Handouts for research data management:

Planning and implementation of research projects

Planning phase

Data management plan

A data management plan (DMP) can be drawn up to plan RDM. A DMP is a document that describes the entire life cycle of research data. It is usually created by researchers, scientists or project teams to plan and organize the handling of data during and after a research project. The main purpose of a DMP is to ensure that research data is managed in an efficient, traceable, long-term accessible and understandable way.

Data management plans are particularly important to promote the reproducibility of research, increase the efficiency of data management and ensure that valuable research data is preserved in the long term. Funding bodies or institutions often require researchers to submit such a plan before a project is approved.

The following should be noted:

- Requirements of the third-party funders
- Funding of research data management
- Legal and ethical issues
- Publication of scientific products

Technical tools for data management planning:

RDMO stands for “Research Data Management Organizer” and is an open source software platform used to create, manage and share data management plans (DMPs). RDMO offers researchers, project teams and institutions the opportunity to create and implement effective data management plans to manage the entire lifecycle of research data. A separate RDMO instance is available for universities in the state of Brandenburg under the name RDMO-BB.[1]

Implementation phase

(Data-) Documentation

Data documentation is an integral part of scientific research and is used during and after the completion of a project. The primary goal is to capture structured information about the how and why of data collection in a processed and digitized form. Good documentation is crucial for short-term and long-term understanding, as well as successful long-term data storage, and also enables errors to be identified and rectified.Datendokumentation findet auf mehreren Ebenen statt.

[1] Link zu RDMO-BB: https://rdmo.fdm-bb.de/ (Stand 19.11.2024)

- At the first level, basic information about a study is collected to provide context to the methodology and data collection. This information can be summarized in the form of a DMP (see).
- The second level comprises further descriptions of individual data folders within a data collection and provides detailed information at a glance on individual data (e.g. collected characteristics, variables) as recorded in ELNs (electronic lab notebooks).
- Metadata (standardized, structured data about the collection) that is often used to catalog and locate objects in the collection. A widely used metadata standard is, for example, the DataCite metadata schema.

ReadMe-File:

ReadMe files are suitable for easy-to-implement documentation of research data. These contain all the relevant information required to understand the research data and are a prerequisite for subsequent use. This includes, among other things:

- Title of the research dataset
- Project name and description
- Identifier
- Place of data collection
- Version
- Creator of the data set
- Description of the research data (e.g. content, method used for data collection, context of origin)
- Description of the method(s) used for data collection and processing
- File list
- File formats
- Keywords
- Information on legal requirements for the research data

ReadMe-Templates:

- Ostdata: https://zenodo.org/records/6956989
- TU Braunschweig: https://www.tu-braunschweig.de/fileadmin/Redaktionsgruppen/Einrichtungen/UB/README_Template_TUBS.txt (Link via: https://www.tu-braunschweig.de/forschung/forschungsdaten-transparenz/forschungsdaten/fdm-services/informationsmaterialien-links)
- Some repositories also provide ReadMe-Templates

Data quality

Data quality control is an integral part of any research and takes place at various stages: data collection, data entry or digitization and data review. It is crucial to assign clear roles and responsibilities for data quality assurance at all stages of research and to develop professional procedures before data collection begins.

Data collection

When collecting data, researchers must ensure that the data collected reflects the actual facts, answers, reflections or events. The quality of the data collection methods used has a major influence on data quality, and the detailed documentation of data collection is evidence of this quality.

Quality control measures during data collection may include:

- Calibration of instruments to check the accuracy, bias and scope of the measuremenDurchführung mehrerer Messungen, Beobachtungen oder Probenahmen
- Verification of the data collection by another person
- Use of standardized methods and protocols for data collection as well as data collection guidelines with clear instructions
- Computer-assisted survey software to standardize surveys, check the consistency of responses, route and adapt questions so that only appropriate questions are asked, confirm responses against previous answers and identify unreliable responses

Data entry und recording

When data is entered into a database or spreadsheet program, coded, digitized or transcribed, quality is ensured and errors are avoided, for example by using standardized and uniform procedures with clear instructions:

- Setting up validation rules or input masks in data entry software
- Use of control vocabularies, code lists and pick lists from which values must be selected to minimize manual data entry - there are internationally agreed conventions for recording information such as ISO 8601, the recommended format for presenting dates and times
- Detailed labeling of variables and data set names to avoid confusion
- Development of an appropriate database structure for the organization of data and data fields

Data control

Bei der Datenkontrolle werden die Daten bearbeitet, bereinigt, überprüft, abgeglichen und validiert. Die Prüfung umfasst in der Regel sowohl automatisierte als auch manuelle Verfahren, wie z. B.:

- Double-checking the coding of observations or responses and of values outside the estimated measurement range
- Checking the completeness of the data
- Checking samples of digital data against the original data
- Double entry of data
- Statistical analyses such as frequencies, means, ranges or clustering to detect errors and anomalous values
- Proofreading of the transcription
- peer review

Further literature:

- Cai, Li and Zhu, Yangyong (2015): "The Challenges of Data Quality and Data Quality Assessment in the Big Data Era", Data Science Journal, Volume 14, pp. 1-10, DOI: http://dx.doi.org/10.5334/dsj-2015-002
- Sidi, Fatimah; Shariat Panahy, Payam Hassany; Affendey, Lilly Suriani; Jabar, Marzanah A.; Ibrahim, Hamidah & Mustapha, Aida (2012): "Data quality: A survey of data quality dimensions," *2012 International Conference on Information Retrieval & Knowledge Management*, Kuala Lumpur, Malaysia, pp. 300-304, DOI: https://doi.org/10.1109/InfRKM.2012.6204995

Storage and processing

The secure storage of research data is one of the essential tasks in research data management. This includes suitable storage services and media as well as a forward-looking storage and backup concept in order to effectively prevent data loss. Data security, especially for data containing sensitive information, must also be guaranteed by suitable measures. Special requirements must also be taken into account for collaborative working, such as the definition of individual access and usage rights.

Data organization forms the basis for the productive **handling of research data**. This includes the following tasks:

- Create a comprehensible folder structure for data
- Schema for uniform and meaningful file naming
- File versioning: use versioning scheme / automatic file versioning
- Use open and standardized file formats

Storage and backup

In principle, the sole use of local storage media and commercial storage services is not advisable. Saving on institutional storage media, on the other hand, offers a certain degree of security, as regular backups and maintenance are guaranteed.

Backup measures should be carried out at regular intervals and according to a defined schedule. The following rule should be observed: The tried-and-tested 3-2-1 backup rule requires that three copies should be saved on at least two different storage media, one of which should be stored decentrally, e.g. in a secure cloud.

Cloud services such as Nextcloud and corresponding software programs are among the most common backup tools.

The storage and backup strategy must be defined in advance/at the start of the project, recorded and, if necessary, agreed within the project group.

Data security and level of protection

Secure data storage also includes ensuring the security of the data. Especially if it contains sensitive information. If the data is related to the GDPR, it should not be possible to access the data from outside (collection phase). The following measures must be taken to protect data from unauthorized access

- Access restrictions (password protection, encryption, etc.)
- Pseudonymization or anonymization of personal data

Data formats

Describe the data type (e.g. raster geodata, questionnaires, interviews) and the file format (TIFF, CSV table, ODT/DOC). If possible, use open file formats and avoid proprietary file formats. Justify the use of proprietary software for data creation (widespread use and acceptance, skills of researchers, existing purchased license).

The choice of file formats has a direct impact on readability, exchangeability and archivability. With regard to the latter in particular, it is best to contact the institution (library) or repository where the data is to be archived, as there may well be different specifications. The difference between recommended and accepted formats can vary greatly.

The following table provides an overview of the most common file types and the formats recommended for you:

Data format	**Recommended Format**
Text	TXT, PDF/A (Typen 2a, 2u und 2b)
„Office Documents“	PDF/A
Tables	CSV
Grid images	TIFF, JPEG2000
Audio	WAVE
Video	MPEG-4
Structured (Text-)Data	XML
Geo data	TIFF + EWF, XML, INTERLIS

Further informations:

- KOST (https://kost-ceco.ch/cms/kad_recommendation_de.html)
- ETH Zürich (https://unlimited.ethz.ch/display/DD/Archivtaugliche+Dateiformate)
- Verbund Forschungsdaten Bildung (https://www.forschungsdaten-bildung.de/dateiformate)
- Library of Congress (https://www.loc.gov/preservation/resources/rfs/RFS%202023-2024.pdf).

Versioning (Software)

During project implementation, data sets are usually subject to constant (further) development (e.g. during selection, aggregation, integration). It has proven to be advantageous to work with versioning, i.e. to mark and document the various versions and keep them for the duration of the project. Especially with text-based data, the use of versioning tools such as Git or SVN facilitates the handling of the different versions.

Contact

Counseling for reseach data management

Blanka Goßner

Forschungsdatenmanagement | Zentrum für Forschung und Transfer

Technische Hochschule Wildau | Hochschulring 1 | 15745 Wildau

Haus 13 | Raum 0.41

Telefon: +49 3375 508 322

Mail: blanka.gossner@th-wildau,de

Counseling on Application for third-party funded research projects

Zentrum für Forschung und Transfer

Mail: forschung@th-wildau.de