NTK
50°6'14.083"N, 14°23'26.365"E
Národní technická knihovna
National Library of Technology

National Centre
for Information Support of Research,
Development, and Innovation

Introduction to Research Data Management

… and how not to get overwhelmed by data

Workshop lecturer: Jan Vališ

Authors of the presentation: Adéla Jílková, Martin Schätz, Jan Vališ

22. 10. 2025

The project National Centre for Information Support of Research, Development and Innovation with the identification code MS2101 is implemented with the support of the Ministry of Education, Youth and Sports.

Question 1: What is your academic affiliation?

A. Czech Technical University in Prague (CTU/ČVUT)

B. University of Chemistry and Technology, Prague (UCT/VŠCHT)

C. Charles University (CUNI/UK)

D. Czech University of Life Sciences Prague (CZU/ČZU)

E. Palacký University (UPOL)

F. Other: **please tell us in the chat.**

Question 2: What is your main research field?

A. DiS

B. Bachelor

C. Master’s

D. Doctoral Candidate

E. Postdoc

F. Other: **please tell us in the chat.**

Question 3: What is your main research field?

A. Natural, medical, and health sciences

B. Engineering and technology

C. Social sciences, business, and law

D. Arts and humanities

E. Other: **please tell us in the chat.**

What is research data and why manage it?

Research data – theoretically

- Any information **collected**, **observed**, **generated,** or **created** during the research process to produce and support research findings
- Primary data
  - Directly from a given research
  - Quantitative (volume) vs. Qualitative (interview)
  - Experimental (pH) vs. Observational (bird migration)
- Secondary data
  - From other research used for different reason

Research data – practically

- Pictures (SEM pictures, photos of birds etc.)
- Videos (movement of particles, interviews etc.)
- Sound records (birds singing, interviews etc.)
- Tables (spectra, observations etc.)
- Texts
- Statistics
- Electronic patient records
- …

Research data are not:

- Project proposal
- Project budget
- Evaluation report
- Conference proceedings
- Journal article

Research data management

- A set of practices, strategies, and activities, including data:
  - **organization,**
  - **documentation**,
  - **storage**, and
  - **sharing**
- Covers all stages of the research process
- Ensures the effectiveness, reproducibility, and reuse of research data

Why manage research data?

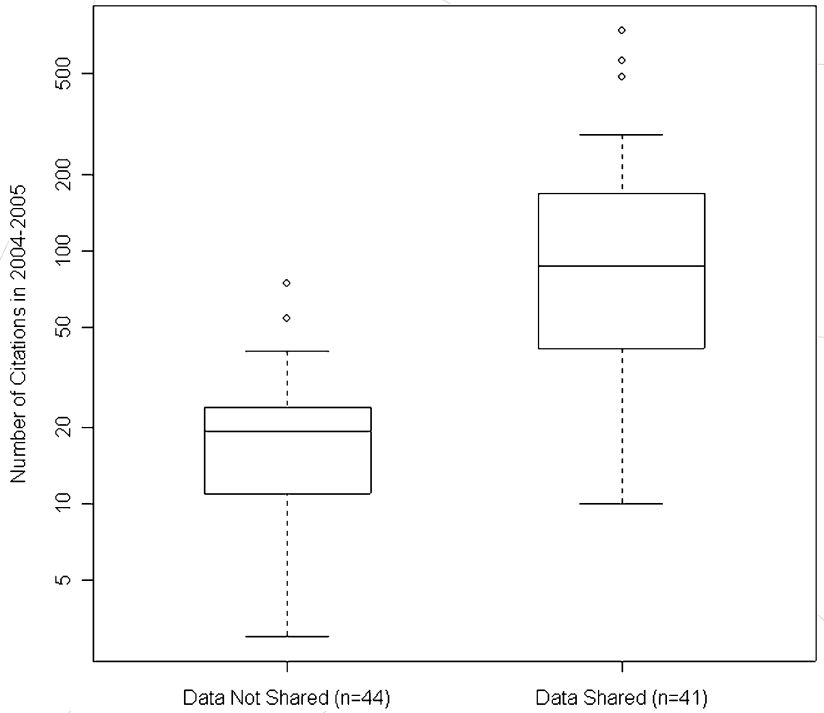
It may be mandatory (institutional, publisher, or research funder requirements)

Keep the research process secure and organized

- Increase efficiency, save time and resources
- Share data with colleagues
- Reduce risk of data loss and improve data security

Enhance global data sharing

- Enable data reuse and enhance collaboration
- Increase the visibility and impact of research
- Increase transparency & improve trust in findings
- Support research integrity & validation of results



2004–2005 citation counts of 85 trials by data availability.
Heather A. et al. 2007. PLOS ONE. License: CC BY 4.0
10.1371/journal.pone.0000308.

Research data

Different fields and disciplines

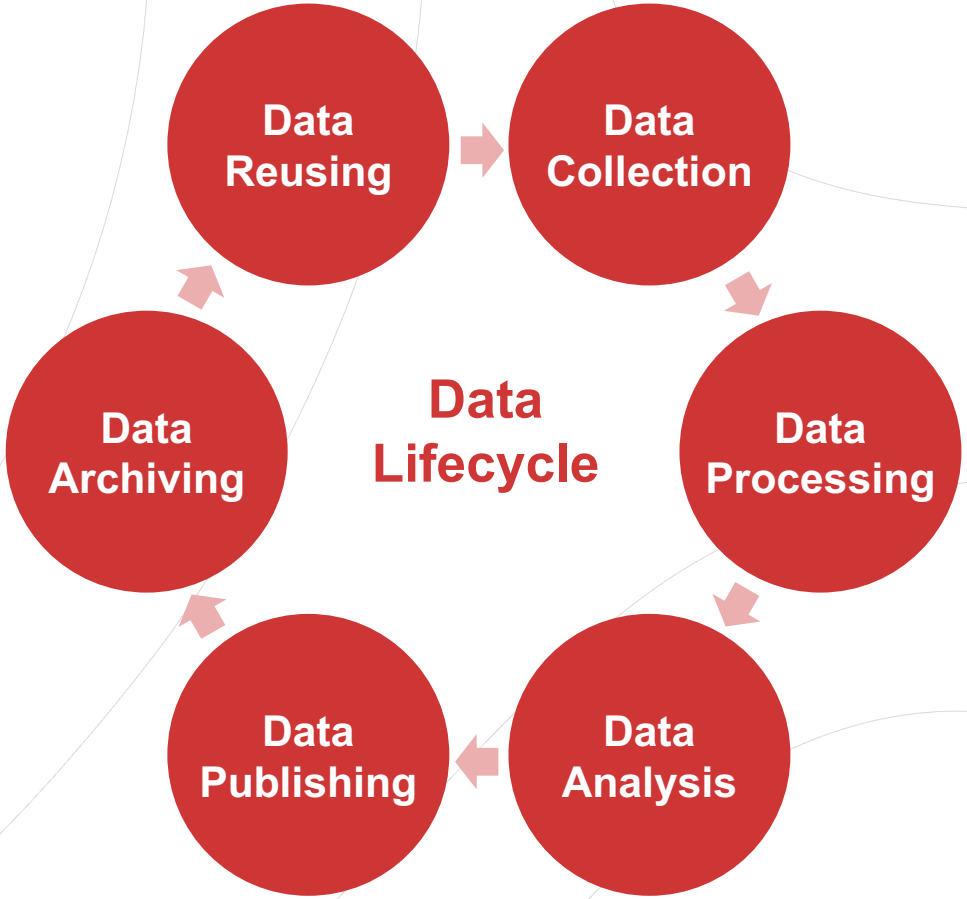
- Natural and life sciences
- Medical and health sciences
- Engineering and technology
- Social sciences
- Arts and humanities

Research data

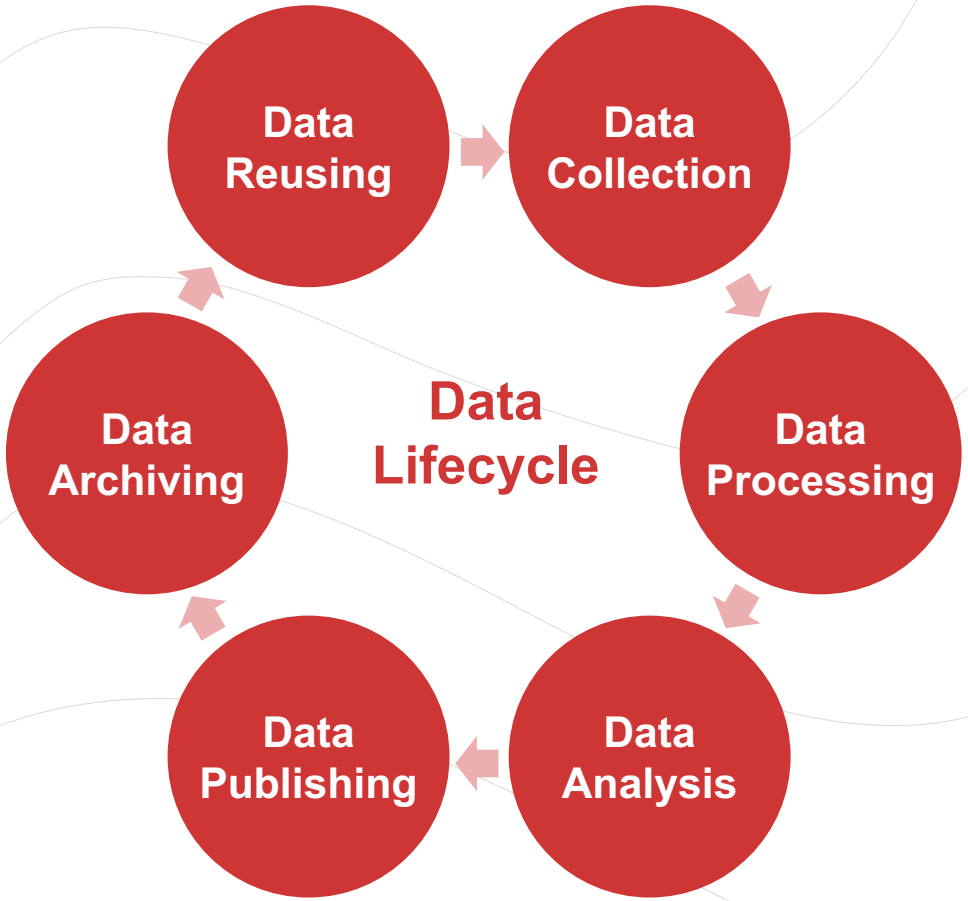
Different fields and disciplines

- Natural and life sciences
- Medical and health sciences
- Engineering and technology
- Social sciences
- Arts and humanities

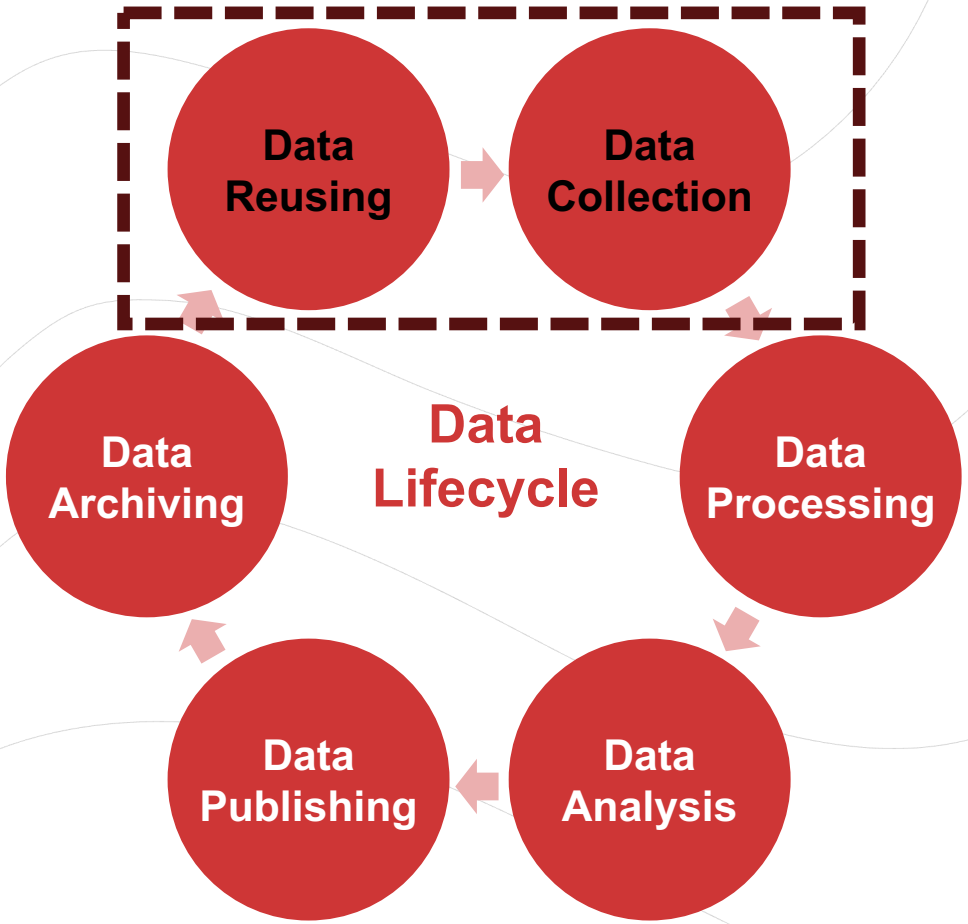
Different stages of research data lifecycle



Research data lifecycle



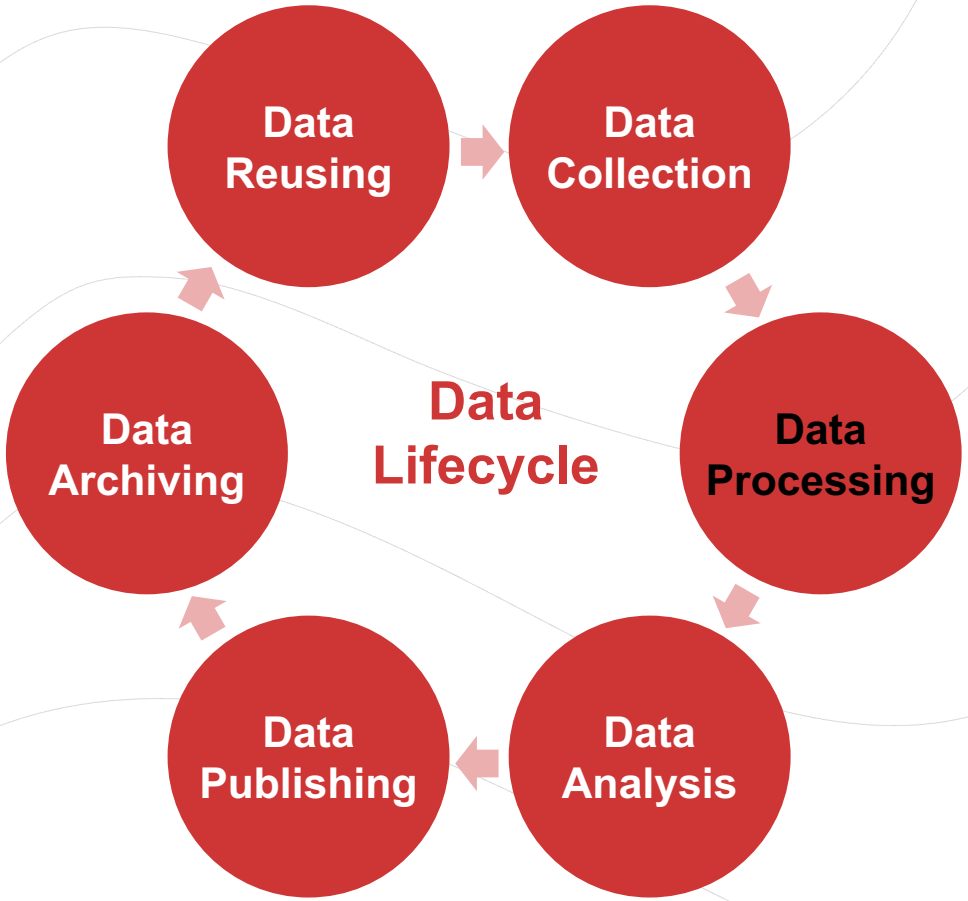
Research data lifecycle



Source Data

Collected/produced “raw data“
Reused data from a database/repository

Research data lifecycle



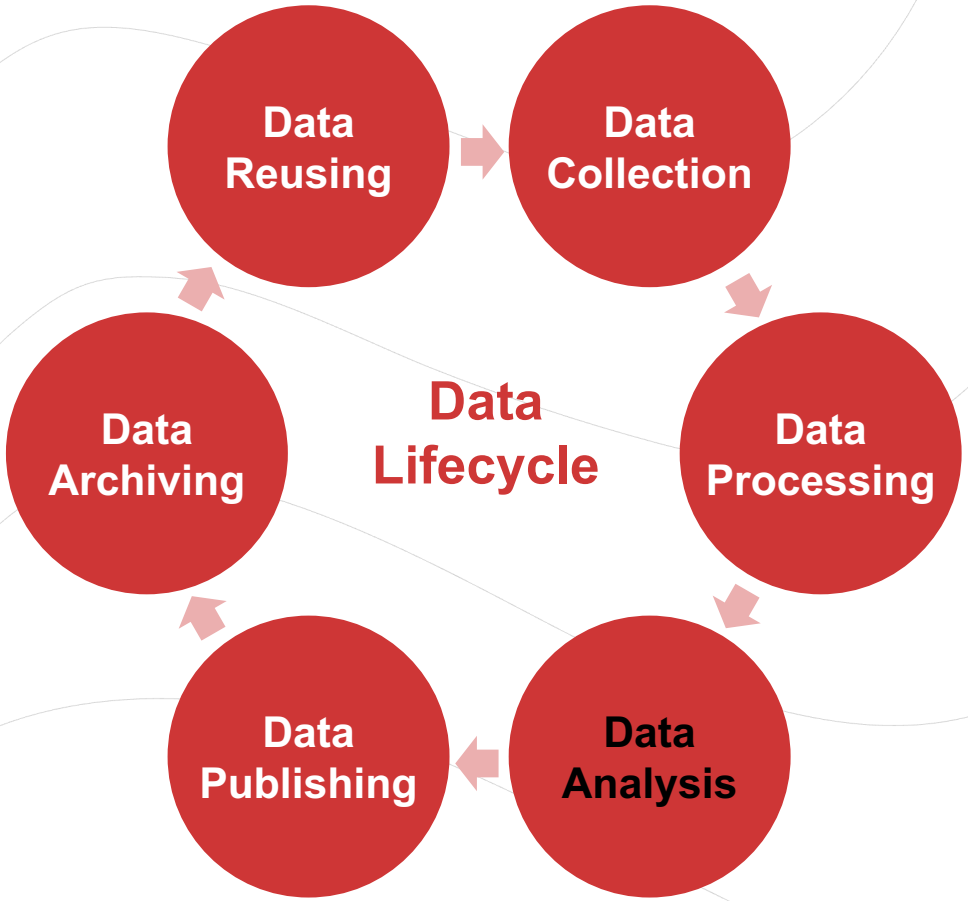
Source Data

Collected/produced “raw data“
Reused data from a database/repository

Data Processing

Transformation of raw data

Research data lifecycle

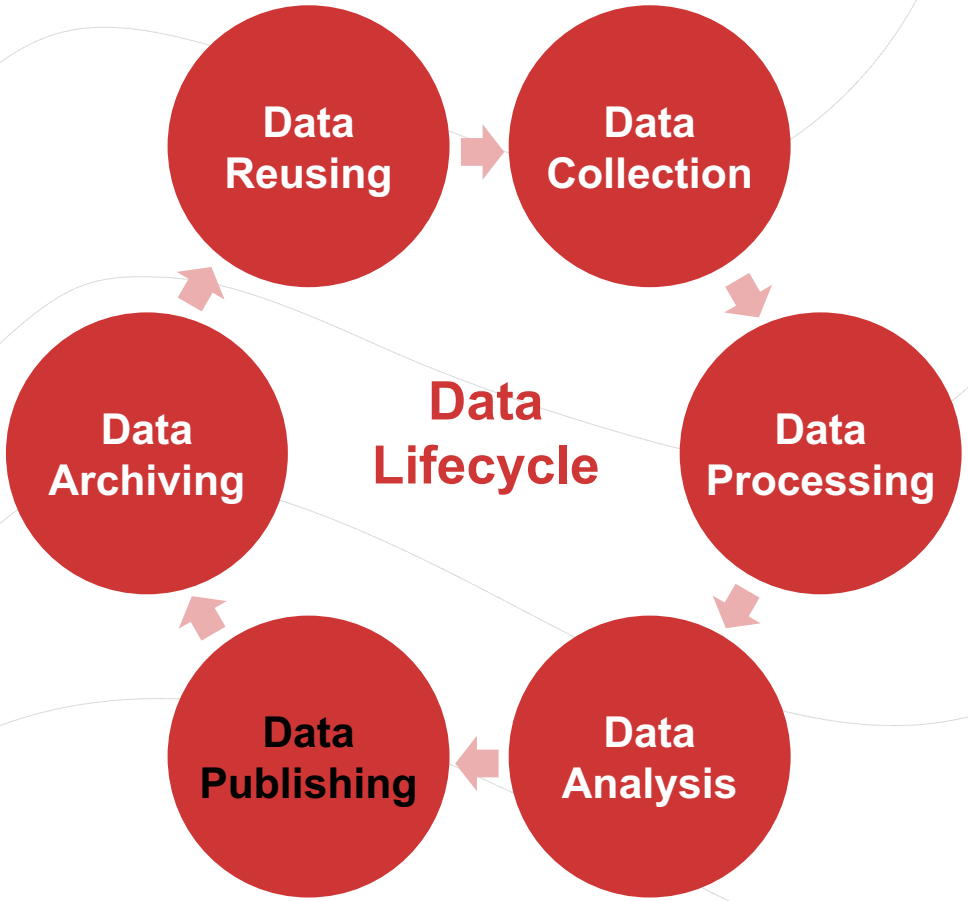


Source Data
Collected/produced “raw data“
Reused data from a database/repository

Data Processing
Transformation of raw data

Data Analysis
Data interpretation
Generation of results and outputs

Research data lifecycle



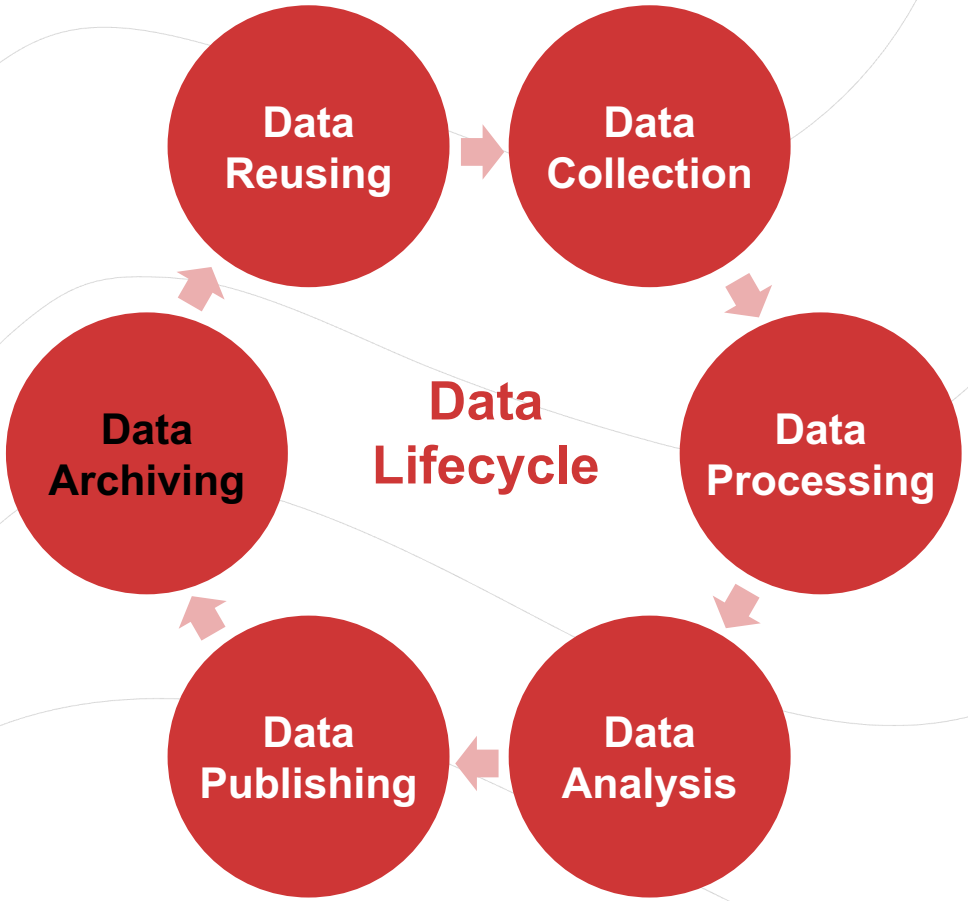
Source Data
Collected/produced “raw data“
Reused data from a database/repository

Data Processing
Transformation of raw data

Data Analysis
Data interpretation
Generation of results and outputs

Data Publishing
Journal article
Manuscript + supplementary information
Databases/repositories
Data underlying publication
Separate datasets

Research data lifecycle



Source Data
Collected/produced “raw data“
Reused data from a database/repository

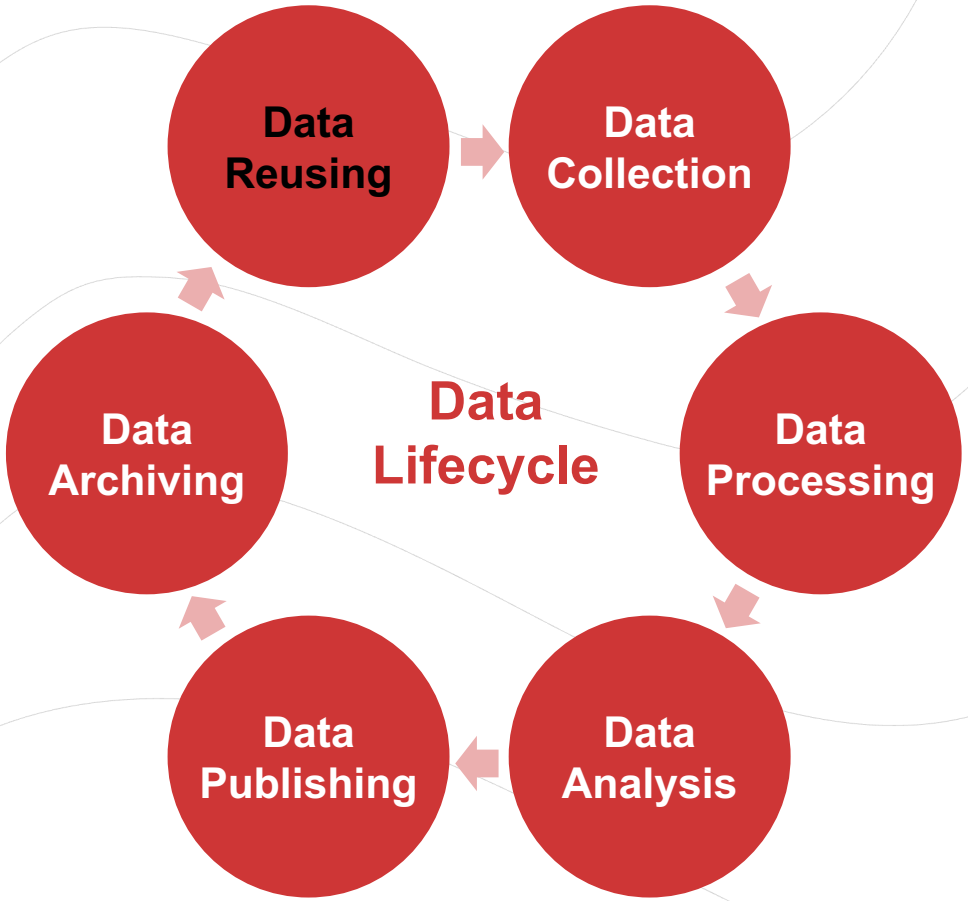
Data Processing
Transformation of raw data

Data Analysis
Data interpretation
Generation of results and outputs

Data Publishing
Journal article
Manuscript + supplementary information
Databases/repositories
Data underlying publication
Separate datasets

Data Archiving

Research data lifecycle



Source Data
Collected/produced “raw data“
Reused data from a database/repository

Data Processing
Transformation of raw data

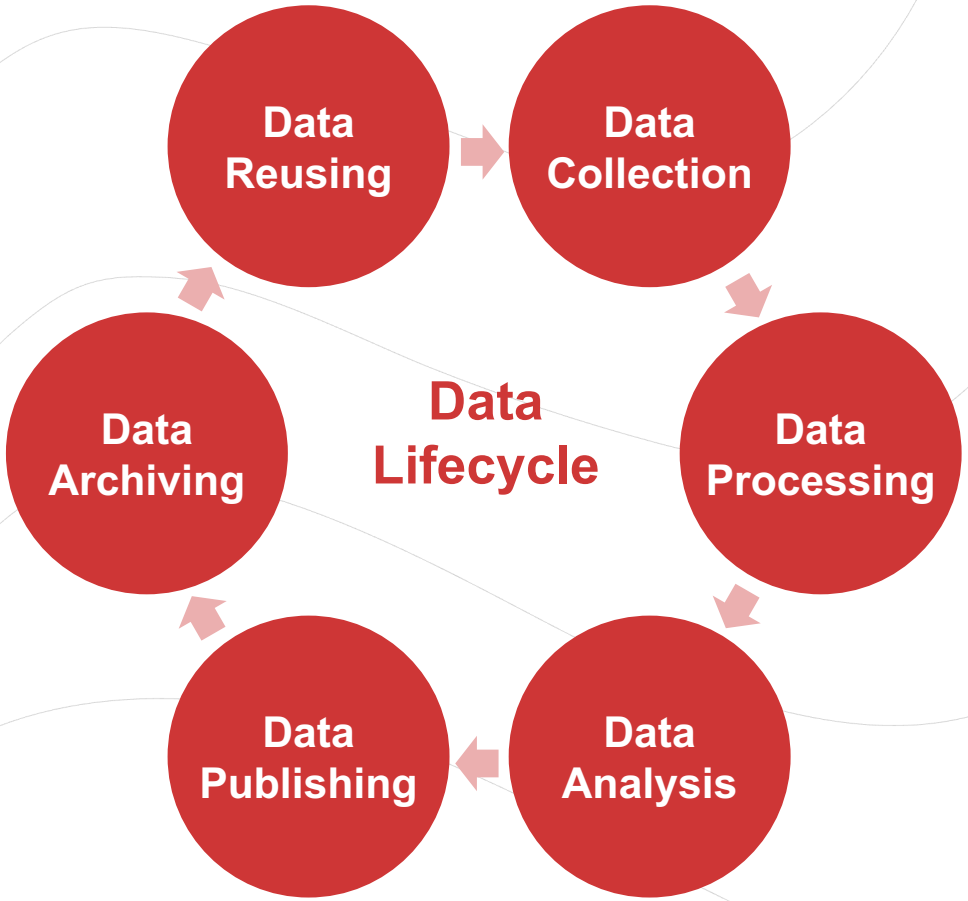
Data Analysis
Data interpretation
Generation of results and outputs

Data Publishing
Journal article
Manuscript + supplementary information
Databases/repositories
Data underlying publication
Separate datasets

Data Archiving

Data Reusing (registries, repositories)

RDM strategies



Organizing
- Directory structure
- Formats, names, versions

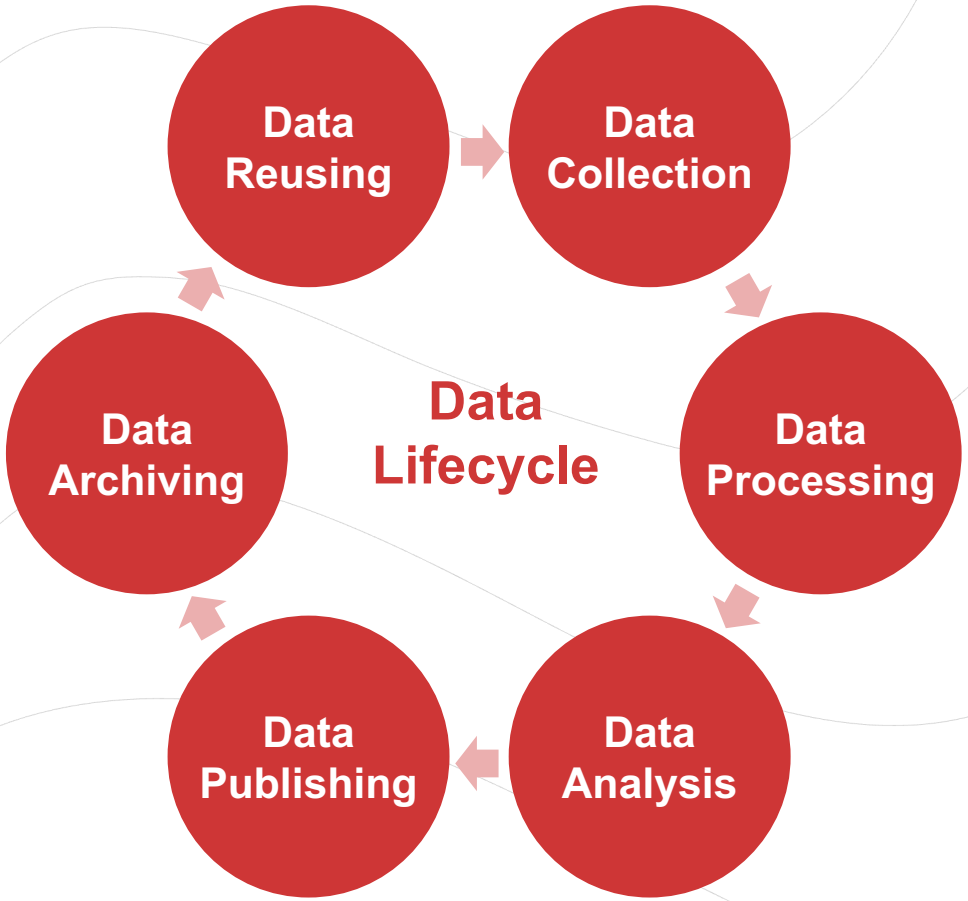
Documentation
- Data description
- Experimental details
- Decisions made
- Metadata

Storage
- Backup
- Long-term preservation

Data access
- Access rights (open, restricted)
- Licenses

RDM strategies

Plan
Generate ideas
Design research
Funding proposal



Organizing
- Directory structure
- Formats, names, versions

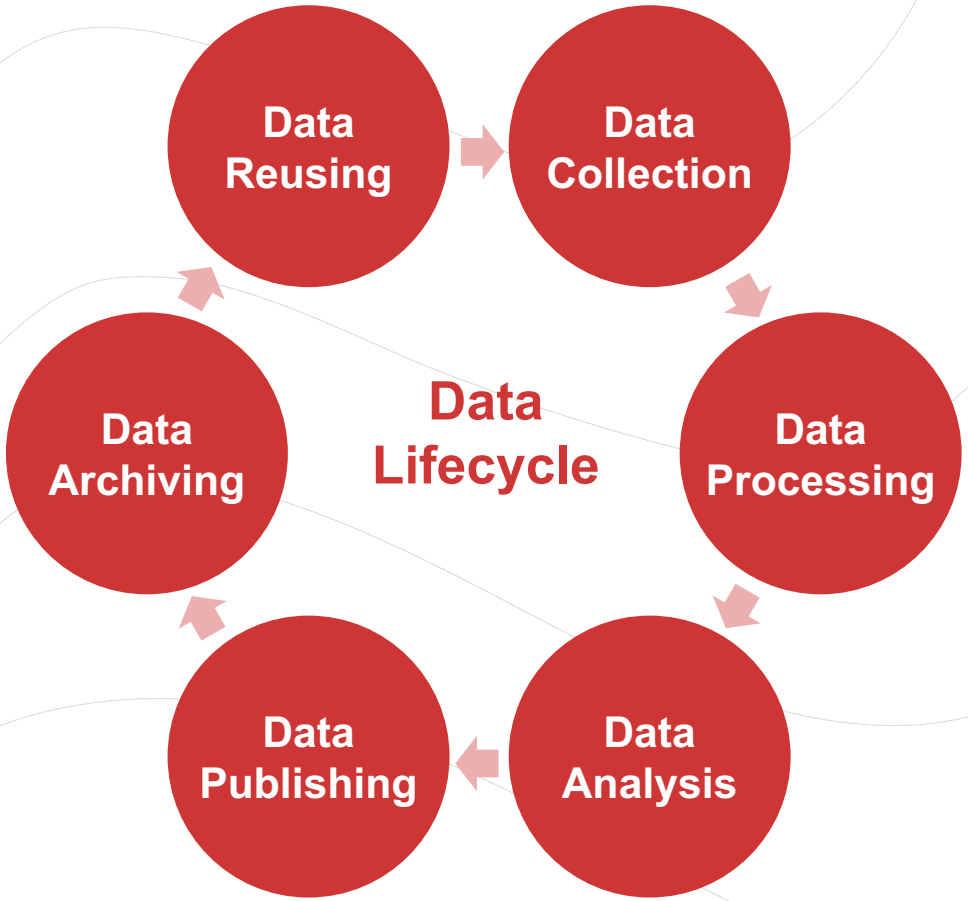
Documentation
- Data description
- Experimental details
- Decisions made
- Metadata

Storage
- Backup
- Long-term preservation

Data access
- Access rights (open, restricted)
- Licenses

Examples of research data requirements and policies

Plan
Generate ideas
Design research
Funding proposal



Funding agency policies

Open Access policy
Data management plan

Legal and ethical requirements

National and European legislation
Ethical framework for researchers
Personal data protection
Intellectual property rights
Commercial use of data

Institutional policies

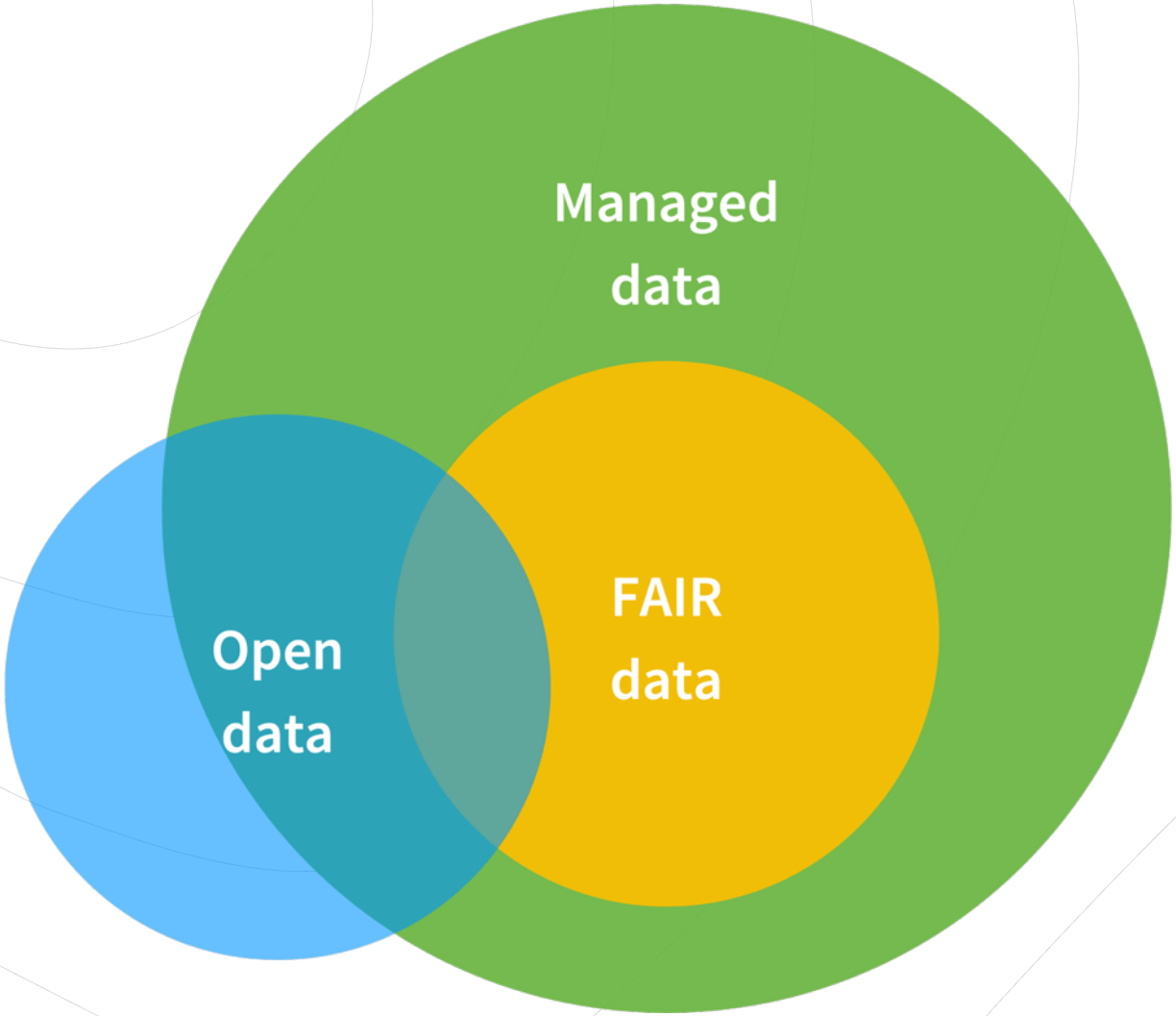
RDM policy
Codes of conduct and ethics
Data protection
Partnership agreement (for collaboration)

Journal & Publisher policies

Data sharing policy

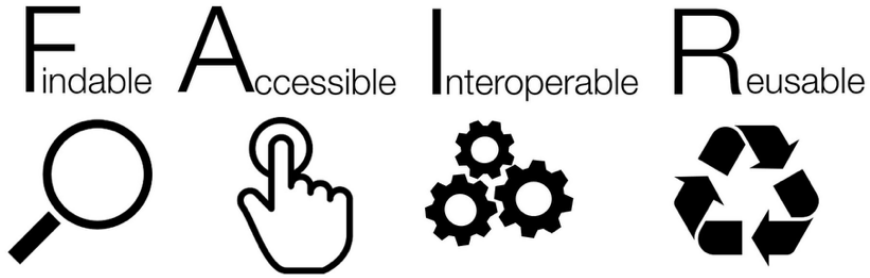
Open vs. FAIR data

Open vs. FAIR Data



Source: Valis, Jan (2024). FAIR vs. Open Data (Venn diagram). doi.org/10.6084/m9.figshare.25648548.v1; License: CC BY

FAIR principles



Source: SangyaPundir, FAIR data principles, CC BY-SA 4.0

Findable

- Metadata
- Persistent Identifiers (DOI, ORCiD, ROR, IGSN…)
- Registration and indexing in searchable repository

Accessible

- Free and open metadata
- Metadata available even when data are not available

Interoperable

- Widely used language
- Preferred formats
- Vocabularies and ontologies

Reusable

- Rich description (Read Me File)
- License
- Field/Community standards

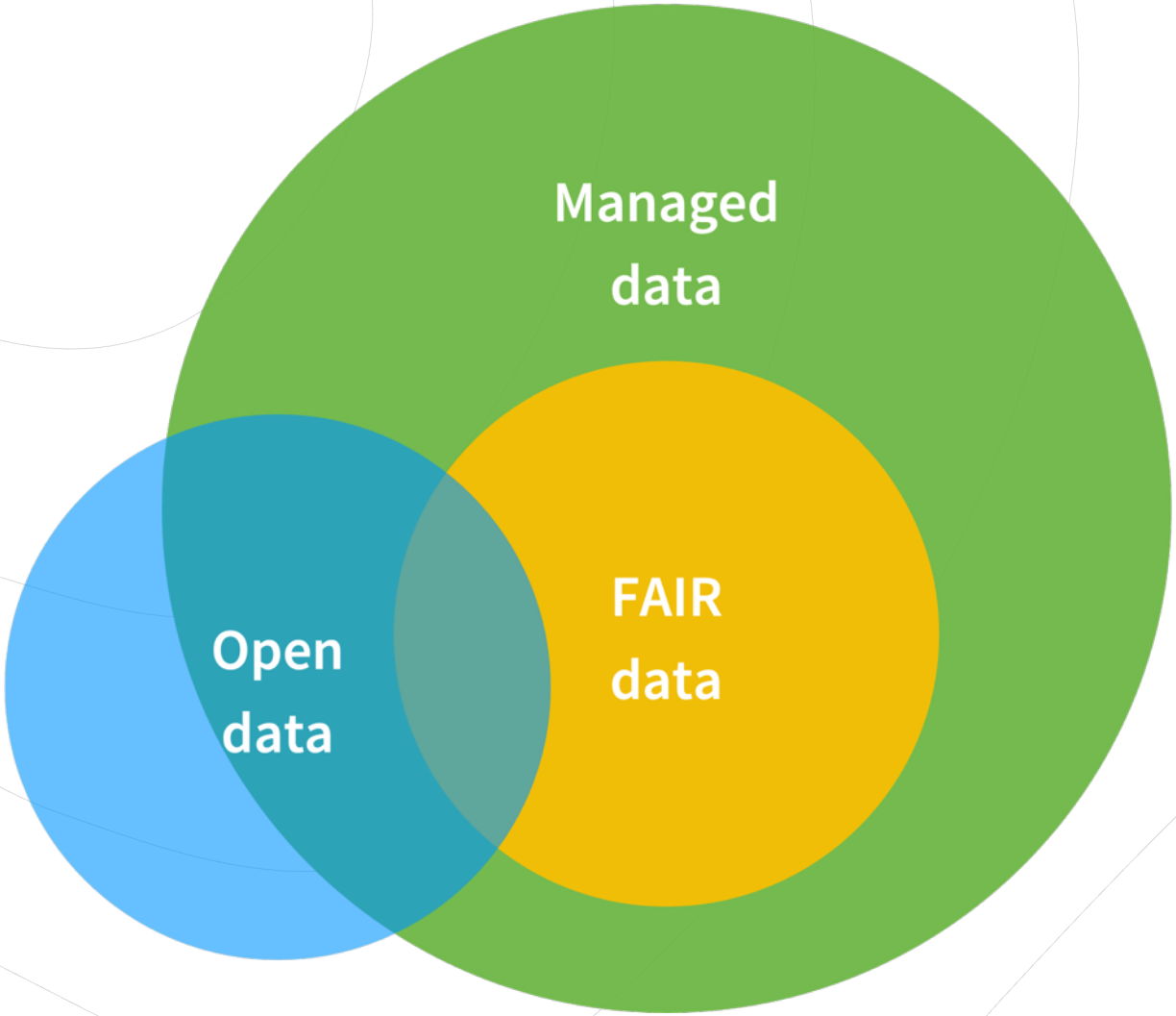
What (not) to publish?

- Personal data
- Sensitive data
- Protected by legitimate interest:
  - Intellectual Property
  - Commercial interests

What to do with these data?

- Can the data be
  - anonymized?
  - shared with informed consent?
  - shared later (embargo)?
  - shared only with selected researchers (controlled access)?
- If not, we can still usually share at least **metadata.**

Open vs. FAIR Data



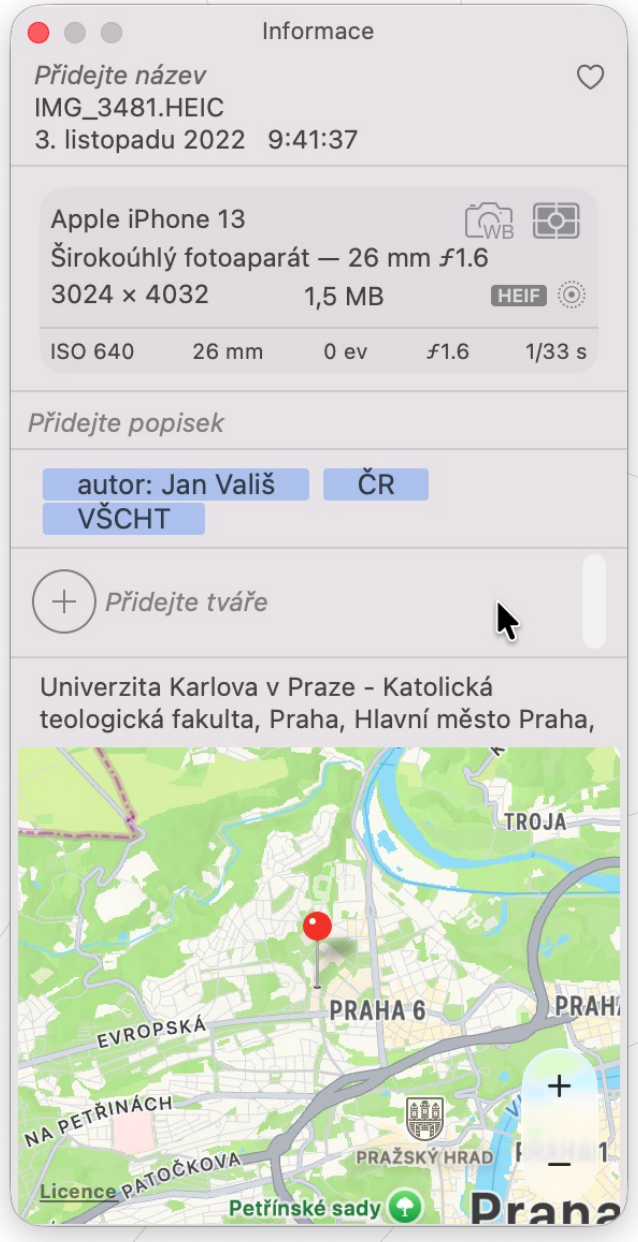
Source: Valis, Jan (2024). FAIR vs. Open Data (Venn diagram). doi.org/10.6084/m9.figshare.25648548.v1; License: CC BY

Metadata

Metadata – theoretically

- Documentation describing data = **“data about data”**
- **Machine readability** → search & retrieval of data from repository.
- Metadata: who, what, when, where, why, how
- Different types
  - descriptive (title, author, date, keywords …)
  - administrative (file formats, owner, license …)
  - field-specific (spectral resolution, CAS number of molecule, species …)

Metadata – practically



Metadata Models

- Universal
  - Dublin Core
  - DataCite
  - …
- Field-specific
  - Open Geospatial Metadata Standard
  - Chemical Research Object Framework

DataCite

- Mandatory
  ID, *Creator, Title, Publisher, Year, Type*
- Recommended:
  Subject, Contributor, Date, Related ID, Description, Geo Location
- Elective:
  Language, Alternative ID, Size, Format, Version, Rights, Funding, Related Item

Questions 4 & 5

4) Select examples of research data:

- ❑ Infrared spectrum of a sample.
- ❑ Grant budget submitted to grant provider.
- ❑ Recording of a structured interview.

5) Select truthful statements regarding metadata:

- ❑ Metadata help in finding research data.
- ❑ Metadata should be machine-readable.
- ❑ Must be human readable.

Data Management Plan (DMP)

DMP – plan before project starts

Administrative

- People involved (qualifications, training, roles & access, ORCiD)
- Requirements (institutional, funder …) & Support
- Budget

Instruments

- Access?
- Documentation?

Software

- Capture/processing/analysis workflows?
- Access to proper software?
- Use of open file formats?

DMP – plan before project starts

Size

- Enough storage for (captured/processed/analyzed) data?
- Many small files or fewer large files?

Backup

- How and where? (HDD, NAS, on/off-campus server)
- Encryption and access control?

Archiving

- What to archive?
- For how long?

DMP – plan before project starts

Legal & Ethical aspects

- Collaboration and services
- Personal/sensitive data
- Ethical committee approval?
- Informed consent?

Copyright License

- How are we legally bound?
- How do we want to license our results?

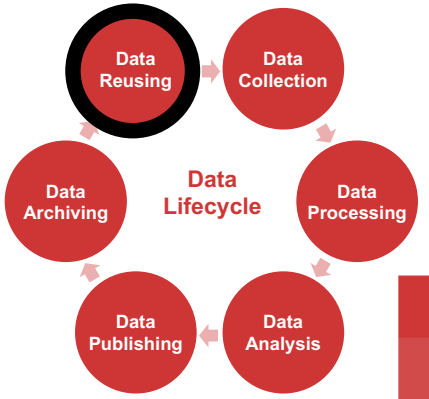
Publishing

- Can we publish data?
- Is there any domain-specific repository?

Tools to help you with DMP

- Data Stewardship Wizard (or FAIR Wizard, e.g. CAS)
- DMP Online
- Argos

Data Reuse



Benefits

- Resource efficiency (time, money, equipment)
- Access to hard-to get data
- Larger datasets

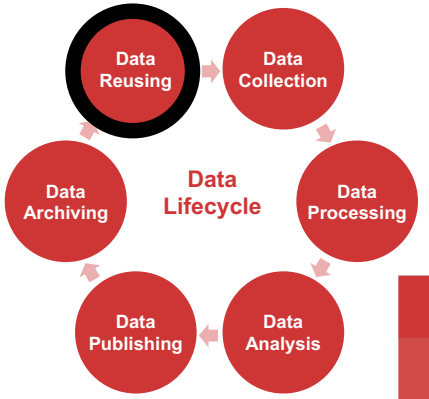
Perils

- Different collection methodologies
- Quality?
- Incompatible formats
- Ethical aspects

Solutions

- Implementation of FAIR principles
  - Community standards
  - Preferred formats
  - Unified ontologies/controlled vocabularies
  - Permissive licensing
- Data harmonization
  - Cleaning
  - Processing
  - Transformation
  - Normalization

Data Reuse



Good practice

- Use of persistent identifiers (DOI, IGSN)
- Use of ethically/legally indisputable
- Reusing established workflows
- Reusing data from other filed

Questionable practices

- Assuming data are correct
- Use without/in breach of license

Reference datasets

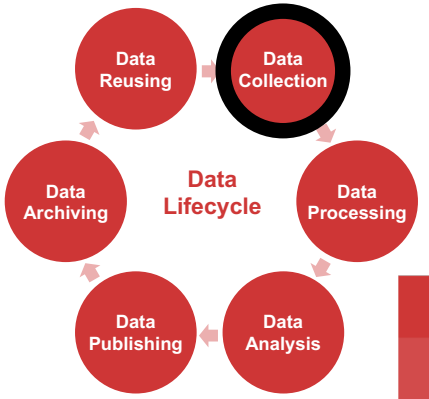
- Trusted data
- Peer-reviewed, validated

(e.g. NIST Webbook of chemistry)

Non-reference datasets

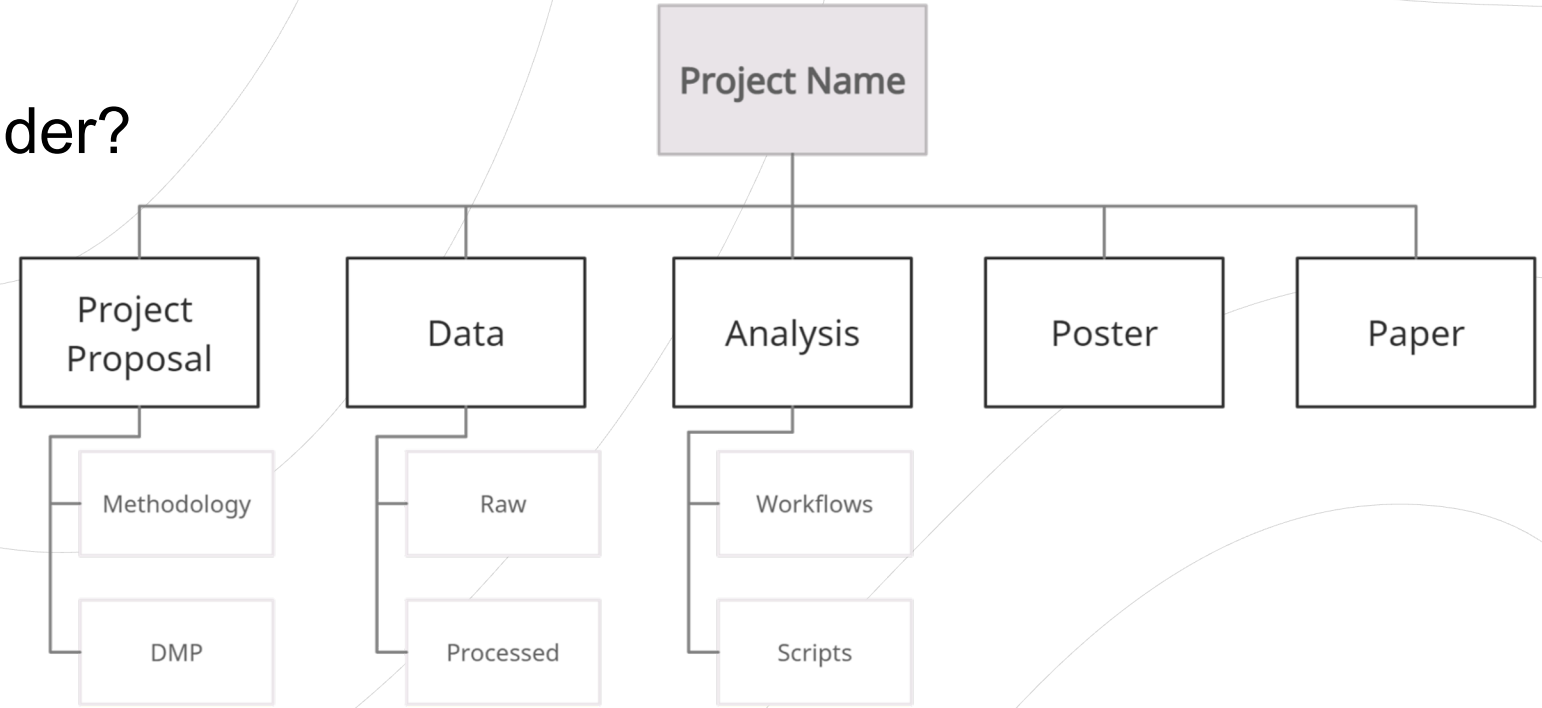
- Experimental/observational data
- Variable quality

Data Collection



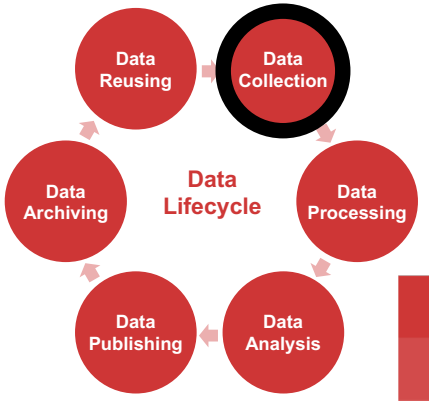
Organization

- Max 4 levels
- Max 10 subfolders to a folder?
- Documentation
  - ReadMe file
  - Codebooks
  - Guides
  - Manuals



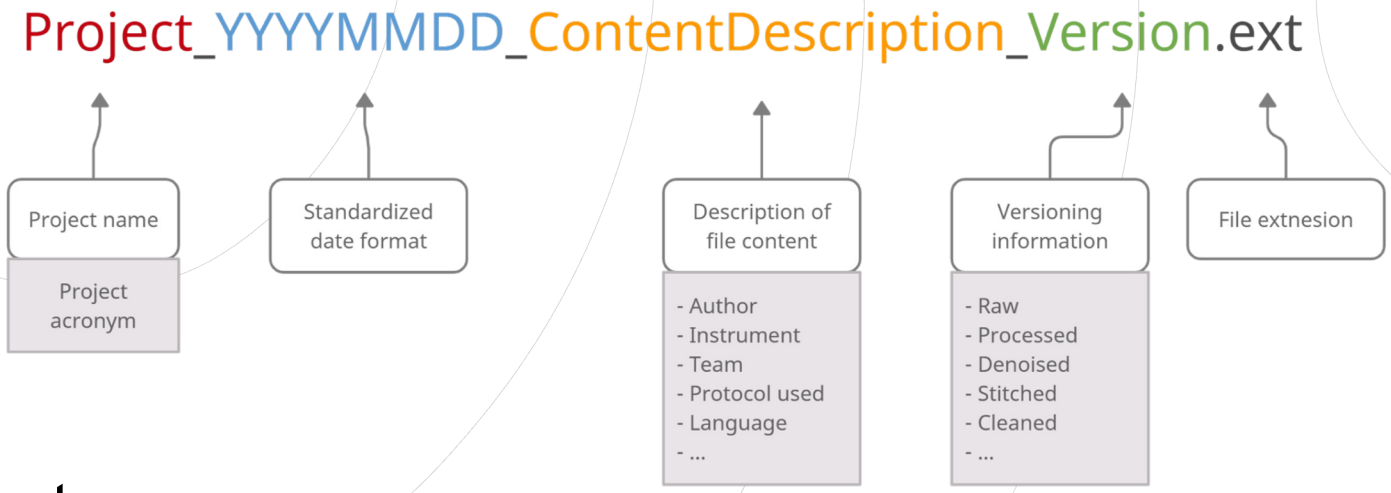
Source: https://biblio.uottawa.ca/en/services/faculty/research-data-management/file-naming-and-organization-data, CC BY

Data Collection



Organization

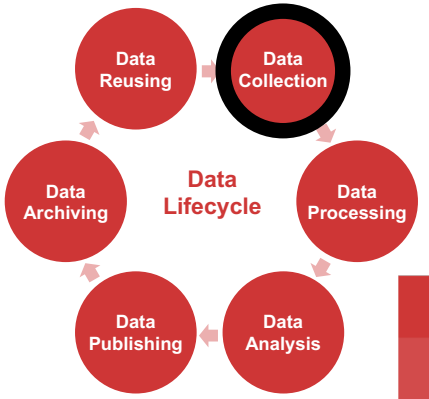
File naming



- Consistent
- No diacritics
- No special characters
- Use
  - hyphens (-) and underscores (_) for separation
  - abbreviations + codebook in ReadMe file
- Include
  - metadata (language, author, instrument, parameters …)
  - transformation stage (raw, processed, analyzed …)

Edited from: biblio.uottawa.ca/en/services/faculty/research-data-management/file-naming-and-organization-data, CC BY

Data Collection



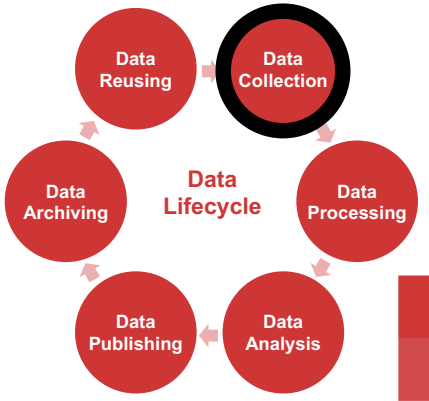
Organization

File naming

Metadata

- Not only for the whole dataset, but also for files
- Electronic Laboratory Notebook
  - Keeps provenance of sample and measurements
  - Keeps metadata structured
  - Helps to keep data FAIR and lowers the threshold for publication
  - Many options – from Google Colaboratory up to all-in-one solutions (e.g. LabFolder)

Data Collection



Organization

File naming

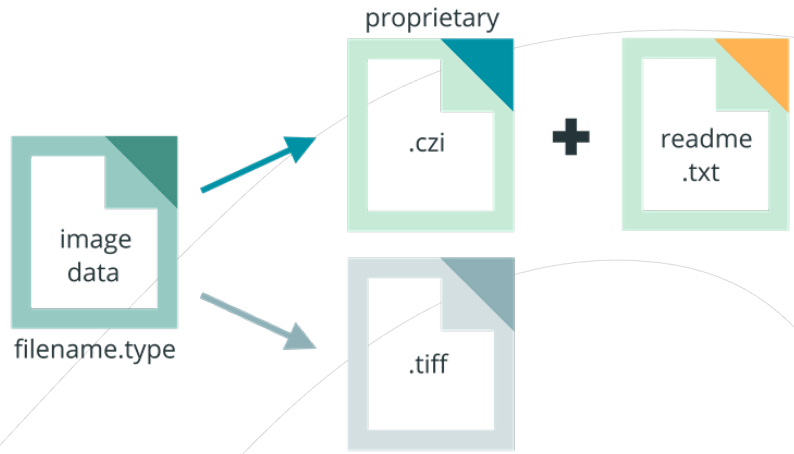
Metadata

File formats

- Preferred vs. popular
- Open vs. proprietary
- If proprietary necessary, include info in ReadMe on how to open/use the file (SW, workflow etc.)
- E.g. TIFF (~~JPEG~~), PDF/A (~~DOCX~~), CSV (~~XLSX~~)

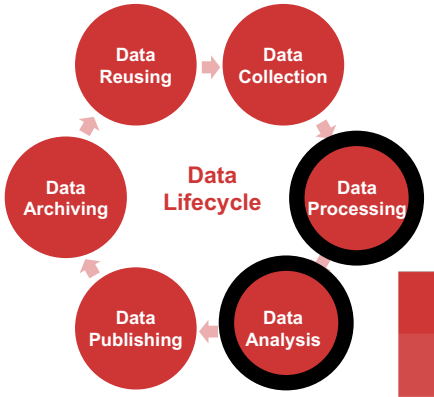
Preferred format

- archiving-friendly
- open
- well-documented
- human & machine readable
- lossless compression
- not dependent on a specific SW



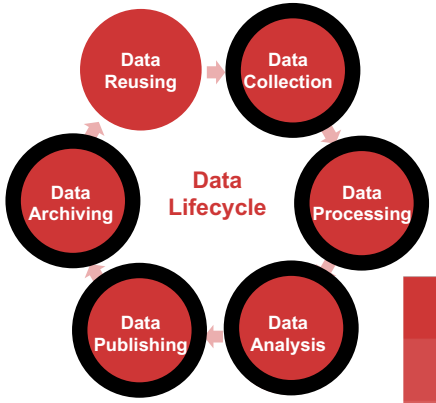
source: howtofair.dk/how-to-fair/file-formats/

Data Processing & Analysis



- Algorithmic workflows preferred (reproducibility, efficiency)
- Always work on a copy: original raw data need to stay untouched
- Minimize work with sensitive data
  (e.g. split sensitive/personal (meta)data and reconstitute them only using a key available to a small group of people)

Data Storage



By stage

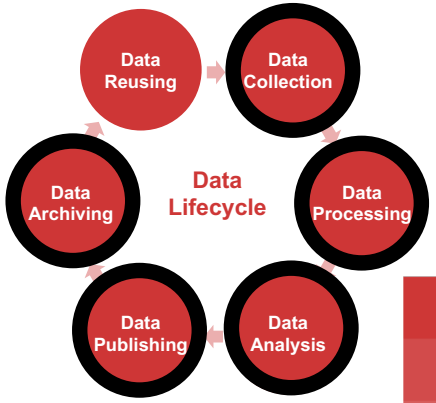
- During collection/processing/analysis (on-campus/off-campus)
- After analysis – **publishing** (e.g. repository)
- After the project finishes – **archiving** (locally, repository, or cold-storage?)

Security aspects

- Encryption
- Training
- Access restriction (by role, read/write etc.)

Financial aspect

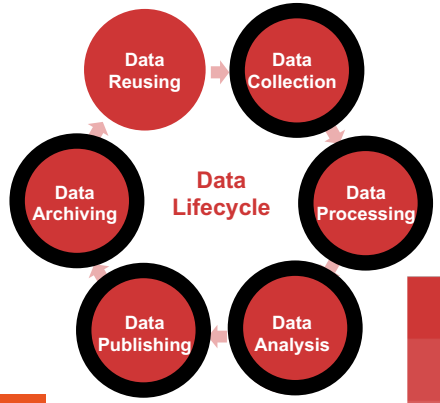
Question 6: How much should your budget be if you need to store 1 TB of data for 10 years?



Assume that a large number of small files is going to be generated. And even though the data will only be archived at the storage facility, the researcher wants to be able to retrieve the whole dataset in one day. As a good custodian, the researcher requires 24/7 support and will tolerate only 1 day of downtime per year. Due to the sensitivity of the data, controlled access will be required

A. **10 000 CZK**
B. **50 000 CZK**
C. **100 000 CZK**
D. **500 000 CZK**
E. **1 000 000 CZK**
F. **5 000 000 CZK**
G. **10 000 000 CZK**

Data Storage – price estimation



DSW Storage Costs Evaluator

Total costs:	TB costs per year:	Result details
2 261 €	**452** €	

Volume

500 GB

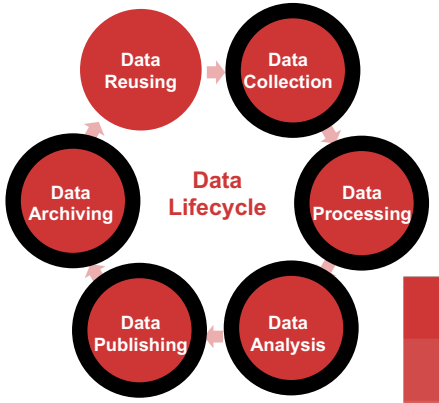
Lifetime

10 years

Detailed storage properties

Source: storage-costs-evaluator.ds-wizard.org/

Data Storage – price of AWS



Standard

First 50 TB/Month	$0.023 per GB →	13 517$ per year
Next 450 TB/Month	$0.022 per GB →	121 651$ per year
Over 500 TB/Month	$0.021 per GB →	129 024$ per year

Archive

Archive Access Tier — All Storage/Month $0.0036 per GB

100TB → 4 424$ per year

Deep Archive Access Tier — All Storage/Month $0.00099 per GB

100TB → 1 217$ per year

Data Publishing

License

- Exclusive
- Non-exclusive
  - SW specific (GPL, MIT, Apache)
  - Creative commons (CC)

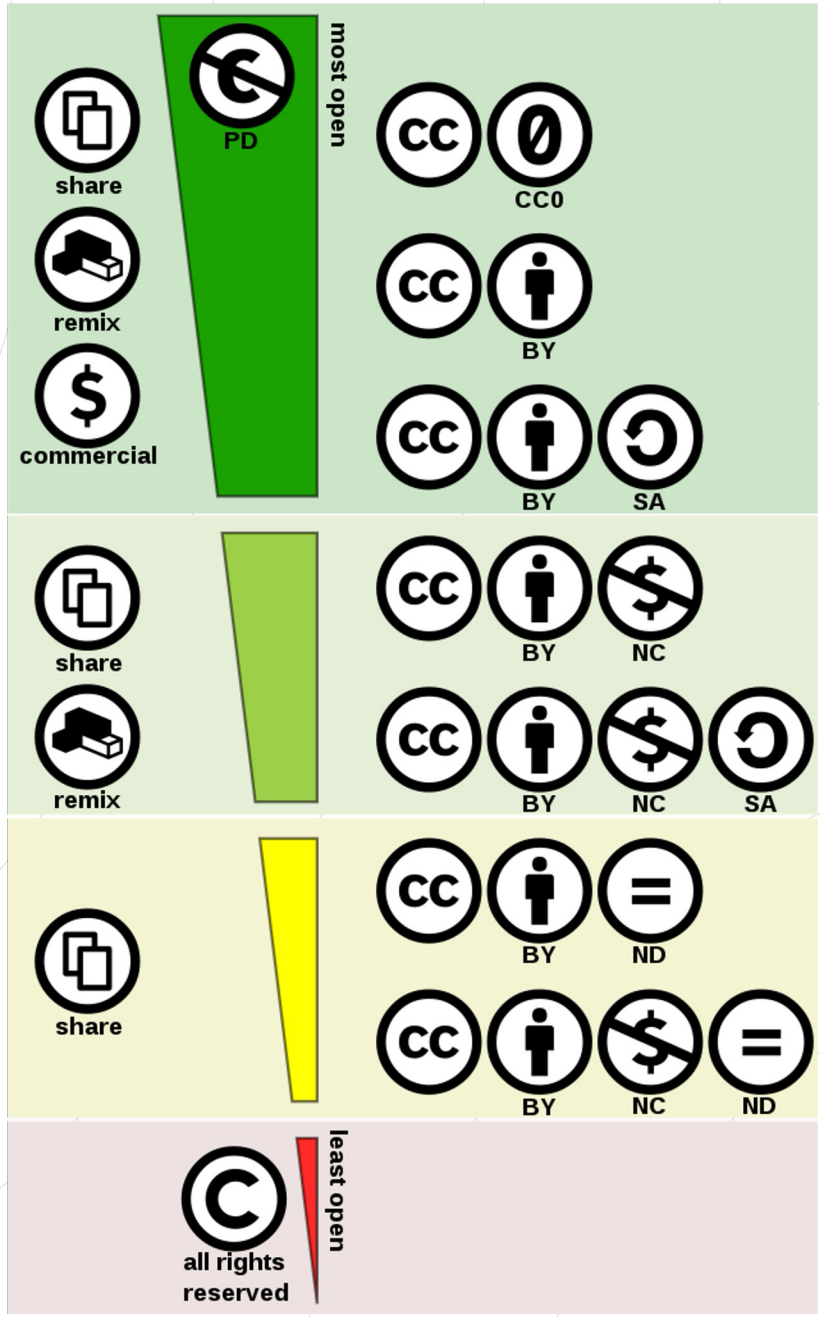
0 - Public domain

BY - By Attribution

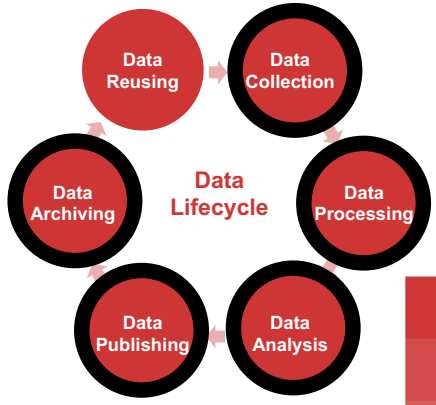
ND - No Derivatives

NC - Non-Commercial

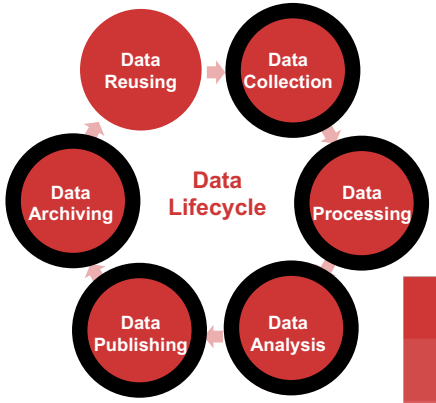
SA - Share Alike



Source: creativecommons.org/



Data Publishing



License

Repository selection

- Institutional/national/generic/field-specific
- Persistent identifiers (DOI, ORCiD, ROR…)
- Metadata – generic (e.g. DataCite) or field specific? Funding/grant field?
- File size/number limitations
- File format compatibility
- Access control/embargo
- Licenses
- Reputation, persistence, sustainability?
- Price

Where to look for?

- All-purpose aggregators (e.g. **re3data**, **FAIRsharing**)
- Societies and grant agencies (e.g. NFDI4Chem, ERC)

How to prepare a dataset and publish it?

Register for a brand-new workshop:

- 8 December 2025 (10:00 – 12:30)
- In-person at NTK
- Topics
  - Dataset preparation (organization, ReadMe file etc.)
  - Repository selection
  - Dataset deposition on Zenodo

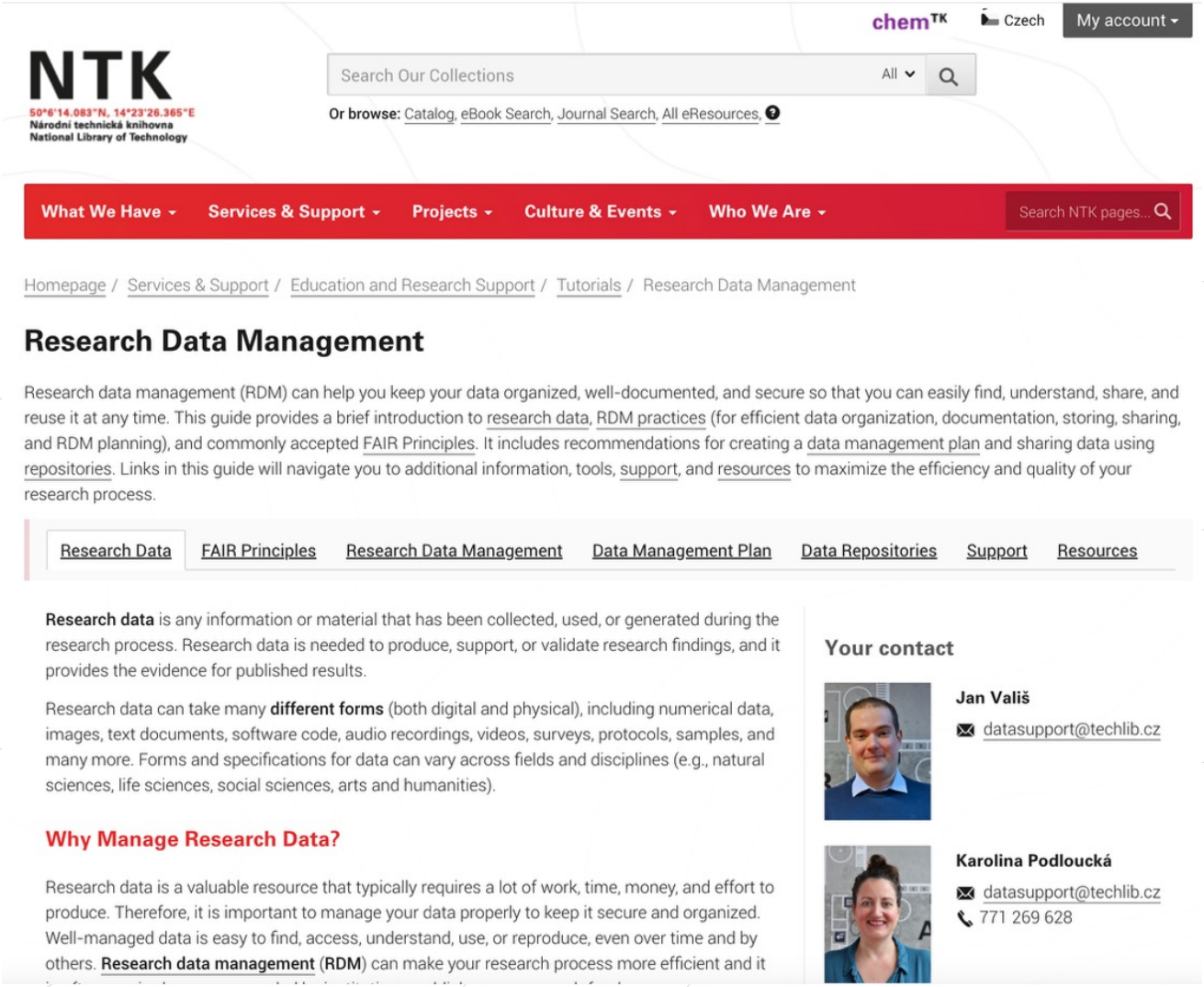
Where to learn more?

NTK

- RDM Guide
- Data Stewardship Course
  - 11 self-guided chapters
  - Finished by self examination & certificate

Self-study

- EOSC CZ – webinars (all)
- NPOS – Czech only (all)
- RDMkit (life-sciences)
- OpenAIRE (all)



Where to get help?

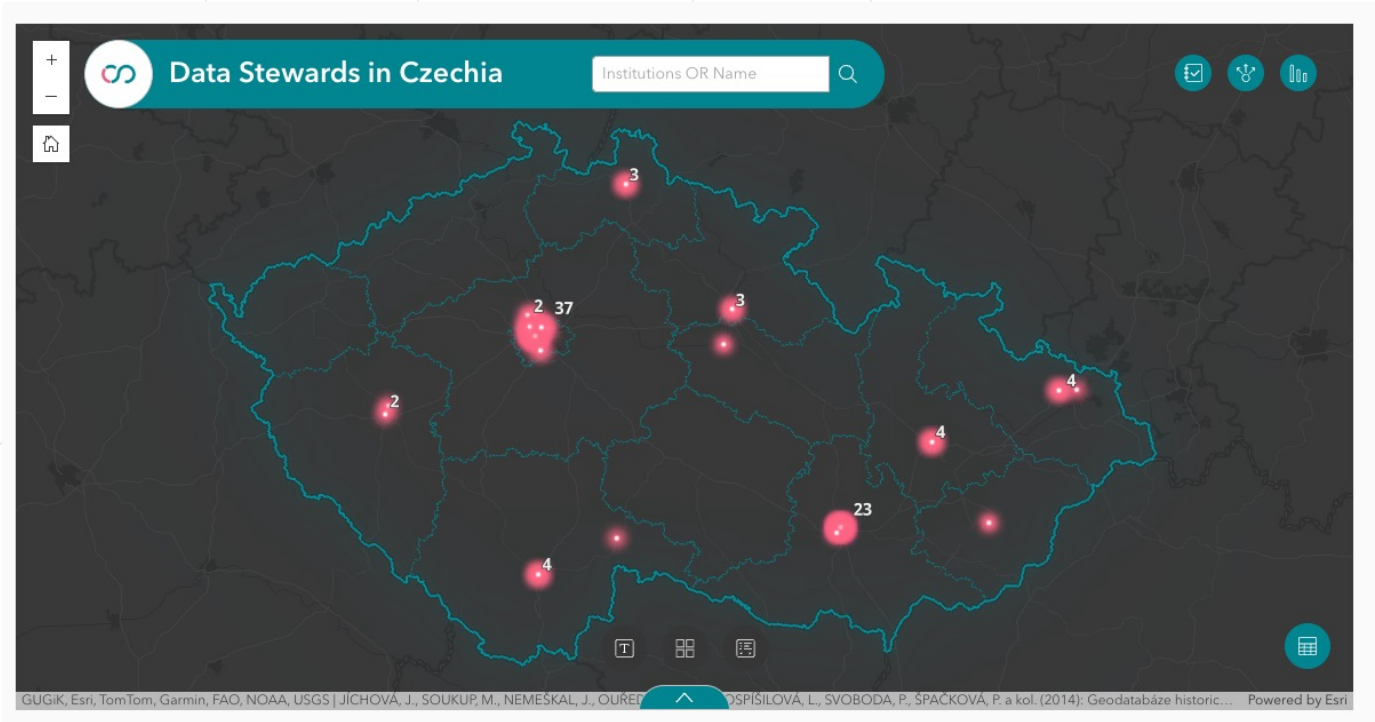
Community

- Map
- Manuals
- Meetings
- Discord

Your institution

- Librarian
- Data Steward

NTK: free consultations



What to take home?

NTK
50°6'14.083"N, 14°23'26.365"E
Národní technická knihovna
National Library of Technology

- Research data management:
  - is a helpful tool, not just an administrative task,
  - is a good scientific practice,
  - can save time and money and
  - ≠ Open Science, but sharing can create impact.
- Communities define standards and they are usually open to others joining their efforts. The same is happening at the national level.
- There are already many online resources to learn from.
- Funding providers already demand and financially support OS and RDM practices, including DMP.

NTK Information Support Team

NTK
50°6'14.083"N, 14°23'26.365"E
Národní technická knihovna
National Library of Technology

1) **Schedule a free consultation with us**

Don't be shy; our team includes doctoral candidates who understand the issues you face.

2) **Attend another webinar**

3) **Explore on your own:** Tutorials, AI tools for research or STEMskiller

4) **Subscribe to our newsletter** for updates on resources, writing support, publishing, research evaluation, and training opportunities.

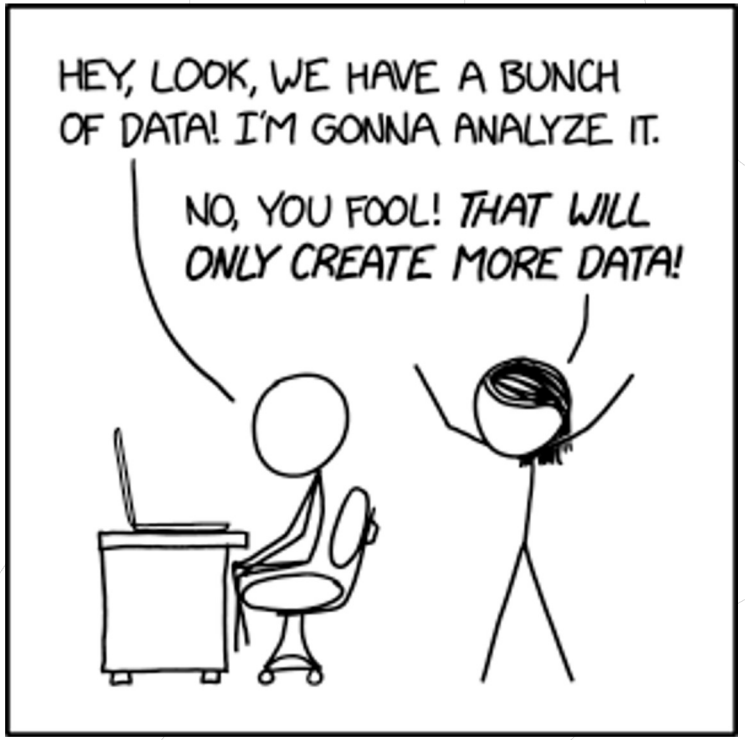
Any questions? Contact us at info@techlib.cz

Contact

Jan Vališ
jan.valis@techlib.cz

NTK
50°6'14.083"N, 14°23'26.365"E
Národní technická knihovna
National Library of Technology

Questions?



Source: https://xkcd.com/2582/;
Available via license: CC BY NC 2.5