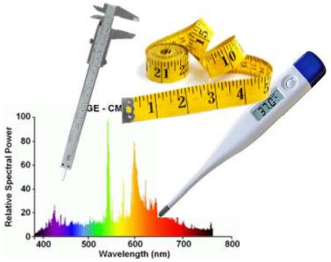
Open and responsible research data management: a few key concepts

UNIC | ONLINE Expert Insights: Open Data

Judith Biernaux, PhD.
RISE - Recherche, Innovation, Support et Entreprises
Research Data Officer
Head of Research Management Unit
Place du XX Août, 7 (Bât. A1)
B-4000 Liège
Tél : +32 (0)4 366 55 14
Jbiernaux@uliege.be

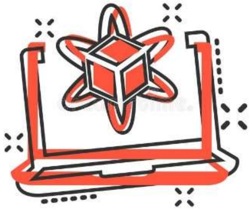
What is research data?

Documents, tables, maps

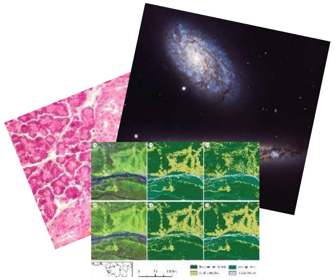


Experimental results

Polls, forms

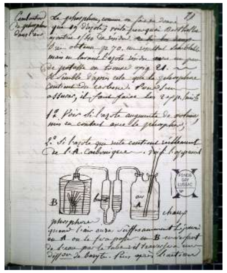


Simulation results

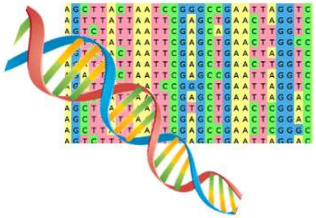


Images, videos

Audio files



Lab notes, field work notes



DNA sequences

Papers, publications

Physical samples

What is research data?

Factual records (numerical scores, textual records, images and sounds) used as primary sources for scientific research and that are commonly accepted in the scientific community as necessary to validate research findings. A research data set constitutes a systematic, partial representation of the subject being investigated.

Recommendation of the Council concerning Access to Research Data from Public Funding (OECD, 2006)

*Any **piece of information** that has been collected, measured, observed or generated*

*-> **to be used** in a research project*

-> in/from experiments, observations, simulations, database compilation, ...

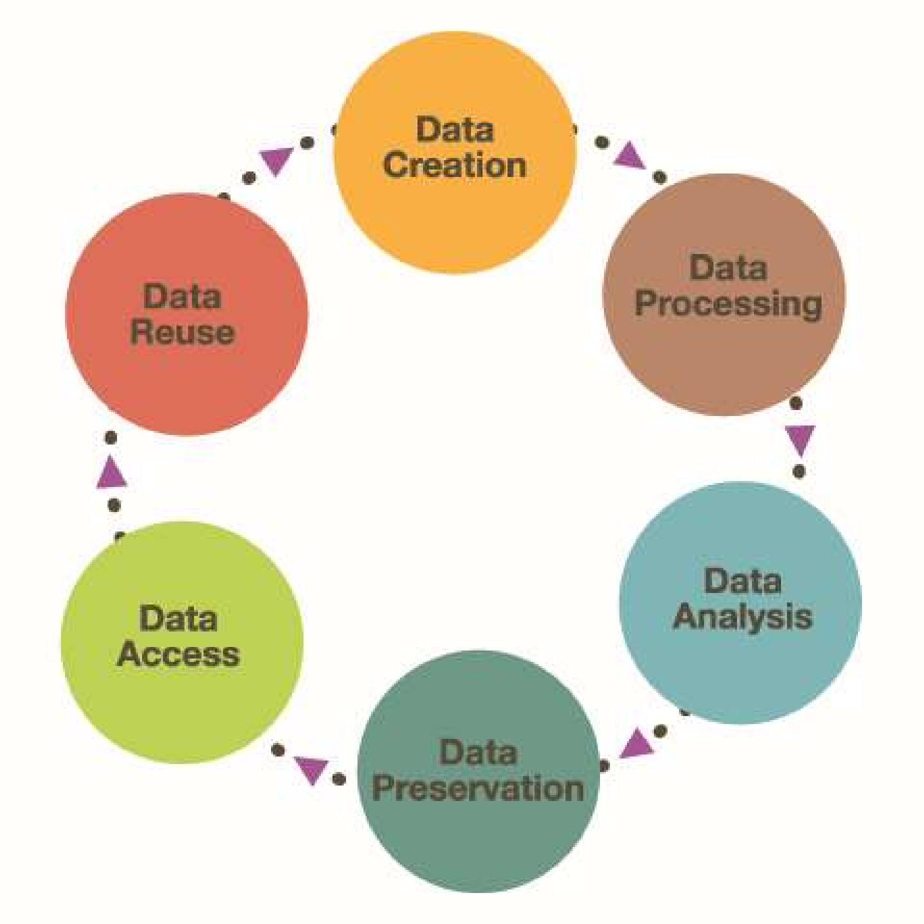
-> quantitative or qualitative

➔ Standardized methods for such a diverse world is an illusion...

➔ ...but common values are not

What is the use of research data?

Data Life Cycle



What usually happens to data:

They are **created**

in a lab, through fieldwork, measurement, on a computer, ...

They are **processed**

cleaned up, sampled, converted, ...

They are **analysed**

statistics, fitting, study, comparison, interpretation, ...

They are **stored**

for long-term preservation

They are **shared**

as open as possible, as closed as necessary

Someone else **re-uses** them

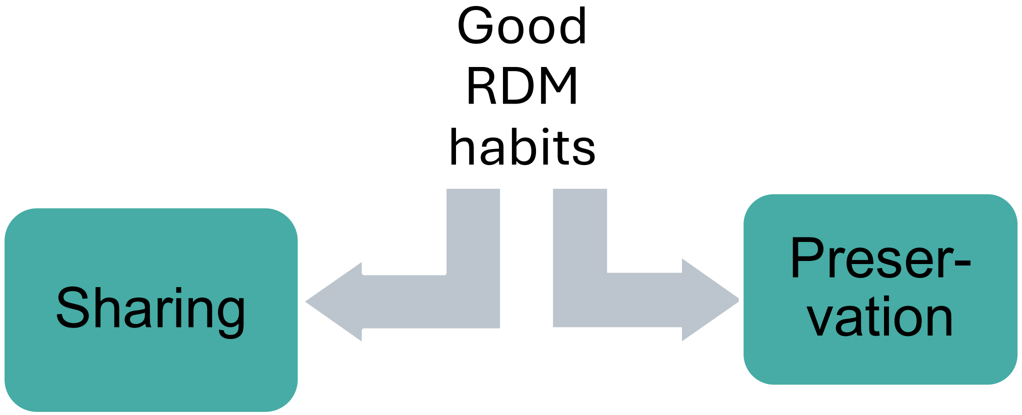
What is the endgame?

Good RDM habits make your data:

- Better **organised**, **protected** and **compliant**
- Easier to **use** and to **understand** for yourself...
- ... but also for your (future) **peers**
- Easier to **share** and **re-use**

Researchers put a lot of effort in collecting or creating data
-> Good RDM prevents these **time and cost** from being wasted by making data **re-useable** (even when not shareable!)

Funders and editors are placing increasing **demands** on RDM

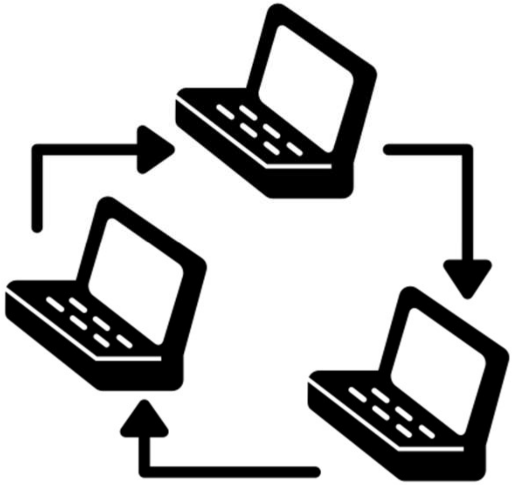


What is the endgame?

Good RDM habits make your data:

- Better **organised**, **protected** and **compliant**
- Easier to **use** and to **understand** for yourself...
- ... but also for your (future) **peers**
- Easier to **share** and **re-use**

It makes your research **reproducible**

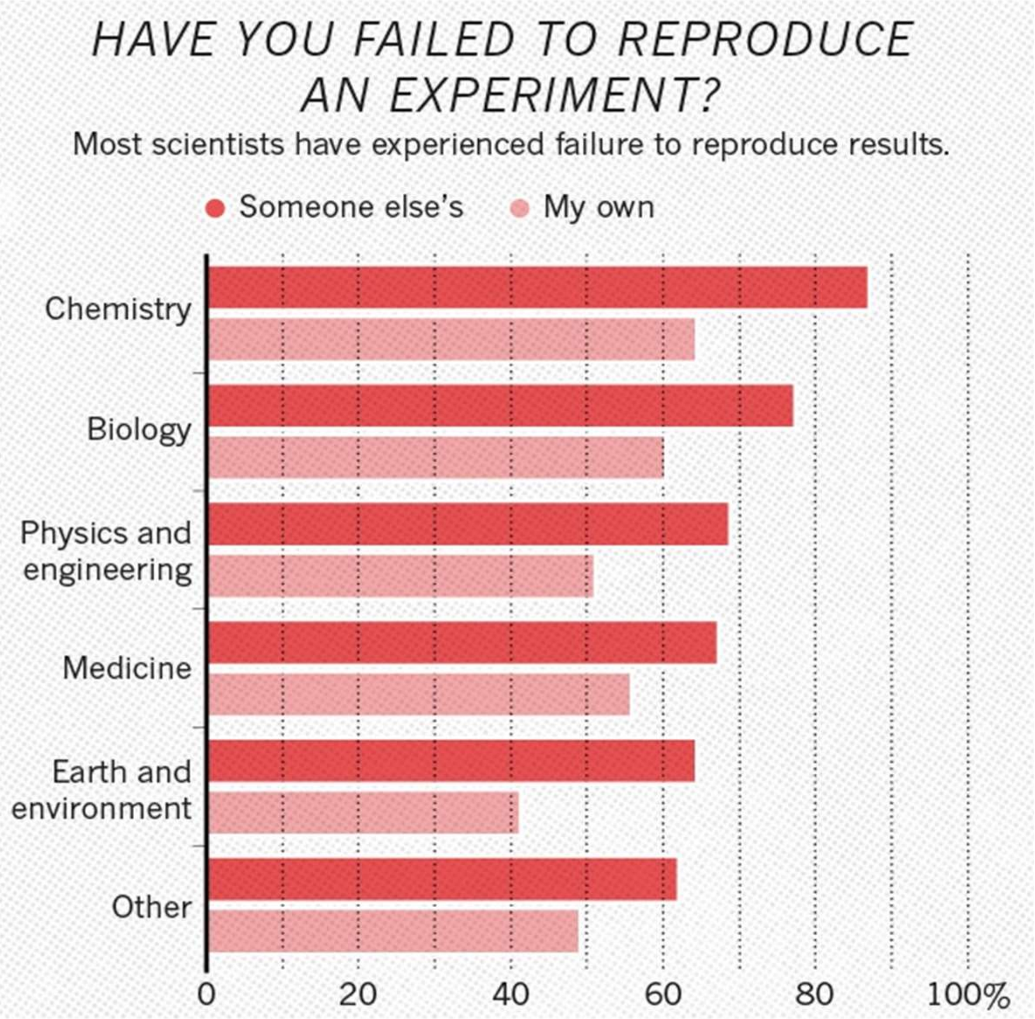


Reproducibility is the possibility for a research paper to be verified, re-used and continued. It applies to both **data and methods.**

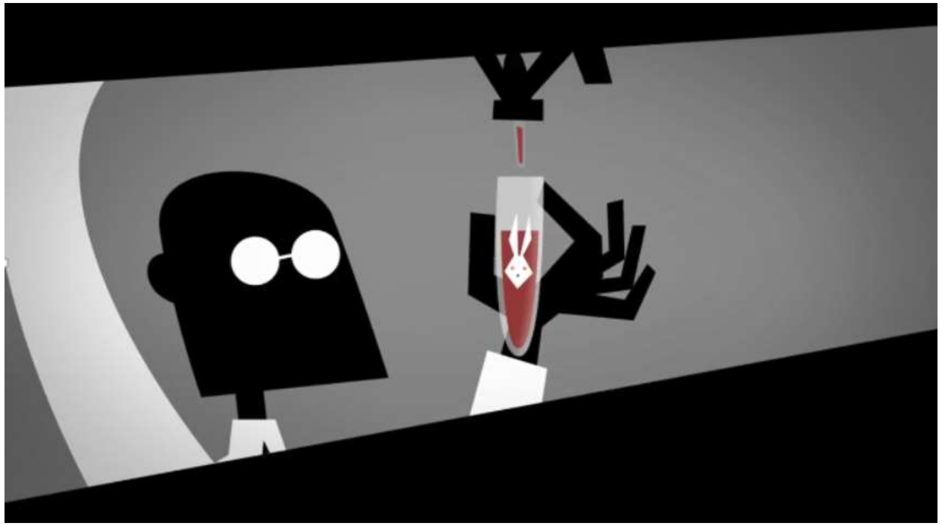
It is what makes your research alive, useful and trustworthy

Why is it so important?

Nature 533, 452–454 (26 May 2016) doi:10.1038/533452a



Data management and ethics



Numerous famous cases:
2014 – The case of Obokata
2020 - Retraction of a paper that held claims on hydroxychloroquine based on fabricated data. This had consequence on COVID-19 gov policies: LancetGate

Retraction Watch

Research lives in a **paradoxical context** that may push us towards questionable practice.

The **pressure to publish results that are always** new and spectacular implicitly encourages **tweaking** our work**, consciously or not**, until you get a result that fits arbitrary criteria,

Between plain fraud to best practice, there are **grey areas** in which we **must make the best choices** possible to ensure reproducibility

Irreproducible science can be suspicious

Fraud = falsification, fabrication, plagiarism -> no tolerance

Data management and ethics

Sometimes it is plain fraud, sometimes it is in a grayer area :

- Pressure to publish
- Pressure to find new, **spectacular** results

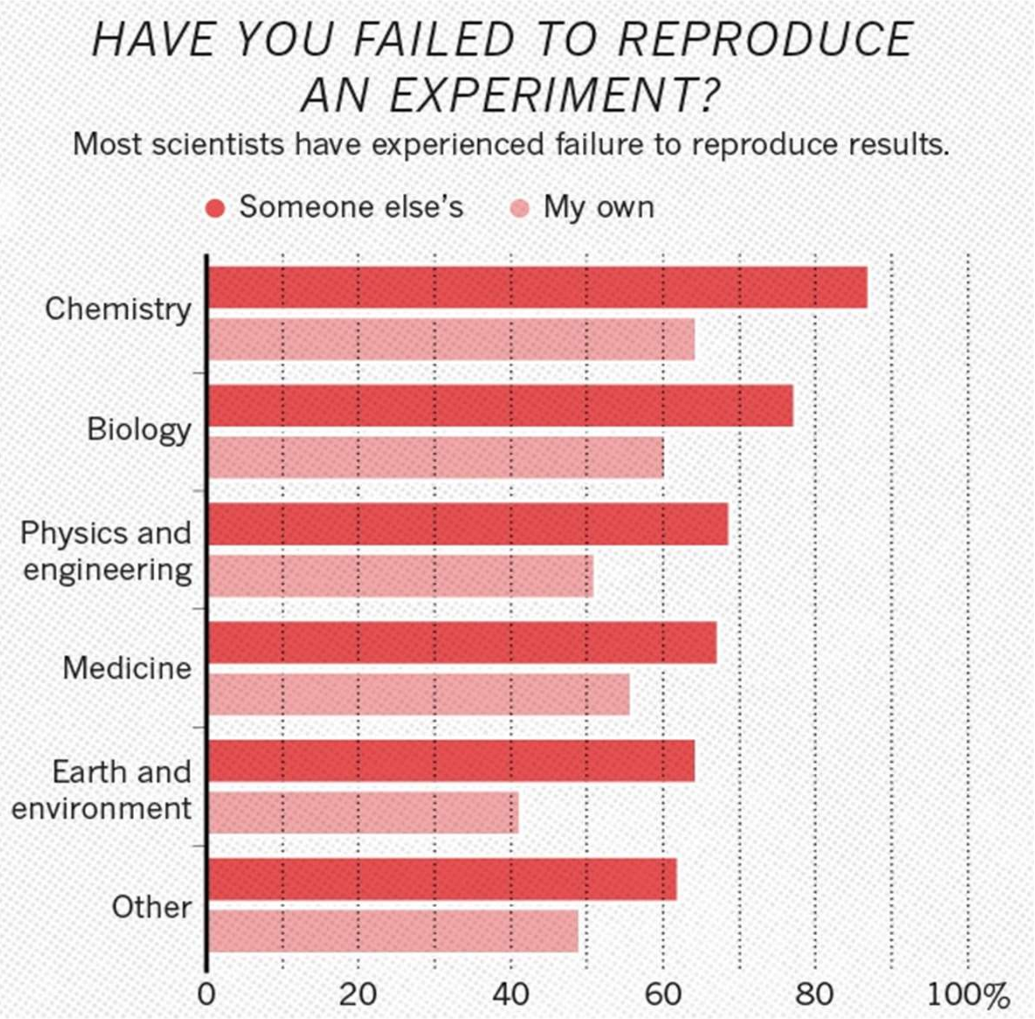
-> Numerous ways to **tweak** your study, **consciously or not**, until you get a result that fits arbitrary criteria

  - Altering how long it lasts
  - Play with the sample size
  - Reporting on only the items in the sample that fit a hypothesis (cherry-picking)
  - Selecting parts of your experimental design to report (outcome switching)
  - P-hacking (collecting lots of variables and playing with data until finding counts as statistically significant)

-> Reproducibility is sometimes much more **trivial** and finds its root in **data re-useability, availability, or traceability of processes**, when data and methodology management is not properly documented and the whole complexity of a few years of project is uneasy to navigate

Why is it so important?

Nature 533, 452–454 (26 May 2016) doi:10.1038/533452a

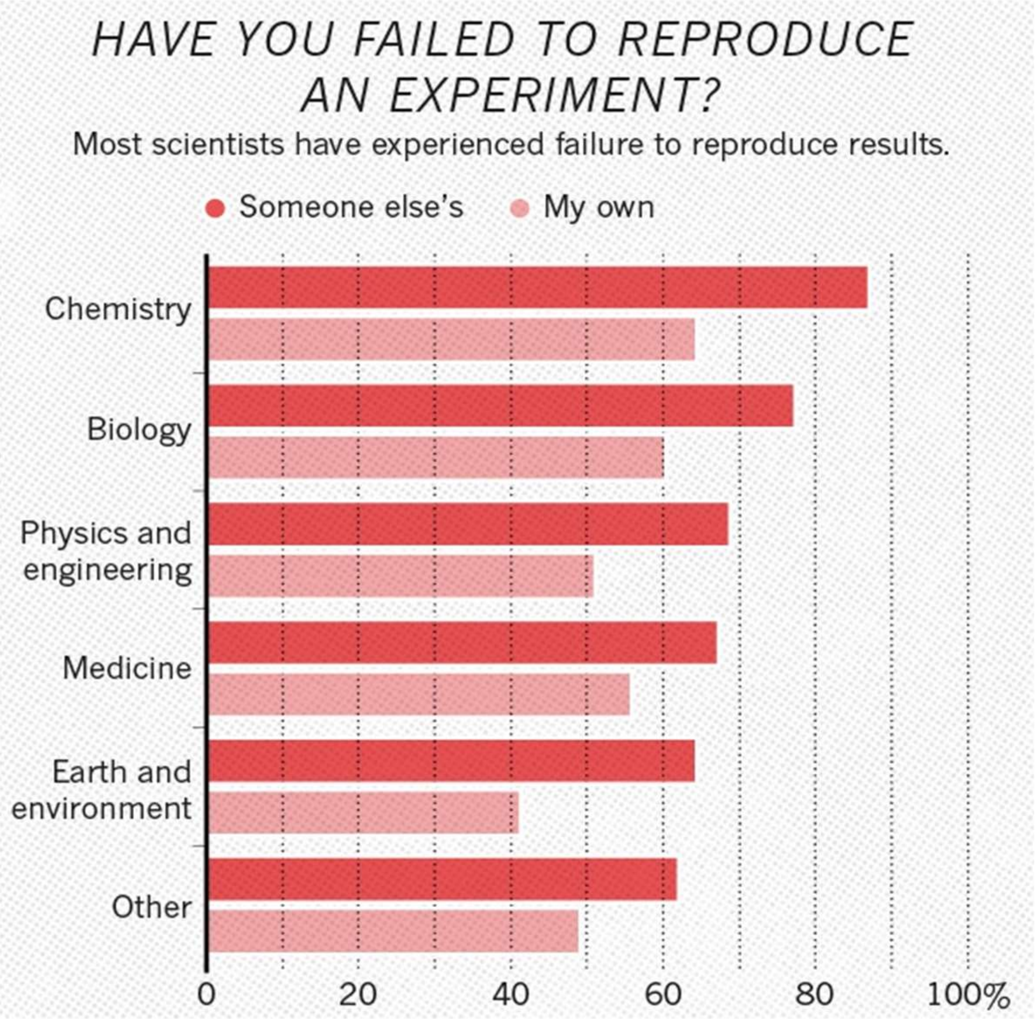


Reproducibility crisis

- Most scientific results are difficult, even **impossible**, to reproduce and/or replicate [*]
- This issue stems from a general **context that does not favour scientific integrity** but can push research towards cutting corners, selective reporting, poor documentation, data unavailability or even fraud

Why is it so important?

Nature 533, 452–454 (26 May 2016) doi:10.1038/533452a



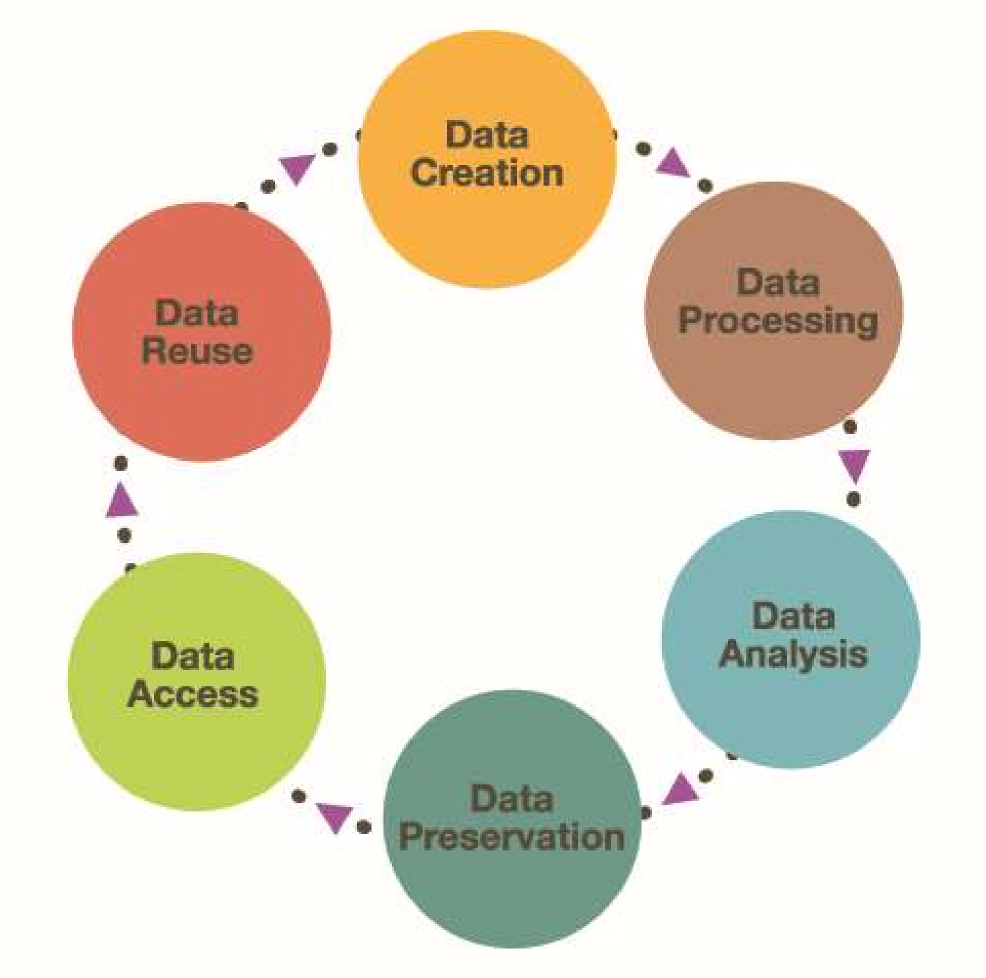
Reproducibility crisis

- This is **not a decrease** in researchers skills but **a cultural phenomenon**, because of the paradoxical system that rules research culture (publish or perish)
- More and more stakeholders are initiating a **cultural change** towards more reproducibility

You can be this change

Data planning

Data Life Cycle



EARLY ON : DATA PLANNING

What is my data like? What are the applicable regulations? Who do they belong to?

Nature, format, volume, source, collection, access…

DURING : DATA HANDLING

How should I store them?

Safety, size, security, backups, documentation…

How am I using my data?

Methodology, quality control

TOWARDS THE END: DATA SHARING

How should I share my data? What happens to my data after my project is over?

Open and FAIR data, licences, data sustainability and re-use

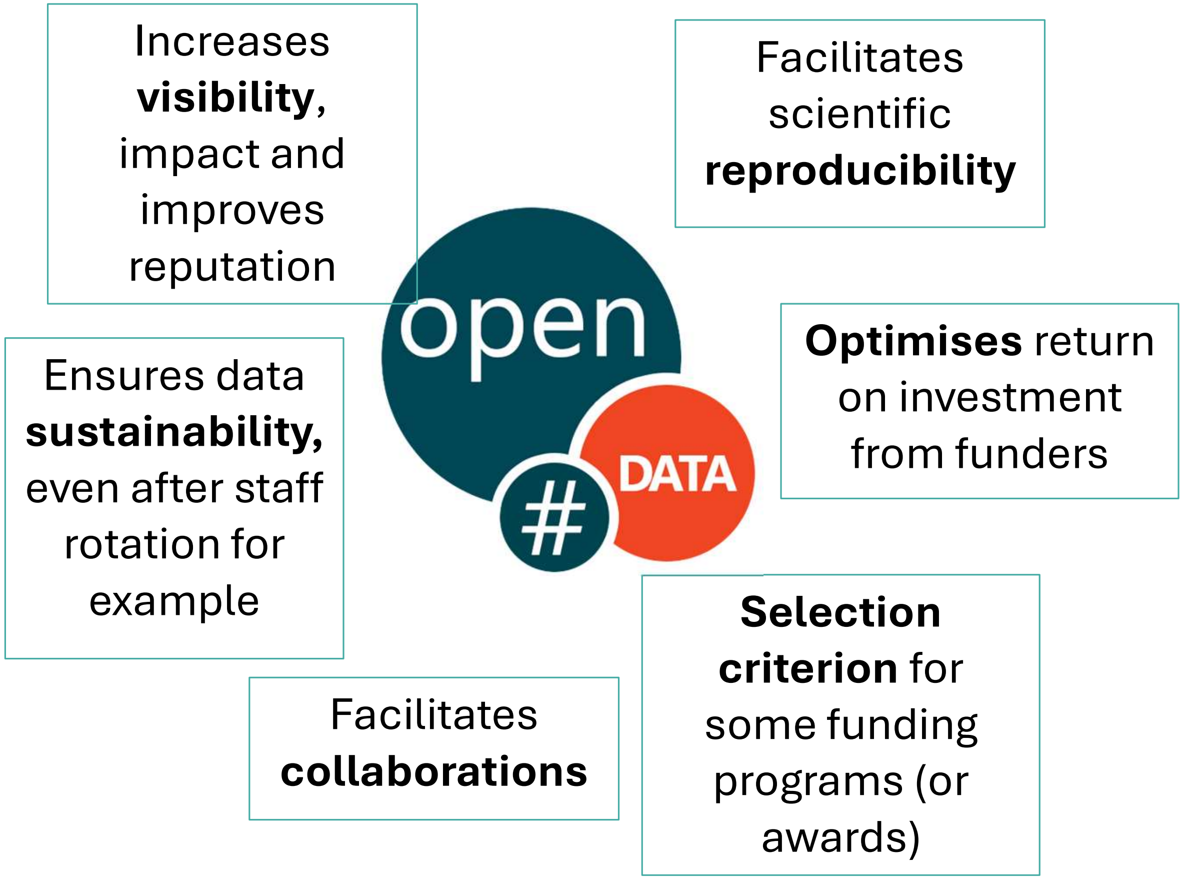
The data FAIRness spectrum

Most European funding agencies encourage sharing scientific results, methods and data.
They refer to the **« as open as possible, as closed as necessary »** principle.

The aim is therefore to practice as much **open data** as possible.

The data FAIRness spectrum

Most European funding agencies encourage sharing scientific results, methods and data.
They refer to the « **as open as possible, as closed as necessary** » principle.



Increases **visibility**, impact and improves reputation

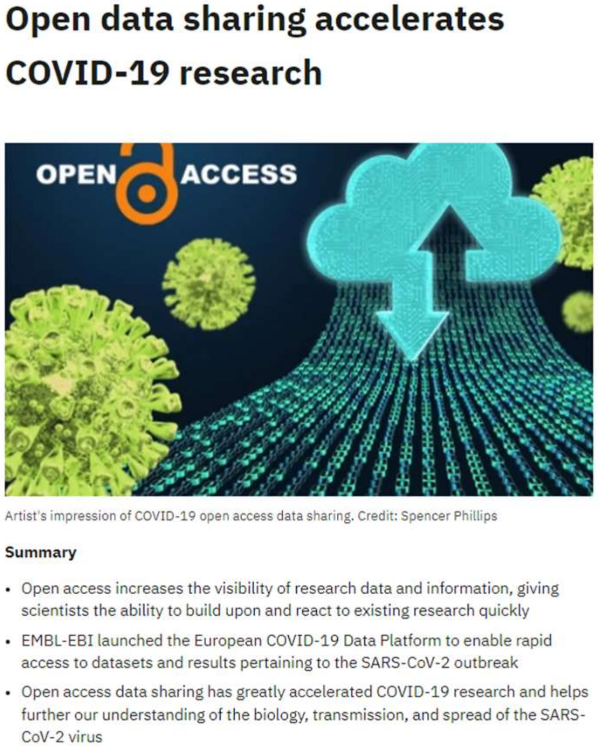
Facilitates scientific **reproducibility**

Ensures data **sustainability,** even after staff rotation for example

Optimises return on investment from funders

Facilitates **collaborations**

Selection criterion for some funding programs (or awards)



Open data sharing accelerates COVID-19 research

Artist's impression of COVID-19 open access data sharing. Credit: Spencer Phillips

Summary

- Open access increases the visibility of research data and information, giving scientists the ability to build upon and react to existing research quickly
- EMBL-EBI launched the European COVID-19 Data Platform to enable rapid access to datasets and results pertaining to the SARS-CoV-2 outbreak
- Open access data sharing has greatly accelerated COVID-19 research and helps further our understanding of the biology, transmission, and spread of the SARS-CoV-2 virus

Victoria Hatch, EMBL-EBI News, Oct 19, 2020

The data FAIRness spectrum

Most European funding agencies encourage sharing scientific results, methods and data.
They refer to the **« as open as possible, as closed as necessary »** principle.

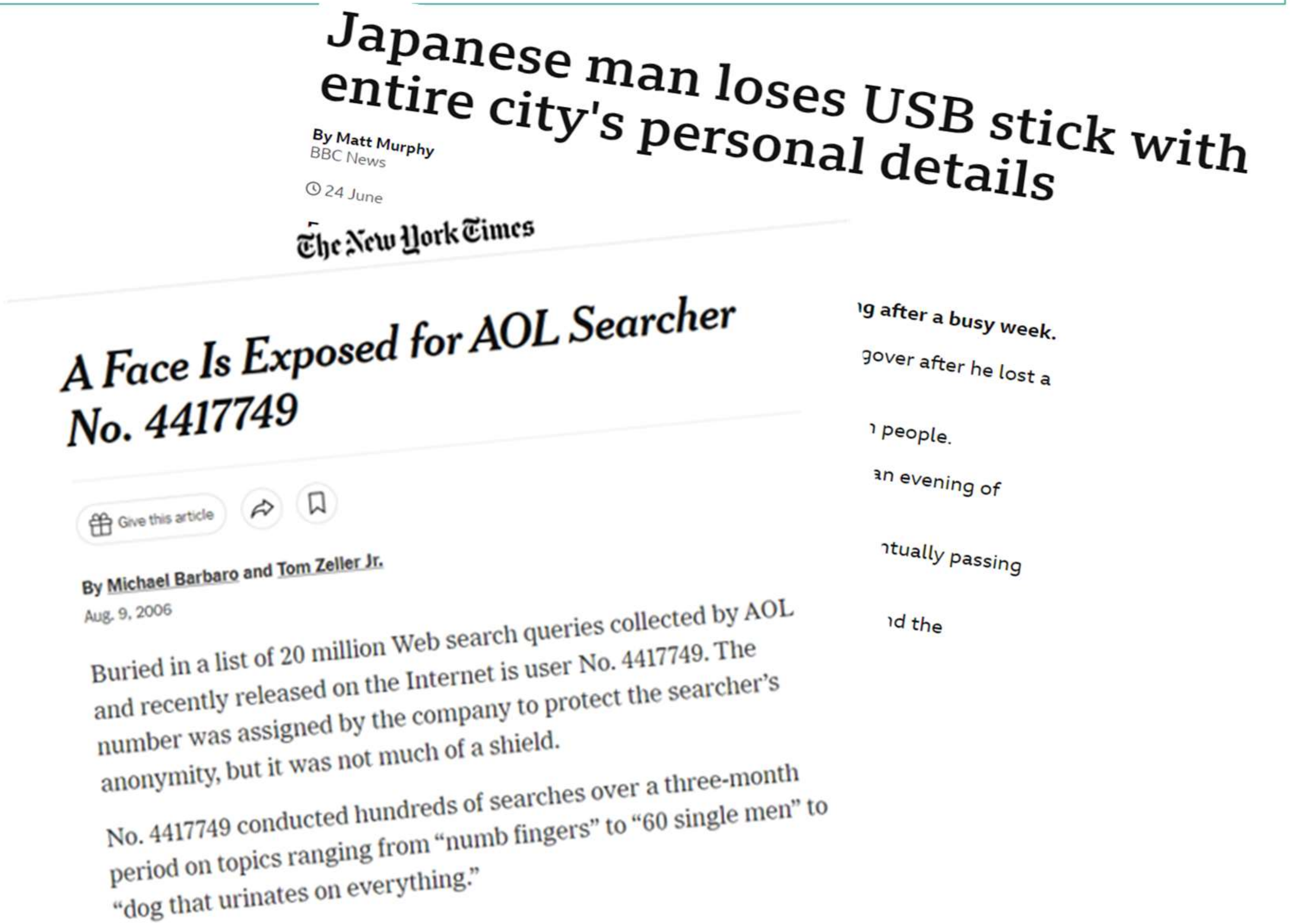
The aim is therefore to practice as much **open data** as possible.

However, open data is **not always possible** or not always the best way to go, or even not the only recommendation that should be observed **(why?)**

The data FAIRness spectrum

Many questions of data management, specifically access, storage, protection and sharing, have **roots in applicable rules and regulations**

-> awareness is a good start !



Japanese man loses USB stick with entire city's personal details

By Matt Murphy
BBC News

24 June

ıg after a busy week.

gover after he lost a

ı people.

an evening of

ıtually passing

ıd the

The New York Times

A Face Is Exposed for AOL Searcher No. 4417749

By Michael Barbaro and Tom Zeller Jr.

Aug. 9, 2006

Buried in a list of 20 million Web search queries collected by AOL and recently released on the Internet is user No. 4417749. The number was assigned by the company to protect the searcher's anonymity, but it was not much of a shield.

No. 4417749 conducted hundreds of searches over a three-month period on topics ranging from "numb fingers" to "60 single men" to "dog that urinates on everything."

The data FAIRness spectrum

Most European funding agencies encourage sharing scientific results, methods and data.
They refer to the « **as open as possible, as closed as necessary** » principle.

The aim is therefore to practice as much **open data** as possible.

However, open data is **not always possible** or not always the best way to go, or even not the only recommendation that should be observed **(why?)**

Data that cannot be shared

For legal reasons (GDPR, NDA, copyright…)
For ethical reasons (risks)
For strategic reasons (patents -> embargo)

Note : good RDM habits are also for oneself ☺

Open data

Not always a token of quality

Not always re-usable straight away (it is not just about posting online)

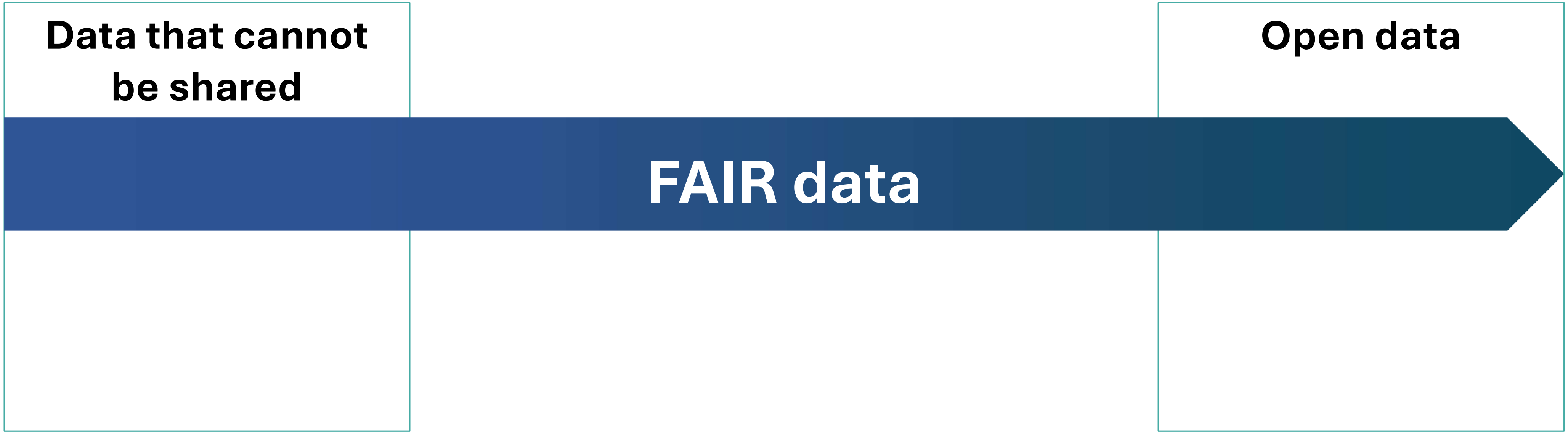
Should be the direction if not the destination

The data FAIRness spectrum

Most European funding agencies encourage sharing scientific results, methods and data.
They refer to the « **as open as possible, as closed as necessary** » principle.

The aim is therefore to practice as much **open data** as possible.

However, open data is **not always possible** or not always the best way to go, or even not the only recommendation that should be observed **(why?)**



Data that cannot be shared

FAIR data

Open data

The data FAIRness spectrum

Findable

Accessible

Interoperable

Reusable

FAIR data

The data FAIRness spectrum

Findable

Data are discoverable and easy to find, by both humans and computers.

- **Metadata : author, date, DOI, contact, keywords, ...**

In most cases, at least the metadata can be shared

Accessible

Data are made available in a **sustainable** way, even after the project is over:

- The (meta)data are retrievable with a flexible protocol in an **open directory** (harvesting)
- If the data cannot be shared, it has to be justified

Interoperable

Data are able to be operated / exchanged / compared between a variety of institutions, workflows, software, applications, systems, ...

- The (meta)data use a broadly compatible format (not proprietary if possible)
- The documentation is in English

Reusable

The data are **sufficiently described** and can be shared with as few restrictions as possible, as the ultimate goal is to optimise data reuse (licenses, formats and docs)

FAIR data

The data FAIRness spectrum

Findable

Data are discoverable and easy to find, by both humans and computers.

- **Metadata : author, date, DOI, contact, keywords, ...**

In most cases, at least the meta

Accessible

Data are made available in a **sustainable** way, even after the project is over:

- The (meta)data th a flexible sting) be justified

Interoperable

Data are able compared bet workflows, softw systems, ...

- The (meta)dat a broadly compatible format (not proprietary if possible)
-

sufficiently described and can be shared with as few restrictions as possible, as the ultimate goal is to optimise data reuse (licenses, formats and docs)

Using a data repository checks most boxes

But how to choose a repository?

FAIR data

Data repositories

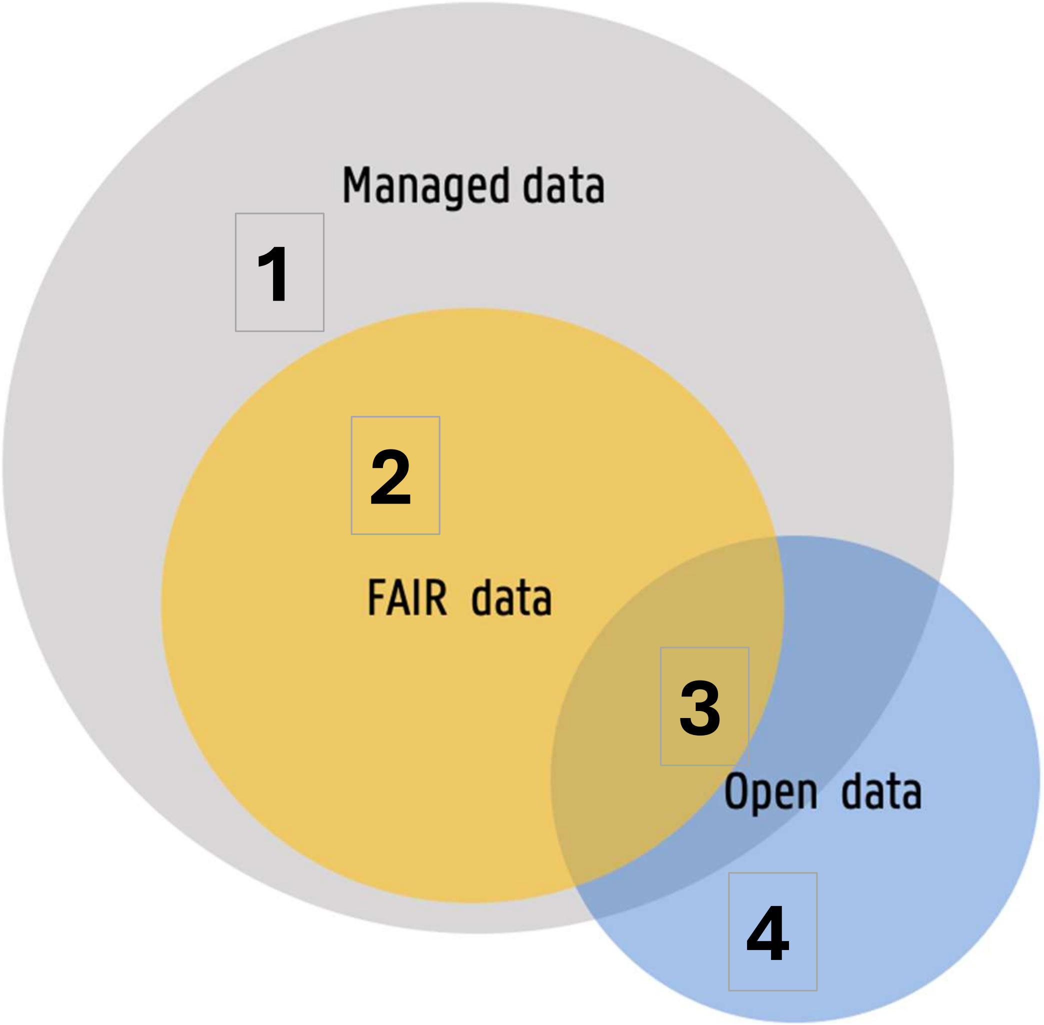
How do I select a data repository?

General	Discipline-specific
Zenodo OSF Figshare Dataverse Institutional repositories (e.g. ULiège Dataverse)	Some examples : The QDR or Bequali (HSS), CDS (astro), NCBI (genomics), .. Catalogs of directories : Re3data, FAIRsharing Ask your peers and supervisor

A good repository:

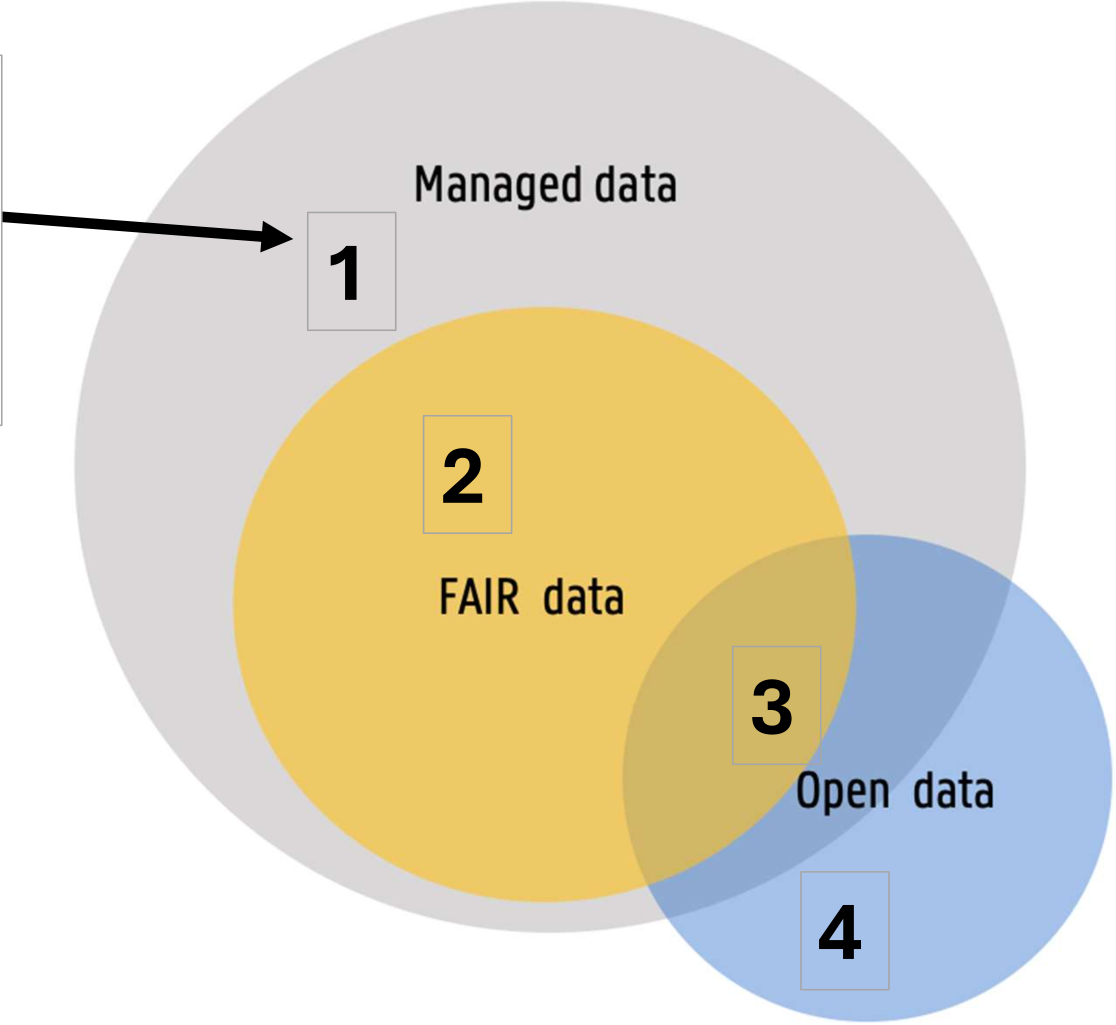
- Is **recognized** by your peers
- Provides a persistent **identifier** such as a DOI or handle
- Comes with a few possibilities for **licenses**
- Has high documentation **metadata standards** with controlled vocabularies (therefore discipline-specific is usually better)
- Lets you **keep all your rights**
- Has a certification such as CoreTrustSeal

The bigger picture



The bigger picture

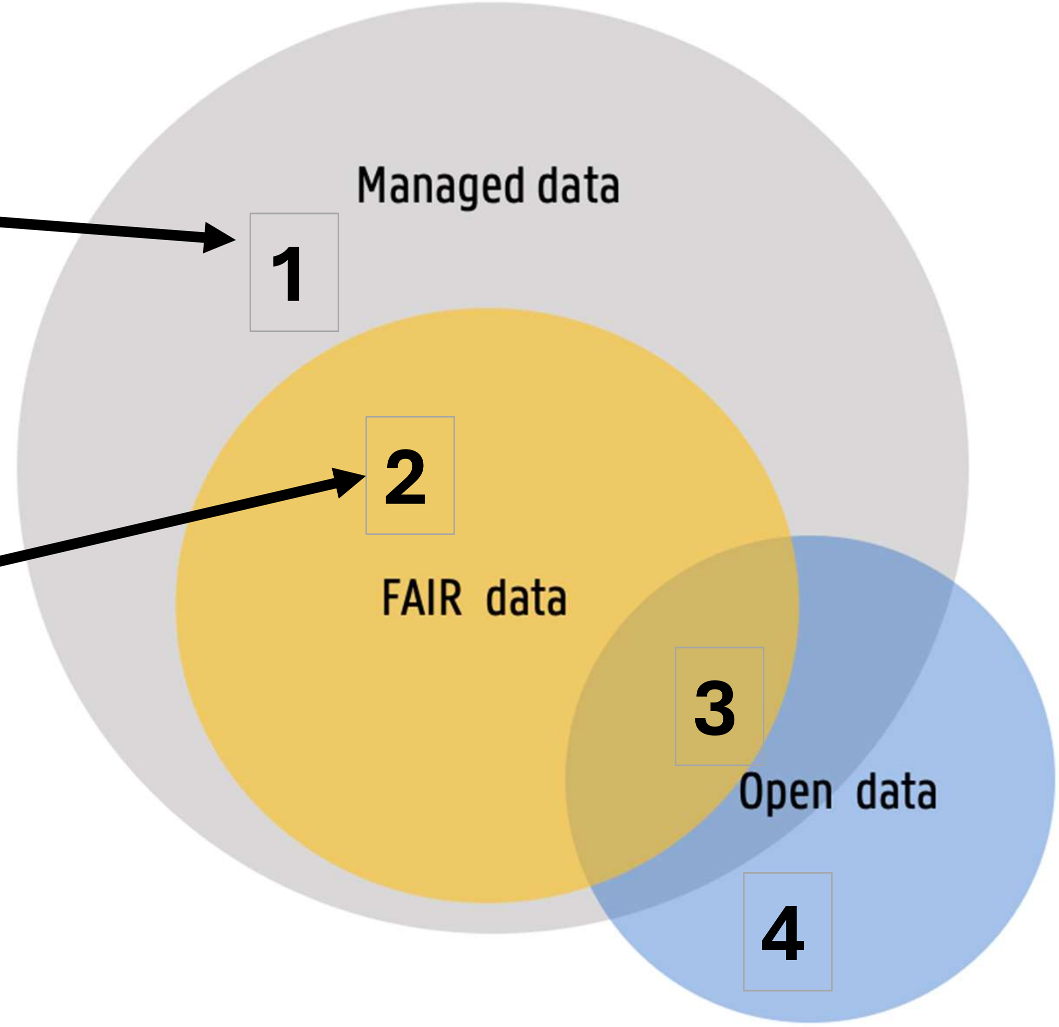
Data are safely and securely stored, in a sustainable way, well documented and in compliance with rules and regulations...



The bigger picture

Data are safely and securely stored, in a sustainable way, well documented and in compliance with rules and regulations...

... they are suitable for sharing and understanding outside of your own lab, or at least the metadata is, so it is made available in a community-standard compliant way...

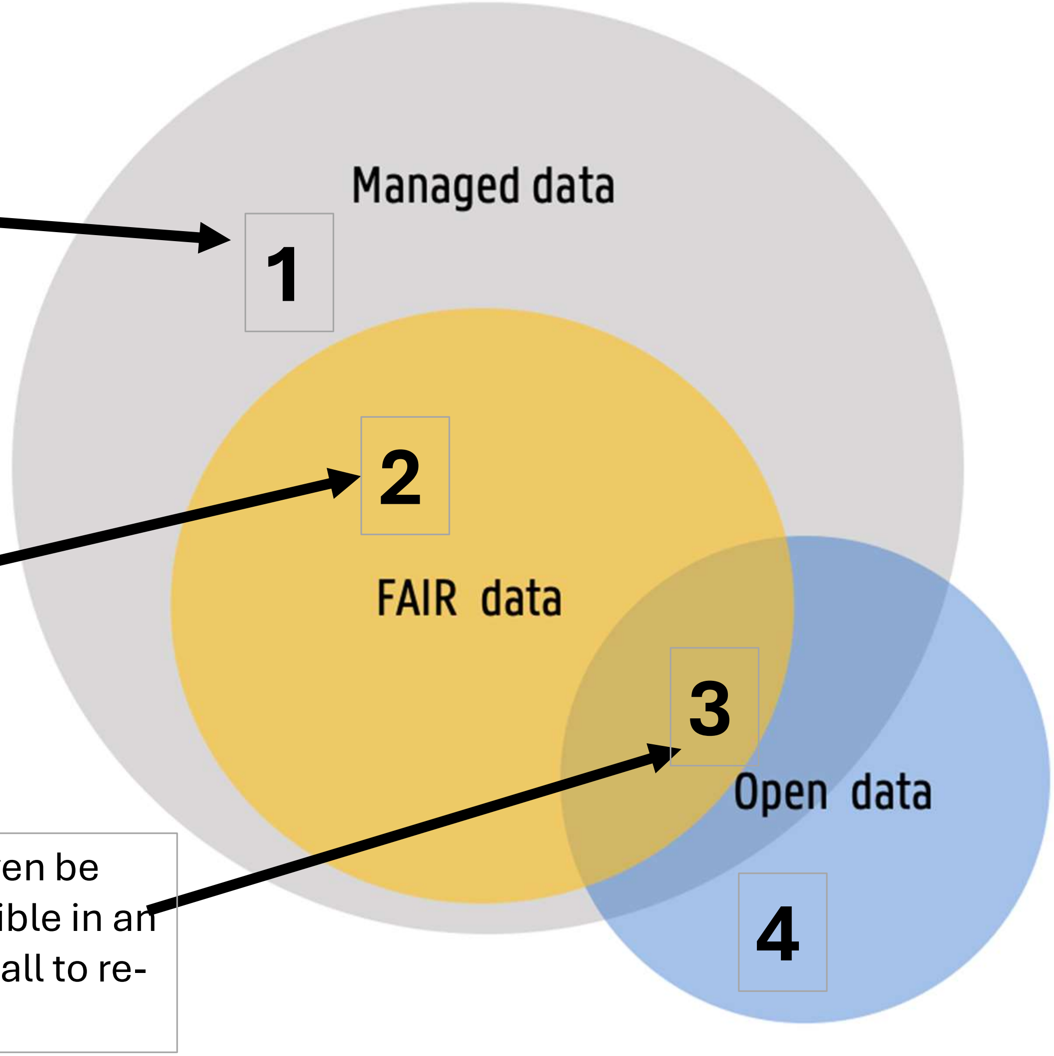


The bigger picture

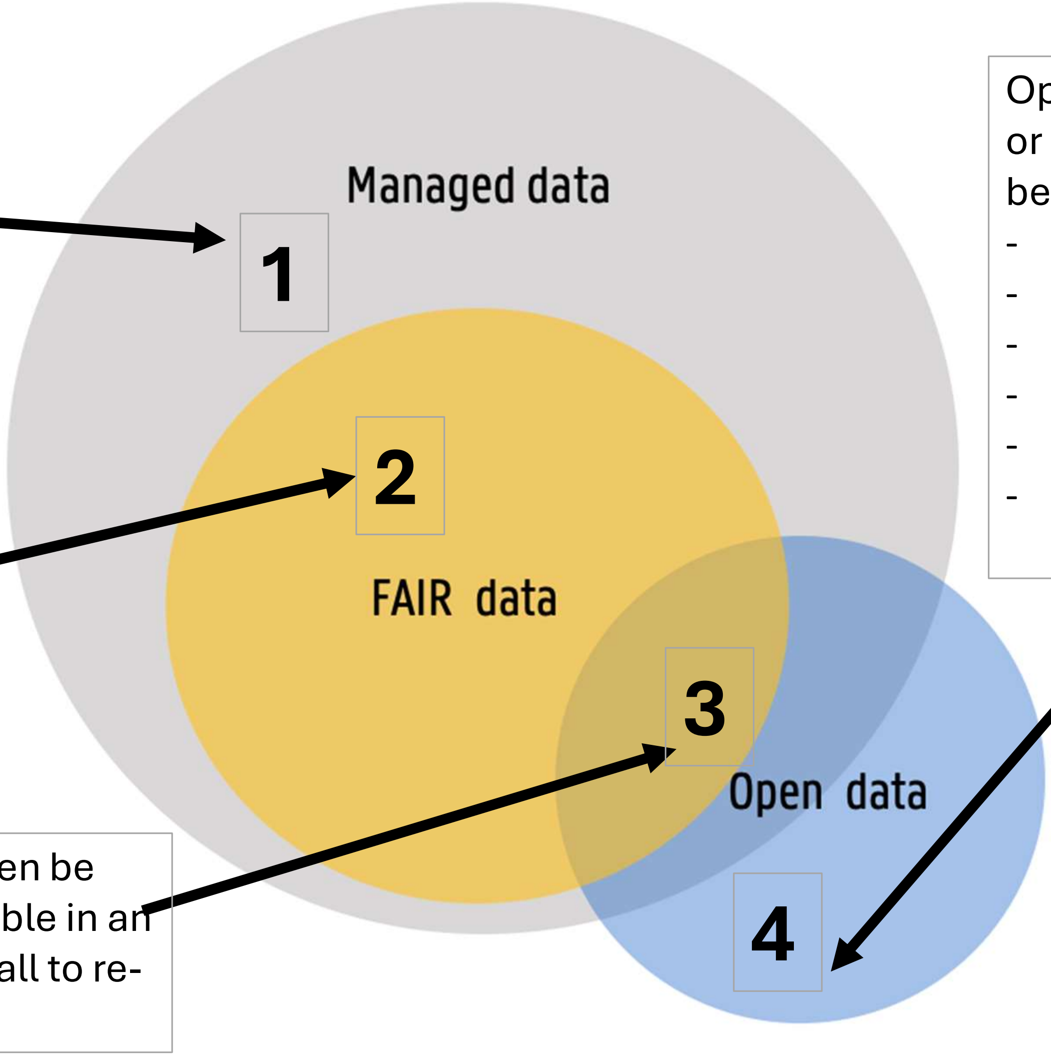
Data are safely and securely stored, in a sustainable way, well documented and in compliance with rules and regulations...

... they are suitable for sharing and understanding outside of your own lab, or at least the metadata is, so it is made available in a community-standard compliant way...

... they can even be made accessible in an open way, for all to re-use



The bigger picture



Data are safely and securely stored, in a sustainable way, well documented and in compliance with rules and regulations…

… they are suitable for sharing and understanding outside of your own lab, or at least the metadata is, so it is made available in a community-standard compliant way…

… they can even be made accessible in an open way, for all to re-use

Open data that is not FAIR or not well managed can be:

- Fraudulent
- Of poor quality
- Poorly described
- Lacking metadata
- In a proprietary format
- …

The bigger picture

No panic:

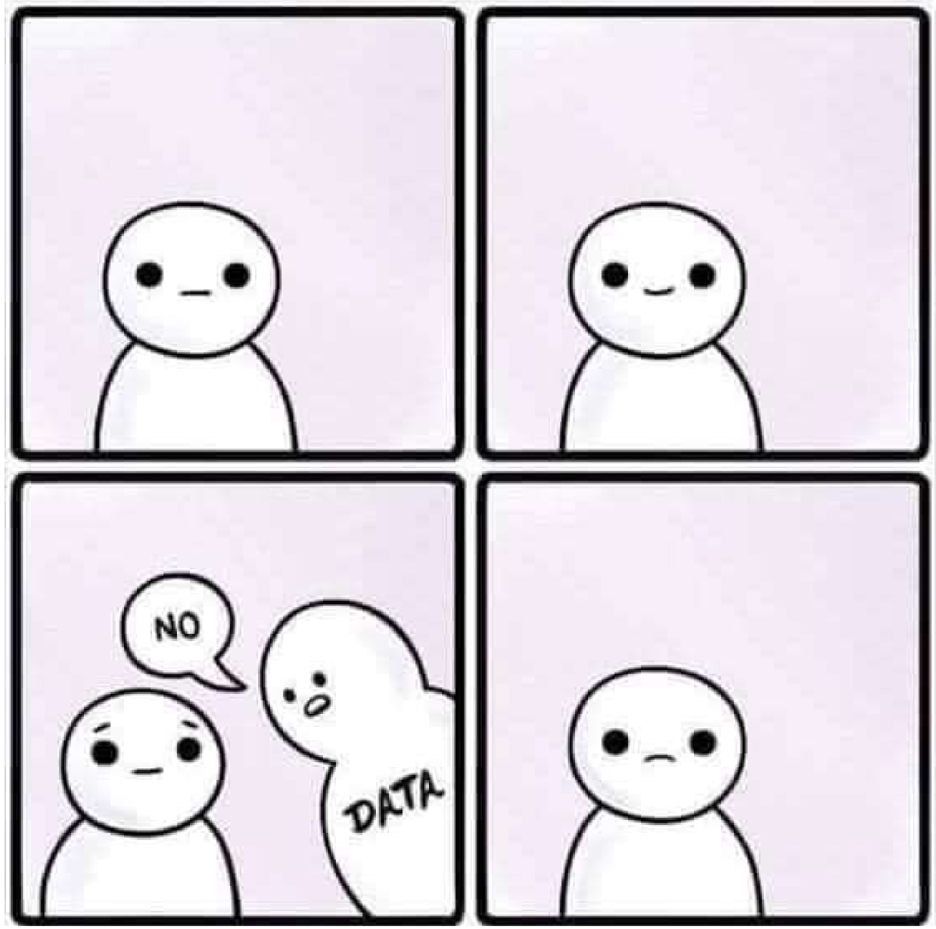
- It is absolutely okay to « play around » with datasets
- The difference between data exploration and misconduct is **transparency** in publication and **traceability** in your day-to-day

Documentation and traceability are the key:

Making raw data, protocols, methodologies…

- **As open as possible, as closed as necessary**
- At least **traceable**

The (real) scientific method.



Additional reading (added after presentation, based on discussions)

- Tips for metadata : https://guides.lib.unc.edu/metadata/home

  Resources from the FOSTER project : https://www.fosteropenscience.eu/content/foster-open-science-training
- Complete guidance from the Data Curation Center (UK) : https://www.dcc.ac.uk/
- FAIR assessment tool: https://www.f-uji.net/
- A list of additional standards for inspiration : https://guides.lib.unc.edu/metadata/standards
- How Anonymous are you really? https://www.ooa.world/
- Main paper about reproducibility crisis:

  https://journals.plos.org/plosmedicine/article?id=10.1371/journal.pmed.0020124
- Nature Poll about reproducibility: https://www.nature.com/articles/533452a
- Evolution of research evaluation: https://coara.eu/