Muster-Datenmanagementplan (DMP)

Team Forschungsdatenmanagement
Universitätsbibliothek, Freie Universität Berlin
https://www.fu-berlin.de/forschungsdatenmanagement

Hinweis: Die Kapitelüberschriften und Inhalte orientieren sich am Fragenkatalog der Checkliste zum Umgang mit Forschungsdaten[1] der DFG. Alle grau hinterlegten Textpassagen dienen nur als Hilfestellung und sollten nicht Teil eines DMPs sein.

Administrative Informationen

Änderungshistorie:

Version	Datum	Änderungen
1.0	01.04.2022	Erster Entwurf

Projekttitel: Muster

Förderkennzeichen: 2022-1234-5678

Projektbeschreibung: Im Projekt „Muster" werden Forschungsdaten erhoben, um mittels <METHODE> zu Schlussfolgerungen zu gelangen.

Principal Investigator:
Maxi Mustermann
Freie Universität Berlin
Institut für Musterwissenschaft
+49 (0)30 1234-5678
maxi.mustermann@fu-berlin.de
https://orcid.org/0001-0002-0003-0004

Beteiligte Forschende und/oder Einrichtungen: Maximilian Mustermann, John Doe, Jane Doe

Forschungsförderer: Deutsche Forschungsgemeinschaft

Förderprogramm: Musterprogramm 2022

Relevante Policies:

- Deutsche Forschungsgemeinschaft. 2019. Leitlinien zur Sicherung guter wissenschaftlicher Praxis (Kodex). doi:10.5281/zenodo.3923602.
- Deutsche Forschungsgemeinschaft. 2015. Leitlinien zum Umgang mit Forschungsdaten. https://www.dfg.de/download/pdf/foerderung/antragstellung/forschungsdaten/richtlinien_forschungsdaten.pdf (zugegriffen: 01. April 2022).
- Freie Universität Berlin. 2021. Forschungsdaten-Policy der Freien Universität Berlin. doi:10.17169/refubium-30560.
- Freie Universität Berlin. 2021. Open-Access-Policy der Freien Universität Berlin. doi:10.17169/refubium-30559.

[1] Deutsche Forschungsgemeinschaft. 2021. Checkliste zum Umgang mit Forschungsdaten (Version: 21.12.2021). https://www.dfg.de/forschungsdaten/checkliste (zugegriffen: 01. April 2022).

1. Datenbeschreibung

- Auf welche Weise entstehen in Ihrem Projekt neue Daten
- Werden existierende Daten wiederverwendet?
- Welche Datentypen, im Sinne von Datenformaten (z. B. Bilddaten, Textdaten oder Messdaten) entstehen in Ihrem Projekt und auf welche Weise werden sie weiterverarbeitet [mit welcher Software]?
- In welchem Umfang fallen diese an bzw. welches Datenvolumen ist zu erwarten?

Für das Vorhaben sind nach Recherchen in gängigen Datenrepositorien keine aktuellen bzw. geeigneten Forschungsdaten zur Nachnutzung verfügbar. Die im Projekt erzeugten Daten werden weitere Erkenntnisse auf dem <GEBIET> ermöglichen. Der erzeugte Datensatz wird durch das Projektteam mit der <METHODE> erstellt.

Hauptsächlich fallen textuelle und tabellarische Daten an. Diese werden nach Möglichkeit in offenen Formaten[2] gespeichert, siehe Aufstellung:

Datentyp	**Datenformat**
Textuelle Daten	txt, rtf, pdf, md
Tabellarische Daten	csv
Skripte für Analysen und Auswertungen	py, ipynb

Während der Projektlaufzeit werden Analysen und Auswertungen mit der frei verfügbaren Programmiersprache Python[3] sowie deren etablierten Open-Source-Bibliotheken (u.a. Pandas[4]) durchgeführt. Außerdem werden demonstrative „Executable Papers“ in Form von Jupyter Notebooks[5] erzeugt.

Das erwartete Datenvolumen wird 50 GB nicht überschreiten.

2. Dokumentation und Datenqualität

- Welche Ansätze werden verfolgt, um die Daten nachvollziehbar zu beschreiben (z. B. Nutzung vorhandener Metadaten- bzw. Dokumentationsstandards oder Ontologien)?
- Welche Maßnahmen werden getroffen, um eine hohe Qualität der Daten zu gewährleisten?
- Sind Qualitätskontrollen vorgesehen und wenn ja, auf welche Weise?
- Welche digitalen Methoden und Werkzeuge (z. B. Software) sind zur Nutzung der Daten erforderlich?

Die erzeugten Forschungsdaten und Skripte werden im institutionellen Repositorium der Freien Universität Refubium[6] veröffentlicht (vgl. Abschnitt 5).

Im Sinne der FAIR-Prinzipien[7] werden die Daten im Repositorium durch Metadaten beschrieben, orientiert am DataCite-Schema[8] (u.a. Abstract, freie Schlagwörter und DDC-Klassifikation). Darüber

[2] Vgl. Böker, Elisabeth. 2021. Formate erhalten. https://www.forschungsdaten.info/themen/veroeffentlichen-und-archivieren/formate-erhalten (zugegriffen: 01. April 2022).
[3] O A. Welcome to Python.org. https://www.python.org (zugegriffen: 01. April 2022).
[4] O A. pandas | Python Data Analysis Library. https://pandas.pydata.org (zugegriffen: 01. April 2022).
[5] O A. Project Jupyter. https://www.jupyter.org (zugegriffen: 01. April 2022).
[6] O A. Refubium. https://refubium.fu-berlin.de (zugegriffen: 01. April 2022).
[7] O A. FAIR Principles. https://www.go-fair.org/fair-principles (zugegriffen: 01. April 2022).
[8] O A. DataCite Schema. https://schema.datacite.org/ (zugegriffen: 01. April 2022).

hinaus wird den Daten eine Dokumentation in Form einer README-Datei[9] (Markdown-Format) hinzugefügt. Diese umfasst u.a. die durchgeführten Arbeitsschritte zur bestmöglichen Nachnutzbarkeit und Reproduzierbarkeit.

Den Metadaten wird durch das Repositorium ein persistenter Identifier (DOI) hinzugefügt, der den Datensatz eindeutig referenzierbar macht.

Die Daten werden nach Typ und Format in verschiedenen Verzeichnissen getrennt gespeichert (z.B. CSV-Dateien in einem „data/tabular"-Verzeichnis, alle Python-Skripte gesammelt in einem „src" oder „scripts"-Verzeichnis).

Die Benennung von Dateien und Verzeichnissen erfolgt nach einem einheitlichen Schema, z.B. werden Datumsangaben nach ISO 2014 formatiert: [JJJJ]-[MM]-[TT] [10]. Das Schema wird zu Projektbeginn mit allen Projektmitarbeitenden gemeinsam festgelegt.

Die Dokumentation der tabellarischen Daten (CSV) erfolgt durch ein bzw. mehrere sogenannte Tabular Data Packages[11]. In Form einer solchen Spezifikation erfolgt eine Dokumentation der Daten, der einzelnen Spalten (Variablen), deren erlaubten Datentypen und Wertebereichen sowie eine Beziehung von Spalten untereinander (auch über einzelne Dateien hinweg). Durch die formalisierte Datenbeschreibung und -dokumentation wird eine toolbasierte Qualitätskontrolle ermöglicht und regelmäßig durchgeführt (z.B. liegen alle Werte konkreter Spalten in erlaubten Wertebereichen). Die Validierung der Daten (anhand der Tabular Data Packages) erfolgt mit dem Frictionless Framework[12].

Die Python-Skripte werden dem Python Style Guide PEP8 [13] entsprechend formatiert. Hierzu werden verschiedene Tools zur Qualitätskontrolle von Software eingesetzt (z.B. Linter). Zur Sicherung der Nachnutzbarkeit wird der Code dokumentiert und kommentiert, ebenfalls den entsprechenden Standards und Best Practices der Python-Community[14] folgend.

Spezifische Versionen verwendeter Bibliotheken (z.B. Pandas in Version X.Y.Z) werden in einem etablierten Format festgehalten (vrsl. requirements.txt [15] oder Pipfile [16]).

Die Nutzung der Daten und Skripte ist durch Open-Source-Standardtools möglich. Kosten für spezialisierte Software zum Lesen/Bearbeiten/Ausführen der Dateien fallen nicht an.

---

[9] O A. 2021. README. In: *Wikipedia*. https://en.wikipedia.org/w/index.php?title=README&oldid=1049034357 (zugegriffen: 01. April 2022).

[10] Vgl. O A. File naming and folder structure | CESSDA TRAINING. https://www.cessda.eu/Training/Training-Resources/Library/Data-Management-Expert-Guide/2.-Organise-Document/File-naming-and-folder-structure (zugegriffen: 01. April 2022).

[11] O A. Tabular Data Package | Frictionless Standards. https://specs.frictionlessdata.io/tabular-data-package (zugegriffen: 01. April 2022).

[12] O A. Describe, extract, validate, and fransform data in Python | Frictionless Framework. https://framework.frictionlessdata.io/ (zugegriffen: 01. April 2022).

[13] O A. PEP 8 -- Style Guide for Python Code. https://www.python.org/dev/peps/pep-0008 (zugegriffen: 01. April 2022).

[14] O A. Our Community. https://www.python.org/community (zugegriffen: 01. April 2022) und O A. Python Software Foundation. https://www.python.org/psf (zugegriffen: 01. April 2022).

[15] O A. Requirements File Format | pip documentation v21.3.1. https://pip.pypa.io/en/stable/reference/requirements-file-format/#requirements-file-format (zugegriffen: 01. April 2022).

[16] O A. Pipfile. https://github.com/pypa/pipfile (zugegriffen: 01. April 2022).

3. Speicherung und technische Sicherung während des Projektverlaufs

- Auf welche Weise werden die Daten während der Projektlaufzeit gespeichert und gesichert?
- [Welche Backup-Strategie gibt es?]
- Wie wird die Sicherheit sensibler Daten während der Projektlaufzeit gewährleistet (Zugriffs- und Nutzungsverwaltung)?

Das erwartete Datenvolumen von maximal 50 GB wird durch einen wissenschaftlichen Speicherbereich[17] (im Folgenden als „Projektnetzlaufwerk" bezeichnet) des Rechenzentrums der Freien Universität (ZEDAT) bereitgestellt.

Während der Projektlaufzeit werden Daten und Skripte auf dem Projektnetzlaufwerk gespeichert. Das Laufwerk wird von allen Projektmitarbeiter*innen als Netzlaufwerk über das jeweilige Betriebssystem eingebunden.

Das Projektnetzlaufwerk unterliegt einer automatisierten, regelmäßigen, dateibasierten Backup-Routine durch das Rechenzentrum. Die Sicherungen werden vom zentralen Backup-Service (ZEDAT)[18] auf Magnetbänder kopiert und in einem Datentresor vorgehalten.

Im Falle, dass Daten und Skripte lokal auf den Arbeitsrechnern der Projektgruppe erzeugt werden, synchronisieren die Mitarbeitenden diese einmal täglich mit dem Projektnetzlaufwerk, um Datenverlust vorzubeugen. Hierzu werden die für das jeweilige Betriebssystem etablierten Open-Source-Tools eingesetzt, z.B. rsync[19].

Da keine sensiblen Daten erhoben werden, erfolgt keine gesonderte Zugriffs- und Nutzungsverwaltung. Der Zugriff zum Projektnetzlaufwerk wird zentral durch das Rechenzentrum der Freien Universität verwaltet (Zugriff haben nur Mitglieder der Projektgruppe).

Darüber hinaus werden Daten und Skripte mit dem Versionskontrollsystem git[20] versioniert. Eine zentrale Organisation des Projekt-Repositories[21] erfolgt über den Anbieter GitHub.

Ein Zugriff auf die Daten durch Dritte ist während der Projektlaufzeit nicht erforderlich; durch die parallele Verwendung von GitHub ist der Zugriff jedoch über das Projekt-Repository möglich.

4. Rechtliche Verpflichtungen und Rahmenbedingungen

- Welche rechtlichen Besonderheiten bestehen im Zusammenhang mit dem Umgang mit Forschungsdaten in Ihrem Projekt?
- Sind Auswirkungen oder Einschränkungen in Bezug auf die spätere Veröffentlichung bzw. Zugänglichkeit zu erwarten?
- Auf welche Weise werden nutzungs- und urheberrechtliche Aspekte sowie Eigentumsfragen berücksichtigt?
- Existieren wichtige wissenschaftliche Kodizes bzw. fachliche Normen, die Berücksichtigung finden sollten?

[17] O A. Speicherbereiche für wissenschaftliche Daten. https://www.fu-berlin.de/sites/scientific-data-storage (zugegriffen: 01. April 2022).
[18] O A. Backup-Service für Server im FU-Netz. https://www.zedat.fu-berlin.de/Backup (zugegriffen: 01. April 2022).
[19] O A. rsync. https://rsync.samba.org (zugegriffen: 01. April 2022).
[20] O A. Git. https://git-scm.com (zugegriffen: 01. April 2022).
[21] Projekt-Repository. https://github.com/1234/5678

In Bezug auf die Daten liegen keine rechtlichen Besonderheiten vor. Die Nachnutzung von Software anderer Urheber*innen wird im Sinne der guten wissenschaftlichen Praxis gemäß der Software citation principles[22] zitiert.

5. Datenaustausch und dauerhafte Zugänglichkeit der Daten

- Welche Daten bieten sich für die Nachnutzung in anderen Kontexten besonders an?
- Nach welchen Kriterien werden Forschungsdaten ausgewählt, um diese für die Nachnutzung durch andere zur Verfügung zu stellen?
- Planen Sie die Archivierung Ihrer Daten in einer geeigneten Infrastruktur? Falls ja, wie und wo?
- Gibt es Sperrfristen?
- Wann sind die Forschungsdaten für Dritte nutzbar?

Die erhobenen Daten und Skripte bieten sich für die Nachnutzung durch Dritte an. Daher Daten sowie Skripte im institutionellen Repositorium der Freien Universität Refubium veröffentlicht. In der Veröffentlichung inbegriffen sind sämtliche erzeugten Rohdaten und Skripte, sowie finale Versionen von Textdaten und Tabellen. Außerdem wird der Veröffentlichung eine Dokumentation beigefügt (vgl. Abschnitt 2). Zwischenergebnisse von Verarbeitungs- und Analyseschritten, die sich sämtlich aus den bereitgestellten Daten und Skripten erzeugen lassen, sind nicht Teil der Veröffentlichung.

Die Veröffentlichung folgt den Empfehlungen der Open-Access-Policy und Forschungsdaten-Policy der Freien Universität. Die Ergebnisse werden unter einer offenen Lizenz (vrsl. Creative Commons 0[23] oder BY[24]) lizenziert.

Durch die Verwendung des Repositoriums Refubium werden mehrere der in den FAIR-Prinzipien[25] adressierten Punkte sichergestellt[26]. So werden die Metadaten über standardisierte Schnittstellen (OAI-PMH) in übergreifenden Nachweissystemen und Suchmaschinen indexiert (z.B. BASE, DataCite Search, OpenAIRE). Dadurch wird eine erhöhte Sichtbarkeit der Forschungsergebnisse erreicht. Die erstellten Metadaten werden durch die Redaktion des Repositoriums geprüft. Des Weiteren wird ein persistenter Identifier (DOI) vergeben.

Im Sinne der Leitlinien zur Sicherung guter wissenschaftlicher Praxis[27] werden die Daten für mindestens zehn Jahre durch das Repositorium öffentlich, d.h. ohne Zugangsbeschränkung, bereitgestellt. Eine separate Archivierung, unabhängig von der Veröffentlichung, ist nicht vorgesehen. Eine Sperrfrist ist nicht erforderlich. Die Veröffentlichung findet so schnell wie möglich, spätestens jedoch innerhalb der letzten drei Monate der Projektlaufzeit statt.

---

[22] Smith, Arfon M., Daniel S. Katz, Kyle E. Niemeyer, und FORCE11 Software Citation Working Group. 2016. Software citation principles. *PeerJ Computer Science* 2: e86. doi:10.7717/peerj-cs.86.

[23] O A. Creative Commons — CC0 1.0 Universell. https://creativecommons.org/publicdomain/zero/1.0/deed.de (zugegriffen: 01. April 2022).

[24] O A. Creative Commons — Attribution 4.0 International — CC BY 4.0. https://creativecommons.org/licenses/by/4.0/deed.de (zugegriffen: 01. April 2022).

[25] 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, Nr. 1 (Dezember): 160018. doi:10.1038/sdata.2016.18.

[26] O A. Leitlinien Forschungsdaten. https://www.fu-berlin.de/sites/refubium/ueber-uns/leitlinien/leitlinien_fd (zugegriffen: 01. April 2022).

[27] Deutsche Forschungsgemeinschaft. 2019. Leitlinien zur Sicherung guter wissenschaftlicher Praxis (Kodex). doi:10.5281/zenodo.3923602.

6. Verantwortlichkeiten und Ressourcen

- Wer ist verantwortlich für den adäquaten Umgang mit den Forschungsdaten (Beschreibung der Rollen und Verantwortlichkeiten innerhalb des Projekts)?
- Welche Ressourcen (Kosten; Zeit oder anderes) sind erforderlich, um einen adäquaten Umgang mit Forschungsdaten im Projekt umzusetzen?
- Wer ist nach Ende der Laufzeit des Projekts für das Kuratieren der Daten verantwortlich?

Hauptverantwortlich für den Umgang mit den im Projekt erzielten Forschungsdaten ist Maxi Mustermann, PI des Vorhabens. Die Einhaltung und Aktualisierung des DMP wird durch <FUNKTION> Jane Doe sichergestellt.

Eine fortlaufende Dokumentation und Aufbereitung der Daten und Skripte erfolgt bereits während der Projektlaufzeit; ihre Finalisierung erfolgt in den letzten drei Monaten.

Nach Ende der Projektlaufzeit werden sämtliche zur Veröffentlichung vorgesehenen Daten am angegebenen Ort publiziert. Über die Laufzeit hinaus findet keine weitere Kuratierung der Daten statt.

Die für das Forschungsdatenmanagement erforderlichen Ressourcen sind im Projektplan unter X aufgeführt.