Research Data Management

Best practice guide with comprehensive information – How to best manage research data

Version 1.0.4

- this best practice guide has been designed as a guidance document to be part of a website on research data management
- the current website is designed without any direct help, support, or training. These sections will be added at a later stage

Author: Matthias Rösslein

ICT-Kommission/AG-Datamanagement

1. December 2019

Content

1 RESEARCH DATA MANAGEMENT – HOME 4

1.1 WHY MANAGE RESEARCH DATA? 4
1.2 OUR GOVERNANCE 4

2 WRITING A DATA MANAGEMENT PLAN 4

2.1 WHAT IS A DATA MANAGEMENT PLAN? 4
2.2 COMPONENTS OF A DATA MANAGEMENT PLAN 4
2.3 DATA MANAGEMENT PLANS FOR FUNDERS 5
2.4 SCIENCE EUROPE – PRACTICE GUIDE 5
2.5 ADDITIONAL RESOURCES 6

3 FORMATTING AND NAMING DATA 7

3.1 CHOOSING A DATA FORMAT 7
3.2 FILE NAMING 7
3.3 VERSIONING 8
3.4 FOLDER STRUCTURE OF A RESEARCH PROJECT 9

4 DATA CAPTURE, ANALYSIS AND DOCUMENTATION 9

4.1 DATA CAPTURE 9
4.2 CODE VERSIONING 9
4.3 DATASET DOCUMENTATION (METADATA) 10
4.3.1 METADATA – DESCRIPTIVE OR STUDY-LEVEL DOCUMENTATION 10
4.3.2 METADATA – STRUCTURAL OR DATA-LEVEL METADATA 10
4.3.3 WHAT SHOULD METADATA CONTAIN 10
4.4 README FILES 11
4.5 METADATA STANDARDS 11

5 STORAGE, BACKUP, AND SECURITY 12

5.1 STORAGE OF ACTIVE DATA 12
5.2 BACKUP 12
5.3 SECURITY 12
5.4 DATA CLASSIFICATION 13

6 ETHICAL AND LEGAL ASPECTS 13

6.1 ETHICS AND DATA PROTECTION 13
6.2 CONFIDENTIALITY 13
6.3 INTELLECTUAL PROPERTY, LICENSING AND RE-USE OF DATA 14
6.3.1 DATA VERSUS DATABASE 14

6.3.2 DATA LICENSING ... 14
6.3.3 SELECTING A DATA LICENSE ... 15
6.3.4 RE-USING EXISTING DATA ... 15
6.3.5 DATA OWNERSHIP @EMPA ... 15

7 DATA SHARING, PRESERVATION, AND CITATION ... 16

7.1 DATA SHARING VERSUS OPEN DATA ... 16
7.1.1 DATA SHARING ... 16
7.1.2 OPEN DATA ... 16
7.2 RESEARCH DATA PRESERVATION ... 16
7.3 FUTURE FILE USABILITY ... 17
7.4 DATA REPOSITORIES ... 17
7.5 RESEARCH DATA CITATION ... 17
7.6 PERMANENT UNIQUE IDENTIFIER ... 18

8 ADDITIONAL RESOURCES ... 18

8.1 USEFUL LINKS ... 18
8.2 ABOUT US ... 19

9 ANNEX ... 19

9.1 SOP ... 19
9.1.1 GENERAL RECOMMENDATIONS ... 19
9.1.2 STRUCTURE AND CONTENT OF SOPS ... 19
9.1.3 CHECKLIST ... 23
9.1.4 REFERENCES ... 23

1 Research Data Management – Home

1.1 Why Manage Research Data?

Research Data Management addresses the organization, capture, storage, preservation, and sharing of research data created during a research project. It comprises researcher's strategies for caring for their data and decisions concerning what to do with the data upon completion of a project. This results in depositing data in a data repository for long-term access and archiving.

For a researcher, having an effective plan and approach for the management of their data is important for several reasons, including:

- Digital data are fragile and can easily get lost
- Good data management can save researcher teams time and resources in the long-run, by making it easier to find and re-use data files
- There are growing research data requirements imposed by funders and publishers concerning the publication of data
- Research data management helps preventing errors and increases the quality of data analysis
- Well-documented data make it easier to write up research results for publications
- Well-managed and accessible data permits others validating and replicating findings
- Data is a scholarly product and, when shared, well-managed data can lead to valuable discoveries by others outside of the original intent.

1.2 Our Governance

This guide is intended to provide Empa researchers with a guideline on how to organize, capture, store, preserve and share research data. As it is stated in the title, this guide provides the "best practice" concerning research data management compiled from various sources. It is not a binding document for Empa employees and every researcher has to take care that he/she complies with the legal requirements – especially from funding agencies such as SNSF but also with the Empa Management Handbook (MHB).

This document should be a living document that is actually used in practice and helpful to our researchers. Any feedback on this best practice guide is therefore welcome and could be included in future versions.

2 Writing a Data Management Plan

2.1 What is a Data Management Plan?

Data management plans (DMPs) are written, living documents that outline what researchers will do with the data during and after research projects. As the Digital Curation Centre (DCC) explains, DMPs "typically state what data will be created and how, and outline the plans for sharing and preservation, noting what is appropriate given the nature of the data and any restrictions that may need to be applied."

Funding agencies are increasingly requiring that grants include a DMP that describes how the data will be handled throughout the research lifecycle and how the data will be disseminated. Even where a plan is not required, having one formalized is good practice and can help to ensure that a research team is following the same approaches to caring for data.

2.2 Components of a Data Management Plan

There is a general set of elements that DMPs should address, while funder requirements for DMPs can differ:

- Roles and responsibilities:
    - Who will be responsible for data management?
    - How will adherence to data management policies be enforced?
- Data production and storage:
    - How and what types of research data will be produced?
    - How is the quality of the research data controlled?
    - How much research data will be produced?
    - How will research data be stored during the active phase of the project?
    - Will publicly-available research data be used (if so, from where)?
- Data organization and documentation:
    - How will data be processed and organized?
    - What file formats will be used?
    - How will data be described or contextualized so that they can be found and re-used by others in the future?
- Data access and sharing:
    - How will data be shared with others?
    - Will release of data be embargoed (if so, why and for how long)?
    - If data are of a sensitive nature, how will sharing be restricted and/or data processed to protect privacy?
- Data re-use:
    - Who can re-use the data?
    - How should others re-use the data?
    - What credit should be given for data re-use?
    - Can others re-disseminate the data?
- Data preservation:
    - Which data will be preserved?
    - How long will data be preserved?
    - Which repository/archive/database will be used?
    - How often will back-ups occur?
    - What metadata or documentation will accompany the data?

For more useful prompts, the DCC in the UK has prepared a Checklist for a Data Management Plan.

2.3 Data Management Plans for Funders

Many granting agencies and funders require data management plans be submitted with grant proposals. They recognize the benefits of data management planning on the integrity of research data and often have policies encouraging data sharing. These data management plans are generally no more than two pages, enough to get researchers thinking about the issues of data management in the early stages of planning when it is most effective. Some funders also require grantees to discuss how they followed their plan at the end of the grant.

2.4 Science Europe – Practice Guide

Science Europe, of which SNFS is a member, has published a practice guide. According to their guidance each DMP has at least to answer the following questions:

1. Data description and collection or re-use of existing data
    a. How will new data be collected or produced and/or how will existing data be re-used?

    b. What data (for example the kinds, formats, and volumes) will be collected or produced?
2. Documentation and data quality
    a. What metadata and documentation (for example the methodology of data collection and way of organizing data) will accompany data?
    b. What data quality control measures will be used?
3. Storage and backup during the research process
    a. How will data and metadata be stored and backed up during the research process?
    b. How will data security and protection of sensitive data be taken care of during the research?
4. Legal and ethical requirements, codes of conduct
    a. If personal data are processed, how will compliance with legislation on personal data and on data security be ensured?
    b. How will other legal issues, such as intellectual property rights and ownership, be managed? What legislation is applicable?
    c. How will possible ethical issues be considered, and codes of conduct followed?
5. Data sharing and long-term preservation
    a. How and when will data be shared? Are there possible restrictions to data sharing or embargo reasons?
    b. How will data for preservation be selected, and where will data be preserved long-term (for example a data repository or archive)?
    c. What methods or software tools will be needed to access and use the data?
    d. How will the application of a unique and persistent identifier (such as a Digital Object Identifier (DOI)) to each data set be ensured?
6. Data management responsibilities and resources
    a. Who (for example role, position, and institution) will be responsible for data management (i.e. the data steward)?
    b. What resources (for example financial and time) will be dedicated to data management and ensuring that data will be FAIR (Findable, Accessible, Interoperable, Re-usable)?

The guidance document further describes the selection criteria of a data repository and the development of a DMP template, which reflects the above outlined set of minimal requirements.

2.5 Additional Resources

The Data Life-Cycle Management (DLCM) project details requirements of DMPs for SNSF project proposals. They provides the corresponding SNSF DMP templates (pdf & docx), which was prepared jointly by teams from the libraries of EPFL and ETH Zurich, with input from DLCM partners.

Further resources can be freely downloaded and used:

DLCM SNSF DMP Template (pdf)
DLCM Data Management Checklist
DLCM Generic DMP Template
EAWAG DMP Guide
DMP Canvas Generator

3 Formatting and Naming Data

3.1 Choosing a Data Format

The research equipment, computer hardware, and software often determine the format of the digital data files. However, converting to a non-proprietary file format improves for preservation and access depending on the format and the tools, which are used. There are software tools, which perform these data format conversation automatically.

Recommended file formats that best support sharing, reuse, and preservation are formats that are open source, software-neutral, unencrypted, if feasible uncompressed otherwise lossless compressed, and in use within disciplinary communities. Some area Stanford University Libraries Data Management Services has made a useful overview of recommended file formats available. The following listing shows the suffix of the related data files:

- Containers: TAR, GZIP, ZIP
- ASCII-Databases: XML, CSV
- Geospatial: SHP, DBF, GeoTIFF, HDF5, NetCDF
- Multidimensional arrays: HDF5, NetCDF
- Moving images: MOV, MPEG, AVI, MXF
- Sounds: WAVE, AIFF, MP3, MXF
- Statistics: ASCII (i.e. .txt), DTA, POR, SAS, SAV
- Still images: TIFF, JPEG 2000, PDF, PNG, GIF, BMP
- Tabular data: CSV
- Text: XML, PDF/A, HTML, ASCII, UTF-8 (i.e. .txt)
- Web archive: WARC

During the planning phase[1] the following considerations might be useful:

- Is the data reliant on proprietary software to access it?
    - If yes, preserving a lossless copy in an open, sustainable file format will help to ensure that everybody can access the data in the future.
    - Beside that it is recommended creating a copy of the original data format with a copy of the (instrument) software to open the original data file.
- If the research data are deposited in a repository at the end of a project, then the repository might have specific guidelines or requirements with respect to file format.
    - If yes, a copy is created in the required format for deposit and the conversion is documented for the users.
- Will converting to another file format modify the data or cause a loss of features?
    - If yes, then it is strongly recommended creating a copy in an open format but preserving the original data format together with a copy of the (instrument) software to open these data files.

3.2 File Naming

Planning the naming of the files makes finding of files easier, avoids duplication, and helps to finalize projects quicker. When naming files, the following should be considered:

- The data files should be named in a consistent manner

[1] For advanced users: If mapping your content to such a convention and directory-structure becomes too complex, you should consider to employ a proper database. In particular, if you feel you spend too much effort in your analysis code to construct the paths for the data files, you are in the process of implementing a primitive database software by yourself and should step back and reconsider.

- The whole project team has to know the naming convention
- Files should receive a meaningful, descriptive name. A file name might include a combination of elements, such as type of equipment used, date, and researcher's surname
- The best order for elements in a file name should be decided; it will affect how the files are sorted
- Brief file names should be used
- Underscores instead of spaces should be used to separate words/dates
- Letters and numbers, rather than special characters like ~ ! @ # $ % ^ & * ( ) ` ; < > ? , [ ] { } ‘ “ are strongly recommended

Repositories, where research data of a certain field might be preferably stored, could have their own naming rules. Therefore, it is advisable check this before the project start and to take them into account, when outlining the naming rules.

3.3 Versioning

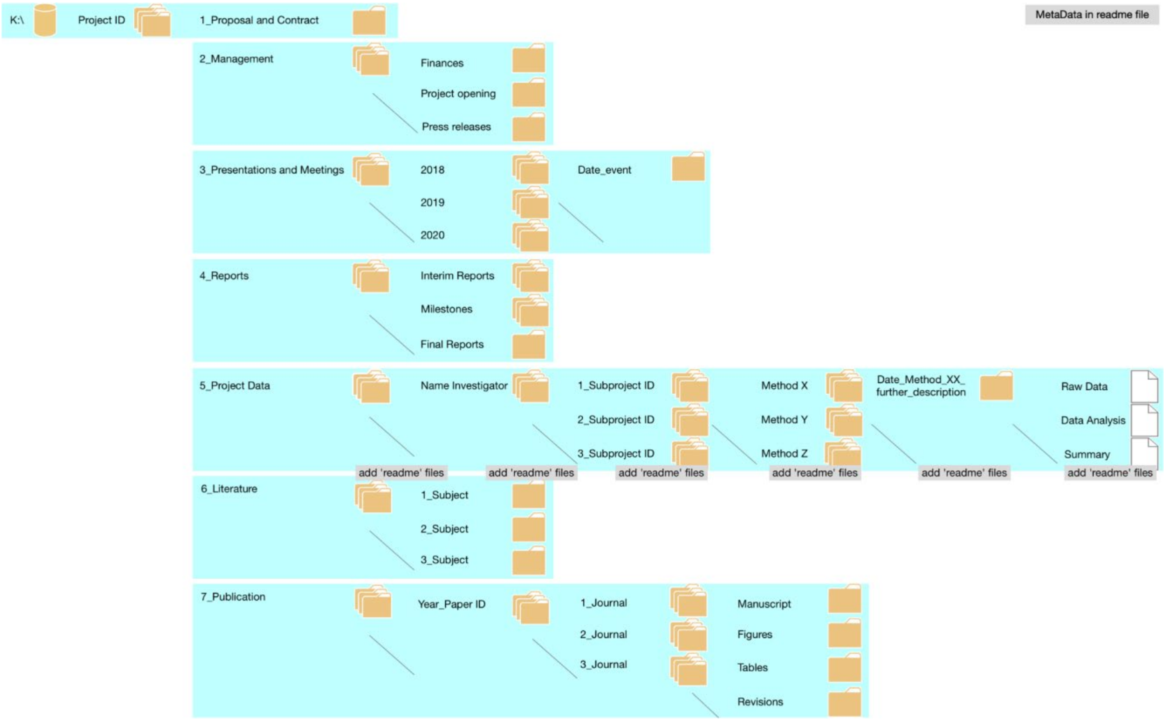
Versioning should be considered when developing the folder and file naming structure. But the following items should be kept in mind:

- A simple method to designate a revision is to note it at the end of the file name. This way, files can be grouped by their name and sorted by version number:
    - image1_v1.jpg
    - image1_v2.jpg
    - image2_v1.jpg
    - image2_v2.jpg
    - ...
- If using sub-versioning then the following naming rules apply:
    - Original document: DMP_1.0
    - Original document with minor revisions: DMP_1.1
    - Document with substantial revisions: DMP_2.0
    - …
- If variable digit version numbers are used, one issue that can arise is that computers will sort files based on the position of the characters. This can lead to undesired sorting:
    - Image0001_v1.jpg
    - Image0001_v10.jpg
    - Image0001_v2.jpg
    - ...
- Therefore, it is recommended to use two digits version numbers:
    - Image0001_v01.jpg
    - Image0001_v02.jpg
    - …
    - Image0001_v10.jpg
    - ….
- A good practice that can help avoiding these problems is to use dates to designate version numbers. If this strategy is chosen, then the dates should be formatted as year-month-day (20190130). Using this order will help avoid confusion when collaborating with other researchers or systems that use a day-month-year or month-day-year as dates, and it will help sort versions in chronological order:
    - Image0001_20181211

  - Image0001_20181214
  - Image0001_20190123
  - ...
- If the files, which are used, are created or edited collaboratively, it is recommended incorporating names or initials into the file naming conventions. In this way one knows, which versions contain updates by each individual of a team:
  - Dataset0001_20180430_RM
  - Dataset0001_20180501_WIP
  - Dataset0001_20180814_HIC
  - …

3.4 Folder Structure of a Research Project

There are different ways to organize the folders and data files in a project. The folder structure shown below is just one possibility:



4 Data Capture, Analysis and Documentation

4.1 Data capture

Data capture should always be done in a consistent way. It is strongly recommended to develop Standard Operating Procedures (SOP) that clearly define the steps to be taken and outlines roles and responsibilities. SOPs are useful even for single person projects to ensure that there is consistency over time. Beside informing on experimental setup and applied procedure, SOPs should also indicate when to create documentation and where and how files are named.

4.2 Code versioning

In cases where data capture and analysis relies strongly on data processing (implemented in any compiled general-purpose programming language such as C/C++, FORTRAN or interpreted languages such as python, R, matlab) a version control software (e.g., svn, git) may provide significant benefits regarding versioning, documentation, bug fixes, and released/applied versions. The use of a version control software is highly recommended for software that is structurally complex (significant number of modules or classes), contains original numerical algorithms, is supposed to be developed and applied over a longer period and is coded simultaneously by several developers. Such versioning systems are NOT thought to be used for data themselves, but can be used for any kind of ASCII file-based coding project (incl. Latex manuscripts or web content).

For Empa internal code development (only Empa users), Empa ICT offers a GitLab server to all Empa users (https://gitlab.empa.ch).

4.3 Dataset Documentation (Metadata)

Describing and documenting data is the only ways to ensure that they will be discoverable, searchable, and re-useable in the future. This documenting is often called metadata and includes all relevant information.

Commonly there are two types of documentation (UK Data Archive – Document your data) that ensure usability far in to the future. They are:

- Descriptive or study-level documentation
- Structural or data-level metadata

4.3.1 Metadata – Descriptive or study-level documentation

These metadata describe the content and purpose of the study in a comprehensive way. They include the goal of the study, the design of the study with all of its experiments and measurements, its major findings and the conclusion. Based on these metadata information it should be feasible to comprehend the study undertaken and its outcome.

4.3.2 Metadata – Structural or data-level metadata

These metadata detail the actual measurement with its samples and references, the employed SOP and the comprehensive data analysis. It is their purpose to illustrate each of the measurement undertaken. Furthermore, they should be so detailed that any other institution is capable to fully reproduce the measurement results. The reproducibility of measurement results and whole studies is one of the essential items of RDM.

4.3.3 What should Metadata contain

Documentation needs will vary by project and by discipline. Many disciplines have developed metadata standards that specify what information should be collected. If there isn't an existing standard, a template should be created that will record all the important details of the data. At the minimum it should include the following keywords:

General Information:

- Title
- *Creator*: names and addresses of data creator(s)
- *Publisher*: addresses of data publishers
- *Identifier*: can be a permanent identifier or an internal project number
- Funder
- *Rights:* Intellectual property or licensing rights for the data
- Access Information
- Language
- *Project description* (e.g. subject, scope etc.)
- *Data Citation*: Preferred format for citing data.

Data and file overview:

- *Data Structure*: including relationships between files
- *File description*: A short description of each file
- *Dates* that the file was created

Methodological information:

- *Measurement process*: Description of measurement setup and measurement method
- *Data capturing*: Description of methods for data capturing
- *Data processing*: Description of methods for data processing (if data is not raw data)

Data specific-information:

- *Variable list*: with full names and definitions of column headings if tabular data
- Measured quantity (measurand)
- *Units of measurement*: IS units of the measurement result
- Location of measurement
- *Definitions*: Definitions for codes or symbols used to record missing information

This list is based on MIT Documentation and Metadata guidance, UK Data Archive Study Level Documentation, and Cornell University, Guide to writing "readme" style metadata. This meta-data information should be included as documentation in a README.txt file in the folder with the data files.

4.4 README Files

These metadata are often collected in readme files, which are plain text files (.txt) or sheets in a spreadsheet. It helps others to understand the research data and interconnections among data files. By titling the file "readme," the date creator informs to users that this file should be looked at first. For researchers depositing data in a data repository, the information in the readme file augments information included in the metadata form. Furthermore, if the deposit includes multiple files, it explains the file naming structure, relationship among the files, and abbreviations used. Cornell University's Research Data Management Service Group has made a useful readme file template available for download.

4.5 Metadata Standards

There are a number of community-maintained lists of disciplinary metadata standards:

- General collection or definition of metadata
  - Research Data Alliance (RDA) - Metadata Directory
  - Digital Curation Center (DCC)
  - Dublin Core - domain independent, basic and widely used metadata standard
  - Fairsharing.org - The standards in FAIRsharing are manually curated from a variety of sources
  - DataCite Metadata Schema (.pdf - V4.1)
- Biological science
  - Minimum Information for Biological and Biomedical Investigations (MIBBI)
  - MINimal information about high throughput SEQeuencing Experiments (MINSEQE) - Genomics standard
- Ecological science
  - Ecological Metadata Language (EML) - specific for ecology disciplines
- Geographic information
  - ISO 19115-1:2014 Geographic information - Metadata - Part 1: Fundamentals
  - Federal Geographic Data Committee's Content Standard for Digital Geospatial Metadata (FGDC-CSDGM)
- Social, behavioral, economic, and health sciences
  - Data Documentation Initiative (DDI) - common standard for social, behavioral and economic sciences, including survey data

5 Storage, Backup, and Security

5.1 Storage of Active Data

A key aspect of a data management plan is a storage strategy for active data and archival data. Data are easily lost, digital files are fragile, and formats and storage media become rapidly obsolete over time.

Active or *working* data are referred to as research data, which is collected and accessed during the course of a project. The datasets will be expanding as the collection of new data continue and so data access is regularly required for processing and analysis. An important component of data management planning is deciding where and how the active data will be stored so that it is readily accessible for data processing but also secure. These issues should be considered:

- Anticipated size of datasets
- Perform capacity planning of CPU, memory and diskspace over lifetime. This ensures that the systems employed are able to scale with the expected requirements.
- Computational requirements: Large-scale analyses may require high-speed processors, network bandwidth (speed) and a substantial amount of disk space
- Backup
- Security, together with risk management
  - How reliable are the systems used?
  - Needed skills regarding data and system handling
- Data classification

5.2 Backup

It must be evaluated which data and which minimum requirements must be met by the backup

- Which data must be saved (raw data, processed data, results)
- Differentiation between backup of the latest state (recovery) or backup of generations (versioning, undoing changes)
- Can the data be retrieved with alternative procedures?
- Is sufficient storage capacity available or additional storage needed?
- Who is responsible for backup and restore?
- How is the data restored in case of an incident?
- How long will the data be retained and preserved?

Documented procedures (3.4.5 Backup and Restore) exist for the central services like network shared storage, ftp-server, private cloud-based solutions (PolyBox, SharePoint). A generation-based backup has been set up for Empa owned infrastructure. There are no standardized solutions for decentralized or external data backup. These must be checked individually.

5.3 Security

MHB-3.4.4 Regulations for the use of informatics at Empa (RUI) These regulations make up general guidelines concerning the proper and secure utilization of information technology systems. Their objective is to optimize the availability of information technology resources for teaching, research and service purposes and to ensure the integrity and confidentiality of all processed and stored information. In addition, the regulations provide directives concerning the misuse of information and the consequences of such misuse.

The responsibility for data security must be clarified on a project-by-project basis and compliance with the requirements ensured. For the use of central IT services, standardized procedures exist to ensure access authorizations and protect data integrity. This includes hardware, software, networks and storage[2].

5.4 Data Classification

The directive MHB "2.3.21 Directive for the Classification of Data at Empa" determines how Empa data is to be classified and how it is to be processed. Any classification of data being generated by third parties will remain independent from the specifications of this directive.

There are other classification features to consider:

- Relevance of the data
- Primary data, secondary data, personal data (MHB 2.3.18) https://www.empa.ch/group/mhb/2.3.18-umgang-mit-dokumenten-und-daten
- Form (Non-electronic data, electronic data)
- Confidentiality (public data, classified data, authorization area)
- Storage (offline, online internal, external e.g. CSCS)
- Capacity (transport and storage, lifetime)

6 Ethical and Legal Aspects

6.1 Ethics and Data Protection

Ethics should be considered early in any research project. Information on data storage, security, availability, and archiving should be included into the DMP. Well-planned informed consent that carefully considers ethics can help make data more shareable.

Art. 36c et seq. of the ETH Act regulate the handling of personal data for research projects at Empa. Personal data that has been rendered anonymous in such a way that the individual is not or no longer identifiable is no longer considered personal data. For data to be truly anonymised, the anonymisation must be irreversible. Personal data that has been de-identified, encrypted or pseudonymised but can be used to re-identify a person remains personal data and falls within the scope of the data protection laws. If a research project uses non-anonymised personal data of any kind (e.g. names, physical and/or IP addresses etc.) or other sensitive data (e.g. medical data, data that allows for military use etc.), a declaration to Empa Legal and/or the Empa Ethics Committee is to be considered, before making this data publicly available.

6.2 Confidentiality

In a research project with external partners (research and/or industrial partners) there is usually an agreement on confidentiality – be it verbal or in writing (e.g. in a Non-Disclosure Agreement). These agreements have to be respected when sharing data and/or uploading data onto an (open) data repository. Regardless of the existence of such agreements, every research project has to comply with the rules of the Empa Management Handbook (MHB). Of particular relevance for research data management are MHB 2.3.21 ("Directive for the Classification of Data at Empa") and MHB 2.3.18 ("Guidelines for Handling Data at Empa").

Furthermore, confidentiality is also important where a research project could potentially lead to patentable inventions. Sharing data (or sharing data too early) might be detrimental

[2] If the data is encrypted, the key management must be regulated. Who is in possession of the key, how is the key kept safe and is it available again during the backup

to the patentability of an invention. In this context, MHB 4.3.1 ("Directive concerning the rights to research results and their exploitation") has to be considered before sharing data.

Whenever confidentiality agreements, patentability of an invention or ethical and data protection concerns prevent data sharing, there has to be an "opt-out" of the affected data; i.e. the affected data cannot be shared or uploaded onto an (open) data repository. Funders of research projects (e.g. EU or SNFS) allow for such an opt-out, as long as it is clearly stated in the DMP – including the reasoning why the data has to be omitted.

6.3 Intellectual Property, Licensing and Re-Use of Data

6.3.1 Data versus database

In any data project, there are likely to be two components: The first is the data collected, assembled, or generated. It is the raw content in the system. This raw data could be hourly temperature readings from a sensor, the age of individuals in a survey, recordings of individual voices, or photographs of plant specimens. The second component is the data system in which the data is stored and managed – the database. In general, raw data (e.g. temperature, humidity etc.) on their own are considered facts and thus are not copyrighted under Swiss law. However, data that are gathered together in a unique and original way, such as databases (e.g. a climate database consisting of temperature, humidity etc. over a given timespan), might automatically fall under copyright protection. Deciding what data needs to be included in a database, how to organize the data, and how to relate different data elements are all creative decisions that may receive copyright protection.

As a clear distinction between raw data and a database can be difficult, clear licensing regarding re-use is important.

6.3.2 Data Licensing

This section covers only the licensing of data and databases. For the licensing of software, please refer either to the Software Declaration Form (SDF) or to the Open Source Software Registration (OSSR) provided by the Empa-Eawag TT-Office. If your software includes a database, the SDF should also be used. For inventions, please use the Invention Disclosure Form (IDF).

There is increased pressure from funders and journals for researchers to release their research data. Applying appropriate licensing when data are released will help ensure proper re-use and attribution. There are many licenses available that represent the range of rights for the creator and licensee of the data. When choosing a license, the (possible) conflict of objectives between confidentiality, patents and ethics on one hand and open data on the other hand should always be kept in mind. As mentioned above, there is the possibility to "opt-out", if researches are obliged to keep certain data confidential or if there might be a patentable invention. Furthermore, it has to be taken into consideration that not all data are in the public domain: A project might, for example, use copyrighted photographs; these photographs are also part of the project’s "data". Therefore, it might be necessary to differentiate between the database and its data content (e.g. images, text, films, music) for licensing. In case of queries or uncertainties, please contact the Technology Transfer Office, either your TT contact person or the administration (email: TT@empa.ch)

In order to facilitate the re-use of data, it is imperative that others know the terms of use for the database and the data content. The Open Data Commons group (ODC) has been developing three standard licenses to govern the use of data sets.

The three ODC licenses are:

1. Public Domain Dedication and License (PDDL): This dedicates the database and its content to the public domain, free for everyone to use as they see fit.
2. Attribution License (ODC-By): Users are free to use the database and its content in new and different ways, provided they provide attribution to the source of the data and/or the database.
3. Open Database License (ODC-ODbL): ODbL stipulates that any subsequent use of the database must provide attribution, an unrestricted version of the new product must always be accessible, and any new products made using ODbL material must be distributed using the same terms. This license applies only to the database itself and not to the data content (which has to be licensed separately, if licensing is possible). It is the most restrictive of all ODC licenses.

Creative Commons (CC) also has a library of standardized licenses, and some of them apply to data and databases. The ODC-By license, for example, is the equivalent of a Creative Commons Attribution license (CC BY). CC BY licenses, however, require copyright ownership of the underlying work (the data content), whereas the ODC-By license applies to works not protected by copyright (such as factual raw data)

The two CC licenses that are of greatest relevance to data management are:

1. CC0 (i.e., "CC Zero"): When an owner wishes to waive the copyright and/or database rights, one can use the CC0 mark. It effectively places the database and data into the public domain. It is the functional equivalent of an ODC PDDL license.
2. Public Domain mark (PDM): It is used to mark works that are in the public domain, and for which there are no known copyright or database restrictions. It is possible to flag factual data as PDM in a database, for example, in order to make it clear it is free to use.

6.3.3 Selecting a data license

There is no single right answer as to which license to assign to a database or content. Note, however, that anything other than an ODC PDDL or CC0 license may cause serious problems for subsequent scientists and other users. This is because of the problem of attribution stacking. It may be possible to extract data from a data set, use it in a research project, and still maintain information as to the source of that data. It is possible to create a data set derived from hundreds of sources with each source requiring acknowledgement. Furthermore, the data in the other databases may not have originated with it, but instead sourced from other databases that also demand attribution. Rather than legally require that everyone provide attribution to the data, it might be enough to have a community norm that says “if you make extensive use of data from this data set, please credit the authors.”

In case of queries or uncertainties when selecting a data license, please contact the Technology Transfer Office, either your TT contact person or the administration (email: TT@empa.ch).

6.3.4 Re-Using Existing Data

When re-using existing data, one has to clarify ownership, obtain permissions if needed, and understand limits set by licenses. In any case, it is important to provide appropriate attribution and citation in accordance with research standards.

6.3.5 Data Ownership @Empa

In accordance with article 36 of the ETH Act, all rights in research results (intellectual property) that have been created by Empa staff in the exercise of their official duties, with the exception of copyrights, are the property of Empa. The author is entitled to the copyright in protected works that are created within the framework of an employment relationship (in particular scientific publications or textbooks). If a work of this kind is developed in fulfilment of a performance obligation undertaken under an employment contract (known as a service work, article 36 of the ETH Act), the assignment of the copyright and of the exploitation rights to Empa can be agreed by contract by Empa with the author. Empa has reached such an agreement with its employees in the employment contracts (excluding publications by employees). For details regarding data ownership, please refer to MHB 4.3.1 ("Directive concerning the rights to research results and their exploitation").

7 Data Sharing, Preservation, and Citation

7.1 Data Sharing versus Open Data

Many funders and journals have data sharing requirements and others call for open data. Managing the research data throughout the project can help ensure that either of these goals can be met. If one is working with a specific funder or journal it is important to check their exact requirements.

7.1.1 Data Sharing

Data Sharing encompasses the spectrum from making data available upon specific request to depositing data in an open and publicly accessible repository. It is important to know specifically what is required by a funder, journal, or institution. For example, one definition of data sharing is making data available to people other than those who have generated them. This can range from bilateral communications with colleagues, up to providing, free unrestricted access to the public through web-based platforms.

7.1.2 Open Data

Open Data is data that is deposited in an open, publicly accessible repository. In particular, the Open Definition summarizes open data as “A piece of data or content is open if anyone is free to use, reuse, and redistribute it- subject only to the requirement to attribute and/or share-alike". Their full definition includes several detailed points that address issues such as access, reuse, redistribution, licensing, technological restrictions and more.

The Panton Principles are a set of recommendations for making research data open in science. They state to support the position on open data:

- Science is based on, reusing and openly criticizing the published body of scientific knowledge.
- For science to effectively function, and for society to reap the full benefits from scientific endeavors, it is crucial that science data be made open.
- By open data in science we mean that it is freely available on the public internet permitting any user to download, copy, analyze, re-process, pass them to software or use them for any other purpose without financial, legal, or technical barriers other than those inseparable from gaining access to the internet itself. To this end data related to published science should be explicitly placed in the public domain.

7.2 Research Data Preservation

Preservation is done at the completion of a project. Ideally the preservation strategy reflects long term thinking and should not be the same as how the data were stored during the project. Things to consider when developing a plan for preservation include:

- What does one need to keep?
- What are requirements of the funders or of the journals?
- How long does the data need to be preserved? (Empa management-handbook, SNFS)
- Who is responsible for the data at the end of the project?
- Does the funder or journal specify a repository?
- Is there sufficient documentation that anyone can use the research data without any assistance, including software needed and file structures?
- Are the file formats open and sustainable?
- If the research data are not deposited into a repository, what is the shelf-life of the hardware and when will data need to be migrated?

The Digital Curation Centre offers further recommendations on these subjects:

- "How to appraise and select research data for curation" guide (.pdf)
- "Five steps to decide what to keep" guide (.pdf)

7.3 Future File Usability

Thinking about future file usability will ensure that the data are still usable and can be shared in the future. Important items to be consider include:

- Is the file format open or closed?
- Is a specific software package required to use the data?
- Do the multiple files comply with the previously defined data file structure?
- Will somebody be able to open the file 10 years from now?

One should select a consistent file format that can be read well in the future. This means that the file format is open, has documented standards, is unencrypted, and is, if feasible, uncompressed, or otherwise using a well-established lossless compression standard. See also section 3.1.

- ASCII formatted files will be readable well into the future
- .docx -> .txt
- .xlsx -> .csv
- .jpg -> .tif
- Save a copy in the original format, just in case

7.4 Data Repositories

Research data repositories host, provide persistent access to, and preserve datasets. For many disciplines, there are repositories familiar to and well-used by researchers in the field. In addition to considering disciplinary practices around data deposit, researchers should determine whether their funder or publisher requires or recommends a specific data repository for archiving and making data available. There are three main types of data repositories available:

- Disciplinary repositories - to check for a repository in a given discipline it is best to search in the Registry of Research Data Repositories (re3data.org). It can be browsed by discipline, data type and country for discovering an appropriate home for their data or to find shared datasets to use in their research.
- General repositories: Zenodo
- Institutional repositories

7.5 Research Data Citation

Citing research data in a manner similar to traditional scholarly works can help ensure proper attribution, improve reproducibility, improve discoverability, and help provide credit for research data as a scholarly output. According to the Force11 examples of the joint declaration of data citation principles should be cited as follows:

- Include an in-text citation near the claims relying on the data in the form of the citation style required by publisher. Additional information may also be included in the in-text citation, such as portion of data set used. e.g. [Author(s), Year, Portion or Subset of Data Used].
- Full citations should be included in the reference list, following the format of the required citation style. The DCC guide "How to Cite Datasets and Link to Publications" (.pdf) provide a comprehensive list of data citation elements. If no format exists, Force11's examples recommend: Author(s), Year, Dataset Title, Data Repository or Archive, Version, Global Persistent Identifier.
- Permanent identifiers, such as DOIs or ARKs, should be given in the form of a linked URL, if possible.
- Data sets are cited at the most detailed level possible and provide a version number, if appropriate.
- When citing a dataset in a publication, the metadata entry in the repository, holding this dataset, should be adjusted to include a link to your publication.

7.6 Permanent Unique Identifier

Permanent Unique Identifiers provide a persistent and, hence, permanent link to a research dataset or other digital object regardless of hardware or domain changes a repository may undergo over time. Persistent identifiers are generally provided when data are deposited into a repository. There are several types of persistent identifiers currently used, including HanDLes (HDL), Archival Resource Keys (ARKs), Persistent URLs (PURLs), and Digital Object Identifiers (DOI). Most researchers are familiar with DOIs as this is the system used for most electronic journal articles. The California digital library's webpage on Understanding Identifiers provides more information on persistent identifiers (DOIs) and allows to generate them.

8 Additional Resources

8.1 Useful Links

The content of following links was incorporated into this website:

- University and research institutes websites:
  - CalTech Library - Research Data Management
  - Cornell University - Research Data Management Service Group
  - Eawag Research Data Management Project
  - Harvard Library – Research Data Management Program
  - John Hopkins Library – Data Management Services
  - MIT Libraries – Data Management
  - Princeton University Library – Research Data Management at Princeton
  - Stanford Libraries – Data Management Services
  - University of Cambridge – Research Data
  - University of Chicago Library – Research Data Management
  - University of Oxford – Research Data Oxford
  - Yale University Library – Research Data Management
  - 4TU – Center for Research Data
- General websites on research data management

    - Digital Curation Center (DCC)
    - (Swiss) Data Life-Cycle Management (DLCM)
    - OpenAIRE – Science set free
    - UK Data Services
- Funder websites:
    - EU – Open Science
    - National Institutes of Health – DataScience@NIH
    - National Science Foundation – Open Data
    - SNSF – Open Research Data

All these websites are listed in purely alphabetical order.

8.2 About Us

Currently research data management is built up at Empa. It will comprise guidance in form of this document, training and direct support and help for individual researchers. So, if you have any questions or need help with your data management plan (DMP) etc., then please contact RDM@empa.ch

9 Annex

9.1 SOP

9.1.1 General Recommendations

This part of the attachment outlines in detail the content a standard operating procedure (SOP). All SOPs should have the same structure and section titles. The actual content of each section is described in the following sections (9.1.2. Structure and Content of SOP). There might be parts of a procedure, which are not covered by the given sections of the document template. Hence additional section and titles can be added to the current document structure. But one should avoid deleting any sections and their titles.

The drafting of SOPs starts after the initial development phase of a method. The main objectives of SOPs are:

- Document the way an operational procedure has to be executed (Operational Procedure).
- Execution of a procedure in an exactly repeatable and hence reproducible way over time and space. (Standardized Operation).
- SOPs permit the determination of performance characteristics of the operational procedure and with it the specification of acceptance criteria.

The drafting process is facilitated, if the content of each of the section is already available. The content of SOPs should be structured in such a way that it focuses on usability. The comprehensibility of the document is essential, so that any other laboratory can easily implement the procedure in their own laboratory without any further in-depth training. Each section should have an appropriate level of details.

SOPs shall be kept under version control. Any refinement of the content and with-it of the text of a SOP should lead to a new version number. All results of a measurement shall be assigned to the respective version number of the respective SOP.

9.1.2 Structure and Content of SOPs

The structure of the sections (chapter titles) of SOPs are listed and their aim and content are outlined below.

9.1.2.1 Chapter Titles of SOPs

1. Introduction

2. Principle of the Method
3. Applicability and Limitations (Scope)
4. Related Documents
5. Equipment and Reagents
    5.1. Equipment
    5.2. Reagents
    5.3. Reagent Preparation
    5.4. Reference Materials
6. Procedure
    6.1. Flow Chart of the Measurement Procedure
    6.2. Step by Step Description of the Measurement Procedure
    6.3. Definition and Equation of the Measurand
    6.4. Statistical Data Evaluation
    6.5. Reporting of the Results
7. Potential Pitfalls
8. Controls, Quality Control Samples and Acceptance Criteria
9. Health and Safety Warnings, Cautions and Waste Treatment
10. Abbreviations
11. References
12. Annex

9.1.2.2 Content of the different sections of SOPs

In general, all SOPs follow similar ideas or outlines, but there might be some differences in the detail structure. Therefore, a description of the content of each section is given below. If there is any section of your measurement procedure missing in the document template, then it should be added at the appropriate location. No section of the template shall be deleted.

9.1.2.2.1 Introduction

The introduction of the SOPs describes shortly its content and aim supporting the user in selection the appropriate measurement procedure.

9.1.2.2.2 Principle of the Method

This section provides a description of the biological or analytical chemical principles, which build the basic of the measurement procedure. It should help the users in their understanding of the method.

9.1.2.2.3 Applicability and Limitations (Scope)

The scope describes the boundaries, within which the measurement procedure of the SOP can safely be applied. For example, the scope describes the materials, for which the measurement procedure is fit for purpose.

It is essential to have the comprehensive understanding, for which area of application the measurement procedure has been tested and hence the measurement model is valid. An appropriated description of the scope of the measurement procedure helps users comprehending under which condition the application of the method is fit for purpose or not.

9.1.2.2.4 Related Documents

This section lists all the SOPs, which are related to this SOP and which have to be in place so that the actual measurement procedure functioning properly.

9.1.2.2.5 Equipment and Reagents

9.1.2.2.5.1 Equipment

A comprehensive list of all the equipment, which is used to perform the measurement procedure described in the SOP.

9.1.2.2.5.2 Reagents

A comprehensive list of all the reagents and chemicals, which are used to conduct the measurement procedure. If feasible specify all chemicals using their CAS number.

9.1.2.2.5.3 Reagent Preparation

A step by step description, how the different reagents are prepared, which are used in the measurement procedure. This does not include ready to use reagents bought from manufacturers.

9.1.2.2.5.4 Reference Materials

All measurements are relative measurements, which are directly linked to reference materials. Here all the used reference materials and their detailed specifications are given. Such information is specified in the reference material certificates, which should also provide information about the traceability of the certified values.

9.1.2.2.6 Procedure

9.1.2.2.6.1 Flow Chart of the Measurement Procedure

It proved very useful in many fields of measurement science to have a detailed flow chart of the measurement procedure. This allows obtaining an overview of the measurement procedure or locating a single step within the whole context of the procedure. Therefore, a detailed flowchart is drawn for each logical unit of steps, which is listed in the written description of a measurement procedure.

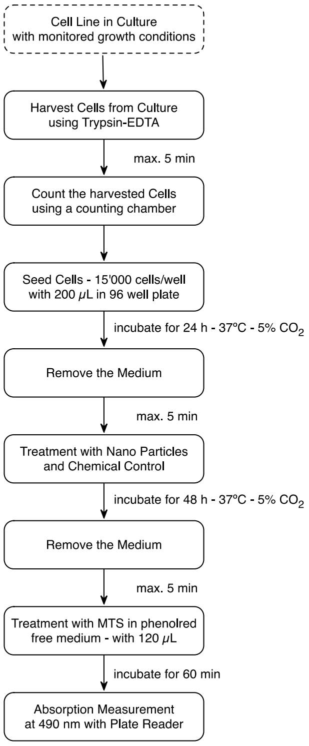

Cell Line in Culture with monitored growth conditions
Harvest Cells from Culture using Trypsin-EDTA
max. 5 min
Count the harvested Cells using a counting chamber
Seed Cells - 15'000 cells/well with 200 μL in 96 well plate
incubate for 24 h - 37°C - 5% CO_2
Remove the Medium
max. 5 min
Treatment with Nano Particles and Chemical Control
incubate for 48 h - 37°C - 5% CO_2
Remove the Medium
max. 5 min
Treatment with MTS in phenolred free medium - with 120 μL
incubate for 60 min
Absorption Measurement at 490 nm with Plate Reader

Figure: the flow chart of the MTS viability assay[1] is given as an example –
The flow chart shows each logical unit of steps of the measurement procedure

9.1.2.2.6.2 Step by Step Description of the Measurement Procedure

A detailed step by step description of the measurement procedure is one of the essential parts of any SOP. This description has to be in such detail that any person familiar with the measurement method and its field can directly perform the measurement without any additional training of information. It should include all the special aspects of the measurement procedure.

This detailed description is the essential precondition so that each measurement can be performed exactly according to the SOP.

9.1.2.2.6.3 Definition and Equation of the Measurand

The measurand[2] describes the value of the quantity, which is intended to be measured with the measurement procedure detailed in the SOP. The comprehensive understanding of the specification of the measurand is essential for all measurement procedures. The definition of the measurand includes the detailed equation of the measurand[2,3]. It shall list all the parameters of the equation and it describes the measurement model.

9.1.2.2.6.4 Statistical Data Evaluation

The section describes step by step the statistical evaluation of the measured values.

9.1.2.2.6.5 Reporting of the Results

The reporting of the measurement results is outlined in this section.

9.1.2.2.7 Potential pitfalls

A listing of potential pitfalls helps users of the SOPs to debug any measurements, which do not pass the acceptance criteria, which show out of specification control results or which results of the quality control samples are not within the range of the control chartings. This section helps the users avoiding the reporting of blunders.

9.1.2.2.8 Controls, Quality Control Samples and Acceptance Criteria

Controls, quality control samples and acceptance criteria allow to establish the confidence that the results obtained with the measurement procedure specified in the SOPs can be trusted. They are essential quality control measures, which should confirm that the performed measurements are according the specification of the measurement procedure and the measurement model is valid. Controls, quality control samples and acceptance criteria should be designed in such a way that they detect any shortcomings, mishap or failures when performing the measurements. Their design shows the standardization bodies that the context of the measurement procedure is understood and the robustness of the method has been established thoroughly. Their effectiveness persuades the regulatory bodies of the quality of the measurement procedure.

9.1.2.2.9 Health and Safety Warnings, Cautions and Waste Treatment

All health and safety warnings together with all precaution measured are outlined in this section. They help the user to perform measurement procedure safely and avoid any health risk. It also tells the users, if they fulfill all the training requirements to handle potentially dangerous goods. In addition, the waste treatment of any hazardous reagents used in the measurement procedure is described in detail. Waste treatment has to comply with international regulation.

9.1.2.2.10 Abbreviations

All abbreviations are defined in this section.

9.1.2.2.11 References

This section lists all references, which are essential for the SOPs. The listing should include any publication of regulatory bodies or of standardization bodies, which are relevant for the SOP.

9.1.2.2.12 Annex

This section lists any annex, which are essential for the implementation of the SOP.

9.1.3 Checklist

The checklist helps to review the completeness of the SOP.

Structure of SOP

1.1. Section of SOP in right order
1.2. Are there missing sections
1.3. Are there any additional sections

Equipment and reagents

1.4. Is the list of equipment comprehensive and specified in enough detail?
1.5. Is the list of reagents itemizing all reagents used in the SOP?
1.6. Are all the reagents specified well enough, so that any confusion can be exclude (CAS #)
1.7. Is the preparation of the reagents described detailed enough?

Reference material

1.8. Exact specifications of the reference material given
1.9. Conditions to store the reference material are described
1.10. Rules to check the stability of the reference material given

Understanding of measurement procedure

1.11. Scope of the measurement procedure clearly defined
1.12. Detailed flowchart of measurement procedure is included
1.13. Comprehensive step by step description of the measurement procedure existing
1.14. Is the level of detail sufficient, so that the description of each step leaves no room for interpretation?
1.15. Measurand clearly defined
1.16. Full equation of the measurand included
1.17. All parameters of the equation of the measurement defined
1.18. List with potential pitfalls existing
1.19. Full set of controls, quality control samples and acceptance criteria with a description, of which part of the measurement procedure is monitored, given
1.20. Are the acceptance criteria in line with relevant publications and requirements of the regulatory bodies?

Statistical evaluation of the measurement

1.21. A full description of the statistical evaluation of the measurement given
1.22. A data set to testing the proper functioning of the statistical evaluation given
1.23. Simple rules to verify the approximate values of the results independently given

Quality control samples

1.24. Exact specifications of the quality control samples given
1.25. Conditions to store the quality control samples are given
1.26. Rules to check the stability of the quality control samples given

9.1.4 References

1 Roesslein, M. *et al.* Use of Cause-and-Effect Analysis to Design a High-Quality Nanocytotoxicology Assay. *Chem. Res. Toxicol.* (2015). doi:10.1021/tx500327y

2 BIPM. International vocabulary of metrology - Basic and general concepts and associated terms (VIPM). *JCGM 200:2012* 1–108 (2012) https://www.bipm.org/en/publications/guides/vim.html

3 BIPM - Joint Committee for Guides in Metrology (JCGM). Evaluation of measurement data - Guide to the expression of uncertainty in measurement. *JCGM 100:2008* 1–134 (2008) https://www.bipm.org/en/publications/guides/gum.html