Sample Data Management Plan (DMP)

Team Research Data Management
University Library, Freie Universität Berlin
https://www.fu-berlin.de/en/sites/forschungsdatenmanagement

> Note: The chapter headings and contents are based on the list of questions in the DFG's checklist regarding the handling of research data[1]. All text passages with a gray background are for guidance only and should not be part of a DMP.

Administrative information

History of changes:

Version	Date	Changes
1.0	01.04.2022	First draft

Project title: Sample

Project acronym: 2022-1234-5678

Project abstract: In the "sample" project, research data is collected to reach conclusions using <METHOD>.

Principal investigator:

Maxi Mustermann
Freie Universität Berlin
Institute for Sample Science
+49 (0)30 1234-5678
maxi.mustermann@fu-berlin.de
https://orcid.org/0001-0002-0003-0004

Researchers and/or institutions involved: Maximilian Mustermann, John Doe, Jane Doe

Research funding organization: Deutsche Forschungsgemeinschaft

Funding program: Sample program 2022

Relevant policies:

- Deutsche Forschungsgemeinschaft. 2019. Guidelines for Safeguarding Good Research Practice. Code of Conduct. doi:10.5281/zenodo.3923602.
- Deutsche Forschungsgemeinschaft. 2015. Guidelines on the Handling of Research Data. https://www.dfg.de/download/pdf/foerderung/grundlagen_dfg_foerderung/forschungsdaten/guidelines_research_data.pdf (accessed: 01 April 2022).
- Freie Universität Berlin. 2021. Research Data Policy of Freie Universität Berlin. doi:10.17169/refubium-32141.
- Freie Universität Berlin. 2021. Open Access Policy of Freie Universität Berlin. doi:10.17169/refubium-31442.

[1] Deutsche Forschungsgemeinschaft. 2021. Checklist Regarding the Handling of Research Data (Version: 21.12.2021). https://www.dfg.de/research_data/checklist (accessed: 01 April 2022).

1. Data description

- How does your project generate new data?
- Is existing data reused?
- Which data types (in terms of data formats like image data, text data or measurement data) arise in your project and in what way are they further processed [with which software]?
- To what extent do these arise or what is the anticipated data volume?

No current or suitable research data are available for re-use for the project, according to searches of common data repositories. The data generated by the project will enable further insights in the <AREA>. The dataset generated will be created by the project team using the <METHOD>.

Mainly textual and tabular data will be generated. These will be stored in open formats[2] where possible, see listing:

Data type	**Data format**
Textual data	txt, rtf, pdf, md
Tabular data	csv
Scripts for analyses and evaluations	py, ipynb

During the project, analyses and evaluations will be performed using the freely available Python[3] programming language and its established open source libraries (including Pandas[4]). In addition, demonstrative "executable papers" will be generated in the form of Jupyter notebooks[5].

The expected data volume will not exceed 50 GB.

2. Documentation and data quality

- What approaches are being taken to describe the data in a comprehensible manner (such as the use of available metadata, documentation standards or ontologies)?
- What measures are being adopted to ensure high data quality?
- Are quality controls in place and if so, how do they operate?
- Which digital methods and tools (e.g. software) are required to use the data?

The generated research data and scripts are published in the institutional repository of Freie Universität Refubium[6] (see section 5).

In line with the FAIR principles[7], the data in the repository are described by metadata, oriented to the DataCite schema[8] (including abstract, keywords and DDC classification). In addition, documentation is added to the data in the form of a README[9] file (Markdown format). This includes, among other things, the steps performed for the best possible reusability and reproducibility.

[2] Data Formats for Preservation. 2019. https://www.openaire.eu/data-formats-for-preservation/ (accessed: 01 April 2022).
[3] Welcome to Python.org. https://www.python.org (accessed: 01 April 2022).
[4] pandas | Python Data Analysis Library. https://pandas.pydata.org (accessed: 01 April 2022).
[5] Project Jupyter. https://www.jupyter.org (accessed: 01 April 2022).
[6] Refubium. https://refubium.fu-berlin.de (accessed: 01 April 2022).
[7] FAIR Principles. https://www.go-fair.org/fair-principles (accessed: 01 April 2022).
[8] DataCite Schema. https://schema.datacite.org/ (accessed: 01 April 2022).
[9] README. 2021. In: *Wikipedia*. https://en.wikipedia.org/w/index.php?title=README&oldid=1049034357 (accessed: 01 April 2022).

A persistent identifier (DOI) is added to the metadata by the repository, making the dataset uniquely referenceable.

The data is stored separately by type and format in different directories (e.g. CSV files in a “data/tabular” directory, all Python scripts collected in a “src” or "scripts" directory).

The naming of files and directories follows a uniform pattern, e.g. dates are formatted according to ISO 2014: [YYYY]-[MM]-[DD][10]. The scheme is defined together with all project members at the beginning of the project.

The documentation of the tabular data (CSV) is done by one or more Tabular Data Packages[11]. In the form of such a specification, a documentation of the data takes place. The individual columns (variables), their permitted data types and value ranges are explicitly defined. Also there is a description of relations between columns (and across individual files). The formalized data description and documentation enable tool-based quality control, which is performed on a regular basis (e.g. all values of concrete columns are within allowed value ranges). The validation of the data (using the Tabular Data Packages) is done with the Frictionless Framework[12].

The Python scripts are formatted according to the Python Style Guide PEP8[13]. For this purpose, various software quality control tools are used (e.g. linters). To ensure reusability, the code is documented and commented, also following the appropriate standards and best practices of the Python community[14].

Specific versions of libraries used (e.g. pandas in version X.Y.Z) are recorded in an established format (likely requirements.txt[15] or Pipfile[16]).

Using the data and scripts is possible through open source standard tools. There are no costs for specialized software to read/edit/execute the files.

[10] See 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 (accessed: 01 April 2022).

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

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

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

[14] Our Community. https://www.python.org/community (accessed: 01 April 2022) and Python Software Foundation. https://www.python.org/psf (accessed: 01 April 2022).

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

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

3. Project storage and technical archiving

- How is the data to be stored and archived throughout the project duration?
- [What is the backup strategy?]
- What is in place to secure sensitive data throughout the project duration (access and usage rights)?

The expected data volume of a maximum of 50 GB is accommodated by a scientific storage area[17] (hereinafter referred to as "project network drive") of the Computing Services of Freie Universität (ZEDAT).

During the project, data and scripts are stored on the project network drive. The drive is mounted by all project employees as a network drive via the respective operating system.

The project network drive is subject to an automated, regular, file-based backup routine by the data center. The backups are copied to magnetic tapes by the central backup service (ZEDAT)[18] and kept in a data vault.

In the event that data and scripts are generated locally on the project group's workstations, the employees synchronize them once a day with the project network drive to prevent data loss. For this purpose, established open source tools for the respective operating system are used, e.g. rsync[19].

Since no sensitive data is collected, there is no separate access and usage management. Access to the project network drive is managed centrally by the computer center of Freie Universität (access is only granted to members of the project group).

In addition, data and scripts are versioned using the version control system git[20]. A central organization of the project repository[21] is done by the provider GitHub.

Access to the data by third parties is not required during the project duration; however, the parallel use of GitHub allows access via the project repository.

4. Legal obligations and conditions

- What are the legal specifics associated with the handling of research data in your project?
- Do you anticipate any implications or restrictions regarding subsequent publication or accessibility?
- What is in place to consider aspects of use and copyright law as well as ownership issues?
- Are there any significant research codes or professional standards to be taken into account?

There are no legal particularities with regard to the data. The re-use of software of other authors is cited in line with good scientific practice according to the software citation principles[22].

[17] Speicherbereiche für wissenschaftliche Daten. https://www.fu-berlin.de/sites/scientific-data-storage (accessed: 01 April 2022).
[18] Backup-Service für Server im FU-Netz. https://www.zedat.fu-berlin.de/Backup (accessed: 01 April 2022).
[19] rsync. https://rsync.samba.org (accessed: 01 April 2022).
[20] Git. https://git-scm.com (accessed: 01 April 2022).
[21] Project-repository. https://github.com/1234/5678
[22] Smith, Arfon M., Daniel S. Katz, Kyle E. Niemeyer, and FORCE11 Software Citation Working Group. 2016. Software citation principles. *PeerJ Computer Science* 2: e86. doi:10.7717/peerj-cs.86.

5. Data exchange and long-term data accessibility

- Which data sets are especially suitable for use in other contexts?
- Which criteria are used to select research data to make it available for subsequent use by others?
- Are you planning to archive your data in a suitable infrastructure? If so, how and where?
- Are there any retention periods?
- When is the research data available for use by third parties?

The collected data and scripts lend themselves to re-use by third parties. Therefore, data as well as scripts are published in the institutional repository of Freie Universität Refubium. Included in the publication are all generated raw data and scripts, as well as final versions of text data and tables. Documentation will also be included with the publication (see Section 2). Intermediate results of processing and analysis steps, which can all be generated from the provided data and scripts, are not part of the publication.

The publication follows the recommendations of the Open Access Policy and Research Data Policy of Freie Universität. The results are licensed under an open license (likely Creative Commons 0[23] or BY[24]).

By using the repository Refubium, several of the points addressed in the FAIR principles[25] are ensured[26]. For example, the metadata are indexed via standardized interfaces (OAI-PMH) in overarching reference systems and search engines (e.g. BASE, DataCite Search, OpenAIRE). This provides increased visibility of research results. The metadata created is reviewed by the repository's editorial team. Furthermore, a persistent identifier (DOI) is assigned.

In accordance with the guidelines for safeguarding good scientific practice[27], the data will be made publicly available by the repository for at least ten years, i.e. without access restriction. Separate archiving, independent of publication, is not provided. No embargo period is required. Publication will take place as soon as possible, but at the latest within the last three months of the project duration.

6. Responsibilities and resources

- Who is responsible for adequate handling of the research data (description of roles and responsibilities within the project)?
- Which resources (costs; time or other) are required to implement adequate handling of research data within the project?
- Who is responsible for curating the data once the project has ended?

[23] Creative Commons — CC0 1.0 Universal. https://creativecommons.org/publicdomain/zero/1.0/deed.en (accessed: 01 April 2022).

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

[25] Wilkinson, Mark D., Michel Dumontier, IJsbrand Jan Aalbersberg, Gabrielle Appleton, Myles Axton, Arie Baak, Niklas Blomberg, et al. 2016. The FAIR Guiding Principles for scientific data management and stewardship. *Scientific Data* 3, No. 1 (December): 160018. doi:10.1038/sdata.2016.18.

[26] Guidelines Research data. https://www.fu-berlin.de/en/sites/refubium/about/guidelines/guidelines_fd (accessed: 01 April 2022).

[27] Deutsche Forschungsgemeinschaft. 2019. Guidelines for Safeguarding Good Research Practice. Code of Conduct. doi:10.5281/zenodo.3923602.

The primary person responsible for handling the research data generated by the project is Maxi Mustermann, PI of the project. Compliance with and updating of the DMP will be ensured by <FUNCTION> Jane Doe.

Ongoing documentation and preparation of data and scripts will occur during the project period; their finalization will occur during the last three months.

At the end of the project term, all data intended for publication will be published at the specified location. No further curation of the data will take place beyond the duration of the project.

The resources required for research data management are listed in the project plan under X.