Introduction to research data management

University of Maribor Open Science Summer School

Maribor, 13. 9. 2022

Dr. Ana Slavec

DOI: 10.5281/zenodo.7069032

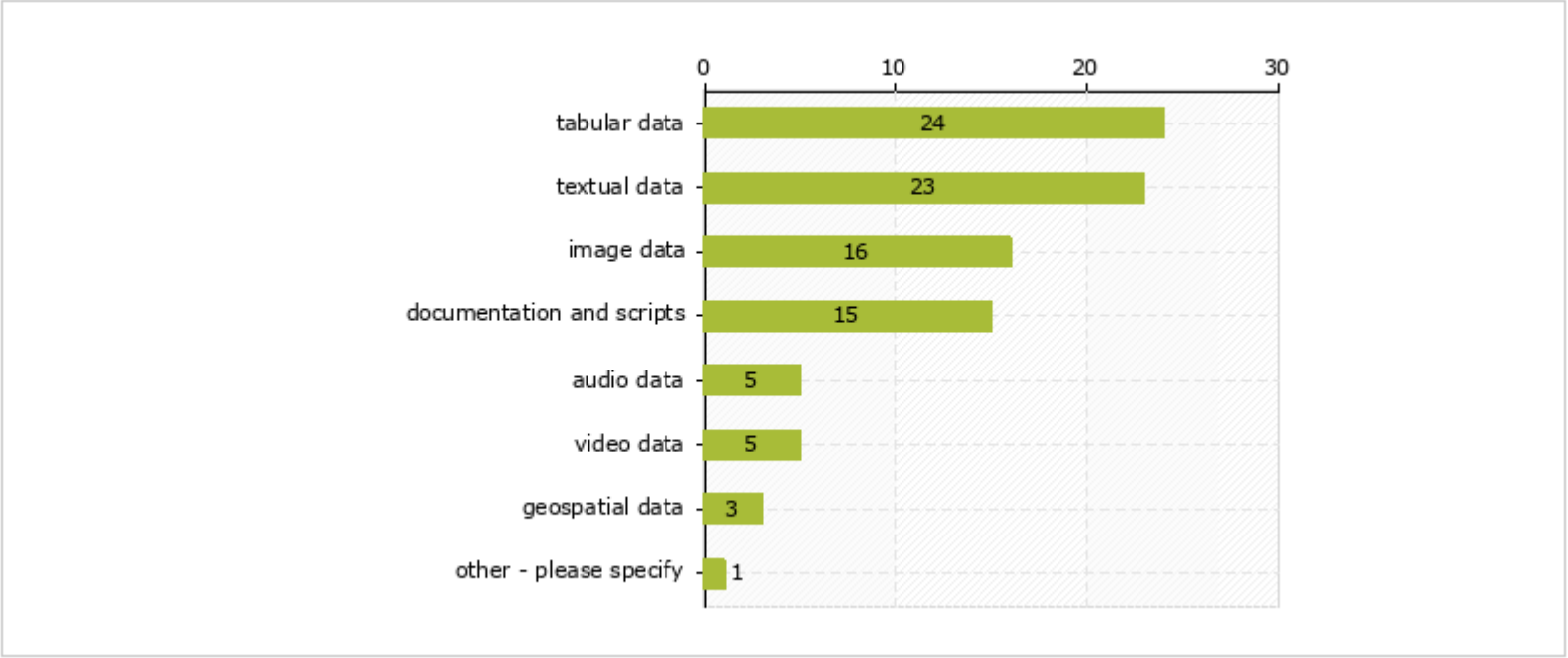
Ice-breaking questions

- On **slido.com with #4255813**
  - What type of digital content do you generate in your research?
  - What formats do you use to save your data?
  - Where do you store your research data?

- In classroom
  - Introduce yourself (name, institution, research field, current research project)

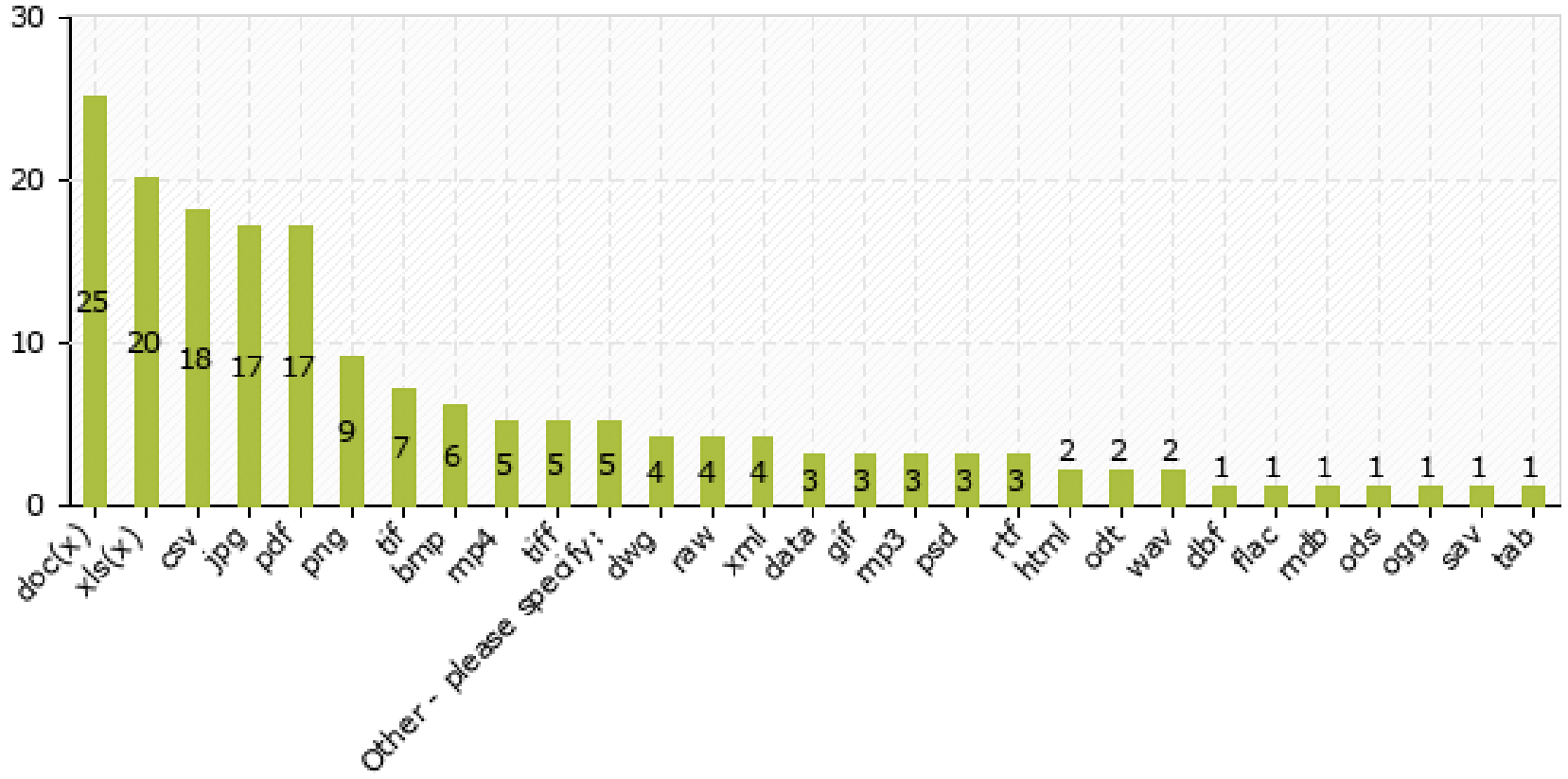
What type of digital content do you generate in your research?

(responses by attendees of one of my previous lectures, n=28)



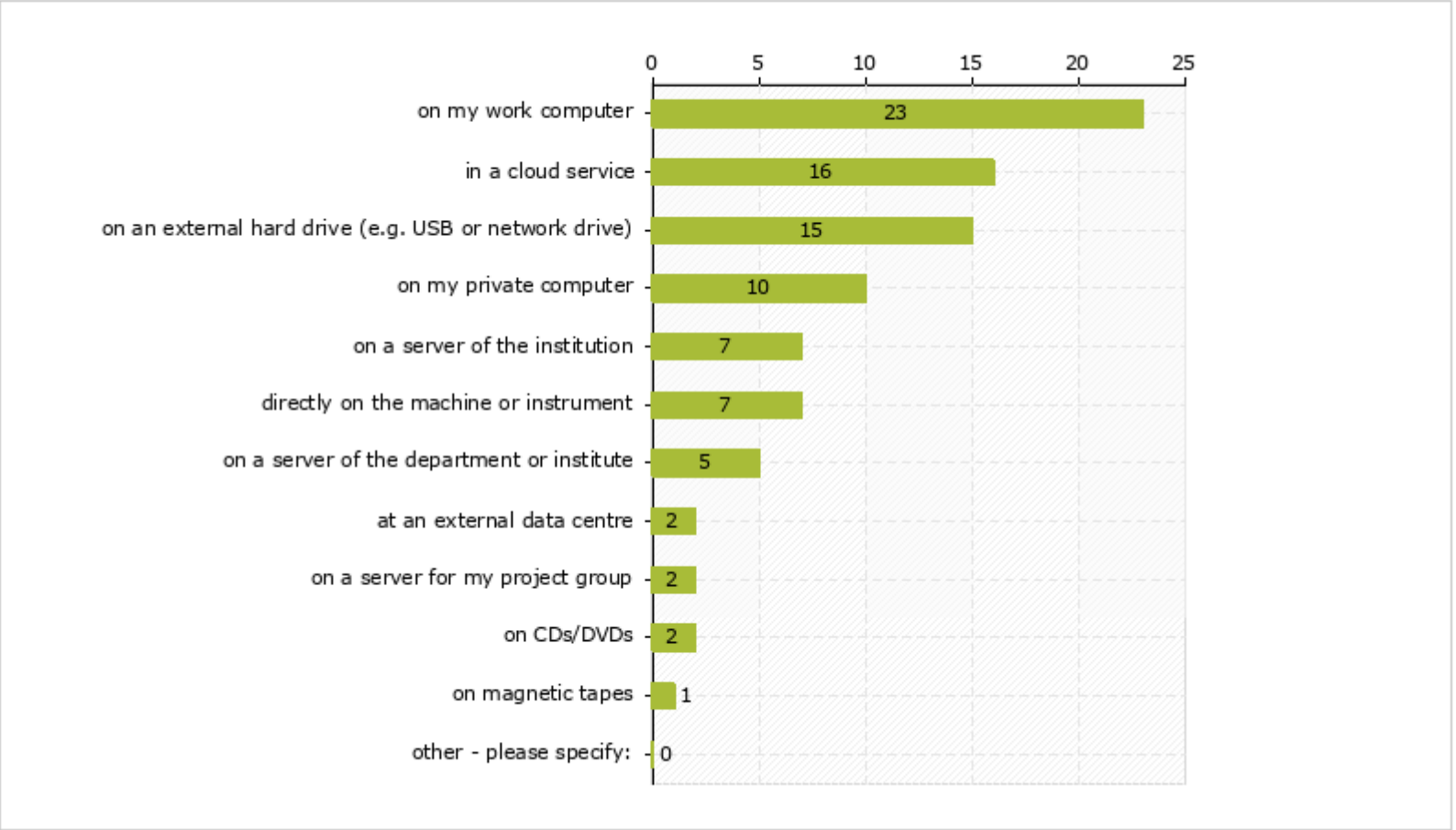
What formats you use to save your data?

(responses by attendees of one of my previous lectures, n=28)



Where do you store your research data?

(responses by attendees of one of my previous lectures, n=28)



Outline of lecture

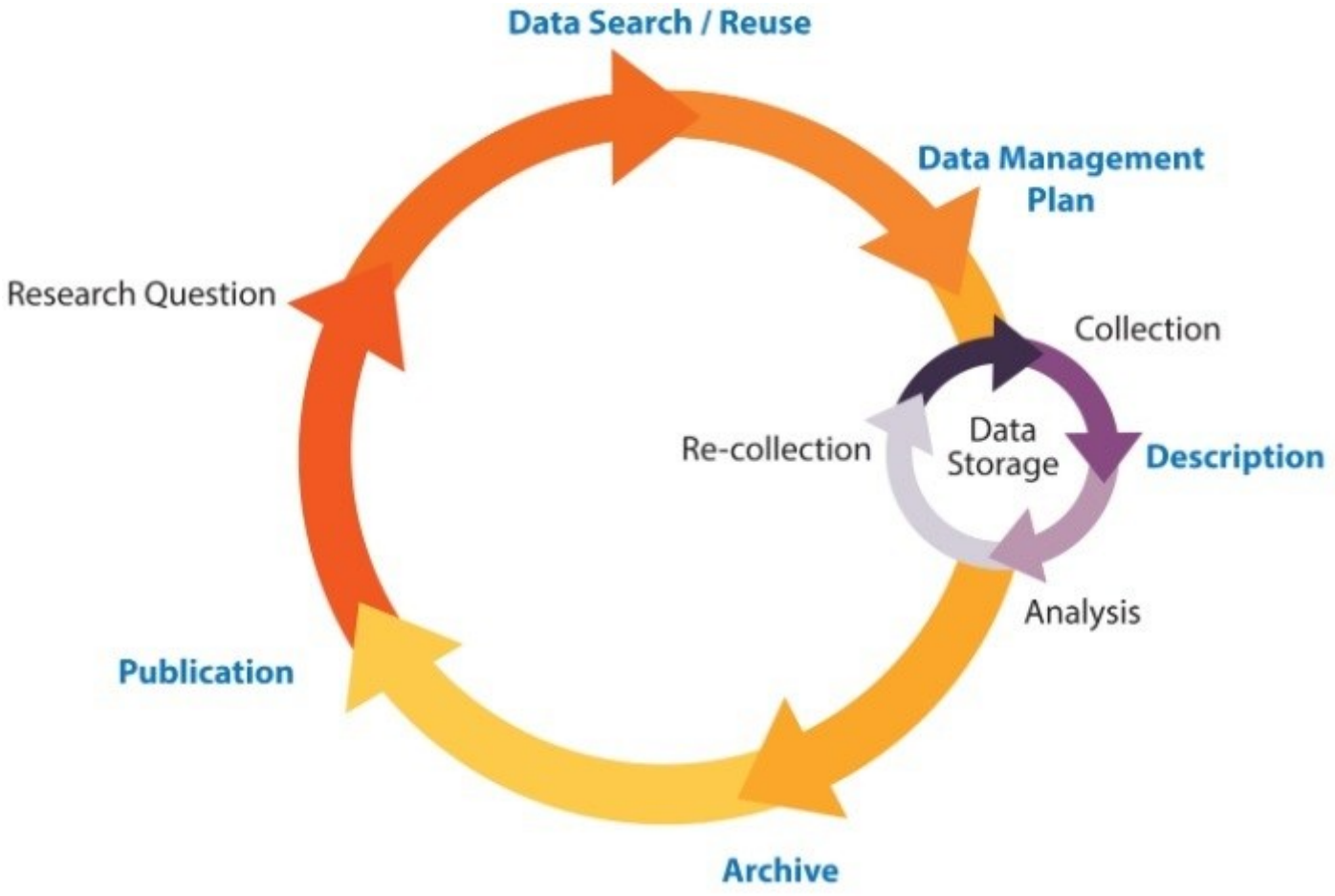
- Research data lifecycle
- FAIR principles
- Openness
- RDM responsibilities

What is/are data?

- **Data** is the plural of *datum* (lat. „something given“) and refers to a set of values of qualitative or quantitative variables
- **Information** is a processed and interpreted outcome of data in a given context
  - Example: temperatures all over the world for the past 100 years is data; the analysis to find that the global temperature is rising is information

- Grammatical note:
  - Information is a mass or uncountable noun (e.g. The information is ready)
  - Data is technically a plural noun that deserves a plural verb (e.g. The data are ready) but since the word datum is not commonly used so the word data is becoming a mass noun (e.g. The data is ready)

Research data lifecycle



Source: University of California, Santa Cruz, Data Management LibGuide

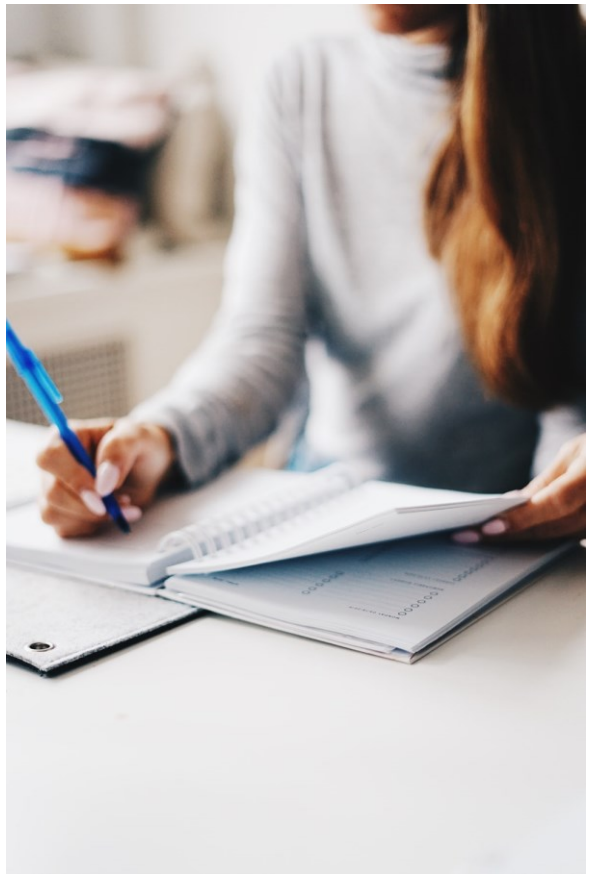
- **Research data management** describes the organisation, storage, preservation, and sharing of data collected and used in a research project.
- It involves decision about how data will be preserved and shared after the project is completed.

Data search/reuse

- Official statistics (national statistical offices, Eurostat…)
- International NGOs (United nations, World bank…)
- Research data repositories (general purpose, institutional, domain specific)
  - E.g. CESSDA Data Catalogue (for social sciences)
- Google dataset search
- …

Data management plan (DMP)

- Description of data to be collected/created
- Standards/methodologies for RDM
- Ethics and intellectual property right
- Plans for data sharing and access
- Strategy for long-term preservation

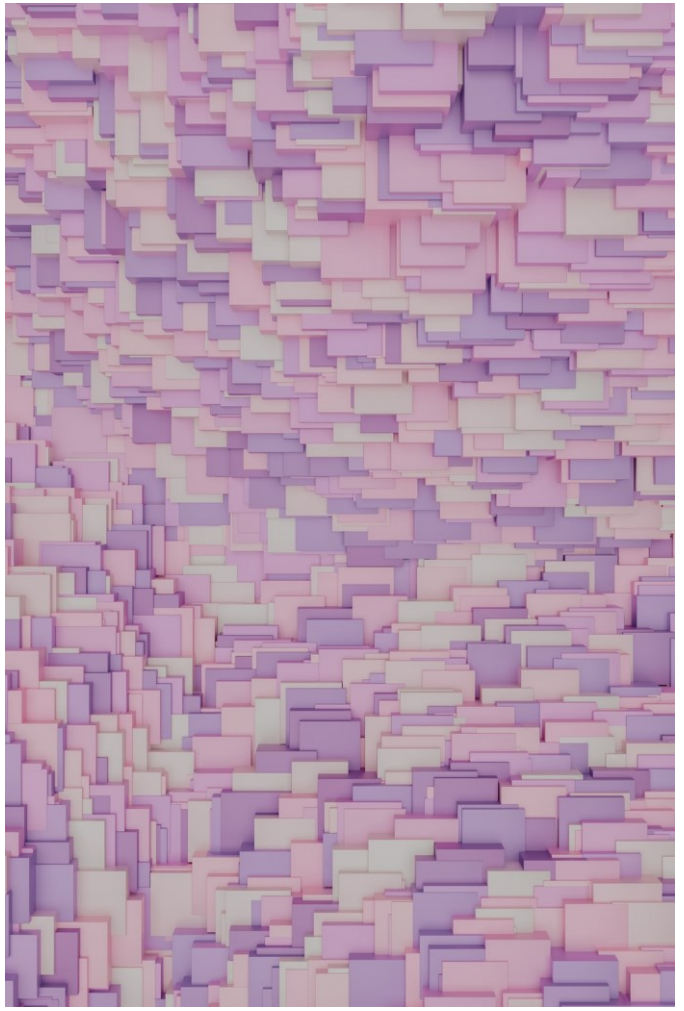


→ 3 lectures on DMPs at this summer school:

- dr. Tea Romih: **Data management plan: step by step**
- Danaja Fabčič Povše: **Data management plan and research with personal data**
- Peter Čerče: **Migrant children and communities in a transforming Europe**

Description

- Source of data (primary data collection and/or reuse of secondary data)
- Type, format and volume of data
- Standards and methodologies
- Structure and name of files
- Versioning
- Quality assurance processes



Data collection

Designed data (active)

- Experiments, Surveys

Organic data (passive)

- Administrative, Transactional

Types of data and recommended formats

Type of data	Recommended formats	Acceptable formats
Tablular data	csv, .tab, .por, .xml	.txt, xls, .dbf, .ods, .sav, .dta, .mdb
Geospatial data	.shp, .shx, .dbf, .prj, .sbx, .sbn, .tif, .tfw, .dwg, .gml	.mdb, .mif, .kml, .ai, dxf, .svg
Textual data	.rtf, .txt, .xml	.html, .doc
Image data	.tif	.jpg, .gif, .tif, .tiff, .raw, .psd, .bmp, .png, .pdf
Audio data	.flac	.mp3, .aif, .wav
Video data	.mp4, .ogv, .ogg, .mj2	.avchd
Documentation and scripts	.rtf, .pdf, .xthml, .htm, .odt	.txt, .doc, .xls, .xml

Source: UK Data Service

Tabular data

Attributes

Country name	Population	Area (square meters)	Year of entry to the EU	…
Austria	8,857,960	301,318	1995	…
Italy	60,404,355	83,858	1957	…
Slovenia	2,070,050	20,273	2004	…
…	…	…	…	…

Entities

Values

Relational data model is the most popular but data can be also organised differently (hierarchical, network, object…).

Data analysis

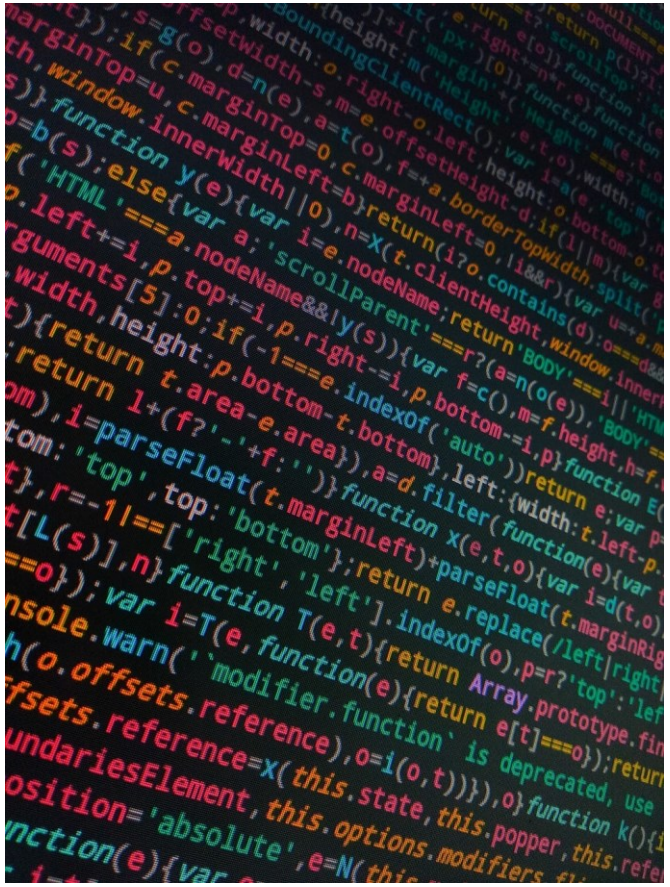
The 80/20 rule (aka the Pareto principle)

80% of time is spent cleansing and organising data

20% left to actually perform analysis

Documentation and metadata

- **Metadata** is „data about data“ (Examples: persistent identifier such as DOI, publication date, title, authors, description, keywords, licence, funding, related identifiers, etc.)
- **Documentation** may also include details on the methodology used, analytical and procedural information, definition of variables, vocabularies, units of measurement, assumptions made, and the format and file type of the data
- Existing community metadata standards: General (e.g. Dublin Core) or discipline specific (e.g. DDI); See RDA Metadata directory



Archive

- **Repository** is a place where records are stored
- **Archive** (type of repository) is a place for preserving primary source material and usually does not allow it to be checked out
- The two terms are often used interchangeably

- Repositories (archives) for data to be held
  - General purpose (e.g. Zenodo, Figshare)
  - Institutional
  - Domain specific (see re3data.org registry)

General purpose and institutional repositories

+ Simple process
+ Fast publication
- The metadata scheme is too general
- Data is not reusable without rich metadata and documentation

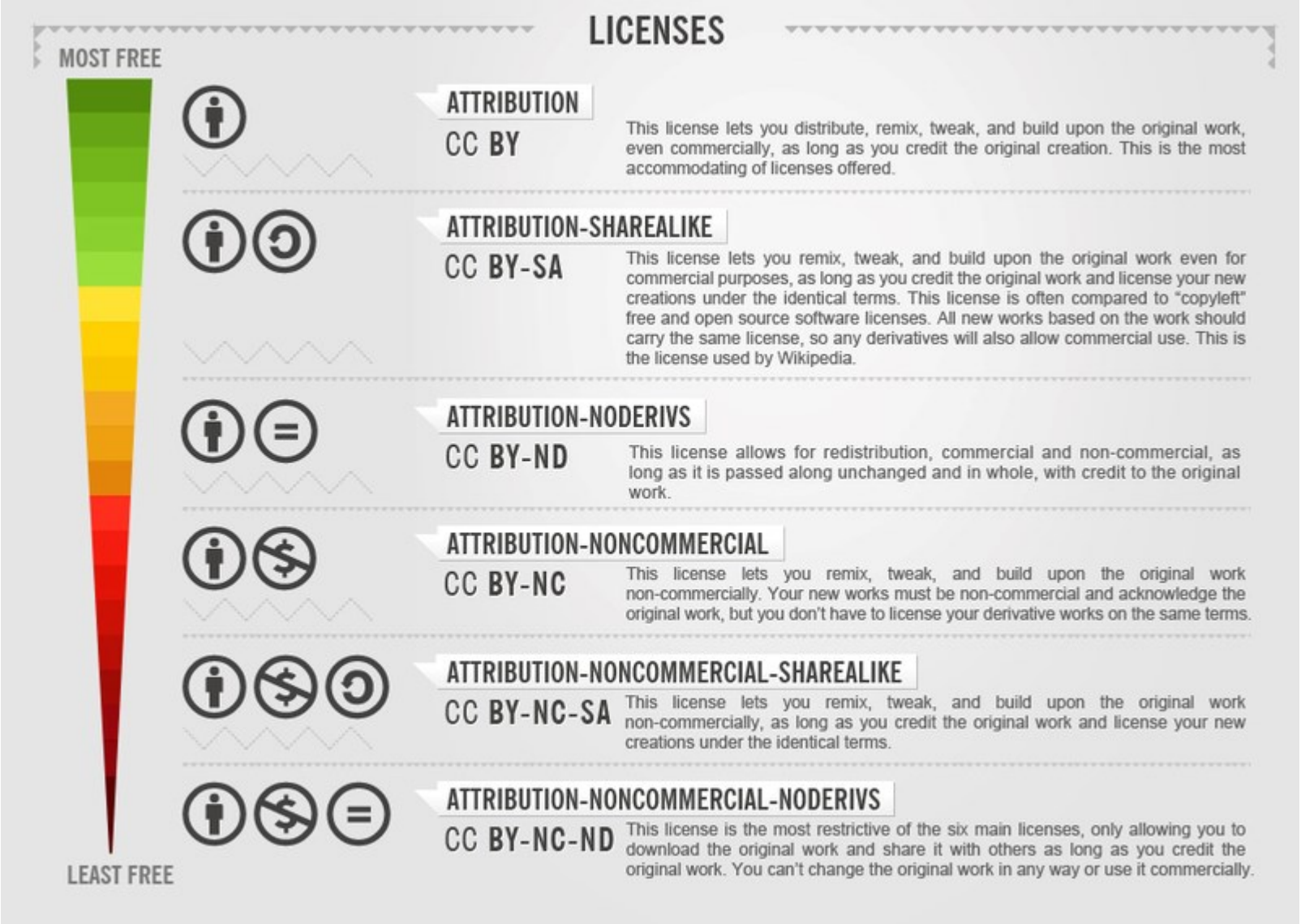
Domain specific repositories

+ Field specific metadata standards
- Lack of repositories for certain fields
- Time consuming to prepare
- Requires domain specific skills

Publication

Licences for reuse

(e.g. Creative commons)



Source: Wikipedia

Ethics and legal compliance

- Informed consent sought for data collection, processing and long-term preservation
- Removal, aggregation, pseudoanonymisation, or anonymisation of direct and indirect identifiers in data files
- Restriction of access do the data in cases, when anonymisation would hinder the reusability of data

→ 3 lectures on ethical and legal issues at this summer school:

- Danaja Fabčič Povše: **Data management plan and research with personal data**
- prof. dr. Janja Hojnik: **Ethics and integrity in research**
- dr. Maja Jančič Bogataj: **Legal issues for open research: selected topics**

The FAIR data principles

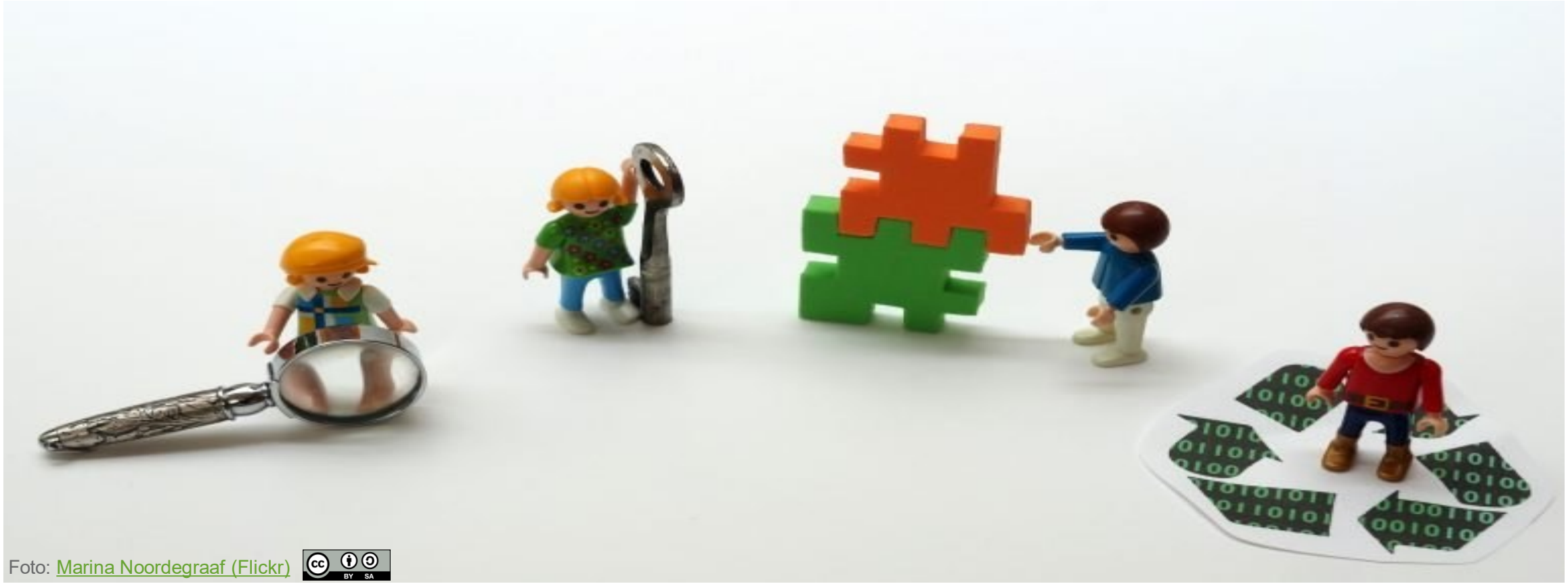
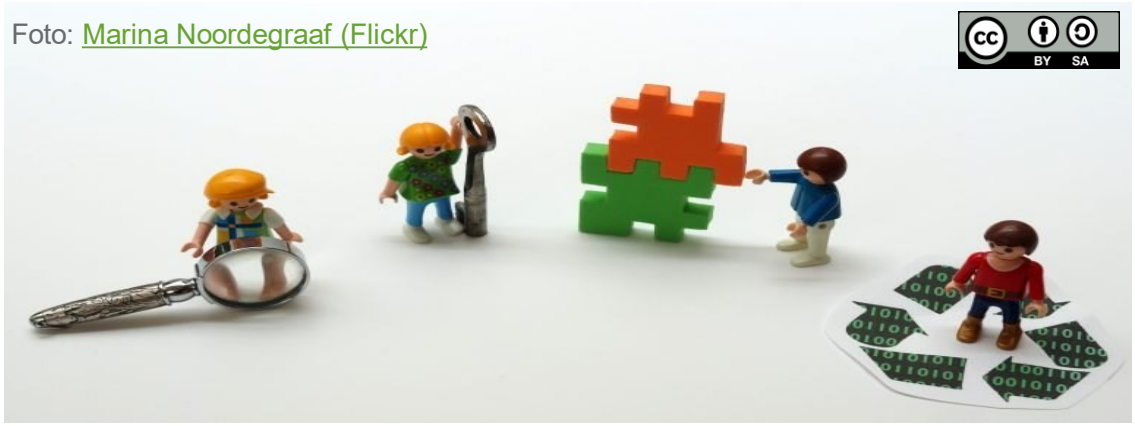


Foto: Marina Noordegraaf (Flickr)

The FAIR data principles

- **Wilkinson et al. 2016**



- **Findable:** metadata and data easy to find for both humans and computers
- **Accessible**: users need to know how can they be accessed, possibly including authentication and authorisation
- **Interoperable**: data can be integrated with other data in applications or workflows for analysis, storage, and processing
- **Reusable**: metadata and data should be well-described so that they can be used in different settings

Findability

F1. (Meta)data are assigned a globally unique and persistent identifier

F2. Data are described with rich metadata (defined by R1 below)

F3. Metadata clearly and explicitly include the identifier of the data they describe

F4. (Meta)data are registered or indexed in a searchable resource

Accessibility

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, where necessary

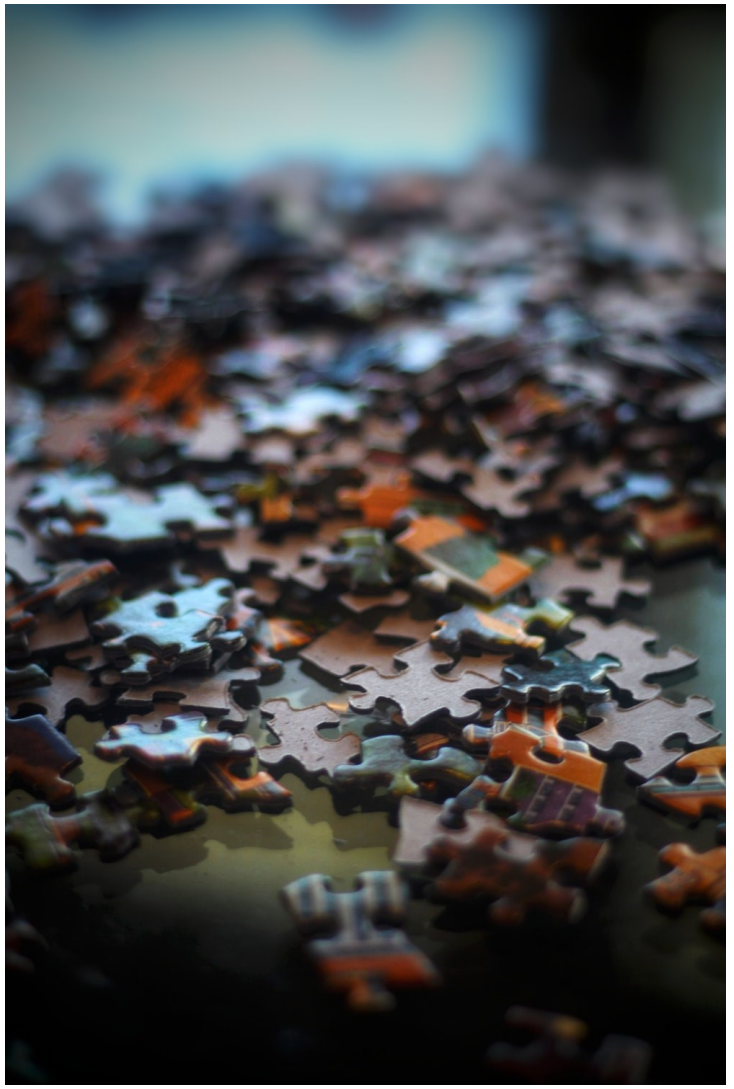
A2. Metadata are accessible, even when the data are no longer available

Interoperability

I1. (Meta)data use a 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



Reusability

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 detailed provenance

R1.3. (Meta)data meet domain-relevant community standards

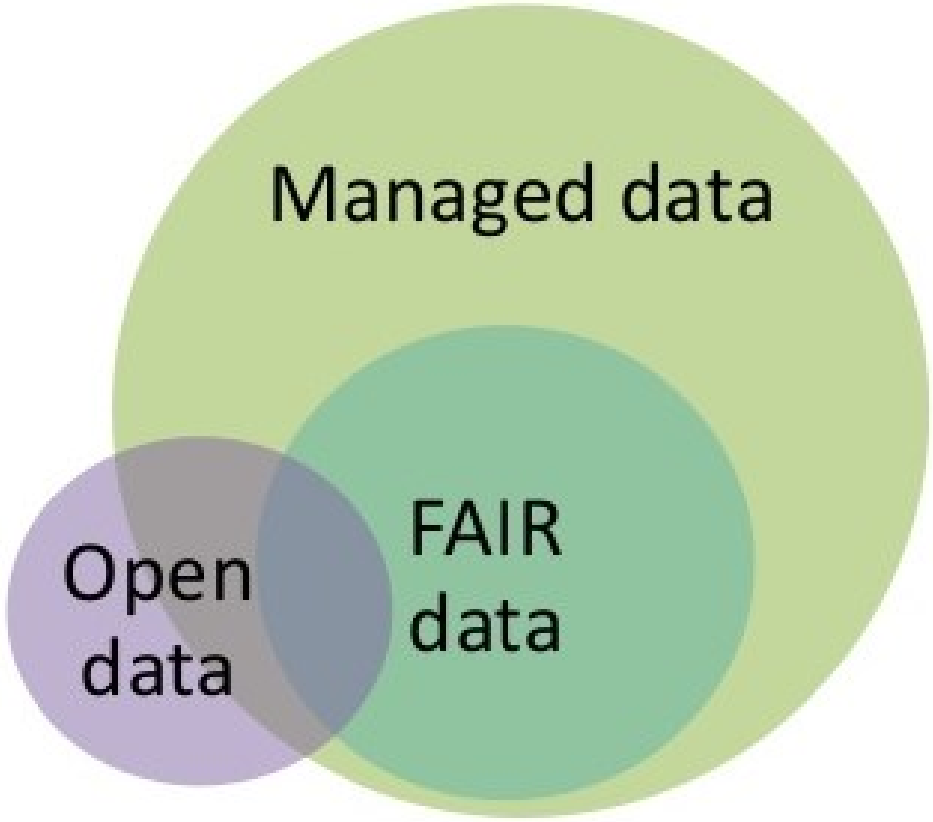
Openness

- **Open data** can be freely used, modified, and shared by anyone for any purpose
- **Legally open**, i.e. available under an open data licence (but it can be subject to requirements to provide attribution (BY) and/or share-alike (SA) licence).
- Technically open, i.e. available for no more than the cost of reproduction and in machine-readable and bulk form (see Open Data Handbook)

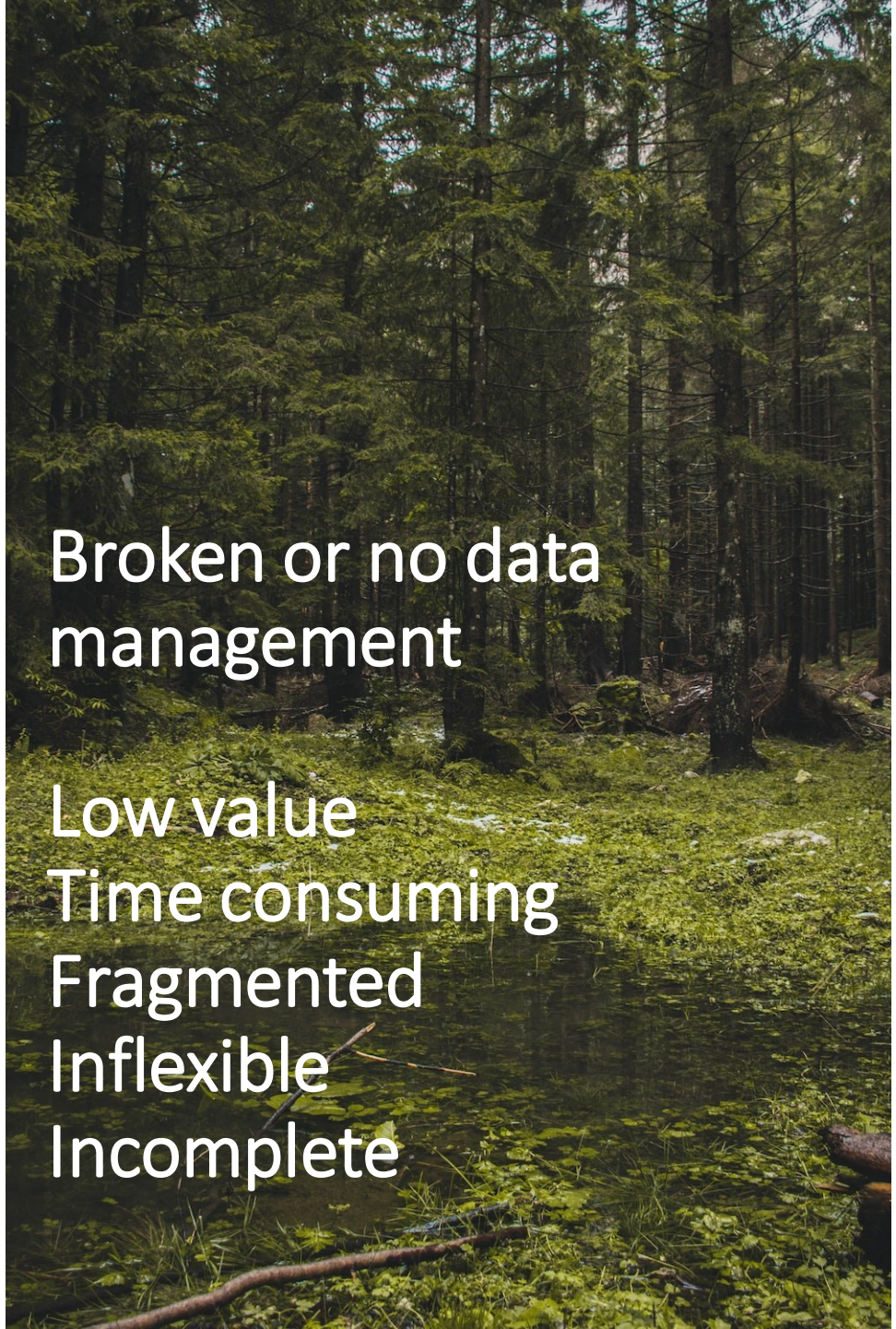
FAIR ≠ Open

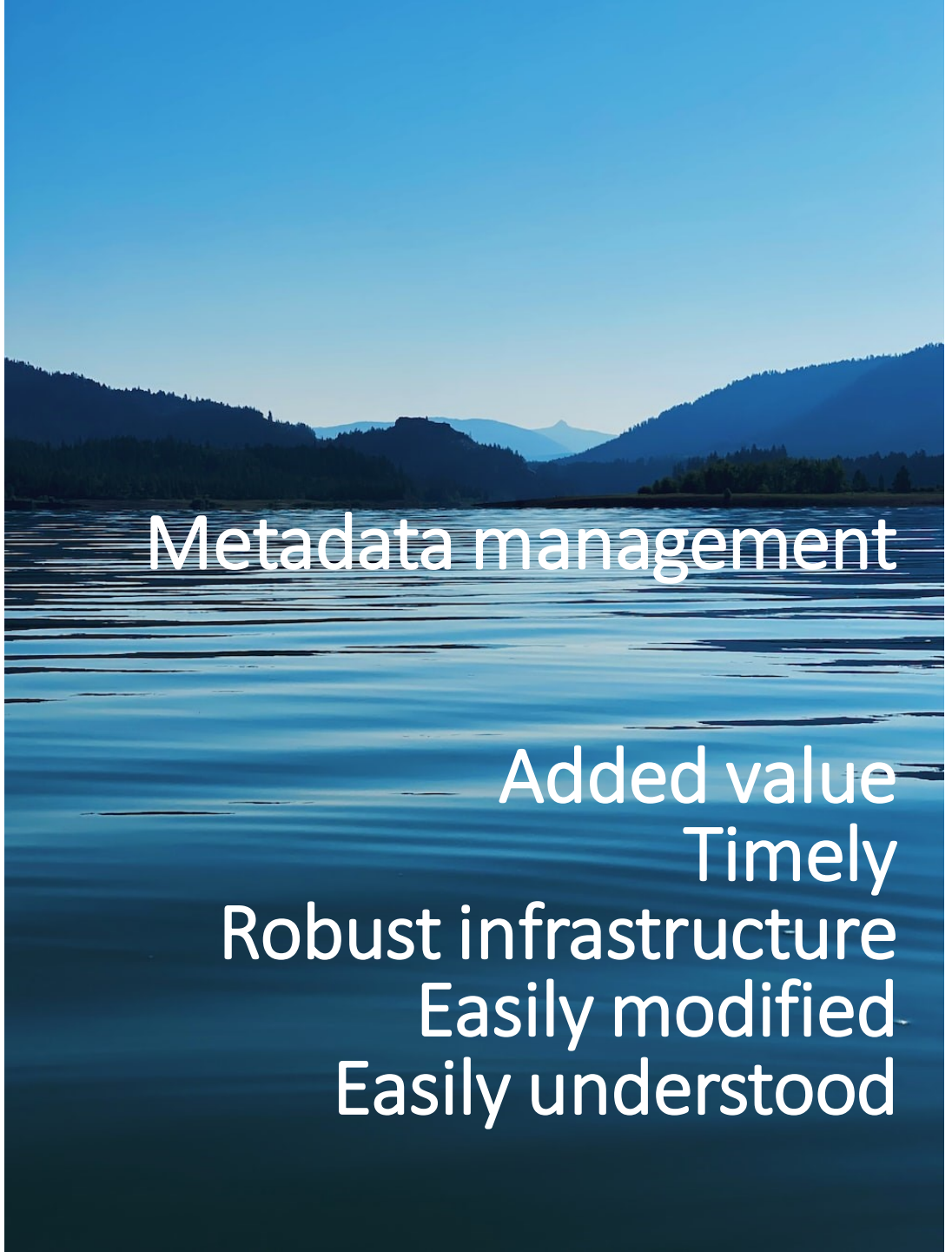
„As open as possible, as closed as necessary“

Understanding the difference between openness and FAIRness



Jones, S. 2018. Open data, FAIR data and RDM: the ugly duckling. Available at: Zenodo.

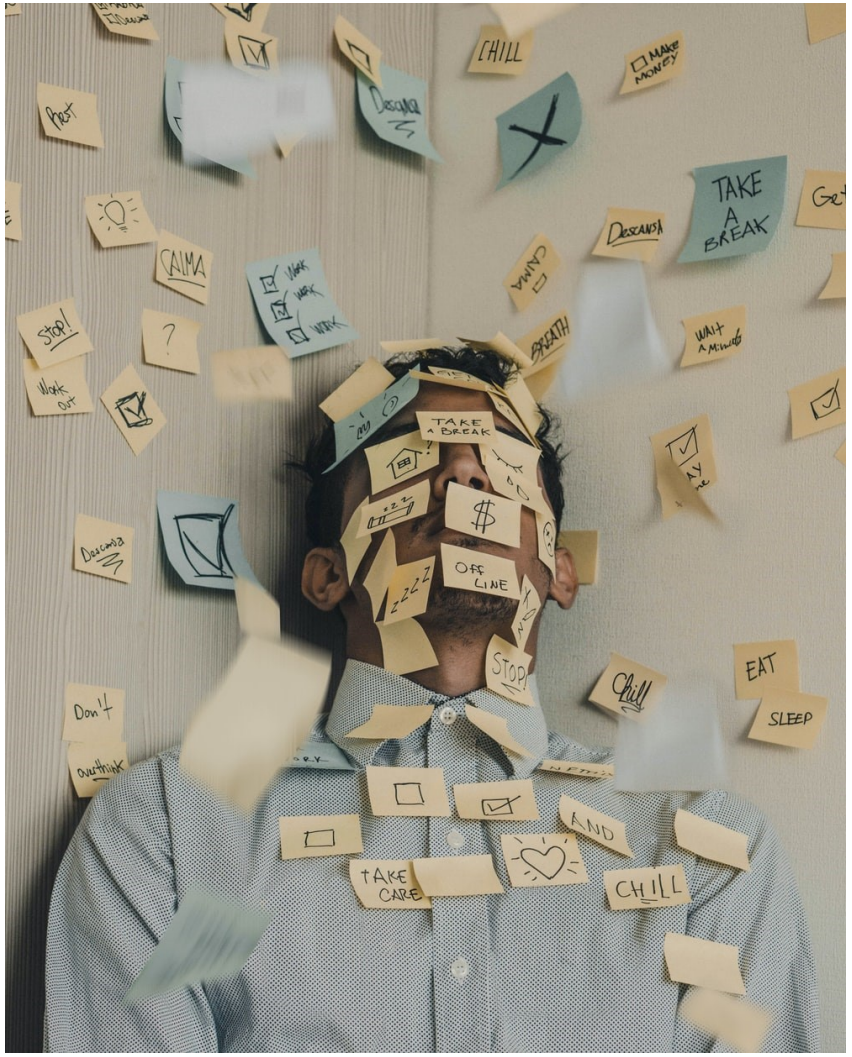

Broken or no data management
Low value
Time consuming
Fragmented
Inflexible
Incomplete


Metadata management
Added value
Timely
Robust infrastructure
Easily modified
Easily understood

Data management responsibilities

- Often specialist expertise is required
- **Research data stewards** advise, support and train researchers in research data management.

- Examples of good practices
  - Data Stewards and Champions at TU Delft
  - Data Champions at University of Cambridge
  - Data Agents and Advisor at Aalto University
- Training
  - CODATA-RDA Summer School for Research Data Science (2019)
  - Data stewards course at the University of Vienna (2020-2022)



Organisations, resources

- Research Data Alliance
- CODATA
- Digital Curation Centre
- UK Data Service
- GO FAIR project
- FOSTER project
- Open Science MOOC
- European Open Science Cloud
- …

Discussion