

02.11.2023

Research Data Management – The Basics

Cantini, Federico
Felder, Fabian
Iosifescu, Ionut

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



Ionut Iosifescu

- Software Engineer
- Technical coordinator and product owner of EnviDat

Who are you and why are you here?

Copyright protected material.

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

Learning Aims

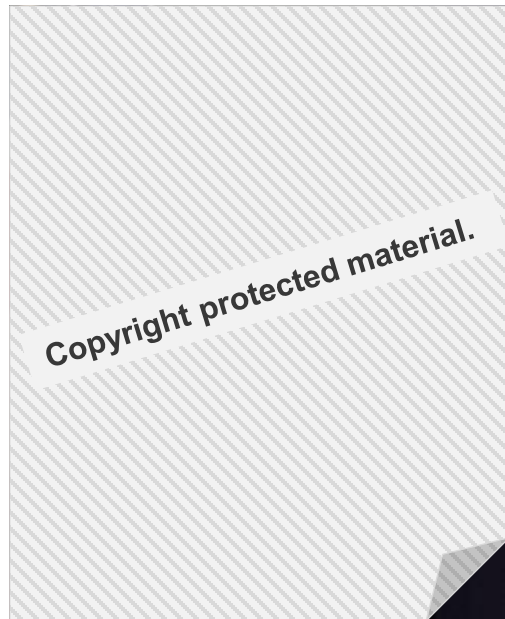
- Life cycle of research data
- Adequate metadata documentation for your code and data
- Storing and publishing data
- Using OpenBIS (ELN) and writing Data Management Plans (DMP)

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 - 10.05
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

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



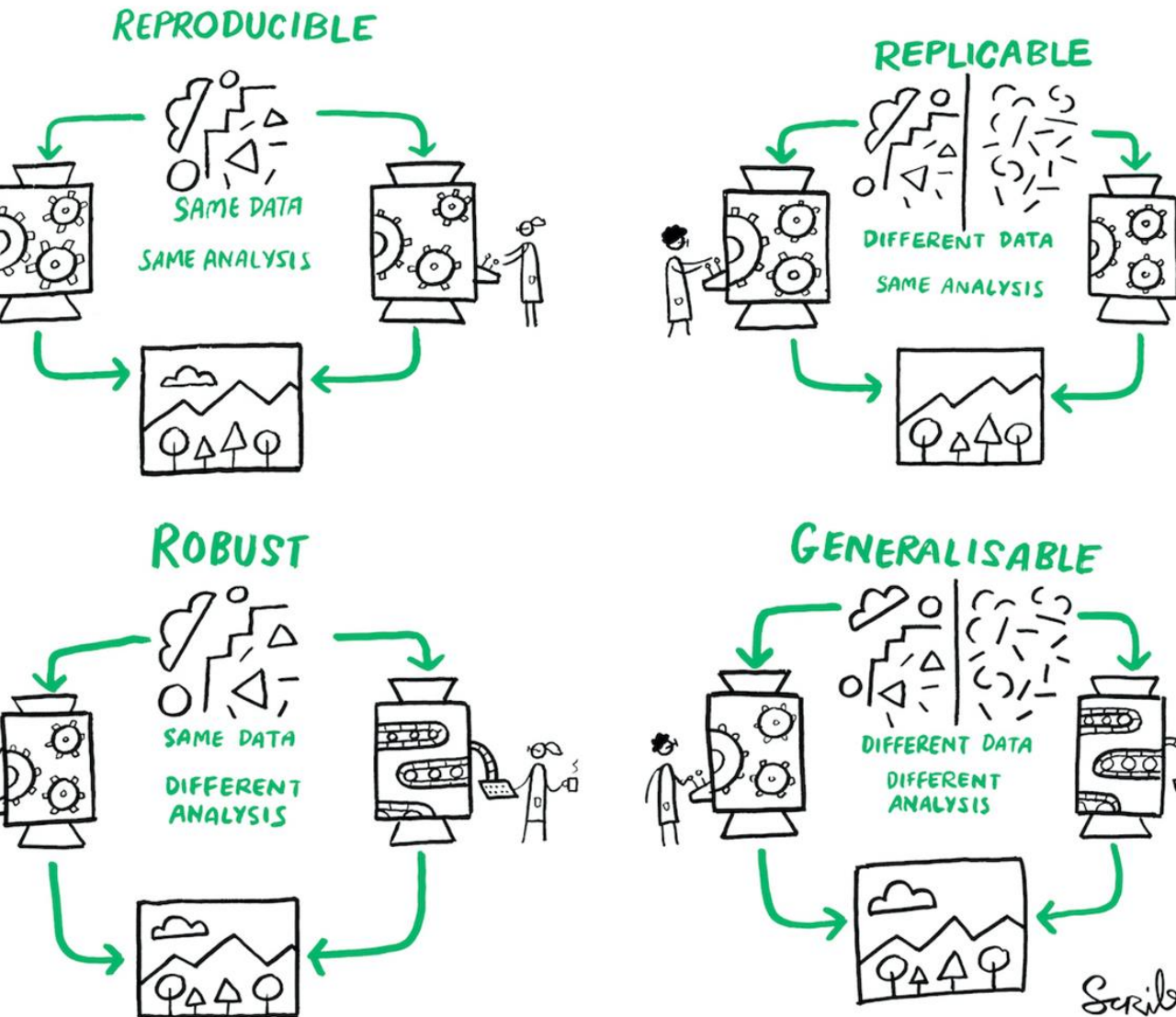
Source: www.fosteropenscience.eu/project

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](https://doi.org/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

Interoperable

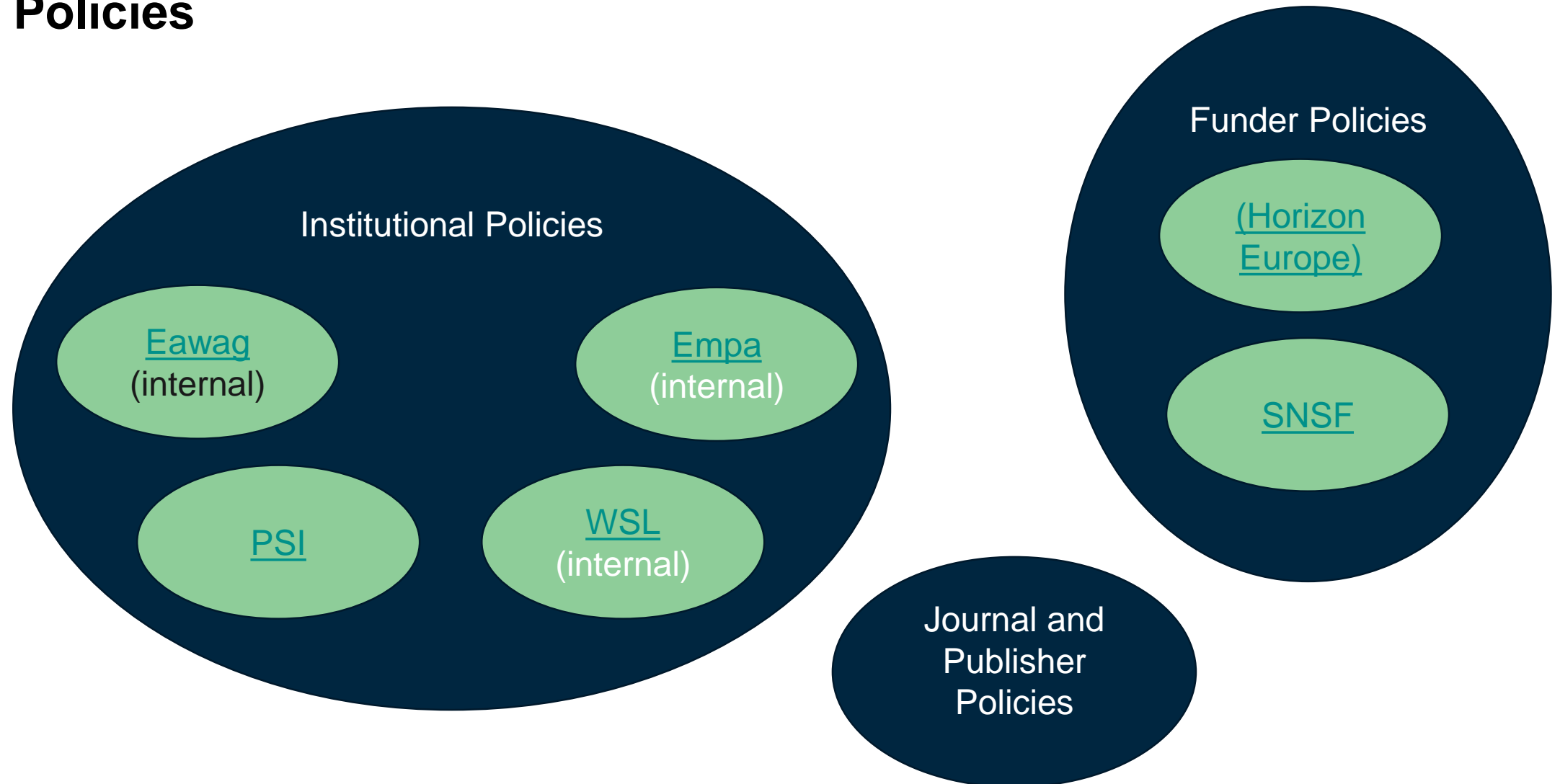
- I1 (Meta)data use formal, accessible, shared, and broadly applicable language for knowledge representation
- I2 (Meta)data use vocabularies that follow FAIR principles
- I3 (Meta)data include qualified references to other (meta)data

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




Policies

Compliance




Project Manager/
Group Leader

DMP

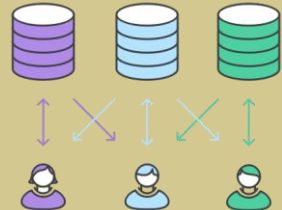


Required for
Funders

As open as possible,
as closed as necessary.

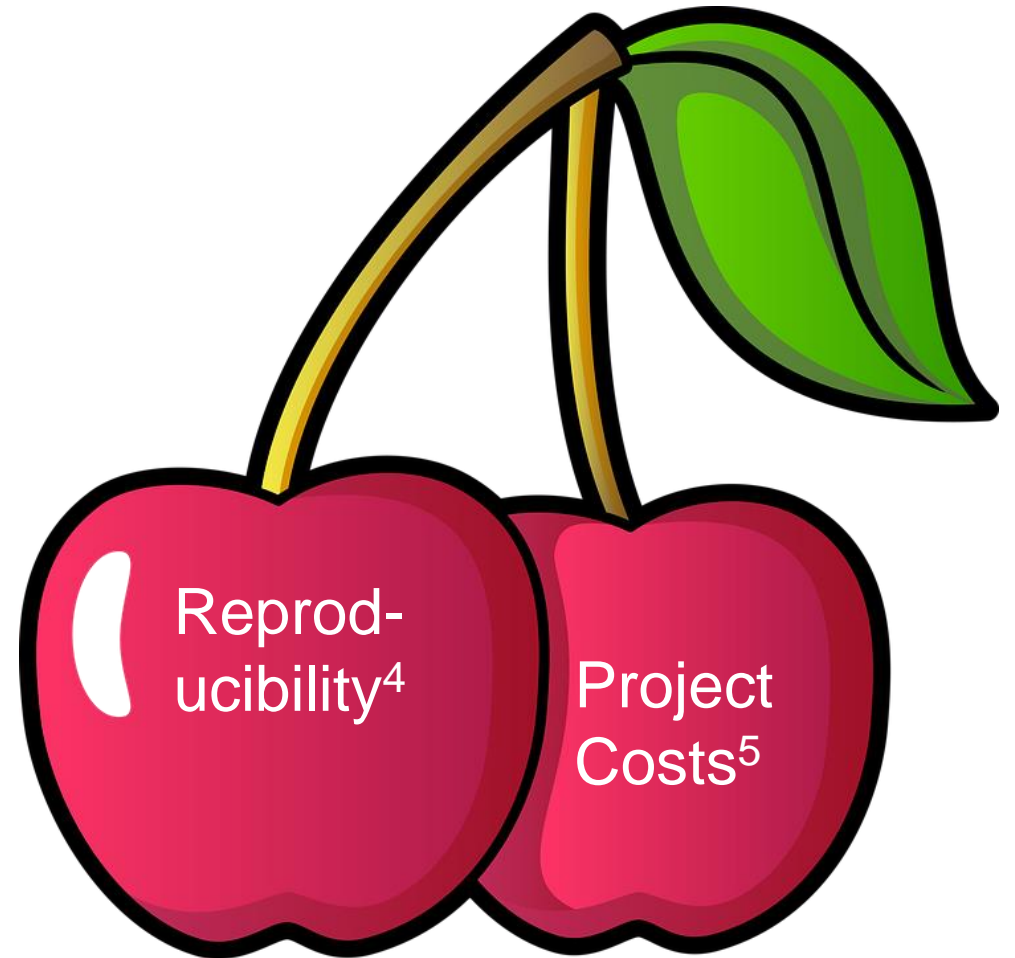


Repositories



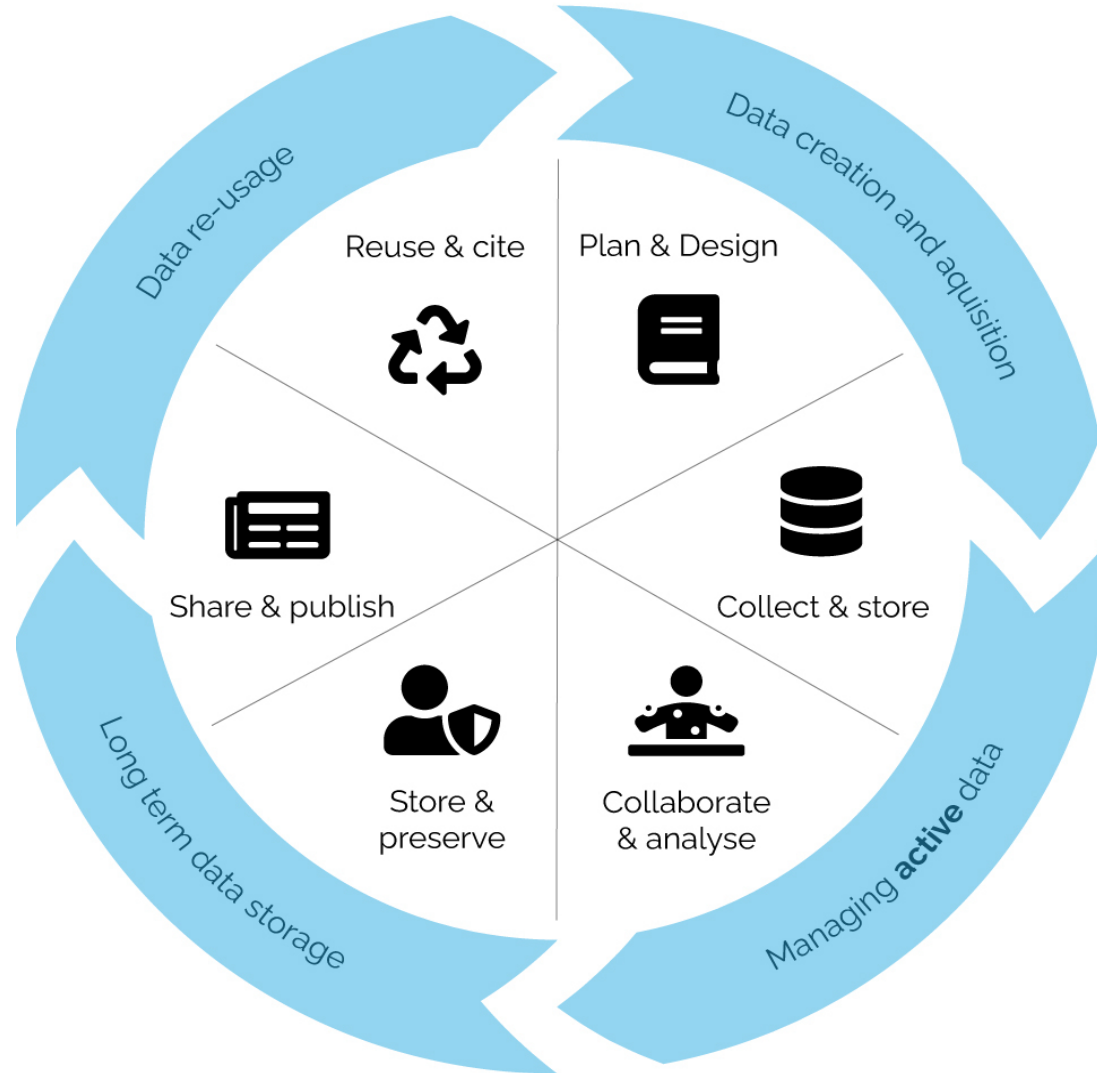
FAIR and non-commercial

Policies



Research Data Life Cycle

Research Data Life Cycle



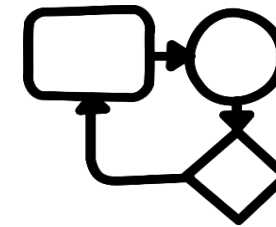
Collect & Store

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...

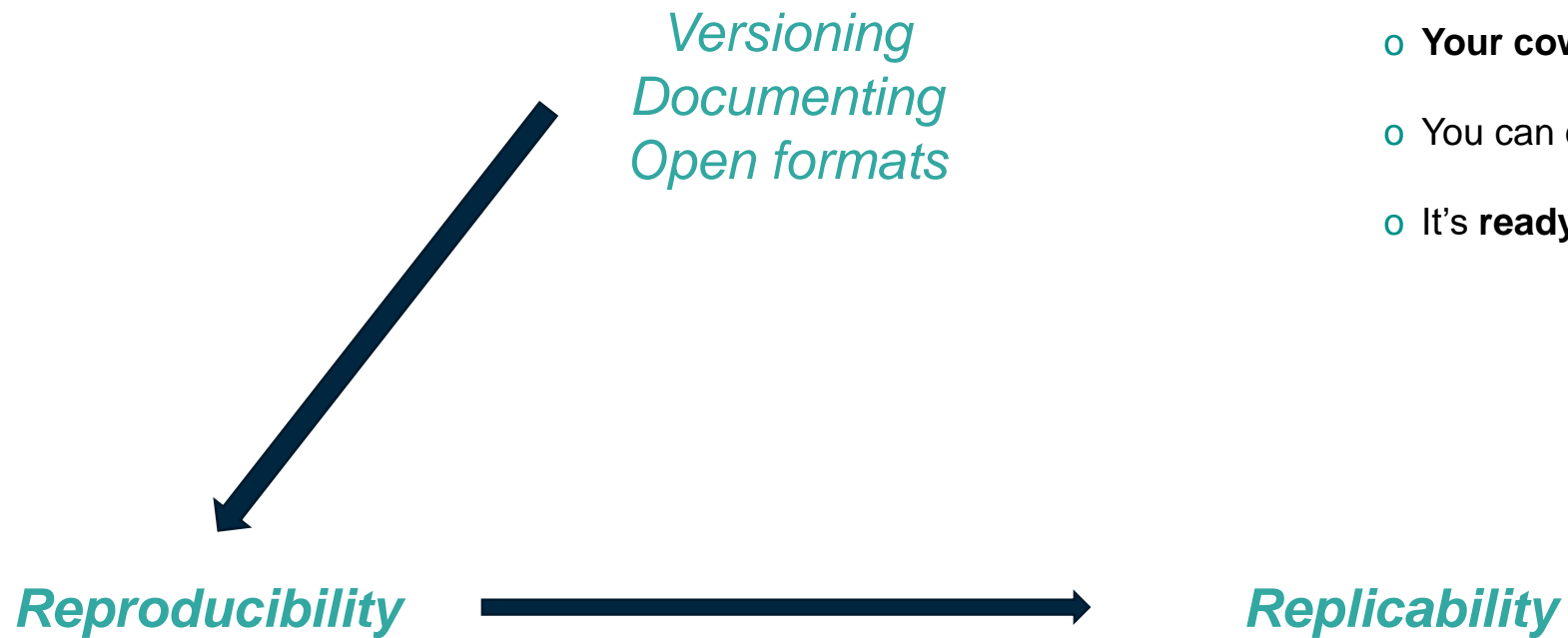


Metadata

Structured information associated with data (and code)

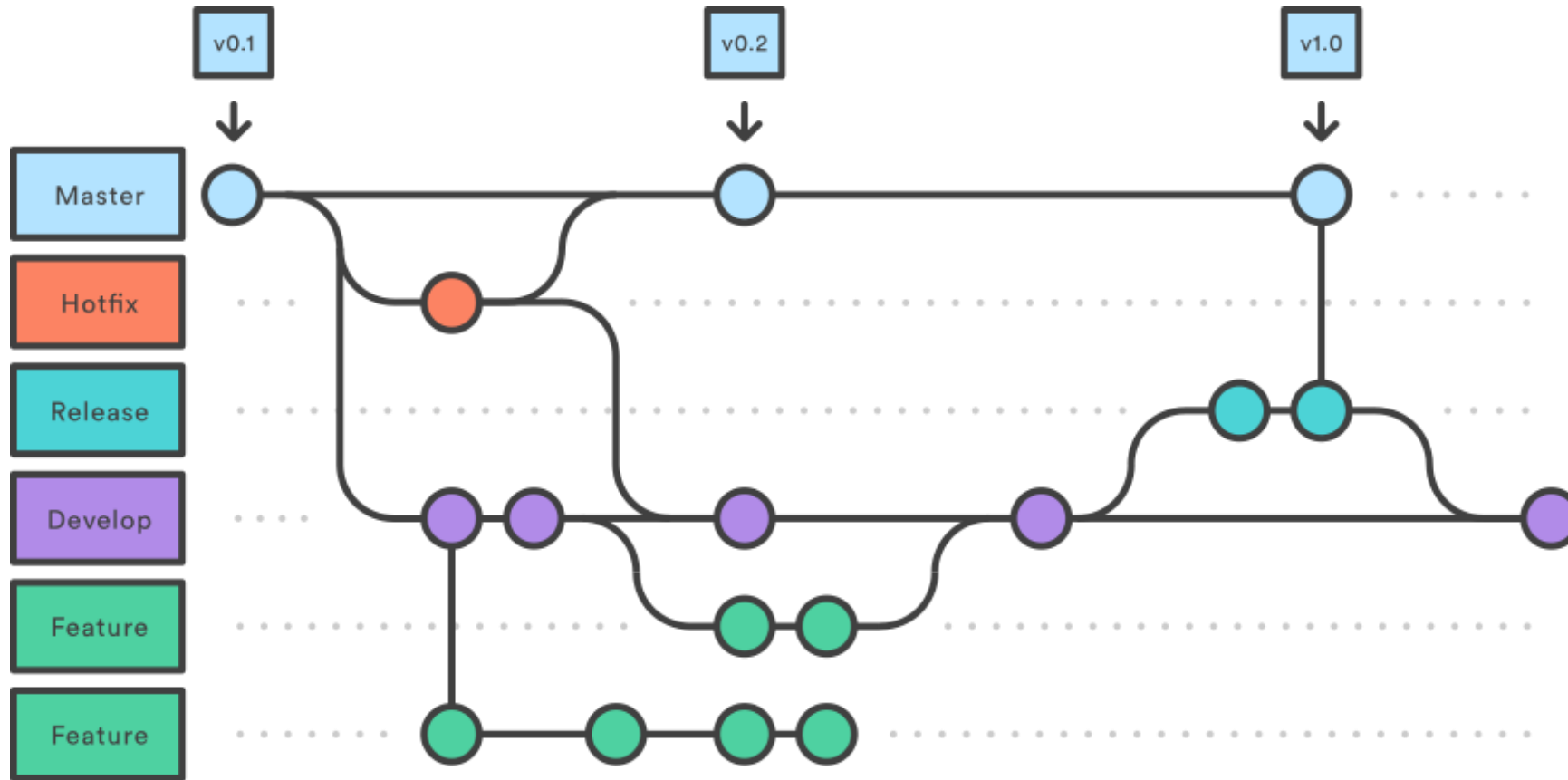
The Who, What, Where, Why & How of data

Collect and Store

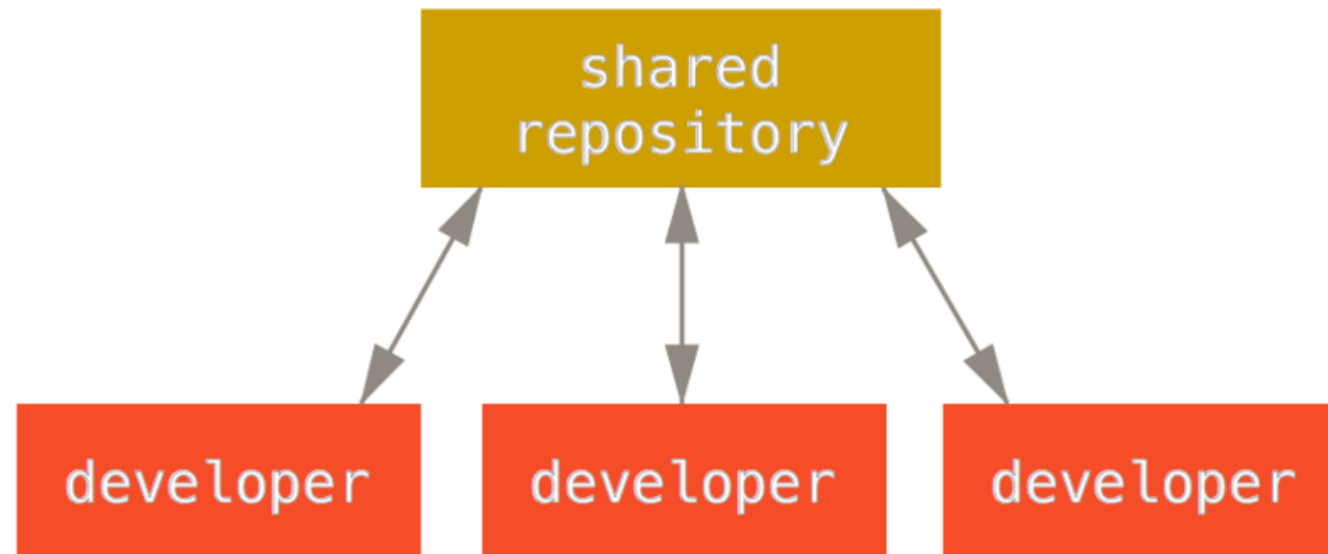


- **You** can find it
- **Your coworkers** can find it
- You can easily **share** it
- It's **ready** for **archiving/publishing**

Collect and Store: Software version control



Collect and Store: Software version control



Collect and Store: Software version control



<https://git-scm.com/>

- o CLI (*Command Line interface*)

- o GUIs (*Graphical User Interfaces*)

<https://git-scm.com/downloads/guis>



Workstation



Your own server



Gitea



GitLab



Internet



Codeberg



GitLab



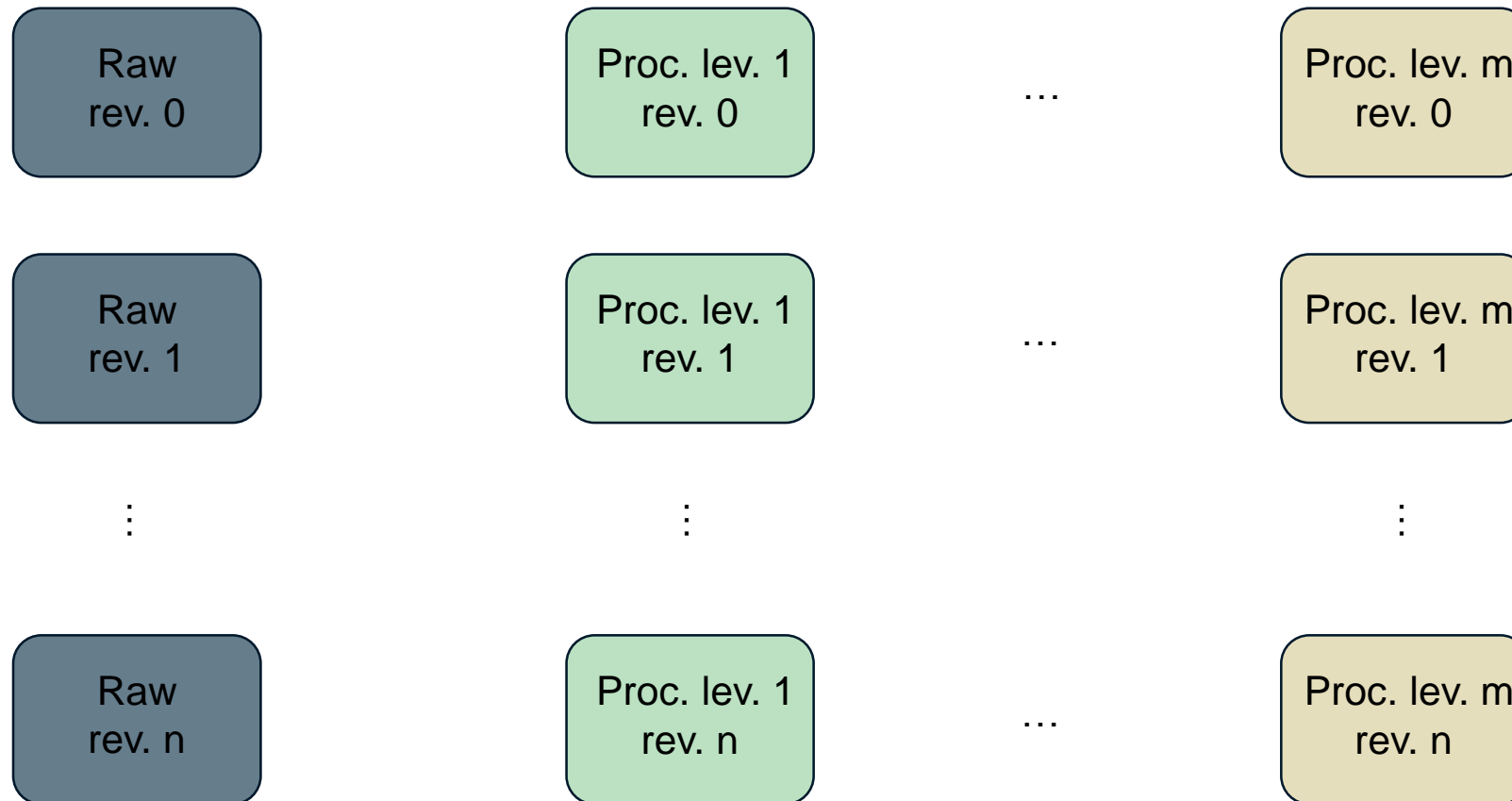
GitHub



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	<ul style="list-style-type: none"> • PDF/A • Plain Text coded as ASCII. UTF-8 or UTF-16 • XML
Spreadsheet	<ul style="list-style-type: none"> • CSV (NEAD)
Images	<ul style="list-style-type: none"> • TIFF (uncompressed or lossless compressed) • PNG
Code	<ul style="list-style-type: none"> • Languages with free environments (e.g. Py or R UTF-8 format of ASCII text)
Audio	<ul style="list-style-type: none"> • FLAC • Wav

Open and lossless formats

If you are using a proprietary format, think about adding an additional format

Collect & Store: Metadata Standards

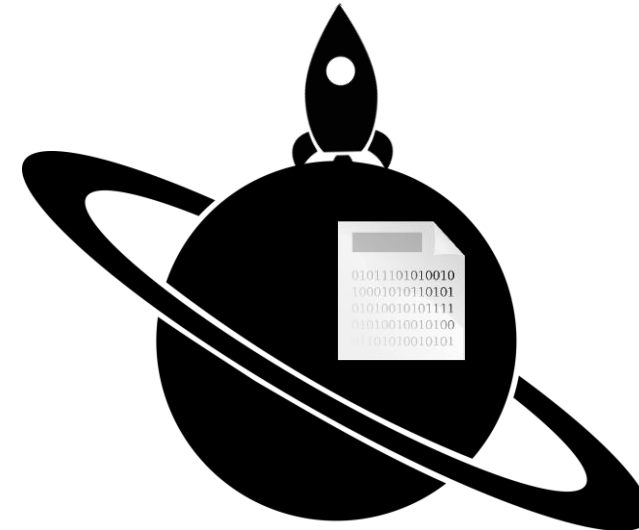
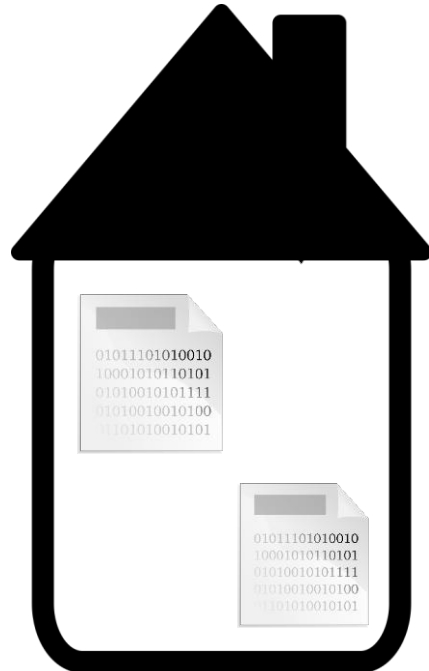
- Definition: Structured data that contains information about other data, but is not the content of the data.
- Metadata is very subject specific. The following directories are helpful:
 - Digital Curation Centre (<https://www.dcc.ac.uk/guidance/standards>)
 - RDA Metadata Standards (<https://rdamsc.bath.ac.uk>)
 - 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	<ul style="list-style-type: none"> • Title of the dataset • Contact information principal investigator • Date of data collection • Geographic location
Data and file overview	<ul style="list-style-type: none"> • Short discription for each file name • Date
Sharing and access informations	<ul style="list-style-type: none"> • Licenses or restrictions
Methodological information	<ul style="list-style-type: none"> • Description of methods for data collection or generation • Description of methods used for data processing
Data specific information (repeat for each dataset)	<ul style="list-style-type: none"> • 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>

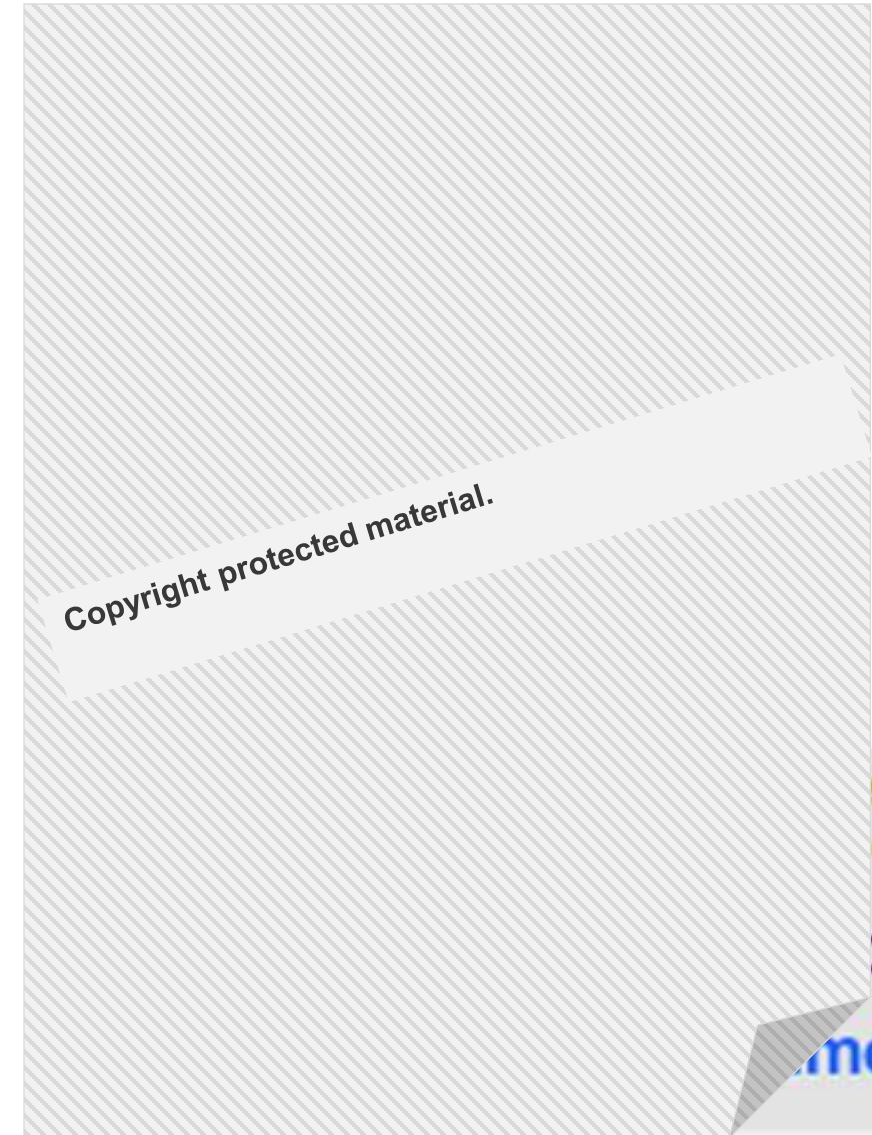
Collect and Store: 3 – 2 – 1 backup



Evaluate & Archive

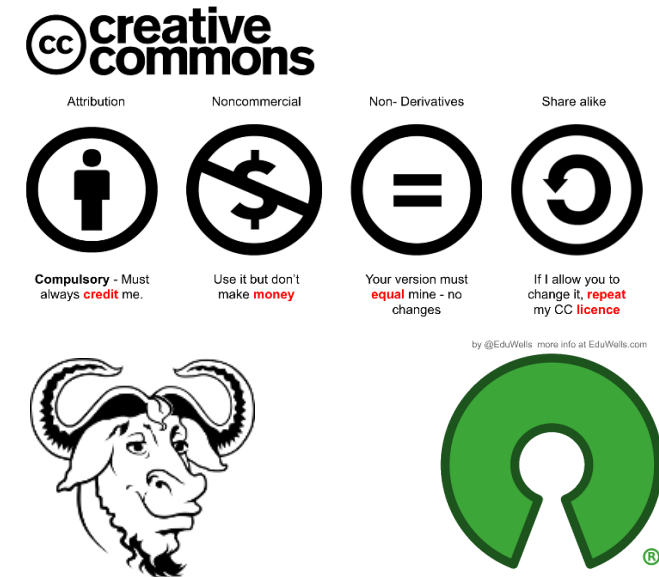
Evaluate & Archive: Data Protection

- **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**



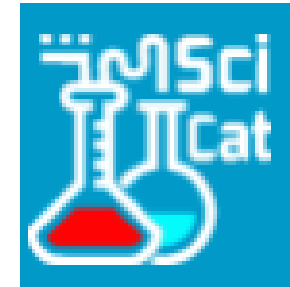
Evaluate & Archive: Data Protection

- 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

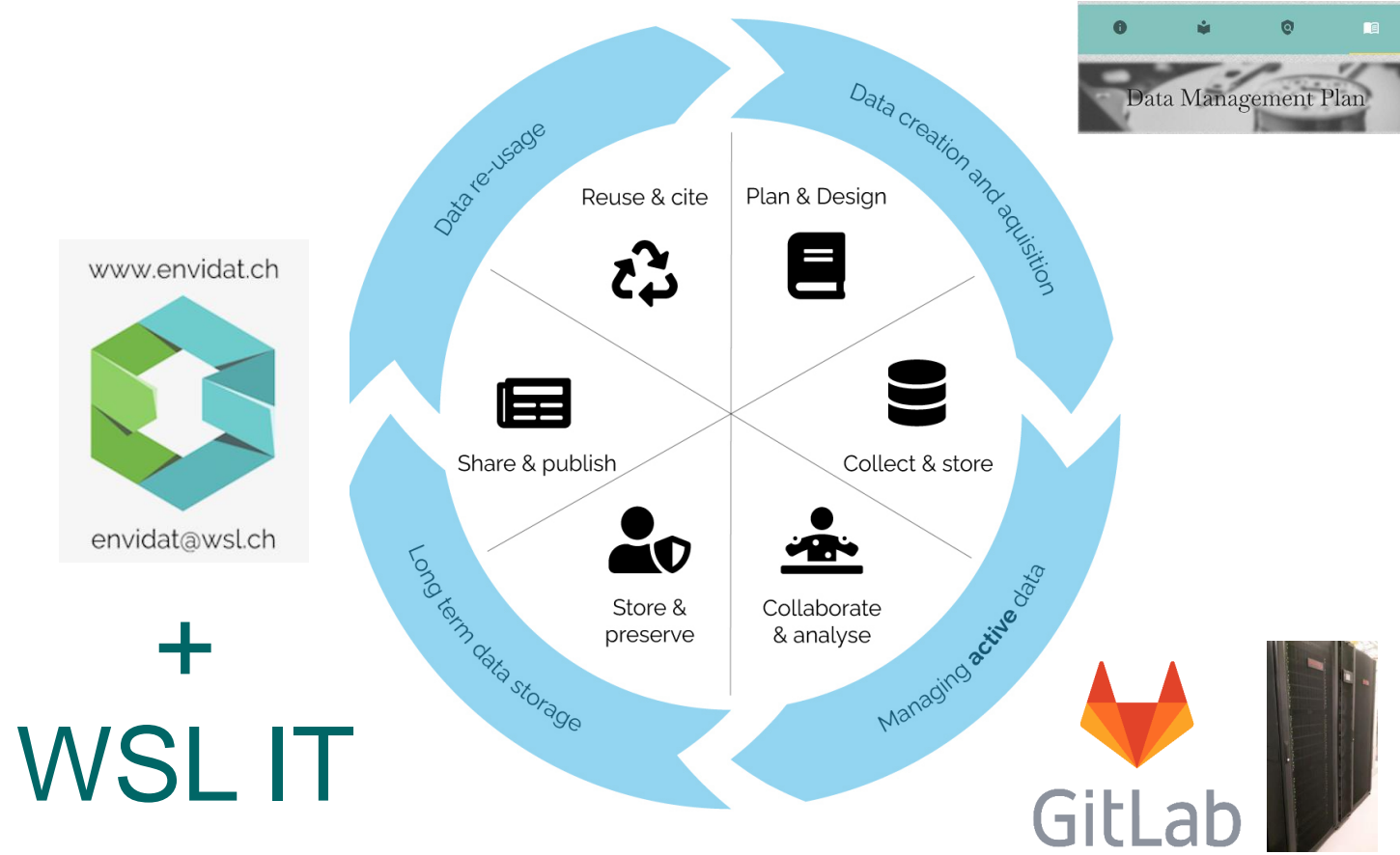
Share & Disseminate: The Choice of Data Repository



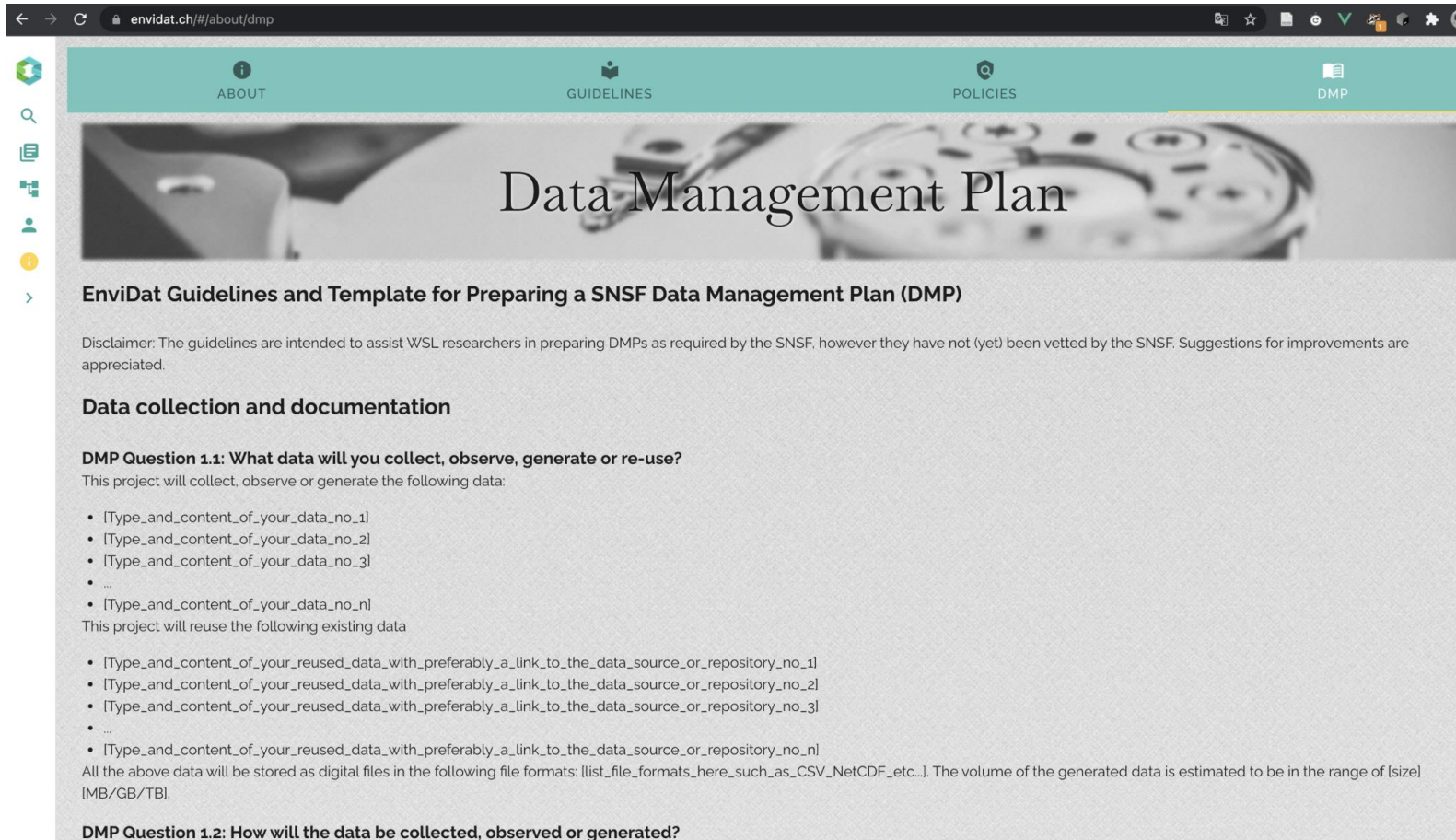
For alternatives: <https://www.re3data.org/>

RDM Services and Support at WSL

RDM Services and Support at WSL



RDM Services and Support at WSL: DMP Template



envidat.ch/#/about/dmp

ABOUT GUIDELINES POLICIES DMP

Data Management Plan

EnviDat Guidelines and Template for Preparing a SNSF Data Management Plan (DMP)

Disclaimer: The guidelines are intended to assist WSL researchers in preparing DMPs as required by the SNSF, however they have not (yet) been vetted by the SNSF. Suggestions for improvements are appreciated.

Data collection and documentation

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

This project will collect, observe or generate the following data:

- [Type_and_content_of_your_data_no_1]
- [Type_and_content_of_your_data_no_2]
- [Type_and_content_of_your_data_no_3]
- ...
- [Type_and_content_of_your_data_no_n]

This project will reuse the following existing data

- [Type_and_content_of_your_reused_data_with_preferably_a_link_to_the_data_source_or_repository_no_1]
- [Type_and_content_of_your_reused_data_with_preferably_a_link_to_the_data_source_or_repository_no_2]
- [Type_and_content_of_your_reused_data_with_preferably_a_link_to_the_data_source_or_repository_no_3]
- ...
- [Type_and_content_of_your_reused_data_with_preferably_a_link_to_the_data_source_or_repository_no_n]

All the above data will be stored as digital files in the following file formats: [list_file_formats_here_such_as_CSV_NetCDF_etc..]. The volume of the generated data is estimated to be in the range of [size] [MB/GB/TB].

DMP Question 1.2: How will the data be collected, observed or generated?

RDM Services and Support at WSL: NEAD Format

Non-Binary Environmental Data Archive (NEAD) format ✕

✉ Ionuț Iosifescu Enescu ✉ Mathias Bavay ✉ Kenneth Mankoff

👤 EnviDat ✉ envidat@wsl.ch 🌐 [10.16904/envidat.187](https://doi.org/10.16904/envidat.187) 📄 Creative Commons Zero - No Rights Reserved (CC0 1.0)

CSV METADATA CONVENTION METEOROLOGICAL DATA FORMAT NEAD RESEARCH DATA MANAGEMENT

Description

Acknowledgement: The NEAD format includes NetCDF metadata and is proudly inspired by both SMET and NetCDF formats. NEAD is designed as a long-term data preservation and exchange format.

The NEAD specifications were presented at the **"WMO Data Conference 2020 - Earth System Data Exchange in the 21st Century" (Virtual Conference)**.

Summary: The Non-Binary Environmental Data Archive (NEAD) format is being developed as a generic and intuitive format that combines the self-documenting features of NetCDF with human readable and writeable features of CSV. It is designed for exchange and preservation of time series data in environmental data repositories.

License: The NEAD specifications are released to the public domain under a Creative Commons 4.0 CC0 "No Rights Reserved" international license. You can reuse the information contained herein in any way you want, for any purposes and without restrictions.

Data and resources 📄 4

NEAD syntax v0.1 (EBNF)		NEAD on GitHub	
📄 txt	3.3 KB	📄 html	4. Nov 2020 13:26
📅 4. Nov 2020 13:22	📅 4. Nov 2020 13:22	📅 4. Nov 2020 13:26	📅 4. Nov 2020 13:26
📅 4. Nov 2020 13:22	📅 4. Nov 2020 13:22	📅 4. Nov 2020 13:26	📅 4. Nov 2020 13:26

MeteoIO (future reference implementation)		PyNEAD (future reference implementation)	
📄	4. Nov 2020 13:38	📄	4. Nov 2020 17:10
📅 4. Nov 2020 13:38	📅 4. Nov 2020 13:38	📅 4. Nov 2020 17:10	📅 4. Nov 2020 17:10
📅 4. Nov 2020 13:38	📅 4. Nov 2020 13:38	📅 4. Nov 2020 17:10	📅 4. Nov 2020 17:10

Citation

Ionuț Iosifescu Enescu; Mathias Bavay; Kenneth Mankoff (2020). Non-Binary Environmental Data Archive (NEAD) format. EnviDat. doi: [10.16904/envidat.187](https://doi.org/10.16904/envidat.187).

Location 📍 🗺

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

RDM Services and Support at WSL: GitLab

WSL/SLF GitLab Repository



Welcome to WSL/SLF GitLab repository!

You must sign up to use this GitLab instance. If you are new and not member of WSL/SLF please [register](#) first. Then contact the project leader to be approved.




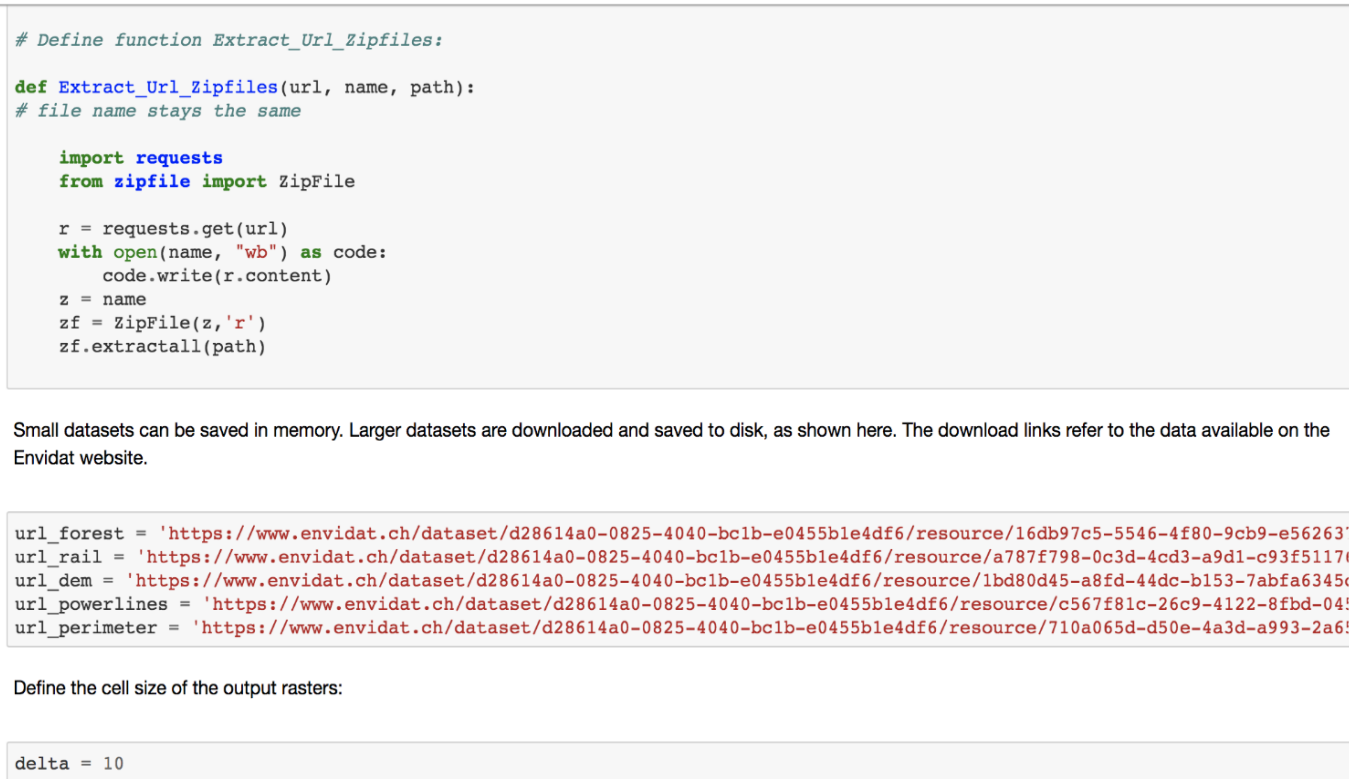
LDAP login is only for WSL/SLF members. External users can sign in on the **Standard** tab.

For public GitLab projects please visit <https://gitlabext.wsl.ch/public>.

LDAP	Standard
<p>LDAP Username</p> <input type="text" value="josifesc"/>	
<p>Password</p> <input type="password" value="....."/>	
<input type="checkbox"/> Remember me	
<input type="button" value="Sign in"/>	

Don't have an account yet? [Register now](#)

RDM Services and Support at WSL: Jupyter Notebooks

[JUPYTER](#) [FAQ](#) [</>](#) [☰](#) [⬇](#)

```

# Define function Extract_Url_Zipfiles:

def Extract_Url_Zipfiles(url, name, path):
    # file name stays the same

    import requests
    from zipfile import ZipFile

    r = requests.get(url)
    with open(name, "wb") as code:
        code.write(r.content)
    z = name
    zf = ZipFile(z, 'r')
    zf.extractall(path)
    
```

Small datasets can be saved in memory. Larger datasets are downloaded and saved to disk, as shown here. The download links refer to the data available on the Envidat website.

```

In [23]: url_forest = 'https://www.envidat.ch/dataset/d28614a0-0825-4040-bc1b-e0455b1e4df6/resource/16db97c5-5546-4f80-9cb9-e56263'
url_rail = 'https://www.envidat.ch/dataset/d28614a0-0825-4040-bc1b-e0455b1e4df6/resource/a787f798-0c3d-4cd3-a9d1-c93f5117'
url_dem = 'https://www.envidat.ch/dataset/d28614a0-0825-4040-bc1b-e0455b1e4df6/resource/1bd80d45-a8fd-44dc-b153-7abfa6345'
url_powerlines = 'https://www.envidat.ch/dataset/d28614a0-0825-4040-bc1b-e0455b1e4df6/resource/c567f81c-26c9-4122-8fbd-04!'
url_perimeter = 'https://www.envidat.ch/dataset/d28614a0-0825-4040-bc1b-e0455b1e4df6/resource/710a065d-d50e-4a3d-a993-2a6!'
    
```

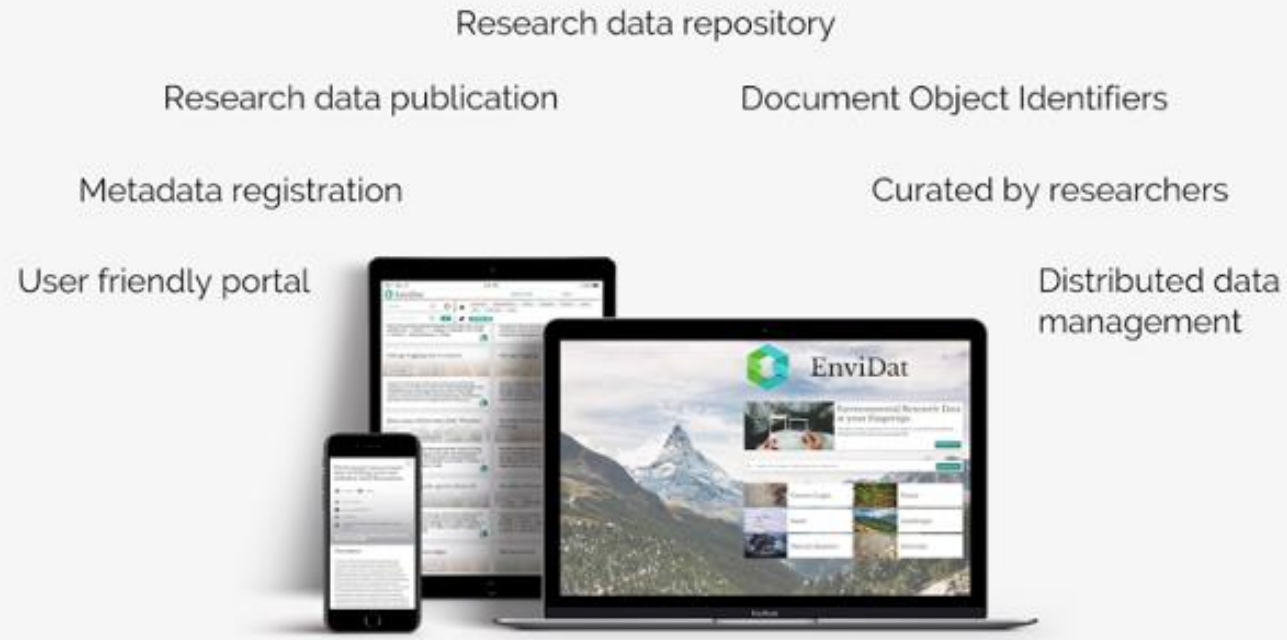
Define the cell size of the output rasters:

```

In [24]: delta = 10
    
```

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.



RDM Services and Support at WSL: EnviDat

CHELSA-TraCE21k: Downscaled transient temperature and precipitation data since the last glacial maximum

Dirk Nikolaus Karger, Michael P. Nobis, Signe Normand, Catherine H. Graham, Niklaus E. Zimmermann

Niklaus E. Zimmermann | niklaus.zimmermann@wsl.ch | 10.16904/envidat.211

AIR TEMPERATURE | GLACIERS | LAST GLACIAL MAXIMUM | OROGRAPHY | PALEO CLIMATE | PRECIPITATION

Description

High resolution, downscaled climate model data are used in a wide variety of applications in environmental sciences. Here we present the CHELSA-TraCE21k downscaling algorithm to create global monthly climatologies for temperature and precipitation at 30 arcsec spatial resolution in 100 year time steps for the last 21,000 years. Paleo orography at high spatial resolution and at each timestep is created by combining high resolution information on glacial cover from current and Last Glacial Maximum (LGM) glacier databases with the interpolation of a dynamic ice sheet model (ICE6G) and a coupling to mean annual temperatures from CCSM3-TraCE21k. Based on the reconstructed paleo orography, mean annual temperature and precipitation was downscaled using the CHELSA V1.2 algorithm.

The data is published under a Creative Commons Attribution 2.0 Generic (CC BY 2.0) license.

Citation

Dirk Nikolaus Karger, Michael P. Nobis, Signe Normand, Catherine H. Graham, Niklaus E. Zimmermann (2020). CHELSA-TraCE21k: Downscaled transient temperature and precipitation data since the last glacial maximum. EnviDat. doi: 10.16904/envidat.211.

DATA CITE | ISO 19139 | GCMD DIF | BIBTEX | RIS

Related Publications

Karger, D.N., Conrad, O., Böhner, J., Kawohl, T., Kreft, H., Soria-Auza, R.W., Zimmermann, N.E., Linder, H.P. & Kessler, M. (2017) Climatologies at high resolution for the earth's land surface areas. Scientific Data 4, 170122.

Karger, D.N., Schmatz, D., Dettling, D., Zimmermann, N.E. (2020) High resolution monthly precipitation and temperature timeseries for the period 2006-2100. Scientific Data.

Data and resources

Data Access via S3

High resolution, downscaled climate model data are used in a wide variety of applications in environmental sciences. Here we present the CHELSA-TraCE21k downscaling algorithm ...

geoliff | 23 Jun 2020 12:07

Technical documentation

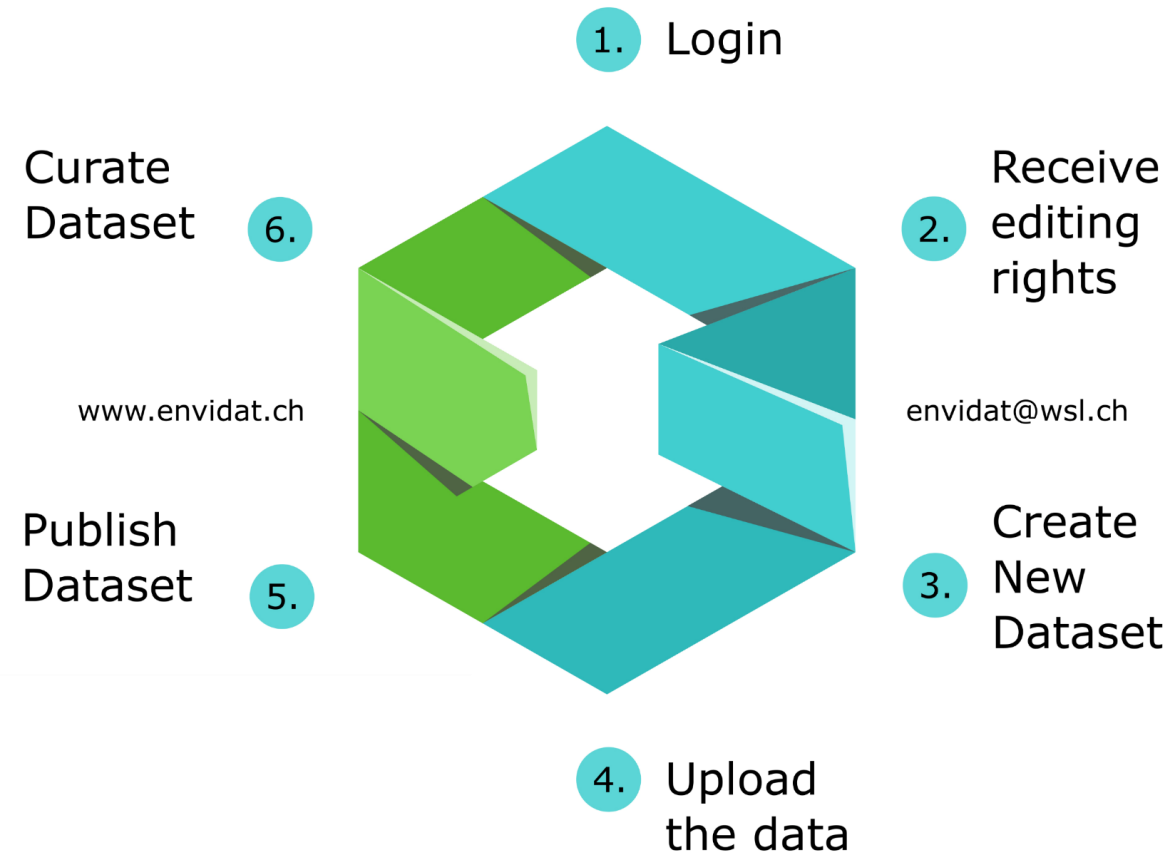
Technical documentation containing file metadata and variable descriptions.

pdf | 479.13 KB | 17 Feb 2021 13:08

Location

doi:10.16904/envidat.211

RDM Services and Support at WSL: EnviDat



RDM Services and Support at WSL: EnviDat

opendata.swiss

Data Organizations Showcases Contact About

Home > Organizations > EnviDat: the Environmental...

EnviDat: the Environmental Data Portal

EnviDat is the WSL data portal providing unified and managed access to environmental monitoring and research data. The portal has the capability to host and publish data sets. Through the sharing of unique and high-value data sets worldwide, EnviDat fosters research and contributes to advancing environmental science.

More information: <https://www.envidat.ch/>

Search datasets...

Reset search

Categories
Education and science 202

Keywords
snow 38 precipitation 31 climate-change 26 temperature 22 raclets 21 wind-speed 20 relative-humidity 18 swiss-forest-lab 17 long-term 16 meteo 16

202 datasets found


Order by Relevance

Snow Drift Station - 3D Ultrasonic
EnviDat: the Environmental Data Portal
Education and science

A Young 81000 sonic anemometer was deployed at Gotschnagratt (LON: 46.859 LAT: 9.849) to record three components of the wind velocity (u, v, w in [m s⁻¹; ¹]) and air...

Last updated: March 27, 2020

SERVICE raclets wind-velocity wind-direction wind-speed anemometer wind



Plan & Design: Data Management Plan (DMP)

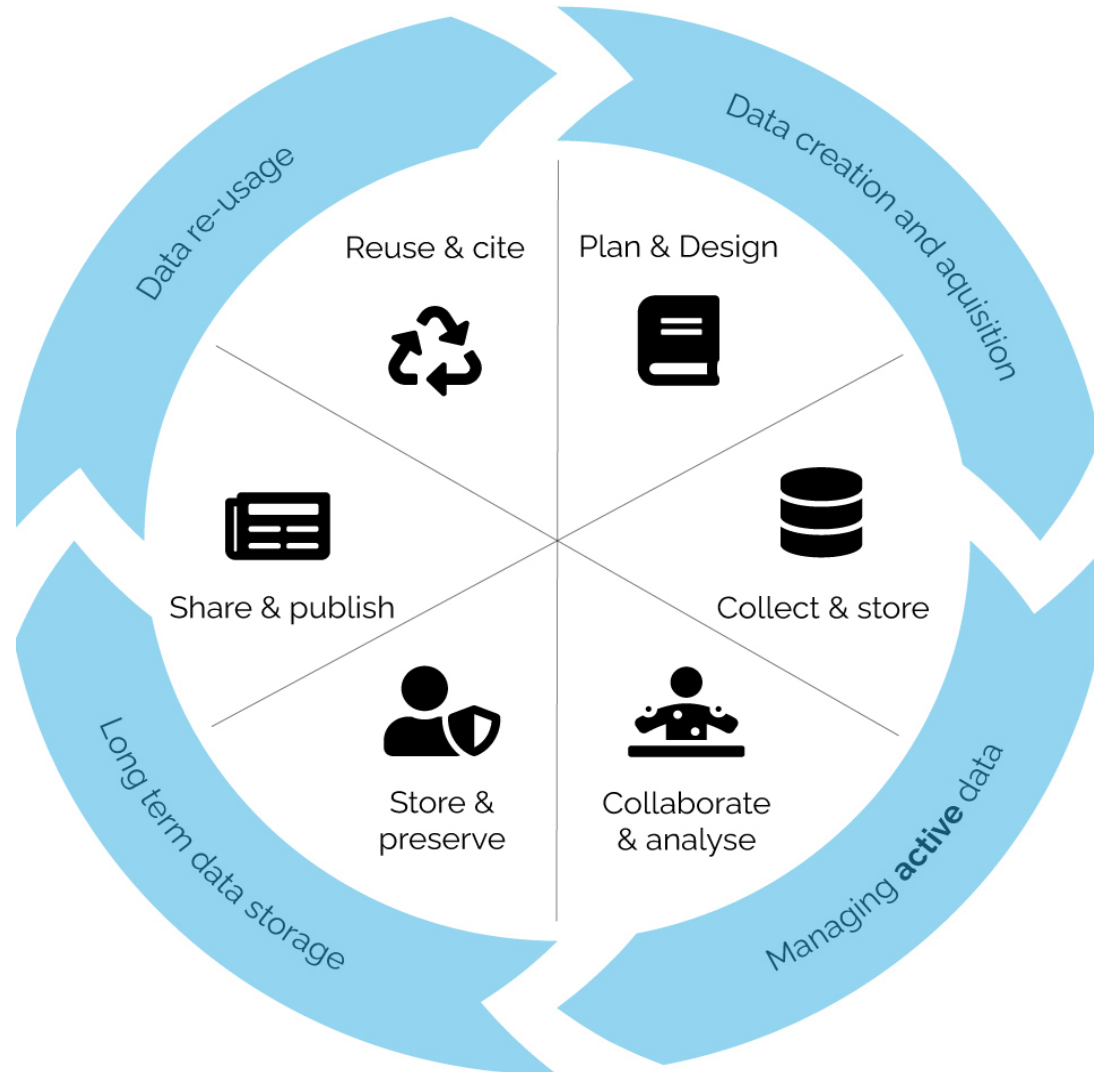
Plan & Design: Why?



Bibliothèque de l'EPFL, «RDM Horror stories | Episode 2 – Stranger Data Things», 11th February 2020.

<https://bit.ly/3qPWMIS>

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

Applications and Projects

Grant application 1

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 major achievements
- Quotes
- Cover letter
- Official certificates
- Weave/Lead Agency and International Co-Investigator Scheme
- Other annexes

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.

Plan & Design: DMP



- Keep it short and simple
- Be stingy with words
- Have one idea per sentence
- Use the active form
- Use positive phrases
- Use concrete terms

«we used the method» not «the method was used»
 «the results are different» not «the results are not the same»
 «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
- Don't nominalise
- Don't use empty modifiers
- Don't use tautologous modifiers

e.g. «It is clear that», «the fact is that», «in an attempt to», «in order to»
 «reduce» not «achieve a reduction in length»
 e.g. «basically», «indeed», «quite», «actually»
 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 as well as rare and protected species
- Requirements for assessments by ethical review boards, permission 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 restriction, IT infrastructure

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

- Consider non-disclosure agreements, potential patents, research collaborations across 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: WSL

1. RDM consulting services (contact envidat@wsl.ch)
2. DMP templates (<https://envidat.ch/#/dmp>) and consulting services (contact envidat@wsl.ch)
3. WSL/SLF GitLab repository: <https://gitlabext.wsl.ch> (allows collaboration with external partners)
4. WSL's Hyperion High-Performance Computing (HPC) Linux Cluster
<https://hyperion.wsl.ch/documentation-master/build/html/> (only available inside the WSL Intranet!)
5. JupyterLab/JupyterHub interactive development environment to combine documentation, code and data
<https://hyperion.wsl.ch/documentation-master/build/html/> (only available inside the WSL Intranet!)
6. Temporary data transfers to external partners for collaboration purposes:
 - For data up to 300GBs – SWITCHfilesender: <https://filesender.switch.ch/> (temporary for 10 days up to max. 60 days)
 - For data above 300GBs and up to several TBs – <https://envicloud.wsl.ch/> (contact envidat@wsl.ch – temporary for max. 10 days)
7. EnviDat – www.envidat.ch – for final data publication with DOI assignment (datasets larger than 1 TB needs to be approved by the WSL directorate)

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

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	<ul style="list-style-type: none"> 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) 	<ul style="list-style-type: none"> 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.) <p>Not accepted for publication, OK for supplementary materials:</p> <ul style="list-style-type: none"> 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) 	<ul style="list-style-type: none"> Word *.doc PowerPoint *.ppt
Spreadsheet or table	<ul style="list-style-type: none"> Comma- or tab delimited text files (*.csv) 	<ul style="list-style-type: none"> Excel *.xlsx (container format) OpenDocument spreadsheets (*.ods) 	<ul style="list-style-type: none"> Excel *.xls, *.xlsb (binary formats)
Raw data and workspace		<ul style="list-style-type: none"> 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) 	<ul style="list-style-type: none"> Binary files such as the standard Matlab files *.mat or the R files *.RData
Raster image (bitmap)	<ul style="list-style-type: none"> 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 	<ul style="list-style-type: none"> TIFF (*.tif) (compressed) GIF (*.gif) BMP (*.bmp) JPEG/JFIF (*.jpg) JPEG2000 (lossy compression) (*.jp2) 	
Vector graphics	<ul style="list-style-type: none"> SVG without JavaScript binding (*.svg) 		<ul style="list-style-type: none"> Graphics InDesign (*.indd), Illustrator (*.ait) Encapsulated Postscript (*.eps) Photoshop (*.psd)
CAD	<ul style="list-style-type: none"> AutoCAD Drawing (*.dwg) Drawing Interchange Format, AutoCAD (*.dxf) Extensible 3D, X3D (*.x3d, *.x3dv, *.x3db) 		
Audio	<ul style="list-style-type: none"> WAV (*.wav) (uncompressed, pulse-code modulated) 	<ul style="list-style-type: none"> Advanced Audio Coding (*.mp4) MP3 (*.mp3) 	
Video ¹	<ul style="list-style-type: none"> FFV1 codec (version 3 or later) in Matroska container (*.mkv) 	<ul style="list-style-type: none"> 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) 	<ul style="list-style-type: none"> 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, 2016 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.