

**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 lecturers: Adéla Jílková, Jan Vališ**

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

**October 30, 2024**

Content of this presentation is licensed via [CC BY 4.0](https://creativecommons.org/licenses/by/4.0/),  
except where otherwise noted for content created  
by third-parties.



# Agenda

## 1. What is research data and why manage it?

- Motivation and benefits of Research Data Management (RDM)
- Research data and RDM overview

## 2. How to approach Research Data Management?

- RDM frameworks (Open Science and FAIR principles)
- RDM strategies and techniques
- RDM plan

# **What is research data and why manage it?**

# Research data and Research data management

## Research data

- Any information **collected, observed, generated, or created** during the research process to produce and support research findings

## 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 can help:

### **Keep the research process organized, secure, and smooth**

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

### **Enhance global data sharing** (Open Science and FAIR principles)

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

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

# Research data

## Different fields and disciplines

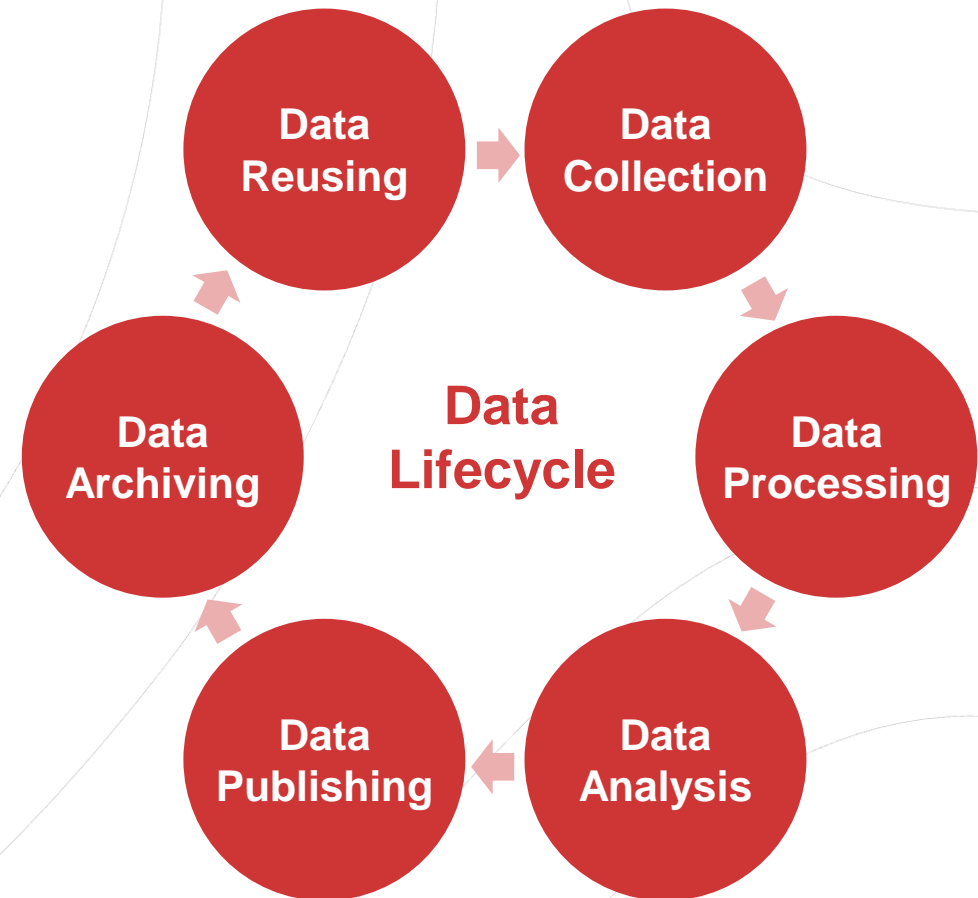
- Natural and life sciences
- Medical and health sciences
- Engineering and technology
- Social sciences
- Arts and humanities

# Research data

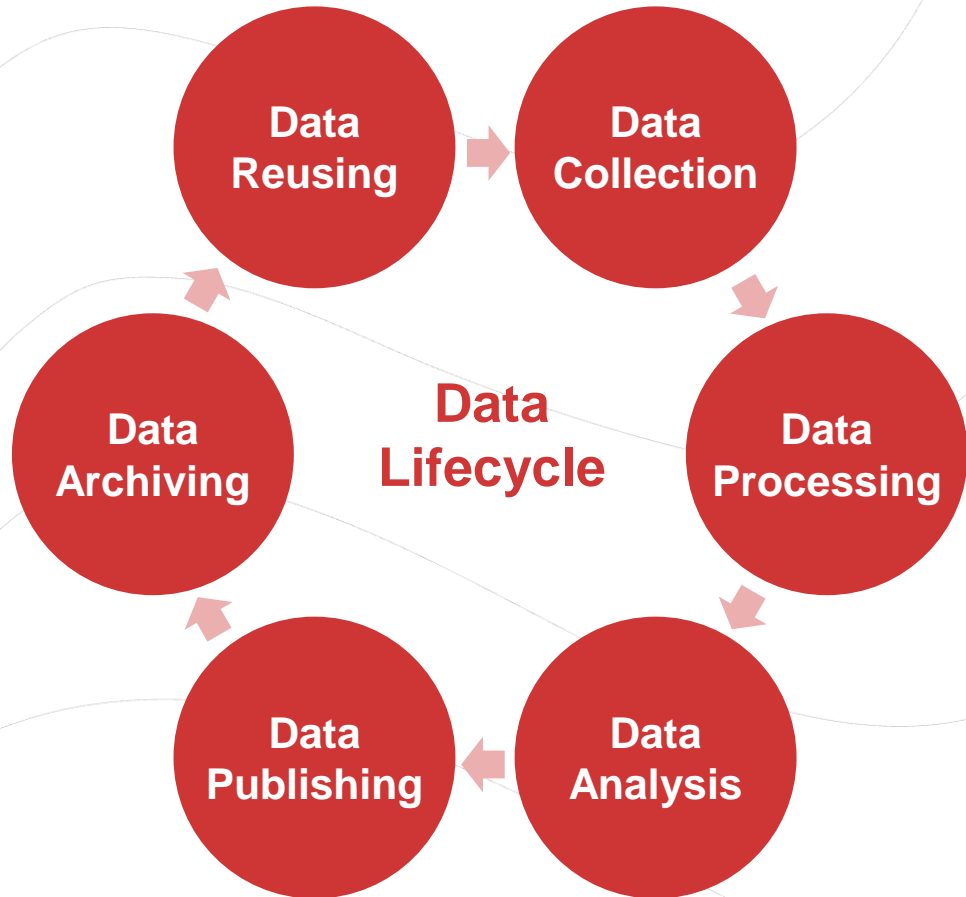
## Different fields and disciplines

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

## Different stages of research data lifecycle

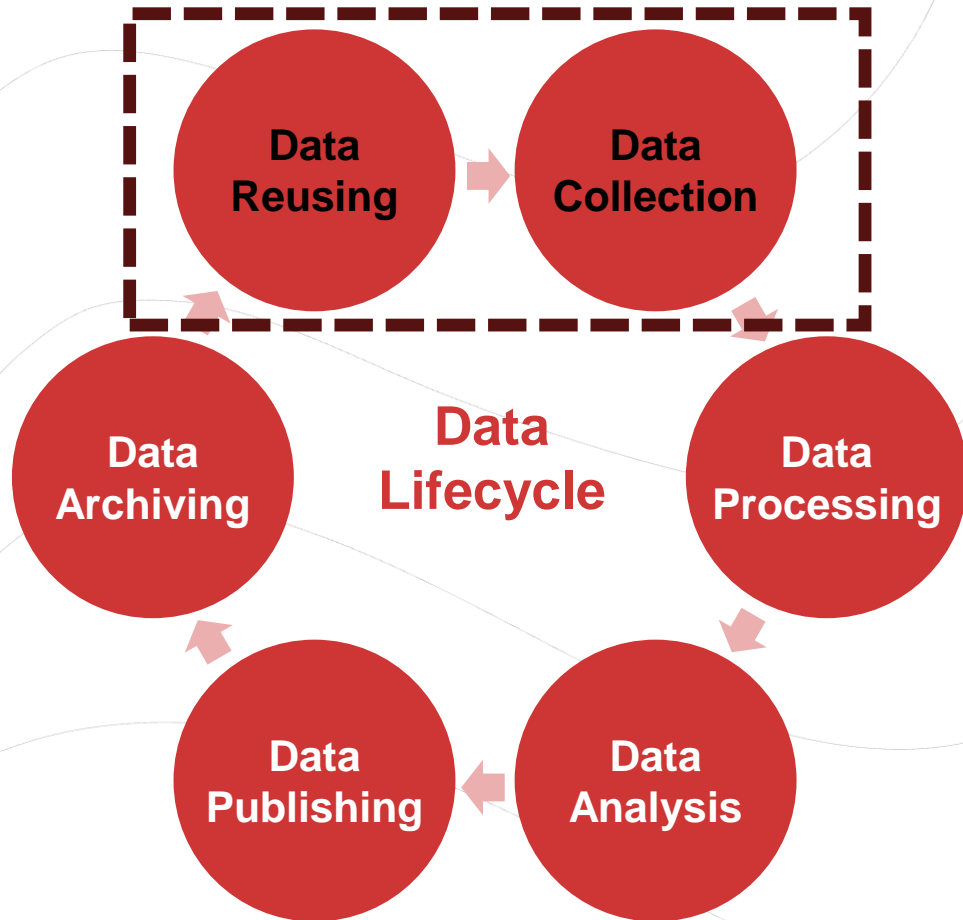


# Research data lifecycle





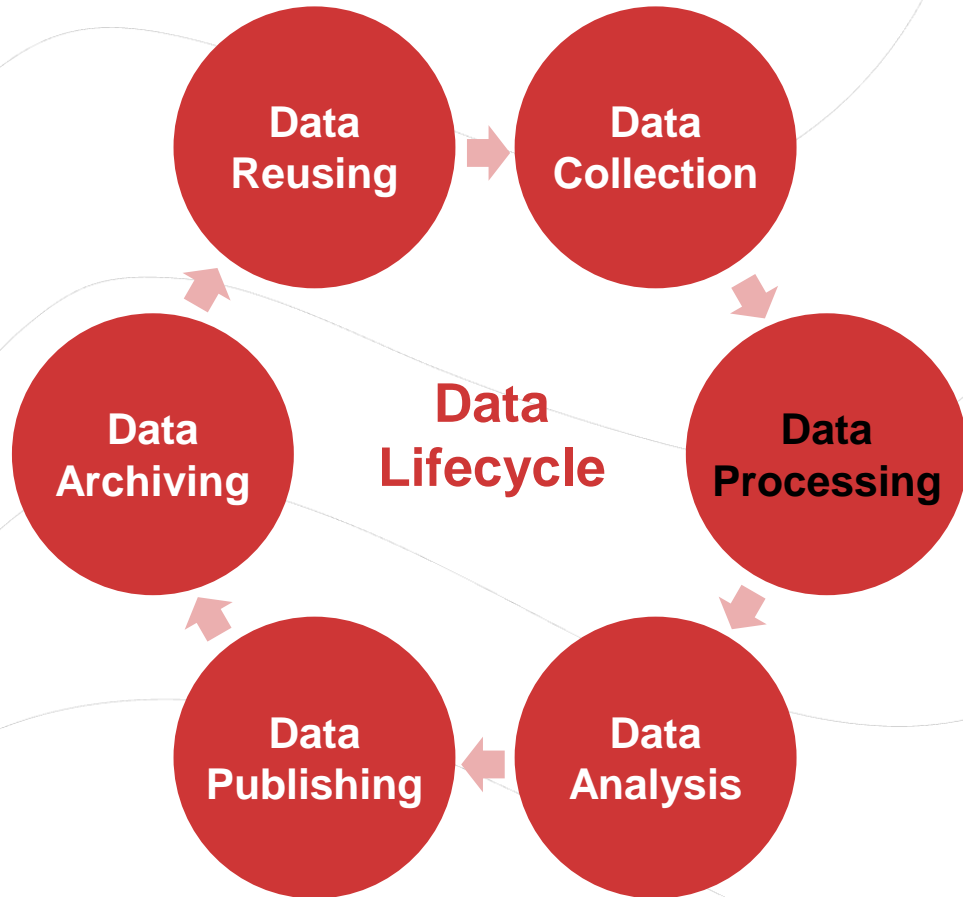
# Research data lifecycle



## Source Data

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

# Research data lifecycle



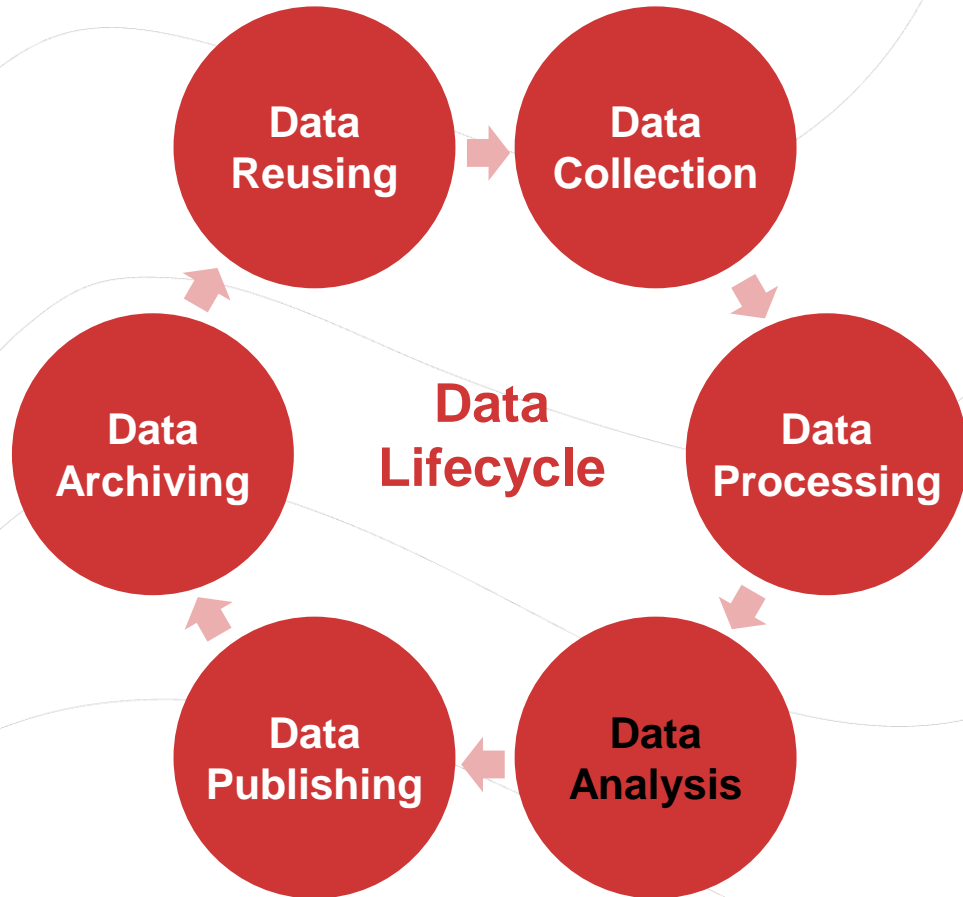
## Source Data

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

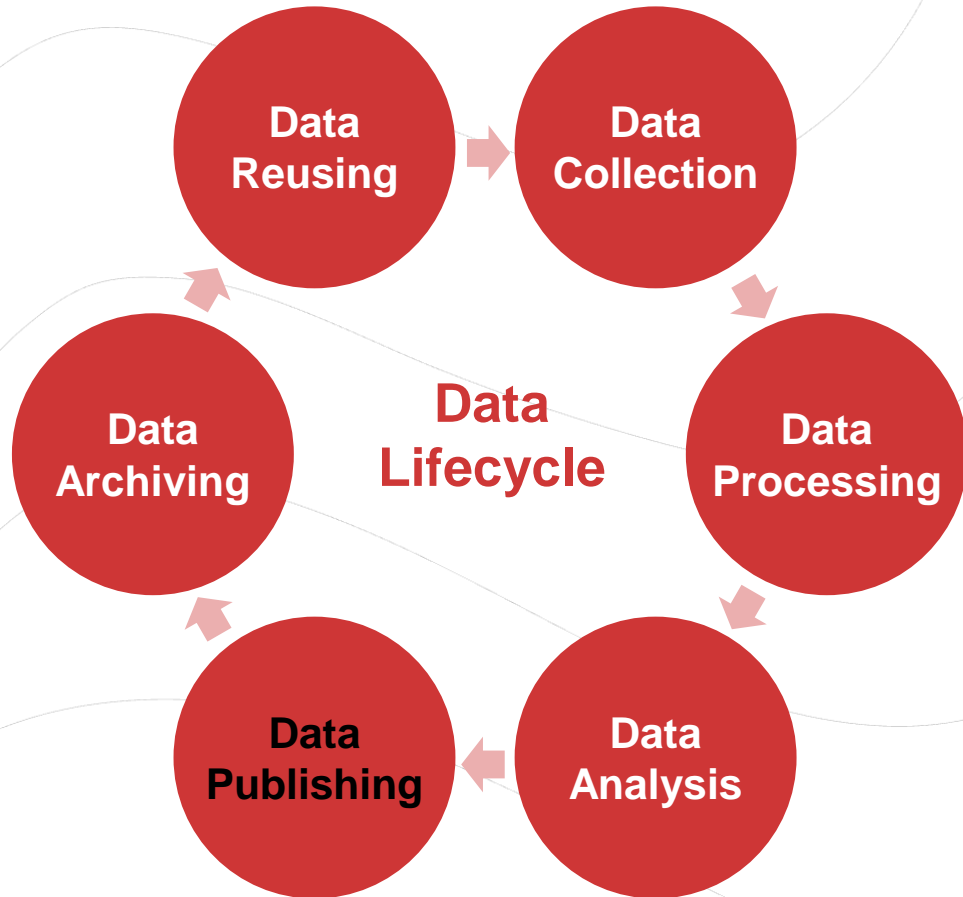
## Data Processing

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

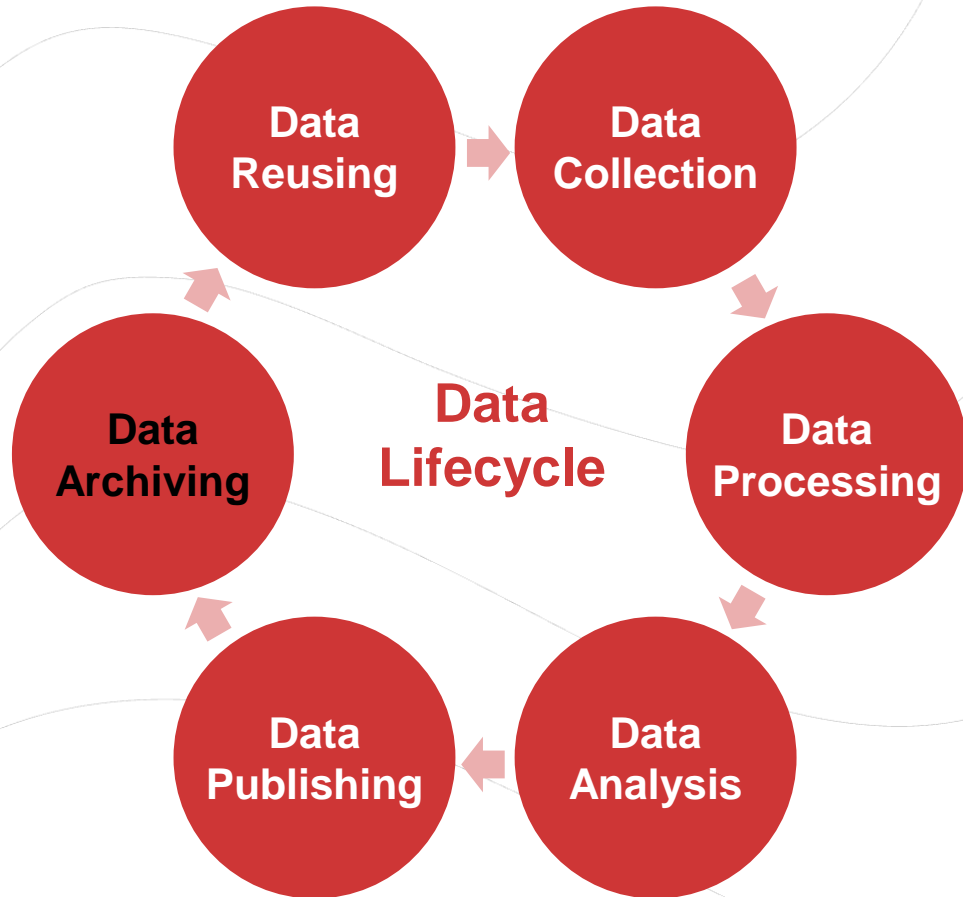
### Journal article

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

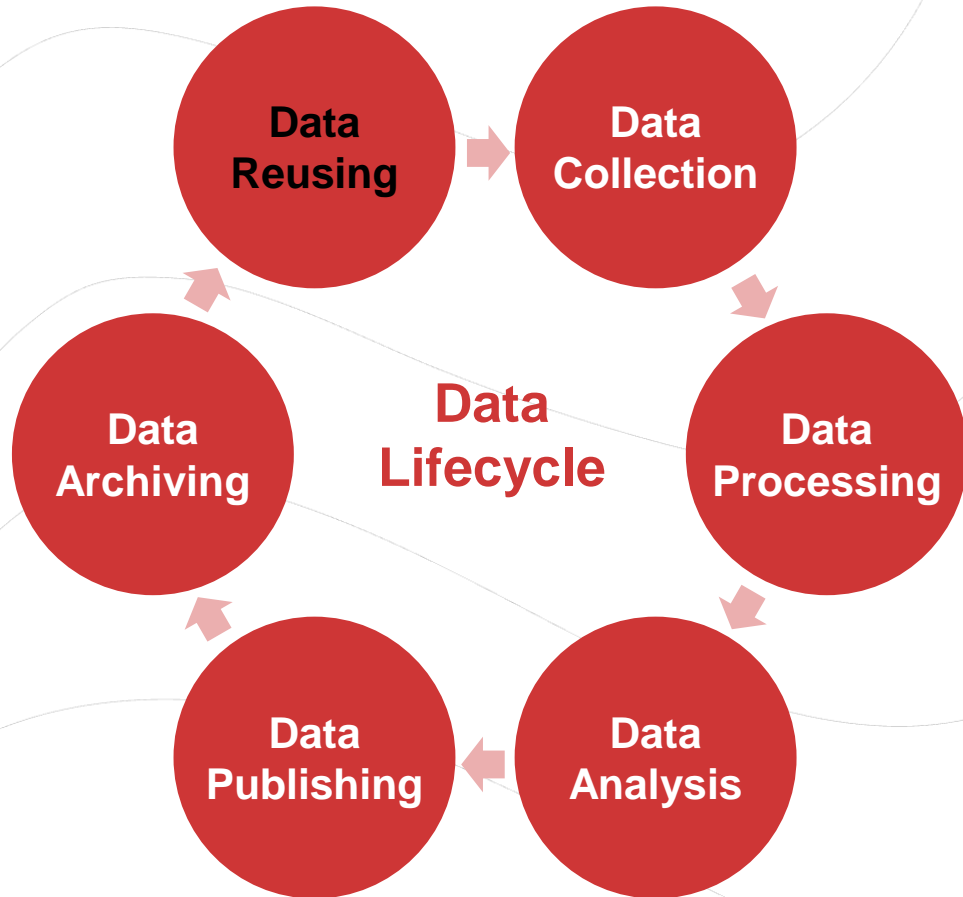
Manuscript + supplementary information

### 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

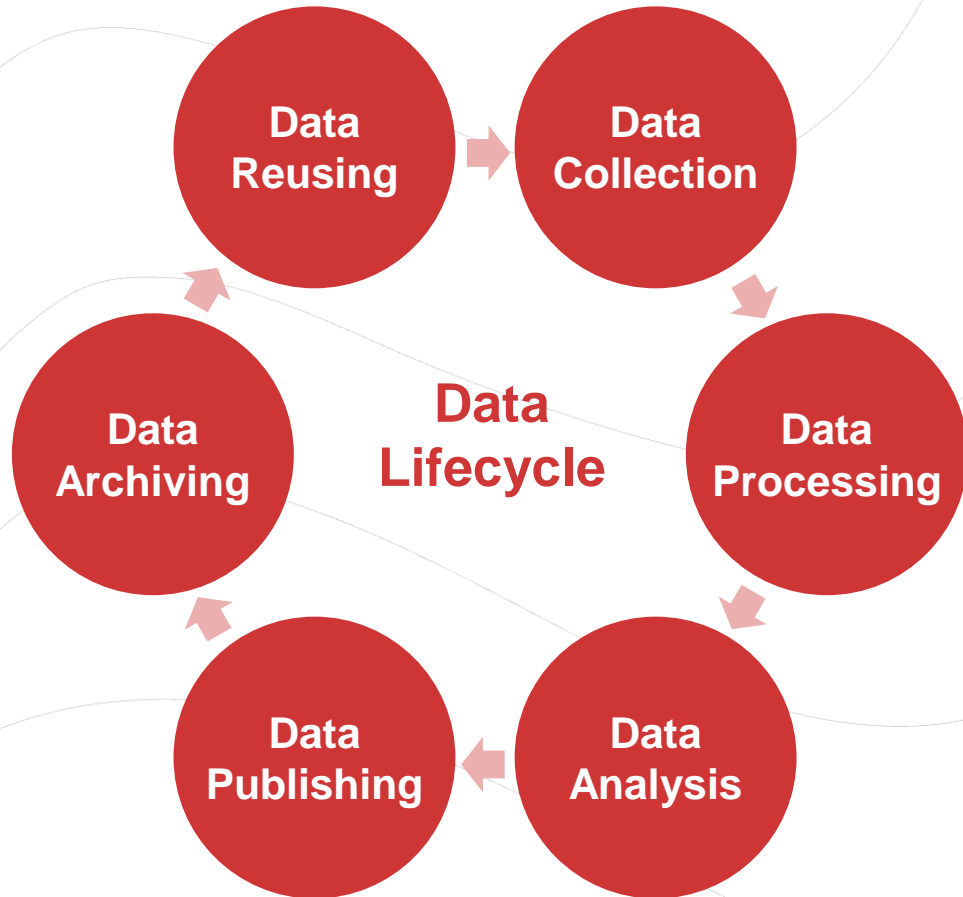
### Databases/repositories

Data underlying publication  
Separate datasets

## Data Archiving

## Data Reusing (registries, repositories)

# Research data management strategies



## Organizing

- Directory structure
- Formats, names, versions

## Documentation

- Data description
- Experimental details
- Decisions made
- Metadata

## Storage

- Backup
- Long-term preservation

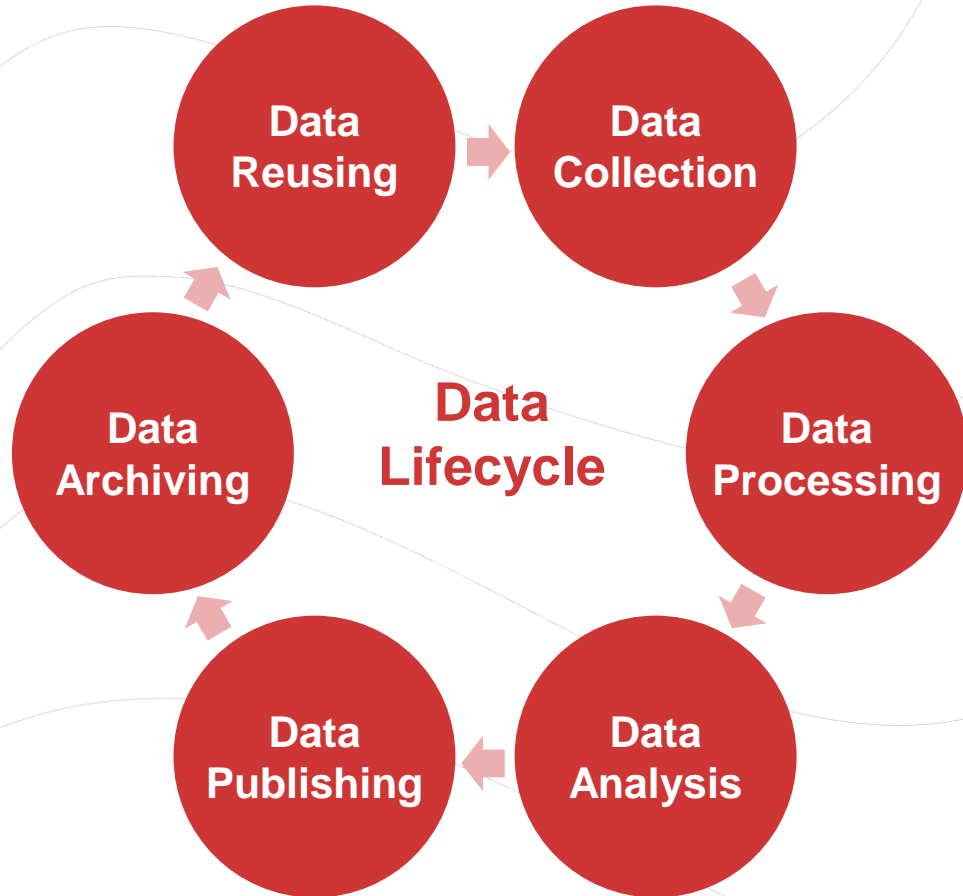
## Data access

- Access rights (open, restricted)
- Licenses

# Research data management strategies

## Plan

Generate ideas  
Design research  
Funding proposal



## Organizing

Directory structure  
Formats, names, versions

## Documentation

Data description  
Experimental details  
Decisions made  
Metadata

## Storage

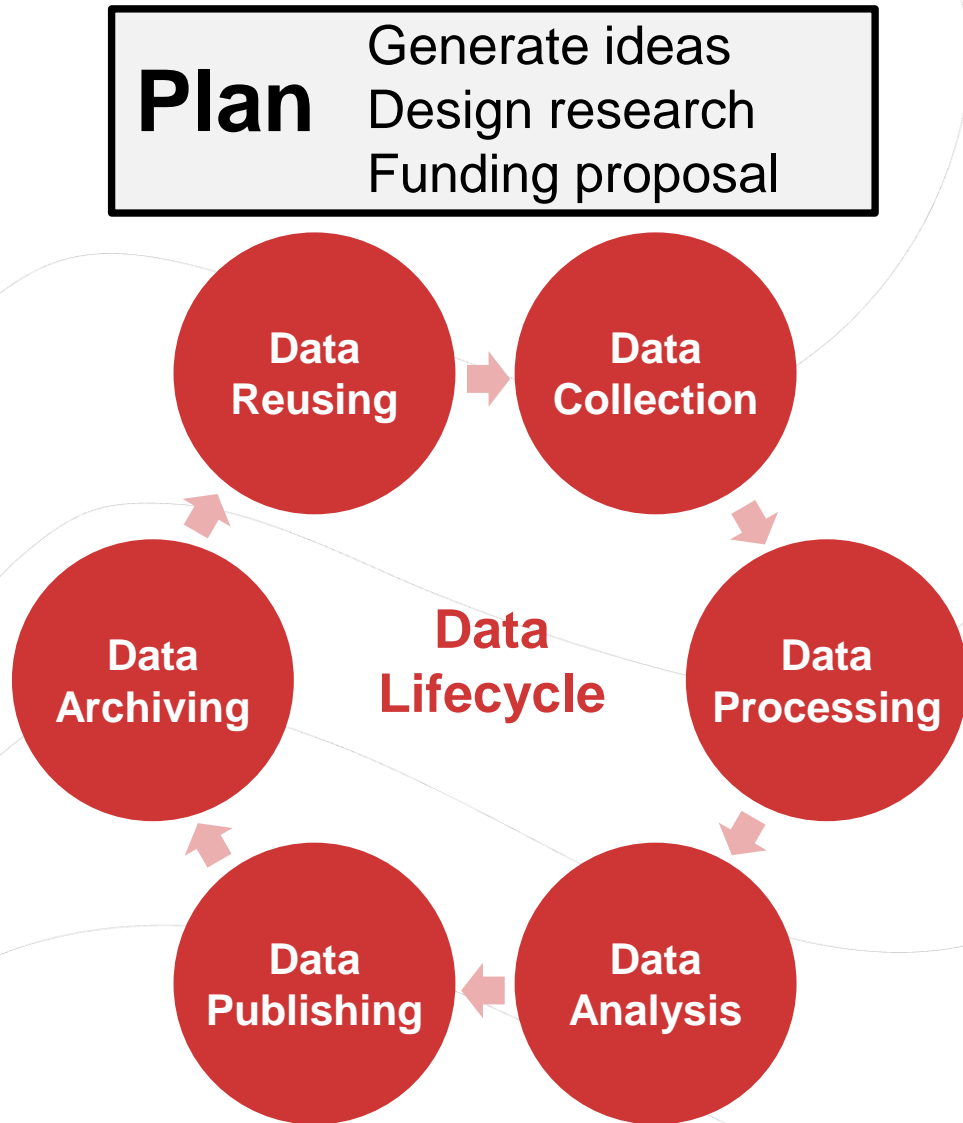
Backup  
Long-term preservation

## 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

Adapted from "Research Data Lifecycle" by UK Data Service, used under [Creative Commons Attribution license](#) (CC BY)

# How to approach Research Data Management

# What is data?

Anything containing information

Some might be self-explanatory:

- Text
- Tables

Other might not:

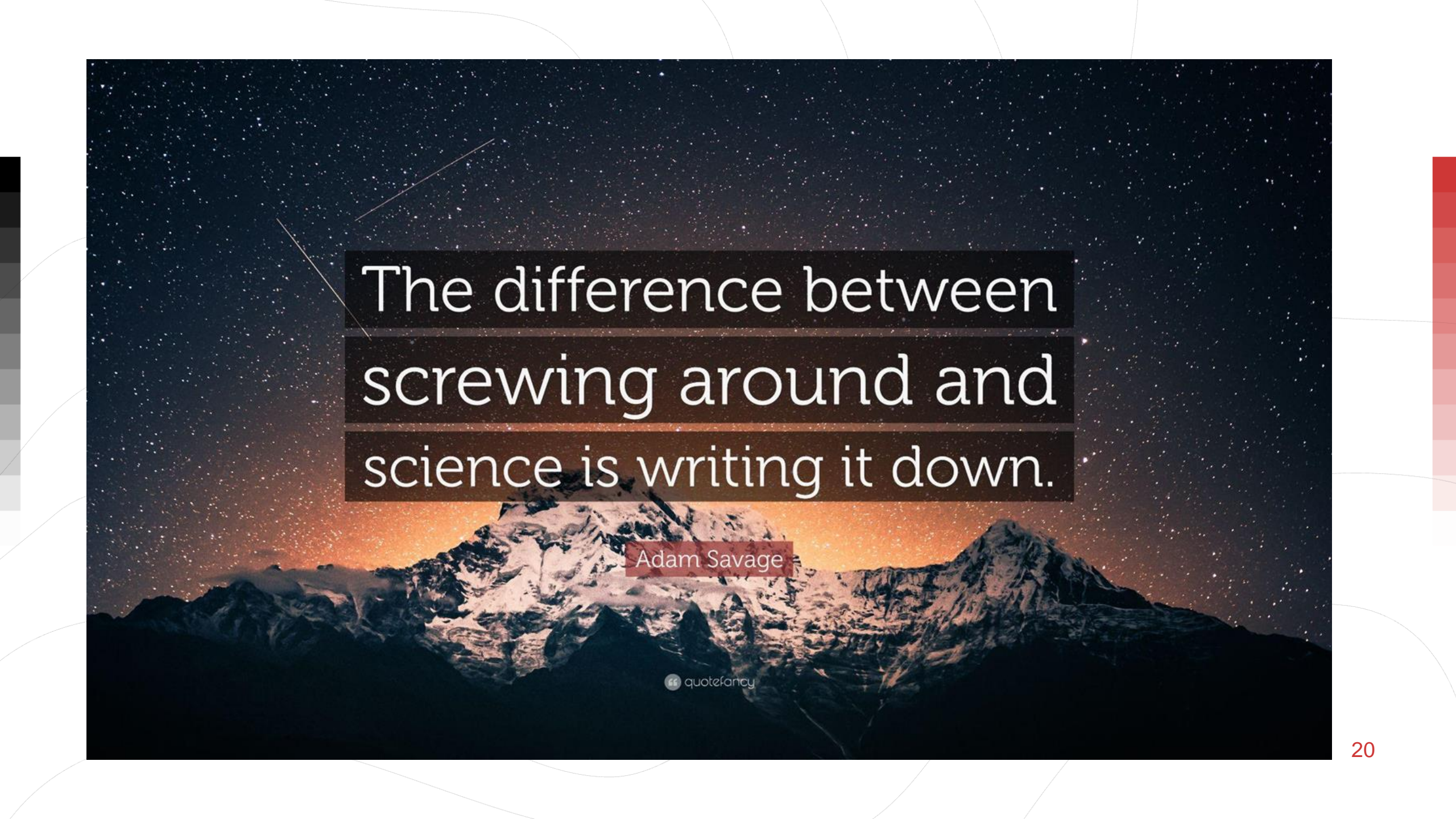
- Measurement results
- Images

Some might not be shared:

- Personal information
- Medical diagnoses

But there is always **metadata**:  
information (data) about data:

- Date of creation
- Author
- License
- Measurement device



The difference between  
screwing around and  
science is writing it down.

Adam Savage

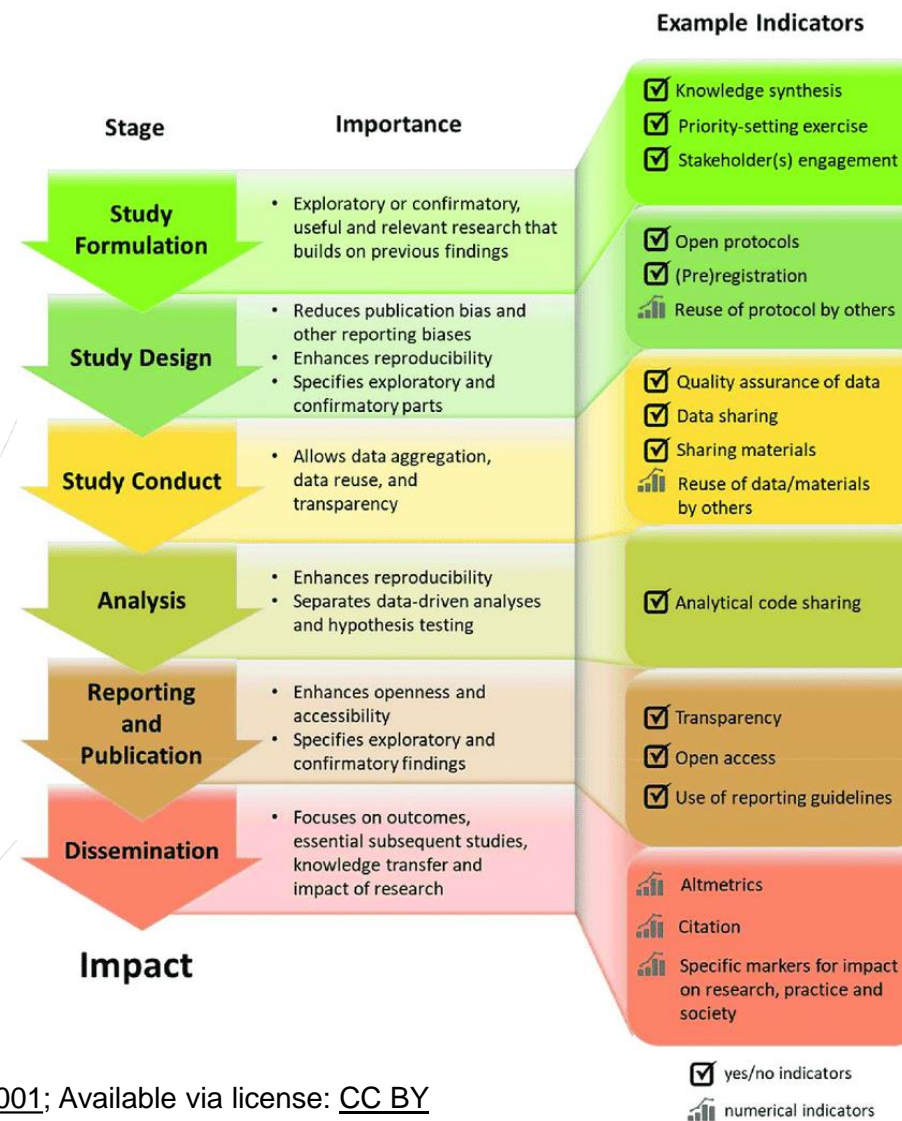
 quote fancy



# Responsible Research Practice

- For knowledge to benefit research and society, it must be trustworthy.
- Trustworthy research is robust, rigorous, and transparent at all stages of design, execution, and reporting.
- Assessment of researchers still rarely includes considerations related to trustworthiness, rigor, and transparency.

## Indicators of responsible research practices



# We need to plan in advance

- Instruments
  - Can we properly document what we are doing, and how?
- Size
  - Do we have enough storage?
- Software
  - Do we have workflow for processing of data?
  - Do we have access to proper software?
  - Can we use open file formats?
- Ethics
  - Are there any set procedures for data processing?
  - Collaboration and services!

# We need to plan in advance

- Backup
  - How and where?
  - Do we need encryption and access control?
- 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?
- Archiving
  - What data to archive?
  - How long?

## Open Science

### Revolution or evolution?





Creating more ways to improve inclusion and access to research and higher education

**Equity**

Research and education are transparent for validation, and all contributions are recognised

**Integrity**

**Open  
Science**

**Collaboration**

Exchanging knowledge and perspectives sooner and in every step, from ideation to communication

**Impact**

Open work is more visible and can be reused and adapted to build new research and educational materials

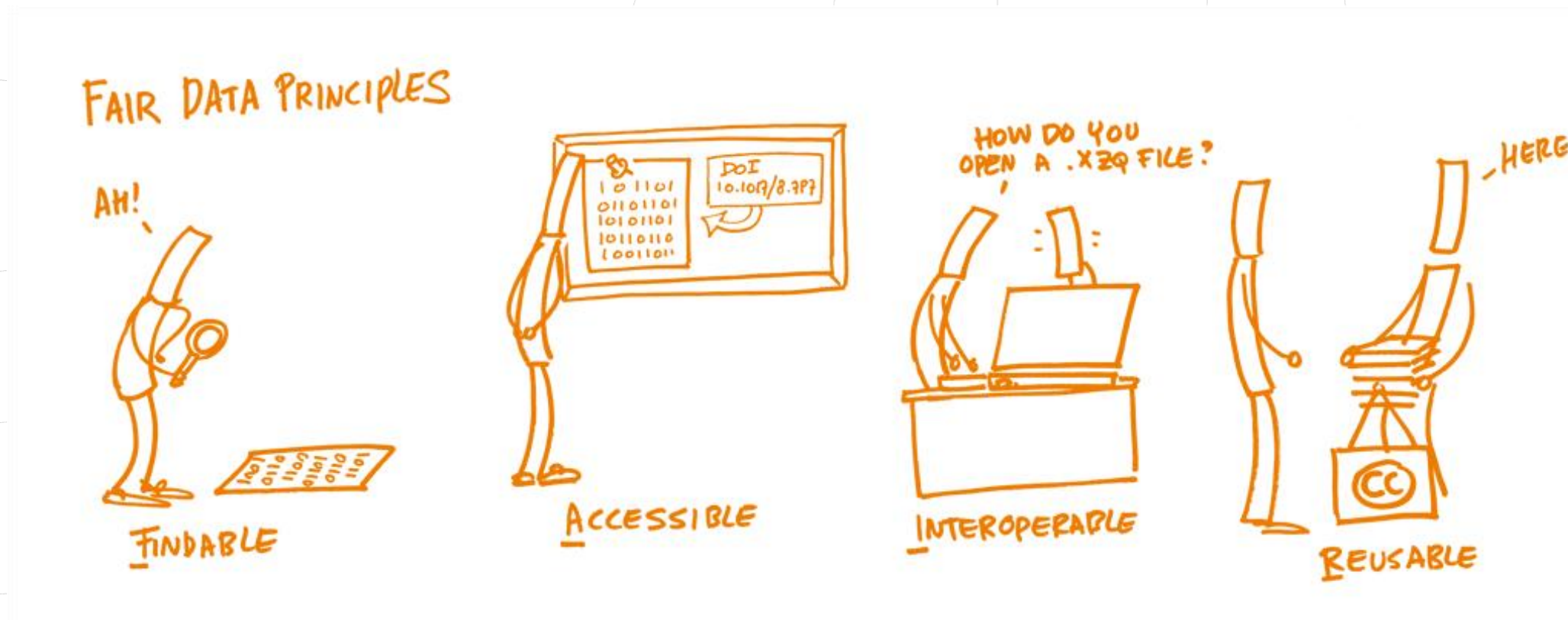
TU Delft | WIM ontwerpers



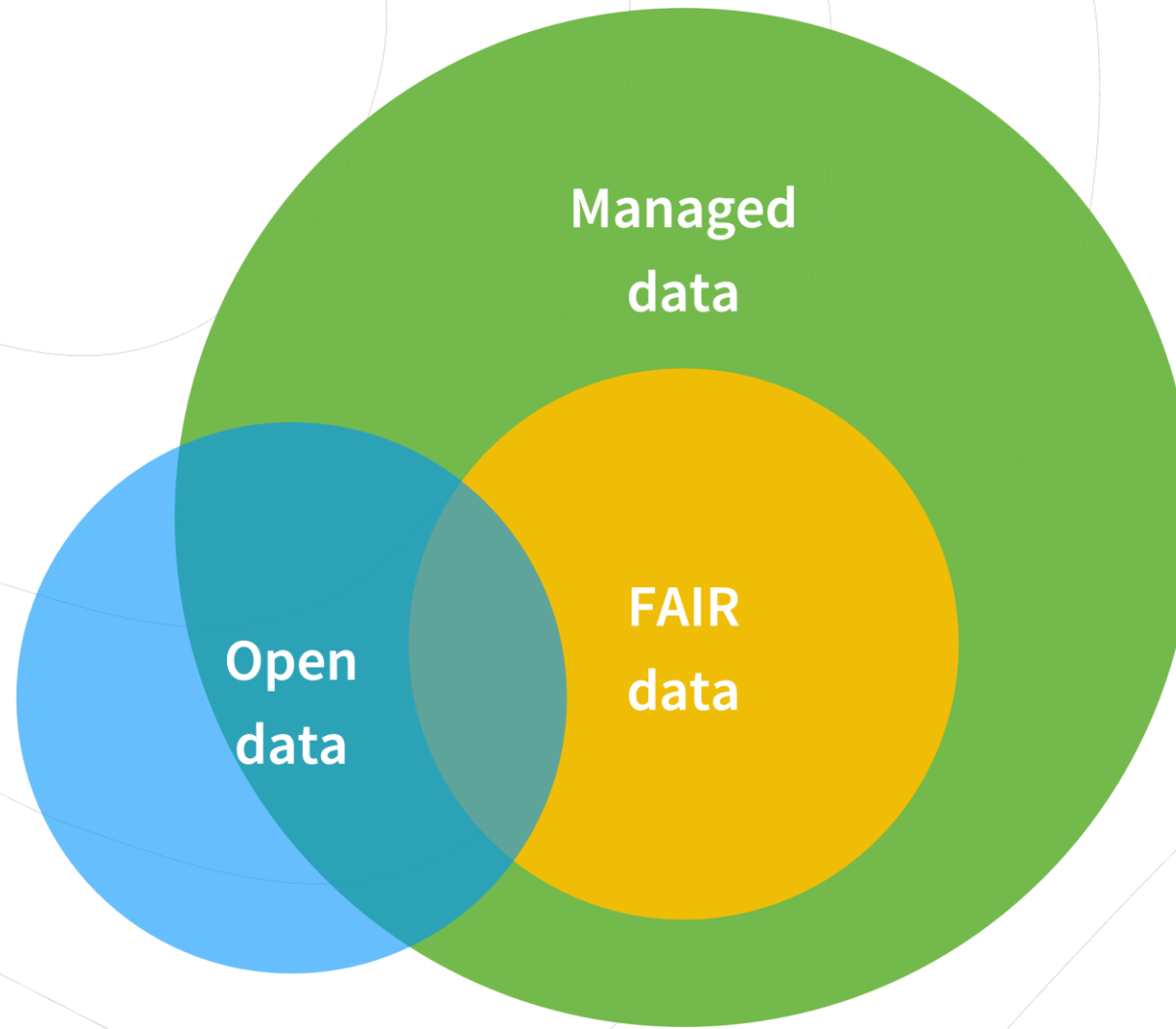
# What we will focus on next:

- FAIR principles
- Data naming conventions
- File formats
- Metadata
- Licensing
- Repositories
- Electronic Laboratory notebook

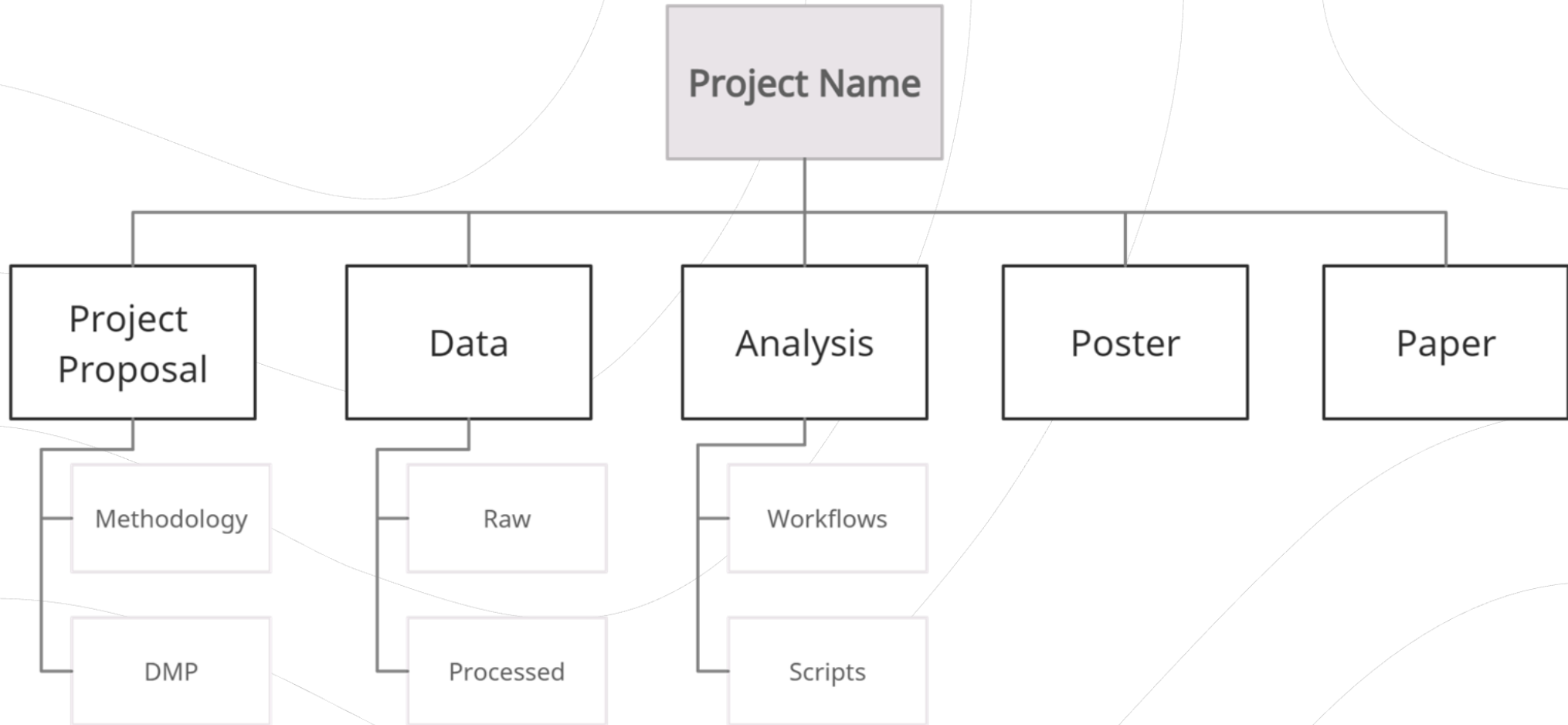
# FAIR - the ultimate goal



# FAIR, Open & Managed Data



# Organizing your data

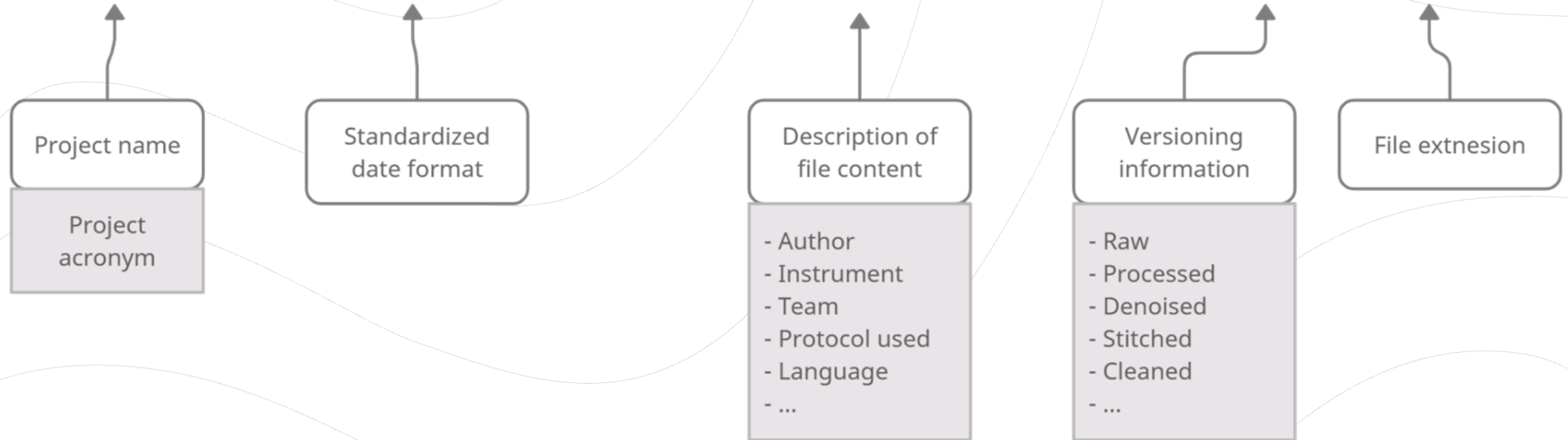


# Organizing your data

- Restrict level of folders to three or four deep
- Consider limiting the number of folders within each folder, to ten
- Include a folder within the folder structure for “documentation”. This might include:
  - Project proposals/protocols
  - Consent and approval forms
  - Methodology documents
  - Data management plan
  - Code used for recodes, analysis, and outputs
  - Readme files with transformation information
  - Readme files with the full names or titles for any abbreviations used in file names
  - Codebooks or guides

# Setup naming convention

Project\_YYYYMMDD\_ContentDescription\_Version.ext

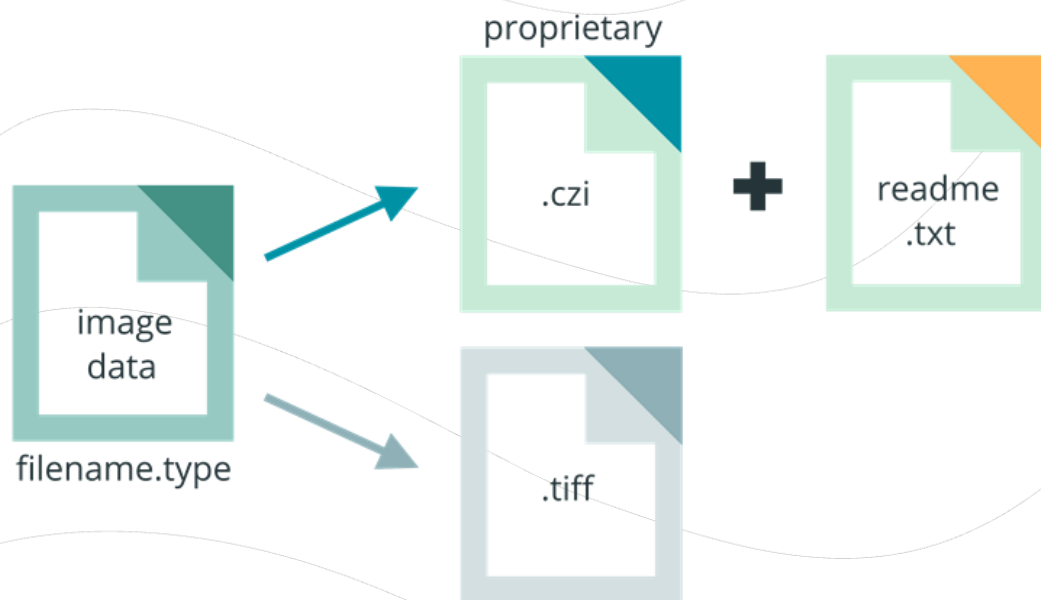


# Setup naming convention

- Avoid using spaces, dots and special characters (& or ? or !)
- Use hyphens (-), underscores (\_), or capitalization (C) to separate elements in a file name
- Include an abbreviation in the file name to identify
  - The instrument used
  - The phase (if research has multiple phases)
  - The transformation phase (i.e., original, raw, compressed, digitized, recoded, restructured, cleaned)
  - The source of third-party data (data provider or principal investigator)
  - The team (if working with multiple teams)
  - The language (if working with multiple languages)
- Include versioning within file names as appropriate



## File formats



- Preferred vs. popular
- Open vs. proprietary

When necessary to use a proprietary format, consider including a `readme.txt` file in your directory that documents the name and version of the software used to generate the file, as well as the company that made the software. This could help you down the road, if you need to figure out how to open these files again.

# Specific file types

Here are some examples of preferred FAIR file formats for preservation:

- **Images:** TIFF, JPEG 2000, PDF, PNG, GIF, BMP, SVG
- **Tabular data:** CSV, TXT
- **Text:** XML, PDF/A, HTML, JSON, TXT, RTF
- **Containers:** TAR, GZIP, ZIP
- **Databases:** XML, CSV, JSON
- **Geospatial:** SHP, DBF, GeoTIFF, NetCDF
- **Video:** MPEG, AVI, MXF, MKV
- **Sounds:** WAVE, AIFF, MP3, MXF, FLAC
- **Statistics:** DTA, POR, SAS, SAV

# Sooo... what are the metadata?

**Metadata is documentation that describes data.**

Properly describing and documenting data allows you to understand and track important details of the work.

Having metadata about the data also facilitates search and retrieval of the data when deposited in a data repository.

**Metadata: the who, what, when, where, why, how of your research.**



## Dublin Core (1999, Dublin, Ohio)

A set of 15 metadata tags:

**Creator**

**Contributor**

**Publisher**

**Title**

**Date**

**Language**

**Format**

**Subject**

**Description**

**Identifier**

**Relation**

**Source**

**Type**

**Coverage**





**Rights**

Element	Definition
Title	A name given to a resource
Creator	An entity primarily responsible for making the content of a resource
Subject	A topic of the content of a resource
Description	An account of the content of the resource
Publisher	An entity responsible for making the resource available
Contributor	An entity responsible for making contributions to the content of a resource
Date	A data of an event in the lifecycle of a resource
Type	The nature or genre of the content of a resource
Format	The physical or digital format of a resource
Identifier	An unambiguous reference to the resource within a given context
Source	A reference to an another resource from which a resource is derived
Language	A language of the content of a resource
Relation	A reference to a related resource
Coverage	The extent or scope of the content of a resource
Rights	Information about rights held in and over a resource

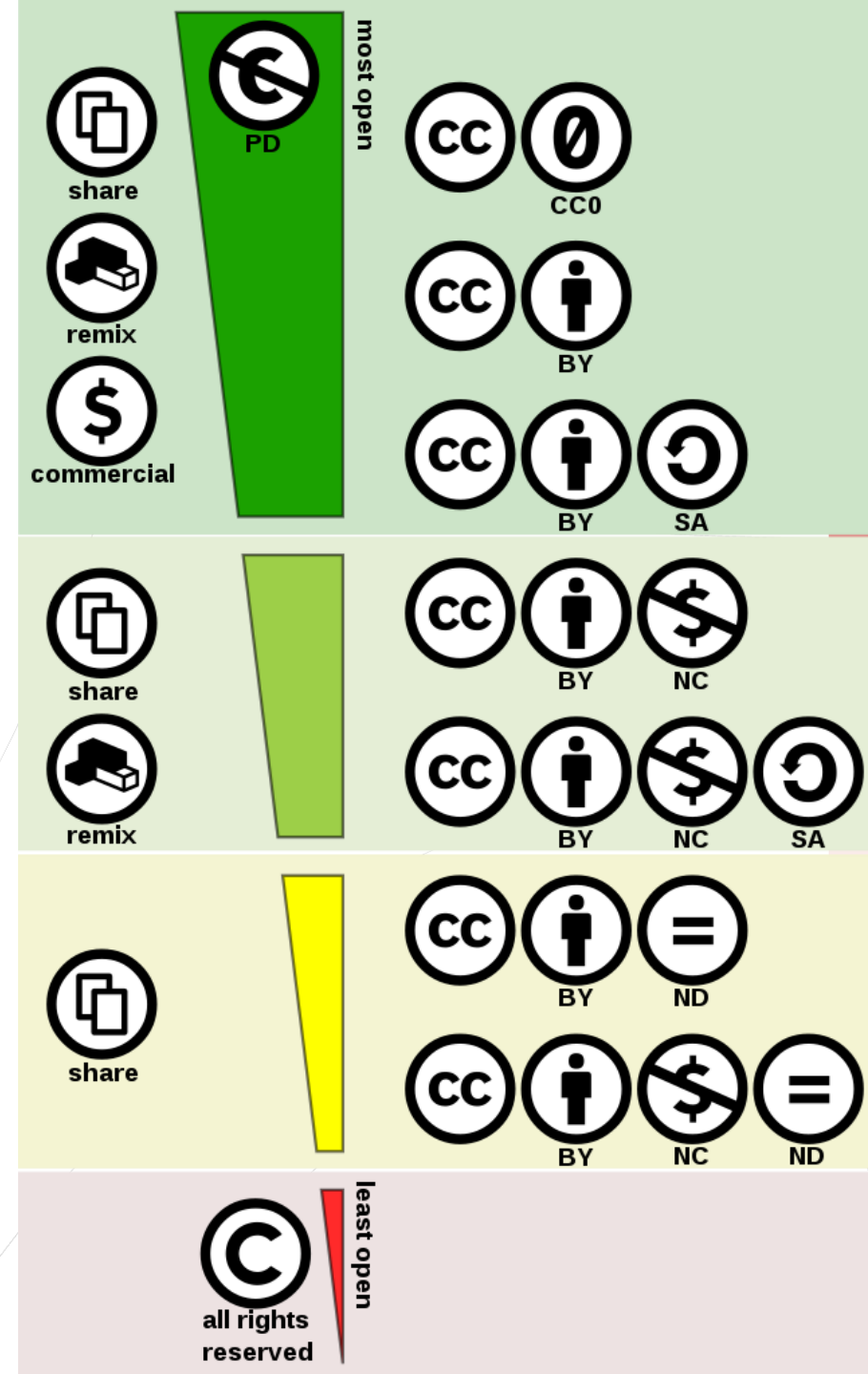
# Creative Commons licence

Easy to understand/easy to use

Meaning of **CC** suffix:

- 0 - Public domain
-  **BY** - By Attribution
-  **ND** - No Derivatives
-  **NC** - Non-Commercial
-  **SA** - Share Alike

Source: <https://creativecommons.org/>

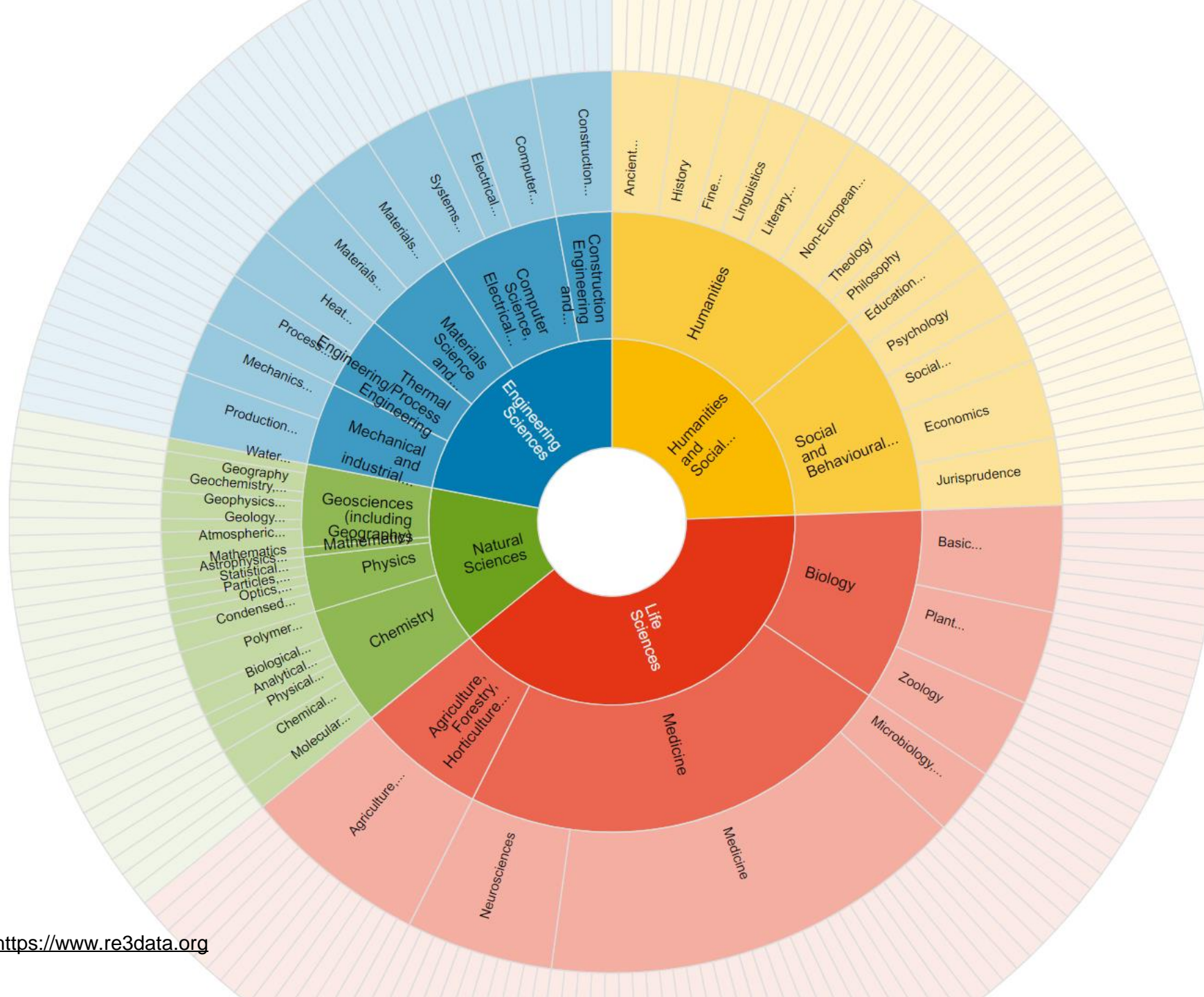


re3data.org  
REGISTRY OF RESEARCH DATA REPOSITORIES

Search...

Search





# Price of storage (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	



# Pricing estimation



## DSW Storage Costs Evaluator

Total costs:  
**2 261 €**

TB costs per year:  
**452 €**

Result details  
▼

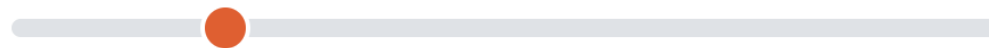
Volume



500

GB

Lifetime



10

years

Detailed storage properties ▼

## Electronic laboratory notebook

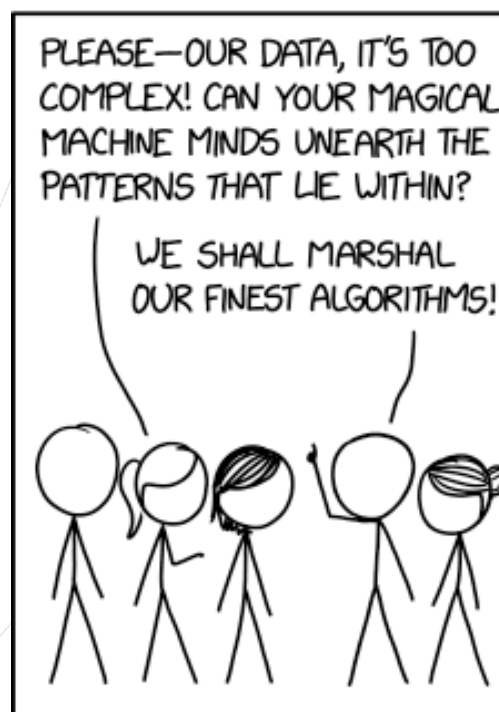
The missing infrastructure for data recording, retrieval, and integrity.

There are many options, from utilizing Google Colaboratory up to all-in-one solutions:

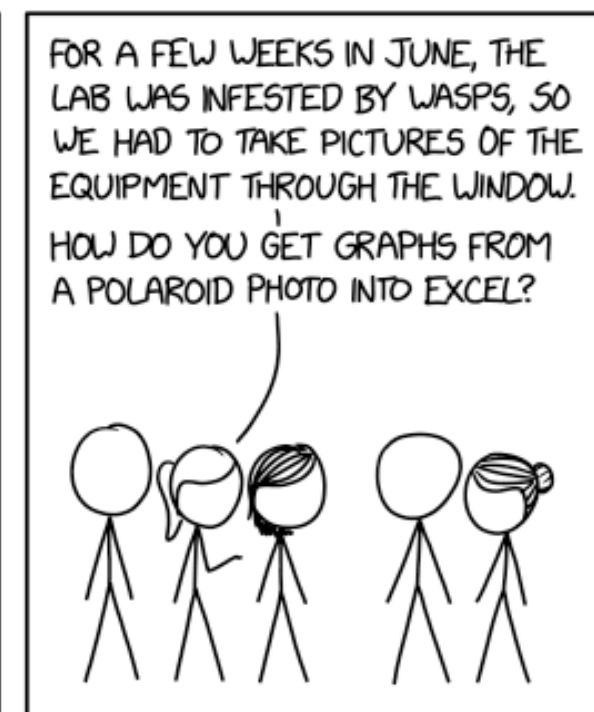
<https://www.labfolder.com/electronic-lab-notebook-eln-research-guide/>



WHAT TECH PEOPLE THINK  
SCIENTISTS NEED HELP WITH:



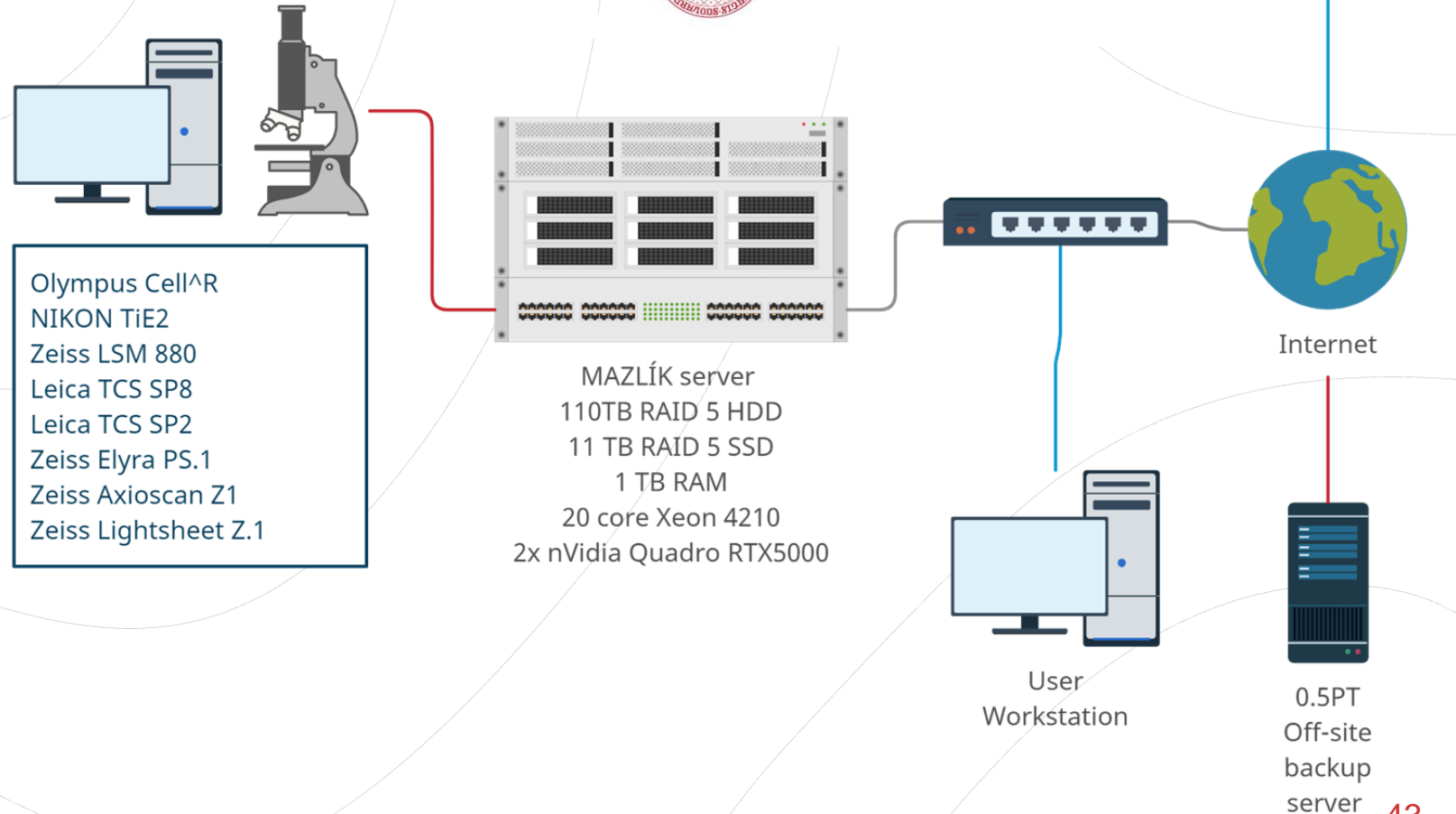
WHAT SCIENTISTS  
ACTUALLY NEED:



# How it looks in practice: VMCF

## Many limitations:

- RDM
- Length of experiments
- Data ownership
- Documenting
- **Ethics**
- Access planning
- Cost management



# Research data management resources

## **Course: Data Stewardship: module 1, DocEnhance (2021)**

<https://moodle.techlib.cz/course/view.php?id=179>

- Developed as part of the DocEnhance project
- The Data Stewardship course was piloted in Norway and Czech republic
- Course was developed for early-career researchers
- Entry level self-guided open course to data stewardship
- 11 modules on various aspects of data management
- Ended by self examination with certificate

# What to take home?

- Open Science is an evolution.
- Managing data is a good scientific practice.
- Managing and sharing data can save time, money, and create impact.
- Communities of researchers worldwide define standards, usually they are open to others joining their efforts. The same is happening at the national level.
- There is already huge amount of resources online to learn from.
- Research data management is a helpful tool, not just an administrative task.
- Funding agency will, in time demand (or already are demanding) Data Management Plans, and support RDM tasks financially.



# Get Assistance

## 1) Schedule a consultation

- Please don't be shy; our team includes doctoral students who know the issues you face
- LaTeX support, Bibliometric services

## 2) Attend other webinars

## 3) Explore by yourself

- STEMskiller: comprehensive skills set map for early career researchers
- Tutorials: NTK instructional materials and recordings and links to more information
- Subject guides



# Research data management guide



The screenshot shows the NTK website interface. At the top left is the NTK logo with coordinates and the full name. A search bar is at the top center. Below it is a navigation menu with categories like 'What We Have', 'Services & Support', 'Projects', 'Culture & Events', and 'Who We Are'. The main content area has a breadcrumb trail: 'Homepage / Services & Support / Education and Research Support / Tutorials / Research Data Management'. The title 'Research Data Management' is followed by an introductory paragraph about RDM. A horizontal menu below the text lists links: 'Research Data', 'FAIR Principles', 'Research Data Management', 'Data Management Plan', 'Data Repositories', 'Support', and 'Resources'. On the right side, there is a 'Your contact' section with two entries: Adéla Jilková and Karolina Podloucká, each with a profile picture, email address, and phone number.

# Contacts

Jan Vališ

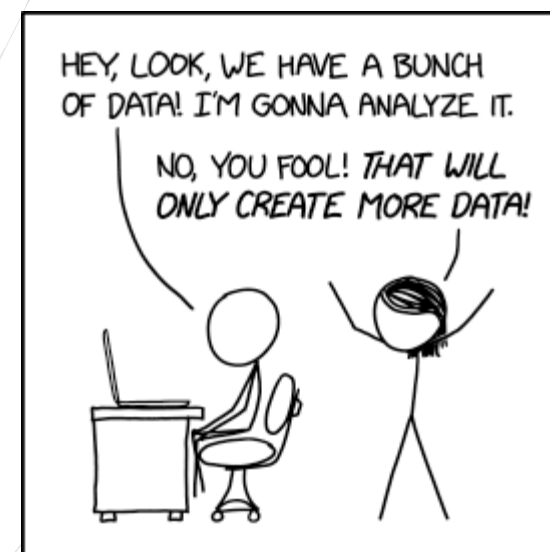
[jan.valis@vscht.cz](mailto:jan.valis@vscht.cz)

Adéla Jílková

[adela.jilkova@techlib.cz](mailto:adela.jilkova@techlib.cz)

Martin Schätz

[martin.schatz@techlib.cz](mailto:martin.schatz@techlib.cz)



# Thank you

## Questions?