EPFL - Ecole Polytechnique Fédérale de Lausanne: EPFL SNSF with examples hosted on web site

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)

Recommandations:

EPFL 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 EPFL Library FastGuide #04 "File formats"*]
- **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 EPFL Library FastGuide #04 "File formats" or List of recommended file formats*]
- **Format of curated data** (if applicable): open standard formats should be preferred [see *EPFL Library FastGuide #04 "File formats" or List of recommended file formats*]
- **Estimation of volume of raw and curated data**

Exemple de réponse:

Examples

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)

Exemple de réponse:

Examples

Recommandations:

EPFL 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 EPFL core facility services (specify their certifications, if any),
- whether you follow double blind procedures (define it),
- 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. Several Git servers are available for ETH domain: ch, gitlab.epfl.ch, gitlab.ethz.ch.
- **Data management system**, such as an Electronic Laboratory Notebook / Laboratory Information System (ELN/LIMS). Within ETH domain, examples of used ELN/LIMS: openBIS, SLims.

Contact for assistance: EPFL Research Data Library team researchdata@epfl.ch

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)

Recommandations:

EPFL 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 EPFL Library Fastguide #05 “Metadata”.

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:
- geographic locations involved (if applicable)
- all relevant information regarding production of data.

Contact for assistance: EPFL Research Data Library team (researchdata@epfl.ch)

Exemple de réponse:

Examples

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)

Recommandations:

EPFL recommendations

Take a look at EPFL Library Fastguide #08 "Personal Data Management". If you need any help, you can contact EPFL Research Office (research@epfl.ch).

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.).

For more information, see the EPFL Research Ethics website.

- 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 (in bold) 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 either the cantonal ethics commission or the institutional ethics commission (EPFL-HREC) is needed. This depends on whether your project is invasive/non-invasive and whether or not health-related

data is collected/used.

More information can be found on the EPFL Research Ethics website or contact the Research Office in case of doubt which ethics commission is the competent authority to review the project (research@epfl.ch).

- 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. Please also see: https://www.epfl.ch/research/ethic-statement/compliance/.
- Research which involves the **collection or use of personal data** needs to be reviewed by the cantonal ethics commission or the institutional EPFL ethics committee (depending on what kind of data is involved). For more information, see the relevant page on the EPFL Research Ethics website or contact the Research Office (research@epfl.ch).
- If **animal experiments** are conducted in the context of the research project, an authorization of the cantonal veterinarian office is needed. (See also: https://www.epfl.ch/research/ethic-statement/compliance/).
- **Dual-purpose technologies** (civil and military): At EPFL, transfer of knowledge, software, demonstrators or prototypes could fall under the scope of the Federal Act on the Control of Dual-Use Goods, Specific Military Goods and Strategic Goods (Goods Control Act, CGA) and its Ordinance (OCB) 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) outside of Switzerland, it must be checked whether the data/information to be transmitted are apt to authorization.
- Please see: https://www.epfl.ch/research/ethic-statement/compliance/ and/or contact the EPFL Legal Counsels of the Technology Transfer Office (TTO) (research@epfl.ch).

Research that may have a **negative impact on the environment**, for example research with Genetically Modified Organisms (GMO), requires an authorization from the Federal Office for the Environment (FOEN). 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. More information can be obtained from the EPFL Safety, Prevention and Health Domain (DSPS) and https://www.epfl.ch/research/ethic-statement/compliance/.

Contact for assistance: EPFL Research Office (research@epfl.ch)

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)

Exemple de réponse:

Examples

Recommandations:

EPFL recommendations

Take a look at EPFL Library Fastguide #08 "Personal Data Management". If you need any help, you can contact EPFL Research Office (research@epfl.ch).

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 researchdata@epfl.ch to request help for storage for research data

3 – If you plan to use the central storage facilities (NAS.epfl.ch), have you checked the price for your necessary storage from the VPSI?

☐ Yes ☐ No

4 – Are you planning on using another storage system?

☐ Yes ☐ No

5 - If yes, have you checked:

☐ The pricing?

☐ The confidentiality?

☐ The ability and conditions to transfer your data to another location?

See also EPFL Library FastGuide #03 "Cost of Research Data Management" and EPFL Library Cost Calculator for Data Management.

Contact for assistance: EPFL: Research Data Library team (researchdata@epfl.ch)

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 licence(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)

Recommandations:

If you need any help, you can contact EPFL Research Office (research@epfl.ch).

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. artistic 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)
    - Permissive
    - No share alike clause
    - 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 authorised, 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".

Contact for assistance: EPFL: Research Data Library team (researchdata@epfl.ch)

Exemple de réponse:

Examples

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.)

Exemple de réponse:

Examples

Recommandations:

EPFL recommendations

See EPFL Library FastGuide #10 “Storage, preservation and publication”, FastGuide #03 “Cost of Research Data Management” and EPFL Library Cost Calculator for Data Management.

Institutional storage solutions:

For EPFL you can find the institutional storage option on this web page: https://support.epfl.ch/help?id=epfl_service_status&service=49a363acdb34c700ef64731b8c96191f

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.

Recommandations:

EPFL 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 EPFL Library FastGuide #10 “Storage, preservation and publication”.

Contact for assistance: EPFL: Research Data Library team (researchdata@epfl.ch)

Exemple de réponse:

Examples

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)

Recommandations:

EPFL recommendations

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

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)
- ETH Zurich researchers are encouraged to publish data in ETH’s own Research Collection repository to ensure full compliance with ETH regulations.
- EPFL Library provides support to guide researchers in the choice of an appropriate (disciplinary) repository.

A project for an EPFL institutional repository is currently ongoing (expected for 2020).

However, EPFL strongly encourages its researchers 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 (120$ for the first 20GB and 50$ for additional GB, Non-profit organization)
- Figshare (free upload, maximum 5GB / dataset, commercial company)

Note: SNSF does not pay for storage in commercial data repositories (even though data preparation costs are eligible). Check the SNSF's criteria for non-commercial repositories (http://www.snf.ch/en/theSNSF/research-policies/open_research_data/Pages/data-management-plan-dmp-guidelines-for-researchers.aspx, section 5.2). If you choose a commercial repository, read carefully the Terms of service to check if they respond to your needs and to your institutions' ones as well as to your institutional (data) policy. See also EPFL Library Cost Calculator for Data Management.

Exemple de réponse:

Examples

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)

Exemple de réponse:

Examples

Recommandations:

EPFL 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…

Contact for assistance: EPFL: Research Data Library team (researchdata@epfl.ch)

4.3 All digital repositories I will choose are conform to the FAIR Data principles

- Yes

Recommandations:

The SNSF requires that repositories used for data sharing are conformed to the FAIR Data Principles. For more information, please refer to the SNSF's explanation of the FAIR Data Principles.

You can find certified repositories in Re3data.org, an exhaustive registry of data repositories.

4.4 I will choose digital repositories maintained by a non-profit organisation

- No
- Yes

Recommandations:

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.

Please note that the SNSF supports the use of non-commercial repositories for data sharing. Costs related to data upload are only covered for non-commercial repositories. Check the SNSF's criteria for non-commercial repositories (http://www.snf.ch/en/theSNSF/research-policies/open_research_data/Pages/data-management-plan-dmp-guidelines-for-researchers.aspx, section 5.2).