Guidance for SNSF DMP template

Purpose

This document is specifically aimed at guiding **EPFL researchers** in understanding what and how to write a Data Management Plan (DMP)[1] for SNSF or other funders using the SNSF DMP template. Drafts can be prepared as documents based on the SNSF template[2] or using DMPonline.[3] The DMP must be submitted to SNSF via the mySNF platform.[4]

The main goal of a DMP is to plan your data strategy ahead of your project in order to **anticipate your needs** in terms of **resources** (servers, hard drives, data curation and preservation, software, etc.) and **good practices** (naming convention, data security, standardized metadata, open science, storage and back-up, etc.), saving time on the long run:

- Data losses will be prevented (either because of hardware failure or lack of documentation)

- Data will be easily reusable by new collaborators or for future projects (due to standardized documentation following the FAIR principles[5])

- Standardized procedures enable any collaborator to understand your processes.

The DMP is required by SNSF, but it is not part of the scientific evaluation process of the project. Its content is assessed for its plausibility and adherence to the SNSF guidelines[6] on Open Research Data (ORD).

Besides, your DMP **is not set in stone**: "The DMP remains editable during the entire lifetime of the grant. Its contents can be adapted as the project evolves." Since you cannot know in advance the exact workflow of your data management, it is recommended to **update** the DMP if the workflow is no longer aligned with the DMP description.

---

[1] https://go.epfl.ch/rdm-fastguide11

[2] https://go.epfl.ch/dmp-snsf

[3] https://dmponline.dcc.ac.uk

[4] https://www.mysnf.ch

[5] http://www.go-fair.org/fair-principles

[6] http://www.snf.ch/en/theSNSF/research-policies/open_research_data/Pages/data-management-plan-dmp-guidelines-for-researchers.aspx

Contents

The current document is the EPFL Library version[7] 6.1, revised in March 2023 by the EPFL Library. The content of this document has been co-written by EPFL Library Research Data team with REO, DPO, TTO, DSI, STI-IT, OSU.

For DMP review or information, please contact researchdata@epfl.ch.

---

[7] The previous version 5.0 by EPFL Library Research Data team and ETH-Bibliothek Digital Curation Team is available under CC-BY-SA license at https://infoscience.epfl.ch/record/295149.

1. Data collection and documentation

1.1 What data will you collect, observe, generate or re-use?

Describe the Research Data you plan to collect, observe, generate or re-use within the research project. Include all datasets that will be created, including but not limited to raw data, processed data, and final data sets. **Software code, algorithms, and protocols are also considered Research Data.**

It is recommended to create a list of datasets (e.g., dataset A, dataset B, dataset C), so you can refer to them in the following sections of the DMP.

For each dataset, list the following:

- **Type**: e.g., imaging data, survey results, simulation code, ML model assets, etc.
- **Format**: e.g., PDF, CSV, TIFF, TXT, etc. A dataset might include more than one format. It is recommended to use open standard formats. If not possible, always select the most widely used format or the format that is supported and used by the applicable research community.
- **Origin and/or Source**: Describe the primary point (e.g., instrument, software, publication, etc.) where the Research Data were generated/collected. Use references, if suitable. If applicable, mention the person/organization/authority who provided the Research Data.
- **Volume**: estimate its volume in MB, GB, etc.

Finally, also indicate the estimated **total volume**.

More information

- EPFL Library FastGuide #01 – Research Data: The Basics[8]
- EPFL Library FastGuide #04 – File Formats[ibid.]
- EPFL Library FastGuide #06 – Code as Data[ibid.]
- DANS file-formats overview[9]

---

[8] https://go.epfl.ch/rdm-fastguide
[9] https://dans.knaw.nl/en/file-formats

.2 How will the data be collected, observed, or generated?

Describe how you collect/observe/generate your data/code. In case you re-use **existing Research Data**, describe how you obtained it. If applicable, mention limitations and permissions, and cite applicable references (e.g., research paper, data provider guidelines, websites).

For each dataset, provide relevant details on how you organize your files, folder structure, and handle versioning. Describe your **naming convention** for folders and file names (for instance, you can use a generic example of naming).

For each dataset, describe the methodologies, standards or quality assurance processes used in the data/code creation/observation/collection. These might include (but are not limited to):

- **Conditions** (e.g., license, terms, agreement, contract) of re-use for pre-existing datasets.
- **Standard** (or internal) pre-processing procedures on acquired Research Data.
- Any previous **literature** explaining the simulation or algorithmic models used to collect/generate your Research Data.
- The **Standard Operating Procedures** (SOP) of the used facility/services. Specify their certifications, if any.

List all **software/code** used to handle your Research Data. For each one, if possible, mention their name, version, source, vendor (with reference websites), and license. Also, describe **how you use them** for your project. These tools might include (but are not limited to):

- **Code Versioning**: e.g., c4science,[10] gitlab.epfl.ch,[11] Git*, Apache Subversion, other code versioning tools
- **ELN/LIMS**: e.g., eln.epfl.ch,[12] SV LIMS,[13] STI RSpace,[14] other ELN tools
- **Computational Notebooks:** e.g., RENKU,[15] noto.epfl.ch,[16] other notebooks
- **Survey tool:** e.g., EPFL RedCap,[17]
- EPFL Google Drive Forms, other surveying tools
- **In-house code**
- **Generic or domain-specific software/code:** e.g., data visualization, analysis software, modelling

---

[10] https://c4science.ch

[11] http://gitlab.epfl.ch

[12] https://eln.epfl.ch

[13] https://www.epfl.ch/schools/sv/it/374-2/applications/eln-lims

[14] https://elnsti.epfl.ch

[15] https://renkulab.io

[16] https://noto.epfl.ch

[17] https://redcap.epfl.ch

More information

- EPFL Library FastGuide #07 – ELN - Electronic Lab Notebook[18]
- EPFL Library FastGuide #08 – Personal Data Management[16]
- Illinois University naming strategy overview[19] and manual versioning approach[20]
- Harvard Medical School ELN website[21]

1.3 What documentation and metadata will you provide with the data?

What information and documentation are required for users (computer or human) to read and interpret your datasets in the future?

For each dataset, indicate all the documentation you create/use that contains the information required to read and interpret your data/code. The documentation might include (but is not limited to):

- **README file** compiled into plain text or markdown. At a minimum, a README includes:[22]
    - authors, dates, links to relevant resources
    - description of relevant parameters range, e.g., geographic, materials characterization, simulation inputs, etc.
    - other relevant information regarding the production of Research Data.
- **Documentation on the software**/code needed to acquire/generate the data and the software used to manipulate/read it (they might be different). Be consistent with section 1.2. Such documentation might include (but is not limited to) instructions, manuals, online documentation, etc.
- **Other types of documentation** that facilitate the reuse of your dataset:
    - Parameter files, log files, protocols, debugging files, input files
    - Automatically generated code documentation (e.g., MkDocs, Swagger, pdoc3, etc.)

State in DMP how you generate such documentation.

Metadata refers to "data about data", i.e., the information that describes the data/code with sufficient context for other users or for your future self.

For each dataset (if possible) provide the metadata standard used to describe it. You may consider:

- Generic metadata standards, e.g., Dublin Core, DataCite. Data repositories such as Zenodo, Figshare, Dryad and many others use these standards.
- Discipline-specific standards
- You may describe the *ad hoc* metadata schema you adopt, if no appropriate (discipline-oriented) existing standard is available

---

[18] https://go.epfl.ch/rdm-fastguide

[19] https://emails.illinois.edu/newsletter/160474.html

[20] https://emails.illinois.edu/newsletter/1864897725.html

[21] https://datamanagement.hms.harvard.edu/analyze/electronic-lab-notebooks

[22] https://go.epfl.ch/rdm-readme

- Metadata may also be embedded in the data (e.g., embedded comments for code),
- A dictionary of some referenced scientific literature (e.g., a glossary).
- If you are using Hierarchical Data Format (HDF5), arbitrary machine-readable metadata can be included directly at any level.

More information

- EPFL Library vademecum and template of README file[23]
- EPFL Library FastGuide #05 – Metadata[24]
- FAIRsharing standards search engine[25]
- University of Illinois Research Data Documentation Examples[26]

---

[23] https://go.epfl.ch/rdm-readme
[24] https://go.epfl.ch/rdm-fastguide
[25] https://fairsharing.org/standards
[26] https://emails.illinois.edu/newsletter/1119537606.html

2. Ethics, Legal and Security Issues

2.1 How will Ethical Issues be Addressed and Handled?

If there are any ethical issues in your project or if you are unsure, consult the EPFL Research Ethics webpage. For support regarding ethics authorizations or ethics compliance inquiries, contact research@epfl.ch.

In this DMP section, describe which **ethical issues** can arise with the research project. Explain how you manage the ethical issues, e.g.:

- Ethical authorizations obtained from the competent ethics committee
- Informed consent form and related procedures in place
- Anonymized[27] or pseudo-anonymized[28] data (data masking)
- Controlled or restricted access to data (access rights management)
- Protective measures in place for data transfers, if applicable
- Data stored or shared using tools or services following apt security policies

Ethical issues can arise (but are not limited to) manipulating **data involving**:

- Human participants (incl. non-medical research, e.g., surveys, class observations, tracking of people's location, etc.)
  - *Personal data:*[29] e.g., name, date of birth, address, photo, video, IP address, GPS coordinates, biometric data, genomic data, etc.
  - *Sensitive data*: e.g., ethnic origin, political opinion, religious belief, trade union membership, health data, sexual orientation, etc.
  - *Vulnerable subjects*: e.g., children, elderly people, pupils/students in educational contexts, employees/collaborators, etc.
- Collection of personal/sensitive/confidential data
- Human cells/tissues or human embryonic stem cells
- Clinical trials
- Animal experimentation
- A confidentiality agreement (a binding contract/agreement in which you agree not to disclose certain information)
- Developing countries (access and benefit-sharing)
- Environmental and/or health and safety issues (e.g., a negative impact on the environment and/or on the health and safety of the researchers)
- The potential for military applications (dual-use technology)

Before proceeding to data collection and processing, note that:

- Not only data that can directly identify a person (e.g., name) is considered personal data, but also data that can make a person identifiable through the **combination of data** (for ex., workplace plus nationality and age).
- Research involving human participants or their personal data cannot be permitted unless the participant has given his/her **consent**: it is necessary to collect participants' consent before proceeding to data collection and processing.

---

[27] Anonymized: the personal identity is sufficiently protected by reducing the information, preventing anyone from re-identifying the person, unless they go to great lengths to do so.
[28] Pseudonymized: the identifying information (surname, date of birth, ...) is replaced by a code (e.g., hashing); the key to recovering the information, while kept secure, can still be used to identify the person.
[29] See Art. 5 FADP for definitions of personal and sensitive data https://www.fedlex.admin.ch/eli/oc/2022/491/fr.

- ▪ **Ethics authorization** from the competent committee (EPFL Human Research Ethics Committee - HREC, or Cantonal ethics commissions) is necessary for research involving human participants and/or personal or sensitive data.
- ▪ If you **reuse existing data**/samples, you might need to obtain ethical approval: find out the compliance requirements for data processing in your research at the EPFL Research ethics webpage.[30]

If your research project deals with personal or sensitive personal data, consult the Privacy in Research[31] EPFL website, contact the EPFL Human Research Ethics Committee[32] for ethics authorizations, or contact the EPFL Data Protection Officer (DPO)[33] about the potential legal requirements related to data processing.

> **Answer example**
>
> If there are no ethical issues in your project, you can use a statement similar to the following: "There are no ethical issues in the generation of results from this project".

More information

- ● Research Ethics Compliance[34] Guidance on assessment for research projects ethics
- ● Swiss law definition of particularly vulnerable persons, subject to special protection[35]
- ● EPFL Library FastGuide #08 – Personal Data Management[36]
- ● EPFL Library FastGuide #09 – Data Masking[33]
- ● DMLawTool decision-tree online tool[37]

2.2 How will data access and security be managed?

In this DMP section, describe: who has access to what dataset on which storage, and how such access rights are managed. Include all **datasets, hardware, online platforms, premises** (e.g., laboratory, office, facility, etc.), plus the roles and **responsibilities.**

Also, list all security concerns related to the handling of your datasets and describe them. If possible, reference existing guidelines, internal procedures, or applicable policies.

Provide factual information on:

- ▪ **Data availability**: Individuals and entities that have access to the data they need, within the maximum time frame agreed by the organization. These principles are embodied by technical and organizational measures, but also by practices to be

---

[30] https://www.epfl.ch/research/ethic-statement

[31] https://www.epfl.ch/campus/services/data-protection/in-practice/privacy-in-research

[32] https://www.epfl.ch/research/ethic-statement/human-research-ethics-committee

[33] https://www.epfl.ch/campus/services/data-protection/data-protections-officer

[34] https://www.epfl.ch/research/ethic-statement/compliance

[35] https://www.fedlex.admin.ch/eli/cc/2013/617/en#chap_3

[36] https://go.epfl.ch/rdm-fastguide

[37] https://dmlawtool.web.app

developed.

- **Data integrity**: How complete and uncorrupted is data, and how it can stay so for at least the entirety of the project. The technical infrastructure of EPFL around data storage and transfers ensures data integrity through different elements.
- **Confidentiality**: Data is not disclosed to unauthorized individuals. It is suggested to use the "need-to-do" principle: only individuals and processes that are duly authorized can access the data, and they can only perform actions corresponding to their work and for which they have been authorized.

Note 1

The **EPFL storage systems** rely on the EPFL identity and access management system. Access to data is therefore managed depending on the status, position, and role of each individual. Each person not affiliated with EPFL who will need access to the EPFL IT infrastructure will have to fill out and sign a document and give an ID for identification purposes.

Note 2

The EPFL Data Protection Officer[38] should be contacted **immediately** in the event of a **data breach** (see: How to notify a data breach[39]).

Answer example

If there no personal or sensitive data involved, then an answer of this type might be adapted to your case and used "All the datasets are accessible by the PI on the online cloud storage [*provider*] where the other group's members can only access the datasets for which they are responsible: in particular [*who*] has access to [*dataset1*], [etc.]. The PI and each member work from their own laptop, where they access the cloud storage with the double factor authentication (2FA). The servers of the cloud storage are located only in Switzerland and are therefore subjected to Swiss laws. As the group's storage on the central EPFL server acts as a back-up for the cloud storage and, given that the datasets do not present any ethical issue, all the members of the research group have access to all datasets backed-up during the project, even by other members. The access on the EPFL server is regulated via the institutional Gaspar accreditation. Moreover, the PI makes monthly back-ups of the entirety of the datasets on an external SSD drive, that is kept under key in her desk and under her sole responsibility."

---

[38] https://www.epfl.ch/campus/services/data-protection/data-protections-officer
[39] https://www.epfl.ch/campus/services/data-protection/how-to-notify-a-data-breach

EPFL

More information

- EPFL Library FastGuide #03 – Cost of Research Data Management[40]
- EPFL Security guidelines[41]
- EPFL privacy in research[42]
- LEX 6.1.4 Directive on the Use of Electronic Infrastructure[43]

2.3 How will you Handle Copyright and Intellectual Property Rights Issues?

In this DMP section, describe the data ownership and authorship: tell who owns the data and with which rights.

If using previously existing data or code whose reuse is applicable, always make sure that you respect all licenses applicable. Check if the reuse of third-party data is restricted (ex. Terms of service of online platforms; Dataset's license from previous research; Digital assets from proprietary software; etc.). If reuse is authorized, then the derived dataset must be shared according to the third party's original requirement or license.

> Note
>
> If your project involves harvesting webpages or other online information, via API or other means, make sure to respect the terms and conditions of the service or data provider (website, dataset, etc.).

When sharing or publishing your own data or code is applicable (i.e. legally possible, as there are no constraints such as privacy issues, commercial contracts, dual-use technology, etc.), choose the right license to communicate all potential restrictions for their future use. In this DMP section, state what license you plan to use for the data you created. Systemically attaching a license to datasets produced in the project is good practice: it facilitates reuse, reproduction, and collaboration and also clarifies authorship and contributions.

As a general recommendation, research at EPFL can be licensed as follows:

- For data: it is recommended to use Creative Commons Attribution (CC-BY-4.0) with **no other constraints**. Use of "no derivative" or "non-commercial" is not recommended, as they are not aligned with scientific good practice or the mission of EPFL. Alternatively, CC0 can be used if citations are not a concern (e.g. for supplementary data).
- For code, it is recommended to use one of the many software licenses with properties similar to those of CC-BY, for example, Apache, BSD, or MIT. The Apache license (or similar), which includes language related to the use of patents, may be needed for larger projects (i.e. if lawyers may be involved).
- The use of "militant" licenses such as CC-SA and GPL is acceptable as part of a commitment to openness or if it is the norm in the community.

---

[40] https://go.epfl.ch/rdm-fastguide
[41] https://www.epfl.ch/campus/services/wp-content/uploads/2018/11/ITsecurity-guide.pdf
[42] https://www.epfl.ch/campus/services/data-protection/in-practice/privacy-in-research
[43] https://www.epfl.ch/about/overview/wp-content/uploads/2020/01/LEX-6.1.4_EN.pdf

We suggest you check this license selector[44] to find the license you want (look for "open data" or "approved by Open Source Initiative"), or TLDRLegal[45] to have simple summaries of the different licenses. You can also use the DMLawTool[46] to navigate the potential legal issues for your Research Data. In case of doubt, contact researchdata@epfl.ch.

If there are any non-disclosure agreements in place, or if Research Data are commercially sensitive (plan to submit a patent application, agreement with industrial partners etc.), use this DMP section to explain how you will handle the aspect and consider contacting the EPFL Technology Transfer Office[47] for support.

> **Answer example**
>
> If there are no patents, nor 3rd party contracts, or other Intellectual Property constraints, then you can use a statement similar to the following: "There are no Intellectual Property issues with the datasets produced during the project, so we choose to disclose our Research Data as soon as possible under CC BY 4.0 license. In case any unforeseen patents or commercial use should arise, then we will contact the EPFL Technology Transfer Office for support, and will update the DMP accordingly."

More information

- EPFL Library FastGuide #12 – Data & Code Licensing[48]
- EPFL TTO Guide - Learn how to share your software[49]
- Creative Commons - Choose the right Creative Commons license for your work[50]
- Comparison between each type of licence - snyk.io blogpost[51]

---

[44] https://ufal.github.io/public-license-selector

[45] https://tldrlegal.com

[46] https://dmlawtool.ccdigitallaw.ch

[47] https://www.epfl.ch/research/services/units/technology-transfer-office

[48] https://go.epfl.ch/rdm-fastguide

[49] https://www.epfl.ch/research/services/protect-intellectual-property/software-licenses

[50] https://creativecommons.org/choose

[51] https://snyk.io/blog/mit-apache-bsd-fairest-of-them-all

3 Data Storage and Preservation

3.1 How will your Data be Stored and Backed up During the Research?

In this DMP section, describe the storage solutions you use: **list all storage solutions** for your active data and code. If needed, consult your Faculty IT (more info later) or the central IT services[52] to understand what is the best storage solution for your research project.

Possible storage solutions can be (but are not limited to): laptops, tablets, desktop computers, cloud services (commercial or non-commercial), external portable storage (USB drive, hard drive), servers (NAS, facility, etc.), etc.

For each storage solution, provide the following details:

- What is its storage **capacity** (or the accessible one)
- Which **datasets** are stored on it (be consistent with section 1.1)
- What **backup** procedure is in place (chosen backup storage, manual/automatic, synchronization method, frequency, responsibility, etc.)

> Note 1
>
> For data integrity and security, the exclusive use of portable storage (e.g., personal laptops, external HDDs, USBs) for Research Data is strongly deprecated. Be consistent with DMP sec. 2.2.

The recommended storage solutions offered at EPFL are as follows:

(A) If the data is subject to official secrecy:[53]

- EPFL NAS, accessible through the whole of EPFL's IT network and via VPN[54]
- SWITCHDrive, servers in Switzerland, authentication via edu-ID login, and the EPFL allows the storage of confidential documents if those are encrypted[55]
- Object Storage service (S3), recommended for cold data (i.e. not expected to be frequently accessed)[56]
- c4science, code version control platform[57]
- GitLab EPFL, code version control platform[58]

(B) If the data is not subject to official secrecy:

- Microsoft 365, MS Office suite and related cloud, does not store data in Switzerland[59]
- Gdrive EPFL,[60] servers not at EPFL, not FADP compliant in case of sensitive data or data under official secrecy
- Follow the EPFL Cloud Guide[61] for other cloud storage providers

---

[52] https://www.epfl.ch/campus/services/en/it-services/it-support

[53] https://www.epfl.ch/campus/services/data-protection/laws-and-regulations/official-secrecy

[54] https://support.epfl.ch/epfl?id=epfl_service_status&service=49a363acdb34c700ef64731b8c96191f

[55] https://www.switch.ch/drive

[56] https://support.epfl.ch/epfl?id=epfl_service_status&service=e9bd3320dbb76f40914f9ec4db961914

[57] https://c4science.ch

[58] https://gitlab.epfl.ch

[59] https://windows.epfl.ch/core/index.aspx?article=300

[60] http://gdrive.epfl.ch

[61] https://inside.epfl.ch/secure-it/en/cloud-guide

> **Note 2**
>
> c4science or GitLab EPFL can be used in conjunction with the EPFL NAS storage, depending on the need for storage capacity.

Although costs and budgeting are information that is not due in the SNSF DMP, keep in mind that there might be costs involved in the storage and backup of Research Data: use the Cost Calculator for Data Management[62] to estimate some of these costs. To budget for them in your project proposal, check out the EPFL Research Office's toolkit to prepare your funding application[63].

> **Note 3**
>
> Data processing on a public Cloud needs a case-by-case analysis. Contact the VPA Legal Affairs or the EPFL DPO.

> **Answer example**
>
> If using EPFL NAS, then you can use a statement similar to the following: "During the project, datasets A, B, C are stored in the [storage solutions] and are automatically synchronized to EPFL NAS accessible by all project collaborators. This is a centrally managed storage system that provides redundancy, replication, controls integrity and is based on the identity and access management system of EPFL. Daily backup is automated, and access is regulated via GASPAR secure authentication."

More information

- EPFL Library FastGuide #10 – Storage, preservation and publication[64]
- EPFL Central IT service's Storage & data page[65]
- EPFL overview of storage options for administrative documents[66]
- FastGuide #03 "Cost of Research Data Management[67]
- Faculty ITs:
  - ENAC-IT[68]
  - SV-IT[69]
  - STI-IT[70]
  - SB-IT[71]
  - IC-IT (CDH, CDM)[72]

3.2 What is Your Data Preservation Plan?

---

[62] https://rdmepfl.github.io/costcalc

[63] https://drive.google.com/drive/folders/1avoAQg2dybKiVPaSXSsw3qX490b-Lrtl

[64] https://go.epfl.ch/rdm-fastguide

[65] https://support.epfl.ch/epfl?id=epfl_index&servicenumber=SVC0061

[66] https://www.epfl.ch/campus/services/en/it-services/storage-of-documents

[67] https://go.epfl.ch/rdm-fastguide

[68] http://www.epfl.ch/schools/enac/fr/a-propos/enac-it

[69] http://www.epfl.ch/schools/sv/it

[70] http://www.epfl.ch/schools/sti/it

[71] https://sb-it.epfl.ch

[72] https://www.epfl.ch/schools/ic/it/en/it-service-ic-it

Not all data that you create/collect/use during your research project need or can be preserved. Keep in mind that the preservation is not a simple dump of datasets, but entails **data curation** procedures to make them reusable in future (see later).

> Note 1
>
> The long-term preservation of datasets follows the EPFL LEX 3.3.2 recommendation on Research Integrity,[73] to "create a transparent and **verifiable archiving system** of the results obtained during the research", as well as the EPFL Compliance Guide's mandate to "be able to prove the veracity of their results and share them with other researchers for **at least ten years** following publication".[74]

In this DMP section, <u>for each dataset</u> specify:

- **If you will preserve it** and to which extent. Be consistent with the information in section 1.1. Consider preserving all software/code (depending on the software license) used to acquire/generate/manipulate your data (be consistent with section 1.2).
- **What criteria** you will use to select the data for preservation.
  Describe how you will select data for preservation. Some criteria might be
  - Long-term value (expensive or difficult to obtain core data assets)
  - Value for re-use (as the basis of scientific claims)
  - Obligation to destroy/publish (compliance with laws or contacts)
  - The cost involved to maintain data (long-term secure storage)
- What **data curation procedures** and methods.
  Describe how you will prepare selected data or code for preservation, mentioning the methods you plan to use: e.g., convert data to open formats; clean tables; anonymize; add documentation; etc.
- On **what platform(s)** you will preserve them.
  For data and code preservation purposes, it is recommended to use EPFL's own ACOUA.[75] Some data dissemination platforms, e.g. Zenodo,[76] can also be considered as preservation solutions, although with more stringent legal and size limitations. Community- or datatype-specific open data platforms can be good solutions, inasmuch as their operator commits to data preservation. If your preservation and dissemination solutions are the same, then be consistent with section 4.1 of the DMP and name them in both sections.
- **When** you will start the preservation.
  This depends on the workflow and research processes. See the below examples:
  - At the end of the project
  - Periodically, e.g., annually
  - At specific milestones, e.g., paper publication
- The **duration** of the preservation.
  Indicate the retention period: this may vary based on the discipline and value of data, but the SNSF recommends storing Research Data for 10 years.[77] This becomes a minimum obligation for some sensitive data such as health data (ex. from clinical trials). Other obligations might exist, depending on possible contracts and applicable laws.

---

[73] https://www.epfl.ch/about/overview/wp-content/uploads/2019/09/3.3.2_principe_integrite_recherche_an.pdf

[74] https://www.epfl.ch/about/overview/regulations-and-guidelines/compliance-guide

[75] https://go.epfl.ch/acoua

[76] https://zenodo.org

[77] https://www.snf.ch/en/dMILj9t4LNk8NwyR/topic/open-research-data

> **Note 2**
>
> Collaborative platforms such as GitHub or GitLab, as well as cloud platforms such as Gdrive or SWITCHDrive, are not preservation solutions.

Although costs and budgeting are information that is not due in the SNSF DMP, keep in mind that there might be costs involved in the preservation of Research Data: use the Cost Calculator for Data Management to estimate some of these costs.[78] In many cases, you can budget for that in your project proposal: check out the EPFL Research Office's toolkit to prepare your funding application.[79]

> **Answer example**
>
> "We plan on using the EPFL ACOUA service to preserve the datasets produced for this project, including the necessary documentation (see section 1.3). Prior to archiving, we will select the datasets for long-term preservation by [state your criteria] and they will be cleaned using [state your methods]. Where still needed (see section 1.1), we will convert file formats into their open-format counterparts for future reusability. If legally possible, we will also include the required software (see section 1.2). The preservation using ACOUA is planned to start towards the end of the funding period in [state approx. date], and the archived datasets will be preserved there for at least 10 years. We will update this DMP with necessary details before the end of the funding period."

More information

- How to Appraise and Select Research Data for Curation[80]
- EPFL Library FastGuide #10 – Storage, preservation and publication[81]
- EPFL Library FastGuide #13 – Data/Code publication platforms[ibid.]
- Identify data with long-term value[82]
- EPFL Library's comparative table of data repositories and related platforms[83]

---

[78] https://rdmepfl.github.io/costcalc

[79] https://drive.google.com/drive/folders/1avoAQg2dybKiVPaSXSsw3qX490b-Lrtl

[80] http://www.dcc.ac.uk/resources/how-guides/appraise-select-data

[81] https://go.epfl.ch/rdm-fastguide

[82] https://dataoneorg.github.io/Education/bestpractices/identify-data-with

[83] https://go.epfl.ch/datarepo

4 Data Sharing and Reuse

4.1 How and Where will the Data be Shared?

In this DMP section, the terms "share" has to be intended as "disseminate" or "publish" (publicly share), i.e., not sharing as for active data. Describe in this DMP section:

- **Which dataset(s)** or subsets of datasets you will publish. It is not always legally possible to publish all Research Data: in case limitations apply to data/code publication, please describe these **exceptions** in the DMP section 4.2.
- **Where** these datasets will be published. Mention the name and provide a link to the **data repository** or other dissemination platform (ex. data journals[84]) for publishing your Research Data. If your preservation and dissemination solutions are the same, then name them in both sections.
- **How** your choice ensures discoverability and ease of access for potential reuse by other researchers. E.g., by providing a DOI, using standardized keywords, using a data repository renowned in your scientific community, etc.

> Note 1
>
> While costs for enabling access to Research Data are eligible, SNSF does not reimburse the cost of dissemination on commercial platforms (e.g., Figshare).

It is recommended to publish Research Data in a data repository and/or other dissemination platform aligned with FAIR principles[85]. If applicable, describe if and how you will:

- Reserve the DOI
- Provide a human-readable description of the dataset(s)
- Select a license that is the least restrictive, in line with section 2.3
- Cross-link your dataset publication to the related scientific output (article, other datasets, etc.)
- Input the identifiers of all the authors of the dataset (e.g., ORCID[86])

For a curated list of data dissemination platforms, check EPFL Library's comparative table.[87] Relevant platforms include Zenodo,[88] GEO DataSets,[89] Envidat,[90] Dryad,[91] OSF,[92] or Figshare,[93] but many others exist. If you plan on using Zenodo, to make your dataset more easily findable, it is strongly recommended to submit it to the EPFL Community.

Commonly used collaborative platforms, such as GitHub and GitLab, or cloud solutions such as Google Drive and SWITCHDrive, do **not implement the FAIR principles** and associated practices endorsed by the SNSF. A good strategy is to additionally

---

[84] https://go.epfl.ch/datajournals

[85] http://www.go-fair.org/fair-principles

[86] https://www.orcid.org

[87] https://go.epfl.ch/datarepo

[88] https://zenodo.org/communities/epfl

[89] https://www.ncbi.nlm.nih.gov/gds

[90] https://www.envidat.ch

[91] https://datadryad.org

[92] https://osf.io

[93] https://figshare.com

share/publish "snapshots" of data/code on a FAIR compliant platform.

> ### Note 2
>
> If using GitHub, you can refer to GitHub's documentation to make your code citable via Zenodo.[94]

> ### Note 3
>
> Some restrictions exist in terms of privacy for the dissemination of datasets: they need to be anonymized. However, personal data that is not sensitive and cannot be anonymized because of the purpose of its processing, can be shared, but on the condition that the third party anonymizes it as soon as possible and that it does not share it further without an EPFL's agreement that regulates its transfer. If the sharing involves collaborators outside EPFL, it is necessary to draw up an agreement that indicates: what categories of data are affected, for which purposes it may be processed, and what guarantees are provided by the laboratory. In order to set the project's terms, contact the legal services of the Vice Presidency for Academic Affairs.[95]

> ### Answer example
>
> Simple case: "After data curation, we will make all datasets underlying the published results openly available on Zenodo at the time of the related publications. Zenodo offers a DOI and implements standardized metadata and keywords that allow for the discoverability of the datasets, as it's harvested by the major search engines. We will further improve the discoverability of our datasets by indicating the affiliation to the EPFL Community on Zenodo.[96] We don't foresee Intellectual Property issues, nor any privacy concerns, so our datasets will be disseminated in Zenodo under a CC-BY-4.0 license."

More information

- EPFL Library FastGuide #01 – FAIR Data Principles[97]
- Search for data repositories also on re3data[98]
- ORCID use for SNSF: Your curriculum vitae – all about the CV format[99]
- University of Illinois File Formats Considerations for Data Sharing[100]
- SNSF's report on ORD, containing criteria to choose a data repository[101] (also highlighted in section 4.3 of this guidelines)

---

[94] https://docs.github.com/en/repositories/archiving-a-github-repository/referencing-and-citing-content
[95] https://search.epfl.ch/?filter=unit&q=VPA-EM-AJ
[96] https://zenodo.org/communities/epfl
[97] https://go.epfl.ch/rdm-fastguide
[98] http://www.re3data.org/browse/by-subject
[99] https://www.snf.ch/en/gKcnwW6aEft4bMPF/page/your-curriculum-vitae-all-about-the-cv-format
[100] https://emails.illinois.edu/newsletter/154676263.html
[101] https://doi.org/10.5281/zenodo.3618123

4.2 Are There Any Necessary Limitations to protect Sensitive Data?

> **Note**
>
> SNSF accepts only restrictions based on legal, ethical, copyright, confidentiality, or other contractual clauses.

Examples of restrictions might be (but are not limited to):
- If a **commercial use** could be considered (ex. patents), or for issues of double-use technology, then contact the TTO.[102]
- If your dataset derives from scraping websites or 3rd party datasets with unclear licenses, then contact researchdata@epfl.ch as first support.
- If you cannot assure the proper **anonymization** of personal or sensitive data, then you cannot publish it. If in doubt, consult the EPFL Research Ethics service[103] and be consistent with section 2.3 of the DMP. Contact researchdata@epfl.ch for support on data anonymization.

Finally, provide a timeframe for **when** your datasets will be published. The SNSF expects that researchers will make them available **as soon as possible**, but no later than the time of publication of the respective scientific output.

> **Answer example**
>
> If no restriction is expected, then you can use a statement similar to the following: "Data which underpins any publication will be made available at the time of publication. All unpublished data will be archived via the ACOUA[104] service at the end of the project." Always follow the principle "as open as possible, as restricted as necessary". In case of no limitations, simply mention that. In case any legal limitation should apply, list all limitations that prevent the sharing of your Research Data as a whole, or part of it.

4.3 I will choose digital repositories that conform to the FAIR data principle [CHECK BOX]

Check if the repository or other dissemination platform you mention in section 4.1 is aligned with the FAIR Data Principles. Tick YES if your answer to each of the questions below is "yes":
- Are datasets given globally unique and persistent identifiers (e.g. DOI)?
- Does the repository allow the upload of intrinsic (e.g. author's name, content of dataset, associated publication, etc.) and submitter-defined (e.g. definition of variable names, keywords, etc.) metadata?
- Is it clear under which license the data will be available, or can the user upload/choose a license?
- Are the citation information and metadata always publicly accessible, even in the case of datasets with restricted access?
- Does the repository provide a submission form requesting intrinsic metadata in a specific format (to ensure machine readability/interoperability)?

---

[102] https://tto.epfl.ch
[103] https://www.epfl.ch/research/ethic-statement
[104] https://go.epfl.ch/acoua

> ▪ Does the repository have a long-term preservation plan for the archived data?
>
> If the repositories or other dissemination platforms you mention in section 4.1 follow these criteria, then tick "YES".

More information

- [Data repositories and related platforms comparative table](#)[105]
- [SNSF's list of most used platforms](#)[106]

4.4 I will choose digital repositories maintained by a non-profit organization [RADIO BUTTON YES/NO]

You are free to select either of the options, but SNSF encourages the use of non-commercial repositories. If there is a payment required for publishing in the non-commercial repository of your choice, this is an eligible cost, and you can still select "YES".

You can either use the EPFL Library's [comparative table](#)[107] or consult [re3data.org](#)[108] to learn if a repository is considered non-commercial:

- Under the tab "Institutions", check if a commercial entity is involved in 'general' or 'technical' responsibility (categories "Type of institution" and "Type(s) of responsibility")
- If not, SNSF considers the repository to be non-commercial (even if 'funding' or 'sponsoring' is provided by a commercial entity).
- If yes, the SNSF considers the solution to be a commercial repository

> Note
>
> You can publish datasets on more than one platform, as long as your choice of licenses is compatible.

Examples of **non-commercial repositories**: [Zenodo](#),[109] [B2SHARE](#),[110] [Dryad](#),[111] [Harvard Dataverse](#),[112] [Materials Cloud](#)[113]

Examples of **commercial repositories**: [Figshare](#),[114] [IEEE DataPort](#),[115] [NeuroMorpho](#),[116] [Mendeley Data](#)[117]

---

[105] https://go.epfl.ch/datarepo
[106] https://www.snf.ch/en/WtezJ6qxuTRnSYgF/topic/open-research-data-which-data-repositories-can-be-used
[107] https://go.epfl.ch/datarepo
[108] http://www.re3data.org
[109] https://zenodo.org
[110] https://b2share.eudat.eu
[111] https://datadryad.org
[112] https://dataverse.harvard.edu
[113] https://archive.materialscloud.org
[114] https://figshare.com
[115] https://ieee-dataport.org
[116] https://neuromorpho.org
[117] https://data.mendeley.com

If you select "NO", you must provide a reason for choosing a commercial repository. One 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. If there is a payment required for publishing in a commercial repository, this is not an eligible cost.

> Answer example
>
> If you use Zenodo for data dissemination and/or ACOUA for preservation, then select "YES".