

EJP RD

European Joint Programme on Rare Diseases

H2020-SC1-2018-Single-Stage-RTD
SC1-BHC-04-2018
Rare Disease European Joint Programme Cofund



Grant agreement number 825575

Del 1.24

Initial Data Management Plan

Organisation name of lead beneficiary for this deliverable:

Partner 1 – INSERM

Due date of deliverable: month 06

Dissemination level:

Public

Table of Contents

INTRODUCTION	3
1. Data collection and documentation	3
1.1. What data will you collect, observe, generate or re-use?	3
1.2. How will the data be collected, observed or generated?	6
1.3. What documentation and metadata will you provide with the data?	9
2. Ethics, legal and security issues	12
2.1. How will ethical issues be addressed and handled?	12
2.2. How will data access and security be managed?	16
2.3. How will you handle copyright and Intellectual Property Rights issues?	19
3. Data storage and preservation	21
3.1. How will your data be stored and backed-up during the research?	21
3.2. What is your data preservation plan?	23
4. Data sharing and re-use	25
4.1. How and where will the data be shared?	25
4.2. Are there any necessary limitations to protect sensitive data?	27
4.3. All digital repositories I will choose are conform to the FAIR Data principles	28
4.4. I will choose digital repositories maintained by a non-profit organisation.....	28
5. APPENDICES	29
5.1. Appendix 1 – RESEARCH DATA: the basics	29
5.2. Appendix 2 – FAIR DATA PRINCIPLES ¹	30
5.3. Appendix 3 – File formats	31
5.4. Appendix 4 – METADATA.....	32
5.5. Appendix 5 – Code	33
5.6. Appendix 6 – Data masking	34
5.7. Appendix 7 – Storage, publication and preservation	35

INTRODUCTION

This Data Management Plan (DMP) is a reference document that describes how data of the European Joint Programme on Rare Diseases (EJP RD) is acquired, generated and managed during the lifecycle and what mechanisms will be used at the end of the project to share and preserve data.

The DMP also completes the EJP RD results (with the information on data, software, protocols, sources, etc.); allows the anticipation of costs (materials and software) for a better allocation of resources and identifies risks (e.g. data loss, incompatible formats, security). It boosts data reuse by reducing the risks of data loss and the efforts of reverse engineering for new collaborators.

Overall the DMP targets the reproducibility of research results and anticipate questions about data in EJP RD.

The DMP is a "living" document, annual updates will be performed.

1. DATA COLLECTION AND DOCUMENTATION

1.1. WHAT DATA WILL YOU COLLECT, OBSERVE, GENERATE OR RE-USE?

Questions you might want to consider:

- What type, format and volume of data will you collect, observe, generate or reuse?
- Which existing data (yours or third-party) will you reuse?

Briefly describe the data you will collect, observe or generate. Also mention any existing data that will be (re)used. The descriptions should include the type, format and content of each dataset. Furthermore, provide an estimation of the volume of the generated datasets.

(This relates to the FAIR Data Principles F2, I3, R1 & R1.2)

Recommendations:

For each dataset in your project, **including data you might re-use**, mention:

- € **Data type:** briefly describe categories of datasets you plan to generate or use, and their role in the project (code and scripts are considered as data) [See Appendix 3]
- € **Data origin** if you are reusing existing data (yours or third-party one). Add the reference of the source if relevant

- € **Format of raw data** (as created by the device used, by simulation or downloaded): open standard formats should be preferred, as they maximize reproducibility and reuse by others and in the future [see Appendix 3]
- € **Format of curated data** (if applicable): open standard formats should be preferred [see Appendix 3]
- € **Estimation of volume of raw and curated data**

Example of response:

Example 1:

The data produced from this research project will fall into two categories:

1. The various reaction parameters required for optimization of the chemical transformation.
2. The spectroscopic and general characterization data of all compounds produced during the work.

Data in category 1 will be documented in [file format].

Data collection and documentation

Spectroscopic data in category 2 will be produced as [file format] and converted to [file format] for further use.

Other characterization data in this category will be collected in [file format].

We anticipate that the data produced in category 1 will amount to approximately 10 MB and the data produced in category 2 will be in the range of 4 - 5 GB.

Example 2:

This project will work with and generate three main types of raw data.

1. Images from transmitted-light microscopy of giemsa-stained squashed larval brains.
2. Images from confocal microscopy of immunostained whole-mounted larval brains.
3. Western blot data.

All data will be stored in digital form, either in the format in which it was originally generated (i.e. Metamorph files, for confocal images; Spectrum Mill files, for mass spectra with results of mass spectra analyses stored in CSV files; TIFF files for gel images; MariaDB SQL dump files for genetics records), or will be converted into a digital form via scanning to create tiff or jpeg files (e.g. western blots or other types of results).

Measurements and quantification of the images will be recorded in excel files (for long term preservation, they will be converted in CSV files. Micrograph data is expected to total between 100GB and 1TB over the course of the project. Scanned images of western blots are expected to total around 1GB over the course of the project. Other derived data (measurements and quantifications) are not expected to exceed 10MB.

Example 3:

The data are health records auto-generated by users of the application X. They are subjected to a contract with the company X.

All fields contain user observations and entered manually, except for temperature which is measured by a Bluetooth connected thermometer.

Data fields per user (anonymized by X): User identifier; Age; Weight, Size.

Data fields per users per day of observation:

- € Temperature and time at which temperature is taken,

- € Cervical fluid quality (none, sticky, creamy, egg white, watery) and quantity (little, medium, lots),
- € Cervix height (low, med, high), cervix openness (closed, med, open), cervix firmness (firm, med, soft),
- € Sexual intercourse (protected or unprotected),
- € Menstruation (light, medium, heavy), spotting, starting a new cycle,
- € Custom data (notable predetermined fields are pregnancy test results or ovulation test results).

Data will be received in CSV format and consists of the record of 2 million users. It will amount to maximum 1GB.

Example 4: from EAWAG DMP

There will be two categories of data: NEW data from this project and EXISTING data from the FOEN Lake Monitoring program.

The NEW data will consist of several file types, all CSV real number format, which are all organized along the same principle: matrixes of times series with various channels, each corresponding to a sensor (number of sensors varies from 1 to 10) and very different length, as the sampling frequency varies by several orders-of-magnitudes.

- € 6 files of CO2, DO, PAR and temperature (24 files at a time; Figure 2), each file only 1 sensor (Delta = 10 min; continuous),
- € Thetis profiles corresponding to time series (equivalent to depth series) of 10 sensors (Delta = 1 s; 5-10 times per day).
- € 5 files of CO2 time series for short-term surface flux measurements (several files, one per month),
- € Meteo data file (eight sensors; continuous),
- € T-Microstructure profiles files (6 sensors at 512 Hz; several files, once per month) and
- € excel files for individual chemical samples (such as alkalinity, sediment trap estimates, etc.; sporadic).

The EXISTING data is already available (CIPAI, CIPEL) in excel sheets with matrices for the individual samplings and a variable number of parameters (~10 to ~25). The EXISTING data will not be modified and remains with the organizations. We will keep a copy on our computers during the project. We anticipate the data produced in category 1 to amount to several hundred MB for the moored and profiled sensor files and ~100 GB for the T-microstructure profiles; the EXISTING data in category 2 is in the range of ~20 MB.

1.2. HOW WILL THE DATA BE COLLECTED, OBSERVED OR GENERATED?

Questions you might want to consider:

- What standards, methodologies or quality assurance processes will you use?
- How will you organize your files and handle versioning?

Explain how the data will be collected, observed or generated. Describe how you plan to control and document the consistency and quality of the collected data: calibration processes, repeated measurements, data recording standards, usage of controlled vocabularies, data entry validation, data peer review, etc.

Discuss how the data management will be handled during the project, mentioning for example naming conventions, version control and folder structures. (This relates to the FAIR Data Principle R1)

Recommendations:

What standards, methodologies or quality assurance processes will you use?

For **each dataset** in your project (including data you might re-use) mention:

- € the use of EJPRD services if applicable,
- € the use of standards or internal procedures; describe them briefly.

If you are working with **personal data**, confirm the following:

- € have the subjects of your data collection (persons) been fully informed (what data do you collect, what will you do with the data, and who will receive it; when will they be deleted) and have the subjects given their informed consent?
- € have the subjects of your data collection (persons) been informed about their rights on information, data deletion and data correction?

How will you organize your files and handle versioning?

Indicate and describe the tools you will use in the project. You may rely on the following tools depending on your needs:

- € **Naming convention**, i.e. the structure of folders and file names you will use to organize your data.

For example: "Project_Experiment_Scientist_YYYYMMDD_HHmm_Version.format"
(concretely: Atlantis_LakeMeasurements_Smith_20180113_0130_v3.csv)

- € **Code revision management system**, such as Git.
- € **Data management system**, such as an Electronic Laboratory Notebook / Laboratory Information System (ELN/LIMS).

Example of response:

Example 1:

The reaction conditions will be recorded and collated using a spreadsheet application and named according to each generation of reaction as follows:

ProjectW_ReactionX_GenerationY_ScientistZ_YYYYMMDD_HHmm.csv

The various experimental procedures and associated compound characterization will be written up using the Royal Society of Chemistry standard formatting in a Word document, each Word document will also be exported to PDF-A. The associated NMR spectra will be collated in chronological order in a PDF-A document.

Example 2:

All samples on which data are collected will be prepared according to published standard protocols in the field [*cite reference*]. Files will be named according to a pre-agreed convention. The dataset will be accompanied by a README file which will describe the directory hierarchy.

Each directory will contain an INFO.txt file describing the experimental protocol used in that experiment. It will also record any deviations from the protocol and other useful contextual information.

This should allow the data to be understood by other members of our research group and add contextual value to the dataset should it be reused in the future.

Example 3:

Experiments will include appropriate controls to ensure validity [*brief description*]. Data consistency will be assessed by comparing repeated measures.

Example 4:

Quality of analytical data will be guaranteed through calibration of devices, repetition of experiments, comparison with literature/internal standards/previous data, by a peer review.

Example 5:

All experimental data will be automatically imported into the institutional electronic Laboratory Information System (LIMS) from the measurement device. Methods and materials will be recorded using the institutional Electronic Lab Notebook (ELN).

Example 6:

The experimental records and observations are recorded by hand-written notes followed by digitization (scanning). The analytical data are collected by the instruments that generated them; they are processed by the native programs [*please specify program name, version and file format*] associated with the instruments. A periodic quality control process will be applied to remove errors and redundancies. Errors include for example incorrect handling and machine malfunction. The quality control process will be documented.

The quality of experimental records and observations will be controlled by repeating experiments.

For NMR and X-ray, the data collection is done through instrument standardized data acquisition programs. For E-chem, UV-Vis, IR, GC, GC-MS, lab-standardized protocols will be used.

Example 7: from EAWAG DMP

The data from the moored sensors is sensor-internally stored and recovered every two months, when sensors will be cleaned and recalibrated if data indicates quality loss.

The CO₂ sensors will be cross calibrated against atmospheric pressure.

The DO and PAR sensors in the mooring will be compared to profiled sensors and deviations detected.

Temperature sensors are extremely stable and are only calibrated before and after the two years using the laboratory temperature bath which is calibrated against the Office of Metrology in Bern every few years to 0.001°C.

The Thesis sensor data is transmitted when surfacing via GSM communication system directly to the lab where sensors deterioration is weekly checked. The instrument will be retrieved every month and sensors cleaned.

The optical sensors will be calibrated according the manual every six months.

The T-microstructure sensors do not need calibration as the data is matched to (very accurate) CTD temperature. Small T shifts are irreverent, as only the spectra matter.

The sensors deterioration (or frequency loss) will visually be checked and is seen in the quality of the Batchelor spectra.

The very simple structure of the CSV files holding the raw data will be documented in a plain text README file. This file, and all raw data files as they become available, will be uploaded to the Eawag Research Data Institutional Collection into one "data package", which is annotated with general metadata. Copies of the raw data files as well as set of calibrated, quality-controlled files stored on the group computers at EPFL will be organized in a folder structure that is also documented in a README file. At the end of the project, the entire set of calibrated, quality-controlled files will be annotated and stored on the Eawag institutional repository as well.

1.3. WHAT DOCUMENTATION AND METADATA WILL YOU PROVIDE WITH THE DATA?

Questions you might want to consider:

- What information is required for users (computer or human) to read and interpret the data in the future?
- How will you generate this documentation?
- What community standards (if any) will be used to annotate the (meta)data?

Describe all types of documentation (README files, metadata, etc.) you will provide to help secondary users to understand and reuse your data. Metadata should at least include basic details allowing other users (computer or human) to find the data. This includes at least a name and a persistent identifier for each file, the name of the person who collected or contributed to the data, the date of collection and the conditions to access the data.

Furthermore, the documentation may include details on the methodology used, information about the performed processing and analytical steps, variable definitions, references to vocabularies used, as well as units of measurement.

Wherever possible, the documentation should follow existing community standards and guidelines. Explain how you will prepare and share this information. (This relates to the FAIR Data Principles I1, I2, I3, R1, R1.2 & R1.3)

Recommendations:

Indicate all the information required in order to be able to read and interpret the data (context of data) in the future. General documentation of the data is often compiled into a plain text or [markdown](#) README file. These formats may be opened by any text editor and are future proofed.

Also check Appendix 4.

In addition, for each data type:

- € Provide the **metadata standard** used to describe the data (for concrete examples see: [Research Data Alliance Metadata Standards Directory](#)).

If no appropriate (discipline oriented) existing standard is available, you may describe the *ad hoc* metadata format you will use in this section.

Metadata may also be embedded in the data (e.g. embedded comments for code). Or, when for example using Hierarchical Data Format [HDF5](#), arbitrary machine-readable metadata can be included directly at any level.

(Metadata refers to "data about data", i.e., it is the information that describes the data that is being published with sufficient context or instructions to be intelligible for other users. Metadata must allow a proper organization, search and access to the generated information and can be used to identify and locate the data via a web browser or web-based catalogue).

- € Describe:
 - € the **software** (including its **version**) used to produce the data and the software used to read it (they can be different),
 - € the format and corresponding filename extension and its version (if possible).

The used software should be archived along with the data (if possible, depending on the software license).

- € Describe the automatically generated metadata, if any.
- € Provide the data analysis or result together with the raw data, if possible.

Additional information that are helpful in a README file:

- € description of the used **software**,
- € description of the used **system environment**,
- € description of relevant **parameters** such as:
 - o geographic locations involved (if applicable),
 - o all relevant information regarding production of data.

Example of response:

Example 1:

The data will be accompanied by the following contextual documentation, according to standard practice for synthetic methodology projects:

1. Spreadsheet documents which detail the reaction conditions.
2. Text files which detail the experimental procedures and compound characterization.

Files and folders will be named according to a pre-agreed convention YXZ [please state your convention], which includes for each dataset, identifications to the researcher, the date, the study and the type of data (see section 1.2).

The final dataset as deposited in the chosen data repository will also be accompanied by a README file listing the contents of the other files and outlining the file-naming convention used.

Example 2:

Metadata will be tagged in XML using the Data Documentation Initiative (DDI) format. The codebook will contain information on study design, sampling methodology, fieldwork, variable-level detail, and all information necessary for a secondary analyst to use the data accurately and effectively.

It will be responsibility of:

- € each researcher to annotate data with metadata,
- € the Principal Investigator to check weekly (during the field season, monthly otherwise) with all participants to assure data is being properly processed, documented, and stored.

Example 3:

IFS and OpenIFS model integrations will be run and standard meteorological and computing performance data output will be generated. Both will be run at ECMWF, and only performance data will be made available to the public. The meteorological output will be archived in MARS, as it is standard research experiment output. The data will be used for establishing research and test code developments and will enter project reports and generally accessible publications. The IFS will not be made available, OpenIFS is available through a dedicated license.

IFS meteorological output (incl. metadata) and format follows the World Meteorological Organization (WMO) standards. Compute performance (benchmark) output will be stored and documented separately. Data will be in ASCII and maintained locally. The output will be reviewed internally, and the ECMWF facilities allow reproduction of this output if necessary.

Example 4:

Two types of metadata will be considered within the frame of the project [*name of the project*]: that corresponding to the project publications, and to the published research data. In the context of data management, metadata will form a subset of data documentation that will explain the purpose, origin, description, time reference, creator, access conditions and terms of use of a data collection.

The metadata that would best describe the data depends on the nature of the data. For research data generated in project [*name of the project*], it is difficult to establish a global criteria for all data, since the nature of the initially considered data sets will be different, so that the metadata will be based on a generalized metadata schema as the one used in Zenodo, which includes elements such as:

- € Title: free text
- € Creator: Last name, first name
- € Date
- € Subject: Choice of keywords and classifications
- € Description: Text explaining the content of the data set and other contextual information
- € needed for the correct interpretation of the data,
- € Format: Details of the file format,
- € Resource Type: data set, image, audio, etc.,
- € Identifier: DOI,
- € Access rights: closed access, embargoed access, restricted access, open access.

Additionally, a readme.txt file could be used as an established way of accounting for all the files and folders comprising the project and explaining how all the files that make up the data set relate to each other, what format they are in or whether particular files are intended to replace other files, etc.

Example 5: from EAWAG DMP

For every data stream (sequences of identical data files) over the entire 2-year period of data acquisition a README File will be generated which contains: (a) the sensors used (product, type, serial number), (b) the temporal sequence of the sensors (time and location, sampling interval), (c) the observations made during maintenance and repairs, and (d) details on the physical units, as well as the calibration procedure and format. This is a standard procedure which we have used in the past.

2. ETHICS, LEGAL AND SECURITY ISSUES

2.1. HOW WILL ETHICAL ISSUES BE ADDRESSED AND HANDLED?

Questions you might want to consider:

- What is the relevant protection standard for your data? Are you bound by a confidentiality agreement?
- Do you have the necessary permission to obtain, process, preserve and share the data? Have the people whose data you are using been informed or did they give their consent?
- What methods will you use to ensure the protection of personal or other sensitive data?

Ethical issues in research projects demand for an adaptation of research data management practices, e.g. how data is stored, who can access/reuse the data and how long the data is stored. Methods to manage ethical concerns may include anonymization of data; gain approval by ethics committees; formal consent agreements. You should outline that all ethical issues in your project have been identified, including the corresponding measures in data management. (This relates to the FAIR Data Principle A1)

Recommendations:

Description and management of ethical issues

- ⊘ Describe which **ethical issues** are involved in the research project (for example, human participants, collection/use of biological material, privacy issues (confidential/sensitive data), animal experiments, dual use technology, etc.).
- ⊘ Explain how these ethical issues will be managed, for example:
 - The necessary ethical authorizations will be obtained from the competent ethics committee.
 - Informed consent procedures will be put in place.
 - Personal/sensitive data will be anonymized.
 - Access to personal/sensitive data will be restricted.
 - Personal/data will be stored in a secure and protected place.
 - Protective measures will be taken with regard to the transfer of data and sharing of data between partners.
 - Sensitive data is not stored in cloud services (e.g. data related to individuals, data under a non-disclosure agreement, data injuring third party rights or legal expertise).

Please check if your project involves data relating to one of the following ethical issues:

- ⊘ Human participants (this includes all kinds of human participation, incl. non-medical research, e.g. surveys, observations, tracking the location of people)

- € Human cells/tissues
- € Human embryonic stem cells
- € A clinical trial
- € The collection of personal/sensitive/confidential data
- € Animal experimentation
- € Developing countries (access and benefit sharing)
- € Environmental and/or health and safety issues (for example, a negative impact on the environment and/or on the health and safety of the researchers)
- € The potential for military applications (dual-use technology).

If you consider that there are no ethical issues in your project, you can use the following statement:

“There are no ethical issues in the generation of results from this project”.

Ethical authorizations:

If your project involves **human subjects**, an ethical authorization from the relevant “national” authority/institution is needed. This depends on whether your project is invasive/non-invasive and whether or not health-related data is collected/used.

- € For research involving work with **human cells/tissues**, a description of the types of cells/tissues used in the project needs to be provided, together with copies of the accreditation for using, processing or collecting the human cells or tissues.
- € Research which involves the **collection or use of personal data** needs to be reviewed by the ethics committee (depending on what kind of data is involved).
- € If **animal experiments** are conducted in the context of the research project, an appropriate authorization must be obtained (if applicable) (e.g. authorization of the veterinarian office).
- € **Dual-purpose technologies** (civil and military): transfer of knowledge, software, demonstrators or prototypes could fall under the scope of the Control of Dual-Use Goods, Specific Military Goods and Strategic Goods in the context of technology transfer or research proposals, but also in informal personal contacts. Before transmission of information, research results, prototypes, etc. to a company, person or institution (even academic), it must be checked whether the data/information to be transmitted are apt to authorization.

Research that may have a **negative impact on the environment**, for example research with Genetically Modified Organisms (GMO) may require an authorization from an Office for the Environment or equivalent according to the beneficiary national law. If the research project has a negative impact on the **health and safety of the researchers** involved (for example if the research proposal involves the use of

elements that may cause harm to humans), authorizations for the processing or possession of harmful materials must be requested.

Example of response:

Example 1:

This project will generate data designed to study the prevalence and correlates of DSM III-R psychiatric disorders and patterns and correlates of service utilization for these disorders in a nationally representative sample of over 8000 respondents. The sensitive nature of these data will require that the data be released through a restricted use contract, to which each respondent will give explicit consent. An ethical authorization will be obtained from the ethics committee for this project.

Example 2:

All data are anonymized, and as such, we are in line with the "National" Act on Data Protection as described on the webpage page of the "[link to Data Protection National Authority](#)".

Example 3:

The project respects all the constraints and requirements as laid down in the "National" Act on Data Protection and supervised by the "[National" Data Protection and Information Commissioner](#).

Indeed, as the finality of the project does not relate to individuals and the published results do not allow to identify the participants nominatively, we have communicated with all participants giving them the following basic information:

- € Author/ Responsible person
- € Type and extent of collected/processed data
- € Objectives of the processing
- € Any communication to be made to third-parties /recipients categories / planned trans borders communications, with all necessary guaranties related to "[Cite the relevant article](#)"
- € The facultative nature of the participation to this project and the possibility to resign at all times
- € The consequences, if any, in case of refusal to participate (No inconvenience should result)
- € Access and correction rights

Example 4:

The project is a medical research project and respects all the rules and regulations laid down in the "National" Act on Data Protection and supervised by the "[National" Data Protection and Information Commissioner](#). We are only using and processing data for individuals who have given their explicit consent.

Example 5: from EAWAG DMP

Dataset X was obtained from the BAFU and is subject to a confidentiality agreement to keep information about the sampling locations secret. We are allowed to share this information among researchers involved in the project. The dataset is being stored in a location to which only project members have access. Please refer to [Section 2.2](#) for technical details about access restrictions. All project members will be informed about sensitivity of this data and

agree not to copy it to other places. This dataset and intermediate datasets containing the sampling locations will be excluded from the data package published along with the final report and replaced with instructions about how to obtain them from the BAFU.



2.2. HOW WILL DATA ACCESS AND SECURITY BE MANAGED?

Questions you might want to consider:

- What are the main concerns regarding data security, what are the levels of risk and what measures are in place to handle security risks?
- How will you regulate data access rights/permissions to ensure the security of the data?
- How will personal or other sensitive data be handled to ensure safe data storage and transfer?

If you work with personal or other sensitive data you should outline the security measures in order to protect the data. Please list formal standards which will be adopted in your study. An example is ISO 27001-Information security management. Furthermore, describe the main processes or facilities for storage and processing of personal or other sensitive data. (This relates to the FAIR Data Principle A1)

Recommendations:

The main concerns regarding data security are data availability, integrity and confidentiality.

- € Define whether:
 - the level of the data availability risk is: low/medium/high.
 - the level of data integrity risk is: low/medium/high.
 - the level of data confidentiality is: low/medium/high.
- € You may choose some of the following options:
 - *Regarding anonymization / encryption:*
 - All personal data will be anonymized in such a way that it will be impossible to attribute data to specific persons.
 - All personal data will be pseudonymized. The correspondence table will be encrypted and access restricted to the project leader.
 - All sensitive data will be encrypted and encryption keys will be managed only by authorized employees.
 - Sensitive data transfers will be end-to-end encrypted.
 - *Regarding access rights:*
 - Sensitive data will be accessible only by authorized participants to the project. The list of authorized participants will be managed by...
 - Data access rules will be detailed in before starting the project.
 - Access to the data/database will be logged, thus each access is traceable.
 - Access to laboratory and offices will be restricted to authorized persons. The list of authorized persons will be managed by...

- Regarding storage and back-up:
 - 1 - Have you planned on where to store your data?
 - Yes No
 - 2 – If no, contact EJP RD to request help for storage for research data
 - Yes No
 - 3 – Are you planning on using another storage system?
 - Yes No
 - 4 - If yes, have you checked:
 - The pricing?
 - The confidentiality?
 - The ability and conditions to transfer your data to another location?

Example of response:

Example 1:

The data will be processed and managed in a secure non-networked environment using virtual desktop technology.

Example 2:

All interviewees and focus group participants will sign a Consent form agreed to by the "institution" ethics committee. We have guaranteed anonymity to our interviewees and focus group participants. Therefore, we will not be depositing .wav files as this would compromise that guarantee. However, anonymized transcripts of the interviews and focus groups will be deposited. We will make sure consent forms make provision for future sharing of data. All identifying information will be kept in a locked filing cabinet and not stored with electronic files.

Example 3:

Data will be stored on the centralized file storage system managed by our institutional [please specify, e.g. life sciences or basic sciences] IT department. The access to the data is managed through the "institution" identity management system, which is a secured system following the best practices in terms of identity management. Our central storage facility has redundancy, mirroring and is monitored.

Example 4: from EAWAG DMP (modified from see below)

Research records will be kept confidential, and access will be limited to the PI, primary research team members, and project participants. Data will be housed on a local server controlled by the PI and will be accessible via SSH and VPN. Data containing identifiable information, or information covered by an NDA, will be held in an encrypted format (symmetric, AES256, key on local server, passphrase only know to PI and primary research team members).

Example 5: from EAWAG DMP

The data we are generating, processing and storing in this project does not pose a particular data security risk. Day-to-day work is conducted on standard-issue workstations in the "institution"-environment with standard enterprise-grade access control. The institution network is a secured system following the best practices in terms of identity management and central storage facility has redundancy, mirroring and is monitored. At different stages,

data will be stored in the Eawag Institutional Collection (see section 1.3). This system is accessible only from within the Eawag network and is comprised of several virtualized Linux systems that receive real-time security patches. Access control is handled according to recognized best practices of server administration.



2.3. HOW WILL YOU HANDLE COPYRIGHT AND INTELLECTUAL PROPERTY RIGHTS ISSUES?

Questions you might want to consider:

- Who will be the owner of the data?
- Which licenses will be applied to the data?
- What restrictions apply to the reuse of third-party data?

Outline the owners of the copyright and Intellectual Property Right (IPR) of all data that will be collected and generated including the license(s). For consortia, an IPR ownership agreement might be necessary. You should comply with relevant funder, institutional, departmental or group policies on copyright or IPR. Furthermore, clarify what permissions are required should third-party data be re-used. (This relates to the FAIR Data Principles I3 & R1.1)

Recommendations:

Attaching a clear license to a publicly accessible dataset allows other to know what can legally be done with its content. When copyright is applicable, [Creative Commons licenses](#) are recommended. This applicability of copyright goes to data which has itself a creative content (e.g. photos) or databases which are the result of a creative work (e.g. innovative collection of data) as well as to final data, which underlies scientific publications. However, a database composed of raw data with engineering values (temperatures, resistances, voltages...) would not qualify as copyrightable. Creative Commons licenses are not recommended for software.

Regarding code licenses, you can use:

- € [GNU-GPL](#) (Open Software)
- € [Apache 2.0](#) (smaller codes, libraries)
 - o Permissive
 - o No share alike clause
 - o Preservation of copyright notice
- € [3clause BSD](#)

[Comparison between each type of licenses](#)

Amongst all Creative Commons licenses, CC0 "no copyright reserved" is recommended for scientific data, as it allows other researchers to build new knowledge on top of a data set without restriction. It specifically allows aggregation of several data sets for secondary analysis. Several data repositories impose the CC0 license to facilitate reuse of their content.

In order to enable a data set to get cited, and therefore get recognition for its release, it is recommended to attach a CC-BY "Attribution" license to the record, usually a description of the dataset (metadata). To get recognition, data sets can be cited directly. However, to increase their visibility and reusability, it is recommended to describe them in a separated document licensed under CC-BY "Attribution", such as a data paper or on the institutional repository. When the data has the potential to be used as such for commercial purposes, and that you intend

to do so, the license CC-BY-NC allows you to keep the exclusive commercial use.

Reuse of third-party data may be restricted. If authorized, the data must be shared according to the third party's original requirement or license. If you have any questions regarding this point, do not hesitate to contact the TTO: andrea.crottini@epfl.ch.

If you need guidance in the publication and license choice, you can check the suggested "[Data publication decision tree](#)".

Example of response:

Example 1:

The research is not expected to lead to patents. Other Intellectual Property Rights (IPR) issues will be dealt in line with the institutional recommendation. As the data is not subjected to a contract and will not be patented, it will be released where possible as open data under Creative Commons CC0 license.

Example 2:

This project is being carried out in collaboration with an industrial partner. The intellectual property rights are set out in the collaboration agreement. The intellectual property generated from this project will be fully exploited with help from the institutional Technology Transfer Office (TTO). The aim is to patent the final procedure and then publish the work in a research journal and to publish the supporting data under an open Creative Commons Attribution (CC-BY) license.

Example 3:

Data is suitable for sharing. They are observational data (hence unique) and could be used for other analyses or for comparison for climate change effects among many things. Reuse opportunities are vast. For this reason, we aim to allow the widest reuse possible of our data and will release them under Creative Commons CC0.

Example 4: from EAWAG DMP

The source code for analysis will most likely utilize the GNU Scientific Library (GSL), which is licensed under the GNU General Public License (GPL). Therefore, we will make our analysis software available under the GPL as well.

3. DATA STORAGE AND PRESERVATION

3.1. HOW WILL YOUR DATA BE STORED AND BACKED-UP DURING THE RESEARCH?

Questions you might want to consider:

- What is your storage capacity and where will the data be stored?
- What are the back-up procedures?

Please mention what the needs are in terms of data storage and where the data will be stored.

Please consider that data storage on laptops or hard drives, for example, is risky. Storage through IT teams is safer. If external services are asked for, it is important that this does not conflict with the policy of each entity involved in the project, especially concerning the issue of sensitive data.

Please specify your back-up procedure (frequency of updates, responsibilities, automatic/manual process, security measures, etc.)

Recommendations:

See [Appendix 7](#)

Example of response:

Example 1:

Storage and back up will be in three places:

- € On Laptop of [Name of Researcher]
- € On institutional collaborative storage
- € Other (please specify)

[Name of Researcher] will be responsible for the storage and back up of data. This will be done weekly. Backups on the institutional infrastructure are automated using the RSYNC tool.

Example 2:

Original notebooks and hardcopies of all NMR and mass spectra are stored in the PI's laboratory. Additional electronic data will be stored on the PI's computer, which is backed up daily. Additionally, the laboratory will make use of the PI's lab server space at institution's storage facility for a second repository of data storage. The PI's lab has access to up to 1 terabyte of information storage, which can be expanded if needed.

All the project data will be stored using the institution's Collaborative Storage, which is backed-up on a regular basis.

Example 3:

All our data will be uploaded to our Electronic Laboratory Notebook (*please specify which one, and where it is installed*). The data is stored on institutional storage facilities and it is set up by our IT support to be automatically backed up daily.

Example 4:

The (*institution*) centralized file storage service follows the best practices and standards regarding storage, for instance high availability, multiple levels of data protection, partnership with providers for support. The service is managed centrally by the hosting department of the Vice Presidency for Information Systems (VPSI) and ensures security, coherence, pertinence, integrity and high availability.

Two distinct storage locations can be found on the (*institution*) campus with replication between the two. (Please note that these two different storage options correspond to different payments.) Physical servers' pairing and clustering guarantees local redundancy of data. Moreover, volume mirroring protects data in case of disaster on the primary site. The copy is asynchronous and automatic and runs every two hours.

The file servers are virtualized for separation between logical data and physical storage, RAID groups ensure physical storage protection: data is split in chunks written on many disks with double parity. Moreover, volume snapshots are used and can allow user restoration of previous versions if need be. For specific needs, optional backup on tape can also be done.

Access to the data is managed by the owner of the volumes through the identity management system of (*institution*). Any person who needs access to data has therefore to be a registered and verified user in the identity management system.

Example 5: from EAWAG DMP

Our team stores the data to be analyzed along with the results using Eawag file services. To easily share data with our collaborators in Fribourg, we synchronize those data with a folder on SWITCHdrive. Since this is sensitive personal data, the folder being synchronized contains encrypted files (public key encryption, key-pairs specifically created for this project).

3.2. WHAT IS YOUR DATA PRESERVATION PLAN?

Questions you might want to consider:

- What procedures would be used to select data to be preserved?
- What file formats will be used for preservation?

Please specify which data will be retained, shared and archived after the completion of the project and the corresponding data selection procedure (e.g. long-term value, potential value for re-use, obligations to destroy some data, etc.). Please outline a long-term preservation plan for the datasets beyond the lifetime of the project.

In particular, comment on the choice of file formats and the use of community standards.

Recommendations:

Describe the procedure, (appraisal methods, selection criteria ...) **used to select data to be preserved**. Note that preservation does not necessarily mean publication (e.g. personal sensitive data may be preserved but never published), but publication means generally preservation.

This section should answer the following questions:

- € What data will be preserved in the long term - **selection criteria**, in particular:
 - **Reusability of the data**: quality of metadata, integrity and accessibility of data, license allowing reuse, readability of data (chosen file formats)
 - **Value of the data**: indispensable data, completeness of the data or data set, uniqueness, possibility to reproduce the data in the same conditions and at what cost, interest of the data, potential of reuse
 - **Ethical considerations**
 - **Stakeholders requirements**
 - **Costs**: additional costs that come for depositing data in a repository or data archive of your choice (costs anticipation and budgeting)

Selection basically has to be done together with or by the data producer or someone else with deep specialist knowledge.

- € What data curation process(es) will be applied, i.e.: anonymization (if necessary), metadata improvement, format migration, integrity check, measures to ensure accessibility.
- € Data retention period (0, 5, 10, 20 years or unlimited)
- € Decision to make the data public
- € Use of sensitive data (i.e. privacy issues, ethics, or intellectual property laws)
- € Definition of the responsible person for data (during the process of selection and after the end of the project)

Other **criteria from the Digital Curation Center** (UK). In addition, select appropriated preservation formats (see section 1.1) and data description or metadata (see section 1.3).

See [Appendix 7](#).

Example of response:

Example 1:

Data will be stored for a minimum of three years beyond award period, per funder's guidelines. If inventions or new technologies are made in connection data, access to data will be restricted until invention disclosures and/or provisional patent filings are made with the institutional Technology Transfer Office.

Example 2:

Where possible, we will store files in open archival formats e.g. Word files converted to PDF-A or simple text files encoded in UTF-8 and Excel files converted to CSV. In case this is not possible, we will include information on the software used and its version number.

Example 3:

Data will be stored on EPFL servers and will be preserved for the long term on [Zenodo](#)

4. DATA SHARING AND RE-USE

4.1. HOW AND WHERE WILL THE DATA BE SHARED?

Questions you might want to consider:

- On which repository do you plan to share your data?
- How will potential users find out about your data?

Consider how and on which repository the data will be made available. The methods applied to data sharing will depend on several factors such as the type, size, complexity and sensitivity of data.

Please also consider how the reuse of your data will be valued and acknowledged by other researchers.

(This relates to the FAIR Data Principles F1, F3, F4, A1, A1.1, A1.2 & A2)

Recommendations:

This part depends on whether your data are sharable. If you are in doubt, when patents could be considered for instance, contact the institutional Technology Transfer Office.

It is recommended to **publish data in well established** (or even certified) domain specific **repositories**, if available:

- € [re3data](#) is a repository directory allowing to select repositories by subject and level of trust (e.g. certifications)

Researchers are strongly encouraged to use disciplinary repositories when they exist. In domains for which no suitable subject repositories are available, generalist repositories are available.

Among the most common used:

- € [Zenodo](#) (free, maximum 50GB/dataset, hosted by CERN)
- € [Dryad](#) (Non-profit organization)

[Figshare](#) (free upload, maximum 5GB /dataset, commercial company)

Example of response:

Example 1:

Some of the ongoing data will be shared on [Researcher]'s Github repository (results and code from the project, data from twitter searches). Major revisions of this page will be backed up using the Github-Zenodo connection (see: <https://guides.github.com/activities/citable-code/>). All other data we will be, where possible and if no further exploitation can be made, published on Zenodo under CC0 license.

We chose Zenodo because it supports the FAIR principles (<http://about.zenodo.org/principles/>). The immediate publication at the end of the project aims to minimize the data loss risk, while the 2 years embargo guarantees us to be first to exploit our data. Zenodo implements long-term preservation features, notably bitstream preservation.

Example 2:

For this project, the National Geoscience Data Centre (NGDC) is the most suited repository. As it is adapted to geodata, it facilitates storage and allows interactive geographical search. In addition, many other researchers in our field are familiar with it.

This repository requires the deposition under Open Government Licence (see: <http://www.nationalarchives.gov.uk/doc/open-government-licence/version/3/>), which demands attribution when the data is reused (our dataset must be cited, similarly to the CC-BY license).



4.2. ARE THERE ANY NECESSARY LIMITATIONS TO PROTECT SENSITIVE DATA?

Questions you might want to consider:

- Under which conditions will the data be made available (timing of data release, reason for delay if applicable)?

Data have to be shared as soon as possible, but at the latest at the time of publication of the respective scientific output.

Restrictions may be only due to legal, ethical, copyright, confidentiality or other clauses.

Consider whether a non-disclosure agreement would give sufficient protection for confidential data.

(This relates to the FAIR Data Principles A1 & R1.1)

Recommendations:

You may mention specifically the conditions under which the data will be made available:

- € there are no sensitive data
- € the data are not available at the time of publication
- € the data are not available before publication
- € the data are available after the embargo of ...
- € the data are not available because of the patent of ... for a period of...

Example of response:

Example 1:

Data which underpins any publication will be made, wherever possible, available at the time of publication.

All unpublished data will be deposited in a data repository 12 months after the end of the award, if there are no other exploitation to be made.

Example 2:

Astronomical data will be diffused but under an embargo of one year for priority of exploitation reasons.

Example 3: from EAWAG DMP

The extensive household survey about water-borne diseases poses severe challenges with regard to anonymization, since simple pseudonymization might not be sufficient to guard against the identification of individual households by an inference attack that uses other available information.

Therefore, we will be only able to publish summary statistics together with the associated article. If a sufficiently anonymized dataset turns out to still hold scientific value, we will publish it no later than one year after completion of the project.

4.3. ALL DIGITAL REPOSITORIES I WILL CHOOSE ARE CONFORM TO THE FAIR DATA PRINCIPLES Yes**Recommendations:**

It is required that repositories used for data sharing are conformed to the FAIR Data Principles

You can find certified repositories in [Re3data.org](https://re3data.org), an exhaustive registry of data repositories.

4.4. I WILL CHOOSE DIGITAL REPOSITORIES MAINTAINED BY A NON-PROFIT ORGANISATION Yes No**Recommendations:**

If you do not choose a repository maintained by a non-profit organization, you have to provide reasons for that.

One possible reason would be to ensure the visibility of your research, if your research community is standardly publishing data on a well-established but commercial digital repository.

5. APPENDICES¹

5.1. APPENDIX 1 – RESEARCH DATA: the basics

RESEARCH DATA DEFINITIONS

- ✓ Material generated or collected during the course of conducting research¹.
- ✓ Factual records used as primary sources for scientific research, commonly accepted in the scientific community as necessary to validate research findings².
- ✓ Information collected, observed, or created, for purposes of analysis to produce original research results³.
- ✓ Any information in binary digital form derived from the research process⁴.

RESEARCH DATA LIFECYCLE

- 1 Creating / Re-using:** planning data collection, locating existing data sources; producing, collecting or documenting data.
- 2 Processing / Analyzing:** validating, cleaning, transforming data; creating metadata; using, creating analysis tools; interpreting the data.
- 3 Preserving / Publishing:** reviewing the data; getting data into a format suitable for preservation; depositing data and metadata in archive / repository; promoting data re-use.



RESEARCH DATA TYPES

- **Observational Data:** data captured in-situ, can't be recaptured, recreated or replaced. Examples: Sensor readings, sensory (human) observations, survey results, interview notes, transcripts
- **Experimental Data:** data collected under controlled conditions, in situ or laboratory-based, should be reproducible, but can be expensive. Examples: gene sequences, chromatograms, spectroscopy, microscopy
- **Simulation Data:** result from using a model to study the behaviour and performance of an actual or theoretical system, models and metadata, where the input can be more important than output data. Examples: climate models, economic models, biogeochemical models
- **Derived/Compiled Data:** reproducible, but can be very expensive. Examples: derived variables, compiled database, 3D models
- **Reference or canonical Data:** static or organic collection [peer-reviewed] datasets, most probably published and/or curated. Examples: gene sequence databanks, chemical structures, census data, spatial data portals⁵

Raw Data

Raw data refer to data that have not been changed since acquisition, eg. a real-time GPS-encoded navigation file, and the initial time-series file of temperature values from a heat probe.

Processed Data/Active Data

Editing, cleaning or modifying the raw data results in processed data, eg. raw multibeam data files can be processed to remove outliers and to correct sound velocity errors⁶.

Credits and sources

[1] <https://www.ed.ac.uk/information-services/research-support/research-data-service>

[2] <https://www.oecd.org/sti/sci-tech/38500813.pdf>

[3] <http://www.ed.ac.uk/information-services/research-support/data-management>


[4] <https://www.degruyter.com/view/product/430793>

[5] <http://guides.library.stonybrook.edu/research-data-services/types>

[6] http://www.marine-geo.org/help/data_FAQ.php

¹ <https://infoscience.epfl.ch/record/265349?&ln=en>

5.2. APPENDIX 2 – FAIR DATA PRINCIPLES¹

<p>Data and metadata are easy to find by both humans and computers.</p>	<p>Humans and computers can readily access or download datasets.</p>	<p>Data from different datasets are prepared to be combined or exchanged.</p>	<p>Published data can be easily combined or replicated in future research.</p>
<h1>F</h1> <h2>FINDABLE</h2>	<h1>A</h1> <h2>ACCESSIBLE</h2>	<h1>I</h1> <h2>INTEROPERABLE</h2>	<h1>R</h1> <h2>REUSABLE</h2>
<p>F1 [Meta]data are assigned a globally unique and persistent identifier.</p> <p>F2 Data are described with rich metadata.</p> <p>F3 Metadata clearly and explicitly include the identifier of the data they describe.</p> <p>F4 [Meta]data are registered or indexed in a searchable resource.</p> <p>DESCRIBE Describe provenance, usage and organization of data with standardized metadata (DataCite, RDA standards, DublinCore). Make metadata available even if data are not.</p>	<p>A1 [Meta]data are retrievable by their identifier using a standardized communication protocol:</p> <ul style="list-style-type: none"> A1.1 the protocol is open, free and universally implementable; A1.2 the protocol allows for an authentication and authorization procedure where necessary. <p>A2 Metadata are accessible, even when the data are no longer available.</p> <p>OPEN Open your data using standardized licenses (ex. Creative Commons). Limitations may apply to the openness (ex. embargo). Disclose files in open formats, even alongside proprietary formats.</p>	<p>I1 [Meta]data use a formal, accessible, shared and broadly applicable language for knowledge representation.</p> <p>I2 [Meta]data use vocabularies that follow FAIR principles.</p> <p>I3 [Meta]data include qualified references to other [meta]data.</p> <p>LINK Use persistent identifiers for datasets (ex. DOI, HANDL, URN) and tag all the metadata with the same identifiers. Cross-link datasets with linked-data standards (RDF).</p>	<p>R1 [Meta]data are richly described with a plurality of accurate and relevant attributes:</p> <ul style="list-style-type: none"> R1.1 [meta]data are released with a clear and accessible data usage license; R1.2 [meta]data are associated with detailed provenance; R1.3 [meta]data meet domain-relevant community standards. <p>PUBLISH Deposit datasets in data repositories, favoring services with user-friendly interfaces.</p>
<p>“Data should be as open as possible, as closed as necessary.” Carlos Moedas EU Commissioner</p>		<p>Did you know? 40% of researchers are aware of the existence of FAIR principles³ 20-50% increased citation for articles linked to associated data⁴</p>	
<p> How FAIR are your data? Take the FAIR self-assessment test²</p>			

Credits and sources

[1] FAIR principles: <https://www.go-fair.org/fair-principles>

[2] FAIR self-assessment tool: <https://www.gnds-nectar-rds.org.au/fair-tool>

[3] State of Open Data 2018: https://fiqshare.com/blog/State_of_Open_Data_2018/440

[4] Open Data Citation Advantage: <https://sparceurope.org/open-data-citation-advantage/>

5.3. APPENDIX 3 – FILE FORMATS

Definition

A **file format** is a standard way to encode data for storage in a computer file. It specifies how bits are used to encode information in a digital storage medium. File formats may be either proprietary or free and may be either unpublished or open¹.

When listing out the data formats you will be using, make sure to include:

- The necessary software to view the data (e.g. SPSS v.3; Microsoft Excel 97-2003).
- Information about version control.
- If data are stored in one format during collection and analysis and then transferred to another format for preservation: list out features that may be lost in data conversion such as system specific labels.

When selecting file formats for archiving, the formats should ideally be:

- Non-proprietary, unencrypted, uncompressed, commonly used by the research community.
- Compliant to an open, documented standard: interoperable among diverse platforms and applications, fully published and available royalty-free, fully and independently implementable by multiple software providers on multiple platforms without any intellectual property².

File formats extensions for reusability/preservation:

Type of data	APPROPRIATE	ACCEPTABLE	NOT SUITABLE
Tabular data with extensive metadata	.csv - .hdf5	.txt - .html - .tex - .por	
Tabular data with minimal metadata	.csv - .tab - .ods - SQL	.xml if appropriate DTD - .xlsx	.xls - .xlsb
Textual data	.pdf - .txt - .odt - .odm - .tex - .md - .htm - .xml	.pptx - .pdf with embedded forms - .rtf	.doc - .ppt
Code	.m - .R - .py - .iypnb - .rstudio - .rmd - NetCDF	.sdd	.mat - .rdata
Digital image data	.tif - .png - .svg - .jpeg	jpg - .jp2 - .tif - .tiff - .pdf - .gif - .bmp	.indd - .ait - .psd
Digital audio data	.flac - .wav - .ogg	.mp3 - .mp4 - .aif	
Digital video data	.mp4 - .mj2 - .avi - .mkv	.ogm - .webm	.wmv - .mov
Geospatial data	NetCDF, tabular GIS attribute data, .shp - .shx - .dbf - .prj - .sbx - .sbn - PostGIS - .tif - .tfw - GeoJSON	.mdb - .mif	
CAD/vector and raster data	.x3d - .x3dv - .x3db - PDF3D .pdf	.dwg - .dxf	
Generic data	.xml - .json - .rdf		

Credits and sources

[1] https://en.wikipedia.org/wiki/File_format

[2] <https://library.stanford.edu/research/data-management-services/data-best-practices/best-practices-file-formats>

5.4. APPENDIX 4 – METADATA

“Metadata is structured information associated with an object for purposes of discovery, description, use, management, and preservation”

[National Information Standards Organization, 2008]

METADATA IS
UBIQUITOUS AND
PROLIFERATIVE

METADATA IS
EMBEDDED
OR SUPPLEMENTAL

METADATA RESULT
FROM AUTOMATIC
OR MANUAL INPUT

INTEROPERABILITY
IS BASED ON
METADATA

● **Technical metadata**
[ex. version of producing device]

5

● **Use metadata**
[ex. number of downloads]

● **Administrative metadata**
[ex. publishing date, rights and licenses]

METADATA FAMILIES

● **Descriptive metadata**
[ex. title, author, keywords]

● **Preservation metadata**
[ex. last checksum date]

From Excel to databases and semantic web knowledge bases, the more metadata you have, the better **data management system** you need.

FAIR data, good quality linked [open]data, mainly relies on rich, detailed, qualified, shared, standardized metadata.

HOW TO?

1. Be systematic, adopt rules, use controlled values
2. Describe your data completely and consistently
3. Use standards

Metadata and metadata standards creation, adoption and maintenance is a **JOINT EFFORT** within and between interest-based communities.

TOOLS TO BUILD YOUR OWN STRONG METADATA

FORMAT, TECHNICAL, INTERCHANGE STANDARDS : [exif](#), [IPTC](#), instrumentation specific standards...

VALUE NORMS, STANDARDS AND REFERENCES : [ISO 8601](#), [ISO 639-1](#), [ISO 3166-1](#), thesaurii, vocabularies, lists of authorities...

CONTENT MODELS AND STANDARDS : [ISA \[Investigation-Study-Assay\] framework](#), [Force11 Software citation principles](#)

STRUCTURE STANDARDS AND SCHEMAS : [INSPIRE](#), [SDMX](#), [Darwin Core](#), [Dublin Core](#), [PROV model](#), [Datacite](#)

More resources

[1] <http://www.dcc.ac.uk/resources/metadata-standards/list>

[2] <http://rd-alliance.github.io/metadata-directory/standards>

5.5. APPENDIX 5 – CODE

When working with code, good practices are also needed. In particular the publication of code is needed in order to understand, reuse and repeat the operation.

TIPS AND TRICKS FOR A BETTER EFFICIENCY IN CODE MANAGEMENT

● VERSIONING

Versioning systems are powerful tools for code management. The most used is **Git**, it's free and open :

- It allows to **track changes** and to undo changes if needed. You can manage easily different versions of your code
- Connected to a repository your code and its modifications are **automatically backedup**
- You can also **work in team** easily on the same code

● SHARING

In order to **share your code and make it visible**, repositories provide various services like version management system, wikis, task management and issues tracking, one of the most known is **Github**.

● DESCRIBING

README documentation is a really important part of coding. It allows you to **explain your code**, for you and others. You should add rich metadata and documentation (README, LICENSE, comments on code...) on any publication of the code.

Some tools like sphinx-doc.org and doxygen.nl can help you by going through your code and generating a preformatted documentation.

● LICENSING

It is important to explain **how your code can be used** by others (and related restrictions). You have at least three options :

- Open source licenses (permissive as MIT or GPL)
- Academic licenses (restrict commercial usage)
- Commercial licenses (reserve commercial usage)

● PUBLISHING

Don't forget to **generate a DOI** to uniquely identify a version of your software and to easily cite it.

Most code repository (like Zenodo or c4science) generate a DOI for your deposit.

TIP : Github provides an integration with Zenodo.

● PRESERVING

Preservation is important for keeping your work secure and also for scientific validation.

C4science is a solution to **preserve your code** for the long term. If you are using another code repository, you can **always make a copy on c4science for preservation**.

5.6. APPENDIX 6 – DATA MASKING



ADVANTAGES WHY IT'S WORTH

- Complies with law
- Makes data sharable
- Prevents data misuse
- Makes data publishable

APPLICABILITY

TESTS ON HUMANS / SENSITIVE DATA

- Name, identification number, location data, online identifier, etc.
- Factors specific to the physical, physiological, genetic, mental, economic, cultural or social identity

TECHNIQUES

PSEUDONYMIZATION

REVERSIBLE
(FOR WORKING DATA)

REPLACING

Replace data by identifiers. The key is stored separately and securely.

ENCRYPTING

Encrypt the data and store the key securely. Appropriate for long-term preservation, not for data publishing.

ANONYMIZATION

IRREVERSIBLE
(FOR PUBLISHED DATA)

GENERALIZING

Diminish granularity by generalizing the variables. Appropriate for data too specific or unique records.

SHUFFLING

Shuffle data over one / several columns without compromising their utility.

FAKING

Prevent the identification of specific records, adding fake data while preserving correlations.

REMOVING

Suppress data or part of the outlier records. Appropriate for processing identifiers.

3RD PARTY DATA

Using commercial datasets or collaborating in a joint research? Then, define a **contract** for data sharing or publication, and distinguish between research **authors** and data **owners**.



HINT

Mitigate the identification risk, but preserve the data utility for research.

SOME TOOLS

TO MASK IDENTITY OR ASSESS IDENTIFICATION RISKS

- [ARX Data Anonymization Tool \[Java\]](https://arx.deidentifier.org)¹
- [Amnesia \[online\]](https://amnesia.openaire.eu)²
- [ARGUS \[Java\]](https://aosient.com/argus/anonymization.shtml)³
- [sdcMicro \[R\]](https://cran.r-project.org/web/packages/sdcMicro/index.html)⁴
- [Differential privacy queries \[SQL\]](https://github.com/uber-archive/sql-differential-privacy)⁵
- [Faker \[Python\]](https://faker.readthedocs.io/en/master/)⁶

SUPPORT AND LAWS



Credits and sources

[1] <https://arx.deidentifier.org>

[2] <https://amnesia.openaire.eu>

[3] <https://aosient.com/argus/anonymization.shtml>

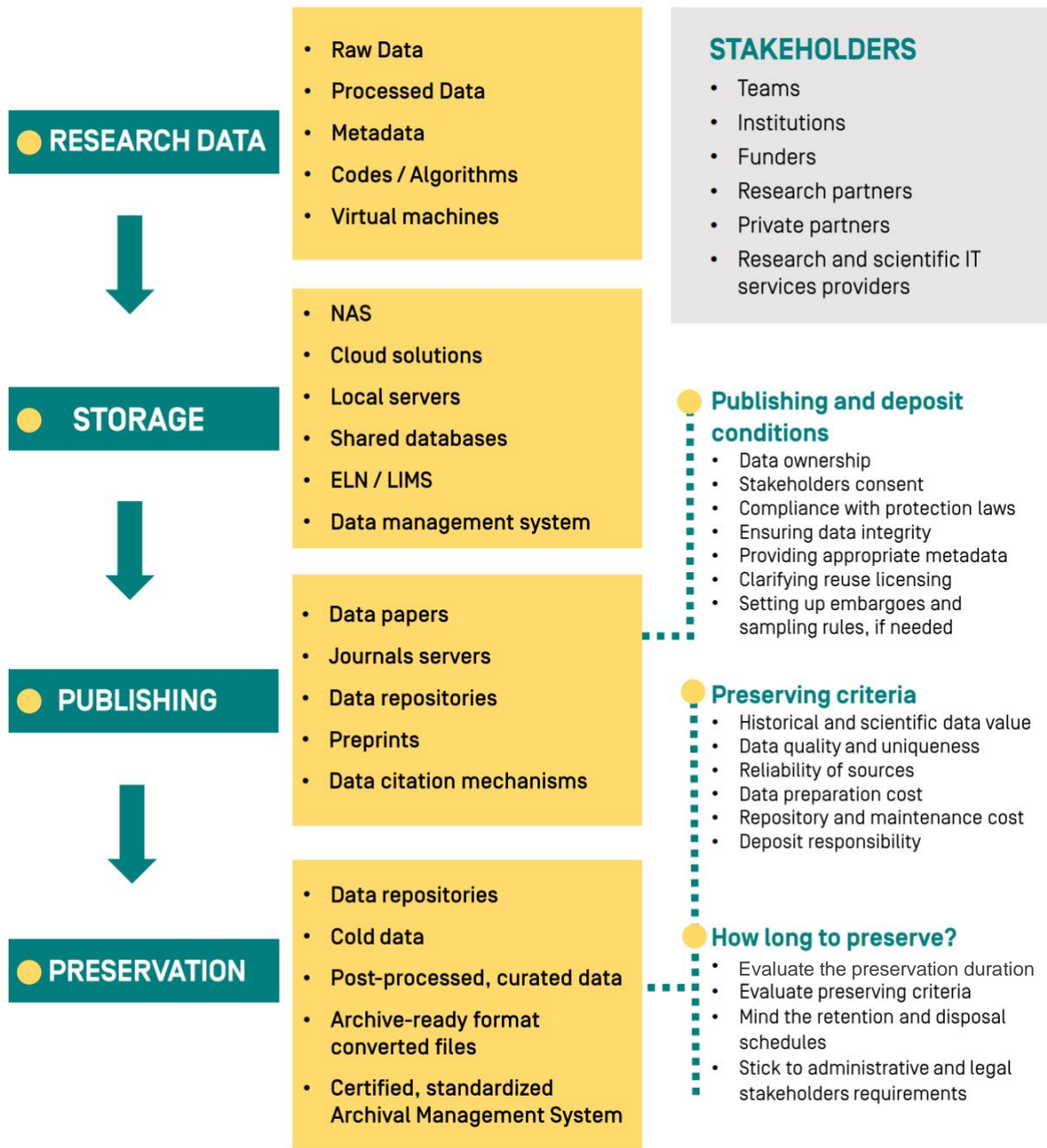
[4] <https://cran.r-project.org/web/packages/sdcMicro/index.html>

[5] <https://github.com/uber-archive/sql-differential-privacy>

[6] <https://faker.readthedocs.io/en/master/>

[7] <https://eur-lex.europa.eu/eli/reg/2016/679/oj>

5.7. APPENDIX 7 – STORAGE, PUBLICATION AND PRESERVATION



NAS: Network Attached Storage
ELN: Electronic Laboratory Notebook
LIMS: Laboratory Information Management System