UNIVERSITÄT DES SAARLANDES

Forschungsdatenmanagement
Grundlagen und Empfehlungen

Version 0.1 | Stabsstelle DN | 30.01.2023

Einleitung

Der Umgang mit Forschungsdaten gehört zu den alltäglichen Aufgaben in der Forschung. Hierbei gilt es jedoch, einige Aspekte und Regularien zu befolgen, damit diese Forschungsdaten auch einen langfristigen und nachhaltigen Nutzen für die Wissenschaft haben. Ein Datenmanagementplan (DMP) ist ein wichtiges Instrument, das den Umgang mit Forschungsdaten im Laufe des Datenlebenszyklus strukturiert und beschreibt. Viele Drittmittelgeber stellen bereits Anforderungen an DMPs für die Fördermittelvergabe. Selbst wenn dies nicht der Fall sein sollte, ist die Erstellung eines DMP für die Arbeit an einem Forschungsvorhaben gewinnbringend. Insbesondere im Hinblick auf technologischen Fortschritt und damit einhergehend die stetig wachsende Fülle an Forschungsdaten ist ein DMP ein unerlässliches Mittel, um die FAIR-Prinzipien umzusetzen. Die UdS stellt eine Policy im Umgang mit Forschungsdaten zur Verfügung unter FDM_policy_UdS_v02.docx.

Diese FAIR-Prinzipien sind 2016 formulierte generische Regeln im Umgang mit Forschungsdaten, die eine verbesserte Auffindbarkeit (Findable), Zugänglichkeit (Accessible), Interoperabilität (Interoperable) sowie Nachnutzbarkeit (Reusable) zum Ziel haben[1].

1 Dateiformate

1.1 Empfohlene Dateiformate

Die erhobenen Daten sollten nach Möglichkeit in Formaten gespeichert werden, die eine langfristige Lesbarkeit und Zugriff gewährleisten. Diese Dateiformate weisen oft einen offenen, gut dokumentierten Quellcode auf und sind nicht-proprietär. Ebenso sollte darauf geachtet werden, dass die Datenformate auf allen (gängigen) Computersystemen bearbeitet werden können. Forschungsdaten sollten überdies unkomprimiert vorliegen. Als Hilfestellung werden im Folgenden einige Dateiformate genannt:

- Bilddateien:
    - Vektorgrafiken (Skizzen, Zeichnungen, Piktogramme) zeichnen sich durch eine unbegrenzte Skalierungsfähigkeit aus (.svg, *scalable vector graphics*), was sich bei großen Bildgrößen in einer signifikant kleineren Dateigröße widerspiegelt. Jedoch können diese Vektorgrafiken nicht von allen gängigen Textverarbeitungsprogrammen fehlerfrei verarbeitet werden.

[1] https://www.go-fair.org/fair-principles/, abgerufen am 02.12.2022.

- Rastergrafiken (Fotos, Mikroskopie) können im Gegensatz zu Vektorgrafiken nicht beliebig skaliert werden. Bei hoher Bildauflösung entstehen somit große Dateigrößen, welche oftmals in einem nicht verlustfreien Dateiformat (z. B. .jpg) gespeichert werden. Stattdessen sollte darauf geachtet werden, die Bilddaten verlustfrei zu komprimieren, was beispielsweise durch die Formate .tiff (tagged image file format), .bmp (bitmap) oder .png (portable network graphics) sichergestellt ist.

- Videodateien:
  - Videodateien sollten ebenso vorzugsweise in einem plattformübergreifenden Format aufgenommen werden. Entscheidend hierbei ist, dass die verwendeten Codecs nicht-proprietär sind, wie bspw. .h264 oder dessen Nachfolger .h265. Als standardisiertes Containerformat ist .mp4 unter der Norm ISO 14496-14 eingetragen.
- Textdateien:
  - Als offene Dokumentenstandards sind die unter ISO 26300 spezifizierten offenen Formate von OpenOffice (.odt, .ods, .odp) zu empfehlen; darüber hinaus bietet sich mit dem Rich Text Format (.rtf) sowie dem .docx ein gut dokumentierter, weit verbreiteter Standard an, der jedoch proprietär ist. ASCII-Text kann in .txt-Formaten gespeichert werden oder als Hypertext (.html, .xml), .csv für Komma-separierte Variablen, oder .tex für LaTeX-Dateiformate.
  - Für nicht editierbaren Text bietet sich das weit verbreitete .pdf (ISO 32000) an.

Diese Liste behandelt lediglich ein paar ausgewählte Szenarien. Eine umfassendere Übersicht empfohlener Dateiformate für Forschungsdaten findet sich bspw. bei der Max-Planck-Gesellschaft [2].

1.2 Konvertierung

Sollten die von Ihnen generierten Forschungsdaten in einem Format vorliegen, welches von den oben genannten Empfehlungen abweicht (z. B. proprietäres Format) und Sie die Daten konvertieren möchten um offene Standards zu gewährleisten, bewahren Sie immer eine Kopie der Daten im Ursprungsformat auf. Kann sichergestellt werden, dass sämtliche Daten inklusive aller Metadaten korrekt konvertiert wurden, kann von dieser Regel abgewichen werden.

Es gilt zu beachten, dass bei manchen der oben angegebenen Dateiformaten nur dann eine verlustfreie Speicherung erfolgt, wenn die ursprünglichen Rohdaten ebenfalls in diesem Format generiert wurden, wie bspw. im Falle von .bmp oder auch .mp3.

[2] https://rdm.mpdl.mpg.de/before-research/file-formats/, abgerufen am 02.12.2022.

2 Erstellung eines Datenmanagementplans

Bedingt durch die Heterogenität der unterschiedlichen Forschungsdisziplinen gibt es keine einheitliche Struktur eines DMPs. Es lassen sich jedoch Kernelemente definieren, die ebenfalls die Empfehlungen vieler Drittmittelgeber widerspiegeln, wie beispielsweise die der Deutschen Forschungsgemeinschaft (DFG).

So enthält ein DMP strukturierte Informationen über das Forschungsprojekt zu jedem Zeitpunkt. Gewisse Informationen sind hierbei statisch, wie bspw. Angaben zu Projektverantwortlichkeiten, andere Informationen sind dynamischer Natur, wie bspw. die Zugriffsrechte im Falle von personellen Änderungen im Projekt.

2.1 Elemente eines Datenmanagementplans

Die allgemeinen Bestandteile eines Datenmanagementplans sind im Folgenden aufgezählt:

1. Administrative Informationen zum Forschungsvorhaben: Ziele, Partner, Projektverantwortliche:r und Laufzeit des Forschungsvorhabens
2. Datengenese
    - Auf welche Weise entstehen in Ihrem Projekt neue Daten?
    - Werden existierende Daten wiederverwendet und wie werden diese migriert?
    - Welche Datentypen, im Sinne von Datenformaten (z. B. Bilddaten, Textdaten oder Messdaten) entstehen in Ihrem Projekt und auf welche Weise werden sie weiterverarbeitet?
    - In welchem Umfang fallen diese an bzw. welches Datenvolumen ist zu erwarten?
    - Welche Hard- und Software wird voraussichtlich verwendet um die Daten zu verarbeiten?
    - Wie ist die Datenstruktur organisiert?
    - Wie werden Metadaten generiert?
3. Datensicherheit und Workflow
    - Wo werden die Daten während des Datenlebenszyklus gespeichert?
    - Werden die Daten zusätzlich gesichert (Backup)?
    - In welchen Intervallen erfolgen die Backups?
    - Werden die Dateispeichersysteme auf Fehlerfreiheit überprüft?
    - Gibt es Datensätze, die besonders sensible Informationen beinhalten?
    - Wie wird der Zugriff auf die Daten verwaltet?
    - Wie werden die Daten benannt (Schema)?
    - Wie werden die Daten versioniert?

- Wie können die Daten zwischen Projektpartnern synchronisiert werden?

4. Ingestplan und Archiving
    - Angaben zur Datenauswahl: Welche Primärdaten werden weiterverarbeitet?
    - Wie erfolgt die Datenübergabe/der Datenaustausch?
    - Wie werden die Daten validiert?
    - Wie werden die Daten archiviert?
    - Informationen zu Metadaten
5. Konsolidierung
    - Welche Eigentums- und Urheberrechte gelten für die Daten?
    - Welche Nutzungsrechte sind vorhergesehen?
    - Sind alle Vorgaben der Fördermittelgeber berücksichtigt?
    - Welche Backup-Strategie wird verfolgt?
    - Sind alle datenschutzrelevanten Punkte berücksichtigt?

2.2 Electronic Lab Notebook/Elektronisches Laborjournal (ELN)

Gemäß den Grundsätzen guter wissenschaftlicher Praxis sind Forschende dazu verpflichtet, Notizen, experimentelle Verfahrensweisen, Protokolle sowie Parameter und Metadaten bei der Durchführung und Planung von Experimenten zu dokumentieren. Im Gegensatz zu traditionellen Laborjournalen bieten ELNs eine Reihe von Vorteilen. So können bei elektronischen Laborjournalen die Einträge nach Schlagwörtern durchsucht werden, aber auch die sichere Speicherung sowie die Verfügbarmachung und Teilen mit Kollaborationspartnern und Mitgliedern der Arbeitsgruppe. Oftmals bieten diese Dienste auch eine Kalenderfunktion für gemeinsam verwendete Laborgeräte sowie die Möglichkeit, diverse Geräte zu verwalten und zu koppeln (laboratory information management system, LIMS).

Ein browserbasiertes OpenSource-Tool zur elektronischen Verwaltung von diesen Labornotizen, welches weltweit an vielen Forschungsinstituten zum Einsatz kommt, ist eLabFTW[3].

Einige der hierin enthaltenen Möglichkeiten für Forschende sind:

- Zeitstempel für durchgeführte Experimente (RFC 3161-konform)
- Importieren von und Exportieren in PDF, ZIP, CSV, JSON
- Umfangreiche Vorlagenbibliothek für Experimente
- Gliederung der Experimente und Protokolle in Einzelschritte möglich
- REST API für den programmatischen Zugriff auf Daten mit einem externen Programm

[3] https://www.elabftw.net/, abgerufen am 02.12.2022.

- Hochladen von Dateien und Einbettung von Fotos in Text unterstützt
- Zeichnen von Molekülen möglich (Chemdoodle)
- Große Entwickler- und Anwendergemeinschaft weltweit
- Einrichtung einer To-do-Liste möglich
- Eingebauter Terminplanungsassistent
- Plattformübergreifend (Mac, Linux, Windows)
- Weltweiter Zugriff möglich, auch mit mobilen Endgeräten
- Speicherung und Verwaltung von Reagenzien, Protokollen, Zelllinien, usw. (LIMS)

Die UdS prüft die Einführung eines zentralen Angebots für ein Laborbuch, wenn es dafür genügend Bedarfsmeldungen gibt.

3 Speicheroptionen

Im Folgenden werden einige Möglichkeiten genannt, wo Forschungsdaten, die innerhalb der Arbeitsgruppe der UdS anfallen, gespeichert werden können.

3.1 M365 – Cloud-Speicher

Die UdS bietet mit dem universitätsweiten Einsatz von Microsoft365 eine cloudbasierte Lösung an um Daten abzulegen. Diese Möglichkeit der Datenablage bietet weiterhin den Vorteil, die Daten mit (externen) Kollaborationspartnern einfach zu teilen und zu bearbeiten. Eine Handreichung zur Einrichtung von OneDrive bzw. SharePoint als Dateiablage kann unter dem Link Handreichung SharePoint-OneDrive.pdf abgerufen werden.

3.2 hizCloud

Dies ist ein universitätsinterner Sync&Share-Dienst auf Basis von Nextcloud, den das Hochschul-IT-Zentrum des Saarlandes (HIZ) anbietet[4]. Der Dienst kann zum Austausch und zur Synchronisation von Dateien auf verschiedenen Endgeräten benutzt werden.

Die hizCloud eignet sich für den Einzelnutzer wie auch für Anwendungen einer Arbeitsgruppe. Hierbei steht jedem Mitarbeiter ein persönliches Speicherkontingent von 50 GB zur Verfügung. Darüber hinaus gibt es Gruppenordner, mit der die Mitglieder einer Arbeitsgruppe die gemeinsamen Dokumente verwalten und bearbeiten können.

[4] https://www.hiz-saarland.de/dienste/hizcloud, abgerufen am 02.12.2022.

Je nach gewähltem Speicherkontingent des Gruppenordners fallen hierbei jedoch jährliche Gebühren an, im Gegensatz zu den Angeboten in M365.

3.3 Universitätsinterner Datenspeicher für Projekte mit >50 TB

Die UdS hat beabsichtigt einen eigenen Forschungsdatenspeicher einzurichten, der für Arbeitsgruppen mit sehr großen Datenmengen zur Verfügung stehen soll. Bitte nehmen Sie dazu mit der Stabsstelle DN Kontakt auf.

3.4 Weitere Tools

Die vorgenannten Möglichkeiten zur Datenablage und -verwaltung bieten viele Vorteile bei der Speicherung von Primärdaten, auf die im Laufe des Projektvorhabens häufig zugegriffen wird bzw. die um Sekundärdaten erweitert werden. Sollen die Daten jedoch archiviert und/ oder publiziert werden, gibt es Repositorien, die ebenfalls die Option aufweisen, einen persistenten Identifikator, wie bspw. eine DOI, zu vergeben. Ein weit verbreitetes Repository ist Zenodo[5], welches vom CERN verwaltet wird und eine Vorreiterrolle für Open Science darstellt. Eine Übersicht aller Repositorien für Forschungsdaten diverser Disziplinen wird auf der Internetpräsenz www.re3data.org gelistet.

4 Metadatenverwaltung

Forschungsdaten unterliegen einem komplexen Datenlebenszyklus. Innerhalb dieses Zyklus müssen entsprechend den Regeln guter wissenschaftlicher Praxis jedoch zu jedem Zeitpunkt gewisse Metadaten ersichtlich sein, um eine eindeutige Zuordnung der Daten zu Personen und Projekten möglich ist. Die NFDI hat zum Ziel fachspezifische Schemata für Metadaten zu entwickeln. Ein allgemeiner Ansatz findet sich z.B. bei dem globalen Konsortium Datacite, dass es sich zum Ziel gemacht hat, Forschungsergebnisse nachhaltig und transparent zu gestalten durch einen einfachen Zugang zu den Forschungsdaten. Als Basis dient ein standardisiertes Metadatenschema mit notwendigen Angaben (**m**andatory), empfohlenen Angaben (**r**ecommended) und **o**ptionalen Angaben. Eine weitere Standardisierung bietet das Dublin Core-Metadatenschema (ISO 15836) mit der Einführung von 15 Kernelementen (*core elements*) zur Deklaration elektronischer Ressourcen. Basierend auf diesen Standards empfehlen wir die Verwendung eines Metadatenschemas wie folgt:

[5] http://www.zenodo.org/, abgerufen am 02.12.2022.

ID #	Name/Eigenschaft	Beschreibung
1	Identifier	Eindeutige Identifizierung eines Objekts durch adäquate Zuordnung zu einem Typenkatalog (bspw. DOI, ISBN, ISSN)
2	Creator	Der Name des verantwortlichen Verfassers oder des Urhebers des Dokuments in seiner aktuellen Fassung. Statt einer natürlichen Person kann dies auch eine Organisation sein.
3	Title	Angabe des Titels des beschriebenen Dokuments.
4	Publisher	Name der Einrichtung die die Daten veröffentlicht, vertreibt, freigibt oder archiviert.
5	PublicationYear	Angabe des Datums, wann die Daten publiziert wurden bzw. voraussichtlich publiziert werden, sofern absehbar.
6	Subject	Angabe von (mehreren) suchtauglichen Schlagwörtern oder keywords, die das Thema des Inhalts widerspiegeln.
7	Contributor	Angabe weiterer an der Dokumentenentstehung beteiligter Personen, typischerweise Kollaborationspartner und Co-Autoren.
8	Date	Charakteristisches Datum im Datenlebenszyklus. Dies kann sowohl das Start- bzw. Enddatum der Datengenerierung sein, aber auch Zeitspannen abdecken. Die Notation sollte der Norm entsprechend (ISO 8601) in der Form YYYY-MM-DD erfolgen.
9	Language	Angabe der Sprache(n) der Ressource, falls zutreffend.
10	ResourceType	Gibt an, um welche Art von Ressourcen es sich handelt, bspw. audiovisuelle Daten, Bildmaterial, Video- oder Textdokumente.
16	Rights	Information zur Rechte am geistigen Eigentum, die an den Daten gehalten werden. Je nach Datentyp bieten sich hierfür bspw. die Creative Commons 'CC-BY' an oder GNU Affero Public License (Quellcode).
17	Description	Kurzzusammenfassung des Inhalts des Dokuments in freitextlicher Form *(abstract)*. Zusätzlich Angaben, wie die Daten generiert

		wurden mitsamt der Angabe aller verwendeten Ressourcen (Hard-/Software) sowie des etwaigen Versuchsprotokolls.
19	FundingReference	Angabe zu Fördereinrichtungen und Behörden, die die betreffende Forschung finanziert haben.

Die Nomenklatur der ID und der zugehörigen Eigenschaft (Deskriptor) in obiger Tabelle orientiert sich hierbei an dem DataCite-Metadatenschema[6], wo auch weitere Definitionen und Handlungsempfehlungen zu finden sind. Wir weisen darauf hin, dass die Angabe obiger Metadateneinträge keinesfalls erschöpfend ist und lediglich einen Minimalkonsens darstellt zur Zuordnung und Verwaltung der betreffenden Forschungsdaten. Ebenso ist die Reihenfolge, in der einzelne Metadatenfelder benannt werden können, vom jeweiligen Zeitpunkt im Datenlebenszyklus abhängig und somit ist auch die Gesamtheit der Metadaten als dynamisch anzusehen.

Die erhobenen Metadaten können mittels diversen Softwaretools in einer Eingabemaske erfasst und dann im betreffenden Dateiordner gemeinsam mit den eigentlichen Forschungsdaten gespeichert werden[7].

5 Verwaltungsempfehlungen zu Forschungsdaten

Um einen nachhaltigen Umgang von generierten Forschungsdaten zu gewährleisten empfiehlt es sich, in jeder Arbeitsgruppe eine einheitliche Ablagestruktur zu implementieren, die gut dokumentiert ist. Auf diese Weise ist sichergestellt, dass einer Auffindbarkeit und Wiederverwendbarkeit der Daten keine Hindernisse im Weg stehen und auch für Dritte klar ersichtlich ist, wie die Datensätze abgelegt sind.

5.1 Benennung von Datensätzen

Bei der Benennung von Forschungsdaten gibt es gewisse Grundsätze und -prinzipien, die einen nachhaltigen Umgang gewährleisten. Im Folgenden werden einige Benennungsempfehlungen gegeben:

[6] https://schema.datacite.org/meta/kernel-4.4/doc/DataCite-MetadataKernel_v4.4.pdf, abgerufen am 02.12.2022.

[7] https://dhvlab.gwi.uni-muenchen.de/datacite-generator/, abgerufen am 02.12.2022.

- Verwenden Sie bei der Benennung immer eine Versionsnummer oder das Datum im Format „YYYY-MM-DD“, an dem die Daten generiert wurden.
- Achten Sie hierbei auf die alphanumerische Sortierung und ggf. führende Nullen bei mehreren ähnlich benannten Dateien oder Ordnern im Rahmen von z.B. einer systematischen Messreihe.
- Wählen Sie Ihr Benennungsschema möglich so, dass Informationen zu den Daten im Dateinamen enthalten sind.
- Verzichten Sie bei der Benennung auf Sonderzeichen und Leerzeichen. Nutzen Sie stattdessen als Worttrenner einen Bindestrich „-“, oder einen Unterstrich „_“ oder Majuskeln.
    - **Ungeeignete** Benennungsbeispiele:
        - ‚Test-v1.txt‘
        - ‚Messung Datum.txt‘
        - ‚Ölfilm_v=5mms^-1.bmp‘
    - **Geeignete** Benennungsbeispiele:
        - ‚Oelfilm_2022-07-29.bmp‘
        - ‚RoteBlutzelle_u1_0mm-s_v01.csv‘
        - ‚GlycerolViscosity_eta_20mPas_2021-04-21.csv‘

5.2 Ordnerstrukturen

Im Forschungsbetrieb gibt es eine heterogene Mischung von anfallenden Daten gemäß ihrer Bedeutung für die Forschungsprojekte. Diese Bedeutung ist bei der Genese der Primärdaten im Allgemeinen noch nicht ausreichend abzuschätzen und erst im weiteren Verlauf der Analyse dieser Primärdaten einzuordnen. Auf diesen Überlegungen aufbauen unterscheiden wir die folgenden Empfehlungen zur Strukturierung der Forschungsdaten nach

1. Primärdaten (Rohdaten, Messergebnisse, Zwischenergebnisse, Simulationsdaten, Beobachtungen, etc., inkl. Metadaten)
2. Sekundärdaten (Analysierte Primärdaten, die von Wissenschaftler:innen der jeweiligen Disziplin interpretiert werden können)

Bei der Ablage der Primärdaten sollte ein sich schnell erschließendes Ablageschema und Nomenklatur vorliegen, sodass unter Zuhilfenahme der ebenfalls abgespeicherten Metadaten die Forschungsergebnisse reproduzierbar sind und die Datengenese nachvollzogen werden kann. Da diese

Datenstrukturen ja nach Fachdisziplin sehr unterschiedlich ausfallen kann keine über diese Tatsachen hinausgehende Empfehlung gegeben werden.

Bei den Sekundärdaten hingegen lässt sich eine genauere Ablagestruktur definieren, basierend auf dem Datenlebenszyklus der sekundären Forschungsdaten von deren Genese bis zur finalen Publikation. Als Grundgerüst empfehlen wir eine Verzeichnisstruktur gemäß folgendem Schema:

Verzeichnisname	Beschreibung
„00_Publication“	○ In diesem Verzeichnis sollte die finale Version der Forschungsarbeit gespeichert werden, sowohl als Quelltext als auch als .pdf. ○ Angabe der Metadaten gemäß Dublin Core in ASCII-Form. ○ Ebenso bei Veröffentlichung die Angabe der DOI. ○ In Unterverzeichnissen können zusätzliche Dokumente zu „Supplementary Materials“ abgelegt werden.
„01_Manuscripts“	○ In diesem Verzeichnis sollten alle vorherigen Versionen der Publikation gespeichert werden, ebenso als Quelltext und .pdf
„02_Figures“	○ Ablage aller für die Publikation erforderlichen Bilder und Grafiken mit Angabe der Nomenklaturregeln sofern nicht selbsterklärend. Zusätzlich die Angabe der zugrundeliegenden Datenpunkte der Grafiken, die aus den Primärdaten extrahiert wurden. Gegebenenfalls Unterordner für jede Version des Manuskripts erstellen. Die Grafiken sollten in ausreichend hoher Auflösung und vorzugsweise mindestens als .pdf vorliegen, sofern es sich nicht um Rastergrafiken wie z.B. bei Fotos handelt.
„03_PrimaryData“	○ Ablage der in die Publikation eingeflossenen Primärdaten inklusive Metadaten. ○ Insofern die Nutzung eines ELN gegeben ist, ein Auszug aller betreffenden Einträge.
„04_Programs“	○ Angabe aller verwendeten Softwaretools und Quellcodes inkl. Parametern für die jeweiligen Simulationen oder Berechnungen.
„05_Bibliography“	○ Ablage aller verwendeten Quellen als ASCII-Text, z.B. mittels Zotero oder BibTex. Nach Möglichkeit sollten auch die beschriebenen Quellen selbst abgelegt werden, um bei nicht frei

	zugänglichen Daten (bspw. Master- oder Bachelorarbeiten) einen Zugriff zu gewährleisten.
„06_Miscellaneous“	○ Ergänzende Dateien (bspw. Gutachten zur Publikation, E-Mail-Verkehr, Videodateien).