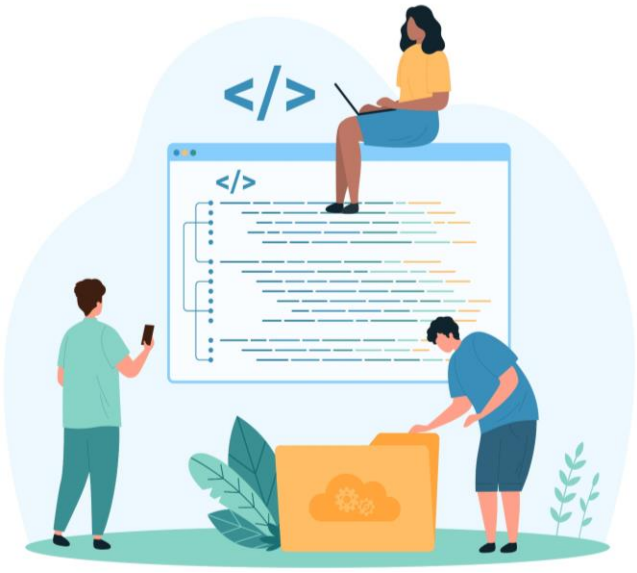
WG Research Data Management



HANDOUTS FROM THE WORKING GROUP RESEARCH DATA MANAGEMENT

STORAGE STANDARDS FOR DATA ORGANISATION

Exposée

Data quality | Data organisation | Folder structure | File conventions | Version control | Documentation | Data formats | Metadata schemas

Edited

17 December 2025 (V 1.7) | Richter Anja, Keyvan Djahangiri

Contact details

forschungsdaten@uni-passau.de
Website

Background

Professional, quality-assured handling of research data is crucial for the connectivity of research and the verifiability of scientific findings.

Research data should be FAIR (**F**indable, **A**ccessible, **I**nteroperable and **R**eusable). Maintaining the authenticity, integrity, accessibility, and comprehensibility of your research data, therefore, plays a key role in its long-term reusability. Well-thought-out planning, precise documentation, and description of your data, consideration of subject-specific (metadata) standards, and provision of the technical infrastructure can make a significant contribution to this.

Before archiving or publishing data, it can be useful to precisely select which data should be stored permanently through the data validation process. At this point, you should also transfer your data to open file formats at the latest to increase long-term readability and usability.

Storage standards for data and file organisation

Data quality

A well-planned data and file organisation is essential for data quality, which includes aspects such as the completeness and accuracy of the data, but also criteria such as the data format.

Intrinsic data quality

- *Objectivity and accuracy:* Data should be unbiased and neutral.
- *Reputation and credibility:* The source of the data influences its quality, e.g. interviewees must be genuine and authentic, while field research providers must be trustworthy and reputable.

Contextual data quality

Certain aspects of data quality can only be assessed in the context of the respective task and therefore require particularly thorough planning before conducting a research project.

- *Completeness*: Data should contain all relevant information to enable a comprehensive analysis.
- *Timeliness*: Data must always be up to date to serve as a foundation for accurate conclusions.
- *Relevance*: Data must be relevant to the specific research context.

Data organisation

To keep the data traceable, make it easier to find, and thus avoid delays in the research process as well as additional work, you should establish a consistent data organisation from the outset, defining clear rules and specifying responsibilities. This includes a clear folder structure, ontologies, and conventions for folder and file naming that are adhered to by all project participants, as well as the consideration of special requirements for data organisation, e.g. in the case of very large datasets, special data protection guidelines for handling personal data, and regulations on access rights within a project group.

- Folder structure: To ensure a clear and consistent folder structure, it is ideal to create a traceable, hierarchical data organisation with a maximum of three subfolder levels. Data that are structurally or thematically related should be stored together, and the folder should be named accordingly in terms of system and content. After the project is completed, check which data – and therefore which folders – are still needed, and which can be deleted to keep the folder structure as simple and comprehensible as possible.
- File conventions: Naming conventions for files and folders, as well as the meaning of used abbreviations, should be carefully selected, plausibly defined, and equally well documented. Use meaningful components such as creator names, versions and processing stages, and make sure to keep the naming consistent.
  Date formats are to be specified in the form YYYYMMDD, numbers always in two or three digits with 0 as a placeholder, and different versions of a file as V with numbering (V01, V02, etc.). Spaces, dots, special characters, umlauts and excessively long file names are not recommended; instead, capital letters and underscores should be used. Avoid repeating information from folder names in the file names.

Project name
MusicCulture _Germany

1. **Level: Category**
   /[Folder number]_[Data status]
   /01_RawData
   /02_ProcessedData
   …
2. **Level: Subcategory**
   /01_RawData/[Period]_[Data]_[Creator]
   /01_RawData/2025_Interviews_Schmitt
   /01_RawData/2025_Interviews_Meier
   …
3. **Level: File**
   /01_RawData/2025_Interviews_Schmitt/[YYYYMMDD]_[Region]_[Interviewee] _[Topic].[File format]
   /01_RawData/2025_Interviews_Schmitt/20250326_Bayern_Müller_TraditionalMusic.mp3

- Version control: If you work with file versions to better track development steps, carry out a version control on a regular basis, store obsolete versions separately after a backup, or delete them if necessary. This reduces the risk of continuing to work with outdated files. A clear labelling of versions in the file name in the form of V01, V02, etc., or in a standardised header within the file provides a simple and quick overview. Additionally, you can use version control software such as Git and Apache Subversion to support this process.

Documentation

A well-maintained, accurate, and constantly updated documentation helps to retrieve data, trace its origin, provide contextual information for interpretation, and prevent confusion with alternative versions. Moreover, accurate documentation is an integral part of good research practice [content in German] (GWP, §9, p.6).

In addition to a permanent contact and general information about your project, your documentation should also include details of the data collection methods, structure and relationships of the data, quality measures, a codebook if applicable, file versions, and information regarding access, confidentiality and terms of use.

Documentation can be managed through a data management plan (DMP), a ReadMe file, a well-structured metadata repository, or an internal project wiki. ReadMe files are a particularly simple form of documentation, providing details on data, folders, and possible changes. Unlike metadata, which is intended for machine analysis, ReadMe files aim to provide other researchers with a clear overview and important contextual information about the data.

Regardless of whether the (meta) data was collected automatically or manually, it is important to use standardised and open formats, subject-specific vocabularies, and common programming languages during the storage process. You should also use structured metadata to ensure that your research data is machine-readable and maintain a clear link between metadata and datasets through persistent identifiers (PIDs), such as DOI.

```
This README.txt file was generated on 2024-02-19 by Naiara Korta Martiartu (naiara.korta@unibe.ch)

------------------
GENERAL INFORMATION
------------------

Contributor information:

      Name: Naiara Korta Martiartu
      Role/Function: Principal Investigator, Data Collector (main contact person)
      Institution: Institute of Applied Physics, University of Bern
      Address: Sidlerstrasse 5, 3012 Bern, Switzerland
      Email: naiara.korta@unibe.ch

      Name: Parisa Salemi Yolgunlu
      Role/Function: Data Collector
      Institution: Institute of Applied Physics, University of Bern
      Address: Sidlerstrasse 5, 3012 Bern, Switzerland

      Name: Martin Frenz
      Role/Function: Sponsor
      Institution: Institute of Applied Physics, University of Bern
      Address: Sidlerstrasse 5, 3012 Bern, Switzerland

      Name: Michael Jaeger
      Role/Function: Principal Investigator (alternative contact person)
      Institution: Institute of Applied Physics, University of Bern
      Address: Sidlerstrasse 5, 3012 Bern, Switzerland
      Email: michael.jaeger@unibe.ch


Date of data collection: December 2022 - May 2023

Location of data collection: Institute of Applied Physics, University of Bern, Bern, Switzerland

Keywords: pulse-echo ultrasound, attenuation imaging, tissue characterization, ultrasound tomography,
```

Extract from a ReadMe-file

File formats

For long-term usability and long-term archiving, it is crucial to pay attention to compatibility (interoperability) and loss-free conversion to alternative formats when selecting a file format. Open, non-proprietary formats are therefore particularly recommended.

The suitability of certain file formats always depends on the software used and the specific objectives of the research projects. However, many file formats can already be easily converted into open formats. Additionally, file formats provide information about the structure of the data stored in the file, as well as its purpose and affiliation.

File type	*Recommendable formats*	*Less recommendable formats*
Table data	CSV, TSV, SPSS portable	Excel (.xlsx/.xls)
Text	TXT, HTML, RTF, PDF/A	Word (.docx/.doc)
Presentation	PDF/A	PowerPoint (.pptx/.ppt)
Multimedia	MP4, WAV, AVI	WMV, QuickTime, H264
Images	TIFF, JPEG2000, PNG	GIF, JPG
Structured data, data exchange	XML, RDF, JSON	SDXF, RDBMS

The CSV format is a particularly good choice due to its openness and compatibility. In contrast, Microsoft's XLS format is proprietary and therefore less accessible.

Metadata schemas

Metadata is 'data about data', i.e., it provides structured information about the available data so that humans and machines can better understand and process the data.

Different types of metadata (e.g. bibliographic, administrative or descriptive) serve various functions. Among other things, they document who collected, processed, and published the data, the methods used to generate the data or the legal conditions under which it can be used. Metadata standards facilitate compliance with specifications, such as the FAIR principles, and ensure that key elements are properly described.

Metadata can have both interdisciplinary and subject-specific characteristics. It is advantageous to structure data according to metadata schemas established by your research community

Additional Hints

- Simply storing data on your personal computer or an external storage medium does not constitute professional archiving after a project's completion. Instead, depositing data in a domain-specific or general repository is recommended.
- You can use the RDMO instance (Research Data Management Organiser) of forschungsdaten.info or the GFBio DMP Tool to plan, organise and manage your data.
- The research data service OstData (in German) and Cornell University (in English, including instructions) provide a template for creating a ReadMe file.
- When selecting research data for publication or archiving, you can use the checklist 'Five steps to decide what data to keep' from the Digital Curation Centre (DCC).
- For renaming large quantities of files and folders, you can use supporting software, such as Ant Renamer Portable, ReNamer (for Mac), GPRename (for Linux).

- A standard for bibliographic metadata is, for example, DataCite Metadata Schema; for administrative metadata in long-term archiving, there is e.g. PREMIS and for subject-specific metadata there is a multitude of standards. An overview of existing metadata standards can be found at Metadata Standards Catalog of the Research Data Alliance, at FairSharing.org, and at DDC.

Links

- Overview of free version control software
- Further examples of file naming conventions and documentation guidelines [content in German] from VerbundFDB
- Example of a dataset featuring a more concise ReadMe file
- Information on the long-term usability of file formats at KOST (Coordination Centre for the Permanent Archiving of Electronic Records) [content in German], and NRW State Initiative on Long-term Availability [content in German]
- See also the handouts and information sheets from the Working Group Research Data Management