28.11.2023

Research Data Management – The Basics

Bachofner, Anusch
Cantini, Federico
Felder, Fabian
Förster, Christian

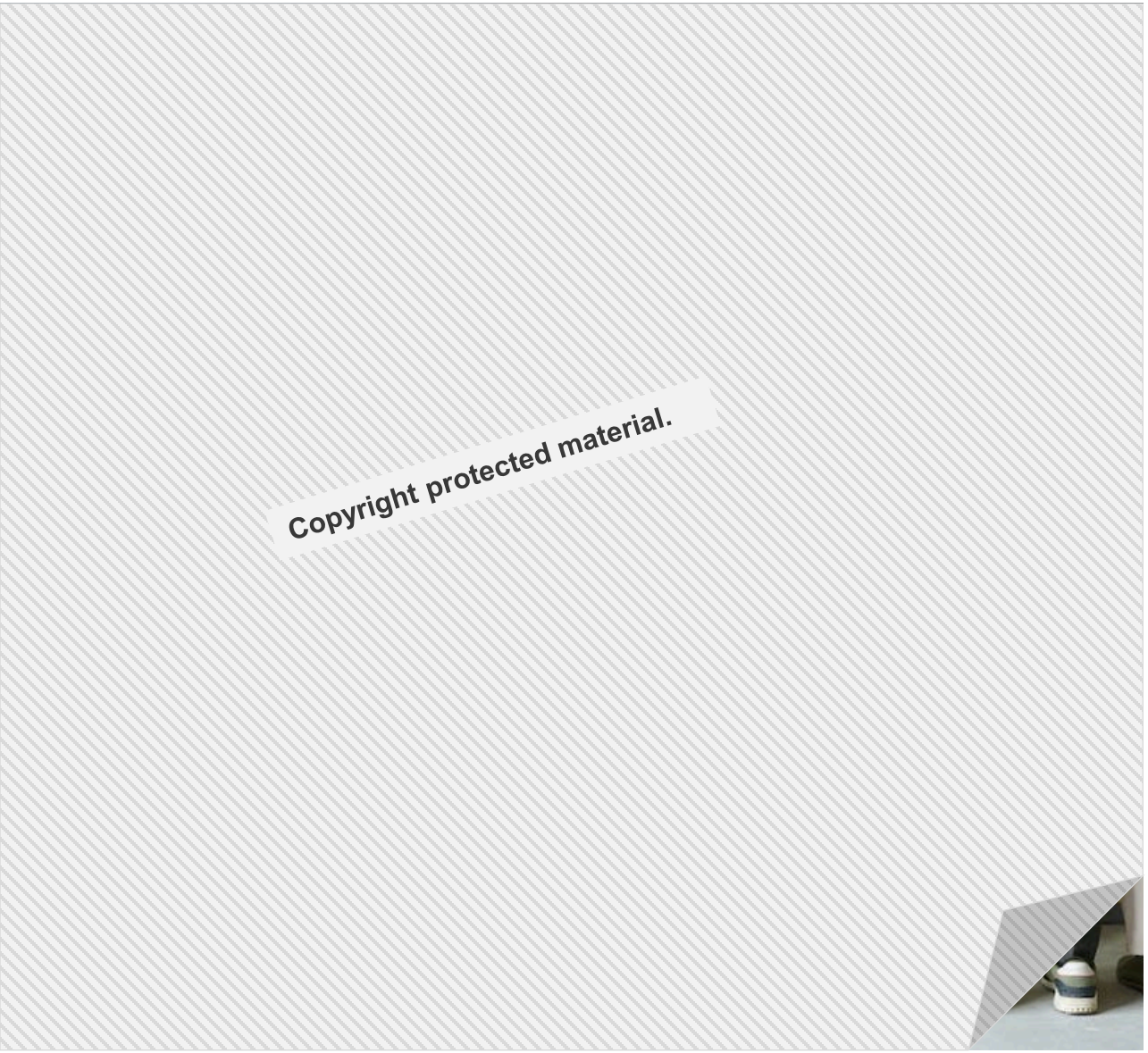
These are your trainers today!

Federico Cantini

- Software Developer
- Technical Lead at Lib4RI

Fabian Felder

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



Who are you and why are you here?

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

Learning Aims

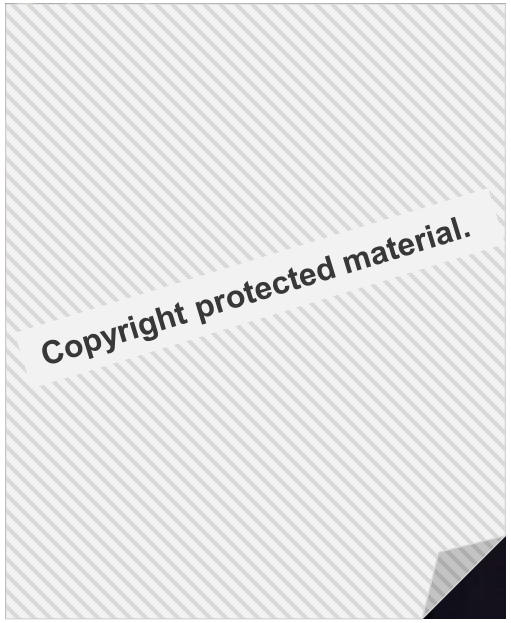
- Life cycle of research data
- Adequate metadata documentation for your code and data
- Storing and publishing data
- Using OpenBIS (ELN) and writing Data Management Plans (DMP)

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 - 10.00
Evaluate & Archive Share & Disseminate	Fabian Felder	10.00 - 10.10
Break		10.10 - 10.25
RDM Services & Support at Eawag	Christian Förster	10.25 - 10.45
RDM Services & Support at Empa	Anusch Bachofner	10.45 - 11.05
Plan & Design	Everyone	11.05 - 11.45

Why is data and associated metadata important?

Why is data and associated metadata important?



«We kill people based on metadata» (2014), Michael V. Haden, director of CIA 2006-2009

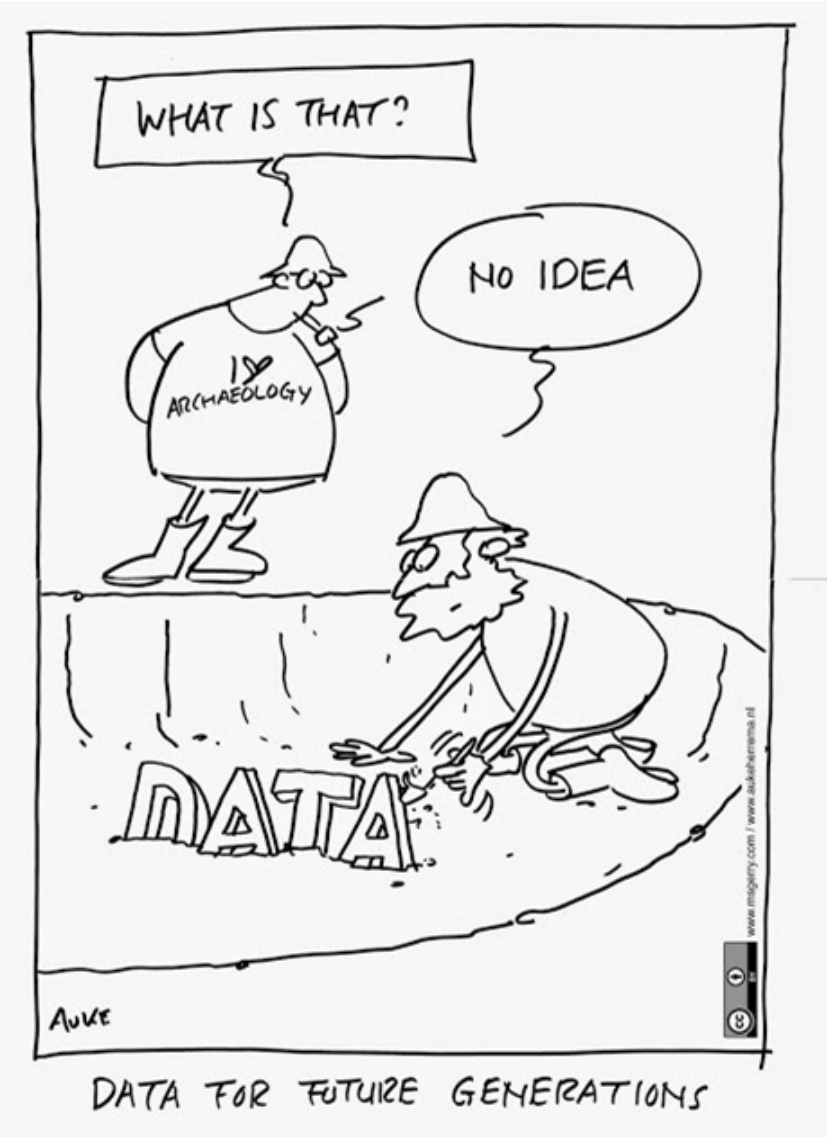
Cham, J. G., «Scratch: A context-changing framework for contextualizing nano informatic structures» (2014), International Journal of Temporal Deflective Behaviour, 4 (1689), p. 432.

Why is data and associated metadata important?

No clean metadata

=

Limited access to Data



Source: www.fosteropenscience.eu/project

Why is data and associated metadata important?

Proper metadata tagging

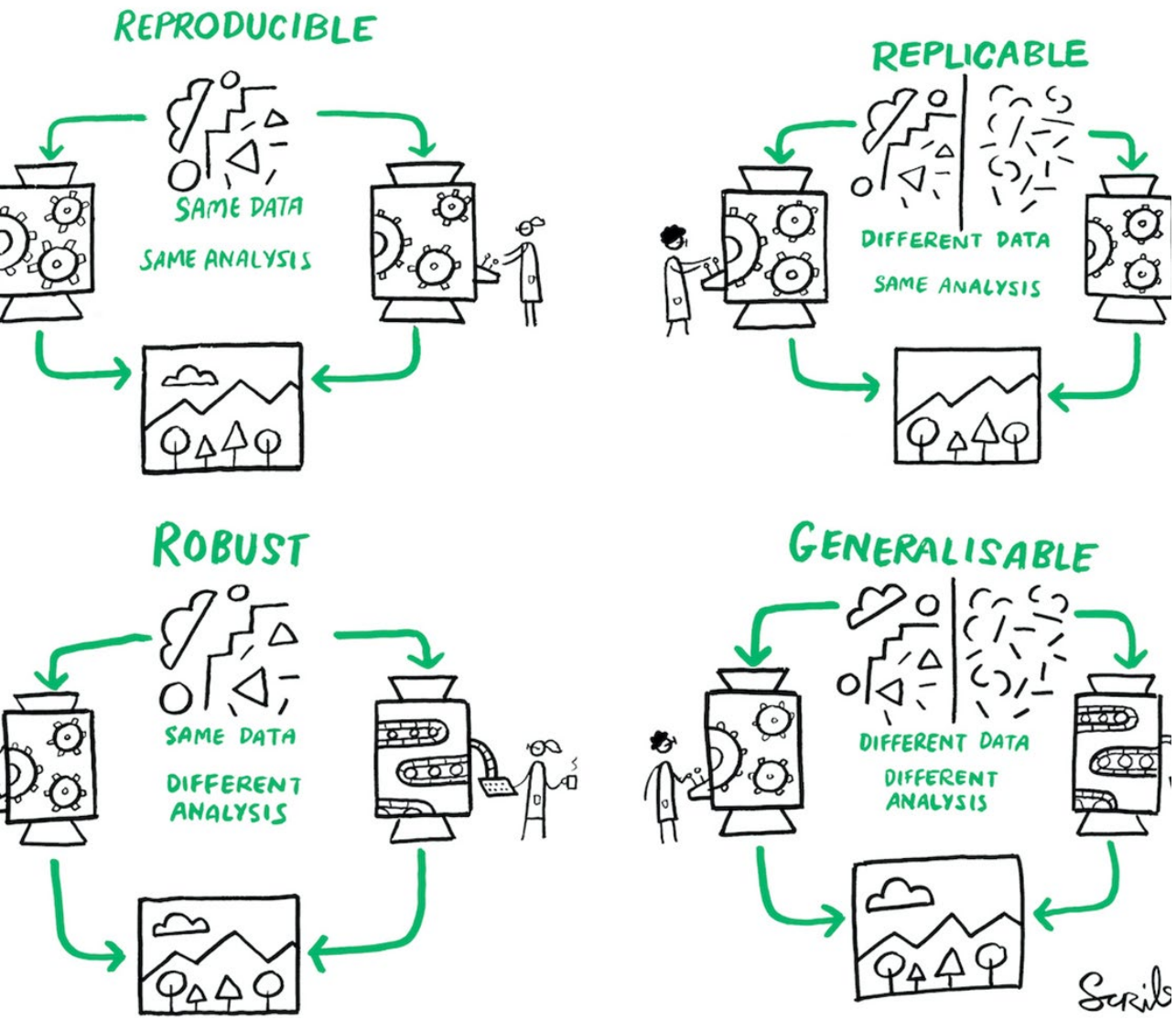
and

research description

is

time consuming

Reproducibility



Reproducibility

Scriberia, "Reproducible Research", *The Turing Way*, CC-BY, DOI: 10.5281/zenodo.3332807

A Handbook for Reproducible Data Science, https://the-turing-way.netlify.app/welcome.html

LEGO® Metadata for Reproducibility

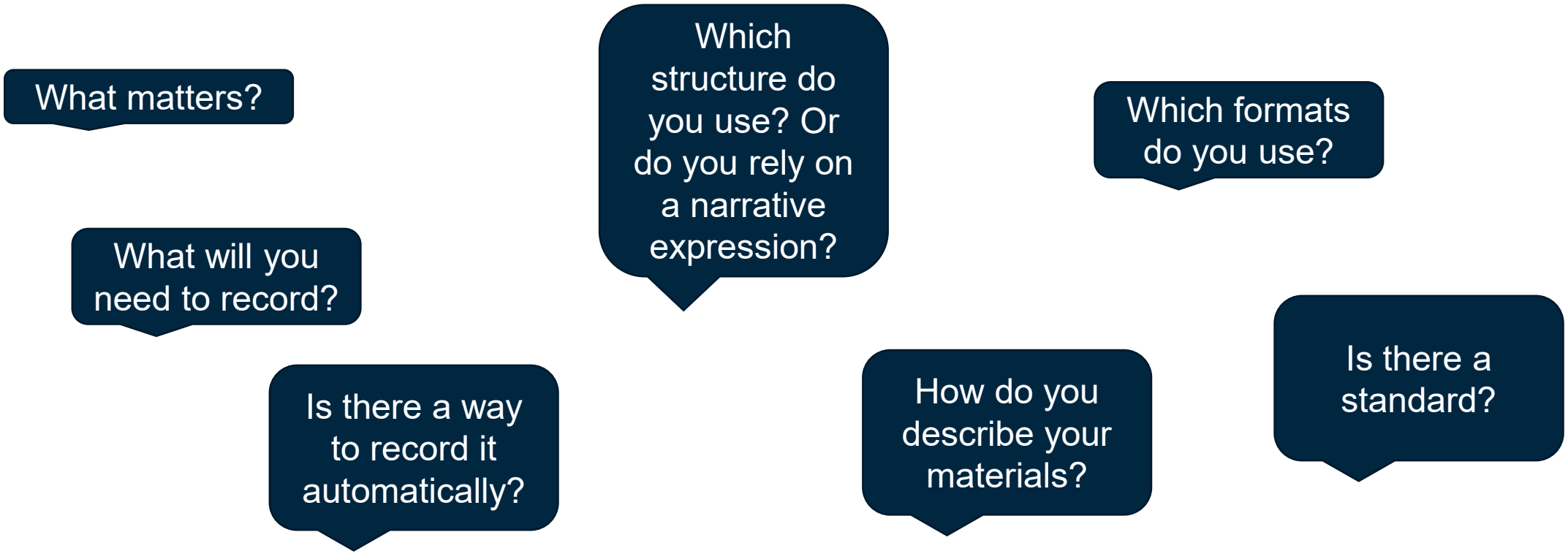
Group A builds Car

Group A documents build

Group B rebuilds the car

Donaldson, Mary and Matt Mahon, «Lego: Metadata for reproducibility», 10.5281/zenodo.3685685.

LEGO® Metadata for Reproducibility



Donaldson, Mary and Matt Mahon, «Lego: Metadata for reproducibility», 10.5281/zenodo.3685685.

FAIR principles – A lot of Metadata

Findable

F1 (Meta)data are assigned a globally unique and persistent identifier
F2 Data are described with rich metadata
F3 Metadata clearly and explicitly includes the identifier
F4 (Meta)data are registered or indexed in a searchable resource

Accessible

A1 (Meta)data are retrievable by their identifier using a standardised communications protocol
- A1.1 The protocol is open, free, and universally implementable
- A1.2 The protocol allows for an authentication and authorisation procedure, when necessary

A2. Metadata are accessible, even when the data are no longer available

Interoperale

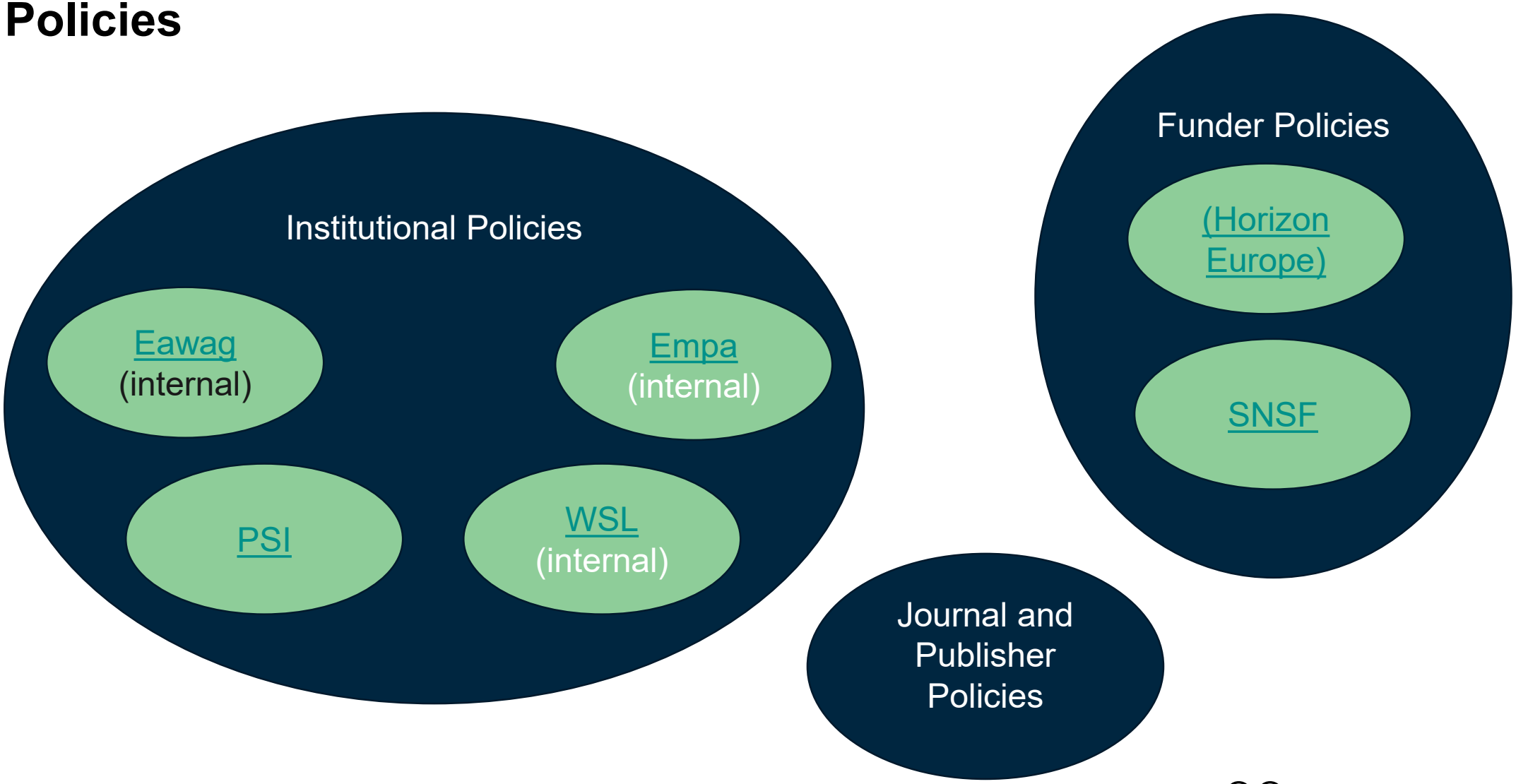
I1 (Meta)data use formal, accessible, shared, and broadly applicable language for knowledge representation
I2 (Meta)data use vocabularies that follow FAIR principles
I3 (Meta)data indlude qualified references to other (meta)data

Reusable

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

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

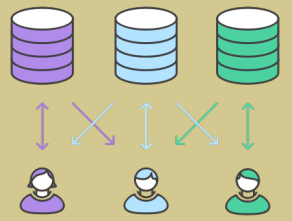
DMP

Required for
Funders

As open as
possible,

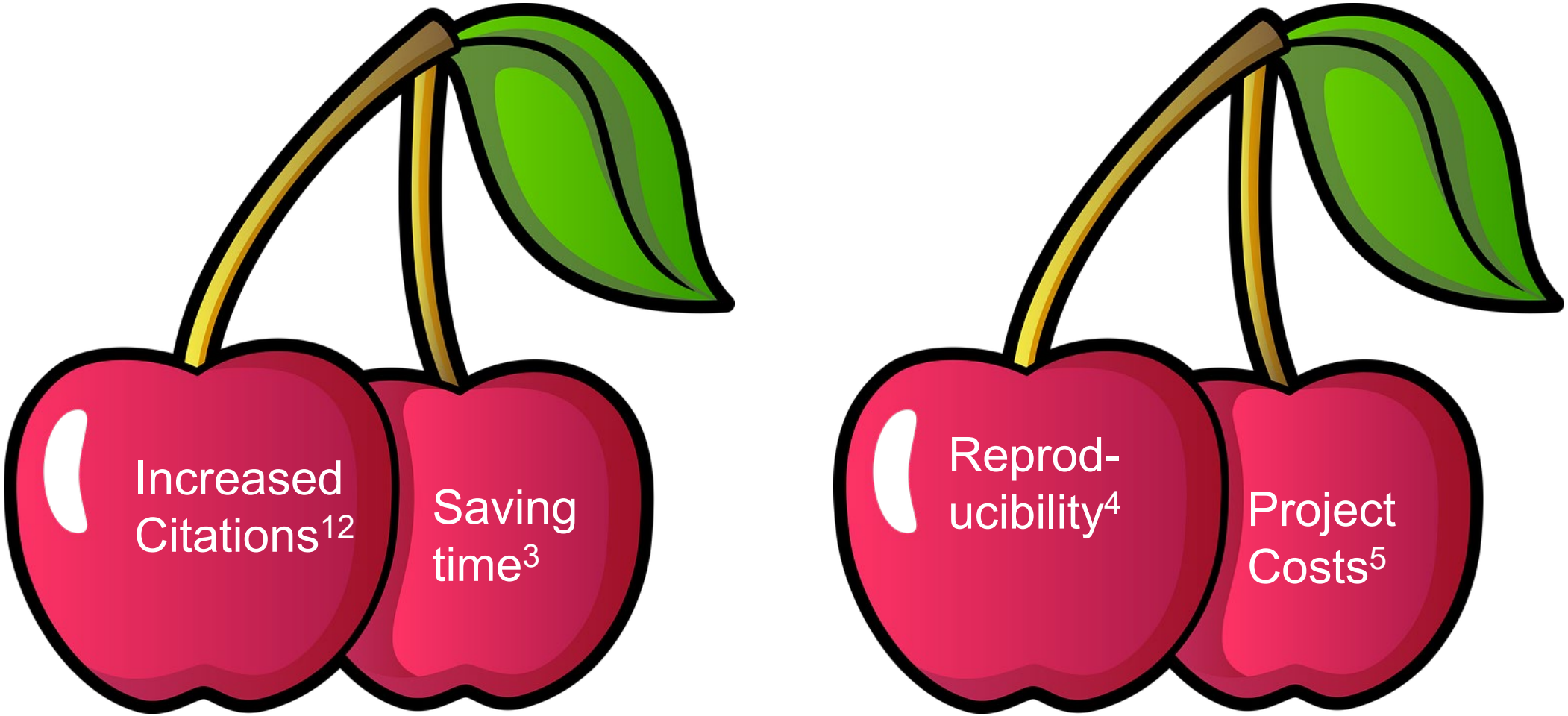
as closed as
necessary.

Repositories



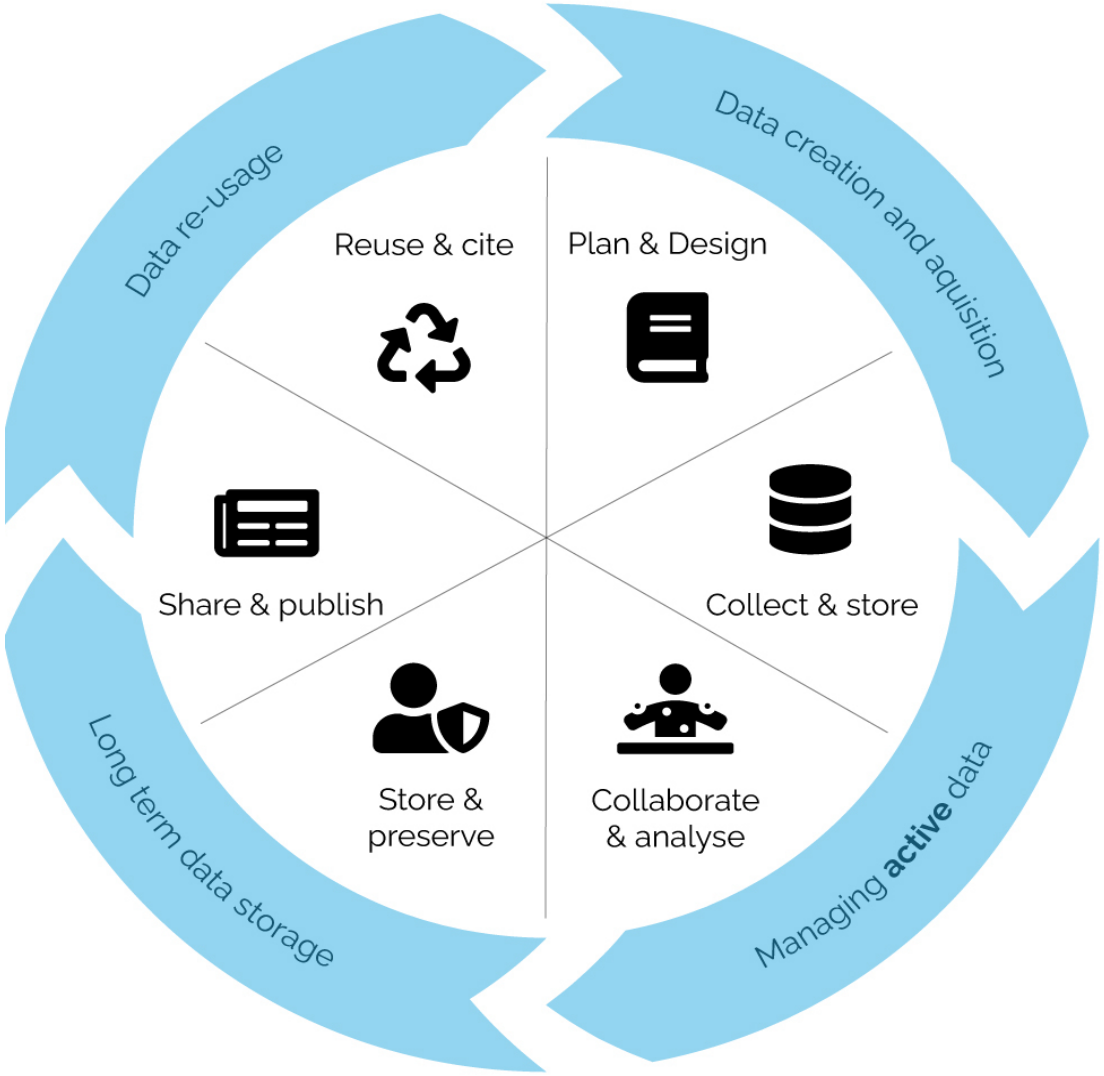
FAIR and non-
commercial

Policies


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

Research Data Life Cycle

Research Data Life Cycle



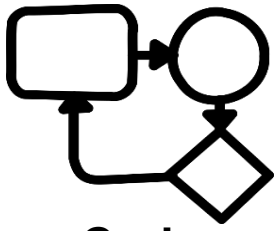
Collect & Store

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…

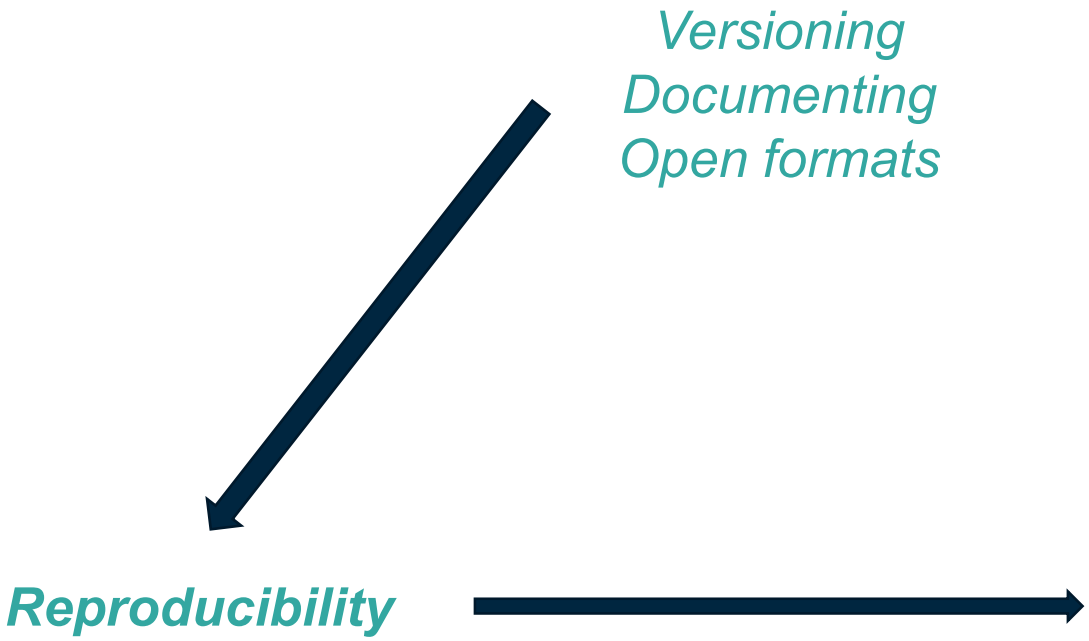
</>

Metadata

Structured information associated with data (and code)

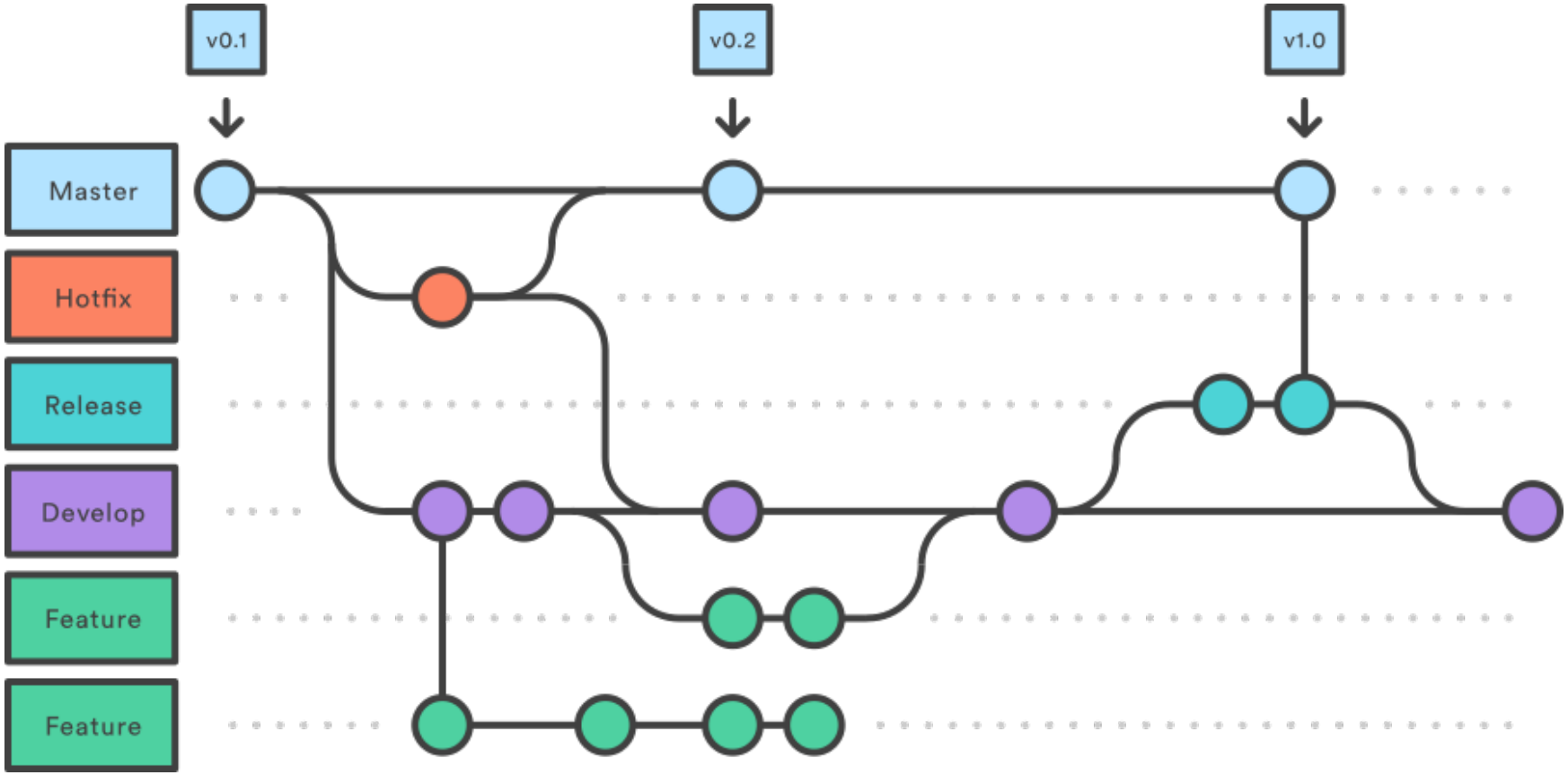
The Who, What, Where, Why & How of data

Collect and Store

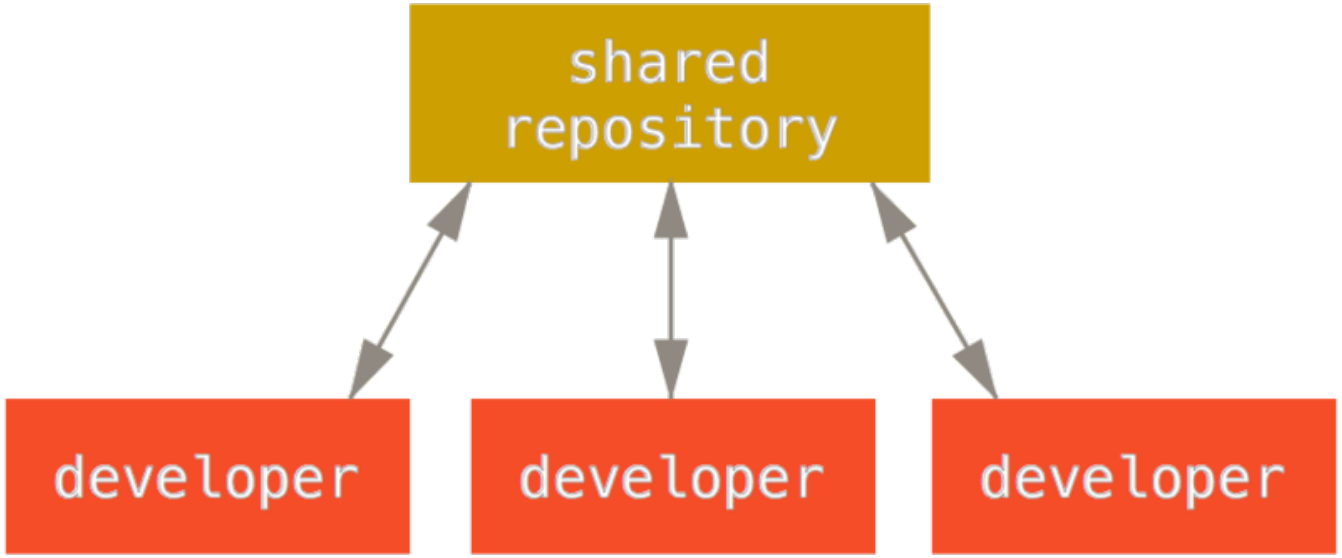


- **You** can find it
- **Your coworkers** can find it
- You can easily **share** it
- It's **ready** for **archiving/publishing**

Collect and Store: Software version control



Collect and Store: Software version control



Collect and Store: Software version control

https://git-scm.com/

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

Workstation

Your own server

Gitea

Forgejo

GitLab

Internet

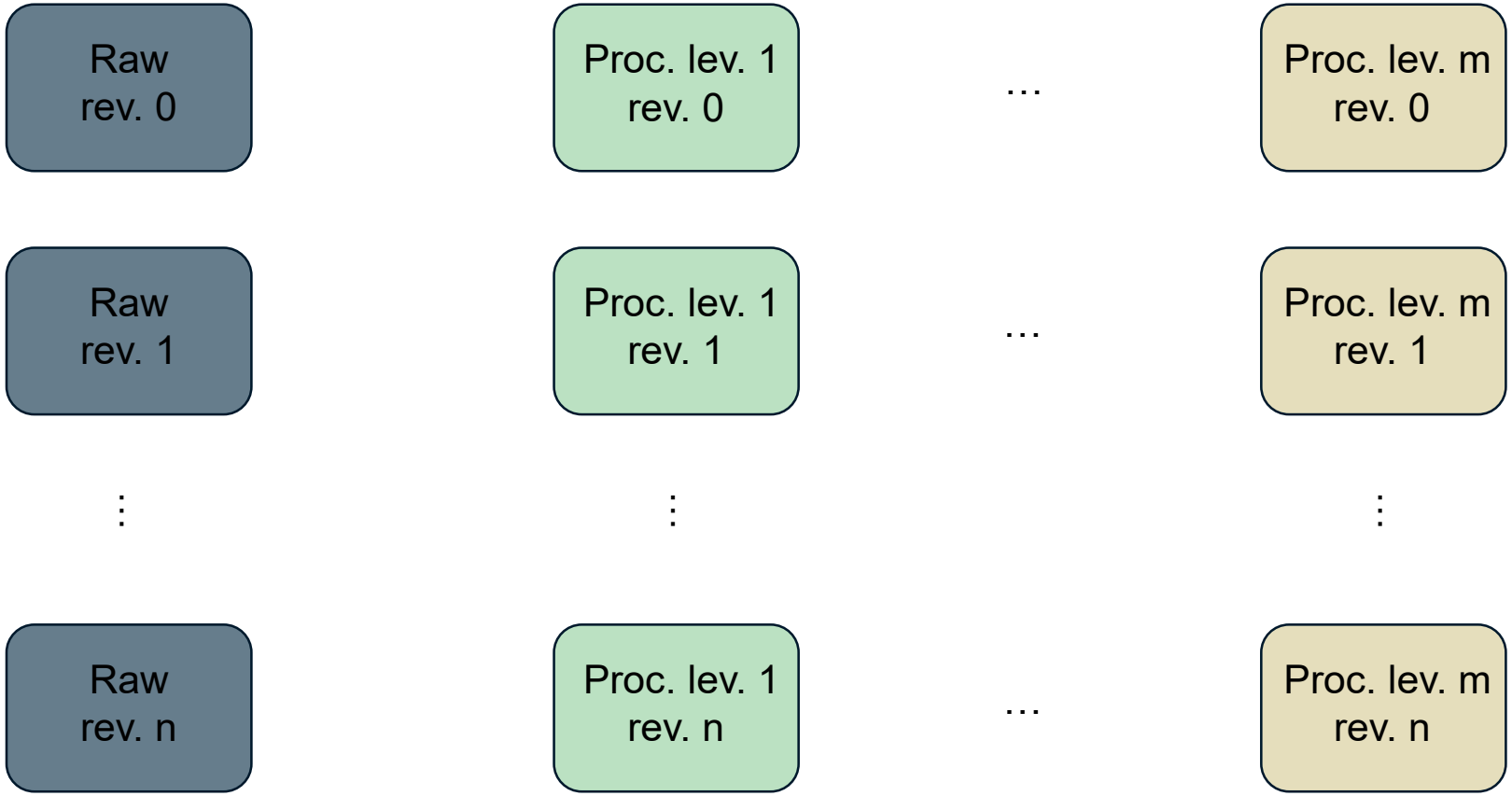
Codeberg

GitLab

GitHub

Bitbucket

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*)

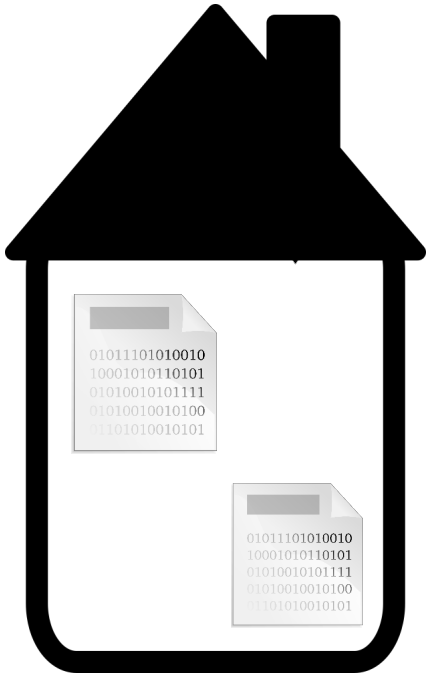
- Recommendation: Stick to a list of defined terms (controlled vocabulary) and don’t use synonyms to describe the same object (e.g. picture or image)

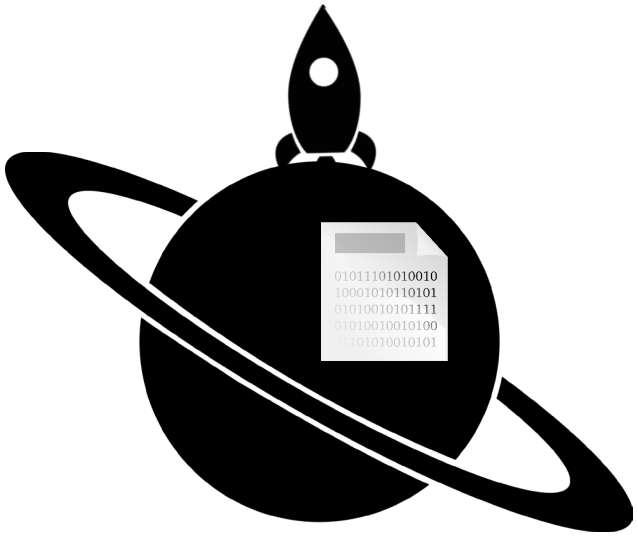
Collect & Store: README File

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

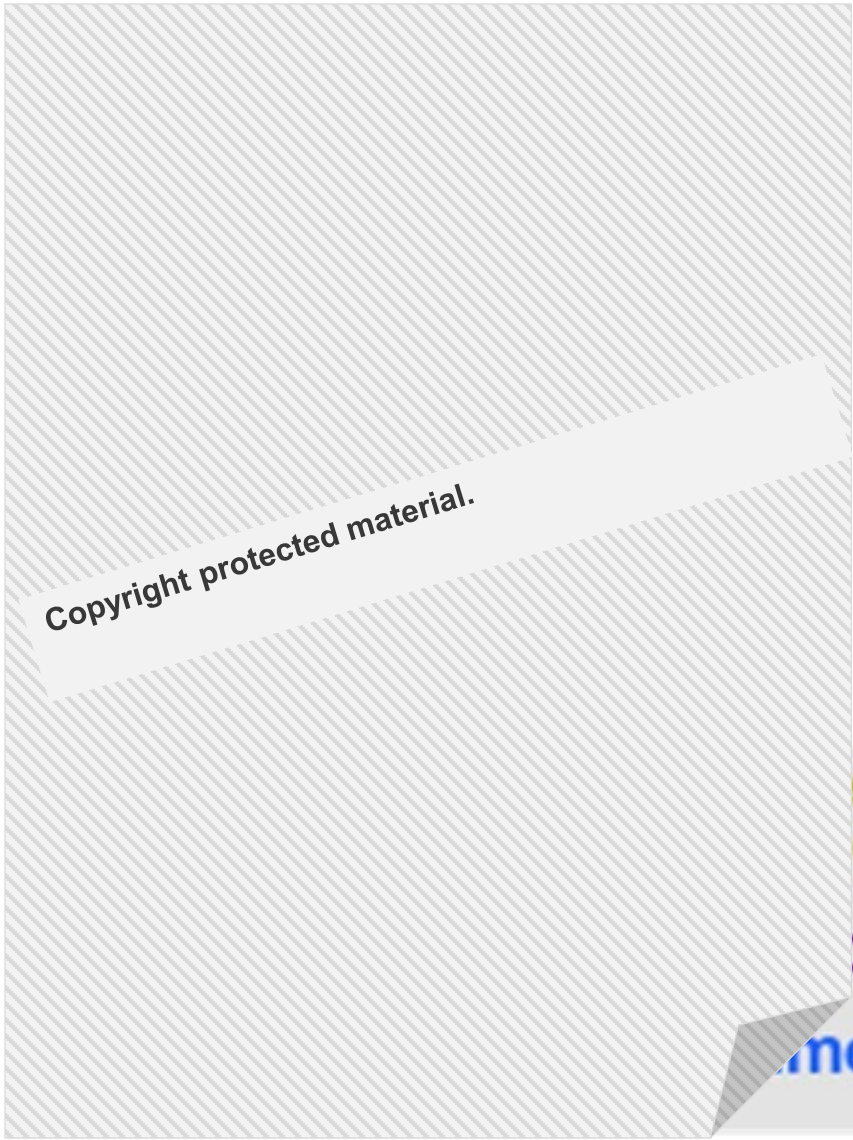
- **Relates to identified or identifiable person**
- **Solutions (https://dmlawtool.ccdigitallaw.ch/) :**
  - Identity irrelevant -> anonymisation
  - Identity relevant -> Ask for consent

    -> Pseudoanomization

    -> Manage access rights

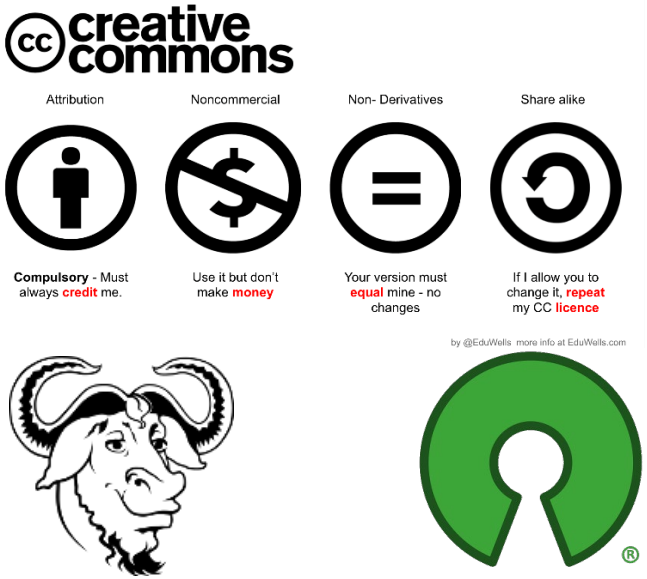
    -> Ability to address

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



Evaluate & Archive: Data Protection

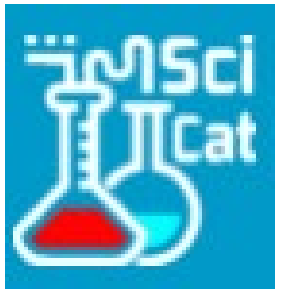
- Processed Data has copyright according to Swiss law
- Use CC licences when publishing factual data on data repositories (ideally CC 0)
- For software use licences specifically designed for software:
- Free Software (Open Source) licences like GPL, Apache, BSD and MIT.
- Exceptions! If you collaborated 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.

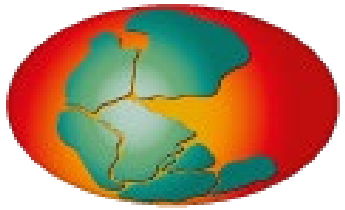


Share & Disseminate

Share & Disseminate: The Choice of Data Repository

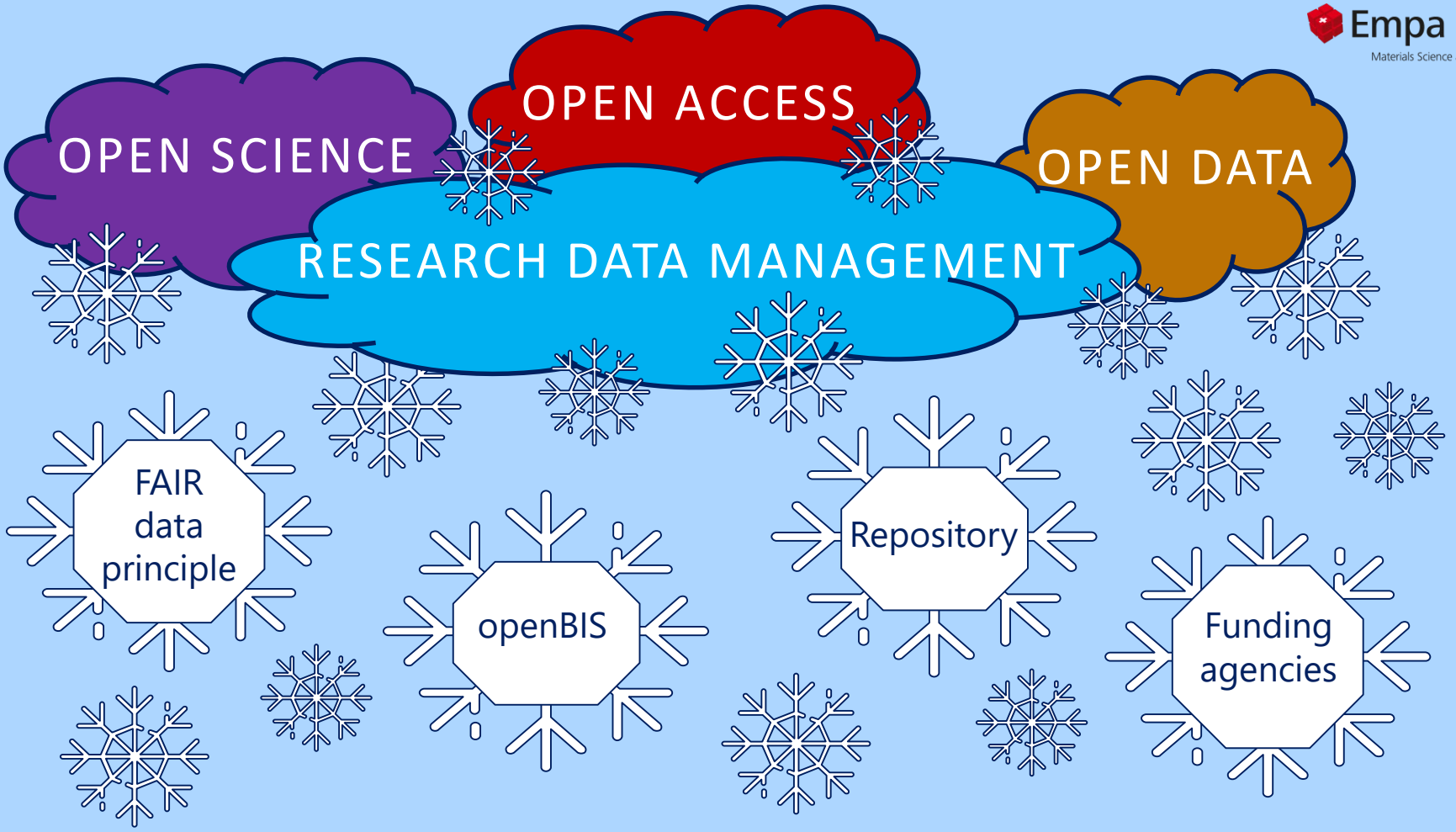
ERIC open

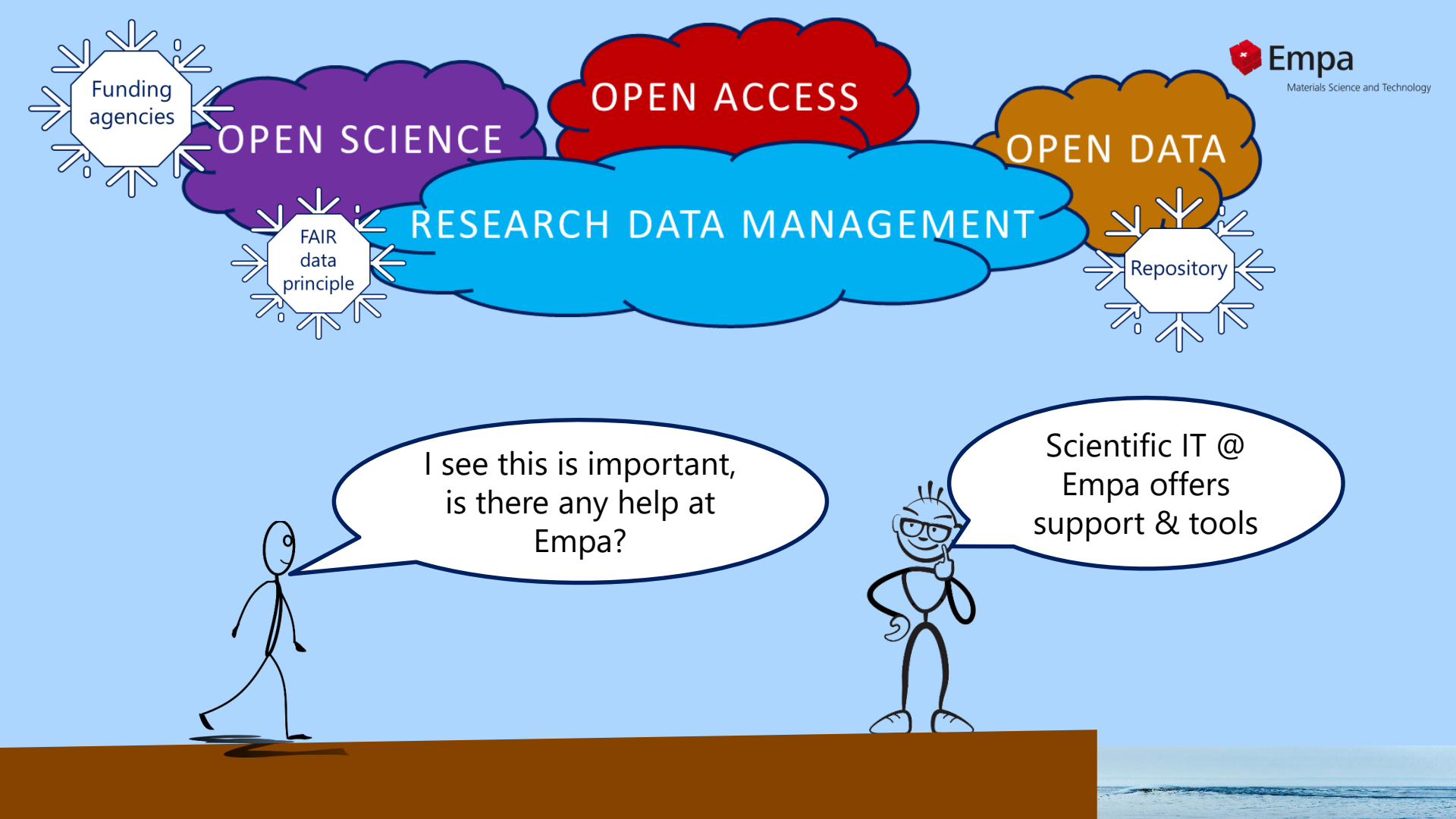


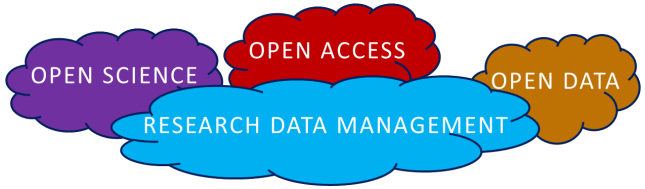


For alternatives: https://www.re3data.org/

RDM Services and Support at Eawag and Empa


Empa
Materials Science and Technology
OPEN SCIENCE
OPEN ACCESS
OPEN DATA
RESEARCH DATA MANAGEMENT
FAIR data principle
openBIS
Repository
Funding agencies


Empa
Materials Science and Technology
Funding agencies
OPEN SCIENCE
OPEN ACCESS
OPEN DATA
FAIR data principle
RESEARCH DATA MANAGEMENT
Repository
I see this is important, is there any help at Empa?
Scientific IT @ Empa offers support & tools



DigitalScience@Empa
Intranet Plattform

Empa
Materials Science and Technology

On Intranet main page

RESEARCH INFOS

Find here a list of Research Calls

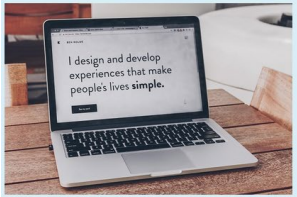
Find here a list of Research Awards

DigitalScience@Empa - find here:

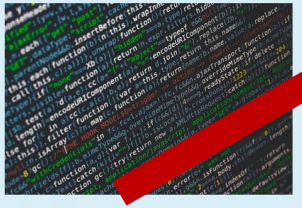
Tools and Platforms, Data Science, Modeling & Simulation, Open Science, Events & Trainings, Community

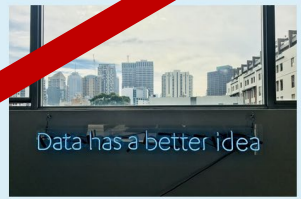
DigitalScience@Empa

Welcome to the platform about digitalization topics at Empa.



Tools & Platforms


Scientific IT



Modeling & Simulation


Open Science


Events & Training


Community

Our new documentation pages:

Scientific IT team

Computing & HPC

Data Management & openBIS

Data Science

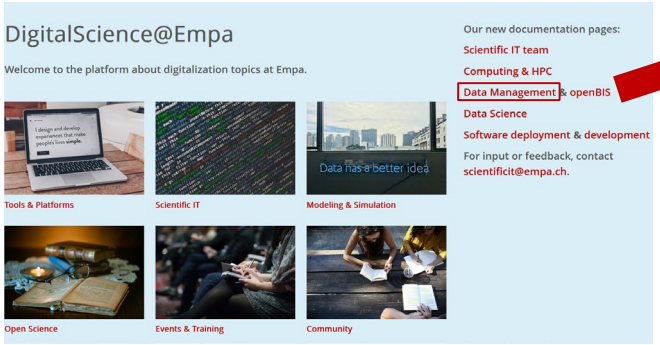
Software deployment & development

For input or feedback, contact scientificit@empa.ch.

https://www.empa.ch/group/s909/overview

RDM Services & Infrastructure @ Empa

Empa
Materials Science and Technology



How to manage your research data?

Check our information and support pages:

1. **RDM guidlines** to see the advantages of managing your data and possible horror stories if you do not: here
2. **openBIS** at Empa helps to manage your research data digitally, which includes save data storage, archiving and publishing in data repositories: here
3. **Data Management Plan (DMP)** to fulfill the requirements of funding agencies: here
4. **Open Access (OA)** to fulfill the requirement of publishing your paper openly: here
5. **Open Data Licenses** learn how licenses help to publish your research data openly: here
6. **Support & training** information here for:
   - openBIS & Data Management
   - DMP templates & how to get funding for RDM costs
   - Open Access by Lib4RI
   - Scientific IT weekly hours for Software development & Data Science questions
   - Python tutorial twice a year by Scientific IT

- ❑ Support for RDM
- ❑ SNF Data Management Plan template
- ❑ RDM guides, Best practice guide
- ❑ Handling of software licenses
- ❑ Open Access @ Empa
  - ❑ Policy
  - ❑ Publication fund

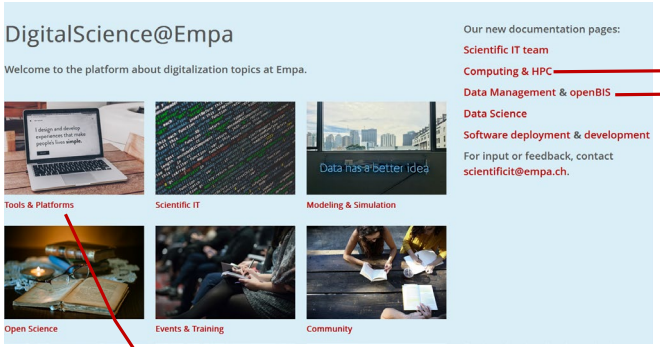
Text snippets available

SNF Data Management Plan Template

- ❑Sections analog to the SNF DMP
- ❑Different categories for
  - ❑Sensitiv data
  - ❑Using the data management system openBIS
  - ❑Without using openBIS
- ❑Description of backup system at Empa
- ❑Recommendation of FAIR repository

RDM Services & Infrastructure @ Empa

Empa
Materials Science and Technology



Data Management System @ Empa

openBIS, the **E**lectronic **L**aboratory **N**otebook (ELN) and **I**nventory **M**anagement **S**ystem (LIMS) at Empa, which enables easy connection

- ✓ **Zenodo** – data repository according FAIR principles
- ✓ **Jupyter Hubs** – for programming in Python, R etc.
- ✓ **Longterm Archiving** – of research data at ETH

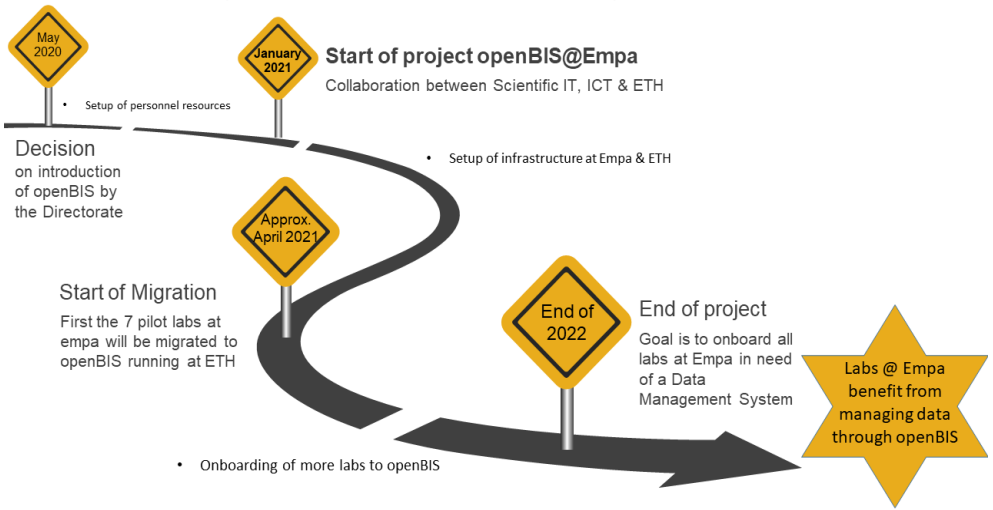
Collaboration platform – Code versioning

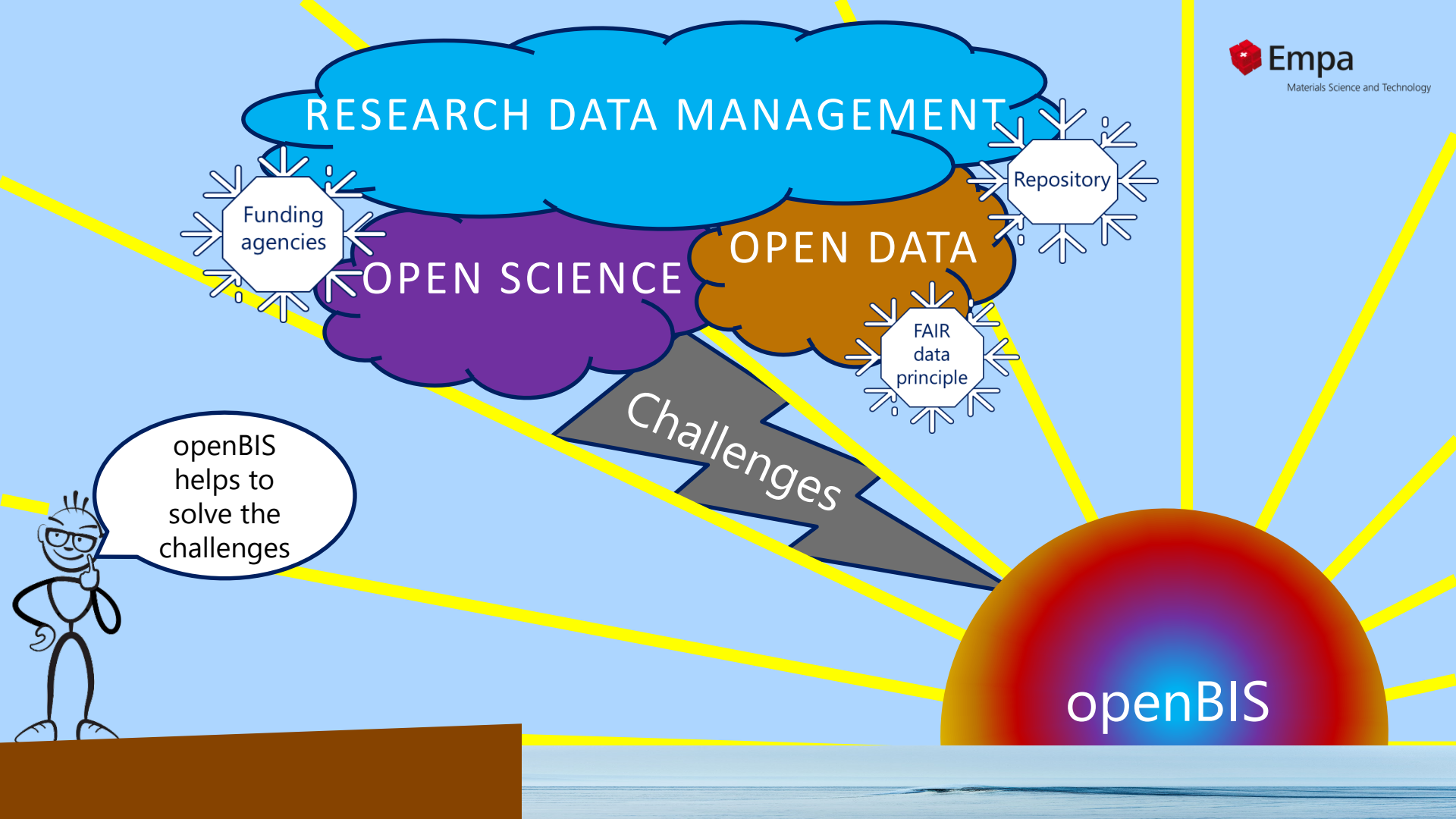
GitLab @Empa

High performance computing (HPC)

Provision of researchers with technical support regarding HPC-related common problems and long-term projects.

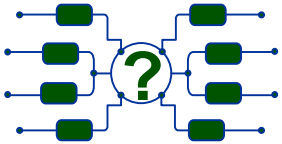
Timeline of openBIS introduction@Empa




RESEARCH DATA MANAGEMENT
Funding agencies
OPEN SCIENCE
OPEN DATA
Repository
FAIR data principle
Challenges
openBIS helps to solve the challenges
openBIS

Landscape – Challenges – Solution

Empa
Materials Science and Technology



paper notebook

competition

regulations

More data

data loss
missing features

need of
collaboration and
visibility

publication of data
data management plan

handling of data
storage cost

RDM &
Open Science

reliability
reproducability

interoperability
accessability
documentability

traceability
reusability
archiving

unified solution

Data Management System for Empa

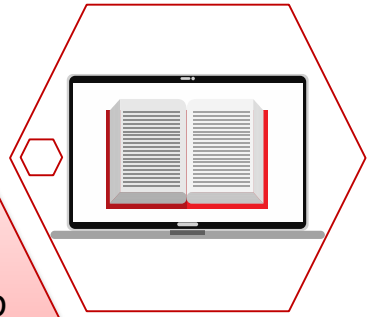
openBIS

openBIS - What is it?

Empa
Materials Science and Technology

Inventory of
- *Samples*
- *Materials*
- *Reagents*
- *SOPs*
- *Equipment*

Electronic Lab Notebook



Description of
- *Experiments*
- *Measurements*
- *Processes*
- *Data analysis*

Data Management

Inventory Management

Storage of research data connected to the experiments described in the electronic lab notebook

A data management system – Connecting lab inventory, research data & lab notebook in 1 tool

openBIS - What is it?

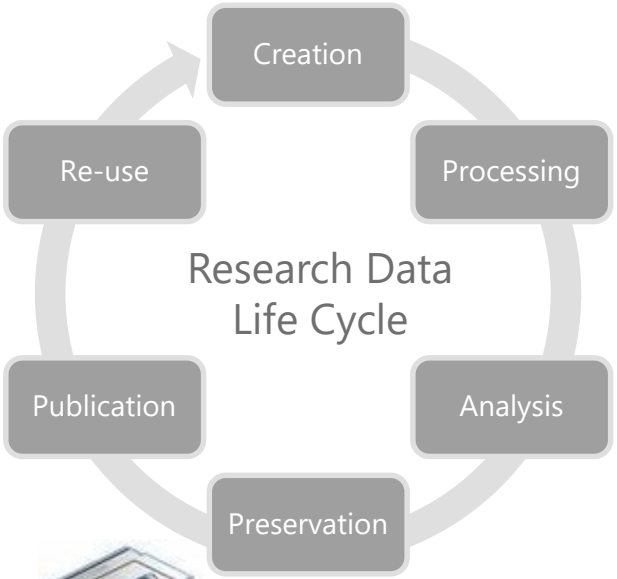
Empa
Materials Science and Technology

Experiment description in electronic lab notebook

openBIS

zenodo

Integration with Data Repositories



Inventory of Protocols/SOP's

MATLAB

jupyterhub

Integration with analysis tools

Archiving

A data management system – Covering most of the data life cycle in 1 tool

openBIS - What is it?

Empa
Materials Science and Technology

Methods & Materials recorded in electronic lab notebook

Experimental data imported into inventory

Barcode reader

Data immunability

Big Data Link

Data collection & documentation

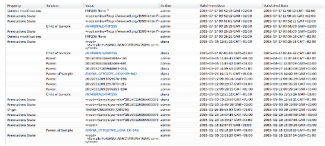
Data sharing & reuse

Ethics, legal & security issues

DMP
Data Management Plan

Link to DMP support with template

Data storage & preservation

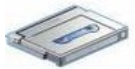


Audit trail

User rights management

Integration with FAIR Data Repositories

Import/Export



Back-up & Archiving

Use the **text snippets** for openBIS in the DMP template

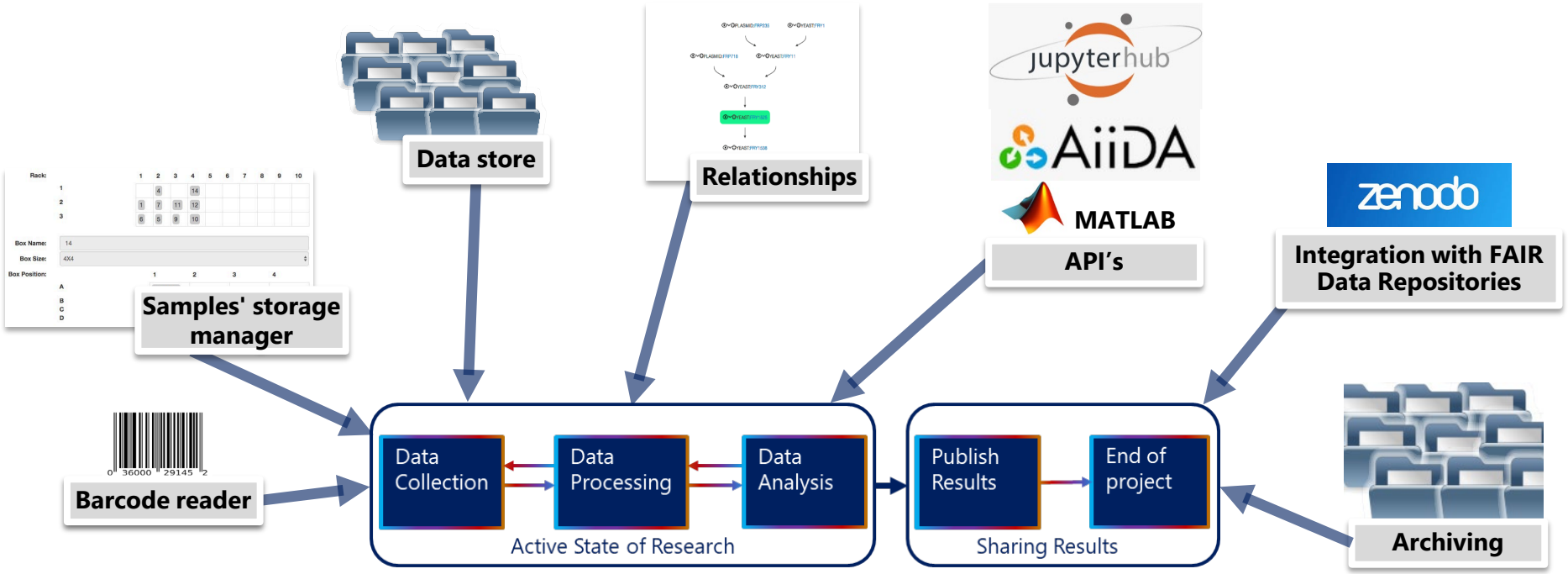
A data management system – Solving DMPlan requirements of funding agencies with 1 tool

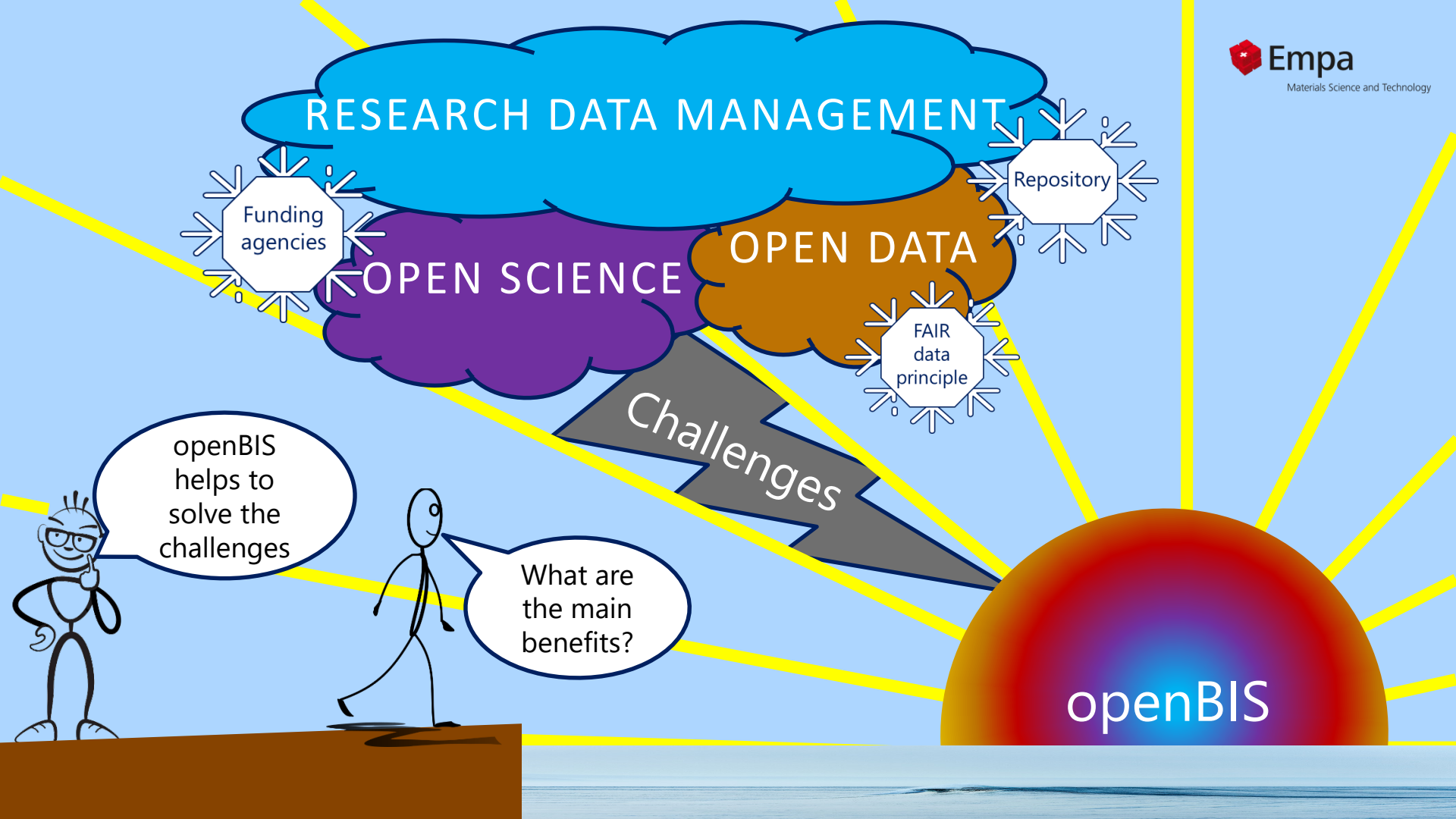
openBIS

- Open source software developed by ETH since 2007
- ELN/LIMS system
  Electronic **L**ab **N**otebook & **L**aboratory **I**nformation **M**anagement **S**ystem

Empa
Materials Science and Technology

For managing research data from "bench" to publication in a central storage

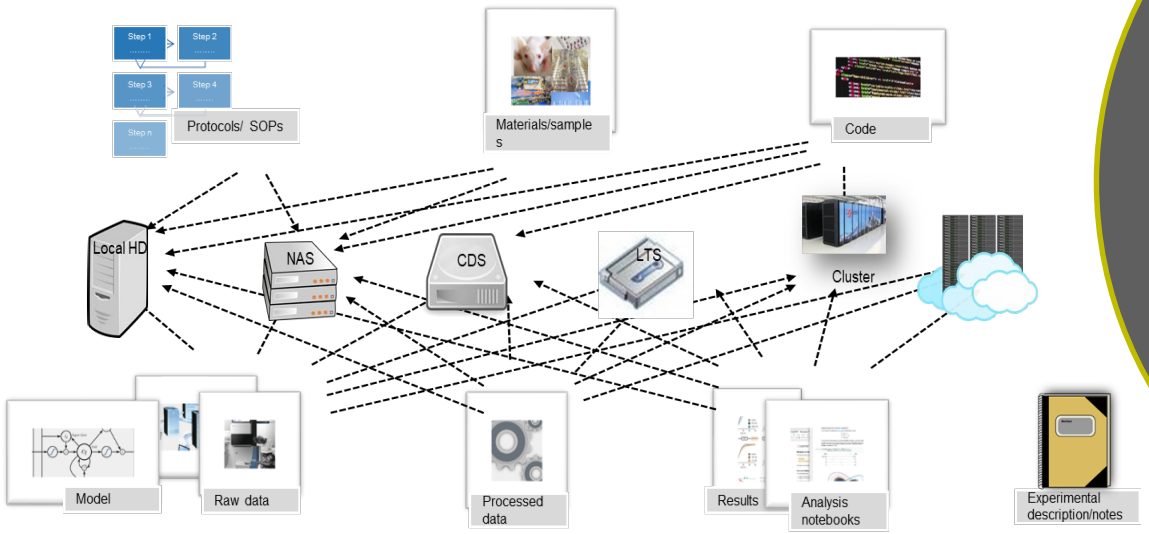



RESEARCH DATA MANAGEMENT
Funding agencies
OPEN SCIENCE
OPEN DATA
Repository
FAIR data principle
Challenges
openBIS helps to solve the challenges
What are the main benefits?
openBIS
Empa
Materials Science and Technology

openBIS - How does it help?

Empa
Materials Science and Technology

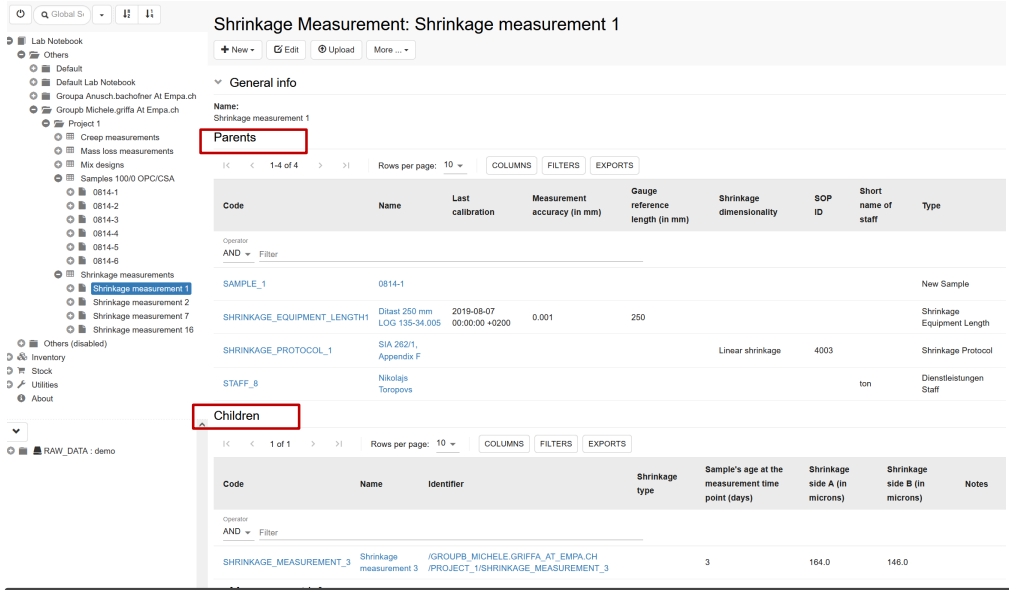
A common scenario
The "*data spread*"

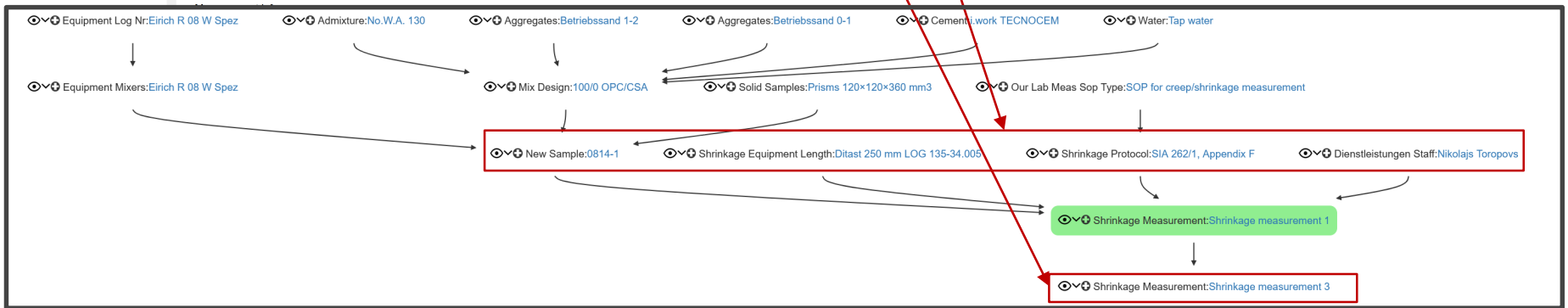


The ideal solution
Having everything connected in 1 place

openBIS - Hierarchy graph

Empa
Materials Science and Technology





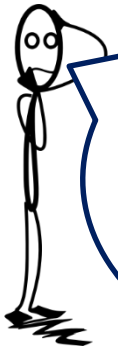
openBIS

A data management system
connecting
lab inventory, research
data & lab notebook

openBIS benefits

Empa
Materials Science and Technology

1. **Prevention of loss of research data & knowhow via structured documentation & storage**
2. **Easy & automatic data archiving over a long period**
3. **Easy & fast connection to repository Zenodo**



Challenges at Empa

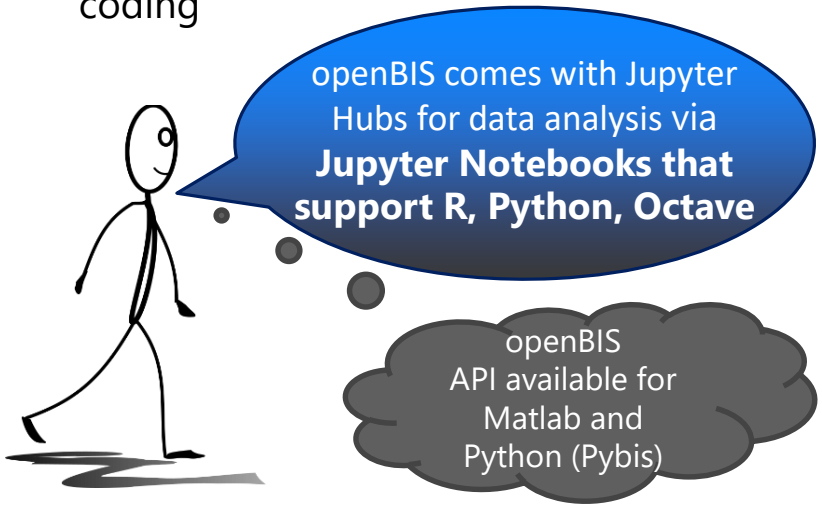
1. **Data loss** due to change of personnel & no proper documentation of data
2. **No archiving solution** easy accessible & reliable
3. **No repository** available

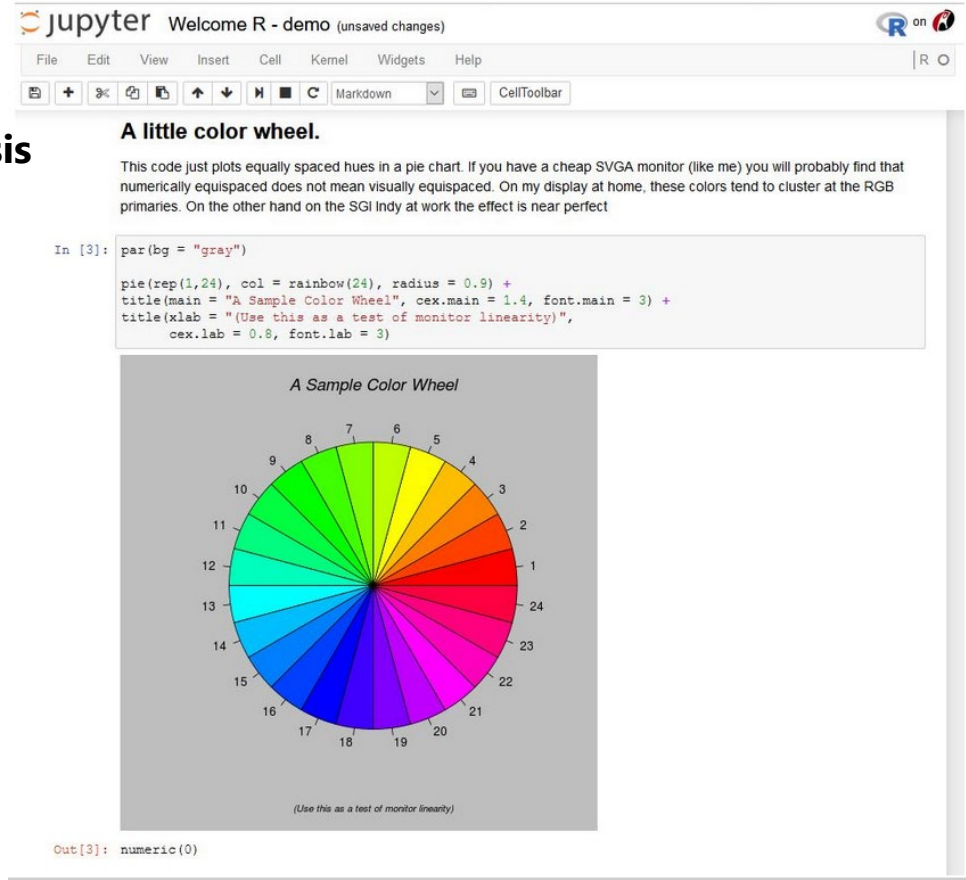
- **Without openBIS** you need to solve these challenges on your own
- **Without proper data management** you risk loosing funding money

Data Science with openBIS

Empa
Materials Science and Technology

- **Jupyter notebooks combine code, documentation & outputs** like plots, images, videos etc.
- Useful for **interactive/exploratory data analysis** and **reproducability**
- **Easy sharing** of code with documentation and results
- Like a **modern lab notebook** for reproducible coding



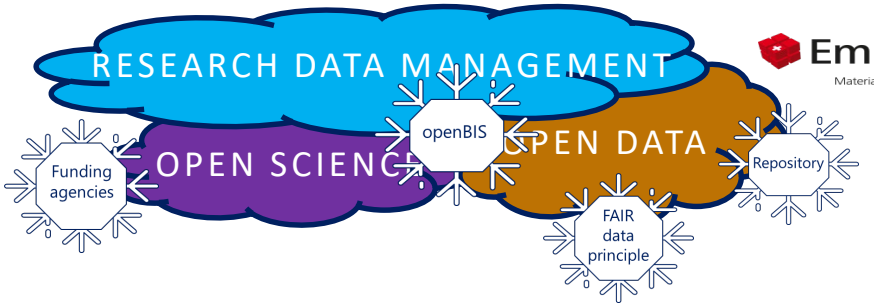


OPEN ACCESS

OA POLICY OF EMPA

Golden way

Green way



Empa
Materials Science and Technology

OA support

Lib4RI
Library for the Research Institutes within the ETH Domain: Eawag, Empa, PSI & WSL

DMP support

Scientific IT
RDM@empa.ch

DATA MANAGEMENT PLAN (DMP)

REPRODUCIBLE

RELIABLE

openBIS

STORING
BACK UP
SECURING

COLLECTING
ORGANIZING
DOCUMENTING

openBIS

F.A.I.R.

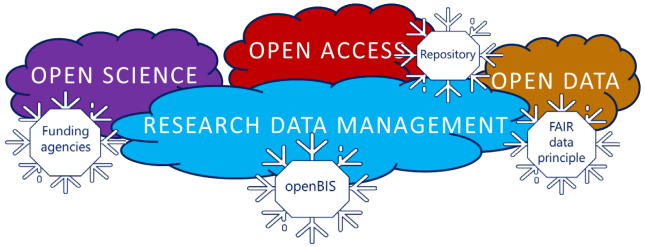
ARCHIVING
SHARING

Repository Zenodo

Reproducible Data Analysis

REUSABLE

Solutions @ Empa for RDM & Open Science



DigitalScience@Empa

Intranet Plattform https://www.empa.ch/group/s909/overview

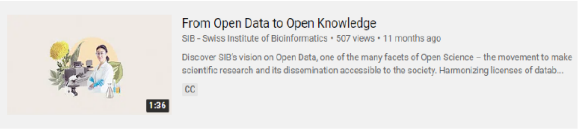
Empa
Materials Science and Technology

Support

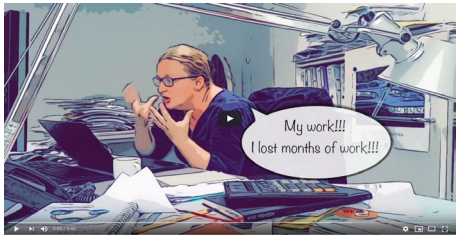
scientificit@empa.ch

https://scientificit.empa.ch/

Videos



YouTube



https://youtu.be/t_rEXpfCTrg
https://youtu.be/tFWd2M2OXwQ
https://youtu.be/6kHGbbdFuDE
https://youtu.be/LCZijZP916o
https://youtu.be/NdklWkRi-ZQ

eawag aquatic research ooo RDM

Eawag **R**esearch Data Management

RESEARCH DATA MANAGEMENT @ EAWAG

— SERVICES & SUPPORT —

Christian Foerster <christian.foerster@eawag.ch> — 2022-12-08

eawag aquatic research ooo RDM

Eawag Research Data Management

WHAT I DO?

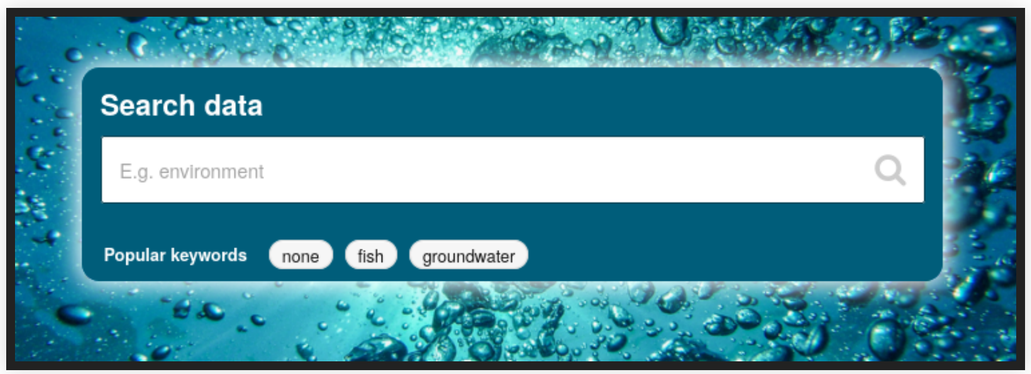
- Data management
- Software development
- DevOps
- Improve ERIC
- Teaching best practices

eawag aquatic research ooo RDM

Eawag Research Data Management

DATA PRESERVATION

ERIC internal



Publishing and archiving of research data

an Eawag-flavored hands-on guide (v1.0)

Eawag
Überlandstrasse 133
P.O.Box 611
8600 Dübendorf
Switzerland
Phone +41 (0)58 765 55 11
Fax +41 (0)58 765 50 28
https://www.eawag.ch

eawag aquatic research ooo

Directive No. 19-03

Dübendorf, January 1, 2019

Directive on the archiving of research data at Eawag

Eawag Research Data Institutional Collection

- Development, Maintenance
- No review, quality control or curation
- Support
  - data organization, formats and annotation
  - process automation

eawag aquatic research RDM
Eawag Research Data Management

DATA PUBLICATION

ERIC open

- Development, Maintenance
- DOI reservation & registration
- Support reg. workflow in sync with article review & publishing
- (Meta)data dissemination, interlinking (ORCIDs, article-DOIs, ...)
- Review, quality control and curation

eawag aquatic research RDM

Eawag Research Data Management

CONSULTING

Data Management Plans

- Guide: Eawag SNF / SNSF Data Management Plan Guide
- DMP reviews

Data management planning

- Sampling, surveys
- Data transmission/transport
- Backup strategy, data safety
- Data security, encryption
- Legal issues, licensing, personal data, anonymization

eawag aquatic research RDM

Eawag Research Data Management

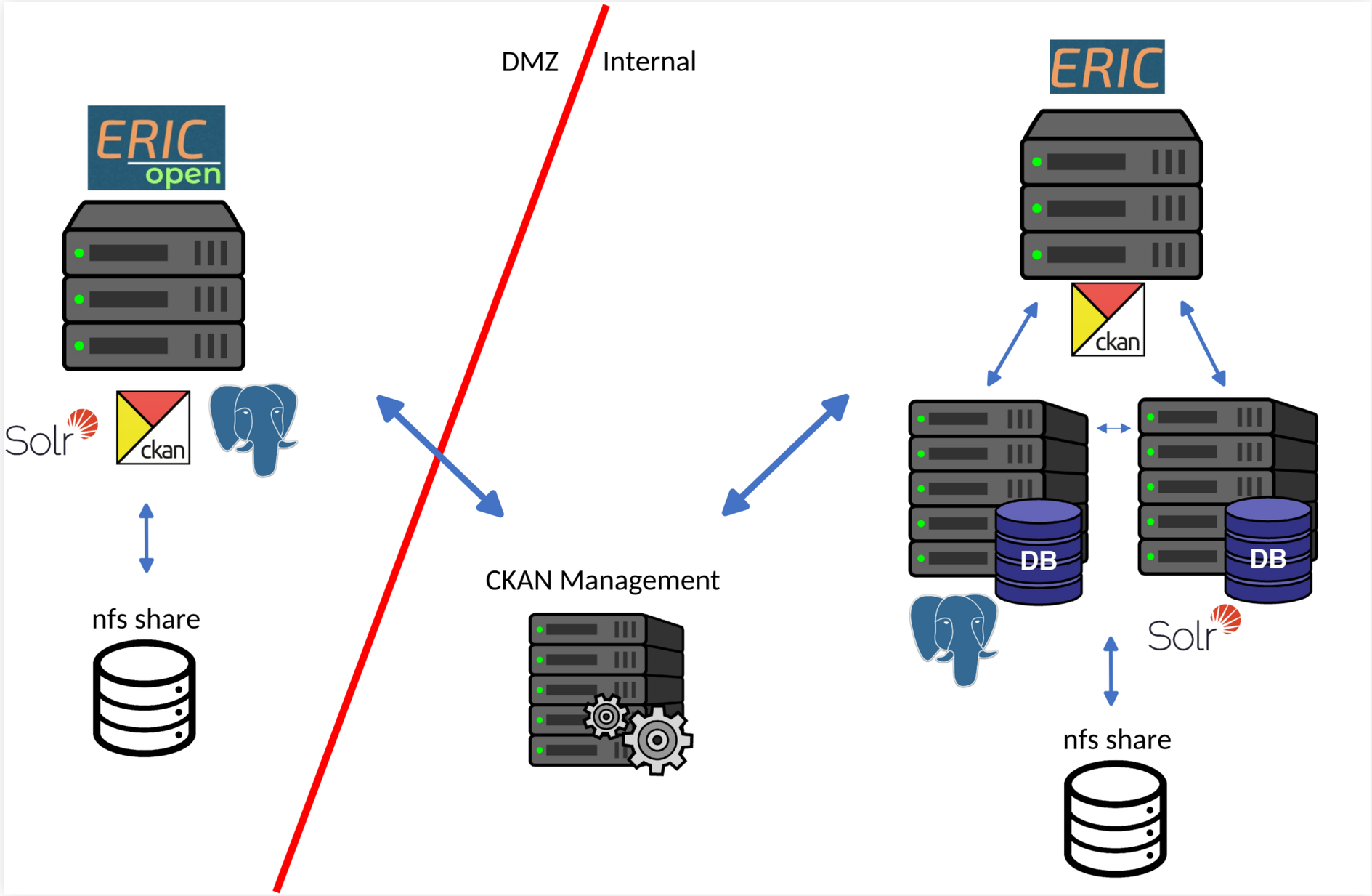
ERIC BASICS

- Comprehensive Knowledge Archive Network (CKAN)
- Findable Accessible Interoperable Reusable (FAIR)
- SNF approved
- Secure infrastructure
- RDM knowledge base
- FAQ (here)
- Eawag branding
- Meta data schema from DataCite
- All extensions on GitHub in the eawag-rdm organization

eawag aquatic research RDM

Eawag Research Data Management

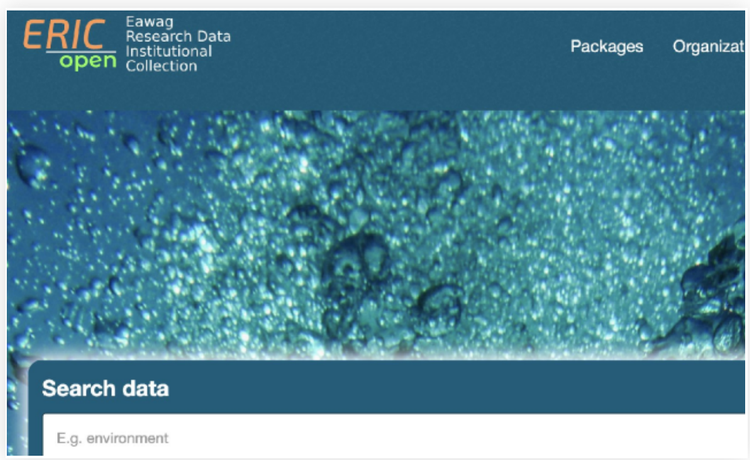
INFRASTRUCTURE



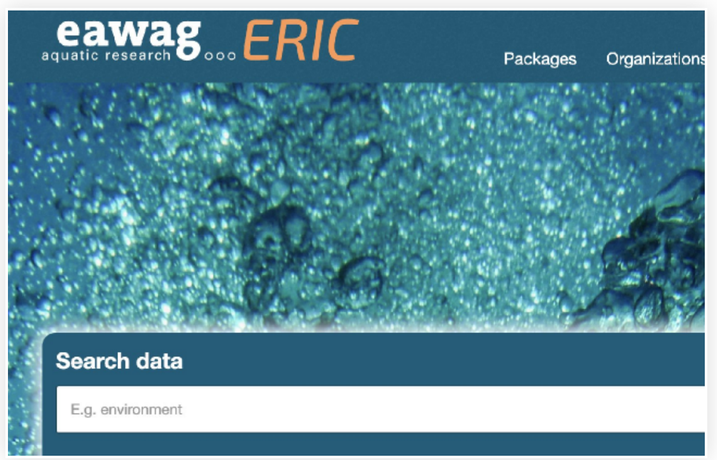
eawag aquatic research ooo RDM

Eawag **R**esearch Data Management

OPEN VS INTERNAL



- opendata.eawag.ch
- Openly accessible
- Strict checking
- Pure read data portal
- Immutable datasets



- data.eawag.ch
- Eawag internal
- Unchecked data
- Read/write platform
- Mutable datasets

eawag aquatic research ooo RDM
Eawag Research Data Management

PUBLICATION WORKFLOW

Good practices throughout

- Keep your data save and clean (minimal guide)
- Version your data (renku)
- Use version control (git, ...)

eawag aquatic research ooo RDM
Eawag Research Data Management

PUBLICATION WORKFLOW

First time ERIC use

- Log in with your Eawag credentials
- Contact the data manager of your department to get editor privileges

Upload to ERIC internal

- Create a package/group
- Upload your neat and organized data
- Write rdm@eawag.ch (naming your package)

eawag aquatic research ooo RDM
Eawag Research Data Management

PUBLICATION WORKFLOW

RDM initial feedback

- Reserved DOI for your publication
- Data publication (checklist)

Iterative package improvement

- Review of checklist
- Rigorous checks (data protection, checksums, data usage, basic code review, check links, ...)
- Feedback

We are aiming for an iteration of 1! ;-)

eawag aquatic research ooo RDM
Eawag Research Data Management

PUBLICATION WORKFLOW

You should

- Add your current code
- Add links to repositories, related packages
- Provide accurate and plentiful meta data

You should not

- Add your paper to the data package
- Incorporate Eawag specific information like paths on a network share
- Violate data protection guidelines

PUBLICATION WORKFLOW

Now the data can be made public!

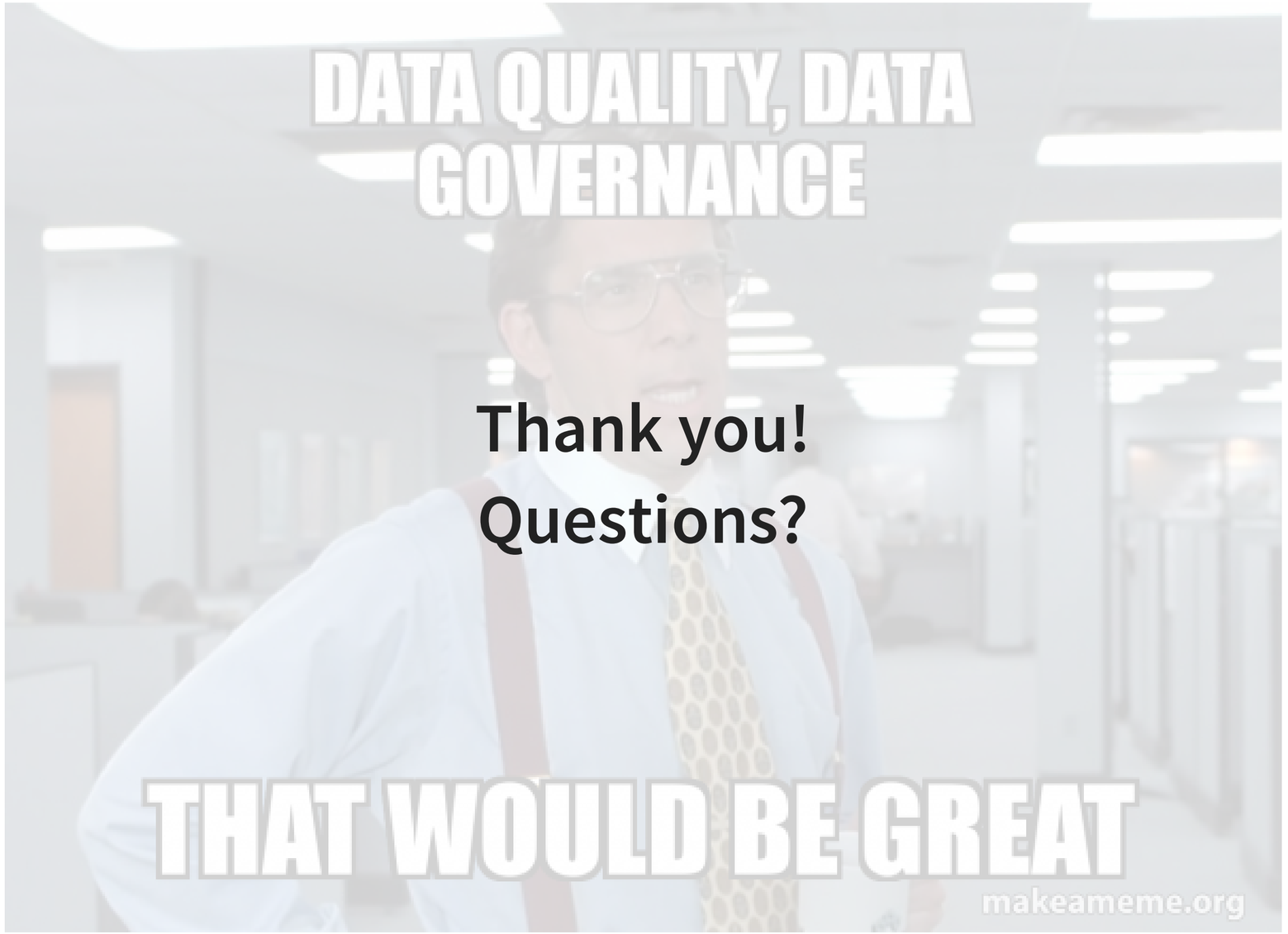
eawag aquatic research ooo RDM
Eawag Research Data Management

FUTURE PLANS

- Update ERIC
- Extend functionality
- Improve and standardize documentation
- Better support for other parts of the DLC like "life" database clusters

eawag aquatic research RDM

Eawag Research Data Management



Thank you!

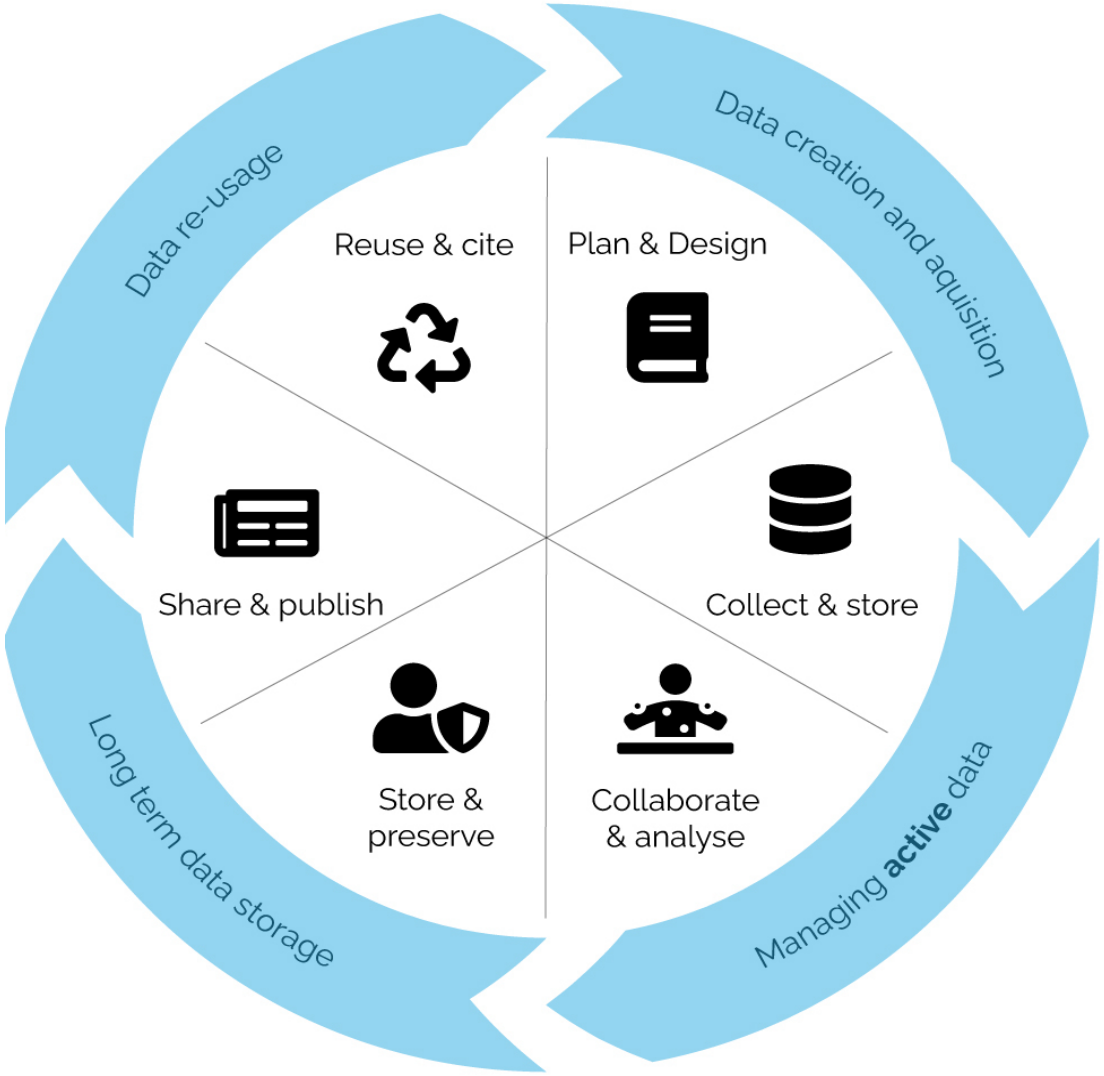
Questions?

Plan & Design:

Data Management Plan (DMP)

Plan & Design: Why?

Bibliothèqe de l'EPFL, «RDM Horror stories | Episode 2 – Stranger Data Things», 11th February 2020. https://bit.ly/3qPWMlS



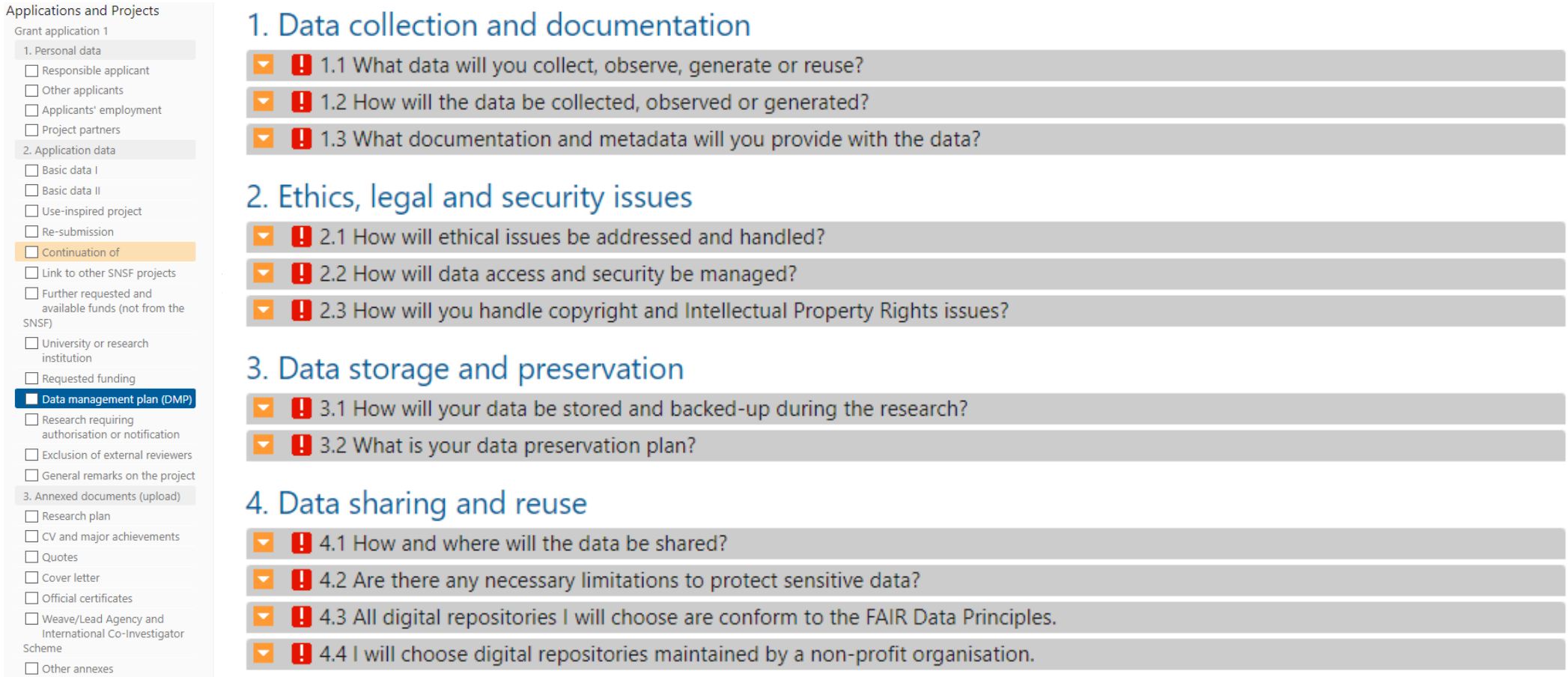
DMP

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.