Guidelines for the publication and archiving of research data at the TU Braunschweig

The TU Braunschweig offers its researchers the opportunity to share their research data [Leopard](#) archive and, if desired, publish. The most important points are discussed below.

1) Prepare research data

Checklist

- ☐ I have saved my data in a comprehensible folder structure.
- ☐ My folder and file names are also understandable for third parties, meaning they are systematic, content-related and clear.
- ☐ The data is designed in such a way that it enables independent traceability and reusability.
- ☐ I have identified data that is worthy of archiving or has selected published data that reflects the scientific relevance of my work.
- ☐ At best, my data is available in standardized, common file formats for free display programs (e.g. XML, CSV, TXT, TIFF).
- ☐ I made sure that the data was not read-only.

Folder structuring

In order to structure the research data as comprehensibly and clearly as possible, a hierarchical structure is suitable. What this looks like depends on the data set and must be determined individually each time. If the data structure is heterogeneous (many different file types), you can create folder types, e.g.:

- Materials →   Data, documentation, publications
- Data →   Texts, photos, models, tables

If the data is more homogeneous, it can also be divided by time, location or method, e.g.:

- Project_XY →   Study_area_01 →   Measurements

It is important that the desired files can be found quickly and intuitively.

File naming

When naming files, a naming convention should be established from the start that is applied consistently throughout the project, especially when multiple parties are involved in a project. The naming should provide information about the content of the data and help to clearly identify it. The creation date, a short description, the location, project name, sample number, analysis method or version number are particularly suitable for this.

The following rules should also be observed:

1. File names should be short but descriptive.
2. No special characters, spaces, punctuation marks or umlauts should be used.
3. Capital letters and underscores should be used instead.
4. Capital letters affect the sorting, which is why you should ensure consistent naming.

5. The most suitable date format is: YYYYMMDD.
6. If numbers are specified, they should always have two digits or even three digits, so that the 0 serves as a placeholder (e.g. Probe01 and not Probe1).
7. The individual elements of the file name should be arranged so that the file can be found as quickly as possible; beginnings of the name such as "Test" or "Draft" should be avoided.
8. If there are different versions of a file, a numbered V should be used to keep order between versions (V01, V02, etc.).
9. Repeating information from folder names in file names should be avoided.

Useful elements for file naming

Meaningful elements	Notes and example
Author's name or initials	LG
Content/technical description	Tags
Date of creation primarily for version control	YYYYMMDD
Processing status	Original, draft, part
version	V02

Examples of file naming

Good	Bad
20240813Agenda	13Aug2024Agenda
2024UBWorkshopRDM	The_2024_Workshop_on_Research Data-management_of_university_library!
20200804_ProjectA_Protocol_LG	Protocol Laura

In-depth information: http://www.data.cam.ac.uk/files/gdl_tilsdocnaming_v1_20090612.pdf

The naming convention of the files and folder structure should be explained in the accompanying README file (see Section 2) of the data set if abbreviations are used.

Data selection

In order to decide which data is suitable for archiving/publication, it can be divided into the following categories.

Category 1: Very valuable research data

  Research data that
  • are unique or very difficult to restore and

- are interesting for a large community and/or for a wide range of possible uses.

→ Be sure to archive/publish!

Category 2: Limited valuable research data

Research data that either
- can be restored with reasonable effort or
- are only interesting for a short time, a small target group or only to prove research results.

→ Archive/publish if effort and costs do not speak against it.

Category 3: Little valuable research data

Research data that
- only created for testing purposes,
- are illegible or not sufficiently described.

→ Data is not suitable for archiving/publication.

File formats

Not every file format is suitable for long-term archiving. To ensure that research data can be read and used even after several years, open, non-proprietary file formats are recommended.

	Recommended	Less recommended
text	PDF/A, TXT, HTML	Word (.docx/.doc)
Tables	CSV, SPSS portable	Excel (.xlsx/,xls)
presentation	PDF/A	PowerPoint (.pptx/.ppt)
multimedia	MP4, WAV, AVI	WMV
Pictures	TIFF, JPEG2000, PNG	GIF, JPG
Data exchange	XML, RDF, JSON	SDXF

Additional information:

https://forschungsdaten.info/themen/veroeffentlichen-und-archivieren/formate-erhalten/

2) Prepare documentation

Checklist

☐ I have maintained comprehensive documentation that enables reproducibility and reuse of the data set. This includes
  o Collection of research data
  o Processing of research data
  o Structure of the data set
☐ I have created a separate README file (TXT or PDF/A) with descriptive metadata.

Metadata

Metadata describes research data. They serve to improve the documentation and findability of data. Typical meat data is title, publisher, author or date. In addition, metadata contains descriptive

information about the context of the data, such as the measuring device, measuring location or software used. These help to understand the data set better. Metadata facilitates the long-term understanding of the research data collected and promotes its subsequent use, which is why they are essential for archiving or publication.

The linking of data and metadata can e.g. This can be done, for example, via a persistent identifier or a README file.

README file

A README file is a text file that is provided alongside a data set and helps describe and understand it. It primarily contains contextual or structural metadata to facilitate interpretability and reuse. Such information can e.g. B. units of measurement parameters, information about processing steps or an overview of individual files of a data package. Overall, keep the README file short and simple, with the information necessary to understand the dataset.

For very heterogeneous data sets (e.g. consisting of geodata, Matlab scripts and tables with measurement data), a) all information can be described in a README file, with reference to individual data specifics, or b) a "main" README be created that describes the context of the entire data set and additionally specific READMEs for sub-data sets or individual files, which are stored in the respective subdirectory. If there are several README files, care should be taken to ensure that the naming of the READMEs indicates which data set/file they describe. Care should be taken to ensure that the README files are formatted consistently.

We have a README-Template created for you. Use these and adapt them to your data set by deleting or adding points.

3) Excursus research software

Checklist

- ☐ My software is versioned.

- ☐ I have provided my software with the associated source code.

- ☐ I have formulated the purpose of the software in the included README file.

- ☐ The code is modularized and designed to be comprehensible, for example using self-explanatory names or comments.

Publishing research software

Software is becoming increasingly important in research. It not only helps analyze data, but is often also crucial for its production. You often need the underlying software in order to properly understand and reproduce research data. In addition, self-developed research software can represent an independent research result. Therefore, research software and research data should be published for reusability.

Metadata for research software

Research software requires its own metadata in order to describe it well and make it comprehensible. An example of a metadata standard is CodeMeta.

Archiving of research software

Long-term archiving of software is difficult. On the one hand, it must be described in an understandable way and, on the other hand, it must be ensured that it is executable. This is a problem, because software is often heavily dependent on the computing environment in or for which it was developed, i.e. on the operating system, hardware or software libraries. These change very quickly as technology advances, meaning that developed software can no longer be read or executed. There is no final solution to solve this problem yet. One approach is to use emulation, in which the executing computer and operating system are simulated in addition to the actual software.

Additional information

Project Suresoft: https://suresoft.dev/

4) Clarify legal aspects and choose a license

Checklist

- ☐ I have ensured that I am not violating any third-party rights (e.g. data protection, copyright, confidentiality agreements) by publishing it.
- ☐ I have selected a suitable license (e.g. Creative Commons) for subsequent use.

There are many legal issues to consider when handling research data. It is advisable to clarify legal issues and responsibilities at the beginning of a project. The following figure provides an initial overview of the questions that need to be considered before publishing research data.

taken from: Elisabeth Böker, & Peter Brettschneider. (2020, May). Publishing research data - organizational and legal issues. Zenodo.http://doi.org/10.5281/zenodo.3864901, CC BY 4.0

Copyright

Copyright protects intellectual creations and services. If there is copyright, the use, distribution and processing of the work is only possible with the consent of the author. Research data is rarely subject to copyright, but there may be exceptions. However, published research data is often provided with licenses that clarify possible subsequent use. Incidentally, databases are an exception and they enjoy specific protection under German law, whereby the creators have the sole right to distribute and reproduce (§ 87b UrhG). Generally, copyright protection ends 70 years after the death of the author. The works are then considered to be in the public domain.

Descriptive metadata of research data should generally not be considered protected by copyright, as they are usually relatively short, purely descriptive representations.

In addition to copyright, research data can also be subject to patent law in certain cases if the research data describes technical teaching. In such cases, only the patent holder may use the patented data.

Employment and service contract requirements

Open Access guidelines of the TU Braunschweig:

https://doi.org/10.24355/dbbs.084-202008280945-0

Guidelines for handling research data from TU Braunschweig:

https://doi.org/10.24355/dbbs.084-201911211552-0

In general, it is important to pay attention to what is stated in the employment contract regarding the rights of use (usually belonging to the employer) of the research data generated and whether, for example, there is a trade secret.

Funders of research projects can also have claims, for example when it comes to the rights to software created in the course of the research project. If in doubt, you should ask before publication in order to obtain consent for distribution if necessary.

Data protection

When handling personal data, various data protection aspects must be taken into account. Data protection restrictions can be avoided through anonymization. "Anonymizing is the changing of personal data in such a way that the individual information about personal or factual circumstances can no longer be assigned to a specific or identifiable natural person or can only be assigned to a specific or identifiable natural person with a disproportionate amount of time, costs and manpower" (Section 3 Para. 6, BDSG). There is also the option of pseudonymization. This describes "the processing of personal data in such a way that the data can no longer be assigned to a specific data subject without the use of additional information, provided that this additional information is kept separately and is subject to technical and organizational measures that ensure that the data is not can be assigned to the person concerned" (Section 46 Para. 5, BDSG new).

If there is a personal reference, research data may only be collected and further processed if the test subjects sign a declaration of consent.

Sample declarations of consent:
https://www.econstor.eu/dspace/bitstream/10419/97181/1/785263330.pdf

The area of data protection is a broad field. The TU Braunschweig has its own [Data protection officer](#), who is available to answer any questions about data protection.

Publication agreements with the repository

Repositories draw up user contracts/agreements for the publication/archiving of research data. It is important to read this carefully in advance and make sure that all the specified requirements are met.

Terms and conditions of LeoPARD:
https://leopard.tu-braunschweig.de/content/publish/contract.xml

Data post-use agreements

Access permission

It is advisable to make research data available as openly as possible. However, the data can also be embargoed (access delayed for a certain period of time) or restricted who has access to the data. Sensitive research data that, for example, contains personal information that cannot be completely anonymized would be a case for such restricted access.

[LeoPARD](#) offers the following variants for research data:

1. Publication (unrestricted access)

   With this variant there are no restrictions on access to the data. Anyone can view and download the data. By choosing a suitable usage license, you can determine how the data may be reused.

2. Embargo (delayed access)

   An embargo can be set for publishing the dataset. Until the embargo ends, the data set is locked and inaccessible to others. The data set will only be published after the self-selected embargo period has ended.

   The use of an embargo can be useful if the publication of research data is to be delayed, e.g. until the publication of an associated magazine article or for commercial reasons when registering a patent.
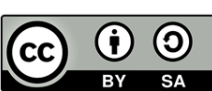
3. Publication (limited access)

   Different types of access conditions can be chosen. For example, LeoPARD makes it possible for research data to only be viewed by members of the TU Braunschweig. If that is not enough, limited access can also be selected, in which only the metadata of the data record can be viewed, but not the data record itself. If a researcher is interested, they can contact the author and the author will decide individually whether the researcher should have access to the data set. This can be helpful if legal or ethical reasons do not allow the data set to be made freely available. It is also possible to clarify with the interested researcher what he would like to use the data for.

4. Archiving
   Of course, it is also possible for the research data to be archived for at least 10 years in accordance with good scientific practice without being made public. In this case, only UB employees can view the data record in order to manage it.

Licenses for data

If you want to publish your research data or reuse third-party research data, the reuse rights must be clarified. Published research data is often provided with licenses that specify the conditions under which research data can be used without having to first ask the author of the data for permission. At the same time, they explain how subsequent use must be declared (citation note). They have established themselves for this Creative Commons Licenses, which also apply to research data from version 4.0.

License	abbreviation	Description	Open Access compliant
	CC0	No rights reserved, public domain	✓
	CC BY	Attribution: unrestricted further processing and commercial use permitted	✓
	CC BY-SA	Attribution + same license (SA = share alike): unrestricted further processing and commercial use under the same license	✓
	CC BY-NC	Attribution + non-commercial: unrestricted further use but not for commercial purposes is permitted	✗
	CC BY-NC-SA	Attribution + Non-Commercial + Same License: Unrestricted redistribution for non-commercial purposes is permitted under the same license	✗
	CC BY-ND	Attribution + no editing: commercial use is permitted, but no editing	✗
	CC BY-NC-ND	Attribution + non-commercial + no editing: Further processing and commercial use are not permitted	✗

DANGER: Once a license has been granted, it cannot be effectively withdrawn!

Licenses for software

There are separate open source software licenses for created software. Common licenses are:

GPL: GNU General Public License
LGPL: GNU Lesser General Public License
MIT: MIT license
Apache license
EUPL: European Union Public License

Detailed list of open source licenses:

https://ifross.github.io/ifrOSS/Licensecenter

Help choosing the right license:

https://creativecommons.org/choose/?lang=de
https://choosealicense.com/

5) Publish

Have all of the points mentioned been considered and implemented? Then nothing stands in the way of archiving or publishing your research data. The data is submitted either independently via Shibboleth and the TU Braunschweig identifier Leopard (Instructions for self-submission) or an employee of the research data team (forschungsdaten@tu-braunschweig.de). We will be happy to support you (e.g. with larger amounts of data). In both cases, you first sign the application form, which you send digitally to the research data team (forschungsdaten@tu-braunschweig.de) can send.

After the data has been submitted, it will be checked by the UB staff and if everything is in order, the research data will be archived or provided with a DOI and published.