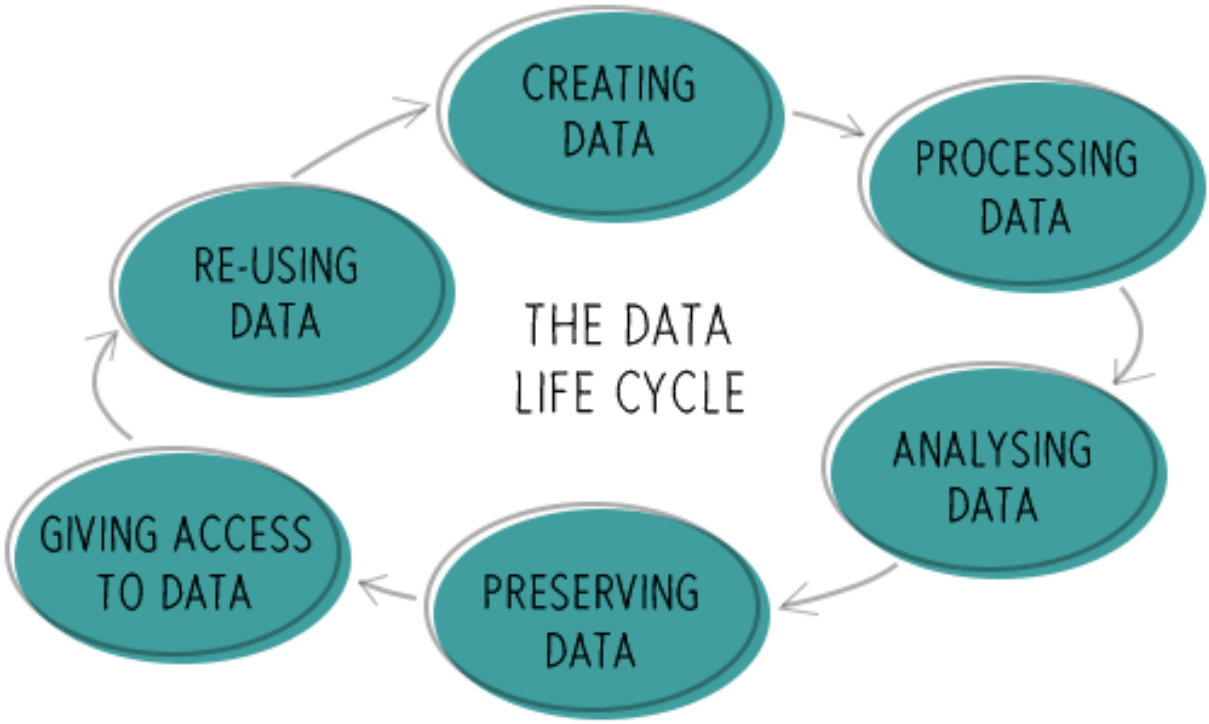
# RDM Walkthrough Guide

# Contents

# Get started with RDM

Research data are “used as primary sources for scientific research, and [...] are commonly accepted in the scientific community as necessary to validate research findings” (OECD Principles and guidelines for access to research data from public funding).




*Research data lifecycle inspired by the UK Data Archive website: http://www.data-archive.ac.uk/create-manage/life-cycle*

## What is research data management?

According to the discipline, data can be for example experimental, observational, computer simulations, textual records, and also physical artefacts. Research data management (RDM) refers to all the decisions and activities related to how researchers handle data throughout its whole life-cycle, from the planning stage of the project to the long term preservation strategies.

Good data management is an essential part of the research process, and contributes to guarantee integrity, transparency and reproducibility of research, as well as to meet institutional expectations and funders’ requirements.

# RDM practical tools

EPFL SNSF DMP Template

ERC DMP Template

EPFL DMP Cost Calculator

EPFL Recommended File Formats

EPFL Data Publication Decision Tree

EPFL Data Management Plan Checklist

RDM Fast Guides

## Love Data Week

The Love Data Week (LDW) is a social media event coordinated by research data specialists, mostly working in academic and research libraries or data archives or centres (discover more on lovedataweek.org). The purpose of the LDW event is to raise awareness and build a community to engage on topics related to **research data management**, like sharing, preservation, reuse, and library-based research data services.

At the EPFL Library, we embrace the spirit of the LDW, as we believe that research data are at the foundation of the scholarly record, and crucial for advancing our knowledge of the world around us. During the LDW, we will share **practical tips**, **resources**, and **stories** to help researchers at any stage in their career use good data practices.

Love Data Week 2018 at EPFL

Love Data Week 2019 at EPFL

## Research Data Management survey results

At the end of 2017, the EPFL Library launched a survey aimed at evaluating the feasibility of data curation/data stewardship services among EPFL researchers. The results give you an overview of the most important needs required by EPFL researchers in terms of research data management.

Take a look at the results

# RDM support and training

## Research data management workshops

Take part in the next training sessions organized by EPFL Library Research Data team. Should you have any question, please feel free to contact researchdata@epfl.ch.

- *Research data management: introduction*

  In this course, you will get an introduction to the main concepts of Research Data Management to apply them to your specific situation.

  Learning outcomes:

  - Know the stakes around Research Data Management
  - Discover how a Data Management Plan (DMP) can help you be more efficient in your research
  - Get an overview of good practices to work with your data at the different stages of your project

- *Research data management: from plan to action*

  In this personalized workshop, you will check the consistency of your data management plan (DMP) and how to implement it in your lab.

  Before you attend the workshop, you will need to fill a form about your current practices. If you are not familiar with research data management already, we suggest that you register to our introduction course.

Your data partners

- EPFL Library
- Research office

# Plan and fund


## Data Management Plan (DMP)

## Data Management costs

## Funder's data requirements

Before starting a new project, planning the management of research data is an important step to formalize this process, anticipate the different measures to put in place at any stage of the data life cycle, estimate the related costs, and identify the possible weaknesses in the plan.

Funding bodies (such as the Swiss National Science Foundation or the European Commission with the Horizon 2020 program) are more and more requesting researchers to provide a Data Management Plan along with their grant applications and to apply good practices in research data management during the whole funded project.

Good research data management is not only part of responsible research, but ensure various benefits for researchers and institutions.

It could be a complex issue, but done early enough it helps in saving time and avoiding problems during the course of the research project. Moreover, tools like data management plan templates and checklists help in streamlining this task and provide a formal framework to organize the work.

# Data management plan (DMP)

## What is a DMP?

A Data Management Plan or DMP is a written document that describes how a laboratory manages its data: how research data will be produced during a research project, how it will be collected, and what to do with it during and after the research process. A DMP describes every step of the data lifecycle, starting from data creation, to transformation, analysis, storage, sharing and reuse.

## Why should you write a DMP?

Writing a DMP will help to identify specific issues and areas of improvement regarding research data management practices. Proper data management is a key prerequisite for effective data sharing throughout the scientific community. This increases the visibility of scholarly work and the citation rates.

Preparing and writing a DMP has numerous benefits:

1. Saving time, resources and efforts
2. Optimizing the research
3. Encouraging collaboration and increasing research impact and visibility
4. Anticipating costs that can be reimbursed by the funding agencies
5. Complying with the funders', publishers' and institutions' requirements

An increasingly number of funders requires now to develop and implement DMPs (Horizon 2020, Swiss National Science Foundation, etc.), as described more in detail in the funder's data requirement section.

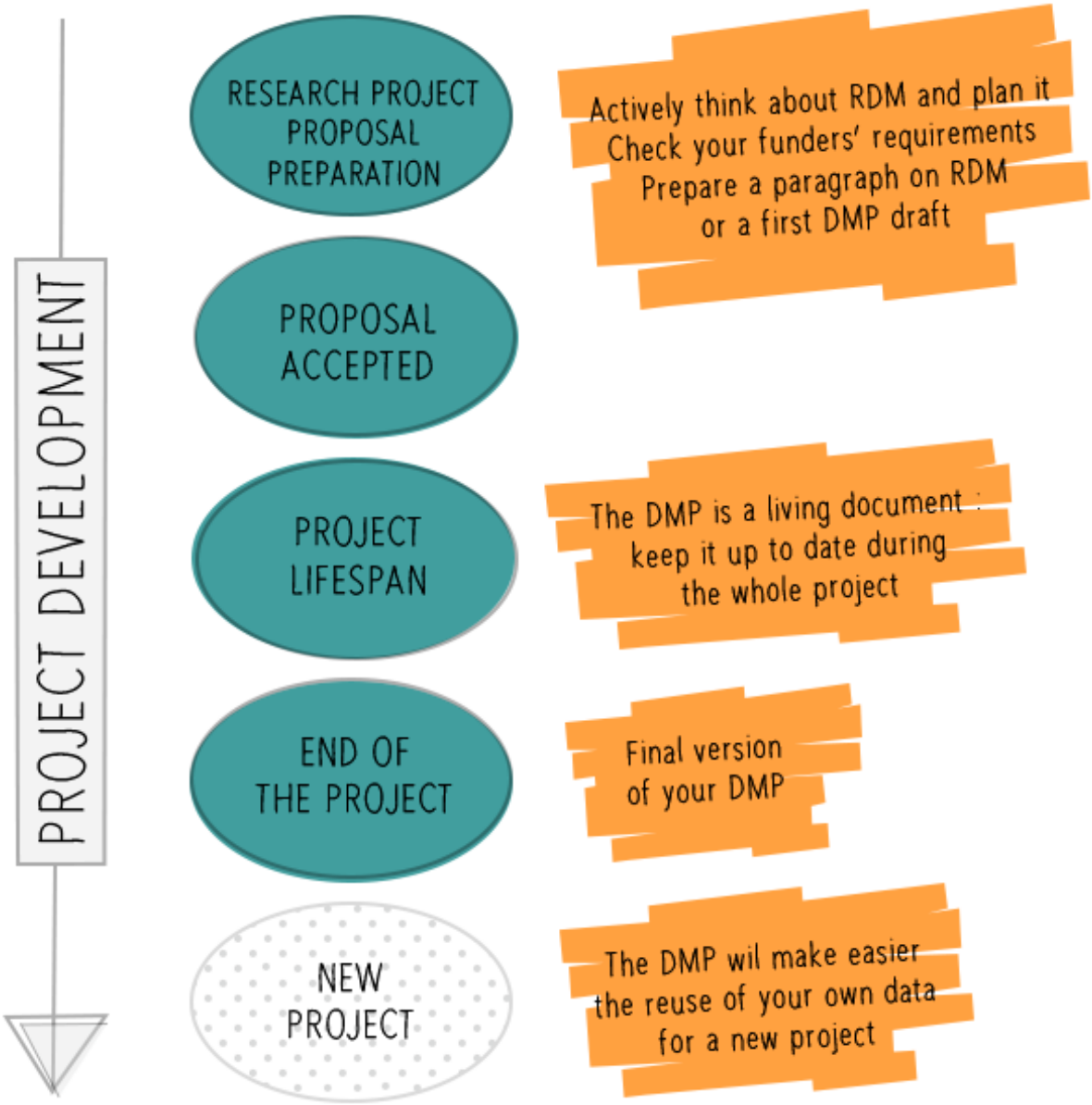
## When should you write a DMP?

A DMP can be written at the beginning of the research project or, even better, during the preparation of the research project proposal. Essentially, preparing a DMP ahead of time will make a significant difference, saving money and effort afterwards. Keep in mind that it is never too late to do it, and that it is also a living document, which as to be updated throughout the research project.

Funding agencies have different requirements about DMPs (more information on page "Funders' data requirements"). For instance, the SNSF will publish the latest release of DMPs. And H2020 requires at least three DMP versions: the first one is requested within the first 6 months of the project, the second one at the middle and a final one at the very end of the research project.

## How writing a DMP?

1. Start by asking yourself questions about your research data using this EPFL-ETH DMP checklist
2. Review your funder's guidelines for DMP and research data management using Sherpa Juliet
3. After reviewing the checklist with your team and agreeing on a strategy, fill out your DMP
4. You can use this EPFL template, except for SNSF (DMP content on mySNF)
5. Update your DMP regularly to reflect any changes in your data management.
6. Apply what you wrote in your DMP since an audit may arise in the future.
7. In case you wish to have additional guidance in creating your DMP, feel free to contact the research data library team at researchdata@epfl.ch and register for the RDM SFP training.




## Useful external resources

Concrete examples of Data Management Plans:

- http://www.dcc.ac.uk/resources/data-management-plans/guidance-examples (prepared by the Digital Curation Centre)
- https://dmptool.org/public_plans (collection of public DMPs created using the DMPTool)

A short video that illustrates how a DMP works concretely, prepared by the Research Data Netherlands:

- https://www.youtube.com/watch?v=gYDb-GP1CA4

# Data management costs

## Costing data management activities

Data management not only has an impact on the project planning, but also on the project budget, because all the related activities can incur costs.

Therefore, researchers are strongly encouraged to take these activities into account when preparing a funding application as the related costs are considered as eligible by many funders. Although it could be a hard task, the following actions can help to assess the cost implication:

- Analyse research data management costs incurred in previous projects and use it as a basis for future planning
- Plan the data management early enough in order to reduce the costs
- Identify data management activities to be performed in your project and estimate the resources needed to perform them (staff time, HR costs, training, storage space…)

According to the High Level Expert Group on the European Open Science Cloud, **on average about 5% of research expenditure** should be spent to ensure a proper data management and stewardship.

To identify which research data management activities should be weighted in the cost estimation.

*Data management cost overview*

This overview, based on the UK Data Management Costing Tool (UK Data Service) and on the Data Management Cost Guide (Utrecht University), is intended as guidance to formulate a more detailed project budget.

0. Preparing

- Make a Data Management Plan (time cost)

1. Data Collection

- Acquiring external databases and software (e.g.: licence cost)
- Formatting and organising data (e.g.: HR cost)
- Transcription (e.g.: time cost)
- Consent for data sharing (e.g.: HR cost)
- Data transfer (e.g.: software cost)

2. Data documentation

- Data description and metadata (e.g.: time, HR cost)
- Documentation (e.g.: HR cost)

3. Data storage and back-up

- Data back-up (e.g.: cost for institutional solutions)
- Data storage (e.g.: cost for institutional solutions)

4. Data access and security

- Data access
- Data security

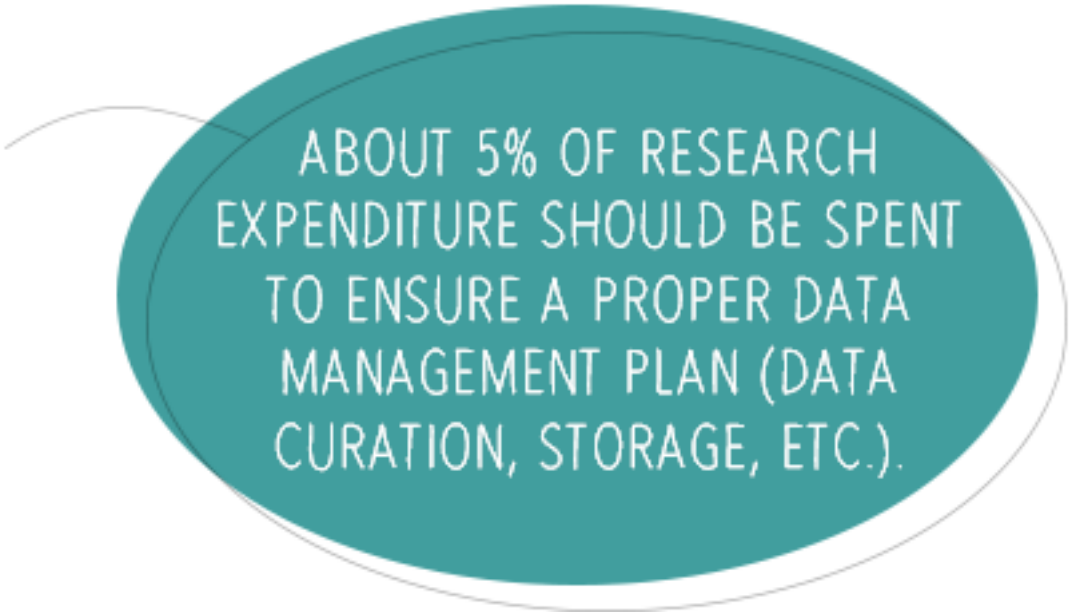
5. Data preservation and archiving

- File format (e.g.: HR cost)
- Access to LTP and data repository

6. Data sharing and reuse

- Anonymization (e.g.: HR cost)
- Copyright (e.g.: juridical advice)
- Data sharing (e.g.: time cost + possible data repository cost)
- Data cleaning (e.g.: HR cost + possible data cleaning service cost)

It is important to avoid double funding as some of this expenditure (especially those related to data storage and back-up) could be included in the indirect costs.




## Eligible costs – Funders' regulations

*Swiss National Science Foundation (SNSF)*

As of October 2017, researchers applying for a SNSF grant will have to:

- Include a data management plan (DMP) in their funding application
- Make data generated by funded projects publicly accessible in non-commercial, digital databases (unless they are bound by legal, ethical, copyright, confidentiality or other clauses)

The SNSF will contribute to the costs related to data archiving (included data preparation), if the research data is deposited in FAIR, digital, recognized and not for profit repositories. It is allowed to upload data to commercial repositories, but in this case only the data preparation costs will be covered by the SNSF.

Costs eligibility conditions are detailed in the article 2.13 of the General implementation regulations for the Funding Regulations.

*Material costs*

2.13 – Material costs: costs of data storage and providing access to data (open data)

1. The costs of enabling access to research data that was collected, observed or generated under an SNSF grant are eligible if the following requirements are met:

    a. the research data is deposited in recognised scientific, digital data archives (data repositories) that meet the FAIR principles and do not serve any commercial purpose.
    b. the costs are specifically related to the preparation of research data in view of its archiving, and to the archiving itself in data repositories pursuant to letter a.

2. All costs charged to the grant must be linked to archiving of data that is thematically related to research that was funded by the SNSF.
3. The maximum charge per grant is generally CHF 10,000.
4. The costs must be considered at the time of submission of the application. Other requirements set by the SNSF concerning the accessibility of research data must be met during the submission of the application in mySNF. This holds in particular for the submission of a data management plan (DMP).

## Horizon 2020 (including ERC grants)

In Horizon 2020, the European Research and Innovation funding programme, open access to scientific publication and research data is an obligation for all beneficiaries, as set out in the article 29.3 of the Model Grant Agreement. The related costs are therefore considered as eligible by the European Commission to support researchers in complying with these requirements, ensuring the necessary resources and the budgetary planning.

The eligibility for reimbursement during the duration of the project is defined in the Model Grant Agreement mainly in the Article 6 (Eligible and ineligible costs), and more specifically in the clause 6.2.D.3.

*Specific conditions for costs to be eligible*

6.2.D.3 – Specific conditions for costs to be eligible – Other direct costs

Costs of other goods and services (including related duties, taxes and charges such as non-deductible value added tax (VAT) paid by the beneficiary) are eligible, if they are:

(a) purchased specifically for the action and in accordance with Article 10.1.1 or

(b) contributed in kind against payment and in accordance with Article 11.1.

**Such goods and services include**, for instance, consumables and supplies, **dissemination (including open access)**, protection of results, certificates on the financial statements (if they are required by the Agreement), certificates on the methodology, translations and publications.

[...]

This budget category covers the costs for goods and services that were purchased for the action (or contributed in-kind against payment), including [...]:

- dissemination costs (including regarding open access to peer-reviewed scientific publications, e.g. article processing or equivalent charges, costs related to open access to research data and related costs, such as data maintenance or storage and conference fees for presenting project-related research)

[...]

Specific cases (costs for other goods and services) [...]:

- **Plan for the exploitation and dissemination of results** — Costs for drawing up the plan for the exploitation and dissemination of the results are normally **NOT eligible since they will have been incurred before the start of the action, to prepare the proposal**. Costs that occur when revising or implementing this plan may be eligible.
- Open access — Costs related to open access to peer-reviewed scientific publications and research data are eligible, if the eligibility conditions are fulfilled. With explicit agreement by the Commission/Agency, it can also include fees levied for a membership scheme (if this is a requirement for publishing in open access or if membership is a pre-condition for significantly lower article processing charges).

## Specific costs at EPFL

The EPFL Library team can support you in the cost evaluation process. For some activities, as for example the active data management, the storage, as well as the data deposition in a repository or publication in a journal, the EPFL provides tools and infrastructure or financial support.

The Research Office provides also a tailored budget calculator for different funding programs. More information is available in the Research Funding webpage (Gaspar login needed).

Feel free to contact us to get more information on these options and the related costs: researchdata@epfl.ch.

# Funders' data requirements

From 2017 onwards, all researchers must implement best practices in scientific research and prepare a Data Management plan (DMP) to apply to selective grants from innovative research programs such as Horizon 2020 (including ERC grants) and the Swiss National Science Foundation (SNSF).

## Swiss National Science Foundation (SNSF)

Starting at submission date October 2017, the SNSF now requires the following from all researchers :

1. A full dynamic Data Management Plan (DMP), when completing their grant application on mySNF.

The DMP is not included in the evaluation of the project, however the funds will not be released until the DMP is fully completed. The DMP must be updated throughout the project and the final version of the DMP will be published in the P3 data bank (SNSF).

The online SNSF DMP form includes four sections:

1. Data collection and documentation
2. Ethics, legal and security issues
3. Data storage and preservation
4. Data sharing and reuse

A DMP pre-fill template specifically designed for the SNSF has been prepared by EPFL and ETHZ to guide and to help researchers save effort and time while completing the SNSF form.

### 2. Depositing and archiving data produced during the project in a FAIR repository

The SNSF requires to deposit data in an open-access and non-commercial repository, which complies with the FAIR principles. Of course, this requirement has to be respected as long as no ethical restrictions apply to it.

*What does FAIR mean?*

FAIR principles mean that the repository meets standards ensuring that data sets are Findable, Accessible, Interoperable, and Re-usable:

- **Findable:** Data and metadata are easy to find by both humans and computers. Machine readable metadata is essential for automatic discovery of relevant datasets and services, and as such are essential to the FAIRification process.
- **Accessible:** Limitations on the use of data, and protocols for querying or copying data are made explicit for both humans and machines.
- **Interoperable:** The computer can interpret the data, so that they can be automatically combined with other data. There is a historical trend in computer science toward increased interoperation (for instance, between different hardware designs, operating systems, programming languages, and communication protocols).
- **Reusable:** Data and metadata are sufficiently well described for both humans and computers in order for them to be replicated or combined in future research.

To understand in more depth what this means concretely, feel free to explore this practical guide. The SNSF provides guidelines for assessing the suitability of repositories as well as examples of suitable repositories. Useful resources regarding relevant repositories based on your area of expertise can for instance be found on https://www.re3data.org.

### 3. Publishing research data

In addition, SNSF expects data underlying a publication to be shared as soon as possible, and at the latest at the time of publication of the respective scientific results. Complementary data can also be shared and more information about this can be found on our website here.

## Horizon 2020 (including ERC grants)

Research funded by H2020 now requires a DMP as a deliverable. At a minimum, one version must be provided at the beginning, a second version during the project and the third at the final stage of the research project. Moreover, data underlying publications must be deposited in a repository, and made accessible and exploitable by third parties.

For more information on this topic, feel free to review the following guidelines:

1. H2020 Guidelines

2. ERC Guidelines

3. FAIR Data Principles

## Overview of funders' data policies

| | Policy Coverage | | Policy Stipulations | | | | | Support Provided | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| Research Funders | Published Outputs | Data | Time Limits | Data Plan | Sharing/ Access | Long-Term Curation | Monitoring | Guidance | Repository | Data Centre | Costs |
| SNSF | ● | ● | ● | ● | ● | ● | ○ | ◐ | ○ | ○ | ◐ |
| European Commission: H2020 | ● | ● | ◐ | ● | ◐ | ◐ | ◐ | ● | ● | ◐ | ● |
| European Commission: ERC | ● | ● | ◐ | ● | ◐ | ◐ | ◐ | ● | ● | ◐ | ● |
| NIH | ● | ● | ◐ | ● | ◐ | ◐ | ◐ | ● | ● | ◐ | ● |
| NSF | ● | ● | ◐ | ● | ◐ | ◐ | ◐ | ● | ● | ◐ | ● |
| Bill & Melinda Gates Foundation | ● | ● | ◐ | ● | ◐ | ◐ | ◐ | ● | ● | ◐ | ● |
| AHRC | ● | ● | ● | ● | ● | ◐ | ○ | ● | ○ | ◐ | ◐ |
| BBSRC | ● | ● | ● | ● | ● | ● | ● | ● | ● | ◐ | ● |
| EPSRC | ● | ● | ● | ◐ | ● | ● | ● | ◐ | ○ | ○ | ● |
| ESRC | ● | ● | ● | ● | ● | ● | ● | ● | ● | ● | ◐ |
| MRC | ● | ● | ● | ● | ● | ● | ○ | ◐ | ● | ○ | ◐ |
| NERC | ● | ● | ● | ● | ● | ● | ● | ● | ● | ● | ◐ |
| STFC | ● | ● | ● | ● | ● | ● | ● | ◐ | ● | ◐ | ◐ |
| Cancer | ● | ● | ● | ● | ● | ● | ● | ◐ | ● | ○ | ● |
| Wellcome | ● | ● | ● | ● | ● | ● | ● | ● | ● | ● | ● |

Inspired by the Digital Curation Centre (DCC) website

# Work with data


## Active data management

## Analysis and visualisation of data

## Metadata and documentation

## Storage and back-up

During the lifespan of a project, researchers have to deal with data management on a daily basis. Different needs may arise according to the discipline and to the phase of the project, however it is possible to identify some common elements.

Active data management is essential in the daily research work, and discipline specific tools could support it in a digital and highly collaborative environment. Analysis and visualization of data are the core activities: from discovering meaningful information from datasets to present them in a visually interpretable form, it is key for researchers to have appropriate software and computing environment available to perform these tasks.

Along with all that, data documentation and storage on a regular basis are very important to ensure security and reusability of data.

# Active data management

Discipline specific tools may assist the management of your active data, from creation, to processing and analysis phases of the data life cycle. Particularly:

- Source code version control systems
- Electronic laboratory notebooks (ELN) and laboratory information management systems (LIMS)
- Computational workflow engines
- Computer science notebooks and environments (see "Data analyse and visualization" page)

## Source code version control systems

c4science: EPFL project with many useful features, notably:

Version control system (support Git, Subversion and Mercurial)

- Unlimited number of public and private project/repositories
- Hosted by SWITCH (in Lausanne with a backup in Zurich) and accessible to the whole Swiss academic community
- Repositories can easily be made accessible to research partners outside Switzerland
- Additional features:
  - Documentation (wiki);
  - Project management (tasks);
  - Continuous integration (Jenkins)

gitlab.epfl.ch: git-based collaborative platform:

- Hosted at EPFL and available for the EPFL community
- User-friendly interface

## ELN / LIMS

Electronic Laboratory Notebooks (ELN) are software replacing paper laboratory notebooks and more. They allow collaborative work and support native digital content (such as microscopy, gels images, DNA sequences, etc.). Depending on the tool, they may have the same legal value as signed paper notebooks.

Laboratory Information Management Systems (LIMS) are information management software supporting modern laboratory operations, such as laboratory equipment and samples' management, including their location and associated data.

ELN/LIMS are tools combining the two sets of functions.

At EPFL, the following systems are available:

- **Life sciences:** SLims which is an ELN/LIMS (actually also used by some labs in STI)
- **Chemistry:** eln.epfl.ch, an in house developed system accessible to all EPFL members

## Computational workflow

A scientific workflow is a formal definition of the research process. In addition of automating tasks, such formalization increases research reproducibility. Workflows are made of a series of computational or data manipulation steps and are machine-readable. Scientific workflow management software allows to easily manage complex or repetitive operations.

*SnakeMake*

SnakeMake is a workflow management system that aims to reduce the complexity of creating workflows by providing a fast and comfortable execution environment, together with a clean and modern specification language in python style. Snakemake workflows are essentially Python scripts extended by declarative code to define rules (for more information you can refer to the Snakemake's documentation page).

Snakemake supports:

- Remote files handling (http-s, sftp, Dropbox, Google Drive)
- Data provenance and rule versions
- Parallelization
- Suspend / resume
- Logging
- Graphical workflow generation

*AiiDA*

"AiiDA is a flexible and scalable informatics' infrastructure to manage, preserve, and disseminate the simulations, data, and workflows of modern-day computational science.

Able to store the full provenance of each object, and based on a tailored database built for efficient data mining of heterogeneous results, AiiDA gives the user the ability to interact seamlessly with any number of remote HPC resources and codes, thanks to its flexible plugin interface and workflow engine for the automation of complex sequences of simulations" (AiiDA website). AiiDA is developed at EPFL.

*Taverna*

"Taverna is an open source multi-platform tool for designing and executing workflows. Taverna is discipline independent and used in many domains, such as bioinformatics, cheminformatics, medicine, astronomy, social science, music, and digital preservation" (Wikipedia). It is composed of several tools, among which:

- Taverna Workbench: desktop application enabling to graphically create, edit and run workflows
- Taverna Command Line: enables to run commands form prompt, e.g. for automated execution
- Taverna Server: remote workflow execution service, enabling to set up a dedicated server

Taverna is also modular: many plugins are available. Taverna supports myExperiment, and thus allows re-using or sharing workflows in a few clicks. Existing workflows are a great source of inspiration to develop your own workflows, either through embedding them directly as sub-workflows or by simply using them as starting points for your own designs.

*Pegasus*

Pegasus runs on various environments including personal computers, campus clusters, grids, and clouds. It is quite flexible, but more difficult to learn than Taverna. No graphical design tool is available.

Pegasus helps constructing workflows in abstract terms without worrying about the details of the underlying execution environment or the particulars of the low-level specifications required by the middleware (Condor, Globus, or Amazon EC2).

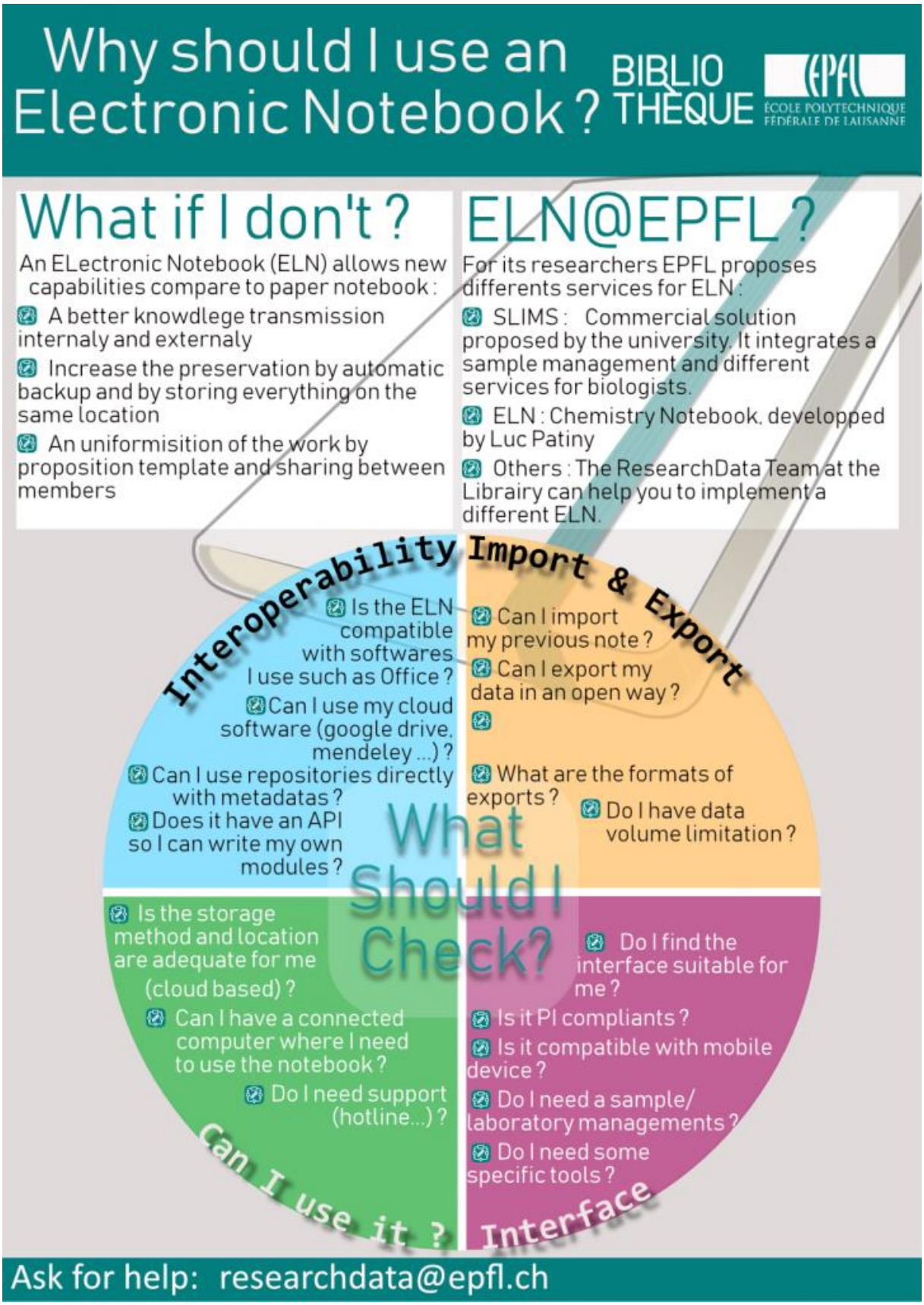
Pegasus is used in many of scientific domains including astronomy, bioinformatics, earthquake science, gravitational wave physics, ocean science, limnology, and others.

Pegasus keeps track of what has been done (provenance) including the locations of data used and produced, and which software was used with which parameters.

Pegasus has a number of features that contribute to its usability and effectiveness:

- portability and reuse
- performance and scalability
- provenance and data management
- reliability and error recover

# Electronic lab notebook



## To know more:

*Interoperability*

- Is the ELN compatible with software I use?

Do you use software regularly such as word, excel or others and you want my ELN to be able to open them and integrate them easily?

- Can I use my cloud software (i.e. Google Drive, Mendeley)?

If you use cloud-based software such as Google Drive, Dropbox, Mendeley, some ELN can import/export them, it might increase the storage space and also sharing possibilities.

- Can I use repositories with metadata?

Open science/open data means sharing your data/codes, a direct integration with repositories may simplify the publication by directly exporting, formatting the notebook to the repository.

- Does it have an API so I can develop my own module?

The perfect system doesn't exist, but an API of the ELN allows you to develop your own modules in order to improve your ELN in the way you want.

*Can I use it? Does the ELN match my needs or obligation?*

- Is the storage method and location adequate for me (cloud-based)?

Most of the ELN use a cloud storage in their facilities. This means your data might be stored in a different country than you are and you don't "host" the data locally. This might be a problem if you use sensitive data, for instance.

- Can I have a connected computer where I need to use the notebook?

You will need to use your notebook where you are doing experiments can you have a connected computer where you need it (for example the lab)? Some ELN can work offline and upload the modification when connected.

- Do I need support (hotline ...)?

Are you ease with going to a system where you manage everything (setup, training...) or do you want some support? Some ELN might be free, but support will be an extra cost. Some ELN doesn't offer anything. You should check therefore carefully who is behind the ELN and what security/support you can get with each specific ELN.

*Interface*

- Do I find the in interface suitable for me?

Each system is different, you have to check that the ELN interface is fine for you, this includes the functionalities for redaction but also the way things are organized graphically and whether they are easy to use or not.

- Is the ELN GLP/ISO Compliant?

1. Some ELN are compliant with some regulation you might need if you are writing patterns or for regulated research (GLP…)
2. You might need to sign and authenticate the data/input stored in your notebook
3. If you work with confidential information, you may also want to check where your data are stored (see 2.1)

- Is the ELN compatible with mobile devices?

You might want to access or use mobile devices, such as tablets or smartphones. This might be in particular useful for taking pictures, hand drawing or to work on this outside.

- Do I need sample or laboratory management?

You might need to have an inventory of your samples or other tools to manage your lab/equipment, some ELN integrates such functionalities.

*Import, Export*

- Can I import my previous notes/documentation?

You might have the need to import your previous note/documentation (from a previous ELN or other systems).

- Can I export my data in an open way?

Does the ELN offer the possibility to export your data, and maybe combine them by using an open format (i.e. XML) that can be used after everywhere (Like a data repository for instance)?

- What are all the formats of exports?

Most of ELN offer PDF as the export format, but with PDF you lose the access to raw data. That's why you might want to check if you can export in a more open format such as XML HTML or others.

- Do I have a volume limitation?

In the case of cloud-based storage, the storage is managed by the ELN company, you might want to check what are the options/limitations and also the cost for the storage. In the case of a local hosted ELN, you want to check the bandwidth and also the external access of such ELN in case of sharing or outside access (a mobile device for example).

*Links*

- Harvard Listing

- Kanza, S., Willoughby, C., Gibbins, N., Whitby, R., Frey, J. G., Erjavec, J., … Kovač, K. (2017). Electronic lab notebooks: can they replace paper? Journal of Cheminformatics, 9(1), 1–15. https://doi.org/10.1186/s13321-017-0221-3
- Kanza, S. (2018). What influence would a cloud based semantic laboratory notebook have on the digitisation and management of scientific research? Retrieved from https://eprints.soton.ac.uk/421045/
- Kwok, R. (2018). Lab Notebooks go Digital. Nature (Vol. 560). https://doi.org/10.1038/d41586-018-05895-3

# Data analysis and visualisation

## What is data visualization?

Data visualization contributes to data analysis and results' communication. "Bringing research data to life by David McCandless (TEDx)" is an inspiring video presenting data visualization.

## Data analysis and visualization environments

"Jupyter Notebook is a free interactive computing environment that enables users to author notebook documents that include: live code, interactive widgets, plots, narrative text, equations, images and video. These documents provide a complete and self-contained record of a computation that can be converted to various formats and shared with others" (Jupyter documentation). Over 50 languages are supported, including Python, R, Matlab, Octave, Scala, Lua and BASH. Many data analysis and visualization libraries are available in this environment, some are listed below.

"R-Studio is a free and open-source integrated development environment (IDE) for R, a programming language for statistical computing and graphics [...]. RStudio is available in two editions: RStudio Desktop, where the program is run locally as a regular desktop application; and RStudio Server, which allows accessing RStudio using a web browser while it is running on a remote Linux server" (Wikipedia). Automated generation of documents is supported using R Markdown and Knittr (see below).

## Software and libraries

Visualisation and analyse tools may be categorised by their flexibility and simplicity of use. Here is a short selection:

| Low-level (library) | Mid-level (library) | High-level (application) |
|---|---|---|
| Matplotlib | Seaborn | Gephi |
| NetworkX | Pandas | Tableau |
| Scipy | Scikit-learn | Matlab |
| Numpy | ggplot2 | Octave |
| D3.js | | |

*Python libraries*

These open Python packages may:

- be combined,
- be used within a Jupyter Notebook, for convivial interactive work and easy sharing.

Numpy: fundamental library for scientific computing with Python, it contains among other things a N-dimensional array object.

Scipy: library providing scientific computing tools, like easy-to-use statistic functions.

Pandas: library providing high-performance, easy-to-use data structures and data analysis tools; the main feature is the use of dataframes.

Matplotlib: plotting library with a great flexibility; it has comparable features to Matlab plotting.

Seaborn: visualization library based on matplotlib; it provides a high-level interface for drawing attractive statistical graphics.

NetworkX: library for the creation, manipulation, and study of the structure, dynamics, and functions of complex networks.

Scikit-learn: library providing easy-to-use tools for machine learning: classification, regression, dimensionality reduction, etc.

*Applications*

Gephi: free multi-platform data analysis software. This tool is very useful to draw networks and relations between elements. For more information and examples see the documentation of Gephi and this video tour.

Matlab: product of the MathWorks company. This platform is built around the Matlab scripting language. It provides tools for data analysis and visualization.

Octave is a free software using a syntax largely compatible with Matlab.

Tableau: commercial software providing an interface to build visualizations and explore data. It exists in different declinations: desktop, server, cloud.

*Some other libraries*

R Markdown: R Markdown is an extension of Markdown, that supports embedded R code, and may be used in conjunction with Knitr to make it easy to create reproducible web-based reports. R Markdown and Knittr packages are available in RStudio.

D3.js: JavaScript library for creating interactive documents based on data. D3 bring data to life using HTML, SVG, and CSS.

ggplot2: visualization package for R language. Very popular when plotting with R.

## Useful external resources

Curated collection of visualization tools chosen by datavisualization.ch:

- http://selection.datavisualization.ch/

Overview of the main data visualization tools available on the web:

- http://www.creativebloq.com/design-tools/data-visualization-712402

# Data documentation and metadata

Data documentation and metadata are essential for research reproducibility. Indeed, undescribed data is generally not reusable by others or even by the researcher who produced it, as it lacks contextualization. For example, columns of numbers are usually cryptic if they are not qualified with at least title and measurement units.

More generally, metadata is defined as data describing data. They provide information about the context, the structure, the provenance and the content of a dataset (or a file) with the aim to increase its usefulness.

The minimum documentation of a dataset is to describe it within a README file and, if appropriated, a naming convention. Both of these should be written in a future proofed and software agnostic format such as simple text or markdown.

A README is however not machine operable. To create machine operable metadata information, many metadata standards are available (see below). Some are generic, others are discipline specific. Depending on the context, various types of tools will assist and even automate metadata creation. This is notably the case for source code version control systems, computational workflow engines, electronic laboratory notebooks and laboratory information management systems (for more information, check our active data management page).

*Common metadata standards*

At EPFL, there are currently no official recommendations for metadata standards. Researchers can refer to the following ones as they are well established and widely used:

Dublin Core: a simple set of 15 terms that can be used to describe datasets and more generally electronic resources. The Qualified Dublin Core is an extension of the terms, adding notably the ability to refine the semantics via standard controlled vocabularies.

Comma Separated Values on the Web (CSV on the web): is a recommendation for documenting CSV files. CSV is the most commonly found data type in repositories. A great strength of this tabular data format resides in its compatibility with most spreadsheets (Excel, Google Docs, LibreOffice) and scientific (R, Python, Matlab, etc.) applications. However, CSV files are often difficult to reuse due to the lack of description of their structure, content or relation to other tabular data files. These elements are covered by CSV on the Web.

HDF5 (Hierarchical Data Format version 5): a set of data formats supported by many platforms (including Java, Matlab, Octave, Mathematica, Python, R and Julia). HDF5 has interesting metadata capabilities. First, all datasets are given a type. Secondly, datasets can be organized in groups and subgroups. Thirdly, each group, subgroup and dataset can be described with an arbitrary quantity of metadata in JSON.

DataCite Metadata Schema: one of the most popular descriptive metadata. It is much more precise than Dublin Core (see above). It is listed last in this list as it requires much more work to implement and is most frequently used by professional data repositories.

*Other common standards:*

BioSharing: life sciences metadata

ISA-Tab: life sciences metadata

Gene Ontology: genetics metadata

ISO 191151: geospatial metadata, and the XML implementation: ISO 19139

SDMX: statistical metadata, and SDMX editor

*Metadata standards directories*

If the above listed metadata standards do not fit your needs, you may find more specific ones in the following resources:

Metadata directory: a collaborative, open directory of metadata standards applicable to scientific data.

Available metadata standards (Wikipedia): a community curated list of metadata standards on Wikipedia.

Metadata schema supported by Dataverse: an open source web application to share, preserve, cite, explore, and analyse research data. This project is a collaboration between the Institute for Quantitative Social Science (IQSS), the Harvard University Library and Harvard University Information Technology organization. Dataverse is committed to using standard-compliant metadata.

*Customization of metadata formats*

If you cannot find an adequate metadata standard, you can create your own metadata format.

JSON: a language independent open standard. It is structured as key and values and is easily human-readable. JSON can be converted in RDF, using JSON-LD.

RDF: the Resource Description Framework is a World Wide Web Consortium specification. Flexible and general, it enables interoperability between datasets across disciplines. Some notations of RDF, i.e. way to encode RDF, are easily readable for humans, such as Turtle. Other notations are more adapted for machine-to-machine data exchange, notably XML-RDF.

XML (Extensible Markup Language): a great number of XML documents formats exist, covering all domains. Examples include LibreOffice documents, RSS and Atom, DublinCore, SVG, and XHTML.

# Storage and backup

Researchers are responsible for the safety, security and integrity of the research data they generate or collect. This also includes any correspondence and records that relate to that research.

For digital data, researchers should confirm with their IT department that the back-up process, e.g. frequency of back-ups, number of back-up copies, and their housing in multiple locations (including off-site copies), is appropriate for their research data.

Please note that storage is different from long-term preservation.

The EPFL centralized file storage service follows the best practices and standards. The service is managed centrally by the hosting department of the VPSI and ensures security, coherence, pertinence, integrity and high-availability. This is achieved through different and redundant data protection levels, as well as through partnerships with providers both of equipment and software, and also through data replication onto two distinct storage locations found on the EPFL campus.

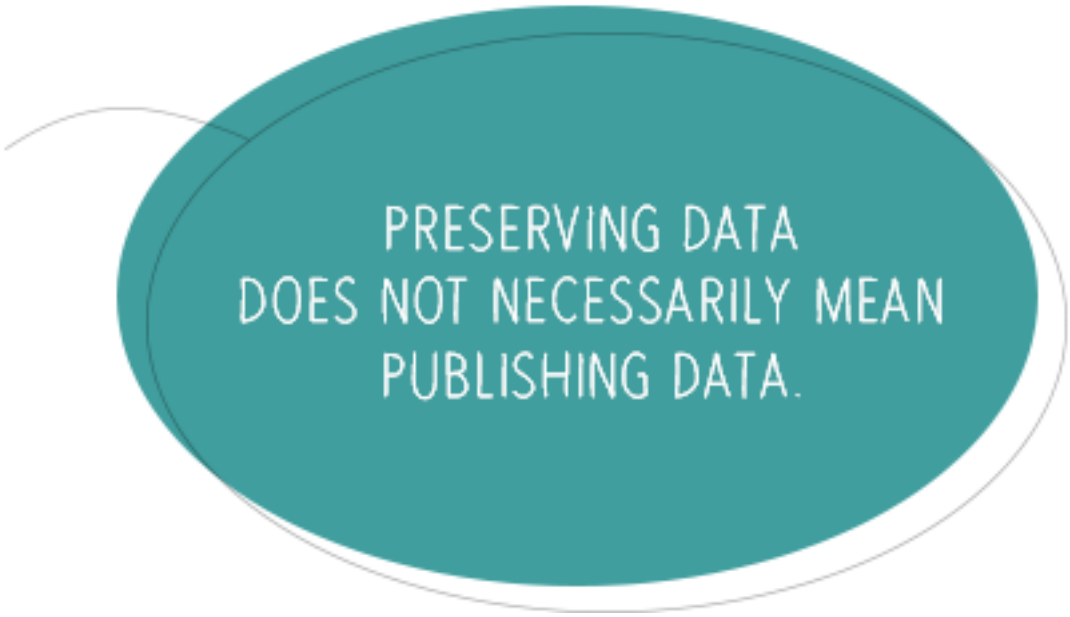
For specific needs regarding data retention, optional tape backups can be proposed; this incurs additional costs.

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

More information on EPFL storage solutions (options, costs, etc.) can be found on the following VPSI webpage (Gaspar login needed). New requests for storage can be addressed through it.epfl.ch – Requester may be staff of the unit or local IT support.

| | Storage & Backup | Long Term Preservation |
|---|---|---|
| Active data | ✓ | ✗ |
| Data recovery | ✓ | ✓ |
| Integrity (monitoring, repair, authenticity) | ? | ✓ |
| Appraisal (what and for how long) | ✗ | ✓ |
| Permanent Identifiers | ✗ | ✓ |
| Description (metadata) | ✗ | ✓ |
| Renderability (format migration, virtualization…) | ✗ | ✓ |

# Publish and preserve




## Data selection for long term preservation

## Data repositories and data journals

## Publishers' requirements

## Personal data protection and anonymization

Research data publishing can be defined as "the release of research data, associated metadata, accompanying documentation, and software code [...] for re-use and analysis in such a manner that they can be discovered on the Web and referred to in a unique and persistent way". It occurs via data repositories and/or (data) journals, which ensure that the data is "well documented, curated, archived for the long term, interoperable, citable, quality assured and discoverable" (Austin, C et al., 2015).

As funding agencies and publishers are increasingly mandating to make data underlying findings and publications accessible, publishing data is now a crucial step for researchers.

# Data selection for long-term preservation

## What is data long-term preservation?

Data preservation does not only mean storing data in a safe manner, but it also implies that data will remain **accessible and reusable in the long term** (for several years or even forever) **ensuring:**

- **Intellectual interpretability** (by providing sufficient metadata and documentation)
- **Technical readability** (by using for example appropriate formats)
- **Integrity** (by replication of the data and checksum usage)

Long-term preservation (LTP) should be planned since the beginning of the project. The Data Management Plan is a useful tool to describe the preservation strategies that the researchers would like put in place, and to make the monitoring during the project lifespan easier.

In order to preserve data correctly, appraisal and selection are needed to determine which data will be ultimately devoted to long-term conservation or eliminated. To well structure a data preservation plan, two questions must be answered:

- What is it worth to be kept?
- For how long?

## Why preserve data?

Providing access to data with adequate metadata is a condition to ensure reproducibility of research results. Moreover, some data are unique and cannot be replaced, so the importance to provide access to them in the long term is even more important.

Long term preservation (and accessibility) of data can also be a funding agencies' requirements for some research funding programs.

## Who decides?

Deciding which data should be preserved and for how long is a decision that belongs to the research team. However, the different stakeholders of a project (funders, research institutions, publishers, etc.) might have specific requirements that should be considered when defining the LTP strategy.

## What does preservation cost?

Preservation costs have to be considered and included in a research project budget as part of the general data management costs:

- Data curation costs include resources needed to manage data during the project, to prepare them before depositing them in a repository and to be the respondent of the data thereafter.
- Repository costs are the charges that can be applied by the repositories for data deposition. These costs depend on different elements, including the dataset size (if the dataset is big these costs could be very high). More information about that can be found in the Data repositories and data journals section.

## How to select data to preserve?

*Prerequisites for data conservation*

In order to deposit data in a repository, a set of conditions must be fulfilled:

- Being the owners of the data (or having the consent of the involved stakeholders)
- Complying with the data protection law (anonymisation, restricted access, etc.)
- Ensuring data integrity and accessibility (no corrupted data, availability of software and hardware, etc.)
- Providing appropriate metadata to ensure data intelligibility
- Clarifying the conditions of reuse with adequate license

It does not make sense to preserve data if any of the above conditions are not fulfilled. In general, good data management and curation throughout the whole project suffice to prevent such limitations.

*Stakeholders requirements*

In Switzerland, there is no legal obligation to preserve or publish research data so far. However, several constraints can be imposed by the stakeholders of a research project.

- Funders: some of them (such as the European Commission with Horizon 2020 and the Swiss National Science Foundation, as described in detail in the "Funder's data requirements" page) require a DMP. One of its section is especially focused on the data preservation strategies, helping the researchers in planning them since the beginning of the project.
- Publishers: an increasing number of them require that the data underlying the articles are made accessible, as indicated in the "Publisher's requirements" page.
- Partners: when a project is carried out in partnership with research teams outside EPFL, it is important to define who will own the data and the final destination of them once the project is completed. Private partners (if any) can also set conditions on the use of the data collected.
- Data repositories: each repository has its own policies regarding the type of accepted data (disciplines, formats, size, etc.). The choice of a relevant one should be made early enough in order to meet its requirements adequately.

The choice of the stakeholders often has an impact on the data management. This question has to be considered when looking for funders, repositories, partners, etc. For example, the Swiss National Science Foundation excludes for profit data repositories.

*Data appraisal*

If the prerequisites are fulfilled and the requirements of the stakeholders are not sufficient to determine the selection of the data to preserve, there are several more qualitative criteria:

- **Ethical issues:** these can either restrict or encourage the publication of data. On one hand, if a misuse of the data is possible, it would be a bad decision to publish them. On the other hand, scientific ethics encourages data transparency, sharing and publication
- **Value of the data** (uniqueness, cost to harvest, links with other dataset, science trends, potential reuse, etc.)
- **Quality of data documentation and metadata**
- **Quality and reliability of the sources and harvest methods**

*Preservation costs*

Preservation costs, especially data curation costs and repositories fees, must be also considered to determine the size of the data to preserve and the duration.

## Does preserving data mean publishing data?

Most often, if datasets have been preserved it is also to be published, but it is not always possible. For example, when there are some restrictions, as copyright or privacy issues for example. Sometimes, researchers decide to work further on a dataset and prefer to restrict the access to these data during this stage.

It is also possible to ensure the preservation of a dataset while retaining control over its use by others. Depending on the chosen data repository, access restrictions, embargoes and sampling can be set up.

## How long to preserve data?

Several repositories define the retention time of data sets, but most of them do not fix a limit. The relevant duration depends on the value of the data and, in particular, on the potential for reuse, which is likely to decrease over time. In a general way, 5 to 20 years of preservation seems reasonable.

## Which data to preserve? Raw data or processed data? What about sampling?

Raw data is data in original state at the time of collection. Processed data is the data transformed and used to analyse the research questions. Which ones to preserve? It depends on the purpose of the data preservation.

If checking the validity of the research results is required, relevant raw data must be at least preserved. However, these is not sufficient to reproduce the results. The code and algorithms used to process these data need to be provided, as well as sufficient metadata to explain how they had been processed.

Sampling is also to be considered to reduce the costs. An option is to only preserve the data directly useful to validate the results or even less if there is no need to be able to prove the results.

# Data repositories and data journals

## Data repositories

Data repositories are infrastructures allowing to preserve and/or publish research outputs.

A distinction can be made between disciplinary and multi-disciplinary repositories:

- Disciplinary repositories are generally a good choice, since they are adapted to subject-specific data and could be more well-known in the disciplinary community. However, data stewardship requires a lot of resources (human, machine time) and some small disciplinary repositories do not always meet basic data management standards.
- Multi-disciplinary repositories accept any type of data, and some of them offer excellent data management services, even for free.

## Finding the right repository

When choosing among different repositories, it is important to consider the following elements to find the most relevant one and maximise the impact of data:

- disciplinary data sharing practices;
- disciplinary/community standard repositories;
- combination of ease of deposit, accessibility, discoverability, curation, preservation infrastructure, organizational persistence and support for used formats and standards.


## Re3data

"Re3data is a global registry of research data repositories from all academic disciplines. It provides an overview of existing research data repositories in order to help researchers to identify a suitable repository for their data" (Wikipedia). Re3data indexes over 1500 repositories and offers search filters.

*Main search filters on Re3data*

Some of them are more of significance than others, notably:

- Subjects: useful to narrow a search to repositories relevant to your discipline. However, take into account that some multi-disciplinary repositories may be better solutions than subject specific ones, especially sizeable well-curated tools such as Zenodo, Dryad or Figshare.
- Certificates: attest that a repository is visible, well curated, and that its data are well described and of good quality. Especially Data Seal of Approval (DSA) and World Data System (WDS) certificates are relevant.
- Data access: an open access to the data will encourage the reuse and citation of your work.
- Data license: the use of acknowledged data licenses implies a clear definition of what users may or may not do with a dataset. Notably Creative Commons licenses (CC-BY, CC0) allow to give or retain various rights on datasets. They are relatively easy to understand, and at the same time, legally well-defined and machine-readable. For computer code, the following licenses are to be considered: Apache, Berkeley Software Distribution (2 and 3 close BSD Licenses), GNU Public Licenses (GPL, LGPL, AGPL), Public Domain.
- Metadata standards: used to describe datasets efficiently, which is essential for their reuse and discoverability. The support of Dublin Core (DC) offers a minimal simple description.
- PID Systems: persistent identifiers enable to cite efficiently a data set, and are built to avoid broken links. The Digital Object Identifier (DOI) and handle system (HDL) are the most common PID Systems.
- AID Systems: author identifiers facilitate the discovery of an author's work through an unambiguous identification. Among them ORCID is valuable.

## Most commonly used multi-disciplinary repositories

Zenodo is a repository operated by CERN covering all scientific disciplines. It offers free data submission for any research as long as it is openly published. In addition, DOI are systematically attributed to records, making them cleanly citable. Another notable feature is its integration with GitHub, enabling to capture, preserve and cite Git repositories.

Dryad is a curated general-purpose scientific data repository. All records in Dryad are associated to published articles, and a data publishing fee is requested for deposition (more details available here). DOI are attributed systematically.

Figshareoffers free data deposition and access for all disciplines, and attributes systematically DOI. Unlike Zenodo and Dryad, Figshare is a commercial repository, belonging to Macmillan Publishers.

## Data journals

Data journals are emerging publications whose main purpose is to make research data discoverable, interpretable and reusable, providing impact and recognition for authors.

Datasets are now being recognized as a primary research outputs, so it can be an interesting option to present them in a data paper. This allows the author to focus on the description of the data, its context, the acquisition methods, as well as its actual and potential use (rather than presenting new hypothesis or interpretations).

Moreover, authors can get credit as data article are peer-reviewed publications and citable. As data journals are always Open Access, an Article Processing Charge (APC) has to be paid by the author for the publication costs. It is possible to request for the Library financial support to cover part of the APC.

## Examples of data journals:

- Scientifc Data (Springer Nature Group)
- Data in Brief (Elsevier)
- Journal of Physical and Chemical Research Data (AIP)
- Journal of Open Research Software (Ubiquity Press)
- GeoScience Data Journal (Wiley)

# Publishers' requirements

An increasing number of scientific publishers is developing policies about sharing of research data underlying the published articles. These requirements are often included in the authors' information or in the ethical guidelines.

Publishers can simply encourage the authors to make their data available (as for example Wiley or ACS) or mandate the data sharing as a condition *sine qua non* for article publication (as for example PLoS).

In this second case, the data needed to validate a publication has to be made available with no (or minimal) restrictions.

In the journal's data policy, a repository or a data journal is often suggested to make the data available, and in some cases, an integrated data submission system is proposed within the article submission process.

The Library can provide advice on how to comply with journals' requirements on research data, help in the interpretation of policies, and with the choice of a suitable data repository and/or data journal.

## Examples of publishers' requirements

- American Chemical Society (encouragement)
  "When requested, the authors should make every reasonable effort to provide data, methods, and samples of unusual materials unavailable elsewhere, such as clones, microorganism strains, antibodies, etc., to other researchers, with appropriate material transfer agreements to restrict the field of use of the materials so as to protect the legitimate interests of the authors. Authors are encouraged to submit their data to a public database, where available".
- Nature Journals (obligation)
  "A condition of publication in a Nature journal is that authors are required to make materials, data, code, and associated protocols promptly available to readers without undue qualifications. Any restrictions on the availability of materials or information must be disclosed to the editors at the time of submission. Any restrictions must also be disclosed in the submitted manuscript".
- PLOS (obligation)
  "PLOS journals require authors to make all data underlying the findings described in their manuscript fully available without restriction, with rare exception. When submitting a manuscript online, authors must provide a Data Availability Statement describing compliance with PLOS's policy. If the article is accepted for publication, the data availability statement will be published as part of the final article. Refusal to share data and related metadata and methods in accordance with this policy will be grounds for rejection".

# Personal data protection and anonymization

## Work with personal data

Personal data is all information related to an identified or identifiable person. Handling such data requires special precautions in order to be compliant with the law (see this Research Office page).

If a project involves personal data it is important to:

- Document the processes and the uses of the data (data management plan, metadata, code used to process data, etc.)
- Get the approval for the project:
  - from HREC (EPFL Human Research Ethics Committee) or
  - from CER-VD (Commission cantonale d'éthique de la recherche sur l'être humain) for most medical studies.
- Get valid consent of individuals. This consent must be given expressly in the case of processing of sensitive personal data.
- Inform participants on their personal data
- Anonymize data, as soon as the purpose of the processing permits
- Secure data against any violation
- Avoid transferring data abroad (be sure to comply with the law if you want to do it)

## Data anonymization

### *Why data anonymization matters*

Data anonymization offers several advantages. In particular, it enables to:

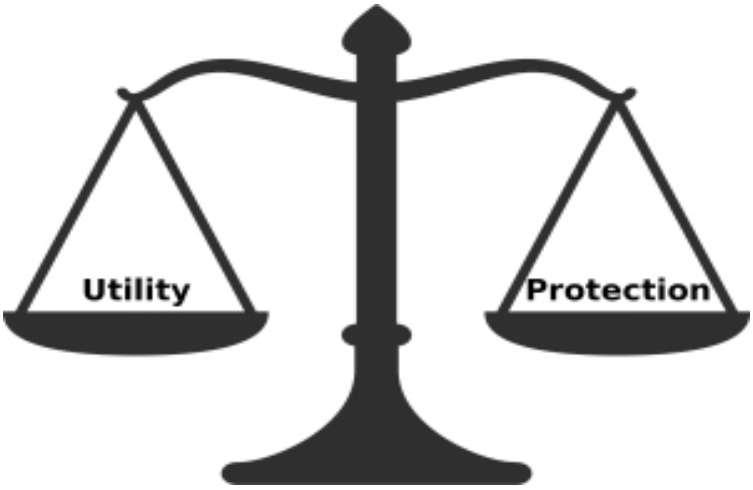
- Prevent violations and misuse of the data
- Comply with legal obligations
- Publish the data
- Make data reusable

### *Pseudonymization vs anonymization*

- Pseudonymization: data directly identifying people (names, IP addresses, phone number, etc.) is replaced by identifiers or encrypted. The key of the masked data is kept separately and securely. Pseudonymization is a good practice in order to work on personal data. It limits the risks related to a data leak. It allows to retrieve the original data, too.
- Anonymization: when time comes to publish this data, pseudonymization is seldom enough. By crossing the data, it is often possible to reidentify persons in pseudonymized datasets. Several methods prevent these risks (see below). Anonymization often means a loss of information and is not reversible. Completely anonymized data is not anymore considered as personal data and can be published.

| Pseudonymization | Anonymization |
|---|---|
| Active data | Active data and published data |
| Reversible | Not reversible |
| Key separately saved | No key needed |
| No loss of information | Loss of information |
| Strong risks of reidentification | Controlled risks of reidentification |
| Security mesures still required | No security mesure required |

Data anonymization is all about the balance between mitigating the risk of reidentification and preserving the utility of the data. The principle of proportionality applies here.




*Dilemma of data anonymization*

### *Methods*

- Removing: simply suppressing the data. It is often the appropriate solution to process direct identifiers like names, phone numbers, email addresses, IP addresses, etc. To suppress part of the outlier records is often necessary too.
- Encrypting: preserve the whole data by encrypting the identification data and keep the key secure. It is a good option for long term preservation but not for publishing the data.
- Generalizing: if the data is too specific and has unique records, the variables may be generalized in order to have less granularity.
- Shuffling: sometimes, it is possible to shuffle data over one or several columns without compromising the utility of the data. For example, if you shuffle IP addresses, you can still analyse globally these addresses but you cannot associate a record with the correct IP address.
- Adding fake data: it is possible to add fake data to a dataset and to preserve correlation factors for example. The presence of fake data may prevent individual records to be identified even if we know that a specific record is part of the dataset.

There are several variables to evaluate the anonymization level of a dataset.

### *How to evaluate the anonymization level*

- k-anonymity: a release of data is said to have the k-anonymity property if the information for each person contained in the release cannot be distinguished from at least k-1 individuals whose information also appear in the release. source
- l-diversity: l-diversity is an extension of the k-anonymity model which reduces the granularity of data representation using techniques including generalization and suppression such that any given record maps onto at least k-1 other records in the data. source
- t-closeness: t-closeness is a refinement of l-diversity group-based anonymization that is used to preserve privacy in data sets by reducing the granularity of a data representation. source
- Differential privacy: differential privacy is a process that introduces randomness into the data, for example by adding fake data or shuffling them.

### *Tools*

- sdcMicro: Statistical Disclosure Control Methods for Anonymization of Microdata and Risk Estimation (R package)
- ARX Data Anonymization Tool: Java application
- ARGUS: Java application

## Need more help?

You can contact us, if you need further advices about data anonymization.

For ethical and legal questions, the Research Office is the main respondent. It provides several useful resources:

- Dedicated webpage to Research involving work with personal data
- Ethical issues checklist (connection needed)

## Bibliography:


- Raghunathan, Balaji. (2013). *The complete book of data anonymization: From planning to implementation*. Boca Raton: CRC Press. [online at EPFL]
- Jordi Soria-Comas, Josep Domingo-Ferrer, & David Sánchez. (2016). *Database anonymization: Privacy models, data utility, and microaggregation-based inter-model connections*. Morgan & Claypool. [online at EPFL]
- Khaled El Emam Luk Arbuckle. (2013). *Anonymizing health data: Case studies and methods to get you started*. O'Reilly Media. [online at EPFL]

# Contact

researchdata@epfl.ch