



Optimizing your Research Data Management

Name1 Surname1

Name2 Surname2

Research Data Management Team,
EPFL Library

SPF

3rd March 2020

About this training material



This work is licensed under a Creative Commons Attribution-NonCommercial 4.0 International License

CC-BY-NC

Today's plan

Morning

- Introductions
- Context, FAIR & Data Management Plan
- (Break ~10h30)
- Documentation & Metadata

Afternoon

- Data formats & software
- Storage & Publication
- (Break ~14h30)
- Self-Assessment & wrap-up



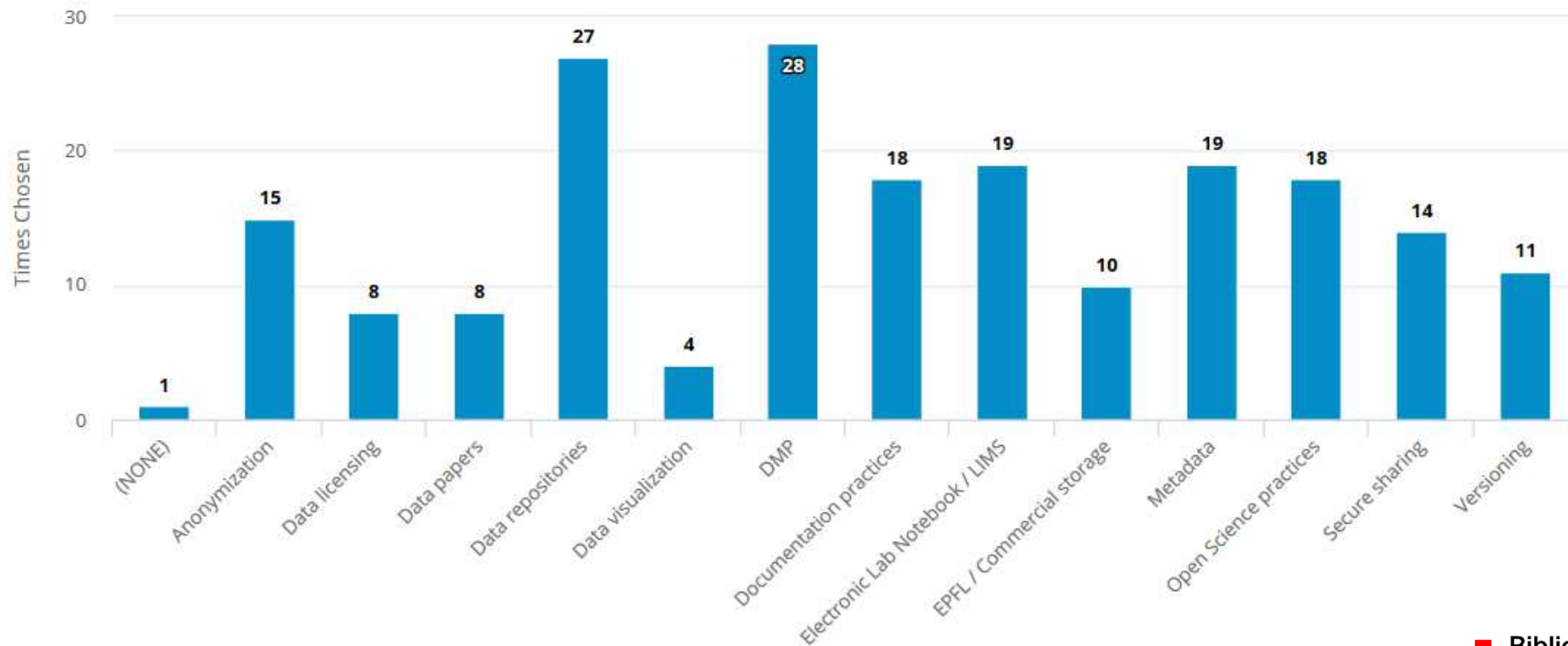
Who are you?

☐ NAME *John Carpenter*
☐ ROLE *Film director*
☐ DOMAIN *Cult cinema*
☐ DATA VOLUME *> 120 TB*
☐ DATA FORMATS *RAW, AVI, FLAC*
☐ DATA MANAGEMENT CHALLENGES
 *Share films scenes*
 *avoiding alien pirates*
☐ DATA MGMT STATUS IN MY LAB / GROUP
 *#\$@&%*!*

Participants Objectives (?)

Which RDM subject is the most important for you?

Number of responses: 74



RDM ... Let's get started



Introduction to RDM

DATA

Factual records: numerical scores, textual records, images, sounds, protocols, **source code**, etc.

RESEARCH DATA

Data used as primary sources for scientific research, and commonly accepted in the scientific community to validate research findings (OECD)

RESEARCH DATA MANAGEMENT (RDM)

The care and maintenance of research data during the research cycle (UC Berkeley Library)



NYU Health Sciences Library, youtu.be/66oNv_DJuPc

RDM also includes legal, ethical & political aspects

Research Data Lifecycle

Creating / Re-using

- Data production
- Documentation
- Data collection
- Data sources
- ...



Processing / Analyzing

- Validate data
- Cleaning data
- Transform data
- Analyse data
- Interpret data
- ...

Preserving / Publishing

- Review data
- Convert formats
- Decide IP license
- Depositing data
- Promote data re-use
- ...

FAIR Principles

- **F indable**
Data and metadata are easy to find by both humans & computers.

 - Use metadata
 - Deposit (meta)data in repository/registry
 - Assign a persistent identifier (eg. DOI, HANDL, URN)
- **A ccessible**
Machines & humans can readily access or download (meta)data.

 - As-open-as-possible access to your data (licensing, ...)
 - Services with user-friendly interfaces
 - Leave the metadata available after data deletion
- **I nteroperable**
Data from different datasets are ready to be exchanged or combined.

 - Use open file format(s), whenever possible
 - Use standardized vocabularies/tags
 - Use cross-linking as much as possible
- **R eusable**
(Meta)data are easily replicated / combined in future research.

 - Attach standardized license to your data (CC, GPL, ...)
 - Capture provenance information as precisely as possible

Download our [Fast Guides](#) ☺

More from the [GO FAIR Initiative](#)

Funding requirements

Horizon 2020

- The biggest EU research programme: ~€80 billion over 7 years (2014-2020)
- The preparation of a DMP is mandatory to receive research funding
- The research data is **open by default**, while allowing opt-outs

SNSF

- Submission of DMPs is mandatory for (most) grant applications (since October 2017)
- Researchers must **share** (at least) the data underlying their publications, to ensure reproducibility



EPFL compliance guide (p.30)

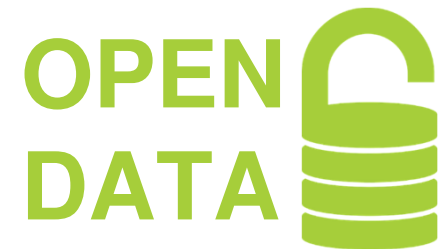
Publisher's requirements on Open Data

Many journals require authors to **publish the data underlying the published results**

Examples:

- PLoS (obligation)
- Nature journals (obligation)
- American Chemical Society (encouragement)
- Wiley journals (encouragement)
- ...

(List of editorial policies on the Dryad website¹⁹)



BE(A)WARE OF *PLAN S*

Open Data logo by the EPFL Research Data Library Team:

- <https://pixabay.com/fr/donn%C3%A9es-ouvertes-base-de-donn%C3%A9es-1518223/>
- Open Sans: <https://fonts.google.com/specimen/Open+Sans?selection.family=Open+Sans>

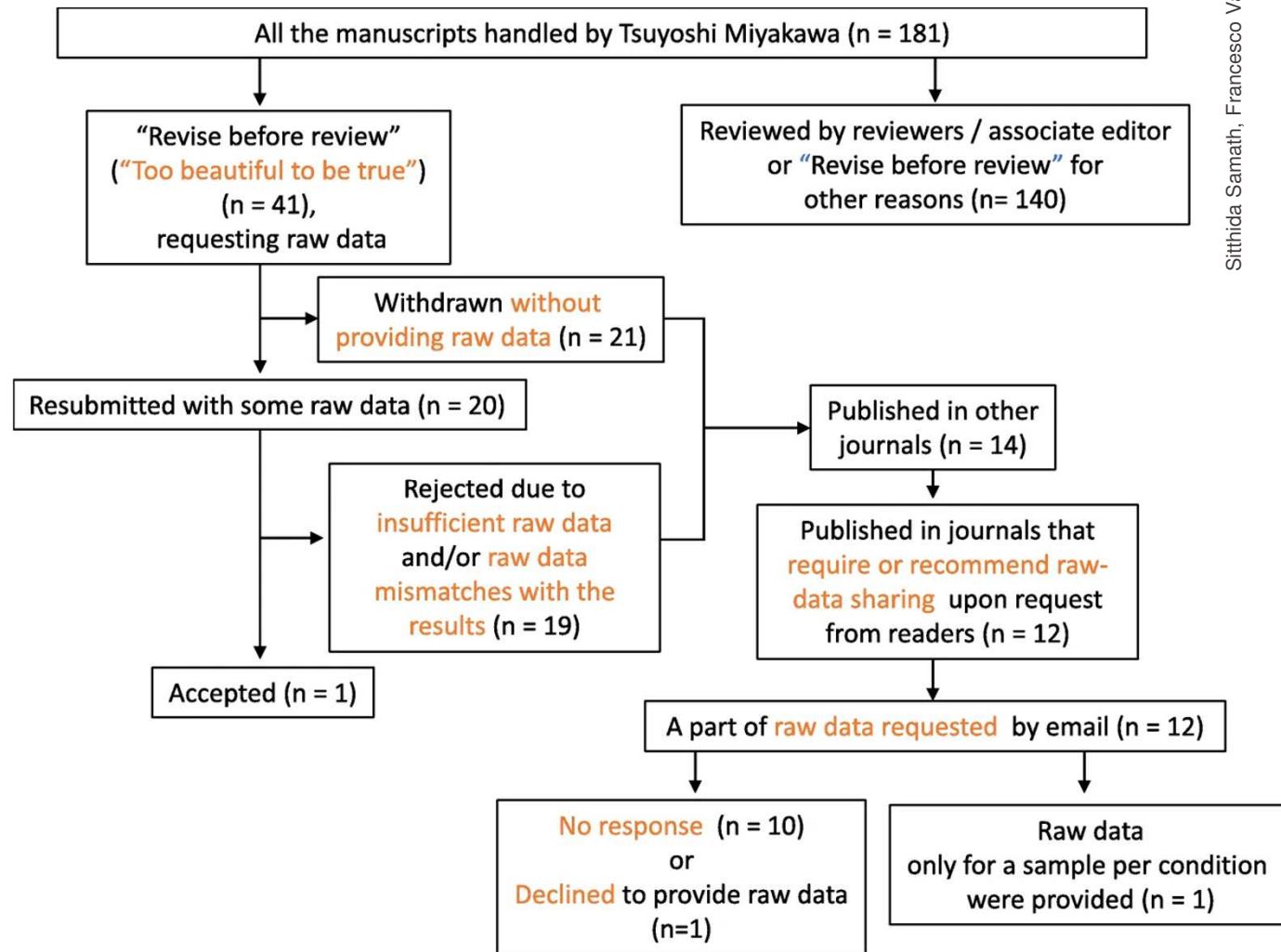
Importance of raw data (!)

EDITORIAL

No raw data, no science: another possible source of the reproducibility crisis

Tsuyoshi Miyakawa

- Lack of raw data: another possible **cause of irreproducibility**
- Many researchers did **not provide the raw data**
- Data fabrication: raw data may **not even exist** in some cases
- Good faith: the insufficiency or mismatch between raw data and results can be **honest mistakes**
- Systematic review and meta-analysis: estimated that 1.97% of authors admitted to have **fabricated, falsified, or modified data** or results at least once [...] the admission rate was 14.12% for falsification when asked about the colleagues



DOI: 10.1186/s13041-020-0552-2

RDM policies, guidelines ...

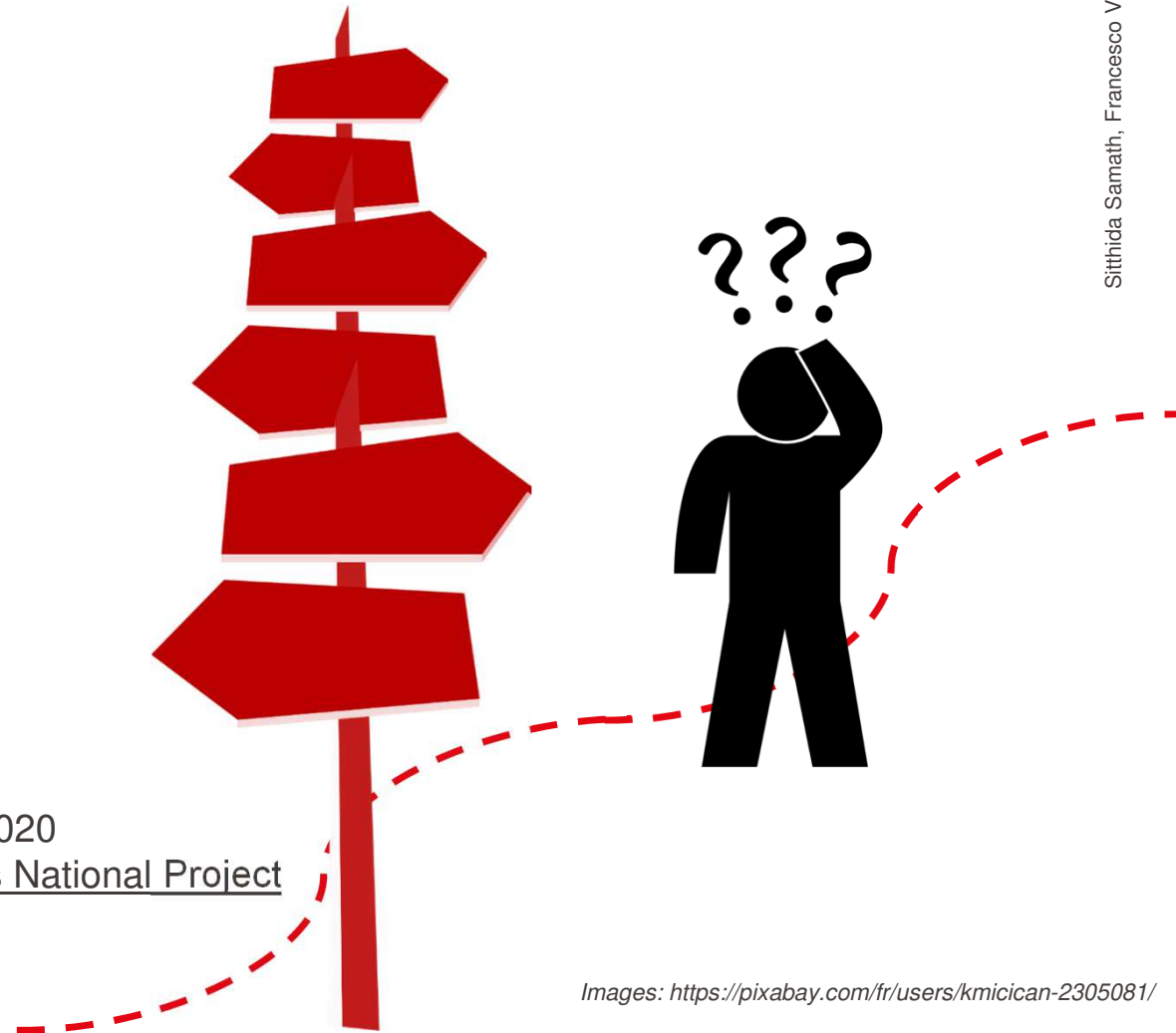
EPFL compliance guide (p.30)

Institutional policies

- Humboldt-Universität zu Berlin
- MIT
- TU Delft
- UNIGE
- University of Cambridge
- University of Edinburgh
- University of Oxford

General guidelines

- SNSF Open Research Data policies
- EC Data Management manual for Horizon 2020
- Digital Lifecycle Management (DLCM) Swiss National Project



Images: <https://pixabay.com/fr/users/kmicican-2305081/>

Openness

“As open as *necessary*, as closed as *possible*”

A.True

B.False

Why a Data Management Plan (DMP)?

- **Plan:** future needs (material, software, HR ...)
- **Science:** impact, better reproducibility, posterity
- **Data reuse:** better use of public funds
- **Openness:** impact, transparency, accountability
- **Visibility:** citations, collaborations, career
- **Compliance:** respect laws, get funds (SNSF, EC ...)
- **Efficiency:** ROI for your lab and beyond
- **Modernity:** world scale digital research, big data
- ...

What is a DMP?

- A **live document**
- &
- A **roadmap**

Describes

- strategy to manage data
- actions to take
- needed resources (time, money, people, tools ...)

Planning all along

Applications and Projects

Grant application 2

1. Personal data

- ☐ Responsible applicant
- ☐ Other applicants
- ☐ Applicants' employment
- ☐ Project partners

2. Application data

- ☐ Basic data I
- ☐ Basic data II
- ☐ Use-inspired project
- ☐ Re-submission
- ☐ Continuation of
- ☐ Link to other SNSF projects
- ☐ Further requested and available funds (not from the SNSF)
- ☐ University or research institution
- ☐ Requested funding
- ☒ Data management plan (DMP)
- ☐ Research requiring authorisation or notification
- ☐ Exclusion of external reviewers
- ☐ General remarks on the project

3. Annexed documents (upload)

- ☐ Research plan
- ☐ CV and research output list
- ☐ Quotes
- ☐ Cover letter
- ☐ Official certificates
- ☐ Lead Agency and International Co-Investigator Scheme
- ☐ Other annexes
- ☒ Administrative part of the application

Information/documents

Grant application 2

New application

Project funding in Mathematics, Natural sciences and Engineering (division II)

Deadline: **01 October 2018 17:00 Swiss local time**

Start: -

In preparation

Data management plan (DMP)

Import DMP

Please describe how you plan to make the research data Findable, Accessible, Interoperable and Reusable ([FAIR data principles](#)) in the following sections. Each of the four topics should be addressed with a level of detail appropriate to the project and research field. Sub-questions and help texts are available for each issue. The "questions you might want to consider" will help you to complete the form. However, depending on the project and research field, you may not need to address each of these questions in your DMP.

Complete the DMP form in the same language as your research plan.

The information provided in this template is not part of the scientific evaluation and will not be shared with external reviewers. Note, however, that the final version of the DMP will be published on [P3 \(public database of the SNSF\)](#) at the end of the project.

Detailed [guidelines](#) are available about the DMP. Furthermore, answers to a set of [frequently asked questions \(FAQs\)](#) about open research data (ORD) are also available.

☐ I do not submit a DMP for the following reason:

1. Data collection and documentation

- ☒ 1.1 What data will you collect, observe, generate or reuse?
- ☒ 1.2 How will the data be collected, observed or generated?
- ☒ 1.3 What documentation and metadata will you provide with the data?

2. Ethics, legal and security issues

- ☒ 2.1 How will ethical issues be addressed and handled?
- ☒ 2.2 How will data access and security be managed?
- ☒ 2.3 How will you handle copyright and Intellectual Property Rights issues?

3. Data storage and preservation

- ☒ 3.1 How will your data be stored and backed-up during the research?
- ☒ 3.2 What is your data preservation plan?

4. Data sharing and reuse

- ☒ 4.1 How and where will the data be shared?
- ☒ 4.2 Are there any necessary limitations to protect sensitive data?
- ☒ 4.3 All digital repositories I will choose are conform to the FAIR Data Principles.
- ☒ 4.4 I will choose digital repositories maintained by a non-profit organisation.

H2020 portal

ec.europa.eu/info/funding-tenders/opportunities/portal/screen/home

The screenshot displays the H2020 portal interface. At the top, there's a navigation bar with 'Grant Management' and 'Project Continuous Report'. Below this, a project summary for '723106 (RetroNets) ERC-COG' is shown, including call and topic details. A row of icons indicates the status of various deliverables: Summary for publication (green check), Deliverables (blue 'i'), Publications (red X), Dissemination (green check), Patents (IPR) (red X), Open Data (blue 'i'), and Gender (green check). Below this, a 'Deliverables' section contains a table with columns for WP No, Del Rel. No, Del No, Title, Lead Beneficiary, Nature, Dissemination Level, Est. Del. Date, Receipt Date, Approval Date, Status, and a small icon. The table lists one deliverable: WP1, D1.1, D1, Data Management Plan, EPFL, ORDP: Open Research Data Pilot, Confidential, only for members of the consortium, 30 Sep 2017, Pending, and a green icon.

List of deliverables

Deliverable Number ¹⁴	Deliverable Title	Lead beneficiary	Type ¹⁵	Dissemination level ¹⁶	Due Date (in months) ¹⁷
D1.1	Data Management Plan	1 - EPFL	ORDP: Open Research Data Pilot	Confidential, only for members of the consortium (including the Commission Services)	6

Check out:

- Statistical tool about H2020 proposals.
- Real examples of DMPs published on the EC website

(Minimum) Content of a DMP

- Institution and contacts
- Data collection and documentation
- Ethics, legal and security issues
- Resources and responsibilities
- Data storage and preservation
- Data sharing and reuse

Lots of tools and partners
to implement your DMP!

DMP horror stories

- “I am submitting a proposal for ... **if it** gets accepted then I’ll access the data concerning ...”
- “I will publish everything **open** to the public ... I will consider **not** publishing everything”
- “The dangerous **material** will be handled following the laws of ...”
- “We do not expect to produce **much** data”
- “We will publish **on line**”

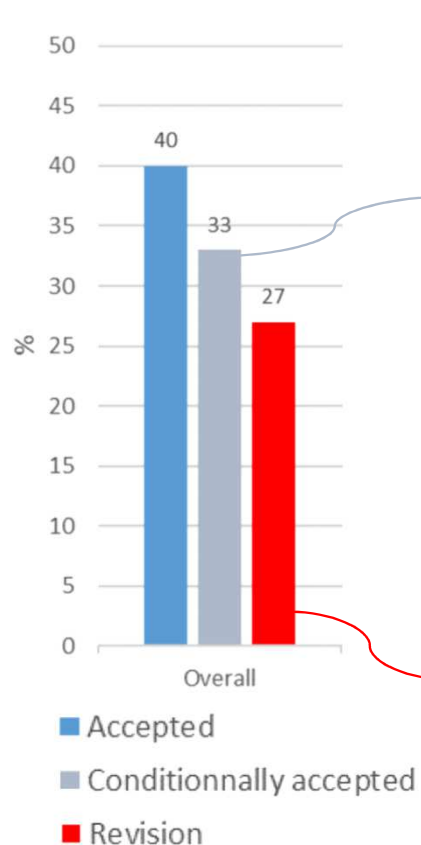
-
- No mention of metadata
 - No mention of access rights
 - No mention of data repositories
 - No mention of laptops, working stations, etc.
 - No mention of filename, or data structure conventions



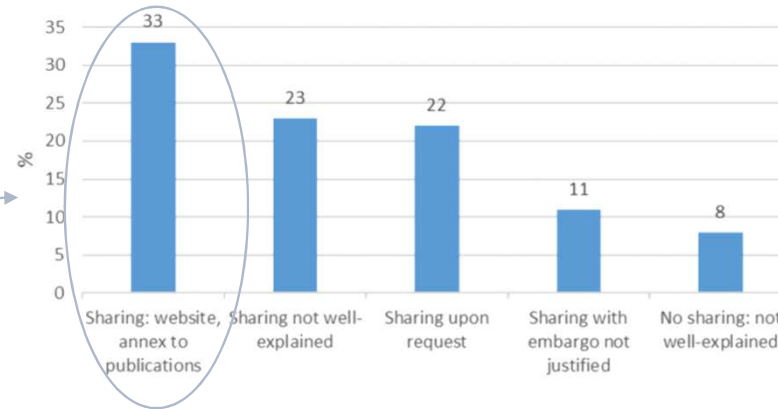
Image: <https://pixabay.com/fr/session-science-pictogramme-fatigue-1989711/>

Open Research Data

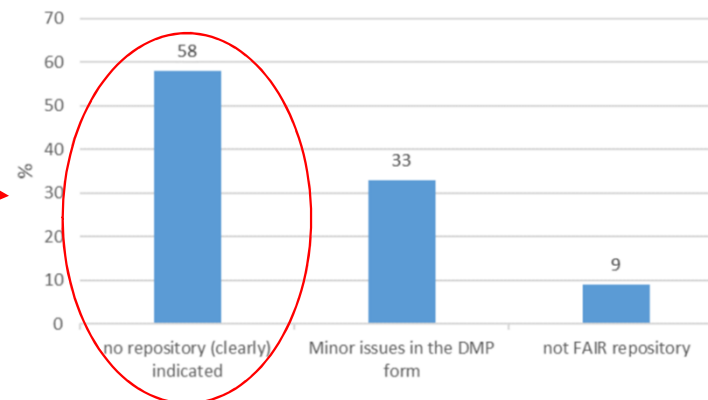
Assessment of DMP



Main reasons for DMPs sent for revision



Main reasons for DMP conditional acceptance



SNSF report 2017-2018

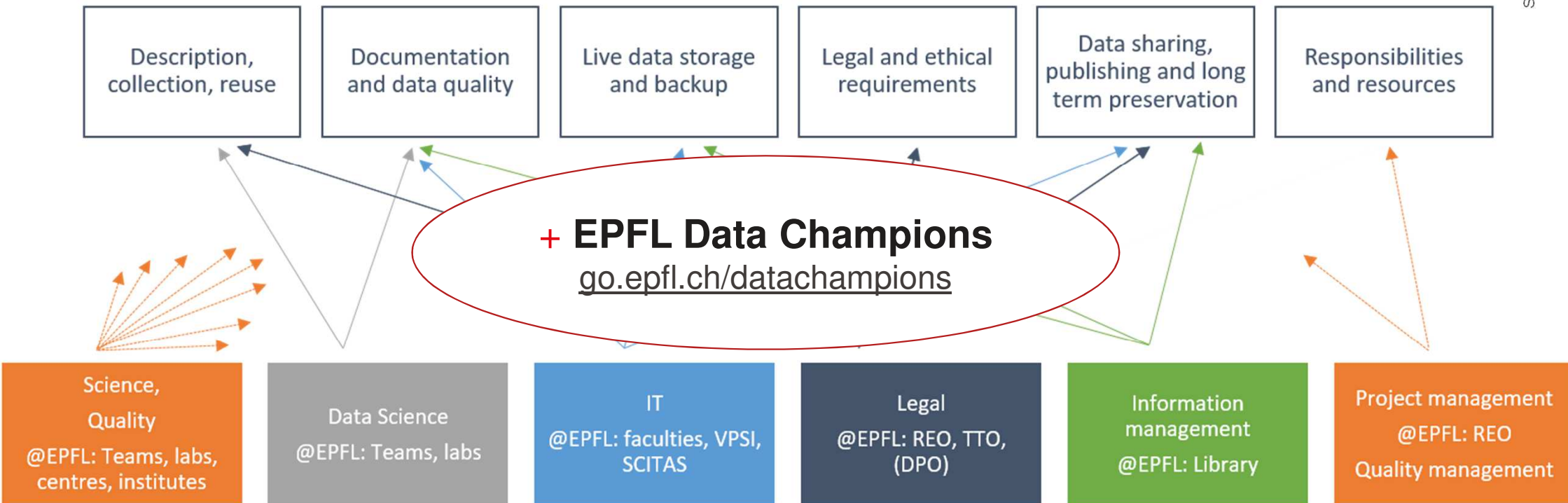
- 16% applicants requested ORD funds
- 0.2% annual costs budgeted for ORD
- 21% applications budgeted > 10k CHF
- 55% mentioned at least one data repository
- 146 different data repositories mentioned
- 6 minimal criteria for repositories

“[...] the SNSF was **flexible in the application of its criteria** [...] and that some data repositories now meet the criteria, which was not the case when the DMPs were analyzed. Therefore, data about the FAIRness of data repositories **should be interpreted in an indicative way.**”

[10.5281/zenodo.3618209](https://doi.org/10.5281/zenodo.3618209)

Beyond the DMP: Skills & Partners

Science + IT + Information Mgmt + Legal knowledge +
+ Project Mgmt + Quality Mgmt + Data Science



RDM self-evaluation

go.epfl.ch/RDM-check



➔ Download your results ➔



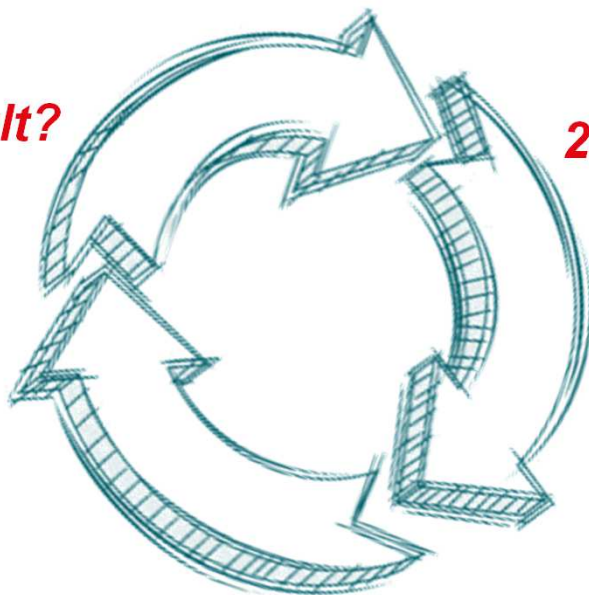
Sources: QR Code generator library of the [Project Nayuki](https://github.com/ProjectNayuki). Images: <https://pixabay.com/vectors/download-pdf-pdf-symbol-download-3660827/>

RDM self-evaluation (feedback)

1. What's your result?

2. What is already OK?

3. What to improve?



Documentation



Image source: Digitalbevaring.dk ([CC BY 2.5 DK](#))

Documentation

WHAT?

- **Description** of your data / code
- **Planned** before starting the data collection

HOW?

- Project-dependent level and **specificity**
- **Methods** every project should follow

WHY?

- Data more understandable **for yourself**
- Data more understandable **for others**
- Saves some time upon publication

Documentation methods

- **Readme** files
- Metadata **standards** & vocabularies
- **In-file** metadata (eg. *.docx* author, creation date, file tagging, etc.)
- Data **dictionaries** / Codebooks
- Folders & Files structure / **naming** convention
- **Versioning**
- **Discovery** metadata (eg. publication keywords)
- **DMP**



Data actions in your project ... to fill in!

ACTIVITIES	COLLEAGUE / PARTNER	TOOLS	TO-DO
FUNDING PLANNING			
CREATION			
ACQUISITION			
ANALYSIS			
STORING			
SHARING			
ARCHIVING			
PUBLISHING			
LEGAL CLEARANCE			
ETHICAL CLEARANCE			

Documentation all along

README

A README provides **info about data file(s) and enables reusability**

README content:

- General information
- Data and file overview
- Sharing and access information
- Methodological information

Best practices

- Write the README as a plain **text file** (open format)
- Follow your **discipline's** scientific taxonomic conventions
- 1 README per data **folder** (whenever possible)
- Name the README in accordance with described **files**
- Use the same **template** for multiple READMEs
- Use standardized **date** formats [*W3C/ISO 8601 date standard*]:
YYYY-MM-DD or YYYY-MM-DD-hh:mm:ss
- Write for human readers (does not replace metadata)

See EPFL Library README [vademecum](#) and [template](#)

```
# (README TEMPLATE, recommended fields are marked with
a *)

# Dataset title

## General information or Introduction section

author(s) info (name, affiliation, persistent id) \*
date of collection (use format) \*
geolocation data (use format)
funding or sponsorship info \*

## Sharing / Access information or License section

licenses \*
terms of use \*
citation instructions \*
links to related publications
links to other research outputs and datasets
url in repository
persistent identifiers

## Data and file(s) overview or Data section

files and folders structure description \*
file formats \*
additional related data
original source if any
dataset version, update description/changelog

## Methodological info or Preparation section and
acknowledgment section

link to publications used as base for methods
methods for processing data \*
technical requirements: necessary instruments and
software, hardware and version numbers, parameters or
calibration data \*
quality assurance process applied
people involved in experiments, surveys, processing,
analysis etc

## Data specific info
```

(README TEMPLATE, recommended fields are marked with a *)

Dataset title

General information or Introduction section

author(s) info (name, affiliation, persistent id) *

date of collection (use format) *

geolocation data (use format)

funding or sponsorship info *

Sharing / Access information or License section

licenses *

terms of use *

citation instructions *

links to related publications

links to other research outputs and datasets

url in repository

persistent identifiers

Data and file(s) overview or Data section

files and folders structure description *

file formats *

additional related data

original source if any

dataset version, update description/changelog

Methodological info or Preparation section and acknowledgment section

link to publications used as base for methods

methods for processing data *

Importance of metadata (!)

2012 – Project of officially **launched**:
Venice's State Archive + Ca' Foscari Univ. + EPFL (DHLAB)

2014 – Non-binding agreement signed. But ... didn't specify the licensing that would regulate researchers' use of the digitized data



2017 – At stake: 1,000 years of records in dynamic digital form: special high-speed scanners, thousands HD images per hour

2019 – **Allegedly**, the digitization of ~190,000 documents (8 TB) **didn't follow a common metadata policy**: archival-science guidelines (require records of provenance for each document)

Now – ... data collection has been paused, amid doubts on the usability of the data already collected!

DOI: [10.1038/d41586-019-03240-w](https://doi.org/10.1038/d41586-019-03240-w)

30




MENU **nature** [Subscribe](#)  


NEWS • 25 OCTOBER 2019

Venice 'time machine' project suspended amid data row

Disagreements among international partners leave plans to digitize the Italian city's history in limbo.

Davide Castelvocchi



[PDF version](#)

RELATED ARTICLES

The 'time machine' reconstructing ancient Venice's social networks

Saving Venice

SUBJECTS

[Databases](#) [History](#)

Historians want to use archive documents to create a virtual time machine for Venice, pictured here in the 18th century. Credit: DEA/Getty

Like the city itself, an ambitious effort to digitize ten centuries' worth of documents that record the history of Venice is at risk of sinking. Two key partners have suspended the Venice Time Machine project after reaching an impasse over issues surrounding open data and methodology. The State Archive of Venice and the Swiss Federal Institute of Technology in Lausanne (EPFL) say they have had to pause data collection, and the archive's director has raised questions about the usability of the 8

Metadata

- **Is not** mere text/char strings, **is** typed and formatted
- **Is** both machine-readable AND human-readable

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

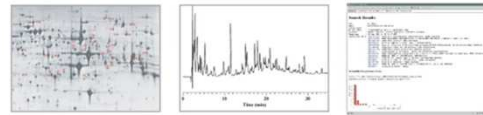
NISO. (2008) framework.niso.org/24.html

- In this instance, research data, code...
- Supporting the research data **lifecycle**

(*) Taken from Taylor, C. F., Paton, N. W., Lilley, K. S., Binz, P.-A., Julian, R. K., Jones, A. R., Zhu, W., Apweiler, R., Aebersold, R., Deutsch, E. W., Dunn, M. J., Heck, A. J. R., Leitner, A., Macht, M., Mann, M., Martens, L., Neubert, T. A., Patterson, S. D., Ping, P., ... Hermjakob, H. (2007). The minimum information about a proteomics experiment (MIAPE). Nature Biotechnology, 25(8), 887-893. <https://doi.org/10.1038/nbt1329>

Example 1

Add MIAPE metadata to proteomics experiments (*)



Data and metadata generated



Data and metadata collected by software



MIAPE-specified data and metadata

Example 2

Fill-in the form when depositing a dataset in data repository

- Add title, author, date, DOI, format, version, ...
- Info stored in repository's internal database

Publication date:
August 7, 2018

DOI:
DOI [10.5281/zenodo.1345472](https://doi.org/10.5281/zenodo.1345472)

Keyword(s):
Fixed point, complete metric space, set-valued mapping, G -metric space

Published in:
RESEARCH REVIEW International Journal of Multidisciplinary: 03 pp. 309-313.

License (for files):
[Creative Commons Attribution 4.0](https://creativecommons.org/licenses/by/4.0/)

- As html for **humans**
- In various schemas & formats for **machines**

Metadata standards

Dataset description (metadata)

- Identifier
- Title
- Creator
- Subject
- Description
- Publisher
- Date
- Formats
- Rights
- ...

Basic & General schemas

- [Dublin Core](#)
- [DataCite](#)
- ...

Disciplinary schemas

- [Digital Curation Centre](#)
- [Linked Open Vocabularies \(LOV\)](#)
- [Fairsharing](#)
- ...



Search Standards

fairsharing.org/standards

Data Repository integration (example)

October 1, 2016

Dataset Open Access

A multi-resolution, multi-epoch low Radio Frequency Survey of the Kepler K2 Mission Campaign 1 Field

Indexed in

Versions

Version 1 10.5281/zenodo.1345472
Aug 7, 2018

Cite all versions? You can cite all versions by using the DOI [10.5281/zenodo.1345471](https://doi.org/10.5281/zenodo.1345471). This DOI represents all versions, and will always resolve to the latest one. [Read more.](#)

Publication date:
August 7, 2018

DOI:
DOI [10.5281/zenodo.1345472](https://doi.org/10.5281/zenodo.1345472)

Keyword(s):
Fixed point, complete metric space, set-valued mapping, G-metric space


Published in:
RESEARCH REVIEW International Journal of Multidisciplinary: 03 pp. 309-313.

License (for files):
[Creative Commons Attribution 4.0](#)

Export

[BibTeX](#)
[CSL](#)
[DataCite](#)
[Dublin Core](#)
[JSON](#)
[JSON-LD](#)
[MARCXML](#)
[Mendeley](#)

Metadata standard usage (examples)



SNP Data Center

Home Data Publications Projects Contact Add Data Centers ▾

Entry no. 5452

Further info

Private URL

Datatype

Filename

Path

Alternative/Online Name

Author/Owner

Medium

Year created

Vegetation units of the SNP and the neighborh

http://www.parcs.ch/snp/pdf_public/2014/5452_20140814_143710_main_snp

GIS Vector Layer

veg_zoller_fullextent

Q:\maindata\snp\botany\gis_pub\zoller_veg.gdb

SNP_NALA.veg_zoller

SNP


File (digital)

1994

ISO 19115:2003

Geographic information -- Metadata

This standard has been revised by [ISO 19115-1:2014](#)

UCAR COMMUNITY PROGRAMS  **unidata**
Data Services and Tools for Geoscience

Network Common Data Form (NetCDF)

Unidata Home » NetCDF

NETCDF

Release Notes

FAQs

NetCDF C & C++ Documentation

NetCDF Fortran Documentation

NetCDF Java Documentation

Download

Support

For Developers

Compatible Software

NetCDF CDash Tests

Related Projects

Network Common Data Form (NetCDF)

 NetCDF (Network Common Data Form) is a set of interfaces for array-oriented data access and a freely distributed collection of data access libraries for C, Fortran, C++, Java, and other languages. The netCDF libraries support a machine-independent format for representing scientific data. Together, the interfaces, libraries, and format support the creation, access, and sharing of scientific data.

See the netCDF package overview ▸

NetCDF News & Announcements

NetCDF 4.6.3
3 mars 2019

NetCDF 4.6.2
21 novembre 2018

NetCDF 4.6.1
20 mars 2018

NetCDF news archive ▸

Citing NetCDF

If you use netCDF and want to provide a DOI/citation, see [How to Acknowledge Unidata](#).

NetCDF Fact Sheet

A netCDF fact sheet [is](#) provides a brief overview of the netCDF package and supported languages and platforms.

View the netCDF fact sheet ▸

on map VEG 1:50000 covers the entire area of the Swiss National
/EG differs between 39 vegetation units ranging from montane to a

karte: Vegetation map of the SNP and its surroundings

	DATATYPE	ALTERNATIVE	YEAR
INA - Vegetation Map Zoller	GIS Vector Layer	Kantonale Verwaltung Graubünden, Amt für Langsamverkehr	2009
Vegetationskartierung	Project	SNP	2013
Carex	Project	SNP	2013
Carex: Magerwiesen SNP	GIS Vector Layer	SNP	2013
Carex: Borst- und Blaugrashalden	GIS Vector Layer	SNP	2013
33833 Archivdatensatz: Vegetation units of the SNP and the neighborhood (Zollerkarte)	GIS Vector Layer	SNP	1994

ISO 19115:2003 defines the schema required for describing geographic information and services. It provides information about the identification, the extent, the quality, the spatial and temporal schema, spatial reference, and distribution of digital geographic data.

ISO 19115:2003 is applicable to:

- the cataloguing of datasets, clearinghouse activities, and the full description of datasets;
- geographic datasets, dataset series, and individual geographic features and feature properties.

ISO 19115:2003 defines:

- mandatory and conditional metadata sections, metadata entities, and metadata elements;
- the minimum set of metadata required to serve the full range of metadata applications (data discovery, determining data fitness for use, data access, data transfer, and use of digital data);
- optional metadata elements - to allow for a more extensive standard description of geographic data, if required;
- a method for extending metadata to fit specialized needs.

Though ISO 19115:2003 is applicable to digital data, its principles can be extended to many other forms of geographic data such as maps, charts, and textual documents as well as non-geographic data.

NOTE Certain mandatory metadata elements may not apply to these other forms of data.

Data dictionaries / Codebooks

Explain variables used in a dataset, within a table

Sheet_1

Show rows with cells including:

Variable	Variable name	Mesaurement unit	Allowed values	Description
Participant ID number	ID	Numeric	001-999	ID number assigned to participant in sequential order
Group number	GROUP	Numeric	1-30	Group assigned to participant based on ID number
Age in years	AGE	Numeric	18.0-65.0	Age of participant in years
Date of birth	DOB	mm/dd/yyyy	1-12/1-31/1951-1998	Participant's date of birth
Gender	SEX	Numeric	1 = male 2 = female	Participant's gender
Date of survey	SURVEY	mm/dd/yyyy	01/01/2015 – 01/01/2016	When the participant completed the survey
Self-reported consumer spending	SPEND	Numeric	0-100,000,000	Self-reported average yearly expenditure
Market sentiment	SENTIMENT	Numeric	1 = negative 2 = neutral 3 = positive	Sentiment towards US domestic economy
Actual GDP growth	GDP	Numeric	-5.0-5.0	Average US yearly GDP growth

Example of content:

variable name, variable label, variable definition, units of measure, allowed ranges, value code, missing data, etc.

Discover more on how to **Create a Codebook** on the Data Documentation Initiative (DDI) [Alliance website](#).

Data categorization: Example

Following data are generated in order to investigate neural processing that produce behavior.

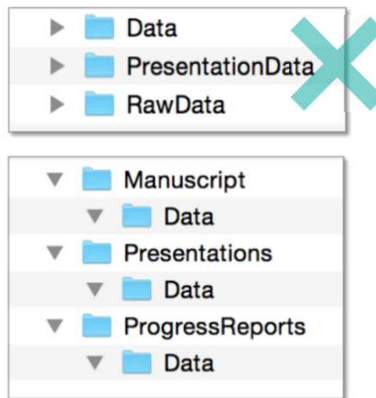
[A] New experimental data (format, size)			[B] Analyzed data (format, reuse, origin, size)		
1.	Cortical Imaging Data:	.mat, 20 TB	1.	Imaging Data (A1):	.mat, 10 TB
2.	Behavioral Filming Data 1:	.tif, 20 TB	2.	Filming Data (A2, A3):	.bin, 5 TB
3.	Behavioral Filming Data 2:	.avi, 10 TB	3.	Filming Data (A2, A3):	.mat, 5 TB
4.	Behavioral Task Data:	.txt, 500 GB	4.	Behavioral Data (A4, A5):	.mat, 500 GB
5.	Behavioral Log Data:	.bin, 1 TB	5.	Histology Data (A8):	.mat, 500 GB
6.	Optical Control Data:	.txt, 1 TB	6.	EPhys Data (A9):	.mat, 500 GB
7.	Experiment Log:	.xlsx, 1 GB			
8.	Histology Data:	.tiff, 1 TB			
9.	Electrophysiology Data:	.mat, 5 TB			

Source: DMP draft by Keita Tamura, Marie Curie fellowship application

Dataset organization

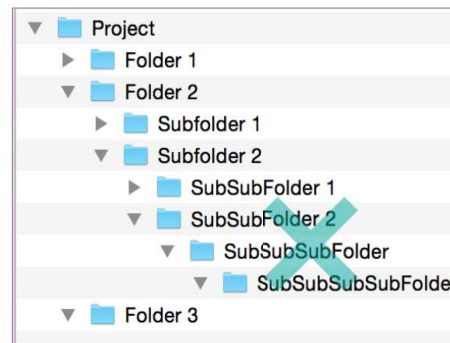
Try to **avoid** ...

overlapping categories



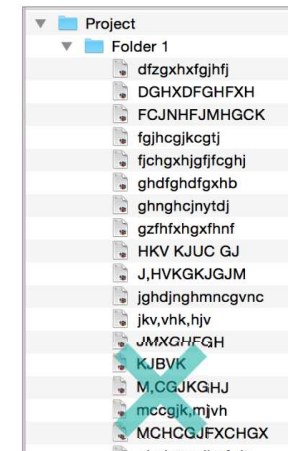
Rule of thumb:
“*sure of the right subdirectory*”

too deep structures



Rule of thumb:
“*no more than 3 clicks*”

too crowded folders



Rule of thumb:
“*fit in one screen*”

Check out:

<https://library.stanford.edu/research/data-management-services/data-best-practices/best-practices-file-naming>

<https://libraries.mit.edu/data-management/files/2014/05/file-organization-july2014.pdf>

Discussion: file naming

1. **One year from now**, will you recognize what your files contain?
2. **What information** needs to be contained in a file name?
3. **What would you change** in the following names?

My passwords.doc	My data.xls
IMPORTANT.doc	My study.doc
My Thesis final final.doc	Doc.1.doc
My Thesis version 12.doc	New doc.doc
Data 01/08/2016.xls	Int 1 (2).doc
Data 10 jan. 2016.xls	Interview 1.doc

Naming convention(s)

For both **folders** and **files**

Limit file name to **32 characters** (better less)

32CharactersLooksExactlyLikeThis.csv

Use leading zeros for multi-digit versions.

For a sequence of 1-10: 01-10

For a sequence of 1-100: 001-010-100

NO ProjID_1.csv ProjID_12.csv
YES ProjID_01.csv ProjID_12.csv

Don't use spaces. Some software read file names with spaces enclosed in quotes when used in the command line.

NO Proj ID 1.csv
YES Proj_ID_01.csv
YES Proj-ID-01.csv

Don't use special characters:

~ ! @ # \$ % ^ & * () ` ; < > ? , [] { } ' "

NO name&date@location.doc

Use a good format for date designations:

YYYYMMDD or YYMMDD.

ProjID_01_20180305.csv

Use only one period (before the file extension)

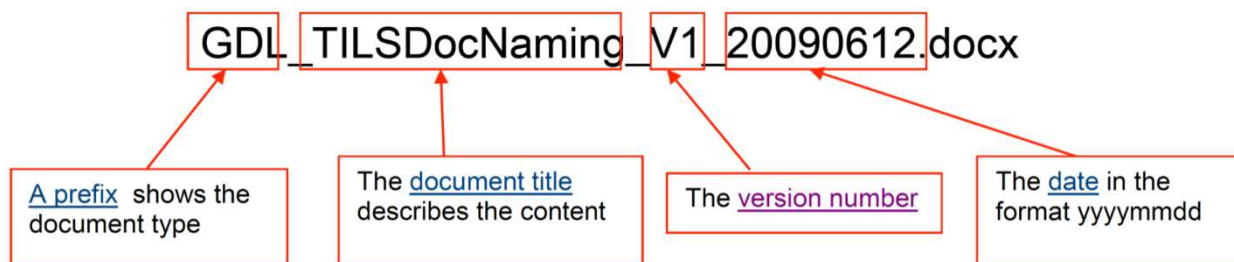
NO name.date.doc
NO name_date..doc
YES name_date.doc

Use specific file names: generic ones may conflict when moved between locations.

NO MyData.csv
YES ProjID_data.csv

See EPFL Library File organization and file naming [vademe cum](https://hmd.youmi-lausanne.ch/QcqQzvhwS3m3YucOG0A6MA#)
<https://hmd.youmi-lausanne.ch/QcqQzvhwS3m3YucOG0A6MA#>

TILS Document Naming Convention



Some Automatic Renaming Tools

- Bulk Rename Utility (Win; free)
- Renamer 4 (Mac)
- PSRenamer (Linux, Mac, Win; free)

File naming example

The researchers wanted to track several things about the tiles:

1. **Study site.** Indicated by the name, ex. FR3, FR7, FR9.
2. **Depth of the water.** Indicated by S (shallow), M (middle), or D (deep).
3. **Date.** Indicated by YYMMDD.
4. **Tile number.** Indicated on the tile.
5. **Tile treatment.** Indicated by C (caged) or U (uncaged).
6. **Number assigned to photo by camera.**
7. **Whether the post-removal photo was of the entire tile or a tile section.**
Indicated by W (whole area), A (upper right), B (lower right), C (lower left), or D (upper left).

Example: FR3S.140623.129C.2653.W.JPG

This was image 2653 of whole, uncovered tile 129 from study site 3 in shallow water, taken on June 23, 2014.

Source: <https://libraries.mit.edu/data-management/files/2014/05/file-organization-july2014.pdf>

Ex.: Your own naming convention

For both **folders** and **files**

Limit file name to **32 characters** (better less)

32CharactersLooksExactlyLikeThis.csv

Use leading zeros for multi-digit versions.

For a sequence of 1-10: 01-10

For a sequence of 1-100: 001-010-100

NO ProjID_1.csv ProjID_12.csv
YES ProjID_01.csv ProjID_12.csv

Don't use spaces. Some software read file names with spaces enclosed in quotes when used in the command line.

NO Proj ID 1.csv
YES Proj_ID_01.csv
YES Proj-ID-01.csv

Don't use special characters:

~ ! @ # \$ % ^ & * () ` ; < > ? , [] { } ' "

NO name&date@location.doc

Use a good format for date designations:
YYYYMMDD or YYMMDD.

ProjID_01_20180305.csv

Use only one period (before the file extension)

NO name.date.doc
NO name_date..doc
YES name_date.doc

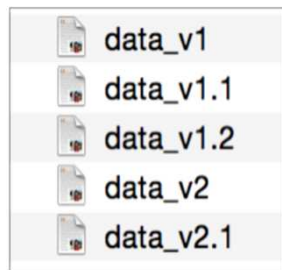
Use specific file names: generic ones may conflict when moved between locations.

NO MyData.csv
YES ProjID_data.csv

1. **Write** your own example of naming convention (5')
2. **Discuss** it in groups of 2-3 peers (2')
3. **Explain** some difficulties to everyone (5')

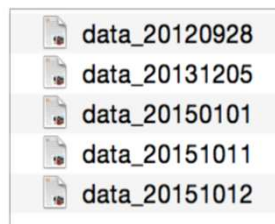
File versioning

Sort



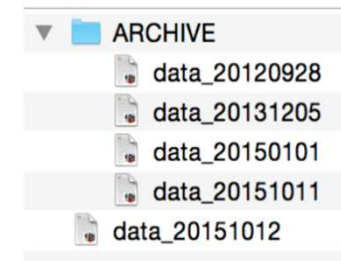
Major changes: ordinal numbers
Minor changes: decimal numbers

Distinguish



Dates distinguish between the different versions.

Separate



Archive older files in a separate folder.

Versioning solutions



Git Project



Git LFS



GitHub



EPFL GitLab



C4Science

EPFL



TortoiseGit



git-annex



Mercurial


Discussion: what information is missing?

Home / Organizations / Magnetic Oxides Group / #151027b

#151027b

Followers
0

Organization



Magnetic Oxides Group

There is no description for this organization

Social

Google+

Twitter

Dataset Groups Activity Stream

#151027b

FMR data for 151027b

Data and Resources

frequency_sweep 8-18GHz

Explore

FMR_10.12GHz

Explore

FMR

Additional Info

Field	Value
Author	Martin Buchner
Maintainer	Martin Buchner
Last Updated	31 mars 2017, 11:22 (UTC+02:00)
Created	31 mars 2017, 11:21 (UTC+02:00)

go.epfl.ch/btF



Source: QR Code generator library of the [Project Nayuki](#).

BOX: Data cleaning: to be documented, too

“60% data scientists say they spend the most time cleaning and organizing data”

Crowdfunder 2016 Datascience report

When	Preprocessing 1st step (if applicable)	Quality assurance Sub-process in the whole process
Motivation	Data ready for analysis	<ul style="list-style-type: none"> ▪ Data ready for analysis / sharing / publishing / preservation / ... ▪ Compliance
How	<ul style="list-style-type: none"> ▪ Transform / Reformat / Clean / Merge / Reconciliate data ▪ Detect errors / aberrations ▪ Try tools such as <u>OpenRefine</u> 	<ul style="list-style-type: none"> ▪ Define expected quality / criteria in a policy (completeness, consistency, accuracy, integrity...) ▪ Implement quality control with human / machine protocols / procedures

GARBAGE IN – GARBAGE OUT

Formats / Software / Storage



Image source: Digitalbevaring.dk (CC BY 2.5 DK)

Research reproducibility issue

“There are two possible outcomes: if the result confirms the hypothesis, then you've made a measurement. If the result is contrary to the hypothesis, then you've made a discovery.”

Enrico Fermi



Nature Methods (2014)



Science & Society (2017)



Science (2018)



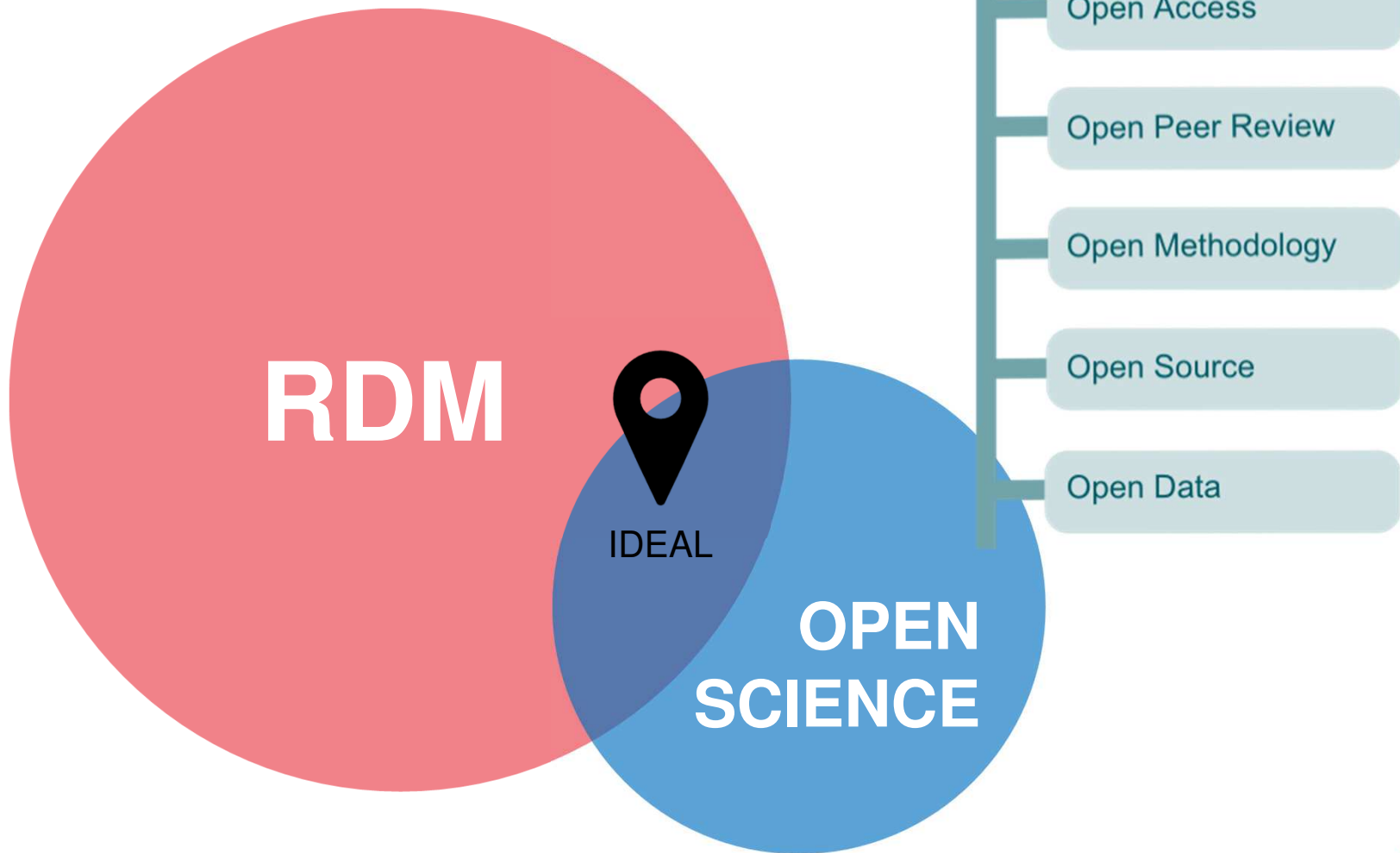
Two projects of the *Open Science Framework*:

- Reproducibility Project: Cancer Biology
- Reproducibility Project: Psychology

Some readings:

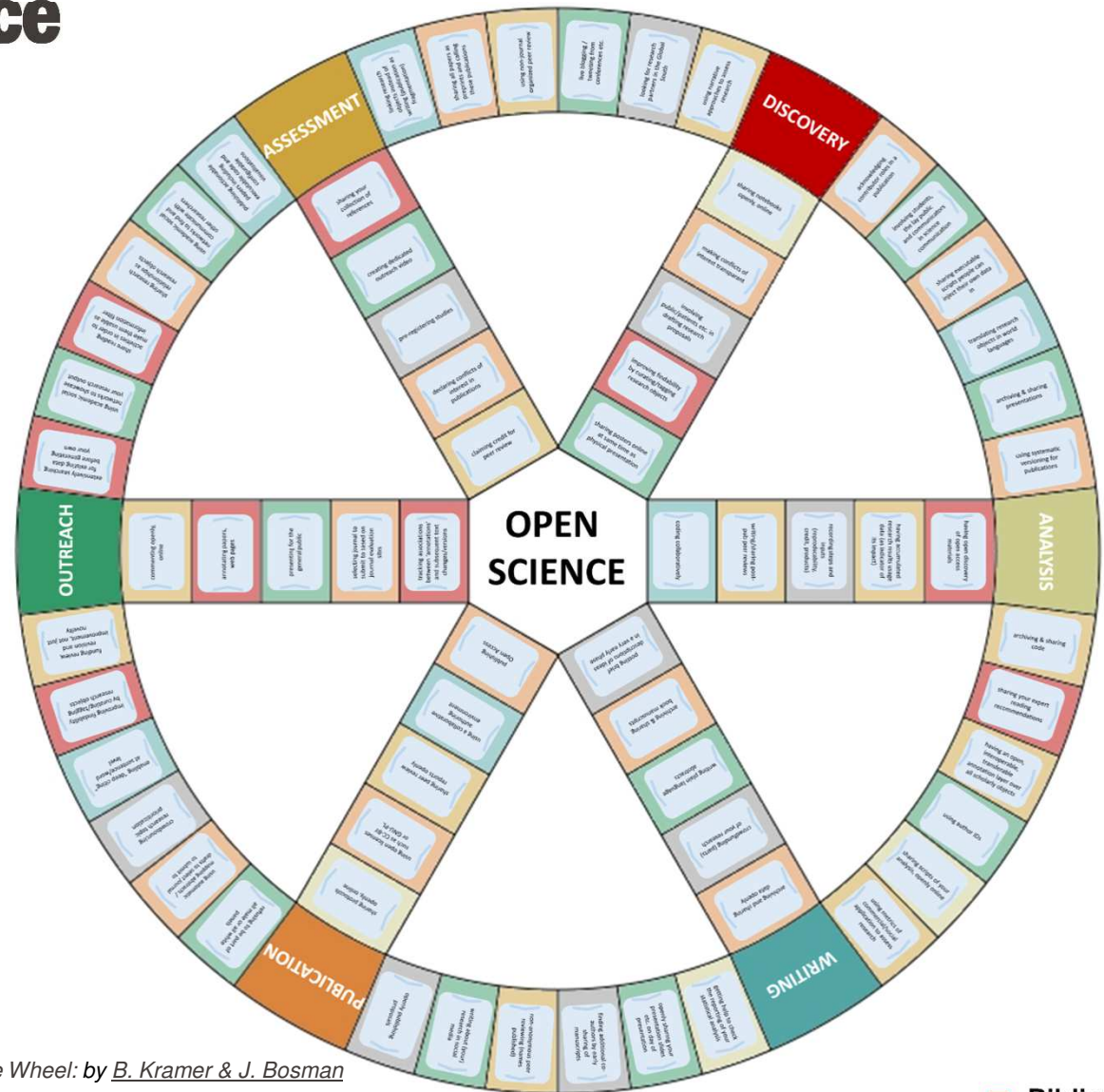
- Implementing Reproducible Research
- Reproducible research with R and Rstudio

RDM \Rightarrow Open Science?



Wheel of Open Science

[go.epfl.ch/OS Wheel](https://go.epfl.ch/OS_Wheel)



Sources: QR Code generator library of the [Project Nayuki](#). Open Science Wheel: by [B. Kramer & J. Bosman](#)

Open Data Decision Tree

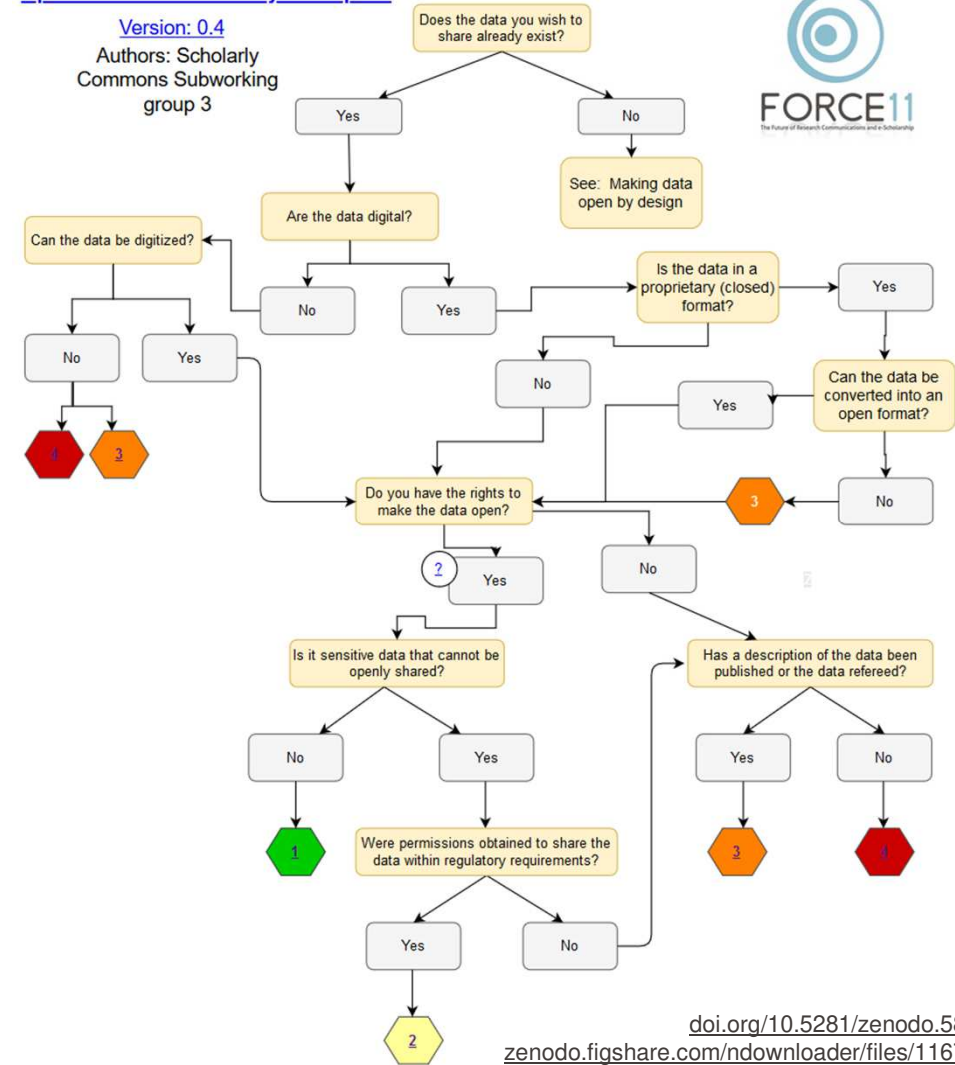
go.epfl.ch/OD_Tree



Open Data: Can I make my data open?

Version: 0.4

Authors: Scholarly
Commons Subworking
group 3



doi.org/10.5281/zenodo.581415

zenodo.figshare.com/ndownloader/files/11673289

Tools: Examples

49

DISCOVERY



Unpaywall.org

For any given DOI, get a OA version of it

Idea: Use online platforms to download articles published in Open Access and more

ASSESSMENT



Altmetric (?)

Assess impact via citations + conversations

Idea: Diversify the way your own research is assessed by including openness metrics

OUTREACH



ORCID

Customize & Update your ORCID account

Idea: Use an open protocol, persistent digital identifier when disseminating your research

WRITING



HackMD.io

Remotely collaborate on written content

Idea: Instant, collaborative online tool based on open source protocol markdown language

PUBLICATION



Zenodo

Disseminate documents + data + code

Idea: Use of online platforms to disseminate your research outputs as openly as possible

ANALYSIS



Protocols.io

Collaborate on lab/code analysis & open them

Idea: Collaborate on / Crowd-source protocols (with versioning), not only experiments

File formats

Standardized, open & widely used formats to:

- ... work on **multiplatform** / multi OS
- ... **collaborate** with more people
- ... avoid **licensing** problems
- ... maximize future research **reusability**
- ... be **independent** of a particular software / company

Examples of Open data formats

- **PDF/A:** ISO standard, archiving, no ciphers, included fonts, ...
- **CSV:** apt for tables, extensible with CSV on the Web
- **SVG:** web friendly, native multiplatform support
- **SQL** (databases communication, Postgresql, PostGIS)
- **MySQL** or MariaDB⁶⁵ (supported by the EPFL central IT)
- **HDF5** (flexible, widely compatibility, Python, R, Matlab, ...)

FAST GUIDE #04

FILE FORMATS

EPFL Library
Research Data Management
FAST GUIDES

Definition

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

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

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

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

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

File formats extensions for reusability/preservation:

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

For further information: [List of EPFL Recommended File Formats³](#)

Credits and sources

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

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

[3] https://researchdata.epfl.ch/wp-content/uploads/2018/05/Recommended_DataFormats_2018_03_05_Final.pdf

Contact and info

researchdata.epfl.ch

researchdata@epfl.ch

BIBLIOTHEQUE

Download our Fast Guides ☺

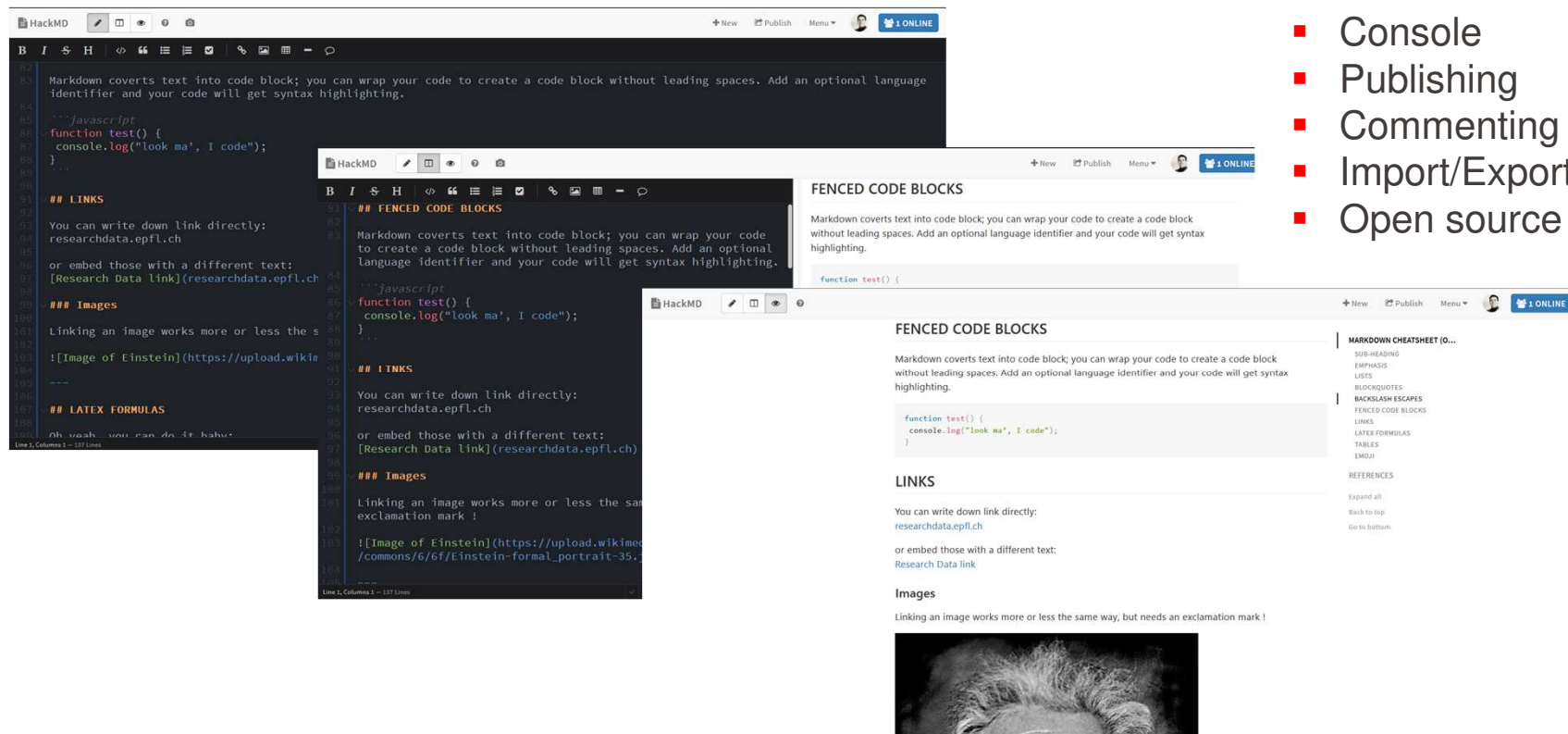
■ Bibliothèque

Open software – Example (hackMD.io)

1. Writing collaborative tool: [hackMD.io](https://hackmd.io)
2. Markdown cheat/sheet: go.epfl.ch/bsX

Why hackMD?

- Multilanguage
- Access management
- WYSIWYG → YSWYG
- Console
- Publishing
- Commenting
- Import/Export/Download
- Open source



What about?

- Office 365
- Google Docs
- Authorea
- Word
- Writer
- LibreOffice
- LaTeX
- ShareLaTeX
- OverLeaf

NOT great ideas?

Ex.: Open format files (1/2)



SUMMARY

[...] Musical scores will be stored in MusicXML or MIDI format [...]

...

4. INCREASE DATA RE-USE

[...] format converters will be employed (or implemented) to keep copies of MusicXML in other formats, such as MEI or Humdrum (which have been stable for a long time already).

Research data is open by default since 2017

By EPFL *Digital and Cognitive Musicology Lab*

1. **Search** for the openness of these formats (5')
2. **Discuss** the reasons behind the choice (5')

Ex.: Open format files (2/2)



musicXML™

The **standard open format** for exchanging digital sheet music [...] designed from the ground up for sharing sheet music files between applications, and for archiving sheet music files for use in the future.

MIDI

An **industry standard** music technology protocol. Wrapper format for MIDI data is relatively transparent, fully documented, **without** any licenses and patents in the underlying technology.



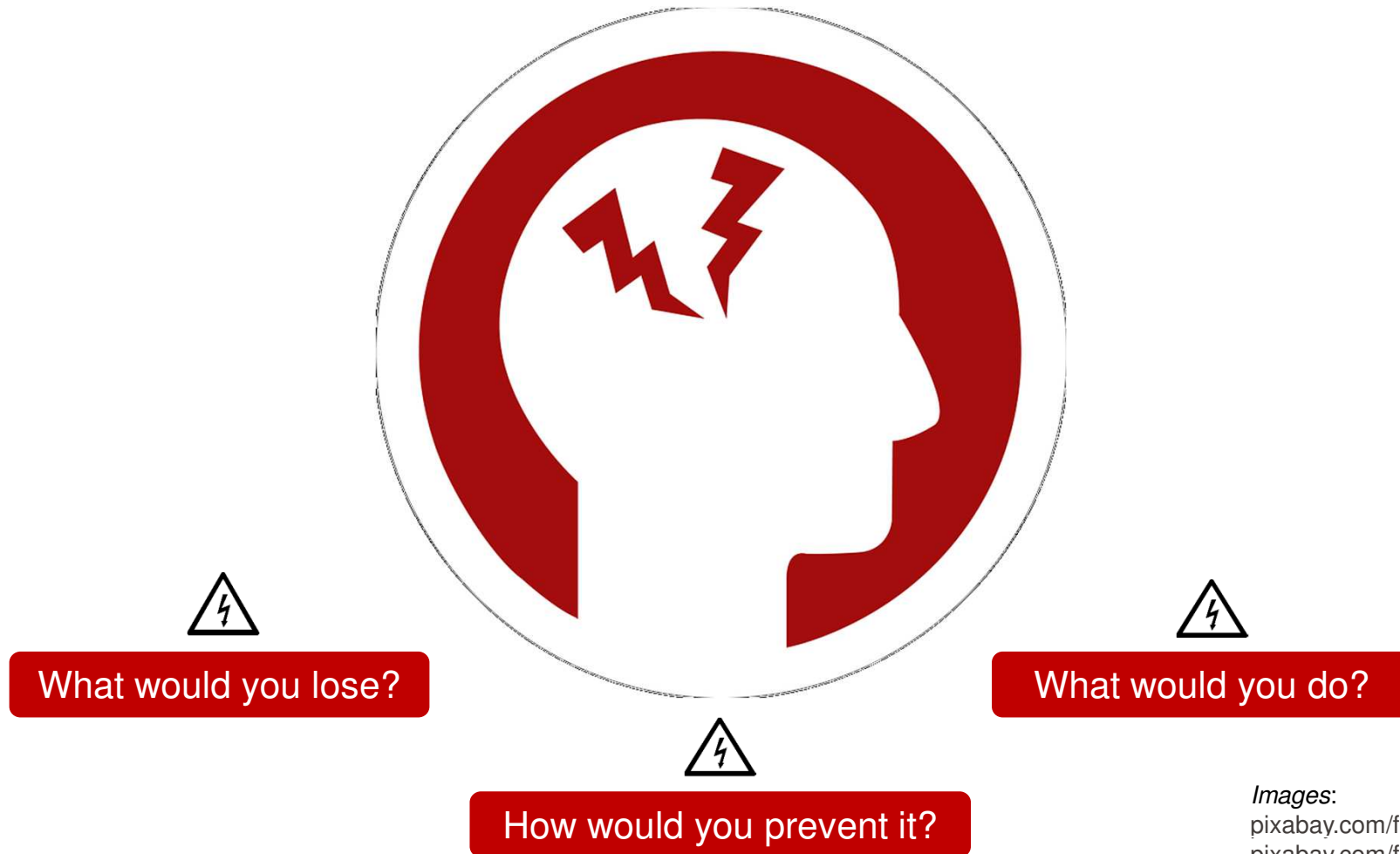
A **public, open standard** controlled by the scholarly community. [...] MEI and MusicXML share some similarities [but] they are guided by two different philosophies.

Humdrum
(logo?)

An encoding syntax to represent sequential data, especially music notation. [...] also refers to a series of software tools. [...] The Humdrum syntax served as an **inspiration** for other music encoding formats such as MusicXML.

NOTE: using open formats does **not** imply that data in that format is public

Discussion: storage IS NOT back-up



Images:
pixabay.com/fr/images/search/headache
pixabay.com/fr/images/search/hazard

Collaborative storage & File synching

SWITCHfilesender



- 50 GB max
- Email a link
- Lasts 20 days

Switchdrive



The academic cloud storage

- 50 GB
- Synchronization

SpiderOak



- Commercial
- Zero-Knowledge security

MS OneDrive



- All file types
- Commercial
- **Not sensitive data!**

Google Drive



- Unlimited (?)
- All file types
- Commercial
- **Not sensitive data!**

Firefox Send



- 2.5 GB max
- Email a link
- Lasts 7 days
- Open source

OwnCloud



- Self-hosted
- Open source
- Secure (SSH)

Not EPFL servers

Synchronization:

Curated comparison list on [Wikipedia](#)



BackInTime



Druva inSync



FreeFileSync



Rsync



Syncthing

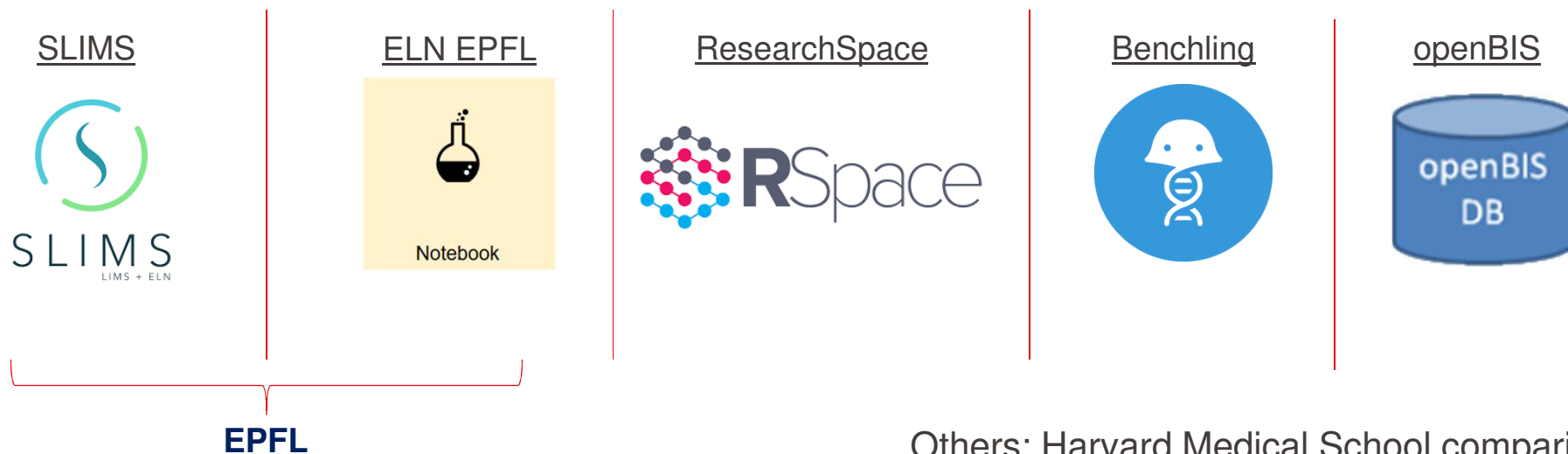


SyncToy

Storage options @EPFL

- **VPSI:** file storage services and backup for individual workstations
- **Faculties IT:** personalized storage option (NAS) for your faculty
- **SCITAS:** Work storage and c4science storage for coders

ELN/LIMS (ELECTRONIC LAB NOTEBOOK / LAB INFORMATION MANAGEMENT SYSTEM)

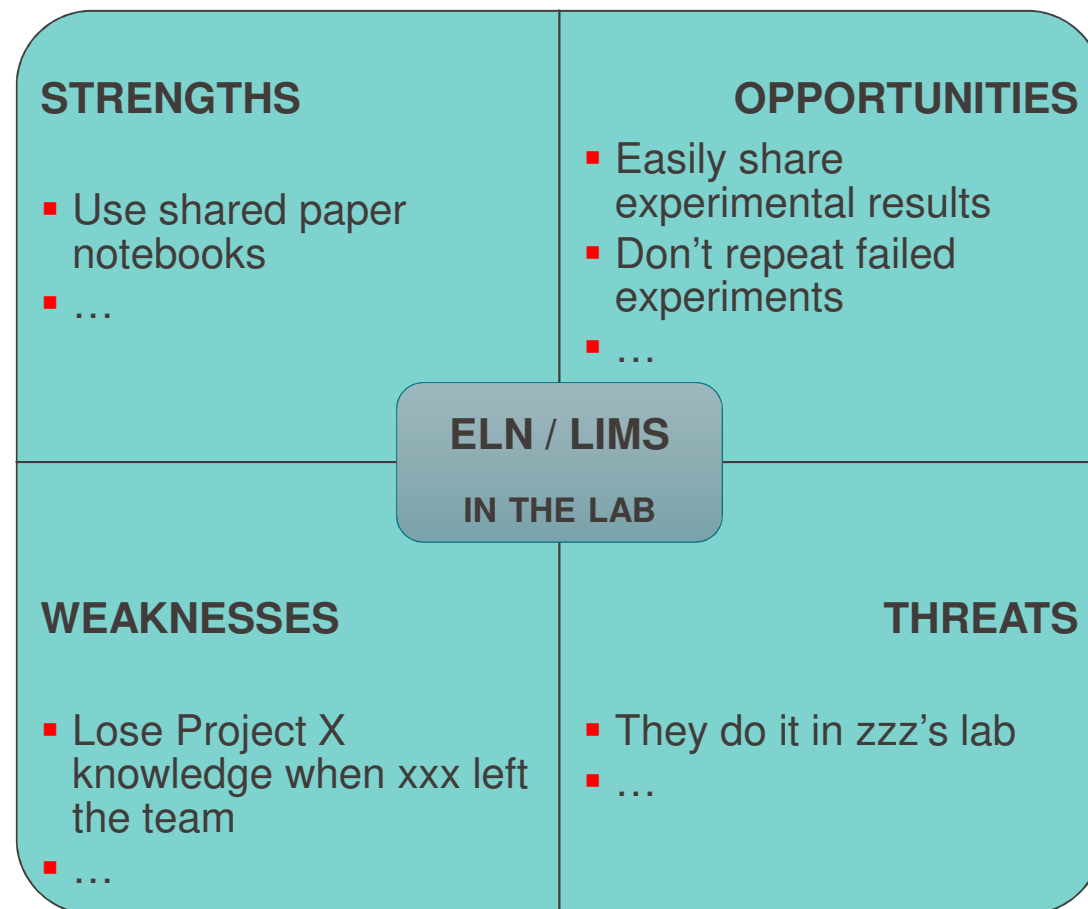


Others: Harvard Medical School [comparison table](#)

Electronic Lab Notebook (ELN) Lab Information Management system (LIMS)

Bundle **together** digital data + notes + logs + code + scripts + stock management + ...

- Why would I need it?
- Is it a tool for me?
- Can I / my team carry on without?



Importance of lab notebooks (!)

1985: George P Smith reports an original phage display method: inserting genes in the DNA for a specific phage protein, allowing the phage to infect and reproduce in bacteria

1993: Frances H Arnold conducts the first directed evolution of enzymes

2016: The €1m Millennium Technology Prize is awarded to Prof Arnold for her pioneering work on "directed evolution"

2018: Nobel: Prof Arnold shares the award with George P Smith and Gregory Winter for their research on enzymes

2019: Cho, Jia & Arnold publish a *Science* report on how to apply the appropriate evolutionary pressure

2020: Retraction: efforts to reproduce the report's work and consequent **examination of the first author's lab notebook** revealed missing entries and raw data for key experiments.

Sources: www.bbc.com/news/world-us-canada-50989423,
www.chemistryworld.com/opinion/frances-arnolds-retraction-and-the-case-for-slow-science/4010994.article

Nobel Prize-winning scientist Frances Arnold retracts paper

3 January 2020

f d t e Share

Nobel Prize



It is painful to admit, but important to do so. I apologize to all. I was a bit busy when this was submitted, and did not do my job well. <https://t.co/gJDU0pzlN8>

– Frances Arnold (@francesarnold)
 January 2, 2020

Files or Database?

File management

- File / folder organization
- File / folder naming
- File / folder versioning system
- File / folder access rights management

Database management

- Data model / Data dictionary
- Metadata design / standards
- Administrative data / logs
- User rights management
- Database administrator

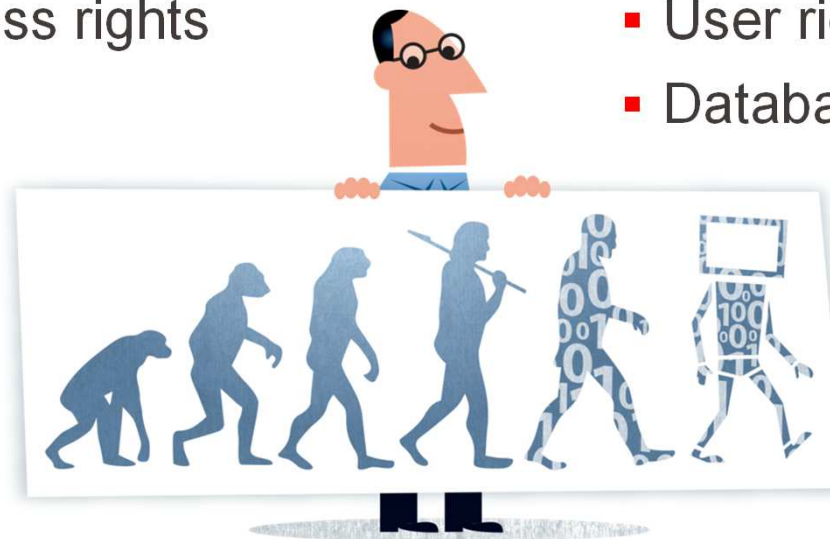


Image source: Digitalbevaring.dk (CC BY 2.5 DK)


EPFL-VPSI storage options

	COLLABORATIVE	ONLINE ARCHIVE	RAW
PERFORMANCE	Very good	Good	Good
REPLICATION		✓	✗
SNAPSHOTS			✗
PROTOCOLS	NFS, SMB/CIFS, WebDAV (optional)	NFS, SMB/CIFS, WebDAV (optional)	NFS, SMB/CIFS
BACKUPS	Optional	Optional	✗
PRICE	CHF 165 /TB /year	CHF 110 /TB /year	CHF 55 /TB /year


Storage ≠ Backup

Additionally, you can chose among **5 tiers of disk speeds**. These prices do not account for HR costs.
More info at the EPFL [File Storage page](#) or contacting [Fabian Figueras](#)

Storage Cost Calculator



EPFL Library
Cost Calculator for Data Management



Welcome to our cost calculator this tool will help researcher to have an estimate of the cost of managing, storing and publishing data.

Many providers are included in the service and you will be able to calculate a cost based on your needs.
Total cost is calculated dynamically based on your inputs.

We hope you will enjoy this tool and it will be useful for you.

[To Know More \(HOWTO\)](#)
[I need help with my DMP](#)

>

>

>

Project Name

My project name

Project Duration

Project Duration : 1 year

Change Currency

CHF

Line controls	Category	Provider information	Cost
<div>1</div> <div>Active Storage</div>	Select a Provider	0 CHF	
<div>1</div> <div>Electronic LabBook</div>	Select a Provider	0 CHF	
<div>1</div> <div>Database</div>	Select a Provider	0 CHF	
<div>1</div> <div>Data Repository</div>	Select a Provider	0 CHF	

rdmepfl.github.io/costcalc



Source: QR Code generator library of the [Project Nayuki](#).

Ex.: Kick off your DMP

SNSF DMP Template

1. Data collection and documentation

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

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

10': Draft your planning (par. 1.1 only)

DMP OPIDOR (prototype)

[Home](#)[Public DMPs](#)[DMP Templates](#)[Help](#)[More ▾](#) [Language ▾](#)

DMP Templates

DMP templates provided by a funder or research organisations, available on DMP OPIDoR. You can download these templates and related guidances, create a plan from these templates.

Template Name ▲	Organisation Name ◆	Organisation Type ◆	Description	Last Updated ◆	Download	Actions
EPFL SNSF	EPFL - Ecole Polytechnique Fédérale de Lausanne	Institution	This template was co-written by EPFL Library and ETH Library in the scope of the DLCM project. The current document is the EPFL version 5.0., revised in July 2019 by the EPFL Library Research Data team. ETH version is available from their own website . For further help, personal feedback or comments, you can contact us at researchdata@epfl.ch .	26-09-2019	 	
EPFL SNSF with examples hosted on web site	EPFL - Ecole Polytechnique Fédérale de Lausanne	Institution	This template was co-written by EPFL Library and ETH Library in the scope of the DLCM project. The current document is the EPFL version 5.0., revised in July 2019 by the EPFL Library Research Data team. ETH version is available from their own website . For further help, personal feedback or comments, you can contact us at researchdata@epfl.ch . Example answers for this template are hosted on a dedicated EPFL web page rather than embedded in DMP OPIDoR.	26-09-2019	 	

https://dmp.opidor.fr/public_templates

Sensitive & Commercial Data

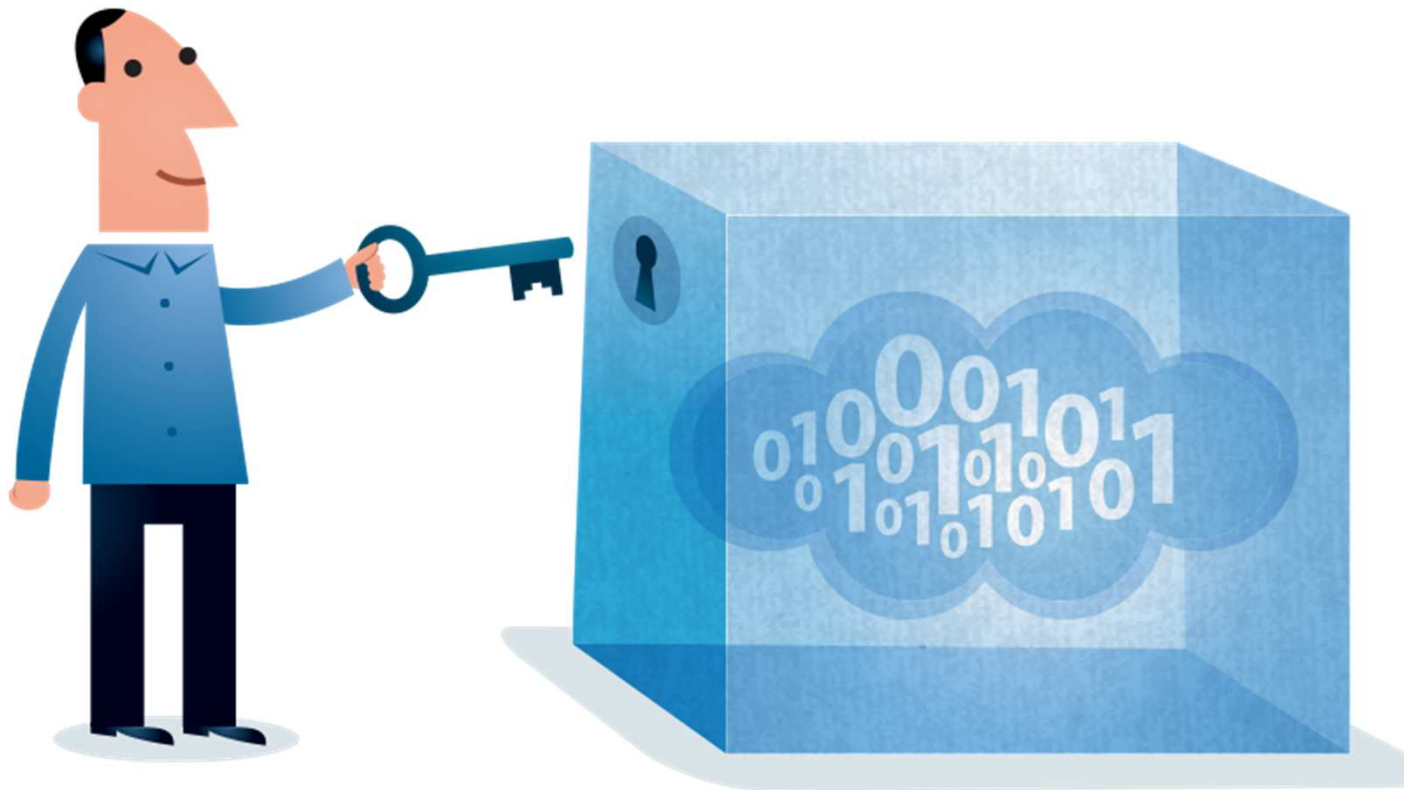


Image source: Digitalbevaring.dk ([CC BY 2.5 DK](#))

Constraints to open data publication

SENSITIVE

- Tests on animals / humans
 - Handle personal data
 - Federal Act on Data Protection (FADP), Human Research Act (HRA), GDPR
 - name, identification number, location data, online identifier, ...
 - factors specific to physical, physiological, genetic, mental, economic, cultural or social identity
- *check the EPFL Human Research Ethics Committee (AREC + HREC forms)*

COMMERCIAL

- **Data from 3rd party** sources? (e.g. commercial datasets, research cooperations, etc.)
- *check out the **contract** for data usage / sharing ... Or make one!*

- Want to potentially submit a **patent**?

→ *check the TTO (Technology Transfer Office) ... Chose the **data license** + tell in the DMP!*

Personal / Sensitive data processing



THE SWISS FEDERAL ACT ON
DATA PROTECTION
[FADP]



THE EU GENERAL DATA
PROTECTION REGULATION
[GDPR]

Any operation with personal data [...] in particular

- the collection
- storage
- use
- revision
- disclosure
- archiving
- or destruction of data

Swiss Federal Act on Data Protection (FADP)
(*Loi sur la Protection des Données* LPD), Art. 3e

Protection: Personal data must be protected **against unauthorised processing** through adequate technical and organisational measures

FADP Art. 7

Disclosure: Making personal data **accessible**, for example:

- by permitting access
- transmission
- or publication

FADP Art. 3f

Collecting consent (online & offline)

*“Research involving human beings may only be carried out if [...] the persons concerned have given their **informed consent** or, after being **duly informed**, have not exercised their **right to dissent**. [...] The persons concerned may withhold or **revoke their consent at any time**, without stating their reasons.”*

HRA, Art. 7

The consent must be

- **Simple**
- **Understandable**
- **Adapted** to the subject (child, teenager...)

HRA Art. 21-22

The screenshot shows the homepage of the swissethics website. The header includes the logo and navigation links: HOME, ABOUT US, ETHICS COMMITTEES, TEMPLATES/RECOMMENDATIONS, EDUCATION TRAINING, LEGISLATION/GUIDELINES, and LINKS. There are also links for RSS and Newsletter, and language options for DE, FR, and IT.

Recent additions / changes

Date	Document
November 11, 2018	New document (v5.3, 06.11.18): Template for writing information for participants for studies according to HRA/ClinO, in German. Versions in French and Italian available in the corresponding language sections of swissethics.ch: .pdf , .docx
November 11, 2018	New document (v2.3, 06.11.18): template for HRO according to HRA/HRO chapter 2 (not: ClinO or HRO chapter 3 "further use"). In German, versions in French and Italian available in the according language sections of swissethics.ch: .pdf , .docx
October 10, 2018	New document (v1.0, 12.09.18): Study protocol template for clinical trials Chapter 4: Other Clinical Trials: .pdf , .docx
October 9, 2018	New document (Swissmedic ref. nr. BW510_00_006e_F0 / V1.0 / sci / bbe / 31.05.2018): Template for the notification of serious adverse events (SAE) and device deficiencies to the ethics committees for clinical trials (ClinO) with medical devices: .pdf
October 2, 2018	New document (v1.0, 26.09.18): Guideline 'application of the General Data Protection Regulation (GDPR)' and template for drafting additional information (addendum) for study participants in line with the General Data Protection Regulation (in German): .docx , .pdf

Please note:

Please send suggestions for the improvement of our templates to Mr [Dr. P. Gervasoni](#)

Please subscribe to our [newsletter](#) or [RSS-Feed](#) to get the latest news.

Quick-Links

[Web portal BASEC](#) for the submission of research projects

Templates

Please always download and use the latest version of the templates.

The templates meet all the aspects required by the Swiss legislation. It is therefore highly recommended to stick to the templates to meet the requirements in full. The use of the templates also facilitates and accelerates the assessment by the Swiss Ethics Committees. The templates shall be applied according to the type of study:

- Information about and template for the synopsis of the study protocol: [.pdf](#), [.docx](#)
- Templates for study protocols
- Patient information and Declaration of consent / ethical evaluation / GPPB

Discussion: What about your research?

- Do you collect, process or store data which is... **sensitive? ... personal?**
- Do you collect, process or store information on... **identifiable persons? ... vulnerable persons? ... children?**
- How do you inform persons/subjects on what you will be doing?

Discussion (5')

Data masking techniques

Pseudonymization

(working data, reversible)



- **PSEUDONYMIZATION**

Replace data by identifiers. The key is kept separately & securely

- **ENCRYPTION**

Encrypt the data & keep the key secure. Also for long-term preservation, not data publishing

Some tools:

- R package: [sdcMicro](#)
- Java application: [ARX Data Anonymization Tool](#)
- Java application: [ARGUS](#)
- Platform: [Amensia](#)

Anonymization

(published data, irreversible)



- **GENERALIZATION**

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

- **SUPPRESSION**

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

- **ADD FAKE DATA**

To prevent the identification of specific records, add fake data while preserving correlations

- **SHUFFLE**

Shuffle data over one / several columns without compromising the utility of the data

(Other: [Differential Privacy](#), [T-closeness](#), ...)

Images:

- <https://www.flaticon.com/packs/general>
- <https://www.flaticon.com/packs/hawcons-documents-filled>

Deletion of identifying data

name	gender	city	age	disease
KELLER Anna	f	Basel	32	no diabetes
BRUNNER Emilia	f	Basel	37	diabetes 2
DURANT Pierre	f	Basel	44	no diabetes
GRAF Julia	f	Basel	45	diabetes 2
GERBER Fritz	m	Basel	20	diabetes 1
FISCHER Urs	m	Basel	23	diabetes 1
WYSS Emilien	m	Geneva	24	no diabetes
STEINER Leo	m	Geneva	28	no diabetes
ROTH Christian	m	Geneva	42	no diabetes
WYSS Rudolf	m	Geneva	48	diabetes 2

	name	gender	city	age	disease
0	*	f	Basel	30 - 39	no diabetes
1	*	f	Basel	30 - 39	diabetes 2
2	*	f	Basel	40 - 49	no diabetes
3	*	f	Basel	40 - 49	diabetes 2
4	*	m	Basel	20 - 29	diabetes 1
5	*	m	Basel	20 - 29	diabetes 1
6	*	m	Geneva	20 - 29	no diabetes
7	*	m	Geneva	20 - 29	no diabetes
8	*	m	Geneva	40 - 49	no diabetes
9	*	m	Geneva	40 - 49	diabetes 2

K-anonymity 2

Deletion of identifying data

name	gender	city	age	disease
KELLER Anna	f	Basel	32	no diabetes
BRUNNER Emilia	f	Basel	37	diabetes 2
DURANT Pierre	f	Basel	44	no diabetes
GRAF Julia	f	Basel	45	diabetes 2
GERBER Fritz	m	Basel	20	diabetes 1
FISCHER Urs	m	Basel	23	diabetes 1
WYSS Emilien	m	Geneva	24	no diabetes
STEINER Leo	m	Geneva	28	no diabetes
ROTH Christian	m	Geneva	42	no diabetes
WYSS Rudolf	m	Geneva	48	diabetes 2

	name	gender	city	age	disease
0	*	f	*	30 - 39	no diabetes
1	*	f	*	30 - 39	diabetes 2
2	*	f	*	40 - 49	no diabetes
3	*	f	*	40 - 49	diabetes 2
4	*	m	*	20 - 29	diabetes 1
5	*	m	*	20 - 29	diabetes 1
6	*	m	*	20 - 29	no diabetes
7	*	m	*	20 - 29	no diabetes
8	*	m	*	40 - 49	no diabetes
9	*	m	*	40 - 49	diabetes 2

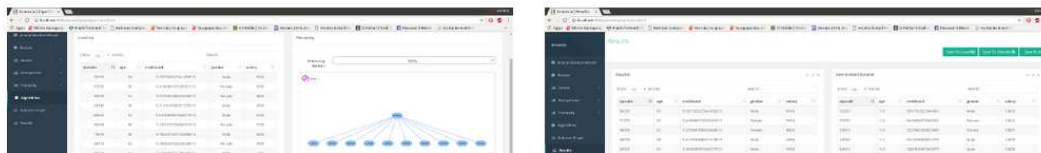
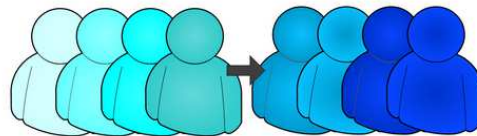
L-diversity 2

Try it out later!

Amnesia

Amnesia is a data anonymization tool, that allows to remove identifying information from data. Amnesia not only removes direct identifiers like names, SSNs etc but also transforms secondary identifiers like birth date and zip code so that individuals cannot be identified in the data. Amnesia supports k -anonymity and k^m -anonymity.

version: 1.0.7 (release date: 21/02/2019)



- Implements data anonymization techniques from the field of **Privacy Preserving Data Publishing (PPDP)**
- Transforms original data to anonymized data by using **generalization** and **suppression**
- Anonymization **not limited to the removal of direct identifiers**; it **also includes removing secondary information** (e.g. like age, zipcode, etc.) that might indirectly lead to identify an individual
- Focuses on **k -anonymity**: guarantees that every record will be indistinguishable from other $k-1$ records
- Supports **2 algorithms for k -anonymity**, Incognito and a parallel version of the Flash algorithm.

Data publication and preservation

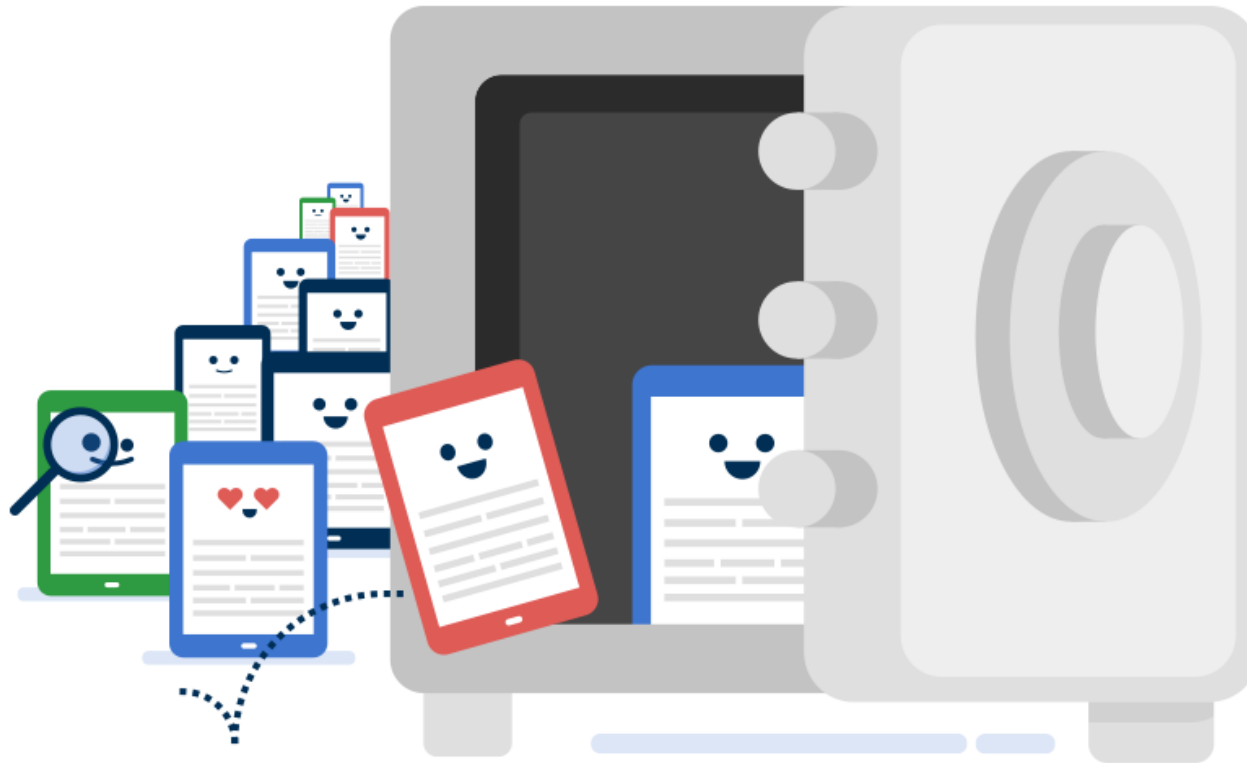


Image source: Digitalbevaring.dk ([CC BY 2.5 DK](https://creativecommons.org/licenses/by/2.5/dk/))

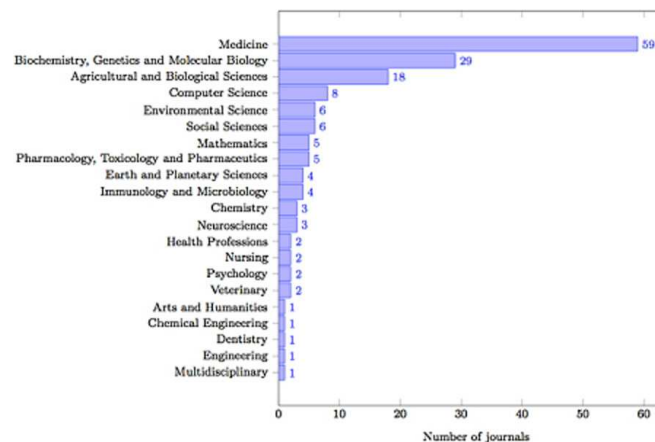
Data publication

Data Papers

A data paper is a peer reviewed document describing a dataset, published in a peer reviewed journal. It takes effort to prepare, curate and describe data (GBIF, 2019)

Data Journals

Data papers are supported by many journals, some of which are "pure", i.e. they are dedicated to publish data papers only, while others – the majority – are "mixed", i.e. they publish a number of articles types including data papers. (Wikipedia, 01.04.2019)



Open Science

F1000Research
Open for Science

Open Science Framework
A scholarly commons to connect the entire research cycle



The Open Science Training Handbook

Data degradation problem

CERN

A 2007 study showed that a bitrot error ratio of 10^{-7} (over 2 months)
Ex.: $\sim 10^9 \cdot 10^{-7} = 10^2 = 100$ bytes of bitrot every 1GB (1024MB)

US FDA

In 2017 the agency added data integrity requirements for the drugs industry
 (FDA 21 CFR, 11 & 211)

Data integrity failure (Possible causes)

- Processing
CPU heat, encryption errors, ...
- Transfer
Network failures, backup errors, ...
- Read / Write
Single bits errors at RAM or ROM levels
- Storage
Aging, background radiation, ...

Countermeasures (Data repositories)

- Redundant hardware
- Uninterruptible power supply
- Certain types of RAID arrays
- Radiation hardened chips
- Error-correcting memory
- Clustered file system
- File systems with block level checksums

Data access sustainability

2014 studies showed in that:

- More than 60% of links to astronomy datasets are **broken after 10 years**
- The bibliography of **1 out of every 5 is impacted** by this phenomenon



SNSF

Researchers must “*share their data according to the FAIR Data Principles on **publicly accessible, digital repositories.***”

ERC

Open Research Data Pilot participants must “*deposit research data [...], including associated metadata, **in the repository as soon as possible.***”



Human Genome Project

A “good example of a large-scale research endeavour in which an openly accessible data repository is being used successfully” [OECD]

Back-up vs. Preservation

	STORAGE & BACKUP	LONG-TERM PRESERVATION
ACTIVE DATA	✓	
DATA RECOVERY	✓	
INTEGRITY (monitoring, repair, authenticity)	?	✓
APPRAISAL (what & for how long)		✓
PERMANENT IDENTIFIERS		✓
DESCRIPTION (metadata)	✗	✓
RENDERABILITY (format migration, virtualization)	✗	✓

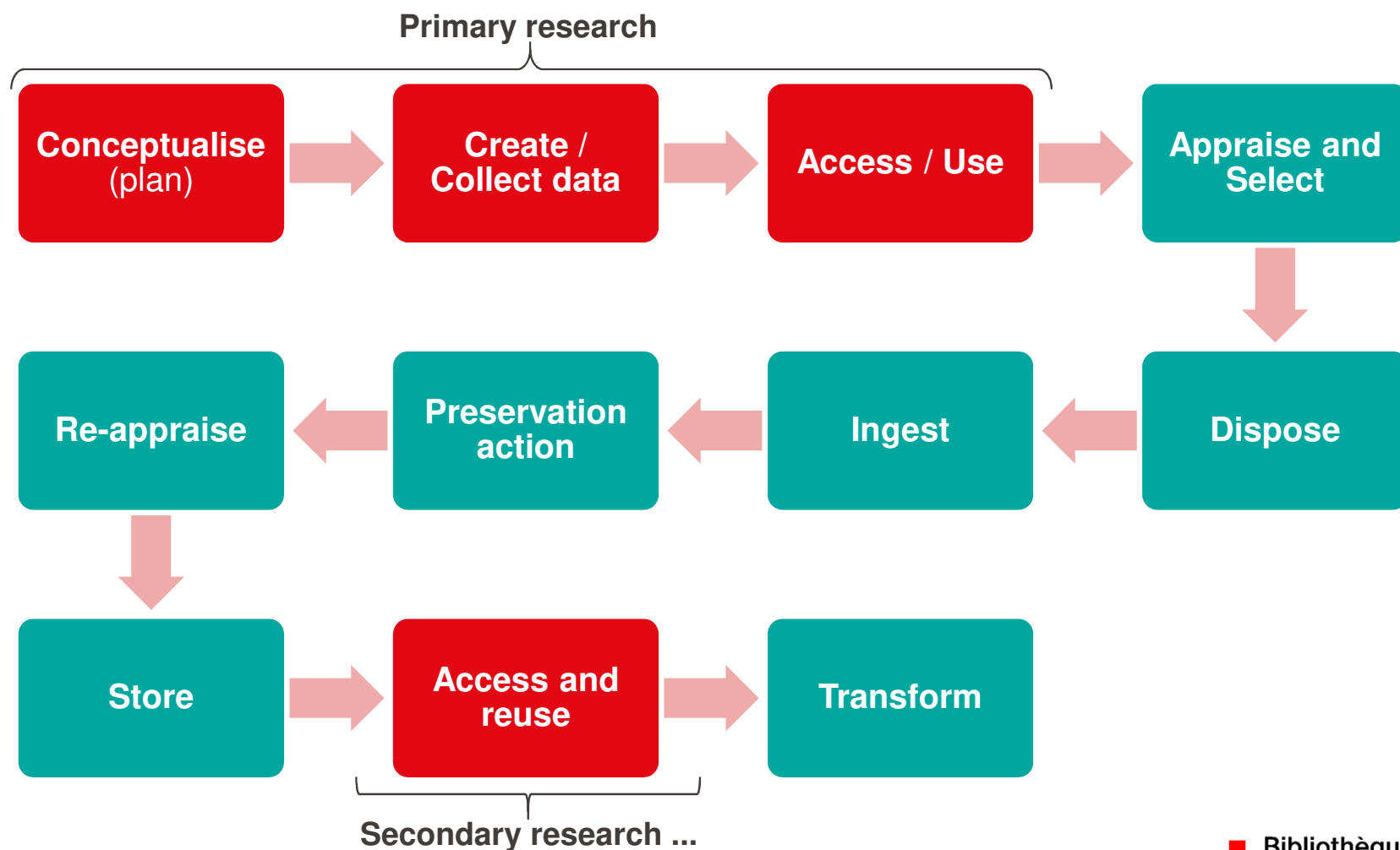
Backup ≠ Preservation

Long-term preservation: Data curation

Maintain, preserve
and **add value** to
research data
throughout its
lifecycle

Digital Curation Centre

If **you** don't do
it, find a data
curator who will



Data repositories: Publication (+ Preservation)

80

Sithida Samath, Francesco Varrato

General
purpose

vs.

Discipline
specific

Institutional

Departmental
project

Personal
page















Publisher

Supplementary
Material

List of Nature's Recommended Data Repositories per discipline

(Non-exhaustive) list of repositories approved by some publishers for hosting data alongside the articles

Data repositories

NAME	DISCIPLINE	NOPROF / INST.	COUNTRY	FREE	MAX VOLUME	LICENSING
	Generic	✓ (CERN)		✓	50GB/dataset, ∞ datasets	CC, GNU, BSD
 MATERIALSCLOUD	STI / Materials	✓ (EPFL)		✓	5GB General, 50GB AiiDa DB	CC-BY (MIT for AiiDa)
	Generic	✗ (Holtzbrinck Group)		Freemium	1 TB per dataset	CC0, CC-BY
 DRYAD	Bio / Medical	✓ (?)		✗	?	CC0
 The Dataverse Project	Generic	✓ (Harvard University)		✓	?	?
 EUDAT	Generic	✓ (HORIZON 2020)		✓	?	CC (DARUP)
 ERIC open	Aquatic	✓ (Eawag)		✓	Unlimited upload size	CC0 default / Changeable

The SNSF encourages the use of re3data.org. Also check the data repositories [recommended by the ERC Scientific Council](#)

Data repository (Example)

zenodo Search Upload Communities Log in Sign up

EPFL - Ecole Polytechnique Federale de Lausanne

Recent uploads

Search EPFL - Ecole Polytechnique Federale de Lausanne

June 13, 2018 (v1) Dataset Open Access View

SPARTAN example data

Christian Sieben; Niccolò Banterle; Kyle M. Douglass; Pierre Gönczy; Sulianna Manley;

Example data to test the features of SPARTAN, a general-purpose particle analysis software for single-molecule localization microscopy (SMLM). The software is integrated into a newly developed framework to perform multi-color 3D reconstruction from 2D SMLM data. SPARTAN is available on [GitHub](#)

Uploaded on June 13, 2018

June 7, 2018 (v1) Dataset Open Access View

Double inverse nanotapers for efficient light coupling to integrated photonic devices - Open Data

Liu, Junqiu;

Available data for the manuscript "Double inverse nanotapers for efficient light coupling to integrated photonic devices"

Uploaded on June 7, 2018

April 28, 2018 (v1) Dataset Closed Access View

A reference airborne LiDAR dataset for forest research

Matthew Parkan; Pascal Junod; René Lugrin; Christian Ginzler;

This repository contains the dataset presented in Parkan et al. (2018). Abstract: The benefits of Airborne Laser Scanning (ALS) to efficiently monitor and manage forests are widely accepted. Products derived from ALS have been successfully used in a range of different domains including ecosystem c

Uploaded on May 25, 2018

New upload

Want your upload to appear in this community?

- Click the button above to upload straight to this community.
- The community curator is notified, and will either accept or reject your upload (see community curation policy above).
- If your upload is rejected by the curator, it will still be available on Zenodo, just not in this community.

EPFL - Ecole Polytechnique Federale de Lausanne

Curated by: infoscience

Curation policy: Datasets submitted by EPFL members.

Created: October 17, 2013

Harvesting API: [OAI-PMH Interface](#)

Zenodo.org has an EPFL Community!

- Hosted by the CERN
 - Free of charges
 - Max 50GB/dataset
 - Unlimited datasets
 - Automated DOI assignement
-
- OpenAIRE integration (EC reporting)
 - GitHub integration
 - ORCID integration
-
- All file formats accepted
 - Usage statistics interface
 - OAI-PMH protocol (content harvesting)
 - 18 petabytes disk cluster
 - Each file has 2 replicas on different servers
 - 2 independent MD5 checksums per file
 - Metadata 12-hourly backup cycle
 - ...

Demo (Zenodo sandbox)

The screenshot shows the Zenodo sandbox interface. At the top is a blue header with the Zenodo logo, a search bar, and links for 'Upload' and 'Communities'. Below the header, there's a 'Recent uploads' section with a list of four items, each with a 'View' button. The items are:

- Kritische Soziale Arbeit und ihr Gegenstand: Eine kritische Auseinandersetzung** by Stalder, Bruno; Vifian, Karin. Uploaded on August 30, 2016.
- Structure Assisted Compressed Sensing Reconstruction of Undersampled AFM Images Dataset** by Oxvig, Christian Schou; Arildsen, Thomas; Larsen, Torben. Uploaded on August 25, 2016.
- Atomic Force Microscopy Images of Cell Specimens** by Christian Rankl. Uploaded on August 18, 2016.
- Measurement and Analysis of Strains Developed on Tie-rods of a Steering System** by Asenov, Stefan. Uploaded on July 12, 2016.

On the right side of the uploads, there are three informational boxes:

- Zenodo now supports usage statistics!** with a link to read more.
- Using GitHub?** with a link to log in and start preserving repositories.
- Zenodo in a nutshell** with a list of features:
 - Research. Shared.** – all research outputs from across all fields of research are welcome! Sciences and Humanities, really!
 - Citeable. Discoverable.** – uploads get a Digital Object Identifier (DOI) to make them easily and uniquely citeable.
 - Communities** – create and curate your own community for a workshop, project, department, journal, into which you can accept or reject uploads. Your own complete digital repository!
 - Funding** – identify grants, integrated in reporting lines for research funded by the European Commission via OpenAIRE.
 - Flexible licensing** – because not everything is under Creative Commons.
 - Safe** – your research output is stored safely for the future in the same cloud infrastructure as CERN's own LHC research data.

At the bottom right, there's a 'Tweets by @ZENODO_ORG' section showing a tweet from @ZENODO_ORG.

sandbox.zenodo.org



Source: QR Code generator library of the [Project Nayuki](#).

H2020 portal

ec.europa.eu/info/funding-tenders/opportunities/portal/screen/home

The screenshot shows the 'Grant Management' interface for project 725675 (UNEARTH) under the ERC-COG program. The top navigation bar includes 'Grant Management' and 'Project Continuous Report'. The main content area displays a progress bar with various milestones: Summary for publication, Deliverables Ethics, DMP, Other Reports, Publications, Dissemination, Patents (IPR), Open Data, Gender, and ABS Regulation. The 'Open Data' section is highlighted, showing that the project does not currently have any Open Datasets. It also displays suggested Open Datasets by OpenAIRE and Project Open Datasets, both of which are currently empty.

Enabling data re-use



Image source: Digitalbevaring.dk ([CC BY 2.5 DK](https://creativecommons.org/licenses/by/2.5/dk/))

Importance of sharing (!)

1976 – Experiments on supercooled water (cooled far below its freezing point) showed a **critical point** at -20°C : its structure fluctuates widely between high- and low-density forms

2011 – Seeking a unified theory of water, simulations on supercooled water by two world-leading groups revealed:

- Chandler et al.: **no critical point** (resembles ordinary water)
- Debenedetti et al.: **critical point** (morphs between two forms)

2014 – Debenedetti&al. **published their code** openly

2016 – At first, Chandler&al. only shared data, then revealed where to find its code and, after lot of reverse engineering ...

2018 – ... the trouble stemmed from an algorithmic trick the Chandler's team used to speed up their code!

[DOI: 10.1063/PT.6.1.20180822a](https://doi.org/10.1063/PT.6.1.20180822a)

PHYSICS TODAY

HOME BROWSE▼ INFO▼ RESOURCES▼ JOBS

DOI:10.1063/PT.6.1.20180822a

22 Aug 2018 in *Research & Technology*

The war over supercooled water

How a hidden coding error fueled a seven-year dispute between two of condensed matter's top theorists.

Ashley G. Smart

14
COMMENTS

< PREV NEXT >



Most people would've seen little reason to quibble with David Chandler's talk at the spring 2011 Statistical Mechanics Conference. Chandler, a chemist at the University of California,

Data / Code licences

COMMONS LICENSES CODE, DATA, TEXT & MULTIMEDIA

The 96/9/EC Directive protects only vs. “substantial” copies of datasets.

Creative Commons:

- Enforced by the author
- Check platform’s policy
- On datasets (no data points)

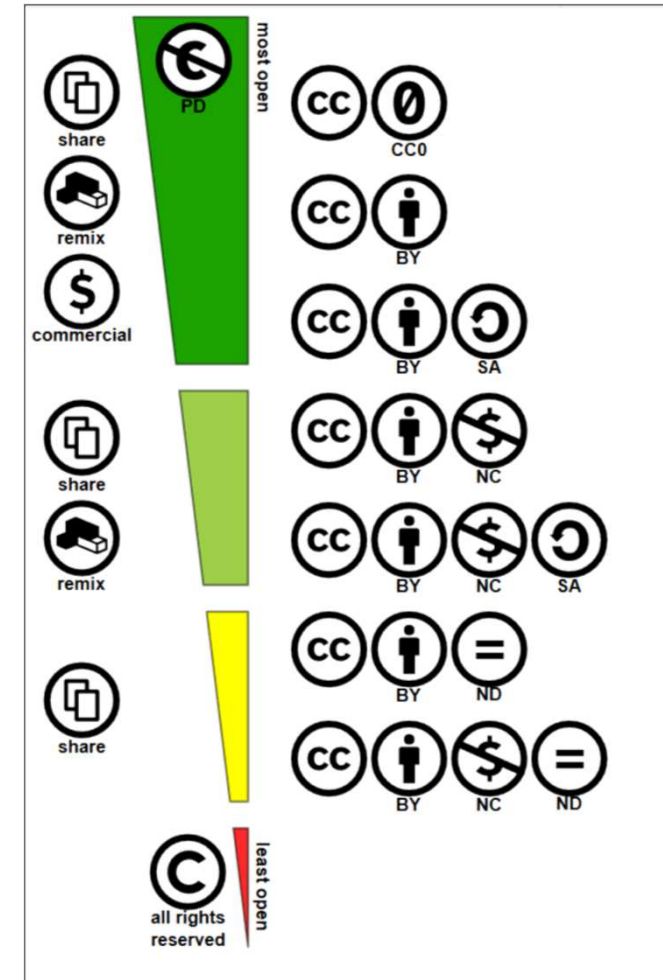
(The Open Data Commons can be a viable option)

SPECIFIC FOR CODE

- GNU-GPL (Open Software)
- Apache2.0 (smaller codes, libraries)
 - Permissive
 - No share alike clause
 - Preservation of copyright notice
- BSD-3clause - similar



Apache



Source: commons.wikimedia.org

Importance of licensing (!)

2012 – Project of officially **launched**:
Venice's State Archive + Ca' Foscari Univ. + EPFL (DHLAB)

2014 – Non-binding agreement signed. But ... **didn't specify the licensing** that would regulate researchers' use of the digitized data

2017 – At stake: 1,000 years of records in dynamic digital form: special high-speed scanners, thousands HD images per hour

2019 – **Allegedly**, the digitization of ~190,000 documents (**8 TB**) didn't follow a common metadata policy: archival-science guidelines (require records of provenance for each document)

Now – ... **data collection has been paused, amid doubts on the usability of the data already collected!**

DOI: [10.1038/d41586-019-03240-w](https://doi.org/10.1038/d41586-019-03240-w)

88

MENU nature Subscribe


NEWS • 25 OCTOBER 2019

Venice 'time machine' project suspended amid data row

Disagreements among international partners leave plans to digitize the Italian city's history in limbo.

Davide Castelvecchi

Twitter Facebook Email



[PDF version](#)

RELATED ARTICLES

The 'time machine' reconstructing ancient Venice's social networks

Saving Venice

SUBJECTS

Databases History

Historians want to use archive documents to create a virtual time machine for Venice, pictured here in the 18th century. Credit: DEA/Getty

Like the city itself, an ambitious effort to digitize ten centuries' worth of documents that record the history of Venice is at risk of sinking. Two key partners have suspended the Venice Time Machine project after reaching an impasse over issues surrounding open data and methodology. The State Archive of Venice and the Swiss Federal Institute of Technology in Lausanne (EPFL) say they have had to pause data collection, and the archive's director has raised questions about the usability of the 8

Potential commercial use

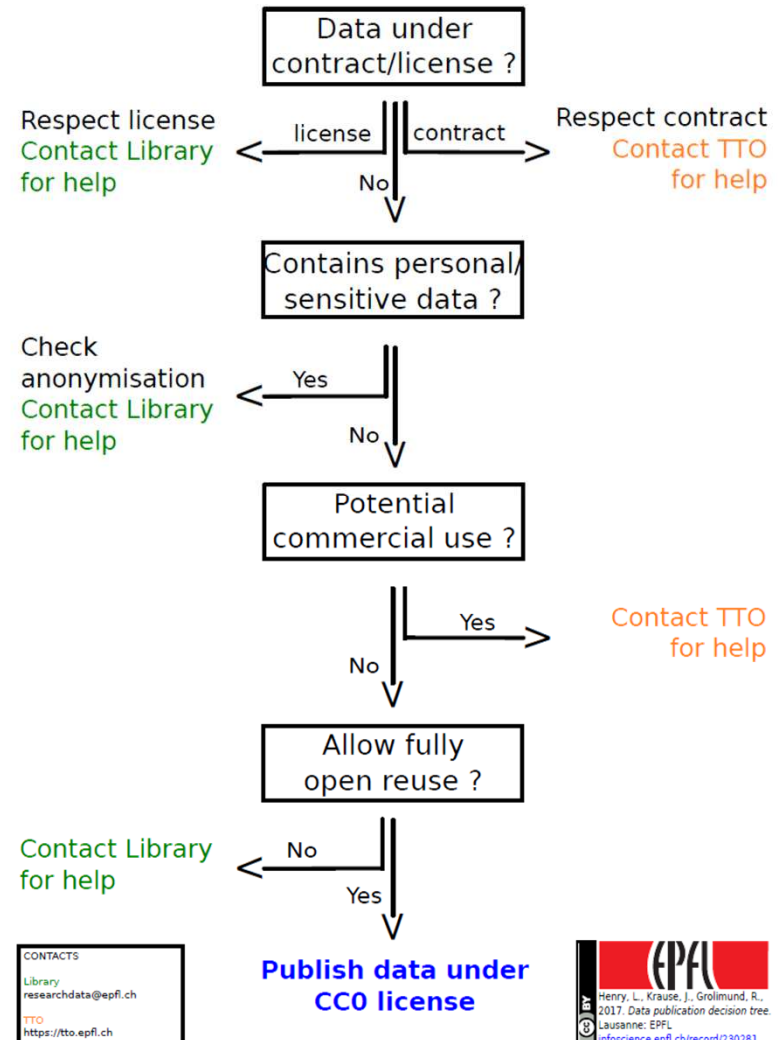
Define the reuse of your data!

Rule of thumb:
“**no CC0 or CC-BY? then TTO**”

TTO = Technology Transfer Office

Source: infoscience.epfl.ch/record/230281

DATA PUBLICATION DECISION TREE



How to cite data(sets)?

Same as any other citation:

- **Author(s)** of the dataset
- **Title** of the dataset / study
- **Year** of online publication
- **Publisher** responsible for distributing the dataset
- **Edition / Version** number associated with the dataset
- **Persistent identifier(s)** as URI, DOI, ORCID, ...

The logo for Data Citation Principles, featuring the letters 'DC' in a large, blue, serif font, with a superscript '1' to the right.

Source: www.force11.org/datacitation *Data Citation Principles*

Find & Reuse

- General lack of a good description makes it **difficult to find** pertinent datasets
- The data reuse requires an even more **precise description**
- The (near) future of data is **linked**

Vertical search engine to find a relevant repository



re3data.org

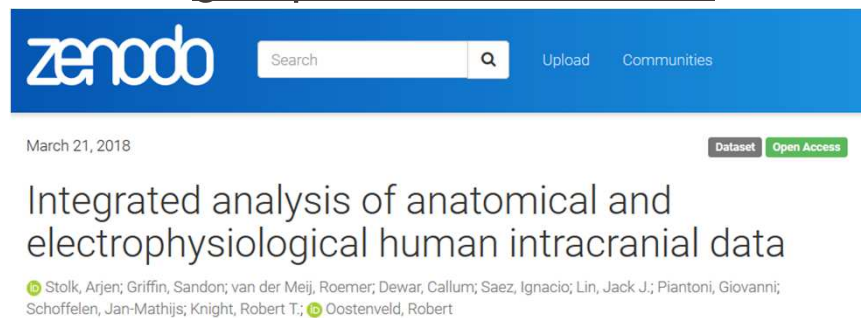


Source: QR Code generator library of the Project Nayuki.

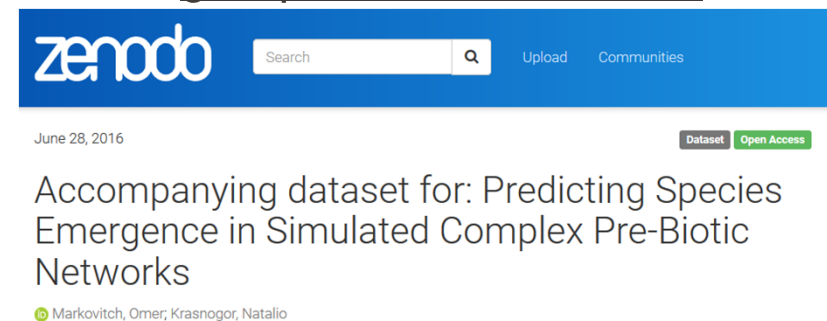
Discussion: Publication (bad dataset)

1. Look at (one of) these datasets ...

go.epfl.ch/badDataset1



go.epfl.ch/badDataset2



2. Answer

- Are those datasets reusable?
- What would you change?

Almost there!

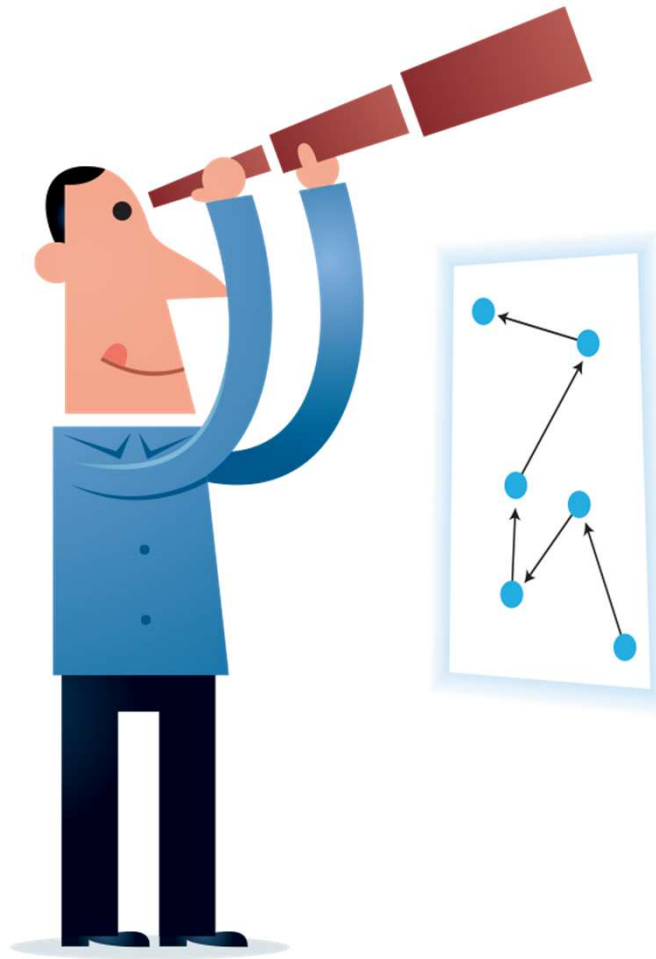


Image source: Digitalbevaring.dk ([CC BY 2.5 DK](#))

Recap: FAIRness self-assessment

1. DISCUSS: how FAIR are **your lab's practices**?

- Is there a data manager?
- Do your directories contain a README file?
- Are data analysis protocols shared?
- Do you produce / convert data in open formats?
- ...

2. TEST: how FAIR are **your datasets**?

DOES	TO-DO

10': Write down what to-do to improve your FAIR score

go.epfl.ch/FAIR

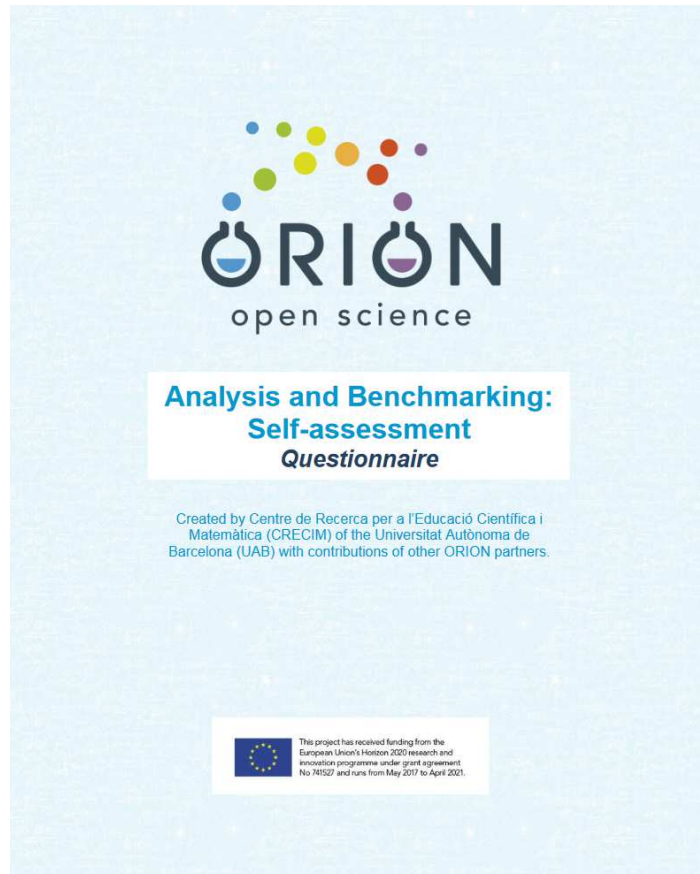


Source: [Australian Research Data Common](#)

Recap: Data actions in your project

ACTIVITIES	COLLEAGUE / PARTNER	TOOLS	TO-DO
FUNDING PLANNING			
CREATION			
ACQUISITION			
ANALYSIS			
STORING			
SHARING			
ARCHIVING			
PUBLISHING			
LEGAL CLEARANCE			
ETHICAL CLEARANCE			

Recap: Go even further? ... Open Science self-assessment



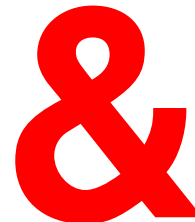
go.epfl.ch/OpenScienceAssessment



Submit us your data needs!

À-LA-CARTE

- Single acces point (researchdata@epfl.ch)
- RDM support (DMP review, follow-up)
- RDM expertise & Lab audit
- Courses & workshops
 - DMP basic/adv. workshops
 - RDM workshops (Fellows/Staff/Ethics)
 - ...
 - [On demand]
- Continuous feedback



BUFFET

- Website (go.epfl.ch/rdm)
- Documents
 - Templates (DMP / RDM strategy)
 - Walkthrough Guide / FastGuides
 - RDM Checklist
 - Funders guidelines
- Software (Storage cost calculator, ...)
- DMP (online) tool
- EPFL Open Science Initiative (collab.)

researchdata@epfl.ch

go.epfl.ch/rdm

go.epfl.ch/training



EPFL

Thank you!

Do you have any questions?

go.epfl.ch/rdm

researchdata@epfl.ch