UNIVERSITÄT DES SAARLANDES

Research data management
Guidelines and recommendations

Version 0.1 | Digitalization and Sustainability unit | 30 January 2023

Introduction

Handling research data is one of the everyday tasks in research. It is important to follow guidelines and regulations to ensure that research can benefit science in a long-term and sustainable way. A data management plan (DMP) is an important tool that describes and sets out how research data is handled throughout the data life cycle. Many third-party funding providers already require DMPs in funding applications. Even if this is not the case, creating a DMP for work on a research project is beneficial. Especially with regard to technological progress and the ever-growing amount of research data associated with it, a DMP is an indispensable means to implement FAIR principles. The policy on data research management at Saarland University is available at FDM_policy_UdS_v02.docx.

The FAIR principles are generic rules set out in 2016 for handling research data that aim to improve findability, accessibility, interoperability and reusability[1].

1 File formats

1.1 Recommended file formats

Where possible, the data collected should be stored in formats that ensure long-term readability and accessibility. File formats often have open, well-documented source code and are non-proprietary. It should also be ensured that data formats can be processed on all (common) computer systems. Research data should also be uncompressed. Some file formats are mentioned below for reference:

- Image files:
    - Vector graphics (sketches, drawings, pictograms) are scalable (.svg, *scalable vector graphics*), and result in a significantly smaller file size for large image sizes. However, vector graphics cannot be processed in all common word processing programs.

[1] https://www.go-fair.org/fair-principles/, retrieved 2 December 2022.

- In contrast to vector graphics, raster graphics (photos, microscope images) cannot be scaled arbitrarily. At high image resolution, this results in large file sizes, which are often stored in a lossy file format (e.g. .jpg). Instead, care should be taken to compress the image data losslessly by using the formats .tiff (tagged image file format), .bmp (bitmap) or .png (portable network graphics).

- Video files:
  - Video files should also preferably be recorded in a cross-platform format. This means using non-proprietary codecs, such as .h264 or its successor .h265. As a standardized container format, .mp4 is registered under the standard ISO 14496-14.
- Text files:
  - Open document standards are recommended as the open formats of OpenOffice (.odt, .ods, .odp) specified under ISO 26300; in addition, the rich text format (.rtf) and .docx offer a well-documented, widely used standard, however these are proprietary formats. ASCII text can be saved in .txt formats or as hypertext (.html, .xml), .csv for comma-separated variables, or .tex for LaTeX documents.
  - The widely used .pdf (ISO 32000) format is suitable for non-editable text.

This list only covers a few selected scenarios. A more comprehensive overview of recommended file formats for research data can be found, for example, on the Max Planck Society website[2].

1.2 Conversion

If the research data you generate is in a format that differs from the recommendations above (e.g. proprietary format) and you want to convert the data to ensure the use of open standards, always keep a copy of the data in the original format. If it can be ensured that all data, including all metadata, have been converted correctly, this rule can be deviated from.

It should be noted that in some of the file formats specified above, lossless storage occurs only if the original raw data was also generated in this format, such as in the case of .bmp or .mp3.

[2] https://rdm.mpdl.mpg.de/before-research/file-formats/, retrieved 2 December 2022.

2 Creating a data management plan

Due to the heterogeneity of the different research disciplines, there is no uniform structure for a DMP. However, core elements can be defined which also reflect the recommendations of many third-party funding providers, such as the German Research Foundation (DFG).

A DMP contains structured information about the research project at all times. Certain information is static in this case, such as information on project responsibilities, other information is updated dynamically, such as the access rights in the event of personnel changes in the project.

2.1 Components of a data management plan

The general components of a data management plan are listed below:

1. Administrative information on the research project: Objectives, partners, project managers and duration of the research project
2. Data generation
    - How does your project generate new data?
    - Are existing data reused and how are they migrated?
    - In terms of data formats like image data, text data or measurement data, which data types arise in your project and how are they processed further?
    - To what extent do these arise or what is the anticipated data volume?
    - Which hardware and software will be used to process the data?
    - How is the data structure organized?
    - How are meta data generated?
3. Data security and workflow
    - Where is the data stored during the data life cycle?
    - Are the data additionally backed up (backup)?
    - At which intervals are backups performed?
    - Are the file storage systems checked for errors?
    - Are there records that contain particularly sensitive information?
    - How is access to data managed?
    - How are the data named (scheme)?
    - How are data versioned?
    - How can data be synchronized between project partners?

4. Data ingestion plan and archiving
    - Data selection: Which primary data are processed further?
    - How is data transferred/exchanged?
    - How are data validated?
    - How are data archived?
    - Information on metadata
5. Consolidation
    - What ownership rights and copyrights apply to the data?
    - What rights of use are foreseen?
    - Are all requirements of the funding providers taken into account?
    - Which backup strategy is followed?
    - Are all aspects of data protection addressed?

2.2 Electronic Lab Notebook (ELN)

In accordance with the principles of good scientific practice, researchers are required to document notes, experimental procedures, protocols, parameters and metadata when conducting and planning experiments. Unlike traditional lab journals, ELNs offer a number of benefits. Electronic laboratory journal entries can be searched for by keywords, stored securely and shared easily with partners and members of the working group. Often these services also offer a calendar for shared laboratory devices as well as the possibility to manage and link various devices (laboratory information management system, LIMS).

ELabFTW[3] is a browser-based open source tool for the electronic management of laboratory notes that is used at many research institutes worldwide.

Some of the options available to researchers include:

- Time stamp for experiments carried out (RFC 3161-compliant)
- Import and export to PDF, zip, CSV, JSON
- Extensive template library for experiments
- Supports structure of experiments and protocols in individual steps
- REST API for direct access to data through a programmable interface

[3] https://www.elabftw.net/, retrieved 2 December 2022.

- Supports uploading files and embedding images
- Supports chemical drawings (Chemdoodle)
- Large developer and user community worldwide
- To-do lists
- Built-in scheduling assistant
- Cross-platform (Mac, Linux, Windows)
- Worldwide access possible, even with mobile devices
- Storing and managing reagents, protocols, cell lines, etc. (LIMS)

Saarland University will investigate the option of implementing a central ELN service if sufficient demand is indicated.

3 Storage options

Below are some ways in which research data generated within a Saarland University working group can be stored.

3.1 M365 – Cloud Storage

As part of the university-wide use of Microsoft365, UdS offers a cloud-based solution for storing data. This data storage facility also offers the advantage of easily sharing and editing the data with (external) collaboration partners. For instructions on how to set up OneDrive or SharePoint for file storage, see the SharePoint-OneDrive.pdf Guide link.

3.2 hizCloud

This is a Nextcloud-based internal university sync and share service offered by the Saarland University IT Centre (HIZ)[4]. The service can be used to exchange and synchronize files on different devices.

The hizCloud is suitable for individual users and working groups. A personal storage quota of 50 GB is available to each employee. In addition, there are group folders that allow members of a working group to manage and edit the shared documents.

Depending on the selected storage quota of the group folder, annual fees are incurred, unlike with M365.

[4] https://www.hiz-saarland.de/dienste/hizcloud, retrieved 2 December 2022.

3.3 University-internal data storage for projects with >50 TB

Saarland University intends to set up its own research data storage facility for working groups with very large amounts of data. Please contact the Digitalization and Sustainability staff unit for more information.

3.4 Further tools

The aforementioned data storage and management options offer many advantages for storing primary data that is frequently accessed in the course of the project or expanded by secondary data. However, if the data is to be archived and/or published, there are repositories that also have the option of assigning a persistent identifier, such as a DOI. A widely used repository is Zenodo[5], which is managed by CERN and is regarded as a pioneer in open science. An overview of all repositories for research data of various disciplines is listed on the website www.re3data.org.

4 Metadata management

Research data is subject to a complex data life cycle. Within this cycle, however, in accordance with the rules of good scientific practice, certain metadata must be visible at all times in order to be able to clearly attribute data to persons and projects. The NFDI aims to develop subject-specific schemes for metadata. A general approach can be found, for example, in the global consortium DataCite, which has set itself the goal of making research results sustainable and transparent through easy access to research data. The basis is a standardized metadata schema with necessary information (**m**andatory), recommended information (**r**ecommended) and **o**ptional information. The Dublin Core metadata schema (ISO 15836) offers further standardization with the introduction of 15 *core elements* for declaring electronic resources. Based on these standards, we recommend using a metadata schema as follows:

[5] http://www.zenodo.org/, retrieved 2 December 2022.

ID	Name/Property	Description
1	Identifier	Unambiguous identification of an object by adequate assignment to a type catalogue (e.g. DOI, ISBN, ISSN)
2	Creator	The name of the responsible author or the author of the document in its current version. Instead of a natural person, this can also be an organization.
3	Title	Indication of the title of the document described.
4	Publisher	Name of the institution that publishes, distributes, releases or archives the data.
5	PublicationYear	Indication of the date when the data was published or is expected to be published, if foreseeable.
6	Subject	Indication of (multiple) searchable keywords or keywords that reflect the topic of the content.
7	Contributor	Indication of other persons involved in the creation of documents, typically collaboration partners and co-authors.
8	Date	Definitive date in the data life cycle. This can be both the start or end date of the data generation, but can also cover periods of time. The notation should be made in accordance with the standard (ISO 8601) in the form YYYY-MM-DD.
9	Language	Indication of the language(s) of the resource, if applicable.
10	ResourceType	Specifies the type of resources involved, such as audiovisual data, images, video or text documents.
16	Rights	Information on intellectual property rights reserved for the data. Depending on the data type, the Creative Commons' CC-BY 'or GNU Affero Public License (source code) may be suitable.
17	Description	Summary of the content of the document in free text form *(abstract)*. In addition, information on how the data was generated,

		including all resources used (hardware/software) and any test protocol.
19	FundingReference	Information on funding bodies and authorities that have funded the research in question.

The nomenclature of the ID and the associated property (descriptor) in the above table is based on the DataCite metadata schema[6], where further definitions and recommendations for action can also be found. We would like to point out that the above metadata entries are by no means exhaustive and represent only a minimum consensus for the assignment and management of the relevant research data. Likewise, the sequence in which individual metadata fields can be named depends on the respective point in time in the data life cycle and thus the completeness of the metadata is also to be regarded as dynamic.

The collected metadata can be entered in an input screen using various software tools and then stored in the relevant file folder together with the actual research data[7].

5 Administrative recommendations for research data

In order to ensure the sustainable handling of generated research data, it is advisable to implement a uniform and well-documented filing structure in each working group. This ensures that there are no obstacles to finding and reusing the data and that it is clear to third parties how the data records are stored.

5.1 Naming of records

When naming research data, there are certain principles that ensure a sustainable approach. A number of naming recommendations are provided below:

[6] https://schema.datacite.org/meta/kernel-4.4/doc/DataCite-MetadataKernel_v4.4.pdf, retrieved 2 December 2022.

[7] https://dhvlab.gwi.uni-muenchen.de/datacite-generator/, retrieved 2 December 2022.

- When naming files, always use a version number or the date on which the data was generated in the format "YYYY-MM-DD".
- Pay attention to alphanumeric sorting and, if necessary, leading zeros for several similarly named files or folders within a systematic measurement series, for example.
- You can choose your file naming scheme so that information on the data is included in the file name.
- Do not use special characters or spaces in file names. Instead, use a hyphen "-" or an underscore "_" or capital letters to separate words.
    - **Unsuitable** file naming examples:
        - 'Test-v1.txt'
        - 'Measurement Date.txtÄ'
        - 'Oilfilm_v=5mms^-1.bmp'
    - **Suitable** file naming examples:
        - 'OilFilm_2022-07-29.bmp'
        - 'RedBloodCell_u1_0mm-s_v01.csv'
        - 'GlycerolViscosity_eta_20mPas_2021-04-21.csv'

5.2 Folder structures

In research operations, there is a heterogeneous mix of data generated according to their importance for research projects. When generating primary data, their importance generally cannot yet be sufficiently estimated and they can only be classified in the further course analysing these primary data. Based on these considerations, we propose the following recommendations for structuring the research data according to

1. Primary data (raw data, measurement results, intermediate results, simulation data, observations, etc., including metadata)
2. Secondary data (analysed primary data that can be interpreted by scientists in the respective discipline)

When storing the primary data, a rapidly evolving storage schema and nomenclature should be available so that the research results can be reproduced and the data genesis can be traced using the stored metadata. Since these data structures vary greatly according to specialist discipline, further recommendations cannot be given beyond this point.

However, in the case of secondary data, a more precise storage structure can be defined based on the data life cycle of the secondary research data from generation to final publication. As a basic structure, we recommend a directory structure using the following schema:

Directory name	Description
00_Publication	o The final version of the research paper should be stored in this directory, both as source code and as .pdf. o Indication of metadata according to Dublin Core in ASCII form. o Likewise, upon publication, the indication of the DOI. o Additional documents relating to "Supplementary Materials" can be stored in subdirectories.
01_Manuscripts	o The final version of the research paper should be stored in this directory, both as source code and as .pdf.
02_Figures	o Storage of all images and graphics required for publication with indication of the nomenclature rules unless self-explanatory. In addition, the specification of the underlying data points of the graphics that were extracted from the primary data. If necessary, create subfolders for each version of the manuscript. The graphics should be in sufficiently high resolution and preferably at least in .pdf format, unless they are raster graphics such as photos.
03_PrimaryData	o Storage of the primary data included in the publication, including metadata. o If an ELN is used, an extract of all relevant entries.
04_Programs	o Specification of all software tools and source code used including parameters for the respective simulations or calculations.
05_Bibliography	o Storage of all sources used as ASCII text, e.g. by using Zotero or BibTex. If possible, the described sources should also be stored in order to ensure access to data that is not freely accessible (e.g. Master's or Bachelor's theses).
06_Miscellaneous	o Additional files (e.g. reports on publication, e-mail correspond-ence, video files).