Originalveröffentlichung in: Helling, Patrick/Speer, Andreas/Eide, Øyvind (Hrsg.): FORGE 2021 - Forschungsdaten in den Geisteswissenschaften: MAPPING THE LANDSCAPE - Geisteswissenschaftliches Forschungsdatenmanagement zwischen lokalen und globalen, generischen und spezifischen Lösungen. Konferenzabstracts. – Köln, 2021. S. 174-179.
DOI: *https://doi.org/10.5281/zenodo.5379571*

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.