18.06.2024

Research Data Management – The Basics

Cantini, Federico
Felder, Fabian
Iosifescu, Ionut
Nordén, Klara

These are your trainers today!

Federico Cantini

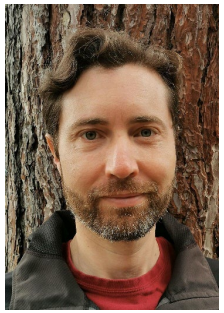
- Software Developer
- Technical Lead at Lib4RI

Fabian Felder

- Open Science specialist
- Group Leader IT services and E-resources at Lib4RI

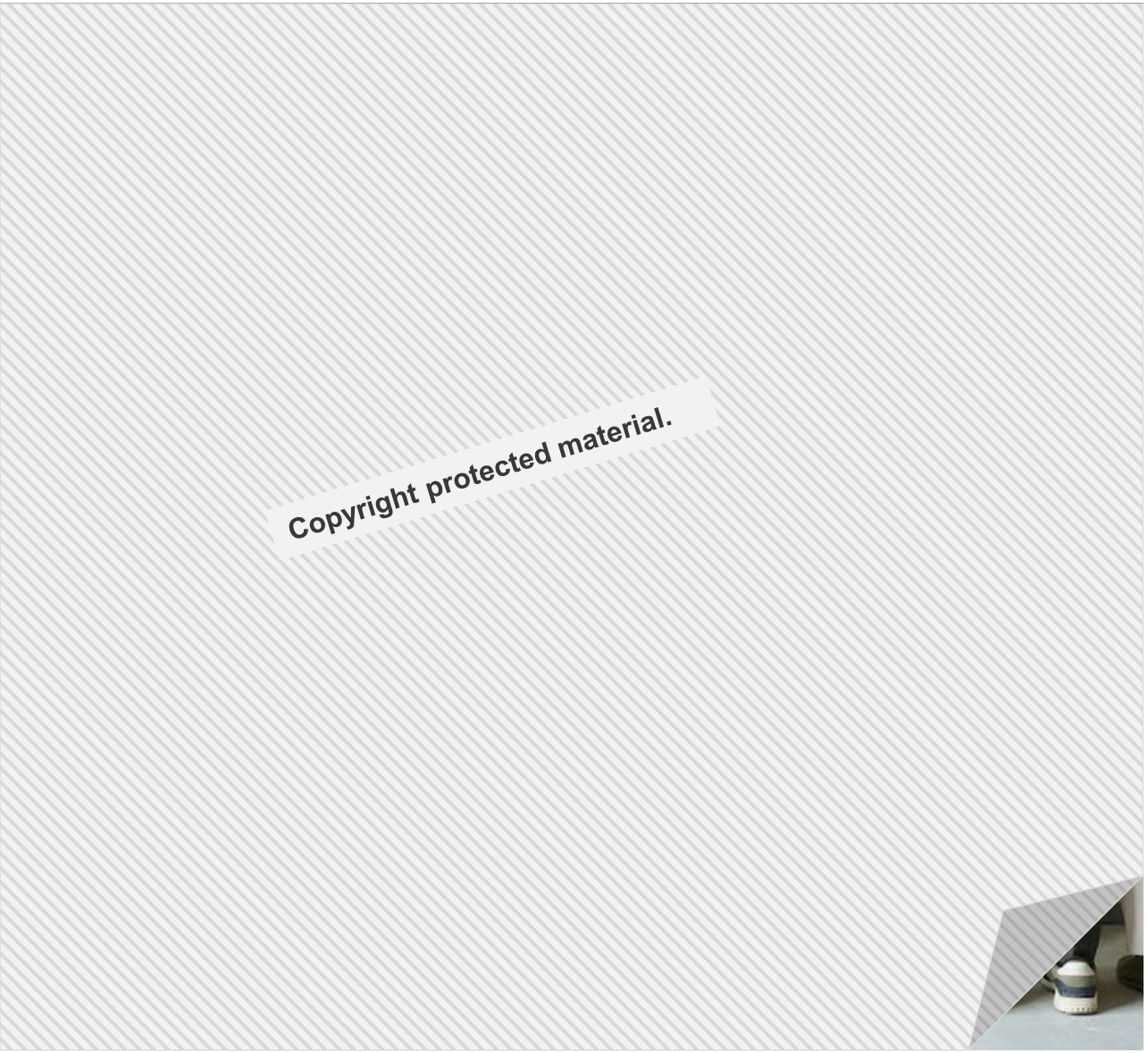
Klara Nordén

- Project manager Open Research Data and Research Data Management at Lib4RI



Ionut Iosifescu

- Software Engineer
- Technical coordinator and product owner of EnviDat



Who are you and why are you here?

https://www.pexels.com/photo/group-of-people-standing-indoors-3184396/

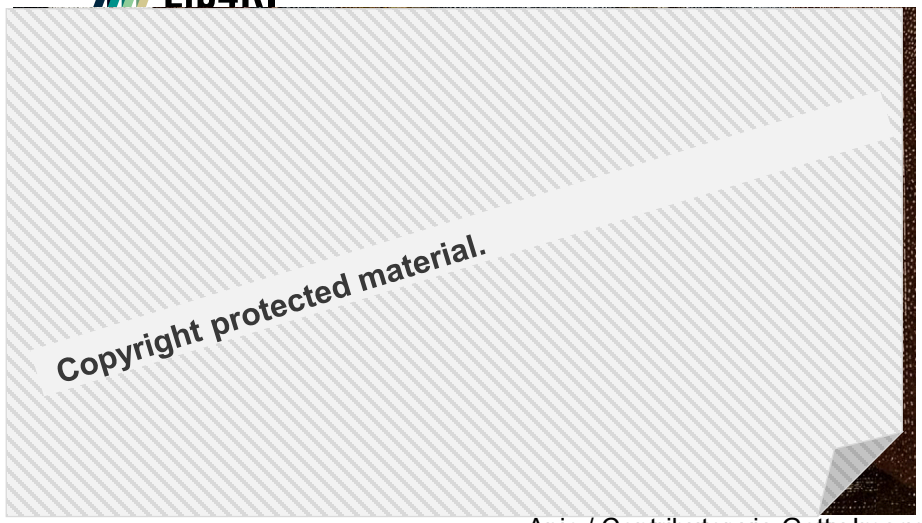
Learning Aims

- Life cycle of research data
- Adequate metadata documentation for your code and data
- Storing and publishing data
- Writing Data Management Plans (DMP)

Program

Topic	Speaker	Time
Introduction	Fabian Felder	9.00 - 9.15
Why Open Science?	Klara Nordén	9.15 - 9.25
Policies and the Research Data Life Cycle	Fabian Felder	9.25 – 9.30
Collect & Store	Federico Cantini	9.30 - 10.05
Evaluate & Archive Share & Disseminate	Fabian Felder	10.05 - 10.15
Break		10.15 - 10.30
RDM Services & Support at WSL	Ionut Iosifescu	10.30 - 11.00
Plan & Design	Everyone	11.00 - 11.45

Why Open Science?



Apic / Contributor via Getty Images

```
%% tail and head distribution
addpath('D:\KIT3');
clearvars; %close
myKsDir = uigetdir('Z:\locker\Fede\1Fish_mimic_obj\');
files2=dir([myKsDir, '\LODdata2*']);
files3=dir([myKsDir, '\obj_num*']);
files4=dir([myKsDir, '\video*.avi']);
load(['Z:\locker\Fede\1Fish_mimic_obj','\Fish_5_',myKsD

%% for obj on and mimics on
tic
Ang=nan(30,16); NRe=nan(30,16); NRtotal=nan(30,16); TAl
a=1; b=1;
for i=22:2:52
    AUX3=[]; AUX2=[]; AUX4=[]; AUX1=[];
    [AUX1, ~]=find(Obj_idx(:,1)==i);
    for j=1:size(AUX1,1)
        [AUX2]=find(Obj_idx(:,2)==Obj_idx(AUX1(j),2));
        M = readtable([myKsDir,'\CUT_' files4(Obj_idx(A

        AUX4=Obj_idx(AUX2,1);
```

Fast forward 320 years….

1704 2024

Lib4RI

“The study should be reproducible from the paper alone”

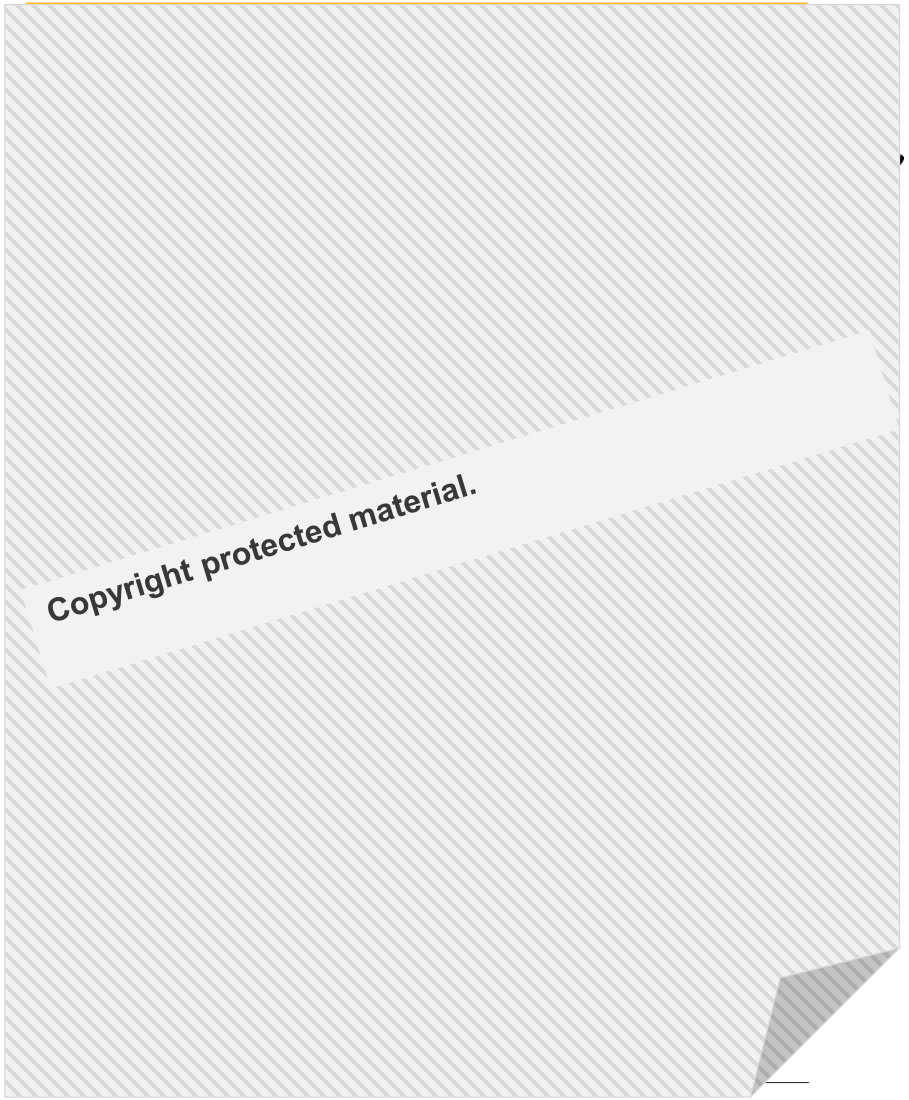
10-25 bugs per 1000 lines of code

(Applications Division at Microsoft, Code complete, Steve McConnell)

Spreadsheets show a typical error rate of 2-7%

(Panko 2005)

Why care about open science?



Data availability
The data that support the findings of this study are available from the corresponding author upon reasonable request.

Code availability
All numerical codes are available from the corresponding authors on request

Lib4RI

1 month later….

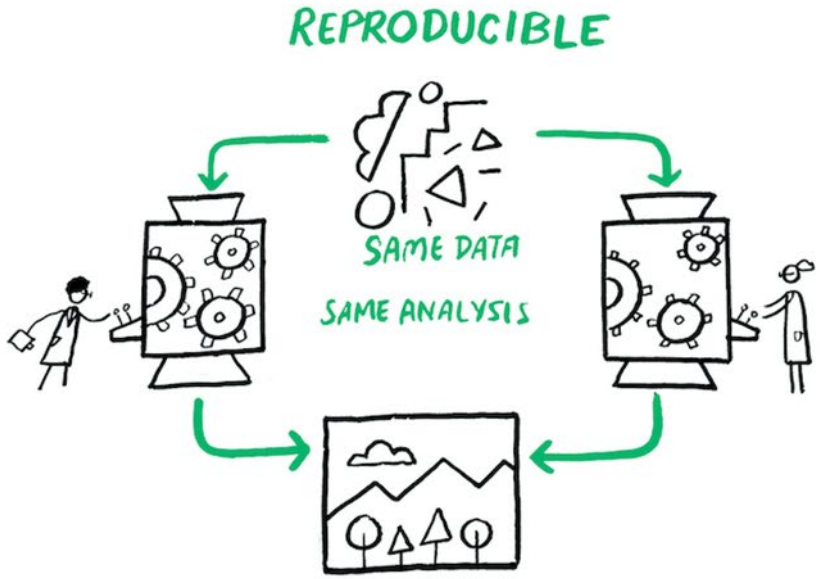
Illustration by Ainsley Seago, CC-by 4.0

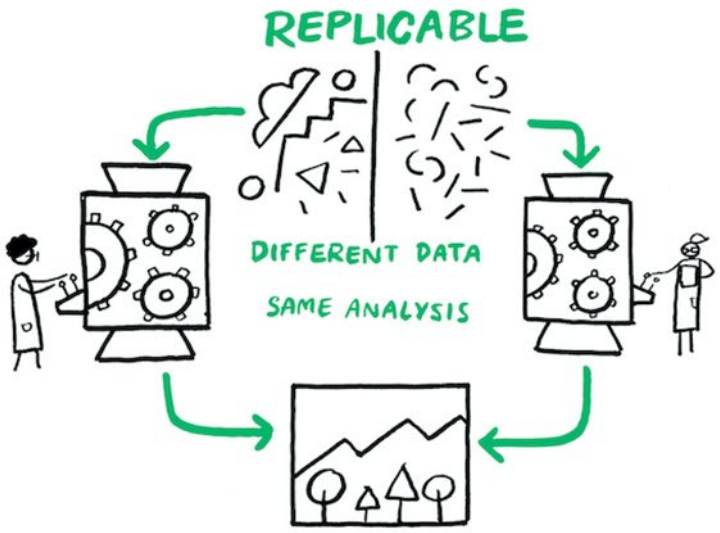
sphera™

2 months later….

Are your results different because you asked a different question
OR
because

- you used a different set-up in your experiment?
- you used a different software?
- you normalized your values differently?
- you've misunderstood the variables in the original data?
- the original study had errors?





"Non-replicable single occurrences are of no significance to science"

(Karl Popper, 1959). The Logic of Scientific Discovery

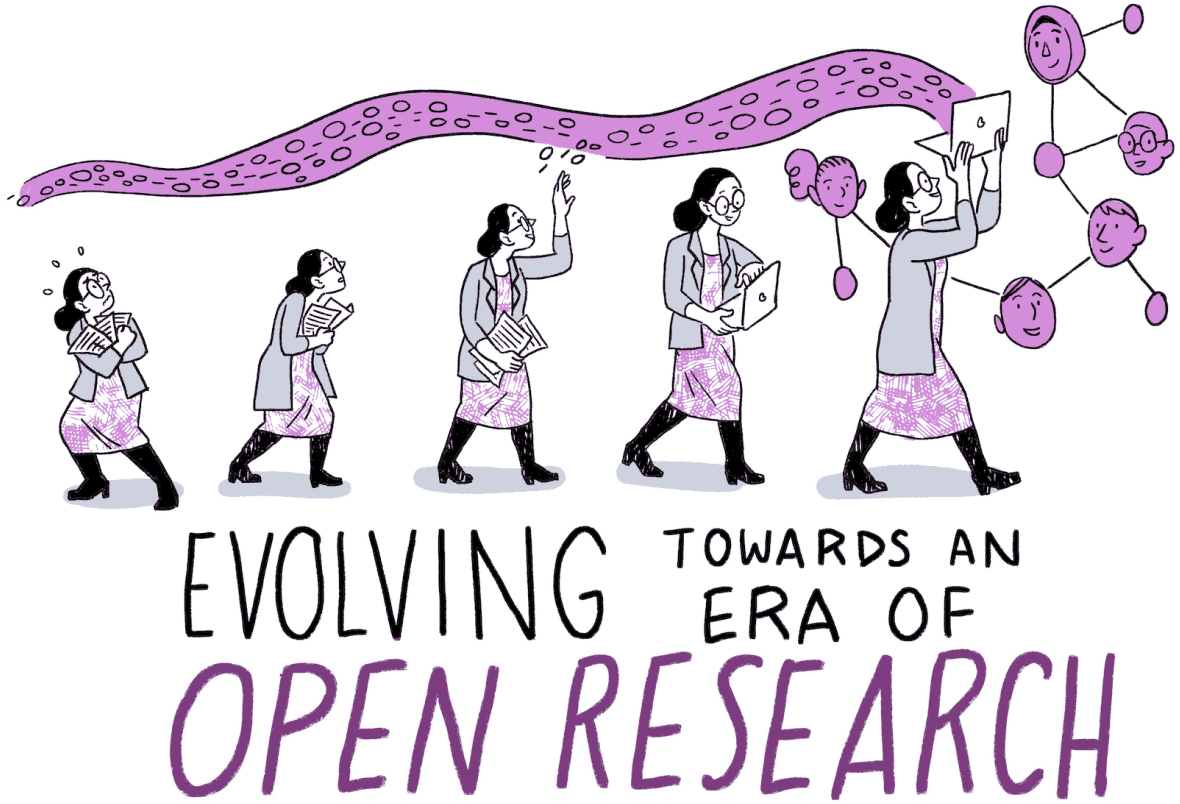
Lib4RI

Why care about open science?

*Open science is about making **your** contribution to the scientific project sustainable, lasting and impactful.*

Science is a communal project, and open science practices creates building blocks that makes it easy for others to build on your results.

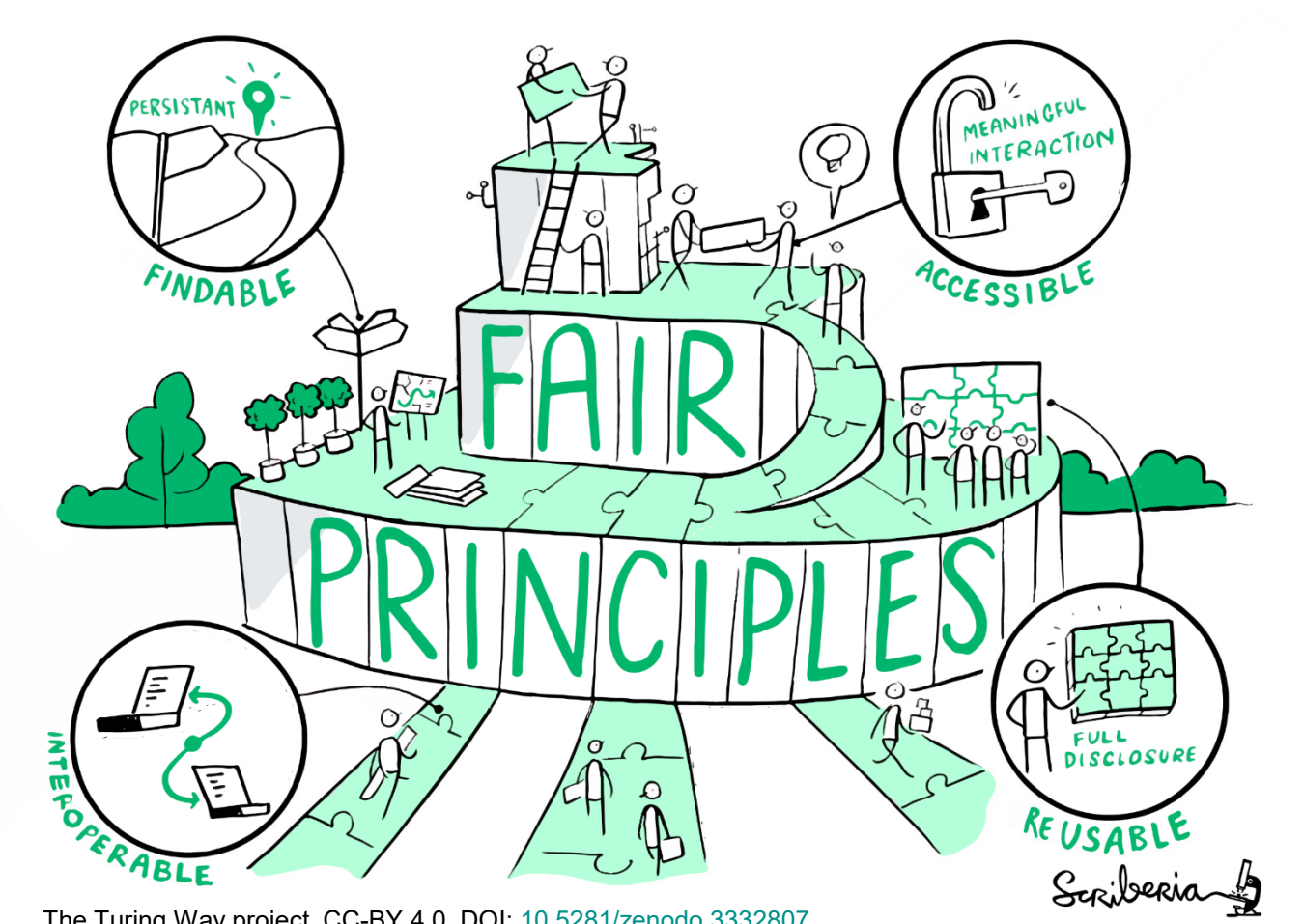
Your most likely future collaborator is… YOU



Scriberia

Open Access *Open Data* *Open Source Software* *Open Source Notebooks*

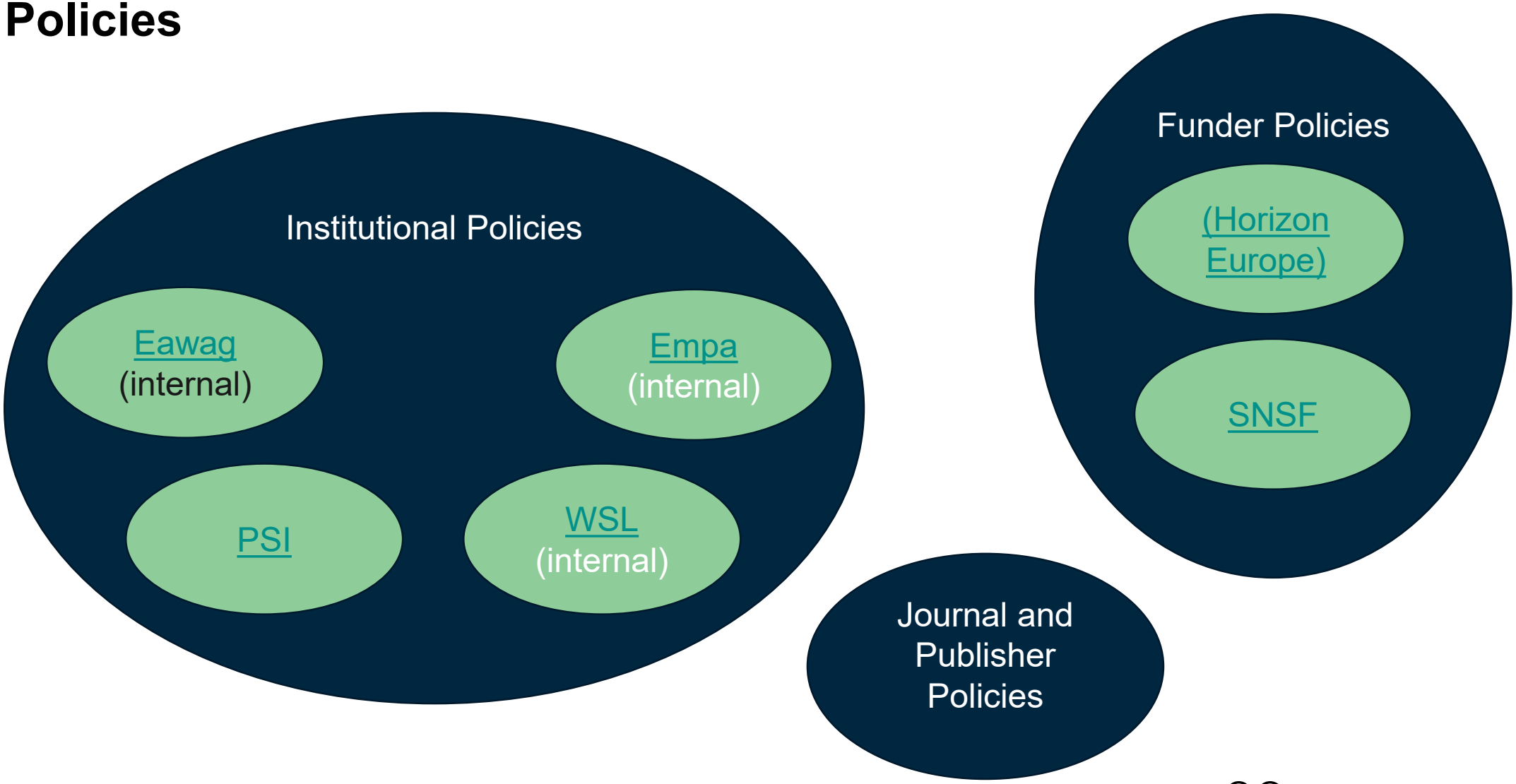
The FAIR data principles



The Turing Way project. CC-BY 4.0. DOI: 10.5281/zenodo.3332807.

Policies

Policies


Institutional Policies
Eawag (internal)
Empa (internal)
PSI
WSL (internal)
Funder Policies
(Horizon Europe)
SNSF
Journal and Publisher Policies

Policies

Compliance

Project Manager/
Group Leader

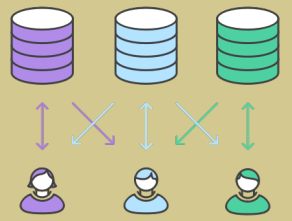
DMP

Required for
Funders

As open as
possible,

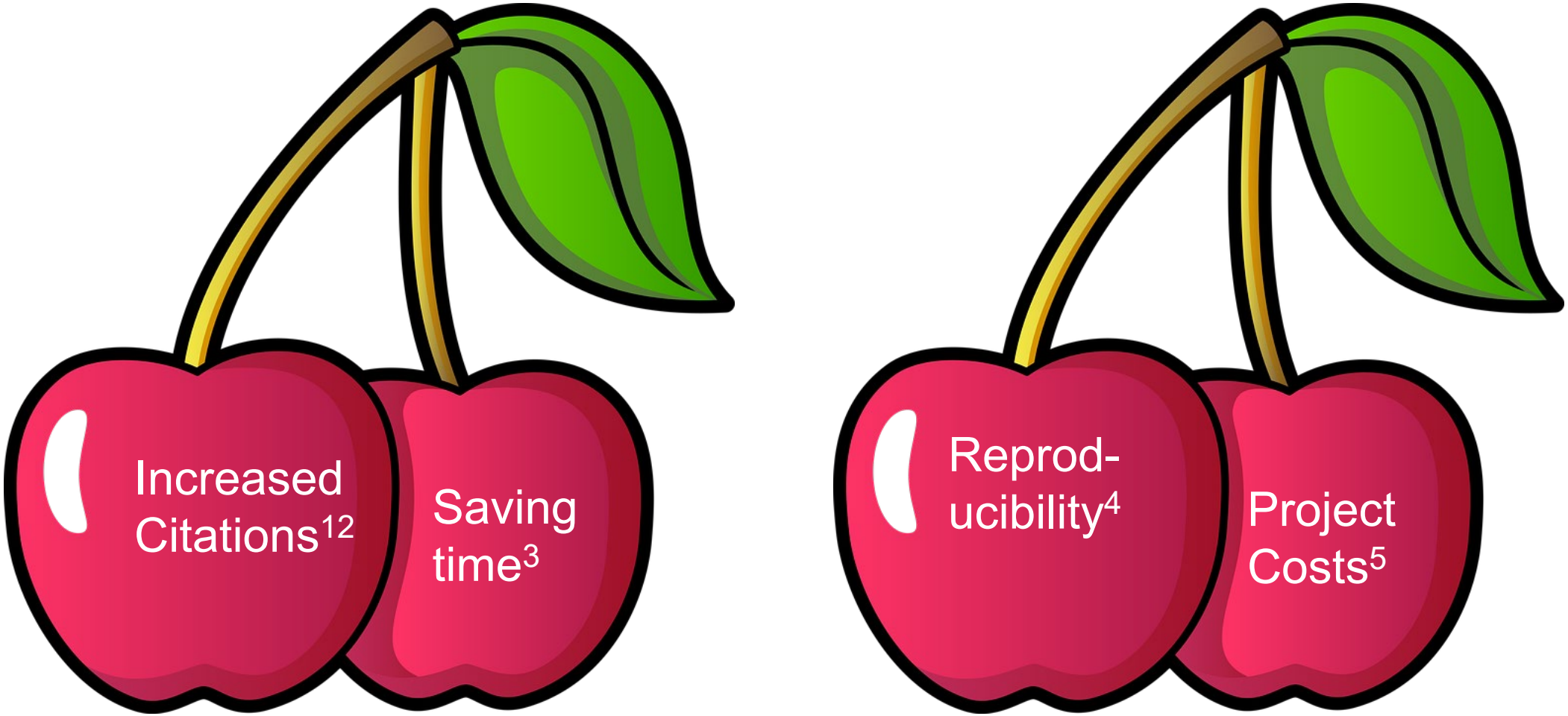
as closed as
necessary.

Repositories

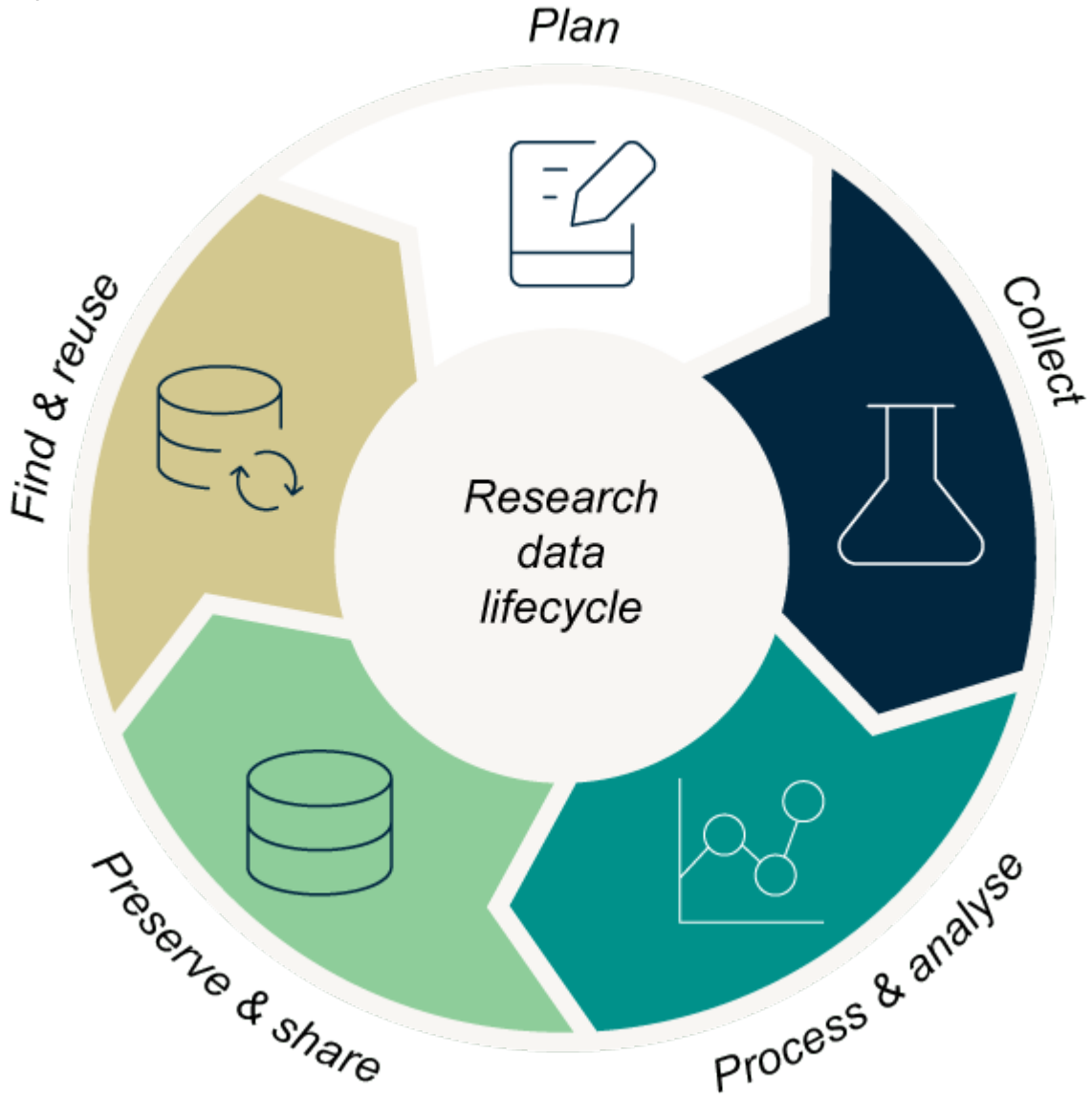


FAIR and non-
commercial

Policies


Increased Citations[12]
Saving time[3]
Reprod-ucibility[4]
Project Costs[5]

Research Data Life Cycle



Research Data Life Cycle

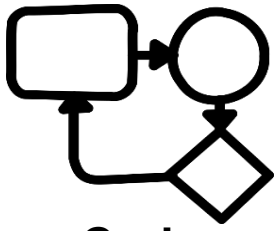
Collect & Store

Collect & Store

```
01010100 01101000
01101001 01101110
01101011 00100000
01100100 01101001
01100110 01100110
01100101 01110010
01100101 01101110
01110100 00101110
```

Data

observational, experimental, simulation…



Code

Applications, scripts…

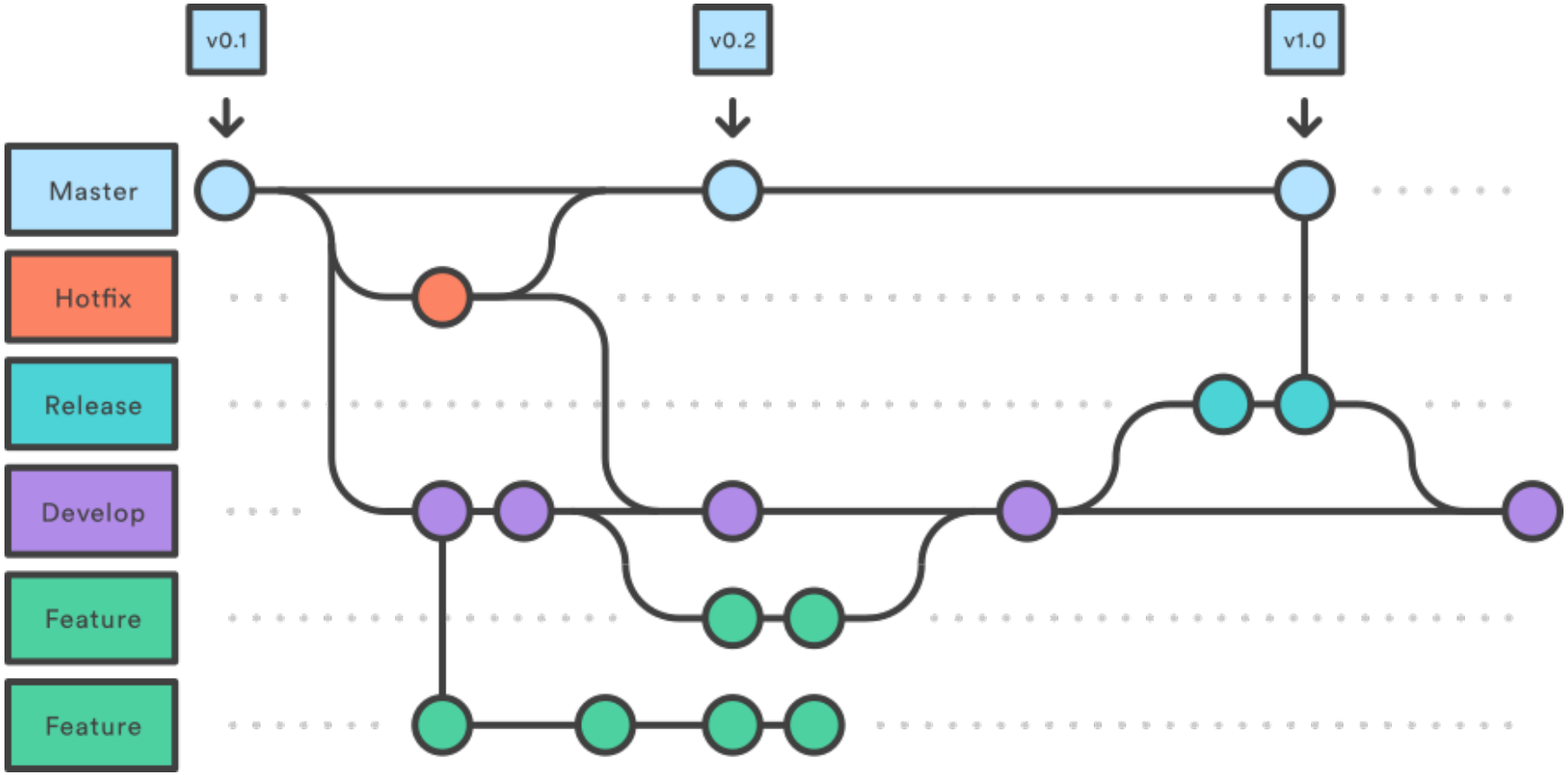
</>

Metadata

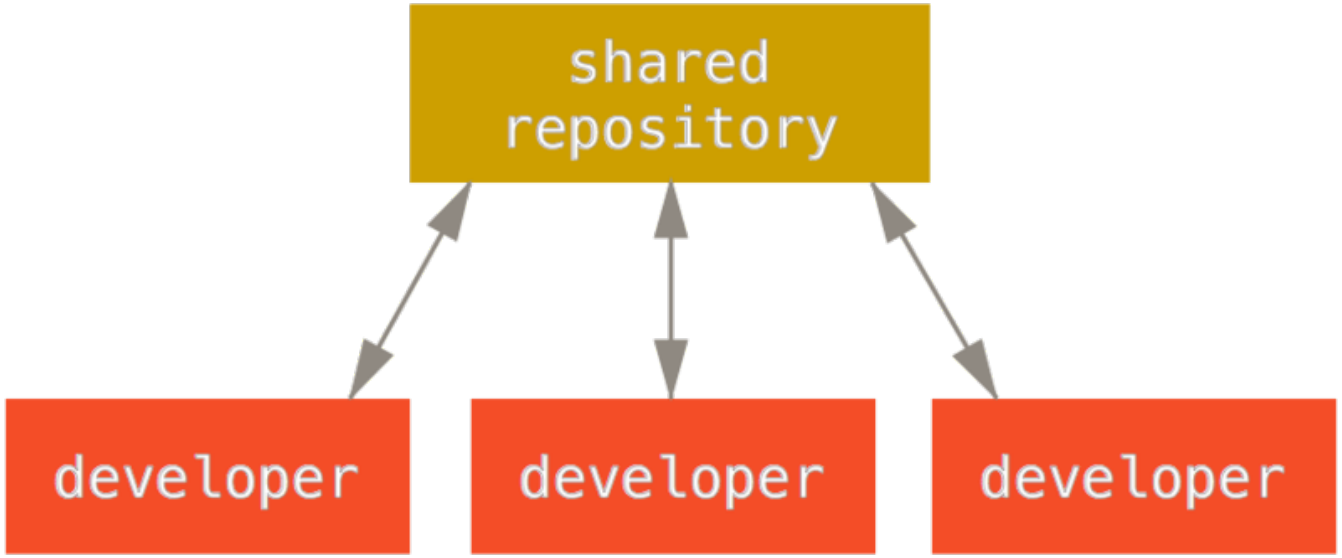
Structured information associated with data (and code)

The Who, What, Where, Why & How of data

Collect and Store: Software version control



Collect and Store: Software version control



Collect and Store: Software version control

https://git-scm.com/

Workstation

Your own server

Gitea

Forgejo

GitLab

Internet

Codeberg

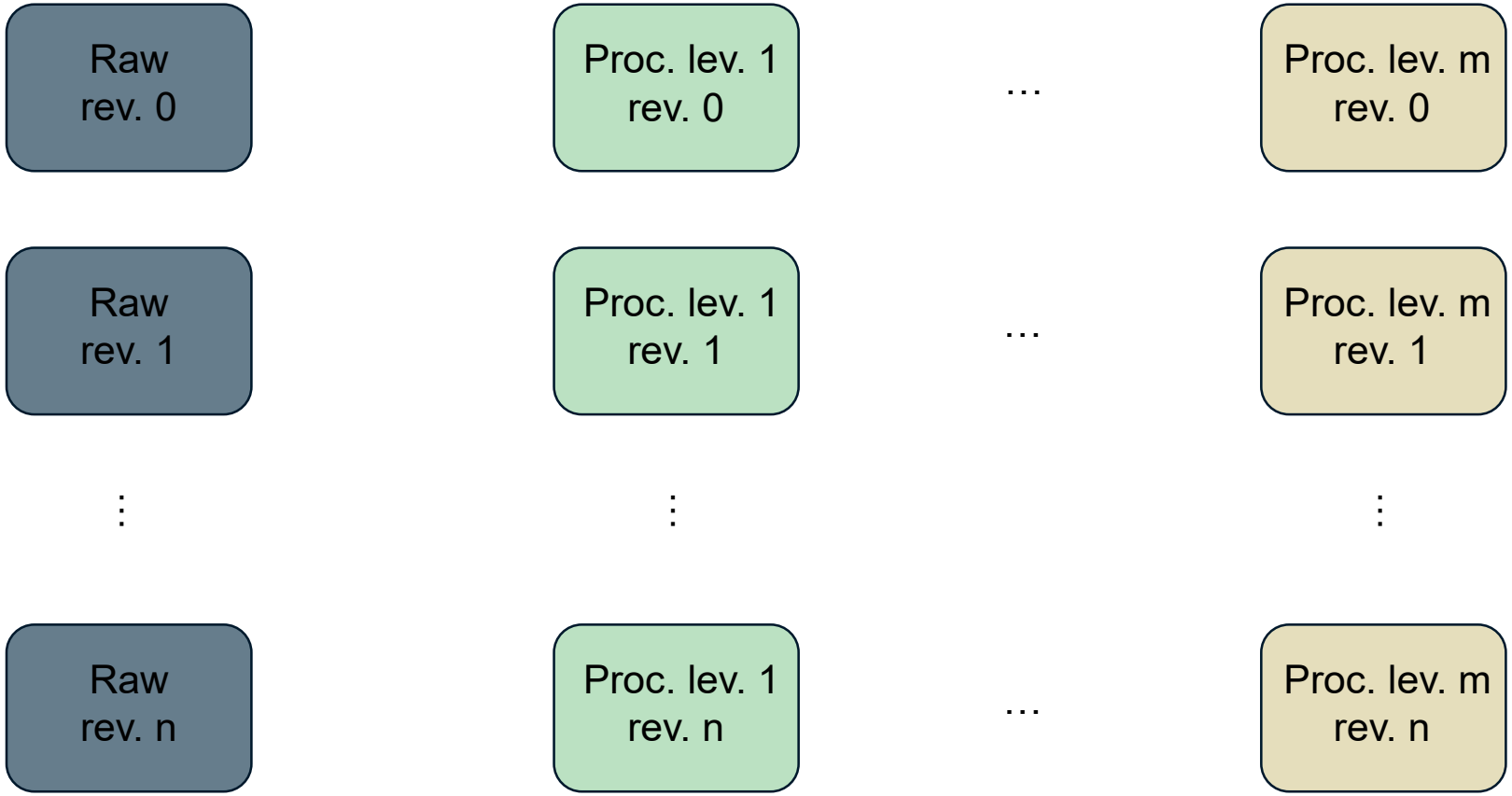
GitLab

GitHub

Bitbucket

- CLI (*Command Line interface)*
- GUIs (*Graphical User Interfaces)*
  https://git-scm.com/downloads/guis

Collect and Store: Data versioning



Collect and Store: Data versioning tools

Renku (*https://renku.readthedocs.io/en/stable/index.html*)

Data Version Control (*https://dvc.org*)

Git Large File Storage (*https://git-lfs.com*)

Lake FS(*https://docs.lakefs.io*)

Collect and Store: File Naming

- Use unique names referencing content
- Limit to 42 characters (preferably less)
- Use ASCII characters, no spaces, points or special characters, e.g. ~!@#$%^&*()[]{}<>’;,’»/
- Include dates and label versions
- Use names to order files:
  - Either, use Dates YYYY-MM-DD or YYYYMMDD (according to ISO 8601) at the beginning to enable chronological order
  - Or, use Versioning with leading zeroes to enable numerical order (enables versions to go beyond 9 without disrupting order)

- If you have started with your project use *Bulk Rename Utility* (Windows) or *Renamer 6* (Mac), *Rename/Thunar Bulk Rename* (GNU/Linux)

Collect and Store: File Formats (recommendation)

Data type	Recommended file formats
Text	• PDF/A • Plain Text coded as ACII. UTF-8 or UTF-16 • XML
Spreadsheet	• CSV (NEAD)
Images	• TIFF (uncompressed or lossless compressed) • PNG
Code	• Languages with free environments (e.g. Py or R UTF-8 format of ASCII text)
Audio	• FLAC • Wav

Open and lossless formats
If you are using a proprietary format, think about adding an additional format

Collect & Store: Metadata Standards

- Definition: Structured data that contains information about other data, but is not the content of the data.

- Metadata is very subject specific. The following directories are helpful:
  - Digital Curation Centre (*https://www.dcc.ac.uk/guidance/standards*)
  - RDA Metadata Standards (*https://rdamsc.bath.ac.uk*)
  - Fairsharing (*https://fairsharing.org*)

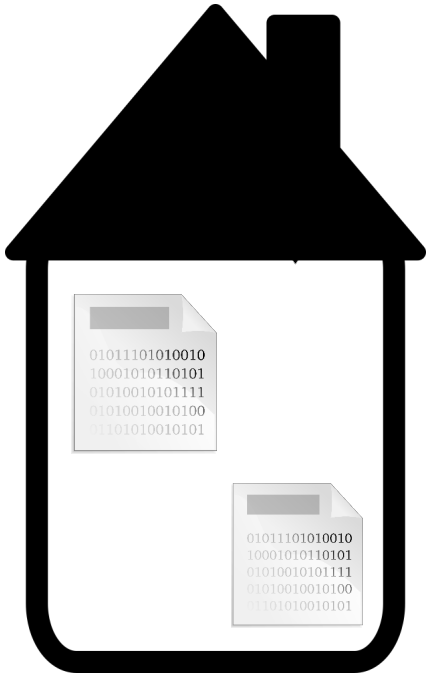
- Recommendation: Stick to a list of defined terms (controlled vocabulary) and don’t use synonyms to describe the same object (e.g. picture or image)

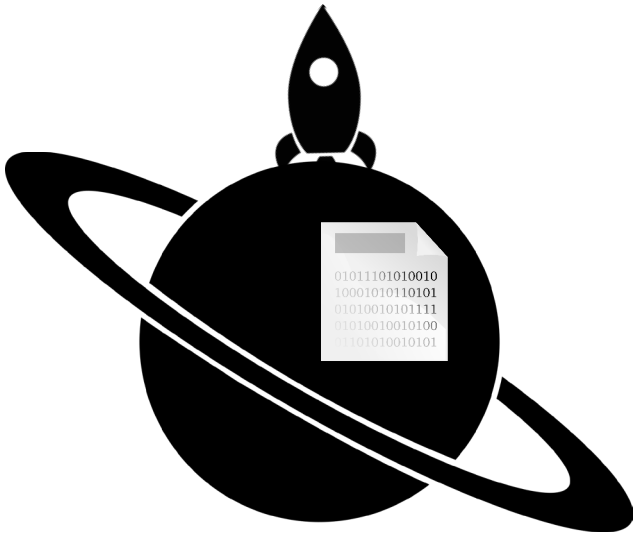
Collect & Store: README File

General information	• Title of the dataset • Contact information principal investigator • Date of data collection • Geographic location
Data and file overview	• Short discription for each file name • Date
Sharing and access informations	• Licenses or restrictions
Methodological information	• Description of methods for data collection or generation • Description of methods used for data processing
Data specific information (repeat for each dataset)	• Variable list, including names and definitions • Units of measuments • Definition for codes or symbols to record missing data

Cornell University: Minimal viable content. For recommended visit: *https://data.research.cornell.edu/content/readme*

Collect and Store: 3 – 2 – 1 backup





Evaluate & Archive

Evaluate & Archive: Data Protection

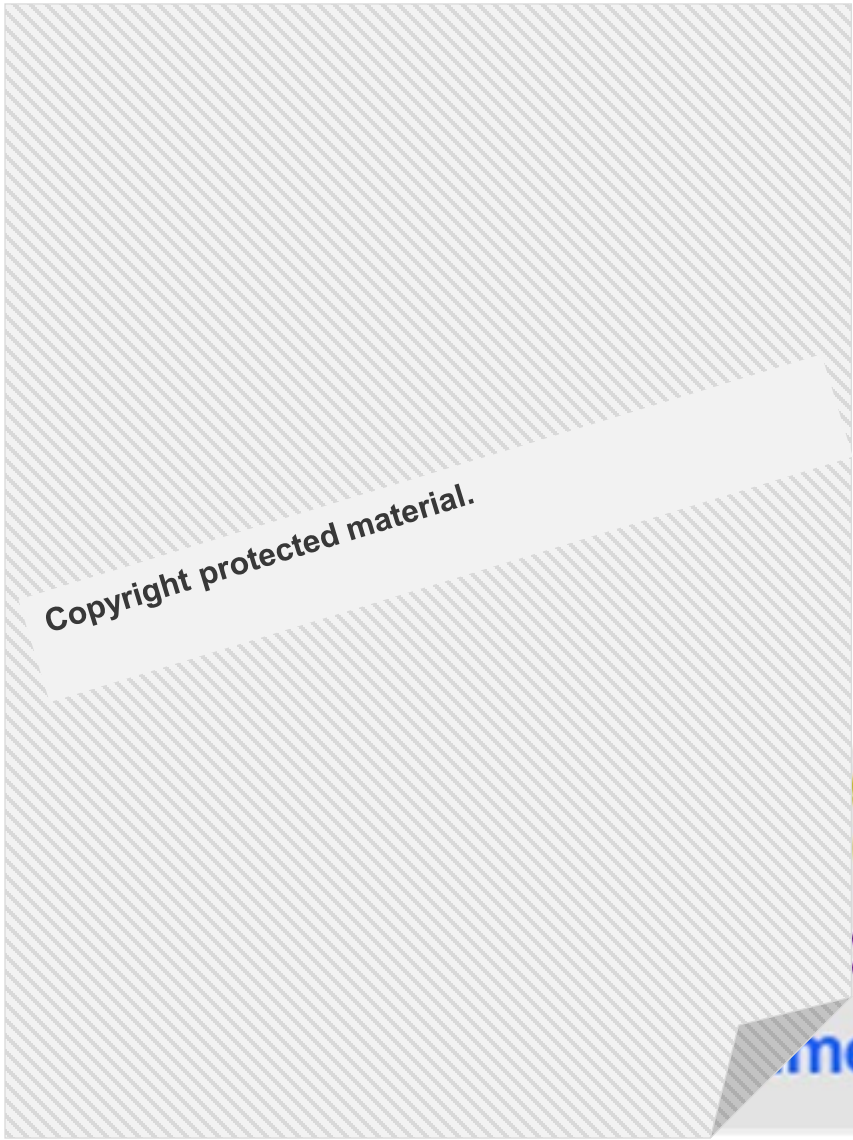
- **Relates to identified or identifiable person**
- **Solutions (https://dmlawtool.ccdigitallaw.ch/) :**
  - Identity irrelevant -> anonymisation
  - Identity relevant -> Ask for consent

    -> Pseudoanomization

    -> Manage access rights

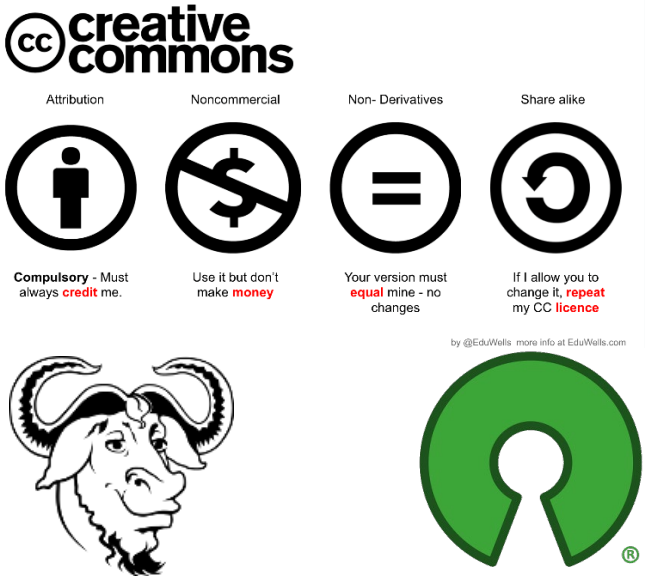
    -> Ability to address

    subject's rights
- **Always contact Data Protection Officers at your Research Institute if your research involves personal data**



Evaluate & Archive: Data Protection

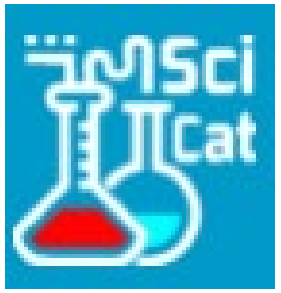
- Processed Data has copyright according to Swiss law
- Use CC licences when publishing factual data on data repositories (ideally CC 0)
- For software use licences specifically designed for software:
- Free Software (Open Source) licences like GPL, Apache, BSD and MIT.
- Exceptions! If you collaborated with external partners in your research project, you need to clarify together with them how and if data can be published.
- Contact the legal teams at your research institute if you feel lost.

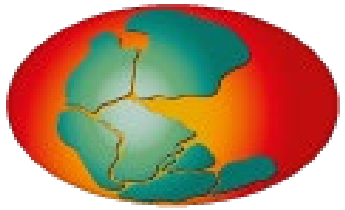


Share & Disseminate

Share & Disseminate: The Choice of Data Repository

ERIC open





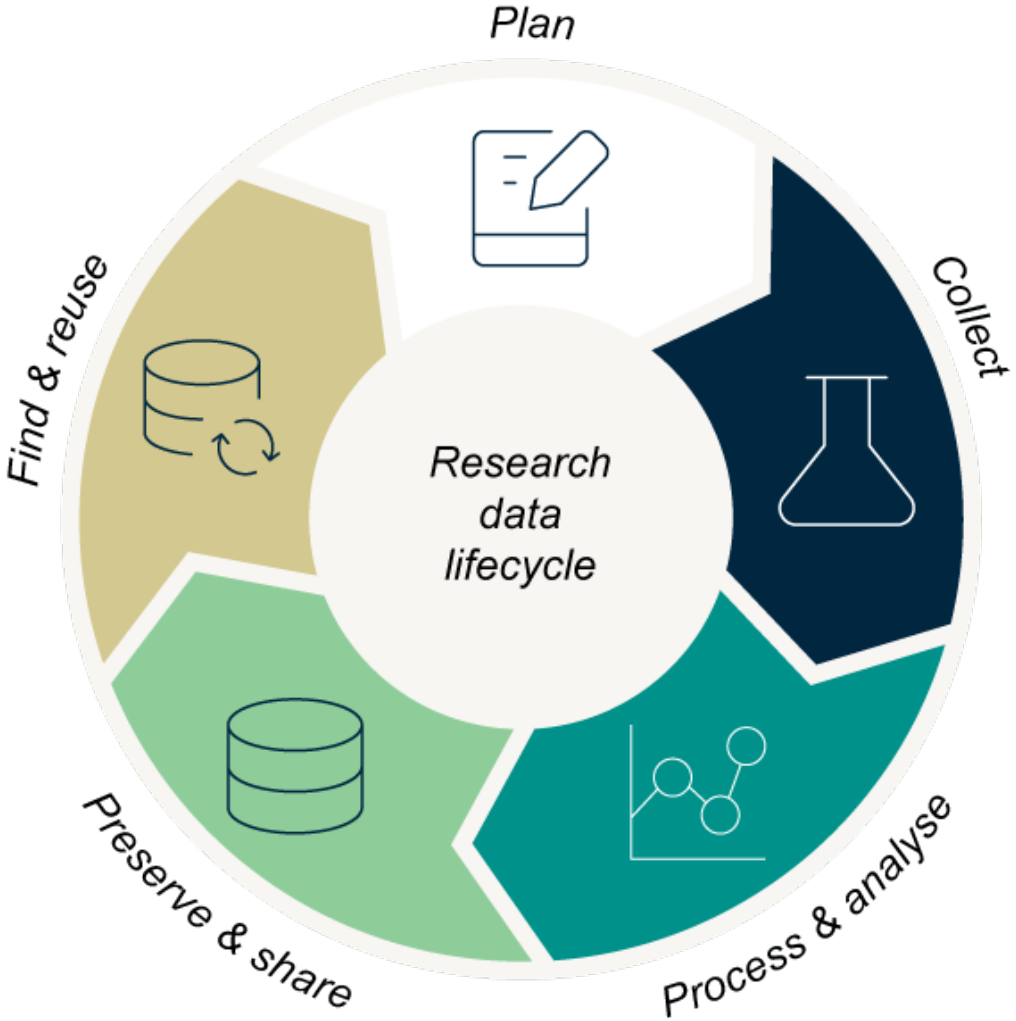
For alternatives: https://www.re3data.org/

RDM Services and Support at WSL

Program

Topic	Speaker	Time
Introduction	Fabian Felder	9.00 - 9.15
Policies, Incentives & the Research Data Life Cycle	Fabian Felder	9.15 - 9.30
Collect & Store	Federico Cantini	9.30 - 10.05
Evaluate & Archive Share & Disseminate	Fabian Felder	10.05 - 10.15
Break		10.20 - 10.30
RDM Services & Support at WSL	Ionut Iosifescu	10.30 - 11.00
Plan & Design	Everyone	11.00 - 11.45

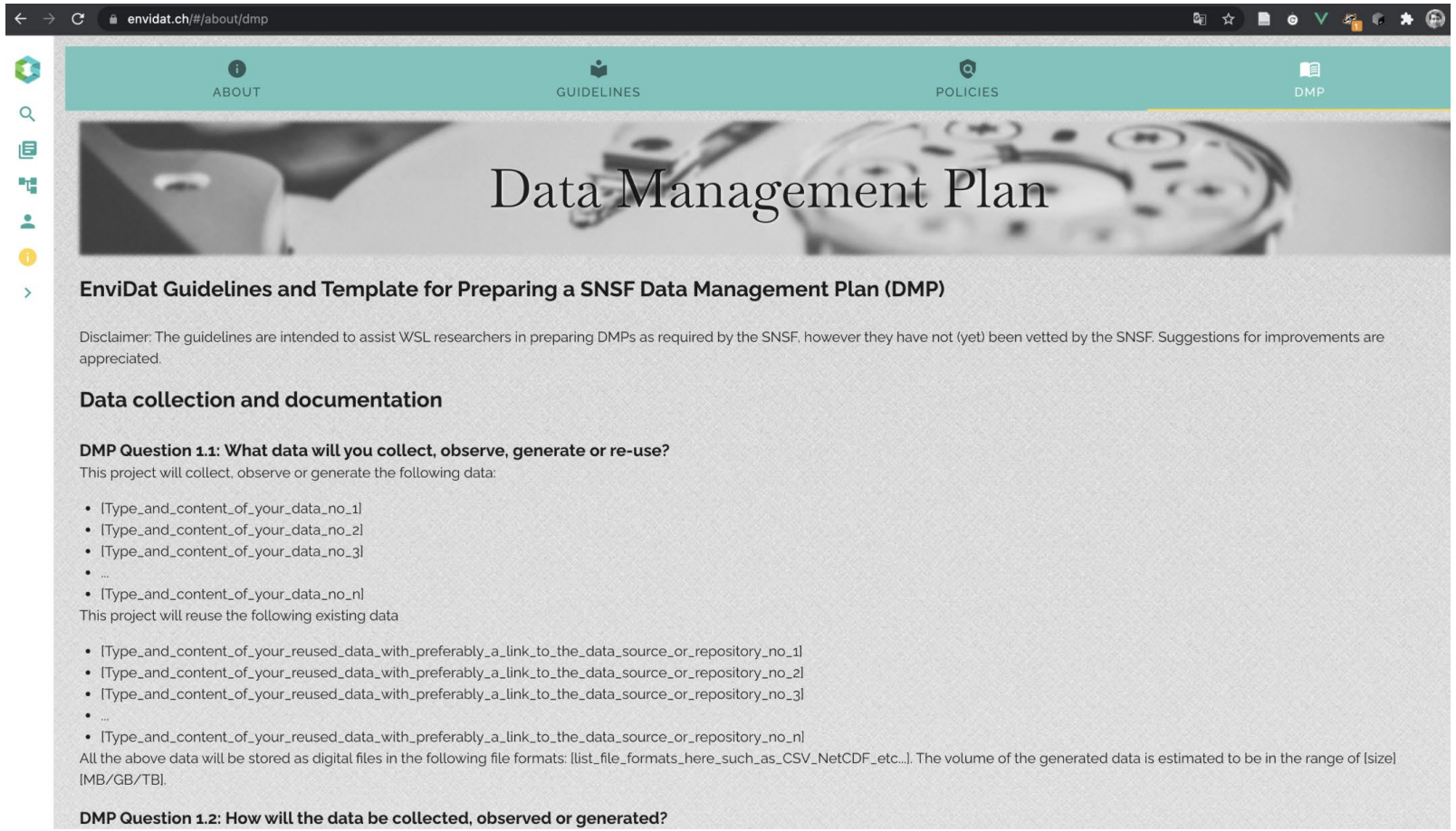
RDM Services and Support at WSL



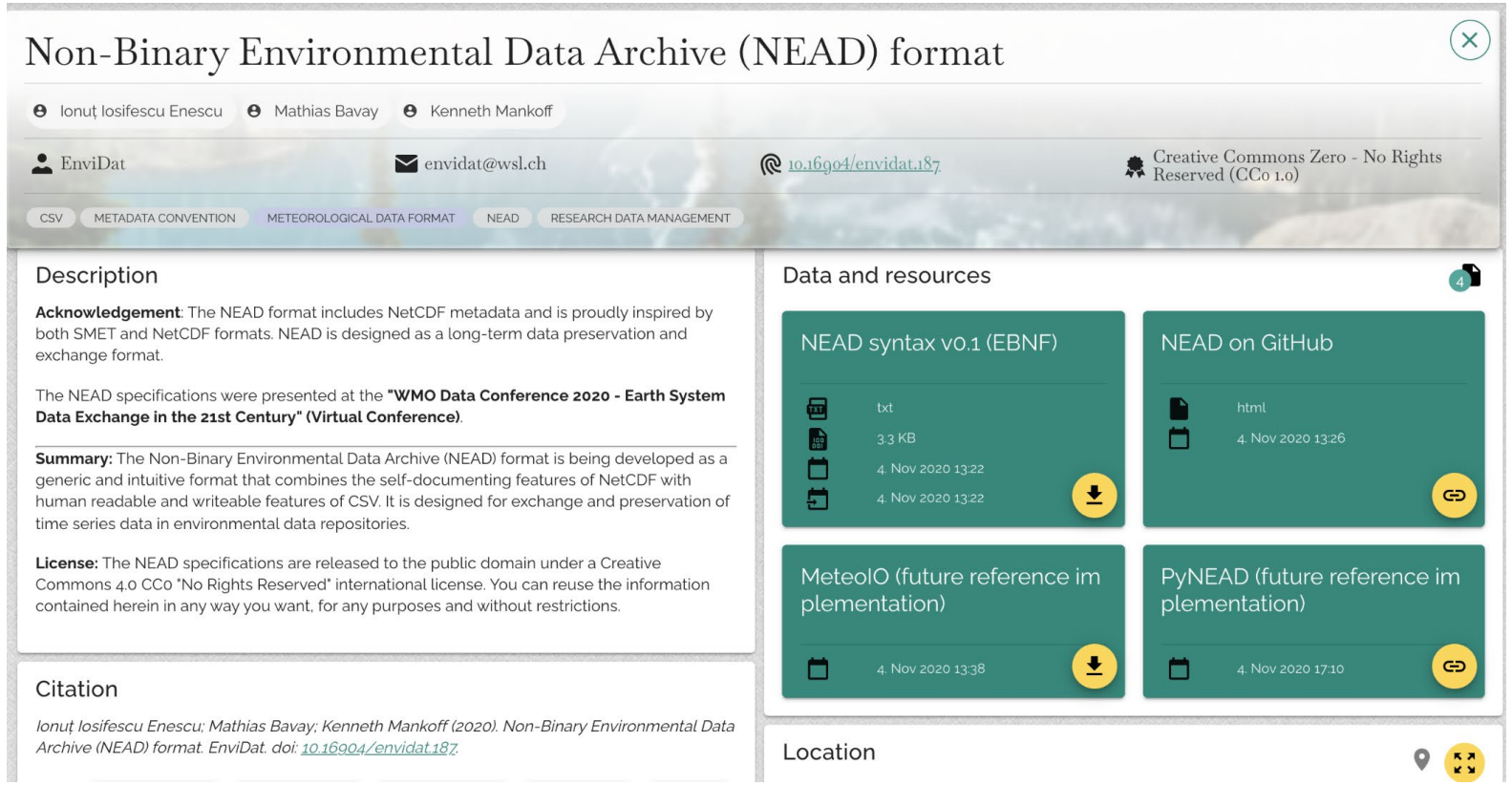
DMP

GitLab

RDM Services and Support at WSL: DMP Template

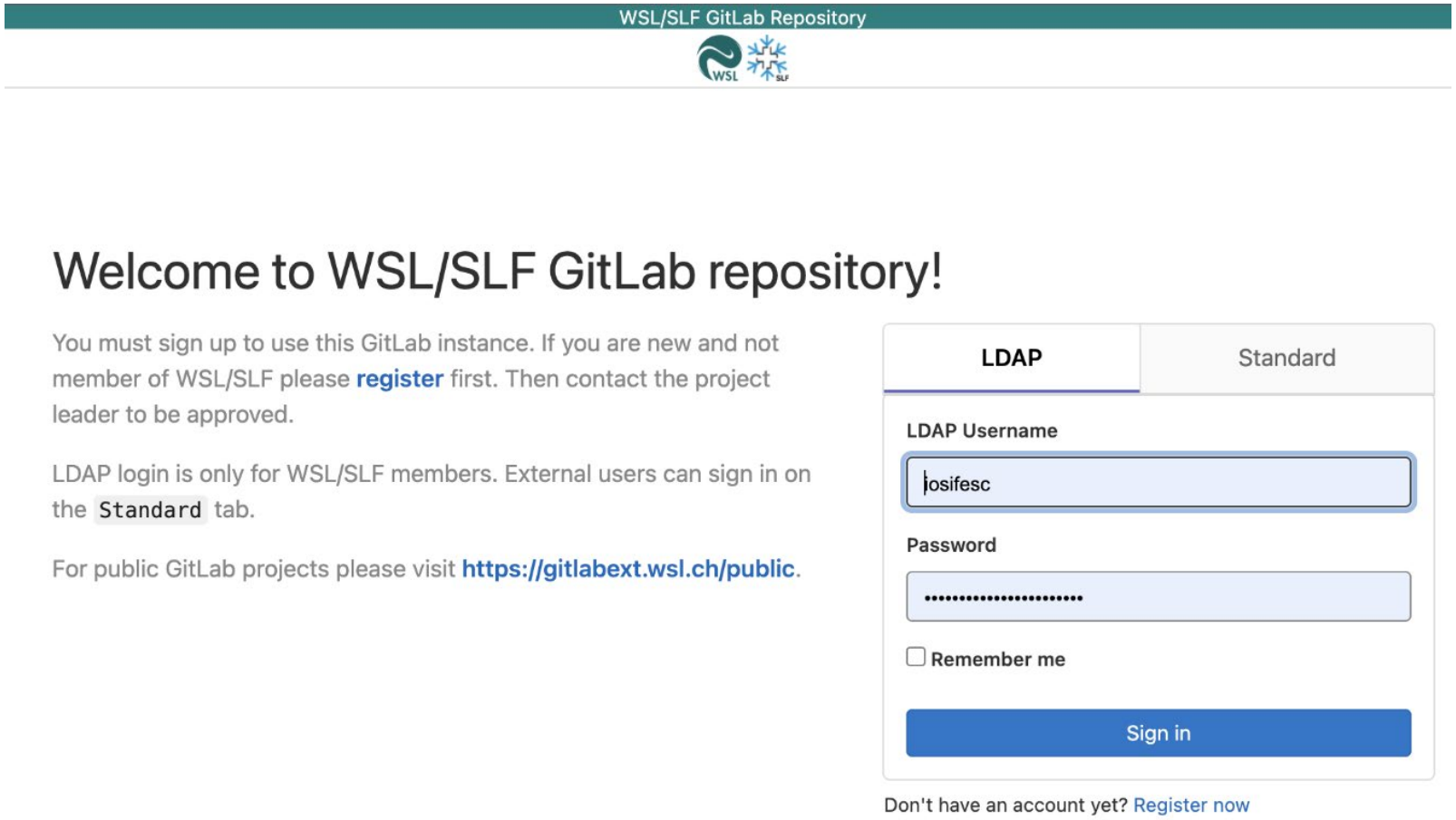


RDM Services and Support at WSL: NEAD Format



https://www.doi.org/10.16904/envidat.187

RDM Services and Support at WSL: GitLab



RDM Services and Support at WSL: Jupyter Notebooks

RDM Services and Support at WSL: EnviDat

EnviDat is the environmental data portal and repository developed by the Swiss Federal Research Institute WSL. EnviDat is designed to host, publish, connect and search environmental research data.

Research data repository

Research data publication

Document Object Identifiers

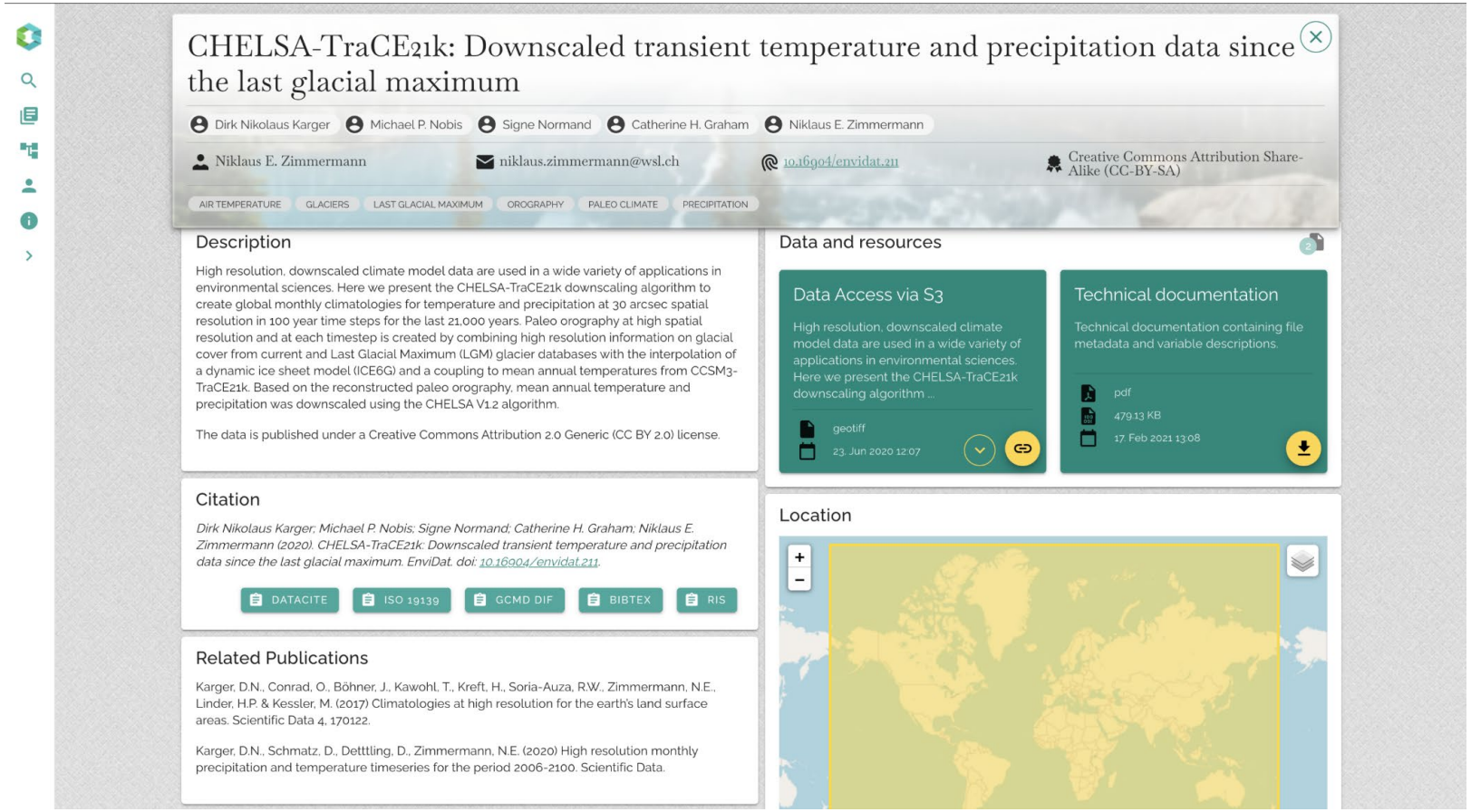
Metadata registration

Curated by researchers

User friendly portal

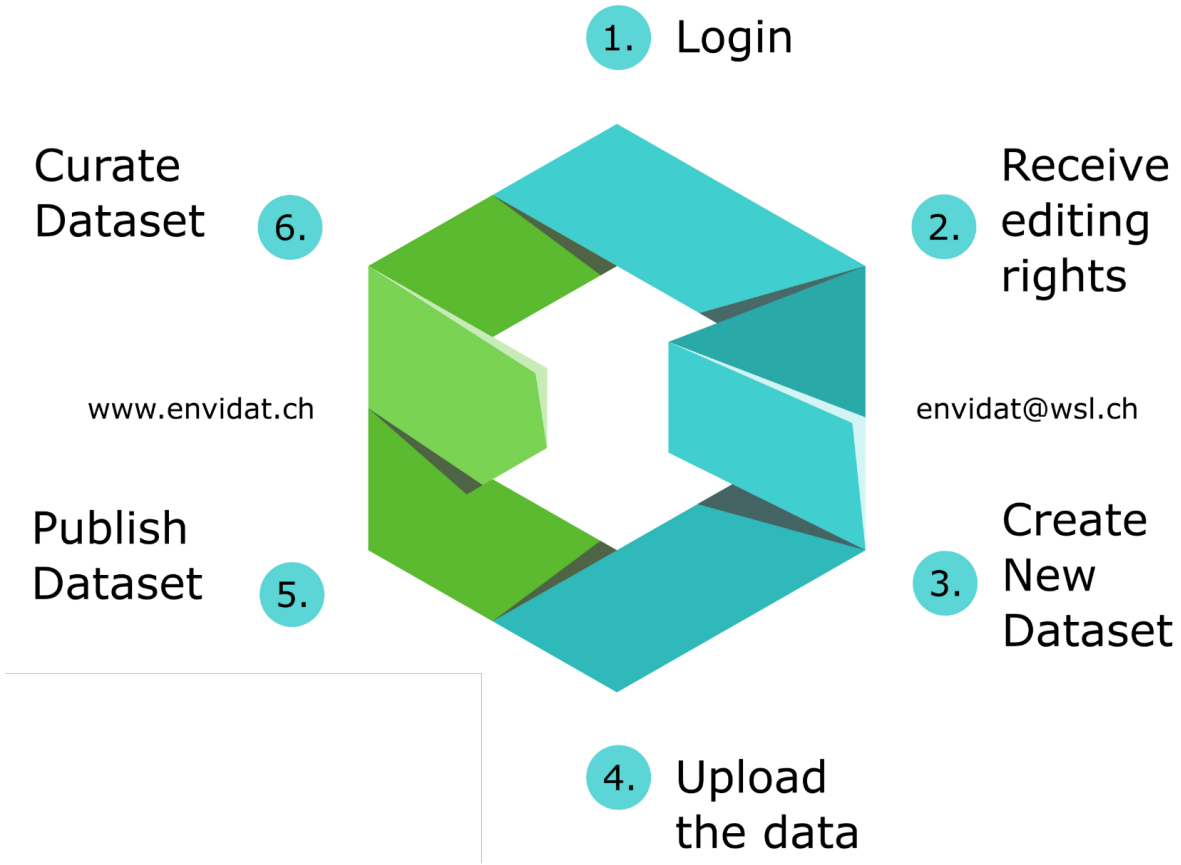
Distributed data management

RDM Services and Support at WSL: EnviDat



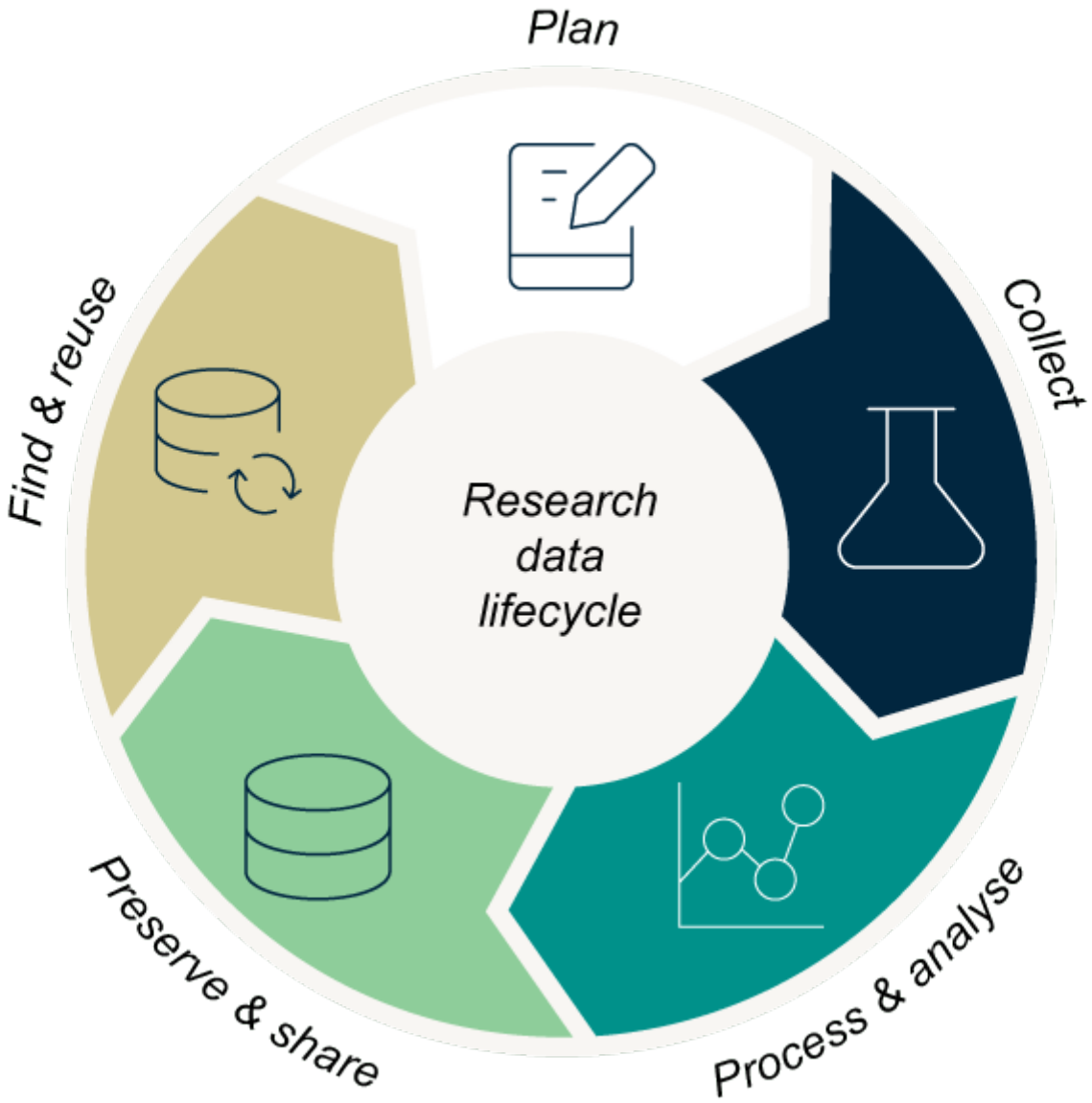
doi:10.16904/envidat.211

RDM Services and Support at WSL: EnviDat



Plan & Design:

Data Management Plan (DMP)



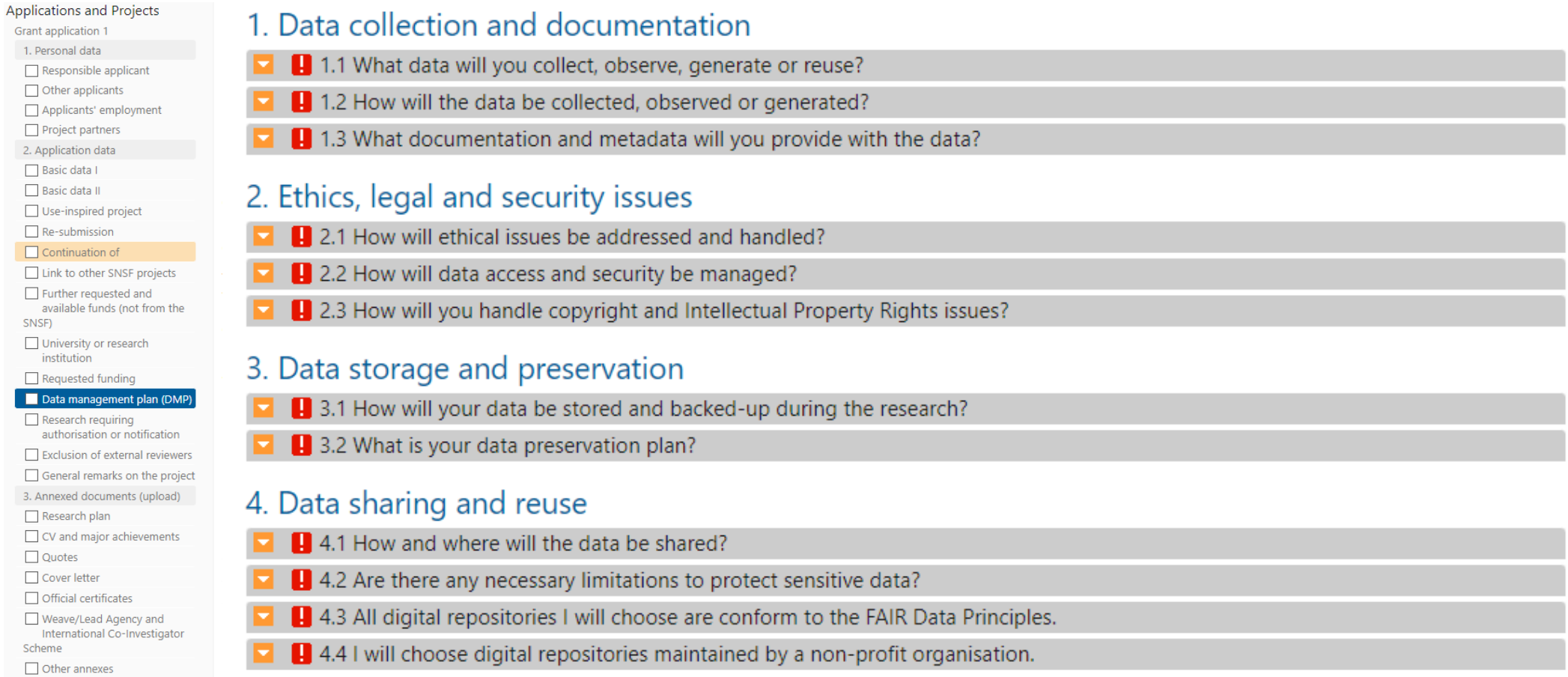
DMP

Covers the whole Research Data Life Cycle

Plan & Design: DMP

- What types of data will be collected and which code (incl. software) will be created or used?
- How will you document the data used and code programmed?
- Where will data and code be stored?
- Who owns the data and code is responsible for security and backup?
- Which data and code will be shared and preserved?
- How will data be shared and with whom?

Plan & Design: DMP



Plan & Design: DMP

- Keep it short and simple
- Be stingy with words
- Have one idea per sentence
- Use the active form — «we used the method» not «the method was used»
- Use positive phrases — «the results are different» not «the results are not the same»
- Use concrete terms — «it will be published in Nature» not «it will be published in a reputable journal»

- Don’t write in «sophisticated style»
- Save on adjectives and adverbs
- Avoid unnecessary constructions — e.g. «It is clear that», «the fact is that», «in an attempt to», «in order to»
- Don’t nominalise — «reduce» not «achieve a reduction in length»
- Don’t use empty modifiers — e.g. «basically», «indeed», «quite», «actually»
- Don’t use tautologous modifiers — e.g. «completely finish», «may potentially», «ultimate result», «blue in colour»

Plan & Design: DMP

1. **Organize yourselves in groups of two (5 minutes)**

2. **Each group will engage with the first section of the SNSF DMP (20 minutes)**
   - Read requirements
   - Write answers and questions
   - Discuss with other group members
   - Designate presenter

3. **Presentation and discussion of findings (20 minutes)**

Plan & Design: DMP - Data Collection and Documentation

1.1 What data will you collect, observe, generate or reuse?

- Type, format (NEAD), content, volume of data, reference to data (if reused)

1.2 How will the data be collected, observed, generated?

- Standards methodology, quality assurance
- File organisation and versioning (folder structures, git, ELN/LIMS, etc.)

1.3 What documentation and metadata will you provide?

- Scientific Metadata (README, metadata standards)
- General Metadata (Depending on choice of data repository)

Plan & Design: DMP - Ethics, Legal and security issues

2.1 How will ethical issues be addressed and handled?

- Information and consent to using personal data, location of critical infrastructure ase well as rare and protected species
- Requirements for assessments by ethical review boards, premission by third parties
- Description of Pseudonymisation or Anonymisation Methods

2.2 How will the data access and security be managed?

- Distinguish datasets according to the level of risk (cf. §2.1) and use an adverb to describe the level of risk («high», «medium», «low»)
- State Storage Location, secure transmission, access restruction, IT infrastructure

2.3 How will you handle copyright and Intellectual Property Rights Issues?

- Consider non-dislosure agreements, potential patents, research collaborations accross institutions
- Recommendation to use CC0 where possible

Plan & Design: DMP - Data Storage and Preservation

3.1 How will your data be stored and backed-up during the research?

- Backup strategy for work at all stages of research (amount of storage needed, frequency of updates, responsibilities, security measures)

3.2 What is your data preservation plan?

- Data formats
- Selection mode for data to be preserved (all relevant data related to reported results, long term preservation of unique datasets)

Plan & Design: DMP - Data Sharing and Reuse

4.1 How and where will the data be shared?

- Repository of choice (non-commercial preferred and required for contribution of up to 10’000 CHF for storage)
- Metadata Policy of said repository

4.2 Are there necessary limitations to protect sensitive data?

- Reasons data cannot be published at certain times (Section §2.1)

4.3 All Digital Repositories I will choose conform to FAIR Data?

- Check box

4.4 All Digital Repository I will choos are mainained by a non-profit oranisation?

- If no, provide justification (costs will not be covered)

Thank you for your attention!

Feedback!

Please give us a short feedback

Questions?

Presentation slides: lib4ri.ch > Learn > Trainings

Appendix

Appendix: Eawag

- **Four links under data.eawag.ch:**
  - https://opendata.eawag.ch/eawagrdm/help/quickstart.html
  - https://opendata.eawag.ch/eawagrdm/help/opendata.html
  - https://doi.org/10.25678/000066
  - https://www.internal.eawag.ch/fileadmin/intranet/informatik/datenman/rdm/directive_archiving_of_researchdata.pdf
- **Difference between ERIC/internal (data.eawag.ch) and ERIC/open (opendata.eawag.ch)**
- **Services are in the form of guides and consulting. Most notable guides in addition to the one mentioned above are**
  - https://doi.org/10.25678/000033
  - https://opendata.eawag.ch/eawagrdm/software-licensing.html
- **Finally the list of resources can be helpful:**
  - https://opendata.eawag.ch/eawagrdm/resources.html

Appendix: Empa

- General overview of topics:
  https://www.empa.ch/web/s909/overview
- Support topics like DMP template of Empa:
  https://www.empa.ch/web/s909/support1

OpenBIS

- General overview: https://www.empa.ch/group/s909/openbis
- Documentation & trainings info:
  https://www.empa.ch/group/s909/documentation-tutorials

Appendix: File Formats EPFL

Bibliothèque de l'EPFL, Research Data, fast guide #4», 2019, https://bit.ly/3NFloYx

TYPE OF DATA	APPROPRIATE	ACCEPTABLE	DEPRECATED
Tabular (extensive metadata)	CSV – HDF5	TXT – HTML – TEX – FASTQ [3] – POR	
Tabular (minimal metadata)	CSV – TAB – ODS – SQL – TSV	XML (if appropriate DTD) – XLSX	XLS – XLSB
Textual / Presentation	TXT – PDF – ODT – ODM – TEX – MD – HTM – XML – EXTXYZ [4] – ODF	PPTX – RTF – DOCX – PDF (with embedded forms) – EPS – IPF	DOC – PPT – DVI – PS
Code / Computation	M – R – PY – IYPNB – RSTUDIO – RMD – NETCDF – AIML	SDD	MAT – RDATA
Image & Spectroscopy	TIF – PNG – SVG – JPEG – FITS	JCAMP – JPG – JP2 – TIF – TIFF – PDF – GIF – BMP – DM3 – OIR – LSM [5]	INDD – AIT – PSD – SPC
Audio	FLAC – WAV – OGG – MXL – MIDI – MEI – HUMDRUM	MP3 – AIF	
Video	MP4 – MJ2 – AVI – MKV	OGM – MP4 – WEBM	WMV – MOV – QT
Geospatial	NETCDF – tabular GIS attribute data – SHP – SHX – DBF – PRJ – SBX – SBN – POSTGIS – TIF – TFW – GEOJSON	MDB – MIF	
3D structures & images	X3D – X3DV – X3DB – PDF3D – POV – PDBML	DWG – DXF – PDB	PXP
Generic	XML – JSON – RDF		

Appendix: File Formats ETH Zürich

ETH-Library, File formats for archiving, 2022, https://bit.ly/3DBqXmb

Assessment of various file formats

Table 1: Our assessment of future readability of some common file formats. (For more detailed information we refer to the recommendations of the Bundesarchiv (German), the KOST (German or French), the Memoriav, the Forschungsdatenzentrums Archäologie & Altertumswissenschaften IANUS (Germany), the Library of Congress and the Harvard Library.)

File type	Recommended	Suitable to only a limited extent	Not suitable for archiving
Text	• PDF/A (*.pdf, preferred subtypes 2b and 2u) • Plain Text (*.txt, *.asc, *.c, *.h, *.cpp, *.m, *.py, *.r etc.) coded as ASCII, UTF-8, or UTF-16 using byte order mark • XML (inclusive XSD/XSL/XHTML etc.; with included or accessible schema and character encode explicitly specified)	• PDF (*.pdf) with embedded fonts • Plain text (*.txt, *.asc, *.c, *.h, *.cpp, *.m, *.py, *.r etc.) (ISO 8859-1 coded) • Rich Text Format (*.rtf) • HTML and XML (The ASCII text is readable over long term; try to avoid external links.) Not accepted for publication, OK for supplementary materials: • Word *.docx • PowerPoint *.pptx • LaTeX, TeX (The ASCII text is readable over long term; open source software required for formatting and the resulting PDF should be included.) • OpenDocument formats (*.odm, *.odt, *.odg, *.odc, *.odf)	• Word *.doc • PowerPoint *.ppt
Spreadsheet or table	• Comma- or tab delimited text files (*.csv)	• Excel *.xlsx (container format) • OpenDocument spreadsheets (*.ods)	• Excel *.xls, *.xlsb (binary formats)
Raw data and workspace		• ASCII Text is suitable for long-term use, but the data import may be time-consuming. • S-Plus files (*.sdd) may be saved as text files. • Matlab *.mat files may be saved in HDF Format. Saving nontrivial ASCII Matlab *.mat files should be avoided because they are not readable with the Matlab load command (see table 2). • Network Common Data Format or NetCDF (*.nc, *.cdf) • Hierarchical Data Format (HDF5) (*.h5, *.hdf5, *.he5)	• Binary files such as the standard Matlab files *.mat or the R files *.RData
Raster image (bitmap)	• TIFF (*.tif) (uncompressed, preferentially TIFF 6.0, Part 1: baseline TIFF). TIFF is preferred as compared to PNG or JPEG2000. • Portable Network Graphics (*.png, uncompressed) • JPEG2000 (*.jp2, lossless compression) • Digital-Negative-Format (*.dng) to keep raw data of digital fotos in addition to an second copy in TIFF format	• TIFF (*.tif) (compressed) • GIF (*.gif) • BMP (*.bmp) • JPEG/JFIF (*.jpg) • JPEG2000 (lossy compression) (*.jp2)	
Vector graphics	• SVG without JavaScript binding (*.svg)		• Graphics InDesign (*.indd), Illustrator (*.ait) • Encapsulated Postscript (*.eps) • Photoshop (*.psd)
CAD	• AutoCAD Drawing (*.dwg) • Drawing Interchange Format, AutoCAD (*.dxf) • Extensible 3D, X3D (*.x3d, *.x3dv, *.x3db)		
Audio	• WAV (*.wav) (uncompressed, pulse-code modulated)	• Advanced Audio Coding (*.mp4) • MP3 (*.mp3)	
Video [1]	• FFV1 codec (version 3 or later) in Matroska container (*.mkv)	• MPEG-2 (*.mpg,*.mpeg) • MP4, which is also called MPEG-4 Part 14 (*.mp4) • QuickTime Movie (*.mov) [2] • Audio Video Interleave (*.avi) • Motion JPEG 2000 (*.mj2, *.mjp2)	• Windows Media Video (*.wmv)

Footnotes

[1] In addition to the file format (or container format), also the codec and the compression method are important. See Ianus, Memoriav and KOST for further information.

[2] In the Version of Nov 21, 2018 of the current document, the format QuickTime Movie was downgraded from „Recommended" to „Suitable to only a limited extent". Apple discontinued the support of Windows QuickTime Player in the year 2016. Windows Media Player thus only supports file format versions 2.0, or earlier, of QuickTime Movie files.

Appendix: References (Slide 18)

[1] SPARC Europe, «The Open Data Citation Advantage», 2017, https://sparceurope.org/open-data-citation-advantage/.

[2] Digital Science, «The state of Open Data Report», 2019, https://digitalscience.figshare.com/articles/report/The_State_of_Open_Data_Report_2019/9980783/2

[3] European Commission and PwC, «Cost-Benefit analysis fro FAIR research Data», 2019.

https://op.europa.eu/en/publication-detail/-/publication/d375368c-1a0a-11e9-8d04-01aa75ed71a1

[4] Baker, M., “1,500 scientists lift the lid on reproducibility”. *Nature* 533, 452–454 (2016). https://doi.org/10.1038/533452a

Appendix: Icon References

Slide 4:

- Le Moign, Vincent, «Lab Scientist Icon», https://icon-icons.com/icon/lab-scientist/101049, free for commercial use.
- Flaticon, «Checkliste», https://www.flaticon.com/de/kostenloses-icon/checkliste_2666469, free for personal and commercial use.
- PLoS, «Open Access logo», https://de.wikipedia.org/wiki/Datei:Open_Access_logo_PLoS_white.svg, CC-0.
- «Databases and People», https://freesvg.org/databases-and-people, CC-0.

Slide 8

- Felixmh, «Krischen-Früchte-Natur-Symbol», free commercial use.