



Data Management Guidelines: How To Handle Data

January 19, 2026

Norbert Riefler

Preface

‘On July 20, 1969, Neil Armstrong climbed out of his spacecraft and placed his feet on the moon. The landing was broadcast live all over the world and was a significant event in both scientific and human History. Today, we can still watch the grainy video of the moon landing but what we cannot do is watch the original, higher quality footage or examine some of the data from this mission. This is because much of the data from early space exploration is lost forever. (...) The tapes were likely wiped and reused for data storage sometime in the 1970s.’

- Kristin Briney[5]

This document holds for all employees of the Martian Mind Excellence Cluster and serves as a reference for how to save data. It delivers reasons for the question of why one should think about data management and gives instructions and examples. Data management in the information age is an absolutely required skill, and a base to cope with the wealth of data and to avoid an information overload. It helps organizing everyone’s own data, but it is also mandatory for data sharing so that others can understand the data structure and are able to interpret the information derived from your data. In this way, your data meet the FAIR standard (Findable, Accessible, Interoperable and Reusable), basis for the discipline of data science. Data sharing and re-usability are crucial points for our Cluster. They are prerequisites for the seamless data flow across projects and research areas.

There are many topics in data management like data governance and security or storage management, but we are treating here only a condensed extract, tailored to data handling of research data.

Contents

1	Why Data Management	3
2	File Organization	4
2.1	Different Data Types and Storage Strategy of Large Data	4
2.2	File and Folder Naming	5
2.3	File Formats	6
2.4	Folder Hierarchy	7
2.4.1	Daily Data	8
2.4.2	Data for Publication	9
3	Documentation and Metadata	11
3.1	Important things to do while you collect or create your data	12
3.2	Documentation of general data in the DataCite-Schema	12
3.3	Generation of a metadata scheme	13
4	Electronic Lab Notebook	14
4.1	General information	14
4.2	EXC standard for documentation and structure	14
5	Git Software repository	15
6	Reference Management Software	16
7	Data and Knowledge Publication Considerations	16
8	Further Tools for Data Management	16
	Bibliography	17
A	Appendix	18
A.1	How to Store Papers – ‘Dummy Paper’	18
A.2	How to Write a DMP	18
A.2.1	Motivation	18
A.2.2	Elements of a DMP	19
A.3	Glossary	22

1 Why Data Management

Data management is a holistic process including design of data structures, data collection, storage and accessibility of data. This guideline presents instructions and tools to perform data management tailored to the needs of our research cluster. The goal is to fulfill the FAIR principles. FAIR (Findable, Accessible, Interoperable, Reusable) is a label for data which are findable in the Web, downloadable, and can be used for your own purposes. This means for example that data from measurements must be clearly understandable (description of the used device with all adjustments, results with units as ASCII files, etc.) and data from simulations must be reproducible (source code, mesh, initial and boundary conditions, etc.). This method also guarantees the fulfillment of Good Scientific Practice. Further benefits of good data management are:

Save time

Planning your data management will save you time and resources [8], e.g., by faster locating data. And well managed data requires less preparation for sharing or writing a paper.

Preserve your data

Depositing your data in a file server (repository) safeguards your investment of time and resources while preserving your research contribution for you and others to use.

Increase your research impact

Making your data available to other researchers can impact discovery and relevance of your research.

Maintain data integrity

Managing and documenting your data throughout its life cycle will allow you and others to understand and use your data in the future.

Meet grant requirements

Many funding bodies now require that researchers deposit data collected as part of a research project, and the DFG demands to share research data as long as privacy and owner rights are not violated.

Promote new discoveries

Sharing your data with other researchers can lead to new and unanticipated discoveries and provide research material for those with little or no funding.

Support open access

Be a catalyst for research and discovery. Show your support for open access and open data by sharing your data (see chapter 7 for issues with sharing data).

This document is a living document because data management is still under development, and due to the change of framing condition. Major revisions are communicated and announced to the Cluster members.

2 File Organization

Access to a file server is provided by your local research group or by your IT administration. The kind of usage or rules may differ, so please ask there for further information.

2.1 Different Data Types and Storage Strategy of Large Data

Depending on the origin of the data, we distinguish between two data types [7]:

- Primary data (sometimes called raw data) is generated by measurement devices, sensors/cameras, simulation programs, observations etc. and are untreated; it is collected for the first time by the researcher.
- Secondary (or analyzed) data are processed primary data. (In social science, secondary data are already collected or produced data by others).

Data from projects are stored in the according project directory (see section 2.4). Every experiment and simulation documented in the Electronic Laboratory Notebook gets a report as a PDF protocol which are also saved in this directory. Very large primary data cannot be stored in the ELN. However, the origin and the storage location (e.g. the precise location on the file server) of these data can be documented there. The rules are:

- data < 100 MByte is stored and documented on the ELN
- all data > 100 MByte is stored on the file server and documented in the ELN.

High-speed video movies can be highly compressed and stored as a mp4-file and uploaded to the ELN instead of the primary data; same for large simulations.

Large amount of data needs specific care (more can be found for instance in Hart et al. [4] as **Rule 10**). Depending on the amount of the data, there are different storage strategies shown in figure 1.

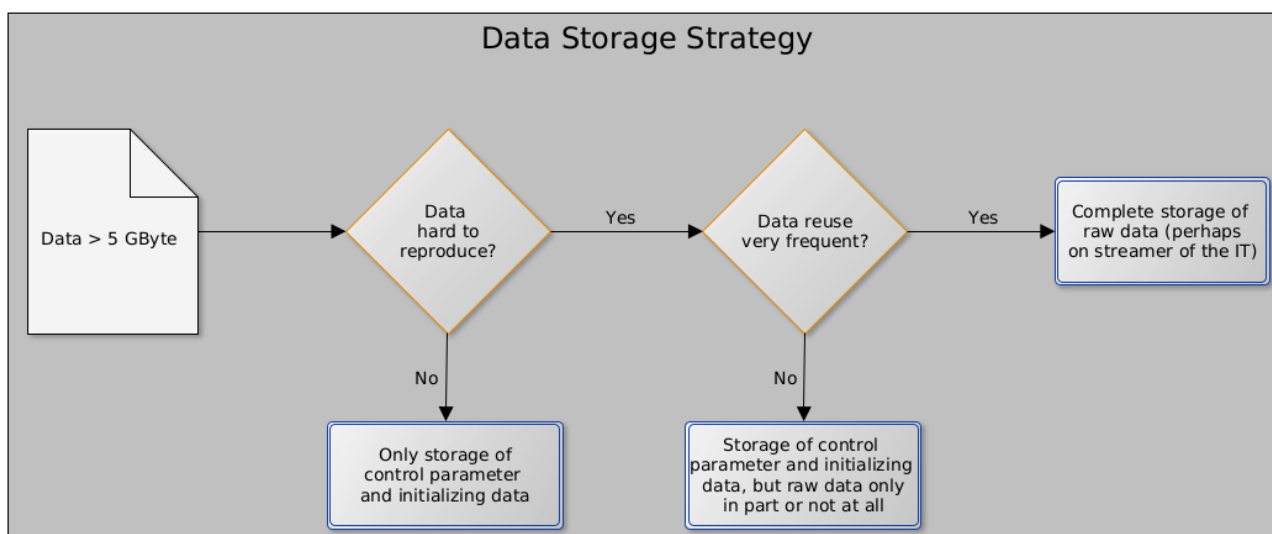


Figure 1: Pfade der Speicher-Strategie

Most important for the strategy is to consider the expense of the reproduction of the data. Simulations can be easily repeated as long as all input parameter, the complete software package and the underlying operating system is stored. However, complex experiments cannot be reproduced that easy. Therefore, large amount of data from experiments are in principle assessed to be more valuable as data from simulations. The generation of data for a single transient simulation using High Performance Computing (HPC), e.g. using the NHR (Nationale Hochleistungsrechner), often exceeds 1 TByte. The storage of these data should be in accordance to the local IT of the group either on a fileserver or an USB hard drives, and the content of the data is documented on the ELN.

The storage period of data is typically limited to 10 years. However, data of publications (sub-directory 'Publications' under the project directory) persists on a Cluster server unless they are stored on a repository (for instance on zenodo.org), with a remark and link to that repository to retrieve the data.

Security Note: It is absolutely necessary to **check all data from external sources** (e.g. from an USB stick) manually against computer viruses before usage or storage of the data. Above that, data loss is common on USB sticks if they are not used over a longer time span, so they should only be used for short term data storage.

2.2 File and Folder Naming

Spend time planning out file naming conventions in the beginning of a project. How do you or others will look for and access files at a later date? Do you think about them by type, location, study or something else?

Naming conventions:

- File names are self-explanatory - they tell what data is in!
- Do not use spaces or special characters - use CamelCase and hyphen '-' or underscore '_' as separator:
 - NO** 'name date v1.txt'
 - YES** 'name_date_v01.txt'
 - YES** 'nameDateV01.txt'
 - YES** 'simpleFoam_u0.2m-s_etha2mPas_v01'
- Develop a file naming scheme that includes information about the data. Example:
[Date]_[Run]_[SampleType]
- Consider sorting when deciding what element of the file name will go first:
 - Use 'YYYY-MM-DD' to save dates.
 - Use leading zeros:
 - NO** 'ProjID_v1.csv' ... 'ProjID_v11.csv'
 - YES** 'ProjID_v01.csv' ... 'ProjID_v11.csv' (sequence up to 99)
 - YES** 'ProjID_v001.csv' ... 'ProjID_v111.csv' (sequence up to 999)

2.3 File Formats

As technologies change, researchers should plan for both hardware and software obsolescence and consider the longevity of their file format choices to ensure long term readability and access.

File formats more likely to be accessible in the future have the following characteristics:

- Non-proprietary
- Open, documented standard
- Common usage by research community
- Standard representation (ASCII, Unicode)
- Unencrypted
- Uncompressed (if not too large...)

Examples of **preferred** file format choices include:

- Open Document Files: .odt (Text), .ods (Spreadsheet), .odp (Presentation); primary format of OpenOffice/Libre Office, standardized in ISO 26300
- Office Open XML: .docx (Text), .xlsx (Spreadsheet), .pptx (Presentation); primary format of Microsoft Office, standardized in ISO 29500
- Rich Text Format: .rtf; proprietary, but well documented format
- Plain Text or Markup Language: .txt, .md (Markdown), .html (Hypertext), .tex (L^AT_EX), .csv (comma separated values)
- Portable Document Format: .pdf; standardized in ISO 32000
- Bibliographic Data: .enl (EndNote; this is the standard software for literature management at the IWT); otherwise, please use human readable formats such as .bib (BibTex), .ris (Research Information System Format)
- Video files in data formats .mp4 and .mpg generated by codecs H.264 and H.265; standardized format (ISO 14496-10) for video compression; **DO NOT USE proprietary codecs such as Quicktime (.mov, .qt)**
- Images: .jpg is the standard file format of most cameras, it is readable by almost any image-related tools, but is not ideal in all cases (e.g. line art). It offers no transparency. Moreover, it is a lossy format which can lead to a permanent degradation of the image, according to the chosen compression. Thus it should be avoided if this is not tolerable. The succession format JPEG2000 .jp2 allows lossless compression and thus may be used for image archiving. Due to licensing issues it is not widely used. The .tiff (Tagged

Image File Format) is an important format for file exchange with publishers. It allows lossless compression and high depth (32 bit) as well as CMYK color. It may be used for image archiving, however, the file size remains comparably large. The 'Portable Network Graphic' format .png is a raster image format with good lossless compression. It yields larger file sizes than .jpg for photos but is very well suited for line art.

- Drawings: It is preferable to use vector based file formats for sketches and drawings, as these allow for an unlimited scaling of the image without affecting the quality. For large printouts, vector graphics are significantly smaller than raster graphics. However, vector-based images are not handled well by all word processors and even if they are usable, may yield varying results on different computers. It is recommended to prepare a vector graphic in an open format (.svg, scalable vector graphics) to ensure long-term readability and it export it to the file format best suited for your word processing software (e.g. .emf for Word, .eps or .pdf for L^AT_EX).
- Binary data: sometimes, it is most efficient to save data in a binary format, as text-based representations would dramatically increase the file size. In this case, it is vital to document how the data is stored and preferably give an example of how it can be read again.
- Large data sets can be saved also as an HDF (Hierarchical Data Format).

If you consider exporting your (measurement) data into a format with the above characteristics, keep a copy in the original software format unless you can assure that all data and metadata is correctly converted. If you deposit your data in a repository, your files may be migrated to newer formats, so that they're usable to future researchers.

2.4 Folder Hierarchy

The specific value of data has a wide variance. Many measurements like those from preliminary experiments were very important to design your experiments, but they are more or less worthless for publications. In contrast, some of your data are highly valuable and find their way into a publication. Due to these differences, the location where you save your data has different valence, according to these three data kinds:

- I. Data with a low valence, see: 2.4.1 Daily Data
- II. Data which are going to be published, see: 2.4.2 Data for Publication
- III. Data of accepted publications: These data are stored on a special area within your file-server where you can only write your data one time – no one should be able to change or delete anything there. It is like a mail box: Once you have thrown in your tax declaration, it is over. This data area is for accepted publications only with all data stored understandable for others!

2.4.1 Daily Data

Every day you generate data, for instance by measuring, simulations, preparing a paper etc. A local fileserver in your research group serves as a file storage for your data, and all these daily data can be for instance saved under an Employees directory and your <family name>:

/<local fileserver>/<department>/Employees/<family name>/

The names enclosed by <department> could be the family name of your PI or a departments, for instance:

- IWT-WT-WB (Wärmebehandlung)
- IWT-VT-RST (Reaktive SprühTechnik)
- ZARM-MaMBA (Moon and Mars Base Analog)

according to your affiliation. An example directory is:

/VT-Server/RST/Mitarbeiter/riefler/

Here, you save the following data:

→ **For Experimental Data:**

- All experiments of the project
- Relation between these experiments (which parameters are varied) in a JSON based text file, see subsection 3.2
- intermediate results
- List of all used devices with adjustment parameters

→ **Experimental Setups:**

- Drawings of all parts
- A list of all required parts and devices

→ **For Simulation Data:**

- List of the used software together with used source code
- All required input files to reproduce simulations
- The required parameters, initial and boundary conditions and mesh information to recalculate all (in case these are not specified in the input files) simulations of the project
- Relationship between these simulations (which parameters are varied) in a JSON based text file, see subsection 3.2

There is no prespecified directory structure for your daily data, so that's up to you. Note that these data under your name are distinct from project data like, e.g., important experimental findings or simulations which are published, which are described in the following.

2.4.2 Data for Publication

Here, you save all data relevant for every publication of the project. Depending on the size of the data, and the decision within your department (see above), primary data files larger than, say, 10 GByte must not be saved within the directory specified here. Instead, a reference has to be given within the documentation in the experiment description on the ELN about the corresponding origin (e.g. high-speed camera measurement or CFD simulation) and where exactly the primary data are stored. The decisive **basic principle** for all your raw and primary data is: **All data must be documented understandably regarding their origin!**

When you prepare a document, say a presentation or a project report, the data are saved in the kind of this exemplary directory:

```
/VT-Server/RST/Projekte/DFG/<project name>
```

→ Resulting Documents:

- Documents from your scientific work: papers, dissertations, presentations, reports, proposals, etc.
- All results (figures, tables, movies, animations, ...) and the underlying measured or simulated data together with evaluation programs, source code and input files

The directory name for published works in a project, the <project name>, is compound on:

```
<year of project start>_<department>_<project handle>_<family name>
```

See subsection A.1 for an example.

Papers and Dissertations

In agreement with all EXC PIs, the directory naming scheme for published data of the File-Server is based on departments, the funding source and then the corresponding project name (see above for <project name> definition):

Text and data of a paper are saved in a directory (please adjust all the given examples):

```
/VT-Server/<department>/Projekte/<funding>/<project name>/Papers/
```

with a directory name like that:

```
<year>_<department>_<paper-name>_<first author>
```

Text and data of a dissertation are saved in this kind:

```
/VT-Server/<department>/Projekte/<funding>/<project name>/Dissertation/
```

with a directory name like that:

```
<year>_<department>_<dissertation_title>_<author>
```

In these directories, you save all the data of your **paper** in subdirectories, whose names are given in the left row of table 1, with the corresponding files explained on the right.

In case of a **dissertation**, you save your presentation of the doctoral examination together with the manuscript under the directory '01_Manuscript+Presentation'. If you do not have supplementary information for your paper/dissertation, or you did not write specific computer

Table 1: The data of every paper has to be saved in eight subdirectories; further remarks:

*When the paper is ready for press the corresponding author loads the last and revised version (i.e. only the real used but absolutely complete data) into the protected directory

**If the data reproduced in a figure/table is from distributed primary data directories, it is sufficient to save here only the generating program

00_FinalPublication*	- The very last version of your paper as a pdf with volume, year and page numbers on a dedicated server directory - EndNote *.ris ASCII export fulfilling the FAIR principles
01_Manuscript	the *.pdf and *.docx or *.tex
02_Figures**	one sub-directory for every figure (figure_01, ...) with data points as *.csv or *.txt, or the generating file/program for the figure points (Matlab, Python, Origin,...); the results are from data stored in or extracted from the 04_PrimaryData directory
03_Tables**	one sub-directory for every table (table_01, ...) with data stored in as a csv-file
04_PrimaryData	- measured or simulated data, each sub-directory with a 'readme.json' file describing the data - ELN pdf pages with QR-code (see below)
05_ProgramSources	- e.g. OpenFOAM-Solver, Python/Matlab code, etc. - case files (mesh) for OpenFOAM, Fluent, ABAQUS etc.
06_References	- the references *.bib or *.enl together with *.ris - Cited papers/books (if publication rights are permitted)
07_Supplement	Supplementary files
08_Misc	Reviewer comments / rebuttals

programs for your research, the '05_ProgramSources', '07_Supplement' and '08_Misc' directories must still be there but with no entries. Changes on the manuscript due to the revision process can be understood by subdirectories "01_Manuscript/01_Submission" and "01_Manuscript/02_Revision" together with storage of the respective manuscript and rebuttal files.

An important directory is the '02_Figures'. There, you save every figure in the manuscript in good quality, and you also save a *.csv file with the data in the figure, or you save a file which generates the corresponding figure – for instance a Matlab or Python file – from the primary data in the '04_RawData' directory. This directory includes the measured or simulated data in *.txt or *.csv format from the measuring device or the simulation software, and a 'readme.json' file generated by the Readme-File-Creator described below with information and metadata about your measured / simulated data. The relation of each data point in a figure to the primary data must be given either implicit, for instance in the Matlab file, or better explicit in the 'readme.json' file. Don't forget: Figures and tables have to be reordered if there are changes in your manuscript. To complete your publication, every cited paper should be added to the directory

'06_References' for internal use. All too often the ideas of a paper cannot be tracked and reproduced because there are references which you can't get because your institution do not offer access to that specific journal, or the reference are from a master thesis which is not available. Therefore, saving all references with suitable file names for a direct relation – together with all the previously mentioned data and information – presents your paper in its entirety. Writing a paper is a multistep process. With the decision of the appropriate journal, the process comes to a crucial point where all the data have to be saved in the structure given above and in the following screenshot, and the review process starts. If the paper is accepted, all changes are integrated and the final pdf with volume, year and page numbers is released, then the last actions are:

- Do not save more than the really required data - but also not too less.
- The numbering of figures and tables in the manuscript must agree with that in the directories (please have a look at the Dummy Paper). Tables without a relation to measured or simulated data do not need an entry in the '03_Tables' directory.
- Check if every ELN-file which describes the experiment/simulation is saved.
- The very last step: Make a copy of the complete paper data into the directory 'Published' on a dedicated file server (tbd.). It's a special directory: You can copy something there, but afterwards you cannot change anything. It's like a post box: Once you have thrown something in, it is gone!

During the review process of publications, the order of figures and tables are often changed and, therefore, have to be renamed accordingly. A Python program ('ImageSorter' for Word or $\text{\LaTeX}$), may be used to simplify manuscript rearrangements and can be sent to you upon request.

Posters and Presentations

Posters or presentations are saved with the same scheme given above except that the directory name '01_Manuscript' is replaced by '01_Poster+Abstract'.

3 Documentation and Metadata

Metadata are data about data and they describe data in more detail. A pressure value without any context has no meaning; it requires information about the system (e.g. which fluid), the location, the pressure sensor (company and model), the used measurement device, etc.

Data documentation and the generation of metadata is essential to understand your data in detail, and also helps other researchers find, use and properly cite your data. With this, data are understandable and reusable, important for FAIR data requirements (see chapter 7).

Various and particular metadata standards are available. Every discipline has own domain specific schemas and you have to ask for or find them by yourself. If there are no schema

for your domain or application, you may develop your own metadata schema. Further general guidelines are provided below.

3.1 Important things to do while you collect or create your data

- Make a note of all file names and formats associated with the project, how the data is organized, how the data was generated (including any equipment or software used), and information about how the data has been altered or processed.
- Note version and producer of measurement devices and computer programs (also of the control programs of measurement devices).
- Include an explanation of codes, abbreviations, or variables used in the data or in the file naming structure.

3.2 Documentation of general data in the DataCite-Schema

The used metadata scheme follows the recommendation of the world wide DataCite consortium (TIB Hannover, British Library, Denmark Technical Information Center, ETH Zurich, CalTech, Australian National Data Service, ...). These data are stored in a 'readme.json' file in **EVERY DIRECTORY** of your data! Please, use our 'Readme File Creator' described below to generate this file. Explanation of some extracted properties are given in the following; please refer to Table 3 in [3] for more information:

ID 1: Identifier

Number used to identify the data. Can be a DOI (Digital Object Identifier) in case of a paper, or an ID from the ELN (eLabFTW generates a long number for every experiment, see below)

ID 2: Creator

Name of the person responsible for the described data (ORCID).

ID 3: Title

May be the title of a dataset, or the name of a piece of software, or the title of a paper.

Figure 2: Data Input Tool: Readme-File-Creator

ID 4: Publisher

The name of the entity that produces, holds, archives, publishes prints, distributes, releases or issues the resource (e.g. 'Leibniz Institute for Materials Engineering IWT').

ID 5: Publication Year

The year when the data was or will be made publicly available.

ID 8: Date

Key date associated with the data, e.g. project start, end date, data modification, data release date, or time period covered by the data

ID 9: Language

Language(s) of the intellectual content of the resource, when applicable

ID 10: Resource Type

A description of the resource; mostly 'DataSet', but may be 'Audiovisual', 'High-Speed Images', 'DataPaper', etc.

ID 19: Funding Reference

Organizations or agencies who funded the research

ID 16: Rights

Any known intellectual property rights held for the data. The best choice is either the Creative Commons license 'CC-BY' (BY attribution) or 'CC-0' for data and 'CC Version 4.0' for written documents.

ID 17: Description

How the data was generated, including equipment or software used, experimental protocol, other things you might include in a lab notebook

ID 18: Geolocation

When the data relates to a physical location, record information about its spatial coverage

The tool displayed in figure 2 can be get after request for every operating system to simplify metadata entry (figure on the right). You can create automatically a file called 'readme.json' with your metadata, embedded in a JSON (Java Script Object Notation) format, which is used as documentation as well as metadata for Data Science Methods. This file is stored in every directory which contains data or program code. Field entries can be empty if they are not applicable.

The same DataCite metadata structure can be included into the ELN (see chapter 4).

3.3 Generation of a metadata scheme

A schema is a logical plan showing the relationships between metadata elements, normally through establishing rules for the use and management of metadata. An example of the development of a metadata schema is given in the paper of Elberskirch et al. [2]. There, all relevant parameters are collected to describe nanoparticles physically as well as biologically (toxico-

logically) in a comprehensive manner which yield a metadata scheme. The development of suitable scheme for your specific domain can be helpful for structuring your data as well as your knowledge.

4 Electronic Lab Notebook (ELN)

4.1 General information

Electronic Lab Notebooks (ELNs) enable researchers to organize and store experimental procedures, protocols, notes and data using their computer or mobile device. ELNs can offer several advantages over the traditional paper notebook in documenting research during the active phase of a project, including searchability within and across notebooks, secure storage with multiple redundancies, remote access to notebooks, and the ability to easily share notebooks among team members and collaborators.

eLabFTW  is an OpenSource, generic, browser based ELN, developed/initialized by Nicolas Carpi (Institut Pasteur, Paris) 2012. Data are stored in MySQL/MariaDB. It is maintained by many developers and is used worldwide (Berkeley, Indian Institute of Technology, KIT, ...).

- Access: t.b.a
- Free definable status of the experiments ('finished', 'running', ...)
- Definition of templates for a step-by-step procedure description of measurements
- Every experiment gets a unique ID and can be summarized easily as a pdf
- Free definable categories/tags
- Graphical text editor to describe your experiments and simulations
- File attachments of data with preview of common formats (pdf, tiff, png, ...)
- Linking of experiments
- Time stamp service (RFC 3161, e.g. DFN)
- Data import/export (csv, zip, json, ...)
- Access using, e.g., Python via an Application Programming Interface (API)

4.2 EXC standard for documentation and structure

To document scientific findings clearly and distinctly and trace them easily even after many years, the title and tags of an experiment or a simulation consists on a system shown in table 2. The information for "material" or "process", "method" and "project" or "project abbreviation" are absolutely imperative. "Material" is seen as an extensible term and can be physical substances

used in experiments and simulations as well. You may also include more entries if required. The exactly same systematics are also applied for the tags (i.e. markers). The reason for this redundancy is the SQL database of eLabFTW which is much faster searching for tags.

tags	description	examples
mandatory information:		
material or process designation	DIN norm + material number; composition in mass-%; collaboration survey	20MnCr5, TiO ₂
project	project abbreviation, number, forerun	EXC-RA3-Tracht
methods	analysing method, test method, process, simulation method	XRD, tensile test, UV-vis, OpenFOAM
optional information:		
device/system	label/name	Shimadzu UV-2600
sample label, sample number	for unique classification	01-10036-22, E0165
base material, material type	chemical elements written-out	Iron, ceramic
department	abbreviation like that of the team	ZARM-MaMBA, IWT-VT-RST

Table 2: Tagging system with the always available tags for every experiment or simulation and the optional tags, which can be completely different according to the research domain.

You have to create a pdf for every experiment, every simulation or every survey simply by a click ('Make a pdf') in the edit mode (see '04_PrimaryData' in Table 1). This kind of documentation includes a unique QR-code and is required for data which is published! The QR-code enables a direct link to the data together with unique ID.

The **maximum size** of uploaded data is limited to **100Mbyte per file**. Data can be uploaded either via the Web browser or an API. A template for the JSON fields available in every experiment with entries for the metadata (see chapter 3) can be sent to you after request.

5 Git Software repository

Git  is a web-based tool that provides a repository for the version control. It is used to manage, plan, create, verify and monitor your software development. Furthermore, it can be used as a document versioning system. For example, during the creation of a larger piece of code on writing (a dissertation, papers, proposals, ...), you can access all the previous versions.

With that, every changed line of code or sentence can be tracked and reconstructed, provided that changes are committed to the system..

A Git instance for the EXC will be launched soon.

6 Reference Management Software

We do not have one central software for managing our scientific literature. But we need to document all of our papers as a result of our research. Therefore, we will collect BibT_EX entries for every publication of your project (either paper or presentation etc.) and beg you to provide a BibT_EX file (perhaps as an export of your used program like, e.g., EndNote).

7 Data and Knowledge Publication Considerations

For all projects funded by public sponsors there is the principle that all gained data and developed methods have to be provided to the public. Therefore, the directory structure of publications described above can be uploaded to zenodo <https://zenodo.org/>, a repository of scientific data, where you get a DOI for your data.

The reuse of your data by others in the sense of the FAIR principles requires the link to a licence (without a statement the default term is "all rights reserved" which only allows viewing but no further usage). Therefore, common licences are the 'CC-BY' licence for text and the '3-clause BSD' license (or 'BSD-3') for software.

The FAIR principles belong to Open Data and is one part within the framework of Open Science, with Open Access of scientific papers and Open Source of software as the three main components.

However, data publication underlies at the same time the restriction of scientific responsibility concerning security relevant research [9]. The freedom of research is certainly granted by law, but free and transparent research bears the risk of direct misuse by third parties for harmful purposes, in particular in military technology. Therefore, a risk analysis should be performed in an early state of the research process. In cases with possible risk for life, health or human dignity, every researcher has to decide under a trade-off to the chances.

Last, sensitive data are data with a relation to personal information. Data privacy rules are described on the web page of the University Bremen¹. Information about data protection means can be found on the web page of the Research Data Centre Qualiservice².

8 Further Tools for Data Management

- Search for Data Repositories: <https://www.re3data.org/>

¹<https://www.uni-bremen.de/informations-und-serviceportal-zu-den-themen-datenschutzrecht-und-informationssicherheit/datenschutz>

²<https://www.qualiservice.org/en/the-helpdesk/data-protection.html>

- <https://zenodo.org/> → scientific data, publications, reports, presentations, videos, etc.
- <https://www.ukdataservice.ac.uk/manage-data/format/recommended-formats>
→ list of recommended data formats
- <http://rd-alliance.github.io/metadata-directory/standards/> → Metadata Discipline Standard Formats
- Renaming Tools: PSRenamer, ExifToolGUI

References

- [1] Deutsche Forschungsgemeinschaft (DFG). *Umgang mit Forschungsdaten*. 2021. URL: https://www.dfg.de/foerderung/grundlagen_rahmenbedingungen/forschungsdaten/index.html.
- [2] Linda Elberskirch et al. "Digital research data: from analysis of existing standards to a scientific foundation for a modular metadata schema in nanosafety." In: *Part. Fibre Toxicol.* 19 (2022), pp. 1–1-19.
- [3] DataCite Metadata Working Group. *DataCite Metadata Schema Documentation for the Publication and Citation of Research Data*. Version 4.3. 2019. URL: <https://doi.org/10.14454/7xq3-zf69>.
- [4] Edmund M. Hart et al. "Ten Simple Rules for Digital Data Storage." In: *PLOS Comput. Biol.* 12 (2016), e1005097-1–12.
- [5] Kristin Briney. *Data Management for Researchers*. Pelagic Publishing, 2015.
- [6] Leibniz Universität Hannover. *Leitfaden zur Erstellung eines Datenmanagementplans*. Version 2.3. Aug. 18, 2020. URL: https://www.fdm.uni-hannover.de/fileadmin/fdm/Dokumente/Leitfaden_DMP_LUH_v2.3.pdf.
- [7] Lisa Gitelman. *"Raw Data" Is an Oxymoron*. MIT Press, 2013.
- [8] Alicia Martinez-Garcia et al. "FAIR principles to improve the impact on health research management outcomes." In: *Heliyon* 9 (2023), e15733-1–10.
- [9] Nationale Akademie der Wissenschaften Leopoldina und Deutsche Forschungsgemeinschaft. *Scientific Freedom and Scientific Responsibility – Recommendations for Handling of Security-Relevant Research*, 2022.

A Appendix

A.1 How to Store Papers – ‘Dummy Paper’

The following screenshot shows a complete data structure of a dummy paper. You may find this dummy paper on our exchange platform (t.b.a). Basically, the idea is that every data in the paper, whether it is a figure or a table, can be related to the original source of the measurement or the simulation. This is realized in the Dummy Paper by Matlab scripts, but can be done using other means as well (Python, Excel). The point is: No matter how, but you have to give a relation between the data saved in the '04_PrimaryData' directory and their representation in the publication. All figures and their generating scripts – here Matlab files – are given together with the measurement data so that running a Matlab (Octave) script within that directory generates the according figure. In the end, someone else can reproduce your results given in your publication. In case of other data like SEM images for instance, you can refer to them by a PDF generated by the ELN including a link to the experiment stored there.

The ‘Dummy Paper’ contains a Word and a $\text{\LaTeX}$ template in '01_Manuscript'. The EndNote or BibTeX literature references are in '06_References'. Both, the EndNote file '*.enl' or the BibTeX file '*.bib', should include links between the references and the freely accessible PDFs of the papers and books also saved in '06_References'.

Name	Date Modified
00_FinalPublishedPaper	2020-11-19 16:35:53
Riefler-ImpedanceCharacterization+CoupledPiezoTubeFluidSystem+...	2020-12-16 15:06:19
01_Manuscript	2023-10-10 12:53:26
01_Submission	2023-10-10 12:53:26
DummyPaper.docx	2020-12-16 15:06:20
DummyPaper.tex	2021-04-26 09:07:26
DummyPaper.tex.pdf	2020-12-21 15:19:27
DummyPaperWord-Export.pdf	2020-12-21 15:21:24
02_Revision	2023-10-10 12:53:16
02_Figures	2020-11-27 10:35:31
Figure0_Format_template.m	2020-12-16 15:06:23
Figure1_from_Literature.png	2020-12-16 15:06:23
Figure2_BesselFcts.m	2020-12-16 15:06:26
Figure2_BesselFcts.png	2020-12-16 15:06:24
Figure3_Sketch.png	2020-12-16 15:06:25
Figure3_Sketch.pptx	2020-12-16 15:06:25
Figure4_Resistance_AirHeptane.m	2021-06-17 13:35:52
Figure4_Resistance_AirHeptane.png	2020-12-16 15:06:26
Figure5_FluidFlow_SpherePacking_StreamLines.eps	2020-12-16 15:06:26
Figure5_FluidFlow_SpherePacking_StreamLines.png	2020-12-16 15:06:26
Figure6_REM+Particles+FiberFilter.png	2020-12-16 15:06:27
03_Tables	2021-06-17 13:23:27
Table2_Bessel_Zeros_fct.m	2020-12-16 15:06:27
Table2_RadialResonanceFrequencies.m	2021-06-17 13:23:53
04_PrimaryData	2022-04-06 09:36:18
Figure4_Resistance_AirHeptane	2021-06-17 13:36:28
MeasuringSeries-2018-05-18	2021-06-17 13:32:32
20190506 - Droplet-Generator-Impedance-Measurements.pdf	2020-12-16 15:06:28
readme.json	2020-12-16 15:06:29
Figure5_FluidFlow_SpherePacking_StreamLines	2020-11-27 13:26:29
Figure6_REM+Particles+FiberFilter	2020-11-27 13:08:23
Figure7_Data_External	2020-11-27 13:24:15
Link to eLabFTW-json-template.json	2022-02-25 10:29:04
05_ProgramSources	2020-11-27 09:48:19
Figure5_FluidFlow_SpherePacking_StreamLines	2020-11-27 10:25:16
Table2_ChebyshevFunctionToolbox-2017	2020-11-27 16:47:47
06_References	2021-04-27 11:01:19
EndNoteLibrary.Data	2020-12-21 13:44:10
Bird.-Transport.Phenomena.-Wiley.2001.pdf	2020-12-16 15:06:46
EndNoteLibrary.enl	2020-12-21 13:39:15
Literatur.ris	2020-12-16 15:06:21
Literature.bib	2020-12-21 14:55:49
Literature.log	2020-12-21 14:10:57
07_Supplement	2020-10-29 12:42:37
08_Misc	2020-11-19 13:12:09
ReviewerResponse+Rebuttal-year-month-day.docx	2020-12-16 15:06:34

Figure 3: Complete data structure of a dummy paper.

A.2 How to Write a DMP

A.2.1 Motivation

The handling of the data, created during a project, contains several issues which are described in a Data Management Plan (DMP). A DMP contains structured information about the research

process of the corresponding project. These information should to be estimated before a project is started. That means you should think about how you will treat the expected data. The following topics belongs to typical DMPs and might be of interest to you; more details can be found in [1, 6]:

1. Administrative information: project name, time period.
2. Methods of data generation: simulations, experiments (devices), data size, kind of data documentation.
3. Data safety: location, interval and capacity of storage device, who gets access.
4. Archiving: which data are archived where with what kinds of metadata.
5. Data sharing: which repository, license condition, required metadata.
6. Resources and responsibility: who is responsible for processes, IT, defining defaults and formats, monitoring; required personal resources; costs.

Please see [1, 6] and the Appendix for more details; some exemplary DMPs can be found here:

https://www.cms.hu-berlin.de/de/dl/dataman/arbeiten/dmp_erstellen

A Data Management Plan (DMP) documents the thoughts of an applicant about the data which will be generated during the applied research project. Research funders (DFG etc.) expects that publicly promoted projects makes their data accessible to the public, where a simple publication was sufficient in the past.

There is no unified structure of DMPs due to the vast number of different research tasks. In the following, the main elements of a DMP from the TIB Hannover are listed, which can be found also in the recommendation of the DFG:

(https://www.dfg.de/download/pdf/foerderung/grundlagen_dfg_foerderung/forschungsdaten/forschungsdaten_checkliste_de.pdf)

A.2.2 Elements of a DMP

0) Administrative information

- Project name
- Project participants
- Project description
- Project ground (doctorate, third party funding, etc.)
- Project duration
- Version of this DMP

1) Methods and kinds of data gathering

- Are there primary data generated, or will be secondary data used?

- Which data types / formats are generated and processed?
- How large are the data?
- Which equipment (instruments, hardware, software, etc.) will be used?
- How are the data organized? (Directory of file oriented structure? Version control?)
- How is the research process and the data been documented?
- Which (technical) standards will be used for description/documentation (metadata, classification)?
- How are the metadata been generated (e.g. automatically, manually, after a guideline, self-defined)?

2) Backup and data safety

- Where are the data stored?
- Which capacity is required?
- Interval of data safety?
- Are there any protective measures required for sensitive data?
- Are there any third party of project partners (e.g. in joint projects) who need access to the data?

3) Archiving

- Which data is archived?
- On which data carrier?
- Are there any requirements for the operators of the infrastructure? E.g. data curation?
- Which metadata have to be provided to find the archived data?
- Which information is additionally required to understand the context of the data?
- How long are the data archived?
- Are there any legal technicalities for the data archiving?
- What are the costs for which service?

4) Data sharing and publication

- Are there data which must be shared with others?
- What are the possible systems for data sharing?
- Which metadata and documentation are additionally required for third parties to use the data?
- Where (data repository, data journal) will the data be published, and how (e.g. open access, with embargo time, restricted access)?

- What are the license condition for the published data? (E.g. 'CC BY 4.0')

5) Resources and responsibilities

- How is the distribution of responsibility governed in this project?
- How is responsible for the data management (processes, IT, guidelines, formats, monitoring)?
- Required personal resources for a successful implementation/realization?
- What are the costs within the project phase and possibly thereafter?
- Which infrastructure resources are required, with additionally costs?

A.3 Glossary

AiF	...	Arbeitsgemeinschaft industrieller Forschungsvereinigungen
ASCII	...	American Standard Code for Information Interchange
BMFTR	...	Bundesministerium für Forschung, Technologie und Raumfahrt
BSD	...	Berkeley Software Distribution
CC-BY	...	Creative Commons - BY attribution
DFG	...	Deutsche Forschungs-Gemeinschaft
DMP	...	Data Management Plan
ECR	...	Early Career Researcher
ELN	...	Elektronic Laboratory Notebook
FAIR	...	Findable Accessible Interoperable Reusable
GWP	...	Gute Wissenschaftliche Praxis
ISO	...	International Organization for Standardization
JSON	...	Java Script Object Notifier
NAS	...	Network Access Storage
PI	...	Principal Investigator
XML	...	eXtended Markup Language

III