Forschungsdatenmanagementplan Geisteswissenschaften
Historisches Datenzentrum Sachsen-Anhalt (Hist-Data)

Dr. Katrin Moeller, Historisches Datenzentrum Sachsen-Anhalt
R. 2.05.0, Emil-Abderhalden-Str. 26/27, 06108 Halle
E-Mail: hinfo@geschichte.uni-halle.de

Was ist ein Forschungsdatenmanagementplan?

Ein forschungsbezogener Data Management Plan ((F)DMP) ist eine strukturierte Dokumentation und Planung des Forschungsprozesses, der vor allem den Umgang mit entstehenden Daten bereits vor dem Start eines Projektes strategisch hinterfragt und plant – auch finanziell. Dieser Plan soll gewährleisten, dass wichtige und zentrale Daten eines Forschungsprozesses nach Projektende und Ergebnissicherung nicht verlorengehen, sondern nach den sogenannten FAIR-Prinzipien (https://www.force11.org/group/fairgroup/fairprinciples) auffindbar, zugänglich, interoperabel und nachnutzbar publiziert werden. Ein FDMP beschreibt daher den Forschungsprozess von der Entwicklung einer Fragestellung bis zur Archivierung der Daten. Für die Beantragung von Drittmitteln bei Forschungsförderern sind solche FDMPs Pflichtbestandteil der guten wissenschaftlichen Praxis.

Mittlerweile gibt es einige zentrale Angebote, um für Projekte einen allgemeinen FDMP zu erstellen. So bietet das Projekt „Research Data Management Organiser“ (RDMO) etwa einen umfassenden Fragekatalog und eine darauf aufbauende Software zur Dokumentation von Arbeitsprozessen an. Auch das Projekt Liber stellt allgemeine Muster für einen FDMP zur Verfügung. An dieser Stelle soll auf zentrale Aspekte des geisteswissenschaftlichen Forschungsdatenmanagements hingewiesen und auf wichtige Standards des forschungsgeleiteten Datenmanagements eingegangen werden. Beratungen und Services hierfür leistet das Historische Datenzentrum Sachsen-Anhalt für geisteswissenschaftliche Fächer. Dabei kooperieren wir eng mit dem Forschungsdatenmangement der Universitäts- und Landesbibliothek Sachsen-Anhalt.

1 Projekt, Methode und Fragestellung und ihre Dokumentation

Zentral für die Entwicklung eines Forschungsdatenmanagementplanes ist die Formulierung einer wissenschaftlichen Fragestellung, die Klärung von Methoden und vor allem die Frage der Verwendung von Quellen bzw. anderer Daten im Projekt, die letztlich zur Erzeugung eigener Daten führen.

1.1 Forschungsfrage

Die Formulierung einer möglichst präzisen wissenschaftlichen Fragestellung und ihre konzentrierte Erläuterung, ist Basis der guten wissenschaftlichen Praxis und Tätigkeit. Je besser eine Forschungsfrage konkret formuliert wird, desto genauer können Methoden und Arbeitsprozesse beschrieben werden. Daher sollte diese Beschreibung in der späteren Dokumentation von Forschungsdaten ebenso wenig fehlen wie die Nennung zentraler Forschungsergebnisse (z. B. als Literaturliste, vorzugsweise auch als Datenbanken in Citavi oder Zotero).

1.2 Methoden

Welche Methoden werden verwendet? Mit welchen Programmen oder Arbeitsmitteln wird dazu gearbeitet? Die Dokumentation von Arbeitsschritten der Datenerhebung oder Nachnutzung stellt darüber Überlegungen an, wie die Fragestellung eines Projekts mithilfe von Daten und Quellen operationalisiert und modelliert werden kann. Dazu wird ein Codebuch zur Dokumentation des

Datenerhebungsprozesses und der Regeln der Datenmodellierung geführt. Dies gilt nicht nur für quantitative Daten, sondern selbstverständlich auch für qualitative Forschungsmethoden! Nur eine ausführliche Dokumentation der Forschungskategorien bzw. Themenkomplexe (Variablen), deren Merkmalsbeschreibungen und Definitionen sowie eine Dokumentation der methodischen Überlegungen und Vorgehensweisen sichert später auch die Verständlichkeit und Nachnutzung von Daten. Dies bringt auch dem jeweiligen Forschungsprojekt Klarheit, vor allem bei kooperativen Forschungsprozessen. Virtuelle Forschungsumgebungen oder geeignete Software (bspw. FuD des eScience Servicezentrums Trier auf einer Instanz des Historischen Datenzentrums Sachsen-Anhalt) und Beratungen können dabei unterstützen, ein geeignetes Datenmodell für eigene Fragestellungen zu entwickeln. Wichtig ist nicht, ob besonders komplexe Datenbanken entwickelt oder Programme genutzt werden, wichtig ist die Beachtung von fundamentalen Grundregeln der Datenerfassung und eine Dokumentation dieser Prozesse.

1.3 Datentypen und Sicherung

Bereits am Beginn eines Arbeitsprozesses sollten Überlegungen zur langfristigen Perspektiven von Daten angestellt werden. Grundlegend ist zu klären, welche Daten tatsächlich für eine transparente und nachhaltige Sicherung in Betracht kommen und wie hoch der Aufwand hierfür ist. Wie auch bei der Ergebnissicherung können bzw. müssen nicht alle Daten langfristig gesichert werden. Daher ist zu bestimmen, welche Daten ein spezifisches Forschungsergebnis angemessen repräsentieren. Insgesamt sind Kriterien für diesen Auswahlprozess zu formulieren. Dabei können auch unterschiedliche Versionen von Arbeitsprozessen gespeichert werden (Rohdaten, bereinigte Daten, Analysedaten, hochqualitätsgesicherte Daten, angereicherte Daten etc.).

Für die langfristige Nutzung von Daten sind zwei Aspekte entscheidend:

A) Der Aspekt der Nachnutzung: Eine schnelle Nachnutzung ermöglichen möglichst freie Daten mit hoher Interoperabilität (also ein rtf-Dokument statt einer PDF/A; statt eines Images einer historischen Karte die Shape-Dateien eines GIS).

B) Andererseits ist die Langzeitarchivierung zu bedenken. Dafür kommen nur ausgewählte Datenypen in Betracht. Eine Liste der möglichen Langzeitformate wird auf Opendata.uni-halle.de gegeben: https://opendata.uni-halle.de/DatenFormate_Share_it_2018_DE.pdf).

Für einen FDMP ist zu berücksichtigen, wie die Analysedaten in die entsprechen Langzeitformate überführt werden. Dabei kann es überlegenswert sein, verschiedene Formate für die unterschiedlichen Nutzungsaspekte und User anzubieten. So ist ein xml-TEI-File zwar eine hervorragende maschinenlesbare und langzeitarchivierbare Quelle, der Ortschronisten ist aber dennoch für die Lesefassung in Form einer PDF dankbar … (z.B. sowohl xml, rtf wie pdf-Dokument). Grundsätzlich muss über die Verwendung von kommerzieller Software nachgedacht werden, da diese häufig keine langzeitarchivierbaren Formate produzieren (z.B. f4Transkript oder MAXQDA).

1.4 Quellendigitalisierung und Transkription

Falls Quellen in digitalisierter Form genutzt werden, sollten diese den Qualitätsstandards der DFG-Praxisregeln (Metadaten, technische Parameter) entsprechen und genauso wie Forschungsdaten mithilfe von persistenten Identifiern (DOI, Purl, Handle etc.) verlässlich verlinkbar sein. Dazu müssen die Nutzungsrechte mit den Gedächtnisinstitutionen geklärt und dokumentiert werden. Im Falle von analogen Quellen können Kooperationen mit Archiven und Bibliotheken hilfreich sein, solche Bestände gemäß der oben genannten Qualitätsanforderung zu digitalisieren und von Seiten der anbietenden Gedächtnisinstitutionen webbasiert und persistent zur Verfügung zu stellen. Ansonsten sind eventuell alternative Strategien mit weiteren Digitalpartnern zu entwickeln. Solche Prozesse sollten bei der Projektplanung bedacht werden.

Falls Quellen im Forschungsprozess transkribiert oder mithilfe von Abstracts erfasst werden, muss sowohl auf die Zitierfähigkeit der einzelnen Daten wie auch die Dokumentation wesentlicher Transkriptionsregeln geachtet werden. Im Sinne der fachübergreifenden Nachnutzbarkeit von Daten sollten Normierungen historischer Schreibweisen bei der Transkription unterbleiben. Aus sprachwissenschaftlicher Perspektive wurden bereits Basisregeln zur Transkription von Texten (http://www.deutschestextarchiv.de/doku/basisformat/index.html) formuliert, die aus Sicht der historischen Forschung ergänzt werden müssen. Ähnliche Regeln gibt es auch für den Umgang mit Tondokumenten und Transkripten von Interviews in den Sozialwissenschaften (https://www.dfg.de/download/pdf/foerderung/grundlagen_dfg_foerderung/informationen_fachwissenschaften/geisteswissenschaften/standards_sprachkorpora.pdf). Mittlerweile erlauben OCR-Techniken die automatische Erkennung von Druck- und Handschriften (Transcribus) für die Kosten einzuplanen sind. Hier ist über den Umgang mit Fehlern und Formen der Fehlerbereinigung nachzudenken. Bei Verfahren des Textminings sind rechtliche Aspekte – besonders bezüglich der Publikation von Daten - zu klären.

2 Umgang mit Daten während des Forschungsprozesses

2.1 Metadaten und Datentransfer

Innerhalb des Forschungsprozesses spielen die Fragen des Datenzugangs bereits ebenso eine wichtige Rolle wie später für die Publikation von Daten. Dabei sind Fragen der Bezeichnung von Daten sowie der regelbasierten Sicherung von Metadaten und Dokumentationen zu klären. Orientierung bieten dazu die allgemeinen Regeln zur Zitation bibliografischer Angaben. Metadaten zu Autoren, Titeln einer Studie, Dateinamen, Zeit und Ort von Erhebungen, Sach- und Schlagwörtern, Herausgebern, Förderern, Kurzbeschreibungen und Beziehungen von Daten (möglichst auch abgebildet in Dateibezeichnungen) untereinander sind hier relevant. Zur Identifikation von Autoren dienen heute in der Regel die GND-Nummern von Personen (Open GND) oder die OrcID-Nummer. Für einzelne Datentypen gibt es momentan eine Vielzahl von Standards und fachspezifischen Beschreibungsformen, die zu beachten sind.

Der Datenserver Share_it der Universität Halle-Wittenberg nutzt Mets/Mods zur Wiedergabe von Metadaten. Andere Projekte nutzen das vereinfachte Schema von DataCite (https://schema.datacite.org/). Mittlerweile gibt es komplexe Überlegungen, wie solche Metadaten in Forschungsprojekten bereits mit der Erhebung von Daten automatisiert oder teilautomatisiert erzeugt werden können.

Fast selbstverständlich, aber oft nicht genügend geklärt, sind innerhalb eines Projektes ebenso Sicherungsstrategien und Backups zentral, die nicht allein auf die zukünftige Datensicherung verschoben werden sollte.

2.2 Rechtemanagement, Datenrecht und Datenschutz

Ebenso sind Strategien des Rechtemanagements, Datenschutzes oder des Umgangs mit Anonymisierung oder Pseudonymisierung wichtig. Personenbezogene Daten können nach dem Archivgesetz erst 30 Jahre nach dem Tod einer Person veröffentlicht werden. Zudem erfordern manche Fragestellungen das Votum eines Ethikrates. Daher müssen Studiendaten evtl. konsequent anonymisiert werden. Das Historische Datenzentrum Sachsen-Anhalt hat beispielsweise ein Anonymisierungskonzept für Oral History-Dokumente entwickelt, das Sie nachnutzen können. Wichtig ist ebenso, dass bei der Führung von Interviews die Zustimmungen der Probanden dokumentiert werden. Im Zweifel sollten dazu Rechtsberatungen bspw. bei den Datenschutzbeauftragten der Universität Halle eingeholt werden. Im Projekt muss geklärt und dokumentiert werden, wem die Daten gehören, wer die Urheber von Daten sind und wer während

und nach einem Projekt Zugang und Verantwortlichkeiten übernimmt. Für die spätere Nachnutzung sind ebenfalls die Lizenzen zu Nutzungsbestimmungen (etwa Creativ Commons) anzugeben.

2.3 Qualitätssicherung

Während und nach der Erhebung von Daten sind Strategien zur Sicherung der Datenqualität wichtig. Hierfür sind von der einfachen, automatisierten und daher häufig fehlerhaften Erhebung bis zum Verfahren mit Double Keying unterschiedliche Stufen der Qualitätssicherung denkbar, die jedoch auch unterschiedlich hohe finanzielle und personelle Ressourcen binden. Manchmal kann es effektiver sein, nicht alle Fehler in Massendaten zu bereinigen. Solche Entscheidungen sind sichtbar zu machen und zu dokumentieren.

Bereits mit der Datenerhebung können Kontrollvariablen zur Datenbereinigung erfasst oder Regeln zum Umgang mit Ungenauigkeit oder Datenverlust definiert werden, um die Datenqualität zu erhöhen. Qualitätssicherung kann auch über die Verwendung von Normdaten und kontrollierten Vokabularen bzw. fachübergreifenden Standards erfolgen (z.B. Generalnormdatei (GND) der Nationalbibliothek, TEI-Regeln bei der Auszeichnung von xml-Daten bei elektronischen Editionen oder internationalen Standards der Klassifikation (z.B. Klassifikation der Berufe 2010 / Ontologie der historischen Berufsbezeichnungen), Standards der Geokodierung)), die häufig auch für die Analyse und Visualisierung von Forschungsdaten wichtige Aufgaben erfüllen. Je besser Daten mit solchen Mitteln erschlossen werden, desto hochwertiger werden sie. Das Historische Datenzentrum kann hierzu beraten und Werkzeuge empfehlen bzw. bereitstellen. Hierfür sind in der Regel finanzielle Ressourcen in Projekten ebenso einzuplanen wie für den Gesamtprozess der Metadatenerhebung und Dokumentation und für die Langzeitarchivierung und Publikation.

3 Datenpublikation

3.1 Vertrauenswürdige Datenarchive

Nach Abschluss eines Projekts steht mittlerweile nicht nur die Veröffentlichung der Forschungsergebnisse, sondern ebenso die Publikation der Forschungsdaten. Dafür sind bisher zahlreiche Forschungsdatenrepositorien geschaffen worden, die zum Teil fachspezifisch oder datenspezifisch differenziert Daten langfristig archivieren. Damit sichern Hostingdienste für Daten mittlerweile eine Speicherung für mindestens 10 Jahre zu, wofür in den meisten Fällen auch Kosten anfallen. Diese Arbeitsschritte erfordern häufig ihre eigene Aktivität (Selfpublishing). Vertrauenswürdige (zertifizierte) Archive können über das zentrale Verzeichnis re3data.org (Registry of Research Data Repositories) aufgefunden werden. Innerhalb der Fachcommunity der historisch arbeitenden Disziplinen besteht allerdings perspektiv der Anspruch, Daten analog zu Büchern dauerhaft zu sichern. Insgesamt gibt es im Rahmen des Aufbaus von Nationalen Forschungsdateninfrastrukturen Bemühungen zur Entwicklung eines solchen zentralen Datenspeichers für historische Daten.

3.2 Hist-Data auf dem Open-Data Server der Universitäts- und Landesbibliothek (share_it)

Forschungsdaten können über das Historische Datenzentrum Sachsen-Anhalt im Open-Access-Publikations- und Forschungsdatenspeicher der Landes- und Universitätsbibliothek Sachsen-Anhalt Share_it abgelegt werden. Hier wird ein begleiteter Dokumentations- und Publikationsprozess angeboten, in dem das Datenzentrum als Herausgeberin fungiert. Das Archiv ermöglicht die Speicherung der Daten und das Auffinden von Ressourcen durch geeignete Metadaten und persistente Referenzierung (z.B. DOI) sowie regelbasierter Dokumentationen. Das DINI-zertifizierte Repository nutzt verschiedene Schnittstellen, um die Auffindbarkeit über Online-Recherchen zu ermöglichen. Dabei werden verschiedene Lizenzmodelle wählbar, die sowohl freie

Nutzungsmöglichkeiten (Creativ Commons) wie auch urheberrechtlich geschützte Open-Access-Angebote unterstützen. Ebenso ist die Publikation von geschützten Daten möglich.

3.3 Datenpaper und Big Data

Publikationen von Daten werden heute zum Teil mit sogenannten Datenpapern begleitet, die nicht einfach als Dokumentationen von Daten betrachtet werden können, sondern Daten umfassender in den Erhebungsprozess, in Nutzungsszenarien und Auswertungsmethoden einbinden. Zugleich werden größere Datenbestände ähnlicher Daten aufgebaut (z.B. Deutsches Textarchiv (Clarin), Archiv der gesprochenen Sprache, Sammlung genealogischer Daten (Verein für Computergenealogie)). Auch ihre Daten können Teil solcher Datensammlungen werden.