Handreichungen für das Forschungsdatenmanagement: Planung und Durchführung von Forschungsvorhaben

Planungsphase

Datenmanagementplan

Zur Planung von FDM kann ein Datenmanagementplan (DMP) aufgesetzt werden. Ein DMP ist ein Dokument, das den gesamten Lebenszyklus von Forschungsdaten beschreibt. Er wird i. d. R. von Forschenden erstellt, um den Umgang mit den Daten vor, während und nach einem Forschungsprojekt zu planen und zu organisieren. Der Hauptzweck eines DMP besteht darin, sicherzustellen, dass Forschungsdaten effizient, nachvollziehbar, langfristig zugänglich und verständlich verwaltet werden.

Datenmanagementpläne sind besonders wichtig, um die Reproduzierbarkeit von Forschung zu fördern, die Effizienz der Datenverwaltung zu steigern und sicherzustellen, dass wertvolle Forschungsdaten langfristig erhalten bleiben. Oftmals verlangen Fördergeber oder Institutionen von Forschenden, einen solchen Plan vorzulegen, bevor ein Projekt bewilligt wird.

Zu beachten sind hier:

- Die Vorgaben der Drittmittelförderer
- Finanzierung des Forschungsdatenmanagements
- rechtliche und ethische Fragegestellungen
- Publikation von wissenschaftlichen Erzeugnissen

Technische Tools zur Unterstützung der Datenmanagementplanung:

RDMO steht für "Research Data Management Organizer" und ist eine Open-Source-Softwareplattform, die dazu dient, Datenmanagementpläne (DMPs) zu erstellen, zu verwalten und zu teilen. RDMO bietet Forschenden, Projektteams und Institutionen die Möglichkeit, effektive Datenmanagementpläne zu erstellen und umzusetzen, um den gesamten Lebenszyklus von Forschungsdaten zu verwalten. Für Hochschulen im Land Brandenburg steht eine eigene RDMO Instanz unter dem Namen RDMO-BB zur Verfügung.[1]

Durchführungsphase

(Daten-) Dokumentation

Die Datendokumentation ist fester Bestandteil wissenschaftlicher Forschung und wird während und nach Abschluss eines Vorhabens angewendet. Primäres Ziel ist, strukturierte Informationen über das Wie und Warum der Datenerhebung aufgearbeitet und digitalisiert zu erfassen. Gute Dokumentation ist entscheidend für ein

[1] Link zu RDMO-BB: https://rdmo.fdm-bb.de/ (Stand 19.11.2024)

kurzfristiges und langfristiges Verständnis, sowie eine erfolgreiche längerfristige Datenhaltung und ermöglicht weiterhin Fehlererkennung und -behebung.

Datendokumentation findet auf mehreren Ebenen statt.

- Auf der ersten Ebene werden grundlegende Informationen über eine Studie gesammelt, um den Kontext zu Methodik und Datensammlung herzustellen. Diese Informationen können in Form eines DMPs zusammengefasst werden.
- Die zweite Ebene umfasst weiterführende Beschreibungen über individuelle Datenordner innerhalb einer Datensammlung und bietet auf einen Blick detaillierte Informationen über einzelne Daten (z. B. erhobene Merkmale, Variablen), wie sie in ELN (electronic lab notebooks) erfasst werden.
- Metadaten, d. h. standardisierte, strukturierte Daten über die Sammlung, die häufig zur Katalogisierung und zum Auffinden der Objekte in der Sammlung verwendet werden. Ein weit verbreiteter Metadatenstandart ist beispielsweise das DataCite-Metadatenschema.

ReadMe-Datei:

Zur einfach umsetzbaren Dokumentation von Forschungsdaten eignen sich ReadMe-Dateien. Diese enthalten alle relevanten Informationen, die für das Verstehen der Forschungsdaten notwendig und Voraussetzung für eine Nachnutzung sind. Hierzu zählen u. a.:

- Titel des Forschungsdatensatzes
- Projektname und -beschreibung
- Identifier
- Ort der Datenerhebung
- Version
- Ersteller*in des Datensatzes
- Beschreibung der Forschungsdaten (z. B. Inhalt, verwendete Methode zur Datenerhebung, Entstehungskontext)
- Beschreibung der Methode(n) zur Datenerhebung und -verarbeitung
- Dateiliste
- Dateiformate
- Schlagwörter
- Angaben zu rechtlichen Ansprüchen an die Forschungsdaten

ReadMe-Vorlagen:

- Ostdata: https://zenodo.org/records/6956989
- TU Braunschweig: https://www.tu-braunschweig.de/fileadmin/Redaktionsgruppen/Einrichtungen/UB/README_Template_TUBS.txt (Link via: https://www.tu-braunschweig.de/forschung/forschungsdaten-transparenz/forschungsdaten/fdm-services/informationsmaterialien-links)

- auch Repositorien stellen z. T. eigene Vorlagen zur Verfügung

Datenqualität

Die Qualitätskontrolle der Daten ist ein integraler Bestandteil jeder Forschung und findet in verschiedenen Phasen statt: bei der Datenerhebung, der Dateneingabe oder Digitalisierung und der Datenprüfung. Es ist von entscheidender Bedeutung, klare Rollen und Verantwortlichkeiten für die Datenqualitätssicherung in allen Phasen der Forschung zuzuweisen und professionelle Verfahren zu entwickeln, bevor die Datenerfassung beginnt.

Datenerfassung

Bei der Datenerhebung müssen die Forschenden sicherstellen, dass die erfassten Daten die tatsächlichen Fakten, Antworten, Überlegungen oder Ereignisse widerspiegeln. Die Qualität der verwendeten Datenerhebungsmethoden hat einen großen Einfluss auf die Datenqualität, und die detaillierte Dokumentation der Datenerhebung ist ein Beweis für diese Qualität.

Zu den Qualitätskontrollmaßnahmen während der Datenerhebung können gehören:

- Kalibrierung von Instrumenten zur Überprüfung der Genauigkeit, Verzerrung und Umfang der Messung
- Durchführung mehrerer Messungen, Beobachtungen oder Probenahmen
- Überprüfung der Datenerhebung durch eine weitere Person
- Verwendung von standardisierten Methoden und Protokollen für die Datenerhebung sowie von Erfassungsvorgaben mit klaren Anweisungen
- Computergestützte Befragungssoftware zur Standardisierung von Befragungen, zur Überprüfung der Konsistenz von Antworten, zur Weiterleitung und Anpassung von Fragen, so dass nur geeignete Fragen gestellt werden, zur Bestätigung von Antworten anhand früherer Antworten und zur Erkennung unzulässiger Antworten

Dateneingabe und Datenerfassung

Wenn Daten in eine Datenbank oder ein Tabellenkalkulationsprogramm eingegeben, kodiert, digitalisiert oder transkribiert werden, wird die Qualität sichergestellt und Fehler werden vermieden, indem beispielsweise standardisierte und einheitliche Verfahren mit klaren Anweisungen verwendet werden:

- Einrichtung von Validierungsregeln oder Eingabemasken in Dateneingabesoftware
- Verwendung von Kontrollvokabularen, Codelisten und Auswahllisten, aus denen Werte ausgewählt werden müssen, um die manuelle Dateneingabe zu minimieren - es gibt international vereinbarte Konventionen für die Aufzeichnung von Informationen wie die ISO 8601, das empfohlene Format für die Darstellung von Daten und Zeiten
- detaillierte Kennzeichnung von Variablen und Datensatznamen, um Verwechslungen zu vermeiden
- Entwicklung einer zweckmäßigen Datenbankstruktur zur Organisation von Daten und Datenfeldern

Datenkontrolle

Bei der Datenkontrolle werden die Daten bearbeitet, bereinigt, überprüft, abgeglichen und validiert. Die Prüfung umfasst in der Regel sowohl automatisierte als auch manuelle Verfahren, wie z. B.:

- doppelte Überprüfung der Kodierung von Beobachtungen oder Antworten und von Werten außerhalb des veranschlagten Messbereichs
- Überprüfung der Vollständigkeit der Daten
- Überprüfung von Stichproben der digitalen Daten anhand der Originaldaten
- doppelte Eingabe von Daten
- statistische Analysen wie Häufigkeiten, Mittelwerte, Spannen oder Clustering, um Fehler und anomale Werte zu erkennen
- Korrekturlesen der Transkription
- Peer-Review

weiterführende Literatur:

- Cai, Li and Zhu, Yangyong (2015): "The Challenges of Data Quality and Data Quality Assessment in the Big Data Era", Data Science Journal, Volume 14, pp. 1-10, DOI: http://dx.doi.org/10.5334/dsj-2015-002
- Sidi, Fatimah; Shariat Panahy, Payam Hassany; Affendey, Lilly Suriani; Jabar, Marzanah A.; Ibrahim, Hamidah & Mustapha, Aida (2012): "Data quality: A survey of data quality dimensions," *2012 International Conference on Information Retrieval & Knowledge Management*, Kuala Lumpur, Malaysia, pp. 300-304, DOI: https://doi.org/10.1109/InfRKM.2012.6204995

Speicherung und Verarbeitung

Die sichere Speicherung von Forschungsdaten zählt zu den essentiellen Aufgaben im Forschungsdatenmanagement. Hierzu zählen passende Speicherdienste und -medien sowie ein vorausschauendes Speicher- und Backup-Konzept, um Datenverlust wirkungsvoll vorbeugen zu können. Auch die Sicherheit der Daten, insbesondere bei Daten, die sensible Informationen erhalten, muss durch geeignete Maßnahmen gewährleistet werden. Für kooperatives Arbeiten sind darüber hinaus besondere Bedarfe zu berücksichtigen, etwa das Festlegen individueller Zugriffs- und Nutzungsrechte.

Die Grundlage für einen produktiven Umgang mit Forschungsdaten bildet die **Datenorganisation**. Diese beinhaltet die folgenden Aufgaben:

- nachvollziehbare Ordnerstruktur für Daten anlegen
- Schema für einheitliche und aussagekräftige Dateibenennung
- Dateiversionierung: Versionierungsschema verwenden / automatische Dateiversionierung
- Verwendung offener und standardisierter Dateiformate

Speicherorte und Backup

Grundsätzlich ist von der *alleinigen* Nutzung lokaler Speichermedien oder kommerzieller Speicherdienste abzuraten. Das Speichern auf institutionellen Speichermedien bietet hingegen ein höheres Maß an Sicherheit, denn regelmäßige Backups und Wartung sind hier sichergestellt.

Backup-Maßnahmen sollten in regelmäßigen Abständen und gemäß einem festgelegten Zeitplan durchgeführt werden. Folgende Regel sollte hierbei Beachtung finden: Die bewährte 3-2-1-Backup-Regel sieht vor, dass drei Kopien auf mindestens zwei verschiedenen Speichermedien gespeichert werden, wovon eine Kopie dezentral, z. B. in einer sicheren Cloud, abgelegt wird.

Cloud-Dienste wie Nextcloud sowie entsprechende Softwareprogramme zählen zu den gängigen Backup-Tools.

Speicher- und Backup-Strategie sind vorab/zu Projektbeginn festzulegen, zu dokumentieren und ggf. innerhalb der Projektgruppe abzustimmen.

Datensicherheit und Schutzniveau

Zur sicheren Speicherung von Daten zählt auch, die Sicherheit der Daten zu gewährleisten. Insbesondere dann, wenn diese sensible Informationen enthalten. Stehen die Daten in einem Zusammenhang mit der DSGVO, sollte ein Zugriff auf die Daten von außen nicht möglich sein (Erhebungsphase). Um Daten vor unbefugtem Zugriff zu schützen sind folgende Maßnahmen zu ergreifen:

- Zugangsbeschränkungen (Passwortschutz, Verschlüsselung etc.)
- Pseudonymisierung bzw. Anonymisierung von personenbezogenen Daten

Datenformate

Beschreiben Sie den Datentyp (z. B. Rastergeodaten, Fragebogen, Interviews) und das Dateiformat (TIFF, CSV-Tabelle, ODT/DOC). Wenn möglich, verwenden Sie offene Dateiformate und vermeiden Sie proprietäre Dateiformate. Begründen Sie die Verwendung proprietärer Software für die Datenerstellung (weit verbreitete Verwendung und Akzeptanz, Kompetenzen der Forschenden, vorhandene erworbene Lizenz).

Die Wahl von Dateiformaten hat direkte Auswirkungen auf die Lesbarkeit, Austauschbarkeit und die Archivfähigkeit. Vor allem in Bezug auf letzteres setzen Sie sich am besten mit der Institution (Bibliothek) oder dem Repositorium in Verbindung, bei der/dem die Daten archiviert werden sollen, da es durchaus verschiedene Vorgaben geben kann. Der Unterschied zwischen empfohlenen und akzeptierten Formaten variiert teilweise stark.

Eine Übersicht über die gängigsten Dateitypen und die für Sie empfohlenen Formate gibt folgende Tabelle:

Dateityp	**Empfohlenes Format**
Text	TXT, PDF/A (Typen 2a, 2u und 2b)
„Office Dokumente“	PDF/A
Tabellen	CSV
Rasterbilder	TIFF, JPEG2000
Audio	WAVE

Video	MPEG-4
Strukturierte (Text-)Daten	XML
Geodaten	TIFF + EWF, XML, INTERLIS

Weitere Informationen:

- KOST (https://kost-ceco.ch/cms/kad_recommendation_de.html)
- ETH Zürich (https://unlimited.ethz.ch/display/DD/Archivtaugliche+Dateiformate)
- Verbund Forschungsdaten Bildung (https://www.forschungsdaten-bildung.de/dateiformate)
- Library of Congress (https://www.loc.gov/preservation/resources/rfs/RFS%202023-2024.pdf).

Versionierung (Software)

Während der Projektdurchführung unterliegen Datensätze meist einer ständigen (Weiter-)Entwicklung (z. B. während Auswahl, Aggregation, Integration). Es ist von Vorteil, mit Versionierungen zu arbeiten, d. h. die verschiedenen Versionen zu kennzeichnen, zu dokumentieren und während der Projektlaufzeit aufzubewahren. Insbesondere bei textbasierten Daten erleichtert die Verwendung von Versionierungstools, wie z. B. Git oder SVN, die Handhabung der verschiedenen Versionen.

Ansprechpersonen

Beratung zum Forschungsdatenmanagement und Datenmanagentplänen

Blanka Goßner

Forschungsdatenmanagement | Zentrum für Forschung und Transfer

Technische Hochschule Wildau | Hochschulring 1 | 15745 Wildau

Haus 13 | Raum 0.41

Telefon: +49 3375 508 322

Mail: blanka.gossner@th-wildau,de

Beratung bei der Beantragung drittmittelgeförderter Forschungsprojekte

Zentrum für Forschung und Transfer

Mail: forschung@th-wildau.de