RDMG
Research Data Management Group

Langzeitarchivierung von Daten aus Forschungsprojekten

Aktualisierte Version 1.0 vom 19. Mai 2021

Susanne Mocken, Jan Leendertse, Dirk von Suchodoletz

Dieses Paper adressiert Aspekte des Forschungsdatenmanagements (FDM) in Drittmittelprojekten an der Universität Freiburg. Es nimmt dabei Bezug auf Vorarbeiten und Präsentationen der Research Data Management Group (RDMG).[1] Das Rechenzentrum und die Universitätsbibliothek hat bereits seit längerer Zeit mit vorbereitenden Arbeiten im Zusammenhang mit bwSFS (Storage-for-Science) und dem Datenmanagement für die Universität begonnen.[2] Die Basis-Speicherinfrastruktur ist inzwischen allgemein verfügbar, die darauf aufsetzenden FDM-Dienste werden schrittweise angepasst und getestet. Die vorliegenden Ausführungen dienen parallel als Input für die mit den Fakultäten vorgesehenen Gespräche zur Umsetzung lokaler FDM-Policies und zur eventuellen zukünftigen Bearbeitung Fragen des FDM in verschiedenen Phasen eines Projekts. Eine etwaige Übergabe von Daten an das Archiv der Universität Freiburg wird hier nicht thematisiert, nur allgemein auf die Vorgabe des Archivgesetzes, Akten inklusive Daten vor Vernichtung dem zuständigen Archiv vorzulegen.

Ziele der Archivierung

Das Primärziel besteht in einer DFG-konformen Ablage von Forschungsdaten für geforderte Zeiträume. Es soll beim Begriff „Forschungsdaten“ von der Definition ausgegangen werden, die im vom Rektorat verabschiedeten Dokument „Grundsätze zum Umgang mit Forschungsdaten an der Albert-Ludwigs-Universität Freiburg“ zugrundegelegt wird.[3] In ihnen wird an Forschungsdaten der Anspruch formuliert, Forschungsdaten „FAIR“ zu machen.[4] Es wird davon ausgegangen, dass eine Speicherung nach den Grundsätzen der DFG ebenso die Bedingungen weiterer Fördergeber erfüllt. Zu nennen sind die Förderprogrammen Horizon2020 (bald Horizont Europa) sowie das BMBF. Es ist jeweils zu diskutieren, ob eine Daten-Publikation und damit öffentliche Verfügbarkeit der Daten oder die Ablage in einem Dark Archive (Ablage ohne öffentliche Referenz) anzustreben ist. Diese Entscheidung bestimmt zu einem guten Teil das Lizenzmodell, das für Forschungsdaten gewählt wird.

Eine frühzeitige Planung für die zukünftigen Daten erleichtert den späteren Umgang. Deshalb ist es sinnvoll, erste Überlegungen, Anforderungen und Workflows anhand konkreter Daten in der

---

[1] https://rdmg.uni-freiburg.de.

[2] Vgl. hierzu beispielsweise das Forum der RDMG in Weiterbildungs-ILIAS, zu erreichen über https://www.rz.uni-freiburg.de/go/rdmg.

[3] Vgl. http://www.uni-freiburg.de/forschung/uni-freiburg-grundsaetze-forschungsdaten-senat.pdf

[4] 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 (1): 160018. https://doi.org/10.1038/sdata.2016.18.

Praxis zu testen. Dabei sollten schrittweise die verschiedenen Aspekte des Forschungsdatenmanagements (FDM) angesprochen werden.

Rahmen

Für die Übernahme von Daten sind die Bedingungen konstituierend, die in den Förderanträgen fixiert wurden. Weitere Festlegungen ergeben sich durch die Grundsätze der Universität Freiburg, die zu einer Policy weiterentwickelt werden, auf der eine Governance für die Mitglieder der Universität beruht. Elemente dieser Governance sind weitere Übereinkünfte auf Fakultätsebene oder innerhalb wissenschaftlicher Communities, die als „Code of Conduct“ (CoC) oder in anderer Form festgehalten sind. Datenmanagementpläne (DMP) als Konkretisierung von CoC setzen den Rahmen für einzelne Forschungsvorhaben.

Der Rahmen, in dem FDM betrieben wird, ist nicht ausschließlich vorgegeben. Er sollte von den Forschenden, die im Zentrum des Forschungsprozesses stehen, wesentlich gestaltet werden. Dokumente wie ein CoC oder Vorschläge zu einer Governance sollten die Ansprüche und Bedarfe konkretisieren, die von der Universität Freiburg oder externen Repositories zu decken sind. Am Ende dieses Dokuments finden sich Hilfen in Form von Leitfragen, solche Bedarfe zu formulieren. Zur Konkretisierung gibt es einen Erfassungsbogen, der auch zur Dokumentation dient.

Beteiligte

An der Übernahme von Daten sind verschiedene Einrichtungen beziehungsweise Forschungsgruppen und Personen beteiligt. Auf der Seite der Datenlieferanten werden verschiedene Funktionen zu erfüllen sein. Es wird eine Person benötigt, die verbindliche Absprachen zum Fördergeber, zur Leitung der Universität und zu Serviceeinheiten wie dem Rechenzentrum treffen kann. Für die Teilprojekte, deren Daten potenziell zu Übernahme anstehen, werden Personen benötigt, die im jeweiligen Kontext über FDM inhaltlich entscheiden können. Perspektivisch wird eine Person benötigt, die in den Kontexten der am Forschungsverbund oder -vorhaben beteiligten Fachbereichen Hilfestellungen zu technischen und fachbezogenen Metadaten und zur Organisation von Daten geben kann. Durch die Mitarbeit der Research Data Management Group (RDMG) und den Aufbau von dedizierten Diensten auf bwSFS wird in 2021 weitere Expertise für mögliche Use-Cases bereit stehen.

Planung des Datenmanagements

Es ist davon auszugehen, dass ungeachtet aller möglichen Überlegungen zu FDM bereits Daten aus und um die Forschung auf IT-Systemen gespeichert werden. Dies ist für Forschungsaktivitäten auf individueller Basis als auch für Gruppen oder Projekte anzunehmen. Es ist zu überprüfen, ob die gewählten Organisationsprinzipien den Anforderungen genügen, die durch die Grundsätze der Universität Freiburg oder die Leitlinien der DFG formuliert sind.

Die RDMG bereitet Materialien vor, die in einem Selfassessment ermitteln, inwieweit die Datenorganisation die genannten Anforderungen erfüllen und wo Maßnahmen ergriffen werden sollten, um einer Übereinstimmung mit ihnen näher zu kommen.

Mit einer konformen Dateiorganisation lassen sich Umsetzungen zu folgenden Fragen zum eigentlichen Management von Forschungsdaten planen:

- Überlegung zur Speicherart (konzeptionell: Data Publication vs. Dark Archive; technisch: Dateisystem, Object-Storage oder Bandlaufwerk (TSM))
- Kuratierung der Daten (inklusive Überprüfung auf sensible Daten) durch die jeweils involvierten Forschenden
- Art der Quellsysteme (Geteilte Speichersysteme wie NextCloud, gemeinsame Ordner in lokalen Netzwerken, Dienstrechner, persönliche Laptops, ...)
- Anreicherung mit notwendigen Metadaten
- Eventuelle Qualitätskontrolle der Metadaten und Kuratierung durch die Projektleitungen
- Paketierung von Daten mit inhaltlicher Zuordnung zu einzelnen Forschungsvorhaben
- Ausfüllen eines Übernahmeprotokolls, Überlassungsvertrag, falls eine Archivierung andernorts

Aus diesen Schritten ergeben sich längerfristige Überlegungen, die in Vorbereitung auf neu entstehende Daten begonnen werden sollten. Generell sind dabei unter anderem folgende Fragestellungen zu erörtern:

- Disambiguierte Zuordnung von Daten und Verantwortlichkeiten (vgl. hierzu "Leitfaden – Verantwortungsvoller Umgang mit Forschungsdaten")[5]
- Zuständigkeiten während Archiv- und Publikationsphase, Regelung des Zugriffs und der Durchsuchbarkeit (Archiv- bzw. FDM-Beauftragter)
- Kostenumlage, falls der Umfang der Daten über die Grundversorgung in bwSFS hinausgeht (Betriebs- und Geschäftsmodell bzw. "Leitfaden – Verantwortungsvoller Umgang mit Forschungsdaten")
- Erwägungen zur Governance: Fachgremium für Daten-Entscheidungen (wegwerfen, aufheben mit Finanzierungszusage, ...)

Im Zuge der Governance-Diskussion wären die Verantwortlichkeiten auf den verschiedenen Ebenen zu benennen. Sie müssen eine jederzeit unmissverständliche Zuordnung innerhalb von Forschungsverbünden und -vorhaben und den beteiligten Fakultäten definieren, um im Zeitverlauf anstehende Fragen zu klären. Dazu gehören das Bestimmen von Laufzeiten, die Sicherstellung der Finanzierung innerhalb der zugesagten Zeiträume, die Zuordnung zu Kostenstellen und die Entscheidung, welche Daten veröffentlicht werden und welche in einem „Dark Archive“ vorzuhalten sind.

Speicherformen

Für die Kuratierung und Paketierung der Daten sowie das Versehen mit Metadaten kann im Rechenzentrum Speicherplatz reserviert werden.[6] Vor der eigentlichen Übernahme sind die Daten in einen Kuratierungsordner zu kopieren, so dass die Teilprojektdaten im Quellordner gelöscht werden können. Für laufende Projekte steht zu diesem Zweck Platz auf dem jetzigen Speichersystem oder bei Bedarf auch dem bwSFS zur Verfügung. Die notwendigen Schritte zum Anlegen und Konfigurieren dieses Speicherbereichs fallen in den Arbeitsbereich des Rechenzentrums.

---

[5] https://www.forschungsdaten.info/typo3temp/secure_downloads/104400/0/cda7fa0aa53a45b87c0f97d34c3c96ab7b1e7346/Leitfaden_-_Verantwortungsvoller_Umgang_mit_Forschungsdaten.pdf.

[6] Für den Überblick zu allgemeinen Speicheroptionen siehe: https://wb-ilias.uni-freiburg.de/goto.php?target=frm_130918_18896&client_id=unifreiburgwb

Für den Fall, dass es regelmäßig Anfragen von (externen) Forschenden bezüglich von Altdaten in abgelaufenen Projekten gibt, wäre es sinnvoll, die Archivdaten bis zur Verfügbarkeit des bwSFS-Archivbereichs auf dem Interims-Speicherplatz zwischenzulagern. Die Altdaten stünden dann potenziell auch für Experimente bereit, bei denen die Durchsuchbarkeit und Nachnutzbarkeit von Forschungsdaten durchgespielt werden kann.

Eine weitere Variante ist Archivierung (oder zusätzlichen Sicherung) auf Bandlaufwerken (TSM). Auch hier kann das Rechenzentrum vorbereitende Arbeiten übernehmen, wenn Fragen z.B. nach entsprechenden Funktionsaccounts für die Zuordnung dieser Sicherungen geklärt sind. Das Speichern auf TSM ist für Daten sinnvoll, die als „Dark Archive“ aufgehoben werden sollen, ohne eine permanente schnelle Verfügbarkeit garantieren zu müssen.

Kuratierung der Daten

Für die endgültige Ablage der Daten mit evtl. Option der Publizierung sollten diese vorher geeignet kuratiert werden. Hierzu zählt beispielsweise eine einfache Durchsicht nach verwaisten Dateien sowie dem Entfernen von Redundanzen und temporären Dateien. Dieser Schritt lässt sich halb automatisieren, erfordert jedoch zusätzlich eine inhaltliche bzw. fachliche Beurteilung. Ebenso sollte überprüft werden, ob potenziell sensible und/oder persönliche Daten enthalten sind. Diese Aufgaben sollten sinnvollerweise von einer Person mit entsprechendem fachlichen Hintergrund übernommen werden. Das könnte auch das Identifizieren eines „Data Stewards“ bedeuten und sollte auf der Ebene der Fakultäten in entsprechende langfristige Governance-Strukturen für Entscheidungen über archivierte Daten münden.

Die hier gemachten Überlegungen können als Input zu künftigen Workflows inklusive der Qualitätssicherung in Forschungsvorhaben dienen. Hierzu ist eine gemeinsame Diskussion in den zuständigen Gremien zu führen.

Annotierung mit Metadaten

Bis zur endgültigen Einführung des bwSFS und eines darauf basierenden Archivsystems ist es nötig, die zu archivierenden Daten auf einem bereits existierenden Speichersystem zwischenzulagern. Um die Migration der Daten in das zukünftige bwSFS-Archiv vorzubereiten, müssen dabei auch alle erforderlichen deskriptiven und administrativen Metadaten erfasst werden. Dieses sollte durch eine Person „vor Ort“ mit fachlicher Expertise bzw. Nähe zu den (ehemaligen) Teilprojekten durchgeführt werden.

Aus der Archivierung ergeben sich eine Reihe von Metadaten, die den Wert der eigentlichen Daten für eine längerfristige Nachnutzung erhöhen, z. B.:

- Projektname
- Projekt-ID / Förder-ID
- Projektbeschreibung
- Forschungsförderer (hier DFG)
- Förderprogramm
- Principal Investigator (PI)
- ORCID-ID der/des PI
- Kontaktinformationen des Datenverantwortlichen (Telefon und E-Mail)
- Erstellungsdatum und Vorhaltezeit

Innerhalb von Gruppen mit einem gemeinsamen CoC können Basisdaten niedergeschrieben oder ein Dokumentenkopf geschrieben werden, der Grundlage für Datenmanagementpläne ist, die Forschungsvorhaben beschreiben. Die Fachdiskussion auf Landesebene, die sich im AK FDM bündelt, geht in Richtung ORCID-ID als Empfehlung für Forschende. Der AK hat eine entsprechende Empfehlung im Oktober 2020 veröffentlicht.[7] Es gibt weitere Metadaten, die man in Zukunft sinnvollerweise für Daten gemäß den Gepflogenheiten der zugehörigen Scientific-Community speichern sollte. Deshalb sollte mit einer Überlegung zu einer Metadaten-Strategie baldmöglichst begonnen werden. Schlagworte für diese Diskussion sind: DataCite, Basis-Metadaten, technische, wissenschaftliche, Publikations-, Metadaten.

Paketierung der Daten

Bei einer großen Anzahl von Datensätzen wird es notwendig, diese nach bestimmten Kriterien zu strukturieren. Eine solche Zusammenstellung kann dann mit gemeinsamen Basis-Metadaten versehen und abgelegt werden. Die gewählte Granularität sollte auch dazu geeignet sein, eine Nachnutzung zu erleichtern, also insbesondere das Wiederfinden und Herausnehmen einzelner (logischer) Datensätze möglich zu machen. Für eine solche Strukturierung sollte ein Verpacken der Daten nach den Standards der eigenen Community erfolgen. Als allgemeines Container-Format für die Archivierung bieten sich dabei insbesondere BagIt oder OCFL an.

Aus den Erfahrungen und der Analyse von Workflows lassen sich Empfehlungen für eine zukünftige Datenablage ableiten. Dies beinhaltet auch die Strukturierung der Ablage nach bestimmten Kriterien, wie beispielsweise Rohdaten, abgeleitete Daten, erklärende Dokumente, Schlüsseltabellen etc. sowie zu publizierende Daten. Diese Empfehlungen ebenso wie potenziell sinnvolle und geeignete Werkzeuge hierzu sollten im Forschungsverbund diskutiert werden. Die RDMG kann als Hilfestellung einen Überblick der einschlägigen Fachliteratur geben.

Formale Datenübernahme

Bei der Übernahme von Daten müssen verantwortliche Personen festgelegt werden, sowohl auf der Seite der Daten Abgebenden als auch auf der Empfängerseite. Ebenso ist mittelfristig ein FDM-Verantwortlicher seitens der Fakultät bzw. des Instituts zu benennen ("Leitfaden – Verantwortungsvoller Umgang mit Forschungsdaten"). In diesem Zusammenhang sollte der (Lese-)Zugriff geregelt werden, also welche Person(engruppe) berechtigt bzw. verpflichtet ist, eventuelle Nachfragen zu bedienen oder die Daten einzusehen.

Informationen zu bwSFS

Öffentliche, renommierte und nachhaltige Community-Services sind die erste Wahl vor Nutzung eines lokalen Systems für den Nachweis von Forschungsdaten. Sollte es die jedoch in der Form nicht geben, stehen hierfür zukünftig Uni-Ressourcen (in gewissem Umfang) zur Verfügung. Derzeit wird am Rechenzentrum das zentrales Speichersystem bwSFS (Storage-for-Science, mehrfach oben genannt) für die Universität und ausgewählte wissenschaftliche Communities (aus dem HPC-Umfeld) aufgebaut und schrittweise in Betrieb genommen. Nach aktuellem Stand der Diskussionen und Evaluationen in der RDMG werden die folgenden Punkte in der nächsten Zeit angegangen und umgesetzt:

FDM-Dienste im engeren/weiteren Zusammenhang bwSFS (Storage-for-Science):

[7] Vgl. Bausteine FDM: https://doi.org/10.17192/bfdm.2020.2.8272

- Repository für Publikationen aus der Universität auf Basis von InvenioRDM
- Archivystem / Dark Archive (no publication) als “Archival Storage” (die konkrete Softwareplatform wird bestimmt).
- Universitäts- bzw. landesweite Kollaborationsplattform auf der Basis von Git. Sie ermöglicht Daten- und Code-Versionierung.
- Speicherinfrastruktur (Filesystem, Object-Storage, für NEMO-Community), seit März 2021 nutzbar.
- Generelles Data-Managementsystem, welches in verschiedenen Diensten eingebunden werden kann. Beispiele sind HPC, Cloud. Es ist für Compute und Workflows zur Datenanalyse nutzbar (Integration in BW-Datenföderation).
- Spezieller FDM-Dienst OMERO. Er wird als Bild-Datenbank angeboten, beispielsweise für Microskopie etc.
- Für FDM-Services, die aus Communities heraus entwickelt werden und bwSWS als Backbonedienst verwenden.
- Nutzung als Backend für ausgewählte Projekte (MWK-Projekt SDC BioDATEN, NFDI DataPLANT)

Teildienste hiervon können in Zukunft von den verschiedenen laufenden Projekten genutzt werden. Auf dieses System kann in zukünftigen Anträgen Bezug genommen werden.

Leitfragen für die Formulierung von Bedarfen

Die „Grundsätze zum Umgang mit Forschungsdaten an der Albert-Ludwigs-Universität Freiburg“ sind mit Stand April 2021 noch nicht sehr konkret, wie die Universität die Forschenden bei der Umsetzung der im gleichen Papier genannten Verpflichtungen zu unterstützen beabsichtigt. Folgende Fragen sollen den Einstieg in eine Diskussion innerhalb von Fakultäten, Verbünden oder Projekten erleichtern. Im Ergebnis soll ein universitätsweiter Diskurs initiiert werden, in dem die Interessen der Beteiligten zur Geltung kommen und der in eine konsensorientierte Governance mündet:

- Welche Vorgaben gibt es bereits bei Ihnen zum Umgang mit Forschungsdaten? Solche Vorgaben kommen beispielsweise durch Drittmittelgeber wie die DFG, das Horizon2020-Programm (bald Horizont Europa) oder das BMBF.
- Verfügt Ihre Einrichtung über einen Code of Conduct, der sich am Kodex der DFG oder ähnlichen Dokumente anlehnt?
- Wie werden Promovierende oder Studierende darüber unterrichtet, wie sie mit Forschungsdaten umzugehen haben?
- Gibt es in Ihrer Einrichtung Vorgaben oder Abläufe, wenn bei der Forschung personenbezogene Daten im Sinne der DSGVO verarbeitet werden? Gibt es ein Verfahren, bei dem der Datenschutzbeauftragte oder die Ethikkommission integriert sind?
- Gibt es Überlegungen, wie Nutzungsrechte und Lizenzen geregelt werden?
- Findet in Ihrer Einrichtung oder in Ihrem Fachbereich ein Diskurs statt, welche Standards für Metadaten geeignet sind oder wie diese weiterentwickelt werden sollten?
- Berücksichtigt Ihre Einrichtung Metriken, mit denen die Qualität von Forschungsdaten überprüft werden kann? Ein Beispiel für eine solche Metrik ist der „Kerndatensatz Forschung“ (→ https://www.kerndatensatz-forschung.de/), der vom Wissenschaftsrat empfohlen wird. (→ https://www.wissenschaftsrat.de/download/archiv/5066-16.pdf)
- Gibt es in Ihrer Einrichtung einen Datenmanager, eine Datenmanagerin, die Forschende unterstützt? Diese Funktion ist eine andere als die eines IT-Administrators oder einer IT-Administratorin.

Quellenverweise

Bagit: https://tools.ietf.org/html/rfc8493

OCFL: https://ocfl.io

Datacite: https://schema.datacite.org

ORCID.iD: https://orcid.org

Empfehlungen zu ORCID/ROR:https://bausteine-fdm.de/article/view/8272/8125

Arbeitskreis FDM in BaWü: https://www.forschungsdaten.info/fdm-im-deutschsprachigen-raum/baden-wuerttemberg/arbeitskreis-forschungsdatenmanagement/