

Data Management Plan (DMP)

When planning research, it is important to consider carefully and document the ways of collecting and processing data during the research project, to specify who has access to these data and who is responsible for them, what will happen to the data after the closure of the project, etc. In order to do all this, it is necessary to create the data management plan and to follow it throughout the project.

DATA COLLECTION

What data will you collect or create?

Data type

- data types (experiment, observation data, survey data, video files, etc.)
- how new data integrates with existing data
- which data deserve long-term preservation
- if some datasets are subject to copyright or intellectual property rights, show that you have permission to use the data

Data formats

- point out and explain the data formats you have chosen
- use open formats
- use standard formats
- use machine-readable formats
- find out if the format allows automatic metadata insertion
- check if the repositories support the selected formats
- recommended data formats:
 - [File Formats. Open Data Handbook](#)
 - [File Formats. Data Archiving and Networked Services](#)

Data volume

- estimate the data volume at the end of the project. It implicates several aspects:
 - preservation
 - access
 - backup
 - data exchange
 - hardware and software
 - technical support
 - expenses

Software

- use open-source software when possible
- open-source software keeps hardware and software costs low
- interoperable with other open-source software
- the software is developed and supported by a large community (higher quality, security and modernization; unfortunately, limited documentation and support)
- software should allow to repeat the data analyzes carried out

- documentation when new software is created
- provide technical support for tailored software
- version management system [git](#)
- cloud-based code repository [GitHub](#)
- open-source software licenses
 - [Choose an open-source license](#)
- are there any existing data that you can reuse?

How will the data be collected or created?

Collecting data

- I'll collect it myself
- I buy the data
- keep in mind:
 - which version of data you reuse or purchase
 - what if the author of the data uploads a new version?
 - store the version used and the vendor documentation on your server
 - check copyrights, licenses, restrictions (access, reuse)
 - check machine readability and interoperability with the planned information system
- name the existing standard procedures and methods
- are there any data standards available?
- how to ensure data quality (availability, integrity, confidentiality)
- how do you handle errors (input errors, problematic values)
- (re)use my previously collected data
- I use public open data ([Estonian Open Data Portal](#))
- (re)using data collected by others, ([re3data](#), [Mendeley Data](#), [DataCite Search](#))

Organization of data

- be systematic and consistent
- naming files: simple, logical, without abbreviations or with standard abbreviations (countries, languages, units of measurement, methods)
- abbreviations in one language throughout
- file organization (options: project name, time, place, collector, material type, format, version)
- folder structure should be hierarchical, simple, logical, short
- copying files to multiple locations is not a good practice; store in one location, create shortcuts
- version control system [git](#)
- cloud-based code repository [GitHub](#)
- metadata (who is responsible for adding metadata)
- article:
 - [Data Organization in Spreadsheets](#)

DOCUMENTATION AND METADATA

What documentation and metadata will accompany the data?

Data documentation

- use this guide for data documentation:
 - Siiri Fuchs, & Mari Elisa Kuusniemi. (2018, December 4). Making a research project understandable - Guide for data documentation (Version 1.2). Zenodo.
DOI: <http://doi.org/10.5281/zenodo.1914401>
- a README text file is included with the data files and should contain as much information as possible about the data files to allow others to understand the data.
 - create one README.txt file for each database
 - always name it as README.txt or README.md (Markdown), not readme, ABOUT, etc.
- The README.txt file should contain the following information:
 - title of the dataset
 - dataset overview (abstract)
 - file structure and relationships between files
 - methods of data collection
 - software and versions used
 - standards
 - specific information about data (units of measurement, explanations of abbreviations and codes, etc.)
 - possibilities and limitations of data reuse
 - contact information for the uploader of the dataset
 - [Guidelines for creating a README file](#)

Metadata

- administrative metadata, project details (ID, funder, rights and licences)
- technical metadata (hardware and software, instruments, tools, access rights)
- descriptive metadata (author, title, abstract, subject terms)
- DataCite Metadata (mandatory, recommended, optional metadata) on [DataCite Metadata Schema 4.4](#)
- metadata standards indicate which fields should be filled:
 - [Directory of Metadata standards](#)
 - [Dublin Core](#) (kasutusel EMU DSpace-is)
 - [Schema.org](#)
 - [DCAT](#)
- controlled metadata dictionaries and classifications tell you what to write in these fields, using standard terminology
 - examples:
 - [Estonian Subject Thesaurus](#)
 - [AGROVOC Multilingual Thesaurus](#)

- [Mammal Species of the World](#)
- [GeoName](#)

ETHICS AND LEGAL COMPLIANCE

How will you manage any ethical issues?

Research integrity

- Estonian Research Council: [“Guidelines for Completing Your Ethics Self-Assessment for Grant Application” 2023](#)
- in case the project has no ethical issues mentioned in the guide, it should also be mentioned in the application
 - [Estonian Code of Conduct for Research Integrity](#)
 - [Good Academic Practice and Implementation of Principles of Academic Ethics in the Estonian University of Life Sciences](#)
 - [Annex 1. Establishing the Academic Ethics Committee, the principles, and organization of activities, the procedure for proceeding a suspicion and a complaint of a breach](#)
 - [Annex 2. Submitting an application](#)

How will you manage copyright and Intellectual Property Rights (IPR) issues?

Copyright and intellectual property rights

- who owns the data (personal and proprietary rights [IPR \(Intellectual Property Rights\)](#))
- data always has an owner, even if it is open data
- how data is licensed
- [Procedure for intellectual property created at the Estonian University of Life Sciences 2021](#)
- how data is licensed [Creative Commons licenses](#)
- are there any restrictions on the reuse of third-party data?
- will data sharing be postponed / restricted e.g., to publish or seek patents?

Personal data protection

- describe here whether the project collects personal data and how it is processed following the [General Data Protection Regulation](#) and the [Estonian Personal Data Protection Act](#)
- [Personal Data Processing Policy at Estonian University of Life Sciences](#)

STORAGE AND BACKUP

How will the data be stored and backed up during the research?

- The goal is to maintain data quality:
 - availability and accessibility

- integrity (correctness, completeness and timeliness)
- confidentiality (only available to authorized persons or systems, key management, storage of log files)
- storage:
 - cloud environments
 - central servers
 - sensitive data servers
 - hard disk drive
 - external hard drive
 - mobile devices
- backup: creating a copy of the current status of data and/or programs that, after an security incident, allows you to restore it to its known current state
 - maintaining and backing up the master file
 - rule 3-2-1 (store your data in 3 copies on 2 different memory devices from which 1 is afar)
 - who is responsible, especially for mobile devices
- carry out a risk analysis: what if
 - IT systems are down
 - power outages, water and fire accidents
 - the device is lost or stolen
 - malware is discovered in devices
 - a team member leaves or dies, etc.
- risk weighing (probability and losses)
- risk assessment: threats and their likelihood, weaknesses, measures
- information security standard ISO / IEC 27001

How will you manage access and security?

- management of access rights (same for all, contractual rights, temporary labor rights)
- storing log files
- pseudonymization, encryption, key management
- data exchange, personal data, third countries
- organizational and physical security: training of a new employee, possible problems with the outgoing workers, internal rules of procedure, fire safety, locking the doors
- who is responsible for information security

SELECTION AND LONG-TERM PRESERVATION

Which data are of long-term value and should be retained, shared, and/or preserved?

What is the long-term preservation plan for the dataset?

FAIR Data

- what data has long-term value? Preserving and sharing it for reuse
- preparing data for sharing, FAIR data
- repository selection

How to make data findable (F)

- the data have a permanent identifier DOI. See [DataCite Estonia](#)
- metadata is in the DataCite registry
- standard metadata like [Dublin Core](#) or use other standards
- machine-readable metadata
- data and relevant metadata are in separate files but linked
- keywords and subject terms
- version management

How to make the data accessible (A)

- choose the repository where the data is stored
- which data is open access e. open data
- which data will remain closed and for what reason
- metadata must be open even when the data is not open (exceptions like rare species location)
- technical metadata: required software (version), instrument specifications, software tools

How to make data interoperable with other computer systems (I)

- mainly the task of the repository
- what data and metadata standards, controlled vocabularies and taxonomies are used
- description of data types: if not standard, how interoperability is ensured
- linking to other data, metadata, and specifications
- data exchange standards

How to ensure data reusability (R)

- partly a task of the repository
- add [README.txt](#) file
- is it raw, cleaned or processed data
- embargo period, grounds
- licenses [Creative Commons licenses](#)
- citing: [DataCite Citation Formatter](#)
- standard metadata, which (domain) standards are used
- provenance of the data (who, where, what, when, published)
- which software version is used
- how long is the data available for re-use
- data quality assurance (availability, integrity, confidentiality)
- suggestions who might need this data (in [README.txt](#))

DATA SHARING

How will you share the data?

- is the data shared in a repository, or as a supplementary data of an article, or as a separate data article in a data journal?
- in which repository is the data stored
- who might find this data useful?
- how do you share your data (open data, or you have to ask for data)
- when do you share (at once, after publication of the article, after embargo period)?
- is the data linked to a publication?
- link to your [ORCID](#) account

Are any restrictions on data sharing required?

- which data is open access, open data?
- which data will remain closed and for what reason?
- any encrypted data
- authentication, who gives access rights
- whether you need to create a user account under certain terms

RESPONSIBILITIES AND RESOURCES

Who will be responsible for data management?

- by positions
 - principal investigator (PI): Data Management Policy, DMP, contracts, costs, training
 - researchers: follow and improve DMP, data management, problem solving
 - data manager: training, consulting, information security, backup, hardware and software
 - laboratory assistant, support staff: according to their tasks
- by workflow
 - who is responsible for data collection, documentation, metadata, data security, etc.?
- an example
 - [TU Delft RD Policy](#)

What resources will you require to deliver your plan?

- costs are mainly related to manpower, hardware and software
- guides, training, lawyer and/or DPO consultation, translation service
- data collection: purchase of data, transcription of recorded interviews
- digitization and OCR: hardware and software, manpower
- software development or software purchase, user licenses

- hardware: computers, servers, instruments, field work equipment
- data analysis: hardware and software, outsourced services
- data storage and backup: predictable data volume, rule 3-2-1
- long-term storage of data: preparation for sharing (formatting), anonymization
- data storage in a repository
- partner meetings, conferences
- project data manager
- consideration: 5% of the project budget