Forschungsdatenmanagement

Jochen Apel

https://orcid.org/0000-0002-0395-4120

Abstract Der Beitrag bietet eine Einführung in das Forschungsdatenmanagement. Ausgehend von den FAIR Data Principles werden verschiedene Aspekte und Ziele des Datenmanagements – von der Projektplanung über die Datenorganisation im Projekt bis zur Veröffentlichung und Archivierung von Forschungsdaten – in ihren Grundzügen skizziert.

Keywords Forschungsdaten, Forschungsdatenmanagement, FAIR-Prinzipien, Geisteswissenschaften, Digital Humanities

1. Forschungsdaten und Forschungsdatenmanagement

Ein strukturierter und planvoller Umgang mit Forschungsdaten ist eine zentrale Anforderung an jedes Forschungsprojekt – nicht nur, aber selbstverständlich auch in der Theologie. Mit dem zunehmenden Einsatz digitaler Werkzeuge und Methoden steigen dabei die Anforderungen an ein effektives und nachhaltiges Datenmanagement, das die Qualität, Nachvollziehbarkeit und Nachnutzbarkeit von Forschungsergebnissen sicherstellt. Das vorliegende Kapitel erläutert daher die Grundzüge des Forschungsdatenmanagements (im Folgenden: FDM). Der Forschungsdatenbegriff wird dabei in der folgenden Weise verstanden:

> Forschungsdaten sind digitale Daten, die im Forschungsprozess erzeugt, gesammelt oder zusammengestellt werden und auf deren Grundlage wissenschaftliche Hypothesen, Modelle oder Theorien gebildet sowie bestätigt bzw. widerlegt werden.[1]

Gemäß dieser Begriffsbestimmung werden digitale Daten durch die spezifischen *epistemischen Rollen,* die sie im Forschungsprozess spielen, zu Forschungsdaten. Das scheint gerade für den Bereich der geisteswissenschaftlichen Forschung treffend,

1 Einen derartigen Definitionsvorschlag hat der Autor des vorliegenden Kapitels 2014 erstmals für eine lokale Webseite formuliert. Vgl. https://web.archive.org/web/20230320185206/https://data.uni-heidelberg.de/faq.html, zuletzt aufgerufen am 19.06.2024.

Jochen Apel: Forschungsdatenmanagement, in: Christopher A. Nunn und Frederike van Oorschot (Hrsg.): Kompendium Computational Theology 1. Forschungspraktiken in den Digital Humanities, Heidelberg: heiBOOKS, 2024, S. 427–440. DOI: https://doi.org/10.11588/heibooks.1459.c21925

wo in vielen Fällen Daten nicht im Rahmen eines Forschungsprojektes erzeugt werden, sondern bereits vorliegen, aber durch die wissenschaftliche Auseinandersetzung erst *zu Forschungsdaten werden:* Wird Quellenmaterial zum Gegenstand wissenschaftlicher Auseinandersetzung, werden die entsprechenden digitalen Daten zu Forschungsdaten. Die vorgeschlagene Begriffsbestimmung ist damit relativ breit. Unter sie lassen sich beispielsweise die folgenden Datentypen subsumieren: Digitale Quellen und Digitalisate von Quellen (neben Text auch Bild-, Video- oder 3D-Daten), ebenso aber auch Bearbeitungen einer Quelle (z. B. ein OCR-generierter Text oder eine in TEI-XML codierte Edition), Ergebnisse von Analysen (z. B. statistische Resultate einer quantitativen Korpusanalyse) oder Datenbanken, in denen Informationen strukturiert zusammengestellt sind.

In der einschlägigen Literatur finden sich eine Reihe weiterer Versuche, sich dem Forschungsdatenbegriff anzunähern, die Geiger (2023) in einer instruktiven Übersicht zusammenstellt. Kennzeichnend für die geisteswissenschaftliche Forschungsdatenlandschaft ist dabei, dass es sich um ein heterogenes Feld handelt, in dem unterschiedliche Datentypen, -formate und -strukturen relevant sind und in dem es nur bis zu einem gewissen Grad bzw. in bestimmten Teilbereichen etablierte Standards gibt (vgl. Pempe 2012).

Auf dieses Verständnis von Forschungsdaten aufbauend, lässt sich genauer fassen, was unter FDM verstanden wird. Hier sollen zwei Begriffsbestimmungen aus der einschlägigen Literatur angeführt werden, die gemeinsam das Bedeutungsspektrum des Begriffs aufspannen:

> Unter dem Management von Forschungsdaten werden alle Maßnahmen verstanden, die sicherstellen, dass digitale Forschungsdaten nutzbar sind. Was dafür notwendig ist, variiert aber stark mit den verschiedenen Zwecken, für die Forschungsdaten genutzt werden sollen. Es lassen sich vier Arten von Zwecken unterscheiden:
>
> - die Nutzung als Arbeitskopie für das wissenschaftliche Arbeiten,
> - die Nachnutzung von Forschungsdaten für spätere Forschung,
> - die Aufbewahrung als Dokumentation des korrekten wissenschaftlichen Arbeitens und
> - die Aufbewahrung, um rechtlichen oder anderen forschungsfremden Anforderungen nachzukommen. (Enke & Ludwig 2013, 13)

> Research data management concerns the organisation of data, from its entry to the research cycle through to the dissemination and archiving of valuable results. It aims to ensure reliable verification of results, and permits new and innovative research built on existing information. (Whyte & Tedds 2011)

Während die erstgenannte Definition insbesondere die Nutzbarkeit der Daten als Ziel des FDM identifiziert, ist der wesentliche Aspekt des in der zweiten Definition verwendeten Bildes des Datenlebenszyklus, dass die Verzahnung der einzelnen Phasen des Forschungsprozesses betont wird: Der Umgang mit Forschungsdaten in späteren Phasen eines Projekts hängt von Weichenstellungen in den früheren Phasen ab. Wer z. B. mit Daten arbeitet, an denen Dritte Rechte halten, und diese Daten veröffentlichen möchte, sollte erforderliche Rechteklärungen bereits bei der Datenerhebung vornehmen; wer möchte, dass ein eigenes Folgevorhaben oder Dritte später die eigenen Forschungsdaten sinnvoll nachnutzen können, der muss bereits während des Forschungsprozesses Ressourcen einplanen, um die Daten verständlich zu dokumentieren etc.

2. FAIR Data Principles und die Ziele des FDM

Im vorhergehenden Abschnitt ist in Anschluss an Enke und Ludwig formuliert worden, dass es im FDM darum gehe, Forschungsdaten nutzbar zu halten. Doch was heißt dies im Detail? Über welche Eigenschaften müssen Forschungsdaten verfügen, damit sie nutzbar sind? Eine Antwort auf diese Fragen liefern die FAIR Data Principles (Wilkinson et al. 2016).[2]

FAIR steht dabei für die vier Eigenschaften *Findable, Accessible, Interoperable und Reusable.* Im Detail werden diese folgendermaßen ausbuchstabiert; hier in der deutschen Übersetzung von Angela Kailus:

> Findable – Auffindbar
> F1. (Meta-)Daten wird ein global eindeutiger und persistenter Identifikator zugewiesen.
> F2. Daten werden mit umfangreichen Metadaten (vgl. R1) beschrieben.
> F3. Metadaten enthalten eindeutig und explizit den Identifikator der Daten, die sie beschreiben. F4. (Meta-)Daten werden in einer durchsuchbaren Ressource registriert oder indiziert.
>
> Accessible – Zugänglich
> A1. (Meta-)Daten sind über ihren Identifikator mithilfe eines standardisierten Kommunikationsprotokolls abrufbar.
> A1.1 Das Protokoll ist offen, kostenlos und universell implementierbar.
> A1.2 Das Protokoll unterstützt bei Bedarf Verfahren zur Authentifizierung und Rechteverwaltung. A2. Metadaten bleiben verfügbar, auch wenn die zugehörigen Daten nicht (mehr) verfügbar sind.

2 Vgl. auch https://www.go-fair.org/fair-principles, zuletzt aufgerufen am 19.06.2024.

> Interoperable – Interoperabel
> I1. (Meta-)Daten nutzen eine formale, zugängliche, gemeinsam genutzte und breit anwendbare Sprache für die Wissensrepräsentation.
> I2. (Meta-)Daten enthalten Vokabulare, welche den FAIR-Prinzipien folgen.
> I3. (Meta-)Daten enthalten qualifizierte Verweise auf andere (Meta-)Daten.
>
> Reusable – Nachnutzbar
> R1. (Meta-)Daten sind detailliert beschrieben und enthalten präzise, relevante Attribute. R1.1. (Meta-)Daten enthalten eine eindeutige, zugreifbare Angabe einer Nutzungslizenz. R1.2. (Meta-)Daten enthalten detaillierte Provenienz-Informationen.
> R1.3. (Meta-)Daten entsprechen den fachgebietsrelevanten Community-Standards.
> (Kailus 2023)

Ohne dass hier die verschiedenen Teilaspekte der FAIR-Prinzipien im Detail diskutiert werden könnten, so lassen sich doch die folgenden wesentlichen Charakteristika identifizieren: *Auffindbarkeit („Findability")* basiert maßgeblich auf der umfassenden Dokumentation und Beschreibung anhand von Metadaten sowie der Nutzung von persistenten digitalen Identifikatoren (z. B. *Digital Object Identifier* (DOI)), die die Grundlage für die stabile Referenzierbarkeit in Publikationen, Nachweissystemen und Suchmaschinen bilden. *Zugänglichkeit („Accessibility")* gründet darauf, dass die Daten so offen wie möglich und so geschützt wie nötig verfügbar gemacht werden. Das heißt dass Forschungsdaten im besten Fall als Open Research Data publiziert werden. Sollte dies nicht möglich sein, können die Daten aber ggf. für berechtigte User mittels geeigneter Authentifizierungsmethoden bereitgestellt werden. Wenn auch diese Möglichkeit nicht besteht, sollten mindestens die beschreibenden Metadaten öffentlich verfügbar sein. Die *Interoperabilität („Interoperability")* wiederum basiert wesentlich auf Standards für Daten und Metadaten. Die Verwendung standardisierter Datenstrukturen, nicht-proprietärer Datenformate oder normierter Vokabulare sind die Grundlage für die einfache, im besten Fall durch Maschinenlesbarkeit automatisierbare, Benutzbarkeit der Daten sowie die mögliche Integration der Daten mit weiteren Datenbeständen. Die *Nachnutzbarkeit („Reuseability")* basiert wiederum zum einen auf einer reichhaltigen inhaltlichen Beschreibung und Dokumentation inkl. Provenienzinformationen, zum anderen auf rechtlichen Festlegungen, wie die Daten nachgenutzt werden können. Im besten Fall geschieht dies durch die Verwendung geeigneter Open-Content-Lizenzen (z. B. Creative Commons-Lizenzen).

Die Publikation und Bereitstellung der Daten gemäß der FAIR-Prinzipien ist eine Aufgabe, bei der Forschende z. B. durch entsprechende Forschungsinfrastrukturen unterstützt werden. So muss beispielsweise die Beschreibung und Dokumentation der Daten durch die Forschenden selbst erfolgen. Aber nur, wenn die für die Bereitstellung der Daten verwendeten Datenrepositorien über geeignete Funktionalitäten

verfügen, konkret z. B. geeignete Metadatenstandards unterstützen oder persistente Identifikatoren anbieten, werden die Daten auffindbar im Sinne der FAIR-Prinzipien sein.

3. Rahmenbedingungen und Richtlinien für das Datenmanagement

Neben den sich aus dem Forschungszusammenhang, aber auch aus den jeweiligen fachspezifischen Forschungspraktiken ergebenden Anforderungen an das FDM sind übergeordnete Rahmenbedingungen und Richtlinien zu beachten, die wichtige Leitplanken für Forschungsprojekte formulieren. Forschende tun daher gut daran, sich bereits bei der Planung eines Projekts mit diesen Rahmenbedingungen auseinanderzusetzen. So formulieren Förderinstitutionen mittlerweile fast flächendeckend Anforderungen an das FDM der von ihnen unterstützten Projekte. Dies geschieht entweder in Form zentraler Richtlinien, wie es u. a. die DFG und die EU tun, und/oder in Form spezifischer Modalitäten im Rahmen der einzelnen Programmlinien, wie dies z. B. beim BMBF der Fall ist.[3]

Aber nicht nur Drittmittelgebende, sondern auch Universitäten und andere Forschungseinrichtungen formulieren einschlägige Regeln zum Umgang mit Forschungsdaten.[4] Zum Teil geschieht dies in dezidierten Datenpolicies oder in Kodizes zur Sicherung guter akademischer Praxis. Darüber hinaus formulieren auch die Fachcommunities selbst – häufig über ihre jeweiligen Fachgesellschaften – sowie Verlage und Fachzeitschriften weitere Rahmenbedingungen zum Umgang mit Forschungsdaten.[5]

4. Datenmanagementpläne

Als Ausgangspunkt und Grundlage des FDM kann ein sog. Datenmanagementplan dienen. Ein Datenmanagementplan ist ein Dokument, in dem zusammengestellt ist, welche Daten erhoben bzw. verwendet werden und wie mit diesen Daten im Projekt sowie nach dem Ablauf der Projektlaufzeit umgegangen werden soll. Datenmanage-

3 Vgl. https://www.dfg.de/foerderung/grundlagen_rahmenbedingungen/forschungsdaten/index.html sowie https://www.dfg.de/download/pdf/foerderung/grundlagen_dfg_foerderung/forschungsdaten/forschungsdaten_checkliste_de.pdf, https://www.openaire.eu/rdm-in-horizon-europe-proposals, https://forschungsdaten.info/themen/informieren-und-planen/foerderrichtlinien. Alle genannten Adressen wurden zuletzt am 19.06.2024 aufgerufen.

4 Vgl. https://www.forschungsdaten.org/index.php/Data_Policies, zuletzt aufgerufen am 19.06.2024.

5 Vgl. https://forschungsdaten.info/themen/ethik-und-gute-wissenschaftliche-praxis/leitlinien-und-policies, zuletzt aufgerufen am 19.06.2024.

mentpläne sind dabei im besten Fall fortzuschreibende Dokumente, die im Projektverlauf regelmäßig konsultiert und im Bedarfsfall aktualisiert werden.

Die Zwecke eines Datenmanagementplans bestehen darin, wohlbegründete Entscheidungen für den Umgang mit den Forschungsdaten des eigenen Projekts treffen zu können, Risiken und Herausforderungen frühzeitig zu identifizieren, in kooperativen Projekten einheitliche Vorgehensweisen und Standards für die gemeinschaftliche Datennutzung zu etablieren und damit die Nachhaltigkeit der Daten sicherzustellen sowie durch die Gestaltung effizienter Prozesse, Zeit und Aufwand zu sparen. Die zentralen Themen, die ein Datenmanagementplan adressieren sollte, hat William Michener in Form von zehn Fragestellungen formuliert (Michener 2015).[6] In einer Übersetzung von Jens Dierkes lauten diese Fragen:

1. Was sind die Anforderungen der Förderorganisationen zum FDM?
2. Welche Daten werden gesammelt?
3. Wie werden die Daten organisiert?
4. Wie werden die Daten dokumentiert?
5. Wie wird die Datenqualität gewährleistet?
6. Wie sieht die Datenspeicherungs- und Archivierungsstrategie aus?
7. Wie wird mit Daten im Forschungsvorhaben und darüberhinausgehend umgegangen (Daten-Policy)?
8. Wie werden die Daten disseminiert?
9. Welche Rollen und Verantwortlichkeiten gibt es?
10. Wie sieht ein realistisches Budget für das FDM aus?
    (Dierkes 2021, 308 f.)

Um einen Datenmanagementplan für das eigene Projekt zu erstellen, existieren eine Vielzahl von Templates, Checklisten sowie webbasierten Tools (Dierkes 2021, 310). Insbesondere Tools wie die Dienste RDMO und DMPonline sind hier hilfreich, da sie in Form von umfassenden Fragenkatalogen den Blick auf sämtliche potentiell relevante Aspekte lenken, die es in der Datenmanagementplanung zu beachten gilt.

6 Vgl hierzu auch https://forschungsdaten.info/themen/informieren-und-planen/datenmanagementplan und den Leitfaden von Science Europe: https://www.scienceeurope.org/media/4brkxxe5/se_rdm_practical_guide_extended_final.pdf. Beide Adressen wurden zuletzt am 19.06.2024 aufgerufen. Hier werden – über die untenstehende Auflistung hinaus – u. a. auch Fragen der Datennachnutzung, der rechtlichen und ethischen Rahmenbedingungen sowie des Teilens von Daten behandelt.

5. Datenmanagement im Projekt (hot data)

Wesentliche Aspekte des Datenmanagements betreffen den planvollen Umgang mit den Forschungsdaten während der Laufzeit eines Projekts. Diese sind in vielerlei Hinsicht fach-, methoden- und datenspezifisch und können in einem einführenden Übersichtstext wie dem vorliegenden nicht ausgeführt werden. Sie werden aber in anderen Kapiteln dieses Handbuchs thematisiert, in denen konkrete Fallbeispiele und spezifische Methoden diskutiert werden. Im Rahmen dieses Kapitels sollen vielmehr exemplarisch generische Aspekte aufgegriffen werden, die in jedem Forschungsprojekt von Relevanz sind.

5.1 Datenerhebung und -sammlung

Auf welche Weise die Daten für ein Forschungsprojekt erhoben werden, hängt offenkundig vom jeweiligen Projekt ab. In geisteswissenschaftlichen Projekten werden die Daten häufig nicht selbst erhoben, sondern von Dritten bereitgestellt, z. B. als öffentlich verfügbare Digitalisate von Bibliotheken und Archiven. Nicht nur wegen erforderlicher Provenienzangaben oder eventuellen Anforderungen der Datengeber hinsichtlich der Nutzung und Weitergabe der Daten ist diese Information bedeutsam, sondern auch weil ggf. Fragen nach der Archivierung und Veröffentlichung der Daten anders beantwortet werden können. Wenn die einem Projekt zugrunde liegenden Daten von einem vertrauenswürdigen Anbieter, z. B. einer Bibliothek, publiziert wurden, dann wird diese Institution auch die langfristige Archivierung der Daten sicherstellen. Das heißt hierfür ist im Projekt ggf. keine eigene Lösung zu finden, sondern man kann sich primär auf den Umgang mit den eigenen Arbeitskopien fokussieren, die nur während der Projektlaufzeit benötigt werden. Selbstverständlich kann es aber auch bei der Nutzung von Daten Dritter der Fall sein, dass im Projekt von diesen abgeleitete eigene Forschungsdaten generiert werden, z. B. Annotationen, tabellarische Auswertungen oder statistische Analysen, deren langfristige Archivierung und Veröffentlichung sinnvoll und damit Aufgabe des Forschungsdatenmanagement des eigenen Projekts ist.

Unabhängig davon, auf welche Weise die Datenerhebung im Detail erfolgt, gilt, dass die für die langfristige Nutzbarkeit der Daten erforderliche Dokumentation, im besten Fall unter Verwendung einschlägiger Metadatenstandards, so früh wie möglich erfolgen sollte. Die Dokumentation kann dabei in unterschiedlichen Formen erfolgen, die sich auch kombinieren lassen, z. B. in Read-Me-Dateien, strukturierten Metadatenbanken, einem Wiki oder auch direkt im Datenmanagementplan.[7]

7 Vgl. https://forschungsdaten.info/themen/beschreiben-und-dokumentieren/datendokumentation, zuletzt aufgerufen am 19.06.2024.

5.2 Speicherung, Backup und Löschung

Für jedes Forschungsprojekt stellt sich die Frage, wo und wie die projektrelevanten Forschungsdaten gespeichert werden, z. B. auf der Festplatte eines lokalen PCs, auf einem Server des Instituts oder einem zentralen Speicherdienst der eigenen Universität oder auch eines kommerziellen Anbieters. Insbesondere für große Datenmengen empfiehlt sich die Nutzung zentraler Dienste. Dabei benötigt jedes Forschungsprojekt eine geeignete Backup-Strategie. Als grobe Orientierung dient hier die sog 3-2-1-Regel: Drei Datenkopien auf zwei unterschiedlichen Speichermedien, davon eine an einem externen Standort (vgl. Krogh 2009, Kapitel 6).

Speichert man die Daten selbst auf einem eigenen Rechner, muss man für das Backup selbst Vorsorge treffen, nutzt man institutionelle Strukturen oder anderweitige Dienstleister, gehört dies ggf. zum jeweiligen Serviceumfang. Daher ist letzteren Lösungen im Grundsatz der Vorzug zu geben. Bei kleinen Datenvolumina bieten die mittlerweile relativ flächendeckend von Forschungseinrichtungen angebotenen Sync-and-Share-Dienste eine niedrigschwellige Lösung.

Darüber hinaus gehören zum FDM ggf. bereits in dieser Phase Auswahlprozesse bezüglich der Löschung von Dateien. Selbstverständlich sollten im Projektverlauf niemals die der Forschung zugrunde liegenden Rohdaten gelöscht werden. Durchaus aber können prozessierte Versionen der Daten gelöscht werden, wenn diese nicht mehr benötigt werden. Dabei sollte jedoch immer eine Dokumentation der durchgeführten Schritte erfolgen, um notfalls die entsprechende Prozessierung auf Grundlage der Rohdaten noch einmal nachvollziehen zu können. Bearbeitungsstände der Daten, die die direkte Grundlage für Veröffentlichungen bilden, sollten aufbewahrt werden. Erreicht das Forschungsprojekt sein Ende, sind dann ggf. ergänzende weitere Regelungen zu treffen (z. B. hinsichtlich der Löschfrist für bestimmte Daten).

5.3 Data Sharing

Bei der Wahl eines geeigneten Speichersystems ist neben Datensicherheit, Backup und Kosten relevant, ob die Daten im Forschungsprojekt mit weiteren Personen geteilt werden sollen und in welcher Form dies erfolgen soll. Wird z. B. im Rahmen einer Forschungsgruppe oder eines Projektverbunds gemeinschaftlich mit den Daten gearbeitet? Sind hierbei parallele Zugriffe auf und/oder zeitgleiches Arbeiten in Dateien erforderlich? Sollen Kooperationspartner*innen anlassbezogen oder dauerhaft Zugriff auf einen Teil oder die Gesamtheit der Daten erhalten? Hier ist darauf zu achten, dass gewählte Dienste im besten Fall bereits ein passendes, möglichst feingranular justierbares Rechtemanagement bereitstellen, in dem entsprechende Zugriffs- und Bearbeitungsfreigaben abgebildet werden können.

5.4 Datenorganisation

Ebenso wie die zuvor thematisierten Aspekte hängt die Datenorganisation stark vom jeweiligen Forschungsprojekt und dem dort relevanten Datenmaterial ab. Mit Bilddaten ist anders umzugehen als mit Textkorpora oder Daten in Tabellenform; Forschende, die alleine arbeiten, haben andere Anforderungen als kooperative Forschungsvorhaben, in denen eine gemeinsame Datenbasis genutzt wird; große Datenmengen müssen anders behandelt werden als kleine Volumina, usw. Dennoch lassen sich einige generische Empfehlungen für die Datenorganisation formulieren.

Die Basis für eine sinnvolle Form der Datenorganisation bilden Konventionen zur Ordnerstruktur und Dateibenennung. Das Handbuch *The Turing Way Community* (2022, Kapitel „Research Data Management“) bietet hier eine übersichtliche Einführung. Forschungsdaten werden in der Regel in einer Ordnerstruktur abgelegt. Diese sollte einem klaren System folgen, beispielsweise einer chronologischen Sortierung, einer Sortierung nach eingesetzten Erhebungsmethoden, der Zuordnung zu einzelnen Teilprojekten o. ä. Innerhalb der Ordner sollten die Dateien dann in einer systematischen Weise benannt werden, beispielsweise indem man für die Dateinamen das Erstellungsdatum in der Form YYYYMMDD verwendet. Weitere Benennungselemente können dann z. B. der Datentyp bzw. die Erhebungsmethode, Namen des*der Forschenden bzw. Initialen (v. a. bei kooperativen Projekten) und Versionsnummern sein.

Leitgedanke sollte sein, dass die Dateinamen Kontext zur jeweiligen Datei liefern sollen, um sie von ähnlichen Dateien sowie von anderen Versionen der gleichen Datei zu unterscheiden. Darüber hinaus kann es sinnvoll sein, sich mit geeigneten Softwaretools zur Unterstützung der Datenorganisation zu befassen. Insbesondere textbasierte Daten können evtl. effizient in einem Git-System verwaltet und versioniert werden; Dateibenennungstools können helfen, größere Mengen an Dateien gemäß einem einheitlichen Schema zu benennen.

5.5 Wahl von Daten- und Dateiformaten

Im Hinblick auf die Aufbereitung der Daten im Sinne der FAIR-Prinzipen sollte zudem frühzeitig das Augenmerk auf die Wahl geeigneter Daten- und Dateiformate gelenkt werden. Wesentlich ist hier insbesondere die Unterscheidung zwischen proprietären und nicht-proprietären (Open Source-)Formaten sowie zwischen Binär- und Textformaten.

Wann immer möglich sollten nicht nicht-proprietäre Formate gewählt werden. Bei Binär- und Textformaten ist hingegen zu differenzieren. Binärformate sind in der Regel weniger speicherintensiv, viele Softwareprodukte verarbeiten und liefern Binärformate aus. Dennoch kann es im Hinblick auf die Langzeitarchivierung der Daten sinnvoll sein, diese (auch) in einem textuellen Format vorzuhalten (sofern

eine entsprechende Migration möglich ist). Es ist somit durchaus denkbar, dass für unterschiedliche Zwecke, nämlich die aktive Arbeit mit den Daten im Projekt und die spätere Archivierung, unterschiedliche Formate das Mittel der Wahl sind. Hier ist frühzeitig zu prüfen, ob eine entsprechende Formatmigration vor der Überführung der Daten in ein geeignetes Archiv oder Datenrepositorium möglich ist, v. a. auch, ob dies ohne Informationsverlust möglich ist. Wenn eine Konvertierung nur mit Informationsverlust realisierbar ist, ist zu bedenken, ob dieser Verlust signifikante Eigenschaften betrifft oder nicht. Beispielsweise kann eine Tabelle im Excelformat Formatierungen enthalten wie fettgedruckte Spaltenüberschriften, die bei einer Konversion ins CSV-Format verloren gehen. Sofern es sich aber bei der Formatierung nicht um eine für die Archivierung signifikante Eigenschaft handelt, ist ein entsprechender Informationsverlust durch die Migration akzeptabel.

Die Schweizer Koordinierungsstelle für die dauerhafte Archivierung elektronischer Unterlagen (KOST) liefert eine hilfreiche Übersicht über mehr als 50 verbreitete Datenformate und ihre Eignung für die Langzeitarchivierung.[8]

6. Archivierung und Veröffentlichung (cold data)

6.1 Repositorien und Datenpublikation

Forschungsdaten sollten spätestens zum Projektende an ein vertrauenswürdiges Forschungsdatenrepositorium oder Datenzentrum übergeben werden, das die nachhaltige langfristige Bereitstellung der Forschungsdaten übernimmt.

Bei der Wahl eines geeigneten Publikationsortes für die eigenen Forschungsdaten können folgende Fragen leitend sein: Gibt es besondere Schutzanforderungen an die Bereitstellung der Daten? Dürfen die Daten beispielsweise nicht öffentlich, sondern nur unter bestimmten Voraussetzungen auf Nachfrage zugänglich sein? In diesem Fall muss ein Dienst oder Repositorium gefunden werden, der bzw. das diese Form des kontrollierten Zugangs gewährleisten kann. Wenn dies nicht der Fall ist, sollte im nächsten Schritt geprüft werden, ob es passende Fachrepositorien gibt. Diese stellen i. d. R. durch ihre Spezialisierung auf bestimmte Disziplinen und/oder Datentypen die geeignetsten Publikationsorte für Forschungsdaten dar. Datenpublikationen werden dort im fachlichen Kontext in einer gemeinsamen Sammlung mit weiteren einschlägigen Daten aus dem Fach sichtbar. Fachrepositorien unterstützen zudem im Fach etablierte Metadatenstandards, ggf. bieten sie spezifische Such- oder Visualisierungsfunktionalitäten und zudem können die Betreiber*innen eines Fachrepositorium ggf. eine umfassende Kuration und Prüfung der Daten auf Basis einschlägiger fachlicher Expertise gewährleisten.

8 S. https://kost-ceco.ch/cms/kad_intro_de.html, zuletzt aufgerufen am 19.06.2024.

Insbesondere wenn kein passender fachlicher Dienst zur Verfügung steht, können ggf. institutionelle Repositorien oder andere generische, d. h. nicht fachspezifische Dienste genutzt werden. Diese bieten nicht die zuvor beschriebenen fachspezifischen Funktionalitäten, stellen aber dennoch vertrauenswürdige und ebenfalls den FAIR-Prinzipien genügende Publikationsorte dar. Insbesondere der kurze Draht zu lokalen Forschungsdaten-Support-Units und direkte Betreuung vor Ort können zudem Argumente für die Nutzung institutioneller Strukturen sein.

Sofern dem keine spezifischen Gründe entgegenstehen, sollten die Daten dabei als Open Research Data publiziert werden, wobei nach Möglichkeit Open-Content-Lizenzen genutzt werden, um eine möglichst breite und niedrigschwellige Nachnutzung der Daten zu gewährleisten. Im Bereich der Forschungsdaten haben sich hier, ähnlich wie bei Open-Access-Publikationen, die sog. Creative-Commons-Lizenzen etabliert, wobei insbesondere mit Blick auf die maschinelle oder zumindest teilautomatisierte Nutzung von Daten beispielsweise in sog. Big-Data-Analysen es auch Stimmen gibt, die dafür plädieren, Forschungsdaten möglichst gemeinfrei nutzbar zu machen, beispielsweise über den CC0-Waiver (Brettschneider et al. 2021).

Das internationale Verzeichnis von Forschungsdatenrepositorien *re3data* listet im Mai 2023 in Summe 22 Fachrepositorien aus dem Bereich der Theologie.[9] Darüber hinaus können aber, je nach Forschungsfeld, andere geisteswissenschaftliche Repositorien geeignete Publikationsorte sein, z. B. die bereits bestehenden oder im Aufbau befindlichen Angebote der geisteswissenschaftlichen NFDI-Konsortien NFDI4Culture, NFDI4Memory, NFDI4Objects und Text+[10] oder die Repositorien des CLARIAH-Verbundes.[11] Auch die AG Datenzentren im Verband DHd *(Digital Humanities im deutschsprachigen Raum)* bietet eine Anlaufstation bei der Suche nach geeigneten Plattformen.[12] Einen Sonderfall in diesem Kontext stellen individuelle Webseiten oder webbasierte Datenbanken dar, die als „Präsentationsschichten" von Forschungsdaten ein häufiges Resultat geisteswissenschaftlicher Projekte sind. Die Wahl solcher individueller Präsentationsformate mag aufgrund der Heterogenität geisteswissenschaftlicher Fragestellungen in der Sache häufig sinnvoll sein, wirft aber unmittelbar das Problem der Nachhaltigkeit auf: Wie und durch wen sollen solche Datenprodukte über die Projektlaufzeit hinaus langfristig betrieben werden können? Dies ist nur durch frühzeitige Involvierung eines Infrastrukturpartners möglich und sollte mit der Fallback-Option der Abschaltung der Webpräsentation bei weiterer Bereitstellung der Rohdaten über ein Repositorium gekoppelt sein.

9 S. https://www.re3data.org/search?subjects[]=107%20Theology, zuletzt aufgerufen am 19.06.2024.

10 S. https://nfdi4culture.de, https://4memory.de, https://www.nfdi4objects.net und https://www.text-plus.org. Alle Adressen wurden zuletzt am 19.06.2024 aufgerufen.

11 S. https://www.clariah.de/publizieren-archivieren, zuletzt aufgerufen am 19.06.2024.

12 S. https://dhd-ag-datenzentren.github.io, zuletzt aufgerufen am 19.06.2024.

6.2 Langzeitarchivierung

Insbesondere für geisteswissenschaftliche Forschungsdaten, die über sehr lange Zeiträume von Relevanz bleiben, stellt sich mit Nachdruck das Problem der digitalen Langzeitarchivierung. Digitale Langzeitarchivierung hat dabei drei Aspekte (vgl. Liegmann et al. 2010):

- Bitstream Preservation
- Erhalt der Funktionalität
- Erhalt der Benutzbarkeit

Es ist offenkundig, dass die digitale Langzeitarchivierung nicht von einzelnen Forschenden oder Forschungsprojekten sichergestellt werden kann; vielmehr braucht es hierfür technisch und organisatorisch elaborierte, nachhaltige bzw. dauerhafte Infrastrukturen, die diese übergeordnete Aufgabe wahrnehmen. Dennoch können Forschende unmittelbar zur Archivierungsfähigkeit der von ihnen generierten Forschungsdaten beitragen, z. B. indem sie offene, nicht-proprietäre Datenformate verwenden bzw. ihre Daten nach Möglichkeit in solche konvertieren. Dies erleichtert die langfristige Erhaltung der Funktionalität der Daten maßgeblich, da solche Formate zum einen mit gewisser Wahrscheinlichkeit langfristig unterstützt werden und zum anderen Betreiber*innen von Archivierungsdiensten mit höherer Wahrscheinlichkeit in der Lage sein werden, die Daten in neue Formate zu migrieren, wenn die vorliegenden Formate nicht länger unterstützt werden. Durch adäquate Dokumentation und die Vergabe reichhaltiger Metadaten tragen Forschende zudem dazu bei, dass die Daten nicht nur funktional, sondern dass sie auch benutzbar bleiben, weil sie nur so verstanden, adäquat interpretiert und kontextualisiert werden können.

7. Fazit

Forschungsdatenmanagement ist ein genuiner Bestandteil des Forschungsprozesses. In den digitalen Geisteswissenschaften, die (wie andere Forschungsfelder) nur mit gemäß den FAIR-Prinzipien organisierten, qualitativ hochwertigen Forschungsdaten ihr volles Potential entfalten können, kommt dem Datenmanagement daher eine zentrale Bedeutung zu, die Alma Gold treffend zusammenfasst:

> „[…] data is the currency of science, even if publications are still the currency of tenure. To be able to exchange data, communicate it, mine it, reuse it, and review it is essential to scientific productivity, collaboration, and to discovery itself.“ (Gold 2007)

FDM ist dabei wesentlich eine Aktivität der Forschenden selbst, ein immanenter Bestandteil des Forschungsprozesses (vgl. auch Lemaire 2018, 245). Es gibt jedoch breite und vielfältige Serviceangebote, die Forschende dabei durch Beratung sowie die Bereitstellung erforderlicher Infrastrukturen und Tools unterstützen. Hierzu zählen u. a. die bereits im Abschnitt zu Repositorien erwähnten Konsortien der NFDI, der CLARIAH-Verbund und der Verbund DHd, aber auch institutionelle Servicestellen zum FDM sind zentrale Ansprechpartner.

Ein konsequent auf die Umsetzung der FAIR Data Principles ausgelegtes FDM verbessert die Qualität der Forschung und ihrer Ergebnisse und ist für den reibungslosen Ablauf eines Forschungsprojekts sowie für auf das aktuelle Projekt aufsetzende Anschlussforschung unerlässlich. Zugespitzt formuliert: Es gibt keine digitale Forschung ohne FDM. Wer mit digitalen Daten arbeitet, geht mit diesen um. Dies kann besser oder schlechter, effizienter oder ineffizient, FAIRer oder weniger FAIR erfolgen und in ebendiesem Sinne wird dann besseres oder schlechteres, effizienteres oder ineffizienteres, FAIRes oder weniger FAIRes Forschungsdatenmanagement betrieben, aber nie kein Forschungsdatenmanagement.

Literaturverzeichnis

Brettschneider, P., Axtmann, A., Böker, E., & Suchodoletz, D. v. (2021). Offene Lizenzen für Forschungsdaten, *o-bib. Das offene Bibliotheksjournal,* 8(3), 1–22. https://doi.org/10.5282/O-BIB/5749 [zuletzt aufgerufen am 19.06.2024].

Dierkes, J. (2021). Planung, Beschreibung und Dokumentation von Forschungsdaten. In M. Putnings, H. Neuroth, & J. Neumann (Hrsg), *Praxishandbuch Forschungsdatenmanagement* (S. 303–326). Berlin/Boston: De Gruyter Saur. https://doi.org/10.1515/9783110657807-018 [zuletzt aufgerufen am 19.06.2024].

Enke, H., & Ludwig, J. (Hrsg.). (2013). *Leitfaden zum Forschungsdaten-Management.* Boizenburg: Verlag Werner Hülsbusch. URL: https://www.forschungsdaten.org/images/b/bo/Leitfaden_Data-Management-WissGrid.pdf. [zuletzt aufgerufen am 19.06.2024].

Geiger, J. D. (2023). Daten/Forschungsdaten. In AG Digital Humanities Theorie des Verbandes Digital Humanities im deutschsprachigen Raum e. V. (Hrsg.), *Begriffe der Digital Humanities. Ein diskursives Glossar.* Wolfenbüttel: Herzog August Bibliothek [= *Zeitschrift für digitale Geisteswissenschaften. Working Papers,* 2]. https://doi.org/10.17175/WP_2023_003 [zuletzt aufgerufen am 19.06.2024].

Gold, A. (2007). Cyberinfrastructure, Data, and Libraries, 1, *D-Lib Magazine* 23(1/2), o. S. https://doi.org/10.1045/september20september-gold-pt1 [zuletzt aufgerufen am 19.06.2024].

Kailus, A. (2023). Handreichung für ein FAIRes Management kulturwissenschaftlicher Forschungsdaten. V. 1.0.3. URL: https://nfdi4culture.de/go/E3625 [zuletzt aufgerufen am 19.06.2024].

Krogh, P. (2009). *The DAM Book. Digital Asset Managament for Photographers.* 2. Aufl. Sebastopol: O'Reilly Media.

Lemaire, M. (2018). Vereinbarkeit von Forschungsprozess und Datenmanagement in den Geisteswissenschaften. Forschungsdatenmanagement nüchtern betrachtet, *o-bib. Das offene Bibliotheksjournal,* 5(4), 237–247. https://doi.org/10.5282/O-BIB/2018H4S237-247 [zuletzt aufgerufen am 19.06.2024].

Liegmann, H., & Neuroth, H. (2010). Einleitung. In H. Neuroth, A. Oßwald, R. Scheffel, S. Strathmann, & K. Huth. (Hrsg.), *nestor Handbuch. Eine kleine Enzyklopädie der digitalen Langzeitarchivierung.* Version 2(3) (S. 1–10). Boizenburg: Verlag Werner Hülsbusch. URL: https://nbn-resolving.de/urn:nbn:de:0008-2010071949 [zuletzt aufgerufen am 19.06.2024].

Michener, W. K. (2015). Ten Simple Rules for Creating a Good Data Management Plan, *PLOS Computational Biology,* 11(10), 1–9. https://doi.org/10.1371/journal.pcbi.1004525 [zuletzt aufgerufen am 19.06.2024].

Pempe, W. (2012). Geisteswissenschaften. In N. Heike, S. Strathmann, A. Oßwald, R. Scheffel, J. Klump, & J. Ludwig (Hrsg.), *Langzeitarchivierung von Forschungsdaten. Eine Bestandsaufnahme* (S. 137–160). Boizenburg: Verlag Werner Hülsbusch.

The Turing Way Community. (2022). The Turing Way. A handbook for reproducible, ethical and collaborative research. Online: *Zenodo.* https://doi.org/10.5281/ZENODO.7625728 [zuletzt aufgerufen am 19.06.2024].

Whyte, A., & Tedds, J. (2011). *Making the Case for Research Data Management.* In *DCC Briefing Papers.* Edinburgh: Digital Curation Centre. URL: https://www.dcc.ac.uk/guidance/briefing-papers/making-case-rdm [zuletzt aufgerufen am 19.06.2024].

Wilkinson, M. D., Dumontier, M., Aalbersberg, I. J., Appleton, G., Axton, M., Baak, A., Blomberg, N., Boiten, J.-W., da Silva Santos, L. B., Bourne, Ph. E., Bouwman, J., Brookes, A. J., Clark, T., Crosas, M., Dillo, I., Dumon, O., Edmunds, S., Evelo, Ch. T., Finkers, R., Gonzalez-Beltran, A., Gray, A. J. G., Groth, P., Goble, C., Grethe, J. S., Heringa, J., C't Hoen, P. A., Hooft, R., Kuhn, T., Kok, R., Kok, J., Lusher, S. J., Martone, M. E., Mons, A., Packer, A. L., Persson, B., Rocca-Serra, Ph., Roos, M., van Schaik, R., Sansone, S.-A., Schultes, E., Sengstag, Th., Slater, T., Strawn, G., Swertz, M. A., Thompson, M., van der Lei, J., van Mulligen, E., Velterop, J., Waagmeester, A., Wittenburg, P., Wolstencroft, K., Zhao, J., & Mons, B (2016). The FAIR Guiding Principles for Scientific Data Management and Stewardship, *Scientific Data,* 3(1), 1–9. https://doi.org/10.1038/sdata.2016.18 [zuletzt aufgerufen am 19.06.2024].