



**DEFINITIONS [1]**

- ✓ “Research data [...] is collected, observed, or created, for purposes of analysis to produce original research results”
- ✓ “Recorded factual material commonly accepted in the scientific community as necessary to validate research findings...”
- ✓ “Materials generated or collected during the course of conducting research...”

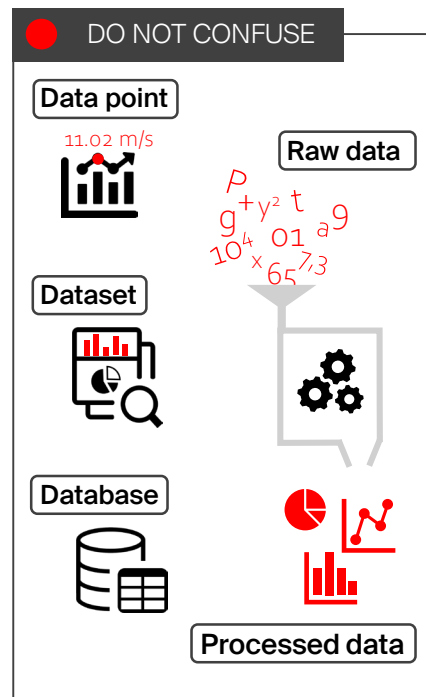
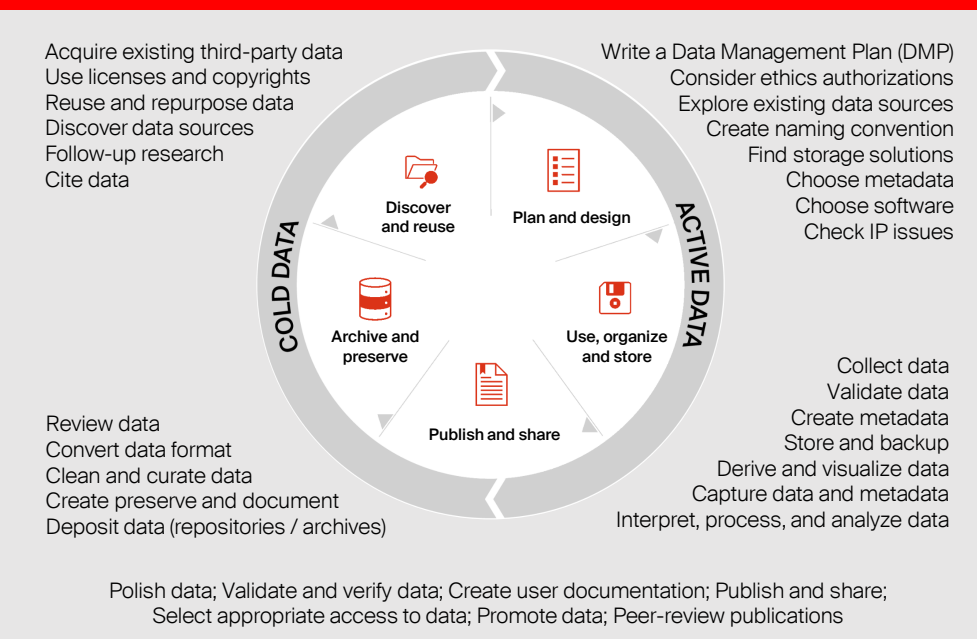
**Code is also data!**



Check the **Fast Guide #6: CODE AS DATA [2]**



**RESEARCH DATA LIFECYCLE [3]**



	Data type	Description	Examples
	<b>Observational</b>	Data captured in-situ, can not be recaptured, recreated or replaced	Sensor readings; Sensory (human) observations; Survey results; Interview notes/transcripts
	<b>Experimental</b>	Data collected under controlled conditions, in situ or lab-based. Should be reproducible, but can be expensive	Gene sequences; Chromatograms; Spectroscopy; Microscopy
	<b>Simulation</b>	The process of taking a large amount of data and using it to mimic real-world scenarios or conditions.	Climate models; Economic models; Biogeochemical models
	<b>Derived / Compiled</b>	Reproducible, but can be very expensive	Derived variables; Compiled database; 3D models
	<b>Reference / Canonical</b>	Static or organic collection [peer-reviewed] datasets, probably published	Gene sequence databanks; Chemical structures; Census data; Spatial data portals
	<b>Metadata</b>	Structured information associated with data for purposes of discovery, description, use, management, and preservation	README files [4]; Publication keywords; File and folder names

Credits and sources  
 [1] [libguides.maclester.edu/data/](https://libguides.maclester.edu/data/)  
 [2] [go.epfl.ch/rdm-fastguide06](https://go.epfl.ch/rdm-fastguide06)  
 [3] [data-archive.ac.uk](https://data-archive.ac.uk)  
 [4] [go.epfl.ch/rdm-readme](https://go.epfl.ch/rdm-readme)

Data and metadata are **easy to find** by both humans and computers

## FINDABLE

- F1** (Meta)data are assigned a globally unique and persistent identifier.
- F2** Data are described with rich metadata.
- F3** Metadata clearly and explicitly include the identifier of the data they describe.
- F4** (Meta)data are registered or indexed in a searchable resource.

### DESCRIBE

Describe provenance, usage, and organization of data with standardized **metadata** <sup>[2]</sup>.  
 Make metadata available **even if** data is not.

Both humans and computers can **readily access** or download datasets

## ACCESSIBLE

- A1** (Meta)data are retrievable by their identifier using a standardized communication protocol:
  - A1.1** the protocol is open, free and universally implementable;
  - A1.2** the protocol allows for an authentication and authorization procedure where necessary.
- A2** Metadata are accessible, even when the data are no longer available.

### OPEN

Open up your data using standardized **licenses** <sup>[3]</sup>. **Limitations** may apply (ex. data protection, commercial datasets, and double-use technology).

Data from different datasets are **prepared to be combined** or exchanged

## INTEROPERABLE

- I1** (Meta)data use a formal, accessible, shared and broadly applicable language for knowledge representation.
- I2** (Meta)data use vocabularies that follow FAIR principles.
- I3** (Meta)data include qualified references to other (meta)data.

### LINK

Use persistent **identifiers** (ex. DOI, HANDL, URN) to **cross-link** datasets. Publish files in **open formats**, even alongside proprietary formats.

Published data can be **easily combined/replicated** in future research

## REUSABLE

- R1** (Meta)data are richly described with a plurality of accurate and relevant attributes:
  - R1.1** (meta)data are released with a clear and accessible data usage license;
  - R1.2** (meta)data are associated with detailed provenance;
  - R1.3** (meta)data meet domain-relevant community standards.

### PUBLISH

Deposit datasets in data **repositories**, favoring services with user-friendly **interfaces**. Make sure to choose a **FAIR-compliant** data repository, also for the relative code.

## MY FAIR DATA?

Check the FAIRness of your dataset with this [self-assessment test](#) <sup>[4]</sup>



### FAIR ≠ Open

- ✓ FAIR ensures data can be found, understood and reused
- ✓ Data can be shared under restrictions & still be FAIR: **"As open as possible, as restricted as necessary"**

Karel Luyben, president of the EOSC <sup>[5]</sup>

## WHERE TO PUBLISH DATA?

To find the appropriate repository for your FAIR data, check the [Data platforms dissemination table](#) <sup>[6]</sup> on the EPFL Library pages



Want to dig more? Check [re3data.org](https://re3data.org) & [fairsharing.org](https://fairsharing.org)



### Credits and sources

- [1] [go-fair.org/fair-principles](https://go-fair.org/fair-principles)
- [2] [go.epfl.ch/rdm-fastguide05](https://go.epfl.ch/rdm-fastguide05)
- [3] [go.epfl.ch/rdm-fastguide12](https://go.epfl.ch/rdm-fastguide12)
- [4] [arcd.edu.au/resources/aboutdata/fair-data/fair-self-assessment-tool](https://arcd.edu.au/resources/aboutdata/fair-data/fair-self-assessment-tool)
- [5] [doi.org/10.5281/zenodo.6807345](https://doi.org/10.5281/zenodo.6807345)
- [6] [go.epfl.ch/datarepo](https://go.epfl.ch/datarepo)

RDM ACTIVITIES TO BE BUDGETED

PEOPLE	HARDWARE	SOFTWARE	SECURITY	PUBLICATION
<b>DATA MANAGEMENT PLAN</b>		DMP writing; DMP revision; DMP publishing		
<b>COLLECTION</b>		Databases and software; Data formatting; Data organization; Data transfer		
<b>ACTIVE MANAGEMENT</b>		Electronic Lab Notebook (ELN); Laboratory Information Management System (SLIMS); Data sharing platform		
<b>DOCUMENTATION</b>		Data description and metadata; Documentation and transcription		
<b>STORAGE / BACK-UP</b>		Data back-up; Data storage; NAS; Cloud		
<b>ACCESS / CONTROL</b>		Access control; Data security; Sensitive data protection; 3 <sup>rd</sup> party data		
<b>SHARING</b>		Anonymization; Copyright assessment; Data cleaning; Data publishing		
<b>PRESERVATION / ARCHIVING</b>		Data preparation; Long-term preservation; Data repository		

RDM costs are listed in the budget templates and toolkits available on the EPFL-ReO website [1]. An overview of the possible costs per research activity is also presented by the Utrecht University [2].

RDM costs can be eligible for funding applications [1]



**SNSF**

Up to CHF 10,000 for Open Research Data (ORD) activities, to prepare datasets and grant access to them in **non-commercial** repositories

**ERC / MSCA / Horizon Europe**

Costs for RDM (e.g., data storage, processing and preservation) and for Open Access to research data (APC, curation, ...) are eligible

**SOME FIGURES**

- 0.2%** Percentage relative to the total funds requested to SNSF in 2018 for ORD activities [3]
- 5%** RDM cost on the total project expenditure expected in 2016 to properly manage and steward data [4]
- 26.2bn €** Per year, cost estimation of not having FAIR data for the EU [5]

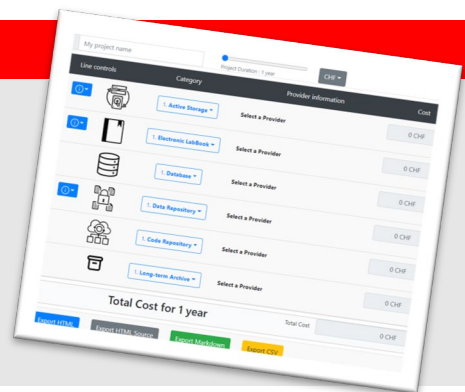
HOW TO ESTIMATE?



The **EPFL Cost Calculator for Data Management** [6] helps researchers to estimate the cost of managing, storing, and publishing data.



At **EPFL Chronos** [7] enables timekeeping for research projects which need to justify the eligible personnel costs to the funding bodies.



Credits and sources

- [1] [go.epfl.ch/ReO-toolkits\\_research-funding](https://go.epfl.ch/ReO-toolkits_research-funding)
- [2] [uu.nl/en/research/research-data-management/guides/costs-of-data-management](https://uu.nl/en/research/research-data-management/guides/costs-of-data-management)
- [3] [zenodo.org/record/3618209](https://zenodo.org/record/3618209)
- [4] [eosc-portal.eu/sites/default/files/realising\\_the\\_european\\_open\\_science\\_cloud\\_2016.pdf](https://eosc-portal.eu/sites/default/files/realising_the_european_open_science_cloud_2016.pdf)
- [5] [op.europa.eu/s/n75s](https://op.europa.eu/s/n75s)
- [6] [costcalc.epfl.ch/](https://costcalc.epfl.ch/)
- [7] [epfl.ch/research/services/fr/gerez-votre-projet/chronos/](https://epfl.ch/research/services/fr/gerez-votre-projet/chronos/)

DEFINITION

A file format is a standard way to encode data as a computer file. It follows a protocol that specifies how bits are used to encode information in a digital storage medium. File formats may be **proprietary** or **open** <sup>[1]</sup>.



DO NOT FORGET

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

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

IDEALLY...

... prefer file formats that are:

-  **Unencrypted, uncompressed**, commonly **used** by the research community
-  Compliant to an **open, documented** standard: **interoperable** among diverse platforms and applications, fully **published** and available **royalty-free**, fully and independently **implementable** by multiple software providers on **multiple platforms** without any intellectual property <sup>[2]</sup>

TYPE OF DATA	APPROPRIATE	ACCEPTABLE	DEPRECATED
<b>Tabular</b> (extensive metadata)	csv - hdf5	txt - html - tex - fastq - por	
<b>Tabular</b> (minimal metadata)	csv - tab - ods - sql - tsv	xml (if appropriate dtd) - xlsx	xls - xlsb
<b>Textual / Presentation</b>	txt - pdf - odt - odm - tex - md - htm - xml - extxyz - odf	pptx - rtf - docx - pdf (with embedded forms) - eps - ipf	doc - ppt - dvi - ps
<b>Code / Computation</b>	m - r - py - iypnb - rstudio - rmd - NetCDF - aiml	sdd	mat - rdata
<b>Image / Spectroscopy</b>	tif - png - svg - jpeg - fits	jcamp - jpg - jp2 - tif - tiff - pdf - gif - bmp - dm3 - oir - lsm	indd - ait - psd - spc
<b>Audio</b>	flac - wav - ogg - mxl - midi - mei - humdrum	mp3 - aif	
<b>Video</b>	mp4 - mj2 - avi - mkv	ogm - mp4 - WebM	wmv - mov - qt
<b>Geospatial</b>	NetCDF - tabular gis attribute data - shp - shx - dbf - prj - sbx - sbn - postGis - tif - tfw - geoJson	mdb - mif	
<b>3D structures &amp; images</b>	x3d - x3dv - x3db - pdf3D - pov - pdbml	dwg - dxf - pdb	pxp
<b>Other</b>	xml - json - rdf		

WANT MORE?

Check out also these tables:

- [fileformats.archiveteam.org/wiki/Scientific\\_Data\\_formats](https://fileformats.archiveteam.org/wiki/Scientific_Data_formats)
- [loc.gov/preservation/resources/rfs/TOC.html](https://loc.gov/preservation/resources/rfs/TOC.html)
- [archives.gov/records-mgmt/policy/transfer-guidance-tables.html](https://archives.gov/records-mgmt/policy/transfer-guidance-tables.html)
- [nationalarchives.gov.uk/information-management/manage-information/digital-records-transfer/file-formats-transfer](https://nationalarchives.gov.uk/information-management/manage-information/digital-records-transfer/file-formats-transfer)

Credits and sources

[1] [en.wikipedia.org/wiki/File\\_format](https://en.wikipedia.org/wiki/File_format)  
[2] [guides.library.stanford.edu/data-best-practices/format-files](https://guides.library.stanford.edu/data-best-practices/format-files)

DEFINITIONS

- ✓ “data that provide information about other data” [1]
- ✓ “structured information associated with an object for purposes of discovery, description, use, management, and preservation” (National Information Standards Organization, 2008) [2]

<b>MANUAL OR AUTOMATIC</b>	Metadata can be created <i>manually</i> or by <i>automated</i> information processing (i.e. captured by computer or machine)
<b>EMBEDDED OR SUPPLEMENTAL</b>	Metadata can be stored in the same file or structure as the data ( <i>embedded metadata</i> ), or in a separate file ( <i>supplemental</i> )

Family	Description	Examples
<b>DESCRIPTIVE METADATA</b>	For finding or understanding a resource	<i>Title, Author, Subject, Keywords, Publication date</i>
<b>ADMINISTRATIVE METADATA</b>		
<b>TECHNICAL METADATA</b>	For decoding and rendering files	<i>File type &amp; size, Device version, Creation date/time, Compression scheme</i>
<b>PRESERVATION METADATA</b>	Long-term files management	<i>Checksum date, Preservation event</i>
<b>RIGHTS METADATA</b>	Intellectual property rights attached to content	<i>Copyright status, License terms, Rights holder</i>
<b>STRUCTURAL METADATA</b>	Relationships of parts of resources to one another	<i>Sequence, Place in hierarchy</i>



Metadata ensures that data are FAIR

Check the **Fast Guide #2: FAIR** [3]



**RESEARCH**

- Find datasets/code using metadata
- Plan & Design metadata
- Acquire / Create metadata
- Clean metadata, control quality
- FAIR-ify (normalize, enrich, reconcile)
- Choose FAIR-compliant repositories
- Carefully fill in forms
- Publish metadata (along with data)
- Allocate / Add PIDs
- Interlink your and others' PIDs
- Hand-over for long-term preservation

Be systematic & consistent

ELEMENTS TO BUILD (YOUR OWN) METADATA

**FORMAT, TECHNICAL, INTERCHANGE STANDARDS**

[exif](#), [IPTC](#), instrumentation specific standards...

**CONTENT MODEL**

[ISA \(Investigation-Study-Assay\) framework](#), [Force11 software citation principles](#)

**NORMS, STANDARDS, REFERENCES**

[ISO 8601](#), [ISO 639-1](#), [ISO 3166-1](#), thesauri, vocabularies, authorities...

**STRUCTURE STANDARDS & SCHEMAS**

[INSPIRE](#), [SDMX](#), [Darwin Core](#), [Dublin Core](#), [PROV model](#), [Datacite](#)

Metadata and metadata **standards** creation, adoption and maintenance is a **joint effort** within and between interest-based communities

**USEFUL RESOURCES**

[www.dcc.ac.uk/resources/metadata-standards](http://www.dcc.ac.uk/resources/metadata-standards)  
[fairsharing.org/](http://fairsharing.org/) / [bartoc.org/](http://bartoc.org/) / [lov.linkeddata.eu](http://lov.linkeddata.eu)

**TRY THEM OUT** [4]

Source code citation<sup>[4]</sup>: [CodeMeta metadata generator](#)

Dataset citation<sup>[5]</sup>: [Datacite metadata generator](#)

Credits and sources

[1] [merriam-webster.com/dictionary/metadata](https://www.merriam-webster.com/dictionary/metadata)  
 [2] [framework.niso.org/24.html](https://www.framework.niso.org/24.html)  
 [3] [go.epfl.ch/rdm-fastguide02](https://go.epfl.ch/rdm-fastguide02)

[4] [codemeta.github.io/codemeta-generator/](https://codemeta.github.io/codemeta-generator/)  
 [5] [github.com/mpaluch/datacite-metadata-generator](https://github.com/mpaluch/datacite-metadata-generator)

As for data, when working with code, good management practices are needed. The publication of code is crucial to understand, validate, reuse and repeat the research.

## VERSIONING



Versioning systems are powerful code management tools.

The best-known is **Git**, it's free and open. It allows to:

- manage different versions of your code, and **track and undo changes** as needed;
- **automatically back up** code and its changes, being connected to a repository;
- collaboratively **work in a team** on the same code.

## SHARING



To **share code and make it visible**, repositories provide various services like versioning systems, wikis, task management, and issues tracking. One of the most used is **Github**.

- EPFL provides [c4science.ch](https://c4science.ch) for code versioning. Data are stored in Switzerland.
- EPFL provides [gitlab.epfl.ch](https://gitlab.epfl.ch) (GitHub open-source alternative). Data are stored at EPFL.

## DESCRIBING



The README file <sup>[1]</sup> is fundamental for coding documentation. It allows you to **explain your code**, to yourself and others. On any publication of the code, you should add documentation and rich metadata (e.g., README.txt, license.md, parameter files, comments on code <sup>[2]</sup>,...). Tools like [sphinx-doc.org](https://www.sphinx-doc.org) or [doxygen.nl](https://doxygen.nl) help by generating preformatted documentation.

## LICENSING



As for data, it is important to explain **how your code can be used or cited** by others (considering related restrictions). Instead of Creative Commons licenses as for data, for code it is recommended to use software-specific licenses, such as:

- Open source licenses (permissive as [MIT](https://opensource.org/licenses/mit-license) <sup>[3]</sup> or [GPL](https://opensource.org/licenses/gpl-license) <sup>[4]</sup>);
- Academic licenses (restrict commercial usage);
- Commercial licenses (reserve commercial usage);

Check the  
**Fast Guide #12:**  
**DATA & CODE**  
**LICENSING** <sup>[5]</sup>



## PUBLISHING



Don't forget to **generate a DOI** to uniquely identify a version of your software and to make it easily citable. Most data repositories <sup>[6]</sup> automatically generate a DOI for your code release.

**TIP:** Github provides a quick and easy integration with Zenodo <sup>[7]</sup>

## PRESERVING



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

- c4science is an EPFL solution to preserve your code **as a backup** solution.
- If you use another code repository, you can **always copy it on c4science**.
- For the **longer term**, a generic data repository (like Zenodo) is appropriate.

### Credits and sources

[1] [go.epfl.ch/rdm-readme](https://go.epfl.ch/rdm-readme)

[2] [peps.python.org/pep-0008/#comments](https://peps.python.org/pep-0008/#comments)

[3] [opensource.org/licenses/MIT](https://opensource.org/licenses/MIT)

[4] [opensource.org/licenses/gpl-license](https://opensource.org/licenses/gpl-license)

[5] [go.epfl.ch/rdm-fastguide12](https://go.epfl.ch/rdm-fastguide12)

[6] [go.epfl.ch/datarepo](https://go.epfl.ch/datarepo)

[7] [docs.github.com/en/repositories/archiving-a-github-repository/referencing-and-citing-content](https://docs.github.com/en/repositories/archiving-a-github-repository/referencing-and-citing-content)



## Are you a *handwriter*?

Many interfaces allow to write notes by using a tablet with a stylus. You can whether store the notes as digital canvas or convert them to searchable text.



When considering an ELN for your lab, ask yourself the following questions...

## DEFINITION

An Electronic Lab Notebook (ELN) is a software, local or web-based, that replicates a paper lab notebook and provides advanced functionalities.

	ELN	Paper notebook
Easy of transmission	✓	✓
Automated back-up	✓	✗
Team work uniformization	✓	✗
Interoperability	✓	✗
Storage and accessibility	✓	✗
Security	✓ Risk of hacking	✓ Risk of stealing
Long term preservation	✓ Formats/tools may become obsolete *	✓
Free of charge	✓ Open source available	✓

\* You can always save your ELN in common and open formats such as CSV or PDF/A

### PRACTICAL

- Are the **storage** method and **location** adequate?
- If our ELN is **cloud-based**, where is data hosted
- Who can access the ELN and its underlying data?
- Do we need a **single** machine where the ELN is installed?
- How does it fit in our research group's **workflow**?
- Is the technical **support** online, onsite, via hotline, ...?
- Is the ELN **business plan** the best for our group?



### INTERFACE

- Do we find the interface **suitable** for our group?
- Is it compatible with **mobile** devices?
- Do we need a **sample**/laboratory management?



### IMPORT-EXPORT

- Can we **import** our previous notes?
- Can we **export** our data in an open way?
- What are the import and export options, is there an **API**?
- What are the import and export **formats**?
- Do we have data **volume** limitations?
- Do we need **proprietary software** to reuse the data?



### INTEROPERABILITY

- Is the ELN **compatible** with other software we use?
- Can we integrate other **cloud** software (SWITCHdrive, GitHub, Zotero, ...) with the ELN?
- Is compatibility limited to the **import** of data we generate?
- Can we **integrate** new services with the ELN?
- Can we use **repositories** from within the ELN?
- Data <sup>[2]</sup>: [Zenodo](#), [MaterialsCloud](#), [Figshare](#), ...
- Code <sup>[3]</sup>: [C4science](#), [EPFL GitLab](#), ...



**EPFL ELN** <sup>[1]</sup>: Chemistry Notebook, developed internally at EPFL



✓ **SLIMS**: Commercial solution (EPFL spinoff) for a Laboratory Information Management System (LIMS), with ELN functionalities, integrating sample management and different services for life scientists <sup>[4]</sup>

✓ **ELN comparison table** from the Harvard Medical School <sup>[5]</sup>

#### Credits and sources

[1] [eln.epfl.ch](http://eln.epfl.ch)

[2] [zenodo.org](https://zenodo.org) & [materialscloud.org](https://materialscloud.org) & [figshare.com](https://figshare.com)

[3] [c4science.ch](https://c4science.ch) & [gitlab.epfl.ch](https://gitlab.epfl.ch)

[4] [agilent.com/en/product/software-informatics/lab-workflow-management-software/slims](https://agilent.com/en/product/software-informatics/lab-workflow-management-software/slims)

[5] [zenodo.org/record/472375](https://zenodo.org/record/472375)

DEFINITIONS

**Personal data** is “all information relating to an identified or identifiable person” [1]

*Examples: Name, Date of birth, Address, Photos, Videos, IP address, GPS coordinates, Telephone number, Credit card number, Number plate,...*

**Sensitive personal data** is personal data “revealing racial or ethnic origin, political opinions, religious or philosophical beliefs; trade union membership; genetic data, biometric data processed solely to identify a human being; health-related data; data concerning a person's sex life or sexual orientation” [2]



FEDERAL ACT ON DATA PROTECTION (FADP) [3]

Applies to projects conducted in Switzerland, with additional laws for research involving human beings (Human Research Act) [4]

**Principles:** Good faith, Lawfulness, Proportionality, Exactitude, Security

- **Data collected on the internet** [5] is still submitted to restrictions, even if published by the subjects
- **Hash** the identifiers if the project goals can be reached without them, and restricted access right to the **pseudomisation** key
- You need to assess the **risk of reidentification**
- **Inform the subjects** about the contact details of your unit, the purposes of your data collection, the recipients of the personal data, their right to access their personal data, and the likely consequences if they refuse to provide their personal data
- Anonymized data received from a **third party** still requires the subject to be informed of this new use
- Legal consent for subjects **under 18** years is required to collect their data
- Personal data can be **published** only if the subject consents to publication, but in no case if they are sensitive
- Guarantee these subjects' **minimal rights**: rights of access, modification, erasure



GENERAL DATA PROTECTION REGULATION (GDPR) [6]

Applies to projects involving personal data of subjects who are in the EU, with some derogations for scientific or statistical purposes (art.89)

**Principles:** Lawfulness, Data minimization, Accuracy, Storage limitation, Integrity, Transparency, Privacy-by-design, Confidentiality, Accountability

- Keep a **description** of how you will implement the principles
- If data processing or storage are **outsourced**, document external services' GDPR compliance
- In the event of a **data breach**, notify the VPSI or the DSPS immediately
- **Inform subjects** of their rights to modify their data, restrict the use and withdraw their participation, as well as extensive information about data collection/processing
- Provide a **Data Protection Impact Assessment (DPIA)** [7] if the project may result in a high risk, i.e. if it involves data processed on a large scale, innovative use of data, sensitive data, vulnerable subjects, data transfers outside the EU, ...
- Any transfer of personal **data abroad** is only guaranteed for transfers to countries [8,9] whose legislation ensures an adequate level of protection
- Guarantee subjects' **minimal rights**: rights of access, rectification, portability, objection, erasure



Revised Federal Act on Data Protection (FADP) in 2020 Expected to enter into force in September 2023 [8]

Applicable as of May 2018

Any doubt?

Contact the EPFL Human Research Ethics Committee [10]

Credits and sources

- [1] [admin.ch/opc/en/classified-compilation/19920153/index.html#a3](https://www.admin.ch/opc/en/classified-compilation/19920153/index.html#a3)
- [2] [ec.europa.eu/info/law/law-topic/data-protection/reform/rules-business-and-organisations/legal-grounds-processing-data/sensitive-data/what-personal-data-considered-sensitive\\_en/](https://ec.europa.eu/info/law/law-topic/data-protection/reform/rules-business-and-organisations/legal-grounds-processing-data/sensitive-data/what-personal-data-considered-sensitive_en/)
- [3] [https://www.fedlex.admin.ch/eli/cc/1993/1945\\_1945\\_1945/en/](https://www.fedlex.admin.ch/eli/cc/1993/1945_1945_1945/en/)
- [4] [admin.ch/opc/en/classified-compilation/20061313/index.html](https://www.admin.ch/opc/en/classified-compilation/20061313/index.html)
- [5] [edoeb.admin.ch/edoeb/fr/home/protection-des-donnees/Internet\\_und\\_Computer/services-en-ligne/medias-sociaux.html](https://www.edoeb.admin.ch/edoeb/fr/home/protection-des-donnees/Internet_und_Computer/services-en-ligne/medias-sociaux.html)
- [6] <https://gdpr-info.eu/>

- [7] [ec.europa.eu/newsroom/article29/item-detail.cfm?item\\_id=611236](https://ec.europa.eu/newsroom/article29/item-detail.cfm?item_id=611236)
- [8] [edoeb.admin.ch/dam/edoeb/fr/dokumente/2018/staatenliste.pdf.download.pdf/20181213\\_Staatenliste\\_f.pdf](https://www.edoeb.admin.ch/dam/edoeb/fr/dokumente/2018/staatenliste.pdf.download.pdf/20181213_Staatenliste_f.pdf)
- [9] [ec.europa.eu/info/law/law-topic/data-protection/international-dimension-data-protection/adequacy-decisions\\_en](https://ec.europa.eu/info/law/law-topic/data-protection/international-dimension-data-protection/adequacy-decisions_en)
- [10] <https://www.bj.admin.ch/bj/fr/home/staat/gesetzgebung/datenschutztaerkung.html>



DEFINITION [1]



**Data masking**, also called data obfuscation, is the process of **hiding original data** with modified content

ADVANTAGES

- Why it is worth
- ✓ Complies with law
  - ✓ Makes data sharable
  - ✓ Prevents data misuse
  - ✓ Makes data publishable

APPLICABILITY

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

PSEUDOANONYMIZATION



ANONYMIZATION



- ✓ FOR ACTIVE DATA
- ✓ REVERSIBLE



- ✓ FOR PUBLISHED DATA
- ✓ IRREVERSIBLE

REPLACING

Replace data by identifiers. Store the key separately and securely.

ENCRYPTING

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

REMOVING

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

GENERALIZING

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

SHUFFLING

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

FAKING

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

UTILITY PROTECTION



RESEARCH DATA

HINT

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

CHECK THESE TOOLS

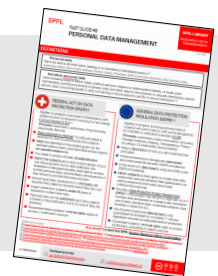
MASK IDENTITY OR ASSESS IDENTIFICATION RISKS

- ✓ [GRAASP insights](#) [2]
- ✓ [ARX Data Anonymization Tool \(Java\)](#) [3]
- ✓ [Amnesia](#) [4]
- ✓ [ARGUS \(Java\)](#) [5]
- ✓ [sdcMicro \(R\)](#) [6]
- ✓ [Differential privacy queries \(SQL\)](#) [7]
- ✓ [Faker \(Python\)](#) [8]
- ✓ [OpenPseudonymiser](#) [9]
- ✓ [AES Crypt](#) [10]



Do you deal with personal data?

Check the **Fast Guide #8: PERSONAL DATA MANAGEMENT** [12]



EPFL [EPFL Research Ethics](#) [13]

[Federal Act on Data Protection \(FADP\)](#) [14]

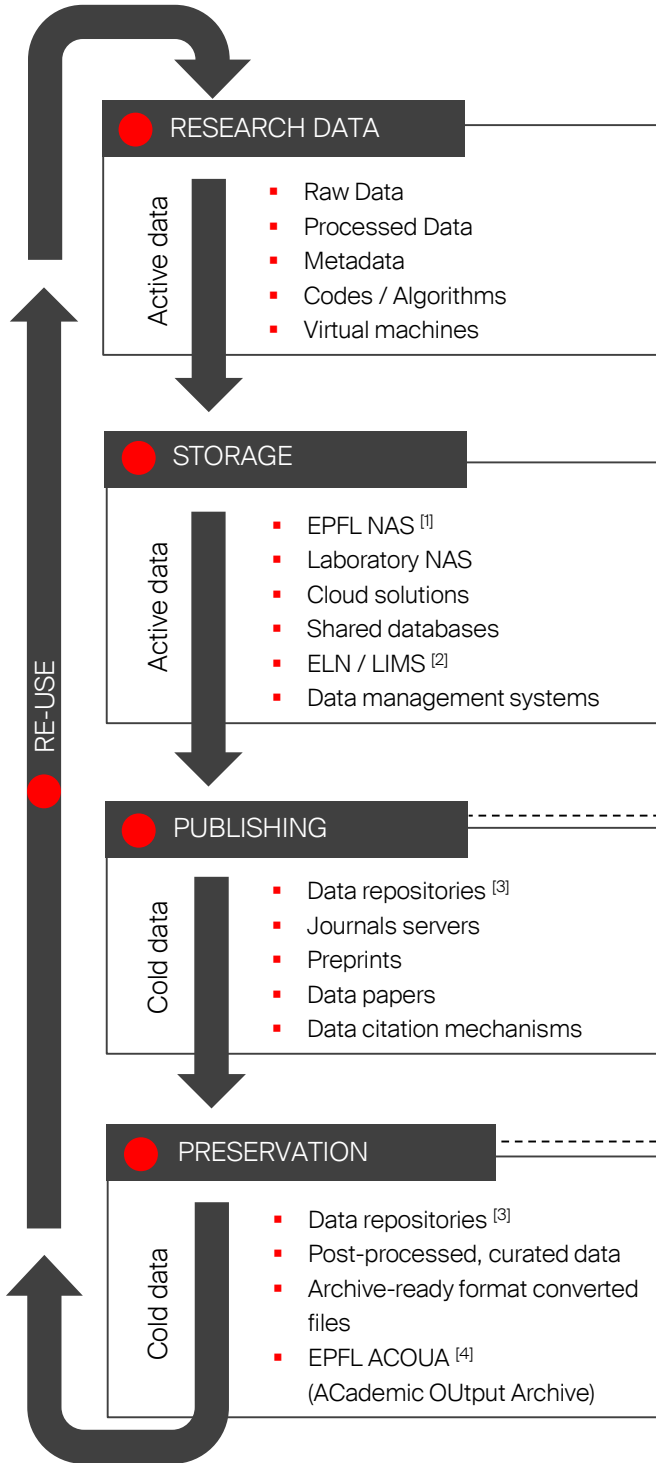
[Human Research Act \(HRA\)](#) [15]

[General Data Protection Regulation \(GDPR\)](#) [16]

Credits and sources

- [1] [en.wikipedia.org/wiki/Data\\_masking](https://en.wikipedia.org/wiki/Data_masking)
- [2] [insights.graasp.org](https://insights.graasp.org)
- [3] [arx.deidentifier.org](https://arx.deidentifier.org)
- [4] [amnesia.openaire.eu/](https://amnesia.openaire.eu/)
- [5] [qosient.com/argus/anonymization.shtml](https://qosient.com/argus/anonymization.shtml)
- [6] <https://cran.r-project.org/web/packages/sdcMicro/index.html>
- [7] [github.com/uber/sql-differential-privacy](https://github.com/uber/sql-differential-privacy)
- [8] [hfaker.readthedocs.io/en/master/](https://hfaker.readthedocs.io/en/master/)
- [9] [openpseudonymiser.org](https://openpseudonymiser.org)
- [10] [www.aescrypt.com](https://www.aescrypt.com)
- [11] [go.epfl.ch/rdm-fastguide08](https://go.epfl.ch/rdm-fastguide08)
- [12] [https://www.fedlex.admin.ch/eli/cc/1993/1945\\_1945\\_1945/en](https://www.fedlex.admin.ch/eli/cc/1993/1945_1945_1945/en)
- [13] [admin.ch/opc/en/classified-compilation/20061313/index.html](https://admin.ch/opc/en/classified-compilation/20061313/index.html)
- [14] [admin.ch/opc/en/classified-compilation/20061313/index.html](https://admin.ch/opc/en/classified-compilation/20061313/index.html)
- [15] [admin.ch/opc/en/classified-compilation/20061313/index.html](https://admin.ch/opc/en/classified-compilation/20061313/index.html)
- [16] <https://gdpr-info.eu>

**KEEP IN MIND: Store ≠ Backup ≠ Preserve ≠ Publish ≠ Archive**



**Who is involved?**



- ✓ Research teams
- ✓ Institutions
- ✓ Funders
- ✓ Research partners
- ✓ Private partners
- ✓ IT service providers

**PUBLISHING CONDITIONS**

- Data ownership
- Stakeholders consent
- Compliance with protection laws
- Ensuring data integrity
- Providing appropriate metadata
- Clarifying reuse licensing
- Setting up embargoes (if needed)

**PRESERVING CRITERIA**

- Historical and scientific data value
- Data quality and uniqueness
- Reliability of sources
- Data preparation cost
- Repository and maintenance cost
- Deposit responsibility

**HOW LONG TO PRESERVE?**

- At least 10 years for the SNSF <sup>[5]</sup>
- Evaluate preserving criteria
- Mind any retention and disposal schedules
- Stick to administrative and legal requirements

Credits and sources

[1] [go.epfl.ch/epfl-nas](https://go.epfl.ch/epfl-nas)  
[2] [go.epfl.ch/rdm-fastguide07](https://go.epfl.ch/rdm-fastguide07)

[3] [go.epfl.ch/datarepo](https://go.epfl.ch/datarepo)

[4] [go.epfl.ch/acoua](https://go.epfl.ch/acoua)

[5] [snf.ch/en/dMILj9t4LNk8NwwR/topic/open-research-data](https://snf.ch/en/dMILj9t4LNk8NwwR/topic/open-research-data) (FAQ)



What is data life cycle?

Check the **Fast Guide #1 : RESEARCH DATA: THE BASICS** [1]

FOR WHOM?

Especially for **YOURSELF** and **YOUR TEAM**, but in practice, many funders now require a DMP. Here is a non-exhaustive list:

- SNSF
- EPFL (some internal projects)
- EC (ERC, FET, MSCA, H2020, ...)
- AXA Research Fund
- U.S. Federal Grants
- Wellcome Trust
- Ligue Vaudoise Contre le Cancer
- CCR-pro

SOME TEMPLATES [5]

**EPFL DMP**  
For EPFL-funded projects or if no other template is provided, aligned with EPFL recommendations.

**SNSF DMP**  
Based on the SNSF Open Research Data Policy, with additional guiding examples.

**ERC DMP**  
Based on the FAIR principles, with additional guiding examples.

**Horizon Europe DMP (or MSCA)**  
Also for Marie Skłodowska-Curie Actions' applicants.

**NCCR RDM STRATEGY**  
Rather than a single project, it describes the data management for all the projects of a NCCR

**EPFL RDM STRATEGY**  
Based on the NCCR RDM Strategy, though mostly targeted to single research groups.

DEFINITION

A DMP - Data Management Plan - is a document describing how data and code of a research project are managed during their life-cycle

WHY

- ✓ **COMPLIANCY** Requested by research funders (public or private), a DMP enhances research reproducibility and the use of public funds.
- ✓ **TRANSPARENCY** Usually published when the funding period ends, a DMP completes the research results with the information on data, software, protocols, sources, etc.
- ✓ **FORECAST** To anticipate costs (materials and software) and identify risks (eg. data loss, incompatible formats, security). DMPs allow institutions to better allocate services.
- ✓ **STREAMLINE** To reduce risks of data loss and the efforts of reverse engineering for new collaborators. A DMP boosts data reuse in the lab and outside.

Target the reproducibility of research results!  
Anticipate questions about data in your projects.

WHAT

- ✓ **DESCRIPTION** Data types, formats, size.
- ✓ **COLLECTION** Sources, experiments, analysis, simulations.
- ✓ **CURATION** Metadata, naming, datasets structures.
- ✓ **STORAGE** Active data, sharing tools, backup, preservation.
- ✓ **RISKS** Access rights, anonymization, ethics assessment.
- ✓ **PUBLICATION** Data repositories [2], IP (ex. data licenses [3]).
- ✓ **COSTS** For RDM: refer to Fast Guide #03 [4].

Not just administrative hurdle! Use your DMP as reference tool for in-lab discussions & decisions.

WHEN

- ✓ **IDEALLY** At the conception of your research project.
- ✓ **USUALLY** When requesting funds.
- ✓ **REALLY** ASAP, but it is never too late.

The DMP is a living document! Keep it up-to-date throughout the project and secure it at the end.

Credits and sources

[1] [go.epfl.ch/rdm-fastguide01](https://go.epfl.ch/rdm-fastguide01)  
[2] [go.epfl.ch/datarepo](https://go.epfl.ch/datarepo)

[3] [go.epfl.ch/rdm-fastguide12](https://go.epfl.ch/rdm-fastguide12)  
[4] [go.epfl.ch/rdm-fastguide03](https://go.epfl.ch/rdm-fastguide03)  
[5] [go.epfl.ch/rdm-guide](https://go.epfl.ch/rdm-guide)

**LICENSING APPLIES TO...**

- Both data and code
- Original work, or from another researcher
- Collected, processed, aggregated, augmented data / code

**AT EPFL?**


EPFL owns the original data & code, but the authors can use it for research and IP [1, Art. 36]

**WHY?**

Licensing data & code is key for

- ✓ **Collaboration** with other researchers, research groups, institutes
- ✓ **Reuse** of others' work, to generate new information or data/code
- ✓ **Clarify** ownership and authorship
- ✓ **Differentiate** authorized use of original vs. derivative work
- ✓ **Share / Publish** your work with clear usage rights


The protection of data by law is not harmonized internationally but varies depending on the specific country.



**Licenses do not all have the same international recognition.**

LICENSES FOR DATA		0110 1001 1010
CC-Zero	No restrictions [2]	
CC-BY	Mandatory citation [2]	
CC-BY-SA	Mandatory citation, Share Alike (Viral) [2]	
CC-BY-ND	Mandatory citation, No Modifications [2]	
CC-BY-NC	Mandatory citation, Non Commercial [2]	
CC-BY-NC-SA	Mandatory citation, Non Commercial, Viral [2]	
ODbL	Open Access specific for databases [3]	
Microdata Research License	For unit-level data (i.e. sets of records containing on individual respondent) [4,5]	

LICENSES FOR CODE [7]		</>
MIT	Short term, Permissive, No warranty	
APACHE	Permissive, Patents allowed, No warranty	
BSD	All code by one organization, GPL mix not allowed	
GPL	Copyleft license, Patents allowed, Viral	
LGPL	Libraries Sharing, Licenses mix allowed	
AGPL	Strong copyleft, Patents allowed, Viral	



**Not sure where to start?**  
 This chooser helps you determine which Creative Commons License is right for you in a few easy steps [6]

**SHARING OR REUSING? [8]**

**Data licenses define conditions about:**

- Data ownership and use
- The treatment of original and derived data

**This is important for researchers who:**

- Receive or collect and compile data from another researcher
- Generate information or data from the other researcher's data on the other researcher's behalf or on their own behalf.

**Moreover, a researcher may want to:**

- Analyze/reuse another researcher's data
- Process/aggregate data for own research, using the processed data
- Licensing the processed, aggregated, augmented data from another researcher

**Any doubt?**  
 Contact the **EPFL Technology Transfer Office** [9]

**Credits and sources**  
 [1] [admin.ch/opc/en/classified-compilation/19910256/index.html#a36](http://admin.ch/opc/en/classified-compilation/19910256/index.html#a36)  
 [2] [creativecommons.org](http://creativecommons.org)  
 [3] [opendatacommons.org/licenses/odbl](http://opendatacommons.org/licenses/odbl)  
 [4] [microdata.worldbank.org/index.php/terms-of-use](http://microdata.worldbank.org/index.php/terms-of-use)  
 [5] [ec.europa.eu/eurostat/cros/content/microdata-access\\_en](http://ec.europa.eu/eurostat/cros/content/microdata-access_en)  
 [6] [chooser-beta.creativecommons.org](http://chooser-beta.creativecommons.org)  
 [7] [choosealicense.com/appendix](http://choosealicense.com/appendix)  
 [8] [doi.org/10.1371/journal.pbio.1002235](https://doi.org/10.1371/journal.pbio.1002235)  
 [9] [epfl.ch/research/services/units/technology-transfer-office](http://epfl.ch/research/services/units/technology-transfer-office)


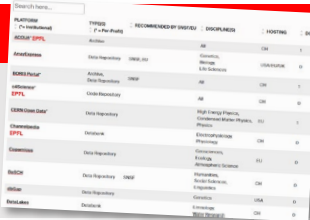
### WHY?

Publishing your research data/code might depend on many factors:

- ✓ Research ethics (transparency, accountability, reproducibility)
- ✓ Funder's request opening data/code (SNSF, Horizon Europe, etc.)
- ✓ Your research group's policy about publication
- ✓ Increase citations (data and code as academic output)

### Need to anonymize before publication?

Check the **Fast Guide #9: Data Masking** <sup>[1]</sup>

### WHERE TO PUBLISH DATA/CODE?

Check the most-used platforms adopted by EPFL researchers in this comparative [Data platforms dissemination table](#) <sup>[2]</sup>. It includes platforms for publication, as well as preservation of data and code, and for different scientific fields.




### CRITERIA TO SELECT A PLATFORM

- Institutional / External
- Data or Code-oriented / Content neutral
- Discipline specific / Discipline neutral
- Offers persistent identifiers (ex. DOI)
- Supports personal identifiers (ex. ORCID)
- Cost is sustainable
- Servers' location compliant with laws
- Suggested by funders
- Max size allowed
- Choice of licenses
- Specific: ex. preview, metadata, community, etc.

### Simple choice suggestion

Filter the comparative table for a platform that is:

- Suggested by [SNSF](#) <sup>[3]</sup> and the [EC](#) <sup>[4]</sup>
- Trusted, listed by [re3data.org](#)
- Hosted in CH or EU
- Able to provide a DOI
- Maintained by a non-profit organization
- Used by your research community




**Publishing your dataset on Zenodo?**  
 Increase the findability of your work: select the **EPFL Community** in the upload page <sup>[5]</sup>



**How to make your GitHub code citable?**  
 Follow GitHub's documentation <sup>[6]</sup> to make a persistent snapshot of your **code on Zenodo**

### 5-STAR SCHEME FOR OPEN DATA <sup>[7]</sup>

- ☆☆☆☆★ Make your stuff available on the Web (whatever format) under an open license
- ☆☆☆★★ Make it available as structured data (ex. Excel instead of scan of a table)
- ☆☆★★★★ Make it available in non-proprietary open format (ex. CSV instead of Excel)
- ☆★★★★ Use URLs to denote things, so that people can point at your stuff
- ★★★★★ Link your data to other data to provide context

*"As open as possible, as restricted as necessary"*

Karel Luyben,  
 president of the EOOSC <sup>[8]</sup>

Credits and sources  
<sup>[1]</sup> [go.epfl.ch/rdm-fastguide09/](https://go.epfl.ch/rdm-fastguide09/)  
<sup>[2]</sup> [go.epfl.ch/datarepo/](https://go.epfl.ch/datarepo/)  
<sup>[3]</sup> [snf.ch/en/WtezJ6qxuTRnSYgF/topic/open-research-data-which-data-repositories-can-be-used/](https://snf.ch/en/WtezJ6qxuTRnSYgF/topic/open-research-data-which-data-repositories-can-be-used/)  
<sup>[4]</sup> [open-research-europe.ec.europa.eu/for-authors/data-guidelines#:~:text=2.2.%20Select%20a%20Repository](https://open-research-europe.ec.europa.eu/for-authors/data-guidelines#:~:text=2.2.%20Select%20a%20Repository)  
<sup>[5]</sup> [zenodo.org/communities/epfl](https://zenodo.org/communities/epfl)  
<sup>[6]</sup> [guides.github.com/activities/citable-code](https://guides.github.com/activities/citable-code)  
<sup>[7]</sup> <https://5stardata.info>  
<sup>[8]</sup> [doi.org/10.5281/zenodo.6807345](https://doi.org/10.5281/zenodo.6807345)