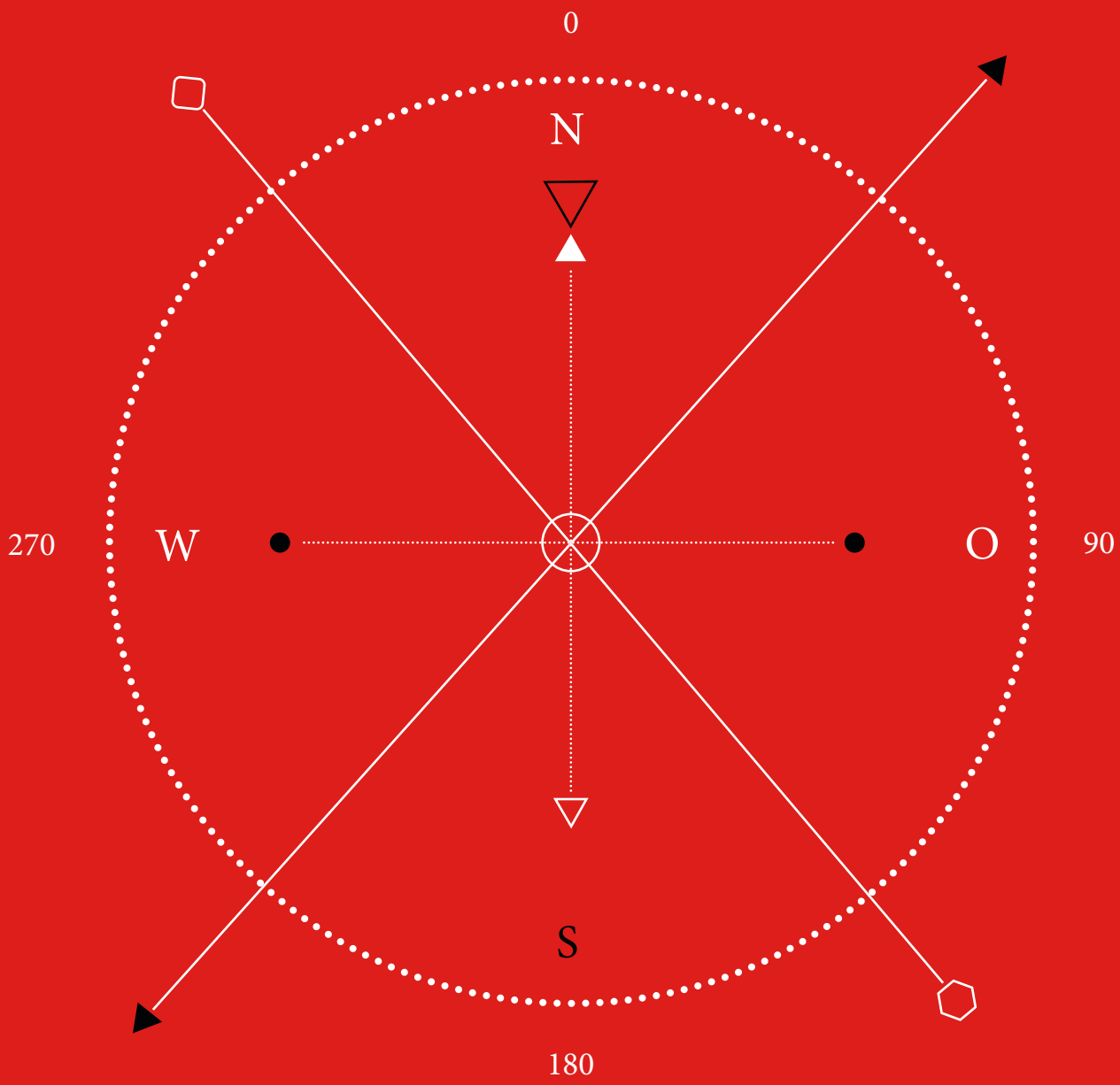
Copyright Julia Sorouri CC-BY 4.0



MAPPING THE LANDSCAPE

FORGE 21

Forschungsdaten in den Geisteswissenschaften

08.–10.09.2021 | Universität zu Köln

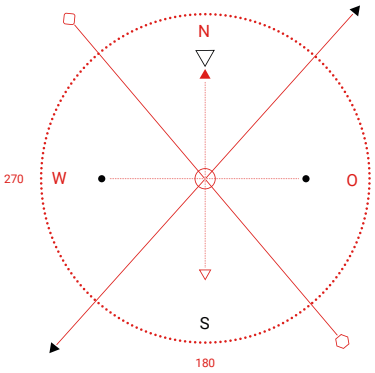
Gefördert und unterstützt durch:

Deutsche Forschungsgemeinschaft

digital humanities im deutschsprachigen raum

FORGE 2021

Forschungsdaten in den Geisteswissenschaften

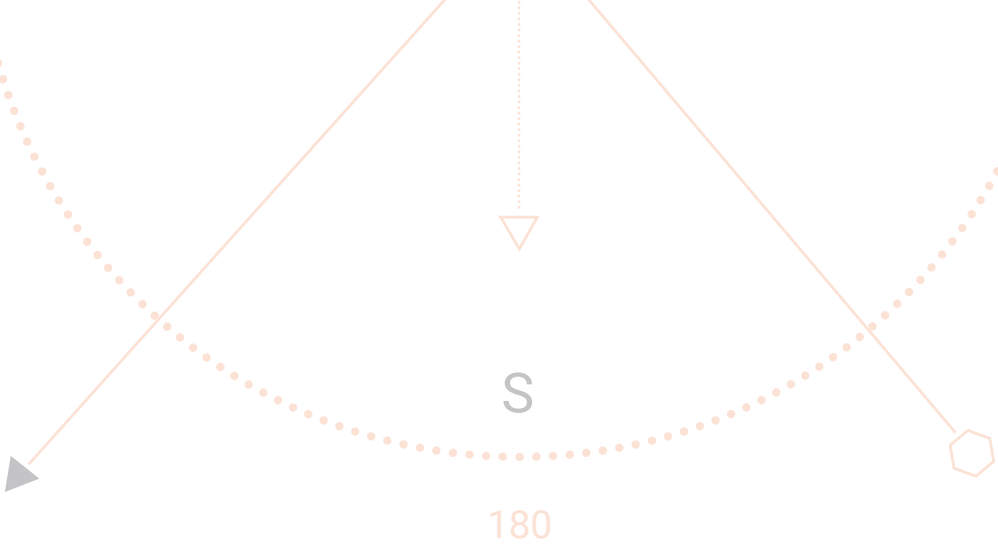


MAPPING THE LANDSCAPE

Geisteswissenschaftliches Forschungsdatenmanagement zwischen lokalen und globalen, generischen und spezifischen Lösungen

– Konferenzabstracts –

Universität zu Köln
08. bis 10. September 2021



Die Beiträge zur FORGE 2021 Konferenz in diesem Book of Abstracts haben ein Peer Review Verfahren durchlaufen. Davon ausgenommen ist der Beitrag „AG Datenzentren: Dokumentation von Forschungsdaten – Erfahrungen und Aufgaben aus der Praxis“ (S. 10).

Alle Beiträge sind unter einer CC-BY Lizenz veröffentlicht und wurden in der FORGE-Community auf Zenodo zusätzlich einzeln referenzierbar publiziert:
https://zenodo.org/communities/forge/.

Das Book of Abstracts zur FORGE 2021 Konferenz ist online verfügbar unter:
https://doi.org/10.5281/zenodo.5289381

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

Andreas Speer
Cologne Center for eHumanities (CCeH)
Universität zu Köln
Albertus-Magnus-Platz
50923 Köln

Øyvind Eide
Institut für Digital Humanities (IDH)
Universität zu Köln
Albertus-Magnus-Platz
50923 Köln

Layout-Redaktion
Rebekka Borges
Nicole Majka
Lukas Sept

Design
Julia Sorouri

Grußwort

Lange Zeit ein Thema nur für Spezialisten, ist das Management digitaler Forschungsdaten auch in den Geisteswissenschaften inzwischen zu einem zentralen Arbeitsfeld geworden. Datengetriebene Forschung, statistische Analysen und Auswertungen sowie grundsätzlich digitale Methoden prägen die akademische Landschaft vieler Fachbereiche. Ein Forschungsdatenmanagementplan ist obligatorischer Bestandteil eines Projektantrages. Nicht zuletzt durch die Nationale Forschungsdateninfrastruktur (NFDI) hat die Thematik eine neue Sichtbarkeit erlangt. Doch ebenso deutlich zeigen sich die damit verbundenen Herausforderungen an die Fachcommunities. Denn es wächst der Bedarf eines adäquaten Umgangs mit digitalen Daten und die Notwendigkeit der Gewährleistung ihrer langfristigen und nachhaltigen Auffindbarkeit, Zugänglichkeit, Interoperabilität und Nachnutzbarkeit.

Um dies sicher zu stellen, wurden in der Vergangenheit im deutschsprachigen Raum verschiedene institutionelle sowie institutionsübergreifende Daten- und Kompetenzzentren an unterschiedlichen Standorten eingerichtet, um (geisteswissenschaftliche) Forschende methodisch, strukturell und technisch bei Fragen des Forschungsdatenmanagements fachspezifisch zu beraten und zu unterstützen. Deren Zusammenschluss stellt die 2014 gegründete Arbeitsgruppe Datenzentren im Verband „Digital Humanities im deutschsprachigen Raum“ e.V. (DHd) dar, die als ein offenes Netzwerk gemeinsame Herausforderungen im geisteswissenschaftlichen Forschungsdatenmanagement adressiert.

Mit der FORGE Konferenz 2021 lebt nun eine Veranstaltungsreihe der AG Datenzentren wieder auf, die in den Jahren 2015 und 2016 an der Universität Hamburg ihren Anfang nahm und ein offenes Forum für den Austausch über zentrale Aspekte des Forschungsdatenmanagements aus einer geisteswissenschaftlichen Perspektive bildet. Der Konferenztitel „Mapping the Landscape – Geisteswissenschaftliches Forschungsdatenmanagement zwischen lokalen und globalen, generischen und spezifischen Lösungen“ verweist darauf, dass auch die FORGE 2021 ein Forum sein soll, in dem das unter einem stetigen Wandel stehende Feld des geisteswissenschaftlichen Forschungsdatenmanagements in seiner Heterogenität abgebildet und die aktuellen Fragen diskutiert werden. Entsprechend sind die Panels gestaltet. Sie wollen einen aktiven und lebhaften Diskurs Forschungsdatenmanagement in den Geisteswissenschaften mit dem Ziel der kritischen Auseinandersetzung mit individuellen Praktiken, Strukturen und Lösungsansätzen im Management digitaler Daten ermöglichen. Hierzu bietet die vorliegende Sammlung der Abstracts einen guten Überblick.

Das Management geisteswissenschaftlicher Forschungsdaten ist an der Philosophischen Fakultät der Universität zu Köln seit vielen Jahren bereits ein zentraler Aspekt wissenschaftlicher Forschungsprozesse. Mit dem Data Center for the Humanities (DCH) als geisteswissenschaftliches Datenzentrum, dem Cologne Center for eHumanites (CCeH) als Kompetenzzentrum für die digitalen Geisteswissenschaften und dem auf die Forschung und forschungsbasierte Lehre fokussierten Institut für Digital Humanities (IDH) verfügt der Standort über drei Einrichtungen, die das Management digitaler Daten immer wieder aus unter-

schiedlichen Perspektiven betrachten und bedienen. Entsprechend freuen wir uns sehr, nach der DHd Konferenz 2018 nun die FORGE Konferenz 2021 an der Universität zu Köln ausrichten zu dürfen und hoffen, dass wir das Kölner Konferenz-Gefühl auch über den digitalen Weg zu den Teilnehmenden und Beitragenden an ihren Heimatinstitutionen transportieren können.

Auch die virtuelle Umsetzung einer digitalen Konferenz bringt einen nicht zu unterschätzenden, organisatorischen Aufwand mit sich. Der Dank gebührt den vielen Helfer*innen vor Ort, die die professionelle Durchführung der FORGE Konferenz 2021 ermöglicht haben, ebenso der Deutschen Forschungsgemeinschaft und dem DHd-Verband/der AG Datenzentren für die großzügige Tagungsförderung.

Ferner möchten wir unseren Dank gegenüber dem Programmkomitee und den Gutachter*innen der Konferenz aussprechen, die durch ihre Reviews einen wichtigen Teil zur inhaltlichen Qualität und Gestaltung der Konferenz beigetragen haben.

Nicht zuletzt möchten wir uns bei allen Autor*innen für ihre spannenden und facettenreichen Beiträge zur FORGE Konferenz 2021 bedanken, die wir mit diesem Book of Abstracts gesammelt veröffentlichen.

Köln, September 2021

Prof. Dr. Dr. h.c. Andreas Speer
Sprecher des Cologne Center for eHumanities (CCeH)

Prof. Dr. Øyvind Eide
Sprecher des Data Center for the Humanities (DCH), Institut für Digital Humanities (IDH)

Inhaltsverzeichnis

Workshops

AG Datenzentren: Dokumentation von Forschungsdaten – Erfahrungen und Aufgaben aus der Praxis S. 10
Moeller, Katrin

Data Literacy Kompetenzprofil für historisch arbeitende Geisteswissenschaftler*innen: Welche Komeptenzen benötigen wir? S. 11
Lehmkuhl, Ursula; Lemaire, Marina

Daten publizieren, auffinden, nachnutzen: die DARIAH-DE Data Federation Architecture S. 17
Jander, Melina; Weimer, Lukas

Forschungsdaten in der Philosophie S. 24
Bartmann, Christoph; Eggert, Eric; Eschweiler, Mark; Geiger, Jonathan D.

Producing and Managing FAIR research data in the Humanities. An attempt using Geovistory & OntoMe. S. 28
Knecht, David; Beretta, Francesco

Über den Tellerrand geschaut: „Data Communities“ jenseits von generischen / fachspezifischen Ansätzen S. 32
Asef, Esther Marie; Huber, Elisabeth; Imeri, Sabine; Ommert, Eva; Rizzolli, Michaela; Wagner, Cosima

Wissenschaftsgeleitetes Forschungsdatenmanagement in der NFDI. Generische und spezifische Perspektiven am Beispiel geisteswissenschaftlicher Querschnittsthemen S. 38
Kammerer, Dietmar; Schrade, Torsten

Vorträge

(Korpus-)linguistische user stories und das UV2 Annotationstool S. 44
Weingart, Anja

A lot of data – Discover, ein Rechercheportal für granulare Forschungsdaten S. 50
Kümmet, Sonja

Aufbereitung und Publikation bibliographischer Forschungsdaten: Omeka S als Werkzeug zur (Meta-)Datenkuration und Wikidata als Plattform zur Forschungsdatendistribution und -publikation S. 57
Frank, Ingo; Weber, Albert

Auszugsweise Veröffentlichung geschützter Werke mit Annotationen steuern S. 65
Gärtner, Markus; Kleinkopf, Felicitas; Andresen, Melanie; Hermann, Sibylle

Der dha-catalogue: Ein Wissensgraph für die Exploration geisteswissenschaftlicher Forschungsdaten S. 71
Ďurčo, Matej; Schneider, Gerlinde

Discuss Data: Community-zentrierter Ansatz für das Forschungsdatenmanagement in den Geistes- und Sozialwissenschaften S. 77
Herrmann, Felix; Kurzawe, Daniel

Disziplinspezifisches Forschungsdatenmanagement – FDM-Bedarfserfassung in den Computational Literary Studies S. 83
Helling, Patrick; Jung, Kerstin; Pielström, Steffen

"Dreistufiges Konzept zur Beratung und Schulung von Forschungsdatenmanagement" - Erfahrungsgebricht S. 96
Jacob, Juliane

Ein Standard in der Praxis: ISO 24624:2016 Transcripton of spoken language S. 100
Schmidt, Thomas; Hedeland, Hanna; Frick, Elena

Einfach FAIR. Geisteswissenschaftliches Arbeiten und nachhaltiges Publizie-ren von Forschungsdaten mit Git S. 107
Arnold, Eckhart; Müller, Stefan; Gaab, Jochen; Landes, Florian; Raaf, Manuel

FDM lokal, gemeinsam und vernetzt gestaltet. Proaktiv zum nachhaltigen FDM S. 112
Sutter, Paul; Marciniak, Katja

Generische und fachspezifische Integration in der Forschungsinfrastruktur CLARIAH-DE S. 118
Buddenbohm, Stefan; Eckart, Thomas; Grandl, Tobias; Helfer, Felix; Jegan, Robin

Inter- und transdisziplinäres Forschungsdatenmanagement in der Ost-, Ostmittel- und Südosteuropaforschung S. 127
Štanzel, Arnošt; Frank, Ingo; König, Sandra; Körfer, Anna-Lena; Valena, Peter

Modellierung als Nachhaltigkeitsstrategie für Forschungsanwendungen S. 132
Fischer, Anna; Schildkamp, Philip; Demmer, Dennis; Neuefeind, Claes; Mathiak, Brigitte

Vom Speziellen ins Generisch und wieder zurück. Wie fachspezifisch muss ein FDM-Zentrum sein? S. 140
Wörner, Kai

Wer macht den Satz? Redaktionsprozesse für Forschungsdaten in den Geisteswissenschaften S. 143
Cremer, Fabian; Klaffki, Lisa; Steyer, Timo

Wie komme ich zu einer Forschungsdatenmanagement-Strategie? Eine Antwort gibt das DIAMANT-Modell S. 150
Lemaire, Marina

Zur Nachhaltigkeit von Visualisierungen in den digitalen Geisteswissenschaften S. 154
High-Steskal, Nicole; Liem, Johannes; Mayr, Eva; Windhager, Florian

Poster

A long-term archiving and dissemination ecosystem for Humanities' research data S. 166
Schneider, Gerlinde; Scholger, Martina; Scholger, Walter; Steiner, Elisabeth; Stigler, Johannes; Stoff, Sebastian; Tosques, Fabio; Vasold, Gunter; Vogeler, Georg; Zuanni, Chiara

Forschungs-Alltags-Infrastruktur S. 174
Fisseni, Bernhard; Arnold, Denis; Lang, Christian

Forschungsdaten.info – FDM-Angebote für die Geisteswissenschaften auf einen Blick S. 180
Böker, Elisabeth; Albrecht, Regina; Gabriel, Vanessa; Meier, Laura; Spenger, Martin; Brettschneider, Peter; Axtmann, Alexandra

Fun with VCS – morer with less. A Tool for Facilitating the Use of Git in Linguistic Research Data Management. S. 184
Ferger, Anne; Jettka, Daniel

Generische und disziplinspezifische Zugänge zur Qualität audiovisueller, annotierter Sprachaten im BMBF-Projekt QUEST S. 190
Wamprechtshammer, Anna; Arestau, Elena

NFDI4Objects: Eine NFDI-Konsortialinitiative zum materiellen Erbe aus etwa drei Millionen Jahren Menschheits- und Umweltgeschichte S. 196
Keller, Christin

Prozessmetadaten für die digitalen Geisteswissenschaften S. 200
Jung, Kerstin; Gärtner, Markus; Schembera, Björn

SDC4Lit – Science Data Center for Literature S. 205
Buck, Nina; Blessing, André; Bönisch, Thomas; Ganzenmüller, Andreas; Hein, Pascal; Hess, Jan; Holz, Alexander; Jung, Kerstin; Kaminski, Andreas; Kamzelak, Roland; Kramski, Heinz Werner; Kuhn, Jonas; Kushnarenko, Volodymyr; Schembera, Björn; Schlesinger, Claus-Michael; Ulrich, Mona; Viehhauser, Gabriel

Sprachwandel digital erfahrbar machen. Forschungsdatenmanagement in der Variationslinguistik S. 211
Ganswindt, Brigitte; Engsterhold, Robert

Virtuelle Forschungsdatenumgebungen – Struktur, Erkenntnis, Nachnutzung S. 216
Rosenkranz, Vinzenz; Schmidt, Karsten

Wir „gärtnern" im FDM und forschungsethische Datenlandschaften. Wilderness and management of research data, including research ethics. S. 222
Strauch, Annette

WORKSHOPS

AG Datenzentren: Dokumentation von Forschungsdaten – Erfahrungen und Aufgaben aus der Praxis

Moeller, Katrin
katrin.moeller[at]geschichte.uni-halle.de
Martin-Luther-Universität Halle-Wittenberg, Deutschland

Hinweis. Der Workshop der AG Datenzentren des Verbands „Digital Humanities im deutschsprachigen Raum“ e.V. ist dem Tagungsprogramm vorgelagert und war nicht Teil des Begutachtungsprozesses.

Die Nachnutzbarkeit und Transparenz von Forschungsdaten fällt umso leichter, je besser und umfassender sie auch dokumentiert und beschrieben werden. Während dem Forschungsdatenmanagementplan schon viel Aufmerksamkeit geschenkt wurde, bleiben Anforderungen und Bausteine zur Dokumentation von Forschungsdaten immer noch zu wenig thematisiert. Grundsätzlich gehören natürlich die Forschungsfragen und -hypothesen, die Erhebungs- und Auswertungsmethoden, die verwendeten Quellen, Tools und Programme sowie Versionierungen, Lizenzen und besondere Maßnahmen zur Qualitäts- und Rechtesicherung hinzu. Welche Anforderungen nun jedoch die einzelnen Datentypen und Forschungsmethoden erbringen, ist gerade in der heterogenen Landschaft der der geisteswissenschaftlichen Fächer und Methoden bisher wenig geklärt. Dabei unterscheiden sich Anforderungen zum Teil von der Dokumentation möglichst vieler Arbeitsschritte bis zur Dokumentation von lediglich publizierten Daten. Brauchen die Geisteswissenschaften etwa einen Standard wie DDI zur Dokumentation sozialwissenschaftlicher Interviews? Im Workshop sollen daher Erfahrungen zur Praxis von Datendokumentationen zusammengetragen und ein Konzeptpapier für ein Musterbeispiel diskutiert und weiterentwickelt werden. Ausgangspunkt sind dabei ganz unterschiedliche Datentypen, wie sie etwa im Rahmen der DHd-Tagung Experimente von der AG Datenzentren „Hands on Research Data“ (https://vdhd2021.hypotheses.org/178) thematisiert wurden. Dabei steht auch die Frage im Raum, wie generisch ein solcher Ansatz sein kann oder muss, um Anforderungen und Nutzungspraxis verlässlich zu gestalten.

Data Literacy Kompetenzprofil für historisch arbeitende Geisteswissenschaftler*innen:

Welche Kompetenzen benötigen wir?

Lehmkuhl, Ursula
Lehmkuhl[at]uni-trier.de
Universität Trier, Deutschland
ORCID-iD: 0000-0003-3100-8809

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

Zusammenfassung. Der als Barcamp organisierte Workshop lädt alle interessierten historisch arbeitenden Geisteswissenschaftler*innen dazu ein, gemeinsam die Frage zu diskutieren, welche Data-Literacy-Kompetenzen die historisch arbeitenden Wissenschaften benötigen. Ausgehend von den Barcamp-Ergebnissen der DHd2020 sehen wir drei größere Themenbereiche, denen sich die Teilgebenden des Barcamps widmen könnten: 1. Data Literacy und geschichtswissenschaftliche Datenkritik: Was können wir von der historischen Methode lernen? 2. Digital Literacy, Data Literacy und andere Literacies: Worin unterscheiden sich Strategien und Herangehensweisen? Wieviel technisches Know-How benötigen Historiker*innen? Wo sind die Grenzen zum Forschungsdatenmanagement? 3. Data Literacy in der geschichtswissenschaftlichen Ausbildung: Welche Datenkompetenzen sollten auch schon im Studium erlernt werden? Wie können wir die Vermittlung von Data Literacy-Kompetenzen in den Curricula verankern? Ist es sinnvoll verschiedene Niveaustufen auszudifferenzieren?

1 Thema

Data Literacy-Kompetenzen sind eine zentrale Voraussetzung für die FAIRe Produktion von Forschungsdaten. Deshalb spielt in den Diskussionen um die Bereitstellung von Forschungsdaten nach den FAIR-Prinzipien[1] das Thema Vermittlung und Aneignung von Data Literacy-Kompetenzen immer wieder eine zentrale Rolle. Aber welche Kompetenzen benötigen Historiker und Historikerinnen eigentlich und

[1] Wilkinson u. a. 2016.

wie können sie vermittelt werden? Zur Beantwortung dieser Fragen wird häufig auf generische Kompetenzprofile mit Bezug auf Data / Digital / Information / Statistical / Code u. v. a. Literacies[2] verwiesen. Doch zeigt sich in der Regel rasch, dass diese den spezifischen Anforderungen der historischen Quellen- und Datenkritik und dem Anspruch der „historischen Methoden" nicht gerecht werden. Die von Ridsdale et al. vorgeschlagene Kompetenzsystematik, die die Bereiche Datensammlung, Datenmanagement, Datenevaluation und Datenanwendung unterscheidet, unterschlägt beispielsweise den Umstand, dass die hier genannten Kompetenzbereiche bereits unter analogen Bedingungen zum Kern historisch-kritischen Arbeitens gehören. Generische Modelle helfen uns deshalb nicht weiter. Vielmehr besteht ein dringender Bedarf, ein disziplinspezifisches Kompetenzprofil für die historisch arbeitenden Wissenschaften zu entwickeln. Dazu ist es nötig, (1) die erforderlichen digitalen Kompetenzen im Umgang mit historischen Forschungsdaten zu identifizieren, (2) die historische Methode um den Bereich der digitalen Datenkritik zu erweitern und (3) historische Datenkritik und Data Literacy als Kernbestandteile der historischen Ausbildung curricular zu verankern. Im Rahmen des geplanten Barcamp wollen wir diese drei Bereich gemeinsam mit der Community – also aus der laufenden Lehr- und Forschungspraxis heraus – ausloten sowie inhaltliche Vorschläge und Umsetzungsstrategien entwickeln.

Ausgangspunkt unserer Diskussionen sind einerseits die methodischen Besonderheiten historischer Forschung und die Spezifik des historischen Quellen- bzw. Datenmaterials, das in der Regel äußerst heterogen und unstrukturiert ist. Andererseits ist die für die historische Methode spezifische Form der Datenkontextualisierung zu berücksichtigen. Zur Quellenkritik gehört stets auch die kritische Kontextualisierung der Entstehungs- und Überlieferungsgeschichte einer Quelle. Die Digitalisierung und die Anreicherung digitalisierter Quellen mit digitalen Merkmalen (Datafication) stellt gerade diesen Schritt der historischen Forschung vor erhebliche Herausforderungen. Neben der klassischen Quellenkritik müssen Historiker und Historikerinnen auch in der Lage sein, Algorithmen, Schnittstellen und digitale Infrastrukturen quellenkritisch zu evaluieren. Schließlich muss der Frage nachgegangen werden, was sich in der methodischen Herangehensweise verändert, wenn wir von der analogen Quelle zu

[2] Ridsdale u. a. 2015; Schüller, Koch, und Rampelt 2021; Schüller, Busch, und Hindinger 2019; GI.DE u. a. 2018; ACRL 2015; Schmidt und Shearer 2016.

Daten und digitalen Methoden übergehen, oder wenn wir analoge und digitale Quellen gemeinsam nutzen und auswerten wollen.
Zur Rahmung unserer Diskussionen können wir u. a. auf die Ergebnisse einer Session des Barcamps „Vermittlung von Data Literacy in den Geisteswissenschaften“[3] auf der DHd2020 zurückgreifen.[4] Hier wurde herausgearbeitet, „dass die traditionelle historische Quellenkritik sehr gut für die Integration digitaler Quellen in die historische Forschung geeignet ist“[5]. Allerdings müsste die Aus- und Weiterbildung von Historikern und Historikerinnen an die zunehmende Verfügbarkeit digitaler Quellen und den sich rapide ausdifferenzierenden digitalen Werkzeugkasten angepasst werden, damit die Geschichtswissenschaft nicht nur die digitalen Verfahren in ihren Methodenkoffer aufnehmen kann, sondern die infrastrukturellen Voraussetzungen dafür schafft, dass zukünftige Forschergenerationen die Geschichte einer digitalisierten Gesellschaft auf der Grundlage ihrer spezifischen Überlieferungen – in der Regel born-digital data – erforschen und beschreiben kann.

1.1 Workshopformat und Durchführung

Der geplante 3-stündige Workshop soll in Form eines Barcamp umgesetzt werden. Im Zentrum der Community-orientierten gemeinsamen Arbeit steht die Frage: Welche Data-Literacy-Kompetenzen brauchen historisch arbeitenden Wissenschaftler*innen? Da bei einem Barcamp die Teilgebenden selbst die Themen und Formate einer Session bestimmen, dienen die drei im Folgenden skizzierten Themenbereiche lediglich als grobe thematische Orientierung für die inhaltliche Gestaltung des Barcamps:

1. **Data Literacy und geschichtswissenschaftliche Datenkritik**: Was können wir von der historischen Methode lernen? → Datenkompetenz als Methodenkompetenz
2. **Digital Literacy, Data Literacy und andere Literacies**: Worin unterscheiden sich Strategien und Herangehensweisen? Wieviel technisches Know-How benötigen Historiker*innen? Wo sind die Grenzen zum Forschungsdatenmanagement?

---

[3] Lemaire and Wuttke 2020.
[4] Sicherung des Etherpads der Session 1 in Raum 2 „Kompetenzen in Data Literacy“ <https://github.com/dhd-ag-datenzentren/dhd-ag-datenzentren.github.io/blob/master/etherpad/Session1_Raum2_Barcamp_Data_Literacy-2020-03-31.pdf> (15.07.2021).
[5] Geiger u. a. 2020.

3. **Data Literacy in der geschichtswissenschaftlichen Ausbildung**: Welche Datenkompetenzen sollten auch schon im Studium erlernt werden? Wie können wir die Vermittlung von Data Literacy-Kompetenzen in den Curricula verankern? Ist es sinnvoll verschiedene Niveaustufen auszudifferenzieren?

Unabhängig von den hier skizzierten Themenbereichen erhalten die Teilgebenden zu Beginn des Barcamp die Möglichkeit, Themen und Formate für die gemeinsame Arbeit vorzuschlagen. Dazu richten wir einen Marktplatz ein, der den Teilgebenden ermöglicht, Ideen, Thesen und Inhalte, die sie gern diskutieren wollen, vorzustellen und Arbeitsformate vorzuschlagen. Auf diese Weise entwickeln wir gemeinsam das Arbeitsprogramm und den Ablaufplan. Die Ergebnisse der Sessions werden in einer kollaborativen Arbeitsumgebung (z.B. Etherpad oder Conceptboard) von den Teilgebenden gemeinsam dokumentiert. Es ist geplant, die Ergebnisse des Workshops in den eSciences Working Papers der Universität Trier zu publizieren. Auch hieran können sich die Teilgebenden beteiligen.

1.2 Ablaufplan

45 Minuten	**Begrüßung, Vorstellungsrunde, Marktplatz und Session-Planung**
5 Minuten	Pause
45 Minuten	**1. Session-Block**
5 Minuten	Pause
45 Minuten	**2. Session-Block**
5 Minuten	Pause
30 Minuten	**Abschlussrunde**

2 Bibliografie

ACRL, Association of College and Research Libraries. 2015. *Framework for Information Literacy for Higher Education*. Chicago, Illinois. http://www.ala.org/acrl/sites/ala.org.acrl/files/content/issues/infolit/framework1.pdf.

Geiger, Jonathan, Aline Deicke, Marina Lemaire, und Stefan Schmunk. 2020. „Digitale Quellenkritik: Quellenkritik 1.1 oder besser 2.0?“ *DHd-Blog* (blog). 8. Dezember 2020. https://dhd-blog.org/?p=14726.

GI.DE, Gesellschaft für Informatik e.V., Deutschen Mathematiker-Vereinigung e.V. DMV, Deutschen Physikalischen Gesellschaft e.V. DPG, und Gesellschaft Deutscher Chemiker e.V. GDCh. 2018. *Data Literacy und Data Science Education. Digitale Kompetenzen in der Hochschulausbildung*. Herausgegeben von Gesellschaft für Informatik e.V. GI.DE. Berlin. https://gi.de/fileadmin/GI/Hauptseite/Aktuelles/Aktionen/Data_Literacy/GI_DataScience_2018-04-20_FINAL.pdf.

Ridsdale, Chantel, James Rothwell, Michael Smit, Hossam Ali-Hassan, Michael Bliemel, Dean Irvine, Daniel Kelley, Stan Matwin, und Bradley Wuetherick. 2015. „Strategies and best practices for data literacy education. Knowledge synthesis report“. https://dalspace.library.dal.ca/bitstream/handle/10222/64578/Strategies%20and%20Best%20Practices%20for%20Data%20Literacy%20Education.pdf.

Schmidt, Birgit, und Kathleen Shearer. 2016. *Librarians' competencies profile for research data management*. Herausgegeben von Joint Task Force on Librarians' Competencies in Support of E-Research and Scholarly Communication. [Online]. https://www.coar-repositories.org/files/Competencies-for-RDM_June-2016.pdf.

Schüller, Katharina, Paulina Busch, und Carina Hindinger. 2019. *Future Skills: Ein Framework für Data Literacy.*

Kompetenzrahmen und Forschungsbericht. Herausgegeben von Geschäftsstelle Hochschulforum Digitalisierung beim Stifterverband für die Deutsche Wissenschaft e.V. Hochschulforum Digitalisierung. Arbeitspapier 47. Berlin. https://doi.org/10.5281/zenodo.3349864.

Schüller, Katharina, Henning Koch, und Florian Rampelt. 2021. „Data-Literacy-Charta". Herausgegeben von Bildung. Wissenschaft. Innovation Stifterverband. https://www.stifterverband.org/charta-data-literacy.

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: Artikelnr. 160018. https://doi.org/10.1038/sdata.2016.18.

Daten publizieren, auffinden, nachnutzen: Die DARIAH-DE Data Federation Architecture

Jander, Melina
jander[at]sub.uni-goettingen.de
Niedersächsische Staats- und Universitätsbibliothek Göttingen, Deutschland
ORCID-iD: 0000-0003-1646-6836

Weimer, Lukas
weimer[at]sub.uni-goettingen.de
Niedersächsische Staats- und Universitätsbibliothek Göttingen, Deutschland
ORCID-iD: 0000-0001-6919-3646

Zusammenfassung. Die DARIAH-DE Data Federation Architecture (DFA) adressiert fünf Aspekte des Research Data Lifecycles und deckt unter Berücksichtigung der FAIR-Prinzipien sowohl Datenpublikation und -speicherung als auch Beschreibung, Mapping und Auffindbarkeit ab. Dadurch geht die DFA weit über die Funktionalität eines reinen Repositoriums hinaus. Eingebettet in die mit BMBF-Förderung aufgebaute Architektur von DARIAH-DE und CLARIAH-DE garantiert die DFA Nachhaltigkeit und Support – bei gleichzeitig niedrigen Nutzungshürden. Denn für deren Verwendung ist lediglich ein DARIAH- oder Föderationsaccount nötig und alle Dienste sind kostenlos. Somit richtet sich das Angebot der DFA an Individualforschende, Forschungsinstitutionen und weitere Datengeber gleichermaßen.
Alle drei Gruppen wird der 90-minütige Grundlagenworkshop adressieren, indem er auf anwendungsbezogener Ebene die Funktionen des DARIAH-DE Publikators vorstellt. Der Publikator ist das in die DFA integrierte Dateneingabewerkzeug, mit dem Forschende ihre Daten für den Import ins DARIAH-DE Repository mit Metadaten anreichern, verwalten und final ins Repository einspielen können. Außerdem werden die Möglichkeiten der Collection Registry und der Generischen Suche sowie deren Einbettung in die DFA auf praktische Weise – im Idealfall mit dem Einspielen eigener Daten – ausprobiert und erläutert.

1 Einleitung

Nicht erst im Zuge der Empfehlungen des Rates für Informationsinfrastrukturen[1] streben Forschungsdatenrepositorien die möglichst FAIRe Bereitstellung ihrer Daten fernab von Individuallösungen an. Diese Forderung nach FAIRen Forschungsprinzipien und -methoden wurde lange auf den Weg gebracht[2] und ist im Wandel hin zum digitalisierten wissenschaftlichen Arbeiten so wichtig wie nie. So ist es nur folgerichtig, dass sich Forschungsdatenrepositorien an dieser Forderung orientieren. Um dabei die Nutzbarkeit zu erleichtern, versuchen Forschungsdateninfrastrukturen verstärkt, möglichst viele Schritte im Datenpublikations- und -archivierungsprozess in einem einzelnen Angebot verfügbar zu machen.
Darum vereint die DARIAH-DE Data Federation Architecture (DFA) insgesamt fünf Dienste in sich, die eine nachhaltige Publikation, aussagekräftige Beschreibung, Mapping und Auffindbarkeit der Forschungsdaten adressieren. Dabei stellt sie schrittweise jene Dienste bereit, die wichtige Aspekte des Research Data Lifecycles[3] abdecken. Der Workshop will diese Schritte gemeinsam mit den Teilnehmenden durchlaufen und somit zur Publikation eigener Forschungsdaten animieren. Dabei bietet er Raum für konstruktives Feedback, welches durch enge Zusammenarbeit der Workshop-Leitung mit den Entwicklern der DFA in deren technische Optimierung einfließt. Perspektivisch kann dies dazu beitragen, den Wandel innerhalb der akademischen Publikationskultur voranzutreiben.

2 Repositorien in den Geistes- und Kulturwissenschaften

Das Angebot an Forschungsdatenrepositorien ist vielfältig. Manche Repositorien beschränken sich auf einzelne Fachdisziplinen (z.B. AMAD), andere auf bestimmte Formate und Communities (bspw. Bildarchiv Foto Marburg oder DTA) oder Publikationstypen (Forschungs(roh)daten vs. Publikationen). Ferner bieten manche nur eine Archivfunktion ohne Veröffentlichung, viele können ihre Dienste nur

[1] RfII 2019.
[2] Schirmbacher & Müller 2009.
[3] Ein Überblick über den Research Data Lifecycle ist bspw. auf dem deutschsprachigen Informationsportal zu Forschungsdatenmanagement (FDM) *forschungsdaten.info* zu finden: forschungsdateninfo.de 2021. In Bezug auf DARIAH-DE bzw. die Geisteswissenschaften vgl. Puhl et al. 2015.

für eine sehr kurze Zeit aufrechterhalten, andere sind kostenpflichtig (z.B. RADAR).
Mit seinen Diensten geht das DARIAH-DE Repository, welches sich aktuell in der Antragsphase für die Zertifizierung mit dem CoreTrustSeal[4] befindet, über reine Archivfunktionen hinaus. Es ist durch die DFA und seine Anbindung an DARIAH- und CLARIAH-DE Teil größerer Infrastrukturen, die Nachhaltigkeit garantieren und an geisteswissenschaftliche NFDI-Konsortien angeschlossen sind. Um die Fachwissenschaft gezielt anzusprechen und einen thematischen Rahmen für die Forschungsdaten zu bieten, ist es zwar geistes- und kulturwissenschaftlich ausgerichtet, hierbei aber nicht an Einzelwissenschaften gebunden. Es ist ferner nicht rein institutionell und bietet neben zitierfähigen Links bspw. auch die Vergabe von Persistenten Identifikatoren (DataCite DOI und ePIC Handle). Technisch von der GWDG und SUB Göttingen betrieben, existiert gleichzeitig via CLARIAH-DE ein Helpdesk,[5] über den Fragen gestellt werden können. Darüber hinaus sind die Verwendung des Repositoriums sowie der gesamten DFA, die Speicherung von Daten und alle zusätzlichen Services kostenlos.[6]

3 Die DARIAH-DE Data Federation Architecture (DFA)

Im Rahmen digitalen Forschens gelten die FAIR-Prinzipien zunehmend als Leitlinien. Oftmals beziehen sich diese allerdings ausschließlich auf die Erzeugnisse von Forschung, bspw. digitalisierte Sammlungsbestände oder auch Forschungsergebnisse. Für Forschende ist es daher wichtig, dass schon der Publikationsprozess auf eine FAIRe Publikation hin ausgerichtet ist.[7]

[4] Siehe: https://www.coretrustseal.org/.
[5] CLARIAH-DE 2021.
[6] Für eine ausführliche Dokumentation des DARIAH-DE Repository siehe DARIAH-DE 2020.
[7] Ivanović et al. 2019.

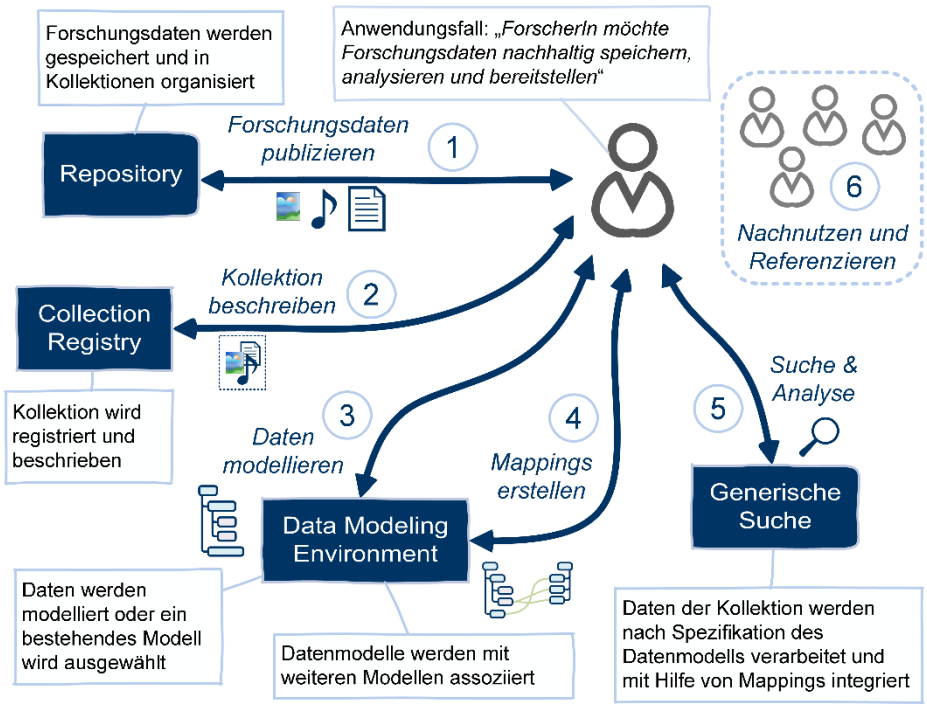


Abb. 1. *Schematische Darstellung der DARIAH-DE DFA.*

Diesem Anspruch wird das DARIAH-DE Repository (vgl. Abb. 1) – unter Nutzung u.a. der DARIAH-DE Authentifizierungs- und Autorisierungsinfrastruktur (AAI) sowie der DARIAH-DE Storage API zur Speicherung von Forschungsdaten[8] – mit seinem modularen Aufbau gerecht. Die DFA setzt sich aus dem Publikator, dem Repository, der Collection Registry, dem Data Modeling Environment und der Generischen Suche zusammen. Während der Publikator es ermöglicht, Forschungsdaten online über eine grafische Benutzeroberfläche ins Repository einzuspielen und dabei mit Metadaten zu versehen, können in der Collection Registry sowohl jene eingespielten Daten als Kollektionen registriert werden als auch externe Sammlungen durch entsprechende Beschreibungen auffindbar gemacht werden. Mit der Generischen Suche beinhaltet die DFA außerdem ein Front-End für die in der Collection Registry und dem Repository verfügbar gemachten Daten und deren Metadaten. Ein besonderer Vorteil ist dabei die kostenfreie Vergabe von Persistenten Identifikatoren (DataCite DOI und ePIC Handle), denn dadurch wird die dauerhafte Referenzierbarkeit der eingespielten Daten sichergestellt. Das Data Modeling Environment ist

[8] Schmunk & Funk 2016.

schließlich eine Komponente, in der Datenmodelle verzeichnet sind. Es ist ein Werkzeug zur Modellierung und Assoziation von Daten, worauf integrative Dienste wie die Generische Suche zugreifen.
Während es zum Kerngeschäft aller Forschenden gehört, Forschungsdaten zu erzeugen und zu analysieren, erfährt das nachträgliche Publizieren der Daten häufig nur wenig Aufmerksamkeit. Die Frage danach, wie mit möglichst geringem Aufwand und technischem Vorwissen die eigenen Forschungsdaten langfristig nachnutzbar gemacht werden können, ist deshalb von zentraler Bedeutung.
Auf diese Frage liefert die DFA mit dem Publikator eine nutzendenfreundliche und zeitsparende Antwort. Mit einem DARIAH- oder Föderationsaccount greifen Nutzende auf den Publikator zu und können mit wenigen Klicks mit dem Einspielen der Daten und Auszeichnung der Metadaten beginnen. Hierfür liegt das Dublin Core Metadatenschema als Standard zugrunde. Das Design von Eingabemaske und Fileupload wird von Forschenden als intuitiv wahrgenommen.[9] Durch die Einbindung des Publikators in die DFA ist dessen Funktionalität nicht auf den Datenupload beschränkt.

4 Durchführung des Workshops

Unser 90-minütiger Workshop wird den Teilnehmenden die Funktionen des Publikators sowie die Möglichkeiten der Collection Registry und die Einbettung in die DFA auf praktische Weise – im Idealfall mit dem Einspielen eigener Daten – nahebringen. Dabei wird nur kurz zum Thema hingeführt, um möglichst viel Zeit für die praktische Anwendung und den Nutzen der DFA zu haben. Er richtet sich an alle am Publikations- und Archivierungsprozess interessierten Forschenden in den Geistes- und Kulturwissenschaften. Technisches Know-How ist nicht von Nöten; der Workshop konzentriert sich auf die praxisnahe Durchführung und streift die zugrundeliegende Technik nur. Voraussetzungen zur Teilnahme sind folglich nur ein funktionierender Rechner und eine stabile Internetverbindung. Um die Tools der DFA nutzen zu können, ist ein Login in die DARIAH AAI nötig, der mit einem DARIAH-Account oder über viele Forschungsorganisationen und -institutionen möglich ist. Eine optionale Voraussetzung sind außerdem mitgebrachte eigene Forschungsdaten, die dann verwendet werden können, wenn wir gemeinsam den Publikationsprozess durchspielen.

[9] Cremer 2018.

Alternativ wird die Workshop-Leitung einen Beispieldatensatz bereitstellen.
Um eine optimale Durchführung zu gewährleisten und möglichst individuell Feedback geben zu können, ist die Anzahl der Workshopteilnehmenden auf 20 beschränkt.

Bibliografie

CoreTrustSeal. Zugriff am 2. Juli 2021. https://www.coretrustseal.org/

Cremer, Fabian. “DARIAH-DE Repository: Notizen zum Nutzen jenseits der Nutzung.” *DHd-Blog*, August 2018. Zugriff am 7. April 2021. https://dhd-blog.org/?p=10368.

CLARIAH-DE. “Support”. Zugriff am 7. April 2021. https://www.clariah.de/support.

DARIAH-DE. *DARIAH-DE Repository Documentation: Release 2020-06-25* (2020). https://repository.de.dariah.eu/doc/services/dhrep_doc.pdf.

Forschungsdateninfo.de. “Der Datenlebenszyklus. Stationen des Forschungsdatenmanagements,” Zugriff am 12. April 2021. https://www.forschungsdaten.info/themen/informieren-und-planen/datenlebenszyklus/.

Ivanović, Dragan, Birgit Schmidt, Rob Grim, and Alastair Dunning. “FAIRness of Repositories & Their Data: A Report from LIBER’s Research Data Management Working Group.” Zugriff am 05. Juli 2021. http://doi.org/10.5281/zenodo.3251593.

Puhl, Johanna, Peter Andorfer, Mareike Höckendorff, Stefan Schmunk, Juliane Stiller, and Klaus Thoden. “Diskussion und Definition eines Research Data LifeCycle für die digitalen Geisteswissenschaften.” *DARIAH-DE Working Papers* 11 (2015). urn:nbn:de:gbv:7-dariah-2015-4-4.

RfII – Rat für Informationsinfrastrukturen. *Herausforderung Datenqualität. Empfehlungen zur Zukunftsfähigkeit von Forschung im digitalen Wandel.* Göttingen 2019.

Schirmbacher, Peter, and Uwe Müller. “Das wissenschaftliche Publizieren – Stand und Perspektiven.” *cms-journal* 32 (2009): 7–12.

Schmunk, Stefan, and Stefan E. Funk. “Das DARIAH-DE- und das TextGrid-Repositorium: Geistes- und kulturwissenschaftliche Forschungsdaten persistent und referenzierbar langzeitspeichern.” *Bibliothek – Forschung und Praxis* 40, no. 2 (2016): 213–221.

Forschungsdaten in der Philosophie

Bartmann, Christoph
bartmann[at]ub.uni-koeln.de
FID Philosophie, Deutschland

Eggert, Eric
eggert[at]ub.uni-koeln.de
FID Philosophie, Deutschland

Eschweiler, Mark
eschweiler[at]ub.uni-koeln.de
FID Philosophie, Deutschland

Geiger, Jonathan
jonathan.geiger[at]adwmainz.de
DGPhil AG Philosophie der Digitalität, Deutschland
ORCID-iD: 0000-0002-0452-7075

Zusammenfassung: Die Digitalisierung erfasst bei ihrer gesamtgesellschaftlichen Durchdringung auch sämtliche Bereiche der Wissenschaft – dies schließt auch die Geisteswissenschaften mit ein. Die Philosophie als akademische Disziplin war bislang sehr zurückhaltend, was die Integration digitaler Methoden, Werkzeuge und Infrastrukturen für ihre Forschungsprodukte und -daten betrifft und es stellt sich die Frage, inwiefern philosophisches Forschen von systematischem Forschungsdatenmanagement profitieren kann und inwiefern dies den Forschungsprozess verändert. Dieser Transformationsprozess kann aber nur einerseits entlang der Bedarfe der Fachcommunity und andererseits entlang des technisch Machbaren erfolgen und muss daher als andauernder Diskurs realisiert werden, der die Digitalisierung (in) der Philosophie auch andauernd begleitet.
Der Workshop adressiert genau dieses Themenfeld und umfasst zu Beginn eine Podiumsdiskussion zwischen Expert*innen sowohl aus der philosophischen Fachcommunity als auch aus dem Bereich Technik und FDM, in der der status quo der Digitalisierung bzw. des FDMs in der Philosophie herausgestellt wird. Im Anschluss werden die Teilnehmenden in Arbeitsgruppen auf break-out-Räume mit spezifischen Problemstellungen verteilt und ihre Arbeitsergebnisse anschließend im Plenum vorgestellt und diskutiert. Die Ergebnisse des Workshops sollen im Nachgang sowohl dem Fachinformationsdienst Philosophie zugute kommen, als auch in die Arbeit der DGPhil Arbeitsgruppe zur philosophischen Digitalitätsforschung einfließen.

Die Digitalisierung erzwingt einen neuen Blick auch auf die geisteswissenschaftliche Forschung, inklusive ihrer Methoden, Praktiken, (Zwischen-)Produkte und Publikationen. Das Arbeiten mit digitalen Umgebungen, Werkzeugen und Daten eröffnet zahlreiche neue Möglichkeiten und Potenziale für die Forschung, gleichzeitig erfordert es aber auch eine Umdenken hinsichtlich Kompetenzen, Workflows und institutioneller Infrastrukturen. Die Digital Humanities sind hierbei als Arbeitsfeld kein neues Phänomen, doch spätestens mit der Nationalen Forschungsdateninfrastruktur (NFDI) und ihren geisteswissenschaftlichen Konsortien gibt es von oberster Stelle aus einen Zugzwang, das digitale Arbeiten auch in die eher traditioneller gelebten Geisteswissenschaften zu integrieren. Dies bringt auch die Möglichkeiten mit sich, Forschung(sprodukte) stärker vernetzen und noch prominenter und nachhaltiger publizieren zu können.

Die Philosophie in Deutschland war bislang recht zurückhaltend, was diesen Perspektivwechsel betrifft. Explizite Ansätze zum Forschungsdatenmanagement lassen sich kaum finden – weder in den Studiengängen, noch in der praktizierten Forschung. Obwohl es einige digitale Editionen philosophischer Klassiker gibt, liegt die primäre Orientierung nach wie vor beim Buchformat. Auch wenn die philosophischen Fachbegriffe mit ihren teilweise immensen Begriffsgeschichten nicht von generisch-geisteswissenschaftlichen Lösungen für eine inhaltliche Feinerschließung hinreichend abgedeckt werden können, ist die Verwendung von Normdaten – wie kontrollierten Vokabularen – äußerst selten. Auch im Bereich Open Access und offene Bildungsressourcen hat die Philosophie die Fülle an Möglichkeiten kaum ausgeschöpft und auch die Daten, die in eher randständigeren Bereichen der Philosophie anfallen (z. B. formal-logische Strukturen in der analytischen Philosophie oder statistische Daten in der Experimentalphilosophie) werden kuratorisch kaum berücksichtigt. Auch VREs (wie zum Beispiel Leselabore oder kollaborative Diskussionsplattformen) oder Forschungssoftware für den Zweig der “computational philosophy” genießen keinen prominenten Status und verharren in der Regel im Stadium von Inselprojekten.

Insgesamt bietet das digitale Arbeiten und proaktives Forschungsdatenmanagement für die Philosophie noch eine ganze Fülle an ungehobenen Schätzen. Für die konkrete Umsetzung stellt sich allerdings auch eine ganze Reihe an Fragen auf unterschiedlichen Ebenen: Was kann in der Philosophie alles als Forschungsdaten begriffen werden? Welche technischen Voraussetzungen braucht es für die Vernetzung der Daten?

Welche Tools und (Web)Services sind hierfür notwendig? Welche institutionellen Rahmenbedingungen braucht es, um Nachhaltigkeit sowohl für die Datenkuratierung als auch für den Betrieb und die Weiterentwicklung der digitalen Werkzeuge zu gewährleisten? Welche Bedarfe gibt es überhaupt von Seiten der fachwissenschaftlichen Community? Wie lassen sich diese Bedarfe erfassen? Welche neuen Kompetenzen sind für die Forschenden erforderlich und müssen hierfür die Curricula angepasst werden?(...)
Diese Fragen erfordern umfangreiche Diskussionen, da die Antworten notwendigerweise komplex ausfallen müssen und viele verschiedene Interessensgruppen beteiligt sind. Dabei kann die Philosophie von den Erfahrungen anderer Fachdisziplinen, die Ähnlichkeiten hinsichtlich der Forschungspraxis aufweisen (z. B. die Germanistik), profitieren, doch ist die Frage, inwiefern es auch Lösungen, Arrangements und Techniken erfordert, die spezifisch für die Forschungsgemeinschaft der Philosoph:innen sind. Der FID Philosophie begreift sich dabei sowohl als Vermittler und Infrastruktur, aber auch als Angebot an die Fachcommunity, den spezifisch philosophischen Ansatz bezüglich Forschungsdaten gemeinsam zu erarbeiten und zu gestalten.

Diesen Themenbereich (Fachinformationen und FDM in der Philosophie) will der hier skizzierte Workshop adressieren. Thematisch stehen die beiden Fragen “Was ist der status quo des FDMs in der Philosophie?” und “Welche Chancen bietet systematisches FDM für die Philosophie?” im Mittelpunkt. Der Ablauf des Workshops ist wie folgt: Nach einer kurzen Einführung beginnt der Workshop mit einer Podiumsdiskussion. Das Panel ist dabei besetzt mit Expert:innen für die philosophische Forschungslandschaft, aber auch für technische Infrastrukturen und praktisches Forschungsdatenmanagement. Die philosophisch-fachliche Seite soll dabei ebenso vertreten sein wie technische Perspektiven. Auch die Podiumsdiskussion widmet sich dem aktuellen Stand, den Möglichkeiten und Bedarfen des digitalen Forschungsdatenmanagements in der Philosophie und soll reichhaltige Impulse für die anschließenden Arbeitssessions geben. Diese schließen direkt an die Podiumsdiskussion an und werden als virtuelle break-out-Räume realisiert, die jeweils einen thematischen Schwerpunkt haben (z. B. “Forschungsdaten”, “Tools und VREs”, “Institutionen und Infrastrukturen”, “nachhaltiges Publizieren”). Die Teilnehmenden verteilen sich eigenständig auf die Räume und halten ihre Überlegungen, Ideen, Vorstellungen und Bedenken zu den jeweiligen Themenkomplexen fest (Etherpads werden zur Verfügung gestellt). Hieran schließt sich im letzten Abschnitt des Work-

shops eine offene Diskussionsrunde im Plenum an, in der die Arbeitsergebnisse aus den Arbeitsgruppen in den break-out-Räumen kurz referiert und diskutiert werden. Funktion dieser abschließenden Diskussionsrunde ist die Zusammenführung der gewonnenen Erkenntnisse und Überlegungen, da die verschiedenen Themenkomplexe zwar unterschieden betrachtet, aber nicht getrennt behandelt werden dürfen. Den Abschluss des Workshops bildet ein kurzes Wrap-up.

Der Workshop ist offen für alle, die ein Interesse an der Thematik haben. Zur Teilnahme sind explizit nicht nur Philosoph:innen, sondern auch Techniker:innen und Forschungsdatenmanager:innen eingeladen, um beide Perspektive maximal gewinnbringend zusammenzuführen. Die Arbeitsergebnisse – insbesondere artikulierte Bedarfe von Seiten der Fachcommunity, aber auch Anforderungen an Institutionen und Technologien etc. – sollen im Anschluss sowohl für die Arbeit des Fachinformationsdienstes (FID) Philosophie, also auch die Arbeitsgruppe philosophische Digitalitätsforschung in der Deutschen Gesellschaft für Philosophie aufgenommen werden. Wie auch die Digitalisierung ein andauernder Prozess ist, braucht es auch bei der Anpassung der Philosophie in Deutschland an eine adäquate digitale Arbeitsweise einen andauernden Diskurs, der diesen Anpassungsprozess reflektiert und community-driven begleitet. Der Workshop soll damit auch einen Beitrag zur Kartierung der Landschaft des Forschungsdatenmanagements der Geisteswissenschaften durch eine Verortung der Philosophie leisten.

Producing and Managing FAIR research data in the Humanities

An attempt using Geovistory Toolbox and OntoME

Knecht, David
david.knecht[at]kleiolab.ch
KleioLab GmbH, Basel, Schweiz

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

Alamercery, Vincent
LARHRA–CNRS/Université de Lyon/ENS de Lyon, Frankreich

Petram, Lodewijk
Huygens Institute for the History of the Netherlands, Amsterdam, Niederlande

Derks, Sebastiaan
Huygens Institute for the History of the Netherlands, Amsterdam, Niederlande

Schneider, Jonas
KleioLab GmbH, Basel, Schweiz

Introduction. Geovistory is a virtual research and data management environment for the humanities & social sciences. It supports the entire workflow of historical knowledge production: From the creation and management of academically validated statements about the past interlinked with textual and tabular sources, and the (spatial) analysis of research data, to the publication of research data in graphical interfaces and as rdf-triples. Coupled with OntoME it produces reusable research data and allows publishing results and data in accordance with the FAIR principles.

With the shift from a primarily analogue-based to a primarily digital-based working context in Humanities research, data modelling should by now have gained the interest of growing numbers of Humanities scholars. After all, in-depth analysis of digital research data is only feasible when the data are properly modelled. Furthermore, knowledge

of shared ontologies and vocabularies is indispensable when it comes to making data ready for reuse by humans and machines alike—an important prerequisite for replication studies and for compliance to the FAIR data principles. Thus far, however, scholars from all but a few sub-disciplines of the Humanities have largely ignored the subject of formal data modelling.[1]

This is set to become a pressing issue; ongoing digitization processes are about to multiply the amount of digital data that researchers in the Humanities have at their disposal. Data, moreover, that often contain vague, fragmentary, uncertain and contradictory observations.

While most members of the Digital Humanities community will be aware of the importance of proper data modelling and compliance with the FAIR principles, we observe that practice lags behind ideals. There is no reason why it should be like this: the task of aligning complex Humanities data with shared ontologies and vocabularies is not as tedious and difficult as scholars might think and brings multiple benefits. Concrete tools, such as Geovistory [2021] and OntoME [2021] are available to support researchers in this process.

Geovistory is an innovative digital tool that allows researchers and heritage professionals to manage their data and make it accessible to both the research community and the wider public. To this end, Geovistory digitally supports the entire workflow of historical knowledge production: From the collection of sources, to the creation and management of academically validated statements about the past or metadata (as research data) interlinked with textual and tabular sources, management of project-specific vocabularies, the creation of information networks and the (spatial) analysis of research data, to the publication of research data in graphical interfaces and as rdf-triples. Geovistory allows managing research data beyond a single project as one can continue building on the created research data or share it with colleagues without creating duplicates due to the innovative project-perspective of a unique information-graph of statements about the past shared among all projects in Geovistory.

In order to produce open and reusable data according to the FAIR principles, each statement in Geovistory is given a unique identifier (findable), modelled according to the CIDOC-CRM standard (ISO 21127:2014) and domain specific extensions (interoperable), produced under creative commons BY-SA 4.0 license (reusable) and will be published in machine and human readable interfaces (accessible). This digital research process is integrated with the “Ontology Management

[1] Gonzalez-Perez 2012.

Environment" OntoME, which is operated by the "Laboratoire de recherche historique Rhône-Alpes" (Lyon) within the Data for History Consortium. A poster of this workflow was presented at the annual Time Machine conference 2019 [Beretta et al. 2019].
This workshop will share a concrete experience of how this workflow has been applied to textual source material and various sets of structured data. It will shed light how a rather generic digital tool is geared towards managing a specific set of research data using a high-level ontology and extensions, how this data can be managed, curated, visualized and be used for further research. The workshop will end with a discussion with the participants on the benefits, but also challenges of working with such tools.
The workshop will present this workflow in three parts building on each other, followed by a Q&A session with the public:

- **Part 1:** Mapping the input data to existing ontologies using OntoME
  - *A collaborative Ontology Management system translating research questions into workflows of semantic data production*.
- **Part 2:** Importing the ontology & data into the data management platform Geovistory
  - *A virtual research environment provides services for FAIRification of historical research data*.
- **Part 3:** Visualizing and publishing the created knowledge graph
  - *Accessing and retrieving the data for further usage*.
- **Discussion**

Bibliographie

Beretta, Francesco, Vincent Alamercery, Sebastiaan Derks, Lodewijk Petram and Jonas Schneider. "Geohistorical FAIR data: data integration and Interoperability using the OntoME platform." *Time Machine Conference 2019*, Oct 2019, Dresden, Germany. https://halshs.archives-ouvertes.fr/halshs-02314003.

Geovistory. "Toolbox." Accessed August 01 2021. https://www.geovistory.com/.

Gonzalez-Perez, Cesar. “A conceptual modelling language for the humanities and social sciences.” *2012 Sixth International Conference on Research Challenges in Information Science (RCIS)* (pp. 1-6), May 2012. IEEE. DOI:10.1109/RCIS.2012.6240430.

OntoME. “Ontology Management Environment.”, Accessed August 01 2021. https://ontome.net/.

Über den Tellerrand geschaut: „Data Communities“ jenseits von generischen / fachspezifischen Ansätzen

Vorstellung und Diskussion eines Konzeptes

Asef, Esther Marie
esther.asef[at]fu-berlin.de
Freie Universität Berlin, Deutschland
ORCID-iD: 0000-0003-2411-4953

Huber, Elisabeth
ehuber[at]uni-bremen.de
Universität Bremen, Deutschland
ORCID-iD: 0000-0003-4084-4315

Imeri, Sabine
sabine.imeri.1[at]ub.hu-berlin.de
Humboldt-Universität zu Berlin, Deutschland
ORCID-iD: 0000-0002-8844-4014

Ommert, Eva
eva.ommert[at]uni-leipzig.de
Universität Leipzig, Deutschland
ORCID-iD: 0000-0001-9093-3477

Rizzolli, Michaela
michaela.rizzolli[at]fu-berlin.de
Freie Universität Berlin, Deutschland
ORCID-iD: 0000-0001-8154-6563

Wagner, Cosima
cosima.wagner[at]fu-berlin.de
Freie Universität Berlin, Deutschland
ORCID-iD: 0000-0003-4957-3478

Zusammenfassung. Alternativ zu einer vereinfachenden Unterscheidung von generischem und fachspezifischem Forschungsdatenmanagement (FDM) diskutiert der Workshop das Potenzial des Konzepts der „Data Communities“[1] für die geistes- und sozialwissenschaftlichen Disziplinen. Mit Blick auf die STEM-Disziplinen formuliert, werden „Data Communities“ als fluide und mehr informelle Gruppen von Forschenden verstanden, die über Praktiken der Arbeit mit und des Teilens von Daten eines bestimmten Typs miteinander verbunden sind – und nicht unbedingt durch disziplinäre Zuordnung. Entsprechend sollten die Praktiken und Interessen solcher „Data Communities“ der Ausgangspunkt sein, um das Teilen von Daten zu unterstützen und anzuregen. Der Workshop thematisiert angesichts heterogener Methoden und Forschungspraktiken in den Geistes- und Sozialwissenschaften Grenzen, Schnittstellen und übergreifende FDM-Praktiken mit Beiträgen aus unterschiedlichen Praxisfeldern. Dabei interessieren insbesondere drei Ebenen: die potenzielle Reichweite des Konzepts „Data Communities“, etwa hinsichtlich der Definition von Datentypen, die strukturellen und politischen Rahmenbedingungen sowie die Ausdifferenzierung der Archivierungslandschaft. Ziel ist es, Raum für die konzeptionelle Diskussion zu schaffen und zu eruieren, welche – strategischen wie operativen – Kriterien sich daraus für ein forschungsadäquat unterstützendes Datenmanagement ableiten lassen.

1 Hintergrund

Forschungsdatenmanagement (FDM) muss neben generischen Lösungen, die nur für einige Aspekte sinnvoll erscheinen, vor allem solche Praktiken und Infrastrukturen anbieten und etablieren, die den spezifischen Besonderheiten einzelner Forschungsbereiche möglichst umfassend gerecht werden – zu diesem Punkt besteht weitgehend Einigkeit. Auch die (bereits bewilligten oder geplanten) NFDI-Konsortien gehen in diese Richtung: Forschungsdatenmanagement ist vor allem spezifisch und richtet sich entlang von Disziplinen, Teildisziplinen oder Gruppen von Disziplinen aus.

Bei näherer Betrachtung ist diese Zuordnung aber keineswegs gleichmäßig oder eindeutig. Vielmehr ist in den Geistes- und Sozialwissenschaften die Ausdifferenzierung innerhalb einzelner Forschungsdisziplinen mitunter größer als es die disziplinäre Ordnung nahelegt – gerade in Wissenschaftsbereichen mit vielen „kleinen Fächern“, breit gefächertem Methodenspektrum und heterogener Forschungspraxis. Das gilt auch für interdisziplinäre Forschungszusammenhänge sowie interdisziplinär angelegte Fächer. Wie spezifisch fachspezifisches Datenmanagement hier sein kann bzw.

[1] Cooper und Springer 2019.

sollte oder mit Blick auf Infrastrukturen und andere Ressourcen so überhaupt realisierbar ist, scheint keineswegs klar. Womöglich sind in der Praxis die Forschungsgegenstände oder die angewandten Methoden entscheidender für Notwendigkeiten und Verfahren im Umgang mit Daten als die Zugehörigkeit zu einer Fachdisziplin?

Ein alternativer Ansatz der Strukturierung von Services und Infrastrukturen wird mit Blick auf die STEM-Fachdisziplinen unter dem Label „Data Communities" in einem Issue Brief der US-amerikanischen Non-Profit-Organisation Ithaka S+R diskutiert[2].

Als „emergent data communities" begreifen die Autorinnen „a loosely connected group of scholars who all work with a particular type of data, often linked by professional relationships"[3]. Data Communities zeichnen sich dadurch aus, dass es bereits gewisse formelle oder informelle Praktiken des Teilens / Nachnutzens von Daten gibt (bspw. auch die Nutzung von öffentlichen Daten). Folglich bestehe unter den Forschenden auch eine höhere Motivation, die Praktiken des Datenteilens voranzutreiben und zu institutionalisieren (bzw. daran mitzuwirken). Als wichtige Eigenschaften von Data Communities sehen die Autorinnen bereits vorhandene Standards, etwa in Bezug auf Dateiformate, Metadaten und ggf. Dokumentation / Kontextualisierung. Als Beispiele werden u.a. eine Datenbank mit spezifischen Informationen zu einer Fliegenart sowie die Forschung zum Thema Luftverschmutzung angeführt. Die Autorinnen argumentieren, dass in einem Bottom-up-Ansatz geeignete Communities identifiziert und gezielt gefördert werden sollten, etwa durch den Aufbau von an spezifische Bedürfnisse angepassten Repositorien.

2 Inhalte des Workshops

Der Workshop möchte das Konzept der Data Communities in die Diskussion im deutschsprachigen Raum einbringen, sein Potenzial für die Geisteswissenschaften eruieren und aus verschiedenen Perspektiven diskutieren. Er richtet sich gleichermaßen an FDM-Praktizierende wie an Forschende und schließt ausdrücklich auch die Sozialwissenschaften mit ein, zumal eine klare Abgrenzung in vielen Fällen ohnehin schwierig ist. Idealerweise sind möglichst verschiedene Forschungsbereiche / Fächer im Workshop vertreten. Beiträge können

[2] Cooper und Springer 2019.
[3] Cooper und Springer 2019, 16.

einen oder mehrere der folgenden Aspekte aufgreifen oder eigene Schwerpunkte setzen:

Data Communities in den Geistes- und Sozialwissenschaften: Hier ist von Interesse, wie die Teilnehmenden das Potenzial des Konzepts einschätzen und ob sie in ihren Arbeitsfeldern solche Communities identifizieren können. Was könnte z.B. als „a particular type of data" angesehen werden? Wären dies eher methodisch definierte Kategorien, wie z.B. Sprachkorpora, digitale Quelleneditionen, qualitative Interviewdaten, ethnografische Forschungsmaterialien, oder wäre hier eher an eine thematische Orientierung zu denken, wie z.B. Daten aus dem Bereich der Migrationsforschung? Oder auch beides, etwa Interviewdaten in Bezug auf Migration?

Strukturelle Bedingungen und politische Rahmungen: Unseres Erachtens besteht eine Diskrepanz zwischen einerseits vorhandenen Strukturen, die auf disziplinären Grenzen aufbauen (repräsentiert z.B. durch die Fachgesellschaften als „natürliche" Ansprechpartner) und hier weitgehend unhinterfragt reproduziert werden, und andererseits komplexen Forschungsfeldern, multiperspektivischer sowie inter- / transdisziplinärer Bearbeitung von Forschungsfragen und daraus entstehenden interdisziplinär orientierten Forschungsverbünden oder Fächern (z.B. Game Studies oder Regionalstudienfächer). Hieraus resultiert ein fast schon anachronistischer Zuordnungsdruck, gerade für die „kleinen Fächer", die durch eine Heterogenität an Methoden und Forschungsgegenständen gekennzeichnet sind. Es stellt sich die Frage, inwieweit das Konzept der Data Communities geeignet ist, um der Vielfalt der Forschungsrealität gerecht zu werden und das Forschungsdatenmanagement auf strategisch-operativer Ebene voranzubringen.

FDM-Praxisschwerpunkt Archivierung von Forschungsdaten: Zumeist werden – nicht zuletzt vor dem Hintergrund struktureller Bedingungen – drei Ansätze zur Archivierung bzw. zur Schaffung von Repositorien diskutiert und umgesetzt: generisch, institutionell und fachspezifisch. Auch hier stellt sich die Frage, ob nicht weitere Ansätze verfolgt werden sollten, die sich z.B. an Data Communities orientieren könnten. Könnten etwa Formen verteilter Archivierung von Forschungsdaten weiterführen? Für die Kuration und die Qualitätssicherung bei der Datenarchivierung ist ein umfassendes Verständnis für das jeweilige Material und den Entstehungskontext wesentlich. Verschiedene Data Communities formulieren unterschiedliche Bedingungen und

Anforderungen an Archivierung und Nachnutzung. Die ursprünglich datenproduzierende und datenarchivierende Data Community kann sich dabei durchaus auch von nachnutzenden Data Communities unterscheiden[4].

3 Ablauf

Der Workshop beginnt im ersten Teil mit einem Input-Vortrag von Danielle Cooper und einem Kommentar von Erzsébet Tóth-Czifra (in Englisch) zum Konzept der Data Communities, mit Raum für Fragen und Diskussion im Anschluss. Im zweiten Teil diskutieren verschiedene Expert:innen, wie Forschungsdatenmanagement für eine Data Community in Bezug auf die drei genannten Schwerpunkte praktisch aussehen kann (in Deutsch). In Kleingruppenarbeit sollen im Anschluss die Chancen und Herausforderungen, die sich aus der Praxis ergeben, herausgearbeitet und für einen partizipativen Debattenbeitrag aufbereitet werden.

4 Beitragende

- Danielle Cooper (New York), Managerin für Kollaborationen und Forschung bei Ithaka S+R.

- Erzsébet Tóth-Czifra (Berlin), Open Science Officer bei DARIAH-EU, verantwortlich für die Konzeption und Stärkung der Dissemination von Forschungsergebnissen in den Geisteswissenschaften.

- Maria Effinger (Heidelberg), Leiterin der Abteilung “Publikationsdienste” der Universitätsbibliothek Heidelberg, Co-Sprecherin NFDI4Culture, Teilprojektleiterin eines INF-Projektes am SFB933.

- Dirk Goldhahn (Leipzig), Wissenschaftlicher Mitarbeiter Digital Humanities an der Sächsischen Akademie der Wissenschaften zu Leipzig.

[4] siehe z.B. Steiner, Zumbühl und Bauder 2008.

- Christoph Schmitt (Rostock), Leiter der Wossidlo-Forschungsstelle für Europäische Ethnologie/Volkskunde an der Universität Rostock und des digitalen Archivs WossiDiA.

- Samuel Mund (Hildesheim und Hannover), Wissenschaftlicher Mitarbeiter Sammlungsmanagement und Erschließung am Center for World Music und zuständig für Digital Humanities, EDV und Technik am Europäischen Zentrum für Jüdische Musik.

Bibliografie

Cooper, Danielle, and Rebecca Springer. 2019. “Data Communities: A New Model for Supporting STEM Data Sharing.” *Ithaka S+R* (May 2019). https://doi.org/10.18665/sr.311396.

Steiner, Daniel, Heinz J. Zumbühl, and Andreas Bauder. “Two Alpine Glaciers over the Past Two Centuries: A Scientific View Based on Pictorial Sources.” In *Darkening Peaks: Glacier Retreat, Science, and Society*, edited by Ben Orlove, Ellen Wiegandt, and Brian H. Luckman, 83–99. Berkeley: University of California Press, 2008.

Workshop: Wissenschaftsgeleitetes Forschungsdatenmanagement in der NFDI

Generische und spezifische Perspektiven am Beispiel geisteswissenschaftlicher Querschnittsthemen

Kammerer, Dietmar
dietmar.kammerer[at]staff.uni-marburg.de
Phillips Universität Marburg, Deutschland

Schrade, Torsten
torsten.schrade[at]adwmainz.de
Akademie der Wissenschaften und der Literatur Mainz, Deutschland

Zusammenfassung. Forschungsdaten in den Geistes- und Kulturwissenschaften erstrecken sich über ein großes Spektrum an Fachdisziplinen und umfassen sehr unterschiedliche Datentypen. In der Nationalen Forschungsdateninfrastruktur (NFDI) geht es darum, die Datenbestände von Wissenschaft und Forschung für das gesamte deutsche Wissenschaftssystem systematisch zu erschließen, zu vernetzen und nachhaltig nutzbar zu machen. Ein übergreifendes Forschungsdatenmanagement im Sinne der FAIR- und CARE-Kriterien wird in der NFDI dann erfolgreich gelingen, wenn neben fachspezifischen Aspekten auch generische Ansätze und Lösungen im Rahmen von Querschnittsthemen erarbeitet werden. Im Workshop sollen in Zusammenarbeit zwischen Vertreter:innen aus den NFDI-Initiativen NFDI4Culture, NFDI4Objects, NFDI4Memory und Text+ und den Workshop-Teilnehmer:innen fachspezifische und generische Aspekte des Forschungsdatenmanagements in den Geistes- und Kulturwissenschaften anhand von konkreten Beispielen aus den Querschnittsthemen “Metadaten, Normdaten & Terminologien”, “Recht & Ethik”, “Provenienz” und “Data Literacy” herausgearbeitet und deren Bedeutung für die NFDI diskutiert werden.

1 Stand und Entwicklung

Zum Oktober 2020 hat mit der Gründung des NFDI e.V. die operative Phase der Nationalen Forschungsdateninfrastruktur (NFDI) begonnen. Ziel ist es, die Datenbestände von Wissenschaft und Forschung für das

gesamte deutsche Wissenschaftssystem systematisch zu erschließen, zu vernetzen und nachhaltig nutzbar zu machen. Erste Konsortien haben ihre Arbeit aufgenommen und konzeptionelle sowie strukturelle Schritte zu einer transdisziplinären Zusammenarbeit im Forschungsdatenmanagement sind eingeleitet. Gleichzeitig schreitet mit der zweiten und dritten Ausschreibungsrunde der NFDI die Entwicklung des fachspezifischen Ausbaus kontinuierlich voran. In den Geistes- und Kulturwissenschaften wird dieser hochdynamische Prozess bereits seit mehreren Jahren von einer Gruppe kooperierender NFDI-Initiativen eng begleitet, die sich mit dem Ziel der Entwicklung eines bestmöglich an spezifischen Bedarfen und Datentypen ihrer jeweiligen Communities ausgerichteten Forschungsdatenmanagements einerseits und der Erarbeitung übergreifender Perspektiven andererseits regelmäßig abstimmen und unterstützen.

Forschungsdaten in den Geistes- und Kulturwissenschaften erstrecken sich über ein breites Spektrum an Disziplinen und umfassen sehr unterschiedliche Datendomänen, die sich in Bezug auf Quantität, Abstraktionsgrad, Dynamik, Qualität und historische sowie methodische Entstehungskontexte deutlich unterscheiden.[1] Gleichzeitig sind forschungsgetriebene digitale Methoden und Verfahren bislang noch in unterschiedlichem Maß in den beteiligten Communities etabliert. Bei aller fachlichen Spezifik kann ein übergreifendes Forschungsdatenmanagement im Sinne der FAIR- und CARE-Kriterien in der NFDI nur dann erfolgreich gelingen, wenn im Rahmen integrativer Abstraktionsprozesse auch die generischen Aspekte gemeinsam herausgearbeitet und operationalisiert werden. Neben spezifischen methodischen und erkenntnistheoretischen Anforderungen und Praktiken, die in komplementären Handlungsfeldern und Maßnahmen zu bearbeiten sind, müssen also auch bedarfsorientierte Lösungen für fachübergreifende Querschnittsthemen entwickelt werden.

Welche Querschnittsbereiche sind aus geistes- und kulturwissenschaftlicher Perspektive besonders relevant? Welche konkreten Fallbeispiele verdeutlichen spezifische fachliche Anforderungen und Bedarfe, welche generischen Ansätze (z.B. im Bereich des Text- und Data-Minings) lassen sich zur Integration von fachlichen Lösungen finden? In welcher Perspektive stehen solche Ansätze mit Blick auf die aktuellen Entwicklungen in der NFDI? Diese

[1] Vgl. das Memorandum of Understanding der genannten Initiativen, Version 2, 28.09.2020, https://doi.org/10.5281/zenodo.4045000.

Fragen sollen im Rahmen eines Workshops entlang der vier Querschnittsthemen (1) „Metadaten, Normdaten, Terminologien“, (2) „Provenienz“, (3) „Rechtliche und datenethische Fragen “ und (4) „Data Literacy“ gemeinsam erörtert werden.

2 Geisteswissenschaftliche Querschnittsthemen

Die NFDI-Initiativen NFDI4Culture[2], NFDI4Objects[3], NFDI4Memory[4] und Text+[5] haben im Rahmen eines schon 2018 initiierten, wissenschaftsgeleiteten Prozesses mehrere Querschnittsthemen identifiziert, denen eine besondere Bedeutung in datengetriebenen Forschungsprozessen der Geistes- und Kulturwissenschaften zukommt. Die Anreicherung digitaler Forschungsgegenstände mittels standardisierter Metadaten, Normdaten und fachspezifischer Terminologien zur Verbesserung der Auffindbarkeit (Findability) und Interoperabilität (Interoperability) gehört ebenso dazu wie Fragen nach der Provenienz sowie rechtliche und ethische Gesichtspunkte mit Blick auf den wissenschaftlichen Zugang (Accessibility) und die Nachnutzbarkeit (Reusability) geistes- und kulturwissenschaftlicher Forschungsdaten. Auch der „Data Literacy“, also der Fähigkeit, planvoll und kompetent mit Daten umzugehen und sie im jeweiligen Forschungskontext bewusst einzusetzen sowie die jeweiligen wissenschaftlichen und gesellschaftlichen Zusammenhänge, in denen sie stehen, kritisch einschätzen und reflektieren zu können, kommt mit Blick auf alle vier FAIR-Kriterien eine wesentliche Bedeutung zu.

[2] Vgl. https://nfdi4culture.de/.
[3] Vgl. https://www.nfdi4objects.net/.
[4] Vgl. https://4memory.de/.
[5] Vgl. https://www.text-plus.org/.

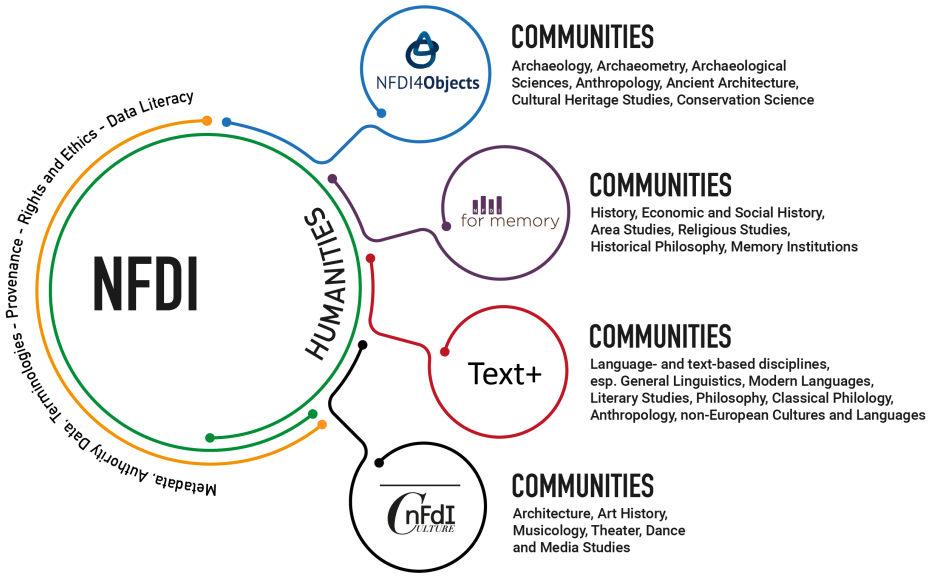


Abb. 1. Identifizierte Querschnittsthemen und Fachcommunities in den geisteswissenschaftlichen Initiativen NFDI4Culture, NFDI4Objects, NFDI4Memory und Text+

Daher bilden diese vier Querschnittsthemen den Kern der Zusammenarbeit der im Rahmen des „Memorandum of Understanding" kooperierenden geistes- und kulturwissenschaftlichen NFDI-Initiativen.

3 Perspektiven des Workshops

Im Workshop sollen anhand konkreter Beispiele aus den genannten Querschnittsbereichen generische und fachspezifische Aspekte des Forschungsdatenmanagements in den Geistes- und Kulturwissenschaften herausgearbeitet und zusammen mit den Workshopteilnehmer:innen perspektiviert werden. Vertreter:innen aus den vier genannten NFDI-Initiativen werden zu Beginn des Workshops im Zusammenspiel kurzer Impulsreferate einen Blick auf die fach- und communityspezifischen Aspekte jedes Querschnittsthemas werfen. Im gemeinsamen Dialog untereinander und in Form digitaler Interaktionsformate mit den Workshop-Teilnehmer:innen (z.B. Online-Whiteboards, Live-Polls etc.) werden die übergreifenden Perspektiven und Synergiepotentiale herausgearbeitet. Eine Panel-Diskussion der Vertreter:innen aus den NFDI-Initiativen unter Einbeziehung der Workshop-Teilnehmer:innen sichert die Ergebnisse und eröffnet Perspektiven für die vertiefte Zusammenarbeit und Weiterentwicklung

des geistes- und kulturwissenschaftlichen Forschungsdatenmanagements in der NFDI. Für die Moderation des Workshops ist das Geisteswissenschaftliche Forum NFDI des DHd Verbands[6] angefragt.

Bibliografie

Memorandum of Understanding by NFDI Initiatives from the Humanities and Cultural Studies, Version 2, 28.09.2020, https://doi.org/10.5281/zenodo.4045000.

Leipzig-Berlin-Erklärung zu NFDI-Querschnittsthemen der Infrastrukturentwicklung, Version 1, 15. Juni 2020, https://doi.org/10.5281/zenodo.3895209.

NFDI Cross-cutting Topics Workshop Report, Version 1, 11.03.2021, https://doi.org/10.5281/zenodo.4593770.

[6] Das Forum geisteswissenschaftliche NFDI ist eine Plattform zum Austausch und zur Koordination zwischen den geisteswissenschaftlichen Fachverbänden im Kontext der Etablierung einer die Anforderungen der Forschenden in den Geisteswissenschaften erfüllenden Komponente in der Nationalen Forschungsdaten-Infrastruktur (NFDI). Siehe: https://nfdi.hypotheses.org/.

VORTRÄGE

(Korpus-)linguistische user stories und das UV2 Annotationstool

Weingart, Anja
anja.weingart[at]uni-konstanz.de
Universität Konstanz, Deutschland

Zusammenfassung. Im Vortrag werden Anforderungen an (korpus-)linguistische Forschungssoftware in Form von *user stories* formuliert, die sich aus einer Beschreibung der Datenpraktiken innerhalb der DFG Projekte *'Uncovering verb-second effects. An interface-based typology* (UV2)' und *'The structure of wh-utterances and the interpretation of wh-words in Romance (and Germanic) languages*' ergeben. Der Begriff Datenpraktiken umfasst im weitesten Sinne den Umgang mit (Sprach)daten von der Erhebung über die Bearbeitung bis zur Analyse. Es soll gezeigt werden, dass das Verfassen von *user stories* eine kleinteilige Beschreibung der entsprechenden Workflows und insbesondere auch eine Reflexion über die Datenpraktiken und Workflows selbst ermöglicht und die Erwartungen an ein Softwaresystem leitet. Abschließend wird die freie Webapplikation *UV2 Annotationstool* (EUPL Lizenz) kurz vorgestellt, die eine generische Implementierung dieser user stories anstrebt.

1 Einleitung

User stories sind ein etabliertes Konzept zur Beschreibung von Anforderungen an ein Softwaresystem.[1] Eine user story ist aus Sicht des Nutzers, in unserem Falle aus der Perspektive des/der Linguist*in formuliert und besteht im besten Fall aus einem Satz. Anhand der user stories werden Akzeptanzkriterien für eine Implementierung festgelegt. Das Verfassen von user stories ermöglicht eine kleinteilige Beschreibung der entsprechenden Workflows, die nicht nur die Verständigung zwischen den Domänenexpert*innen und den Softwareentwickler*innen erleichtert, sondern auch eine Reflexion über die Workflows selbst bzw. die Erwartungen an ein Softwaresystem leitet. Die funktionalen Anforderungen an (korpus-)linguistische Forschungssoftware werden anhand von user stories formuliert, die sich aus einer Beschreibung der Datenpraktiken innerhalb der DFG Projekte *'Uncovering verb-second effects. An interface-based typology* (UV2)' und *'The structure of wh-utterances and the interpretation of wh-words in Romance (and Germanic) languages*'

[1] vgl. Wirdemann 2011, Cohn 2004.

ergeben. Der Begriff Datenpraktiken umfasst im weitesten Sinne den Umgang mit Daten und umfasst die Erhebung, Bearbeitung und Analyse. In diesem Vortrag sollen Aspekte der Datenbearbeitung vorrangig besprochen werden.

2 (Korpus-)linguistische user stories

Ziel obengenannter Forschungsprojekte ist eine vergleichende syntaktische Analyse, wobei das UV2 Projekt grammatikalische Bedingungen für die Verbzweitstellung in den großen romanischen Sprachen und Varietäten untersucht und das Projekt zu den *W*-Fragen die Wortstellung in Interrogativsätzen, insbesondere in *Warum*-Fragesätzen, analysiert. Die Datenbasis ist ein Korpus von Paralleltexten (20 Asterix Comics und das Alte und Neue Testament (Gen-2Könige und Matthäus-Johannes)) sowie eine Sammlung von Fragesätzen aus '*The Complete Sherlock Holmes*' und Daten aus Produktions- und Perzeptionsexperimenten.[2] Im Rahmen der Projekte werden die Sprachdaten nach theoriespezifischen Kriterien annotiert, um grammatische Muster auffindbar zu machen. Das heißt, es entsteht kein vollständig annotiertes Korpus, sondern es werden nur diejenigen Phänomene annotiert, die für die Bearbeitung der Forschungsfragen relevant sind. Wie zum Beispiel in (1) für Spanisch (Sp), Deutsch (Dt), Französisch (Fr.) und europäisches und brasilianisches Portugiesisch (EP/BP) gezeigt, sind die Positionen und die morphosyntaktische Kategorie der Konstituenten im Satz annotiert.

(1)a. Sp. ¿[Qué]**WH-1** [ha]**Vfin-2** [occurrido]**Vpart-3**?"(AST01-6:3a2)
b. Dt. [Was]**WH-1** [ist]**Vfin-2** [euch]**S-3** [zugestoßen]**Vpart4**?
c. Fr. [Que]**WH-1** [vous]**S-cl** [est]**Vfin-2**-[il]**-cl** [arrivés]**Vpart3**?
d. EP/BP [O que é que]**WH-1** [aconteceu]**Vfin2** [com vocês]**PP**?

Allein aus dieser kurzen Beschreibung der Projekte lassen sich mindestens 20 user stories für die Bereiche Sprachdaten, Abfrage und Annotation ableiten. In (2) seien vier stories beispielhaft genannt.

(2)a. Als Nutzer möchte ich Sätze, Texte eingeben und speichern können.
b. Als Nutzer möchte ich Texte oder Sätze als Paralleltexte anzeigen lassen.
c. Als Nutzer möchte ich die Quelle des Satzes angeben können.
d. Als Nutzer möchte ich Annotationen selbst definieren können.

[2] vgl. Kaiser et al. 2019 und Schmid et al. to appear.

Der zentrale Aspekt des Annotationsworkflows ist, dass neue Erkenntnisse eine Veränderung der relevanten Annotationen und/oder eine Integration neuer Daten nötig machen. Das bedeutet, dass die Sprachdatensammlung und die Annotationen erweiterbar und veränderbar sein sollen. Das Problem der Unveränderbarkeit von Annotationen bei (automatisch) annotierten Webkorpora ist kürzlich auch in Pintzuk (2019) formuliert worden. Gerade die Anforderung an eine flexible Datenstruktur führt dazu, dass die Daten oft in tabellarischer Form gespeichert und mit Office-Programmen (Textverarbeitung oder Tabellenkalkulation) bearbeitet werden. Diese Programme haben den Vorteil, dass sie verfügbar und ohne Expertenwissen bedienbar sind. Die Forschenden können damit unabhängig von personellen oder finanziellen Ressourcen Daten selbst strukturieren und verwalten.[3] Diese Programme verleiten jedoch zu einer "visuell orientierten" Datenstrukturierung und Annotation. Auf solche Daten sind keine Suchanfragen möglich und sie sind nicht mehr automatisch in andere Formate überführbar und damit mit anderen Werkzeugen bearbeitbar. Für eine (korpus-)linguistische Software bedeutet dies, dass diese Datenpraktik insoweit berücksichtigt werden sollte, als dass ein Import/Export der Sprachdaten in diese Formate (xml, csv) ermöglicht wird.

3 Das UV2 Annotationstool

Bevor die Implementierung der user stories (2a) und (2b) vorgestellt wird, soll die Motivation für die Entwicklung des UV2 tools kurz dargelegt werden. Ziel ist es eine freie, webbasierte Anwendung zu schaffen, die Basisfunktionalitäten für das Verwalten, Annotieren und Veröffentlichen von textbasierten Sprachdaten bereitstellt und anders als bei bestehenden Annotationstools, wie zum Beispiel WebLicht[4] oder Webanno[5], wird kein eigenes Datenformat erzeugt, sondern, aus oben genannten Gründen, xml/csv verwendet. Ziel ist es insbesondere auch eine Softwarearchitektur und Projektstruktur zu schaffen, die eine möglichst leichte Anpassung an andere Anwendungsfälle erlaubt und somit für andere Projekte wiederverwendbar ist. Aus diesem Grund werden nur offene und freie Technologien verwendet sowie das Konzept der losen Kopplung anhand der Entwurfsmuster Boundary-Control-Entity,

[3] Zum häufigen Einsatz dieser Formate und Programme siehe die Studie der AG Romanistik von 2015.

[4] https://weblicht.sfs.uni-tuebingen.de/weblichtwiki/index.php/Main_Page.

[5] https://webanno.github.io/webanno/.

Data-Transfer-Object und RESTful web services umgesetzt.[6] Wie in Abb. 1 gezeigt, sind projektunabhängige Komponenten wie die Benutzerverwaltung und Bibliographieverwaltung in getrennten Datenbanken verfügbar, die unabhängig von der projektspezifischen Datenbank wiederverwendet werden können.

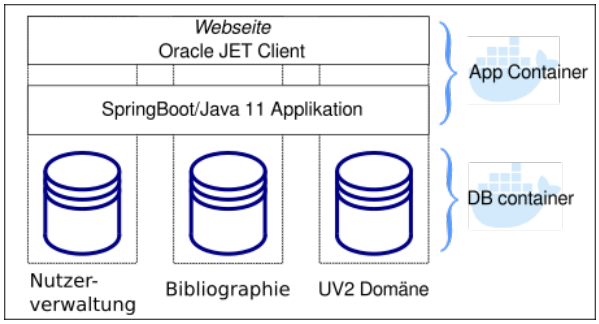


Abb. 1. Ausschitt der Datenstruktur des UV2 Annotationstools

Des Weiteren ermöglicht die Softwareprojektstruktur zusammen mit einer Containerisierung mit docker auf GitLab/CI eine getrennte Entwicklung des Oracle JET clients, der PostgreSQL Datenbanken (liquibase) und der Applikation, die aufgrund einer Teststrategie mit Junit und java testcontainers kontrollierbar bleibt.
Für das UV2 tool werden die user stories für die Bereiche Sprachdaten, Annotation, Abfrage und Import/Export implementiert. Für die Umsetzung der stories (2a) und (2b) zur Speicherung der Daten und der Anzeige als Paralleltext werden die Sprachdaten in der Datenbank (UV2 Domäne) als *snippet* gespeichert. Dieser Basiseinheit kann ein frei definierbarer Typ zugewiesen werden z.B. der Typ Text, Satz, Vers oder Wort. Jedem *snippet* ist eine Referenz mit einem Sprachattribut zugeordnet. Die Datenstruktur ist im entity-relationship Diagramm in Abb.2 illustriert.

[6] vgl. Bien 2012, Fielding 2000.

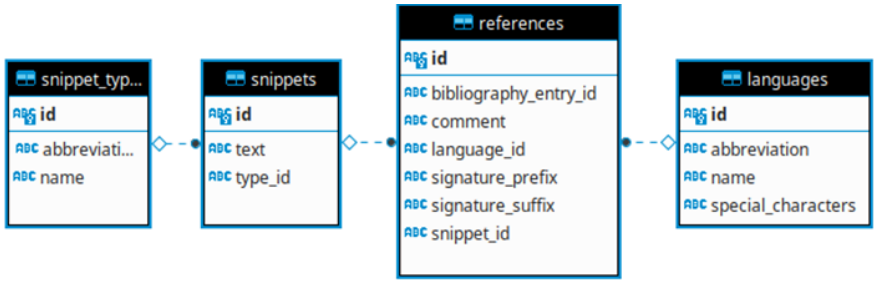


Abb. 2. Datenstruktur für *snippets*

Im Falle des UV2 Projekts setzt sich die Referenz eines *snippets* aus den bibliographischen Angaben (*bibliography_entry_id*), einem Sprachattribut und den Stellenangaben wie AST01-6:3a2 zusammen, die in der manuellen Transkription der Asterix Bücher verwendet wurden. AST01 wird als *signature_prefix* gespeichert und ist eine abstrakte Repräsentation des ersten Asterix Buchs. 6:3a2 bezeichnet die Stelle des *snippets* im Dokument und wird als *signature_suffix* gespeichert. Beide Attribute ergeben eine Signatur, die für alle Übersetzungen des *snippets* gleich ist. Anhand von url queries auf diese Attribute der Referenz lassen sich Paralleltexte anzeigen. Dies soll abschließend vorgeführt werden.

Bibliografie

Bien, Adam. *Real World Java EE Patterns*. press.adam-bien.com. 2012.

Cohn, Mike. *User Stories Applied*. Addison-Wesley, 2004.

AG Digitale Romanistik: “Ergebnisse der Umfrage der AG Digitale Romanistik zur Langzeitarchivierung von digitalen Forschungsdaten für die Romanistik”. Mitteilungsheft DRV 2015 no.1 (2015). https://www.deutscher-romanistenverband.de//wp-content/uploads/2017/12/Auswertung_Forschungsdaten-Umfrage.pdf

Fielding, Roy Thomas. *Architectural Styles and the Design of Network-based Software Architectures*. PhD Diss. University of California. 2000.
https://www.ics.uci.edu/~fielding/pubs/dissertation/top.htm

Kaiser, G. A., K. von Heusinger, and S. Schmid. “Word order variation in Spanish and Italian interrogatives. The role of the subject in ‘why’-interrogatives”. In Proceedings of the IX Nereus International Workshop: “Morphosyntactic and Semantic Aspects of the DP in Romance and Beyond”. edited by N. Pomino, 69–90. Fachbereich Linguistik, Universität Konstanz (Arbeitspapier 131), 2019.

Osmani, Addy. *Learning JavaScript Design Patterns*. Sebastopol: O'Reilly Media. 2012.

Pintzuk, Susan. “Adding linguistic information to parsed corpora”. Linguistic Issues in Language Technology 18, no. 4 (2019): 1-34. https://www.aclweb.org/anthology/2019.lilt-18.5.

Schmid, S., K. von Heusinger, and G. A. Kaiser. “On word order variation and information structure in Spanish and Italian ‘why’-interrogatives.” Cadernos de Estudos Linguísticos (to appear).

Wirdemann, Ralf. *Scrum mit user stories*. München: Hanser, 2015.

A lot of data –

Discover, ein Rechercheportal für granulare Forschungsdaten

Kümmet, Sonja
Sonja.Kuemmet[at]ub.uni-muenchen.de
Universitätsbibliothek der LMU München, Deutschland

Zusammenfassung. Gegenstand des Vortrags ist die Vorstellung eines im Rahmen des Modellprojekts „eHumanities – interdisziplinär" entwickelten lokalen Rechercheportals für (granulare) Forschungsdaten – Discover. Anhand eines Pilotprojekts aus den digitalen Geisteswissenschaften, VerbaAlpina, werden Anforderungen für die Erschließung und Präsentation granularer Daten ermittelt und deren (technische) Umsetzung skizziert. Den Hauptteil des Vortrags bildet eine Systemvorführung mit Fokus auf Funktionen, die den Nutzer*innen den Umgang mit der aus der Granularität (und Versionierung) resultierenden beträchtlichen Datenmenge erleichtern sollen. Hierzu gehören nicht zuletzt spezifische Facetten zur Eingrenzung der Suchmenge, z.B. nach Beschreibungsebene, sowie Sprungmarken, um innerhalb der Hierarchie, d.h. z.B. von einem übergeordneten zu einem untergeordneten Datensatz, navigieren zu können. An die Systemvorführung schließt sich eine kritische Auseinandersetzung mit der implementierten Lösung an. Diese betrifft v.a. den nicht unerheblichen Ressourcenaufwand, der nur bedingt skaliert werden kann: Während sich der Ingest-Workflow (zumindest in Teilen) auf neue Projekte übertragen lässt, sind die Analyse und das Mapping der Ausgangsdaten projektspezifisch und müssen somit immer aufs Neue durchgeführt werden. Unabhängig von der Ressourcenfrage stellt sich für die Portalbetreiber auch die Frage nach den Bedarfen in den Geisteswissenschaften: Für welche Projekte ist eine granulare Erschließung der Forschungsdaten sinnvoll und für welche ist eine Erschließung und Publikation auf Projektebene ggf. ausreichend?

Ziel des Vortrags ist es, eine lokale Lösung für den Umgang mit granularen Forschungsdaten vor- und zur Diskussion zu stellen.
Ausgehend von einem Problemaufriss werden zunächst konkrete Anforderungen an ein Rechercheportal für granulare Forschungsdaten abgeleitet. In einem zweiten Schritt wird die Umsetzung der Anforderungen, d.h. die technische Entwicklung, skizziert. Die Herausforderungen, die hierbei zu bewältigen waren, leiten über zur Diskussion, zum einen der vorgestellten Lösung im Speziellen und zum anderen des Umgangs mit Granularität (in den Geisteswissenschaften) im Allgemeinen.

1 Granularer Nachweis von Forschungsdaten

Das vorzustellende Rechercheportal Discover[1] wurde in enger Kooperation mit dem sprachwissenschaftlichen DFG-Projekt VerbaAlpina[2] im Rahmen des bayerischen Modellprojekts „eHumanities – interdisziplinär“[3] entwickelt. Um die Auffindbarkeit und die Referenzierbarkeit zu verbessern, wurde entschieden, nicht nur das Projekt als Ganzes mit Metadaten zu beschreiben, sondern auch die darin enthaltenen Datensätze. Daraus resultieren 3 Beschreibungsebenen: das Projekt als Ganzes (1 Forschungsdatum), die Einzeldatensätze (130.000 Forschungsdaten) und 3 verschiedene Aggregationstypen von Einzeldatensätzen (11.000 Forschungsdaten). An die Stelle eines Forschungsdatums traten somit über 140.000 Forschungsdaten – eine Herausforderung in vielerlei Hinsicht.
Als konkrete Anforderungen für das zu entwickelnde Rechercheportal ergaben sich hieraus:

- Eindeutige Referenzierbarkeit und Persistenz aller Forschungsdatenobjekte, unabhängig von ihrer Beschreibungsebene
- Abbildung der hierarchischen Struktur der Forschungsdaten in den Metadaten, um zwischen den Beschreibungsebenen navigieren zu können
- Intuitive Bedienbarkeit
- Versionierung der Forschungsdaten
- Bereitstellung der Metadaten per OAI-PMH-Schnittstelle

2 (Technische) Umsetzung und Systemvorführung

Der erste Umsetzungsschritt war die Festlegung eines geeigneten Datenmodells. In Abwägung der Vorteile der Nachnutzung eines Standards einerseits und projektspezifischer Erfordernisse (v.a. auch hinsichtlich der Granularität) andererseits wurde ein DataCite Anwendungsprofil, rdUB[4], entwickelt. rdUB erweitert DataCite beispielsweise um ein Element *Hierarchy*, das die Beschreibungsebene des jeweiligen Forschungsdatums angibt.

[1] https://discover.ub.uni-muenchen.de/.
[2] Krefeld, Thomas | Lücke, Stephan (Hrsgg.) (2014–): VerbaAlpina. Der alpine Kulturraum im Spiegel seiner Mehrsprachigkeit, München, online, https://dx.doi.org/10.5282/verba-alpina.
[3] https://www.fdm-bayern.org/ehumanities-interdisziplinaer/.
[4] https://github.com/UB-LMU/rdUB.

Ausgehend von diesem Datenmodell wurde ein komplexer (Ingest-)Workflow implementiert, der die Ausgangsdaten, d.h. die Datensätze des sprachwissenschaftlichen Pilotprojekts, nach rdUB transformiert und hierbei noch fehlende Relationsinformationen (isPartOf, isPreviousVersionOf, isNewVersionOf) ergänzt. Grundsätzlich wurde eine modulare, aus den Komponenten Fedora, Solr und Project Blacklight bestehende Systemarchitektur gewählt, um flexibel auf neuartige Use Cases reagieren zu können.
Aufgrund der nicht zuletzt auch durch die Versionierung beträchtlichen Datenmenge setzt das Projektteam hinsichtlich der Referenzierbarkeit auf ein zweistufiges Verfahren: Alle Forschungsdaten erhalten, unabhängig von ihrer Beschreibungsebene, einen globalen institutionsspezifischen Identifikator. Zusätzlich besteht die Möglichkeit der Vergabe eines DataCite DOI, wobei sich die Modalitäten hier nach der hierarchischen Verortung des Forschungsdatums richten: Für Forschungsdaten auf Projektebene wird standardmäßig ein DOI registriert; für Forschungsdaten unterhalb der Projektebene können bei Bedarf DataCite DOIs per Webformular (s. Abb1) beantragt werden.

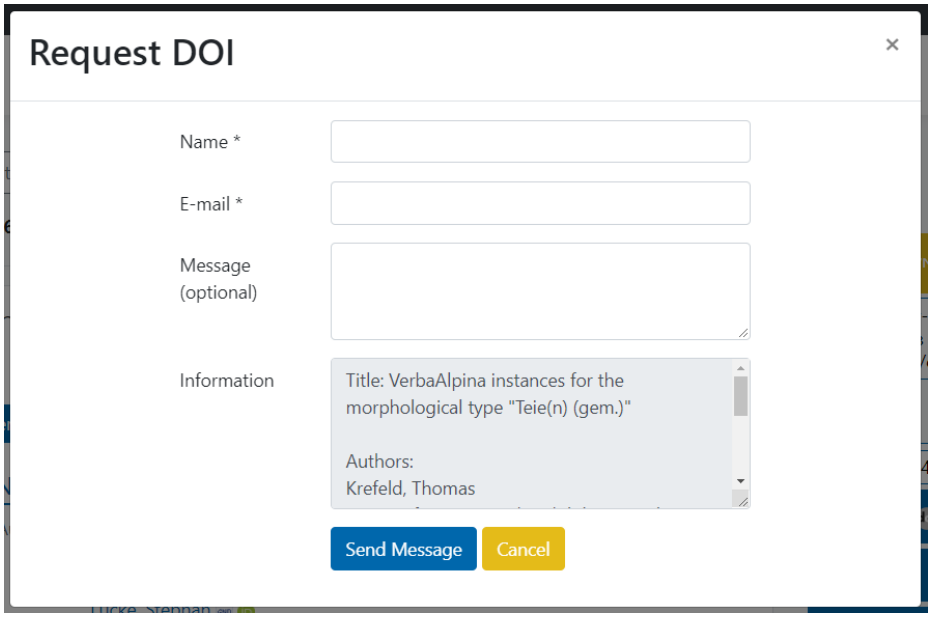


Abb. 1. Request DOI Webformular.

Das Rechercheportal bietet seinen Benutzer*innen drei verschiedene **Sucheinstiege**:

- **Suchfeld**; Indexiert sind alle für die Auffindbarkeit eines Forschungsdatums relevanten Informationen: Schlagwörter, DDC-Notationen, Orte, Personen, Institutionen, PIDs (z.B. Wikidata, GeoNames, GND, …), Titel und Beschreibungstexte. Darüber hinaus können Forschungsdaten auch direkt durch Eingabe eines Identifikators (institutionsspezifischer Identifikator, ggf. DataCite DOI, projektspezifischer Identifikator) recherchiert werden.
- **Facetten** für einen Schnelleinstieg (s. Abb2);

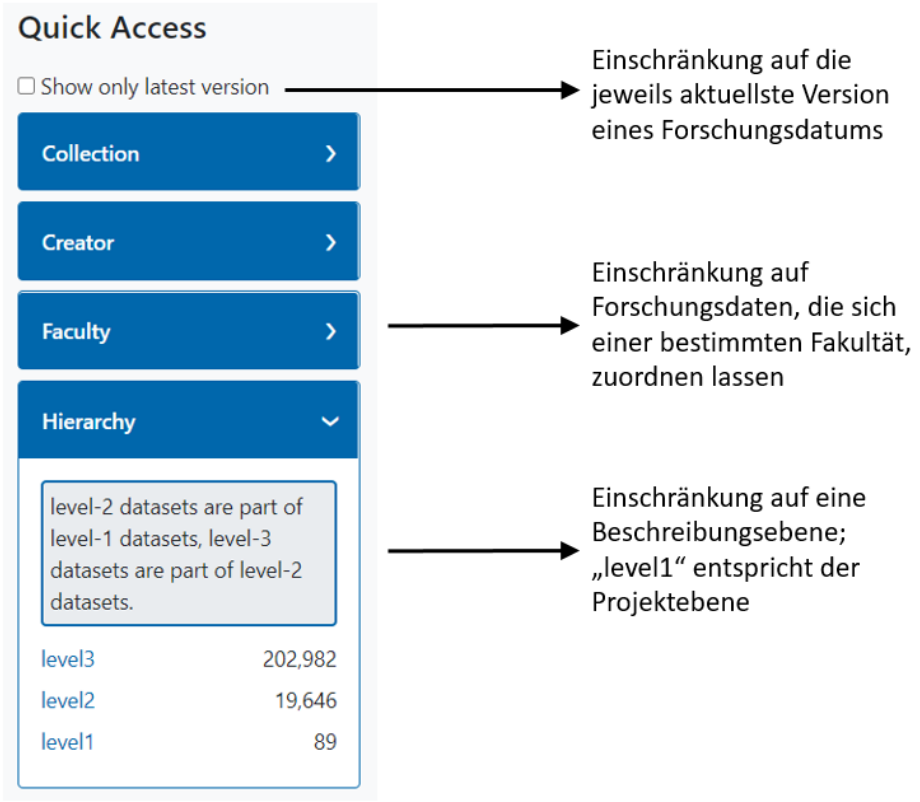


Abb. 2. Schnelleinstieg.

- **Collections**; Die Forschungsdaten werden in Collections organisiert, wobei ein granular erschlossenes Forschungsprojekt als eine Collection gewertet wird.

Auf der **Trefferlistenseite** stehen, abhängig von der Vollständigkeit der Metadaten, 8 weitere Facetten zur Eingrenzung der Suche zur Verfügung: Contributor, DDC (Dewey Decimal Classification), (File) Format, Keyword, Language, Location, Rights und Year of Publication.

Buttons auf der **Einzeltrefferanzeige** (s. Abb3) ermöglichen den Nutzer*innen nicht nur, zwischen den Versionen eines Forschungsdatums hin- und her zu wechseln, sondern auch zwischen den Beschreibungsebenen zu navigieren. So startet ein Klick auf „Start Search (Is Part Of)“ eine Suche nach allen Forschungsdatensätzen, zu denen der betreffende Datensatz gehört. Umgekehrt liefert ein Klick auf „Start Search (HasPart)“ alle Datensätze, aus denen sich der Ausgangsdatensatz zusammensetzt.
Erwähnenswert sind auch die Icons hinter Personennamen, Keywords und Locations. Diese zeigen an, dass die betreffende Information mit einem Eintrag in einem Normdatenvokabular, wie z.B. ORCID, GND, GeoNames und Wikidata, verknüpft ist. Für Nutzer*innen bietet dies den Vorteil, dass sie zum einen, durch Klick auf das Icon den jeweiligen Normdateneintrag einsehen können, und zum anderen, dass sie durch Anwählen des Namens eine Suche nach der individualisierten Entität (z.B. die Person „Krefeld, Thomas“, anstelle des Namens „Krefeld, Thomas“) anstoßen können.

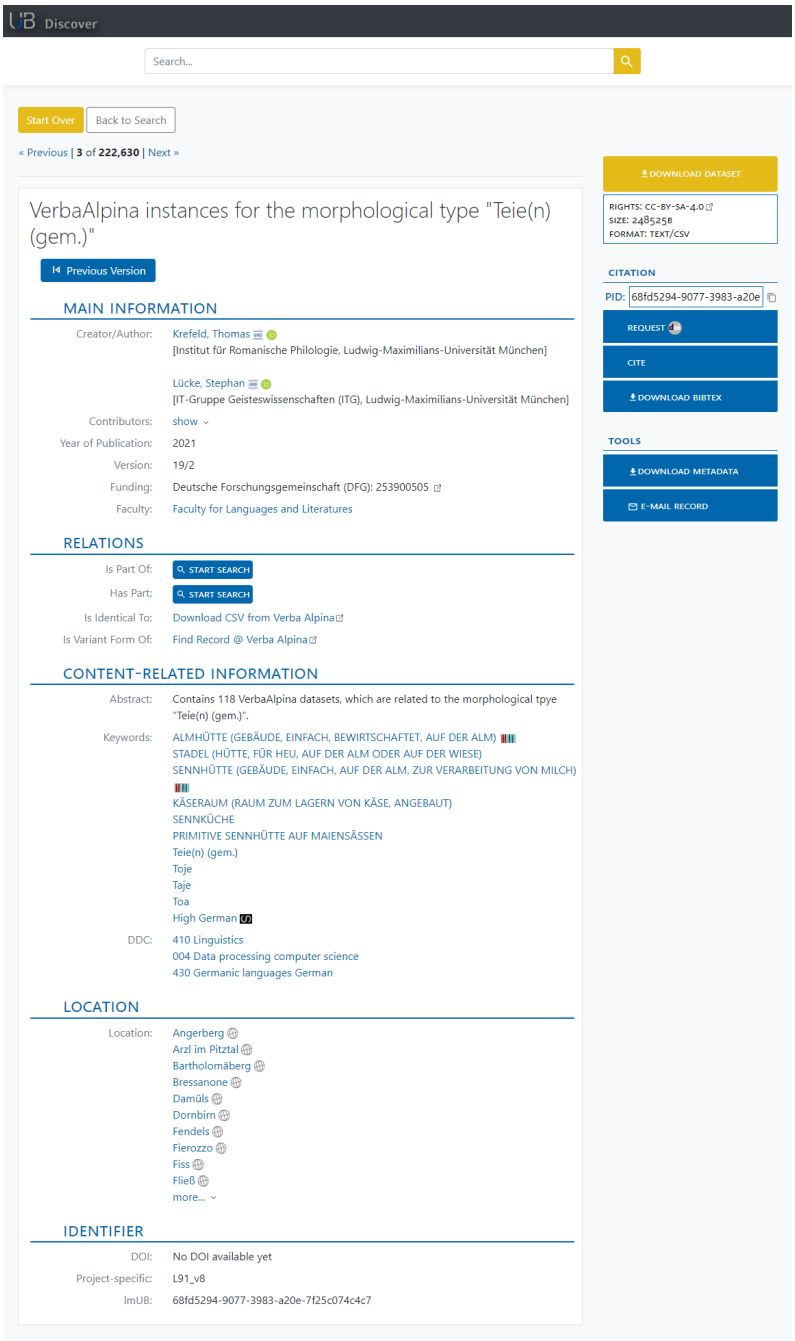


Abb. 3. Einzeltrefferanzeige.

3 Granularer Nachweis von Forschungsdaten

Das Pilotprojekt steht mit seiner feingranularen Erschließung und halbjährlichen Versionierung für ein Ende eines Spektrums, dessen anderes Ende die klassische Erschließung auf Projektebene (d.h. ein Metadatensatz pro Projekt) bildet. Die 2-jährige Arbeit an diesem Projekt konnte auf der einen Seite aufzeigen, was prinzipiell an (Tiefen-)Erschließung möglich ist: Jeder Datensatz innerhalb des Projekts wurde mit einem umfangreichen Set an Metadaten ausgezeichnet und damit separat auffindbar und referenzierbar gemacht. Auf der anderen Seite hat das Modellprojekt aber auch verdeutlicht, wie aufwändig die Umsetzung ist. Zwar kann der einmal erarbeitete Ingest-Workflow (modifiziert) auf Folgeprojekte übertragen werden aber das Mapping der Ausgangsdaten in das interne Format rdUB muss für jedes Projekt neu durchgeführt werden. Diese **Ressourcenintensität** hat die Betreiber von Discover dazu veranlasst, ihr Service-Angebot zu staffeln: Standardmäßig werden Forschungsdaten nur auf Projektebene erschlossen. Sollen auch Datensätze innerhalb des Projekts mit Metadaten versehen und damit auffindbar und referenzierbar gemacht werden und braucht es hierfür konzeptionelle und v.a. technische Unterstützung, kann diese vorerst nur im Rahmen einer Projektbeteiligung angeboten werden.
Das Pilotprojekt hat aber nicht nur wegen des hohen Umsetzungsaufwands den Charakter eines Testballons: Die Nachfrage nach einer granularen Erschließung von Forschungsdaten hält sich, zumindest bei den Plattformbetreibern, (noch) in Grenzen. Gut möglich, dass sich diese Zurückhaltung als „Henne-Ei-Problem" interpretieren lässt. Grundsätzlich fehlt es, nach Ansicht der Vortragenden, an einem Diskurs darüber, wie granular Forschungsdaten in den Geisteswissenschaften publiziert werden sollen, für welche Arten von Projekten eine Tiefenerschließung sinnvoll und für welche wiederum eine Beschreibung des Forschungsprojekts als Ganzes ausreichend ist. Der eingereichte Vortrag versteht sich als Impulsbeitrag zu dieser notwendigen Diskussion.

Aufbereitung und Publikation bibliographischer Forschungsdaten

Omeka S als Werkzeug zur (Meta-)Datenkuration und Wikidata als Plattform zur Forschungsdatendistribution und -publikation

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

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

Zusammenfassung. Im Vortrag wird eine generische Vorgehensweise zur Aufbereitung gedruckter Bibliographien in bibliographische Daten zur Publikation und Distribution in Wikidata vorgestellt. Die bibliographischen Angaben in gedruckter Form werden dazu mit maßgeschneiderten Parsern extrahiert. Bei der weiteren Aufbereitung erfolgt eine Normalisierung der bibliographischen Entitäten und der extrahierten und identifizierten Erscheinungsorte, Herausgeber, etc. Die bibliographischen Daten werden zu diesem Zweck als tabellarische CSV-Daten in Omeka S als Werkzeug zur Metadatenkuration und Datenmodellierung importiert. Die Modellierung erfolgt anhand eines mittels der *Resource Templates* von Omeka S zusammengestellten Anwendungsprofils mit FRBR-Elementen. Vor dem Importieren der aufbereiteten und FRBRisierten bibliographischen Daten in den Wikidata-Datenbestand können diese bereits vorab mit dem Wikidata-Modul für Omeka S – oder ggf. auch schon zuvor mit OpenRefine – mit Wikidata-Einträgen zu den in der Bibliographie identifizierbaren Entitäten angereichert werden. Die in Wikidata integrierten und kontextualisierten bibliographischen Daten können als Forschungsdaten zur Analyse und Visualisierung nachgenutzt, aber auch weiter angereichert und kontextualisiert werden, was den wesentlichen Mehrwert von Wikidata als Distributions- und Publikationsplattform für die im Projekt aufbereiteten und angereicherten bibliographischen Daten darstellt.

1 Einleitung

Hintergrund des Vortrags ist ein neues Digitalisierungsprojekt zu deutschsprachigen Zeitungen aus dem östlichen Europa. Das Projekt zielt einerseits auf die Digitalisierung historischer Zeitungen, andererseits auf eine umfangreiche Erfassung bibliographischer Daten zu sämtlichen nachweisbaren Pressepublikationen (ca. 6.000 Titel)

deutscher Bevölkerungsgruppen in der Region aus dem Zeitraum 1727 bis zur Gegenwart.
Grundlage für die Datenerfassung ist eine bereits im volltextdurchsuchbaren PDF-Format vorliegende, 1.348-seitige Bibliographie. PDF-Dokumente gewährleisten zwar durch ihre buchsimulierende Funktionalität rasche Übersicht und selbsterklärende Nutzbarkeit, können jedoch – wie ihre gedruckten Pendants – den Informationsbedarf Forschender nicht mehr vollauf bedienen: Die solitäre Zusammenstellung von Quellen in einer Fachbibliographie genügt nicht mehr, um das Material adäquat zu ordnen und auch Zugänge nachzuweisen.
Von zentraler Bedeutung für bibliographische Hilfsmittel sind vermehrt direkte Zugriffsmöglichkeiten auf das gesuchte Material,[1] darüber hinaus auch auf Kontextualisierungsmöglichkeiten mit vergleichbaren Referenzen, Personen oder Orten[2]. Daraus resultiert die Forderung nach einer flexibleren Integration und Vernetzung von bibliographischen Ordnungssystemen und von Repositorien, so dass einerseits Parallelrecherchen unnötig und andererseits Wissenskontexte sichtbar werden.[3]

2 Vorgehensweise

Um diesen Forderungen entsprechen zu können, wird die Periodika-Bibliographie in einem ersten Schritt in ein maschinenlesbares Tabellenformat (CSV) konvertiert. Dazu bauen wir individuell auf die Struktur und die Formatierung der bibliographischen Angaben im PDF-Dokument zugeschnittene Parser.[4]
Zur weiteren Aufbereitung und Anreicherung werden die CSV-Daten in Omeka S[5] importiert. Im Vortrag zeigen wir, wie Omeka S nicht nur als als Werkzeug zur Metadatenkuration, sondern auch zur Modellierung der Daten selbst verwendet werden kann – und zwar komplexer bibliographischer Daten inkl. ungenauer und unsicherer Angaben (z.B. zu Erscheinungszeiträumen und -orten), wie sie typisch beim Umgang mit historischen Quellen sind. Omeka S ermöglicht mit den *Resource Templates* quasi Anwendungsprofile[6] mit Elementen aus importierten

[1] "[…] providing 'discovery is not enough'; scholars want direct access to the materials they identify […]" (Kraus and Stöhr 2015, S. 170).
[2] Vgl. Buckland 2015, S. 36.
[3] Vgl. Hartmann & Thänert 2015, S. 35.
[4] Eine Alternative wäre das Trainieren von CRF-Modellen für Informationsextraktion, die sich bereits bei der semantischen Aufbereitung von bibliographischen Daten in gedruckter Form bewährt haben (siehe Lüschow 2020).
[5] https://omeka.org/s/.
[6] Heery & Patel 2000.

Vokabularen wie etwa FaBIO (FRBR-aligned Bibliographic Ontology)[7] zu erstellen und diese u.a. mit Hilfe des Wikidata-Moduls[8] anzureichern. FRBR (Functional Requirements for Bibliographic Records) fungiert dabei als konzeptuelles Rahmenwerk[9] für die feingliedrige Modellierung bibliographischer Entitäten und deren Zusammenhänge untereinander. Im letzten Schritt werden die aufbereiteten, angereicherten und *FRBRisierten* Daten veröffentlicht. Um die Langzeitverfügbarkeit und Zitierbarkeit der Daten sicherzustellen, werden diese in unserer wissenschaftlichen Forschungsdateninfrastruktur LaMBDa und OstData[10] publiziert. Als Zielsystem mit der derzeit größten Reichweite zur Distribution der Daten bietet sich die freie Wissensbasis Wikidata an.[11] Ursprünglich für die Repräsentation der faktischen Informationen in der Online-Enzyklopädie Wikipedia entworfen[12], hat sich Wikidata zu einem universellen Linking-Hub für Wissensbasen entwickelt, der gerade auch für bibliographische Daten ein ideales Umfeld darstellt. So berücksichtigt das Wikidata-Datenmodell bei der Modellierung bibliographischer Entitäten – wenn auch eingeschränkt bzw. vereinfacht – FRBR-Konzepte.[13] Der Import der Daten wird durch ausgereifte Tools unterstützt.[14]

3 Implikationen von Wikidata als Lösung

Das Potential von Wikidata für die Publikation und Distribution der Periodika-Bibliographie wirkt sich auf folgenden Ebenen positiv aus:

1. Datenhaltung: Der Import der bibliographischen Daten in die offene Infrastruktur von Wikidata macht den Betrieb einer eigenen Datenbankinfrastruktur überflüssig. Darüber hinaus ist im Rahmen des ‚Wikimedia-Universums' die Nachhaltigkeit des Vorhabens durch langfristige Verfügbarkeit der Daten gewährleistet.
2. Kontextualisierung: Die große Stärke von Wikidata besteht darin, dass die Datenbank von bibliographischen Daten über

[7] Peroni & Shotton 2012.
[8] https://omeka.org/s/modules/Wikidata/.
[9] Vgl. Tillett 2005.
[10] https://gepris.dfg.de/gepris/projekt/413708228.
[11] https://www.wikidata.org/.
[12] Vrandečić & Krötzsch 2014.
[13] Siehe: https://www.wikidata.org/wiki/Wikidata:WikiProject_Books/de#Bibliographische_Eigenschaften (1. Mai 2021).
[14] Mithilfe des Tools QuickStatements lassen sich Daten per Stapelverarbeitung nahtlos in Wikidata laden (https://www.wikidata.org/wiki/Help:QuickStatements (13. Jan. 2021).

Personendaten bis hin zu abstrakten Konzepten global und multilingual enzyklopädisches Wissen repräsentiert. Daher lassen sich die Angaben zu den Periodika mit Informationen zu Personen (Herausgebern, Redakteuren etc.), Erscheinungsorten, historischen Ereignissen, Quellen u.a. verknüpfen.

3. Informationsaufbereitung und Nachweis: Wikidata ist weder eine geschlossene noch eine abgeschlossene Ressource. Datensätze können ausgebaut, korrigiert und weiterführend verlinkt werden. Dadurch lassen sich die in der Bibliographie angelegten Informationen nicht nur als 'living document' auf neuestem Stand halten, sondern sie können in Wikidata die Funktion eines wachsenden zentralen Nachweisinstruments übernehmen, indem Zugangslinks zu Digitalisaten unterschiedlicher Institutionen erfasst werden. Dieser Funktion kommt entgegen, dass Datensätze in Wikidata Normdaten und Identifikatoren (z.B. GND, ZDB, LOCSH etc.) integrieren, die wiederum auf weiterführende Ressourcen in Bibliotheken und anderen Kulturerbeeinrichtungen verweisen.[15]

4. Open Science / Citizen Science: Ebenso wie Wikipedia basiert auch Wikidata auf der Beteiligung interessierter Forscherinnen und Forscher bzw. Bürgerinnen und Bürger. Dadurch eröffnen sich für die im Projekt digitalisierten Ressourcen sowie die Metadaten aus der Bibliographie zahlreiche Chancen zur Steigerung ihrer Wahrnehmung, zur aktiven Nachnutzung und zur kollaborativen Wissenserzeugung. Darauf aufbauend ließen sich etwa Szenarien zur Transkription von Digitalisaten, zum Verfassen neuer Wikipedia-Artikel zu den Zeitungen, Zeitschriften und Volkskalendern oder auch zu Ereignissen oder Abbildungen, die in Einzelartikeln der digitalisierten Quellen wiedergegeben werden, realisieren[16].[17]

---

[15] Die Bedeutung von Wikidata für Kulturerbeeinrichtungen (GLAMs) lässt sich an zahlreichen Kooperationen zwischen Nationalbibliotheken u.a. und Wikimedia ablesen, z.B. das Projekt der Deutschen Nationalbibliothek "Authority Control meets Wikibase" (https://wiki.dnb.de/display/GND/Authority+Control+meets+Wikibase (13. Jan. 2021)). Vgl. überdies das White Paper zu Wikidata der Association of Research Libraries (ARL) (ARL Task Force on Wikimedia and Linked Open Data 2019).

[16] Bemme & Munke 2021.

[17] Mit großem Erfolg wird dies z.B. im Rahmen der Digitalisierung des Journals „Die Gartenlaube" von der SLUB Dresden praktiziert: https://diedatenlaube.github.io/ (13. Jan. 2021).

5. Wissensgenerierung: Wikidata hält nicht nur offene Schnittstellen zum Export oder zur Überprüfung der Datenqualität bereit.[18] Über einen SPARQL-Endpoint lassen sich darüber hinaus komplexe Anfragen an die Datenbank stellen. Mit der SPARQL-Abfragesprache werden Fragen mit verschiedenen Variablen möglich, wie z.B. nach Zeitungen, die Mitte des 19. Jahrhunderts in der Bukowina auf Deutsch erschienen und deren Herausgeber aus dem heutigen Österreich stammten. Der Wikidata SPARQL Query Service[19] unterstützt Nutzer bei der Formulierung von Anfragen und bildet die Basis für Informationsvisualisierungswerkzeuge: Suchergebnisse lassen sich auf Karten lokalisieren, zeitliche Abläufe in Zeitleisten abbilden und Quantitäten oder Relationen diagrammatisch darstellen, so dass implizite Zusammenhänge hervortreten und u.U. neue Wissenskontexte offensichtlich werden. Die in Wikidata kontextualisierten bibliographischen Daten können somit als Forschungsdaten insbesondere für quantitative Analysen und Visualisierungen zum Einsatz kommen.

4 Fazit und Ausblick

Aus Sicht des *Data Curation Continuum*[20] lässt sich der vorgestellte Arbeitsablauf wie folgt verorten: Omeka S dient als lokale Lösung zur automatischen und kollaborativen Aufbereitung der Daten innerhalb der *Projektdomäne* und Wikidata ist unsere globale Lösung als Distributionsplattform in der *Publikationsdomäne*. Wikidata ist darüber hinaus ein *Collaboration Data Store*[21] für die Kooperations- und Publikationsdomäne zur Bearbeitung der veröffentlichten Daten als ‚Living Data' mit der Möglichkeit, daraus nach substanzieller Änderung oder Erweiterung eine neue Version des Datensatzes im institutionellen Forschungs-datenrepositorium LaMBDa und dem Forschungsdatendienst OstData zu publizieren.
Als generische Lösung eignet sich die Vorgehensweise zur Aufbereitung von Bibliographien aus verschiedenen Fachbereichen, wobei die kollaborative Sammlung, Aufbereitung und Anreicherung innerhalb des Projektteams auch gleich in Omeka S begonnen werden kann oder die initialen Daten z.B. aus dem Literaturverwaltungsprogramm Zotero

[18] Vgl. Voß 2016.
[19] https://query.wikidata.org/.
[20] Treloar & Klump 2019.
[21] Siehe Treloar & Klump 2019.

importiert werden können. Um projektspezifische Forschungsfragen zu beantworten, kann die Bibliographie anstatt in Wikidata als Forschungsdaten in eine Forschungsumgebung wie ResearchSpace[22] importiert werden.[23] Die Integration der Daten in Wikidata folgt jedoch der Annahme, dass Fachbibliographien in erster Linie in die vielfältigen Kontexte von Digital Humanities eingebunden werden müssen, um ihre Potentiale auszuschöpfen – als „[…] Scharnier innerhalb eines vernetzten wissenschaftlichen Arbeitsprozesses".[24]

Bibliographie

ARL Task Force on Wikimedia and Linked Open Data. 2019. "ARL White Paper on Wikidata: Opportunities and Recommendations." Association of Research Libraries. https://www.arl.org/resources/arl-whitepaper-on-wikidata/.

Bemme, Jens, und Martin Munke. 2021. "Open Citizen Science: Leitbild für kuratorische Praktiken in Wissenschaftlichen Bibliotheken." In *Bibliotheken Als Orte Kuratorischer Praxis*, edited by Klaus Ulrich Werner, 165–200. https://doi.org/10.1515/9783110673722-013.

Buckland, Michael. 2015. "Data Management as Bibliography." *Bulletin of the American Society for Information Science and Technology* 37 (6): 34–37. https://doi.org/10.11588/dco.2015.1.18641.

Hartmann, Andreas, und Sabine Thänert. 2015. "Vom Thesaurus zum semantischen Netz. Potenziale von Data Mining in bibliographischen Datensätzen." *Digital Classics Online* 1 (1): 34–45. https://doi.org/10.11588/dco.2015.1.18641.

Heery, Rachel, und Manjula Patel. 2000. "Application Profiles: Mixing and Matching Metadata Schemas." *Ariadne* 25. http://www.ariadne.ac.uk/issue/25/app-profiles/.

[22] Oldman & Tanase 2018.
[23] Siehe dazu z.B. Kräutli & Valleriani 2017.
[24] Schröter 2013, S. 110.

Kraus, Eva, und Matti Stöhr. 2015. “On the Way to a ‘German Historical Bibliography’: Current State and Perspectives – The DFG Project ‘Continued Cooperative Development of Historical Subject Bibliographies’.” In *Historical Bibliography as an Essential Source for Historiography*, edited by Kristina Rexová, Bernadette Cunningham, Václava Horčáková, und Věra Hanelová, 161–79. Cambridge Scholars Publishing.

Kräutli, Florian, und Matteo Valleriani. 2017. “CorpusTracer: A CIDOC Database for Tracing Knowledge Networks.” *Digital Scholarship in the Humanities* 33 (2): 336–46. https://doi.org/10.1093/llc/fqx047.

Lüschow, Andreas. 2020. “Automatische Extraktion und Semantische Modellierung der Einträge einer Bibliographie französischsprachiger Romane.” In *DHd 2020 Spielräume: Digital Humanities Zwischen Modellierung und Interpretation. Konferenzabstracts*, edited by Christof Schöch, 80–84. Verband Digital Humanities im deutschsprachigen Raum e.V. https://doi.org/10.5281/zenodo.3.

Oldman, Dominic, und Diana Tanase. 2018. “Reshaping the Knowledge Graph by Connecting Researchers, Data and Practices in ResearchSpace.” In *The Semantic Web – ISWC 2018*, edited by Denny Vrandečić, Kalina Bontcheva, Mari Carmen Suárez-Figueroa, Valentina Presutti, Irene Celino, Marta Sabou, Lucie-Aimée Kaffee, und Elena Simperl, 11137:325–40. Lecture Notes in Computer Science. Springer. https://doi.org/10.1007/978-3-030-00668-6_20.

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.

Schröter, Marcus. 2013. “Recherchieren und Kommunizieren: Die Fachbibliographie in der digitalen Zukunft – zu aktuellen Entwicklungen am Beispiel der Geschichtswissenschaft.” Freiburger Universitätsblätter, 202: 95–120.

Tillett, Barbara. 2005. “FRBR and Cataloging for the Future.” *Cataloging and Classification Quarterly* 39 (3): 197–205.

Treloar, Andrew, und Jens Klump. 2019. “Updating the Data Curation Continuum: Not Just Data, Still Focussed on Curation, More Domain-Oriented.” *International Journal of Digital Curation* 14 (1): 87–101. https://doi.org/10.2218/ijdc.v14i1.643.

Voß, Jakob. 2016. “Wikidata als Universalbibliographie: ein Kommentar.” *LIBREAS. Library Ideas* 29: 95–129. https://libreas.eu/ausgabe29/02voss.

Vrandečić, Denny, und Markus Krötzsch. 2014. “Wikidata: A Free Collaborative Knowledgebase.” *Communications of the ACM* 57 (10): 78–85. https://doi.org/10.1145/2629489.

Auszugsweise Veröffentlichung geschützter Werke mit Annotationen steuern

Gärtner, Markus
markus.gaertner[at]ims.uni-stuttgart.de
Universität Stuttgart, Deutschland
ORCID-iD: 0000-0002-2687-4350

Kleinkopf, Felicitas
felicitas.kleinkopf[at]kit.edu
Karlsruher Institut für Technologie (KIT), Deutschland
ORCID-iD: 0000-0001-8670-2668

Andresen, Melanie
melanie.andresen[at]ims.uni-stuttgart.de
Universität Stuttgart, Deutschland
ORCID-iD: 0000-0002-3913-1273

Hermann, Sibylle
sibylle.hermann[at]ub.uni-stuttgart.de
Universität Stuttgart, Deutschland
ORCID-iD: 0000-0001-9239-8789

Zusammenfassung. Die Herausgabe oder Veröffentlichung zugrundeliegender Daten zum Zwecke der Evaluation oder Nachnutzung durch Dritte ist ein wesentlicher Bestandteil der guten wissenschaftlichen Praxis. Dem stehen diverse Einschränkungen des geltenden Urheberrechts entgegen, wodurch insbesondere an literarischen Werken interessierte Forschungsdisziplinen massiv beeinträchtigt werden, da nur auf gemeinfreien Texten basierende Korpora oder andere Forschungsdaten uneingeschränkt weitergegeben werden können. Zwar existieren Möglichkeiten, geschützte Ursprungsdaten derart zu verändern und weiterzuverwenden, dass die endgültige Form nicht mehr dem ursprünglichen Urheberrecht unterliegt, allerdings ist dies für viele Anwendungsfälle keine zufriedenstellende Lösung. Wir stellen als Alternative das XSample Konzept vor, welches auf urheberrechtlichen Schranken basiert und es ermöglicht, kleine Auszüge geschützter Werke und Annotationen in Reinform zugänglich zu machen. Da selbige Auszüge durch ihren geringen Umfang die Gefahr bergen, nicht die für konkrete Interessen oder Forschungsfragen relevanten Datenpunkte zu enthalten, wird zusätzlich eine Korpusanfrageschnittstelle eingesetzt. Somit können Nutzer*innen ihre Bedarfe gezielt mittels einer formalen Anfragesprache definieren und die Nützlichkeit der

generierten Auszüge maximieren. Die prototypische Implementierung unseres Ansatzes setzt auf weit verbreitete Standards und Technologien und ist mit geringem Aufwand in existierende Infrastruktur integrierbar.

1 Problematik

Ein Großteil der Forschung in den Digitalen Geisteswissenschaften und anderen textbasierten Disziplinen sieht sich seit langer Zeit mit dem Problem konfrontiert, dass das geltende Urheberrecht die Weitergabe der gewonnenen Forschungsdaten erheblich erschwert, oder mitunter gänzlich unmöglich macht. Dies führte zu der immer noch andauernden Situation, dass Forschungsfragen in textverarbeitenden Disziplinen oftmals nicht am eigentlichen wissenschaftlichen Interesse und Bedarf ausgerichtet sind, sondern vielmehr basierend auf der Verfügbarkeit und (Weiter-) Verwendbarkeit der zugrundeliegenden oder angestrebten Textressourcen formuliert werden. Insbesondere Annotationsprojekte arbeiten seit jeher nahezu ausschließlich auf gemeinfreiem Material statt an zeitgenössischen Werken, da nur aus gemeinfreiem Inhalt aufgebaute Korpora auch veröffentlicht werden dürfen, was die gute wissenschaftliche Praxis erfordert. Insbesondere zum Zwecke späterer Nachprüfung und Evaluation für Anschlussforschungen ist eine öffentliche Verfügbarkeit erstrebenswert, wenn nicht sogar zwingend erforderlich. Zwar lassen sich individuelle Absprachen mit Rechteinhabern erzielen, um Arbeiten auf geschützten Werken zu ermöglichen und die resultierenden Forschungsdaten weitergeben zu können. Allerdings steht der damit häufig verbundene zeitliche Aufwand in direktem Widerspruch zur vergleichsweise kurzen Laufzeit typischer Forschungsprojekte. Im Folgenden wird auf Basis der aktuellen (deutschen) Rechtslage ein Konzept vorgestellt, das als Kompromiss die Interessen der Rechteinhaber schützen und gleichzeitig die auszugsweise Herausgabe von Textkorpora zur Nachnutzung ermöglichen soll.

2 Rechtliche Möglichkeiten

Neben expliziten Absprachen zwischen Rechteinhabern und Forschenden boten sich bisher nur begrenzte Möglichkeiten, um Forschung auf geschützten Texten durchzuführen und dabei nicht automatisch die Nachnutzbarkeit durch Nichtherausgabe von Forschungsdaten zu beeinträchtigen. Das kürzlich vorgestellte Konzept der abgeleiteten Textformate[1] beschreibt Verfahren, um Originaltexte geschützter Werke

[1] Schöch 2020.

durch verschiedene Methoden derart zu obfuskieren, dass die endgültige Repräsentation nicht mehr dem ursprünglichen Urheberrecht unterliegt und somit ohne Einschränkungen weitergegeben werden darf. Während dieser Ansatz für bestimmte Anwendungsfälle nutzbar ist, erfordern insbesondere Disziplinen mit einem hohen Anteil interpretativer Analysen zwingend Zugriff auf den originalen Kontext textueller Untersuchungsgegenstände. Dies gilt auch für die Evaluation von Anwendungsfällen mit ursprünglich rein quantitativen Untersuchungsmethoden.
Mittlerweile wurde die Problematik auch seitens des Gesetzgebers aufgegriffen: § 60d des Urheberrechtsgesetzes (UrhG) erlaubt seit 2018 die Nutzung urheberrechtlich geschützter Werke zum Zwecke der nichtkommerziellen Forschung mittels Text und Data Mining. Integriert wurden daneben auch Regularien zur Zugänglichmachung für Dritte während eines laufenden Projekts, womit zumindest in Veröffentlichungsverfahren innerhalb der aktiven Projektlaufzeit die Weitergabe im Kontext von Review-Prozessen ermöglicht wird. Nachdem der europäische Gesetzgeber im Jahr 2019 die sog. DSM-Richtlinie verabschiedet hat, erfolgten zum Juni 2021 rechtliche Änderungen im deutschen UrhG. Leider wurde in diesem Zuge keine ausdrückliche Möglichkeit geschaffen, wie Korpora zu Zwecken von Anschlussforschungen nachgenutzt werden können.

3 Auszugsweise Veröffentlichung

Im aktuell laufenden Projekt „XSample“ wurde unter Berücksichtigung der wissenschaftlichen Bedarfe und juristischen Rahmenbedingungen[2] ein Workflow-Konzept erarbeitet, wie bibliothekarische Einrichtungen als zentrale Akteure die Nachnutzbarkeit von geschützten Textkorpora unterstützen können. Dieses Konzept stützt sich auf die Verknüpfung der urheberrechtlichen Schranken des § 60c[3] und § 60d UrhG. Mittels einer als Webinterface implementierten Benutzerschnittstelle soll es Interessierten so auf einfachem Wege ermöglicht werden, basierend auf inhaltlichen Kriterien auszugsweise Zugang zu geschützten Korpora zu erhalten, um beispielsweise damit zusammenhängende Veröffentlichungen oder Ergebnisse zu validieren oder deren Eignung für Anschlussforschung im Kontext eigener Forschungsfragen zu testen. Eine

[2] Kleinkopf 2021.
[3] Diese urheberrechtliche Schranke erlaubt die anteilige Weitergabe von bis zu 15% geschützter Werke zum Zwecke der nichtkommerziellen Forschung.

prototypische Implementierung des Konzepts als JSF-Web-Applikation befindet sich derzeit in der Entwicklung und ist quelloffen[4] verfügbar. Abbildung 1 zeigt eine grobe Übersicht der am Workflow beteiligten Akteure und den Datenfluss von Primärdaten über erzeugte Forschungsdaten oder Annotationen bis hin zu generierten Auszügen zur Nachnutzung durch weitere Projekte. Aus Platzgründen wird hier auf eine detailliertere Beschreibung der Architektur[5] verzichtet. In der oberen Schleife erzeugen Forschende auf Basis geschützter Werke Forschungsdaten und pflegen diese in ein nicht-öffentliches Repositorium ein, bzw. stellen Metadaten in einem öffentlichen Teil des Repositoriums bereit, um die Auffindbarkeit der Ressourcen sicherzustellen. Durch den hohen Verbreitungsgrad von Repositorien-Systemen wie Dataverse[6] mit feingranularem Rechtesystem und der Anbindbarkeit an lokal existierende Authentifizierungsschnittstellen erfüllen Bibliotheken oder ähnliche Institutionen an Universitäten häufig bereits die infrastrukturellen Voraussetzungen für diesen Workflowabschnitt. Ausgehend von den öffentlich verfügbaren Metadaten können Interessierte nach ebenfalls erfolgter Authentifizierung die Generierung individueller Auszüge anstoßen.

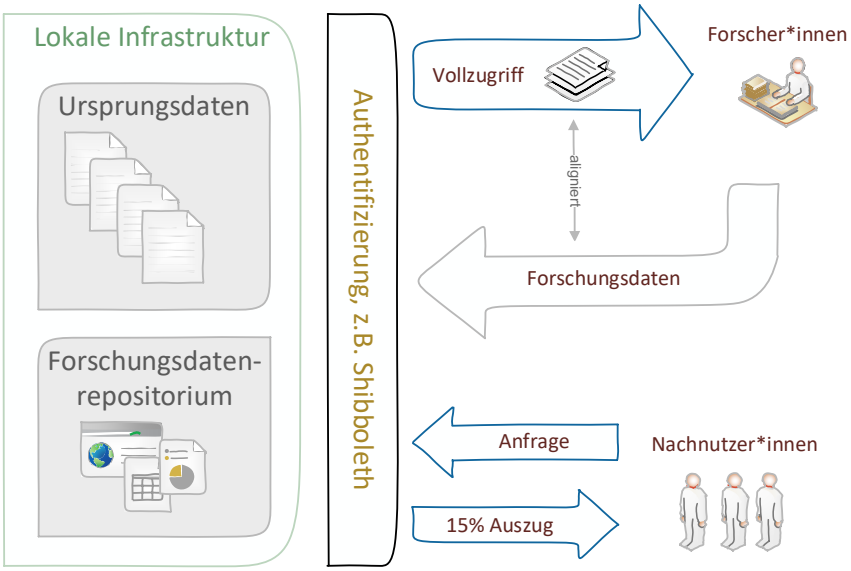


Abb. 1. Akteure und Architekturüberblick

[4] https://github.com/ICARUS-tooling/xsample-server.
[5] Gärtner 2021.
[6] The Dataverse Project, https://dataverse.org/.

4 Anfragegesteuerte Auszugerstellung

Je nach Anwendungsfall oder Forschungsfrage kann es sich als wenig zielführend herausstellen, wenn eine statische (z.B. die ersten 15%) oder zufällig zusammengestellte Teilmenge eines Textkorpus als Auszug herausgegeben wird. Um die Nützlichkeit der im Auszug enthaltenen Daten für die Nachnutzung zu maximieren, wird bei der Anfrage zur Auszuggenerierung ermöglicht, eine integrierte Korpusanfrageschnittstelle[7] zu nutzen. Anhand der im Korpus enthaltenen Annotationen kann das Interesse an bestimmten Phänomenen formal spezifiziert werden, beispielsweise an bestimmten linguistischen, aber auch allen anderen in Annotationen hinterlegten Merkmalen. Zusätzlich stehen verschiedene weitere Optionen zur Verfügung, um den Inhalt erzeugter Auszüge optimal auf die individuellen Bedürfnisse anzupassen, ohne jedoch voreilig Zugriff auf die geschützten Daten an sich zu gewähren. Auf diese Weise kann sichergestellt werden, dass Nutzer*innen im Idealfall genau den Anteil als Auszug erhalten, der für sie am interessantesten ist und sie somit daraus den größtmöglichen Nutzen unter aktuell geltenden Einschränkungen ziehen können.

Bibliografie

Gärtner, Markus. 2020. “The Corpus Query Middleware of Tomorrow – A Proposal for a Hybrid Corpus Query Architecture“, *Proceedings of the 8th Workshop on Challenges in the Management of Large Corpora*: 31-39. https://www.aclweb.org/anthology/2020.cmlc-1.5

Markus Gärtner, Felicitas Kleinkopf, Melanie Andresen and Sibylle Hermann: "Corpus Reusability and Copyright - Challenges and Opportunities", *Proceedings of the 9th Workshop on Challenges in the Management of Large Corpora*: S.10-19. https://doi.org/10.14618/ids-pub-10467.

Kleinkopf, Felicitas, Janina Jacke, Markus Gärtner. 2021. “Text- und Data-Mining“, *MMR 2021*: 196-200.

Schöch, Christof, Frédéric Döhl, Achim Rettinger, Evelyn Gius, Peer Trilcke, Peter Leinen, Fotis Jannidis, Maria Hinzmann, and Jörg Röpke. 2020. “Abgeleitete Textformate:

[7] Gärtner 2020.

Text und Data Mining mit urheberrechtlich geschützten Textbeständen.“ *Zeitschrift für digitale Geisteswissenschaften (ZfdG)* 5 (2020). http://dx.doi.org/10.17175/2020_006.

Der *dha-catalogue*: Ein Wissensgraph für die Exploration geisteswissenschaftlicher Forschungsdaten

Ďurčo, Matej
matej.durco[at]oeaw.ac.at
Österreichische Akademie der Wissenschaften, Österreich

Schneider, Gerlinde
gerlinde.schneider[at]uni-graz.at
Universität Graz, Österreich
ORCID-iD: 0000-0002-4942-9225

Zusammenfassung. Der Beitrag beschreibt die gemeinsamen Aktivitäten des österreichischen Repositorienverbundes für geisteswissenschaftliche Forschungsdaten, im Hinblick auf eine Harmonisierung der Systeme auf Ebene der Metadaten und Schnittstellen. Auf dieser Grundlage soll ein gemeinsamer Metadaten-Katalog Forschenden und Interessierten eine einheitliche Suche über die distribuierten Datenbestände einzelner Repositorien ermöglichen. Eine gemeinsam entwickelte Ontologie zur Beschreibung geisteswissenschaftlicher Datenobjekte soll Interoperabilität bezüglich der deskriptiven wie auch administrativen und strukturellen Metadaten herstellen, sowie eine inhaltliche Vernetzung durch die Repräsentation von in den Primärdaten referenzierten Entitäten erreichen. Die Ontologie soll dabei als gemeinsame Sprache zwischen den Repositorien dienen und eine semantische Vernetzung zwischen den Datenobjekten ermöglichen. In einem gemeinsamen Projekt erstellen einzelne Partner Mappings für die Überführung der lokalen Metadaten in eine RDF-Repräsentation auf Grundlage der Ontologie. Diese Daten werden jeweils über OAI-PMH zum Harvesting bereitgestellt und von einer zentralen Stelle aggregiert, indiziert und über eine Web-Benutzeroberfläche zum Suchen und Browsen bereitgestellt. Hierzu werden verschiedene Ansätze zur Exploration der Daten erprobt und evaluiert.

1 Einleitung

Dieser Beitrag beschreibt die gemeinsamen Aktivitäten des österreichischen Repositorienverbundes für geisteswissenschaftliche Forschungsdaten, im Hinblick auf eine Harmonisierung der Systeme auf der Ebene der Metadaten und Schnittstellen. Auf dieser Grundlage soll ein gemeinsamer Metadaten-Katalog Forschenden und Interessierten eine einheitliche Suche über die distribuierten Datenbestände einzelner Re-positorien ermöglichen. Dieser Katalog soll einerseits die

Sichtbarkeit und Auffindbarkeit einzelner Datenobjekte erhöhen, andererseits aber auch einen Anreiz schaffen, Forschungsdaten in einem der beteiligten Repositorien langzeitverfügbar zu archivieren.

2 Ausgangslage

In Österreich werden an mehreren Standorten institutionelle Repositorien betrieben, die der Verwaltung und Langzeitarchivierung vielfältiger geisteswissenschaftlicher Forschungsdaten dienen. Bereits 2014 wurde in einem Maßnahmenkatalog, der "DHA-Strategie", zur Förderung der digitalen Geisteswissenschaften in Österreich ein Repositorienverbund angestrebt, der das Ziel verfolgt, "digitale Ressourcen bestandsunabhängig und -übergreifend als Datenbasis für wissenschaftliche Forschungsvorhaben zu verwenden"[1]. Die im Jahr 2020 im Rahmen von CLARIAH-AT aktualisierte Strategie erneuert und konkretisiert dieses Vorhaben.[2] Der Verbund soll neben der Schaffung des gemeinsamen Datenkataloges auch als Plattform für Austausch und Kommunikation über Technologien, Best Practices, Standards im Bereich Forschungsdatenmanagement und -repositorien dienen.
Diese Aktivitäten stehen im Einklang mit zahlreichen Initiativen auf europäischer Ebene zur Metadaten-Aggregierung und Harmonisierung, welche auf eine breitere Dissemination und somit bessere Auffindbarkeit von Forschungsdaten im Sinne der FAIR Prinzipien[3] abzielen. Hier sind auf der europäischen Ebene insbesondere Europeana zusammen mit ihren nationalen und themenspezifischen Sub-Aggregatoren (in Österreich Kulturpool[4]), OpenAIRE[5], oder auch das CLARIN Virtual Language Observatory[6] zu nennen.

3 *dha-ontology*

In den letzten Jahren wurden im Rahmen des Konsortiums CLARIAH-AT, sowie im Rahmen des Projekts KONDE[7] mehrere Vorarbeiten in Richtung eines gemeinsamen Discovery-Service durchgeführt,

[1] Alram et al. 2015.
[2] CLARIAH-AT 2021.
[3] https://www.force11.org/group/fairgroup/fairprinciples.
[4] http://www.kulturpool.at/.
[5] https://www.openaire.eu/.
[6] https://vlo.clarin.eu/.
[7] http://www.digitale-edition.at/.

insbesondere wurde eine gemeinschaftliche Ontologie, die *dha-ontology* erarbeitet.[8]
Diese Ontologie dient der Beschreibung geisteswissenschaftlicher Datenobjekte mit dem Zweck Interoperabilität auf der Ebene der deskriptiven wie auch administrativen und strukturellen Metadaten herzustellen, sowie eine inhaltliche Vernetzung durch die Repräsentation von in den Primärdaten referenzierten Entitäten zu erreichen, und in weiterer Folge einen Austausch der Metadaten über die Grenzen eines Repositoriums hinweg zu ermöglichen. Die *dha-ontology* soll dabei als "gemeinsame Sprache" zwischen den Repositorien dienen und eine semantische Vernetzung zwischen den Datenobjekten ermöglichen.

Die Eckpunkte dieser Ontologie sind:

- Abbildung des Forschungskontextes
- Abbildung des Kontextes des digitalen Objektes im Repositorium (Digital Preservation)
- Explizite Unterscheidung zwischen dem digitalen Objekt und dem von ihm repräsentierten Artefakt (z.B. das Digitalisat vs. das gedruckte Buch aus dem 18. Jahrhundert)
- Hierarchische (und andere) Beziehungen zwischen Objekten
- Repräsentation von Konzepten und Entitäten aus den Primärdaten des Objektes
- Modellierung und Standardisierung von unterschiedlichen Präsentationsformen und Disseminationskanälen

Das Vorhaben fügt sich dabei in eine Reihe von aktuellen Projekten ein, welche eine graphenbasierte bzw. semantische Datenrepräsentation für die Aggregation und Vereinheitlichung von Datenressourcen wählen. Genannt seien hier der OpenAIRE Research Graph mit dem zugehörigem Datenmodell, welcher den Fokus auf die Verlinkung und Nachvollziehbarkeit von Forschungsergebnissen und deren Relationen setzt und somit die dafür relevanten Informationen als Entitäten abbildet[9], wie auch Europeana mit dem Europeana Data Model mit Fokus auf die Beschreibung und Aggregation von Cultural Heritage Objects.[10]
Die genannten Vokabularien sowie andere gängige Formate zur Beschreibung digitaler Objekte wie etwa Dublin Core/DC Terms, decken zwar Teilaspekte eines geisteswissenschaftlichen Datenobjektes ab,

[8] Ďurčo et al. 2019.
[9] Manghi et al. 2019.
[10] Isaac 2013 & Europeana 2017.

erfüllen aber nicht alle Anforderungen. Es wurde deshalb eine eigene Ontologie entwickelt, die sich dezidiert auf den Anwendungsfall bezieht. Bestehende Schemata wurden allerdings als konzeptuelle Grundlage verwendet und ein Mapping auf diese ist vorgesehen.

4 *dha-catalogue*

Im Projekt DiTAH (Digitale Transformation der österreichischen Geisteswissenschaften)[11] erstellen einzelne Partner Mappings für die Überführung der lokalen Metadaten in eine RDF-Repräsentation auf Grundlage der *dha-ontology*. Diese Daten werden jeweils über OAI-PMH zum Harvesting bereitgestellt und von einer zentralen Stelle aggregiert, indiziert und über eine Web-Benutzeroberfläche zum Suchen und Browsen bereitgestellt. Dabei kommt einerseits die bewährte Apache Solr Technologie für eine schnelle facettierte und Volltext-Suche zum Einsatz, andererseits bietet sich die Bereitstellung der Daten als RDF mit der dahinterliegenden Ontologie an, diese direkt in einen Triplestore zu importieren und die Discovery-Lösung nativ auf den RDF-Daten aufzubauen.
Das Erproben beider Ansätze sowie die Evaluierung der jeweiligen Vor- und Nachteile bzw. Stärken und Schwächen ist explizit ein Ziel des Projektes. Während die Solr-basierte Lösung hohe Performanz und Vertrautheit der Benutzer*innen mit dem Erkundungsmechanismen bei minimalen Implementierungskosten garantiert, erlaubt die Tripelstore-basierte Lösung eine uneingeschränkte Exploration des gesamten Wissensgraphen, sowie komplexe Abfragen. Die geplante inhaltliche Erschließung unterstützt diese Form des Retrieval. In den Primärdaten vorkommende Konzepte und Entitäten sollen explizit in den Metadaten erfasst und mit externen Referenzressourcen (WikiData, Geonames, etc.) verknüpft werden und dadurch eine interne Vernetzung der Datensätze ermöglichen sowie diese zu einem Teil der LOD-Cloud machen.

5 Verhältnis zu anderen Aggregatoren

Angesichts der Vielzahl an Initiativen, und sogar eines nationalen Aggregators (Kulturpool) ist die Frage nach der Notwendigkeit eines weiteren Katalogs legitim. Die Motivation ist zum einen die Spezialisierung auf geisteswissenschaftliche Forschungsdaten (im Gegensatz zu Objekten des Kulturellen Erbes bei Kulturpool/Europeana) sowie der Fokus auf Österreich als überschaubareren Aktionsraum (im Gegensatz z.B. zu

[11] https://www.ditah.at/.

OpenAIRE), welcher eine reichhaltigere und tiefere Erschließung ermöglicht. Gleichzeitig ist gerade das Ökosystem der Metadaten in der Forschung ein inhärent multipolares und es gibt gute Gründe für ein Nebeneinander von Aggregatoren mit unterschiedlichen Spezialisierungen und Aggregierungsgraden/Aktionsradien und entsprechenden Zielgruppen. Die Dissemination über möglichst viele Kanäle ist dabei ganz im Sinne des Auffindbarkeits-Gebots. So kann der *dha-catalogue* als ein Zwischenknoten in einem Aggregierungsnetzwerk dienen, und potentiell von anderen Aggregatoren geharvested werden.

6 Perspektive und Herausforderungen

Die größte Herausforderung bei der Erstellung des *dha-catalogue* liegt in der Heterogenität der Daten entlang zweier Dimensionen: Zum einen im Hinblick auf die breite Palette der Datenformate, zum anderen bezüglich der Disziplinen, in deren Kontext die Daten erstellt wurden (Die Bandbreite reicht hier von museologischen bis zu linguistischen Kontexten). Diese Heterogenität muss durch das Datenmodell abbildbar sein und auch im User Interface für die Benutzer*innen nachvollziehbar dargestellt werden. Die konzeptuelle Kongruenz muss beim Zusammenführen der Metadaten von verschiedenen Partnern mit ihren jeweils eigenen Schemata, Vokabularien und Nomenklaturen sichergestellt werden, was einen Konsens zwischen den Partnern und entsprechende lokale Implementierungen erfordert.
Der *dha-catalogue* soll 2022 online gehen, ein Prototyp wird bei der Tagung vorgestellt.

Bibliografie

Alram, M., C. Benda, M. Ďurčo, K. Mörth, S. Wentker, T. Wissik, G. Budin, M. Kaiser, S. Majewski, B. Palme, W. Scholger, H. Stigler and M. Thaller. *DH-AUSTRIA-STRATEGIE* (2015), https://doi.org/10.1553/DH-AUSTRIA-STRATEGIE-2015s1 (accessed 30 April 2021).

CLARIAH-AT Konsortium. *Digital Humanities Austria Strategie 2021+* (2021), http://gams.uni-graz.at/o:clariah.dha-strategie-2021.draft (accessed 30 April 2021).

Ďurčo, M., C. Fritze, P. Hinkelmanns, R. Hörmann, C. Koch, E. Lenhart, U. Lobis, C. Pollin, D. Schlager, D. Schopper, C.

Steiner, E. Steiner, C. Steindl, J. Stigler and M. Trognitz. *Dha-ontology* (2019), https://github.com/ditah-at/dha-ontology (accessed 30 April 2021).

Europeana Foundation. *Definition of the Europeana Data Model v5.2.8* (2017), https://pro.europeana.eu/files/Europeana_Professional/Share_your_data/Technical_requirements/EDM_Documentation//EDM_Definition_v5.2.8_102017.pdf (accessed 30 April 2021).

Isaac, A. *Europeana Data Model Primer (*2013), http://pro.europeana.eu/files/Europeana_Professional/Share_your_data/Technical_requirements/EDM_Documentation/EDM_Primer_130714.pdf (accessed 30 April 2021).

Manghi, P., A. Bardi, C. Atzori, M. Baglioni, N. Manola, J. Schirrwagen and P. Principe. *The OpenAIRE Research Graph Data Model* (2019), http://doi.org/10.5281/zenodo.2643199 (accessed 30 April 2021).

Discuss Data: Community-zentrierter Ansatz für das Forschungsdatenmanagement in den Geistes- und Sozialwissenschaften

Herrmann, Felix
felix.herrmann[at]uni-bremen.de
Forschungsstelle Osteuropa, Universität Bremen, Deutschland
ORCID-iD: 0000-0002-5134-4378

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

Zusammenfassung. Im Vortrag soll der Communiy-zentrierte Forschungsdatenmanagement (FDM) Ansatz vorgestellt werden, der im Projekt Discuss Data, einer Plattform für die Veröffentlichung, Annotierung und die Diskussion von Forschungsdaten für die Area Studies zum post-sowjetischen Raum, umgesetzt wird. Discuss Data ist als Zwischenschicht zwischen vorhandenen FDM-Diensten für Authentifizierung, Langzeitarchivierung und DOI Vergabe und den Forschenden konzipiert. Statt ganzer wissenschaftlicher Disziplinen oder Institutionen werden wissenschaftliche (Sub-)Communities für das FDM ihrer jeweiligen knowledge domain adressiert, die über große Schnittmengen in Bezug auf Daten und mit diesen zusammenhängende Methoden und Problemstellungen verfügen. Dabei wird davon ausgegangen, dass Forschende hochqualifizierte Expert/innen für die Forschungsdaten ihres Bereichs sind und aktiv an deren Kuratierung beteiligt sein sollten. Zudem ist die qualitative akademische Diskussion ein wesentlicher Teil der Qualitätssicherung von Forschungsdaten und darauf aufbauender Analysen und sollte eng angelehnt an die Daten-Veröffentlichung erfolgen. Discuss Data stellt daher interaktive, aus social media Plattformen bekannte Interfaces und Funktionen bereit, um die Einstiegshürde für Forschende in das FDM ihrer Bereiche möglichst niedrig zu halten.

Es gibt eine breite Spannweite von Angeboten zum FDM. Diese reicht von institutsspezifischen-, über fachspezifische-, bis hin zu generischen Diensten für die Speicherung und Bereitstellung von Forschungsdaten, wobei einzelne Angebote sich mitunter auf spezifische Datentypen und Methoden fokussieren.

Discuss Data[1] geht einen alternativen Weg: Wissenschaftliche Communities werden direkt angesprochen und erhalten einen eigenen Diskussionsraum für Daten. Communities werden in Discuss Data nicht als *scientific communities* mit einem direkten Fachbezug bzw. über den direkten Kontext zu wissenschaftlichen Theorien[2] bzw. Theorienetzen[3] oder etablierten Arbeitsgruppen[4] verstanden, sondern über eine gemeinsame Schnittmenge der verwendeten Methoden, Fragestellungen, Vokabularien und Datentypen definiert. Dies mag mit Fachbereichen zusammenfallen, ermöglicht jedoch auch breitere Zugänge über diese hinaus und bezieht explizit den wachsenden Trend zu trans- und interdisziplinären Forschungsansätzen mit ein. Insbesondere durch den von Digitalisierung und Kommunikation maßgeblich beeinflussten methodischen Wandel wird so ein übergreifendes Wirken in der Forschung ermöglicht, welches traditionelle Trennungen zwischen Ablage- und Wirkungsorten aufbricht.[5]

Discuss Data bietet eine Plattform für Forschende aus unterschiedlichen Fachbereichen, die ein Interesse an bestimmten Forschungsthemen, den entsprechenden Daten und mit diesen verbundenen Methoden und Problemen verbindet. Daten werden nicht bloß präsentiert, sondern bleiben ein aktiver Gegenstand im Diskurs und können darüber hinausgehend zeitlich unabhängig geteilt und diskutiert werden. Der klassische Publikationsbegriff wird um digitale Aspekte erweitert.[6] Discuss Data ermöglicht die Veröffentlichung, Annotierung und die Diskussion von Forschungsdaten und ist initial als Plattform für die Community der Area Studies zum post-sowjetischen Raum angelegt. Es wird nicht eine bestimmte Forschungs(teil)disziplin angesprochen, sondern es wird anhand eines geografischen Schwerpunktes ein Kommunikationsraum geschaffen, an dessen thematischen Kategorien (z.B. Corruption, Social Policy, Education) spezifische (Sub-)Communities Daten aus ihrem Bereich teilen und diskutieren können.

Der in Discuss Data umgesetzte community-zentrierte Ansatz für das FDM[7] geht von folgenden Grundannahmen aus:

---

1 https://discuss-data.net/.

2 Kuhn 1996, S. 44f; S. 261.

3 Netze im Sinne zusammenhängenden Theorieelementen, wie etwa in Moulines, „Theory-nets" 1979, S. 419f beschrieben.

4 Gläser 2016.

5 Könneker 2017.

6 Breuer und Trilcke 2021.

7 Herrmann und Kurzawe 2020.

1. Die qualitative akademische Diskussion ist wesentlicher Teil der Qualitätssicherung von Forschungsdaten und darauf aufbauender Analysen. Nur durch die Einbeziehung von Kontext und Methoden der Datensammlung bzw. -generierung lassen sich Daten angemessen einordnen. Die Diskussion von Daten findet in der Regel an anderen Orten, als deren Veröffentlichung statt. Dies ist eine Trennung, die im digitalen Raum keine Berechtigung mehr hat. Stattdessen werden virtuelle Kommunikationsorte benötigt, die Daten und die damit verbundene Diskurse zusammenbringen.
2. Forschende sind hochqualifizierte Expert/innen im Bezug auf die Forschungsdaten ihres Bereichs und sollten an deren Kuratierung aktiv beteiligt sein. Viele vorhandene FDM-Infrastrukturen sehen die Nutzerbeteiligung allerdings nicht vor und bieten zumeist weder einfachste Funktionen für Kommentare, noch für andere Interaktionen. Um die Einstiegshürde gering zu halten, müssen entsprechende Systeme interaktive Funktionen bieten und gleichzeitig deren Bedienung so intuitiv wie möglich gestalten.
Der Aufbau von Discuss Data wurde von 2017 bis 2021 durch die DFG gefördert. Der Quellcode steht in einem Git-Repository als Open Source bereit.[8] Die Plattform ist seit Oktober 2020 in einer Betaversion online. Sie ist als Zwischenschicht zwischen Forschenden und vorhandenen FDM-Diensten konzipiert. Diese ermöglichen die Vergabe von DOIs für eine bessere Zitier- und Auffindbarkeit und die zusätzliche Nutzung etablierter Daten-Repositories für die Langzeitarchivierung. Künftig sollen weitere Dienste angebunden werden, z.B. Bibliothekskataloge und jene von Fachinformationsdiensten zum Austausch von Metadaten oder weitere Repositorien.
Aktuell zählt Discuss Data 67 Forschende und 51 Datenpublikationen. Daten können in mehreren Formen und orientiert an den Prinzipien für FAIR Data[9] publiziert werden, im Idealfall als Open Access. Entsprechend lizensierte Daten werden dann über die DARIAH-DE[10] Infrastruktur öffentlich bereitgestellt. DARIAH-DE ist eine nationale digitale Infrastruktur für Geistes- und Kulturwissenschaften und Teil der

8 https://gitlab.gwdg.de/discuss-data/discuss-data/.
9 https://www.go-fair.org/fair-principles/.
10 https://de.dariah.eu/repository/.

europäischen Forschungsinfrastruktur DARIAH-EU.[11] Die Integration in dieses Netzwerk ist ein Baustein zur Nachhaltigkeit von Discuss Data.[12] Alle auf Discuss Data gespeicherten Daten erhalten eine DOI[13] über DataCite[14]. Daten können jedoch auch zunächst nur eingeschränkt publiziert werden, etwa wenn eine Embargofrist gilt oder Gründe, wie etwa der Schutz von Respondent/innen, gegen Open Access als Publikationsmodell sprechen. Dabei sind die Metadaten öffentlich, die Daten jedoch nur zugriffsbeschränkt einsehbar. Die Zugriffsrechte lassen sich feingranular von den Datengeber/innen steuern. Zudem können auch lediglich die Metadaten von Datensammlungen in Discuss Data veröffentlicht werden, etwa wenn die Daten nur innerhalb einer bestimmten Einrichtung eingesehen werden können. So können auch diese Daten von Forschenden gefunden sowie diskutiert und die Datengeber/innen über die Plattform kontaktiert werden.

Datasets auf Discuss Data können kommentiert werden, was das Vorhalten von Daten und das Führen von Diskursen zu diesen auf einer Plattform ermöglicht. Über Nutzer/innenprofile, die Following-Funktion und weitere Kommunikationsfeatures besteht die Möglichkeit, sich auf der Plattform zu vernetzen. Die Benutzerauthentifizierung kann auf zwei Arten geschehen: Discuss Data ist über die DARIAH-AAI[15] angebunden und erlaubt es Nutzenden sich sowohl über Shibboleth mit dem Account ihre Heimatinstitution, als auch über einen DARIAH Account anzumelden.

Das User Interface von Discuss Data erinnert bewusst an bekannte social media Plattformen und nutzt die inzwischen verbreitete Vertrautheit, um die Einstiegshürde möglichst niedrig zu gestalten. Aus dem Kreis der Nutzenden sollen im Laufe der Zeit Kurator/innen gewonnen werden, welche in thematischen Kategorien redaktionelle Aufgaben, wie die technische Prüfung der Daten und als fachliche Ansprechpartner für die Publikationen, übernehmen.

Nach einigen Monaten im Beta-Betrieb zeichnet sich ein differenziertes Bild von der Annahme der Plattform durch die Forschenden ab. Die Zahl der Neuanmeldungen und auch der Datenpublikationen nimmt

[11] DARIAH-DE wird über eine Betriebskooperation mit 16 Institutionen betrieben und ist der deutsche Beitrag zu DARIAH-EU.

[12] Die aus DARIAH-DE entwickelten Angebote werden in CLARIAH (https://www.clariah.de/) eingebracht und sind ein Baustein der kommenden NFDI Text+ (https://www.text-plus.org/).

[13] https://www.doi.org/.

[14] https://datacite.org/.

[15] https://de.dariah.eu/aai/.

kontinuierlich zu und Institutionen aus dem Forschungsfeld fragen aktiv das Projekt für Info-Veranstaltungen und potentielle Kooperationen an. Allerdings wird die Kommentarfunktion noch wenig genutzt. In einer ersten Communityumfrage unter Nutzenden und Interessierten wurde interessanterweise jedoch genau diese Funktionalität als wünschenswert und besonderes Alleinstellungsmerkmal herausgestellt. Aus den bisherigen Erfahrungen wird deutlich, dass auf die Aufbauphase eine des verstärkten Community Engagements folgen muß, in der etwa durch Workshops und Summer Schools die Userbasis erweitert und Diskurse angestoßen werden.
Discuss Data plant die Ausweitung des Modells auf weitere Communities (u.a. aus den Geschichtswissenschaften und den Digital Humanities). Perspektivisch soll weiteren interessierten Communities mit dedizierten *Community Spaces* die Möglichkeit geboten werden, eine Discuss Data Instanz zu nutzen. Durch die Bereitstellung von Schnittstellen werden Community-spezifische Mehrwertfunktionen (etwa Datenviewer oder Filterfunktionen) ermöglicht. Mittelfristiges Ziel ist die Bereitstellung einer Plattform, die generische Kernfunktionen mit flexibel zuschaltbaren Community-spezifischen Features kombiniert und damit ein nachhaltiges Entwicklungs- und Betriebsmodell ermöglicht.

Bibliografie

Breuer, C, und P Trilcke. „Die Ausweitung der Wissenschaftspraxis des Publizierens unter den Bedingungen des digitalen Wandels". Herausgegeben von der Arbeitsgruppe »Wissenschaftspraxis« im Rahmen der Schwerpunktinitiative »Digitale Information« der Allianz der deutschen Wissenschaftsorganisationen, 15 p. (2021) https://doi.org/10.48440/allianzoa.041.

Könneker, Carsten. „Wissenschaftskommunikation in vernetzen Öffentlichkeiten". In *Forschungsfeld Wissenschaftskommunikation,* herausgegeben von Bonfadelli, Heinz, Birte Fähnrich Corinna Lüthje, Jutta Milde, Markus Rhomberg, Mike S. Schäfer (2017)

Gläser, Jochen. *"Wissenschaftliche Produktionsgemeinschaften - Die soziale Ordnung der Forschung* (Frankfurt/Main; New York: Campus, 2006).

Herrmann, Felix und Daniel Kurzawe. „Discuss Data: Community-zentrierter Ansatz für das Forschungsdatenmanagement in den Geistes- und Sozialwissenschaften“, *Bausteine Forschungsdatenmanagement*: Nr. 2 (2020), https://doi.org/10.17192/BFDM.2020.2.8277.

Kuhn, Thomas S. *The structure of scientific revolutions*. 3rd ed. (1996), (Chicago, IL: University of Chicago Press, 1962).

Moulines, C.U. „Theory-nets and the evolution of theories: the example of Newtonian mechanics“. *Synthese* 41, Nr. 3, 417–39 (1979): https://doi.org/10/dksqb2.

Disziplinspezifisches Forschungsdatenmanagement
FDM-Bedarfserfassung in den Computational Literary Studies

Helling, Patrick
patrick.helling[at]uni-koeln.de
Institut für Deutsche Philologie, Universität Würzburg, Deutschland
ORCID-iD: 0000-0003-4043-165X

Jung, Kerstin
kerstin.jung[at]ims.uni-stuttgart.de
Institut für Deutsche Philologie, Universität Würzburg, Deutschland
ORCID-iD: 0000-0002-9548-8461

Pielström, Steffen
pielstroem[at]biozentrum.uni-wuerzburg.de
Institut für Deutsche Philologie, Universität Würzburg, Deutschland

Zusammenfassung. Die außerordentliche Heterogenität geisteswissenschaftlicher Forschungsfragen und Methoden stellt das Forschungsdatenmanagement (FDM) vor besondere Herausforderungen. Die DFG fördert seit 2020 ein eigenes Schwerpunktprogramm für die Computational Literary Studies (SPP-CLS), ein junges, interdisziplinäres Forschungsfeld unter Beteiligung der Literaturwissenschaften, der Informatik und der Computerlinguistik. Mit zehn geförderten Forschungsprojekten an verschiedenen Universitäten in Deutschland und der Schweiz kann dieses Schwerpunktprogramm als Blaupause aufgefasst werden, an der sich die FDM-Bedarfe dieser noch jungen Fachdisziplin untersuchen lassen. Hierzu wurde mit allen Einzelprojekten eine 2-stufige Befragung basierend auf 47 meist offen gehaltenen Leitfragen durchgeführt. Die Ergebnisse stellen die Basis für eine umfassende FDM-Landschaftsvermessung und Bedarfserfassung dar, die als Ausgangspunkt für die effektive Beschreibung und Entwicklung von FDM-Lösungsstrategien für das Programm, als auch perspektivisch für die gesamte Fachdisziplin der Computational Literary Studies dient.

1 Ausgangssituation

Forschungsdatenmanagement ist ein zentraler Bestandteil ausgezeichneter und nachhaltiger Forschung.[1] Aufgrund der Diversität geisteswissenschaftlicher Forschungsfragen, Methoden und Ziele sowie einer ausdifferenzierten, fachdisziplinären Forschungslandschaft sind geistes- und kulturwissenschaftliche Forschungsdaten typischerweise sehr heterogen.[2] Entsprechend bedarf es (fach)-spezifischer Lösungsstrategien, die eine passgenaue Bedienung fachlich geprägter Bedarfe im geisteswissenschaftlichen Forschungsdatenmanagement (FDM) im Sinne der FAIR-Prinzipien[3] ermöglichen.
Es existieren bereits unterschiedliche Angebote, die auf verschiedene fachliche oder strukturelle Communities spezialisiert sind: FDM-Services für die Breite der Geisteswissenschaften werden u.a. vom Infrastrukturkonsortium CLARIAH entwickelt.[4] Unterschiedlich angesiedelte Infrastrukturen mit ausdifferenzierteren Angeboten stellen standortunabhängige, fachspezifische Lösungen zur Verfügung.[5] Darüber hinaus konnten sich standortspezifisch einzelne fach- und disziplinspezifische FDM-Angebote etablieren[6] und nicht zuletzt sind die Entwicklungen hin zu einer Nationalen Forschungsdateninfrastruktur (NFDI) in ihrer konsortialen Organisationsform stark daten- und fachgetrieben.[7]
Ein zentrales Element bei der Entwicklung und Etablierung von solchen spezifischen und strukturell unterschiedlich angegliederten FDM-Servicestrukturen sind zu Beginn i.d.R. Bedarfserfassungen und -abfragen[8] sowie die Orientierung an Fachtraditionen.[9] Die adäquate Adressierung bestimmter (Teil)-Communities stellt sowohl bei der Bedarfserfassung

[1] Bryant, Lavoie & Maplas 2017.
[2] Pempe 2012.
[3] Wilkinson et al. 2016.
[4] „CLARIAH-DE", letzter Zugriff 30. März 2021, https://www.clariah.de/.
[5] vgl. "Language Archive Cologne", letzter Zugriff 29. März 2021, https://lac.uni-koeln.de/. "GerDraCor", letzter Zugriff 29. März 2021, https://dracor.org/ger. "Deutsches Textarchiv", letzter Zugriff 29. März 2021, https://www.deutschestextarchiv.de/. "DIPF - Leibniz-Institut für Bildungsforschung und Bildungsinformation", letzter Zugriff 29. März 2021, https://www.dipf.de/de/dipf-aktuell. "CrossAsia", letzter Zugriff 29. März 2021, https://crossasia.org/.
[6] u.a. Blumtritt et al. 2018; Helling, Moeller & Mathiak 2018.
[7] "Nationale Forschungsdateninfrastruktur", letzter Zugriff 29. März 2021, https://www.nfdi.de/.
[8] u.a. Kronenwett 2016; Akers & Doty 2013; Vock 2019.
[9] "NFDI4Memory - Problem Stories", letzter Zugriff 29. März 2021, https://4memory.de/problem-stories-overview/. "NFDI Text+ - User Stories", letzter Zugriff 30. März 2021, https://www.text-plus.org/forschungsdaten/user-stories/.

als auch bei der Einrichtung und Verbreitung von Services in diesem Zusammenhang häufig eine wesentliche Herausforderung dar.

2 Digitale Literaturwissenschaften in a Nutshell - das DFG Schwerpunktprogramm "Computational Literary Studies"

Unter dem Begriff „Computational Literary Studies" hat sich in den letzten Jahren ein Forschungsfeld etabliert, das computergestützte, quantitative Forschungsmethoden zur Analyse digitalisierter, literarischer Texte einsetzt, um literaturwissenschaftliche Forschungsfragen zu untersuchen. Es handelt sich damit um ein interdisziplinäres Forschungsfeld zwischen Informatik, Computerlinguistik und Literaturwissenschaft. Die DFG fördert seit 2020 ein eigenes Schwerpunktprogramm "Computational Literary Studies" (SPP-CLS).[10] Dieses setzt sich aus zehn Teilprojekten, einem Zentralprojekt sowie einem assoziierten Projekt an unterschiedlichen Standorten in Deutschland und der Schweiz zusammen.
Wenngleich das SPP-CLS selbst nicht die gesamte Forscher*innen-Landschaft der Fachdisziplin abbildet, so stellt es dennoch einen klar abgesteckten Raum dar, der sich durch eine Konzentration fachlich naher Forschungsprojekte auszeichnet und als Blaupause für die noch junge Fachdisziplin verstanden werden kann.

3 FDM Landschaftsvermessung im SPP-CLS

Aufgrund der Konzentration fachlich naher Forschungsprojekte bietet es sich für das Programm an gezielt spezifische FDM-Empfehlungen und Best Practices zu formulieren und Lösungsstrategien für die im Programm angesiedelten Projekte und modellhaft für die gesamte Fachdisziplin zu identifizieren bzw. zu entwickeln.

Um entsprechende FDM-Bedarfe und -Bedingungen innerhalb des SPP-CLS zu erfassen, wurden qualitative Interviews via Zoom mit Vertreter*innen aller Teilprojekte geführt. Die Durchführung der Interviews orientierte sich an einem Gesprächsleitfaden, der aus 47 meist offenen Fragen besteht (Helling et al. 2020). Die Gespräche wurden

[10] SPP 2207 „Computational Literary Studies", letzter Zugriff: 27. Juli 2021, https://dfg-spp-cls.github.io/.

aufgezeichnet. Zur Steigerung der Qualität der Antworten wurden jedem Projekt seine Ergebnisse inkl. gegebener Antwortmöglichkeiten aus Interviews mit anderen Teilprojekten für eine Überarbeitung zur Verfügung gestellt.
Die Ergebnisse der Interviews stellen die Grundlage zur bedarfsorientierten Entwicklung von Best Practices und Empfehlungen für den Umgang mit Forschungsdaten sowie dem Aufbau von passgenauen und disziplinspezifischen FDM-Lösungsstrategien dar.

3.1 Erste Ergebnisse der Landschaftsvermessung

Um einen grundsätzlichen Überblick über Methoden und disziplinspezifische Strukturen innerhalb des SPP-CLS zu erhalten, wurde unter anderem die Anwendung von Analyseverfahren und methodischen Werkzeugen erfragt (Abb. 1). Dies ermöglicht es Trends und Dynamiken innerhalb des Programms abzulesen, die schließlich als beispielhafte Beschreibung für die Definition eines State-of-the-Art der gesamten Fachdisziplin verwendet werden können.

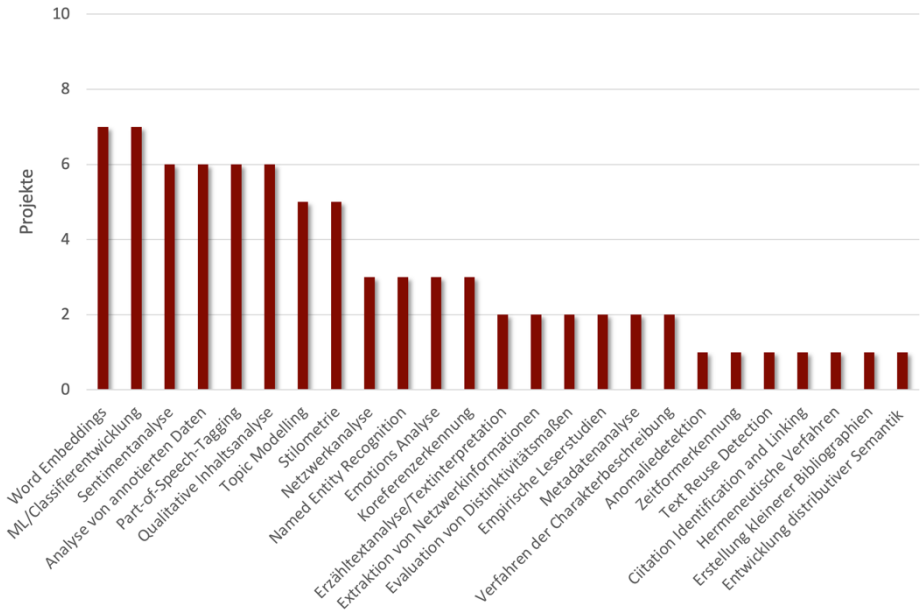


Abb. 1. Analyseverfahren und methodische Werkzeuge im SPP-CLS.

Ähnliches gilt auch in Bezug auf die Verwendung virtueller Umgebungen und technischer Infrastruktur zur Projektverwaltung (Abb. 2) sowie

zentraler Tools, bspw. zur manuellen und automatischen Annotation (Tabelle 1).

Tabelle 1. Genutzte Annotationstools innerhalb des SPP-CLS.

Annotationstool	**Projekte**
Catma	5
TreeTagger	1
CorefAnnotator	1
Sentiment Analyzer	1

Dabei stellt die erfasste Relevanz einiger Tools wie u.a. Github und Gitlab als Versionsverwaltungssysteme bereits eine wichtige Information in Bezug auf das aktive FDM innerhalb des SPP-CLS dar (Abb. 2).

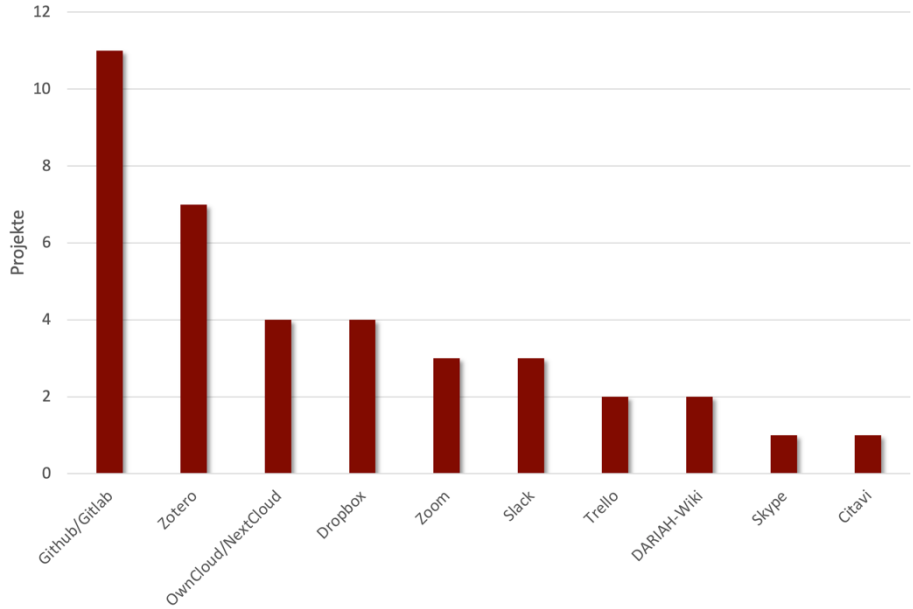


Abb. 2. Virtuelle Umgebungen und technische Infrastruktur im SPP-CLS.

Neben der Verwendung bestimmter Programmiersprachen (Tabelle 2) sowie Bibliotheken (Abb. 3), die auf Technologie-Stacks hinweisen, mit denen am Ende der Projektlaufzeiten umgegangen werden muss, sind auch die Nutzung und Produktion von bestimmten Datentypen (Tabelle 3) und -formaten (Abb. 4) wichtige Indikatoren für ein passgenaues FDM.

Tabelle 2. Programmier- und Skriptsprachen im SPP-CLS.

Programmier-/Skriptsprache	**Projekte**
Python	9
R	4
Shell-Skripte	4
Java	3
X-Technologien	2
JavaScript	2
HTML	2
CSS	2
SQL	1

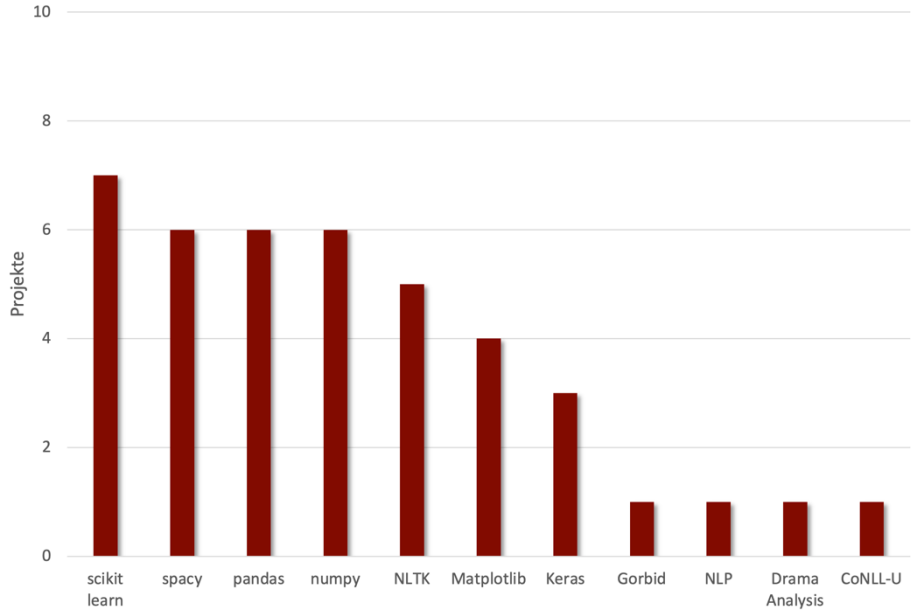


Abb. 3. Genutzte Bibliotheken im SPP-CLS.

Tabelle 3. Genutzte Datentypen innerhalb der Teilprojekte im SPP-CLS.

Datentypen	**Projekte**
Text	10
Softwarecode	9
Numerische Daten	6
Bilddaten	5
Bibliographische Daten	4
Wörterbücher/Listen	2
Interviewdaten	2
Netzwerkdaten	1

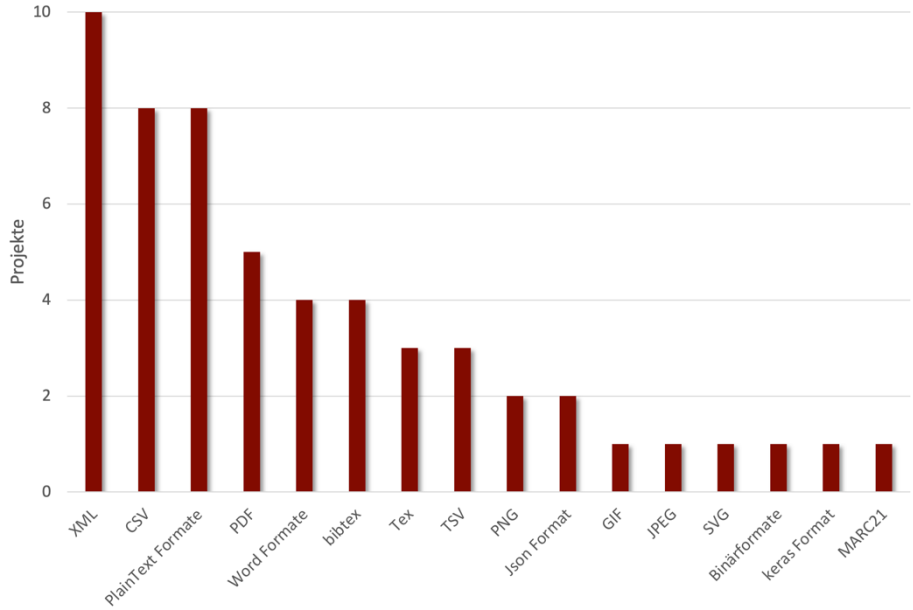


Abb. 4. Datenformate, die in den Teilprojekten eine wichtige Rolle spielen.

Allerdings stellen nicht nur Forschungsdaten Ergebnisse wissenschaftlicher Forschungsprozesse dar, sondern auch lebende Systeme wie bspw. Software, Websites oder dynamische Datenbanken (Tabelle 4).

Tabelle 4. Geplante lebende Systeme im SPP-CLS.

Lebende Systeme	**Projekte**
Website	7
Tools/Anwendungen	3
Bibliotheken	1
Dashboard	1

Zudem gilt es bei der Entwicklung einer FDM-Gesamtstrategie, insbesondere bei einer verteilten Organisationsstruktur wie im SPP-CLS, bereits in den Einzelprojekten etablierte FDM-Routinen und -strukturen (Abb. 5 und 6) zu integrieren bzw. zu berücksichtigen.

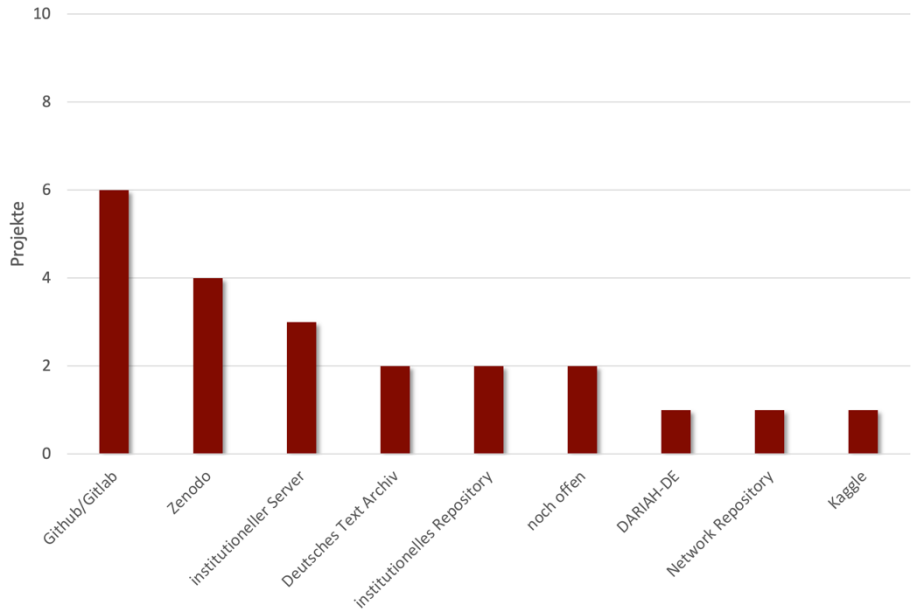


Abb. 5. Bereits bestehende Archivierungslösungen/-ansätze innerhalb der Teilprojekte.

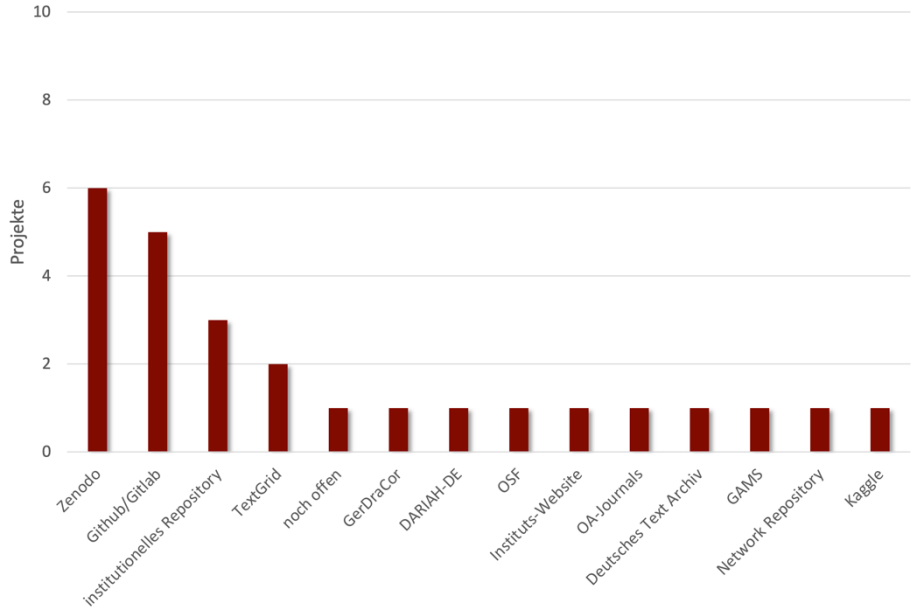


Abb. 6. Bereits bestehende Publikationslösungen/-ansätze innerhalb der Teilprojekte.

3.2 Maßnahmen, Herausforderungen und Perspektiven

Im Rahmen der Auswertungen der Interviewergebnisse konnten bereits einige Bedarfe und Anforderungen an das generelle FDM innerhalb des SPP-CLS und entsprechend perspektivisch für die Fachdisziplin identifiziert werden: Services wie Github und Gitlab stellen zentrale Instrumente der Verwaltung von und Kollaboration in Projekten dar. Aufgrund individueller Kapazitätsgrenzen und Zugriffsregelungen gab es allerdings den Bedarf einer geschlossenen, institutionell betriebenen Lösung für die Projektorganisation und den Transfer größerer Datenmodelle. Hierfür wurde eine gemeinsame Gitlab-Instanz aufgesetzt.
Trotz einiger projektübergreifend verbreiteter Datentypen und -formate weisen die Teilprojekte dennoch eine gewisse Heterogenität auf. Neben einer formatspezifischen Datenstrategie bedarf es am Ende schließlich auch eines gemeinsamen Metadatenmodells, um die technische und inhaltliche Heterogenität der Forschungsergebnisse als gemeinsames Gesamtergebnis des SPP-CLS abzubilden.

Eine besondere Herausforderung stellt der Umgang mit lebenden Systemen dar, die in einzelnen Teilprojekten auf Basis verschiedener Technologie-Stacks und mit unterschiedlichen Funktionalitäten entwickelt werden. Um diesen hochdynamischen Prozess zu kanalisieren, sollen zukünftig entsprechende Handreichungen und Best Practices formuliert und die Teilprojekte aktiv bei der Vorbereitung einer langfristigen Archivierung und Verfügbarmachung unterstützt werden.

4 Ausblick

Das SPP-CLS, die Ergebnisse der FDM-Landschaftsvermessung und die daraus entstehenden FDM-Services, -Maßnahmen und Best Practices haben das Potential als Blaupause für die gesamte Fachdisziplin zu fungieren.
In unserem Vortrag werden wir die beschriebenen Ergebnisse komplettieren und ein umfassenderes Bild der Forschungs- und FDM-Landschaft innerhalb der Computational Literary Studies zeichnen.

Bibliografie

Akers, Katherine G. und Jennifer Doty. 2013. „Disciplinary differences in faculty research data management practices and perspectives." *The International Journal of Digital Curation*, Volume 8, Issue 2: 5-26. https://doi.org/10.2218/ijdc.v8i2.263.

Blumtritt, Jonathan, Patrick Helling, Brigitte Mathiak, Felix Rau und Andreas Witt. 2017. „Forschungsdatenmanagement in den Geisteswissenschaften an der Universität zu Köln." *o-bib - Das offene Bibliotheksjournal*, Bd. 5 Nr. 3 (2018): 104-117. https://doi.org/10.5282/o-bib/2018H3S104-117.

Bryant, Rebecca, Brian Lavoie und Constance Malpas. 2017. "A Tour of the Research Data Management (RDM) Service Space. The Realities of Research Data Management, Part 1." https://doi.org/10.25333/C3PG8J.

Helling, Patrick, Katrin Moeller und Brigitte Mathiak. 2018. „Forschungsdatenmanagement in den Geisteswissenschaften – der Dienstekatalog der AG-Datenzentren des Verbands Digital Humanities im deutschsprachigen Raum." *ABI Technik*, Band 38 Heft 3: 251–261. https://doi.org/10.1515/abitech-2018-3006.

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

Kronenwett, Simone. 2016. „Forschungsdaten an der Philosophischen Fakultät der Universität zu Köln". M.A.-Arbeit, Technische Hochschule Köln. https://publiscologne.th-koeln.de/frontdoor/index/index/docId/925. Letzter Zugriff: 08.04.2021.

Pempe, Wolfgang. 2012. „Geisteswissenschaften." In *Langzeitarchivierung von Forschungsdaten. Eine Bestandsaufnahme*. Hrsg. Heike Neuroth, Stefan Strathmann, Achim Oßwald, Regine Scheffel, Jens Klump, Jens Ludwig, 137-60. Boizenburg: Verlag Werner Hülsbusch.

Wilkinson, Mark D., Michel Dumontier, IJsbrand Jan Aalbersberg, Gabrielle Appleton, Myles Axton, Arie Baak, Niklas Blomberg, et al. 2016. "The FAIR Guiding Principles for scientific data management and stewardship." *Scientific Data*, 3(1): 160018. https://doi.org/10.1038/sdata.2016.18.

Vock, Rainer. 2019. „Evaluation der FDM-Beratung 2019 – Evaluation des Beratungsangebots der Kontaktstelle Forschungsdatenmanagement (KS FDM) an der Friedrich-Schiller-Universität Jena." Bericht 4.2. eeFDM-Projekt (BMBF), Jena. https://doi.org/10.22032/dbt.40382.

„Dreistufiges Konzept zur Beratung und Schulung von Forschungsdatenmanagement“

Erfahrungsbericht

Jacob, Juliane
juliane.jacob[at]uni-hamburg.de
Universität Hamburg, Deutschland

Zusammenfassung. Als zentrale Einrichtung ist das Zentrum für nachhaltiges Forschungsdatenmanagement an der Universität Hamburg der Ansprechpartner für Forschungsdatenmanagement. Der Ursprung lag 2013 in dem Projekt "Geisteswissenschaftliche Infrastruktur für Nachhaltigkeit", das eine langfristige Verfügbarkeit von Forschungsdaten und Forschungsanwendungen an der Fakultät für Geisteswissenschaften initiierte. Bedarfe und Anforderungen rund um geisteswissenschaftliche Forschungsdaten wurden untersucht. In dem Projekt konnte identifiziert werden, dass Forschungsanwendungen häufig Einzelfalllösungen darstellen und somit eine sehr zeitaufwendige Kuration und Migration in eine geeignetere Anwendung stattfindet.
Gerade in den Geisteswissenschaften sind die erhobenen Forschungsdaten sehr heterogen und die technischen Vorkenntnisse zu Softwarelösungen anders zu bewerten als in technisch/naturwissenschaftlichen Bereichen. Eine frühzeitige Lehre über Forschungsdatenmanagement gerade im Hinblick auf Kuration, aber auch Datenlebenszyklen, Datenspeicherung etc. ist eine unumgängliche Voraussetzung für Qualitätssicherung, Langzeitarchivierung und eine erhebliche Zeitersparnis. Ebenso das stärker in den Fokus gerückte kollaborative Arbeiten an von aneinander getrennten Standorten erfordert Kenntnisse über Tools, die FAIR-Prinzipien wie bspw. Datendokumentation.
Im Vortrag soll auf das dreistufige Angebot des Zentrums zum Aufbau von Forschungsdatenmanagement-Skills eingegangen werden. Hierbei ist das Angebot zum Selbststudium, interaktiven (digitalen) Schulungen und individuelle Beratungen zu erwähnen, was jeweils auf die Vorkenntnisse der/die Studenten/innen oder Wissenschaftler/innen in Bezug auf die vorhandenen Daten oder Anwendungen angepasst wird.

1 Hintergrund

Mit dem Projekt „Geisteswissenschaftliche Infrastruktur für Nachhaltigkeit“ wurde 2013 eine langfristige Verfügbarkeit von Forschungsdaten und Forschungsanwendungen an der Fakultät für Geisteswissenschaften an der Universität Hamburg initiiert. Bedarfe und Anforderungen rund um geisteswissenschaftliche Forschungsdaten konnten während

der Projektlaufzeit identifiziert werden. Es bestätigte sich, dass Forschungsdaten und Anwendungen sehr heterogen sind und bspw. eine zeitintensive Kuration oder Migration von Einzelfalllösungen droht.
Des Weiteren ist im vergangenen Jahr durch isoliertes, aber dennoch kollaboratives Arbeiten der Bedarf bspw. an aussagekräftigen Dokumentationen und Speicherung der Daten weiter in das Bewusstsein gerückt. Mehr denn je müssen sich wissenschaftlich arbeitende Personen nicht nur inhaltlich mit ihren eigenen Daten und der Organisation derer beschäftigen, sondern auch bezogen auf nachhaltiges Forschungsdatenmanagement (FDM). Die weiterentwickelte Forschungslandschaft, die Anforderungen der Forschungsförderer und bspw. auch die wachsende Anzahl von Graduiertenkollegs rücken die Notwendigkeit und die Vorteile von effizientem FDM in den Fokus.

2 Vorstellung dreistufiges Konzept

Das Zentrum für nachhaltiges Forschungsdatenmanagement bietet daher ein dreistufiges Angebot zum Aufbau von FDM-Skills an.
Erstens ist die Webseite des Zentrums für nachhaltiges Forschungsdatenmanagement mit hilfreichen Angeboten und externen Links zum Selbststudium und Recherche zu nennen. Dieses niedrigschwellige Angebot spricht nicht nur Einsteiger an, sondern dient auch Erfahrenen bspw. bei der Antragsstellung (RDMO-Instanz vom Zentrum) oder bei der Veröffentlichung (Forschungsdatenrepositorium mit DOI-Vergabe vom Zentrum).
Zweitens sind fachspezifische Weiterbildungsangebote für Geisteswissenschaften in Form von interaktiven (digitalen) Schulungen ein effektives Angebot, dass alle Erfahrungsstufen anspricht. Zweistündige Einführungsveranstaltungen werden regelmäßig angeboten und auch über Newsletter veröffentlicht. Außerdem finden Schulungen für spezifische Gruppen wie Graduiertenkollegs statt, in denen durch die Gruppenzusammengehörigkeit die Relevanz der Lehrinhalte potenziert wird. Komplettiert wird das Angebot durch Schulungen zu speziellen Anwendungen wie u.a. die Webbasierte Datenbank „Heurist“ oder Datenbereinigung mit „Open Refine“.
Drittens wird eine individuelle Beratung angeboten. Einerseits in einer niedrigschwelligen offenen Sprechstunde, die aktuell digital stattfindet, und andererseits Terminberatung für spezifische Fragestellungen.

3 Fokus interaktiven (digitalen) Schulungen

Insbesondere auf das Schulungsangebot soll detailliert eingegangen werden, da hierdurch eine größere Teilnehmerzahl angesprochen und mit dem gewonnenen Wissen wiederum gezieltes Selbststudium erleichtert wird.
Das Schulungsmaterial ist fachlich modular aufgebaut und in Markdown geschrieben; als Zeichen für aktiv gelebte Nachhaltigkeit und Unabhängigkeit vom Betriebssystem bzw. der Software. Somit können alle Mitarbeitern*innen des Zentrums bedarfsgerecht zugreifen, durchsuchen und anpassen. Außerdem verzahnen sich die fachlichen mit den methodischen Inhalten, so dass gerade in der derzeitigen Online-Lehre die Didaktik angepasst ist, um die Interaktivität zu erhalten. Beispielhaft soll an einer Schulung für Geisteswissenschaftler vorgestellt werden, wie sich die fachlichen und didaktischen Inhalte verzahnen lassen und wie Teilnehmende befähigt werden, Gelerntes unmittelbar zu implementieren.
Für das Gelingen einer Online-Schulung muss der/die Vortragende sehr vertraut mit der Online-Präsentationsumgebung sein, um notfalls zu technischer Hilfestellung fähig zu sein. Außerdem sollten Online-Tools so gewählt sein, dass die Usibility der eingebundenen, externen Anwendungen sowohl für den Vortragenden als auch für die Teilnehmenden praktisch schnell und ohne komplizierte Erläuterungen umsetzbar sind. Auch hierbei sollte die Online-Veranstaltung inhaltlich und didaktisch so konzipiert sein, dass die Aufmerksamkeit und Mitarbeit gezielt aktivieren und das Schulungsziel klar definiert ist.
Genauer thematisiert werden die Erfahrungen von didaktisch/methodischen Elementen wie Breakout-Sessions für Diskussionen und kollaboratives Arbeiten, integrierte Umfragen, externe Umfragetools (PollEv, Pingo), Whiteboards, Padlet, Mentimeter, sowie verschiedenen Sozial- und Arbeitsformen.
Die Inhalte reichen von grundlegenden Begriffen wie dem FAIR-Prinzip und allen Schritten des Forschungsdatenzyklus, Definition von geisteswissenschaftlichen Forschungsdaten und Gründe für FDM, über relevante Dienste (wie RDMO) und Anwendungen (Datenbank „Heu-rist“, Sammlungsportal FUNDus!), bis hin zu Open Access Veröffentlichungen im Repositorium mit DOI-Vergabe. Das Thema der Datensicherheit und Verwaltung (Stichwort Versionierung) ist aus dem Home-Office heraus ebenfalls eine besondere Herausforderung.
Eine Hürde stellt die Sichtbarkeit und damit auch die Zahl der Kursanmeldungen innerhalb einer Einrichtung dar. Nicht immer ist es sinnvoll Kurse Institutsübergreifend zu öffnen, sondern eine gezielte Suche nach

Multiplikatoren (z.B. Clustern/ Sonderforschungsbereichen/ Mailinglisten von der Fakultät für Geisteswissenschaften/ Graduiertenkollegs) ist eine hilfreiche Lösung. Des Weiteren schafft es ein „Wir-Gefühl“, wenn sich eine vordefinierte Gruppe mit FDM beschäftigt und auch im Anschluss noch an FDM weiterarbeiten kann und bspw. ein DMP gemeinsam erstellen wird oder Daten begutachtet/bereiniget werden.

Ein Standard in der Praxis: ISO 24624:2016 Transcription of spoken language

Schmidt, Thomas
thomas.schmidt[at]ids-mannheim.de
Leibniz-Institut für Deutsche Sprache (IDS), Mannheim, Deutschland

Hedeland, Hanna
hedeland[at]ids-mannheim.de
Leibniz-Institut für Deutsche Sprache (IDS), Mannheim, Deutschland

Frick, Elena
frick[at]ids-mannheim.de
Leibniz-Institut für Deutsche Sprache (IDS), Mannheim, Deutschland

Zusammenfassung. Der Beitrag beschreibt die Entwicklung und Anwendung des TEI-basierten ISO-Standards ISO 24624:2016 Transcription of spoken language, der seit einigen Jahren für gesprochensprachliche Forschungsdaten aus unterschiedlichen Kontexten eingesetzt wird. Ein standardisiertes Dateiformat ermöglicht Interoperabilität zwischen verschiedenen Werkzeugen und weiteren Angeboten von Datenzentren und Infrastrukturen. Durch die methodologisch fundierte Abwägung zwischen Standardisierung und Flexibilität kann der ISO/TEI-Standard zudem Forschungsdaten aus verschiedenen Forschungskontexten abbilden, und so interdisziplinäre Vorhaben erleichtern. Der Beitrag stellt einige Anwendungsbereiche aus dem Lebenszyklus gesprochensprachlicher Forschungsdaten vor, in denen auf dem ISO/TEI-Standard basierenden Erweiterungen existierender Softwarelösungen erfolgreich umgesetzt werden konnten, und zeigt weitere Beispiele für die zunehmende Verbreitung des Formats.

1 Einführung

Der Standardisierung von Dateiformaten – über einzelne Werkzeuge, Datenzentren und möglichst auch Disziplinen hinweg – kann eine Schlüsselrolle im Aufbau digitaler Infrastrukturen zukommen. Bei ausreichender Verbreitung und hinreichender technischer Unterstützung haben Standards das Potential, Arbeitsabläufe zu effektivieren, die Nachnutzbarkeit und Interoperabilität von Tools zu steigern und die nachhaltige Archivierung und Bereitstellung von Forschungsdaten zu vereinfachen. Nicht-wissenschaftsspezifische Standards wie XML (als

W3C-Recommendation für strukturierten Text) oder MPEG (für audiovisuelle Daten) belegen schon seit langem, dass und wie sich dieses Potential realisieren lässt. Für wissenschaftsspezifische Standards ist teilweise noch offen, welche Form und Art der Verbreitung am besten geeignet sind.
Im Jahr 2016 hat die ISO unter dem Titel „ISO 24624:2016 Language resource management — Transcription of spoken language“ (ab hier: ISO/TEI-Standard) einen Standard für die digitale Repräsentation von Transkripten gesprochener Sprache veröffentlicht. Dieser basiert auf Kapitel 8 der Richtlinien der Text Encoding Initiative und orientiert sich an den in Schmidt (2011) formulierten Ideen, die TEI-Richtlinien mit der Praxis, in der die Formate verschiedener Transkriptionstools wie ELAN, EXMARaLDA oder Praat als De-Facto-Standards fungieren, in Einklang zu bringen.

2 Anwendungskontexte des ISO/TEI-Standards

In diesem Beitrag werden exemplarisch verschiedene Kontexte vorgestellt, in denen der ISO/TEI-Standard zum Einsatz kommt. Dabei wird jeweils kurz der Kontext selbst beschrieben und dann auf spezifische Herausforderungen eingegangen, die sich in Bezug auf die praktische Nutzung, die Akzeptanz und die Verbreitung des Standards stellen.

2.1 Datenerzeugung: EXMARaLDA

Aus Sicht der Forschenden ist die Nutzung eines Standards nur dann eine Option, wenn sie sich mit etablierten Arbeitsabläufen vereinbaren lässt und möglichst wenig zusätzlichen Aufwand verursacht. Für Korpora gesprochener Sprache, wo die manuelle Transkription und Annotation des audiovisuellen Quellmaterials einen wesentlichen Anteil der Arbeit ausmacht, wird kein Standard Aussicht auf Akzeptanz haben, der sich nicht von mindestens einem einer kleinen Menge etablierter Tools erzeugen lässt. EXMARaLDA[1] – neben ELAN[2], Praat[3] und Transcriber[4] eines der weiter verbreiteten solchen Tools – kann in seiner neuesten Fassung nun ISO/TEI-Daten in verschiedenen Varianten importieren und exportieren. Etablierte Arbeitsabläufe müssen dafür nicht verändert werden.

[1] Schmidt u. Wörner 2014, https://exmaralda.org.
[2] https://www.mpi.nl/corpus/html/elan/.
[3] http://www.praat.org.
[4] http://transag.sourceforge.net/.

2.2 Manuelle Annotation: WebAnno

Im Rahmen eines Projekts zur Förderung innovativer Lehrkonzepte wurde das webbasierte Annotationswerkzeug WebAnno[5] um ein Modul für die Unterstützung audiovisueller Daten sowie die Visualisierung von Transkriptionsdaten im ISO/TEI-Format erweitert[6]. Im Unterschied zu den gängigen Transkriptions- und Annotationswerkzeugen erlaubt WebAnno einen erweiterten Fokus auf größere Abschnitte des Diskurses, sowie komplexe Annotationstypen in Form von hierarchischen Bäumen oder Ketten. In der Lehre wurde die in WebAnno bereits vorhandenen Funktionen der kollaborativen Annotation und Darstellung der (Nicht-)Übereinstimmung der Annotation (Inter-Annotator-Agreement) genutzt.

2.3 Automatische Annotation: WebLicht / WebMAUS

In der europäischen Forschungsinfrastruktur CLARIN[7] wurde der Ansatz verfolgt, automatische Annotationsaufgaben über standardisierte Web-Schnittstellen zu realisieren. Waren die Schnittstellen anfangs noch auf schriftsprachliche Daten einerseits (das Format TCF und der Dienst WebLicht[8], der zahlreiche NLP-Webservices anbietet) und „phonetische" Daten andererseits (die Webservices WebMaus[9] für phonetische Segmentierung) ausgerichtet, können WebLicht und WebMaus mittlerweile auch Daten in ISO/TEI verarbeiten bzw. erzeugen. Damit erweitern sich der Einsatzbereich der Webservices und die Möglichkeiten, Aufgaben in der Erstellung und Analyse mündlicher Korpora zu automatisieren.[10]

2.4 Datenqualität: QUEST

Im BMBF-geförderten Projekt QUEST[11], das mittels Qualitäts- und Kurationskriterien für audiovisuelle annotierte Sprachdaten deren Nachnutzung fördern und vereinfachen möchte, spielt der ISO/TEI-Standard eine zentrale Rolle. Da Transkriptionsdaten aus verschiedenen Disziplinen in verschiedenen Dateiformaten erstellt werden, ermöglicht eine Abbil-

[5] Eckart de Castilho et al. 2014.
[6] Remus et al. 2019, eine Demo-Version steht unter http://ltdemos.informatik.uni-hamburg.de/webanno-mm zur Verfügung.
[7] https://www.clarin.eu/.
[8] https://weblicht.sfs.uni-tuebingen.de.
[9] https://clarin.phonetik.uni-muenchen.de/BASWebServices/interface/WebMAUSBasic
[10] vgl. Schmidt et al. 2017, Fisseni u. Schmidt 2020.
[11] https://www.slm.uni-hamburg.de/ifuu/forschung/forschungsprojekte/quest.html.

dung auf das ISO/TEI-Format einerseits überhaupt eine einheitliche Betrachtung und Verwendung der Daten, andererseits ist für diese Abbildung eine genaue Dokumentation der Originaldaten notwendig. Die erforderliche Dokumentation benutzter Formate, Schemata und Konventionen erfüllt zugleich entsprechende Anforderungen der FAIR-Prinzipien und dient daher als Operationalisierung dieser für Nachnutzende entscheidenden Aspekte der Datenqualität.[12]

2.5 Datenzugang (Query, Anzeige): ZuMult

Für die Visualisierung und Query mündlicher Korpora werden im Projekt ZuMult[13] webbasierte Zugänge entwickelt, die auf einer weiter spezifizierten Fassung des ISO/TEI-Standards basieren. Neben XSL-Stylesheets, die verschiedene Ansichten der XML-basierten Ausgangsdaten generieren, entwickelt ZuMult vor allem eine auf dem MTAS-System[14] basierende Query Engine, über die mündliche Korpora mit vielfältigen Annotationen mittels der in der Korpuslinguistik weit verbreiteten CQP Query Language[15] abgefragt werden können.[16]

2.6 Weitere Entwicklungen und Einsatzbereiche

Die genannten Kontexte sind eine Auswahl derjenigen, in denen wir selbst Anwendungen entwickeln, für die der ISO/TEI-Standard eine zentrale Rolle spielt. Fünf Jahre nach Veröffentlichung des Standards lässt sich jedoch durchaus beobachten, dass es weitere Zusammenhänge gibt, in denen der Standard sich zu etablieren beginnt. Exemplarisch genannt seien das im Umfeld des französischen CORLI-Konsortiums (CORpora, Languages and Interaction) entwickelte System TEICorpo[17], das Spoken Torlak Corpus[18] oder die Verwendung im Rahmen des Langzeitvorhabens INEL[19], das sich mit der Dokumentation indigener nordeurasischer Sprachen befasst.

[12] vgl. Hedeland 2021.
[13] http://zumult.org.
[14] Brouwer et al. 2016.
[15] http://cwb.sourceforge.net/files/CQP_Tutorial/.
[16] Frick u. Schmidt 2020.
[17] Parisse et al. 2018, https://ct3.ortolang.fr/ct3/teicorpo-doc/.
[18] Vuković 2021, 2020.
[19] Ferger u. Jettka 2020, https://www.slm.uni-hamburg.de/inel/.

3 Diskussion

Als vorläufige Erkenntnis zeigt sich in der Summe, dass sich durch die Einführung des ISO/TEI-Standards neue Möglichkeiten auftun, in denen sich das versprochene Potential zu realisieren beginnt. Entscheidend für die Praxis sind dabei nach unserer Ansicht erstens eine adäquate Unterstützung durch geeignete Tools, zweitens ein für ForscherInnen greifbarer Mehrwert durch verbesserte Interoperabilität. Beides erfordert aufwändige Entwicklungsarbeit und kontinuierliche Abstimmung mit AnwenderInnen, anderen EntwicklerInnen und Infrastrukturinitiativen.

Bibliografie

Brouwer, Matthijs, Hennie Brugman und Marc Kemps-Snijders. 2017. "MTAS: A Solr/Lucene based Multi Tier Annotation Search solution." In *Selected papers from the CLARIN Annual Conference 2016, Aix-en-Provence, 26–28 October 2016. Linköping Electronic Conference Proceedings*. 136, Nr 2 (2017): 19–37.

Eckart de Castilho, Richard, Chris Biemann, Irina Gurevych und Seid Muhie Yimam. 2014. "WebAnno: a flexible, web-based annotation tool for CLARIN." In *Proceedings of the CLARIN Annual Conference 2014, October 24–25, 2014, Soesterberg, The Netherlands*.

Ferger, Anne und Daniel Jettka. 2020. "Use Cases of the ISO Standard for Transcription of Spoken Language in the Project INEL." In *Proceedings of CLARIN Annual Conference 2020. Virtual Edition*, hrsg. v. Navarretta, Constanze und Maria Eskevich, 126–130. Utrecht: CLARIN.

Fisseni, Bernhard und Thomas Schmidt. 2020. "CLARIN Web Services for TEI-annotated Transcripts of Spoken Language." In *Selected Papers from the CLARIN Annual Conference 2019. Linköping Electronic Conference Proceedings*. 172, Nr. 3 (2020): 12–22.

Frick, Elena und Thomas Schmidt. 2020. "Using Full Text Indices for Querying Spoken Language Data." *In Proceedings of*

the LREC 2020 Workshop, Language Resources and Evaluation Conference, 11–16 May 2020, 8th Workshop on Challenges in the Management of Large Corpora (CMLC-8), 40–46. Paris: European Language Resources Association, 2020.

Hedeland, Hanna. 2021. “Towards Comprehensive Definitions of Data Quality for Audiovisual Annotated Language Resources.” In *Selected papers from the CLARIN Annual Conference 2020. Linköping Electronic Conference Proceedings*. 180, Nr. 11 (2021). 93–103.

Parisse, Christophe, Céline Poudat, Ciara R. Wigham, Michel Jacobson und Loïc Liégeois. 2018. “CORLI: A linguistic consortium for corpus, language, and interaction.” In *Selected papers from the CLARIN Annual Conference 2017, Budapest, 18–20 September 2017. Linköping Electronic Conference Proceedings*. 147, Nr. 2 (2018): 15–24.

Remus, Steffen, Hanna Hedeland, Anne Ferger, Kristin Bührig und Chris Biemann. 2019. “WebAnno-MM: EXMARaLDA meets WebAnno.” In *Selected papers from the CLARIN Annual Conference 2018, Pisa, 8–10 October 2018. Linköping Electronic Conference Proceedings*. 159, Nr. 17 (2019): 166–172.

Schmidt, Thomas. 2011. “A TEI-based approach to standardising spoken language transcription.” *Journal of the Text Encoding Initiative* 1 (2011). doi:10.4000/jtei.142.

Schmidt, Thomas und Kai Wörner. 2014. “EXMARaLDA.” In *The Oxford Handbook of Corpus Phonology*, hrsg. v. Durand, Jacques, Ulrike Gut und Gjert Kristoffersen, 402–419. Oxford: OUP

Schmidt, Thomas, Hanna Hedeland und Daniel Jettka. 2017. “Conversion and annotation web services for spoken language data in CLARIN.” In *Selected papers from the CLARIN Annual Conference 2016, Aix-en-Provence, 26–28 October 2016. Linköping Electronic Conference Proceedings*. 136, Nr 9 (2017): 113–130.

Vuković, Teodora. 2020. *Spoken Torlak dialect corpus 1.0 (transcription)*. Slovenian language resource repository CLARIN.SI. http://hdl.handle.net/11356/1281.

Vuković, Teodora. 2021. “Representing variation in a spoken corpus of an endangered dialect: the case of Torlak.” *Language Resources and Evaluation*. https://doi.org/10.1007/s10579-020-09522-4.

Einfach FAIR

Geisteswissenschaftliches Arbeiten und nachhaltiges Publizieren von Forschungsdaten mit Git

Arnold, Eckhart
arnold[at]badw.de
Bayerische Akademie der Wissenschaften, Deutschland
ORCID-iD: 0000-0002-1138-499X

Müller, Stefan
mueller[at]badw.de
Bayerische Akademie der Wissenschaften, Deutschland

Gaab, Jochen
Gaab[at]kbl.badw.de
Bayerische Akademie der Wissenschaften, Deutschland

Landes, Florian
Landes[at]kbl.badw.de
Bayerische Akademie der Wissenschaften, Deutschland

Raaf, Manuel
raaf[at]badw.de
Bayerische Akademie der Wissenschaften,

Zusammenfassung. Im Vortrag soll erläutert werden, inwieweit geisteswissenschaftliche Forschungsdaten mit Git und einem Git-Repositorium gemeinschaftlich und in verschiedenen Fassungen und Erstellungstufen bearbeitet, gesichert und veröffentlicht werden können. Wie wir im Vortrag darlegen werden lässt sich durch die Verwendung eines Git-Repositoriums als Veröffentlichungsplattform bereits ein großer Teil der Anforderungen für FAIR-Data abdecken, ohne dass auch nur eine Zeile Programmcode dafür geschrieben werden muss. Damit rücken git-Repositorien (und dasselbe gilt auch für Repositorien mit anderen Versionsverwaltungssystemen wie mercurial, subversion etc.) schon beinahe in die Nähe einer Jedermanns-Technologie für die Veröffentlichung von geisteswissenschaftlichen Forschungsdaten.

Wie das möglich ist, welche geringe Mehrarbeit dennoch geleistet werden muss und was man bei der Veröffentlichung von Forschungsdaten mit git noch beachten muss (z.B. Datenschutz), diskutieren wir in diesem Vortrag, wobei wir auch auf unsere eigenen Erfahrungen mit git und gitlab als Veröffentlichungsplattform eingehen.

1 Einleitung: Wofür Versionskontrollsysteme eingesetzt werden können

Für die Arbeit an Programmen, zumal die gemeinsame Arbeit an größeren Programmen, ist seit langem die Verwendung von Austausch- und Versionierungswerkzeugen üblich geworden. Von diesen Werkzeugen am meisten verbreitet ist mittlerweile Git; doch das Folgende trifft auf andere wie Mercurial, Fossil oder Subversion mutatis mutandis zu.
Die wesentlichen Vorteile, die der Einsatz von Git u. a. hat, sind aber nicht auf die Arbeit mit Programmtexten beschränkt, sondern lassen sich bei der Arbeit mit Texten überhaupt genießen, also auch bei der Arbeit mit geisteswissenschaftlichen Textdaten – wie Editionen, Transkriptionen, Korpora und Katalogen. Andererseits sind natürlich auch Nachteile in Kauf zu nehmen.
Grundsätzlich lassen sich diese Arten von Daten mit git verwalten:

- Pflegen und Veröffentlichung in Auflagen von zeilenbasierten Textdaten: Die Daten sind zeilenbasierte Textdaten, angereichert mit Leseansichten, werden hin und wieder verbessert und dann in neuer Auflage zur Verfügung gestellt.

- Erarbeitung und Bearbeitung von XML-Daten oder anderen Markup-Daten: Die Daten sind XML-Daten und werden fortlaufend geändert und ergänzt. Git wird im Hintergrund benutzt, aber von den Redakteuren über eine Schnittstelle, die z. B. automatisch Bearbeitungsschritte in Git-Aktionen (namentlich Commit und Push) übersetzt.

- Mit Einschränkungen kann git auch für Binärdaten eingesetzt werden. Zumindest Rohdaten, wie z.B. Bildscans von Texten, die sich nicht mehr ändern, können ganz gut noch mit im Repositorium untergebracht werden. Bei Binärdaten, deren Bearbeitungsschritte mit der Versionsverwaltung mitverfolgt werden sollen, können notfalls Erweiterungen wie GitLFS helfen.

2 Was einem der Einsatz von git bringt (Vorteile)

- Die Arbeit am Text kann in den gewohnten Ordnern und Dateien und mit den gewohnten Eingabewerkzeugen erfolgen. Der Einsatz von Git u. ä. erfordert weder die Überführung der Daten in ein neues System noch die Einübung neuer Eingabewerkzeuge.
- Die Arbeit am Text kann auf dem eigenen Rechner erfolgen und erfordert keinen Netzanschluss. (Das kann für die Arbeit z. B. in Archiven nützlich sein.)
- Der Austausch der Einzelarbeit in einer Gruppe, die an einem gemeinsamen Textgesamt arbeitet, wird so unterstützt, dass die Einzelbeiträge automatisch zusammengeführt werden und nicht verschiedene Fassungen entstehen, von denen jede einen anderen Fortschritt gegenüber der anderen enthält. Das ist ja eine der größten Gefahren beim gemeinschaftlichen Arbeiten.
- Kommt es zu einer Überschneidung, die nicht automatisch vermieden oder aufgelöst werden konnten, wird sie beim Austausch angezeigt, ihre Behebung wird unterstützt, und der Austausch wird erst abgeschlossen, wenn die Überschneidung behoben ist.
- Man kann seine Arbeit aufwandslos versionieren, also Bearbeitungsstufen festhalten, zwischen denen man nach Belieben wechseln kann.
- Man kann seine Arbeit aufwandslos in verschiedene Fassungen gliedern – z. B. für gefahrloses Ausprobieren größerer Änderungen oder zum Erstellen von Ausgaben für Vorveröffentlichungen, aus denen Unfertiges gestrichen ist.
- Man kann diese Bearbeitungsstufen und Fassungen auf Dauer sichern und in beliebiger Auswahl veröffentlichen und zum Herunterladen unter Permanentlinks anbieten.
- Man kann den Bearbeiterkreis schrittweise öffnen und Beiträge von außen bequem prüfen und eingliedern.
- So lässt sich eine Bereitstellung nach den FAIR-Prinzipien, soweit die Datenhaltung selbst betroffen ist, einfach umsetzen: Der Einsatz von Git führt von selbst zu einer versionierten Bereitstellung unter permanentlinkfreundlichen URLs, die mit jedem Git-Repositorium kompatibel ist, entsprechend leicht umgezogen und von allen Nutzungswilligen leicht und verlustfrei kopiert und auch wieder lokal gespeichert werden kann, zudem zur Verwendung von menschenlesbaren Textdaten und zur Kollaboration einlädt. Man

bekommt gewissermaßen folgende Punkte auf der FAIR-Checkliste geschenkt:

- Menschen und Maschinengerechter Zugriff auf die Forschungsdaten, einschließlich der Möglichkeit des Gesamtdownloads (in git-Terminologie des „klonens“) aller Daten auf einmal.
- Versionierung und Links auf die einzelnen Versionen, die mit Hilfe von Tagging, semantisch plausibel gestaltet werden können.
- Um daraus Permalinks zu generieren ist höchstens noch das Vorschalten einer zusätzlichen Link-Auflösungsschicht (Resolver) erforderlich.
- Damit sind der Zugreifbarkeit (Accessibility) keine Schranken mehr in den Weg gestellt, und die Nachnutzbarkeit (Reusability) und Interoperabilität soweit vorbereitet, wie das auf der Abstraktionsebene eines inhaltsunabhängigen Datenverwaltungssystems überhaupt möglich ist.
- Es ist ein Werkzeug, das Entwickler und Wissenschaftler gemeinsam nutzen könnten, was für die gemeinsame Arbeit z. B. an einer Netzveröffentlichung günstig sein mag.

3 Herausforderungen und Grenzen beim Einsatz von git (Nachteile)

- Die Datenhaltung ist vor allem für Textdateien geeignet und nicht oder nur weniger gut für Audio- und Bild- oder andere Binärdateien.
- Git hält genau fest, wer wann welche Änderungen durchführt, was zu Datenschutzproblemen führen kann. Ein Lösungsansatz besteht darin, mit zwei Repositorien zu arbeiten, einem nichtöffentlichen Arbeitsrepositorium und einem öffentlichen Publikationsrepositorium. Wichtig ist, dass beide Repositorien unabhängig voneinander sind, also keine sog. „Klone“, und dass in das Veröffentlichungsrepositorium nur die jeweils jüngste Schicht der Versionsgeschichte einer freigegebenen Version übernommen wird.
- Wie einfach ist Git wirklich? Die Benutzung bei der alltäglichen Datenarbeit könnte den Arbeitsablauf hemmen und den Wunsch nach einer Schnittstelle entstehen lassen (wie im Fallbeispiel 3 benutzt), was aber wieder den Entwicklungsaufwand erhöht. Eine Lösung kann die Verwendung eine git-Shell oder der eingebauten git-Unterstützung des Texteditors sein.

4 Resumé

Git bringt von Haus aus fast alles mit, um die Veröffentlichung von (textförmigen) Forschungsdaten entsprechend der FAIR-Prinzipien zu ermöglichen. Die notwendigen Anbauten beschränken sich auf einen Link-Auflöser (um Permalinks zu gewährleisten) und ggf. ein Zweit-Repositorium für die datenschutzkonforme Veröffentlichung. Damit sind auch schon zwei der vier FAIR-Prinzipien, nämlich Accessibility und Reusability weitgehend abgedeckt. Die beiden verbleibenden Prinzipien Findability und Interoperability betreffen dagegen Aspekte, die auf der Ebene der Repositorien-Technologie allein ohnehin nicht gelöst werden können.

5 Bibliografie

- Perkel, J. Democratic databases: science on GitHub. Nature 538, 127–128 (2016). https://doi.org/10.1038/538127a.

- Kaden, Ben, Kleineberg, Michael, Lubetzki, Nathalie, & Walk, Martin. (2018, May). Der Umgang mit dissertationsbezogenen Publikationen von Forschungsdaten. Ergebnisse des eDissPlus-Projektes. Zenodo. http://doi.org/10.5281/zenodo.1254902.

FDM lokal, gemeinsam und vernetzt gestaltet

Proaktiv zum nachhaltigen FDM

Sutter, Paul
paul.sutter[at]bbaw.de
Berlin-Brandenburgische Akademie der Wissenschaften, Deutschland

Marciniak, Katja
katja.marciniak[at]bbaw.de
Berlin-Brandenburgische Akademie der Wissenschaften, Deutschland
ORCID-iD: 0000-0002-8031-4813

Zusammenfassung. Die interdisziplinäre Initiative „Forschungsdatenmanagement“ (FDM) hat das Ziel, alle an der Berlin-Brandenburgischen Akademie der Wissenschaften (BBAW) generierten Forschungsdaten zu ordnen, zu sichern und langfristig verfügbar zu halten. Über einen Stakeholder-orientierten Ansatz wird eine einheitliche FDM-Strategie entwickelt, welche die lückenlose Betreuung aller Projekte und Vorhaben sichert. Die Initiative sucht proaktiv das Gespräch mit den WissenschaftlerInnen um frühzeitig Kontakt aufzubauen und den aktuellen Stand des Datenmanagements abzufragen. Gemeinsam mit den weiteren Beteiligten (IT, DH-Einheit, Archiv, Bibliothek) wird mit jedem Vorhaben ein lokal angepasster Datenmanagementplan (DMP) mittels eines Fragenkatalogs in RDMO (Research Data Management Organiser) erstellt. So werden Vereinbarungen und Maßnahmen explizit festgehalten, die zur Datenqualität beitragen, das Risiko des Datenverlustes verringern oder die Veröffentlichung der Daten betreffen. Ziel ist es, die sichere Aufbewahrung und Zugänglichmachung des kulturellen Erbes zu unterstützen, das im Rahmen der Akademienvorhaben teils über Jahrzehnte erforscht, erschlossen und bearbeitet wird. Flankiert wird das Angebot der kollaborativen DMP-Erstellung von einem Beratungs- und Schulungsangebot der Initiative. Die Summe der systematisch abgefragten Informationen aus allen Datenmanagementplänen dient der BBAW außerdem als Wissensbasis, ermöglicht dadurch ein passgenaues Vorgehen und ist der weiteren Strategieentwicklung dienlich.

1 Die Initiative Forschungsdatenmanagement

Im Juli 2020 hat die interdisziplinäre Initiative „Forschungsdatenmanagement“ (FDM) der Berlin-Brandenburgischen Akademie der Wissenschaften (BBAW) die Arbeit aufgenommen. Sie stellt sich der Herausforderung, alle an der BBAW generierten Forschungsdaten zu ordnen, zu sichern, zu publizieren und langfristig verfügbar zu halten. Übergeordnetes Ziel ist die Entwicklung einer Forschungsdatenmanagementstrategie, die den komplexen Anforderungen der digitalen geisteswissenschaftlichen Forschungsdaten an der BBAW gerecht wird und dabei die Organisationsstruktur und Projektformen der Akademie berücksichtigt.

Viele der Daten, die an den Akademien erzeugt und bearbeitet werden, sind einzigartig und nicht oder nur mit großem Ressourcenaufwand reproduzierbar. Die Aufbewahrung und Zugänglichmachung der Forschungsdaten sind als Weg zur Erschließung, Sicherung und Vergegenwärtigung weltweiter kultureller Überlieferungen Auftrag des Akademienprogramms und damit von höchster Priorität. Eine besondere Herausforderung stellt dabei der Erhalt komplexer digitaler Ressourcen dar, die als Präsentations- und Analyseumgebungen der geisteswissenschaftlichen Daten entwickelt wurden.[1] Digitale Ressourcen sind ein Amalgam aus (Meta-)Daten, Datenbanken und Software, für deren langfristige Handhabung und Weitergabe bisher nur unzureichende Lösungen existieren.

Mit ihrer Arbeit trägt die Initiative Forschungsdatenmanagement zur Umsetzung der Selbstverpflichtung der BBAW bei, die Ergebnisse öffentlich geförderter Forschung frei und langfristig zugänglich zu machen, wie zuletzt im „Leitbild Open Science“[2] festgehalten. Dabei folgt die Initiative den FAIR-Prinzipien und legt hohen Wert auf die Verfügbarkeit und Nachnutzbarkeit der Daten und (soweit möglich) der Software. Als Ergebnis der Selbstevaluation durch RISE-DE[3] liegt eine Bewertung und Zielbestimmung für die Akademie vor, die als Grundlage zur Entwicklung der beschriebenen FDM-Strategie dient.

[1] vgl. Sahle und Kronenwett 2013.
[2] Berlin-Brandenburgische Akademie der Wissenschaften, „*Leitbild Open Science*“.
[3] Hartmann et al. 2019.

2 Vernetztes Vorgehen

Die Initiative knüpft insbesondere an die FDM-Arbeit des Teams von TELOTA-IT/DH[4] an, das durch die digitale Umsetzung und Beratung von über 50 Projekten an der BBAW umfangreiche Erfahrungen gesammelt hat und durch das sich die Akademie seit zwei Jahrzehnten aktiv an den Entwicklungen der Digital Humanities und des FDMs beteiligt.

Darüber hinaus arbeitet die Initiative eng mit weiteren Stakeholdern der Akademie zusammen, um alle Aspekte des Datenmanagements abzudecken. Die Initiative versteht FDM nicht als neuartige, losgelöste Aufgabe, sondern als Summe bestehender datenfokussierter Dienstleistungen sowie deren Ausbau, Konsolidierung und strukturierter Verflechtung.

Als Schnittstelle zwischen den WissenschaftlerInnen der BBAW und den Teams von TELOTA, der IT-Abteilung, der Bibliothek und dem Archiv fördert die Initiative die (frühzeitige) Einbindung der Infrastruktureinheiten in die Beratung und Erstellung von DMPs. Das gemeinsame Verständnis von Forschungsdatenmanagement als Teamaufgabe trägt zu einer transparenten Kommunikation innerhalb der Akademie bei und stellt die lückenlose Betreuung aller Vorhaben über den gesamten Datenlebenszyklus sicher. Der enge Kontakt mit den etablierten Abteilungen verbessert außerdem die Akzeptanz und Integration der Initiative sowie der durchgeführten und geplanten FDM-Maßnahmen.

Über den Horizont der BBAW hinaus steht die Initiative mit den weiteren Wissenschaftsakademien der Akademienunion im Austausch und stellt ihre Arbeitsergebnisse zur Nachnutzung innerhalb der FDM-Community und darüber hinaus bereit.

3 Proaktives Vorgehen

Forschungsdatenmanagement ist ein kontinuierlicher Prozess, der bereits vor der eigentlichen Forschung beginnt und zeitlich über den Abschluss der Projekte hinausgeht. Um die Entwicklung der Prozesse abzubilden wird ein Datenmanagementplan (DMP) idealerweise mit dem Projektantrag angefertigt und fortlaufend aktualisiert. Die Initiative FDM begleitet neue Vorhaben bei der Erstellung und Überarbeitung ihrer

[4] Berlin-Brandenburgische Akademie der Wissenschaften, „*TELOTA*“.

DMPs. Neben der Betreuung neu gestarteter BBAW-Projekte, liegt ein Schwerpunkt der Initiative auf der Unterstützung bereits laufender Projekte.

Die Initiative geht dafür proaktiv auf die Akademienvorhaben zu und stellt so eine systematische und umfassende Betreuung aller Projekte sicher. Ziel der jeweils ersten Konsultation ist es, durch strukturierte Interviews mittels der Software RDMO (Research Data Management Organiser)[5] gemeinsam mit den Akademienvorhaben und den weiteren beteiligten Stakeholdern Datenmanagementpläne zu erstellen. Die DMPs werden im Vorfeld der kollaborativen Bearbeitung durch die Initiative FDM vorausgefüllt. Sie bieten einen aktuellen und von allen Beteiligten geprüften Überblick zu den bisher produzierten und zukünftig noch zu sammelnden oder zu erhebenden Daten und dokumentiert die individuellen FDM-Bedarfe. Durch die Erstellung der DMPs wird sichergestellt, dass mögliche Hindernisse und Engpässe frühzeitig erkannt und adressiert werden können. Das gemeinsame Besprechen der Fragen und Antworten, sowie zahlreiche Hilfestellungen im Fragebogen, ermöglichen eine niedrigschwellige Auseinandersetzung mit dem Thema FDM. Der lokale Bezug der Fragen auf die bekannten Stakeholder erleichtert die Identifizierung und Adressierung notwendiger Schritte.

Auch laufende Forschungsvorhaben profitieren von der Erstellung eines DMPs und werden für Fragen des FDMs sensibilisiert. Gemeinsam können mögliche Handlungsfelder definiert und Verantwortlichkeiten festgelegt werden sowie zeitliche und ggf. nachträglich auch finanzielle Ressourcen eingeplant werden, um ausgewählte Forschungsdaten nach Projektende zu veröffentlichen.

Ergänzend bietet die Initiative offene Schulungen für die Mitarbeitenden der BBAW an, in denen Konzepte des Forschungsdatenmanagements vorgestellt und Good Practices, wie Regeln der Dateibenennung, Backup-Strategien oder die Verwendung deskriptiver Metadaten, vermittelt werden. Das Informationsangebot wird durch eine umfangreiche Website der Initiative komplettiert.[6]

[5] Leibniz-Institut für Astrophysik Potsdam, „*RDMO*“.
[6] Marciniak und Sutter, „*Initiative Forschungsdatenmanagement*“.

4 Zentrale Wissensbasis

Mit der Software RDMO stellt die Initiative ein zentrales Tool für die Erstellung und Verwaltung von Datenmanagementplänen bereit. Der lokal angepasste Fragenkatalog kann so kollaborativ und benutzerfreundlich über einen Internetbrowser ausgefüllt werden. Durch die Anpassung werden die AnsprechpartnerInnen an der BBAW direkt benannt und verlinkt. Auf die Geisteswissenschaften zugeschnittene Erläuterungstexte, Antwortmöglichkeiten und Links zu weiterführenden Informationen helfen den WissenschaftlerInnen dabei informierte Entscheidungen zu treffen und regen dazu an, sich selbstständig weiterzubilden.

Die DMPs sind an der BBAW nicht nur Grundlage für das Datenmanagement der einzelnen Vorhaben, sondern dienen auch als Wissensbasis für die IT-Abteilung, das Digital Humanities Team und die Bibliothek: Durch die systematische Erfassung aller Vorhaben wird eine neue Form der Querschnittsansicht möglich. Die gesammelten Informationen lassen beispielsweise Rückschlüsse auf Anforderungen an ein institutionelles FD-Repositorium zu. Die neu gewonnenen Erkenntnisse sind so für die weitere FDM-Strategieentwicklung dienlich. Über die API-Schnittstelle können Angaben automatisiert ausgelesen und weiterverwendet werden.

5 Ausblick

Nach Ablauf der zweijährigen Laufzeit möchte die Initiative einen nachhaltigen und langfristigen Workflow für das Management der Forschungsdaten an der BBAW bereitstellen und so eine verbesserte Nachhaltigkeit, Verfügbarkeit, Nachnutzbarkeit und Finanzierbarkeit digitaler Arbeitsprozesse und deren Ergebnisse sicherstellen sowie deren Einbindung in die entsprechenden NFDI-Konsortien gewährleisten.

Mit dem Vortrag möchte die Initiative FDM ihre Arbeit vorstellen und dabei insbesondere auf die Vorgehensweise zur DMP-Erstellung, die dafür eingesetzten Werkzeuge und BBAW-spezifische Anpassungen eingehen.

Bibliografie

Berlin-Brandenburgische Akademie der Wissenschaften. 2019. „Das Leitbild Open Science der Berlin-Brandenburgischen Akademie der Wissenschaften“. https://edoc.bbaw.de/frontdoor/index/index/docId/3136.

Berlin-Brandenburgische Akademie der Wissenschaften. „TELOTA - The Electronic Life Of The Academy“. Zugriff am 19.04.2021. https://www.bbaw.de/bbaw-digital/telota.

Hartmann, Niklas K., Boris Jacob, und Nadin Weiß. 2019. „RISE-DE – Referenzmodell für Strategieprozesse im institutionellen Forschungsdatenmanagement“, https://doi.org/10.5281/ZENODO.2549343.

Leibniz-Institut für Astrophysik Potsdam. „RDMO. Research Data Management Organiser.“ Zugriff am 19.04.2021. https://rdmorganiser.github.io/.

Marciniak, Katja, Paul Sutter. „Initiative Forschungsdatenmanagement.“ Zugriff am 19.04.2021. https://forschungsdatenmanagement.bbaw.de/de.

Sahle, Patrick, Simone Kronenwett. „Jenseits der Daten: Überlegungen zu Datenzentren für die Geisteswissenschaften am Beispiel des Kölner ‘Data Center for the Humanities’.“ *LIBREAS. Library Ideas* 9, no. 2 (2013). https://libreas.eu/ausgabe23/09sahle/.

Generische und fachspezifische Integration in der Forschungsinfrastruktur CLARIAH-DE

Buddenbohm, Stefan
sbudden[at]gwdg.de
SUB Göttingen, Deutschland
ORCID-iD: 0000-0002-3469-6101

Eckart, Thomas
teckart[at]informatik.uni-leipzig.de Universität Leipzig, Deutschland ORCID-iD: 0000-0001-6692-0713

Gradl, Tobias
tobias.gradl[at]uni-bamberg.de
Universität Bamberg, Deutschland
ORCID-iD: 0000-0002-1392-2464

Helfer, Felix
helfer[at]informatik.uni-leipzig.de
Universität Leipzig, Deutschland
ORCID-iD: 0000-0002-5739-2047

Jegan, Robin
robin.jegan[at]uni-bamberg.de
Universität Bamberg Deutschland
ORCID-iD: 0000-0002-0388-7220

Zusammenfassung. CLARIAH-DE führt die beiden etablierten Forschungsinfrastrukturen CLARIN-D und DARIAH-DE im Rahmen eines BMBF-geförderten Projekts zu einem gemeinsamen Angebot für geistes- und sozialwissenschaftlich Forschende zusammen. Mit seiner nachhaltigen Bereitstellung von Forschungsdaten, technischer Infrastruktur, digitalen Werkzeugen, virtuellen Forschungsumgebungen sowie Informations- und Schulungsmaterialien trägt CLARIAH-DE auch zur Nationalen Forschungsdateninfrastruktur (NFDI) bei. Die Integrationserfahrungen des Projektes liegen insbesondere auf organisatorischer und technischer Ebene, wobei sich dieser Vortrag auf die technische Integration und die Konsequenzen bezieht, die sich aus unterschiedlichen Sichtweisen auf Forschungsdaten und deren Nutzung in verteilten Umgebungen für integrative Maßnahmen ergeben. Problem- und

Erfolgsfälle werden anhand konkreter Beispiele illustriert. Als eine zentrale Schlussfolgerung ergibt sich: je fachspezifischer Anforderungen, und damit Komponenten, sind, desto komplexer eine Zusammenführung. Dabei sind zukünftige Kontexte ebenfalls zu berücksichtigen, bspw. die NFDI oder EOSC, und gefällte Entscheidungen gegenüber der Community plausibel zu begründen.

1 Einleitung

Die Zusammenführung zweier etablierter Forschungsinfrastrukturen zum gemeinsamen Angebot CLARIAH-DE verbindet sich mit Herausforderungen auf (mindestens) drei verschiedenen Ebenen:

- Strukturelle Aspekte, bspw. bestehende organisatorische Strukturen oder Verpflichtungen Dritten gegenüber, die den Gestaltungsspielraum begrenzen und die globale vs. lokale Perspektiven von Forschungsinfrastrukturen (und Forschungsdatenmanagement) zu berücksichtigen haben. Im Fall von CLARIAH-DE sind dies bspw. die ERIC- oder NFDI-Ebenen.
- Infrastrukturelle Aspekte, bspw. inkompatible technische Komponenten oder unterschiedliche Verfahrensweisen und Methoden beim Umgang mit Forschungsdaten, wobei generische und fachspezifische Ausprägungen unterschiedlich komplex umzusetzen sind.
- Kulturelle Aspekte, bspw. die Art und Weise wie miteinander kommuniziert wird und wie Entscheidungen getroffen werden.

Im Folgenden stehen die infrastrukturellen Aspekte im Vordergrund. Die durch CLARIAH-DE adressierten Infrastrukturbereiche lassen sich in drei Kategorien mit jeweils spezifischen Charakteristika (siehe Abbildung 1) einteilen.

	Integrationsebenen			
	Komplexität / Aufwand	Standardi-sierung	Sichtbarkeit / Nutzenerwartung	Externe Abhängigkeiten
Basisinfrastruktur-Komponenten	standardisierte Komponenten mit geringerem Aufwand zusammen-führbar (AAI, Monitoring)	häufig Einsatz standardi-sierter, marktgängiger Lösungen	gering, selbst substanzieller Nutzen für die Forschenden wird nicht unbedingt mit der Forschungs-infrastruktur verbunden	interne Abhängigkeiten gering; Abhängigkeit von externen Dienstleistern (“marktgängige” Standardl-ösungen) stärker
Anwendungen	hoch; je “forschungs-näher”, desto komplexer, da forschungs-spezifische Anforderungen	Standardi-sierung bzgl. Forschungs-daten möglich, Werkzeug-ebene jedoch häufig mit fach-spezifischen Anpassungen	hoch; Nutzende erhoffen sich eine Verbesserung durch neue Werkzeuge	fallbezogen; Werkzeuge der einen Fachdomäne auch in der anderen Fachdomäne einsetzbar oder anpassbar
(Forschungs-) Daten	hoch, sofern die Daten abhängig von spezifischen Forschungsfragen erhoben wurden	Standardi-sierung häufig in einem Zielkonflikt mit Granularität	hohe Erwartungshaltung insb. in SSH Forschungs-kontexten	i.d.R. abhängig von der Nutzung angepasster, fachspezifischer Werkzeuge oder Standards

Abb. 1. Infrastrukturelle Integrationsebenen der Forschungsinfrastruktur CLARIAH-DE.

Ausgehend von den spezifischen Charakteristika können Entscheidungen zur Zusammenführung, Anpassung oder zum Beibehalt ausgewählter Komponenten getroffen werden, die ein angemessenes Verhältnis von Aufwand und Nutzen aufweisen. Der Nutzen kann sich auf der Seite der Infrastrukturbetreibenden wie auch auf Seiten der Forschenden materialisieren. Die Aufwand-Nutzen-Abwägung muss in die Entscheidungsfindung einfließen, wobei für eine Forschungsinfrastruktur die Nutzenden – in unserem Fall Forschende aus den Geistes-, Kultur- und Sozialwissenschaften – im Vordergrund stehen. Bei der operativen Umsetzung wurde in der Regel auf den gleichen Standardprozess zurückgegriffen: (1) auf Basis einer konkreten Zielspezifikation wurde durch kleine Arbeitsgruppen mit geringem organisatorischem Overhead Lösungen entwickelt und (2) die

konkreten Implementationsergebnisse in regelmäßigen Treffen mit dem Projekt abgestimmt bzw. diskutiert.

Als eine zentrale Schlussfolgerung dieser Arbeiten ergibt sich: je fachspezifischer Anforderungen, und damit Komponenten, sind, desto komplexer eine Zusammenführung. Dabei sind zukünftige Kontexte ebenfalls zu berücksichtigen, bspw. die NFDI oder EOSC, und gefällte Entscheidungen gegenüber der Community plausibel zu begründen.

Ausgehend von diesen drei Integrationsebenen werden im Folgenden konkrete Beispiele vorgestellt.

2 Basisinfrastruktur

In CLARIAH-DE wurden unter anderem das Monitoring, die Authentifizierungs- und Authorisierungsinfrastrukturen (AAI) sowie der Helpdesk integriert. Für Basisdienste ist die Verwendung einer überschaubaren Anzahl von Standardwerkzeugen üblich. Abgesehen von einer besseren Usability für die Nutzenden (Einheitlichkeit), verringert dies den Aufwand technischer Integration und verschiebt Integrationsarbeit in das Konfigurationsmanagement. Die Querbezüge zu kulturellen und strukturellen Integrationsaspekten lassen sich allerdings auch hier nicht ignorieren. Deutlich wird dies zum Beispiel im Bereich Monitoring, wenn unterschiedliche Herangehensweise an das Thema Diensteverfügbarkeit in einer einheitlichen Plattform berücksichtigt werden müssen[1]. Im Gegensatz dazu sind geringere Aufwände zu erwarten, wenn Dienste hinsichtlich ihrer Struktur und Nutzung weitgehend identisch bzw. komplementär aufgebaut/konfiguriert sind[2]. Querbezüge, die direkt in den Forschungs- und Arbeitsbereich der Nutzenden wirken, bspw. mit Blick auf das Forschungsdatenmanagement, sind bei den in CLARIAH-DE integrierten Basisinfrastrukturkomponenten weniger relevant.

3 Anwendungen

Integrationsbemühungen auf Anwendungsebene erfordern einen differenzierten Blick auf die Komponenten und deren Einsatzradius: Während die Erwartungshaltung seitens der Nutzenden höher ist als bei

[1] Buddenheim, 2021.
[2] Sambale et al. 2019.

der Basisinfrastruktur, setzt eine erfolgreiche Integration ein gewisses Maß an Kompatibilität beziehungsweise Anpassbarkeit der jeweiligen Werkzeuge voraus, um diese auch für alternative Domänen oder Anwendungsfälle zu ertüchtigen. Ein erfolgreiches Beispiel in CLARIAH-DE ist u.a. die Integration des Language Resource Switchboards von CLARIN in das TextGrid Repository[3] (siehe Abbildung 2)[4], die im Hintergrund die Verknüpfung von anderen Basisinfrastrukturkomponenten – hier AAI – voraussetzt.

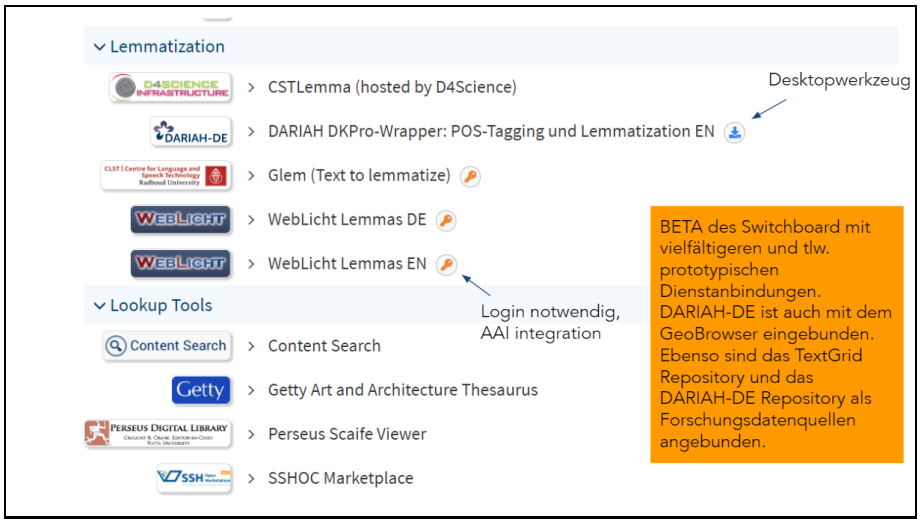


Abb. 2. Umsetzung auf Anwendungsebene: Verknüpfung von Werkzeugen und Diensten im Switchboard (Beta, https://beta-switchboard.clarin.eu/tools).

Auch für die Umsetzung des CLARIAH-DE Tutorial Finders (siehe Abbildung 3) finden aus beiden Initiativen hervorgegangene Komponenten eine integrative Anwendung. Mit Blick auf Medientypen, Zugriffsschnittstellen und konkrete Inhalte stark heterogene Lehr- und Lernmaterialien werden flexibel auf Basis der DARIAH-DE Datenföderationsarchitektur registriert, modelliert und integriert[5]. Sprachanalytische Funktionalität bspw. zur Geo-referenzierung und Textklassifikation wird durch die Integration von CLARIN Diensten bereitgestellt.

[3] Exemplarisch hier nachgewiesen: https://textgridrep.org/browse/vqmz_0 (Switchboard-Link in der Rubrik “Werkzeuge” aufrufen.).
[4] Weimer, 2021.
[5] Gradl, Jegan 2021.

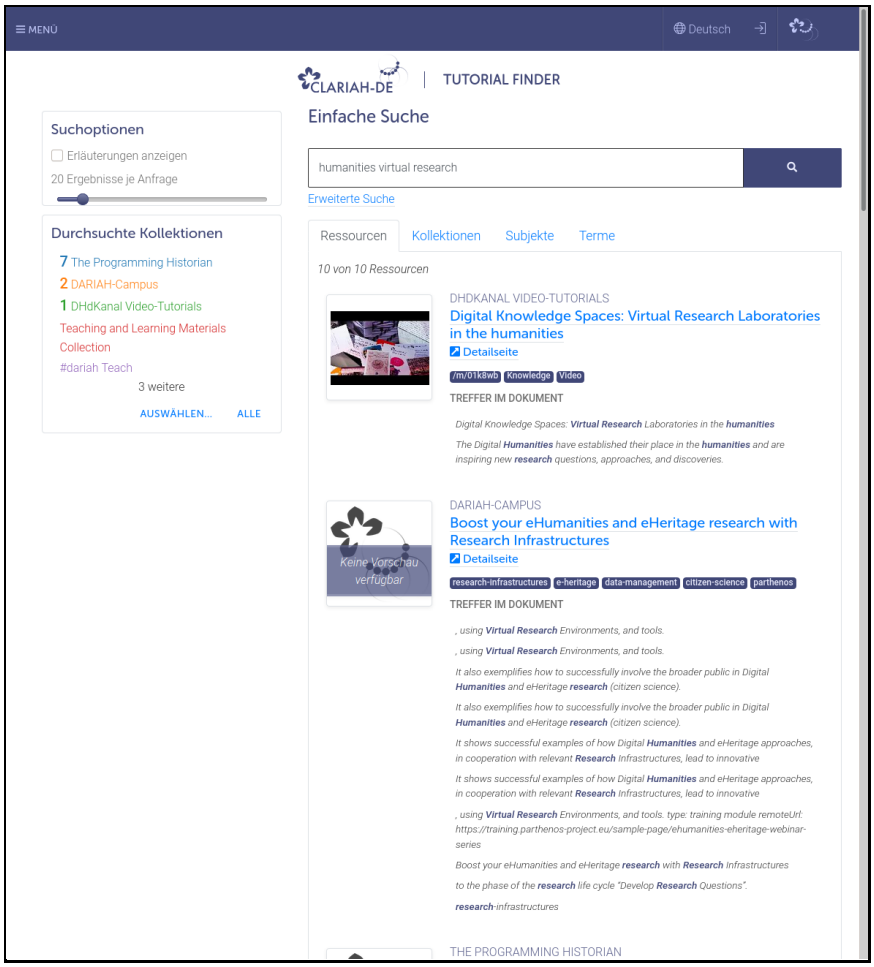


Abb. 3. Prototyp des CLARIAH-DE Tutorial Finder als integratives Anwendungsbeispiel.

4 Forschungsdaten

Der Bereich Forschungsdaten/Suchlogiken war in CLARIAH-DE mit vergleichsweise hohen Erwartungen durch die Nutzenden verbunden. Vereinfacht ist es der Wunsch nach einem zentralen Einstiegspunkt für die verschiedenen Suchräume von CLARIN und DARIAH, konkret:

- CLARIN: Virtual Language Observatory (VLO): Metadaten zur Beschreibungen von Text- und Sprachkorpora, lexikalischen Ressourcen, über 1,2 Mio. Einträge, Lucene-basiert; Federated Content Search (FCS): Schwerpunkt auf Volltexten und Korpora, 4.000 Kollektionen, FCS Query Language; Component Metadata

Initiative (CMDI) mit über 180 Metadaten-Schemata, Metadaten-Aufbereitung[6],

- DARIAH: Generic Search: 47 Kollektionen mit über 1,2 Mio. Einträgen, Kollektionen recht heterogener, vorwiegend textueller Forschungsdaten; DARIAH Collection Description Data Model (DCDDM), Elasticsearch (DCsimple für Facettierung),

ohne dabei deren Spezifika zu verlieren.

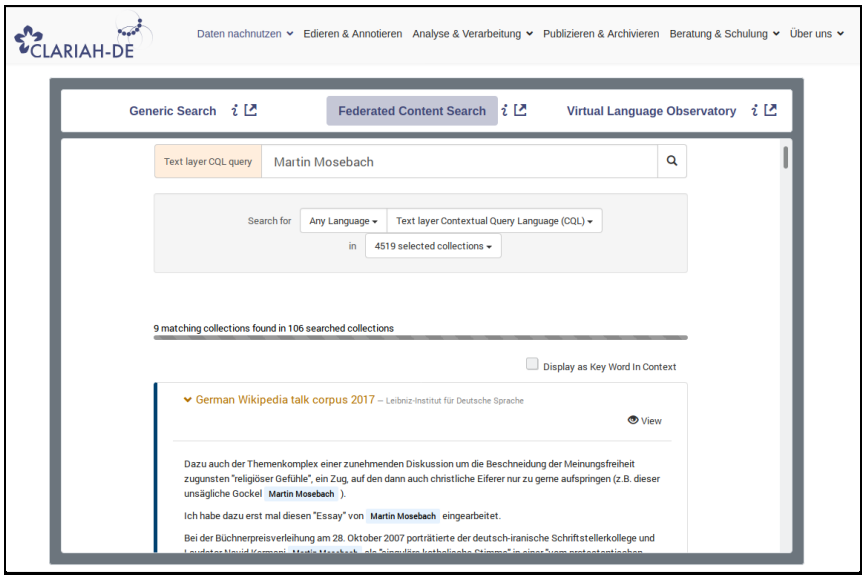


Abb. 4. Beispiel für ein eingebundenes CLARIN Zentrum, hier ein IDS-Korpus in der Federated Content Search; eingebettet in die integrierte CLARIAH-DE-Suche.

Der infrastrukturelle Rahmen dieser Suchen ist auf beiden Seiten teilweise vergleichbar aufgebaut, d.h. es gibt Komponenten für Mapping in unterschiedliche Schemata, Verfahren zur Registrierung von Kollektionen und Endpunkten sowie Anwendungen zur Verarbeitung oder Visualisierung von Forschungsdaten. Die tiefere Integration in die europäische Ebene (ERIC) bei CLARIN ist auffällig. Auf DARIAH-Seite liegt eine Herausforderung in der anderen Organisationsweise: anders als bei CLARIN sind die Forschungsdaten nicht in Zentren organisiert bzw. in der Regel sind Einzelforschende die Urheber.

[6] https://github.com/clarin-eric/VLO-mapping.

Folgende Erfahrungen stechen für den Bereich Forschungsdaten/Suchlogiken heraus[7]:

- vergleichsweise hoher Integrationsaufwand, da unterschiedliche Zielgruppen und Informationsbedürfnisse in einer funktionalen Verkapselung der Suchdienste mit jeweils unterschiedlichen inhaltlichen Schwerpunkten resultieren,
- Maßnahmen: einfache Angleichung verwendeter Style-Informationen bis hin zu komplexen und aufwendigen Anpassungen verwendeter Datenmodelle und Anfragesprachen fließen in die Aufwand-Nutzen-Abwägung ein,
- Nutzerperspektive muss im Zentrum stehen, so dass als Lösung auch eine "Benutzerschnittstelle als Fassade vor einem potenziell hochintegrierten Datenbestand" das Mittel der Wahl sein kann,
- integrierte Zugänglichkeit aller Bestände an einer Stelle – wie in der CLARIAH-DE-Suche (Abbildung 4) prototypisch gezeigt – ist wichtig, um Anschlussfähigkeit zu anderen Akteuren (bspw. EOSC oder NFDI) zu gewährleisten; ein derartiger Zugangspunkt funktioniert dabei als technische und "organisatorische" Einladung an andere Forschungs-infrastrukturen.

Bibliografie

Buddenbohm, S., Das Monitoring von CLARIAH-DE, *DHdBlog*, abgerufen 12.04.2021, https://dhd-blog.org/?p=14881.

Eckart, T., Gradl, T., Jegan, R., Margaretha, E., Werthmann, A., Helfer, F., Buddenbohm, S., CLARIAH-DE Cross-Service Search: Prospects and Benefits of Merging Subject-specific Services. *DARIAH-DE Working Papers Nr. 41*. Göttingen: DARIAH-DE, 2021. URN: urn:nbn:de:gbv:7-dariah-2021-1-9.

Gradl, T., Jegan, R., Nachnutzung Git-basierter Sammlungen im Rahmen der Infrastrukturdienste von CLARIAH-DE. *DARIAH DE Working Papers Nr. 42*. Göttingen: DARIAH-DE, 2021. URN: urn:nbn:de:gbv:7-dariah-2021-2-5.

Sambale, H., Hedeland, H., Pirinen, T., User Support for the Digital Humanities. *Selected Papers from the CLARIN Annual*

[7] Eckart et al. 2021.

Conference 2019. Linköping Electronic Conference Proceedings 172:14, s. 119-125. https://doi.org/10.3384/ecp2020172014.

Weimer, L., Die Einbindung externer Werkzeuge in das TextGrid Repository, *DHdBlog*, abgerufen 12.04.2021, https://dhd blog.org/?p=14999.

Inter- und transdisziplinäres Forschungsdatenmanagement in der Ost-, Ostmittel- und Südosteuropaforschung

Štanzel, Arnošt
arnost.stanzel[at]bsb-muenchen.de
Bayerische Staatsbibliothek, Deutschland
ORCID-iD: 0000-0002-4686-8185

Frank, Ingo
frank[at]ios-regensburg.de
IOS Regensburg, Deutschland

König, Sandra
sandra.koenig[at]leibniz-gwzo.de
GWZO Leipzig, Deutschland

Körfer, Anna-Lena
anna-lena.koerfer[at]herder-institut.de
Herder-Institut Marburg, Deutschland

Valena, Peter
peter.valena[at]collegium-carolinum.de
Collegium Carolinum, München, Deutschland

Zusammenfassung. Anhand der Erfahrungen aus dem Aufbau von *Ost-Data*, dem Forschungsdatendienst für die Ost-, Ostmittel- und Südosteuropaforschung, fokussiert der Beitrag auf die Bedeutung fachspezifischer Lösungen für das Forschungsdatenmanagement. Der Beitrag thematisiert insbesondere drei Herausforderungen, die beim Aufbau eines regionalwissenschaftlichen Forschungsdatendienstes eine Rolle spielten: 1) Der inter- und transdisziplinäre Charakter der Regionalforschung stellt hohe Anforderungen an eine adäquate Ansprache von Forschenden, die mit einer Vielzahl an Forschungsdatentypen aus unterschiedlichen wissenschaftlichen Disziplinen arbeiten. 2) Der Bedarf an angepassten Lösungen zeigt sich vor allem beim Blick auf die institutionelle Landschaft der Regionalforschung, die in Deutschland aus einzelnen Lehrstühlen und einer Vielzahl außeruniversitärer Einrichtungen besteht, die stark in ihrer Größe variieren. Die Diversität der Wissenschaftslandschaft bringt einerseits unterschiedliche Wissensstände in Bezug auf das Forschungsdatenmanagement hervor und stellt das Projektteam andererseits vor die

Herausforderung, diese heterogenen Zugänge durch entsprechende Angebote aufzufangen, zu bereichern. 3) In *OstData* wird ein auf DataCite basierendes Metadatenschema mit regionalwissenschaftlichem Schwerpunkt entwickelt. Anhand von *User Stories* zu Projektbeginn hat sich gezeigt, dass in der Regionalwissenschaft ein hoher Bedarf besteht, Forschungsdaten dem regionalwissenschaftlichen Schwerpunkt entsprechend kommentieren, erschließen und publizieren zu können (u. a. ausgedehnte Sacherschließung, erweiterte Angaben zur Datenprovenienz). Eine solches regionalwissenschaftliches Metadatenschema muss einerseits auf die Bedarfe der Community hinwirken, andererseits generische Lösungen implementieren wie Normvokabularien und Fachklassifikationen, was zu gesteigerten Herausforderungen im Mapping mit anderen Datenschemata führt (z. B. für den Metadatenaustausch).

1 Forschungsdatenmanagement in der Regionalforschung

In den Area Studies werden Regionen aus geistes-, sozial- und wirtschaftswissenschaftlicher Perspektive untersucht. Disziplinär umfasst dies insbesondere die Geschichtswissenschaften, aber auch politikwissenschaftliche und wirtschaftswissenschaftliche Ansätze, ergänzt um vielfältige Zugänge aus dem Kanon der sogenannten „Kleinen Fächer". Passend zum Tagungsmotto „Mapping the Landscape" soll im vorgeschlagenen Beitrag vorgestellt werden, wie die Community der Ost-, Ostmittel- und Südosteuropaforschung im Rahmen des Forschungsdatendienstes *OstData* kartiert wurde, und welche Erfahrungen bei der Einführung des Forschungsdatenmanagements gesammelt wurden. Aufgrund der Besonderheiten der Community sind wir überzeugt, wertvolle Beiträge zu den von der Tagung aufgeworfenen Fragen machen zu können, insbesondere bezüglich der Gegenüberstellung von fachspezifischen und generischen Lösungen. Anhand der folgenden drei Besonderheiten unserer Fachcommunity sollen Lösungsansätze zum Forschungsdatenmanagement in einer interdisziplinären, geisteswissenschaftlich geprägten Regionalwissenschaft präsentiert werden und darüber hinaus *lessons learned* für Forschungsdatenmanagement insgesamt gegeben werden. Inter- und Transdisziplinarität ist prägend für Area Studies, die ein gemeinsames Forschungsinteresse teilen, nämlich Forschung zu einer konkreten Region, die aus unterschiedlichen methodischen, disziplinären und theoretischen Ansätzen untersucht wird. Die Ost-, Ostmittel- und Südosteuropaforschung ist zudem durch eine relativ diverse institutionelle Ausformung sowie durch eine überdurchschnittlich hohe Zahl außeruniversitärer Forschungseinrichtungen unterschiedlicher Größe

geprägt. Die entsprechende Regionalwissenschaft zeichnet sich in hohem Maß durch Mehrsprachigkeit aus, die etwa bei der Nutzung von Normvokabularien zur Beschreibung von Daten und Metadaten zu großen Herausforderungen führt.

1.1 Forschungsdatenmanagement für die Ost-, Ostmittel- und Südosteuropaforschung

Wie Forschungsdatenmanagement für eine spezifische Community aussehen kann, soll anhand der Projekterfahrungen von *OstData* dargestellt werden. Wie die Forschung zur Region selbst ist auch dieser zwischen lokalen und globalen Forschungsdatenangeboten und -infrastrukturen verortet. Getragen von einschlägigen, außeruniversitären Institutionen (Bayerische Staatsbibliothek (BSB), Collegium Carolinum – Forschungsinstitut für die Geschichte Tschechiens und der Slowakei, Leibniz-Institut für Geschichte und Kultur des östlichen Europas (GWZO), Herder-Institut für historische Ostmitteleuropaforschung – Institut der Leibniz-Gemeinschaft, Leibniz-Institut für Ost- und Südosteuropaforschung) wird ein Forschungsdatenrepositorium für die deutsche Fachcommunity aufgebaut. Dieses wird flankiert von Aktivitäten, um eine Publikationskultur für Forschungsdaten zu etablieren. Dazu gehört, grundlegendes Wissen zum Forschungsdatenmanagement mittels Schulungen zu vermitteln und Materialien zum Selbststudium (Best Practices, Handreichungen, Informationsmaterial) anzubieten. Gleichzeitig wurden durch Projektmitarbeiter Gespräche mit Forschenden, Zeitschriftenredakteur*innen und Institutsleitungen geführt, um Vorbehalte gegenüber der Veröffentlichung von Forschungsdaten zu identifizieren und entsprechend begegnen zu können. Auf Basis der alltäglichen Projekterfahrungen wird die Bedeutung fachspezifischer Perspektiven auf geisteswissenschaftliches Forschungsdatenmanagement hervorgehoben. Neben allgemeinen Herausforderungen in den Area Studies wie Mehrsprachigkeit oder der Arbeit in zum Teil autoritären politischen Systemen zeigt sich vor allem, dass für einzelne Disziplinen passgenaue Angebote und Lösungen entstehen müssen. Die Sprachwissenschaften sind beispielsweise bei der Nutzung von Daten weiter fortgeschritten als Historiker*innen. Noch deutlicher zu Tage treten die Unterschiede mit Blick auf die Sozial- und insbesondere Politikwissenschaften, in welchen bereits ein konkreter Datenbegriff etabliert ist, etwas, woran es vielen geisteswissenschaftlichen Disziplinen bisher mangelt. Im Verlauf verschiedener Schulungen wurden kontinuierlich Erfahrungen zu Erwartungen und Wissensständen Forschender aus einzelnen

Disziplinen gesammelt, welche wiederum genutzt wurden, um die Präsentationen anzupassen und zu verbessern. Gleichzeitig kann die interdisziplinäre Auseinandersetzung und der Transfer von Begriffen fruchtbar für alle Disziplinen sein und interdisziplinäre Nachnutzungsperspektiven von Forschungsdaten aufzeigen: Was der*dem Historiker*in Oral History ist, ist der*dem Politikwissenschaftler*in ein semistrukturiertes Interview. Um dieses Potential zu heben, müssen Forschungsdaten in einem ersten Schritt überhaupt erst veröffentlicht werden, um einen inter- und transdisziplinären Austausch zu befördern, damit Forschende über Fachgrenzen hinweg von einer Publikation profitieren.

1.2 Mapping the landscape: eine diverse institutionelle Landschaft mit spezifischen Anforderungen an Infrastruktur und Beratung

Eine weitere Besonderheit der Ost-, Ostmittel- und Südosteuropaforschung liegt darin, dass sie relativ stark an außeruniversitären Institutionen erfolgt, die von großen Leibniz-Einrichtungen mit eigener Infrastruktur und Personal für das Forschungsdatenmanagement bis hin zu kleinen Forschungseinrichtungen mit wenigen Mitarbeiter*innen reichen. *OstData* muss daher auf unterschiedliche Bedingungen und Voraussetzungen in der Community eingehen. So gibt es auf der einen Seite institutseigene Forschungsdatenmanager*innen, auf der anderen Seite Einrichtungen, die zunächst einen großen Bedarf an einführenden Schulungsangeboten zu Grundbegriffen des Datenmanagements haben. Hier hat es sich als vorteilhaft erwiesen, Schulungsangebote für Forschende im Vorfeld intensiv mit an den jeweiligen Einrichtungen tätigen Personen vorzubereiten und beispielsweise mit den zuständigen Forschungsdatenmanager*innen von Instituten oder Universitäten, Bibliothekar*innen, Projektleiter*innen oder Geschäftsführer*innen Wissensstände, Erwartungen und Bedarf zu besprechen. Ein nicht zu unterschätzender Faktor ist auch die enge Zusammenarbeit mit dem Fachinformationsdienst Ost-, Ostmittel- und Südosteuropa, in welchem aufgrund seiner langjährigen Arbeit und Kommunikation mit der Fachcommunity eine sehr gute Einschätzung über Wissensstände und Bedarfe der Forschenden verfügt und zudem über die entsprechenden Kontakte zu diesen verfügt.
OstData hat zudem die Funktion, die Community als Ganzes gegenüber nationalen und supranationalen Forschungsdateninfrastrukturen wie den sich konstituierenden NFDI-Konsortien oder der European Open Science Cloud zu sensibilisieren und als Aggregator zu vertreten.

1.3 Spezifische Lösungsansätze: Das projektspezifische Metadatenschema

Einen tieferen Einblick in die individuellen Praktiken, Strukturen und Lösungsansätze von *OstData* gibt die Vorstellung des Metadatenschemas, das den Bedarfen der Community entsprechend um fachspezifische Zonen zur Sacherschließung, Methodik und Datenprovenienz ergänzt wurde. *OstData* hat sich dazu entschieden bestehende Metadatenstandards wie Dublin Core oder DataCite nicht Eins-zu-eins zu übernehmen, sondern aufgrund von User Stories aus der Community erweiterte Fachzonen auszuweisen. Die ausführlichere Beschreibung von Forschungsdaten soll Akzeptanzfaktoren schaffen, damit Forschungsdaten nicht nur im Rahmen der regionalwissenschaftlichen Area Studies nachgenutzt werden, sondern auch in anderen Forschungskontexten Anwendung finden. Generell soll eine detaillierte Auszeichnung von Forschungsdaten durch Metadaten in geisteswissenschaftlichen Communities zu einem tieferen Verständnis im Umgang mit Daten führen (*Data Literacy*). Auch wenn es sich beim projektspezifischen Metadatenschema um ein stark erweitertes Application Profile des DataCite-Schemas handelt, ist sich das Projektteam um die problematischen Aspekte eines solchen Entwicklungsprozesses bewusst:

- Ob und wie werden Forschende von den umfangreichen Datendokumentationsmöglichkeiten Gebrauch machen? In welchem Umfang benötigen sie hierfür Anleitung und Betreuung?
- Inwiefern kann Im- und Export von Metadaten in und aus anderen Schemata durch entsprechende Mappings möglichst umfangreich sichergestellt werden?
- Trägt die Verwendung allgemeiner Normvokabulare und Klassifikationen wie der GND und der DDC sowie fachspezifischer Ansätze wirklich dazu bei, die Interdisziplinarität der Ost-, Ostmittel- und Südosteuropaforschung widerzuspiegeln und wird diese seitens von Forschenden auch bei der Recherche nach Forschungsdaten gewinnbringend eingesetzt?

1.4 Schluss

Der Beitrag zeigt den Wert fachspezifischer Lösungen auf, da diese besser von der Community angenommen werden und so mehr Daten zur Verfügung gestellt werden und macht gleichzeitig deutlich, dass generische Lösungen eine zentrale Rolle in der Forschungsdatenlandschaft spielen, da diese u. a. Grundlagen liefern, um Metadaten inter- und transdisziplinär vergleichbar zu halten und Forschungsdaten nachnutzbar zu machen.

Modellierung als Nachhaltigkeitsstrategie für Forschungsanwendungen

Fischer, Anna
anna.fischer[at]uni-koeln.de
Data Center for the Humanities (DCH), Universität zu Köln, Deutschland

Schildkamp, Philip
philip.schildkamp[at]uni-koeln.de
Data Center for the Humanities (DCH), Universität zu Köln, Deutschland

Demmer, Dennis
dennis.demmer[at]uni-koeln.de
Data Center for the Humanities (DCH), Universität zu Köln, Deutschland

Neuefeind, Claes
c.neuefeind[at]uni-koeln.de
Cologne Center for eHumanities (CCeH), Universität zu Köln, Deutschland

Mathiak, Brigitte
bmathiak[at]uni-koeln.de
Data Center for the Humanities (DCH), Universität zu Köln, Deutschland

Zusammenfassung. Die Entwicklung und Bereitstellung von Software-Anwendungen wie z. B. Recherche-Datenbanken, interaktiven Visualisierungen oder digitalen Editionen ist zu einem integralen Bestandteil geisteswissenschaftlicher Forschungspraxis geworden. Um den wissenschaftlichen Anforderungen an die Nachvollziehbarkeit und Reproduzierbarkeit von Forschungsergebnissen gerecht zu werden, stellt sich die Frage, welche Strategien sich eignen, um Anwendungen in den (digitalen) Geisteswissenschaften auf nachhaltige Weise konzipieren, implementieren, bereitstellen und vor fremden Eingriffen schützen zu können. Vor diesem Hintergrund erforscht das DFG-geförderte Kooperationsprojekt *SustainLife* seit 2018 die Übertragbarkeit von Cloud-Deployment-Verfahren auf die Gegebenheiten und Bedürfnisse von DH-Forschungsanwendungen, um diese langfristig erhalten und bereitstellen

zu können. Die Modellierungssprache *TOSCA (Topology Orchestration Specification for Cloud Applications)* sowie die Open-Source-Implementierung *OpenTOSCA* nutzen wir dabei als Methode, um existierende Nachhaltigkeitsansätze für Forschungsanwendungen miteinander zu verbinden. In unserem Beitrag möchten wir aufzeigen, dass TOSCA die Diversität von Forschungsanwendungen in den DH auffangen kann und viele Ansätze für deren ressourceneffizienten Erhalt und Betrieb bietet. Weiterhin zeigen wir Anknüpfungspunkte für zukünftige Forschungsarbeiten auf, die den Umgang mit dem OpenTOSCA-Ökosystem und dessen Bereitstellung vereinfachen und verbessern können.

1 Einleitung

Die Entwicklung und Bereitstellung von Software-Anwendungen wie z. B. Recherche-Datenbanken, interaktiven Visualisierungen oder digitalen Editionen ist zu einem integralen Bestandteil geisteswissenschaftlicher Forschungspraxis geworden. Um den wissenschaftlichen Anforderungen an die Nachvollziehbarkeit und Reproduzierbarkeit von Forschungsergebnissen gerecht zu werden, stellt sich die Frage, welche Strategien sich eignen, um Anwendungen in den (digitalen) Geisteswissenschaften auf nachhaltige Weise konzipieren, implementieren, bereitstellen und vor fremden Eingriffen schützen zu können. Vor diesem Hintergrund erforscht das DFG-geförderte Kooperationsprojekt *SustainLife* seit 2018 die Übertragbarkeit von Cloud-Deployment-Verfahren auf die Gegebenheiten und Bedürfnisse von DH-Forschungsanwendungen, um diese langfristig erhalten und bereitstellen zu können.[1]

2 Anwendungen als Publikationsformat

Die an Anwendungen gekoppelten Forschungsergebnisse zu erhalten bedeutet mehr als die Archivierung der enthaltenen Daten und des Programmcodes. Der Verlust der Funktionen und der Entwicklungshistorie einer Anwendung führt unweigerlich auch zum Verlust möglicher Erkenntnisse, die erst aus der Kopplung zwischen den Daten und der Funktionalität der Anwendung entstehen. Somit ist es notwendig, Forschungsanwendungen als eigenes Publikationsformat zu betrachten, dessen Nachvollziehbarkeit und Nachnutzbarkeit Grundvoraussetzungen für gute wissenschaftliche Praxis sind.

[1] Neuefeind u. a. 2018.

Die Herausforderung für den nachhaltigen Betrieb von Forschungsanwendungen besteht jedoch darin, dass diese in ihrer Funktionsweise von ihrer digitalen Umgebung abhängig sind. Dieses „digitale Ökosystem" einer Anwendung kann u. a. das Betriebssystem, einen Webserver oder eine Datenbankanbindung umfassen und ist, aufgrund technologischer Fortschritte und *deprecations*, unweigerlich Veränderungsprozessen unterworfen. Damit Anwendungen zugänglich und betriebsfähig bleiben, muss also ihre stetige Anpassung an das sich weiterentwickelnde Ökosystem sichergestellt werden, was für zeitlich begrenzte Forschungsprojekte eine enorme logistische, technische und nicht zuletzt finanzielle Herausforderung darstellt.

Technisch orientierte Strategien und Standards zur nachhaltigen Sicherung von Forschungsanwendungen in den DH sind bisher stark institutionell geprägt und entsprechend heterogen. Zudem ist ein Rückgriff auf etablierte Wartungsstrategien aus dem Bereich der Informatik und Software-Entwicklung nur bedingt umsetzbar für die DH, weil die nötige Expertise sowie infrastrukturelle Voraussetzungen nicht immer gegeben sind. Weiterhin sind die industrietauglichen und technisch ausgereiften Lösungsansätze wie z. B. Emulation, vollständige Virtualisierung und Containerisierung entweder herstellerabhängig oder an bestimmte Hardware gebunden und somit als Nachhaltigkeitsstrategie für DH-Forschungsanwendungen nicht ausreichend interoperabel.

Innerhalb der DH existieren unterschiedliche Ansätze diesem Problem zu begegnen, darunter: (1) finanzielle und personelle Mittel in die Wartung (und ggf. Reimplementierung) der Anwendung investieren[2], (2) die Anwendung stilllegen und nur noch als statische Momentaufnahme bereitstellen, wobei der Quellcode und die Daten archiviert werden[3], (3) den Technologie-Stack eingrenzen, d.h. die Verwendung bestimmter Komponenten vorschreiben oder verbieten[4] sowie (4) die Anwendung virtualisieren oder containerisieren und somit technisch einkapseln[5].

Die Nachhaltigkeitsstrategie des Projekts SustainLife basiert auf dem OASIS-Standard *TOSCA (Topology Orchestration Specification for Cloud Applications)*. TOSCA ist eine Modellierungssprache, mit Hilfe derer man (Cloud-) Anwendungen auf eine hersteller- und technologieunabhängige Weise beschreiben, bereitstellen und

[2] Smithies u. a. 2019.
[3] Arneil, Martin, und Newton 2019.
[4] Arneil, Martin & Newton 2019.
[5] Smithies u. a. 2019.

verwalten kann[6,7]. Sie dient dabei als Abstraktion weit verbreiteter Deployment-Technologien wie Docker oder OpenStack.

3 Unsere TOSCA-basierte Nachhaltigkeitsstrategie

In unserem explorativen Forschungsprojekt haben wir zunächst konkrete Bedarfe für Forschungsanwendungen in den Digital Humanities erhoben, indem wir (1) lokale Bedarfe identifizierten und abbildeten[8], (2) quantitative Umfragen, u. a. auf einem Workshop zur Nachhaltigkeit Digitaler Editionen durchführten[9], (3) die technische Struktur gängiger Forschungsanwendungen qualitativ analysierten[10], und darauf basierend (4) repräsentative Use-Case-Anwendungen zur Modellierung in TOSCA auswählten. Unsere Erhebungen zeigten, dass Anwendungen in den DH sowohl methodisch als auch technologisch von starker Heterogenität gekennzeichnet sind. Andererseits konnten wir aufzeigen, dass bestimmte Technologien (wie bspw. der LAMP-Stack oder X-Technologien) überdurchschnittlich häufig zum Einsatz kommen[11].
Die zur Modellierung ausgewählten Use Cases, wie z. B. die Musical Competitions Database[12], wurden daran orientiert und mithilfe des Open-Source-Ökosystems *OpenTOSCA*[13] umgesetzt. OpenTOSCA, entwickelt vom IAAS (Institut für Architektur von Anwendungssystemen, Universität Stuttgart), wurde im Laufe unseres Forschungsprojekts stetig an die sich ergebenden methodischen und technologischen Bedarfe der DH angepasst. Das *DCH* (*Data Center for the Humanities*, Universität zu Köln) verfügt nun über eine OpenTOSCA-Instanz sowie ein Repositorium mit bereits in TOSCA modellierten Anwendungsbausteinen und Anwendungen. Während die Instanz bisher nur im lokalen Netzwerk zur Verfügung steht, ist das Repositorium[14] öffentlich zugänglich.
Zur (Langzeit-) Archivierung laufender Anwendungen inklusive ihrer Abhängigkeiten verwenden wir das im TOSCA-Standard definierte

[6] „Topology and Orchestration Specification for Cloud Applications Version 1.0“ 2013.
[7] „TOSCA Simple Profile in YAML Version 1.3“ 2020.
[8] Neuefeind, Schildkamp, Mathiak, Marčić, u. a. 2019.
[9] Neuefeind, Schildkamp, Mathiak, Harzenetter, u. a. 2019.
[10] Helling, Schildkamp & Mathiak 2019.
[11] Neuefeind, Schildkamp, Mathiak, Harzenetter, u. a. 2019.
[12] Neuefeind, Schildkamp, Mathiak, Marčić, u. a. 2019.
[13] Breitenbücher u. a. 2016.
[14] Das Repositorium ist unter der URL https://gitlab.cceh.uni-koeln.de/sustainlife/tosca-definitions-dch abrufbar.

Paketierungs- und Austauschformat *CSAR* (Cloud Service Archive). CSARs enthalten entweder nur die Anwendungslogik oder sind vollständig in sich geschlossen und bündeln alle Komponenten, vom Betriebssystem bis zur eigentlichen Programmlogik, die für die Bereitstellung der enthaltenen Anwendung erforderlich sind. Beim Einfrieren einer Anwendung enthält das resultierende CSAR nicht nur die Komponenten der Anwendung, sondern auch ihren internen Zustand zum Zeitpunkt des Einfrierens[15].
Unser TOSCA-basierter Nachhaltigkeitsansatz ermöglicht es, Anwendungen von ihren Ökosystemen und somit von technologischen *deprecations* zu entkoppeln sowie Komponenten unabhängig voneinander und von der jeweiligen Anwendung zu aktualisieren. Hinzu kommt, dass Nutzer*innen die Möglichkeit haben, bereits vorhandene Komponenten wiederzuverwenden, sodass Synergieeffekte zwischen modellierten Anwendungen mit gleichen Komponenten entstehen, was die Instandhaltungskosten einer Anwendung reduziert. Denn je mehr Komponenten bereits vorhanden sind, umso weniger aufwendig ist es, neue Anwendungen zu modellieren oder bereits modellierte Anwendungen aktuell zu halten[16]. Die stetige Weiterentwicklung von OpenTOSCA durch das IAAS stellt außerdem neue Funktionalitäten zur Verfügung, u. a. zur Versionierung von Anwendungen[17] oder zum „Einfrieren“ und „Wiederauftauen“ von Anwendungen[18].

4 Forschungsdesiderate

Basierend auf den Erfahrungen und Ergebnissen des Projektes sind wir der Auffassung, dass sich TOSCA als ein institutionsübergreifender und community-weiter Nachhaltigkeitsstandard für Forschungsanwendungen in den DH in besonderer Weise eignet, da TOSCA sowohl die Heterogenität der Anwendungen auffangen kann als auch deren ressourceneffizienten Erhalt und Betrieb ermöglicht. Es hat sich auch gezeigt, dass es nicht möglich ist, heterogene Forschungsanwendungen in den DH mit einer umfassenden Modellierungsstrategie aufzufangen ohne einen gewissen Grad an Komplexität einzuführen. Somit sollten weitere Arbeiten zu TOSCA als Nachhaltigkeitsstrategie darauf abzielen, die Nutzerfreundlichkeit weiterhin zu erhöhen und eine automatisiert unterstützende

[15] Harzenetter u. a. 2020.
[16] Schildkamp u. a. 2020.
[17] Harzenetter 2018.
[18] Harzenetter u. a. 2020.

Benutzerführung zu ermöglichen. Ziel sollte es stets sein, dass Schwierigkeitsgrad und Aufwand des Modellierungsprozesses in Relation zur Komplexität der Anwendung stehen, also skalieren. Um eine community-weite Strategie zu etablieren, ist es wünschenswert, Bedienungsanleitungen, Best Practices und ein die Modellierung anleitendes Assistenzsystem zu schaffen.
Nicht zuletzt hängt die Anwendbarkeit eines technologieorientierten Nachhaltigkeitsstandards nicht nur von den technologischen und methodologischen Gegebenheiten eines Projektes ab, sondern auch von dessen finanziellen, logistischen und politischen Umständen[19]. TOSCA als Strategie oder Standard ersetzt also nicht den Bedarf an individueller Beratung, den Forschungsdatenzentren bieten, sondern sollte den Beratungsprozess ergänzen und verstärken.

Bibliografie

Arneil, Stewart, Holmes Martin, und Greg Newton. „Project Endings: Early Impressions From Our Recent Survey On Project Longevity In DH“. In *Book of Abstracts of the 30th Annual Conference on Digital Humanities (DH2019)*. Utrecht, 2019. https://dev.clariah.nl/files/dh2019/boa/0891.html.

Breitenbücher, Uwe, Christian Endres, Kálmán Képes, Oliver Kopp, Frank Leymann, Sebastian Wagner, Johannes Wettinger, und Michael Zimmermann. „The OpenTOSCA Ecosystem – Concepts and Tools“. In *European Space project on Smart Systems, Big Data, Future Internet - Towards Serving the Grand Societal Challenges*, 112–30. Rome, Italy: SCITEPRESS - Science and Technology Publications, 2016. https://doi.org/10.5220/0007903201120130.

Harzenetter, Lukas. „Versioning of Applications Modeled in TOSCA“. Masterarbeit, Universität Stuttgart, 2018. http://dx.doi.org/10.18419/opus-9710.

Harzenetter, Lukas, Uwe Breitenbücher, Kálmán Képes, und Frank Leymann. „Freezing and Defrosting Cloud Applications:

[19] Smithies u. a. 2019.

Automated Saving and Restoring of Running Applications“. *SICS Software-Intensive Cyber-Physical Systems* 35, Nr. 1–2 (August 2020): 101–14. https://doi.org/10.1007/s00450-019-00415-8.

Helling, Patrick, Philip Schildkamp, und Brigitte Mathiak. „Nachhaltigkeit von Forschungsdateninfrastrukturen am Beispiel von Digitalen Editionen. Was steht noch in 2039?“. In *Postersession of the 20th Annual DINI Conference (DINI 2019)*. Osnabrück, 2019. https://dini.de/fileadmin/jahrestagungen/2018/DINI_Jahrestagung_2019_Helling.pdf.

Neuefeind, Claes, Lukas Harzenetter, Philip Schildkamp, Uwe Breitenbücher, Brigitte Mathiak, Johanna Barzen, und Frank Leymann. „The SustainLife Project. Living Systems in Digital Humanities“. In *IBM Research Report – Papers from the12th Advanced Summer School On Service-Oriented Computing (SummerSOC'18)*, 101–12. Hersonissos, 2018. https://dominoweb.draco.res.ibm.com/reports/RC25681.pdf.

Neuefeind, Claes, Philip Schildkamp, Brigitte Mathiak, Lukas Harzenetter, Johanna Barzen, Uwe Breitenbücher, und Frank Leymann. „Technologienutzung im Kontext Digitaler Editionen. Eine Landschaftsvermessung“. In *Book of Abstracts zur 7. Tagung des Verbands Digital Humanities im deutschsprachigen Raum (DHd 2019)*. Frankfurt, Mainz, 2019. https://doi.org/10.5281/zenodo.2596094.

Neuefeind, Claes, Philip Schildkamp, Brigitte Mathiak, Aleksander Marčić, Frank Hentschel, Lukas Harzenetter, Uwe Breitenbücher, Johanna Barzen, und Frank Leymann. „Sustaining the Musical Competitions Database: a TOSCA-based Approach to Application Preservation in the Digital Humanities“. In *Book of Abstracts of the 30th Annual Conference on Digital Humanities (DH 2019)*.

Utrecht: DataverseNL, 2019. https://doi.org/10.34894/XMZPBZ.

Schildkamp, Philip, Lukas Harzenetter, Frank Leymann, Brigitte Mathiak, Claes Neuefeind, Uwe Breitenbücher, und Anna Fischer. „Workshop on Modelling and Maintaining Research Applications in TOSCA“. In *Book of Abstracts of the 31th Annual Conference on Digital Humanities (DH 2020)*. Ottawa (virtuelle Konferenz), 2020. https://dh2020.adho.org/wp-content/uploads/2020/07/120_WorkshoponModellingandMaintainingResearchApplicationsinTOSCA.html.

Smithies, James, Carina Westling, Anna-Maria Sichani, Pam Mellen, und Arianna Ciula. „Managing 100 Digital Humanities Projects: Digital Scholarship & Archiving in King’s Digital Lab“. *Digital Humanities Quarterly* 13, Nr. 1 (26. April 2019). http://www.digitalhumanities.org/dhq/vol/13/1/000411/000411.html.

„Topology and Orchestration Specification for Cloud Applications Version 1.0“. Zugegriffen 14. Mai 2021. http://docs.oasis-open.org/tosca/TOSCA/v1.0/os/TOSCA-v1.0-os.html.

„TOSCA Simple Profile in YAML Version 1.3“. Zugegriffen 14. Mai 2021. https://docs.oasis-open.org/tosca/TOSCA-Simple-Profile-YAML/v1.3/os/TOSCA-Simple-Profile-YAML-v1.3-os.html.

Vom Speziellen ins Generische und wieder zurück

Wie fachspezifisch muss ein FDM-Zentrum sein?

Wörner, Kai
kai.woerner[at]uni-hamburg.de
Universität Hamburg, Deutschland
ORCID-iD: 0000-0001-8939-4437

Zusammenfassung. Das Zentrum für nachhaltiges Forschungsdatenmanagement an der Universität Hamburg ist die zentrale Anlaufstelle zu allen Fragen des FDMs und der Forschungsinformation für alle Fakultäten der Universität. Es ist personell und organisatorisch aus dem Projekt „Geisteswissenschaftliche Infrastruktur für Nachhaltigkeit“ (gwin) hervorgegangen, das ein entsprechendes Angebot (nur für den Bereich des FDMs) für die geisteswissenschaftliche Fakultät konzipiert und etabliert hat. Dort konnten aus einer Erhebung unter den Professoren, den Kontakten zu den Geisteswissenschaftlern und der konkreten Arbeit – insbesondere in der Kuration von Forschungsanwendungen – Bedarfe und Anforderungen identifiziert werden, die für die geisteswissenschaftliche Domäne typisch zu sein, und die sich in domänenspezifische Workflows und Best Practices übersetzen ließen. Die Erweiterung des Fokus auf weitere Fakultäten und Fächer und die damit einhergehende zusätzliche Menge an zu kuratierenden Daten und Anwendungen hat uns dazu gezwungen, stärker nach Gemeinsamkeiten als nach Spezialitäten in der Struktur der Daten unterschiedlicher Fächer zu suchen. Insbesondere bei der reinen Datenkuration, also der Aufbereitung von Forschungsdaten für die langfristige Speicherung und Nachnutzung, sind wir auch fündig geworden. In diesem Beitrag soll daher die Frage gestellt (und sicher nicht vollständig beantwortet) werden, wie spezialisiert Services für das Forschungsdatenmanagement auf die Bedarfe von Fächer- und Fachcommunities zugeschnitten sein müssen, ohne den dafür nötigen Aufwand in absurde Höhen zu treiben.

1 Genese

Das Projekt „Geisteswissenschaftliche Infrastruktur für Nachhaltigkeit“ (gwin) wurde 2013 an der Fakultät für Geisteswissenschaften der Universität Hamburg eingesetzt, um eine Anlaufstelle für Fragen des Forschungsdatenmanagements zu etablieren, denen sich Forschende vor allem durch Forderungen von Drittmittelgebern konfrontiert sahen.

Das Projekt wurde direkt vom Präsidium der Universität gefördert mit der Aussicht, das Angebot im (nicht näher definierten) Erfolgsfall auf die gesamte Universität auszuweiten. Die Initiative zum Projekt kam direkt aus der GW-Fakultät, die in einem „eHumanities 2020+“ genannten Konzept bereits Vorarbeiten geleistet hatte.

Zu Beginn des Projektes wurde eine Umfrage unter Professoren und Multiplikatoren an der Fakultät durchgeführt, um die Bedarfe mit den Ideen aus dem eHumanities-Konzept abzugleichen. Aus dieser Erhebung und der konkreten, direkt begonnenen Arbeit – insbesondere in der Kuration von Forschungsanwendungen – konnten Bedarfe und Anforderungen identifiziert werden, die für die geisteswissenschaftliche Domäne typisch zu sein schienen und die sich in domänenspezifische Workflows und Best Practices übersetzen ließen.

Zum Ende der Laufzeit lies sich der Erfolg des Projektes unter anderem darin messen, dass mit Kurationsarbeit eine Reihe von Datenbeständen und Forschungsanwendungen vor ihrem sicheren Verschwinden bewahrt werden konnten, Aktivitäten zur Dissemination ein Bewusstsein für das Thema Forschungsdaten schaffen und mit konkreten Handlungsempfehlungen und Services unterfüttern konnten und das Projekt allgemein sehr positiv aufgenommen wurde.

Im Wissen um die Notwendigkeit entsprechender Forschungsunterstützender Services entschied sich das Präsidium 2017, das Projekt als eine Einrichtung für die gesamte Universität zu verstetigen und mit weiterem Personal auszustatten. Auch der Bereich der Forschungsinformation wurde dem neuen „Zentrum für nachhaltiges Forschungsdatenmanagement“ zugeschlagen, auch in der Hoffnung, synergetische Effekt daraus zu erzielen.

2 Status Quo

Die Umfrage innerhalb der GW-Fakultät zeichnete ein Bild von einer höchst heterogenen Datenlandschaft, in der sich selbst Forscher des selben Fachbereiches auf keinen gemeinsamen Nenner bei der Beschreibung ihrer erhobenen Daten einigen können. Die dafür individuell gefundenen (oder besser *er*fundenen) Lösungen schienen dieses Bild zum Teil zu bestätigen – in sehr vielen Fällen wurden für die Datenhaltung spezielle, nur für einen einzigen Erhebungs- und Auswertungszweck angepasste Datenbanken und Auswertungswerkzeuge geschaffen. Dass dies eine Nachnutzung extrem erschwert und auch die spätere Kuration solcher Daten vom Insiderwissen längst entschwundener studentischer Hilfskräfte abhängig macht, ist

inzwischen bekannt. Es führt aber vor allem dazu, dass sich entspechende Dienstleistungen so gut wie gar nicht skalieren lassen: Bei jedem Kurationsprojekt steht man wieder am Anfang.

Die Erweiterung des Fokus auf weitere Fakultäten und Fächer und die damit einhergehende zusätzliche Menge an zu kuratierenden Daten und Anwendungen zwingt uns also geradezu, stärker nach Gemeinsamkeiten als nach Spezialitäten in der Struktur der Daten unterschiedlicher Fächer (oder auch nur Forscher) zu suchen.

Des weiteren ist bei der Kuration von Forschungsanwendungen eine wichtige Erkenntnis, dass es klare Vorgaben zur Aufwands/-Nutzenschätzung bei der Rettung von Anwendungen geben muss, wenn allen Forschenden die Nutzung der Dienste des Zentrums kostenlos gegeben werden soll. Für Projekte, die bereits in der Antragsphase Kontakt mit dem Zentrum aufnehmen, können auch Vorgaben und Empfehlungen gemacht werden, deren Berücksichtigung zumindestens Möglichkeiten der späteren Priorisierung schafft.

Insbesondere bei der reinen Datenkuration, also der Aufbereitung von Forschungsdaten für die langfristige Speicherung und Nachnutzung, werden wir bei der Suche nach Gemeinsamkeiten immer häufiger fündig und stellen auch fest, dass im weiteren Sinne generische Lösungen (zum Beispiel das von uns etablierte “catch all/long tail” Forschungsdatenrepositorium) den Wünschen der Forschenden häufig eher entgegenkommt, als spezialierte Dienste, die einen erheblichen Wartungs- und Pflegeaufwand auf Anbieterseite und von den Forschenden eine weit über das Forschungsinteresse hinausgehende Beschäftigung mit den Daten erfordert.

Die Frage ist also, wie spezialisiert Services für das Forschungsdatenmanagement auf die Bedarfe von Fächer- und Fachcommunities zugeschnitten sein müssen, ohne den dafür nötigen Aufwand in absurde Höhen zu bewegen. Für das FDM-Zentrum ist diese Frage noch nicht abgeschlossen, aber einige interessante Beobachtungen, die sicher auch auf andere FDM-Einrichtungen übertragbar sind, lassen sich aber sehr wohl machen.

Wer macht den Satz?

Redaktionsprozesse für Forschungsdaten in den Geisteswissenschaften

Cremer, Fabian
mail[at]fabiancremer.de
Leibniz-Institut für Europäische Geschichte, Mainz, DH Lab (IEG), Deutschland
ORCID-iD: 0000-0001-8251-9727

Klaffki, Lisa
klaffki[at]hab.de
Herzog August Bibliothek, Wolfenbüttel; Forschungsverbund Marbach Weimar Wolfenbüttel, Deutschland
ORCID-iD: 0000-0002-8791-7422

Steyer, Timo
t.steyer[at]tu-braunschweig.de
Technische Universität Braunschweig, Universitätsbibliothek, Deutschland
ORCID-iD: 0000-0003-0218-2269

Zusammenfassung. Die Veröffentlichung von Forschungsdaten in den Geisteswissenschaften bleibt weiterhin eine Ausnahmeerscheinung. Der Vortrag untersucht die Rahmenbedingungen für die Veröffentlichung von Datensätzen in den Geisteswissenschaften und formuliert Ansätze für redaktionelle Aufgaben und Abläufe. Der Vorgang der Veröffentlichung umfasst von der Vorbereitung der Daten bis zu deren bibliografischen Erfassung viele Prozessschritte. Die Abläufe bei der Publikation von Datensätzen werden herausgearbeitet und dabei definiert, welche Aufgaben zwischen Forschenden und unterstützenden Einheiten wie Verlagen, Bibliotheken, Infrastrukturen und anderen Stellen aufgeteilt werden können. Eine Standardisierung und Arbeitsteilung könnte die Quantität und Qualität von Datenpublikationen erhöhen. Dazu werden die Rahmenbedingungen und die erforderlichen Prozesse vorgestellt und auf dieser Folie die aktuelle Landschaft der Anbieter (klassische Journals, Data Journals, Datenzentren mit Repositorien, Beratungsstellen) befragt. Dabei bieten die bereits bestehende Strukturen vielfältige Möglichkeiten für die Veröffentlichung von Forschungsdaten, wenngleich die redaktionellen Prozesse unterschiedlich gestaltet sind. Der Vortrag trägt zur weiteren Ausdifferenzierung der Aufgaben und Zuständigkeiten im Forschungsdatenmanagement bei. Damit lässt sich die gegenseitige

Erwartungshaltung klarer formulieren – ein wichtiger Schritt hin zu besseren Angeboten einerseits und einem Kulturwandel andererseits.

1 Der Ausgangspunkt

Die Publikation von Forschungsdaten ist (nicht nur) in den Geisteswissenschaften noch eine Ausnahmeerscheinung. Hinsichtlich datenbasierter Methoden, digitalen Infrastrukturen und veränderten Fachkulturen weisen die geisteswissenschaftlichen Fächer zwar durchaus unterschiedliche Entwicklungsstände und -geschwindigkeiten auf, aber die Datenpublikation hat sich in der Breite noch nicht als Forschungspraxis etablieren können.[1] Die Gründe für die verhaltene Veränderung der Veröffentlichungspraxis werden häufig diskutiert und gelegentlich untersucht,[2] allerdings ergeben sich daraus selten konkrete Maßnahmen für die Akteure des Wissenschaftssystems. Vor allem jenseits der Investition in Infrastruktur und der wohlwollenden Ansprache in Empfehlungen und Leitlinien finden sich nur wenige strukturelle Maßnahmen zur schrittweisen Veränderung der Publikationspraxis und ihrer Rahmenbedingungen. In dem Artikel "Redaktionssache Forschungsdaten"[3] wird die Idee einer Redaktion für Forschungsdaten skizziert, die Forschende vor und bei der Publikation von Daten unterstützt, ausgeführt von informationswissenschaftlich und informationstechnologisch ausgebildetem Personal. In vielen öffentlichen Forschungs- und Infrastruktureinrichtungen ist dies im klassischen Publikationswesen bereits ein etabliertes Arbeitsmodell (umgesetzt als Publikationsabteilung, Redaktion, Lektorat, Universitäts- oder Hausverlag). Dieser Beitrag soll dieses Konzept einer Redaktion weiterentwickeln und es auf die aktuelle Praxis in der Publikationslandschaft projizieren. Die redaktionellen Prozesse sollen dabei ausformuliert, mögliche Akteur:innen identifiziert und ein Arbeitsbündnis samt

[1] Neben den Praxiserfahrungen der Autor:innen lassen sich auch folgende Indikatoren heranziehen: Die kleine Anzahl an Data Journals und deren Output tragen bisher kaum zum Publikationsgeschehen bei. Die wenigen übergreifenden Repositorien wie bspw. "Humanities Commons CORE" (https://hcommons.org/deposits/) enthalten nur wenige Datensets (Anteil <2%). Datenpublikationen sind selbst bei den Konferenzen der Digital Humanities nur untergeordnet und werden in Calls oder Abstracts selten thematisiert - wobei einige Disziplinen wie die Computerlinguistik die Bereitstellung von Daten zumindest im Reviewverfahren bereits standardmäßig etabliert haben.

[2] Kaden 2018.

[3] Cremer, Klaffki und Steyer 2019.

Aufgabenteilung zwischen Forschenden und Redaktion herausgearbeitet werden.

2 Die Bedingungen

Die Rahmenbedingungen für die Veröffentlichung von Forschungsdaten in den Geisteswissenschaften werden von mehreren Faktoren beeinflusst:

1) Für die Datenpublikation sind die Produktions- und Analyseprozesse entscheidend: In den Geisteswissenschaften ist das Sammeln und Erzeugen von manuellen und intellektuellen Prozessen dominiert, die in komplexen Datensammlungen resultieren, deren Interpretation Kontextwissen voraussetzt, weil der Kontext den eigentlichen Gegenstand der Forschung darstellt.[4] Komplexe und kontextabhängige Informationsobjekte haben kein standardisiertes Dokumentations- und Repräsentationsformat und erschweren damit standardisierte Publikationsprozesse.
2) Die Aufbereitung nicht-strukturierter und komplexer Datensätze zu interoperablen und nachnutzbaren Daten erfordert informationswissenschaftliche Fachkompetenz (z. B. für Normdatenauszeichnung, Formatvalidierung, formale Beschreibung), die bei Forschenden nicht vorausgesetzt werden können.
3) Das Reputationssystem der Geisteswissenschaften basiert vorrangig (noch) auf der Monografie und traditionellen Verlagen.[5] Datenpublikationen sind in diesen Veröffentlichungsprozessen nicht vorgesehen.[6]
4) Institutionalisierte Serviceangebote im Forschungsdatenmanagement sind in geisteswissenschaftlichen Forschungseinrichtungen noch im Aufbau oder noch nicht eingerichtet. Während bei institutionellen Sammlungen standardisierte Prozesse zur Qualitätssicherung und Bereitstellung von Daten

[4] Vgl. Borgman 2007, 223: "While scholarly publications in all fields set data within a context, the context and interpretation are the scholarship in the humanities."
[5] Vgl. Ferwerda, Frances und Stern 2017; Terras, 2014.
[6] Vgl. Borgman 2010, 7: "In most fields, the rewards come from publication, not from data management. Scholars are hired and promoted based on their publication record rather than on the quality of their metadata."

erfolgen, ist eine Unterstützung für einzelne Forschende bei der Datenverarbeitung ein Ausnahmefall.[7]

5) Auch unabhängig von der Entwicklung in der geisteswissenschaftlichen Forschung haben sich bei Fördergebern bereits detaillierte Anforderungen an ein Forschungsdatenmanagement etabliert. So fordert der DFG-Kodex ein effektives und verlässliches Forschungsdatenmanagement ausdrücklich auch für die Geistes- und Sozialwissenschaften.[8] Die Geisteswissenschaften sind daher gefordert, ihre Arbeitsprozesse entsprechend dieser Vorgaben umzugestalten und die Qualitätskriterien zu entwickeln.

3 Die Prozesse

Aus diesen Rahmenbedingungen lässt sich ableiten, dass Redaktionsprozesse für Forschungsdaten mehrere Stufen umfassen müssten. Eine erste Aufgabenliste für eine "Forschungsdatenredaktion" ist in Cremer, Klaffki und Steyer 2019 aufgestellt, mit der Perspektive auf eine Implementierung ist die Aufzählung jedoch stärker zu strukturieren.[9] Dieser Beitrag schlägt eine Unterteilung vor, in formale Qualitätssicherung (bspw. Validierungen), Kurations- und Sicherungsmaßnahmen (bspw. Dateiformatkonversion), semantische Anreicherung (bspw. Normdaten) sowie in logistische Abläufe (bspw. Upload) und Prozesse des Ordnens und Verzeichnens (bspw. Metadatenvergabe). Darauf aufbauend lässt sich bei bestehenden Angeboten jeweils ermitteln, welche der Aufgabenschritte von welchen Akteur:innen gemeinsam oder arbeitsteilig übernommen werden. Dadurch wird explizit, welche Last bei den Forschenden verbleibt (oder entfällt) und welche Arbeiten durch unterstützende Stellen übernommen (oder ergänzt) werden können – und welche Schritte nicht berücksichtigt sind. Hinsichtlich möglicher Entwicklungen stehen sich zwei Pole gegenüber: Zum einen wäre im Zuge eines Kompetenzaufbaus bei den Forschenden verbunden mit einer erhöhten Reputation der Forschungsdatenpublikation eine Übertragung von Aufgaben von Seiten der Anbieter auf die Forschenden denkbar. In diesem Szenario wäre das Forschungsdatenmanagement eine Kernkompetenz des

[7] Diese Einschätzung wurde zuletzt im Workshop des Arbeitskreises Forschungsdaten der Leibniz-Gemeinschaft bestätigt, https://escience.aip.de/ak-forschungsdaten/veranstaltungen/9-workshop/3661-2/.

[8] Deutsche Forschungsgemeinschaft 2019.

[9] Cremer, Klaffki und Steyer 2019, 121–122.

wissenschaftlichen Arbeitens. Zum anderen könnte aber auch eine Übertragung der Publikationsaufgaben auf die unterstützenden Stellen stattfinden. Forschungsdatenmanagement wäre dann eine Form der Spezialisierung im Bereich der Wissenschaftsunterstützung. Zwischen den beiden extremen Szenarien liegen die kooperativen Arbeitsformen bei der Datenpublikation.

4 Die Landschaft

In Ergänzung der konzeptuellen Ebene werden konkrete Beispiele für Forschungsdatenpublikationen und die damit verbundenen Prozesse vorgestellt.[10] Ein Blick auf die bestehende Landschaft, die Akteur:innen und Praktiken gibt auch Aufschlüsse darüber, wo Prozesse skalieren oder sich erweitern lassen. Auch der Einfluss und die Beachtung der fachspezifischen oder lokalen Rahmenbedingungen kann dabei hervortreten. Wir untersuchen und diskutieren in dem Beitrag folgende Angebote und Akteure:

1) Zeitschriften (Journals) im Bereich der digitalen Geisteswissenschaften sind in der Regel auch mit einem Innovationsanspruch hinsichtlich der Publikationsprozesse verbunden. So bieten einige Journals interaktive und multimediale Oberflächen, erproben offene Reviewformate oder neue Formate. Wie sehen die Journals den Bedarf und die eigenen Möglichkeiten, Forschungsdaten zu publizieren?[11]
2) Data Journals veröffentlichen Data Papers, die als natursprachliche Textdokumente die betreffenden Forschungsdaten dokumentieren, beschreiben und kontextualisieren. Damit bieten sie ein Publikationsformat für Daten, das sich als Äquivalent zu klassischen Texten in Zitations- und Reputationsmechanismen einfügt. Die Forschungsdaten selbst werden parallel in Repositorien hinterlegt. Wie sind bei Data Journals im Bereich der Geisteswissenschaften die Redaktionsprozesse organisiert und wo finden Kontrollen oder Anreicherungen statt?[12]

[10] Die Prozesse hängen von unterschiedlichen Publikationsstrategien ab: 1) Daten als eigenständiges Informationsobjekt, 2) Daten im Rahmen eines Data Papers in einem Data Journal, 3) Daten als Supplement einer klassischen Publikation. Vgl. Pampel und Elger 2021, 527–528.

[11] Siehe u.a. ZfdG – Zeitschrift für digitale Geisteswissenschaften, https://zfdg.de/e-wie-editorial.

[12] Siehe u.a. Journal of Open Humanities Data (JOHD), https://openhumanitiesdata.metajnl.com/about/editorialpolicies/.

3) Datenzentren mit institutionellem oder disziplinärem Auftrag bieten oft die Möglichkeit, Daten in einem Repositorium zu veröffentlichen und sichern zusätzlich deren Archivierung und Erhaltung.[13] Welche redaktionellen Abläufe erfolgen hier beim Ingest? Skalieren die Angebote und Prozesse für mehr Daten, mehr Anfragen und breitere Nutzungsgruppen?
4) Beratungsstellen zu Forschungsdaten und zum Forschungsdatenmanagement gehören zu den verbreitetsten Serviceangeboten und sind mittlerweile an zahlreichen Einrichtungen und auf unterschiedlichen Ebenen implementiert, sowohl in universitären wie außeruniversitären Kontexten. In diese Kategorie fallen auch Stellen mit Unterstützungsaufträgen in großen Verbundprojekten. Mit welchen Anfragen zur Datenpublikation werden diese Beratungsstellen konfrontiert? Gibt es Stellen, die neben Beratung auch konkrete Unterstützung und Datenverarbeitung realisieren (könnten)?

5 Diskussion

Im Mittelpunkt der Diskussion soll die Leitfrage stehen, ob sich ein arbeitsteiliger Redaktionsprozess bei Datenpublikationen modellieren und definieren lässt. Dabei ist besonders wichtig, ob eine Aufgabenteilung zwischen Forschenden und Anbietern umsetzbar ist und als standardisierter Prozess entwickelt werden kann.
Ebenfalls ist zu hinterfragen, ob ein solcher Publikationsprozess den Spezifika der Geisteswissenschaften gerecht werden kann und wie dies zur Qualitätssicherung beiträgt: Denn nur wenn die Publikation von Forschungsdaten mit Reputation verbunden ist, wird eine Arbeitsteilung auch umsetzbar sein. Qualitätsgesicherte, aber für die Forschenden niedrigschwellig angebotene, Unterstützungsformen könnten zu anderem Publikationsverhalten beitragen.

[13] Siehe u.a. heiDATA, Universität Heidelberg, https://data.uni-heidelberg.de/datenaufnahme.html.

Bibliografie

Borgman, Christine L. Scholarship in the digital age: information, infrastructure, and the Internet. Cambridge, Mass: MIT Press, 2007.

Borgman, Christine L. „Research Data: Who will share what, with whom, when, and why?“, 2010.

Cremer, Fabian, Lisa Klaffki, und Timo Steyer. „Redaktionssache Forschungsdaten“. *Bibliothek Forschung und Praxis* 43, Nr. 1 (2019): 118–25. https://doi.org/10.1515/bfp-2019-2018.

Deutsche Forschungsgemeinschaft. „Guidelines for Safeguarding Good Research Practice. Code of Conduct“. Zenodo, 15. September 2019, https://doi.org/10.5281/zenodo.3923601.

Ferwerda, Eelco, Frances Pinter, und Niels Stern. „A Landscape Study on Open Access and Monographs: Policies, Funding and Publishing in Eight European Countries“. Report. Knowledge Exchange, 5. Oktober 2017. https://apo.org.au/node/117361.

Kaden, Ben. „Warum Forschungsdaten nicht publiziert werden“. *LIBREAS. Library Ideas*, Nr. 33 (2018). https://libreas.eu/ausgabe33/kaden-daten/.

Pampel, Heinz, und Kirsten Elger. „5.6 Publikation und Zitierung von digitalen Forschungsdaten“. In *Praxishandbuch Forschungsdatenmanagement*, 521–36. De Gruyter Saur, 2021. https://doi.org/10.1515/9783110657807-028.

Wie komme ich zu einer Forschungsdatenmanagement-Strategie?

Eine Antwort gibt das DIAMANT-Modell

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

Zusammenfassung. Das DIAMANT-Modell wurde im Rahmen des BMBF-geförderten Projekts „Prozessorientierte Entwicklung von Managementinstrumenten für Forschungsdaten im Lebenszyklus“ (PODMAN) entwickelt. Das grundsätzliche Ziel des Modells besteht darin, Forschungseinrichtungen strategische Instrumente bereitzustellen, um Forschungsdatenmanagement (FDM)-Technologien und -Services aufbauen, pflegen, optimieren und/oder weiterentwickeln zu können. Auf diese Weise wird es Forschenden erleichtert, für ihre Projekte möglichst einfach eine FDM-Strategie zu entwickeln und umzusetzen. Das DIAMANT-Modell besteht aus dem FDM-Referenzprozess, der den gesamten FDM-Prozess und die zugehörigen Akteur:innen darstellt und ihren Grad der Beteiligung / Verantwortung an einem FDM-Prozessschritt benennt. Ein weiterer Bestandteil des Modells ist die FDM-Kompetenzmatrix. Sie dient der IST- / SOLL-Analyse, um den Referenzprozess auf den eigenen konkreten Anwendungsfall (z. B. die eigene Institution, Abteilung oder Forschergruppe) übertragen zu können. Die Visualisierung der Prozesse mit Hilfe der ARIS Geschäftsprozessmodellierung erleichtert zu dem die Kommunikation zwischen den Beteiligten und hilft dabei Schnittstellenprobleme zu identifizieren. Der Vortrag wird anhand eines konkreten Beispiels exemplarisch vorführen, wie das DIAMANT-Modell angewendet wird und dabei die Kernelemente, ihre Zusammenhänge und Wirkung erläutern.

Die Ansprüche von Seiten der Drittmittelgebenden in Form von Leitlinien zum Umgang mit Forschungsdaten (FAIR-Prinzipien) erscheinen nicht nur den Forschenden sehr hoch, sondern stellen auch Hochschulen und außeruniversitäre Forschungseinrichtungen vor große Herausforderungen, um ihre Wissenschaftler:innen bei der Entwicklung und Umsetzung von digitalen Forschungskonzepten zu unterstützen. Oftmals stellt sich die Frage – und die nicht nur bei den Forschenden – was die Erfüllung der FAIR-Prinzipien im Praktischen genau bedeuten, welche Kompetenzen und Ressourcen eigentlich benötigt werden und

wie die identifizierten Bedarfe adäquat bedient werden können? Sowohl die FAIR-Prinzipien als auch Beschreibungsmodelle wie z. B. der Datenlebenszyklus formulieren optimale Abläufe und Ergebnisse, die jedoch nicht dazu geeignet sind, sie eins zu eins auf ein konkretes Forschungsprojekt oder gar eine ganze Forschungseinrichtung zu übertragen. Erschwerend kommt hinzu, dass viele Aktivitäten im Forschungsdatenmanagement (FDM) nicht durch eine Einzelperson oder eine einzelne Abteilung einer Forschungseinrichtung durchgeführt werden können. D. h. es müssen Kooperationen geschlossen und Arbeitsabläufe definiert werden, die eine optimale, reibungslose Funktionsweise garantieren, sodass die Forschenden zuverlässig die gewünschten Ergebnisse für ihre projektspezifische FDM-Strategie erhalten.

Um eine FDM-Strategie (egal auf welcher Organisationsebene) zu entwickeln, muss analysiert werden, welche Bedarfe eine Forschungseinrichtung, ein Institut, eine Abteilung oder andere Forschungseinheiten (z. B. Forschergruppen, Forschungsprojekte) im konkreten institutionellen Kontext haben; welche Kompetenzen bereits vorhanden sind, welche fehlen und wo gegebenenfalls Schnittstellenprobleme bei der Erfüllung bestimmter FDM-Aufgaben bestehen. Genau genommen geht es also darum den Informationsverarbeitungsprozess im FDM für einen konkreten Anwendungsfall systematisch zu analysieren, ggf. die Lücken zu identifizieren und darauf aufbauend Lösungswege zu finden. Genau diesen Analyseprozess unterstützt das DIAMANT-Modell, das vom BMBF geförderten Projekt „Prozessorientierte Entwicklung von Managementinstrumenten für Forschungsdaten im Lebenszyklus“ (PODMAN) entwickelt wurde.

Das DIAMANT-Modell dient dazu, eine Informationsarchitektur für das Management von Forschungsdatenservices zu kreieren bzw. weiterzuentwickeln: DIAMANT steht für „**D**esigning an **I**nformation **A**rchitecture for Data **MAN**agement **T**echnologies“.[1] Das Modell besteht zum einen aus dem FDM-Referenzprozess, der – angelehnt an die ARIS-Prozessmodellierung[2] – den gesamten FDM-Prozess mit den einzelnen Prozessschritten (FDM-Funktionen) und die zugehörigen Akteur:innen (FDM-Rollen) in einem Prozessdiagramm darstellt und in der sogenannten FDM-Steuerungsmatrix[3] angibt, in welcher Beziehung und Verantwortung eine Akteur:in in einen Prozessschritt involviert ist. Dabei

[1] vgl. Lemaire, Gerhards, Kellendonk, Blask, & Förster 2020.
[2] vgl. Scheer, August-Wilhelm 2002.
[3] vgl. Lemaire et al. 2020, S. 26.

setzt der FDM-Referenzprozesses direkt beim Forschungsprozess an und stellt somit die Perspektive der Forschenden ins Zentrum.

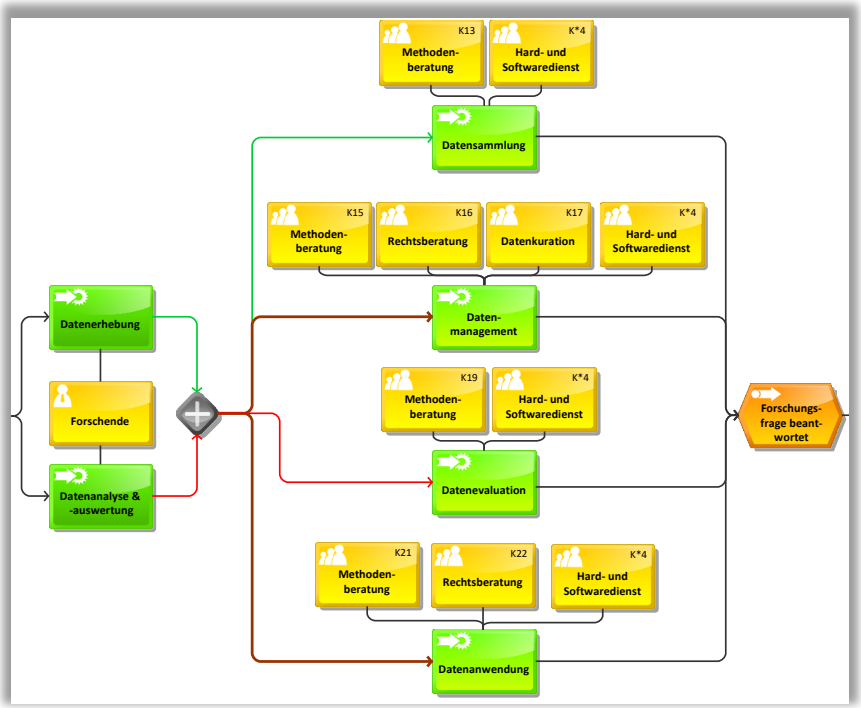


Abb. 1 Ausschnitt aus der Ereignisgesteuerten Prozesskette des FDM-Referenz-prozesses (Vgl. Lemaire et al. 2020, 23)

Ein weiterer Bestandteil des DIAMANT-Modells ist die FDM-Kompetenzmatrix[4], die basierend auf dem FDM-Referenzprozess für jeden Prozessschritt und den in erster Linie daran beteiligten Akteur:innen definiert, welche Kompetenzen zur Erfüllung des Schrittes benötigt werden.
Diese Instrumente (Steuerungs- & Kompetenzmatrix) dienen dazu eine IST- / SOLL-Analyse, ein sogenanntes Benchmarkingverfahren, auf einen konkreten Anwendungsfall zu übertragen. Es wird dabei helfen herauszufinden, welche FDM-Funktionen gebraucht und welche Kompetenzen dafür benötigt werden. Es hilft dabei Schnittstellenprobleme zu identifizieren bzw. auch Aufgabenstellungen für bestimmte Akteur:innen transparenter zu machen und klarer zu artikulieren. Auf diese Weise werden auch die Verantwortlichkeiten unter den Akteur:innen eindeutiger verteilt. Vor allem die Visualisierung

[4] vgl. Lemaire et al. 2020, S. 30.

durch die Prozessmodellierung mit ARIS hilft in der Kommunikation zwischen verschiedenen Beteiligten, wenn es sich um komplexere Prozesse mit vielen Akteur:innen handelt.

Das DIAMANT-Modell kann auf den verschiedenen Ebenen entweder zur Weiterentwicklung der FDM-Strategie einer gesamten Forschungseinrichtung, aber auch für kleinere Forschungseinheit (z. B. Forschungsinstitute, Fachabteilungen, größere Forschergruppen etc.) verwendet werden.

Der Vortrag wird exemplarisch anhand von Ausschnitten aus dem Archivierungsworkflow der Universität Trier vorführen, wie das DIAMANT-Modell angewendet wird und dabei seine Kernelemente und deren Zusammenhänge und beabsichtigten Wirkungsweisen erläutern.

Bibliografie

Lemaire, Marina, Lea Gerhards, Stefan Kellendonk, Katarina Blask, and André Förster. Das DIAMANT-Modell 2.0. Modellierung Des FDM-Referenzprozesses Und Empfehlungen Für Die Implementierung Einer Institutionellen FDM-Servicelandschaft. ESciences Working Papers 05. Trier, 2020. https://doi.org/10.25353/ubtr-xxxx-f5d2-fffb.

Scheer, August-Wilhelm. ARIS - Vom Geschäftsprozess zum Anwendungssystem. 4. durchges. Aufl. Berlin: Springer Berlin, 2002.

Zur Nachhaltigkeit von Visualisierungen in den digitalen Geisteswissenschaften

High-Steskal, Nicole
nicole.high-steskal[at]donau-uni.ac.at
Donau-Universität Krems, Österreich
ORCID-iD: 0000-0002-9823-5507

Liem, Johannes
johannes.liem[at]donau-uni.ac.at
Donau-Universität Krems, Österreich
ORCID-iD: 0000-0001-8696-0245

Mayr, Eva
eva.mayr[at]donau-uni.ac.at
Donau-Universität Krems, Österreich
ORCID-iD: 0000-0001-8402-5990

Windhager, Florian
florian.windhager[at]donau-uni.ac.at
Donau-Universität Krems, Österreich
ORCID-iD: 0000-0002-5170-2243

Zusammenfassung. Die zunehmende Verfügbarkeit großer Mengen an digitalen Forschungsdaten macht nachhaltiges Datenmanagement auch zu einer omnipräsenten Herausforderung in den digitalen Geisteswissenschaften. Dies gilt für die nachhaltige Verfügbarkeit digitaler Infrastrukturen ebenso wie für die Funktionstüchtigkeit und Langzeitverfügbarkeit von Tools zur Auswertung und Vermittlung der wachsenden Datenbestände. Zu diesen Werkzeugen zählen in prominenter Weise auch interaktive Visualisierungen, welche vielfach die Analyse und Kommunikation von Forschungsdaten unterstützen, und welche auch als Zwischen- und Endergebnisse von digitalen Forschungsprozessen eine substantielle Rolle spielen. Trotz der wachsenden Bedeutung solcher Instrumente fehlen bis heute allerdings Strategien, welche deren Erfassung und langfristige Verfügbarmachung im Rahmen von Datenmanagementplänen ermöglichen würden, was auch der Transparenz und Provenienz von Forschungsergebnissen dienlich wäre. Vor diesem Hintergrund reflektieren wir in dem Beitrag die multiplen Komponenten von Visualisierungen, welche durch Forschungsdateninitiativen gesichert werden sollten, und regen die Formulierung von FAIR-Prinzipien für die Visualisierung an.

1 Einführung – Digitale Nachhaltigkeit

Die zunehmende Digitalisierung kultureller Bestände und die entsprechende Verfügbarkeit großer Mengen an Forschungsdaten macht nachhaltiges Datenmanagement auch zu einer eminenten Herausforderung in den (digitalen) Geisteswissenschaften. In diesem Kontext ist die Langzeitverfügbarkeit digitaler Infrastrukturen (i.e. von Systemen zur Speicherung, Organisation und Verteilung digitaler Daten) ebenso zu gewährleisten, wie die Funktionstüchtigkeit von Tools zur Auswertung und Vermittlung der wachsenden Datenbestände. Zu diesen Werkzeugen zählen auch interaktive Visualisierungen, die in unterschiedlichen Forschungsphasen zur Analyse und Kommunikation von Daten zum Einsatz kommen. Die folgenden Überlegungen sind motiviert von dem Umstand, dass trotz der wachsenden Bedeutung solcher Tools bis dato Strategien fehlen, um diese im Rahmen von Datenmanagementplänen zu erfassen und entsprechend der FAIR-Prinzipien langfristig zu erhalten. Solche Verfahren würden nicht nur eine gute wissenschaftliche Praxis stärken, sondern auch mit Blick auf die Transparenz und Provenienz von Forschungsergebnissen deren Nachvollziehbarkeit erhöhen. Im Folgenden reflektieren wir kurz die Rolle von Visualisierungen im Feld der Geisteswissenschaften, unterscheiden Komponenten von Visualisierungen, die gesichert werden sollten, und liefern einen Abriss über ähnliche Debatten für die Dokumentation von Software, um im Bereich der Visualisierung ähnliche Diskussionen anzuregen. Die praktische Motivation hierzu erwächst aus der Arbeit an einem nachhaltigen Umgang mit Visualisierungen im H2020 Projekt InTaVia (https:/intavia.eu), welcher auch anderen Visualisierungsprojekten zur Orientierung dienen kann.

2 Visualisierungen in den Digital Humanities

In der letzten Dekade kam es zu einer Intensivierung der Entwicklung und Nutzung von interaktiven Visualisierungen in den digitalen Geisteswissenschaften.[1] Prominente Beispiele in diesem Kontext inkludieren Interfaces zur Darstellung und Exploration von mehrdimensionalen geschichtlichen Daten[2], historischen Netzwerken[3],

[1] Benito-Santos & Sanchez 2020; Bradley et al. 2018.
[2] Posner 2014.
[3] Molinero et al. 2017.

kulturellen Sammlungen[4], sowie von großen Textkorpora[5]. Mit Blick auf traditionelle hermeneutische Forschungsmethoden erschließt sich die Nützlichkeit solcher Ansätze oft als Big Data-Technologien, welche etablierte Verfahren des close readings um makroanalytische Optionen des "distant readings"[6] oder "distant viewings"[7] erweitern. In zunehmendem Maße finden auch Reflexionen über die Rolle von Visualisierungen im Rahmen geisteswissenschaftlicher Forschungsprozesse statt.[8] In auffälliger Weise kommen hierbei aber bislang Reflexionen auf nachhaltige Entwicklungsstrategien zu kurz, die die Langzeitverfügbarmachung, Reproduzierbarkeit, und Nachvollziehbarkeit von Visualisierungstechnologien gewährleisten könnten.

3 Nachhaltigkeit verschiedener Visualisierungskomponenten

In der Forschung zu und mit Visualisierungen spielen multiple Komponenten und Prozesse zusammen: (a) Erzeugung, Prozessierung und Selektion von Daten, (b) Designprozesse, (c) Implementierungsprozesse und (d) user-seitiges Nutzungsverhalten.

3.1 Datenerzeugung, -selektion und -aufbereitung

Während die Erfassung und Erhaltung von vorhandenen Daten in der Regel im Fokus von Datenmanagementplänen liegen, sind transiente Schritte der Datenselektion, Datenerzeugung, und Datenaufbereitung oftmals der Nachvollziehbarkeit entzogen. In diesem Kontext bedarf es neuer Ansätze der Dokumentation, welche einen kritischen Nachvollzug von digitalen Translationsketten (als Serien von "material-medialen Translationen")[9] und repräsentativer "Standpunktgebundenheit"[10] überhaupt erst möglich machen.[11]

[4] Windhager et al. 2018a.
[5] Jänicke et al. 2015.
[6] Moretti 2007.
[7] Arnold & Tilton 2019.
[8] Bradley et al. 2018; Drucker 2015; Hinrichs et al. 2019.
[9] Latour 1999.
[10] Drucker 2017.
[11] Windhager et al. 2018b.

3.2 Visualisierungsdesign

Gestalterische Entscheidungen in der Erstellung von Visualisierungen sind für VisualisierungsnutzerInnen selten transparent. In diesem Kontext bieten design study papers ein etabliertes Mittel um ein nachhaltiges Verständnis der entsprechenden Forschungs- und Entwicklungsprozesse zu ermöglichen.[12] Darüber hinaus bieten sich neue Möglichkeiten zur Dokumentation von Designprozessen durch Notebookumgebungen (z.B. Observable https://observablehq.com/).[13] Diese stellen sicher, dass das Schreiben von Programmcode mit der genauen Beschreibung der Designentscheidungen einhergeht und somit neben dem Tool auch sein Werdegang dokumentiert wird.

3.3 Implementierung

Implementierungen von Visualisierungen werden in der Regel über Code-Repositories dokumentiert, geteilt und langfristig verfügbar gemacht. Idealerweise geschieht dies im Zusammenspiel mit der Dokumentation des Quellcodes, Informationen zur Versionshistorie, und mit Medien der technischen Visualisierungsdokumentation, sowie Anleitungen, Tutorials, oder Onboarding-Techniken.[14] Eine weitere Möglichkeit der nachhaltigen Verfügbarmachung ist die Verwendung einer deklarativen Syntax zur Formalisierung von Visualisierungen. Vega und Vega-Lite beschreiben beispielsweise das Mapping zwischen Daten und deren visueller Repräsentation anhand eines maschinen- sowie menschenlesbaren Textformates (i.e., JSON Spezifikation).[15]
Aufgrund der rasanten Entwicklung von Nutzungs- und Entwicklungsumgebungen ist die Erhaltung der Lauffähigkeit von älteren Visualisierungen ein eigenes Aufgabengebiet.[16] Java-basierte Visualisierungen vertragen sich bereits nach wenigen Jahren kaum noch mit aktuellen Web-Technologien und Browserumgebungen, was evident macht, dass nicht nur der Quellcode, sondern auch die jeweils gängigen Kompilier- bzw. Interpretationsmechanismen eine große Bedeutung für die nachhaltige Verfügbarkeit von Visualisierungen haben. Eine im Forschungskontext kaum realisierbare Strategie ist das stetige Aktualisieren von Visualisierungs-Quellcodes entsprechend moderner Technologien. In Analogie zur Erhaltung von alten Computer-

[12] Sedlmair et al. 2012.
[13] Wood et al. 2019.
[14] Stoiber et al. 2019.
[15] Vgl. Satyanarayan et al. 2016.
[16] Storey et al. 2008.

und Konsolenspielen erscheinen hier eher “Time Capsule”-Emulationen von Hard- und Softwareumgebungen als erfolgversprechende Strategien. Der Einsatz weit verbreiterter, etablierter Open Source Werkzeuge zur Umsetzung von Visualisierungsprojekten vereinfacht deren Wartbarkeit und kann somit ebenfalls zur Nachhaltigkeit einer Visualisierung beitragen.

3.4 User-Verhalten

Eine mögliche Gefährdung der historischen Nachvollziehbarkeit und Reproduzierbarkeit von visualisierungs-basierten Forschungsergebnissen ergibt sich nicht zuletzt auch aus den üblicherweise undokumentierten Explorations- und Analyseschritten von ForscherInnen und NutzerInnen, welche - ebenso wie Daten und Tools - mit zukünftigen Forschungsberichten und -ergebnissen verfügbar gemacht werden könnten.[17]

4 Diskussion

Die Inklusion von weiteren Aspekten des Forschungsprozesses in das Konzept des Forschungsdatenmanagements wird erst jüngstens für die Erhaltung von Forschungssoftware verstärkt diskutiert.[18] Lamprecht und KollegInnen fordern in diesem Kontext prägnant, dass die Erhaltung von Forschungssoftware entsprechend der FAIR-Prinzipien anzustreben sei, diese jedoch für Forschungssoftware adaptiert werden müssen.[19] Konkrete Anleitungen oder allgemeine Standards sind hierfür allerdings noch kaum zu finden – oder decken selten die Komplexität der Forschungssoftware ab. So wurden auf forschungsdaten.info bereits Empfehlungen publiziert, doch sind diese im Umfang knapp bemessen und sehr allgemein formuliert. Im Februar 2021 hat zuletzt die FAIR for Research Software -Forschungsgruppe der Research Data Alliance einen Zwischenbericht ihrer Tätigkeiten publiziert mit Empfehlungen zu Nachhaltigkeit und Forschungssoftware.[20] Bislang gibt es nur vereinzelte Beispiele, wo Überlegungen zum nachhaltigen Umgang mit Visualisierungen oder Forschungssoftware formalisiert, etwa in Forschungs”daten”plänen oder Muster-Richtlinien, wie an den

[17] Shrinivasan and van Wijk 2008.
[18] Bach et al. 2019a; Anzt et al 2020.
[19] Lamprecht et al. 2020.
[20] Katz et al. 2021.

Helmholtz-Zentren[21], festgehalten und im Forschungsprozess eingeplant werden.
Forschungssoftware wird von Lamprecht definiert als “software that is used to generate, process or analyse results that you intend to appear in a publication (either in a journal, conference paper, monograph, book or thesis)”.[22] Diese Definition könnte auch auf viele Visualisierungstools angewandt werden, jedoch bringen diese als Multikomponentensysteme eine Reihe von eigenen Herausforderungen mit sich. Durch eine vertiefte Auseinandersetzung mit diesen Herausforderungen sind aber nicht nur langfristigere gesellschaftliche Beiträge der Geisteswissenschaften zu erhoffen, sondern auch, wie zuletzt von Drucker formuliert, eine bessere Balancierung von technischer Komplexität und tatsächlicher Nachhaltigkeit.[23]

Dieses Projekt wird von der Europäischen Kommission im Rahmen des H2020 Research and Innovation Programme, Grant Agreement No. 101004825 gefördert. – This project has received funding from the European Union’s Horizon 2020 research and innovation programme under grant agreement No 101004825.

[21] Bach et al. 2019b.
[22] Lamprecht et al. 2020, 39.
[23] Drucker 2021.

Bibliografie

Anzt, Hartwig, Felix Bach, Stephan Druskat, Frank Löffler, Axel Loewe, Bernhard Y. Renard, Gunnar Seemann, et al. 2020. “An Environment for Sustainable Research Software in Germany and Beyond: Current State, Open Challenges, and Call for Action.” F1000Research 9 (April): 295. https://doi.org/10.12688/f1000research.23224.1.

Arnold, Taylor, and Lauren Tilton. 2019. “Distant Viewing: Analyzing Large Visual Corpora.” Digital Scholarship in the Humanities, no. fqz013 (March). https://doi.org/10.1093/digitalsh/fqz013.

Bach, F., Druskat, S., Katerbow, M., Loewe, A., & Seemann, G. (2019a). Herausforderungen für die nachhaltige Entwicklung, Bereitstellung und Pflege von Forschungssoftware in Deutschland. deRSE-Konferenz, Potsdam.

Bach, F., Bertuch, O., Busse, C., Castell, W. zu, Celo, S., Denker, M., Dinkelacker, S., Druskat, S., Faber, C., Finke, A., Fritzsch, B., Hammitzsch, M., Haseleu, J., Konrad, U., Krupa, J., Leifels, Y., Mohns-Pöschke, K., Moravcikova, M., Nöller, J., … Wortmann, D. (2019b). Muster-Richtlinie Nachhaltige Forschungssoftware an den Helmholtz-Zentren. https://doi.org/10.2312/os.helmholtz.007

Benito-Santos, Alejandro, and Roberto Therón Sánchez. 2020. “A Data-Driven Introduction to Authors, Readings, and Techniques in Visualization for the Digital Humanities.” IEEE Computer Graphics and Applications 40 (3): 45–57. https://doi.org/10.1109/MCG.2020.2973945.

Bradley, Adam James, Mennatallah El-Assady, Katharine Coles, Eric Alexander, Min Chen, Christopher Collins, Stefan Jänicke, and David Joseph Wrisley. 2018. "Visualization and the digital humanities." IEEE Computer Graphics and Applications 38(6): 26-38. https://doi.org/10.1109/MCG.2018.2878900

Dörk, Marian, Sheelagh Carpendale, and Carey Williamson. 2011. “EdgeMaps: Visualizing Explicit and Implicit Relations.” In Visualization and Data Analysis 2011, 7868:78680G. International Society for Optics and Photonics. https://doi.org/10.1117/12.872578.

Drucker, Johanna. 2015. “Graphical Approaches to the Digital Humanities.” In A New Companion to Digital Humanities, 238–50. John Wiley & Sons, Ltd. https://doi.org/10.1002/9781118680605.ch17.

———. 2017. “Information Visualization and/as Enunciation.” Journal of Documentation 73 (5): 903–16. https://doi.org/10.1108/JD-01-2017-0004.

———. 2021. “Sustainability and Complexity: Knowledge and Authority in the Digital Humanities.” Digital Scholarship in the Humanities, no. fqab025 (March). https://doi.org/10.1093/llc/fqab025.

Hinrichs, Uta, Stefania Forlini, and Bridget Moynihan. 2019. “In Defense of Sandcastles: Research Thinking through Visualization in Digital Humanities.” Digital Scholarship in the Humanities 34 (Supplement_1): i80–99. https://doi.org/10.1093/llc/fqy051.

Jänicke, Stefan, Greta Franzini, Muhammad Faisal Cheema, and Gerik Scheuermann. 2015. “On Close and Distant Reading in Digital Humanities: A Survey and Future Challenges.” https://doi.org/10.2312/eurovisstar.20151113.

Katz, Daniel S., Morane Gruenpeter, Tom Honeyman, Lorraine Hwang, Mark D. Wilkinson, Vanessa Sochat, Hartwig Anzt, Carole Goble, and for FAIR4RS Subgroup 1. 2021. “A Fresh Look at FAIR for Research Software.” ArXiv:2101.10883 [Cs], February. http://arxiv.org/abs/2101.10883.

Lamprecht, Anna-Lena, Leyla Garcia, Mateusz Kuzak, Carlos Martinez, Ricardo Arcila, Eva Martin Del Pico, Victoria Dominguez Del Angel, et al. 2020. “Towards FAIR Principles for Research Software.” Data Science 3 (1): 37–59. https://doi.org/10.3233/DS-190026.

“Lamprecht et Al_2020_Towards FAIR Principles For Research Software.Pdf.” n.d.

Latour, Bruno. 1999. Pandora’s Hope: Essays on the Reality of Science Studies. Cambridge, Mass: Harvard University Press.

Molinero, Vanessa Serrano, Benjamin Bach, Catherine Plaisant, Nicole Dufournaud, and Jean-Daniel Fekete. 2017. “Understanding the Use of The Vistorian: Complementing Logs with Context Mini-Questionnaires.” Visualization for the Digital Humanities, October, 6.

Moretti, Franco. 2007. Graphs, Maps, Trees: Abstract Models for Literary History. Paperback edition. London New York: Verso.

Posner, Miriam. 2014-11-23. “Getting Started with Palladio – Miriam Posner’s Blog.” Accessed May 2, 2021. https://miriamposner.com/blog/getting-started-with-palladio/.

Satyanarayan, Arvind, Dominik Moritz, Kanit Wongsuphasawat, and Jeffrey Heer. 2017. “Vega-Lite: A Grammar of Interactive Graphics.” IEEE Transactions on Visualization and Computer Graphics 23 (1): 341–50. https://doi.org/10.1109/TVCG.2016.2599030.

Sedlmair, Michael, Miriah Meyer, and Tamara Munzner. 2012. “Design Study Methodology: Reflections from the Trenches and the Stacks.” IEEE Transactions on Visualization and Computer Graphics 18 (12): 2431–40. https://doi.org/10.1109/TVCG.2012.213.

Shrinivasan, Yedendra Babu, and Jarke J. van Wijk. 2008. “Supporting the Analytical Reasoning Process in Information Visualization.” In Proceedings of the SIGCHI Conference on Human Factors in Computing Systems, 1237–46. CHI ’08. New York, NY, USA: Association for Computing Machinery. https://doi.org/10.1145/1357054.1357247.

“Softwareentwicklung Und Gute Wissenschaftliche Praxis | Ethik Und Gute Wissenschaftliche Praxis | Themen | Forschungsdaten Und Forschungsdatenmanagement.” n.d. Accessed May 2, 2021. https://www.forschungsdaten.info/themen/ethik-und-gute-wissenschaftliche-praxis/softwareentwicklung-und-gute-wissenschaftliche-praxis/.

Stoiber, Christina, Florian Grassinger, Margit Pohl, Holger Stitz, Marc Streit, and Wolfgang Aigner. 2019. “Visualization Onboarding: Learning How to Read and Use Visualizations.” OSF Preprints. https://doi.org/10.31219/osf.io/c38ab.

Storey, Margaret-Anne, Chris Bennett, R. Ian Bull, and Daniel M. German. 2008. “Remixing Visualization to Support Collaboration in Software Maintenance.” In 2008 Frontiers of Software Maintenance, 139–48. https://doi.org/10.1109/FOSM.2008.4659257.

Windhager, Florian, Paolo Federico, Günther Schreder, Katrin Glinka, Marian Dörk, Silvia Miksch, and Eva Mayr. 2018a. “Visualization of Cultural Heritage Collection Data: State of the Art and Future Challenges.” IEEE Transactions on Visualization and Computer Graphics 25 (6): 2311–30.

Windhager, Florian, Katrin Glinka, Eva Mayr, Günther Schreder, Marian Dörk. 2018b. „Zur Weiterentwicklung des “cognition support”: Sammlungsvisualisierungen als Austragungsort kritisch-kulturwissenschaftlicher Forschung“, in: Georg Vogeler (Hg.): Kritik der digitalen Vernunft. DHd 2018 Köln (Book of Abstracts): 341-345.

Wood, Jo, Alexander Kachkaev, and Jason Dykes. 2019. “Design Exposition with Literate Visualization.” IEEE Transactions on Visualization and Computer Graphics 25 (1): 759–68. https://doi.org/10.1109/TVCG.2018.2864836. The University of Chicago Press Editorial Staff, The Chicago Manual of Style, 17th Edition (Chicago: University of Chicago Press, 2017), https://doi.org/10.7208/cmos17.

POSTER

A long-term archiving and dissemination ecosystem for Humanities' research data

Schneider, Gerlinde
gerlinde.schneider[at]uni-graz.at
Zentrum für Informationsmodellierung - Austrian Centre for Digital Humanities, Universität Graz, Österreich
ORCID-iD: 0000-0002-4942-9225

Scholger, Martina
martina.scholger[at]uni-graz.at
Zentrum für Informationsmodellierung - Austrian Centre for Digital Humanities, Universität Graz, Österreich
ORCID-iD: 0000-0003-1438-3236

Scholger, Walter
walter.scholger[at]uni-graz.at
Zentrum für Informationsmodellierung - Austrian Centre for Digital Humanities, Universität Graz, Österreich
ORCID-iD: 0000-0002-9256-0958

Steiner, Elisabeth
Elisabeth.steiner[at]uni-graz.at
Zentrum für Informationsmodellierung - Austrian Centre for Digital Humanities, Universität Graz, Österreich

Stigler, Joahnnes
Johannes.stiegler[at]uni-graz.at
Zentrum für Informationsmodellierung - Austrian Centre for Digital Humanities, Universität Graz, Österreich
ORCID-iD: 0000-0003-0803-1496

Stoff, Sebastian
Sebastian.stoff[at]uni-graz.at
Zentrum für Informationsmodellierung - Austrian Centre for Digital Humanities, Universität Graz, Österreich
ORCID-iD: 0000-0001-6941-113X

Tosques, Fabio
Fabio.tosques[at]uni-graz.at
Zentrum für Informationsmodellierung - Austrian Centre for Digital

Humanities, Universität Graz, Österreich

Vasold, Gunter
gunter.vasold[at]uni-graz.at
Zentrum für Informationsmodellierung - Austrian Centre for Digital Humanities, Universität Graz, Österreich

Vogeler, Georg
georg.vogler[at]uni-graz.at
Zentrum für Informationsmodellierung - Austrian Centre for Digital Humanities, Universität Graz, Österreich
ORCID-iD: 0000-0002-1726-1712

Zuanni, Chiara
chiara.zuanni[at]uni-graz.at
Zentrum für Informationsmodellierung - Austrian Centre for Digital Humanities, Universität Graz, Österreich
ORCID-iD: 0000-0003-4027-0278

Introduction. For almost 20 years, the digital repository GAMS (Geisteswissenschaftliches Asset Management System) has been providing long-term preservation of Humanities' resources offering both a technical solution and a way of achieving sustainability in the handling of research data. Such domain-specific digital repositories rely on specialised design principles and a strategic orientation to balance the need for long-term sustainability and the use of advanced technologies. The repository solely relies on open source software due to its commitment to open access and open science. Its design principles intend to foster compliance with international best practices like the COAR Community Framework for Good Practices in Repositories (COAR 2020) and the certification guidelines of the CoreTrustSeal, and following the FAIR principles of findability, accessibility, interoperability and reusability. Long-term preservation entails migration to new software versions and reflecting advancements in technology as well as responding to security issues; increase in projects and services require a more elaborate and stable system architecture. This submission describes the conceptual framework and architecture of a large digital repository infrastructure that specialises in the preservation, management and presentation of research data in the Humanities, as well as the maintenance and migration of this exemplary research data ecosystem.

1 Introduction

Domain-specific digital repositories rely on specialised design principles and a strategic orientation to balance the need for long-term sustainability and the use of advanced technologies. This proposal describes the conceptual framework and architecture of GAMS (Geisteswissenschaftliches Asset Management System), a large digital repository infrastructure maintained by the Centre for Information Modelling – Austrian Centre for Digital Humanities (University of Graz) that specialises in the preservation, management and presentation of research data in the Humanities. Further, issues of maintenance and migration of this exemplary ecosystem are discussed.
For almost 20 years, the digital repository GAMS has been providing long-term preservation of Humanities' resources offering both a technical solution and a way of achieving sustainability in the handling of research data. Currently, the repository contains about 115.000 annotated digital objects in more than 60 projects. This ranges from digital scholarly editions or text collections (e.g. letters, medieval accounting books) to image collections (e.g. historical photographs and postcards) as well as digitised cultural heritage collections (e.g. archaeological artefacts or museum objects)[1].

2 Design Principles

The OAIS-compliant[2] repository follows a largely XML-based content strategy utilising domain-specific data models and recognised international standards[3]. All projects are realised within the same infrastructure, using a limited set of preferred formats and technologies to ensure maintainability over a long period.
One of the unique features is the management of digital resources as complex digital objects based on content models (classes) for specific research data types, including any number of data streams (properties), partly automatically generated metadata records and disseminators (methods), specified in the WSDL[4]. These disseminators are implemented as services in the sense of service-oriented architectures[5].

[1] Zentrum für Informationsmodellierung 2021.
[2] Space Data Systems 2012.
[3] TEI (TEI Consortium 2021), LIDO (ICOM 2010), SKOS (Miles and Bechhofer 2009), RDF (Brickley and Guha 2014), EDM (Europeana Foundation 2014), Dublin Core (Dublin Core Metadata Initiative 2014) etc.
[4] Chinnici et al. 2007.
[5] Erl 2016.

Pipeline configuration is available for data storage, metadata extraction and dissemination strategies. The digital representation of a medieval manuscript for example consists of descriptive, technical, structural and administrative metadata, a plurality of facsimiles, a TEI-encoded transcription, and RDF representations to support linked data. Methods bound to the object allow for synoptic presentations of multiple variants, visualizations and provide project specific search and analysis functions. Further content models for common Humanities' data are natively implemented in the infrastructure, examples are 3D, LIDO, SKOS/OWL/RDF, RTI, METS/MODS. All data streams and methods are stored and managed within a single completely self-descriptive object, addressable by a unique identifier, which offers significant advantages for preservation.

The ecosystem solely relies on open source software due to its commitment to open access and open science[6]. The design principles intend to foster compliance with international best practice like the COAR Community Framework for Good Practices in Repositories[7] and certification guidelines (GAMS is currently certified with the CoreTrustSeal and as a CLARIN B-Centre), following the FAIR principles of findability, accessibility, interoperability and reusability[8].

3 Migration: Reaching the next Level of the Infrastructure

Long-term preservation entails migration to new software versions and reflecting advancements in technology as well as responding to security issues; increase in projects and services require a more elaborate and stable system architecture.

The new version of the core system is based on Fedora 6, which implies a fundamental switch from the object-oriented paradigm to a linked data architecture based on the Linked Data Platform (LDP)

[6] Core technologies are Fedora Commons (currently version 3, after migration version 6) (Lagoze et al. 2005) for storage and management of digital objects, Apache Lucene (Apache Software Foundation 2011-2021) and Apache Solr (Apache Software Foundation 2021) for full text search, Blazegraph Triplestore (Systap 2018) as graph database, RDF4J (Eclipse RDF4J 2021), PostgreSQL Database Server 1996-2016 as relational database, Spring (Spring 2021), Apache Cocoon (Apache Software Foundation 2008) as platform for web services used as object disseminators, and Loris IIIF Image Server (Stroop 2013-2015) to provide access to images via the IIIF Image API (Appleby 2012-2021).

[7] COAR 2020.

[8] Force 11 2011-2020.

specification[9] and a stronger focus on restful APIs for services[10]. This change, especially the disappearance of the content model-based approach, posed considerable challenges[11]. The former object-oriented model had to be adapted to the new graph-based model without violating existing data and interfaces. Special attention had to be paid to conservation and re-implementation of the interfaces already in use to guarantee that all surface components will function without adaptations. The system was massively overhauled at the infrastructure level. The individual service components now run in isolated Docker containers in a Kubernetes-driven cluster. This allows for better scaling and shorter downtimes. With regard to storage infrastructure the new version relies on Ceph, a redundant, distributed and almost infinitely scalable storage solution. The effective migration to the new version is scheduled for the second half of 2021 given that Fedora 6 was recently released[12].

4 Conclusion

Running a digital repository of complex data from the Humanities has to take into account that data and software changes over time. Deciding on a limited list of preferred formats and standards and choosing an appropriate underlying software are thus important factors with regard to long-term preservation and sustainable data curation. Replacing core components of the software infrastructure can be seen as the ultimate proof of a successful long-term strategy.

5 Bibliografie

Apache Software Foundation. 2008. Apache Cocoon. http://cocoon.apache.org/ (accessed 26 July 2021).

Apache Software Foundation. 2011-2021. Apache Lucene. http://lucene.apache.org/ (accessed 26 July 2021).

Apache Software Foundation. 2021. Apache Solr. https://solr.apache.org/ (accessed 26 July 2021).

[9] Speicher, Arwe and Malhotra 2015.
[10] Fielding 2000.
[11] DuraSpace 2016.
[12] Lyrasis 2021.

Appleby, M., Sanderson, R., Snydman, S., Stroop, J. and Warner, S. 2012-2021. IIIF Image API 2.0. http://iiif.io/api/image/2.0/ (accessed 26 July 2021).

Brickley, D., Guha, R. V. 2014. RDF Schema 1.1. https://www.w3.org/TR/rdf-schema (accessed 26 July 2021).

Chinnici, R., Moreau, J., Ryman, A. and Weerawarana, S. 2007. Web Services Description Language (WSDL) Version 2.0. https://www.w3.org/TR/wsdl20/ (accessed 26 July 2021).

Confederation of Open Access Repositories (COAR). 2020. COAR Community Framework for Good Practices in Repositories. https://doi.org/10.5281/zenodo.4110829 (accessed 26 July 2021).

Dublin Core Metadata Initiative. 2014. DCMI Metadata Terms. http://dublincore.org/documents/dcmi-terms (accessed 26 July 2021).

DuraSpace. 2016: Design - API Extension Architecture. https://wiki.duraspace.org/display/FF/Design+-+api+extension+architecture (accessed 26 July 2021).

Eclipse RDF4J 2021: Scalable RDF for Java https://rdf4j.org/ (accessed 26 July 2021).

Erl, T. 2016. Service-Oriented Architecture: Analysis and Design for Services and Microservices. Upper Saddle River, NJ: Prentice Hall.

Europeana Foundation. 2014. Europeana Data Model Documentation. https://pro.europeana.eu/page/edm-documentation (accessed 26 July 2021).

Fielding, R.T. 2000. Architectural Styles and the Design of Network-based Software Architectures. Ph.D. thesis, University of California, Irvine.

Force 11. 2011-2020. The Fair Data Principles. https://www.force11.org/group/fairgroup/fairprinciples (accessed 26 July 2021).

ICOM. international committee for documentation. 2010. Lightweight Information Describing Objects – LIDO. http://network.icom.museum/cidoc/working-groups/lido/lido-technical/specification/ (accessed 26 July 2021).

Lagoze, C., Payette, S. and Shin, E. 2005. Fedora: An architecture for complex objects and their relationships, Berlin-Heidelberg, http://arxiv.org/ftp/cs/papers/0501/0501012.pdf (accessed 26 July 2021).

Lyrasis. 2021. Fedora 6.0.0. https://duraspace.org/fedora/download/ (accessed 26 July 2021).

Miles, A. and Bechhofer, S. 2009. SKOS Simple Knowledge Organization System Reference. In: W3C (eds.). https://www.w3.org/TR/2009/REC-skos-reference-20090818 (accessed 26 July 2021).

Speicher S., Arwe J., Malhotra, A. 2015. Linked Data Platform 1.0. https://www.w3.org/TR/ldp/ (accessed 26 July 2021).

Spring 2021: https://spring.io/ (accessed 26 July 2021).

Stroop, J. et al. 2013-2015. Loris IIIF Image Server. https://github.com/loris-imageserver/loris (accessed 26 July 2021).

Systap. 2018. Blazegraph. https://www.blazegraph.com (accessed 26 July 2021).

TEI Consortium. 2021. TEI P5: Guidelines for Electronic Text Encoding and Interchange. Version 4.2.2. https://www.tei-c.org/Vault/P5/4.2.2/doc/tei-p5-doc/en/html/ (accessed 26 July 2021).

The PostgreSQL Global Development Group. 1996-2016. PostgreSQL 9.1.24 Documentation. https://www.postgresql.org/docs/9.1/index.html (accessed 26 July 2021).

The Consultative Committee for Space Data Systems. 2012. Reference Model for an Open Archival Information System (OAIS). https://public.ccsds.org/pubs/650x0m2.pdf (accessed 26 July 2021).

Zentrum für Informationsmodellierung – Austrian Centre for Digital Humanities. 2021. GAMS. Geisteswissenschaftliches Asset Management System. http://gams.uni-graz.at (accessed 26 July 2021).

Forschungs-Alltags-Infrastruktur

Fisseni, Bernhard
bernhard.fisseni[at]uni-due.de
Universität Duisburg-Essen, Deutschland

Arnold, Denis
arnold[at]ids-mannheim.de
Leibniz-Institut für Deutsche Sprache, Mannheim, Deutschland

Lang, Christian
lang[at]ids-mannheim.de
Leibniz-Institut für Deutsche Sprache, Mannheim, Deutschland

Zusammenfassung. In unserem Beitrag diskutieren wir Aspekte einer Forschungsdateninfrastruktur für den wissenschaftlichen Alltag auf Projektebene und argumentieren für eine Unterstützung von Projekten während der Erfassung und Bearbeitung von Daten, d. h. vor deren endgültiger Veröffentlichung. Dabei differenzieren wir zwischen Projekten, deren primäres Ziel es ist, eine Ressource aufzubauen (*ressourcenschaffende Projekte*, kurz RP) und solchen, die zur Beantwortung einer konkreten Forschungsfrage Daten sammeln und auswerten (*Forschungsprojekte*, kurz FP). Wir argumentieren dafür, dass bei den offenkundigen Unterschieden zwischen beiden Projektarten die grundsätzlichen Ansprüche an das alltägliche Forschungsdatenmanagement im Kern sehr ähnlich (wenn auch unterschiedlich akzentuiert und skaliert) sind. Diese Ähnlichkeit rührt nicht zuletzt daher, dass im Rahmen von FP gesammelte Daten in Bezug auf das Projektziel primär Mittel zum Zweck sein mögen, sie jedoch bereits im Arbeitsprozess in unterschiedlichem Maß von unterschiedlichen Beteiligten genutzt werden. Wir gehen konkret auf die Aspekte **Datenorganisation und -verwaltung**, **Metadaten**, **Dokumentation** und **Dateiformate** und deren Anforderungen in den verschiedenen Projekttypen ein. Schließlich diskutieren wir Lösungsansätze dafür, Aspekte des Forschungsdatenmanagements auch in (kleineren) Forschungsprojekten nicht post-hoc, sondern bereits in der Projektplanung als Teil der alltäglichen Arbeit zu berücksichtigen und entsprechende Unterstützung in der Forschungsinfrastruktur vorzusehen.

1 Problemstellung

Die wesentliche These unseres Beitrags ist, dass vielen Forschenden mit großen, allgemeinen Forschungsinfrastrukturen, wie sie gerade ent-

wickelt werden, nur partiell geholfen ist: Diese zielen auf Verarbeitungspipelines für standardisierte Aspekte bestimmter Forschungsschritte und auf Veröffentlichungsworkflows für deren Abschluss. Auch die Beratung von Forschungsdatenmanagementabteilungen in Universitäten oder Forschungsinstituten zielt häufig auf den Umgang mit Daten im Nachhinein. Der Umgang mit Daten im Forschungsalltag bleibt von den Angeboten im Regelfall unberührt. Informationsangebote wie https://www.forschungsdaten.info/ helfen beim Einstieg ins Thema Forschungsdatenmanagement, geben aber keine Hilfestellung bei konkreten Fragestellungen.

Im *Forschungsalltag* stehen projektspezifische Anforderungen im Vordergrund und Daten werden in Arbeitsformaten gespeichert, die unter Umständen nicht interoperabel sind, aber die Informationen für die Arbeit im Projekt optimal bündeln. Herausforderungen treten bei der Datei- und Dokumentverwaltung auf, beim Austausch über den eigenen Festplattenrand hinaus und bei der ‚Vorveröffentlichung' von Daten, aber auch bei der Vorbereitung der Integration der Daten in große Infrastrukturen und Nutzung von Verarbeitungspipelines. Es geht also um die Ausgestaltung des *collaborative working space*,[1] allerdings in Hinblick auf Perspektiven der Datenhaltung und des Forschungsdatenmanagements, idealerweise ergänzt durch institutsübergreifende *Virtual Research Environments (VRE)*.[2]

Für diese Herausforderungen gibt es bekannte Lösungen, aber deren Zusammenstellung und Implementierung ist gerade für kleine Projekte – und um diese geht es hier hauptsächlich – nicht trivial und stellt einen beträchtlichen Zusatzaufwand dar. Unterstützung in diesem Bereich würde insbesondere kleineren geisteswissenschaftlichen Projekten helfen, die personell eher technikfern sind und sich auf inhaltliche Aufgaben der Forschung bzw. der Aggregation und Kuration von Daten konzentrieren.[3]

---

1 Wissik & Ďurčo 2016.

2 Vgl. z. B. Candela, Castelli & Pagano 2013.

3 Ein anonymes Gutachten weist darauf hin, dass alle Probleme, die dieser Beitrag benennt, bereits gelöst seien und nicht mehr auftreten dürften. Ein anderes Gutachten stimmt der Problembeschreibung vollumfänglich zu. Dies sehen wir als Hinweis darauf, dass sich der Forschungsalltag in verschiedenen Fachdisziplinen und/oder Institutionen darin unterscheidet, inwieweit Überlegungen zum Forschungsdatenmanagement Berücksichtigungen finden, und als Bestätigung dafür, dass in der Community breiterer Diskurs notwendig ist.

2 Zwei Pole

Wir unterscheiden solche Projekte, die Material rekursiv erschließen und zur Verfügung stellen (*ressourcenschaffende Projekte*, kurz RP), von solchen Projekten, die eine konkrete Forschungsfrage bearbeiten, wobei nebenher Daten anfallen, die Nachnutzungspotential haben (*Forschungsprojekte*, kurz FP).[4] Beider Bedarf ist in einigen Punkten ähnlich, aber die Schwerpunkte sind sehr unterschiedlich. (Viele Projekte liegen zwischen beiden Extremen.) Zunächst scheint offensichtlich, dass RP grundsätzlich Überlegungen zum Forschungsdatenmanagement anstellen müssen, für FP scheint dies Forschenden jedoch nicht immer nötig. Unsere Erfahrung legt jedoch nahe, dass die Unterschiede zwischen beiden Arten von Projekten geringer sind, als man denken könnte; im Folgenden erläutern wir Aspekte einer Infrastruktur für den Forschungsalltag.

3 Aspekte einer Alltagsinfrastrukturnutzung

Eine Alltagsinfrastruktur muss eine breitere Perspektive auf Datennutzung einnehmen, als durch den Begriff *Nachnutzung* abgedeckt wird. Nachnutzung bezeichnet die Nutzung nach der in einem FP vorgesehenen Nutzung. Für RP, bei denen keine Nutzung definiert ist, ist der Begriff inadäquat. Gemeinsam ist jedoch, dass es ein Kontinuum von Nutzbarkeit (und Nachnutzbarkeit) gibt bezüglich der Breite, in der Daten geteilt werden; folgende Gradierung bietet eine Orientierung:

1. Daten können innerhalb des Projekts nutzbar sein, wobei sich Projekte auch häufiger über mehrere Abteilungen und Institutionen erstrecken,
2. Daten können mit Kooperationspartner:innen geteilt werden,
3. Daten können von Forscher:innen genutzt werden, die (a) in einem entfernteren Feld arbeiten oder (b) andere Forschungsfragen verfolgen.

Es ist offensichtlich, dass (3a) insbesondere für RP relevant ist. Da RP grundsätzlich Ressourcen erstellen, die allgemein genutzt werden, sind (3a) und (3b) dort generell im Blick. Bei FP stehen (1) und (2) im Vordergrund. (3) erscheint oft als nebensächlich, dennoch gewinnen z. B. Metastudien, aber auch weitere Formen der Nachnutzung immer mehr an Bedeutung. RP und FP benötigen jedoch dieselben infrastrukturellen Komponenten zur Daten- und Informationsorganisation:

[4] Flanders und Jannidis (2015) verwenden eine ähnliche Unterscheidung in *curation-driven* und *research-driven*.

Zunächst werden grundsätzliche Techniken und Werkzeuge der **Datenorganisation und -verwaltung** benötigt: zur Versionierung und Dateibenennung und generell zur Verwaltung zusammengehöriger Daten (Dokumentenmanagement, ggf. über ein Repositorium oder eine Datenbank).
Ohne **Metadaten** sind Daten nicht auffindbar und im Detail nicht nutzbar. Metadaten werden in Infrastrukturen oft auf die gesamten Daten eines Projekts bezogen, also zum Beispiel ein ganzes Korpus, eine Bildersammlung, eine Gesamt-Edition. Im Forschungsalltag aber auf einzelne Teile, etwa ein einzelnes Bild, einen einzelnen Brief oder eine Zeitungsmeldung. Für die Verwendung in Repositorien oder für die Integration in Infrastrukturen sind je nach Zielgruppe sehr verschiedene Formate gefragt, die auch einen sehr großen Gestaltungsspielraum bieten.
Weiterhin müssen Daten und Verfahren dokumentiert werden. Bei RP steht die **Dokumentation** der Daten und der erfolgten Aufbereitung im Vordergrund; bei FP die Dokumentation der erfolgten Forschung, ggf. im Sinne einer Qualitätskontrolle (Reproduzierbarkeit). Hierher gehört auch die Dokumentation von Workflows für die N(achn)utzung. Beides benötigt jedoch dieselbe Art Werkzeuge und Vorgaben. In den Naturwissenschaften gibt es die Tradition von Laborbüchern, die in den Geisteswissenschaften so nicht besteht. Für diese werden zunehmend digitale Implementierungen erarbeitet, die sich die Geisteswissenschaften eventuell an die eigenen Bedürfnisse anpassen könnte.
Ein letzter wichtiger Aspekt der Dateiorganisation sind **Dateiformate**. Je nach Fach sind verschiedene Formate im Gebrauch. Verschiedene Workflows und Werkzeuge erfordern verschiedene Formate. Zum Beispiel verwendet das CLARIN-Angebot WebLicht ein eigenes Dateiformat (Text Corpus Format[5]) für seine Pipelines, während z. B. CLARIAH-DE sich auf das DTA-Basisformat[6] als ein wesentliches Pivot-Format festgelegt hat. Verschiedene Formate enthalten unter Umständen nicht dieselbe Information; in einem uns bekannten Projekt wurden daher drei nicht inhaltlich deckungsgleiche Varianten derselben Daten abgelegt.

4 Aspekte einer Alltagsinfrastruktur

Die folgenden Aspekte verstehen sich als Diskussionsanregungen im Sinne der Tagungsleitlinien. Im erlaubten Umfang ist es natürlich nicht möglich, eine umfängliche Lösung zu skizzieren.

5 Siehe WebLicht Wiki (2021).
6 Haaf, Geyken & Wiegand 2014.

Digitale Aspekte der Kuration (Versionierung, Metadatenmodell) von Ressourcen werden nicht überall als natürlicher Bestandteil einer Ressourcenerstellung mitgedacht, sondern als zusätzliche oder nachträgliche Punkte betrachtet, da fachliche Überlegungen im Vordergrund stehen und Dateien unter Umständen als weißes Blatt Papier wahrgenommen werden, das menschenlesbar, nicht unbedingt maschinenlesbar befüllt werden muss. Daher braucht es für Fachwissenschaftler:innen gezielte Fortbildungsangebote; diese können für Wissenschaftler:innen in Ausbildung auch in die Curricula integriert werden.
Konkrete Umsetzungen der Konzepte und der Aufbau lokaler Systeme geht aber für viele kleine Projekte über das Leistbare hinaus, da in FP und sogar in RP grundsätzlich fachliche Kompetenzen im Vordergrund stehen müssen und die Entwicklungen insbesondere im technischen Bereich des FDM zu schnell sind, um ausschließlich ‚nebenher' verfolgt zu werden. Wichtig sind daher Anlaufstellen in spezifischen Bereichen, die an der Schnittstelle zwischen Forschungsdatenmanagement und Forschungsalltag den Überblick behalten und dadurch beraten können. Ein Beispiel für die Sensibilisierung ist EdMa, die Editionsmatrix,[7] aus dem CLARIAH-DE-AP1: Sie bietet die Möglichkeit, Editionen zu klassifizieren und daraus Empfehlungen für die Datenhaltung (Format, Konvertierungen) abzuleiten.

Bibliografie

Candela, Leonardo, Donatella Castelli, and Pasquale Pagano. 2013. "Virtual Research Environments: An Overview and a Research Agenda." *Data Science Journal*, GRDI–013.

Julia Flanders and Fotis Jannidis. 2015. Knowledge Organization and Data Modeling in the Humanities. White paper. https://wwp.northeastern.edu/outreach/conference/kodm2012/index.html.

Haaf, Susanne, Alexander Geyken, and Frank Wiegand. 2014. "The DTA 'Base Format': A TEI Subset for the Compilation of a Large Reference Corpus of Printed Text from Multiple Sources." *Journal of the Text Encoding Initiative* 2014/15.

[7] Schulz, Fisseni & Sendler 2021.

Schulz, Daniela, Bernhard Fisseni und Simon Sendler. 2021. „EdMA: eine Matrix zur Kategorisierung digitaler Editionen.“ CLARIAH-DE-Arbeitsberichte Nr. 5. CLARIAH-DE. https://doi.org/10.14618/ids-pub-10501.

WebLicht Wiki. 2021. “The TCF Format.” 2021. https://weblicht.sfs.uni-tuebingen.de/weblichtwiki/index.php/The_TCF_Format.X.

Wissik, Tanja, and Matej Ďurčo. 2016. “Research Data Workflows: From Research Data Lifecycle Models to Institutional Solutions.” In *Selected Papers from the CLARIN Annual Conference 2015, October 14–16, 2015, Wrocław, Poland*, edited by Koenraad De Smedt, 94–107. Linköping: Linköping University Electronic Press.

Forschungsdaten.info – FDM-Angebote für die Geisteswissenschaften auf einen Blick

Böker, Elisabeth
elisabeth.boeker[at]uni-konstanz.de
bw2FDM, KIM, Universität Konstanz, Deutschland
ORCID-iD: 0000-0002-6025-3144

Albrecht, Regina
regina.albrecht[at]ub.uni-muenchen.de
Universitätsbibliothek der LMU München, Deutschland

Gabriel, Vanessa
vanessa.gabriel[at]ub.uni-muenchen.de
Universitätsbibliothek der LMU München, Deutschland
ORCID-iD: 0000-0002-2058-5160

Meier, Laura
laura.meier[at]ub.uni-muenchen.de
Universitätsbibliothek der LMU München, Deutschland
ORCID-iD: 0000-0003-1368-2306

Spenger, Dr. Martin
martin.spenger[at]ub.uni-muenchen.de
Universitätsbibliothek der LMU München, Deutschland
ORCID-iD: 0000-0002-8841-5985

Brettschneider, Peter
peter.brettschneider[at]uni-konstanz.de
bw2FDM, KIM, Universität Konstanz, Deutschland
ORCID-iD: 0000-0003-0551-1363

Axtmann, Alexandra
alexandra.axtmann[at]kit.edu
bw2FDM, KIT-Bibliothek, Deutschland
ORCID-iD: 0000-0001-5303-5352

Zusammenfassung. Forschungsdaten.info ist das deutschsprachige Informationsportal zu Forschungsdatenmanagement (FDM). Mit

praxisnahen Artikeln führt die Seite an das Thema heran. Die Plattform wird von einem überregionalen Team von Spezialistinnen und Spezialisten betreut und bietet eine Vielzahl von Einstiegsmöglichkeiten in das komplexe Thema FDM. Mit dem Menü kann direkt über ein bestimmtes FDM-Thema, einen Wissenschaftsbereich oder über regionale Angebote in die Recherche einsteigen. Zusätzlich gibt es den Menüpunkt "Praxis kompakt", der bei Fragen in der alltäglichen Arbeit schnell zu einer Antwort führen soll."

Ein zentrales und etabliertes Angebot ist dabei der Menüpunkt „Wissenschaftsbereiche". Mit der speziellen Fachunterseite „Geisteswissenschaften" bietet forschungsdaten.info einen umfangreichen und informativen Überblick über generische und spezifische FDM-Angebote in den geisteswissenschaftlichen Disziplinen. Die Unterteilung der Fachseite „Geisteswissenschaften" erfolgt in vier Bereiche: „NFDI", „Projekte, Initiativen und Netzwerke", „Repositorien und Datenjournale" sowie „Tools und Services". Auf der Fachseite werden sowohl aktuelle Informationen berücksichtigt als auch zusammen mit Forschenden neue Inhalte zu Forschungsprojekten und forschungsunterstützenden, praxisnahen Hilfsmitteln erstellt. Zu den Hilfsmaterialien zählen beispielsweise nachnutzbare Schulungsmaterialien, die speziell für die Geisteswissenschaften erstellt wurden, sowie Links zu internationalen Ressourcen. Trotz der deutschsprachigen Ausrichtung der Plattformen wird ein globaler Fokus eingenommen und es werden nationale und internationale Standards berücksichtigt, die auf die geisteswissenschaftlichen Fächer anwendbar sind. Ebenfalls bietet forschungsdaten.info disziplinübergreifende Informationen zum FDM an, bspw. in Form von FAQs, Veranstaltungsreihen oder einem Quiz aus dem Menüpunkt „Praxis kompakt". Das Poster zeigt, welche Ressourcen forschungsdaten.info für die Geisteswissenschaften bietet und wie die geisteswissenschaftliche Community Inhalte beisteuern kann.

1 Langabstract Poster

Forschungsdaten.info ist das zentrale deutschsprachige Informationsportal zu Forschungsdatenmanagement (FDM). Mit praxisnahen Artikeln führt die Seite ins FDM ein. Die Seite gliedert sich in die Menüpunkte: „Themen", „Wissenschaftsbereiche", „FDM im deutschsprachigen Raum" und „Praxis kompakt". Die Beiträge im Menüpunkt „Themen" umfassen dabei die Schritte des Forschungsdatenlebenszyklus von der Antragsplanung eines Forschungsprojekts, der Arbeit mit Forschungsdaten im Forschungsalltag, der Umsetzung des Antrags bis hin zur Publikation und Nachnutzung der erzeugten Daten. Auch Rechte und Pflichten im Umgang mit Forschungsdaten werden behandelt. Zusätzlich liefern Best-Practice-Beispiele und Informationsmaterialien aus den

einzelnen Wissenschaftsbereichen Anregungen, um Daten besser (nach-)nutzbar zu machen.

Der umfassende Ansatz von forschungsdaten.info ist – gerade auch im Vergleich zu anderen FDM-Angeboten – eine Besonderheit: 30 FDM-Expertinnen und Experten aus dem D-A-CH-Raum erstellen die Inhalte und tauschen sich bei monatlichen Redaktionstreffen aus, um das Angebot auszubauen. Die Seite dient einerseits als Anlaufstelle für erste Fragen zum Bereich FDM und bietet gleichzeitig gebündelt und gut strukturiert Informationen zu speziellen Anliegen. Dabei wird auf die Angebote mit redaktionell erstellten Texten und Verlinkungen verwiesen. Ein zentrales Angebot, welches sehr sich bereits sehr gut in der Community etabliert hat, ist dabei der Menüpunkt „Wissenschaftsbereiche". Mit der dort enthaltenen Fachseite „Geisteswissenschaften" bietet forschungsdaten.info einen umfangreichen und informativen Überblick über generische und spezifische FDM-Angebote in den geisteswissenschaftlichen Disziplinen.

Charakteristisch für das Forschungsdatenmanagement in den Geisteswissenschaften ist dabei die Heterogenität der materiellen und immateriellen Forschungsgegenstände und damit der im Forschungsprozess entstehenden Daten und deren Formate. Darunter fallen beispielsweise Kunst- und Kulturobjekte, Personen, Geodaten, digitalisierte Quellen und annotierte Texte. Vor dem Hintergrund dieser Vielfältigkeit an Objekten und der beschreibenden Metainformationen bedarf es ausdifferenzierter FDM-Prozesse, die der Vielfältigkeit der Fachdisziplinen gerecht werden. Gleichzeitig werden Wege aufgezeigt, um geisteswissenschaftliche Forschungsdaten in strukturierter Form zu aggregieren und in übergeordnete Suchinfrastrukturen zu integrieren.

Forschungsdaten.info bietet dafür mit der Fachseite „Geisteswissenschaften" einen umfangreichen, klar gegliederten und informativen Überblick. Dort werden beispielsweise die bislang geplanten bzw. bereits geförderten „NFDI-Konsortien" vorgestellt. Unter „Projekte, Initiativen und Netzwerke" finden geisteswissenschaftliche Forschende die wichtigsten Akteure aus dem Bereich auf einen Blick – dazu zählen auch die Angebote der Fachinformationsdienste zu FDM-Themen. „Repositorien und Datenjournale" gibt einen Überblick über fachspezifische Angebote, die vor allem für die Datenpublikation oder für die Suche von Daten zur Nachnutzung relevant sind. Vervollständigt wird das Angebot mit

„Tools und Services“, das Schulungsmaterialien und Unterlagen zum Selbststudium vereint.

Ein zentraler Bestandteil von forschungsdaten.info ist die Einbeziehung von Expert*innen aus den Disziplinen wie auch der FDM-Community. Daraus sind bereits erste Kooperationen zwischen forschungsdaten.info und Initiativen aus dem geisteswissenschaftlichen FDM-Bereich entstanden. Zu nennen ist hier beispielsweise das Modellprojekt "eHumanities - interdisziplinär", innerhalb dessen neue Dienstleistungen rund um das FDM in den Digital Humanities entwickelt werden. Ergebnisse aus dem Projekt werden auch auf forschungsdaten.info präsentiert.

Bereits jetzt ist forschungsdaten.info ein zentraler Anlaufpunkt für Informationssuchende zum Thema FDM. Gegeben ist ein umfassendes Informationsangebot, welches eine wichtige Ressource auch speziell für Studierende und Forschende aus den (digitalen) Geisteswissenschaften darstellt. Ein weiterer großer Vorteil der Seite ist dabei, dass sowohl eigene Artikel veröffentlicht werden als auch eine kuratierte Sammlung von externen Quellen und Angeboten verfügbar gemacht wird. Forschungsdaten.info richtet sich neben Forschenden auch an Infrastruktureinrichtungen, (Verbund-)Projekte und Lehrende, die FDM-Inhalte in ihre Beratungs- und Informationsangebote sowie Veranstaltungen integrieren möchten. Eine Nachnutzung der Inhalte wird durch freie Lizenzen – die Seite verwendet CC0 – unkompliziert ermöglicht.

Mit dem Poster wollen wir bei der FORGE 21 zum einen auf die Informationsplattform hinweisen und diese vorstellen. Zum anderen ist es uns ein Anliegen, den Austausch darüber innerhalb der Fachcommunity zu ermöglichen und das Angebot mit ihr weiterzuentwickeln. Wir laden daher die Community sehr gerne dazu ein, uns Ergänzungen für die bestehenden Kategorien vorzuschlagen, um eine möglichst vollständige Kartierung der geisteswissenschaftlichen FDM-Landschaft in Zukunft auf forschungsdaten.info anzubieten.

Fun with VCS - more with less

A Tool for Facilitating the Use of Git in Linguistic Research Data Management

Ferger, Anne
anne.ferger[at]uni-paderborn.de
Universität Paderborn, Deutschland
ORCID-iD: 0000-0002-1382-2658

Jettka, Daniel
daniel.jettka[at]uni-paderborn.de
Universität Paderborn, Deutschland
ORCID-iD: 0000-0002-2375-2227

Introduction. The work with digital data demands some form of version control. In its simplest and still common appearance version control consists of file naming conventions, and often some creative form of duplication of files. The rising degree of collaboration in data creation and complexity of data, however, brings these basic methods to their limits, which is one of many practical and theoretical reasons why explicit and structured version control with the help of a version control system (VCS) is required. The introduction and use of a VCS for research data in the humanities faces the obstacle of a steep learning curve and poses severe problems with regard to usability, since VCS implementations mostly originate from software engineering. To deal with this problem we discuss the open source script LAMA which is used to perform operations with the VCS Git. Using LAMA does not require in-depth knowledge of the VCS or additional technical expertise, so the focus of the individual researcher can remain on the creation of the research data itself. Additional advantages include messenger integration and the option to run consistency operations with pluggable tools

1 Introduction

Using a version control system (VCS) for organizing and collaborating on research data yields multiple benefits for the quality control, reusability and reproducibility of the created resources. With VCSs originally designed for software development, the acquisition of a certain degree of technical expertise in combination with conceptual and operational specifics are often required for their application. This hampers the application of VCSs in areas which do not have a strong connection to software engineering, like the humanities.

To deal with this problem the open source script LAMA (Linguistic Automation Management Assistant)[1] was created to allow for a neat communication of inexperienced users with the VCS Git[2]. Using LAMA does not require in-depth knowledge of the VCS or additional technical expertise (it runs on Windows, Linux, and Mac), so the focus of the individual researcher can remain on the creation of the research data itself. Additional advantages include messenger integration and the option to run consistency operations with pluggable tools. While the hurdle of using the VCS is minimized for inexperienced users, its advantages with all its functionality can still be leveraged by more technically experienced collaborators.

2 Version control for research data in the humanities

Version control enhances the integrity of research data and can be seen as a prerequisite for producing replicable, high-quality research data[3] and software[4]. It is a technical means to tackle the call for replication of empirical research results in the humanities[5].

2.1 Usability and technical proficiency when employing a VCS

The handling and application of version control should become a natural skill in the digital literacy of empirically working researchers in the humanities and beyond. The transition phase, however, has to be supported by technology and workflows which foster the immediate and correct operation of VCSs and prevent support and maintenance cycles wherever possible.
Most VCSs come with a multitude of clients, often in the form of graphical user interfaces (GUIs)[6] that aim for easier handling of the version control mechanisms, i.e. enhance usability. However, they are also forced to represent the full (or at least most of the) functionality of a VCS because most clients target a broad and discipline-overarching user community.
Many factors favor the most commonly used VCS Git, but there are two main disadvantages, which are the usability and user-friendliness as

[1] Ferger/Jettka, 2021.
[2] Software Freedom Conservancy, Inc. “Git.” - The concept of LAMA could be extended to work with other VCSs as well.
[3] e.g. Klump et al., 2020. and Beckman et al., 2021 for the field of statistics
[4] e.g. Scheliga et al., 2019.
[5] cf. Peels/Bouter, 2018.
[6] e.g. Kernel, “Git Interfaces.”.

well as the handling of binary files[7]. Both play an important role in (linguistic) research data creation and management. While using Git LFS[8] (Large File Storage) can help with the handling of (large) binary files, learning Git and using one of the available GUIs is one of the main hurdles keeping researchers (in linguistics or generally in the humanities) from using Git as a VCS for their research data.

2.2 Meeting the challenges with LAMA

To overcome the usability challenges and the obstacle of a very steep learning curve, a new approach was tested to minimize the communicative problems between users (in this case linguistics researchers) and VCS. The result is the implementation of LAMA. The shell script LAMA serves as a basic Git client adapted to the needs of basic Git users, following minimalistic principles. By focusing purely on the operations needed for (linguistic) research data LAMA facilitates the use of the VCS, while at the same time minimizing the risk of data loss caused by accidental misuse. Dependencies and the visual design also concentrate on the absolute essentials, which enables wider use cases and a straightforward setup of the script.
The user interface of LAMA consists of a text-based menu (see Figure 1) with basic operations[9] sufficient for working collaboratively on (linguistic) research data.

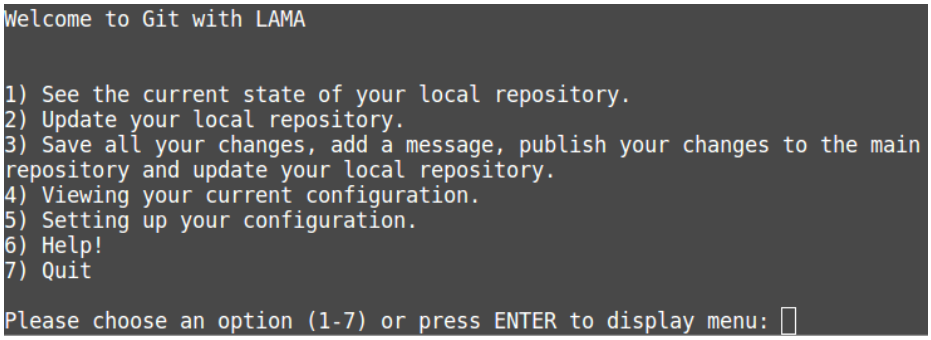


Fig. 1. Screenshot of the LAMA menu.

[7] cf. Hermann et al., 2018: 19.
[8] GitHub, Inc. "Git LFS."
[9] For administration and support purposes a "secret" option 0 is also available which allows for entering any command, e.g. more complex Git operations.

Besides its simplicity, a main advantage of LAMA is the possibility to induce operations into the version control workflow without adding complexity for the users. For instance, a slightly more complex version exists that contains integration with a messenger (tested with Mattermost[10]) to allow for reports or error messages to be sent automatically, which can potentially be used for diagnostic purposes. Furthermore, mechanisms have been successfully implemented for enhanced handling of semi-structured file formats (e.g. XML files) which often pose challenges to VCSs. In this manner, it is possible to add further operations for quality assurance, client-side data cleaning and reporting in LAMA.[11] A German version of LAMA is also available.

3 Use cases for LAMA and further work

While LAMA has already proven useful through its successful application in a research project that produces complex, high-quality linguistic data, some important questions remain. Firstly, it should be discussed if the concept can be transferred directly to other projects, and especially how target groups in the humanities (or elsewhere) can be identified and reached who do not have much contact or direct access to digital/DH experts who could assist with the implementation of reliable version control. Secondly, the actual performance and usability of LAMA would have to be thoroughly tested and compared to other VCS clients. Feedback on those questions, possibly supported by a new feedback option in LAMA, would be highly valuable for the further development.

Currently, there are limits in the complexity of version control operations that are covered. Features of Git which are used in sophisticated data creation workflows and highly collaborative contexts (e.g. branching) are not supported directly (or at least not controllable by the target user). This on the one hand is a conscious design decision (as it clearly adds to the simplicity), but on the other hand certainly poses a limit to the contexts of its application. However, for more advanced operations a multitude of other clients are available.

While in the current version of LAMA automated webhooks are only used for the messenger integration, this option can be used in a much more versatile way to facilitate the integration into any existing infra-

[10] Mattermost, Inc., “Mattermost.”.

[11] Similar quality control methods are available natively in Git, e.g. in the form of Git Hooks. Their main disadvantage is that they have to be configured individually for each user/data repository.

structure, technical setup and data creation workflow. For instance, automatic creation of issues in a project management system or in GitHub[12] could be helpful. By providing automatic issue creation in an issue tracker directly related to the LAMA code, an efficient feedback option could be implemented in the future.

4 References

Beckman, M. D., Çetinkaya-Rundel, M., Horton, N. J., Rundel, C. W., Sullivan, A. J. & Tackett, M. (2021). "Implementing Version Control With Git and GitHub as a Learning Objective in Statistics and Data Science Courses", *Journal of Statistics and Data Science Education*, 29:sup1, S132-S144. https://doi.org/10.1080/10691898.2020.1848485

Ferger, A., Jettka, D. (2021). "LAMA - your friendly and easy git script (3.0)". Zenodo. https://doi.org/10.5281/zenodo.4725651

GitHub, Inc. "GitHub." GitHub. The largest and most advanced development platform in the world. Accessed April 30, 2021. https://github.com

GitHub, Inc. "GitLFS." Git Large File Storage. An open source Git extension for versioning large files. Accessed April 30, 2021. https://git-lfs.github.com/.

Kernel. "Git Interfaces." Git. Interfaces, Frontends, and Tools. Accessed April 30, 2021. https://git.wiki.kernel.org/index.php/InterfacesFrontendsAndTools

Hermann, F., Pietsch, C., & Cimiano, P. (2021). "Conquaire Infrastructure for Continuous Quality Control". *Studies in Analytical Reproducibility: the Conquaire Project*. http://nbn-resolving.de/urn:nbn:de:0070-pub-29517575

Klump, J., Wyborn, L., Downs, R., Asmi, A., Wu, M., Ryder,G. & Martin, J. (2020). "Principles and best practices in data versioning for all data sets big and small". Version 1.1. Research Data Alliance. https://doi.org/10.15497/RDA00042

[12] GitHub, Inc. "GitHub.".

Mattermost, Inc. “Mattermost.” Mattermost. An open-source collaboration tool for developers. Accessed April 30, 2021. https://mattermost.com/

Peels, R., Bouter, L (2018). “The possibility and desirability of replication in the humanities” *Palgrave Commun* **4,** 95. https://doi.org/10.1057/s41599-018-0149-x

Scheliga, K. S., Pampel, H., Konrad, U., Fritzsch, B., Schlauch, T., Nolden, M., zu Castell, W., Finke, A., Hammitzsch, M., Bertuch, O., Denker, M. (2019) “Dealing with research software: Recommendations for best practices”, Potsdam. https://doi.org/10.2312/os.helmholtz.003

Software Freedom Conservancy, Inc. “Git.” Git. Free and open-source distributed version control system. Accessed April 30, 2021. http://git-scm.com/

Generische und disziplinspezifische Zugänge zur Qualität audiovisueller, annotierter Sprachdaten im BMBF-Projekt QUEST

Wamprechtshammmer, Anna
anna.wamprechtshammer[at]uni-hamburg.de
QUEST, Universität Hamburg, Deutschland

Arestau, Elena
elena.arestau[at]uni-hamburg.de
QUEST, Universität Hamburg, Deutschland

Zusammenfassung. In diesem Beitrag stellen wir mit dem BMBF-Verbundprojekt QUEST einen Ansatz zur Bestimmung der Nachnutzbarkeit von audiovisuellen, annotierten Sprachdaten vor. Der Fokus des Vorhabens liegt auf Forschungsdaten, die im Rahmen empirischer Forschung in den Bereichen Sprachdokumentation, Sprachkontakt- und Mehrsprachigkeitsforschung entstehen. Zur Evaluierung des Nachnutzungspotentials solcher Sprachdaten sollen einerseits generisch ausgerichtete datentechnische und dokumentatorische Standards für die verschiedenen relevanten Ressourcentypen sowie deren Metadaten und andererseits disziplinspezifische Kurationskriterien, die auf bestimmte Nachnutzungsszenarien ausgerichtet sind, entwickelt werden. In Bezug auf die sprachwissenschaftliche Sekundärnutzung mehrsprachiger Daten werden für die spezifischen Nachnutzungsszenarien ‚Lernerkorpora' und Korpora gedolmetschter Gespräche beispielsweise Evaluationskriterien für die Übersetzung entwickelt und Fragen nach Metadatenstandards behandelt. Im Rahmen einer Erprobung von Methoden der Qualitätssicherung im Bereich heterogener digitaler Sprachdaten strebt das Forschungsvorhaben an, die Prüfung von Qualitätsstandards und Kurationskriterien als datentechnische Dienstleistung anzubieten. Dazu wird für ausgewählte Ressourcentypen zum Projektende ein System der automatischen kontinuierlichen Qualitätskontrolle implementiert.

1 Einleitung

Wir präsentieren das Projekt "QUEST: Quality – Established: Erprobung und Anwendung von Qualitätsstandards und Kurationskriterien für audiovisuelle, annotierte Sprachdaten"[1], welches als eines von zwölf

[1] https://www.slm.uni-hamburg.de/ifuu/forschung/forschungsprojekte/quest.html.

Projekten von 2019 bis 2022 vom Bundesministerium für Bildung und Forschung gefördert wird, um die Qualität und das Nachnutzungspotential von Forschungsdaten zu verbessern. Mit Fokus auf audiovisuellen, annotierten Sprachdaten erarbeitet das Verbundvorhaben verlässliche Qualitätsstandards und Kurationskriterien. Um Forschende bei der Anwendung der Kriterien und gleichermaßen Datenzentren bei der Beurteilung der Nachnutzbarkeit von digitalen Sprachdaten zu unterstützen, erprobt QUEST darauf aufbauend Verfahren der Qualitätssicherung für die Erstellung und Kuration solcher Ressourcen.
Nach Einordnung des Vorhabens in das Umfeld qualitätssichernder Maßnahmen werden wir unseren Zugang zur Qualitätssicherung audiovisueller, annotierte Sprachdaten vorstellen, indem wir erstens beispielhaft auf Qualitätsstandards und insbesondere Kurationskriterien als Evaluationskriterien eingehen und zweitens die Implementierung der Kriterien im Rahmen qualitätssichernder Maßnahmen darlegen.

2 Hintergrund

Im Zuge der voranschreitenden Digitalisierung spielen qualitätssichernde Maßnahmen in der Behandlung von Forschungsdaten eine immer größere Rolle und werden zunehmend zu den großen Herausforderungen von Forscher*innen und Forschungseinrichtungen gerechnet.[2]
Ausgehend von der vieldiskutierten Frage, wie Datenqualität in Abhängigkeit des jeweiligen Forschungsbereichs messbar wird, nimmt QUEST bezüglich der Beurteilung von Datenqualität das Nachnutzungspotential audiovisueller, annotierter Sprachdaten in den Blickpunkt.
Im Zusammenhang mit der zunehmenden Bedeutung der FAIR-Prinzipien (vgl. Wilkinson et al. 2016) für das Management von Forschungsdaten haben bereits einige Initiativen bzw. Projekte Werkzeuge / Mittel entwickelt und bereitgestellt, um den Grad der FAIRness von Daten manuell und / oder automatisch zu bewerten[3]. Existierende Ansätze, die auf den FAIR-Prinzipien basieren, und es sich neben der Entwicklung von Metriken[4] auch zur Aufgabe gemacht haben, Werkzeuge zur Implementierung der Metriken zu entwickeln, sind in der

[2] vgl. Förderung Kurationskriterien und Qualitätsstandards von Forschungsdaten 2021.
[3] vgl. https: //fairassist.org.
[4] vgl. https://www.go-fair.org/2017/12/11/metrics-evaluation-fairness/.

Regel generisch basiert und zielen auf die Bewertung von Forschungsdaten im Allgemeinen ab. Sie bieten keine detaillierten Anleitungen für das Forschungsdatenmanagement für spezifische Ressourcentypen in Bezug auf bestimmte Disziplinen und referieren lediglich auf Standards einer Community, ohne diese näher zu bestimmen. Der FAIRification-Prozess[5] erfordert jedoch zwingend auch operationalisierbare, ressourcentypspezifische Anforderungen. QUEST setzt an diesem Desiderat an und erarbeitet für audiovisuelle, annotierte Sprachdaten sowohl generische Qualitätskriterien als auch nachnutzungstypspezifische Kurationskriterien, mittels derer das Potential qualitätsgesicherter Daten im Hinblick auf eine interdisziplinäre Nachnutzung oder eine Verwendung der Daten im Rahmen einer Third-Mission sichergestellt werden soll.

3 Zugang zur Datenqualität – generisch und fachspezifisch

Zur Beurteilung der Qualität bzw. Nachnutzbarkeit von audiovisuellen, annotierten Sprachdaten im Sinne einer langfristigen Zugänglichkeit sowie der Öffnung für eine breite wissenschaftliche und nicht-wissenschaftliche Nutzung definiert QUEST im ersten Schritt Evaluationskriterien, um basale Anforderungen an eine digital betriebene Forschung sicherzustellen.
Die Evaluationskriterien nehmen zum einen auf generische Qualitätsstandards Bezug, die unabhängig eines intendierten Nachnutzungsszenarios auf alle Arten audiovisueller, annotierter Sprachdaten zu applizieren sind. Zum anderen wird darauf abgezielt, disziplinspezifsche Kurationskriterien[6] zu erarbeiten, die auf bestimmte Nachnutzungsszenarien zugschnitten sind, d.h. stärker auf die Nachnutzbarkeit in einzelnen Disziplinen wie Sprachdokumentations- oder Mehrsprachigkeitsforschung bezogen sind.
Definierte Qualitätsstandards und Kurationskriterien können als Mindeststandards für die disziplinspezifische Nachnutzbarkeit bei der Konzeption und Begutachtung zukünftiger Projekte herangezogen werden und geben so Auskunft über das Nachnutzungspotenzial von Ressourcen.

[5] vgl. https://www.go-fair.org/fair-principles/fairification-process/.
[6] Kurationskriterien sind operationalisierbare Kriterien, die Aussagen über das Wiederverwendungspotenzial von Ressourcen in bestimmten Disziplinen erlauben.

3.1 Qualitätsstandards und Kurationskriterien

Während das Ziel der Arbeiten zu Qualitätsstandards die Ausarbeitung generisch ausgerichteter Standards ist, erfolgt die Ausarbeitung fachspezifischer Kurationskriterien ausgehend von spezifischen Use Cases. Mit Blick auf konkrete Nachnutzungsszenarien für Forschungsdaten aus den Bereichen Sprachdokumentation, Mehrsprachigkeitsforschung, Gebärdensprache, 'Oral Histroy' und Anthropologie werden sowohl Anforderungen an Daten, deren Struktur und Inhalt im Hinblick auf eine interdisziplinäre Nachnutzung als auch eine Verwendung der Daten im Rahmen einer Third-Mission definiert.
Anhand des Nachnutzungsszenarios 'Lernerkorpora'[7] aus dem Bereich mehrsprachiger Daten soll aus disziplinspezifischer Perspektive beispielhaft dargelegt werden, wie konkrete, operationalisierte Kurationskriterien aussehen könnten und auf welche Weise sie in Abstimmung mit der Community gewonnen werden können, um eine breitestmögliche Akzeptanz von Standards in den betreffenden wissenschaftlichen Zusammenhängen zu erreichen. Bestandsaufnahmen in diesem Bereich und durchgeführte Experteninterviews haben gezeigt, dass beispielsweise für die Nachnutzung und Qualität von Lernerkorpora sogenannte ‚Design-Kriterien' zentral sind.[8] Dazu gehören einerseits eine sorgfältige Auswahl von Daten sowie eine klare Vorstellung von der intendierten Zielgruppe (z.B. die individuellen Voraussetzungen der Lernenden, ihre Muttersprache, Besonderheiten der Spracherwerbssituation und die Einstellungen zum Sprachenlernen). Andererseits spielen auch die Metadaten und eine angemessene Dokumentation, die alle Arbeitsabläufe und Arbeitsschritte bei der Konstitution der Metadaten nachvollziehbar darstellt, eine große Rolle.[9]

3.2 Qualitätssichernde Maßnahmen

Zur fachdisziplinübergreifenden Implementierung der Kriterien wird QUEST zuletzt insbesondere in Kooperation mit den beteiligten Datenzentren Evaluierungsmechanismen zur Verfügung stellen, die den Aufbau und die Aufbereitung audiovisueller, annotierter Sprachdaten im

[7] "A learner corpus […] is an electronic collection of learner produced data formatted for automatic analyses, elicited from L2 or L3/Lx learners or users that provides essential metadata, details critical information on elicitation tasks, and is built around explicit and published design criteria" (Bell / Payant 2020: 54).
[8] vgl. Granger et al. 2016, Tono 2003, Bell / Payant 2020.
[9] vgl. Schütte 2013, Schmidt et al 2013.

Sinne der erarbeiteten Qualitätsstandards und Kurationskriterien leiten und unterstützen und somit schließlich der Feststellung der Einhaltung der Kriterien dienen sollen.
Bereitgestellt wird ein gestaffeltes Instrumentarium bestehend aus Online-Fragebogen und webbasierten Qualitätschecks, das die Bewertung von Sprachkorpora nach vorab definierten Qualitätskriterien in Übereinstimmung mit den FAIR Prinzipien ermöglicht. Auf diese Weise wird Forschenden umfangreiche Unterstützung zur nachhaltigen Datenerstellung und Kuratierung geboten. Insgesamt zielt QUEST darauf ab, auf diese Weise eine Verbesserung existierender und zukünftiger Daten im Hinblick auf die Reproduzierbarkeit von Forschungsergebnissen als auch für die Nachnutzbarkeit im Sinne einer langfristigen Zugänglichkeit zu erreichen.

Bibliografie

Bell, Philippa, Payant, Caroline, "Designing Learner Corpora: Collection, Transcription, and Annotation", in *The Routledge Handbook of Second Language Acquisition and Corpora*, ed. Dans N. Tracy-Ventura et M. Paquot (New York: Routledge, 2020), 53–67).

"Förderung: Kurationskriterien und Qualitätsstandards von Forschungsdaten," *Bundesministerium für Bildung und Forschung (BMBF)*, accessed April 29, 2021. https://www.bildung-forschung.digital/de/foerderung-kurationskriterien-und-qualitaetsstandards-von-forschungsdaten-2281.html.

Granger, Sylviane, Gilquin, Gaetanelle, Meunier, Fanny (2016) (eds.): *The Cambridge handbook of learner corpus research*. Cambridge: University Press.

Schmidt, Thomas, Wörner, Kai, Hedeland, Hanna and Lehmberg, Timm (2013): *Leitfaden zur Beurteilung von Aufbereitungsaufwand und Nachnutzbarkeit von Korpora gesprochener Sprache*, Mannheim, Institut für Deutsche Sprache.

Schütte, Wilfried (2013): Metadaten für Gesprächsdatenbanken: ein Überblick und ihre Verwaltung in der IDS-Datenbank Gesprochenes Deutsch (DGD), in Kratochvílová, Iva / Wolf, Norbert Richard (Hrsg.): *Grundlagen einer sprachwissenschaftlichen Quellenkunde*. Tübingen: Narr (Studien zur Deutschen Sprache 66).

Tono, Yukio (2003): "Learner corpora: Design, development and applications", in Archer, D., Rayson, P., Wilson, A. and McEnery, T. (eds.): *Proceedings of the Corpus* 27 Linguistics 2003 Conference. UCREL technical paper number 16. UCREL, Lancaster University, 800–809), accessed July 29, 2021. http://ucrel.lancs.ac.uk/publications/cl2003/papers/tono.pdf.

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.

NFDI4Objects: Eine NFDI-Konsortialinitiative zum materiellen Erbe aus etwa drei Millionen Jahren Menschheits- und Umweltgeschichte

Keller, Christin
christin.keller[at]dainst.de
Deutsches Archäologisches Institut, Deutschland
ORCID-iD: 0000-0002-5727-0226

Zusammenfassung. Seit 2018 entsteht NFDI4Objects in einem Bottom-up-Prozess und verfolgt das gemeinsame Ziel das Forschungsdatenmanagement in verschiedenen Disziplinen maßgeblich voranzutreiben. NFDI4Objects repräsentiert eine breite Community, die sich mit den materiellen Überresten der Menschheit aus einem Zeitraum von mehr als 2,6 Millionen Jahren beschäftigt. Sie umfasst zahlreiche Fachrichtungen aus den Geistes- und Naturwissenschaften, die durch gemeinsame Anforderungen an das Forschungsdatenmanagement miteinander verbunden sind. Mittlerweile vereint NFDI4Objects alle relevanten Forschungs-, Landes- und Bundeseinrichtungen, Universitäten und Museen, die sich mit dem materiellen Erbe unserer Vergangenheit beschäftigen.

Der Begriff “Objekt", wie er von NFDI4Objects gedacht wird, deckt das gesamte Spektrum der Spuren und Hinterlassenschaften vergangener menschlicher Aktivitäten ab: von Artefakten über Ökofakte, archäologische Befunde und Architektur bis zu anthropogen geformten Landschaften, die durch vielfältige Objektkontexte und komplexe (Objekt-)Biografien eine besondere Herausforderung für das Forschungsdatenmanagement darstellen. So können Objekte nur mit ausreichender Kontextinformation über ihre ursprüngliche Fundsituation, Fundumstände sowie über Erhaltungsbedingungen usw. verstanden werden. Fortwährend verändern sie sich durch Gebrauch, Zerfall und Restaurierung. Das Potenzial Objekte in ihren sich verändernden Kontexten digital zu erforschen ist daher für zahlreiche Disziplinen von zentraler Bedeutung.
Ziel der Nationalen Forschungsdateninfrastruktur (NFDI) ist die systematische Verwaltung von Wissenschafts- und Forschungsdaten, die langfristige Speicherung, Sicherung und Verfügbarkeit sowie die Vernetzung dieser Daten auf nationaler und internationaler Ebene. Die NFDI wird eine Vielzahl von Akteuren in einem koordinierten Netzwerk von Konsortien zusammenführen, die der Aufgabe nachkommen

werden, wissenschaftsbasierte Datendienste für Forschungscommunities bereitzustellen.
NFDI4Objects repräsentiert eine breite Community, die sich mit den materiellen Überresten der Menschheit aus einem Zeitraum von mehr als 2,6 Millionen Jahren beschäftigt. Sie umfasst zahlreiche Fachrichtungen aus den Geistes- und Naturwissenschaften, die durch gemeinsame Anforderungen an das Forschungsdatenmanagement miteinander verbunden sind.
Seit 2018 entsteht das Netzwerk in einem Bottom-up-Prozess. Das gemeinsame Ziel, das Forschungsdatenmanagement in unseren Disziplinen maßgeblich voranzutreiben hat während der Vorbereitung von NFDI4Objects beachtliche Synergien innerhalb unserer Community ausgelöst. Mittlerweile vereint NFDI4Objects alle relevanten Forschungs-, Landes- und Bundeseinrichtungen, Universitäten und Museen, die sich mit dem materiellen Erbe unserer Vergangenheit beschäftigen. Ist das Deutsche Archäologische Institut (DAI) Hauptantragsteller der Konsortialinitiative, besteht der Kreis der Mitantragstellenden aus Forschenden und ausgewiesenen Expert_Innen eines recht heterogenen Forschungs- und Wissenschaftsfeldes. Sie alle eint jedoch, dass sie von einem funktionierenden Datenökosystem in NFDI4Obects immens profitieren würden.
Der Begriff “Objekt", wie er von NFDI4Objects gedacht wird, deckt das gesamte Spektrum der Spuren und Hinterlassenschaften vergangener menschlicher Aktivitäten ab: von Artefakten über Ökofakte, archäologische Befunde und Architektur bis zu anthropogen geformten Landschaften, die durch vielfältige Objektkontexte und komplexe (Objekt-) Biografien eine besondere Herausforderung für das Forschungsdatenmanagement darstellen. So können Objekte nur mit ausreichender Kontextinformation über ihre ursprüngliche Fundsituation, Fundumstände sowie über Erhaltungsbedingungen usw. verstanden werden. Objekte sind also nicht statisch. Sie verändern sich fortwährend durch Gebrauch, Zerfall und Restaurierung. Im Zuge einer Ausgrabung wird zudem ihr Kontext oft unwiederbringlich zerstört. In Aufarbeitungen und Sammlungen werden die Objekte häufig neu kombiniert und geordnet, aber auch immer wieder nach sich ändernden Forschungsparadigmen interpretiert. Das Potenzial, Objekte in ihren sich verändernden Kontexten digital zu erforschen, ist für zahlreiche Disziplinen zentral und könnte vielleicht auch für die Teilnehmenden eine interessante Bereicherung sein.
Das Arbeitsprogramm von N4O wurde bewusst so gestaltet, dass es die vorhandene Expertise nutzt und sich dem aktuellen Forschungsstand

anpasst. Somit hat das Programm seinen Ursprung in aktuellen Standards und Prozessen der analogen Welt und bietet gleichzeitig einen Rahmen, das Potenzial von Arbeitsmethoden der digitalen Zukunft zu erkennen.
Task Area 1 DOCUMENTATION setzt sich aus den Disziplinen zusammen, die direkt an der Erfassung, Entdeckung und Beschreibung von Spuren menschlicher Aktivitäten arbeiten. So wird in der Feldarbeit ein immenser Umfang an Daten generiert. Die im Antrag formulierten Arbeitspakete von Task Area 1 (TA1) orientieren sich zum einen an den notwendigen Schritten zur Harmonisierung von Standards und Spezifikationen von Formaten, Modellen und Metadaten für eine Primärdokumentation. Zum anderen befasst sich TA1 mit den Erfassungstechnologien und der Bereitstellung von Anforderungen zur Integration von Altdaten.
Die in Task Area 2 COLLECTING eingeleiteten Maßnahmen wollen gewährleisten, dass die Informationen über Objektssammlungen der N4O-Community zur Verfügung gestellt werden. So könnten beispielsweise Tontafeln verschiedener Einrichtungen den Forschenden und der Öffentlichkeit zur Verfügung gestellt und in Forschungsprojekten weiter untersucht werden. Folglich muss Task Area 2 (TA2) muss dem allgegenwärtigen Problem des Betriebs lokaler, oft proprietären Sammlungsmanagementsystemen begegnen. Da eine Öffnung der Daten für eine übergreifende infrastrukturelle Aggregation verhindern. In diesem Zusammenhang bearbeitet TA2 Themen zur Provenienz von Sammlungen und entsprechende Anforderungen an die schrittweise Zugänglichkeit.
Task Area 3 ANALYTICS AND EXPERIMENTS ergänzt die Daten um die Ergebnisse naturwissenschaftlicher Analysemethoden, die Informationen zur Herkunft, Herstellung und Nutzung des Objekts hinzufügen. Das Arbeitsprogramm der Task Area 3 (TA3) zielt in erster Linie auf die Planung, Entwicklung und Bereitstellung notwendiger Repositorien und Dienste, die aus bestehenden Datenbanken abgeleitet werden.
Der letzte Schritt in der Biografie eines Objekts ist seine nachhaltige Bewahrung. Diese Aufgabe obliegt den Behörden, wie z.B. den Denkmalschutzämtern, die sich in Task Area 4 PROTECTING engagieren. Sie werden wesentliche Informationen in das Konsortium einbringen und sich an Maßnahmen beteiligen, die Daten zugänglich und interoperabel machen. Expertise in Restaurierung und Konservierung, die einen weiteren wichtigen Forschungsdatenbestand darstellen, wird ebenfalls in Task Area 4 bereitgestellt.

Task Area 5 STORAGE AND ACCESS AND DISSEMINATION entwickelt einen bereits bestehenden Dienst zur Langzeitarchivierung weiter und stellt zuvor in der Community etablierte Terminologiedienste zur Verfügung, die unterschiedliche Vokabulare und Konzepte miteinander verweben. Zu den Maßnahmen gehört auch die Weiterentwicklung wichtiger Zusatzdienste, die den räumlichen und zeitlichen Bezug von Informationen sicherstellen.
Task Area 6 COMMONS AND QUALIFICATION bündelt die fachlich-inhaltliche Qualitätssicherung, die auch die Grundlage für die Community-Arbeit ist.
In Task Area 7 SUPPORT AND COORDINATION werden die operativen Maßnahmen koordiniert und die übergeordneten Themen bearbeitet. Ziel der Task Areas ist es auch die Community stärker im Hinblick auf andere Konsortien sowie national und international einzubinden.

Prozessmetadaten für die digitalen Geisteswissenschaften

Jung, Kerstin
kerstin.jung[at]ims.uni-stuttgart.de
Institut für Maschinelle Sprachverarbeitung (IMS), Universität Stuttgart, Deutschland

Gärtner, Markus
markus.gaertner[at]ims.uni-stuttgart.de
Institut für Maschinelle Sprachverarbeitung (IMS), Universität Stuttgart, Deutschland
ORCID-iD: 0000-0002-2687-4350

Schembera, Björn
schembera[at]hlrs.de
Höchstleistungsrechenzentrum (HLRS), Universität Stuttgart, Deutschland

Zusammenfassung. Prozessmetadaten sind Teil einer nachhaltigen Dokumentation von erzeugten Forschungsdaten und -ergebnissen. Sie dienen dazu, die zum Forschungsprozess getroffenen Entscheidungen möglichst genau festzuhalten und helfen damit den Nachnutzenden zu bewerten, ob sich die Daten für die eigene Forschungsfrage eignen bzw. ob Ergebnisse verglichen werden können. Dabei spielen in den digitalen Geisteswissenschaften sowohl von automatischen Werkzeugen ausgeführte wie auch manuelle Arbeitsschritte eine essentielle Rolle. Im Rahmen einer Abbildung von drei Schemata für Prozessmetadaten aus verschiedenen Projekten wird untersucht, wie spezifisch die Ablaufinformationen für die digitalen Geisteswissenschaften in der Speicherung als Prozessmetadaten sind. Dabei werden die Schemata GRAIN aus der Computerlinguistik, RePlay-DH für die Digital Humanities und EngMeta aus den Ingenieurwissenschaften betrachtet. Allen betrachteten Ansätzen liegt eine Struktur aus Eingabe, Operation und Ausgabe zu Grunde, die eine prinzipielle Abbildung ermöglicht, aber Spielraum für individuelle Ausprägungen lässt und damit jeweils Annahmen für den spezifischen Einsatz enthält.

1 Prozessmetadaten für automatische und manuelle Arbeitsschritte

Für eine nachhaltige Dokumentation ist es nicht nur wichtig, die in einem Forschungsprozess entstandene Forschungsdaten und -ergebnisse abzulegen, sondern auch festzuhalten, wie diese entstanden sind. Solche Prozessmetadaten ermöglichen potentiellen Nachnutzenden zu entscheiden, ob sich eine Ressource für ihre Forschungsfrage eignet und können bei Forschungsergebnissen eine höhere Vergleichbarkeit ermöglichen, auch dann, wenn die eigentlichen Forschungsdaten nicht zugänglich sind. Dabei sind vor allem auch die im ursprünglichen Forschungsprozess nachrangigen Entscheidungen festzuhalten, die, aufgrund ihrer geringen Relevanz für die eigene Forschungsfrage, keinen Eingang in Publikationen finden oder aufgrund technischer Natur zum Zeitpunkt der Erstellung einer Dokumentation nicht mehr in allen Details verfügbar sind. Sollen die Forschungsdaten in anderen Kontexten nachgenutzt werden, können aber gerade kleine, ggf. auch technische Entscheidungen sehr relevant sein, z.B. wenn in der Nachnutzung Frequenzfragen eine Rolle spielen.
Metadatenschemata, die entweder Abschnitte zu Prozessmetadaten enthalten oder vollständig deren Erfassung dienen sind mittlerweile in verschiedenen Disziplinen zu finden und beschreiben Abläufe oft in Form von sequentiell ausgeführten einzelnen Schritten.
Im Folgenden betrachten wir drei Schemata für Prozessmetadaten, die jedem Schritt ein Strukturtrippel aus *Eingabe*, *Operation* und *Ausgabe* zu Grunde legen: EngMeta[1] aus den Ingenieurwissenschaften, GRAIN[2] aus der Computerlinguistik und RePlay-DH[3] mit dem Fokus Digital Humanities.
Prozessmetadaten aus den Natur-, Lebens- und Ingenieurwissenschaften sowie Werkzeuge, die solche Schrittfolgen (Workflows) erfassen, sind oft darauf ausgerichtet, eine automatisierte Ausführung zu ermöglichen (Kepler, n.d.). Im Gegensatz dazu ist es für die digitalen Geisteswissenschaften relevant, automatische und manuelle Schritte gleichrangig zu erfassen und zu behandeln. Den manuellen Schritten sind dabei unter anderem Annotation, Interpretation, Extraktion oder Selektion zuzurechnen, den automatischen Schritten die Ausführung von computationellen Werkzeugen zur Annotation, Analyse oder Visualisierung.

[1] Schembera und Iglezakis 2020.
[2] Jung, Gärtner und Kuhn 2019.
[3] Gärtner 2019.

Die hier betrachtete Abbildung zwischen den drei Schemata soll zeigen, wie ähnlich bzw. unterschiedlich die Annahmen über die benötigten Prozessmetadaten sind und damit, welche Anwendungen sie ermöglichen. Neben der gleichrangigen Erfassung von manuellen und automatischen Schritten wird dabei auch auf die Flexibilität der Strukturinhalte geachtet, so dass z.B. Eingabe und Ausgabe auch leer sein können.

2 An der Abbildung beteiligte Schemata

Die Abbildung erfolgt auf der Ebene der Merkmale für die einfache Werte vergeben werden. Der Pfad zu diesen Merkmalen wird berücksichtigt (z.B. unterscheidet sich “step→actor→name” von “workflowSteps→operators→components→name”) muss aber für eine geeignete Abbildung nicht identisch sein (z.B. kann “step→ environment→compiler→name” auf “workflowSteps→operators→ components→name” abgebildet werden).

2.1 GRAIN

Das GRAIN-Schema ist nach dem computerlinguistisch annotierten Korpus benannt, für dessen Erstellung es verwendet wurde. Es ist aber nicht auf die Korpuserstellung beschränkt, sondern stellt generische Blöcke für Eingabe, Ausgabe und Workflow-Schritte zur Verfügung. Sowohl Werkzeuge als auch Personen können als *Operator* in den Workflow-Schritten auftauchen. Eingabe und Ausgabe können aus leeren Listen bestehen und es kann frei festgelegt werden, ob ein Element Eingabe oder Komponente des Operators ist (z.B. ein verwendetes Lexikon). Das Schema liegt als JSON-Vorlage vor.

2.2 RePlay-DH

Die im Kontext des RePlay-DH Projekts entwickelte Desktop-Anwendung überwacht Arbeitsverzeichnisse und hilft bei der Dokumentation von Änderungen innerhalb einzelner Arbeitsschritte. Der Anwendungskontext sind hierbei Forschungsworkflows in den Digital Humanities, die als solche sehr heterogen ausfallen können und ein großes Spektrum an verschiedensten Arbeitsschritten sowohl manueller Art (z.B Annotationen, Kuration oder Kommentieren) als auch vollautomatischer Prozessierung enthalten können. Das Schema für die bei der semi-automatischen Dokumentation anfallenden und in JSON-Format hinterlegten Prozess-Metadaten ist daher sehr flexibel gehalten und ähnelt in vielen Aspekten dem von GRAIN. Der Fokus liegt ebenfalls

auf der Beschreibung generischer Workflowschritte mit optionalen Ein- und Ausgabedaten und involvierten Personen oder Werkzeugen.

2.3 EngMeta

EngMeta (Engineering Metadata) bezeichnet ein Metadatenschema für die computergestützten Ingenieurwissenschaften.[4] EngMeta wurde anhand der Anwendungsfälle der Aerodynamik und der Thermodynamik, welche innerhalb der computergestützten Ingenieurwissenschaften als repräsentativ gelten können, im Rahmen des Forschungsprojekts DIPL-ING entwickelt.[5] Da in die Workflows dieser Fachdisziplin, die vor allem Computersimulation nutzt, sehr wenig manuelle Arbeitsschritte eingehen, liegt der Fokus vor allem auf der Erfassung der automatisch generierten der Prozessmetadaten. Diese werden in EngMeta unter dem Überbegriff Prozessmetadaten gefasst und umfassen eine Annotation der Forschungsdaten mit Information zur Rechnerumgebung und Software, zum Experimentalgerät, zur Simulationsmethode und ihrer Parameter und Variablen sowie die handelnden Personen und die Ein- und Ausgabedaten, die innerhalb des Prozessschrittes angefallen sind. Viele dieser Informationen werden automatisch im Arbeitsschritt mitgeliefert, z.B. Software- und Compilerversionen. Im Bereich der computergestützten Ingenieurwissenschaften lassen sich neben fachspezifischen und technischen Metadaten insbesondere jene Prozessmetadaten gut erfassen.[6]

Im Rahmen des Posters kann näher auf die Abbildung sowie die spezifischen Anforderungen für die Anwendung in den digitalen Geisteswissenschaften eingegangen werden.

[4] Schembera und Iglezakis 2020.
[5] Schembera 2019.
[6] Schembera 2021.

Bibliografie

FoKUS. n.d. “EngMeta – Beschreibung von Forschungsdaten.” Accessed July 30, 2021. https://www.izus.uni-stuttgart.de/fokus/engmeta/.

Gärtner, Markus. 2019. “RePlay-DH Process Metadata Schema.” https://doi.org/10.18419/darus-474

Jung, Kerstin, Markus Gärtner, and Jonas Kuhn. 2019. “Fine-GRAINed Process Metadata.” *In Metadata and Semantic Research. MTSR 2019*, edited by Emmanouel Garoufallou, Francesca Fallucchi and Ernesto William De Luca, 373-78. Cham: Springer. https://doi.org/10.1007/978-3-030-36599-8_33

Kepler. n.d. “The Kepler Project.” Accessed July 30, 2021. https://kepler-project.org/

Schembera, Björn, Björn Selent, Anett Seeland, Dorothea Iglezakis. 2019. "Datenmanagement in Infrastrukturen, Prozessen und Lebenszyklen für die Ingenieurwissenschaften: Abschlussbericht des BMBF-Projektes Dipl-Ing." https://doi.org/10.2314/KXP:1693393980

Schembera, Björn. 2021. "Like a rainbow in the dark: metadata annotation for HPC applications in the age of dark data." *The Journal of Supercomputing*: 1-21. https://doi.org/10.1007/s11227-020-03602-6.

Schembera, Björn, and Dorothea Iglezakis. 2020. "EngMeta: metadata for computational engineering." *International Journal of Metadata, Semantics and Ontologies* 14(1): 26-38. https://doi.org/10.1504/IJMSO.2020.107792.

SDC4Lit – Science Data Center for Literature

Nachhaltiger Datenlebenszyklus für digitale Literatur

Buck, Nina
nina.buck[at]hlrs.de
Höchstleistungsrechenzentrum (HLRS), Universität Stuttgart, Deutschland

Blessing, André
andre.blessing[at]ims.uni-stuttgart.de
Institut für Maschinelle Sprachverarbeitung (IMS), Universität Stuttgart, Deutschland

Bönisch, Thomas
boenisch[at]hlrs.de
Höchstleistungsrechenzentrum (HLRS), Universität Stuttgart, Deutschland

Ganzenmüller, Andreas
ganzenmueller[at]hlrs.de
Höchstleistungsrechenzentrum (HLRS), Universität Stuttgart, Deutschland

Hein, Pascal
pascal.hein[at]ilw.uni-stuttgart.de
Institut für Literaturwissenschaft (ILW), Universität Stuttgart, Deutschland

Hess, Jan
jan.hess[at]dla-marbach.de
Deutsches Literaturarchiv Marbach (DLA), Deutschland

Holz, Alexander
alexander.holz[at]dla-marbach.de
Deutsches Literaturarchiv Marbach (DLA), Deutschland

Jung, Kerstin
kerstin.jung[at]ims.uni-stuttgart.de
Institut für Maschinelle Sprachverarbeitung (IMS), Universität Stuttgart, Deutschland

Kaminski, Andreas
kaminski[at]hlrs.de
Höchstleistungsrechenzentrum (HLRS), Universität Stuttgart, Deutschland

Kamzelak, Roland
kamzelak[at]dla-marbach.de
Deutsches Literaturarchiv Marbach (DLA), Deutschland

Kramski, Heinz Werner
kramski[at]dla-marbach.de
Deutsches Literaturarchiv Marbach (DLA), Deutschland

Kuhn, Jonas
jonas.kuhn[at]ims.uni-stuttgart.de
Institut für Maschinelle Sprachverarbeitung (IMS), Universität Stuttgart, Deutschland

Kushnarenko, Volodymyr
volodymyr.kushnarenko[at]hlrs.de
Höchstleistungsrechenzentrum (HLRS), Universität Stuttgart, Deutschland

Schembera, Björn
schembera[at]hlrs.de
Höchstleistungsrechenzentrum (HLRS), Universität Stuttgart, Deutschland

Schlesinger, Claus-Michael
claus-michael.schlesinger[at]ilw.uni-stuttgart.de
Institut für Literaturwissenschaft (ILW), Universität Stuttgart, Deutschland

Ulrich, Mona
mona.ulrich[at]dla-marbach.de
Deutsches Literaturarchiv Marbach (DLA), Deutschland

Viehauser, Gabriel
gabriel.viehhauser-mery[at]ilw.uni-stuttgart.de
Institut für Literaturwissenschaft (ILW), Universität Stuttgart, Deutschland

Zusammenfassung. Das interdisziplinäre Science Data Center for Literature (SDC4Lit) hat sich den Aufbau eines Datenzentrums zur Sammlung, Archivierung, Erschließung, Erforschung und Vermittlung von digitalen literarischen Quellen mit computergestützten Methoden als Ziel gesetzt. Im Zuge dessen entsteht eine Forschungsplattform, auf der Quellen, Methoden, Werkzeuge und Vermittlungskompetenzen gebündelt, aufbereitet und strukturiert angeboten werden. Dabei wird auch beachtet, dass Modellierung, Aufbewahrung und Analyse aufgrund der besonderen Form digitaler Literatur zusammengehören. Gegenstand sind genuin digitale literarische Materialien mit einem Schwerpunkt auf Literatur im Netz und Born-digitals in Vor- und Nachlässen. Diese sind in einem Primärdatenrepositorium gespeichert und können in einer Arbeitsumgebung, die computergestützte Methoden und Werkzeuge bereitstellt und unterstützt, gelesen und analysiert werden. Die Forschungsergebnisse können in einem Forschungsdatenrepositorium, das auch über das Portal zugänglich sein wird, gespeichert und zur Verfügung gestellt werden. Eine solche Aufgabe bringt viele Herausforderungen für die Forschung und die Archivarbeit mit sich und benötigt unterschiedliche Expertisen, weshalb SDC4Lit diese Aufgabe interdisziplinär bearbeitet.

Die Digitalisierung verändert die Bedingungen für die Entstehung, Verbreitung, Wahrnehmung und somit Erforschung von Literatur. Die neuen medialen Bedingungen führen nicht nur zur Übertragung von gedruckten Materialien in digitale Formen, sondern schaffen und etablieren selbst neue Literaturformen und –gattungen. Dazu gehören zum Beispiel literarische Hypertexte und Tweets, Blog-Formate sowie Texte und Textgeneratoren, die auf computerlinguistische Methoden setzen. Auch im Bereich der Vor- und Nachlässe gibt es zunehmend elektronische Datenträger, die medienspezifische Formate, Dateistrukturen oder spezifische Dateiformate enthalten. Diese enthalten auch eine große Menge an Metadaten, die beispielsweise für die Bearbeitung von editionsphilologischen oder von kultur- und mediengeschichtlichen Fragestellungen relevant sind.

Einerseits führt gerade diese digitale Form zu einer Reihe von besonderen Anforderungen für die Archivierung und Bereitstellung dieser Materialien. Anderseits erweisen sich digitale Texte als besonders interessant für die Arbeit mit computergestützten Methoden, da sie bereits digital vorliegen.

Das Projekt Science Data Center for Literature (SDC4Lit)[1] möchte die Anforderungen, die Digitale Literatur an ihre Sammlung, Erschließung, Archivierung, Erforschung und Vermittlung stellt, systematisch reflektieren und entsprechende Lösungen für einen nachhaltigen Datenlebenszyklus für Literaturforschung und -vermittlung langfristig umsetzen. Dabei soll beachtet werden, dass Modellierung, Aufbewahrung und Analyse aufgrund der besonderen Form digitaler Literatur eng miteinander verbunden sind.

Ein Schwerpunkt ist der Aufbau von langzeitverfügbaren Repositorien für Digitale Literatur (insbesondere Netzliteratur und Born-digitals[2]) und zugehörigen Forschungsdaten sowie die Entwicklung einer Analyseumgebung für die Arbeit mit dem Material.

Das Primärdatenrepositorium bildet im künftigen Betrieb den zentralen Speicher für digitale Literatur als Zielort für das Harvesting von Literatur im Netz und die Archivierung sowie Bereitstellung digitaler Bestandteile von Vor- und Nachlässen. Für den Aufbau des Repositoriums muss eine geeignete Softwarelösung gefunden und ein Datenmodell entwickelt werden. Angesichts der besonderen Form der Primärdaten stellen sich sehr spezifische Anforderungen. Diese Anforderungen können von keiner zurzeit frei zur Verfügung stehenden Repositoriumssoftware vollständig erfüllt werden, sodass für spezifische Aufgaben eigene Entwicklungsarbeit geleistet werden muss. Für das Datenmodell wird auf bereits vorhandene Standards zurückgegriffen, und diese, wo nötig, an unsere Anforderungen angepasst.

Die geplante Analyseumgebung soll es Forscherinnen und Forschern ermöglichen, freigegebene Primärmaterialien mit den bereitgestellten computergestützten Methoden und Werkzeugen zu analysieren. Um die Bedürfnisse der Forschungscommunity nicht zu verfehlen, werden anhand von Fallstudien und im Austausch mit den wissenschaftlichen Communities der Zielgruppen mögliche Anforderungen an die aufzubauende Analyseumgebung erhoben. Im nächsten Schritt werden die entsprechenden Analysetools ausgewählt, angepasst oder neu entwickelt und schließlich in die Analyseumgebung integriert. Dabei werden bereits existierende Ansätze und Systeme wie The NEXT-Repository[3] (Repositorium zur Elektronischen Literatur), die Rhizome

[1] SDC4Lit Homepage. URL: https://www.sdc4lit.de/, Zugriff am 6.8.2021.
[2] Kramski, Heinz Werner, „Stichwort Born-digitals“. URL: https://edlex.de/idex.php?title=Born-digitals, Zugriff am 6.8.2021.
[3] The NEXT. URL: https://the-next.eliterature.org/, Zugriff am 6.8.2021.

Artbase[4] (Archiv mit Netzkunst), Archivematica[5] (System zur digitalen Erhaltug), SolrWayback[6] (zur Analyse von archivierten Webseiten) oder Webservices aus Infrastrukturinitiativen wie z. B. CLARIN[7] zur Nachnutzung untersucht.

Ein zweites Repositorium bietet Nutzerinnen und Nutzern die Möglichkeit, ihre in der Analyseumgebung generierten Forschungsdaten strukturiert zu speichern. Forschungsdaten können dann für die Nachnutzung in weiteren Forschungsvorhaben zur Verfügung gestellt werden.

Das Onlineportal verbindet die Repositorien und die Analyseumgebung und gewährt den Nutzerinnen und Nutzer einen ihren Berechtigungen entsprechenden Zugang. Perspektivisch sollen auch externe Repositorien an die Forschungsplattform über eigene Schnittstellen angebunden werden können.

Die Sammlung, Archivierung, Erschließung, Erforschung und Vermittlung von Literatur im digitalen Wandel ist eine Aufgabe, die Forschung und Archive gleichermaßen betrifft und die daher nur in der interdisziplinären Zusammenarbeit zu bewerkstelligen ist. Aus diesem Grund sind im Projekt Partner mit unterschiedlichen Expertisen in den Bereichen Archiv, Höchstleistungsrechnen, Maschinelle Sprachverarbeitung und Digital Humanities vereint, die diese Aufgabe gemeinsam bearbeiten.

Bibliografie

Archivematica: open-source digital preservation system, https://www.archivematica.org/en/.

CLARIN ERIC Homepage, https://www.clarin.eu/.

[4] Rhizome Artbase. URL: https://artbase.rhizome.org/wiki/Main_Page, Zugriff am 6.8.2021.
[5] Archivematica: open-source digital preservation system. URL: https://www.archivematica.org/en/, Zugriff am 6.8.2021.
[6] GitHub - netarchivesuite/solrwayback: A search interface and wayback machine for the UKWA Solr based warc-indexer framework. URL: https://github.com/netarchivesuite/solrwayback, Zugriff am 6.8.2021.
[7] CLARIN ERIC Homepage. URL: https://www.clarin.eu/ Zugriff am 6.8.2021.

GitHub - netarchivesuite/solrwayback: A search interface and wayback machine for the UKWA Solr based warc-indexer framework, https://github.com/netarchivesuite/solrwayback.

Kramski, Heinz Werner, „Stichwort Born-digitals“, https://edlex.de/index.php?title=Born-digitals.

Rhizome Artbase, https://artbase.rhizome.org/wiki/Main_Page.

SDC4Lit Homepage, https://www.sdc4lit.de/.

The NEXT, https://the-next.eliterature.org/.

Sprachwandel digital erfahrbar machen

Forschungsdatenmanagement in der Variationslinguistik

Ganswindt, Brigitte
ganswindt[at]uni-marburg.de
Forschungszentrum Deutscher Sprachatlas, Universität Marburg, Deutschland

Engsterhold, Robert
engsterhold[at]uni-marburg.de
Forschungszentrum Deutscher Sprachatlas, Universität Marburg, Deutschland

Zusammenfassung: Am Beispiel des historischen und weltweit einmaligen „Sprachatlas des Deutschen Reichs“ und seiner Datengrundlage (über 50.000 handgeschriebene Fragebögen in Kurrentschrift) werden die Anforderungen an ein jeweils zeitgemäßes Forschungsdatenmanagement in den Bereichen der Sprachgeographie und Variationslinguistik vorgestellt und diskutiert. Der Fokus des Beitrags liegt dabei zum einen auf den Besonderheiten der historischen Materialsammlung und ihrer Bedeutung für die Regionalsprachenforschung sowie für die Analyse von Sprachwandelprozessen. Zum anderen wird die Materialerschließung und die weltweit zugängliche, kostenfreie Bereitstellung der Daten im Internet vorgestellt. Um stets aus dem aktuellen Stand der Technik zu bleiben sind kontinuierliche Modernisierungen im Forschungsdatenmanagement erforderlich, die im Vortrag präsentiert werden sollen.

1 Der „Sprachatlas des Deutschen Reichs“

1.1 Entstehungsgeschichte

In den Jahren 1876 bis 1887 verschickte der Marburger Bibliothekar und Germanist Georg Wenker an alle Volksschulen im ehemaligen Deutschen Reich Fragebogen (so genannte „Wenkerbogen“), die zur Erhebung der Lokaldialekte des Deutschen dienten. Die lokalen Gewährspersonen waren aufgefordert, 40 vorformulierte hochsprachliche Sätze schriftlich in ihren jeweiligen Dialekt zu übertragen. Diese Sätze waren so konstruiert, dass typische lautliche und einige grammatische Eigenschaften der betreffenden Dialekte

durch die Übersetzung hervortreten mussten. Nach Abschluss der in mehreren Etappen mit unterschiedlichen Fragebogen durchgeführten Erhebung hatte Wenker circa 45.000 Fragebogen aus fast 41.000 Schulorten vorliegen. Einschließlich der späteren Nacherhebungen in weiteren deutschsprachigen Regionen umfasst das Korpus historische Forschungsdatensätze zu über 50.000 Orten.
Die Wenkerbogen stellen die Datengrundlage zum „Sprachatlas des Deutschen Reichs“ dar, einem handgezeichneten Sprachatlas, der auf 1.668 Kartenblättern (eine Gesamtkarte besteht jeweils aus drei Teilkarten) phonetisch-phonologische sowie morphologische Besonderheiten der deutschen Dialekte präsentiert. Dieser Atlas und die dazugehörigen Wenkerbogen stellen seit Anbeginn ihrer Entstehung bis heute ein einmaliges Forschungsinstrument der Sprachwissenschaft dar. Heute liegt ihr besonderer Wert darin, dass sie nicht nur den analytischen Vergleich jüngerer und älterer Sprachstufen ermöglichen, sondern auch die oftmals einzige Erschließung heute nicht mehr existenter Sprachvarietäten sichern.
Aufgrund seines übergroßen Formates und der Verwendung von bis zu 22 Farben pro Karte konnte der „Sprachatlas des Deutschen Reichs“ nie gedruckt werden. Viele Forschergenerationen mussten daher stets nach Marburg reisen, um vor Ort mit dem einzigartigen Material – einer Kompletterschließung einer Nationalsprache – zu arbeiten.

1.2 Digitalisierung der historischen Forschungsdaten

Dies änderte sich erst ab dem Jahr 2000, in dem mit der Digitalisierung dieser umfangreichen Forschungsdaten im DFG-geförderten Projekt DiWA („Digitaler Wenker-Atlas“) begonnen wurde. Erst durch die Digitalisierung des Materials konnte der Atlas und seine Datengrundlage überhaupt publiziert und einer breiten fachwissenschaftlichen Community sowie interessierten Laien im Internet frei zur Verfügung gestellt werden.
Für die Digitalisierung der über 50.000 Wenkerbogen und der 1.668 Karten des „Sprachatlas des Deutschen Reichs“ wurde eine externe Firma beauftragt, die die Daten mit einem Gesamtvolumen von über 5TB auf DVDs zur Verfügung stellte. Zu Beginn des Projektes Anfang der 2000er Jahre war ausreichend Festplattenplatz zwar schon theoretisch vorhanden, aber finanziell oft nicht erschwinglich, weswegen für die Bereitstellung der Karten und Bogen nach Möglichkeiten gesucht wurde, die Dateigröße zu reduzieren. In der ersten Version des „Digitalen Wenker-Atlas“ wurde sich für ein proprietäres System von ERDAS entschieden, deren Dateiformat ECW, ein Wavelet-Format, die

Datenmenge stark reduzieren konnte. Dieses Kernmodul des DiWA-Projektes basierte auf einem JAVA-Plugin, das allerdings nach einigen Jahren nicht mehr von modernen Browsern ausgeführt werden konnte.
Ab dem Jahr 2008 konnte in dem Akademiegeförderten Langzeitprojekt „Regionalsprache.de“[1] (REDE) ein Sprachgeographisches Informationssystem (SprachGIS) entwickelt werden, in das alle in DiWA verfügbaren Karten und Daten im Zuge einer Servermigration übertragen werden konnten. Zusätzlich werden im SprachGIS alle historischen Daten mit modernen Sprachatlanten (ca. 12.500 Rasterkarten sowie 14.200 in eine Datenbank überführte Vektorkarten), circa 12.300 Tonaufnahmen deutscher Dialekte sowie einer umfangreichen Online-Bibliographie zusammengeführt, die explorative und vergleichende Analysen der Dialektologie und Variationslinguistik mit Datenquellen verschiedenster Art ermöglichen.
Technisch baut das SprachGIS rein auf open-source-Lösungen auf, da sich hier Standards und frei verfügbare Systeme für die Darstellung von Karten oder großen Bilddateien entwickelt haben. Dementsprechend wurde das SprachGIS als seinerzeit sehr moderne JavaScript Web-App angelegt. Das Herz dieser Anwendung ist das OpenLayers-Framework, ursprünglich in der Version 2[2] verwendet.
Die Bereitstellung der Rasterkarten erfolgt nach dem WMS-Standard und die Anwendung mapserv[3] dient als Kartenserver. Mit der IIPImage-Software[4] werden die Wenkerbogen bereitgestellt und mittels einer leicht modifizierten Version des zugehörigen IIPMooViewer in der Web-App dargestellt. Als interaktive JavaScript-Anwendung ermöglicht das SprachGIS nicht nur ein passives Anschauen von Karten, sondern bietet auch die Möglichkeit, selbst Karten zu erstellen, Analysen mit eigenen linguistischen Daten auszuführen und mit historischen Karten in Kontext zu setzen. Zudem lassen sich die selbsterstellten Karten für Publikationen in angemessenen Bildformaten und Auflösungen exportieren.

2 Modernisierung des Forschungsdatenmanagements

Zwar wird die SprachGIS Web-App kontinuierlich weiterentwickelt und betreut und befindet sich bereits in der dritten Hauptversion,

[1] Kehrein 2020, Ganswindt et al. 2015, https://www.regionalsprache.de.
[2] https://openlayers.org/two/.
[3] https://mapserver.org.
[4] https://iipimage.sourceforge.io.

nichtsdestoweniger ist eine grundlegende Modernisierung mittlerweile erforderlich.
Aktuell wird daran gearbeitet, die bereits verwendeten Standards zu modernisieren sowie weitere Standards zu verwenden, um so eine bessere Integrierung in eine moderne, vernetzte Wissenschaftslandschaft zu ermöglichen. Daher ist eine Umstellung auf OpenLayers 6 in Vorbereitung, die nicht nur eine bessere und schnellere Darstellung von Raster- und Vektorkarten ermöglicht, sondern auch Bilddateien nach dem IIIF-Standard[5] darstellen kann. Außerdem wurden die Rasterbildformate von ECW auf JP2 umgestellt.
Seit 2020 existiert ein Prototyp[6] für einen neuen Wenkerbogen-Viewer, der auf dem IIIF 2.1-Standard basiert und neben Darstellung und Verortung der Bögen auch die Möglichkeit bietet, Transliterationen der Bogen zu betrachten oder selbst zu erstellen. Zunächst als internes Projekt begonnen, wurden circa 1.200 Transliterationen (Überführung der handschriftlichen Kurrentschrift der Bogen in eine Datenbank) normalisiert und online gestellt. Eine Erweiterung zu einem „Citizen Science"-Projekt ist derzeit in Planung.
Die Überarbeitung des SprachGIS bietet neben einer modernen Benutzeroberfläche auch eine REST-API, um Daten zur direkten Einbindung in andere Projekte oder Anwendungen zur Verfügung zu stellen. Zudem sollen Dereferenzierbarkeit und semantische Annotationen eine Einbindung in das Semantische Netz ermöglichen. Damit einhergehend wird im Rahmen der Hessischen Forschungsdateninfrastruktur (HeFDI) unser Repository derzeit überarbeitet und an ein aktuelles Forschungsdatenmanagement angepasst.
Im Vortrag sollen anhand der Wenker-Materialien und dem jeweils erforderlichen Umgang mit diesen Daten die Anforderungen und Herausforderungen an ein zeitgemäßes Forschungsdatenmanagement aufgezeigt werden. Erst die zeitgemäße Materialerschließung hat zu zahlreichen, teils wegweisenden Ergebnissen der dialektologischen und variationslinguistischen Fachgeschichte geführt, die ohne eine Interaktion zwischen Geisteswissenschaften und Informatik nicht möglich gewesen wären. Ganz entscheidend ist dafür natürlich die kostenfreie Verfügbarkeit der Forschungsdaten im Internet, die WissenschaftlerInnen weltweit bequemen Zugriff auf die Daten für ihre eigene Forschung ermöglicht.

[5] https://iiif.io.
[6] https://wenker.online.uni-marburg.de/wenker.

Bibliografie

Ganswindt, Brigitte, Roland Kehrein, Alfred Lameli, „Regionalsprache.de (REDE)“, in *Regionale Variation des Deutschen – Projekte und Perspektiven*, edited by Roland Kehrein, Alfred Lameli Stefan Rabanus (Berlin/Boston: De Gruyter, 2015), 421–453.

Kehrein, Roland, „Das Wenker-Material im REDE SprachGIS“, in Minderheitensprachen und Sprachminderheiten: Deutsche und seine Kontaktsprachen in der Dokumentation der Wenker-Materialien (Hildesheim/Zürich/New York: Olms, 2020), 21–56.

Virtuelle Forschungsdatenumgebungen

Struktur, Erkenntnis, Nachnutzung

Rosenkranz, Vinzenz
vinzenz.rosenkranz[at]uni-tuebingen.de
eScience-Center Universität Tübingen, Deutschland
ORCID-iD: 0000-0002-5387-425X

Schmidt, Karsten
karsten.schmidt[at]uni-tuebingen.de
eScience-Center Universität Tübingen, Deutschland
ORCID-iD: 0000-0003-0337-3024

Zusammenfassung. Das Forschungsdatenmanagement steht am Beginn eines jeden Forschungsvorhabens und sollte bis zum Ende mitgedacht werden. Eine Forschungsdatenumgebung unterstützt die Forschenden bei der Strukturierung und Analyse ihrer Daten. Diese Systeme werden meist erst während des Projektzeitraums entwickelt. Das „Wie" und „Was" muss also bei jedem einzelnen Projektbeginn entweder neu entworfen oder in ähnliche vergangene Formen eingefügt werden. Das am eScience-Center der Universität Tübingen entwickelte System ***Spacialist***, verlagert die Forschungsumgebung in den virtuellen Raum. Dadurch ermöglicht es ein ortsungebundenes und kollaboratives Arbeiten. Die Datenaufnahme in das generisch definierte Datenmodell standardisiert die Datenstruktur und garantiert dadurch eine qualitative Analyse der Daten. Die Forschungsdaten können im System kommentiert, validiert und gegebenenfalls korrigiert werden. Dabei unterstützen ein integriertes Moderationssystem mit Rollenvergabe, eine Kommentarspalte sowie Möglichkeiten zur Referenzierung von Literatur oder auf in der integrierten Dateiverwaltung hinterlegten Dateien. Ebenfalls können Geodaten und benutzerdefinierte Kartenlayer verwendet werden, um raumbezogene Daten zu erfassen und einzubeziehen. Die Entwicklung von *Spacialist* ist von Anfang an transparent gestaltet, um das System auf die Workflows der Fachbereiche auszurichten. Die dabei einbezogenen Fachcommunities der Digital Humanities haben darüber hinaus die Möglichkeit Spacialist durch Plugins zu erweitern und an spezifische Anforderungen anzupassen. Zudem können die Forschungsergebnisse unkompliziert in bestehende (universitäre) Archivsysteme integriert werden.

Eine virtuelle Forschungsdatenumgebung dient der systematischen Erfassung von Forschungsdaten, im besten Fall über den gesamten Lebenszyklus der Daten hinweg. Angereichert mit einer Bandbreite unterschiedlicher Informationen und Formaten, können diese interdisziplinär, individuell, aber auch generisch, eingesetzt werden. Dabei findet die Integration in den Forschungsprozess bereits direkt bei der Aufnahme statt und ermöglicht dort eine Anreicherung mit essenziellen (administrativen, inhaltsbeschreibenden) Metainformationen sowie die Verknüpfung zu wissenschaftlichen Identifikationssystemen (wie beispielsweise ORCID). Dabei stehen die Anforderungen des Forschenden selbst, seine Organisation und individuelle Strukturierung, die gerade bei großen Datenmengen und unterschiedlichen Formaten wichtig ist, im Zentrum unserer entwickelten virtuellen Forschungsumgebung ***Spacialist***[1] (Abb. 1)**.** Zusätzlich ermöglicht es diese Forschungsumgebung direkt bei der Eingabe erste Analysen, Korrekturen und Validierungen vorzunehmen. Dabei kombiniert *Spacialist* sowohl datenbank-technische Aspekte für Erfassung und Quellennachweise als auch Visualisierungskomponenten, wie z.B. Raumdaten und 3D-Objekte.
Für die technische Umsetzung haben wir großen Wert auf Flexibilität gelegt und damit den Fokus auf Software gesetzt, die dem Entwickler diese Möglichkeiten bietet. Deshalb verwenden wir als Grundlage für *Spacialist* das etablierte Datenbanksystem ***PostgreSQL***[2], welches mit der Erweiterung *PostGIS* bestens für das Verarbeiten von Raumdaten gerüstet ist. Bei der Entwicklung des Frontends, stützen wir uns auf etablierte Tools und Frameworks mit großer Community. Für das Erscheinungsbild nutzen wir ***Bootstrap***[3], welches eine schnelle und einheitliche Entwicklung von UI-Komponenten erlaubt. Als Kommunikationselement mit der Datenbank setzt unsere Umgebung auf das *PHP-Webframework* ***Laravel***[4], welches eine Vielzahl an Funktionen anbietet, die über die einfache Kommunikation mit der Datenbank hinaus gehen. Dessen große Community erweitert Laravel stetig durch Plugins, die offen zur Verfügung gestellt werden. Auf der Benutzerseite setzen wir das *JavaScript-Framework* ***Vue.js***[5] ein, welches sich neben den großen Konkurrenten *React* (Facebook) und *Angular*

[1] eScience-Center Homepage, “Spacialist“.
[2] PostgreSQL Homepage, “PostgreSQL: The world's most advanced open source database”.
[3] Bootstrap Homepage, “Bootstrap · The most popular HTML, CSS, and JS library in the world.”.
[4] Laravel Homepage, “Laravel - The PHP Framework For Web Artisans”.
[5] Vue.js Homepage, “Vue.js”.

(Google) etabliert[6] hat und bereits bei großen Anwendungen, wie *Nextcloud*, zum Einsatz kommt. Auch *Vue.js* lässt sich durch Community-Plugins oder eigene Entwicklungen erweitern.
Durch diese Kombination ermöglichen wir nicht nur eine agile Entwicklung, sondern stellen den Forschenden eine Anwendung zur Verfügung, die sie nach Ihren Ansprüchen anpassen können. Dabei setzt unsere Forschungsumgebung auf ein – in relationalen Datenbankstrukturen abgebildetes – Datenmodell, welches von den Benutzern selbst konfiguriert wird. Damit ist die Zuordnung und Ausgestaltung der Attribute der zu beschreibenden Objekttypen frei wählbar. Lediglich die Auswahl an Datentypen (Text, Zahl, …) für die Attribuierung ist dabei von der Umgebung vorgegeben.

Die Beschreibung der einzelnen Attribute und Auswahlmöglichkeiten (beispielsweise für Dropdown-Auswahllisten), können von den Projektbenutzern frei vergeben werden. Darüber hinaus besteht die Möglichkeit diese Zuordnungen und Ausgestaltungen im Nachhinein zu ändern, zu erweitern oder gänzlich zu entfernen. Dieser Teil wird über ein parallel entwickeltes und darauf zugeschnittenes – auf *SKOS*-Standard basierendes – Thesaurus-Tool realisiert. Neue Sprachen können jederzeit hinzugefügt oder entfernt werden. Auch lassen sich Einträge ändern, Definitionen hinzufügen sowie Notizen ergänzen.
Neben der Eingabe neuer Daten können ebenfalls vorhandene Daten in Form von verschiedenen Dateiformaten oder Literaturnachweisen zum System hinzugefügt werden, um die erhobenen Daten mit weiteren (Meta-)daten anzureichern. Die direkte Unterstützung verschiedener Dateiformate ist dabei ein wichtiger Schritt, um alle Bereiche des Forschungsdatenmanagements zu bündeln. So kann *Spacialist* mit einfachen Textdaten über Bilder und Archivdaten bis zu 3D-Dateien umgehen und diese zum Teil weiterverarbeiten.
Durch die oben genannten Möglichkeiten und die Verlagerung in den virtuellen Raum ergeben sich vielzählige Möglichkeiten zur ortsunabhängigen Kollaboration von Wissenschaftler*innen. Über Zugriffskontrollen können verschiedene Rollenelemente vergeben werden, um z.B. im Moderator-Modus Datenbearbeitungen erst nach entsprechender Diskussion aufzunehmen. So lassen sich gesicherte Informationen und eine Versionierung von einzelnen Datenbeständen realisieren. Gleichzeitig können Änderungen im Verlauf nachvollzogen und Fehler korrigiert werden. Dieses Kontrollsystem schützt vor

[6] Daityari, “Angular vs React vs Vue”.

unbefugten Zugriffen, ungewollten Änderungen und ermöglicht die Abfrage ältere Bestände. Ein weiterer Vorteil der Nutzung der hier vorgestellten Forschungsdatenumgebung *Spacialist* ist die durch die Realisierung als Webanwendung einfach zu realisierende Sicherung der Datenbestände in den (universitären) Serverstrukturen. Hier sind die Daten selbst gegen einen Serverausfall oder einen Angriff geschützt. Ein Verlust der Daten durch einen Ausfall der EDV des Bearbeiters an den jeweiligen Standorten ist somit ausgeschlossen und die Erhaltung der zusammengetragenen Forschungsdaten gewährleistet.
Außerdem erlaubt sie eine einfachere Archivierung der Daten in bestehende Systeme, wie beispielsweise Universitäts-Repositorien. Diese zeitintensive Aufgabe wird einfacher, wenn schon während des Forschungsprozesses notwendige Dokumentationen vom System erfolgen.
Spacialist ist seit Beginn als Open-Source-Projekt umgesetzt[7], um die gesamte interessierte Community einzubinden und am Entwicklungsprozess teilhaben zu lassen. Dadurch war es möglich in kürzester Zeit eine Vielzahl an Rückmeldungen zu erhalten, die in die weitere Ausrichtung und Stabilität der Forschungsumgebung eingeflossen sind. So haben wir es in den letzten Jahren geschafft, eine zuverlässige, intuitive und umfängliche Umgebung zu schaffen, die die Forschenden aus den verschiedensten Disziplinen bei ihrer Arbeit unterstützt.

[7] GitHub, „Projektseite Spacialist“.

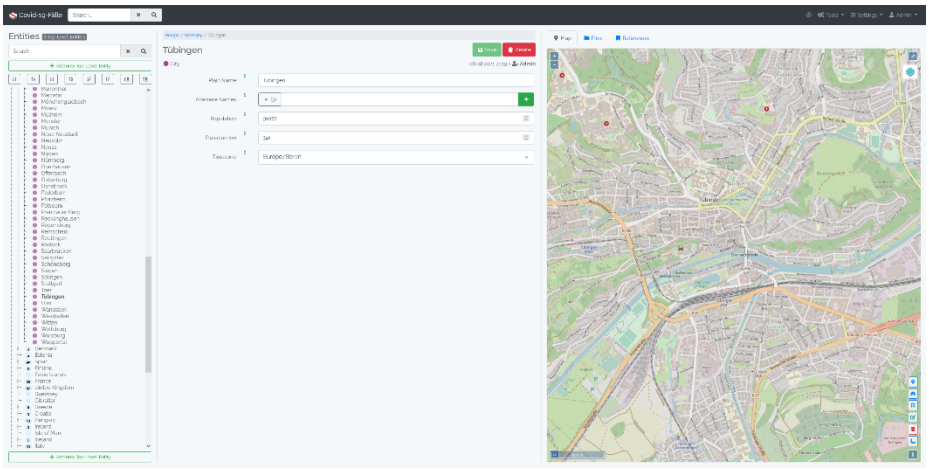


Abb. 1. Spacialist – Eintrag zu Tübingen.

Bibliografie

"Bootstrap · The most popular HTML, CSS, and JS library in the world.", Bootstrap Homepage, aufgerufen am 08.08.2021, https://getbootstrap.com.

Daityari, Shaumik. "Angular vs React vs Vue: Which Framework to Choose in 2021", veröffentlicht am 15.03.2021 auf codeinwp.com, https://www.codeinwp.com/blog/angular-vs-vue-vs-react/#popularity.

"Laravel - The PHP Framework For Web Artisans", Laravel Homepage, aufgerufen am 08.08.2021, https://laravel.com.

PostgreSQL: The world's most advanced open source database", PostgreSQL Homepage, aufgerufen am 08.08.2021, https://www.postgresql.org.

"Spacialist – Die virtuelle Forschungsumgebung für die Spacial Humanities", eScience-Center Uni Tübingen, aufgerufen am 08.08.2021, https://uni-tuebingen.de/forschung/forschungsinfrastruktur/escience-center/spacialist-1.

“Spacialist“, GitHub Projektseite, Stand 08.08.2021,
https://github.com/eScienceCenter/Spacialist.

“Vue.js”, Vue.js Homepage, aufgerufen am 08.08.2021,
https://v3.vuejs.org.

Wir „gärtnern“ im FDM und forschungsethische Datenlandschaften.

Wilderness and management of research data, including research ethics.

Strauch, Annette
annette.strauch-davey[at]uni-hildesheim.de
UB Hildesheim, Deutschland

Zusammenfassung. Das Poster der Stiftung Universität Hildesheim zum Forschungsdatenmanagement zeigt, wie eine wissenschaftliche Bibliothek, so wie die Universitätsbibliothek Hildesheim, geisteswissenschaftliche Disziplinen im Forschungsdatemanagement aktiv unterstützt. Forschungsprojekte werden durch Serviceinfrastrukturen, den technischen Infrastrukturen intern (Rechenzentrum) und extern (GWDG, NFDI4Culture, usw.) in allen Phasen begleitet. Geisteswissenschaftliche Projekte werden außerdem hinsichtlich der forschungsethischen Belange unterstützt. Ende 2020 fand ein Tag der Forschung zur Forschungsethik statt. Das FDM der UB Hildesheim orientiert sich u.a. an den FAIR- und Care-Prinzipien. Methodenorientierte Beratungen gehen über das Dienstleistungsportfolio der UB hinaus. Workshops zum geisteswissenschaftlichen Forschungsdatenmanagement finden jedes Semester statt. Der Kontakt zur FDM-Stelle kann den Weg zu den Ethikkommissionen weisen. Überlegungen zur Ethik gehen über die erforderlichen datenschutzrechtlichen zu den personenbezogenen Daten hinaus. Das gesamte Forschungsfeld wird reflektiert. Der Fachbereich 2, Kulturwissenschaften und Ästhetische Kommunikation, vereint die Institute für bildende Kunst und Kunstwissenschaft, das Institut für Kulturpolitik, das Institut für Literarisches Schreiben und Literaturwissenschaft, das Institut für Medien, Theater und Populäre Kultur, das Institut für Musik und Musikwissenschaft sowie das Institut für Philosophie.

Die Beratungsangebote im aktiven Forschungsdatenmanagement sind heterogen. Sie geschehen im Austausch auf dem Campus der Stiftung Universität Hildesheim mit ganz unterschiedlichen Akteurinnen und Akteuren. Im Fachbereich 1 ist die Erziehungswissenschaft sehr

forschungsorientiert aufgestellt. Es gibt Schwerpunkte in der Forschung und Lehre, die sich mit Pädagogik der Kindheit und ‘Diversity Education‘ befassen. Das Institut für Sozialwissenschaften, gleichfalls im Fachbereich 1 angesiedelt, vereinigt Politikwissenschaft und Soziologie, wobei als Forschungsgegenstände beispielhaft „Politik und Internet“ sowie „Solidarität“ genannt werden. Der Fachbereich 2, Kulturwissenschaften und Ästhetische Kommunikation, vereint die Institute für bildende Kunst und Kunstwissenschaft, das Institut für Kulturpolitik, das Institut für Literarisches Schreiben und Literaturwissenschaft, das Institut für Medien, Theater und Populäre Kultur, das Institut für Musik und Musikwissenschaft sowie das Institut für Philosophie. Letzteres beinhaltet im Bereich der Forschung das DFG-Graduiertenkolleg “Ästhetische Praxis” und das DFG-Reinhart-Kosseleck-Projekt “Geschichten der Philosophie in globaler Perspektive”/”Histories of Philosophy in a Global Perspective”.

Forschungsnahe Dienste werden auch von den wissenschaftlichen Bibliotheken wahrgenommen. Ein konkreter Service für die Geisteswissenschaften wird durch das FDM erstellt. Eine wissenschaftliche Bibliothek, wie die Universitätsbibliothek Hildesheim, kann somit die geisteswissenschaftlichen Disziplinen, z.B. im Fachbereich 2 beim Umgang mit Forschungsdaten unterstützen. Die "Scientific Communities" in den einzelnen Instituten, die auf dem Poster genannt sind, werden durch Serviceinfrastrukturen, den technischen Infrastrukturen intern (Rechenzentrum) und extern (GWDG, NFDI4Culture, usw.) begleitet.

Geisteswissenschaftliche Projekte werden außerdem hinsichtlich der forschungsethischen Belange unterstützt. Obwohl eine wissenschaftliche Bibliothek ein wichtiger Teil des Forschungssupports ist, kann sie nicht für alle Disziplinen die gesamte forschungsnahe Unterstützung leisten, besonders dann wenn finanzielle und personelle Ressourcen knapp sind. Hier können einzelne NFDI-Konsortien in den Task Areas unterstützen und Partner der Forschung und des FDM sein. Bei der Postersession der FORGE 21 werden die Chancen durch die Zusammenarbeit mit NFDI4Culture genannt und abgebildet. Alle Kooperationen sind im Kontext der FAIR-Prinzipien und Open Science zu sehen Beispiele werden gezeigt. Der Beitrag gibt praxisnahe Einblicke in die Unterstützung zum Forschungsdatenmanagement (FDM) in den Geisteswissenschaften und den Bildungswissenschaften für Forschende durch Bibliotheken und FDM-Beratungsstellen. Thematisiert werden die Bedarfe von übergreifenden FDM-Angeboten

und fachspezifischen Bedarfen. Beispiele aus der Praxis werden präsentiert. Im Bereich der Forschungsethik werden folgende Themen behandelt: Genehmigungen und Ethikvoten, die Forscher*innen ggf. einholen und vorlegen müssen, forschungsethische Grundkenntnisse, die zum Grundverständnis des digitalen Wandels in der Wissenschaft gehören, Dienstleistungen der Bibliotheken, die Forscher*innen dazu anregen, ihren Forschungsgegenstand ethisch zu reflektieren. Anderen die Forschungsdaten, auch als OER, zur Verfügung zu stellen, beschäftigt generell viele Akteure innerhalb der globalen E-Science Community, sei es die Forschenden selbst, die Lehrenden oder die Lernenden. Die Ergebnisse wissenschaftlicher Bemühungen, die Ergebnisse der Forschung, sind Teil der gesellschaftlichen Fortentwicklung und bedürfen der Verbreitung und Veröffentlichung. Open Science umfasst Open Access, Open Data, Open Source und OER. Die Arbeiten einzelner Task Areas der NFDI4Culture werden für Forscherinnen und Forscher sehr nützlich sein, weil sie zukünftig den Forschenden Lösungen auf Fragen anbieten können, für die es Anfang 2021 keine hinreichenden Standards gibt.

Bibliografie

Stille, W., Farrenkopf, S., Hermann, S., Jagusch, G., Leiß, C., & Strauch-Davey, A. (2021). „Forschungsunterstützung an Bibliotheken, Positionspapier der Kommission für forschungsnahe Dienste des VDB." O-Bib. Das Offene Bibliotheksjournal / Herausgeber VDB, 8(2), 1–19. https://doi.org/10.5282/o-bib/5718.

Strauch, A. (2020). “Universitätsbibliotheken heute. Partner im Forschungsdatenmanagement in der Praxis,” ABI Technik, 40(2), 177-186. doi: https://doi.org/10.1515/abitech-2020-2008.

Prof. Torsten Schrade (Spokesperson NFDI4Culture). (2020, December 9). “NFDI4Culture im Überblick. Konsortium für Forschungsdaten zu materiellen und immateriellen Kulturgütern. Zenodo.” https://doi.org/10.5281/zenodo.4313368.