonstrates the potential of prompt engineering as a potential tool for improving the efficiency of research data creation in the humanities, particularly in the context of converting unstructured correspondence texts into TEI XML. By harnessing the power of large-scale language models and refining their output through an iterative, prompt-driven process, we can streamline the conversion process, save time and resources, and facilitate more robust and accurate research data in the humanities.

Bibliografie

Ekin, Sabit. ‘Prompt Engineering For ChatGPT: A Quick Guide To Techniques, Tips, And Best Practices’. TechRxiv, 29 April 2023. https://doi.org/10.36227/techrxiv.22683919.v1.

Fries, Jason, Leon Weber, Natasha Seelam, Gabriel Altay, Debajyoti Datta, Samuele Garda, Sunny Kang, et al. ‘BigBio: A Framework for Data-Centric Biomedical Natural Language Processing’. Advances in Neural Information Processing Systems 35 (6 December 2022): 25792–806.

Tian, Haoye, Weiqi Lu, Tsz On Li, Xunzhu Tang, Shing-Chi Cheung, Jacques Klein, and Tegawendé F. Bissyandé. 'Is ChatGPT the Ultimate Programming Assistant -- How Far Is It?' arXiv, 24 April 2023. https://doi.org/10.48550/arXiv.2304.11938.

'The TEI Guidelines'. Accessed 3 May 2023. https://tei-c.org/release/doc/tei-p5-doc/en/html/index.html.

Watkins, Ryan. 'Guidance for Researchers and Peer-Reviewers on the Ethical Use of Large Language Models (LLM) in Scientific Research Workflows'. OSF Preprints, 21 April 2023. https://doi.org/10.31219/osf.io/6uh8p.

Wilkinson, Mark D, Michel Dumontier, IJsbrand Jan Aalbersberg, Gabrielle Appleton, Myles Axton, Arie Baak, Niklas Blomberg, et al. 'The FAIR Guiding Principles for Scientific Data Management and Stewardship'. Scientific Data 3, no. 1 (2016): 1–9.

Windsor, Brad, and Kevin Choi. 'Thistle: A Vector Database in Rust'. arXiv, 25 March 2023. https://doi.org/10.48550/arXiv.2303.16780.

Verzerrte Geschichte durch ungleiche Erschließung?

Eine Untersuchung zum Recording Bias in Münzhortdatenbanken

Rademacher, Philip
mail[at]philip-rademacher.de
Universität Wuppertal, Deutschland

Zusammenfassung. In der *„Coin Hoards of the Roman Empire“* Datenbank der Universität Oxford sind 14.740 Münzhorte aus der römischen Kaiserzeit verzeichnet. Immer mehr Beiträge nutzen diese umfangreiche Datenbasis zur Beantwortung quantitativer Forschungsfragen. Doch wie repräsentativ sind solche Daten, um die Antike zu erforschen? Dieser Beitrag analysiert die Repräsentativität exemplarisch für die Münzdatenverfügbarkeit innerhalb der Datenbank. Sind Hortfunde aus der Nähe antiker Städte oder Militäreinrichtungen häufiger bis auf Münzebene erschlossen als ländliche Funde? Sind Münzhorte aus bestimmten Jahrhunderten häufiger auf Münzebene erschlossen als andere? Als Methode wird ein logistisches Klassifikationsmodell verwendet, welches auch im maschinellen Lernen eingesetzt wird, um Zusammenhänge aufzuzeigen.

1 Problemstellung

Antike Münzhorte stellen eine Quelle dar, deren Bedeutung für Alltags- und Wirtschaftsgeschichte aufgrund der mangelhaften literarischen Überlieferung dieser Themen zunimmt. Einige Forschungsbeiträge haben versucht, mit Hortdaten Antworten auf ökonomische, gesellschaftliche oder kulturelle Fragen zu finden.[1] Münzhorte aus der römischen Kaiserzeit eignen sich aufgrund ihrer zahlenmäßig hohen Erfassung in der „Coin Hoards of the Roman Empire“-Datenbank (CHRE) optimal für solche quantitativen Fragestellungen.[2]

---

[1] Auswahl an datenbasierten Beiträgen zu den Themenfeldern Geldproduktion: Thordeman (1948), Crawford (1974), Buttrey (1993), van Heesch (2011); Geldzirkulation: Creighton (1992), Swan (2020); Krieg und Unruhe: Gazdac (2012), de Callatay (2017), Turchin & Scheidel (2009) und Deponierungspraktiken: Sycamore (2019), Yates & Bradley (2010).

[2] In Mairat, Wilson & Howgego (2022) und Bland et al. (2020) werden die Daten aus der CHRE-Datenbank für unterschiedliche Forschungsfragen genutzt.

Damit die Analyse solcher Daten ein Bild von der historischen Wirklichkeit zeichnen kann, müssen die Daten – statistisch formuliert – repräsentativ sein. Münzhorte in der Datenbank sind jedoch keine zufällig entstandene Teilmenge der ursprünglichen Gesamtheit aller angelegten Münzhorte einer Zeit. Es hängt von anderen Faktoren als vom Zufall ab, welche Horte erhalten blieben, archäologisch gesichtet, publiziert oder in einer Datenbank erschlossen wurden.[3]

Neben der ungleichen Überlieferung unterscheiden sich die Münzhorte auch in ihrem Erschließungszustand. Angaben zu den enthaltenen Münzen sind nicht durchgehend verfügbar. Quantitative Forschungsfragen, die auf die Verfügbarkeit dieser Informationen angewiesen sind, können dadurch nicht alle Münzhortfunde berücksichtigen. Wenn der Erschließungszustand eines Münzhortes nicht vom Zufall, sondern von bestimmten Hortmerkmalen abhängt, beeinflusst dies die Repräsentativität der Forschungsdaten. Ein solcher Recording Bias führt als statistisches Problem dazu, dass die Verteilung der Hortmerkmale in den Forschungsdaten gegenüber der Grundgesamtheit verzerrt ist.

2 Methodischer Ansatz

Ob die Verfügbarkeit von Münzdaten in der CHRE-Datenbank mit bestimmten Hortmerkmalen zusammenhängt, wird in zehn Ländern mittels eines logistischen Klassifikationsmodells analysiert.[4] Dabei werden folgende Fragen untersucht: Sind Hortfunde aus antiken Städten oder Militäreinrichtungen häufiger bis auf Münzebene erschlossen als ländliche Funde? Sind Münzhorte mit bestimmten Nominalen oder Horte eines bestimmten Jahrhunderts häufiger auf Münzebene erschlossen als andere?

Die angewandte Methodik ist nur ein erster Schritt, um das Problem der Datenverzerrungen anzugehen. Sie betrachtet den Recording Bias isoliert von allen vorherigen Verzerrungsstufen. Formal wird eine kleinere Teilmenge (Münzhorte mit Münzdaten) mit einer anderen

---

[3] Einen Überblick über die verschiedenen Verzerrungsursachen in archäologischen Datenbanken bietet Robbins (2013). Zur Diskussion der Repräsentativität von Hortdaten vgl. auch Crawford (1969).

[4] Als Untersuchungsgebiete wurden Bulgarien, Frankreich, Deutschland, Israel, Italien, die Niederlande, Portugal, Rumänien, Spanien sowie das Vereinigte Königreich ausgewählt. Diese Länder haben mehr als 100 Hortfunde in der Datenbank und darüber hinaus ein ausgeglichenes Verhältnis von Hortfunden mit und ohne Münzdaten.

größeren Teilmenge (Münzhorte in der Datenbank) verglichen und überprüft, ob die kleinere Teilmenge für die größere Teilmenge repräsentativ ist. Dabei bleibt unklar, ob die größere Teilmenge überhaupt repräsentativ für ihre ursprüngliche Grundgesamtheit (alle gebildeten Münzhorte) ist.

3 Zentrale Ergebnisse

In den Niederlanden sind Münzhorte außerhalb der römischen Provinzgrenzen deutlich häufiger auf Münzebene erschlossen (Abb. 1), während in Deutschland ein gegenteiliger Effekt feststellbar ist. Dort und in Israel sind Münzhorte aus militärischen Umgebungen häufiger erschlossen als ländliche Münzhorte. In Portugal, Rumänien und Portugal gilt dies für städtische Münzhorte.

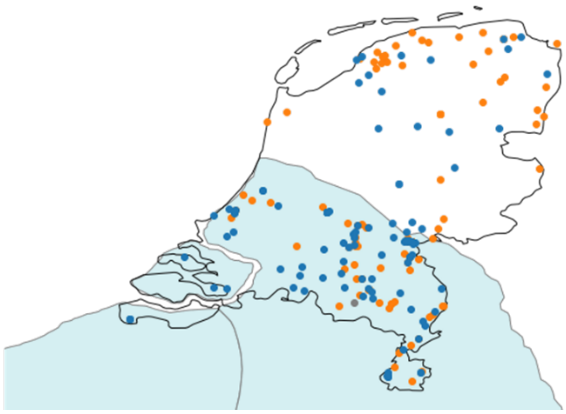


Abb. 1. Im Norden der Niederlande sind prozentual mehr Horte auf Münzebene erschlossen (orange) als im Süden, wo mehr Horte keine Münzdaten haben (blau). Die hellblaue Fläche kennzeichnet die Ausdehnung des römischen Kaiserreiches.

Die Münzdatenverfügbarkeit korreliert zudem häufig mit dem Deponierungszeitpunkt: In den Niederlanden, Portugal, Spanien und dem Vereinigten Königreich sind Münzhorte, die im 1. Jahrhundert gebildet wurden, gegenüber jüngeren Horten überrepräsentiert. Nur in Deutschland und Israel dominieren andere Jahrhunderte.

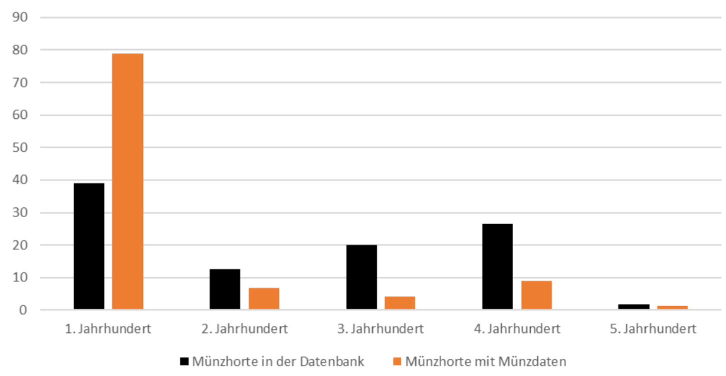


Abb. 2. Im Vereinigten Königreich haben Horte aus dem 1. Jahrhundert unter den auf Münzebene erschlossenen Horten einen deutlich höheren Anteil als in der Datenbank insgesamt.

Starke Verzerrungen zeigen sich im Vereinigten Königreich, wo die Datenerfassung noch nicht abgeschlossen ist: Mit den ausgewählten Hortmerkmalen konnte das Modell die Münzdatenverfügbarkeit in 85% der Fälle – und über beide Klassen hinweg – korrekt klassifizieren. Horte aus militärischen Umgebungen, aus den Provinzen und aus anderen als dem ersten Jahrhundert sind stark unterrepräsentiert (Abb. 2).

Die Münzdatenverfügbarkeit kann daher trotz einer hohen Anzahl an Horten (n = 1.335) und einer Münzdatenverfügbarkeit von 42,56% hinsichtlich der Merkmalsverteilung verzerrt sein kann.

4 Diskussion

Nicht nur die ungleiche Überlieferung von Quellen mit bestimmten Merkmalen kann unser Geschichtsbild verzerren, sondern auch die ungleichmäßige Verfügbarkeit bzw. Erschließung von bestimmten Merkmalen (wie hier die Münzdatenverfügbarkeit bei Horten). Auch große Datenbanken, deren Anspruch darin liegt, systematisch Informationen zusammenzutragen, können hinsichtlich ihrer Merkmalsverfügbarkeit verzerrt sein. Die Ergebnisse regen an, dass der Frage nach der Repräsentativität von Quellenmaterial in der fachwissenschaftlichen Methodendiskussion ein höherer Stellenwert zu kommen sollte. Verzerrungen in Forschungsdaten sollten nicht nur wissenschaftstheoretisch diskutiert, sondern wenn möglich, datenbasiert analysiert werden. Logistische Klassifikationsmodelle sind dazu ein geeignetes Handwerkzeug. Je transparenter der Datenentstehungsprozess anhand der Daten nachvollzogen werden kann, umso mehr wird – besonders beim Einsatz künstlicher Intelligenz – eine unverzerrte Datenanalyse möglich.

Bibliografie

Bland, Roger, Chadwick, Adrian, Ghey, Eleanor, Haselgrove, Colin, Mattingly, David J., Rogers, Adam & Jeremy Taylor. *Iron age and Roman coin hoards in Britain*. Oxford: Oxbow Books, 2020.

Buttrey, Theodore. "THE PRESIDENT'S ADDRESS: Calculating Ancient Coin Production: Facts and Fantasies." *The Numismatic Chronicle* 153 (1993): 335-351.

Crawford, Michael. "Coin hoards and the pattern of violence in the late Republic." *Papers of the British School at Rome* 37 (1969): 76-81.

Crawford, Michael. *Roman Republican Coinage*. Cambridge: Cambridge University Press, 1974.

Creighton, John. *The Circulation of Money in Roman Britain from the First to Third century*. Dissertation, Universität Durham, 1992.

De Callatay, Francois. "Coin deposits and civil wars in a long-term perspective (c. 400 BC–1950 AD)." *The Numismatic Chronicle* 177 (2017): 313-338.

Găzdac, Cristian. "'War and peace'! Patterns of violence through coin hoards distribution. The Middle and Lower Danube from Trajan to Aurelianus." *Istros* 18.1 (2012): 165-198.

Mairat, Jerome, Andrew Wilson, & Chris Howgego (Hg.). *Coin Hoards and Hoarding in the Roman World* (Oxford: Oxford University Press, 2022).

Robbins, Katherine. "Balancing the Scales: Exploring the Variable Effects of Collection Bias on Data Collected by the Portable Antiquities Scheme". *Landscapes*, Vol. 14, Nr. 1 (2013): S. 54-72.

Swan, David. *Cross-Channel Hoarding in the Late Iron Age and Early Roman Periods (200 BC to AD 43).* Dissertation, Universität Warwick, 2020.

Sycamore, Rachael. *Beyond the Objects: Landscape, Spatiality and Romano-British Metalwork Hoarding.* Dissertation, Universität Leicester, 2019.

Thordeman, Bengt. "The Lohe hoard: a contribution to the methodology of numismatics." *The Numismatic Chronicle and Journal of the Royal Numismatic Society* 8.3/4 (1948): 188-204.

Turchin, Peter, & Walter Scheidel. "Coin hoards speak of population declines in Ancient Rome”. *Proceedings of the National Academy of Sciences* 106.41 (2009): 17276-17279.

van Heesch, Johan. Quantifying Roman Imperial Coinage. In:. *Quantifying Monetary Supplies in Greco-Roman Times*, von de Callataÿ, Francois (Hg.), Bari: Edipuliga, 2011.

Yates, David & Richard Bradley. “Still water, hidden depths: the deposition of Bronze Age metalwork in the English Fenland”, *Antiquity*, Vol. 84, Nr. 324 (2010): S. 405-415.

Organisation bestimmt Technik: Persistenz und Veränderung in Infrastrukturen zur langfristigen Sicherung von Forschungsdaten

Schiller-Stoff, Sebastian
sebastian.stoff[at]uni-graz.at
Universität Graz, Österreich
ORCID-iD: 0000-0001-6941-113X

Vasold, Gunter
gunter.vasold[at]uni-graz.at
Universität Graz, Österreich
ORCID-iD: 0009-0001-1636-316X

Steiner, Elisabeth
elisabeth.steiner[at]uni-graz.at
Universität Graz, Österreich
ORCID-iD: 0000-0001-9116-0402

Zusammenfassung. (Digitale) Forschungsdaten nachhaltig und wiederverwendbar zu verwalten, zu archivieren und zur Verfügung zu stellen ist eine wesentliche Herausforderung aktueller Forschung. Der Vortrag stellt die zentrale Frage, ob und wie Forschungsinfrastrukturen im akademischen Kontext im Sinne der Softwarearchitektur nachhaltig entworfen und entwickelt werden können. Durch das tiefere Verständnis und die Bewusstmachung von organisatorischen Einflussgrößen soll die Qualität und Persistenz von technischen Lösungen und ihre langfristige Wartbarkeit verbessert werden. Gerade im Bereich der Langzeitverfügbarkeit kann plakativ auf den Punkt gebracht werden: Personelle und finanzielle Persistenz in einer angemessenen Organisationstruktur führen zu technischer Persistenz.

1 Einleitung

(Digitale) Forschungsdaten nachhaltig und wiederverwendbar zu verwalten, zu archivieren und bereitzustellen, ist eine wichtige Anforderung aktueller Forschung. Die Etablierung entsprechender Infrastrukturen in Form von vertrauenswürdigen digitalen Repositorien schreitet voran. Viele technische Fragen scheinen gelöst, organisatorischen Einflussgrößen wird aber meist zu wenig Beachtung geschenkt. Der vorliegende Beitrag möchte durch Bewusstmachung

dieser Faktoren das Verständnis ihres Einflusses auf Qualität und Persistenz technischer Lösungen verbessern.
Die Softwarearchitektur kennt das Konzept der „Technischen Schulden“. Diese werden als Resultat von im Lauf der Zeit getroffenen, suboptimalen Entscheidungen verstanden. Wir versuchen aufzuzeigen, dass Technische Schulden durchaus auch organisatorisch bedingt sein können.

2 Entwicklung von wartbaren Infrastrukturen im akademischen Kontext

2.1 Thesen

Der Vortrag stellt die zentrale Frage, ob und wie Forschungsinfrastrukturen im akademischen Kontext aus Sicht der Softwarearchitektur nachhaltig entworfen bzw. Technische Schulden dauerhaft minimiert werden können. Dazu werden drei Behauptungen aufgestellt:

1. **Jede (benutzte) Software muss geändert werden**, und zwar nicht nur aufgrund von sich verändernden technologischen Bedingungen (z.B. Sicherheitsrisiken), sondern auch aufgrund sich wandelnder Annahmen und Erwartungshaltungen. Software ist damit auch ein soziales Konstrukt und kein einmaliges Produkt. Wird die Änderungsfähigkeit nicht von Anfang an in der Systemarchitektur berücksichtigt, kann der dauerhafte Betrieb nicht gelingen. Dies betrifft Infrastrukturen zum Zwecke der Langzeitsicherung und -bereitstellung in besonderem Maße.
2. **Jede (benutzte) Software degradiert über Zeit**. Software besitzt den Hang zur Degradierung, was dem nachhaltigen Betrieb eines Systems entgegensteht. Werden zwangsläufige Änderungen am System nicht dauerhaft kontrolliert und systematisiert, droht der Verlust der Änderungsfähigkeit. Technische Schulden entstehen immer und lassen sich nicht vollständig vermeiden - sondern nur minimieren.
3. **Jede (benutzte) Software ist früher oder später entlang sozialer Kommunikationsnetzwerke strukturiert**. Laut Conway’s Law entspricht jede Software in ihrer Struktur (bzw. Architektur) früher oder später der Struktur der arbeitsteiligen Prozesse des Unternehmens bzw. der Organisation. Änderungen der Organisationsstruktur wirken sich über Zeit somit auch auf die Struktur der entwickelten Infrastruktur aus.

Technische Schulden sind entgegen der Bezeichnung oft eher ein organisatorisches als rein technisches Problem.

2.2 Änderungsdruck

Veränderte Organisationsstrukturen erzeugen technischen Änderungsbedarf. Die geschieht manchmal abrupt, häufiger jedoch schleichend durch kleine organisatorische und personelle Änderungen. Werden diese zu lange ignoriert, entspricht die Infrastruktur nicht mehr den Anforderungen. Sie wird zunehmend umgangen oder auf unvorhergesehene Weise genutzt. Beides hat längerfristig negative Auswirkungen auf die Persistenz.

Wie geht man nun damit um? Aus unserer Erfahrung sollten Änderungen stetig, aber nie ad hoc erfolgen, sondern sorgfältig unter Einbeziehung von Forschenden und technischem Personal diskutiert, geplant und auf negative Auswirkungen untersucht werden. Entscheidend ist auch hier die zugrunde liegende Architektur: Ist ein System entsprechend geplant, kann flexibler auf solche Änderungen reagiert werden.

2.3 Personal

Das technische Personal ist nicht nur Teil der Gesamtorganisation, sondern stellt in sich eine Organisation dar. Da diese Gruppe meist klein ist, hängt viel an einzelnen Mitarbeiter:innen. Gerade deshalb ist die Auswahl der Mitglieder und das Miteinander zentral. Technische Entscheidungen sollten in der Gruppe getroffen werden. Das erhöht nicht nur den Bus-Faktor, die Einbindung jüngerer Mitglieder verhindert auch Betriebsblindheit. Umgekehrt kann die Erfahrung älterer Kolleg:innen als Korrektiv dienen.

Flache Hierarchien, offene Kommunikation und ein Grundkonsens über die Softwarearchitektur haben sich bewährt. „Flickwerk“ im Sinne spontaner Lösungen führen unserer Erfahrung nach rasch zu Unzufriedenheit, Zuständigkeitsverweigerung und Konflikten im Team.

Nicht vergessen werden dürfen auch die anderen Stakeholder, also das institutionelle Umfeld. Wenn wir von der Prämisse ausgehen, dass Organisation Technik bestimmt, muss hier hoher Wert auf Austausch gelegt werden, nicht nur um die sich verändernden Anforderungen zu verstehen und zu bedienen, sondern auch weil dadurch finanzielle und organisatorische Möglichkeiten des technischen Teams bestimmt werden.

3 Zusammenfassung und Ausblick

Zusammenfassend lässt sich also festhalten: Technik folgt Organisation. Nur wenn eine Infrastruktur in der Lage ist, organisatorische Einflussgrößen zu berücksichtigen, wird diese Bestand haben. Hier zeigt sich, dass die Anpassbarkeit an geänderte Ansprüche wesentlich ist. Ohne organisatorische und technische Änderungsfähigkeit kann längerfristig nicht angemessen auf den zwangsläufigen Anstieg Technischer Schulden reagiert werden.

Nachnutzbarkeit und Erweiterungsfähigkeit sind wichtig, ergeben sich aber als Nebeneffekt, wenn die Architektur auf die Anpassungsfähigkeit ausgerichtet ist. Der zunächst höhere Planungs- und Implementierungsaufwand macht sich rasch bezahlt, wenn organisatorisch bedingte Anpassungen nötig werden.

Organisation bestimmt Technik aber auch in dem Sinne, dass personelle und finanzielle Ausstattung selbstverständlich die Infrastruktur maßgeblich beeinflusst und dies in einem gesunden Verhältnis zur Organisationsstruktur stehen muss. Gerade im Bereich der Langzeitverfügbarkeit kann plakativ auf den Punkt gebracht werden: Personelle und finanzielle Persistenz in einer angemessenen Organisationstruktur führen zu technischer Persistenz.

Bibliografie

Conway, Melvin. “How do committees invent”. *Datamation* (1968), 28–31.

IEEE Computer Society. *Standard Glossary of Software Engineering Terminology*, 610.12-1990.

Dowalil, Herbert. *Modulare Softwarearchitektur: Nachhaltiger Entwurf durch Microservices, Modulithen und SOA 2.0.* München 2020.

Lilienthal, Carola. *Langlebige Software-Architekturen: Technische Schulden analysieren, begrenzen und abbauen.* Heidelberg 2020.

Nachiappan Nagappan, Brendan Murphy, and Victor Basili. “The influence of organizational structure on software quality: an empirical case study”. In *Proceedings of the 30th international conference on Software engineering (ICSE '08).* Association for Computing Machinery, New York, NY, USA, 2020, 521–530. https://doi.org/10.1145/1368088.1368160.

Neuefeind, Claes et al. “Sustainability Strategies for Digital Humanities Systems”, In *DHN 2020: Digital Humanities*; 2020, 530-537, DOI: 10.17613/NCDT-DN37.

Richards, Mark, and Neal Ford. *Handbuch moderner Softwarearchitektur: Architekturstile, Patterns und Best Practices*; Heidelberg 2021.

Smithies, James, and C. Sichani, and P. Mellen, and A. Ciula. “Managing 100 Digital Humanities Projects: Digital scholarship and archiving in King’s Digital Lab”. *DHQ - Digital Humanities Quarterly*, 13, 1, 2019. URL: http://www.digitalhumanities.org/dhq/vol/13/1/000411/000411.html.

Starke, Gernot et al. *Basiswissen für Softwarearchitekten: Aus- und Weiterbildung nach iSAQB-Standard zum Certified Professional for Software Architecture*.

Starke, Gernot. *Effektive Softwarearchitekturen: Ein praktischer Leitfaden.* München 2020.

The Consultative Committee for Space Data Systems. *Reference Model for an Open Archival Information System (OAIS).* 2012. URL: https://public.ccsds.org/pubs/650x0m2.pdf.

Das Projektende

Zum praktischen Umgang mit Forschungsdaten eines geisteswissenschaftlichen Editionsprojekts

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

Zusammenfassung. Das an der BBAW angesiedelte Akademienvorhaben Corpus Coranicum arbeitet seit 2007 genuin digital an einem mehrere Module umfassenden Portal zum Text des Korans. Da sich das Projekt dem Ende zuneigt, lohnt sich ein Blick auf die Zukunft der während der Laufzeit angesammelten Forschungsdaten. Während sich die Suche nach einem Fachrepositorium vergleichsweise einfach gestaltete, sind Fragen zur Definition und Typisierung der sehr multilingualen Forschungsdaten durch die Datenvielfalt schwieriger zu beantworten. Auch muss bedacht werden, dass nicht alle Forschungsdaten publizierbar sind und dennoch unter FAIR-Bedingungen in der Institution aufgehoben werden sollten. Der Vortrag versucht ausgehend vom praktischen Umgang mit Forschungsdaten im Herbst eines geisteswissenschaftlichen Editionsprojekt theoretische Überlegungen zur Schärfung der Definition von Forschungsdaten im editionswissenschaftlichen Kontext beizutragen.

1 Einleitung

Auch Langzeit-Forschungsprojekte des Akademienprogramms enden. In der langen Projektlaufzeit werden zunehmend Daten gesammelt, ausgewertet und zum Teil publiziert. Trotz fachspezifischer Eigenheiten kristallisieren sich dennoch Prozesse, die sich auf geisteswissenschaftliche Forschungsprojekte im Allgemeinen anwenden lassen. Objekte wie digitalisierte Faksimile, Bibliographien, Transkriptionen, Annotationen et al. werden in den meisten Forschungsprojekten gesammelt und ausgewertet. Wie die im Rahmen der wissenschaftlichen Routinen gesammelten und erstellten Daten auch nach Projektende für die Wissenschaft erhalten werden können, soll im Folgenden am Beispiel des Vorhabens Corpus Coranicum behandelt werden.[1]

[1] (Unsworth 2000)

2 Das Vorhaben

Die Forschungsweise des seit 2007 an der BBAW angesiedelten Akademienvorhabens Corpus Coranicum folgte von Beginn einem digitalen Paradigma. Die Webseite https://www.corpuscoranicum.de repräsentiert mit den dort vier publizierten Kernmodulen (Handschriften, Lesarten, Umwelttext und Kommentar) die primären digitalen Forschungsergebnisse des Vorhabens.

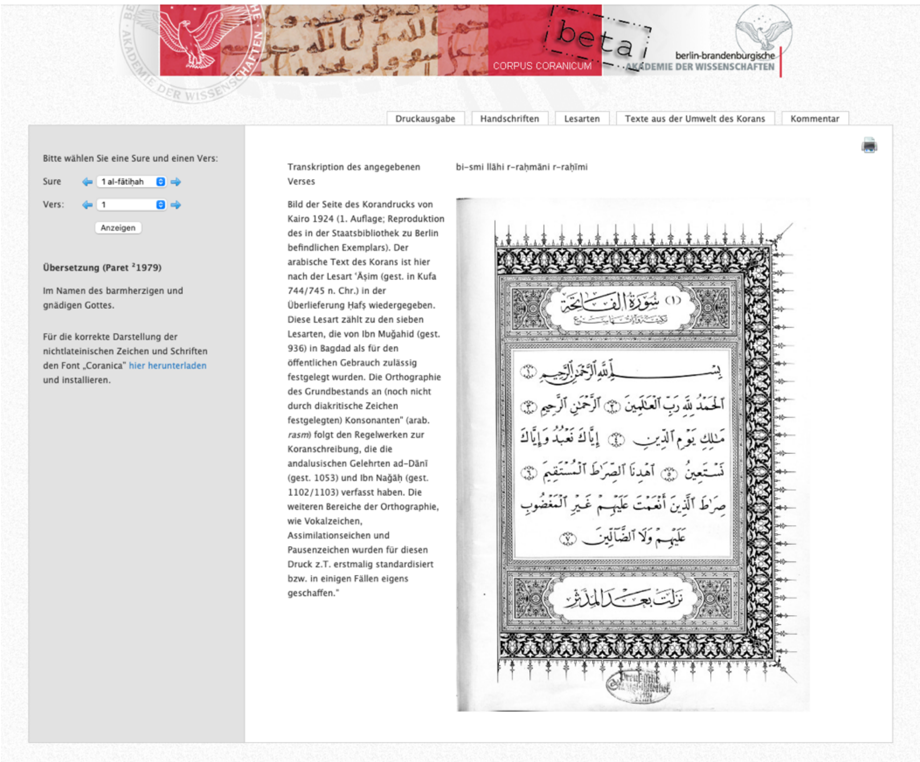


Abb. 1. Die rote Website Corpus Coranicum. Webarchiv vom 3.7.2013.

Während das Modul Handschriften vor allem Metadaten und Transkriptionen und – wo genehmigt – Digitalisate von Koran-Handschriften aus Bibliotheken, Auktionskatalogen und anderen Sammlungen bündelt, halten die Lesarten in SQL-Datenbanken Varianten der mündlichen Überlieferung des Textes transliteriert fest. Die Daten der Umwelttexte bestehen ebenso aus SQL-Tabellen, während Transkriptionstexte TEI-kodiertes XML sind. Dazu gesellen sich in Nebenprojekten C-14-Datierungsanalysen[2], Wachswalzen,

[2] (Marx und Jocham 2019)

TAVO-Karten, Filme und Nachlässe, die analog und digital vorliegen, Publikationen, Konkordanzen, Bibliografien, etc.. Neben dem publizierten Material hat das Vorhaben auch zahlreiche Daten gesammelt, die aus rechtlichen Gründen nicht publiziert werden können, wenn z.B. Nutzungs- oder Urheberrecht oder Persönlichkeitsschutz dem entgegen steht. Datenmanagementpläne für das Vorhaben wurden erstellt.[3]

Abb. 2. Die blaue Website Corpus Coranicum. Webarchiv vom 21.10.2019.

Die Universitätsbibliothek Halle betreibt in dem DFG-geförderten Sondersammelgebiet Naher Osten ein Datenrepositorium für dieses Fachgebiet, weshalb die Suche nach einem geeigneten Repositorium schnell abgeschlossen werden konnte. Es wird angenommen, dass Fachrepositorien in den Disziplinen am ehesten wahrgenommen werden. So wird der bildretrodigitalisierte Nachlass Kellermann in Halle als Pilot abgelegt, perspektivisch sollen dort bei Projektabschluss die publizierbaren Forschungsdaten des Projekts abgelegt werden. Eine Verzeichnung der Forschungsdaten auf dem edoc-Server der BBAW ist

[3] (Sutter und Marciniak 2021)

seit Herbst 2022 möglich und wird dort zur institutionellen Dokumentation vorgenommen, so dass die an der BBAW produzierten Forschungsdaten dort zentral recherchierbar bleiben, auch wenn sie in anderen Repositorien (zenodo, github etc.) abgelegt wurden. Während mit der Repositorienfreiheit zumindest ein hierarchisches Konzept – fachlich, institutionell, allgemein – für die Forschungsdatenablage einen Handlungsrahmen absteckt, bleiben bei der Verzeichnung der Forschungsdaten Fragen nach der Zuschreibung von Autorenschaft auf der granularen Ebene noch nicht gelöst. Wie kann die Tätigkeit der Datenmodellierung in im Vergleich zu TEI-Headern reduzierten Metadatensets wie Dublin Core angemessen abgebildet werden? Taxonomien[4] sind zwar hilfreich für die Klärung, werden jedoch nur in seltenen Fällen von entsprechenden Metadatenformularen unterstützt und müssten zudem in bibliothekarischen Erfassungssystemen abbildbar sein.

3 Die Forschungsdaten

Was sind Daten in einem editionswissenschaftlichen Grundlagenprojekt im digitalen Paradigma? Der Begriff Forschungsdaten lässt sich im digitalen Wörterbuch der deutschen Sprache seit 1975 nachweisen.[5] Die Konnexion zum digitalen ist hier offensichtlich, jedoch mag es eine sprachliche Ungenauigkeit sein, Forschungsdaten nur auf die digitalen Komponenten eines editorischen Forschungsprojekts zu beziehen.[6] Um jedoch diese Daten von handschriftlichen Notizen, Anmerkungen, Exzerptionen, Schreibübungen und weiteren Dokumenten, die sich im typischen Nachlass einer/s Editionswissenschaftler:in – im besten Falle in einem Archiv – finden lassen, zu trennen, ist der Begriff Forschungsdaten sinnvoll und bezieht sich nur auf die digitalen Objekte (um hier nicht digitale Daten schreiben zu müssen), die erst seit Kurzem in die informationswissenschaftliche Perspektive geraten sind.
Forschungsdaten im editionswissenschaftlichen Kontext sind vielfältig, ihnen gemein ist jedoch, dass sie für den Produktionsprozess der (digitalen) Edition wichtig waren. Im konkreten Beispiel sind das:

- AV-Digitalisate
- XML-Dateien und dazugehöriges Schema
- Zotero|Citavi|Endnote-Bibliographien
- SQL-Datenbank

[4] („CRediT“ o. J.)
[5] https://www.dwds.de/wb/Forschungsdaten
[6] Vgl. (Kamzelak 2022)

- Publikationswebsite
- Digitale dienstliche Überlieferung
    - Email
    - Vertragliche Unterlagen

Die im Projekt entwickelte Dateneingabesoftware, sowie die verwendete Software zur Erstellung der Edition gehört hier nicht dazu und ist besser unter dem Stichwort Research Software gesondert zu betrachten. Für deren Sicherung sind insbesondere Fragen über die langfristige Verwendbarkeit kritisch, die bei technologieagnostischen XML-basierten Formaten wegfallen. Forschungssoftware wurde bislang auf github publiziert, veraltet jedoch mit der Zeit, wenn z.B. eingebundene JavaScript-Bibliotheken inkompatibel werden.[7]
Aus der Liste lässt sich zudem entnehmen, dass nicht alle Daten sofort publizierbar sind. Aus den Arbeitsdateien müssen personenbezogene Daten gelöscht werden, also z.B. Login-Daten. Dennoch muss eine Weiterverwendung in zukünftigen Forschungsprojekten bestmöglich nach heutigem Kenntnisstand erreicht werden. Nicht publizierbare Forschungsdaten werden im Akademiearchiv aufbewahrt und nach OAIS verzeichnet und unterliegen den gesetzlich geregelten Zugangsfristen. Hier geben auch im Digitalen archivarische Leitlinien zur Abgabe dienstlicher Unterlagen der BBAW wichtige Kriterien für die digitale Kassation. Wichtig in diesem Zusammenhang ist, dass diese Forschungsdaten trotz Zugangsbeschränkung immer noch den FAIR-Prinzipien folgen, wenn ihre Metadaten FAIR sind. Wenn also publizierte Findbücher diese Daten auffindbar machen, diese interoperabel und reusable sind, dann können die Forschungsdaten über den gatekeeper Archiv auch zugänglich sein. Eine kostenintensive digitale Langzeitarchivierung sollte durch aktives Forschungsdatenmanagement, das zwischen Langzeitspeicherung und -archivierung unterscheidet, auf die digital unikalen Objekte beschränkt werden, also nur auf die Objekte angewendet werden, deren analoges Äquivalent nicht mehr vorhanden ist oder nie vorhanden war (z.B. Emails).

[7] Ein Beispiel mag hier das Tool rasmify sein, das arabischen Text auf seine Grundform ohne Vokalisationszeichen reduziert (rasm). Vgl. https://github.com/telota/rasmify

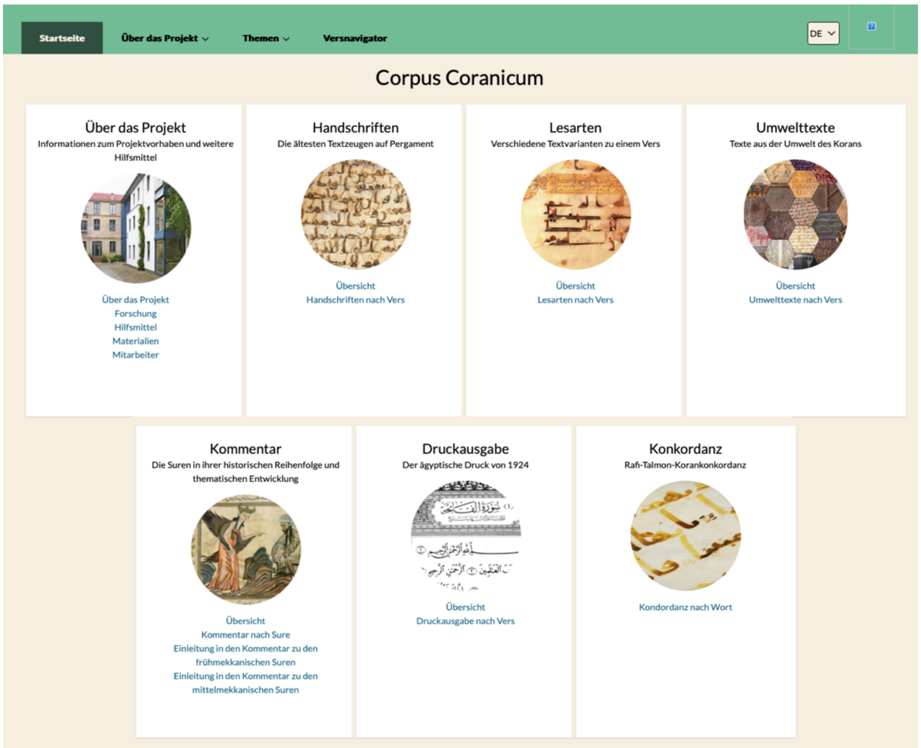


Abb. 3. Die grüne Website Corpus Coranicum: https://www.corpuscoranicum.de.

Ausgehend von diesen Herausforderungen geisteswissenschaftlicher Forschungsdaten in einem inklusiven Verständnis stellt der Vortrag Lösungsstrategien vor, die von der Initiative Forschungsdaten-management an der Berlin-Brandenburgischen Akademie der Wissenschaften für editorische Grundlagenprojekte im Dialog mit Philologie, IT, Digital Humanities und Informationswissenschaften entwickelt werden. Die Erkenntnisse, die durch diesen täglichen Umgang mit den vielfältigen Daten gewonnen werden, lassen es zu, einen Beitrag zur Definition des Begriffs Forschungsdaten im editionswissenschaftlichen Kontext zu leisten. Wenn Roland S. Kamzelak jüngst Forschungsdaten für die Editionsphilologie als „neues Ideenfeld“[8] bezeichnete, so ist ihm dahingehend zuzustimmen, dass die Definition des Begriffs Forschungsdaten Arbeit ist, die in der Editionsphilologie noch geleistet werden muss.[9] Sie kann nur durch Zusammenarbeit unterschiedlicher Fachdomänen geleistet werden und muss Theorie und Praxis vereinen.

[8] (Kamzelak 2022, 115)
[9] (Kamzelak 2022, 114)

Bibliografie

„CRediT“. o. J. CRediT. Zugegriffen 15. Mai 2023. https://credit.niso.org/.

Kamzelak, Roland S. 2022. „Forschungsdaten Und Edition: Herausforderungen Und Chancen“. Editio 36 (1): 106–15. https://doi.org/10.1515/editio-2022-0005.

Marx, Michael Josef, und Tobias J. Jocham. 2019. „Radiocarbon (14C) Dating of Qurʾān Manuscripts“. In Qurʾān Quotations Preserved on Papyrus Documents, 7th-10th Centuries, 188–221. Leiden: Brill. https://doi.org/10.1163/9789004376977_007.

Sutter, Paul, und Katja Marciniak. 2021. „FDM lokal, gemeinsam und vernetzt gestaltet. Proaktiv zum nachhaltigen FDM“. Cologne, September 3. https://doi.org/10.5281/zenodo.5379645.

Unsworth, John. 2000. „Scholarly Primitives: what methods do humanities researchers have in common, and how might our tools reflect this?“ 13. Mai 2000. https://people.brandeis.edu/~unsworth/Kings.5-00/primitives.html.

"FAIR Collections as Data"

Services von Kulturerbeeinrichtungen für die datengetriebene Forschung

Woitas, Kathi
kathi.woitas[at]gmail.com
Zentralbibliothek Zürich, Schweiz
ORCID-iD: 0000-0003-1796-1978

Zusammenfassung. Mit dem Konzept Digital Scholarship wird die digitale Transformation der Wissenschaft beschrieben. Bestände aus Kulturerbe-Institutionen stellen hierbei eine unverzichtbare Grundlage für datengetriebene Forschungsansätze in den Geisteswissenschaften dar. Mit deren breiter Aufbereitung und Kuration als Datenkonvolute könnte die Verfügbarkeit von FAIRen Forschungsdaten im grossen Umfang erhöht werden. Collections as Data als ideeller und praktischer Ansatz im Kulturerbe-Sektor, um die datenbasierte Nutzung der Bestände zu ermöglichen und zu vereinfachen, bietet hierfür einen vielversprechenden Ausgangspunkt. Mit der breiten Umsetzung von «FAIR Collections as Data» können Bibliotheken die zentrale Basis für die Entwicklung von Digital Scholarship Services legen.

1. Digital Scholarship

In der Informationswissenschaft wird die tiefgreifende digitale und datengetriebene Transformation der wissenschaftlichen Praxis durch die Konzepte Digital Scholarship bzw. Data Scholarship[1] diskutiert. Nach Borgman wird Wissenschaft als umfassend verstandenes Funktionssystem in Zukunft daten- und informationsintensiv, verteilt, interdisziplinär und kollaborativ geprägt sein. Daten als "input and output of scholarship" werden zum zentralen Moment der Forschung. Diese Verflechtung von Daten und Wissenschaft ist unausweichlich, und prägt sich in spezifischen Wissensinfrastrukturen aus, die auch von normativen, sozialen und technischen Einflüssen geformt werden. Paradigmatisch betrifft dies etwa die Digital Humanities, deren Datenzentriertheit sich im Kern auf zwei Arten manifestiert: Erstens werden Daten zum zentralem Untersuchungsgegenstand, was ihre Produktion durch Datafizierung von Forschungsobjekten einschließt. Zweitens werden entsprechende datenbasierte Analyse-,

[1] Borgman, "Scholarship in the Digital Age"; Borgman, "Big data, little data, no data".

Modellierungs- und Synthesetechniken als Forschungsmethoden eingesetzt.

Der Begriff der Digital Scholarship dient praktisch auch als Überbegriff für die sich neuformierenden datengetriebenen Disziplinen – und Bibliotheken reagieren mit sogenannten Digital Scholarship Services auf den Bedarf an begleitenden Dienstleistungen.[2] Digital Humanities und Computational Social Science stützen sich zentral auf Bestände des kulturellen Erbes. Die Verfügbarkeit von Ressourcen aus Bibliotheken, Museen und Archiven (LAM) in datafizierter Form stellt somit eine unverzichtbare Grundlage für die datengetriebene Forschung in den Geistes- und Sozialwissenschaften dar.[3] Angesichts dieses Wandels wird die Erzeugung, Aufbereitung und Bereitstellung solcher Forschungsdaten für Bibliotheken eine ähnlich zentrale Bedeutung erlangen wie die angestammte Bereitstellung von wissenschaftlichen Informationen.

2. (FAIR) Collections as Data

Während Bibliotheken seit langem eigene Daten in Form von Metadaten, Normdaten und Digitalisaten produzieren, wird mit dem Konzept Collections as Data[4] nun auch deren aktive Nutzbarmachung verfolgt. Das Santa Barbara Statement on Collections as Data[5] listet zehn Prinzipien auf, um den „computational use“ von Beständen aus Kulturerbeinstitutionen breit zu ermöglichen und zu fördern, im Kern also, digitalisierte und originär digitale Sammlungen maschinenverarbeitbar bereit zu stellen. Neben möglichst niedrigen Nutzungsbarrieren um verschiedenen Zielgruppen gerecht zu werden, sind hierbei ethische Erwägungen, Fragen der technischen Robustheit, Interoperabilität und Nachhaltigkeit zu adressieren. Eine detaillierte Dokumentation soll nicht nur Informationen zu Provenienz und Datafizierung liefern, sondern auch die Nachnutzung vereinfachen. Die erzeugten Daten sind unter so offenen Lizenzen wie möglich bereitzustellen, dies schließt verwandte Datenbestände wie Metadaten

[2] Wilms, „Digital Humanities in European Research Libraries“; Woitas, „Digital Scholarship Services“.

[3] Schöch, „Big? Smart? Clean? Messy? Data in the Humanities“; Bubenhofer und Rothenhäusler, „„Korporatheken‘“; Hawkins, „Archives, Linked Data and the Digital Humanities“.

[4] Padilla u. a., „Final Report --- Always Already Computational“.

[5] Padilla u. a., „Santa Barbara Statement on Collections as Data --- Always Already Computational“.

und Daten aus bereits erfolgter Analyse mit ein. Collections as Data stellt dabei ein dynamisches, community-getriebenes Rahmenwerk aus Empfehlungen, Hilfsmitteln und Best Practices dar, das sich in steter Weiterentwicklung befindet.[6]

Seit der Etablierung des Collections-as-Data-Konzeptes kuratieren und veröffentlichen eine wachsende Zahl von Kulturerbe-Institutionen ihre Daten, oft in dezidiert nutzerorientierter Ausrichtung.[7] Gleichzeitig wächst das Bewusstsein für den Wert von Daten aus LAM-Institutionen für die weitere Verwendung als Forschungsdaten[8], in besonderer Weise auch für Text-Mining-Anwendungen[9] und die Data-Science-Forschung.[10] Daneben werden wissenschaftliche Literatur- und Faktendatenbanken, und damit erworbene Bestände explizit als Desiderate genannt.[11] LAM-Institutionen sind somit nicht nur aufgerufen, Bestände zu datafizieren, sondern die Nutzung im Forschungskontext spezifisch mit zu denken.

Zwischen den Collections-as-Data- und den FAIR-Prinzipien[12] bestehen dabei große Ähnlichkeiten: Primär das Ziel des “computational use” bzw. der “machine-actionable data”, aber ebenso die Sichtweise, dass Metadaten selbst zu kuratierende Daten darstellen, und die Forderungen nach umfassender Erschliessung, Vertrauenswürdigkeit, Nachhaltigkeit, Provenienzinformationen, der Verwendung von (offenen) Standards und der Deklarierung von Nutzungsbedingungen. Die aktuelle Checklist to publish collections as data in GLAM institutions[13] enthält zudem Empfehlungen von FAIR-ähnlichen Prinzipien, etwa die Referenzierbarkeit mit persistentem Identifikator.

[6] Vgl. die Programmseite “Collections as Data: Always Already Computational” https://collectionsasdata.github.io/ bzw. die Fortsetzung “Collections as Data: Part to Whole”, https://collectionsasdata.github.io/part2whole/.

[7] Ames, „Transparency, provenance and collections as data“; Sherratt, „GLAM Workbench“; Candela u. a., „Reusing Digital Collections from GLAM Institutions“.

[8] Koster und Woutersen-Windhouwer, „FAIR Principles for Library, Archive and Museum Collections“; Hauf u. a., „Data Reuse in the Social Sciences and Humanities“.

[9] Senseney u. a., „Transforming Library Services for Computational Research with Text Data“.

[10] Neudecker, „Cultural Heritage as Data“; “BigLAM: BigScience Libraries, Archives and Museums“

[11] Allianz der Deutschen Wissenschaftsorganisationen, „Bedarf und Anforderungen an Ressourcen für Text und Data Mining“.

[12] Wilkinson u. a., „The FAIR Guiding Principles for Scientific Data Management and Stewardship“; Jacobsen u. a., „Fair principles“.

[13] Candela u. a., „A checklist to publish collections as data in GLAM institutions“.

Die Unterschiede zwischen Beiden – Fokus auf eine breite Nutzbarkeit und Vermittlung vs. Orientierung an disziplinspezifischen Erfordernissen, „open by default“ vs. elaborierte Zugangsregime, wo nötig – sind indes nicht trivial. Anknüpfungspunkte von Collections as Data bestehen in Form von ethischen Überlegungen und Provenienzinformationen auch zu den CARE Principles for Indigenous Data Governance.[14] Letztere können dezidiert für Sammlungen aus indigenen bzw. kolonialen Kontexten zur Operationalisierung dieser Punkte herangezogen werden.

Collections as Data spielen eine Schlüsselrolle bei der Entwicklung von Digital Scholarship Services. Um eine der wissenschaftlichen Informationsversorgung analoge „Datenversorgung“ zu bieten, müssen potentielle Datenquellen breit für die datenbasierte Forschung erschlossen werden. Idealerweise leisten LAM eine systematische Abklärung und Kuration ihrer gesamten Ressourcen als Daten: eigene Kataloge und digitale Sammlungen, aber auch der lizenzierten Inhalte. In Auswahl wurde dies an der Universitätsbibliothek Bern durchgeführt. Die geschaffenen generischen Zugänge zu Datenbeständen umfassen dabei

- Jupyter Notebooks zum Datamining von eigenen digitalisierten Sammlungen auf nationalen Plattformen[15]
- lizenzierte Text-and-Data-Mining-Plattformen
- mit rechtlichen und technischen Bedingungen dokumentierte freie und lizenzierte Ressourcen[16] für das TDM.

Die umfassende Implementierung dieser Daten-Services ist eine komplexe, ressourcenintensive Aufgabe, die neue Kompetenzen, Verfahren und z.T. Infrastrukturen erfordert. Dies gilt umso mehr, wenn Daten nicht für allgemeine, sondern für spezifische Zwecke bereitgestellt und publiziert werden sollen. Spezifische Bedarfe von Forschungsprojekten sind evident, etwa für die Erstellung von Datenkonvoluten on demand.[17] Mit den dokumentierten rechtlichen und

---

[14] Carroll u. a., „The CARE Principles for Indigenous Data Governance“; Imeri und Rizzolli, „CARE Principles for Indigenous Data Governance“.

[15] Siehe Python Toolbox, Web-Tools, https://github.com/ub-unibe-ch/ds-pytools/tree/main/web-tools

[16] Siehe Digital Scholarship Services, https://www.ub.unibe.ch/service/digital_scholarship/index_ger.html

[17] Lippincott, „Mapping the Current Landscape of Research Library Engagement with Emerging Technologies in Research and Learning“, 14–15; Pinfield, Cox, und Rutter, „Mapping the Future of Academic Libraries“, 51–52; Senseney u. a., „Transforming Library Services for Computational Research with Text Data“, 18–20.

technischen Abklärungen, ausgestalteten Datenzugängen und lizenzierten Datenplattformen im Rahmen von Collections-as-Data-Bemühungen kann eine wertvolle Basis zur einfachen Gewinnung von spezifischen Forschungsdaten gelegt werden. Mit der weiteren, spezifischen Verarbeitung der ausgewählten Daten gemäß der FAIR Principles und gegebenenfalls der CARE-Prinzipien („FAIR Collections as Data"), könnte so auf Basis von Collections as Data die Verfügbarkeit von Forschungsdaten effektiv und effizient erhöht werden.

Bibliografie

Allianz der Deutschen Wissenschaftsorganisationen, Schwerpunktinitiative Digitale Information. „Bedarf und Anforderungen an Ressourcen für Text und Data Mining: Zusammenfassung der Ergebnisse einer Umfrage aus dem Zeitraum April bis Mai 2015". TIB Hannover, 2015. https://doi.org/10.2314/GBV:836171284.

Ames, Sarah. 2021. „Transparency, provenance and collections as data: The National Library of Scotland's Data Foundry". *LIBER Quarterly: The Journal of the Association of European Research Libraries* 31 (1): 1–13. https://doi.org/10.18352/lq.10371.

„BigLAM: BigScience Libraries, Archives and Museums". 2023. HuggingFace. 31. Januar 2023. https://huggingface.co/biglam.

Borgman, Christine L. 2007. *Scholarship in the Digital Age: Information, Infrastructure, and the Internet*. Cambridge, MA: MIT Press.

———. 2015. *Big data, little data, no data: scholarship in the networked world*. Cambridge, MA: The MIT Press. https://www.jstor.org/stable/j.ctt17kk8n8.

Bubenhofer, Noah, und Klaus Rothenhäusler. 2016. „‚Korporatheken': Die digitale und verdatete Bibliothek". *027.7 Zeitschrift für Bibliothekskultur / Journal for Library Culture* 4 (2): 60–71. https://doi.org/10.5281/zenodo.4705307.

Candela, Gustavo, Nele Gabriëls, Sally Chambers, Thuy-An Pham, Sarah Ames, Neil Fitzgerald, Katrine Hofmann, u. a. 2023. „A checklist to publish collections as data in GLAM institutions“. arXiv. https://doi.org/10.48550/arXiv.2304.02603.

Candela, Gustavo, María Dolores Sáez, MPilar Escobar Esteban, und Manuel Marco-Such. 2022. „Reusing Digital Collections from GLAM Institutions“. *Journal of Information Science* 48 (2): 251–67. https://doi.org/10.1177/0165551520950246.

Carroll, Stephanie Russo, Ibrahim Garba, Oscar L. Figueroa-Rodríguez, Jarita Holbrook, Raymond Lovett, Simeon Materechera, Mark Parsons, u. a. „The CARE Principles for Indigenous Data Governance“. *Data Science Journal* 19 (4. November 2020): 43. https://doi.org/10.5334/dsj-2020-043.

Hauf, Nicolai, Andreas Fürholz, Vanessa Christina Klaas, Jennifer Morger, Elena Šimukovič, und Martin Jaekel. „Data Reuse in the Social Sciences and Humanities: Project Report of the SWITCH Innovation Lab “Repositories & Data Quality”“. Winterthur: ZHAW Zurich University of Applied Sciences with support from SWITCH, März 2021. https://doi.org/10.21256/zhaw-2404.

Hawkins, Ashleigh. 2021. „Archives, Linked Data and the Digital Humanities: Increasing Access to Digitised and Born-Digital Archives via the Semantic Web“. *Archival Science* 22 (Dezember): 319–44. https://doi.org/10.1007/s10502-021-09381-0.

Imeri, Sabine, und Michaela Rizzolli. „CARE Principles for Indigenous Data Governance: Eine Leitlinie für ethische Fragen im Umgang mit Forschungsdaten?“ *o-bib. Das offene Bibliotheksjournal / Herausgeber VDB* 9, Nr. 2 (14. Juni 2022): 1–14. https://doi.org/10.5282/o-bib/5815.

Jacobsen, Annika, Ricardo de Miranda Azevedo, Nick Juty, Dominique Batista, Simon Coles, Ronald Cornet, Mélanie Courtot, u. a. 2020. „Fair principles: interpretations and implementation considerations“. *Data Intelligence* 2 (1–2): 10–29. https://doi.org/10.1162/dint_r_00024.

Koster, Lukas, und Saskia Woutersen-Windhouwer. 2018. „FAIR Principles for library, archive and museum collections: A proposal for standards for reusable collections“. *The Code4Lib Journal*, Nr. 40 (Mai). https://journal.code4lib.org/articles/13427.

Poster

Retrodigitalisierung bibliographischer Daten mit Hilfe von Parser-Technologien

Arnold, Eckhart
arnold[at]badw-muenchen.de
Bayerische Akademie der Wissenschaften, München, Deutschland

Frank, Ingo
frank[at]ios-regensburg.de
Leibniz-Institut für Ost- und Südosteuropaforschung, Regensburg, Deutschland

Weber, Albert
weber[at]ios-regensburg.de
Leibniz-Institut für Ost- und Südosteuropaforschung, Regensburg, Deutschland

Zusammenfassung. Der Vortrag stellt eine generische Vorgehensweise zur Aufbereitung gedruckter Bibliographien in bibliographische Daten vor. Fallbeispiel ist eine über 5.200 Titel umfassende Pressebibliographie, deren Daten über den DHParser extrahiert werden. Dieser basiert auf formalen Grammatiken (EBNF), die regulären Ausdrücken ähneln, aber wesentlich leistungsfähiger sind. Die Unterstützung von Komponenten-Tests für Teile der Grammatik und fehlertolerantem Parsen erleichtern eine inkrementelle Parser-Entwicklung, durch die die Aufbereitung von textförmig vorliegenden Bibliographien in strukturierte Daten in einem relativ planbaren und kontrollierten Prozess möglich ist.

1 Einleitung

Hintergrund des Vortrags ist ein Digitalisierungsprojekt zur deutschsprachigen Presse aus dem östlichen Europa. Das Projekt zielt einerseits auf die Digitalisierung historischer Zeitungen, andererseits auf eine umfangreiche Erfassung bibliographischer Daten zu sämtlichen nachweisbaren Zeitungen und Zeitschriften (ca. 5.200 Titel) deutscher Bevölkerungsgruppen in der Region aus dem Zeitraum 1681 bis 2023. Die bibliographischen Daten, die zuvor teils aus anderen gedruckten Bibliographien, teils aus der Zeitschriftendatenbank recherchiert wurden, sind dabei Grundlage für eine Datenbank, die zu einem offenen Presselexikon ausgebaut werden soll.

Für die Datenerfassung wird auf eine bereits im volltextdurchsuchbaren Word- und PDF-Format vorliegende, 1.150-seitige Bibliographie zurückgegriffen. Die Daten zu den einzelnen Titeln sind strukturiert in bis zu 33 bibliographischen Feldangaben und können somit von einem Parser-Programm im Prinzip exakt erfasst werden. In der Praxis zeigt sich allerdings, dass in solcherart manuell strukturiertem Material Fehler, Abweichungen oder Varianten vorkommen, was die Parserentwicklung vor zusätzliche Herausforderungen stellt. Dennoch ist dieser Ansatz sinnvoll und sogar schon bei vergleichsweise kompliziertem Material wie Druck-Wörterbüchern erfolgreich eingesetzt worden [7].

Für den Bau des Parsers wurde in dem vorliegenden Projekt der Parsergenerator DHParser verwendet [1]. Ursprünglich für die Erstellung domänenspezifischer Notationen an der Bayerischen Akademie der Wissenschaften entwickelt, eignet sich dieselbe Technologie aber auch für die Retrodigitalisierung strukturierter Textdaten.

2 Vorgehensweise

Für die Entwicklung der Grammatik verwendet man zunächst nur Auszüge dieser Textdaten als Testmaterial. Bei der von DHParser unterstützten Test-getriebenen Grammatik-Entwicklung, werden daraus nun zunächst Test-Beispiele für einzelne Komponenten der Bibliographie extrahiert.

Die Grammatikentwicklung erfolgt entweder im Top-Down oder im Bottom-Up-Verfahren:

- Top-Down: Zunächst die grammatischen Regeln für die Grobstruktur des Dokuments formulieren (z.B. Dokument = eine Folge von durch Leerzeilen getrennten bibliographischen Einträgen) und dann nach und nach verfeinern.
- Bottom-Up: Umgekehrt mit den Feinstrukturen beginnen und diese nach und nach zu umfassenderen zusammensetzen.

Durch die bereits vorhandene „Test-Suite“ wird immer sichergestellt, dass die Grammatik für die Komponente, die gerade entwickelt wird, zuverlässig funktioniert, bevor zur nächsten übergegangen wird. Schritt für Schritt werden aus den Fehlerstellen neue Test-Beispiele zur Verbesserung der Grammatik gebildet.

Dabei ist ein fehlertoleranter Parser, der nicht gleich beim ersten Fehler abbricht, sondern alle Fehlerstellen auf einmal ausgibt, sehr hilfreich. Bei wenig verbleibenden Fehlern können die Ausgangstexte von den Fachwissenschaftlern oder sogar von Hilfskräften entsprechend korrigiert werden.

3 Diskussion und Vergleich mit anderen Ansätzen

Die Technologieauswahl bei Retrodigitalisierungsprojekten wie diesem ist nicht auf Parser-Technologien beschränkt. Je nach der Art des Datenmaterials bieten sich insbesondere auch Machine Learning-Verfahren (ML) wie etwa zum Trainieren maßgeschneiderter Parser mit Conditional Random Field-Modellen [5] oder KI-gestützte Methoden wie z.B. GROBID [4] an.

Man kann grob sagen, dass sich Parser-Technologien vor allem bei gut strukturiertem Material anbieten, weil sie den Vorteil haben, dass man keine umfangreichen Trainingsdatensätze präparieren muss, und sie zudem naturgemäß exakter arbeiten als ML-Techniken.

Bei schlecht strukturiertem Material stoßen Parser-Technologien an ihre Grenzen. Hier können KI-Algorithmen ihre Stärken ausspielen. Eine Kombination beider Methoden ist aber durchaus denkbar.

Parsergeneratoren sind klar im Vorteil, weil sie die Verarbeitung wesentlich komplexerer Strukturen ermöglichen. Parsergeneratoren, die wie DHParser formale Grammatiken verarbeiten können [2], zeichnen sich dabei durch bessere Übersichtlichkeit gegenüber solchen aus, die wie PyParsing nur die Definition einer Grammatik in Programmcode erlauben.

4 Fazit und Ausblick

Parser-Technologien als Werkzeug zur Datenaufbereitung führen in den Digital Humanities eher ein Nischen-Dasein. Das kann damit zusammenhängen, dass der Bau von Parsern und die Erstellung formaler Grammatiken als schwierig gelten. Mit DHParser steht jedoch ein Werkzeug zur Verfügung, dass die inkrementelle Entwicklung eines Parsers vergleichsweise einfach macht und zudem auf geisteswissenschaftliche Anwendungen zugeschnitten ist.

Prinzipiell kann die erstellte Grammatik zum Parsen andere Bibliographien angepasst oder in Teilen ggf. direkt nachgenutzt werden. In weiteren Arbeitsschritten können die von DHParser gelieferten XML-Daten etwa unter Verwendung der FRBR-aligned Bibliographic Ontology (FaBiO) [6] zu RDF/Turtle-Daten für das Semantic Web und Linked Data aufbereitet werden.[1]

Bibliographie

Arnold, Eckhart. 2016–2023 "DHParser. A Parser-Generator for Digital-Humanities-Applications", *Division for Digital Humanities Research & Development, Bavarian Academy of Science and Technology, Munich Germany*, https://gitlab.lrz.de/badw-it/dhparser.

Arnold, Eckhart. 2016–2023. "DHParser. User Manual.", https://dhparser.readthedocs.io.

Frank, Ingo, und Albert Weber. 2021. „Aufbereitung und Publikation bibliographischer Forschungsdaten. Omeka S als Werkzeug zur (Meta-)Datenkuration und Wikidata als Plattform zur Forschungsdatendistribution und -publikation." In *FORGE 2021: Forschungsdaten in den Geisteswissenschaften – Mapping the Landscape – Geisteswissenschaftliches Forschungsdatenmanagement zwischen lokalen und globalen, generischen und spezifischen Lösungen (FORGE2021)*, Köln. https://doi.org/10.5281/zenodo.5379617.

Lopez, Patrice. 2008–2023: "GROBID", https://github.com/kermitt2/grobid.

Lüschow, Andreas. 2020. „Automatische Extraktion und Semantische Modellierung der Einträge einer Bibliographie französischsprachiger Romane." In *DHd 2020 Spielräume: Digital Humanities zwischen*

---

[1] Siehe dazu die bereits bei der FORGE 2021 mit allen Arbeitsschritten vorgestellte generische Vorgehensweise [3], bei der übrigens zum Parserbau zuerst der Einsatz von PyParsing vorgesehen war, aber dann aus den in Abschnitt 3 genannten Gründen durch den DHParser ersetzt wurde.

Modellierung und Interpretation. Konferenzabstracts, hrsg. von Christof Schöch und Patrick Helling, 80–84. Verband Digital Humanities im deutschsprachigen Raum e.V. https://doi.org/10.5281/zenodo.4621704.

Peroni, Silvio, und David Shotton. 2012. “FaBIO and CiTO: Ontologies for Describing Bibliographic Resources and Citations.” Journal of Web Semantics 17: 33–43. https://doi.org/10.1016/j.websem.2012.08.001.

Zacherl, Florian 2022. „Digitale Tiefenerschließung traditioneller Lexikographie – am Beispiel des Romanischen Etymologischen Wörterbuchs.“ Korpus im Text, Band 16. Version 1 (22.06.2022, 01:28). URL: http://www.kit.gwi.uni-muenchen.de/?p=82908&v=1.

Eine prosopographische Datenbank zur Geschichte der Mathematik an der Universität Tübingen

Beeley, Philip
philip.beeley[at]history.ox.ac.uk
Faculty of History, University of Oxford, Großbritannien

Kahle, Reinhard
reinhard.kahle[at]uni-tuebingen.de
Universität Tübingen, Deutschland

Zusammenfassung. Es wird ein Datenbankprojekt zu Mathematikern an der Universität Tübingen vorgestellt, mit Daten aus verschiedenen Quellen wie Matrikelregistern, Listen von Lehrstuhlinhabern und Vorlesungsabschriften. Diese Datenbank ist Teil eines Projektvorhabens im Rahmen der anstehenden 550-Jahrfeier der Universität Tübingen unter dem Titel „Disziplingeschichte als Universitätsgeschichte". Die zu konstruierende prosopographische Datenbank wird die Daten interaktiv und öffentlich zugänglich machen und es damit erlauben, individuelle Lebens- und Wissenschaftswege in der Mathematik nachzuvollziehen. In diesem Vortrag geht es um die Herausforderungen, die sich beim Konzept der Datenbank im Hinblick auf die Spezifikation, Datenlage, Nutzbarkeit und nicht zuletzt Kompatibilität und Vernetzbarkeit mit anderen bestehenden Datenbanken ergeben.

1 Hintergrund

Seit der Gründung der Universität Tübingen vor bald 550 Jahren ist die Mathematik als akademische Disziplin kontinuierlich vertreten gewesen. Eine derartige Kontinuität der Mathematik als Lehr- und Forschungsfach kann keine andere vergleichbare Institution in den Grenzen des heutigen Deutschlands vorweisen (vgl. Schöner 1994). Im Rahmen eines Forschungsprojekts unter dem Titel „Disziplingeschichte als Universitätsgeschichte", das die Geschichte der Mathematik an der Universität Tübingen auch im Hinblick auf ihre Wirkung über die Universität hinaus untersuchen will, ist der Aufbau einer prosopographischen Datenbank geplant, in der die Mathematik betreffenden Daten interaktiv und öffentlich zugänglich gemacht werden. Das Universitätsarchiv und die Universitätsbibliothek verfügen über eine in ihrer Breite und Tiefe einmalige Sammlung von Daten zu verschiedenen Aspekten des Fachs Mathematik in dieser über mehrere

Jahrhunderte sich erstreckenden Zeitspanne: Daten zu den Studierenden (Name, Herkunft, Eltern, Alter bei der Matrikulation, usw.), zu den Lehrkräften (Name, Herkunft, Studium, Tätigkeitszeitraum, Vorlesungen, usw.), und zu den Lehrveranstaltungen (Vorlesungstitel, Dozent, Datum, usw.). Teilweise liegen diese Informationen in gedruckter Form vor, teilweise in Form von durch das Universitätsarchiv ins Netz gestellte Dokumente (teils als gedruckte, teils als handschriftliche Quellen). Hinzu kommen Daten, die sich aus anderen Quellen, insbesondere aus zeitgenössischer wissenschaftlicher Korrespondenz erschließen lassen – zum Beispiel aus dem gedruckten Briefwechsel Johannes Keplers oder aus dem Briefwechsel Leonhard Eulers. Immer wieder zeichnete sich die Universität durch namhafte Vertreter des Fachs aus – beispielsweise durch Michael Mästlin und Johann Scheubel im 16. Jh., Johannes Kepler und Wilhelm Schickard im 17. Jh., Georg Bernhard Bilfinger, Georg Wolfgang Krafft, und Johann Kies im 18. Jh., Alexander von Brill und Hermann Hankel im 19. Jh. (Seck 1981; Seck 1995; Hantsche 1996).

2 Das Vorhaben

Das Projekt will auf der Grundlage von datentechnisch exzellenten bis guten Voraussetzungen exemplarisch die Vorteile eines prosopographischen Ansatzes benutzen, um die Geschichte der Mathematik an der Universität Tübingen seit ihrer Gründung zu dokumentieren. Die Konstruktion der interaktiven und öffentlich zugänglichen Datenbank basiert vor allem auf dem mit dem Namen John Bradley verbundenen Konzept der *Factoid Prosopography*. Entsprechende Felder der Datenbank werden mit biographischen Informationen zu Studierenden und Lehrkräften der mathematischen Wissenschaften an der Tübinger Universität so aufgefüllt, dass an die Datenbank gerichtete Fragen Details zu einzelnen Dozenten, Themen, und Studierenden oder auch zu Gruppen derselben sich abrufen lassen. Dabei sollen stark strukturierte Daten nach den Prinzipien von 'linked open data' Konventionen wie RDF erfaßt werden. Nach Absprache mit der Handschriftenabteilung der Tübinger Universitätsbibliothek sowie mit dem Universitätsarchiv soll es auch möglich sein, entsprechende Datenfelder im Einzelfall mit digitalisierten Aufnahmen von Originaldokumenten aus ihren Beständen zu verlinken.

3 Datenlage/Quellenlage

Das Tübinger Universitätsarchiv verfügt zu internen Zwecken über eine eigene Datenbank, in der Angaben zu den Studierenden der Universität seit ihrem Beginn enthalten sind (basierend auf Bürk et al. 1953-4). Nach Auskunft der Archivarin ist es möglich, daraus einen Auszug zu 'Mathematik' zu erstellen. Details zu den Lehrstuhlinhabern sind (Conrad 1960) zu entnehmen und können ggf. durch die *Neue Deutsche Biographie* und andere Nachschlagewerke ergänzt werden. Diese Rohdaten müssen ggf. durch historische Nachforschungen gesäubert werden. Informationen zu den an der Universität Tübingen gehaltenen mathematischen Vorlesungen und Publikationen der Lehrstuhlinhaber sind teilweise vom Universitätsarchiv ins Netz gestellt worden. Weitere Informationen etwa zu den in Tübingen tätigen Privatdozenten müssen noch erforscht werden. Dabei ist zu berücksichtigen, dass die Mathematik fakultativ in der Zeit verschiedentlich zugeordnet wurde: sie war zeitweilig mit der Philosophie, Medizin, und der Naturphilosophie (siehe Köpf 2014) (Homa 2016) (Holtz et al. 2005) verbunden. Detaillierte Vorlesungsverzeichnisse gibt es ab circa 1800, während Studentenakten ab circa 1830 vollständig aufgenommen worden sind. Die Personalakten der Professoren sowie deren Hörerlisten sind ab 1900 aufgenommen worden.

4 Forschungsarbeit und Datenanalyse

Das primäre Ziel des Projekts ist die Produktion einer interaktiven und öffentlich zugänglichen prosopographischen Datenbank, mit der durch eine benutzerfreundliche Schnittstelle gezielt Fragen zu Individuen und Gruppen von Individuen gerichtet und beantwortet werden können. Anhand dieses Instruments wird es möglich sein, eine Reihe neuer Forschungsansätze zu realisieren, bei denen beispielsweise Fragen zur Herkunft und zum Berufsweg von Studierenden der mathematischen Wissenschaften an der Tübinger Universität oder etwa zu fachspezifischen Entwicklungen im Südwesten Deutschlands feststellen und mit Hilfe der Datenanalyse sich modellieren und visualisieren lassen. Die vielfältigen Möglichkeiten eines solchen Ansatzes sind schon durch das an der Universität Oxford angesiedelte Projekt "Networking Archives" nachgewiesen worden (Beeley et al 2023).

Bibliografie

Beeley, Philip (zusammen mit Howard Hotson, Sebastian Ahnert et al.) 2023. ‘Network Analysis and the Early Modern Archive’, Sonderheft der *Huntington Library Quarterly* (im Druck).

Bürk, Albert; Wille, Wilhelm; Hermelink, Heinrich (Hrsg.) 1953-4. *Die Matrikeln der Universität Tübingen, 1477-1827*. 3 Bde, Stuttgart: Kohlhammer.

Conrad, Ernst 1960. *Die Lehrstuhle der Universität Tübingen und ihre Inhaber*, Tübingen: Selbstverlag.

Hantsche, Irmgard (Hrsg.) 1996. Der “mathematicus”: zur Entwicklung und Bedeutung einer neuen Berufsgruppe in der Zeit Gerhard Mercators, Bochum: Brockmeyer.

Homa, Bernhard 2016. *Die Tübinger Philosophische Fakultät 1652-1752. Institution – Disziplinen – Lehrkräfte*, Stuttgart: Franz Steiner Verlag.

Holtz, Sabine Holtz; Betsch, Gerhard; Zwink, Eberhard (Hrsg.) 2005. *Mathesis, Naturphilosophie und Arkanwissenschaft im Umkreis Friedrich Christoph Oetingers (1702-1782)*, Stuttgart: Franz Steiner Verlag.

Köpf, Ulrich (Hrsg.) 2014. *Die Universität Tübingen zwischen Orthodoxie, Pietismus und Aufklärung*, Ostfildern: Jan Thorbecke Verlag.

Schöner, Christoph 1994. *Mathematik und Astronomie an der Universität Ingolstadt im 15. und 16. Jahrhundert*, Berlin: Duncker & Humblot.

Seck, Friedrich (Hrsg.) 1981. *Wissenschaftsgeschichte um Wilhelm Schickard*, Tübingen: J. C. B. Mohr (Paul Siebeck).

Seck, Friedrich (Hrsg.) 1995. *Zum 400. Geburtstag von Wilhelm Schickard. Zweites Tübinger Schickard-Symposium*, Sigmaringen: Jan Thorbecke Verlag.

Data Literacy für die Klassische Philologie

d[AI]dalos – eine interaktive Infrastruktur als Lernangebot

Beyer, Andrea
beyeranz[at]hu-berlin.de
Humboldt-Universität zu Berlin, Deutschland
ORCID-iD: 0000-0002-8468-6411

Schulz, Konstantin
schulzkx[at]hu-berlin.de
Deutsches Forschungszentrum für Künstliche Intelligenz, Berlin, Deutschland
ORCID-iD: 0000-0002-3261-9735

Zusammenfassung. Das Poster informiert über das DFG-geförderte explorative Entwicklungsvorhaben Daidalos, das es Forschenden der Klassischen Philologie und verwandter Disziplinen ermöglichen soll, verschiedene Methoden des Natural Language Processing (NLP) an selbst zusammengestellten Forschungskorpora anzuwenden. Dabei ist Daidalos als interaktive Forschungsinfrastruktur konzipiert, die zugleich den Ausbau wesentlicher Teilfähigkeiten von Data Literacy, z. B. die Zusammenstellung und Analyse von Korpora oder den Umgang mit Annotationen, TEI-XML und graphischen Auswertungen, unterstützt. Hierzu sind vor allem forschungsorientierte, didaktische Lernbausteine und deren Implementierung in die Infrastruktur angedacht, um ein fach- und forschungsbezogenes Lernen zu ermöglichen.

1 Data Literacy in der Klassischen Philologie

Die Klassische Philologie hinkt im deutschsprachigen Raum den methodischen Entwicklungen der Digital Humanities (DH) hinterher. Anders ist es international, wenn u. a. literaturwissenschaftliche (z. B. Autorschaft, Ochab & Essler 2019) oder philologische Fragen (z. B. Textrekonstruktionen, Assael et al. 2022) mit DH-Methoden angegangen werden. Problematisch für Erwerb und Ausbau von Data Literacy in der Fachcommunity ist, dass es einerseits kaum Tools und Lernmaterialien gibt, die didaktisch aufbereitet und forschungsorientiert an die DH-Methoden heranführen. Andererseits fehlen domänenspezifische Grundlagendaten für die Anwendung moderner NLP-Modelle, die nur eingeschränkt auf die sog. Low-Resource-Sprachen Latein und Altgriechisch übertragen werden können (McGillivray 2013). Daher ist zusammen mit der Fachcommunity der fachspezifische Ausbau von Data Literacy durch eine interdisziplinäre Lern- und Forschungsinfrastruktur zu adressieren.

Existierende Infrastrukturen wie Weblicht sind nicht für die Klassische Philologie geeignet, weil sie keine Anbindung an die dort üblichen Textkorpora, Zitationsstandards und Analysekategorien bieten.

2 Das Daidalos-Projekt

Das noch im Aufbau befindliche interdisziplinäre DFG-Projekt Daidalos[1] entwickelt eine Software, die Forschenden der Klassischen Philologie und angrenzender Disziplinen Folgendes ermöglichen soll:

a. Annotation, Analyse und Adaption altsprachlicher Textkorpora,
b. Visualisieren, Speichern, Teilen und Exportieren der Analyseergebnisse sowie
c. Reflexion über Ergebnisse und Methodik, Verfeinerung und Neuausrichtung der Forschungsfragen.

2.1 Aufbau der Infrastruktur

Der interaktive, explorative Zugang zu antiken Texten soll Methodenreflexion fördern und User qualifizieren. Entsprechend bietet die Software folgende Inhalte:

a. Demo-Workflows mit konfigurierbaren Elementen, z. B. Korpusauswahl, wobei die Konfiguration mittels kuratierter, authentischer Beispiele detailliert erklärt und reflektiert wird;
b. eine domänenspezifische Oberfläche, die Forschende bei der Auswahl der Komponenten unterstützt, d. h. mit Interaktionen zu Vorwissen, Forschungsinteresse und Methodenwahl (jeweils speicherbar als personalisierte Einstellungen);
c. freie Konfiguration von Text- und Methodenwahl durch eigenen Quellcode (für erfahrene Forschende).

Langfristig ist die Infrastruktur in einem universitären Rechenzentrum verankert. Ein Geschäftsmodell wird ggf. nach einer Projektverlängerung zu entwickeln sein.

2.2 KI-Didaktik für eine fachspezifische Data Literacy

Orientiert am Kompetenzrahmen „Future Skills“ (Schüller 2019) wollen wir die Forschenden (i.d.R. mit Hochschulstudium, aber ohne technisches Vorwissen) vom Interpretieren der Ergebnisse und Daten (Dekodieren) zum aktiven Umgang mit ihren eigenen

[1] https://hu.berlin/daidalos

Forschungsdaten (Rekodieren) führen. Außerdem steht auch die KI-basierte Funktionsweise ausgewählter NLP-Tools (z.B. neuronale Parser) im Vordergrund. Methodisch folgen wir dem forschenden Lernen (Huber & Reinmann 2019) und Micro-Learning (Taylor & Hung 2022) mit Micro-Credentials (Thi Ngoc Ha et al. 2022). So wollen wir authentische Beispiele und Erläuterungen bieten sowie die Lernschritte individuell zuschneiden. Die Aufbereitung der forschungsnahen Beispiele erfolgt u. a. über Kurztutorials, Code-Snippets in JupyterLab und kleinschrittigen, kompetenzorientiert gestuften[2] Übungen mit automatischem Feedback (H5P[3]). Die Nutzung der Ressourcen wird begleitet durch Workshops und individuelle Use-Case-Partnerschaften.

Bibliografie

Assael, Yannis, Thea Sommerschield, Brendan Shillingford et al., „Restoring and attributing ancient texts using deep neural networks," *Nature* 603, 280–283 (2022).

Huber, Ludwig, Gabi Reinmann, *Vom forschungsnahen zum forschenden Lernen an Hochschulen – Wege der Bildung durch Wissenschaft*, Wiesbaden: Springer, 2019.

Kleinknecht, Marc, Thorsten Bohl, Uwe Maier, Kerstin Metz (Hrsg.), *Lern- und Leistungsaufgaben im Unterricht*, Bad Heilbrunn: Julius Klinkhardt, 2013.

McGillivray, Barbara, *Methods in Latin Computational Linguistics*, Leiden | Boston: Brill, 2013.

Ochab, Jeremi K., Holger Essler, „Stylometry of literary papyri," *Proceedings of the 3rd International Conference on Digital Access to Textual Cultural Heritage*, 139–142 (2019).

Schüller, Katharina, „Ein Framework für Data Literacy," *AStA Wirtschafts- und Sozialstatistisches Archiv* 13, 297-317 (2019).

Taylor, Ai-dung, Woei Hung, „The Effects of Microlearning: A Scoping Review," *Education Tech Research Dev* 70, 363–395 (2022).

[2] Hier folgen wir Kleinknecht et al. (2013). Die Autoren schlagen zur Analyse des kognitiven Aktivierungspotentials von Aufgaben aus allgemeindidaktischer Sicht folgende Stufung kognitiver Beanspruchung vor: 1. Fakten und Reproduktion, 2. Prozeduren und naher Transfer, 3. Konzepte und weiter Transfer, 4. Metakognition und Problemlösen.

[3] https://h5p.org/

Thi Ngoc Ha, Nguyen, Michael Spittle, Anthony Watt & Nina Van Dyke, „A systematic literature review of micro-credentials in higher education: a non-zero-sum game," *Higher Education Research & Development*, 1-22 (2022).

Realitätscheck Reproduzierbarkeit

Ein studentisches Open-Science-Projekt zur Reproduzierbarkeit von Forschungsergebnissen

Blümm, Mirjam
mirjam.bluemm[at]th-koeln.de
Technische Hochschule Köln, Deutschland

Frick, Claudia
claudia.frick[at]th-koeln.de
Technische Hochschule Köln, Deutschland

Zusammenfassung. Open Science als Grundlage einer transparenten und reproduzierbaren Wissenschaft wird zunehmend in Lehrpläne aufgenommen und ist prädestiniert um in studentischen Reproduktionsstudien statt in klassischen Vorlesungen vermittelt zu werden. Die Fallstudie eines Mastermoduls der Technischen Hochschule Köln zeigt, dass der Versuch, eine publizierte Studie anhand der veröffentlichten Forschungsdaten und -ergebnisse zu reproduzieren, eine sehr anschauliche und nachhaltige Lernerfahrung darstellt. Für Studierende, Lehrende und Forschende, ist dies eine Möglichkeit das eigene Verständnis von Reproduzierbarkeit zu überprüfen und gleichzeitig die Qualität des veröffentlichten wissenschaftlichen Wissens zu sichern.

1 Open Science in der Lehre

Wissenschaftliche Studien sollten so durchgeführt und dokumentiert werden, dass andere sie verstehen und reproduzieren können.[1] Es ist allerdings nicht immer auf den ersten, zweiten oder gar dritten Blick ersichtlich und zu beurteilen, ob eine Studie entsprechend konzipiert und beschrieben wurde. Das schlichte Auffinden der Daten und des Codes reichen nicht zwingend aus. Die nächste Herausforderung besteht dann darin, auch die eigenen Studien so zu gestalten und zu beschreiben, dass sie reproduzierbar sind.
Das Modul „Open Science“ im Masterstudiengang „Digital Science“ der Technischen Hochschule Köln adressiert sowohl die Reproduktion der Studien anderer als auch das Erstellen eigener reproduzierbarer

[1] Wir definieren eine Studie als reproduzierbar, wenn ihre Ergebnisse mit der gleichen Methode und den gleichen Daten, die von den Autor:innen verwendet und veröffentlicht wurden, wiederhergestellt werden können (Chiarelli et al. 2021, S. 10-11).

Studien. Zu den Lernzielen des Moduls gehört die Fähigkeit, einen wissenschaftlichen Diskurs zu führen, Werkzeuge und Dienste anzuwenden, Forschungsdaten zu verarbeiten und bereitzustellen sowie Fallstudien zu verstehen und zu reproduzieren. Der Kurs folgt dem „flipped classroom"-Prinzip (Kirch 2016) und besteht aus einer Kombination von Inputs und Diskussionen, die sich auf die Projekte der Studenten konzentrieren.

2 Die Fallstudie

Im Wintersemester 2023 wählte eine Gruppe Studierender die veröffentlichte Studie „Women's Preference for Masculine Traits Is Disrupted by Images of Male-on-Female Aggression" (Li et al. 2014) als Projektarbeit aus, die sie über eine Suche in DataCite gefunden hatten. Die Rohdaten (Li et al. 2015) sind über das Open-Access-Repository Dryad verfügbar. Bei dem Versuch die Studie zu reproduzieren, ergaben sich einige Fragen, so dass die Gruppe den Hauptautor kontaktierte und glücklicherweise weitere Informationen u.a. den ursprünglichen Programmcode erhielt. Es gelang den Studierenden im Laufe des Semesters die Studie erfolgreich zu reproduzieren und die Forschungsergebnisse zu bestätigen. Zusätzlich probierten sie eigene alternative Modelle aus um die Robustheit der Studie zu testen.[2] Auch diese unterstützten die Schlussfolgerungen der Originalarbeit.
Bei der Untersuchung des Originaldatensatzes entdeckten die Studierenden allerdings, dass ein Teil der Daten versehentlich dupliziert worden war und eines der verwendeten Bilderpaare vertauscht interpretiert wurde. Dies hatte zwar keinen Einfluss auf die Ergebnisse der Studie (im Gegenteil wurden sie durch die die Wiederholung mit den korrekten Daten sogar noch eindeutiger) trotzdem hielt die Gruppe erneut Rücksprache mit dem Autor um ihn über ihre Entdeckungen zu informieren. Dieser reagierte vorbildlich und ermutigte sie, eine Korrektur auf PsyArXiv (Randall et al. 2023) zu publizieren und in einem Kommentar zur Originalstudie zu verlinken, auf den er selbst antwortete und die Community zum Lesen ermutigte.

3 Fazit

Die Reproduktion von Forschungsergebnissen findet noch viel zu selten statt. Studentische Reproduktionsstudien als Projektarbeiten können

[2] Eine Studie ist robust, wenn ihre Ergebnisse mit einer neuen Methode, aber den gleichen Daten bestätigt werden können (Eickhoff 2020, 20:11-23:58).

dazu beitragen, diese Lücke zu schließen und gleichzeitig entsprechende Fähigkeiten an künftige Forschende zu vermitteln. Wie das Beispiel der TH Köln gezeigt hat, kann dieser Ansatz erfolgreich durchgeführt werden, erfordert aber viel Flexibilität und Engagement von Lehrenden, Studierenden und der wissenschaftlichen Gemeinschaft insgesamt.

Bibliografie

Chiarelli, Andrea, Loffreda, Lucia and Johnson, Rob, "The Art of Publishing Reproducible Research Outputs: Supporting emerging practices through cultural and technological innovation." *Zenodo*. (2021) DOI: 10.5281/zenodo.5521077.

Eickhoff, Simon, "Forschung im Fokus: Widersprüche in der Wissenschaft." *Online. Haus der Universität*. (2020) https://youtu.be/3g1Ip4NZHwM. (Letzter Zugriff: 23.07.2023)

Kirch, Crystal, *Flipping With Kirch: The Ups and Downs from Inside My Flipped Classroom*. (New Berlin, Wisconsin: The Bretzmann Group, 2016) ISBN 978-0692661901

Li, Yaoran, Bailey, Drew H., Winegrad, Benjamin, Puts, David A., Welling, Lisa L. M. and Geary, David C., "Women's Preference for Masculine Traits Is Disrupted by Images of Male-on-Female Aggression." *PLoS ONE 9 (10) (2014): e110497.* DOI: 10.1371/journal.pone.0110497.

Li, Yaoran, Bailey, Drew H., Winegrad, Benjamin, Puts, David A., Welling, Lisa L. M. and Geary, David C., Data from: "Women's preference for masculine traits is disrupted by images of male-on-female aggression", *Dryad, Dataset,* (2015) DOI: 10.5061/dryad.9bg43.

Randall, Natasha, Küçük, Berrak, Li, Yaoran, Bailey, Drew H., Winegrad, Benjamin, Puts, David A., Welling, Lisa L. M. and Geary, David C., "Correction to: Women's Preference for Masculine Traits Is Disrupted by Images of Male-on-Female Aggression." Online. preprint. *PsyArXiv* (2023) DOI: 10.31234/osf.io/ap38y.

Entwicklung und Implementierung eines Metadaten-Modells für Literatur im Netz

Ein Erfahrungsbericht aus dem Projekt SDC4Lit

Buck, Nina
nina.buck[at]hlrs.de
Höchstleistungsrechenzentrum (HLRS), Universität Stuttgart
ORCID-iD: 0000-0002-4651-6040

Ulrich, Mona
mona.ulrich[at]dla-marbach.de
Deutsches Literaturarchiv Marbach, Deutschland
ORCID-iD: 0000-0001-9591-5614

Jung, Kerstin
kerstin.jung[at]ims.uni-stuttgart.de
Universität Stuttgart, Institut für Maschinelle Sprachverarbeitung (IMS), Deutschland
ORCID-iD: 0000-0002-9548-8461

Ganzenmüller, Andreas
andreas.ganzenmueller[at]hlrs.de
Höchstleistungsrechenzentrum (HLRS), Universität Stuttgart
ORCID-iD: 0009-0002-3451-7334

Kushnarenko, Volodymyr
volodymyr.kushnarenko[at]hlrs.de
Höchstleistungsrechenzentrum (HLRS), Universität Stuttgart
ORCID-iD: 0000-0001-7427-2410

Bönisch, Thomas
thomas.boenisch[at]hlrs.de
Höchstleistungsrechenzentrum (HLRS), Universität Stuttgart
ORCID-iD: 0000-0003-3108-8597

Zusammenfassung. Die Entwicklung eines passenden Metadaten-Modells erfordert Zeit und Aufwand und darf sich nicht nur auf die Beschreibung der Daten selbst beschränken, sondern muss immer auch dem Einsatz auf unterschiedlichen Systemen gerecht werden. Im Projekt SDC4Lit wurde für archivierte literarische Werke aus dem Netz ein Metadaten-Modell basierend auf Standards verschiedener Kataloge und

Datenbanken sowie im Hinblick auf die Implementierung in einem Repositorium entwickelt.

Das Prinzip der Metadaten wird schon seit Jahrhunderten im Bibliothekswesen verwendet und mit der digitalen Verknüpfbarkeit von Daten, z.B. über das World-Wide-Web, spielen sie heute eine zentrale Rolle in der Strukturierung von Informationen. Metadatenstandards ermöglichen Austausch und digitale Zugänglichkeit zu Metadaten. Verschiedene Standards dienen dabei verschiedenen Zwecken und sind dabei auf Informationsbedürfnisse, Anwendungsszenarien und Blickwinkel der jeweiligen Community ausgerichtet. So finden sich stets weitere Daten, die so spezifisch sind, dass zu ihrer angemessenen Beschreibung ein eigens angepasstes Metadaten-Modell notwendig ist.

Die Erfindung des World-Wide-Web hatte auch einen bedeutenden Einfluss auf die Literatur. Neue digitale Formen von Literatur sind entstanden. Um diese Literatur angesichts der Schnelllebigkeit des Internets vom Verschwinden im Netz zu bewahren, werden diese Werke als digitale Objekte archiviert. Im Rahmen des Projektes SDC4Lit[1] wurde dazu ein Konzept entwickelt und die notwendige Infrastruktur aufgebaut, um die bereits vorhandenen Bestände an Literatur im Netz des Deutschen Literatur Archiv in Marbach (DLA)[2] der Wissenschaftlichen Gemeinschaft zur Verfügung zu stellen. Bei diesem Vorhaben spielten die Metadaten eine zentrale Rolle.

Zu den Beständen des DLA gehören literarische Blogs, literarische Onlinezeitschriften und Werke der Netzliteratur. Aktuell umfassen diese Bestände zusammen etwa 500 Quellen, weshalb die möglichst automatisierte Verarbeitung dieses Materials, inklusive der Metadaten, allein aufgrund der Größe der vorhandenen Datenmenge eine hohe Priorität hatte.

Natürlich gibt es auch einige prinzipiell für die Beschreibung von Netzliterarischen Werken geeignete Metadaten-Modelle wie z.B. die CELL-Taxonomie[3], die unter anderem auch die Beschreibung von Webseiten ermöglicht. Da die für das Projekt relevanten Daten grundsätzlich bereits archiviert sind, aber zum Teil in mehreren Repräsentationen vorliegen, welche teils auch noch in verschiedenen Systemen gespeichert

[1] SDC4Lit Homepage. URL: https://www.sdc4lit.de/, Zugriff am 09.05.2023.

[2] Deutsches Literatur Archiv (DLA) Marbach Homepage. URL: https://www.dla-marbach.de/, Zugriff am 09.05.2023.

[3] Cell Project, Taxonomies Definition. URL: https://cellproject.net/taxonomies-definition, Zugriff am 12.05.2023.

sind, war es von Anfang an eine besondere Herausforderung, ein Metadaten-Konzept zu entwickeln, welches der Heterogenität des Materials gerecht wird.

Zum Teil konnte man dabei auf Vorarbeiten aus früheren Projekten der beteiligten Institute zurückgreifen. So konnte bereits ein vorhandenes Anwendungsprofil (DLA - Application Profile[4]) genutzt werden, welches speziell für die Anforderungen der Netzliteratur entwickelt wurde und mit METS[5], MODS[6] und PREMIS[7] technische, bibliografische, strukturelle, administrative, prozessbezogene und erhaltungsbezogene Metadaten umfassen. Seitens des HLRS[8] konnte zudem auf Erfahrungen aus dem Projekt „Dipl-Ing"[9] zurückgegriffen werden, in welchem das Metadaten-Schema EngMeta[10] entwickelt wurde, das im Forschungsdaten-Repositorium der Universität Stuttgart (DaRUS) Verwendung findet.

Da die vorhandenen Metadaten aus unterschiedlichen Katalogen und Datenbanken zusammengetragen werden mussten, und leider nicht immer völlig konsistent und technisch auch nicht immer ordentlich strukturiert waren, war es notwendig ein aufwendiges Metadaten-Mapping vorzunehmen. Dazu wurden die vorhandenen Metadaten evaluiert und kategorisiert, sodass am Ende für alle zu befüllenden Felder des neuen Metadaten-Modells Mapping-Regeln formuliert werden konnten, die anschließend technisch (mit Python-Skripten) umgesetzt werden mussten. So war es möglich die vorhandenen Metadaten auf das Metadaten-Schema MODS abzubilden, und die Beziehungen anschließend in PREMIS darzustellen. Die so aufbereiteten Metadaten konnten schließlich ins Repositorium eingespielt werden.

Die Entwicklung eines eigenen Metadaten-Modells erfordert Zeit und Aufwand. Ein gutes Metadaten-Modell darf sich dabei nicht nur auf die

---

[4] Kuch, Stephanie, „DLA - Application Profile, Version 3". URL: https://wwik-prod.dla-marbach.de/line/images/f/f1/Application_profile_V3_en.pdf, Zugriff am 12.05.2023.

[5] METS (Metadata Encoding and Transmission Standard). URL: https://www.loc.gov/standards/mets/, Zugriff am 12.05.2023.

[6] MODS (Metadata Object Description Schema) Homepage. URL: https://www.loc.gov/standards/mods/, Zugriff am 12.05.2023.

[7] PREMIS (Preservation Metadata) Homepage. URL: https://www.loc.gov/standards/premis/, Zugriff am 12.05.2023.

[8] HLRS: Höchstleistungsrechenzentrum Homepage. URL: https://www.hlrs.de/, Zugriff am 12.05.2023.

[9] Dipl-Ing Projekt. URL: https://www.izus.uni-stuttgart.de/fokus/fdm-projekte/dipling/, Zugriff am 12.05.2023.

[10] EngMeta Metadatenschema. URL: https://www.izus.uni-stuttgart.de/fokus/engmeta/, Zugriff am 12.05.2023.

Beschreibung der Daten selbst beschränken, sondern muss immer auch dem Einsatz auf unterschiedlichen Systemen (Anforderung der Interoperabilität gemäß den FAIR-Prinzipien) gerecht werden.

Bibliografie

Cell Project, Taxonomies Definition. URL: https://cellproject.net/taxonomies-definition, Zugriff am 12.05.2023.

Deutsches Literatur Archiv (DLA) Marbach Homepage. URL: https://www.dla-marbach.de/, Zugriff am 09.05.2023.

Dipl-Ing Projekt. URL: https://www.izus.uni-stuttgart.de/fokus/fdm-projekte/dipling/, Zugriff am 12.05.2023.

EngMeta Metadatenschema. URL: https://www.izus.uni-stuttgart.de/fokus/engmeta/, Zugriff am 12.05.2023.

HLRS: Höchstleistungsrechenzentrum Homepage. URL: https://www.hlrs.de/, Zugriff am 12.05.2023.

Kuch, Stephanie, „DLA - Application Profile, Version 3“. URL: https://wwik-prod.dla-marbach.de/line/images/f/f1/Application_profile_V3_en.pdf, Zugriff am 12.05.2023.

METS (Metadata Encoding and Transmission Standard). URL: https://www.loc.gov/standards/mets/, Zugriff am 12.05.2023.

MODS (Metadata Object Description Schema). URL: https://www.loc.gov/standards/mods/, Zugriff am 12.05.2023.

PREMIS (Preservation Metadata). URL: https://www.loc.gov/standards/premis/, Zugriff am 12.05.2023.

SDC4Lit Homepage. URL: https://www.sdc4lit.de/, Zugriff am 09.05.2023.

EVOKS - Benutzerfreundliche Erstellung kontrollierter Vokabulare für die Geisteswissenschaften

Ernst, Felix
felix.ernst[at]kit.edu
Karlsruher Institut für Technologie, Karlsruhe, Deutschland
ORCID-iD: 0000-0002-2102-4170

Frank, Laura
laura.frank[at]kit.edu
Karlsruher Institut für Technologie, Karlsruhe, Deutschland
ORCID-iD: 0000-0001-6286-2771

Götzelmann, Germaine
germaine.goetzelmann[at]kit.edu
Karlsruher Institut für Technologie, Karlsruhe, Deutschland
ORCID-iD: 0000-0003-3974-3728

Eckhardt, Klara
klara.eckhardt[at]student.kit.edu
Karlsruher Institut für Technologie, Karlsruhe, Deutschland

Maly, Jan
ufrum[at]student.kit.edu
Karlsruher Institut für Technologie, Karlsruhe, Deutschland

Preker, Yannis
uvsfa[at]student.kit.edu
Karlsruher Institut für Technologie, Karlsruhe, Deutschland

Scholz, Jonas
jonas.scholz2[at]student.kit.edu
Karlsruher Institut für Technologie, Karlsruhe, Deutschland

Zusammenfassung. EVOKS ist ein Werkzeug zur benutzerfreundlichen, kollaborativen Erstellung, Bearbeitung und Veröffentlichung von Wissensgraphen im SKOS-Format durch Fachwissenschaftler:innen ohne tiefe Vorkenntnisse in Ontologieentwicklung. Bei der Entwicklung wurden die FAIR-Prinzipien beachtet. Hieraus folgt die Verwendung von standardisierten Schnittstellen, Datenmodellen und Protokollen sowie Persistenz durch feste IDs und eine Versionierung. Durch eine Nutzer:innen- und Gruppenverwaltung wird ein einfacher Reviewprozess ermöglicht sowie die Urheberschaft aller erstellen Inhalte sichergestellt.

EVOKS wird bereits in verschiedenen, größtenteils geisteswissenschaftlichen Forschungsprojekten genutzt bzw. erprobt. Daher ist das Ziel des Posters, EVOKS der Forschungsgemeinschaft vorzustellen und wertvolles Feedback zu erhalten, um die Software weiterzuentwickeln.

Die Anwendung semantischer Werkzeuge wie kontrollierter Vokabulare, Taxonomien und Thesauri wird mit steigender Datenmenge und zunehmender Komplexität in den digitalen Geisteswissenschaften immer wichtiger, um neue Forschungsergebnisse zu gewinnen[1]. Zur Repräsentation kontrollierter Vokabulare eignet sich das weit verbreitete SKOS-Format[2], welches vor allem aufgrund seiner Einfachheit zum Aufbau FAIRer Vokabulare vorgeschlagen wurde[3]. Es ermöglicht semantische Beziehungen zwischen Konzepten[4] wie 'broader', 'narrower' oder 'related' sowie mehrsprachige Bezeichner.
EVOKS[5] (**E**ditor for **Vo**cabularies to **K**now **S**emantics) ist eine Open Source Software zur kollaborativen, benutzerfreundlichen Erstellung, Bearbeitung und Veröffentlichung von SKOS-Wissensgraphen. Es stellt einen wichtigen Baustein bei der Wissensanreicherung und FAIRification[6] von Daten und Metadaten durch semantische Methoden dar.

Alleinstellungsmerkmal von EVOKS ist im Gegensatz zu bereits existierender Software, dass der Fokus bei Softwareentwicklung auf Anwender:innen innerhalb der Geisteswissenschaften gelegt wurde. Somit können ohne tiefe Kenntnisse der Ontologieentwicklung rasch komplexe Wissensgraphen kollaborativ erstellt werden. Es können entweder bestehende Wissensgraphen im SKOS-Format importiert oder neue Wissensgraphen angelegt werden (Abb. 1).
Diese lassen sich unkompliziert auf Knopfdruck im Browser- und Publikationswerkzeug SKOSMOS[7] veröffentlichen, welches eine standardisierte Schnittstelle zur Weiterverwendung der Daten bietet[8] ('accessible'). Hierdurch können die generierten Thesauri der gesamten Forschungsgemeinschaft im SKOS-Format ('interoperable') und mit

---

[1] Hyvönen (2020)
[2] https://www.w3.org/TR/skos-reference/
[3] Cox u. a. (2021)
[4] Konzepte sind in SKOS lose definiert als eine Idee, Vorstellung oder Gedankeneinheit.
[5] https://github.com/ffeelliixx/EVOKS
[6] https://www.go-fair.org/fair-principles/fairification-process/
[7] https://skosmos.org/
[8] Suominen u. a. (2015)

fester ID ('findable') zur Verfügung gestellt werden. Eine Versionierung sorgt dafür, dass URL-Verweise auf einzelne Terme veröffentlichter Wissensgraphen persistent sind. Somit können die Wissensgraphen in einem kontinuierlichen Prozess überarbeitet werden, ohne dass es zu einem nicht aufzulösenden URL-Verweis auf obsolet gewordene Terme kommen kann.

Thesauri sind nicht nur als Mittel zur Beschreibung von Forschungsdaten von Bedeutung, sondern zweifelsohne auch eine wissenschaftliche, schützenswerte Leistung, mit welcher Forschungsfragen beantwortet werden können. Vor allem bei kollaborativer Erstellung eines Thesaurus stellt sich die Frage der Provenienz der einzelnen Einträge. EVOKS löst dies durch eine Nutzer:innenverwaltung. Bei Erstellung von Inhalten werden automatisiert zugehörige Metadaten erstellt, wodurch den Nachnutzer:innen klare Informationen zu Lizenz- und Urheberschaft ('reusable') zur Verfügung gestellt werden. Durch eine Gruppenverwaltung ist es möglich, Nutzer:innen verschiedene Rollen zuzuweisen, um beispielsweise einen Reviewprozess der erstellten Inhalte zu ermöglichen, ohne dass es zu Änderung am Wissensgraphen kommt.

Durch die Nutzung von EVOKS sind bereits erste Forschungsergebnisse entstanden. Bei Teilprojekten des DFG-geförderten *Sonderforschungsbereich 980 - Episteme in Bewegung. Wissenstransfer von der Alten Welt bis in die Frühe Neuzeit*[9] wurde beispielsweise ein Metadaten-Vokabular für digitalisierte Überlieferungen der aristotelischen Schrift 'de interpretatione' entwickelt[10], ebenso für das Werk 'Atalanta fugiens' von Michael Maier. Beide waren und sind Basis für datengetriebene Analysen der Werke.
Beim BMBF-geförderten Verbundprojekt *Materialisierte Heiligkeit: Torarollen als kodikologisches, theologisches und soziologisches Phänomen der jüdischen Schriftkultur in der Diaspora* wird EVOKS zur Nomenklatur von spezifischen Charakteristika der untersuchten Torarollen verwendet[11].
Im Rahmen des NFDI-Konsortiums *Materialwissenschaft & Werkstofftechnik (NFDI-MatWerk)* wurde ein Akronym-Vokabular

---

[9] Söring u. a. (2019)
[10] Krewet u. a. (2022)
[11] Frank u. a. (2023)

entwickelt[12], was auch die disziplinübergreifende Nutzbarkeit von EVOKS unterstreicht.
Im *Sonderforschungsbereich 1475 - Metaphern der Religion. Religiöse Sinnbildung in sprachlichen Prozessen* befindet sich EVOKS aktuell noch in der Erprobungsphase.

Mit dem Poster wollen wir einerseits EVOKS der Forschungsgemeinschaft präsentieren. Anderseits wollen wir unserem Paradigma der nutzerzentrierten Software-Entwicklung treu bleiben und weiterhin möglichst viele potentielle Nutzer:innen der Fachgemeinschaft involvieren, um durch Rückmeldung EVOKS weiterzuentwickeln.

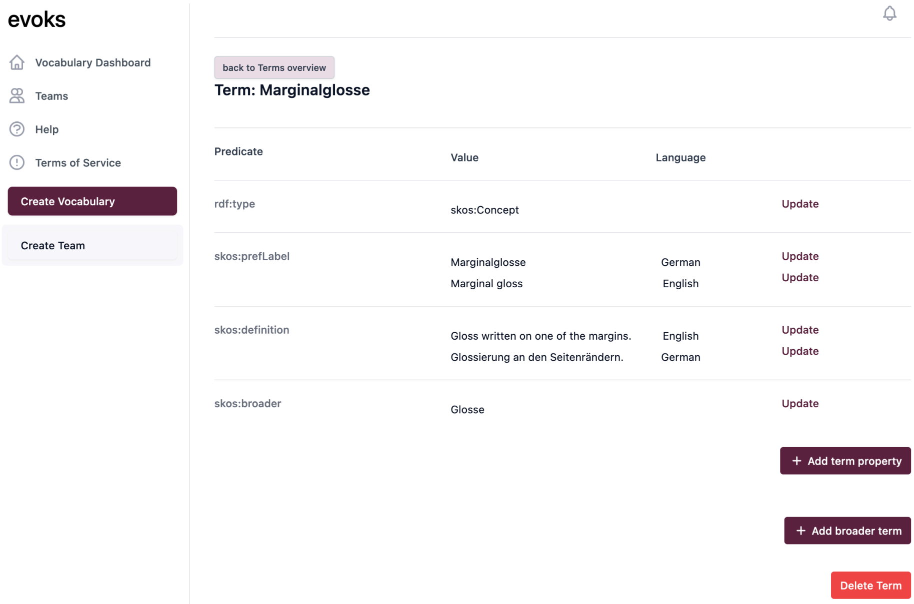


Abb. 1. Einzelterm-Ansicht in EVOKS.

Bibliografie

Abdildina, Gulzaure, Felix Ernst, Rossella Aversa, und Philipp Ost. „A Controlled Vocabulary for Acronyms of NFDI-MatWerk Using the Vocabulary Service EVOKS". Siegburg, Germany, 2023. https://doi.org/10.5445/IR/1000160373.

[12] Abdildina u. a. (2023)

Cox, Simon J. D., Alejandra N. Gonzalez-Beltran, Barbara Magagna, und Maria-Cristina Marinescu. „Ten Simple Rules for Making a Vocabulary FAIR“. *PLOS Computational Biology* 17, Nr. 6 (16. Juni 2021): e1009041. https://doi.org/10.1371/journal.pcbi.1009041.

Frank, Laura, Dana Eichhorst, Rebecca Ullrich, Katharina Haddassah Wendl, Annett Martini, und Danah Tonne. „Schrifttradition digital: Rituell reine Torarollen in der jüdischen Diaspora“. Trier, Luxemburg, 10. März 2023. https://doi.org/10.5281/zenodo.7715864.

Hyvönen, Eero. „Using the Semantic Web in Digital Humanities: Shift from Data Publishing to Data-Analysis and Serendipitous Knowledge Discovery“. *Semantic Web* 11, Nr. 1 (31. Januar 2020): 187–93. https://doi.org/10.3233/SW-190386.

Krewet, Michael, Felix Ernst, Germaine Götzelmann, Philipp Hegel, Torsten Schenk, Sibylle Söring, und Danah Tonne. „Die Aktualität des Unzeitgemäßen“. Gehalten auf der DHd 2022 Kulturen des digitalen Gedächtnisses. 8. Tagung des Verbands „Digital Humanities im deutschsprachigen Raum“ (DHd 2022), Potsdam, 7. März 2022. https://doi.org/10.5281/zenodo.6328015.

Söring, Sibylle, Germaine Götzelmann, Philipp Hegel, Michael Krewet, und Danah Tonne. „An der Schnittstelle von Fach- und Informationswissenschaft: Das INF-Projekt des SFB 980 ‘‘Episteme in Bewegung. Wissenstransfers von der Alten Welt bis in die frühe Neuzeit”“. *Bausteine Forschungsdatenmanagement*, Nr. 2 (28. Oktober 2019): 89–95. https://doi.org/10.17192/bfdm.2019.2.8083.

Suominen, Osma, Henri Ylikotila, Sini Pessala, Mikko Lappalainen, Matias Frosterus, Jouni Tuominen, Thomas Baker, Caterina Caracciolo, und Armin Retterath. „Publishing SKOS Vocabularies with Skosmos“, 1. Juni 2015.

Der krönende Abschluss: Paläographische Besonderheiten im Kontext der automatischen Texterkennung

Frank, Laura
laura.frank[at]kit.edu
Karlsruher Institut für Technologie, Deutschland
ORCID-iD: 0000-0001-6286-2771

Ernst, Felix
felix.ernst[at]kit.edu
Karlsruher Institut für Technologie, Deutschland
ORCID-iD: 0000-0002-2102-4170

Götzelmann, Germaine
germaine.goetzelmann[at]kit.edu
Karlsruher Institut für Technologie, Deutschland
ORCID-iD: 0000-0003-3974-3728

Zusammenfassung. Die automatische Erkennung von Text aus Bildern wird immer leistungsstärker und zuverlässiger, nicht zuletzt durch den erfolgreichen Einsatz von Machine-Learning-Methoden. Dennoch liegt das Ziel auf der Umwandlung in reinen maschinenlesbaren Text. Allerdings verbergen sich in paläographischen und kodikologischen Details in Manuskripten und Drucken oftmals tiefliegende Bedeutungen, die bei automatischer Texterkennung nicht im Fokus stehen. Dieses Poster möchte diese Forschungslücke beleuchten und den Bedarf einer Erweiterung der bestehenden automatischen Texterkennung verdeutlichen, welche paläographische Details fokussiert. Die Ansätze für die technische Umsetzung einer solchen Erkennung sollen präsentiert werden. Ebenso sollen geisteswissenschaftliche Projekte mit ähnlichen Anwendungsgebieten auf das Vorhaben aufmerksam werden.

Die automatische Erkennung von Text aus Bildern ist nach wie vor ein weit genutztes Werkzeug in Wirtschaft, Wissenschaft und Verwaltung. Sie gewinnt durch fortschreitende Digitalisierungsprozesse in Bibliotheken und Archiven insbesondere im Bereich der Handwritten Text Recognition (HTR) in den Geisteswissenschaften an immer größerer Bedeutung. Zudem bietet der Einsatz fortschrittlicher Machine-Learning-Methoden wie Recurrent Neural Networks und Long Short-Term Memory Networks in der automatischen Texterkennung eine

höhere Genauigkeit der Ergebnisse.[1] Im gleichen Zuge steigt die Vielfalt an handschriftlichen Stilen und zugrundeliegenden Sprachen der vortrainierten Models sowie die Auswahl an modernen HTR-Tools wie beispielsweise eScriptorium[2]. Diese ermöglichen die Transkription von digitalisierten Manuskripten in maschinenlesbaren Text und bieten auch für komplexere und alte Handschriften einen Zugang zu deren Inhalt.
Die traditionelle HTR zielt dabei auf die Erkennung des Texts, allerdings ist oftmals nicht nur der reine Inhalt entscheidend. Vielmehr liegen in möglichen kodikologischen und paläographischen Details bedeutungsvolle Informationen versteckt, die den Inhalt um wertvolle Aspekte ergänzen, welche allerdings bei den aktuellen HTR-Tools gänzlich verloren gehen.
Als Beispiel lassen sich Buchstabendekorationen in jüdischen Torarollen nennen. In diesen handschriftlichen Meisterwerken sind die hebräischen Buchstaben oftmals mit zusätzlichen Kronen (*tagin*), Schnörkeln oder Flaggen geschmückt, die so genannten *otijjot meshunnot*. Die Verzierungen werden vom Schreibenden nicht willkürlich gesetzt, sondern verbergen tieferliegende Botschaften[3]. Insbesondere Torarollen aus dem mittelalterlichen aschkenasischen Raum weisen eine breite Vielfalt an Dekorationen auf, welche nach jüdischer Tradition nicht vorgesehen sind. Die Wichtigkeit dieser besonders dekorierten Buchstaben spiegelt sich in der vielfältigen Schreiberliteratur wider, in welcher diese diskutiert und ausgelegt werden.[4]
Obwohl die Verwendung von HTR-Programmen zur Transkription von Bilddigitalisaten von Torarollen dank fortschrittlicher Machine-Learning-Methoden zu guten Textergebnissen führt, werden dabei die feinen Dekorationen vollständig vernachlässigt oder stellen sogar ein Hindernis der korrekten Buchstabenerkennung dar. Selbst wenn optische Besonderheiten in der Erkennung miteinbezogen werden, ist die Darstellung im Output durch fehlende maschinenlesbare Repräsentation von beispielsweise verzierten Buchstaben unklar.
Dementsprechend lässt sich die Schlussfolgerung ziehen, dass in Ergänzung zur HTR ein zusätzlicher Forschungsbedarf mit Blick auf paläographischen Merkmalen besteht. Andere Felder wie die Automatic Handwriting Identification streifen diese Forschungslücke, fokussieren aber Spezifika des Schreibprozesses statt der Bedeutung der verwendeten Dekorationen. Ein kritischer Blick auf die bestehenden

[1] Memon et al. 2020.
[2] Kiessling et al. 2019.
[3] Martini 2022.
[4] Perani 2022.

Methoden zeigt somit, dass eine konzeptionelle Abzweigung aus dem bestehenden HTR-Workflow nötig ist, um eine übergreifende Betrachtung verschiedener Aspekte zu ermöglichen.
Diesen neuen Fokus als Feld zwischen HTR und Automatic Handwriting Identification wollen wir in Angriff nehmen und im Rahmen einer Dissertation in der Informatik eine technische Lösung entwickeln. Dabei sollen sowohl Aspekte aus der HTR genutzt und vertieft als auch neue Methoden integriert werden. Mit diesem Poster wollen wir unsere konzeptionellen Ansätze präsentieren. Des Weiteren erhoffen wir uns einen aktiven Austausch mit anderen Forschenden, um Projekte mit ähnlichem Anwendungsgebiet kennenzulernen.

Bibliografie

Kiessling, Benjamin & Tissot, Robin & Stokes, Peter & Stoekl Ben Ezra, Daniel. “eScriptorium: An Open Source Platform for Historical Document Analysis.” 19-19, 2019. 10.1109/ICDARW.2019.10032.

Martini, Annett. "9 Die Bedeutung der Buchstaben, tagin und otijjot mešunnot in der aschkenasischen Schriftauslegung des Mittelalters" In ›Arbeit des Himmels‹: Jüdische Konzeptionen rituellen Schreibens in der europäischen Kultur des Mittelalters, 246-266. Berlin, Boston: De Gruyter, 2022. https://doi.org/10.1515/9783110722062-009

Memon, Jamshed & Sami, Maira & Khan, Rizwan & Uddin, Mueen. “Handwritten Optical Character Recognition (OCR): A Comprehensive Systematic Literature Review (SLR).”, 2020. IEEE Access. 1-1. 10.1109/ACCESS.2020.3012542.

Perani, Mauro. "Chapter 11 The Tagin: Their Origin, Use, and Oscillating Evolution between Embellishment and Mystical Signifier. New Light from the Ancient Bologna Sefer Torah". In *The Hebrew Bible Manuscripts: A Millennium*, Leiden, Niederlande: Brill, 2022. doi: https://doi.org/10.1163/9789004499331_012

Präsentation archivierter Werke der Literatur im Netz

Erfahrungen zur Wiedergabe von WARCs im Projekt SDC4Lit

Ganzenmüller, Andreas
andreas.ganzenmueller[at]hlrs.de
Höchstleistungsrechenzentrum Stuttgart (HLRS), Universität Stuttgart, Deutschland
ORCID-iD: 0009-0002-3451-7334

Kushnarenko, Volodymyr
volodymyr.kushnarenko[at]hlrs.de
Höchstleistungsrechenzentrum Stuttgart (HLRS), Universität Stuttgart, Deutschland
ORCID-iD: 0000-0001-7427-2410

Buck, Nina
nina.buck[at]hlrs.de
Höchstleistungsrechenzentrum Stuttgart (HLRS), Universität Stuttgart, Deutschland
ORCID-iD: 0000-0002-4651-6040

Ulrich, Mona
mona.ulrich[at]dla-marbach.de
Deutsches Literaturarchiv Marbach (DLA), Deutschland
ORCID-iD: 0000-0001-9591-5614

Jung, Kerstin
eckartkn[at]ims.uni-stuttgart.de
Institut für Maschinelle Sprachverarbeitung (IMS), Universität Stuttgart, Deutschland
ORCID-iD: 0000-0002-9548-8461

Bönisch, Thomas
boenisch[at]hlrs.de
Höchstleistungsrechenzentrum Stuttgart (HLRS), Universität Stuttgart, Deutschland
ORCID-iD: 0000-0003-3108-8597

Zusammenfassung. Am Beispiel der Erfahrungen aus dem Projekt SDC4Lit soll gezeigt werden, welcher Aufwand mitunter nötig ist, wenn Datenbestände nicht nur in einem Repositorium abgelegt werden sollen,

sondern auch eine datenspezifische Präsentation des Materials gewünscht wird, und was dies für den nachhaltigen Betrieb entsprechender Plattformen bedeutet.

Wenn neben der bloßen Bereitstellung von (Forschungs-)Daten in einem Repositorium auch eine datenspezifische Präsentation gewünscht wird, steigt der mit dem Aufbau einer entsprechenden Plattform verbundene Aufwand oft sprunghaft an, und hat auch Auswirkungen auf den langfristigen Betrieb derartiger Dienste. Exemplarisch seien hier deshalb die Erfahrungen im Umgang mit Literatur im Netz aus dem Projekt SDC4Lit[1] geschildert.

SDC4Lit hat es sich u.a. zum Ziel gesetzt eine Plattform aufzubauen, auf welcher archivierte Werke der Literatur im Netz sowohl der wissenschaftlichen Gemeinschaft als auch der interessierten Öffentlichkeit zugänglich gemacht werden sollen. Dazu wird mit Sammlungsmaterial des Deutschen Literaturarchiv Marbach[2] (DLA) gearbeitet. Bei den betreffenden Werken handelt es sich zumeist um gecrawlte und archivierte Webseiten, die im WARC-Format[3] vorliegen.

Im Rahmen des Projekts sollten netzliterarische Werke nicht nur als WARC-Dateien, die mitunter mehrere Gigabyte groß sein können, in einem Repositorium abgelegt und mit Metadaten beschrieben werden, sondern die Werke sollten auch in ihrer ursprünglichen Form präsentiert werden, so wie sie einst im Browser dargestellt wurden. Dadurch soll es Anwendern ermöglicht werden die literarischen Werke in ihrem ursprünglichen Erscheinungsbild betrachten zu können.

Es musste also eine technische Lösung erarbeitet werden, um die im Repositorium abgelegten WARCs wieder als Webseite darzustellen. Die Schwierigkeit hierbei lag darin, dass es sich bei den archivierten Webseiten technisch gesehen um alles Mögliche handeln konnte. Von modernen dynamischen Webseiten mit dahinterliegenden Datenbanken und Content-Management-Systemen, bis hin zu in Handarbeit selbstgebauten Webseiten aus HTML, CSS und JavaScript, mit integrierten externen Inhalten und veralteten Technologien (z.B. Flash), welche von modernen Browser nicht länger unterstützt werden. Für die Wiedergabe

[1] SDC4Lit Homepage. URL: https://www.sdc4lit.de/, letzter Zugriff am 09.05.2023.

[2] Deutsches Literaturarchiv Marbach (DLA) Homepage. URL: https://www.dla-marbach.de/, letzter Zugriff am 09.05.2023.

[3] WARC Spezifikation: URL: https://iipc.github.io/warc-specifications/specifications/warc-format/warc-1.0/, letzter Zugriff am 09.05.2023.

von WARCs werden daher spezielle Programme benötigt, beispielsweise *pywb*[4] [5] oder *SolrWayback*[6]. Und da in der wissenschaftlichen Gemeinschaft häufig *pywb* und *SolrWayback* parallel eingesetzt werden, wurde in SDC4Lit beschlossen, beide WARC-Player als ergänzende Online-Dienste zu betreiben.

Die Herausforderung lag nun darin, eine Schnittstelle zu entwickeln und zu implementieren, über welche *pywb* und *SolrWayback* an das Repositorium angebunden werden. Die Anwendungen *pywb* und *SolrWayback* laufen dabei in ihren je eigenen virtuellen Maschinen, verfügen aber über einen gemeinsamen Speicherbereich (um unnötige redundante Datenhaltung zu vermeiden), auf dem die entpackten WARC-Inhalte für die Wiedergabe abgelegt werden. Die Anbindung an das Repositorium wurde dabei mit Python-Skripten realisiert, welche über die API der eingesetzten Repositoriums-Lösung (Dataverse[7]) den Datenbestand im Repositorium mit dem der Wiedergabe abgleichen, hinzugekommene Werke, sofern diese für die Veröffentlichung freigegeben sind, herunterladen und im Speicherbereich der Wiedergabe entpacken. Abschließend werden die hinzugekommenen Werke noch von *pywb* und *SolrWayback* indiziert, damit auch die Volltext-Suche innerhalb der Werke möglich ist.

Eine derartige Präsentation der Daten ist für die Anwender der Plattform ohne Zweifel ein Gewinn. Doch es gilt zu bedenken, dass durch die erforderliche Infrastruktur (Hardware, aber vor allem auch die eingesetzten Softwarekomponenten), um diese zusätzlichen Dienste betreiben zu können, der zukünftige Wartungsaufwand der ganzen Plattform doch in einem nicht unerheblichen Maße gestiegen ist. Im Hinblick auf den längerfristigen Betrieb derartiger Infrastruktur, müssen daher unbedingt auch entsprechende Ressourcen (inklusive Personal) eingeplant werden.

---

4 GitHub: webrecoder/pywb. URL: https://github.com/webrecorder/pywb/, letzter Zugriff am 09.05.2023.

5 Webrecorder pywb Dokumentation. URL: https://pywb.readthedocs.io/en/latest/, letzter Zugriff am 09.05.2023.

6 GitHub: netarchivesuite/solrwayback. URL: https://github.com/netarchivesuite/solrwayback/, letzter Zugriff am 09.05.2023.

7 The Dataverse Project Homepage. URL: https://dataverse.org/, letzter Zugriff am 09.05.2023.

Bibliografie

SDC4Lit Homepage. URL: https://www.sdc4lit.de/, letzter Zugriff am 09.05.2023.

Deutsches Literaturarchiv Marbach (DLA) Homepage. URL: https://www.dla-marbach.de/, letzter Zugriff am 09.05.2023.

WARC Spezifikation: URL: https://iipc.github.io/warc-specifications/specifications/warc-format/warc-1.0/, letzter Zugriff am 09.05.2023.

GitHub: webrecoder/pywb. URL: https://github.com/webrecorder/pywb/, letzter Zugriff am 09.05.2023.

Webrecorder pywb Dokumentation. URL: https://pywb.readthedocs.io/en/latest/, letzter Zugriff am 09.05.2023.

GitHub: netarchivesuite/solrwayback. URL: https://github.com/netarchivesuite/solrwayback/, letzter Zugriff am 09.05.2023.

The Dataverse Project Homepage. URL: https://dataverse.org/, letzter Zugriff am 09.05.2023.

Research Data Management for Arts and Humanities: Integrating Voices of the Community

The Brand-new, Collective Publication of the DARIAH Working Group on Research Data Management

Gelati, Francesco
francesco.gelati[at]uni-hamburg.de
Universität Hamburg, Deutschland
ORCID-iD: 0000-0002-6066-1308

Wuttke, Ulrike
ulrike.wuttke[at]fh-potsdam.de
Fachhochschule Potsdam, Deutschland
ORCID-iD: 0000-0002-8217-4025

Gietz, Peter
p.gietz[at]daasi.de
DAASI International GmbH, Tübingen, Deutschland

Summary. The poster showcases the publication "Research Data Management for Arts and Humanities: Integrating Voices of the Community" issued in 2023 by the DARIAH (Digital Research Infrastructure for the Humanities and Arts) Working Group on Research Data Management. The collective, open-access volume brings together case studies and consolidated experiences with a strong pan-European approach.

A frequently voiced concern about the Open Science paradigm is that it has not had the same impact on the different disciplines. Research areas and researchers who are heavily dependent on third-party data are still facing complex legal, ethical, technical, infrastructural and interoperability challenges and their needs form a rather underrepresented blind spot in the open research culture. A particular Open Science challenge in research data workflows in the Arts and Humanities domain is their dependence on cultural heritage sources hosted and curated in museums, libraries, galleries and archives. The availability of digital (digitised and born-digital) Cultural Heritage Data is fundamental to research in many disciplines in the Arts and Humanities. Still, Cultural Heritage collections are usually not made available digitally with research/academic reuse in mind. A major difficulty when scholars interact with heritage data is that Cultural Heritage institutions, universities and other research-performing organisations are embedded into very different legal, funding, structural and organisational frameworks.

DARIAH (Digital Research Infrastructure for the Humanities and Arts) is a permanent research infrastructure financed by the European Union which supports digitally-powered research and teaching across the Arts and Humanities (see: https://www.dariah.eu/about/dariah-in-nutshell/). It includes institutions and national networks in the whole European continent and beyond. In DARIAH, several working groups are active. The Working Group Research Data Management (see: https://www.dariah.eu/activities/working-groups/research-data-management/) is composed of volunteering members who are Research Data Management professionals or scholars. Members who have so far

taken part in the working group's activities with continuity are active in the following countries: Belgium, Germany, The Netherlands, Poland, Switzerland.
The working group organised a hybrid (both on-premise and remote) writing session in June 2022 in Warsaw (Poland) with the aim of describing aforementioned challenges, interactions, along with a variety of situations and approaches. The outcome is the publication "Research Data Management for Arts and Humanities: Integrating Voices of the Community".
It brings together case studies and consolidated experiences from the members of the group about:

- How certain Arts and Humanities and Cultural Heritage institutions developed capacities for data support;
- Ways in which Cultural Heritage professionals can be efficiently involved in open and sustainable Arts and Humanities data workflows;
- How to facilitate the reuse, dissemination and solidification of researcher-friendly, FAIR-by-design curation practices of arts and humanities research data, including also sensitive data;
- How multilingualism can be supported throughout this work;
- How to solve the problem of long-term digital preservation.

Furthermore, the publication outlines, with a comparative approach, research data management policies issued by the European Union as well as by some European countries (Belgium, France, Germany and Poland), including information on very recent developments in France and Germany.

The publication is available both as a dynamic GitLab project (https://gitlabce.rrz.uni-hamburg.de/uahh-digitaledienste/rdm-for-arts-and-humanities) and as a static PDF (DOI: 10.5281/zenodo.8059626). It will be soon published in the online book series DARIAH-DE Working Papers.

Acknowledgments: the publication (to be published Open Access in 2023) has been made possible by the third DARIAH Working Groups Funding Scheme Call for the years 2021-2023 and the resulting grant administered by the Digital Humanities Centre at the Institute of Literary Research of the Polish Academy of Sciences (IBL PAN).

F wie Registry

Die Text+ Registry als Hilfsmittel zur Auffindbarkeit von Ressourcen

Genêt, Philippe
p.genet[at]dnb.de
Deutsche Nationalbibliothek, Deutschland
ORCID-iD: 0009-0001-5095-8052

Gradl, Tobias
tobias.gradl[at]uni-bamberg,de
Universität Bamberg, Deutschland
ORCID-iD: 0000-0002-1392-2464

Hensen, Kilian
kilian.hensen[at]uni-koeln.de
CCeH, Universität zu Köln, Deutschland
ORCID-iD: 0000-0001-6708-1237

Kudella, Christoph
kudella[at]sub.uni-goettingen.de
Niedersächsische Staats- und Universitätsbibliothek Göttingen, Deutschland
ORCID-iD: 0000-0002-9645-7122

Schulz, Daniela
schulz[at]hab.de
Herzog August Bibliothek Wolfenbüttel, Deutschland
ORCID-iD: 0000-0003-3167-5089

Zusammenfassung. Im Kontext des NFDI-Konsortiums Text+ entsteht mit der Registry ein übergreifendes Verzeichnis, in dem Ressourcen verschiedener Datendomänen erfasst und vernetzt werden. Die Registry speist sich aus verschiedenen Datenquellen, geht aber in ihrem Ansatz der zentralen Verzeichnung unterschiedlicher Ressourcentypen über bestehende Angebote hinaus. Die Findability von Ressourcen spielt auf mehreren Ebenen eine Rolle. Als zentrales Verzeichnissystem erhöht die Registry die Auffindbarkeit von Ressourcen, diese – oder zumindest deren Metadaten – müssen für eine Aufnahme aber erst identifiziert werden. Hier wird ein Community-basierter Ansatz verfolgt. Im Posterbeitrag sollen Herausforderungen, Möglichkeiten aber auch Grenzen der Registry reflektiert und der Stand der Arbeiten vorgestellt werden.

1 Ziele und Scope

Im Angebotsportfolio des NFDI-Konsortiums Text+[1] stellt die Registry als zentrales, datendomänenübergreifendes Recherche- und Informationsinstrument eine wichtige Komponente dar. Ressourcen der Datendomänen lexikalische Ressourcen, Sammlungen und Editionen werden darin erfasst, miteinander in Beziehung gesetzt und vernetzt (Abb. 1).

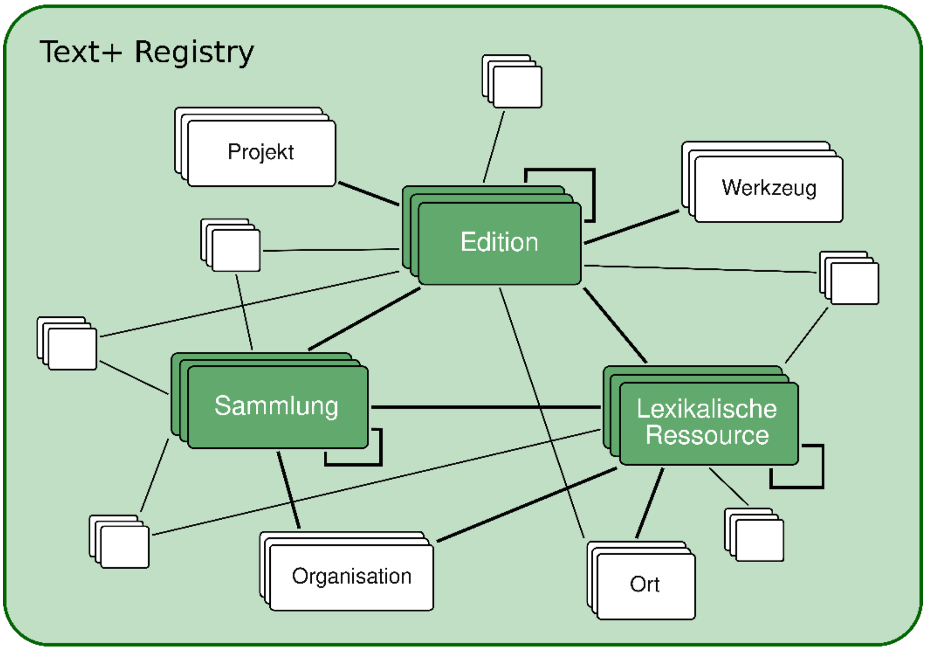


Abb. 1. Vernetzte und kontextualisierte Ressourcen in Text+

Der Mehrwert gegenüber bestehenden Nachweissystemen liegt im übergreifenden Ansatz:

- Datengeber:innen können durch die Integration in die Registry die Sichtbarkeit ihrer Ressourcen und damit deren Findability erhöhen,
- für Nutzende stellt die Registry einen zentralen Zugangspunkt zu vorhandenen Ressourcen dar.

Die Registry möchte bestehende Verzeichnisse und Kataloge dabei nicht ersetzen, sondern durch das Angebot eines forschungsunterstützenden ‘one stop shop’ ergänzen. Schnittstellen werden genutzt, über-

[1] Vgl. für weiterführende Informationen die Webseite des Konsortiums. URL: https://www.text-plus.org/ [Zugriff: 14.07.2023].

gebene Daten erweitert und damit die Anschlussfähigkeit an die übergreifenden Strukturen und Angebote der NFDI insgesamt (z.B. Wissensgraphen) ermöglicht.

2 Die Registry als technische Komponente

Die Registry geht als grundlegende Weiterentwicklung aus der DARIAH-DE Collection Registry[2] hervor. Die Weiterentwicklung erfolgt insbesondere entlang dreier Schwerpunkte:

Datenmodelle sind je Datendomäne an unterschiedlichen Anforderungen ausgerichtet und wurden basierend auf einschlägigen fachlichen Standards und Best Practices, sowie auch in direkter Zusammenarbeit mit den Communities (z.B. Fachinformationsdiensten) entwickelt.[3] Mappings ermöglichen eine übergreifende Recherche.

Durch die Formalisierung von Beziehungen zwischen Sammlungen, Editionen und lexikalischen Ressourcen, aber auch Werkzeugen, Akteuren und weiteren Entitätstypen entsteht ein graphartiges **Ressourcennetzwerk**, das auch über Text+ hinaus anschlussfähig ist.

Die Registry wird als zentrale Infrastrukturkomponente für den Import, die Kuration und die Bereitstellung von Ressourcenbeschreibungen eingerichtet. Weitere Such- und Recherchewerkzeuge wie die CLARIN Federated Content Search (FCS) und die DARIAH-DE Generische Suche (GS)[4], übergreifende Dienste wie das Webportal (Abb. 2) sowie spezifische Dienste können auf dieses zentrale Verzeichnis aufsetzen und auf eine redundante Datenhaltung verzichten. **Schnittstellen** für den Ingest aus bestehenden Verzeichnissen und auch deren Bereitstellung bilden damit einen weiteren Schwerpunkt.

---

[2] Vgl. z. B. die DARIAH Instanz, URL: https://colreg.de.dariah.eu oder die Instanz des CLARIAH-DE Tutorial Finders, URL: https://teaching.clariah.de/colreg-ui/ [Zugriff: 14.07.2023].

[3] Einen konzisen Einblick in den Gegenstandsbereich und spezifische Herausforderungen der Registry der Datendomäne “Editionen” bietet Gradl, Kudella, und Schulz 2021. Vorarbeiten für die Erfassung von Editionen wie die sog. Editionsmatrix “EdMa” stammen u.a. aus dem CLARIAH-DE-Projekt (2019-2021). Vgl. hierzu Schulz, Fisseni, und Sendler 2021.

[4] Vgl. zu den Such- und Recherchediensten und deren Zusammenwirken Eckart et al. 2021.

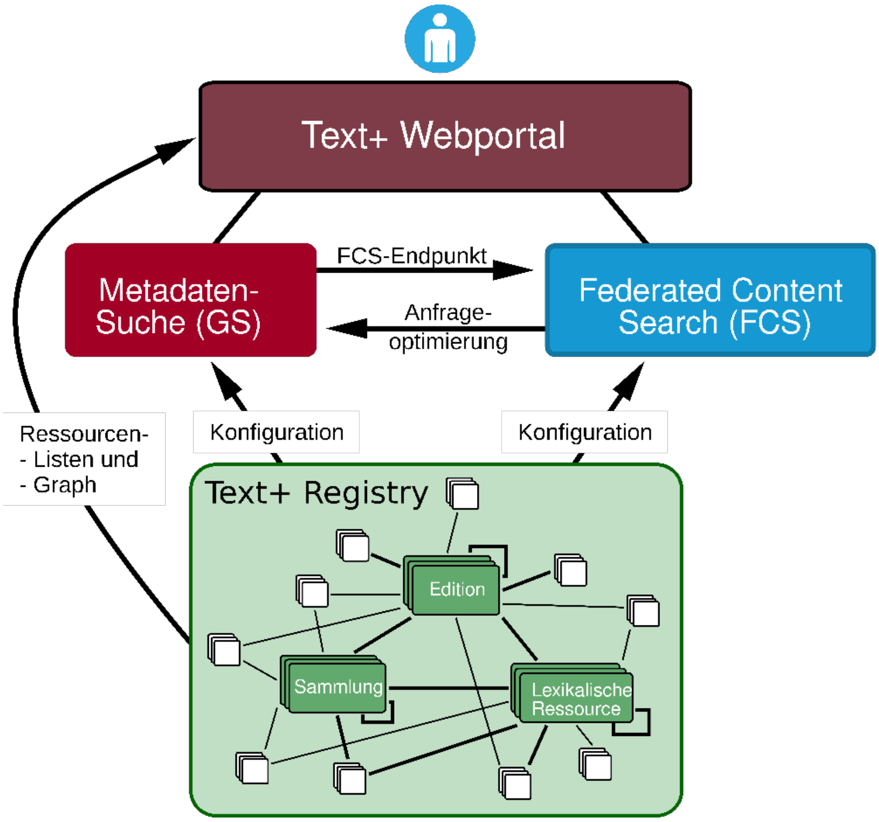


Abb. 2. Zusammenspiel der Such- und Recherchewerkzeuge von Text+

3 Nutzen der Registry

Die Registry ermöglicht Forschenden eine schnelle und treffsichere Suche nach relevanten Sprach- und Textdaten verschiedener Datenzentren und -quellen. Zusätzlich stellen Querverweise und Beziehungen zwischen den Ressourcen – sowohl innerhalb als auch jenseits der einzelnen Datendomäne – diese in einen breiteren Kontext. Solchermaßen ist die Registry vor allem der Findability der FAIR-Prinzipien[5] verschrieben.

Eine intuitive Nutzer:innenführung sowie ein zweckdienliches und hinreichend nachvollziehbares Set an Mindestaufnahmekriterien sollen

[5] Vgl. Wilkinson et al. 2016.

dazu beitragen, in enger Zusammenarbeit mit der Community möglichst viele Ressourcen in unterschiedlicher Tiefe zu verzeichnen.[6]

Bibliografie

Eckart, Thomas, Gradl, Tobias, Jegan, Robin, Margaretha, Eliza, Werthmann, Antonina, Helfer, Felix, Buddenbohm, Stefan (2021), „CLARIAH-DE Cross-Service Search: Prospects and Benefits of Merging Subject-specific Services” (DARIAH-DE Working Paper), https://nbn-resolving.org/urn:nbn:de:gbv:7-dariah-2021-1-9

Gradl, Tobias, Christoph Kudella, Daniela Schulz (2021), „Die Registry der TA Editions – Datenmodell und Prototyp” (Poster), https://doi.org/10.5281/zenodo.7244134.

Schulz, Daniela, Bernhard Fisseni, Simon Sendler (2021), „EdMA – Eine Matrix zur Erfassung und Kategorisierung digitaler Editionen“ (CLARIAH-DE Arbeitsbericht 5), https://doi.org/10.14618/ids-pub-10501.

Wilkinson, Mark D., Michel Dumontier, IJsbrand Jan Aalbersberg, Gabrielle Appleton, Myles Axton, Arie Baak, Niklas Blomberg, et al. „The FAIR Guiding Principles for Scientific Data Management and Stewardship.“ Scientific Data 3, no. 1 (March 15, 2016): 160018. https://doi.org/10.1038/sdata.2016.18.

[6] Anreize werden u.a. durch die Bereitsstellung von Fördermitteln für Kooperationsprojekte geschaffen. Vgl. die Informationen zur jeweils aktuellen Ausschreibungsrunde URL: https://www.text-plus.org/forschungsdaten/kooperationsprojekte/ [Zugriff: 14.07.2023].

Wie FDM einen Beitrag zur Data Literacy Education leisten kann

Erfahrungsbericht zur Verbesserung der Data Literacy an der Universität Hamburg

Jacob, Juliane
juliane.jacob[at]uni-hamburg.de
Universität Hamburg, Deutschland
ORCID-iD: 0000-0002-0443-3570

Zusammenfassung. Data Literacy, also der kritisch-reflexive Umgang mit Daten, ist eine Grundkompetenz für den inhalts- und sinnvollen Umgang mit Daten und Datenmustern. 2018 entstand an der Universität Hamburg ein Data Literacy Education Netzwerk, das seitdem einen Beitrag zur Ermöglichung eines daten- und informationsbezogen selbstbestimmten Umgang mit Daten in Alltag, Beruf und Wissenschaft schafft. Das Zentrum für nachhaltiges Forschungsdatenmanagement bietet verschiedene Schulungskonzepte an und baut diese stetig aus. Welche Erkenntnisse und Herausforderungen bei der Implementierung auftreten, wird in Form eines Erfahrungsberichts aufbereitet.

Die Universität Hamburg (UHH) hat ca. 6.500 Professor_innen und wissenschaftliche Mitarbeiter_innen, sowie ca. 5.700 Doktorand_innen und ca. 43.500 Student_innen. Ihre Datenkompetenzen sind so divers wie die acht Fakultäten und diversen Einrichtungen[1]. Vor dem Hintergrund einer umfassenden Datafizierung und Digitalisierung hat sich 2018 an der UHH ein Data Literacy Education[2] (DLE) Netzwerk[3] konsolidiert, an dem auch das Zentrum für nachhaltiges Forschungsdatenmanagement (ZFDM)[4] beteiligt ist.

Das DLE-Mission Statement umfasst, dass ein fakultätsübergreifender, informeller Zusammenschluss von Lehrenden einen inhaltlich, methodisch, technisch und lebenspraktisch angemessenen Umgang mit Daten vermittelt, um die Frage „was?“ (anwendungsbezogen), „wie?“ (technisch bezogen) und „weshalb?“ (gesellschaftlich-kulturell) mit Daten erzeugt werden soll, beantworten zu können. Das DLE-Netzwerk fördert dadurch einen daten- und informationsbezogen selbstbestimmten Handeln mit Daten in Alltag, Beruf und Wissenschaft. Neben

---

[1] https://www.uni-hamburg.de/uhh/profil/fakten.html (Zugriff am 08.05.2023)
[2] https://www.stifterverband.org/data-literacy-education (Zugriff am 08.05.2023)
[3] https://www.uni-hamburg.de/dle-netzwerk.html (Zugriff am 08.05.2023)
[4] https://www.fdm.uni-hamburg.de (Zugriff am 18.07.2023)

Grundkompetenzen für einen kritisch-reflexiven Umgang mit Daten ist auch Fachexpertise notwendig, denn Daten und vermeintliche Muster müssen bedeutungs- oder sinnvoll interpretiert werden. Technologische (Datenerzeugung, -aufbewahrung und -verarbeitung) und methodische Kompetenzen (Statistik und Wahrscheinlichkeitsrechnung, inkl. deren erkenntnistheoretischer Befragung) verbessern zudem die individuelle Data Literacy.

Das DLE-Netzwerk der UHH bietet konkrete Lehr- und Lernangebote an und entwickelt diese stetig weiter.

Darüber hinaus fördert die Stiftung Innovation in der Hochschullehre[5] für eine Laufzeit von drei Jahren das Projekt „Digital and Data Literacy in Teaching Lab" (DDLitLab)[6]. Das ZFDM hat darin zwei geförderte Teilprojekte, um Kompetenzen in Forschungsdatenmanagement (FDM) strukturiert zu vermitteln. Das erste Projekt („Early Education in Data Management Decisions", Laufzeit April 2022 bis März 2023[7][8]) implementierte einen Workshop[9] für Student_innen, der inhaltlich in Anlehnung an die Lernzielmatrix[10] und methodisch-didaktisch an den Train-the-Trainer-Kurs zum Thema FDM[11] konzipiert wurde. Hinzu kommt eine Begleitforschung in Form von Lehrevaluationen und Interviews von Lehrenden. Das zweite bewilligte Projekt („Early Education in Data Management Decisions an adapted course", Laufzeit April 2023 bis März 2024[12]) erweitert das Portfolio um vier spezifische Module, die von Lehrenden niedrigschwellig gebucht werden können. Neben dem FDM-Grundlagen-Modul gibt es eines für Datenqualität und Bereinigung

[5] https://stiftung-hochschullehre.de (Zugriff am 18.07.2023)

[6] https://www.isa.uni-hamburg.de/ddlitlab.html (Zugriff am 08.05.2023)

[7] https://www.isa.uni-hamburg.de/ddlitlab/data-literacy-lehrlabor/erste-foerderrunde.html (Zugriff am 08.05.2023)

[8] https://www.fdm.uni-hamburg.de/ueber-uns/projekte/ddlitlab.html (im Aufbau, Zugriff am 08.05.2023)

[9] Jacob, Juliane, Neumann, Jana, & Schulz, Sandra. (2023, March). Forschungsdatenmanagement: Workshop für Studierende im DDLitLab-Projekt. http://doi.org/10.25592/uhhfdm.9583 (Zugriff am 18.07.2023)

[10] Petersen, Britta, Engelhardt, Claudia, Hörner, Tanja, Jacob, Juliane, Kvetnaya, Tatiana, Mühlichen, Andreas, Schranzhofer, Hermann, et al. 2022. "Lernzielmatrix zum Themenbereich Forschungsdatenmanagement (FDM) für die Zielgruppen Studierende, PhDs und data stewards". Zenodo. https://doi.org/10.5281/zenodo.7034478.

[11] Biernacka, Katarzyna, Buchholz, Petra, Danker, Sarah Ann, Dolzycka, Dominika, Engelhardt, Claudia, Helbig, Kerstin, Jacob, Juliane, et al. 2021. Train-the-Trainer-Konzept zum Thema Forschungsdatenmanagement (Version 4). Zenodo. doi:10.5281/zenodo.5773203.

[12] https://www.isa.uni-hamburg.de/ddlitlab/data-literacy-lehrlabor/zweite-foerderrunde.html (Zugriff am 08.05.2023)

(Theorie und Praxis Python Pandas mit Jupyter Notebooks), ein weiteres Modul für Datenrecherche und Nachnutzung, inkl. rechtlicher Aspekte und ein Modul für FDM-Kolloquium für Abschlussarbeiten.
Die Schulungsangebote richten sich auch an weitere Interessengruppen wie Doktorand_innen, Post-Docs und thematisieren neben den genannten Inhalten auch die ZFDM-Services[13]. Es wird angestrebt, die Ergebnisse des gesamten DDLitLab-Projektes für die Nachnutzung zu veröffentlichen.
Dass FDM lange Zeit nicht im Studium vorgesehen war, ändert sich aktuell. FDM ist ebenso relevant, wenn noch keine oder kaum Daten vorliegen bzw. der Umfang der Daten verhältnismäßig gering ist, da es zu Zeitersparnis und Qualitätsverbesserung führen kann. In den Evaluationen und Interviews äußern dies Student_innen und auch betreuende Personen. Die größten Herausforderungen sind die Sichtbarkeit der Angebote, Bewertung der Relevanz von den Student_innen und die hohe Absage-/Fehlquote der Teilnehmenden. Scheinbar ist das unabhängig von ECTS, Umfang der Lehrveranstaltung, Tages-/Wochenzeit und Veranstaltungsort/-form (digital/Präsenz). Am fruchtvollsten war bislang, FDM-Angebote nach Absprache mit Dozent_innen in deren Veranstaltungen zu implementieren und ggf. einen Nachweis darüber auszustellen.

[13] https://www.fdm.uni-hamburg.de/service.html (Zugriff am 18.07.2023)

Data Papers

Eine kritische Bestandsaufnahme

Jansky, Caroline
jansky[at]hab.de
Herzog August Bibliothek Wolfenbüttel, Deutschland

de la Iglesia, Martin
iglesia[at]hab.de
Herzog August Bibliothek Wolfenbüttel, Deutschland

Zusammenfassung. Data Papers schlagen die Brücke zwischen der etablierten Publikationsform des Zeitschriftenartikels und neuartigen Forschungsdatenpublikationen. Das Poster visualisiert die Skizze eines Data Papers für die Digitalen Geisteswissenschaften und den Prozess seiner Erstellung. Für diese konzeptuellen Überlegungen zum Publikationsformat Data Paper werden im Rahmen einer Bestandsaufnahme sowohl geisteswissenschaftliche Data Journals / Data Papers als auch solche Publikationsorgane, die sich nicht ausdrücklich als Data Journals verstehen, hinsichtlich ihres Umgangs mit Forschungsdaten betrachtet. So werden die unterschiedlichen Herangehensweisen an und Verständnisse von Data Papers herausgearbeitet. Dabei gilt es, die Bedarfe der beteiligten Akteur*innen zu ermitteln und gegeneinander abzuwägen.

1 Data Papers

In geisteswissenschaftlichen Zeitschriften hält das in den Naturwissenschaften bereits etablierte Format des Data Papers (auch als „dataset articles“ oder „data articles“[1] bezeichnet) Einzug,[2] und als Data Journals widmen sich neue Publikationsorgane diesem Beitragstypus.[3] Grundsätzlich sollen durch Data Papers die Sichtbarkeit von publizierten Forschungsdaten erhöht, deren Nachnutzung und (community-basierte) Qualitätssicherung erleichtert werden. Data Papers schlagen damit die Brücke zwischen herkömmlichen Zeitschriftenartikeln und den relativ neuen Forschungsdaten-

[1] RfII – Rat für Informationsinfrastrukturen 2019, S. 52.
[2] Vgl. Kindling und Strecker 2022; Avanço 2022.
[3] Z. B. *Research Data Journal for the Humanities and Social Sciences* (*RDJ*, seit 2016), *Journal of Open Humanities Data* (*JOHD*, seit 2015).

publikationen, die hinsichtlich ihrer Standardisierung, Qualitätssicherung und Umsetzung derzeit diskutiert werden.
Bei näherer Betrachtung zeigt sich, dass das Verständnis von Data Papers sowie deren praktische Umsetzung bisweilen sehr unterschiedlich ausfallen: Auch hier hat sich bislang noch kein einheitlicher Standard etabliert. Deshalb lohnen sich konzeptuelle Überlegungen zu Data Papers, auch wenn sich diese als ein „Brückenformat“ auf dem Weg zu standardisierten, qualitätsgesicherten und ausreichend kontextualisierten geisteswissenschaftlichen Forschungsdatenpublikationen in entsprechenden Repositorien herausstellen und somit obsolet werden könnten.

2 Perspektiven

Diese Überlegungen müssen sich an den Bedarfen der beteiligten Akteur*innen orientieren und diese Bedarfe im Sinne einer Best Practice gegeneinander abwägen. Für geisteswissenschaftliche Zeitschriften sind die folgenden Perspektiven zu nennen:

- Autor*innenperspektive: Welche inhaltlichen, qualitativen oder formalen Kriterien müssen die Daten erfüllen, um überhaupt als Grundlage für ein Data Paper infrage zu kommen? Wie wird Autor*innenschaft / Urheber*innenschaft der Daten kenntlich gemacht?[4]

- Redaktionelle Perspektive: Welche Eigenschaften müssen Datenpublikationen aufweisen, damit die Veröffentlichung eines Data Papers einen Erkenntnisgewinn bietet? Worin bestehen die Unterschiede und Gemeinsamkeiten zwischen Data Paper, Data Review und Executable Paper, ist es sinnvoll, diese Formate zu unterscheiden und parallel anzubieten?[5] Können im Qualitätssicherungsverfahren Daten- und Beitragsqualität gleichermaßen beurteilt werden? Wie können veränderliche Datenpublikationen angemessen repräsentiert werden? Ist es sinnvoll, als Zeitschrift Daten selbst zu hosten?

- Rezipient*innenperspektive: Kann ich von der Qualität des Papers auf die Datenqualität schließen? Können und sollten Datenpublikation und Data Paper getrennt voneinander zitiert werden?

[4] Vgl. Kratz und Strasser 2015.
[5] Vgl. Guido und Guerard 2023; Lasser 2020.

• Gutachter*innenperspektive: Was soll (vornehmlich) begutachtet werden: die Qualität der Forschungsdaten oder die Qualität der Darstellung im Data Paper? Welche Begutachtungsverfahren finden Anwendung, wie „offen“[6] sind diese und welche Kriterien werden angelegt?

3 Ziel

Das Poster visualisiert die Skizze eines Data Papers für die Digitalen Geisteswissenschaften und stellt den Prozess seiner Erstellung und die dabei zu treffenden Entscheidungen grafisch dar. Dabei finden Aspekte der konkreten Umsetzung und Integration in redaktionelle Workflows besondere Berücksichtigung.
Die Grundlage hierfür bildet eine Bestandsaufnahme des Umgangs mit Forschungsdaten in internationalen geisteswissenschaftlichen Data Papers und Data Journals, aber auch in Publikationsorganen, die sich nicht ausdrücklich als Data Journals verstehen bzw. deren Publikationsformate nicht explizit als Data Papers benannt sind. Diese werden auf Grundlage der oben genannten Bedarfe systematisch analysiert, die Ergebnisse dieser Analyse fließen im Sinne einer Synthese zur Best Practice in die Skizze eines „idealen“ Data Papers für die Digitalen Geisteswissenschaften ein.

Bibliografie

Avanço, Karla. „Data Papers… and FAIR“. *The Road to FAIR*, 17. Juni 2022. https://roadtofair.hypotheses.org/364.

Fadeeva, Yuliya. „Qualitative Sprünge in der Qualitätssicherung? Potenziale digitaler Open-Peer-Review-Formate“. V2. In *Fabrikation von Erkenntnis – Experimente in den Digital Humanities*, herausgegeben von Manuel Burghardt, Lisa Dieckmann, Timo Steyer, Peer Trilcke, Niels Walkowski, Joëlle Weis, Ulrike Wuttke. (= Zeitschrift für digitale Geisteswissenschaften / Sonderbände, 5). Wolfenbüttel: 2023 [2021]. https://doi.org/10.17175/sb005_002_v2.

Guido, Daniele und Elisabeth Guerard. „Executable Papers in the Humanities, or How Did We Land to the Journal of Digital History“. Vortrag am 4. Februar 2023 im Rahmen von *FOSDEM 2023*, Brüssel. Video, 14:09.

[6] Vgl. Fadeeva 2021.

https://fosdem.org/2023/schedule/event/openresearch_executable_papers/.

Journal of Open Humanities Data. (2015–). https://openhumanitiesdata.metajnl.com.

Kindling, Maxi und Dorothea Strecker. *List of Data Journals*. V1. Zenodo. https://doi.org/10.5281/zenodo.7082126.

Kratz, John Ernest und Carly Strasser. „Researcher Perspectives on Publication and Peer Review of Data." *PLoS ONE* 10, Nr. 2 (13. Februar 2015): e0117619. https://doi.org/10.1371/journal.pone.0117619.

Lasser, Jana. „What Is an Executable Paper?" *Sozialwissenschaftliche Methodenberatung*, 18. Juni 2020. https://sozmethode.hypotheses.org/1045.

Research Data Journal for the Humanities and Social Sciences. (2016–). https://brill.com/view/journals/rdj/rdj-overview.xml.

RfII – Rat für Informationsinfrastrukturen. *Herausforderung Datenqualität – Empfehlungen zur Zu-kunftsfähigkeit von Forschung im digitalen Wandel.* 2. Auflage. Göttingen: 2019. https://rfii.de/?p=4043

van Edig, Xenia, Sarah Dellmann, Anna Renziehausen und Frauke Ziedorn. „Data Papers – eine Ode an die Daten". *TIB Blog*, 15. Februar 2022. https://blogs.tib.eu/wp/tib/2022/02/15/data-papers-eine-ode-an-die-daten/.

Walters, William H. „Data Journals: Incentivizing Data Access and Documentation within the Scholarly Communication System". *Insights* 33, Nr. 1 (10. Juni 2020). https://doi.org/10.1629/uksg.510.

Aufbau einer Messaging-Pipeline am ZKM zur Harmonisierung der Datenlandschaft und Umsetzung der FAIR Prinzipien

Kohlbecker, Andreas
kohlbecker[at]zkm.de
ZKM | Zentrum für Kunst und Medien Karlsruhe, Deutschland
ORCID-iD: 0000-0003-1046-8750

Zusammenfassung. Am Beispiel des ZKM | Zentrum für Kunst und Medien Karlsruhe wird gezeigt, wie historisch gewachsene Datenlandschaften in Kulturinstitutionen durch den Einsatz von Messaging-Pipelines harmonisiert werden können. In vielen Kulturinstitutionen besteht historisch bedingt eine heterogene Daten-Infrastruktur. Die Komplexität des Netzwerks aus Systemen und Datenflüssen kann die Umsetzung der FAIR Prinzipien erschweren. Am ZKM | Zentrum für Kunst und Medien Karlsruhe wurde durch die Implementierung einer „Messaging-Pipeline" die Grundlage für eine moderne Daten-Infrastruktur geschaffen, die bestehende Abhängigkeiten entkoppelt und Raum für flexible Lösungen schafft. Komponenten und Prozesse, die zur Realisierung der FAIR Prinzipien erforderlich sind, können dadurch effizienter implementiert werden.

Am Beispiel des ZKM | Zentrum für Kunst und Medien Karlsruhe, wird gezeigt, wie historisch gewachsene Datenlandschaften in Kulturinstitutionen durch den Einsatz von Messaging-Pipelines harmonisiert und flexibilisiert werden können. Diese Strategie kann zu einer Daten-Infrastruktur führen, welche die Erfüllung der FAIR[1] Data Prinzipien wesentlich erleichtert.

1.1 Ausgangslage in vielen Kulturinstitutionen

Für Kulturinstitutionen kann es eine große Herausforderung sein, ihre internen Datenbestände für die Forschung entsprechend der FAIR Prinzipien nutzbar zu machen. Über lange Zeiträume gewachsenen Daten-Infrastrukturen bilden oft ein Konglomerat aus heterogener Systemen und Datenmodellen. Auch wenn Datenschemata sich schon teilweise an Standards wie Spectrum[2], LIDO[3] oder CIDOC CRM[4] anlehnen, wer-

[1] Wilkinson, et al. 2016
[2] https://collectionstrust.org.uk/spectrum/
[3] https://lido-schema.org
[4] https://cidoc-crm.org/

den oft nur in Einzelfällen allgemein anerkannte Terminologien verwendet. Ebenso kann es am Einsatz von Persistent Identifier (PIDs) fehlen. Fehleingaben und missbräuchliche Verwendung von Datenfeldern kommen erschwerend hinzu.

Gerade für größere Institutionen kann diese Ausgangslage eine große Hürde bei der Umsetzung der FAIR Prinzipien bedeuten. Eine Verbesserung der Situation innerhalb der bestehenden Struktur ist aufgrund vieler gegenseitiger Abhängigkeiten oft nur langsam oder nur mit großem personellen und zeitlichem Aufwand möglich, der sich im Tagesgeschäft nur schlecht unterbringen lässt.

Durch den Aufbau einer Messaging-Pipeline kann eine parallele Struktur zu etablieren werden, die sukzessive zu einer Entkoppelung und Entflechtung bestehender Systeme führt und eine iterative Verbesserung der inhaltlichen und strukturellen Datenqualität ermöglicht.

Für die Umsetzung der FAIR Prinzipien muss die bestehende Infrastruktur oft durch weitere Komponenten ergänzt werden: Bereitstellung von Metadaten für Maschinen und Menschen gleichermaßen, effiziente Suchmöglichkeiten, Zuweisung von Persistent Identifier (PIDs) zu Daten-Objekten und die Etablierung von Prozessen, um registrierte PIDs aktuell zu halten. Daten müssen in Austauschformate transformiert und per Webservice bereitgestellt werden. Der Zugriff auf Daten und Objekte muss sicher geregelt sein.

1.2 Implementierung einer „Messaging Pipeline“ am ZKM

Im April 2022 wurde im ZKM mit dem Aufbau einer sog. „Messaging-Pipeline“ auf Basis von Apache Kafka[5] begonnen, die alle Datenquellen erfassen und in parallelen Datenströmen potenziellen Konsumenten zur Verfügung stellen kann. Diese bildet das Rückgrat, an das bestehende und neue Systeme sukzessive angebunden werden.

So werden zum Beispiel alle Änderungen in Archiv- und Sammlungs-Datenbanken tagesaktuell in den Datenstrom überführt und auf ein stringenteres Datenmodell gemappt, das intensiven Gebrauch von allg. akzeptierten Terminologien und Normen, wie VIAF[6], GND[7], Getty AAT[8]

---

5 Apache Kafka, 2023
6 https://viaf.org/
7 https://gnd.network
8 http://vocab.getty.edu/aat/

macht. Hierbei erkannte Mappingprobleme führen zu Fehlermeldungen, die als Datenstrom in die Pipeline einfließen, analysiert und in Form von Berichten verantwortlichen Personen übermittelt werden. So werden die Daten in einem iterativen Prozess bereinigt. Gleichzeitig stärkt dieser Feedback-Mechanismus das Verantwortungsbewusstsein für Datenqualität, denn veröffentlicht wird nur, was sauber ist.

Derzeit entsteht zudem ein umfassendes Daten-Portal mit Suchfunktionalität. Ein Dienst zum Management und Auflösen von PIDs auf Basis des ARK[9] Systems für potenziell alle Objekte des ZKM ist in der Konzeption. Dieser Komponente muss Daten aus diversen Quellen in der Institution integrieren, um z.B. den Zugriff von Nutzern nach Berechtigungsstufen zu regeln und Anfragen auf die entsprechend Metadaten- oder Objekt- URLs auflösen zu können.

Der Implementierung von Kafka als Basisinfrastruktur in Kulturinstitutionen ist ein nicht zu unterschätzender Aufwand, da es sich um eine Basistechnologie handelt, die entsprechend der individuellen Anforderungen ausgestaltet werden muss. Auch wenn es andere Workflow und Pipeline Konzepte gibt, die initial ev. leichter zu implementieren sind, bietet die Flexibilität und Entkopplung durch Kafka mehr Möglichkeiten der Anpassung und somit auch der Übertragung dieses Konzeptes auf vergleichbare Institutionen.

Bibliografie

„Apache Kafka," Apache Software Foundation, accessed April 27, 2023, https://kafka.apache.org/

Kunze, John A., 2003, „Towards Electronic Persistence Using ARK Identifiers"

CIDOC Conceptual Reference Model (CRM), accessed July 20, 2023, https://collectionstrust.org.uk/spectrum/

Getty Vocabularies, accessed July 20, 2023, http://vocab.getty.edu/aat/

VIAF, the Virtual International Authority File, accessed July 20, 2023, https://viaf.org/

[9] Kunze, 2003

Spectrum, the UK collection management standard, accessed July 20, 2023, https://collectionstrust.org.uk/spectrum/

LIDO – Lightweight Information Describing Objects, accessed July 20, 2023, https://lido-schema.org

Wilkinson, Mark D., Michel Dumontier, IJsbrand Jan Aalbersberg, Gabrielle Appleton, Myles Axton, Arie Baak, Niklas Blomberg, et al., "The FAIR Guiding Principles for Scientific Data Management and Stewardship," *Scientific Data* 3 (2016): 160018. https://doi.org/10.1038/sdata.2016.18.

Oral-History.Digital

Eine Erschließungs- und Rechercheumgebung für audiovisuelle, narrative Forschungsdaten

Kompiel, Peter
p.kompiel[at]fu-berlin.de
Universitätsbibliothek der Freien Universität Berlin, Deutschland

Zusammenfassung: Das Projekt Oral-History.Digital hat eine digitale Recherche- und Erschließungsplattform für wissenschaftliche Sammlungen von audiovisuell aufgezeichneten narrativen Zeitzeugen-Interviews entwickelt. Sammlungsinhaber*innen können Audio- und Video-Interviews mit Metadaten und dazugehörigen Transkripten, Biografien und Bildern einstellen, softwareunterstützt nach etablierten Standards bearbeiten, nachhaltig bereitstellen und sicher archivieren. Forscher*innen können verschiedene Interview-Archive sammlungsübergreifend durchsuchen und ihre Inhalte analysieren. Die Plattform bietet ferner Werkzeuge und Empfehlungen für die Transkription, automatische Spracherkennung und Verschlagwortung an. Das Projekt wird mit DFG-Förderung seit 2020 umgesetzt und steht ab September 2023 zur Verfügung.

Abstract: Audiovisuell aufgezeichnete narrative Interviews sind wichtige Quellen in der Geschichtswissenschaft und weiteren Sprach-, Sozial- und Geisteswissenschaften. Dank der Digitalisierung können Oral History-Interviews inzwischen auf breiterer Basis analysiert werden. Digital gespeicherte Interviews können dabei als Musterbeispiel geschichtswissenschaftlicher Forschungsdaten gelten, sind sie doch im Forschungsprozess gemeinsam mit den Interviewten erzeugte Quellen, die für spätere Überprüfungen und Sekundäranalysen zur Verfügung stehen oder stehen sollten. Im digitalen Wandel wird auch ihre FAIRe Archivierung und – auch interdisziplinäre – Sekundärnutzung zunehmend nachgefragt (Apel, Leh, Pagenstecher, *Oral History im Digitalen Wandel*). Es mangelt jedoch an einheitlichen Erschließungsstandards, einem übergreifenden Nachweissystem und einer Forschungsumgebung, die digitale Recherchen und Analysen quellenspezifisch unterstützt. Daher wurde mit DFG-Förderung seit 2020 die Informationsinfrastruktur Oral-History.Digital (oh.d) aufgebaut.
Oral-History.Digital ist eine digitale Recherche- und Erschließungsplattform für wissenschaftliche Sammlungen von audiovisuell aufgezeichneten narrativen Zeitzeugen-Interviews. Im

Sinne der FAIR-Prinzipien macht das Portal die Interviews als audiovisuelle Forschungsdaten auffindbar, zugänglich, verknüpfbar und nachnutzbar. Interviewprojekte und Sammlungsinhaber*innen können Audio- und Video-Interviews mit Metadaten und dazugehörigen Transkripten, Biografien und Bildern einstellen, softwareunterstützt nach etablierten Standards bearbeiten, nachhaltig bereitstellen und sicher archivieren. Dafür stehen Werkzeuge und Empfehlungen für Transkription, automatische Spracherkennung und Verschlagwortung zur Verfügung. Je nach Erschließungszustand und Rechtesituation können die Interviews mittels einer differenzierten Nutzerverwaltung zugänglich gemacht werden. Forschende können verschiedene Interview-Archive sammlungsübergreifend durchsuchen, analysieren und in ihrer Arbeitsmappe annotieren. In der Rechercheumgebung können sie per Volltextsuche über zeitkodierte Transkripte direkt an ausgewählte Interviewstellen springen. Die untertitelte Videoansicht macht Sprechweise, Mimik und Gestik der Auswertung zugänglich. Filterfacetten, Karte und Register erlauben spezifische Recherchen.
Die Universitätsbibliothek der Freien Universität Berlin entwickelt Oral-History.Digital gemeinsam mit erfahrenen Partner*innen aus der FernUni Hagen, der LMU München, der Uni Bamberg, der FAU Erlangen sowie dem FZH Hamburg sowie mit Pilotnutzer*innen aus Forschungsprojekten, Museen, Archiven und Stiftungen. Das archivübergreifende Interviewportal steht seit September 2023 unter https://www.oral-history.digital/ zur Verfügung.
Auf der FORGE können Konferenzteilnehmer*innen während der Postersession das oh.d-Portal und seine Funktionalitäten kennenlernen. Interessierte Inhaber*innen von Interview-Sammlungen können sich über die Möglichkeiten der Sicherung und Erschließung ihrer eigenen Interviews informieren.

Bibliografie

Linde Apel, Almut Leh, Cord Pagenstecher, Oral History im digitalen Wandel. Interviews als Forschungsdaten, in: Erinnern, erzählen, Geschichte schreiben. Oral History im 21. Jahrhundert, hrsg. v. Linde Apel (Berlin: Metropol 2022), 193-222, URL: https://zeitgeschichte-hamburg.de/files/public/FZH/PDF/apel_erinnern_ebook_offen.pdf

Annette Gerstenberg, Cord Pagenstecher, ‚Mi ricordo', ‚je me souviens': ich erinnere mich. Sammlungsübergreifende Interviewanalysen in Oral History und Korpuslinguistik, in: Apropos.

Perspektiven auf die Romania 9 (2022), 213–239, DOI: https://doi.org/10.15460/apropos.9.1902.

Cord Pagenstecher, Interview-Archive zum Nationalsozialismus. Die digitale Erschließung und Analyse von Oral History-Sammlungen am Beispiel des Online-Archivs Zwangsarbeit 1939-1945, in: Nationalsozialismus digital. Die Verantwortung von Bibliotheken, Archiven und Museen sowie Forschungseinrichtungen und Medien im Umgang mit der NS-Zeit im Netz, hrsg. v. Markus Stumpf, Hans Petschar, Oliver Rathkolb (Wien: V & R unipress 2021), 101-118, DOI: https://doi.org/10.14220/9783737012768.101.

Cord Pagenstecher, Oral History und Digital Humanities, in: BIOS – Zeitschrift für Biographieforschung, Oral History und Lebensverlaufsanalysen 30, Nr. 1-2 (2017), 76-91, https://doi.org/10.3224/bios.v30i1-2.07.

Daten sind Daten sind Daten sind Daten

Zu den Auswirkungen datengestützter Analysen auf Forschungsinfrastrukturen und Datenverständnis in der Medienwissenschaft

Matuszkiewicz, Kai
kai.matuszkiewicz[at]uni-marburg.de
Philipps-Universität Marburg, Deutschland

Zusammenfassung: Das Poster möchte illustrieren, welche Rolle Forschungsinfrastrukturen wie Fachrepositorien in der digitalen Transformation der Medienwissenschaft spielen. Hierbei ist es nicht nur essentiell, dass Forschungsinfrastrukturen wissenschaftsgetrieben entwickelt werden, um auf deren Bedarfe mit neuen Dienstleistungen reagieren zu können, darüber hinaus ist es notwendig, dass sich diese Forschungsinfrastrukturen aktiv an diesem fachlichen Aushandlungsprozess beteiligen und ihn mitgestalten. Dies betrifft insbesondere Daten und wie die Arbeit mit diesen die Medienwissenschaft methodisch und praxeologisch verändert. Mit dem Poster soll aufgezeigt werden, was dies konkret für ein Fachrepositorium in der Medienwissenschaft sowie die medienwissenschaftliche Auffassung von Daten bedeutet.

Durch die digitale Transformation werden Repositorien zu zentralen Akteur:innen der Wandlungsprozesse durch Open Science, auf die sie nicht nur reagieren sollten, sondern die sie proaktiv mitgestalten müssen, wollen sie zukünftig nicht nur passiv als Container von Daten fortbestehen.[1] Dies gilt in besonderem Maße für Fachrepositorien, die einen starken Bezug zu ihrer Community unterhalten, indem sie sich am Community-Building des Faches hinsichtlich der Gestaltung und Bereitstellung von Forschungsinfrastrukturen beteiligen, umso wissenschaftsgetriebenen Ansätzen der Infrastrukturentwicklung zu entsprechen. Dementsprechend sind Fachrepositorien auch in die Diskurse ihrer Disziplinen eingebunden, die den Wandel von wissenschaftlichen Arbeitspraktiken im Zuge der digitalen Transformation und einer damit zusammenhängenden zunehmenden Verbreitung von Open-Science-Praktiken betreffen.
Einer dieser Diskurse dreht sich in der Medienkulturwissenschaft, die zu großen Teilen immer noch von einem hermeneutischen Selbstverständnis geprägt ist, um die Frage der Forschungsdaten. Forschungsdaten sind hierbei zumeist ein umstrittener Gegenstand.

[1] Vgl. dazu Matuszkiewicz 2022.

Dies liegt daran, dass sie sich terminologisch (nur) schwer fassen lassen, vermeintlich nicht den fachlichen und gegenständlichen Besonderheiten der Medienwissenschaft als Geisteswissenschaft entsprechen oder ihre Relevanz bzw. Existenz gar bestritten wird.[2] Den Betreiber:innen von Forschungsinfrastrukturen wie Fachrepositorien stellt sich dies aber nicht nur als theoretische, fachlich spannende Auseinandersetzung aus verschiedenen Perspektiven dar, sondern provoziert durch die Notwendigkeit praktischer Umsetzungen Reaktionen hierauf, die oftmals zeitnah und pragmatisch erfolgen müssen. Ein Fachrepositorium, das die Neuausrichtung der Arbeitspraktiken seiner Disziplin mitgestalten möchte, muss dies aber nicht nur als Herausforderung sehen, sondern kann es als Chance der fachlichen Weiterentwicklung begreifen.

Dies bedeutet, Forschungsdaten in das Fachrepositorium aufzunehmen, es eröffnet aber noch weitere Möglichkeiten, indem man den Datencharakter von Entitäten in den Mittelpunkt rückt, die per se keine Forschungsdaten sind. Publikationen und Metadaten lassen sich z. B. selbst zu Forschungsdaten im Zuge fachhistorischer datengestützter Analysen machen. Das medienwissenschaftliche Fachrepositorium media/rep/[3] führt deshalb in Kooperation mit dem Projekt Digital Cinema-Hub[4] eine fachgeschichtliche Fallstudie durch, in dem die jüngere medienwissenschaftliche Fachhistorie im deutschsprachigen Raum datengestützt untersucht wird. Das Poster möchte konkrete Einblicke in die fachgeschichtliche Fallstudie und die wissenschaftsgetriebene Infrastrukturentwicklung sowie deren wechselseitige Bezüge geben. Darüber hinaus soll aber insbesondere reflektiert werden, welche Anforderungen sich hieraus an ein Repositorium ergeben und somit aufgezeigt werden, wie derartige Anwendungsfälle zur wissenschaftlichen Infrastrukturentwicklung beitragen können. Hierdurch sollen auch Hinweise zur Gestaltung wissenschaftlicher Forschungsdateninfrastrukturen gegeben werden. Vielversprechend sind solche Unterfangen, werden Forschungsinfrastrukturen letztlich hierdurch zu Untersuchungsgegenständen[5] und erlauben zugleich eine Fokussierung auf Daten als Forschungsdaten, deren Status nicht aus sich selbst heraus bestimmt wird, sondern sich durch situative Funktionalisierungen ergibt.

---

[2] Vgl. zu Daten in den Geisteswissenschaften Drucker 2011 sowie Gitelman 2013.

[3] https://mediarep.org/ (21.04.2023).

[4] https://www.uni-marburg.de/de/fb09/medienwissenschaft/forschung/forschungsprojekte/dici-hub (21.04.2023).

[5] Vgl. Kammerer 2020.

Bibliografie

Drucker, Johanna, “Humanities Approaches to Graphical Display,” *Digital Humanities Quarterly* 5 (2011). http://www.digitalhumanities.org/dhq/vol/5/1/000091/000091.html.

Gitelman, Lisa, *Raw Data Is an Oxymoron* (Boston: The MIT Press, 2013), https://doi.org/10.7551/mitpress/9302.001.0001.

Kammerer, Dietmar, “Nicht nur suchen und finden, sondern entdecken und erforschen. Über das Open-Access-Repositorium media/rep/ als Werkzeug medienwissenschaftlicher Forschung,” *Open-Media-Studies-Blog. Zeitschrift für Medienwissenschaft,* December 14, 2020. https://zfmedienwissenschaft.de/online/open-media-studies-blog/nicht-nur-suchen-und-finden-sondern-entdecken-und-erforschen

Matuszkiewicz, Kai, “Eine Frage der Einstellung. Von Repositorien, Fächern und Menschen im Zuge der digitalen Transformation,” *o-bib. Das offene Bibliotheksjournal* 9 (2022). https://doi.org/10.5282/o-bib/5846

A Data Pipeline for Digital Humanities

Development of a Solution for Humanities Data Digitization

Mollenhauer, Sabrina
sabrina.mollenhauer[at]uni-vechta.de
Universität Vechta, Deutschland

Summary. The dissertation project focuses on humanities data and humanities researchers that have not been involved in the Digital Humanities, and the proposal of a solution to enable them to prepare their data for digital sharing. In doing so, the poster presents a process beginning with an exploration of the traditional data space, continues with the determination of a target group, which will lead to the formulation of requirements in accordance with the software engineering process. Furthermore, the exploration of existing digital open data tools and platforms will serve to determine technical requirements of a data pipeline solution designed for the target audience and their data.

1 Introduction

As a result of ethically relevant considerations and long-standing legal restrictions, the digital collection and preservation of research data in the humanities is far behind that of digital data and computational methodologies in the natural sciences[1,2]. At the same time, as early as 2012 Fiormonte warned against the domination of a small group of privately funded researchers dominating the field and proposed an inclusive, diverse model of DH[3], an idea that has also been supported by others[4,5]. This creates a pressure to make data available within public infrastructure. Since 2012, how far has the provision of humanities data progressed? Is it possible to determine common issues in different marginalized humanities research that hinder digitization of data? And if so, what would a solution look like to facilitate the digitization of humanities data that has not been made available thus far?

---

1 cp. Collins et al, "*Going Digital*," 10.
2 cp. Rehbein, *On Ethical Issues*, 631-654.
3 Fiormonte, *Towards a Cultural Critique*, 59.
4 cp. Collins et al, "*Going Digital*," 10.
5 Wuttke, *Wege bereiten*, 1-16.

2 Developing a Digital Humanities Pipeline

The presented project proposal follows several steps with the final goal of creating a pipeline for humanities researchers to contribute to existing and future public federated data publication services.

2.1 Exploration of the Humanities Data Space

In a first step, an exploration of the humanities data space shall determine the low hanging fruit of data that has not been made available as research data according to its technical, legal, and ethical characteristics. It is likely, but not necessary that data belonging to this group is used by a group of researchers representing a heterogeneous cross-section of humanities research fields[6].

In a second step, specific non-functional qualitative requirements are determined based on the needs assessments carried out thus far in the Humanities.[7,8,9]

2.2 Exploration of the Public Federated Data Platforms and Tools

In a third step, the technical functional requirements[10] of existing public federated data services and tools will be determined. In this analysis, the focus will be on those tools that follow technical paradigms and are supported by existing national (NFDI) and European (EOSC) initiatives, such as the DARIAH-DE collection registry in order to warrant future application.[11,12,13,14,15]

2.3 Development of a Data Pipeline for Digital Humanities

The initial non-functional requirements are specified further in an iterative user-centric software design and developent process16,17 incorporating functional requirements. Finally, the aim is to develop a

6 cp. Collins et al, "*Going Digital*," 10.
7 Imeri and Danciu 2017, p. 10.
8 Schopfel and Prost 2016, p. 100.
9 ISO/IEC/IEEE 29148, *Requirements Engineering*, 43.
10 ibid., 13.
11 RfII, *Föderierte Dateninfrastrukturen,* 42.
12 cp. Gradl et al, *Heterogene Daten*, 1-10.
13 cp. Patrick 2023.
14 cp. Herklotz et al. 2021.
15 cp. Strathern et al. 2020.
16 Bednar and Welch 2008, 225-236.
17 Raven and Flanders 1996, 1-13.

platform for researchers to participate in open science with shared open data. All the steps that will serve in the development of the pipeline shall be presented through the poster, as a basis for discussion.

Bibliography

Bednar, Peter M., and Christine Welch. "Contextual inquiry and requirements shaping." In Information Systems Development: Challenges in *Practice, Theory, and Education Volume 1*, pp. 225-236. Boston, MA: Springer US, 2008.

Collins, Sandra, Natalie Harrower, Dag Trygve Truslew Haug, Beat Immenhauser, Gerhard Lauer, Tito Orlandi, Laurent Romary, and Eveline Wandl-Vogt. "Going digital: Creating change in the Humanities: ALLEA E-HUMANITIES WORKING GROUP REPORT." [Research Report] ALLEA. 2015._xffff_

Fiormonte, Domenico. "Towards a cultural critique of the digital humanities." *Historical Social Research/Historische Sozialforschung* (2012): 59-76.

Gradl, Tobias, Andreas Henrich, and Christoph Plutte. "Heterogene Daten in den Digital Humanities: Eine Architektur zur forschungsorientierten Föderation von Kollektionen." *Zeitschrift für digitale Geisteswissenschaften* Sonderband 1, Grenzen und Möglichkeiten der Digital Humanities (2015). 10.17175/sb001_020

Herklotz, Markus, Lars Oberländer, Irene Schumm, and Renat Shigapov. "iVA: Ein interaktiver Virtueller Assistent von BERD@ BW zur Aufbereitung von Rechtsfragen im Bereich Open Science." *Tage 2021* (2021): 306.

ISO/IEC/IEEE 29148. "Systems and Software Engineering—Life Cycle Processes—Requirements Engineering." (2011).

Patrick, M. “DMPonline: Introduction and Basics.” In . Bodleian Libraries, University of Oxford. (2023).

Rehbein, Malte. "It’s our department: On ethical issues of digital humanities." In *K. Richts & P. Stadler (eds), "Ei, dem alten Herrn zoll' ich Achtung gern'": Festschrift für Joachim Veit zum 60. Geburtstag. Allitera, München*, 631-654, 2016.

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 – Rat für Informationsinfrastrukturen: *Föderierte Dateninfrastrukturen für die wissenschaftliche Nutzung. NFDI, EOSC und Gaia-X: Vergleich und Anregungen für eine engagierte Mitgestaltung des Ausbaus und der Weiterentwicklung, RfII Berichte No. 4 (*Göttingen 2023)

Strathern, Wienke, Moritz Issig, Kati Mozygemba, and Jürgen Pfeffer. "QualiAnon-The Qualiservice tool for anonymizing text data." en. Tech. rep. TUM-I2087. (2020).

Wuttke, Ulrike. "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 (2022).

PhiWiki - ein semantisches Wiki für die Philosophie

Podschwadek, Frodo
frodo.podschwadek[at]adwmainz.de
Akademie der Wissenschaften und der Literatur Mainz, Deutschland
ORCID-iD: 0000-0003-1248-4228

Vater, Christian
christian.vater[at]adwmainz.de
Akademie der Wissenschaften und der Literatur Mainz, Deutschland
ORCID-iD: 0000-0003-1367-8489

Geiger, Jonathan D.
jonathan.geiger[at]adwmainz.de
Akademie der Wissenschaften und der Literatur Mainz, Deutschland
ORCID-iD: 0000-0002-0452-7075

Zusammenfassung. Das PhiWiki ist eine sich in der Entwicklung befindliche Software-Anwendung, die es zunächst Philosoph:innen, prinzipiell aber auch anderen Geisteswissenschaftler:innen, ermöglichen soll, Daten zu den Ideen und Begriffen ihrer Disziplin semantisch zu erfassen und neue Verbindungen innerhalb dieser Daten zu entdecken. Hierbei werden Technologien wie die Mediawiki-Oberfläche, semantische Datenspeicher und unterschiedliche Normdaten-Formate verwendet, um die Anwendung performant zu halten und anschlussfähig an ein föderiertes Internet zu machen.

Digitale Werkzeuge haben das Publikationswesen radikal verändert und zu einem beträchtlichen Anstieg an veröffentlichtem Material geführt.[1] Die Auswirkungen dieses technologischen Wandels sind in allen akademischen Disziplinen spürbar, sie werden jedoch besonders deutlich dort wahrgenommen, wo Forschungsmaterial und Forschungsprodukte vorwiegend Texte sind.

Eine dieser Disziplinen ist die Philosophie, für die heute eine Vielzahl von internationalen Fachzeitschriften mit zunehmender Tendenz zur Spezialisierung existiert. Es wird daher zunehmend schwieriger, einen Überblick über bestimmte Themenbereiche zu behalten oder sich noch wenig bekannte Themenfelder zu erschließen. Gleichzeitig steigt der

[1] Siehe z.B. Jason Priem, Heather A. Piwowar, and Bradley M. Hemminger. “Altmetrics in the Wild: Using Social Media to Explore Scholarly Impact.” 2012; Bo-Christer Björk and David Solomon. “Developing an Effective Market for Open Access Article Processing Charges.” 2014.

Bedarf nach kollaborativer Textproduktion und der Verwendung von Metadaten und Ontologien.

Hier setzt das PhiWiki-Projekt an, das derzeit von einer Gruppe von Mitarbeiter:innen der Digitalen Akademie (DA) an der Akademie der Wissenschaften und der Literatur Mainz sowie des Open Science Lab der Technischen Informationsbibliothek Hannover (TIB) entwickelt wird.[2] Grundlage der Entwicklung ist hierbei die praktische Erfahrung mit dem Einsatz von Wikis in der Wissenschaft vor dem Hintergrund von Ansätzen zu einer Theorie der digitalisierten Enzyklopädie.[3]

Das PhiWiki ist eine Software-Anwendung, die neue Möglichkeit bietet, philosophische Konzepte effizient und sinnvoll zu katalogisieren und miteinander in Beziehung zu setzen. Sie verwendet hierfür unter anderem:

- Semantic Reasoning zum Aufzeigen von Verbindungen und Mustern in Datenbeständen,
- einen SPARQL-Endpoint zur Einbindung in eine föderierte Dateninfrastruktur,
- eine Mediawiki-basierte Benutzeroberfläche für kollaboratives Wissensmanagement,
- eine Triple-Store-Datenbank für flexible Modellierung komplexer Daten und
- NLP-Methoden wie Entity Linking zu Normdaten, z.B. aus dem Normdatenkatalog (GND) der Deutschen Nationalbibliothek sowie aus der Internet Philosophy Ontology (InPhO).

Das Projekt bietet einen anspruchsvollen Ansatz zur Katalogisierung von Informationen über Entitäten aus der akademischen Philosophie und bietet leistungsstarke Werkzeuge zur Erkundung und Entdeckung neuer Verbindung innerhalb vorliegender Daten. Es stellt zudem eine Arbeitsumgebung bereit, die den bewährten Umfang der editorischen Funktionen des Mediawikis nutzt.

---

[2] Beteiligte Personen zum Zeitpunkt der Einreichung sind: Christian Vater (DA), Kolja Bailly (TIB), Jonathan D. Geiger (DA), Frodo Podschwadek (DA).

[3] Z.B. in der Fokusgruppe “Begriffs- und Wissensgeschichte” der AG philosophische Digitalitätsforschung / Philosophie der Digitalität der Deutschen Gesellschaft für Philosophie. Hinzu kommen verschiedene Experimentalentwicklungen und *after work pet projects*.

Der Educational Resource Finder der Cultural Research Data Academy

Aus- und Weiterbildungsangebote zu FDM sowie Code und Data Literacy

Polywka, Andrea
polywka[at]staff.uni-marburg.de
NFDI4Culture / Philipps-Universität Marburg, Deutschland
ORCID-iD: 0000-0003-0003-6719

Zusammenfassung. Die Cultural Research Data Academy widmet sich als interdisziplinäre und dezentrale Institution des Konsortiums für Forschungsdaten materieller und immaterieller Kulturgüter (NFDI4Culture) der Bündelung bestehender fachspezifische und bedarfsorientierter Aus- und Weiterbildungsmöglichkeiten im Bereich Data und Code Literacy. Demnächst veröffentlicht das Team der CRDA ein kuratiertes Portfolio, welches Informationen zu unterschiedlichen Kurs- und Weiterbildungsangeboten sammelt, die sich thematisch an die Fachcommunities der Kunstgeschichte, Musikwissenschaft, Film- und Medienwissenschaft, Theater- und Tanzwissenschaft, sowie Architektur richten und inhaltlich an der TaDiRAH-Taxonomie orientiert sind.

Die Cultural Research Data Academy und die Konzipierung des Portfolios

Die Cultural Research Data Academy (**CRDA**) ist eines von sieben Aufgabenbereichen des Konsortiums für Forschungsdaten materieller und immaterieller Kulturgüter (NFDI4Culture) und widmet sich als interdisziplinäre und dezentrale Institution den Aus- und Weiterbildungsmöglichkeiten für die Fachcommunities der Kunstgeschichte, Musikwissenschaft, Film- und Medienwissenschaft, Theater- und Tanzwissenschaft, Architektur sowie für die Kulturerbe-Institutionen. Übergeordnetes Ziel der CRDA ist die Entwicklung und Verbesserung "kultureller" Daten- und Code-kompetenz sowie von computergestütztem Denken und Datenmanagementfähigkeiten für die Geistes- und Kulturwissenschaften. In den kommenden Jahren soll die CRDA bestehende fachspezifische und bedarfsorientierte Aus- und Weiterbildungsmöglichkeiten im Bereich Data und Code Literacy bündeln sowie eigene Angebote, Toolkits und Kompetenzrahmen erarbeiten und anbieten.
Eines der Ziele der CRDA besteht darin die Bedarfe der Fachcommunities nach generischen sowie spezifischen Weiterbildungsangeboten abzufragen (im Rahmen der jährlichen Forumsveranstaltungen, eigener Schulungen, sowie Austauschtreffen mit Vertreter:innen der Partnerin-

stitutionen, sowie FDM-Landesinitiativen) und auf dieser Grundlage eigens entwickelte Formate anzubieten und ebenso auf Ausbildungsangebote von externen Institutionen hinzuweisen. Beispielsweise wird ein eigener Basiskurs zum Thema "Forschungsdatenmanagement" angeboten, der auf Anfrage gebucht werden kann und fortlaufend durch fachspezifische Module und nach Zielgruppen durch das CRDA-Team ergänzt wird (derzeit wurden eigene Kursmodule zu "FDM-Basics" für GLAM-Mitarbeitende, "FDM (nicht nur) für Musikbibliothekare" und "Audiovisuelle (Forschungs-)Daten – Anwendungsbereiche und Analysetools" angeboten).
Das Portfolio stellt eine kuratierte Sammlung von Aus- und Weiterbildungsangeboten zur Verfügung. In engem Austausch mit dem Team des Projektes DALIA (Data Literacy Alliance), welches unter anderem Angebote der verschiedenen Konsortien bündeln soll, werden Daten zu den aufgeführten Angeboten systematisch und nach bestimmten Standards erhoben. Diese Erfassung verbessert nicht nur die Durchsuchbarkeit des Portfolios an sich, sondern bietet auch die Möglichkeit die Daten über Schnittstellen zugänglich zu machen, etwa für das DALIA-Projekt. Die Nachnutzung der Daten wird dadurch erhöht und die projektübergreifende Zusammenarbeit zu den Themen Schulungen, Weiterbildungen und Lehr- und Lernmaterialien im Bereich von Datenkompetenzen sowie Forschungsdatenmanagement ermöglicht und vertieft.

Stepping up data literacy and research impact in the Humanities through data publishing

Schmidt, Birgit
bschmidt[at]sub.uni-goettingen.de
Georg-August-Universität Göttingen, Deutschland
ORCID-iD: 0000-0001-8036-5859

McGillivray, Barbara
barbara.mcgillivray[at]kcl.ac.uk
King's College London, Großbritannien
ORCID-iD: 0000-0003-3426-8200

Summary. This poster presents work related to promoting peer-reviewed data papers across the humanities, focusing on experiences of the Journal of Open Humanities Data (JOHD). It provides information on the publication and review process, and points out opportunities to engage with the ongoing reform of research assessment and teaching and training efforts.

The open research movement and initiatives like the FAIR principles have been critical in establishing the importance of data in research (Wilkinson et al., 2016). Attention to openly available data in Humanities and Social Sciences (HSS) research has gradually grown which can be attributed to the increased availability of digital collections, the development of new data-intensive methods, increasingly solid infrastructures, funder requirements and the involvement of research libraries in data curation. Additional efforts are needed to embed research data management skills into educational programmes (Engelhardt et al. 2022).

Focusing on the experience of the Journal of Open Humanities Data (JOHD), this poster examines current work to promote peer-reviewed data papers across the humanities. JOHD publishes two kinds of fully peer-reviewed data papers: short papers containing a concise and structured description of a humanities dataset and full-length papers discussing methods, challenges, and limitations in the creation, collection, management, access, processing, or analysis of humanities research data.

Since its launch in 2015, JOHD has grown significantly and has become an important player in raising awareness of data sharing and open data publishing in the humanities. One such example is the social media campaign #showmeyourdata where JOHD authors were invited to post a tweet showcasing an image of their datasets. Moreover, McGillivray et al. (2022) analyzed the effect that publishing data papers has on the citations of associated research articles and views of associated datasets.

The poster will also look at the place of humanities data publishing in the broader research publication ecosystem. Data publishing can contribute significantly to the development of a framework for evaluating data creation, sharing, and reuse in the context of the recently started reform of research assessment (CoARA, 2022) which encourages institutions to design and implement new assessment criteria that consider diverse research outputs beyond traditional journal publications – data papers should be included as they are well placed to demonstrate efforts of robustness and transparency of research.

Bibliography

Claudia Engelhardt et al. *How to be FAIR with your data*. Göttingen: Universitätsverlag Göttingen, 2022. http://dx.doi.org/10.17875/gup2022-1915.

Coalition for Advancing Research Assessment (CoARA), "Agreement on Reforming Research Assessment", July 20, 2022. https://coara.eu/agreement/the-agreement-full-text/

Wilkinson, Mark D., Michel Dumontier, IJsbrand Jan Aalbersberg, Gabrielle Appleton, Myles Axton, Arie Baak, Niklas Blomberg, et al., "The FAIR Guiding Principles for Scientific Data Management and Stewardship," *Scientific Data* 3 (2016): 160018. https://doi.org/10.1038/sdata.2016.18.

McGillivray, Barbara, Marongiu, Paolo, Pedrazzini, Nilo, Ribary, Marton, Wigdorowitz, Mandy, & Eleanora Zordan, "Deep Impact: A Study on the Impact of Data Papers and Datasets in the Humanities and Social Sciences," *Publications* 10(4) (2022): 39. http://dx.doi.org/10.3390/publications10040039.

FDM im materiellen Erbe von rund drei Millionen Jahren Menschheits- und Umweltgeschichte

Beispiele für die Einbeziehung der Forschungs-Community durch TRAILs in NFDI4Objects

Thiery, Florian
florian.thiery[at]leiza.de
Leibniz-Zentrum für Archäologie, Deutschland

Höke, Benjamin
benjamin.hoeke[at]rps.bwl.de
Landesamt für Denkmalpflege Baden-Württemberg, Deutschland

Keller, Christin
christin.keller[at]dainst.de
Deutsches Archäologisches Institut, Deutschland

Zusammenfassung. NFDI4Objects ist ein Konsortium innerhalb der Nationalen Forschungsdateninfrastruktur (NFDI), das sich dem materiellen Erbe von rund drei Millionen Jahren Menschheits- und Umweltgeschichte widmet. Dieses Paper stellt Aspekte verschiedener NFDI4Objects TRAILs (Task-Related Activities for the Implementation and Launch of services) vor. Dazu zählen Evaluierungen und Umfeldanalysen, die mit Hilfe der Forschungscommunity erarbeitet werden. Des Weiteren werden Methoden und Infrastrukturen vorgestellt, die es ermöglichen, FAIRe und nachvollziehbare Daten zu erzeugen. Zudem wird erörtert, wie Community Hubs (z.B. Wikidata und Wikipedia) zu beitragen können.

1 NFDI4Objects und TRAILs

NFDI4Objects (N4O) ist ein Konsortium innerhalb der Nationalen Forschungsdateninfrastruktur (NFDI), das sich dem materiellen Erbe von rund drei Millionen Jahren Menschheits- und Umweltgeschichte widmet[1]. Derzeit oft dezentral, projektgebundene und temporär gelagerte Forschungsdatenbestände werden im Rahmen von NFDI4Objects für das gesamte deutsche Wissenschaftssystem systematisch erschlossen, zugänglich gemacht und national wie auch international mit Community-Standards vernetzt. Das Konsortium tritt für

[1] vgl. Bibby, Bruhn, Dürhkohp et al. (2021).

die Etablierung der FAIR-Prinzipien[2] ein und begleitet aktiv die digitale Transformation der Arbeitsmethoden. NFDI4Objects stellt sich zudem der Aufgabe, große und komplexe Datenbestände aus Forschungsprozessen zu erschließen und gleichzeitig den nachhaltigen und langfristigen internationalen Zugang zu digitalen Ergebnissen von Forschungsprojekten entsprechend den Bedürfnissen der Nutzenden zu ermöglichen.

Die Zusammenarbeit und die Kommunikation mit und in den Fachcommunities wird über mehrere Strukturelemente durchgeführt: Community Cluster (CC), Temporary Working Groups (TWG) und Task-Related Activities for the Implementation and Launch of services (TRAILs), siehe Abb. 1.

[2] vgl. Wilkinson, Dumontier, Aalbersberg et al. (2016).

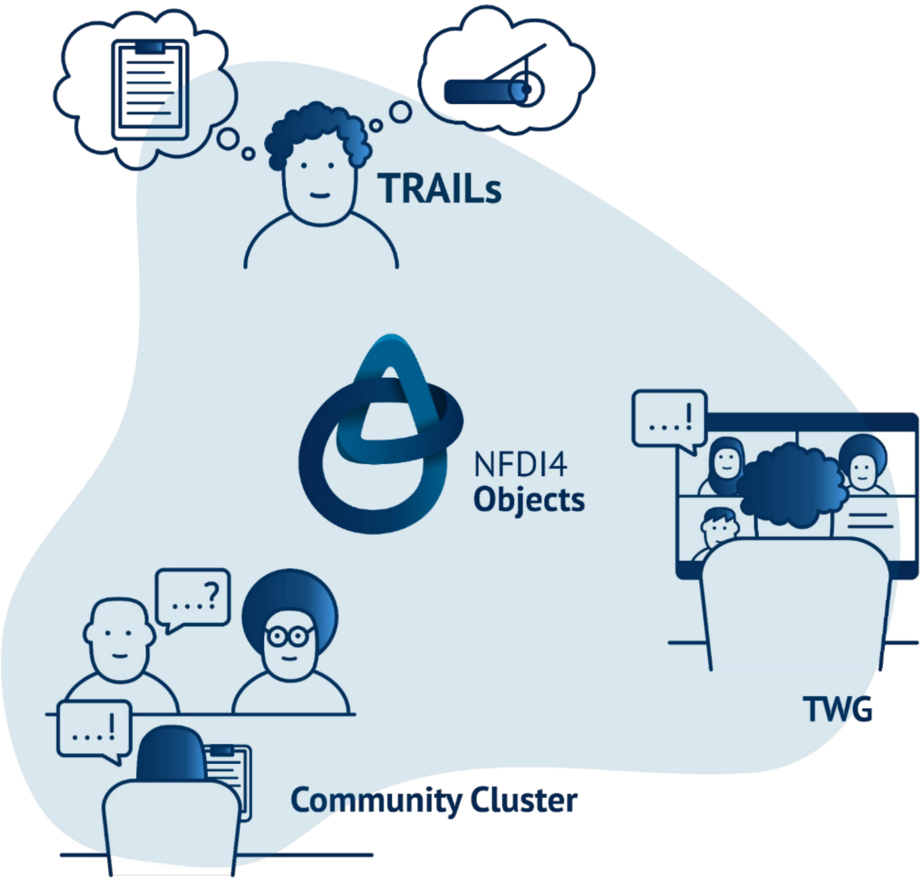


Abb. 1. Community-Interaktion in NFDI4Objects. Lizenz: Vanessa Liebler / i3mainz, CC BY-ND 4.0.

Bei der digitalen Transformation bedarf es einer kritischen Auseinandersetzung mit Forschungsdaten, z.B. in Bezug auf ihre Entstehung, der Provenienz (CARE-Prinzipien[3]) und Qualität. Hierbei entstehen durch die rapide steigende Anzahl heterogener und dezentral gespeicherter Daten besondere Herausforderungen zur FAIRen Modellierung und Bereitstellung von Forschungsdaten. Hierbei stehen insbesondere Ontologien, kontrollierte Vokabulare, Austauschformate und interoperable maschinenlesbare Services im Vordergrund.

[3] vgl. Carroll, Herczog, Hudson et al (2021).

Insbesondere semantische Modellierungen bilden eine strukturierte Basis für die Verwendung von KI-Technologien (Abb. 2).

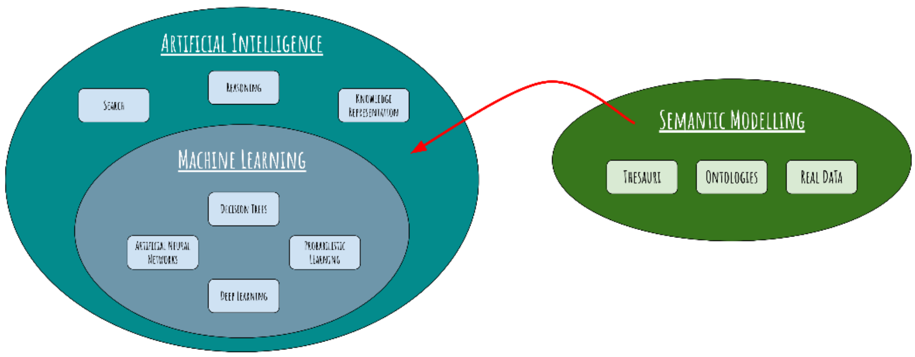


Abb. 2. Artificial Intelligence needs Semantic Modelling. Lizenz: Florian Thiery / LEIZA, CC BY 4.0, via Wikimedia Commons.

2 Beispiele für TRAILs

Verschiedenste Aspekte der Heterogenität geisteswissenschaftlicher Forschungsdaten in Bezug auf den Stand der Umsetzung der FAIR-Prinzipien sowie der Dokumentation und Beschreibung von Forschungsdaten werden in den TRAILs 1.1[4] (N4O TA1 for Documentation) und 2.1[5] (N4O TA2 for Collecting) aufgenommen.

Die FAIRedelung[6] [sic!] von Forschungsdaten und damit deren nachvollziehbare maschinenlesbare semantische Modellierung zur Abbildung fachspezifischer Kriterien und Steigerung der Datenqualität ist eines der Ziele von NFDI. In N4O for Collecting (TA2) geschieht dies vor allem durch Evaluation zur Modellierung von fuzziness and wobbliness, d.h. Vagheiten und Unsicherheiten. Zudem bieten Community Hubs wie Wikidata und Open Street Map auch Citizen Scientists die Möglichkeit zur Partizipation in der Forschungscommunity.

[4] vgl. Bibby, Höke, Lang (2021).
[5] vgl. Schäfer, Weisser, von Hagel et al. (2021).
[6] Wortschöpfung nach Prof. Dr. Cornelis Menke. Danke an Prof. Dr. Kai-Christian Bruhn für den Hinweis.

Ziel des TRAILs 2.2[7] ist das Sammeln, die Erweiterung und die Evaluation von Modellierungsansätzen zu fuzzyness and wobbliness[8] (Abb.3) in Forschungsdaten.

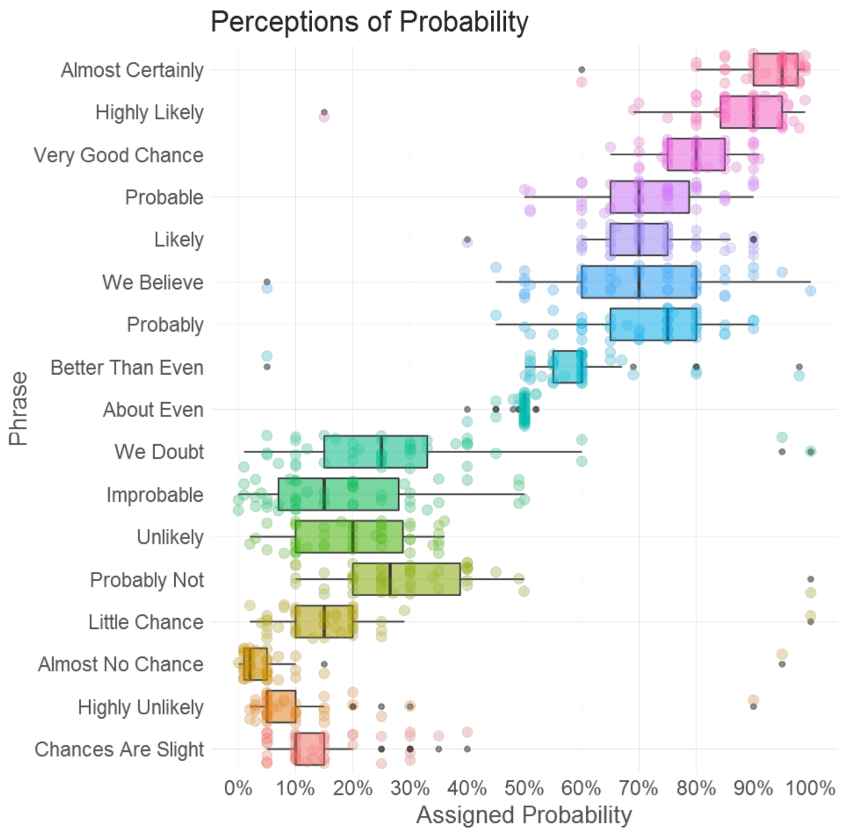


Abb. 3. Perceptions of Probability. This data was gathered using Reddit's /r/samplesize community. Lizenz: Zoni Nation @zonination, CC BY 4.0, via GitHub.

[7] vgl. Unold, Thiery, Mees (2019); Thiery, Mees, Tolle, Wigg-Wolf (2022); Tolle, Wigg-Wolf (2015); Thiery, Mees (2018).
[8] vgl. Thiery, Mees, Tolle, Wigg-Wolf (2021).

Beschreibende Elemente wie Texte und Bilder sind die Voraussetzung für eine vollständige nachvollziehbare Klassifizierung. Ausgewählte community-driven Vokabulare werden durch die Arbeiten in TRAIL 2.5[9] über den DANTE-Vokabularserver zur Verfügung gestellt und zusätzliche beschreibende Elemente wie freie Texte und Medien über das Wikimedia-Universum veröffentlicht. Dieser TRAIL mündet in einem generischen Workflow, der das Zusammenspiel mit Citizen Scientists in Wikidata, Wikipedia und Wikimedia Commons abbildet[10] (Abb. 4).

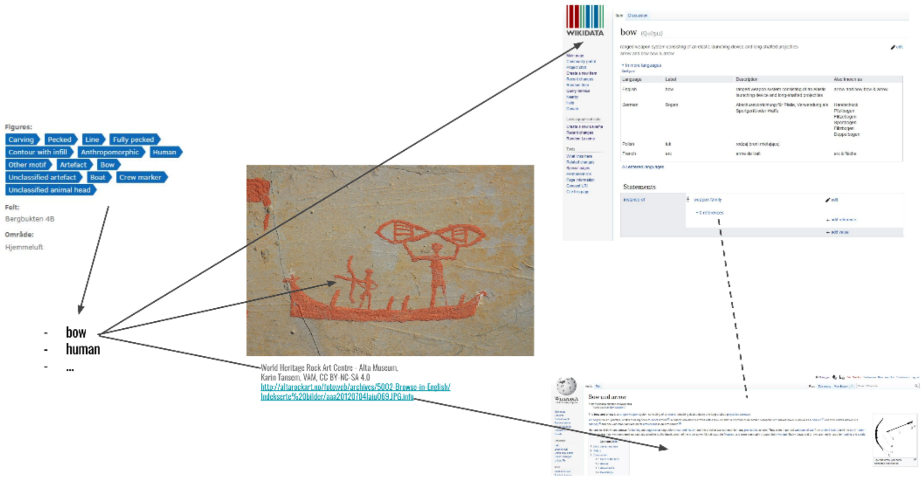


Abb. 4. Rock Art aus Alta und das Zusammenspiel mit dem Wikimedia Universe. Daten: World Heritage Rock Art Centre - Alta Museum, Wikidata und Wikipedia. Lizenz: Florian Thiery, CC BY 4.0.

Bibliografie

Bibby, D., K.-Chr. Bruhn, F. Dürhkohp, Chr. Eckmann, U. Himmelmann, B. Höke, Chr. Keller, u. a. 2021. „Digitales Forschungsdatenmanagement in der Archäologie und die Initiative NFDI4Objects“. *Blickpunkt Archäologie* 2 (2021): 150–63. https://doi.org/10.5281/zenodo.5823867.

[9] vgl. Thiery, Mees, Wienand, Börner (2021).

[10] vgl. z.B. Schmidt, Thiery, Trognitz (2022); Thiery, Veller, Raddatz et al. (2023); Thiery (2022); Thiery F., Distel, Schmidt und Thiery P. (2023), Thiery, Homburg, Schmidt, Voß, Trognitz (2021); Schmidt, Thiery (2022).

Bibby, David, Benjamin Höke, und Matthias Lang. 2021. „TRAIL 1.1: Directory and Evaluation of Existing Tools, Standards and Data". *NFDI4Objects TRAILs* 2021. https://doi.org/10.5281/zenodo.7870210.

Carroll, S.R., Herczog, E., Hudson, M. et al. 2021. "Operationalizing the CARE and FAIR Principles for Indigenous data futures." *Sci Data 8*, 108 (2021). https://doi.org/10.1038/s41597-021-00892-0.

Schäfer, Felix, Bernhard Weisser, Frank von Hagel, Florian Thiery, und Allard W. Mees. 2021. „TRAIL 2.1: Exploring the RDM Landscape in Museums and Collections". *NFDI4Objects TRAILs* 2021. https://doi.org/10.5281/zenodo.5849866.

Schmidt, Sophie C., und Florian Thiery. 2022. „SPARQLing Ogham Stones: New Options for Analyzing Analog Editions by Digitization in Wikidata". *CEUR Workshop Proceedings* 3110 (Graph Technologies in the Humanities 2020): 211–44. https://doi.org/10.5281/zenodo.6380914.

Schmidt, Sophie C., Florian Thiery, und Martina Trognitz. 2022. „Practices of Linked Open Data in Archaeology and Their Realisation in Wikidata". *Digital* 2 (3): 333–64. https://doi.org/10.3390/digital2030019.

Thiery, Florian. 2022. „SPARQLing Ogham – Irische Ogham-Steine als Linked Open Data". *ZfdG - Zeitschrift für digitale Geisteswissenschaften* Fabrikation von Erkenntnis – Experimente in den Digital Humanities. (Sonderband 5). https://doi.org/10.17175/SB005_010.

Thiery, Florian, Anne-Karoline Distel, Sophie C. Schmidt, und Peter Thiery. 2023. „Irische ᚑᚌᚆᚐᚋ Steine in OSM und Wikidata". *Squirrel Papers* 5 (1): No. 3. https://doi.org/10.5281/zenodo.7870480.

Thiery, Florian, Timo Homburg, Sophie C. Schmidt, Jakob Voß, und Martina Trognitz. 2021. „SPARQLing Geodesy for Cultural Heritage

New Opportunities for Publishing and Analysing Volunteered Linked (Geo-)Data". *FIG Peer Review Journal* FIG e-Working Week 2021 – Virtually in the Netherlands 21-25 June 2021. https://doi.org/10.5281/zenodo.5639381.

Thiery, Florian, und Allard Mees. 2018. „Taming Ambiguity - Dealing with doubts in archaeological datasets using LOD“. *Squirrel Papers* 4 (4): No. 2. https://doi.org/10.5281/zenodo.7361759.

Thiery, Florian, Allard Mees, Karsten Tolle, und David Wigg-Wolf. 2021. „TRAIL 2.2: Evaluation of Fuzziness and Wobbliness in Numismatics and Ceramology“. *NFDI4Objects TRAILs* 2021. https://doi.org/10.5281/zenodo.5654897.

Thiery, Florian, Allard W. Mees, Kasten Tolle, und David G. Wigg-Wolf. 2022. „How to Handle Vagueness and Uncertainty in Graph-Based LOD Knowledge Modelling? Dealing with Archaeological Numismatic and Ceramological Real World Data.“ *Squirrel Papers* 4 (1): No. 2. https://doi.org/10.5281/ZENODO.7184523.

Thiery, Florian, Allard Mees, Johannes Wienand, und Susanne Börner. 2021. „TRAIL 2.5: A Workflow for Enhancing Iconography Authority Data in the Wikimedia Universe“. *NFDI4Objects TRAILs* 2021. https://doi.org/10.5281/zenodo.5849809.

Thiery, Florian, Jonas Veller, Laura Raddatz, Louise Rokohl, Frank Boochs, und Allard W. Mees. 2023. „A Semi-Automatic Semantic-Model-Based Comparison Workflow for Archaeological Features on Roman Ceramics“. *ISPRS International Journal of Geo-Information* 12 (4): 167. https://doi.org/10.3390/ijgi12040167.

Tolle, Karsten, und David Wigg-Wolf. 2015. „Uncertainty Handling for Ancient Coinage“. In CAA2014. 21st Century Archaeology. Concepts, Methods and Tools. Proceedings of the 42nd Annual Conference on Computer Applications and Quantitative Methods in Archaeology., herausgegeben von François Giligny, François Djindjian, Laurent Costa, Paola Moscati, und Sandrine Robert, 171–78. Oxford: Archaeopress.

Unold, Martin, Florian Thiery, und Allard Mees. 2019. „Academic Meta Tool. Ein Web-Tool zur Modellierung von Vagheit“. *ZfdG - Zeitschrift für digitale Geisteswissenschaften* Die Modellierung des Zweifels – Schlüsselideen und konzepte zur graphbasierten Modellierung von Unsicherheiten. (Sonderband 4). https://doi.org/10.17175/SB004_004.

Wilkinson, M., Dumontier, M., Aalbersberg, I. et al. 2016. “The FAIR Guiding Principles for scientific data management and stewardship." *Sci Data* 3, 160018 (2016). https://doi.org/10.1038/sdata.2016.18.

Text+ – von der Zusammenkunft von Daten, Werkzeugen und Infrastruktur

Weimer, Lukas
weimer[at]sub.uni-goettingen.de
Niedersächsische Staats- und Universitätsbibliothek Göttingen, Deutschland
ORCID-iD: 0000-0001-6919-3646

Annisius, Marie
m.annasius[at]dnb.de
Deutsche Nationalbibliothek, Deutschland

Dogaru, George
george.dogaru[at]gwdg.de
Gesellschaft für wissenschaftliche Datenverarbeitung mbH Göttingen, Deutschland

Stein, Regine
regine.stein[at]sub.uni-goettingen.de
Niedersächsische Staats- und Universitätsbibliothek Göttingen, Deutschland
ORCID-iD: 0000-0003-3406-5104

Zusammenfassung. Die Task Area Infrastruktur/Betrieb des NFDI-Konsortiums Text+ tritt für die Text+-Datendomänen Sammlungen, Lexikalische Ressourcen und Editionen als Infrastrukturprovider auf, befördert deren Vernetzung und unterstützt dadurch eine verbesserte Interoperabilität unterschiedlicher geisteswissenschaftlicher Teilbereiche mit ihren verschiedenartigen Forschungsdaten sowie in die gesamte NFDI hinein. Dies geschieht in der konsortiumsinternen Zusammenarbeit durch die Bereitstellung von Kollaborationstools, für die gesamte geisteswissenschaftliche Community zum Beispiel durch die gemeinsame Nutzung der GND oder die Bereitstellung eines JupyterHubs. Das Poster stellt die Task Area Infrastruktur/Betrieb sowie die genannten Beispiele und deren praktische Einbindung in die Arbeit anhand von Anwendungsfällen vor.

Das auf sprach- und textbasierte Forschungsdaten fokussierte NFDI-Konsortium Text+ (https://www.text-plus.org/) konzentriert sich zunächst auf die drei Datendomänen digitale Sammlungen, lexikalische Ressourcen und Editionen. Diese drei Datendomänen bearbeiten unterschiedliche wissenschaftliche Teilbereiche der sprach- und textbasierten Forschung und nutzen und produzieren daher teils stark unterschiedliche Forschungsdaten und Methoden. Um die Arbeit der Datendomänen innerhalb des Konsortiums zu unterstützen und wo möglich zu harmonisieren, wirkt die weitere Task Area Infrastruktur/Betrieb (Infrastructure/Operations, IO) als Schnittstelle. Sie stellt auf der einen Seite grundlegende Dienste der projektinternen Zusammenarbeit zur Verfügung, die alle Mitarbeitenden des Konsortiums nutzen können, unabhängig davon, in welcher Datendomäne ihre Arbeit verortet ist (Weimer 2022). Gleichzeitig erarbeitet sie in enger Zusammenarbeit mit den Datendomänen übergreifende Lösungen, die deren Forschungsdaten interoperabel vernetzen und gemeinsam auffindbar machen sollen. Das Poster zeigt Beispiele, welche Wege das Konsortium eingeschlagen hat und wie so der Gefahr der “Versäulung” der Datendomänen entgegengewirkt wird.

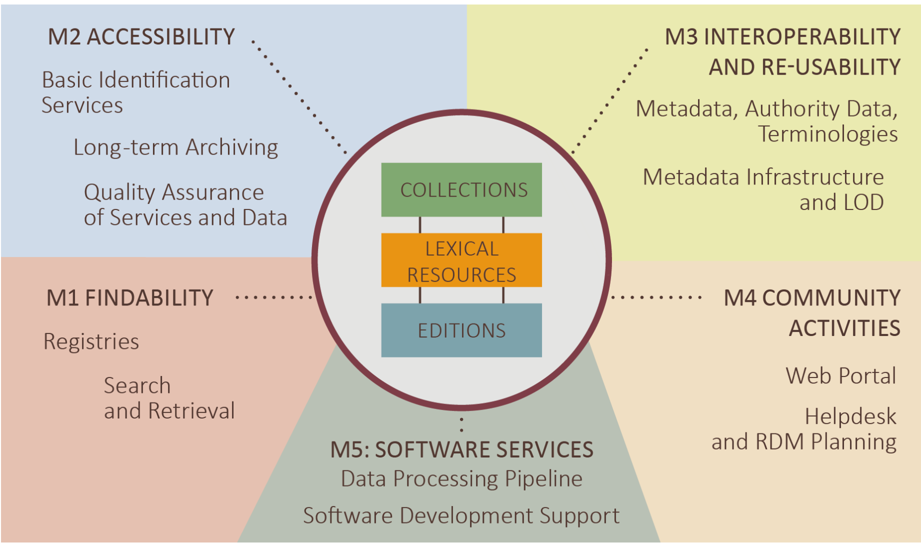


Abb. 1: Die Arbeitsschwerpunkte der Task Area Infrastruktur/Betrieb

Daten vernetzen und verwalten: Die Gemeinsame Normdatei (GND)

Die Gemeinsame Normdatei (GND) stellt die größte Normdatensammlung für Kultur- und Forschungsdaten im deutschsprachigen Raum dar. Durch ihre Öffnung wird sie zunehmend zum zentralen Angebot von Knotenpunkten im Netz und

Sucheinstiegen. Sie ist ihrerseits mit Normdateien anderer Nationalbibliotheken verknüpft.

Die GND fungiert in Text+ als Querschnittsthema und erstreckt sich über alle Datendomänen. Die Vernetzung von Daten mittels GND ermöglicht die kooperative Nutzung, Verwaltung und Verlinkung dieser Daten, unabhängig von Textsorten und domänenübergreifend. Mit dem Aufbau einer GND-Agentur für sprach- und textbasierte Forschungsdaten durch IO wird der beständige Ausbau der Vernetzung mittels GND innerhalb von Text+ weiter vorangetrieben (Annisius et al. 2022).
Aber auch durch die Arbeit anderer GND-Agenturen sowie durch das GND4C-Projekt wird die Nutzung der GND und damit die Vernetzung über das Konsortium hinaus auf NFDI-Ebene und mit anderen Communities bestärkt.
Zahlreiche, wiederkehrende Forums-Veranstaltungen bilden ein verbindendes Element, um die GND thematisch in den Communities zu verankern und auf Entwicklungen sowie deren Bedarfe einzugehen (Annisius/Buddenbohm 2023).

Jupyter Notebooks: Daten analysieren und verarbeiten
Jupyter Notebooks eignen sich für die Analyse, Verarbeitung und Visualisierung kleiner wie großer Datenmengen, für NLP, Machine Learning und vieles mehr. IO stellt einen JupyterHub bereit und arbeitet an domänenspezifischen Notebook-basierten Beispiellösungen und Verfahren (Dogaru/Jander 2022). Diese veranschaulichen die Vorteile der Jupyter Notebooks und ihre Eignung für verschiedenste Aufgaben. Viele der User Stories von Text+ thematisieren die Schwierigkeiten im Umgang mit Daten, die sich oft nur schwer produzieren und wiederverwenden lassen (Rißler-Pipka et al. 2021). Dem kann durch eine Jupyter-Infrastruktur (leichte Bedienung durch Bereitstellung vorinstallierter Software und Notebooks, Einbindung von Storage für möglichst direkten Zugriff auf Daten, Zugriff auf High Performance Computing Ressourcen) und durch die Erstellung von Jupyter-basierten Werkzeugen entgegengewirkt werden. Die Erfahrung bei der Erarbeitung von Notebooks für domänenspezifische Aufgaben (wie etwa Transformation von XML-/JSON-Daten für die Präsentation, Suche/Annotation von Textdaten, interaktive Ansichten) zeigt das

enorme Potential dieser Technologie, die zum Mittel der Wahl für ein breites Spektrum an Aufgaben in allen Task Areas und darüber hinaus in der gesamten geistes- und kulturwissenschaftlichen Community werden kann.

Bibliographie

Annisius, Marie, & Buddenbohm, Stefan. (2023). Werkstattbericht: 2. GND-Forum Text+. Text+ Blog. https://textplus.hypotheses.org/3530

Annisius, Marie, Fischer, Barbara K., & Steckel, Alexander. (2022). Ein Baum in der GND. Text+ Plenary 2022 (TextPlusPlenary), Mannheim. Zenodo. https://doi.org/10.5281/zenodo.7249018.

Dogaru, George, & Jander, Melina. (2022). Jupyter-Notebooks: Ein Angebot für die Text+-Community. Text+ Plenary 2022 (TextPlusPlenary), Mannheim. Zenodo. https://doi.org/10.5281/zenodo.7251753.

Rißler-Pipka, Nanette, Barthauer, Raisa, Buddenbohm, Stefan, Calvo Tello, José, Friedrichs, Sonja, & Weimer, Lukas. (2021). Community Involvement in Research Infrastructures: The User Story Call for Text+ (1.0.0). Zenodo. https://doi.org/10.5281/zenodo.5384085.

Weimer, Lukas. (2022). Zusammenarbeit innerhalb von Text+: Tools und Dienste. Text+ Blog. https://textplus.hypotheses.org/609.

Kompetenzzentrum OCR

Automatische Texterkennung als Serviceangebot

Will, Larissa
larissa.will[at]uni-mannheim.de
Universitätsbibliothek Mannheim, Deutschland
ORCID-iD: 0009-0004-6220-8939

Huff, Dorothee
dorothee.huff[at]uni-tuebingen.de
Universitätsbibliothek Tübingen, Deutschland
ORCID-iD: 0000-0003-0866-9967

Zusammenfassung: Die Möglichkeiten, die verschiedenen Programme im Bereich automatisierter Texterkennung heutzutage bieten, sind vielfältig. Deren Anwendung sowie die Vor- und Nachverarbeitung der Digitalisate ist jedoch 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 zu diesem Thema. Das umfangreiche Know-how im Bereich automatisierte Texterkennung und die verschiedenen Serviceangebote des Kompetenzzentrums sollen hier erläutert werden und Wissenschaftler*innen hinsichtlich der Einsatzmöglichkeiten von Texterkennungssoftware informiert werden.

Abstract

Durchsuchbare Volltexte für historische Drucke und Handschriften bieten einen zeitgemäßen, umfassenden Zugang zum Kulturgut und können als Grundlage für Anwendungen im Bereich der Data Science dienen.[1]
Die Möglichkeiten, die die verschiedenen Programme 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

[1] Weil und Kamlah, 2019.

Informationseinrichtungen und wissenschaftliche Projekte in Baden-Württemberg und darüber hinaus zu diesem Thema.[2]

Das Kompetenzzentrum kann ein breites Know-how für unterschiedliche Programme wie z. B. Tesseract[3], Transkribus[4] und eScriptorium[5] vorweisen. Die UB Mannheim ist aktuell mit zwei Teilprojekten an OCR-D[6] beteiligt, wodurch auch hier Synergien entstehen.[7] Nach Auslaufen des Projekts OCR-BW 2022 werden die Services im Sinne der Nachhaltigkeit 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.[8] 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.[9] 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

[2] Weil und Kamlah, 2020; Projektübersicht OCR-BW, 2023.
[3] Tesseract OCR, 2023.
[4] READ-COOP, 2023.
[5] Scripta/escriptorium, 2023.
[6] OCR-D, 2023.
[7] Projekte der UB Mannheim, 2023.
[8] Gehrlein et. Al, 2020.
[9] Weil, 2018.

des Trainingsmaterials oder sogar der Modelle möglich und wie können diese 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.[10]
Seit November 2022 bietet das Kompetenzzentrum zudem das niedrigschwellige Angebot einer offenen Online-Sprechstunde an.[11] 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.[12]

Auf diesem Poster soll das Serviceangebot der Universitätsbibliotheken Mannheim und Tübingen im Bereich der automatischen Texterkennung für Forschende vorgestellt und Wissenschaftler*innen über die Einsatzmöglichkeiten von Texterkennungssoftware informiert werden.

Bibliografie

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/.

[10] eScriptorium/Universitätsbibliothek Mannheim, 2023.
[11] Will, 2022.
[12] OCR-BW, 2023.

„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/

Zerstörtes Kulturgut

Die kontextualisierte Aufbereitung von kulturellen Forschungsdaten

Wolter, Vivien
s2viwolt[at]uni-trier.de
Universität Trier, Deutschland

Alili, Julia
s2jualil[at]uni-trier.de
Universität Trier, Deutschland

Chudoba, Hendrik
s2hechud[at]uni-trier.de
Universität Trier, Deutschland

Zusammenfassung. Das Forschungsprojekt ‚Zerstörtes Kulturgut' wurde im Rahmen des Masterseminars ‚Praxis der Digital Humanities' im Sommersemester 2022 an der Universität Trier von vier Studierenden entworfen. Im Fokus steht dabei die Erhebung von Forschungsdaten zur Kontextualisierung von Kulturstätten, die durch Kriege zerstört wurden. Durch diese digitale Aufbereitung soll das Bewusstsein für zerstörte Kulturgüter geschärft werden. Wegen der noch immer thematischen Relevanz wird das Projekt weitergeführt und soll künftig um neue Inhalte und zusätzliche wissenschaftliche Standards erweitert werden.

1 Motivation

Aufgrund Putins Angriffskrieg auf die Ukraine rückte das Thema Krieg und die daraus resultierende Zerstörung seit Anfang 2022 in den Fokus. Neben den unabsehbaren Folgen für Menschen und dem Verlust essenzieller Infrastruktur, sind auch die Folgen für Kultur und Kulturgüter immens.

Wir setzten es uns zum Ziel im Rahmen unseres Projektes, Forschungsdaten zu Kulturgütern, die durch Kriege beschädigt oder zerstört wurden, kontextualisiert und gesammelt aufzubereiten. Von Beginn an war es uns wichtig, den Open Science Ansatz bestmöglich zu verfolgen und die FAIR-Prinzipien so gut wie möglich umzusetzen.

2 Status quo

Im Rahmen des Projektes wurde eine Webseite[1] mithilfe von GitHub Pages erarbeitet, die übersichtlich und benutzerfreundlich Informationen zu den einzelnen Kulturstätten und Kriegen aufzeigt. Die Entscheidung fiel auf GitHub, da dieser Dienst die wesentlichen Zwecke (kostenloser Serverspace, kurzfristige und niedrigschwellige Veröffentlichung der Webseite) erfüllte. Alle Dateien, Bilder, der Webseitencode und alle Webseitenänderungen sind in einem öffentlichen GitHub Repository[2] dokumentiert und einsehbar. Dadurch möchten wir den uns zum Ziel gesetzten Open Science Ansatz und das accessible-Prinzip umsetzen.
Wegen der zeitlichen Begrenzung von einem Semester und der geringen Anzahl an Projektmitgliedern, haben wir uns darauf verständigt, eine Auswahl der Kriege und zerstörten Kulturstätten ab 1991 zu betrachten, um den Fokus auf eine qualitative Ausarbeitung der Inhalte zu legen. Die Auswahl der einzelnen Kulturstätten wird vor allem von der Transparenz und Vertrauenswürdigkeit der Quellenlage abhängig gemacht. Die Informationen werden mittels Online-Zeitungsartikeln und Regierungswebseiten zusammengetragen. Alle referenzierten Links werden in der Wayback Machine[3] gespeichert, um die Seiten zu archivieren und hier nach dem reusable-Prinzip zu arbeiten.
Zusätzlich zu den inhaltlichen Informationen der Kulturgüter und Kriegen auf der Webseite, werden XML-Dateien im LIDO-Format zur wissenschaftlichen Nachnutzung zum Download zur Verfügung gestellt. Das LIDO-Schema wurde für den Austausch von objekt- und materialbezogenen Daten entwickelt[4]. Dieser Standard ist für unser Projekt flexibel genug, um die objektbezogenen Informationen der Kulturstätten auch in den Kontext der Kriege setzen zu können, im Rahmen dessen sie zerstört wurden.
Des Weiteren werden zum Zweck der Interoperabilität Wikidata-Nummern der Kulturstätten und Kriege mit einem Hyperlink auf der Webseite und zuzüglich vorhandene GND-Nummern in den XML-Dateien angegeben.

[1] Webseite ‘Zerstörtes Kulturgut’, https://zerstoertes-kulturgut.github.io/, zuletzt aufgerufen am 21.07.2023.

[2] Repository ‘Zerstörtes Kulturgut’, https://github.com/zerstoertes-kulturgut/zerstoertes-kulturgut.github.io, zuletzt aufgerufen am 21.07.2023.

[3] Wayback Machine, https://archive.org/web/, zuletzt aufgerufen am 21.07.2023.

[4] LIDO Schema v1.1, https://www.lido-schema.org/schema/latest/lido.html, zuletzt aufgerufen am 20.07.2023.

3 Weiterentwicklung

An der Umstellung der technischen Infrastruktur wird aktuell gearbeitet. So wird die Webseite von GitHub Pages auf eine dynamische Webseite auf Grundlage von Typo3 umgestellt, da wir von einem proprietären System zu einer Open Source Software wechseln möchten. In Zukunft wird in den XML-Dateien ein standardisiertes Vokabular verwendet, um auch hier eine höhere wissenschaftliche Qualität zu erreichen. Sobald diese fertiggestellt sind, möchten wir sie zur Archivierung, Nachnutzung und zur besseren Zitierbarkeit im DARIAH-DE Repository[5] und auf Zenodo[6] veröffentlichen.
Das Poster wird die Arbeit dieses studentischen Projekts mit besonderem Fokus auf die Erstellung der Metadaten und dem verwendeten LIDO-Schema vorstellen. Darüber hinaus wird das Weiterentwicklungspotenzial im Bereich des Datenmanagements und der Langzeitarchivierung diskutiert.

---

5 DARIAH-DE Repository, https://de.dariah.eu/repository, zuletzt aufgerufen am 23.07.2023.

6 Zenodo, https://zenodo.org/, zuletzt aufgerufen am 23.07.2023.

Migrating Research Data to Another Repository

Zinn, Claus
claus.zinn[at]uni-tuebingen.de
Universität Tübingen, Deutschland
ORCID-iD: 0000-0002-6067-5451

Trippel, Thorsten
thorsten.trippel[at]uni-tuebingen.de
Universität Tübingen, Deutschland
ORCID-iD: 0000-0002-7211-7393

Summary. Five years ago, we crafted a detailed scenario for migrating our research data from our locally-maintained, institutional repository to an external repository for which we had little control over. Now, with the rising cost of updating and maintaining our repository software to the latest version, we decided to realize the scenario step by step. We describe the challenges we encountered in the migration process.

1 Motivation

The maintenance of an institutional research data repository comes with substantial costs. While a large part of the efforts is devoted to data curation and ingestion as well as the communication with data depositors and consumers, there is a significant workload involved in keeping the repository software up-to-date. Costs related to software maintenance are rising, so we decided to migrate our research data to an external, infrastructural organization that is experienced with research data management, already hosts research data from a variety of other disciplines, and has trained staff. Five years ago, we crafted a workflow for this purpose (Trippel and Zinn, 2019). We have now put the workflow into motion and describe the challenges we encountered.

2 Technical background

Since 2010, our department offers a repository system to researchers of a Collaborative Research Centre in Linguistics. It holds around 650 data streams (mostly text-based, from psycholinguistic, experimental data to word embeddings), totaling around 600 gigabytes of storage. We started with a system based upon Fedora-Commons 3, which we extended with a number of bells and whistles (*e.g.,* a GUI to support data ingestion and rights management; an OAI-PMH port for data harvesting; export of ISO 24622-X (CMDI) based metadata with converters to Dublin Core and MARC-21). Due to security reasons, we later updated the system to Fedora-Commons 4. Security patches available for this version where applied whenever possible.

3 Migration Process

All research data is being migrated to the university-wide repository (https://fdat.uni-tuebingen.de/), which organizes all data into research communities. A community to host data from the "Tübingen Archive for LAnguage Resources" (TALAR) has been created in FDAT. The data archivists of the old (source) repository become the data curators of the new community in the (target) repository and keep their role as *data steward* for this community. Upon completion of the migration process, the old system is being shut down.

User authentication and authorization (AAI). A significant amount of research data was not publicly accessible in the source repository. The target repository has only limited AAI capabilities to restrict access. Data still under publication embargo will continue to be inaccessible to all users in the target repository. Interested parties must contact the data steward of the FDAT TALAR community. For new research data, our institution will continue to provide help to researchers who would like to archive their research data in a trusted, sustainable environment.

Metadata harmonization. In the source repository, all research data was described using CMDI-based metadata while the target repository requires DataCite. During conversion, to minimize information loss, resource-specific metadata was written into DataCite's description fields. Moreover, the CMDI metadata becomes part of the data stream so that users can access the original metadata. Also, we will continue to operate an OAI-PMH provider that offers CMDI-based metadata to third parties, *e.g.*, the Virtual Language Observatory (https://vlo.clarin.eu).

Persistent identification. Each dataset in the source repository was accessible through a persistent identifier using the handle system. Part identifiers were used to address individual files of a dataset. The target repository makes use of the DOI system. We therefore mapped the legacy handle to the DOI while omitting all part identifiers.

Bibliography

Thorsten Trippel and Claus Zinn, "Lessons Learned: On the Challenges of Migrating a Research Data Repository from a Research Institution to a University Library". *Language Resources and Evaluation* 55, 191-207 (2021), Springer.