Hands-on-Workshop Datendokumentation

Lemaire, Marina
marina.lemaire@uni-trier.de
Universität Trier, Servicezentrum eSciences

Moeller, Katrin
katrin.moeller@geschichte.uni-halle.de
Martin-Luther-Universität Halle-Wittenberg,
Historisches Datenzentrum Sachsen-Anhalt

Schulz, Julian
Schulz@MaxWeberStiftung.de
Max Weber Stiftung, Geschäftstelle, Digital Humanities und Forschungsdatenmanagement

Söring, Sibylle
sibylle.soering@fu-berlin.de
Freie Universität Berlin, Universitätsbibliothek, Leitung Forschungsdatenmanagement

Wettlaufer, Jörg
jwettla@gwdg.de
Akademie der Wissenschaften zu Göttingen,
Koordination Digitalisierung und Datenkuration

Einführung

Format: Workshop, ganztags
Gruppengröße: max. 30 Teilnehmende
Techn. Ausstattung: Beamer, bevorzugt digitales Whiteboard oder Pinnwände & Medienkoffer, evtl. einen weiteren Raum für Gruppenarbeit, ausreichend Steckdosen für Laptops

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

Workshopkonzept

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

Vor dem Hintergrund einer zunehmenden Zahl an potenziell verwendbaren Forschungsdaten mag es verwundern, dass bislang eher selten auch eine Nachnutzung dieser Daten in neu gelagerten Forschungskontexten erfolgt. Es entsteht der Eindruck, dass „Success Stories" im Bereich der Nachnutzung insbesondere bei geisteswissenschaftlichen Forschungsdaten ein Desiderat darstellen. Ein Grund hierfür mag darin liegen, dass eine strukturierte und detaillierte Form der Datendokumentation bislang wenig im Fokus stand. (Vgl. Daudrich 2018, 13) Entsprechend fehlen fach- bzw. methodenspezifische Best Practice-Modelle, wie sie zunehmend im Kontext generischer Ansätze formuliert werden. (Vgl. dazu z. B. CESSDA 2020, Kap. 2. Organise & Document; Dierkes 2021) Eine strukturierte, standardisierte Dokumentation ist jedoch zwingend erforderlich, um Daten in neuen Projekten nachnutzen zu können. Insbesondere die Grundlagen der Datenerhebung (Auswahl, Begrenzungen, Ursprung, Datenqualität, Prozessierungen usw.) müssen nachvollziehbar sein, um eine spätere Verwendung überhaupt erst zu ermöglichen. Dabei wird deutlich, dass selbst hinsichtlich der Ziele, der Definition und der Grundelemente einer Datendokumentation keine einheitliche Auffassung besteht.

Ein Kernelement im Bereich der Dokumentation stellt die Beantwortung und sukzessive Anpassung eines Datenmanagementplans (DMP) dar. Verstanden als „Living Document", kann ein DMP dazu beitragen, die in einem Projekt verwendeten Daten, Software und Methoden detailliert darzustellen und die aus dem Projekt resultierenden Forschungsdaten damit zu kontextualisieren. Während die Erstellung eines DMP inzwischen vermehrt seitens der Fördermittelgebenden als obligatorisch betrachtet wird, besteht weitgehend noch keine Pflicht, die-

sen zusammen mit den Forschungsdaten zu veröffentlichen. Im Sinne der Nachvollziehbarkeit aller im Projekt unternommenen Schritte wäre die Veröffentlichung des DMP als Beitrag zur Dokumentation jedoch anzuraten.

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

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

Im Rahmen des Workshops soll dieses Desiderat rund um das Thema Datendokumentation aufgegriffen und mit den Teilnehmenden diskutiert werden. In einem ersten Schritt wird das Ziel verfolgt, eine Arbeitsdefinition herzustellen, um eine gemeinsame Vorstellung davon zu erhalten, was unter „Datendokumentation" zu verstehen ist, welche Komponenten (z. B. DMP, Metadaten, Codebook) zwingend erforderlich sind und welche dagegen eher optionalen Charakter besitzen. Darauf aufbauend soll praxisnah ergründet werden, welche Formen der Dokumentation benötigt werden, um nicht nur die Auffindbarkeit von Forschungsdaten, sondern auch ihre Nachnutzung zu vereinfachen bzw. überhaupt zu ermöglichen. In diesem Kontext wird auch zu diskutieren sein, wer – d. h. Forschende oder Kuratierende – für die Dokumentation der Daten verantwortlich zeichnet. Schließlich wird als weiteres Ziel des Workshops vorgegeben, ein besseres Verständnis davon zu erlangen, welche Informationen zwingend Teil einer Datendokumentation sein sollten (z. B. Kontext der Erhebung, Erhebungsmethode, Struktur der Daten und deren Beziehung zueinander). Der Workshop bezieht die Perspektive der Infrastruktureinrichtungen ein (z. B. Repositoriumsbetreibende, Datenzentren) und kann dazu dienen, einen Überblick zu bereits bestehenden Formen der Datendokumentation zu erhalten.

Die DHd-AG Datenzentren hat sich in bisher zwei verschiedenen Veranstaltungen[1] mit der Dokumentation von Forschungsdaten beschäftigt. Dabei wurden vor allem die Herausforderungen der datenhaltenden Institutionen diskutiert, die besonders in der Standardisierung und effizienten Ausgestaltung von Workflows zur Dokumentation von Forschungsdaten liegen. Ziel des hier eingereichten Workshops ist dagegen die Perspektive der Nutzer*innen selbst. Gezielt soll für unterschiedliche, aber typische geisteswissenschaftliche Daten die Dokumentation von Forschungsdaten hinsichtlich ihres Informationswertes, der Verständlichkeit, der Vollständigkeit und ihres tatsächlichen Gebrauchswertes zur Nachnutzbarkeit geprüft werden.

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

Workshop-Programm

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

– Für den Bereich der quantitativen Daten wird Paul Beckus (Historiker an der Martin-Luther-Universität Halle/Wittenberg) aus der datenerstellenden Perspektive berichten, welche Fragen und Probleme sich ihm bei der Dokumentation von Datensätzen ergaben und welche Unterstützungsangebote sich ein historisch arbeitender Wissenschaftler in diesem Prozess erhofft. (Vgl. Beckus 2021)

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

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

lung und -dokumentation im Laufe der Versionen verändert haben.

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

Nach den Berichten sollen in einer halbstündigen Gruppenarbeit alle gesammelten Aspekte zur Datendokumentation gesichtet und versucht werden, eine erste Kategorisierung auf einem gemeinsam zu bearbeitenden Whiteboard vorzunehmen. Die Ergebnisse werden im Plenum diskutiert und zusammengeführt. Hierzu wird ein Schema[2] verwendet, das die Sicht der Datenzentren repräsentiert. Beides dient als Grundlage für die nachfolgenden Gruppenarbeiten am Nachmittag während des Datathons. Beim Datathon stellt jeweils eine Person einen Datensatz vor und macht einen Vorschlag für ein Nachnutzungsszenario, anhand dessen die Gruppen gemeinsam versuchen, den Datensatz zu verstehen und das Dokumentationsschema weiterzuentwickeln. In der Gruppenarbeit sollen sie einerseits feststellen, welche Informationen fehlen und hierfür Anforderungen formulieren, und andererseits überlegen, was sie dokumentieren müssten, damit ihre Ergebnisse wiederum verstehbar und nachnutzbar werden. Auf diese Weise sollen sie die Kategorien und Aspekte der Datendokumentation auf dem Whiteboard überarbeiten und weiterentwickeln. Für die Bereitstellung eines Datensatzes haben sich bislang folgende Personen bereit erklärt:

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

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

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

– Katrin Moeller vom Historischen Datenzentrum Sachsen-Anhalt an der Martin-Luther-Universität Halle/Wittenberg – Interviewdaten der BOLSA-Längsschnittstudie (https://bolsa.uni-halle.de/suche/)

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

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

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

Programmablauf

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

Fußnoten

1. Hands on Research Data, Workshopreihe der AG Datenzentren auf der vDHd2021, https://vdhd2021.hypotheses.org/178, Zugegriffen 9. Dezember 2022; Dokumentation von Forschungsdaten – Erfahrungen und Aufgaben aus der Praxis, Workshop der AG Datenzentren auf der FORGE21, https://forge2021.uni-koeln.de/programm/workshop-ag-datenzentren.html, Zugegriffen 9. Dezember 2022.
2. Das Schema wurde für den oben genannten Workshop auf der FORGE 2021 entwickelt und erprobt. https://docs.google.com/spreadsheets/d/19IBDWw_iBnWU8oy8R92UWKCFuKOZBaONITGQUZRJrLw, Zugegriffen 9. Dezember 2022.

Bibliographie

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

Beckus, Paul. 2021. Der Fürst im Kabinett: Supplikations- und Herrschaftspraxis unter Franz von Anhalt-Dessau (1758-1817). Quellen und Forschungen zur Geschichte Sachsen-Anhalts 24. Halle: Mitteldeutscher Verlag.

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

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

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

Daudrich, Anna. 2018. "Umgang mit Forschungsdaten in den Geistes- und Sozialwissenschaften. Bericht zur Bedarfserhebung an

der Friedrich-Alexander-Universität Erlangen-Nürnberg (FAU)". https://www.fdm-bayern.org/files/2018/11/forschungsdatenmanagement-in-den-geisteswissenschaften-an-der-fau-umfrage.pdf (zugegriffen: 9. Dezember 2022).

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

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

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

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

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

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

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

Jensen, Uwe. 2012. "Leitlinien zum Management von Forschungsdaten. Sozialwissenschaftliche Umfragedaten". Herausgegeben von Leibniz Institut für Sozialwissenschaften GESIS. Technical Reports. Köln. http://www.gesis.org/fileadmin/upload/forschung/publikationen/gesis_reihen/gesis_methodenberichte/2012/TechnicalReport_2012-07.pdf (zugegriffen: 9. Dezember 2022).