Research Data Management and Reproducibility

Good habits for good research

Introduction to scientific computing – GIGA Doctoral School
Oct 17, 2022

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

In most cases, what is the very first step of a PhD research ?

In most cases, what is the very first step of a PhD research ?

You are constantly reusing research results or data

Reproducible research

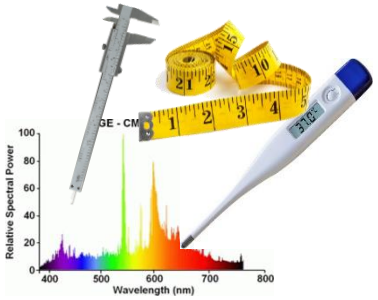
Reproducibility is the possibility for a research paper to be verified, re-used and continued. It applies to both **data and methods.**

Your thesis is meant to continue living after its end and be re-used as much as possible.
It is what makes your research alive, useful and trustworthy

Reproducibility

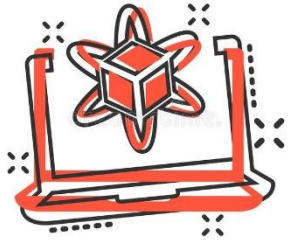
What is research data?

Documents, tables, maps

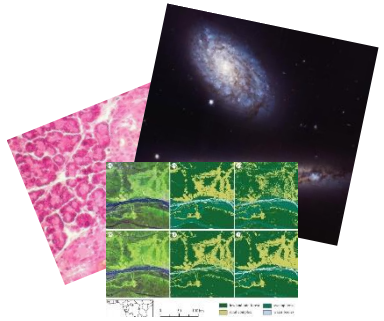


Experimental results

Polls, forms

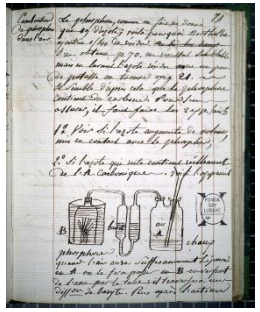


Simulation results



Images, videos

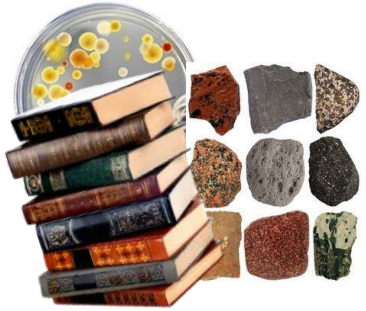
Audio files



Lab notes, field work notes

DNA sequences

Papers, publications



Physical samples

What is the use of research data?

Data are at the **core** of your research:

-> They enable the process of **answering your research question**

-> They provide **validation** or nuance to your working **hypotheses**

-> They usually contribute to the choice/design of your **methodology**

-> They may have an impact on the **quality** of your results

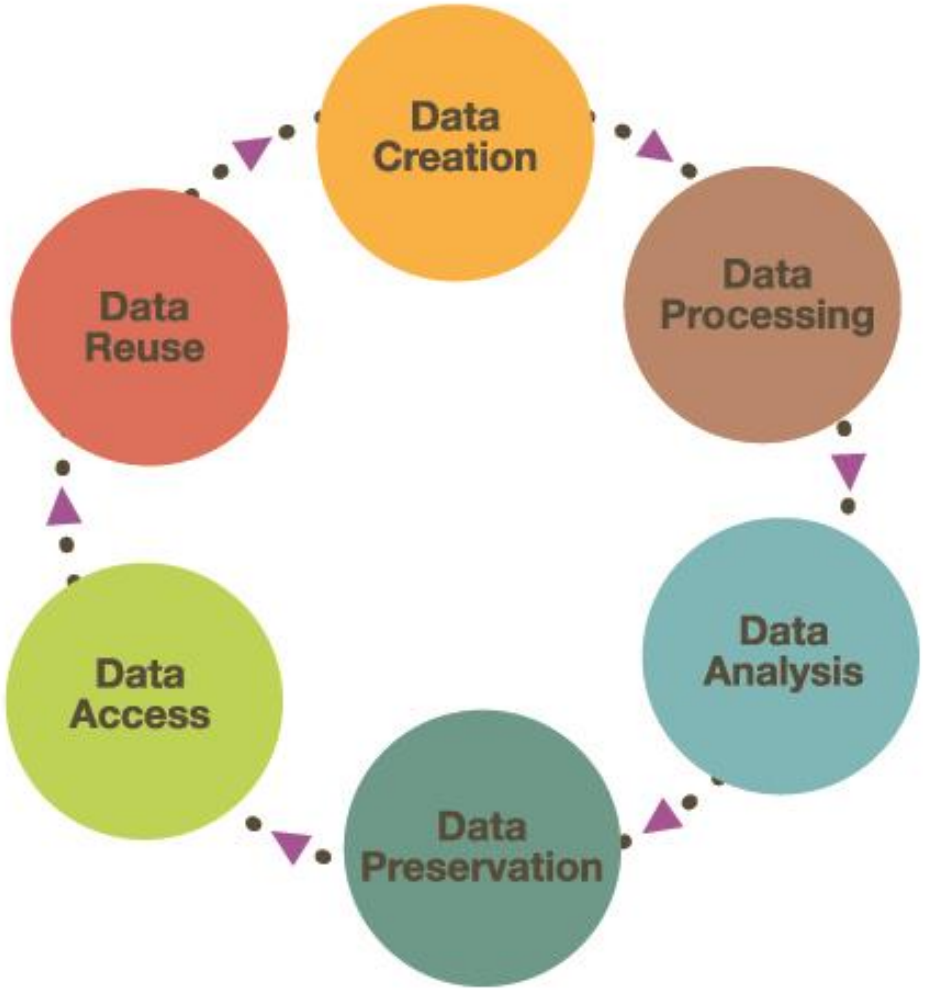
-> They sometimes carry an **economical** value

They ought to be **well-understood**, treated with **care** and go **through high-quality processes**

-> **Reasearch Data Management (RDM)**

What is the use of research data?

Data Life Cycle



Set of practices around research data, including but not limited to:

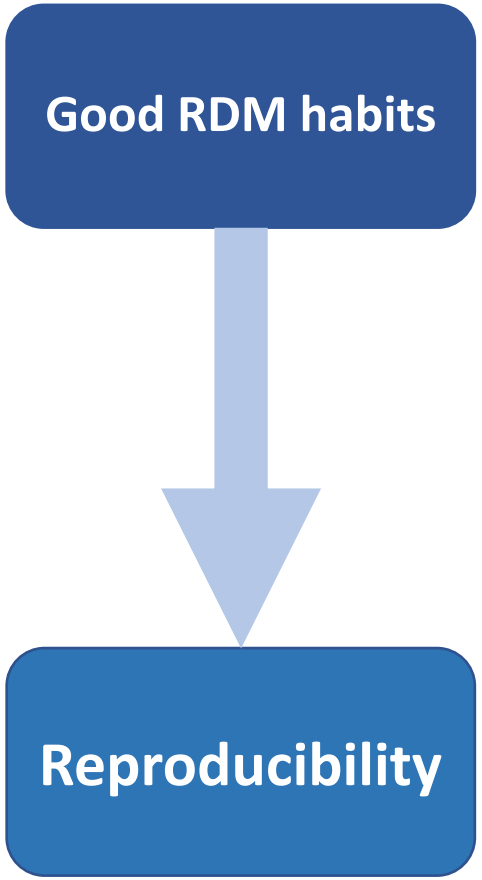
-> collecting (first / second hand)
-> storage, curation
-> documentation
-> formatting
-> filtering, sampling
-> analysis
-> publishing, sharing

Responsible RDM:
-> Adopting **good habits** for each of these tasks, so that research data get easier to use, to share, and to re-use

Reproducible research

Caring about data sustainability automatically makes your data:

- Better **organised**, **protected** and **compliant**
- Easier to **use** and to **understand** for yourself...
- ... **but also for your (future) peers**
- Easier to **re-use** and maybe even to **share**
- To sum up, it makes your research **reproducible**

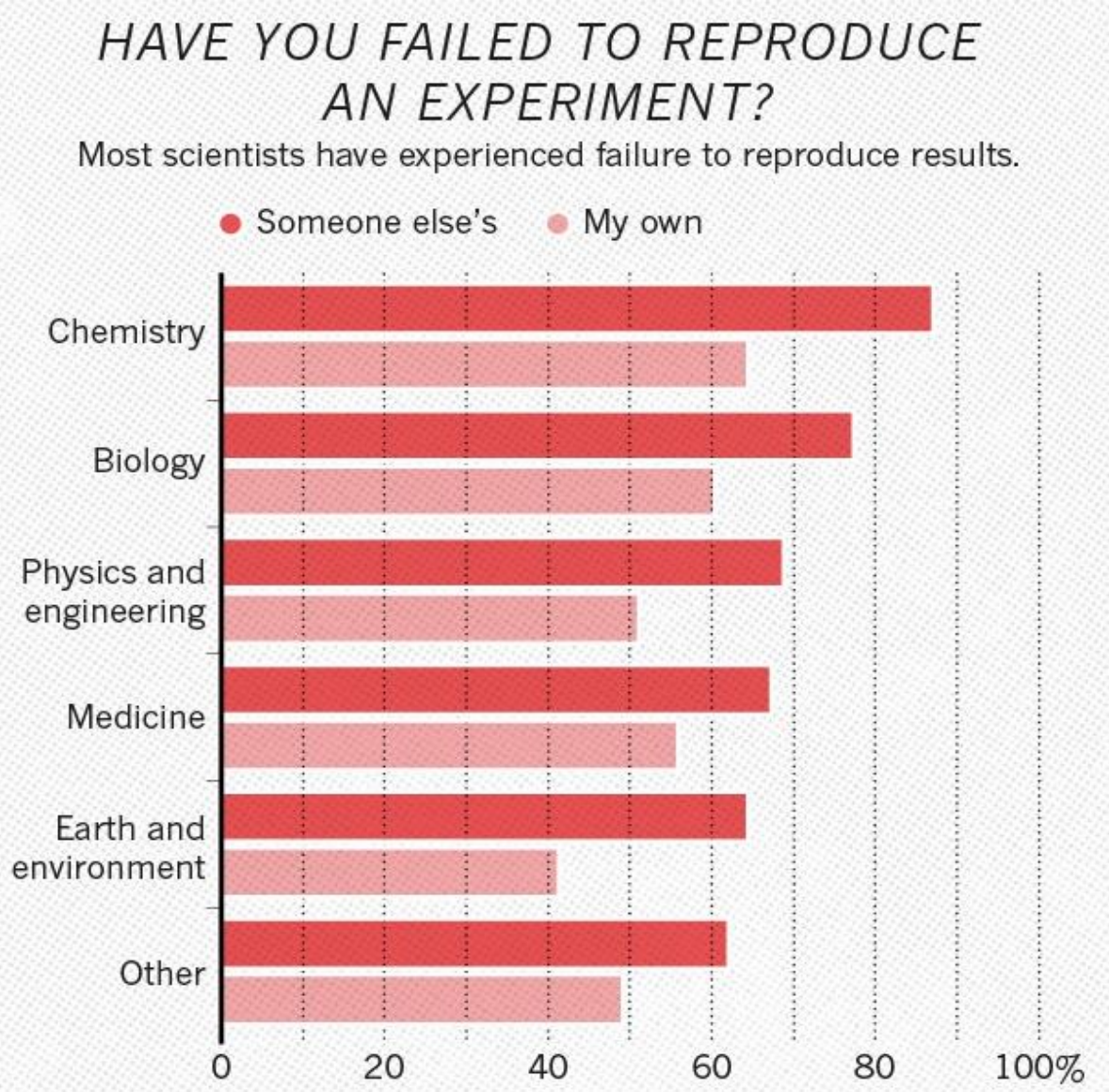


Reproducibility is the possibility for a research paper to be verified, re-used and continued. It applies to both **data and methods.**

Your thesis is meant to continue living after its end and be re-used as much as possible.
It is what makes your research alive, useful and trustworthy

Why is it so difficult?

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



Why is it so important?

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 or even fraud

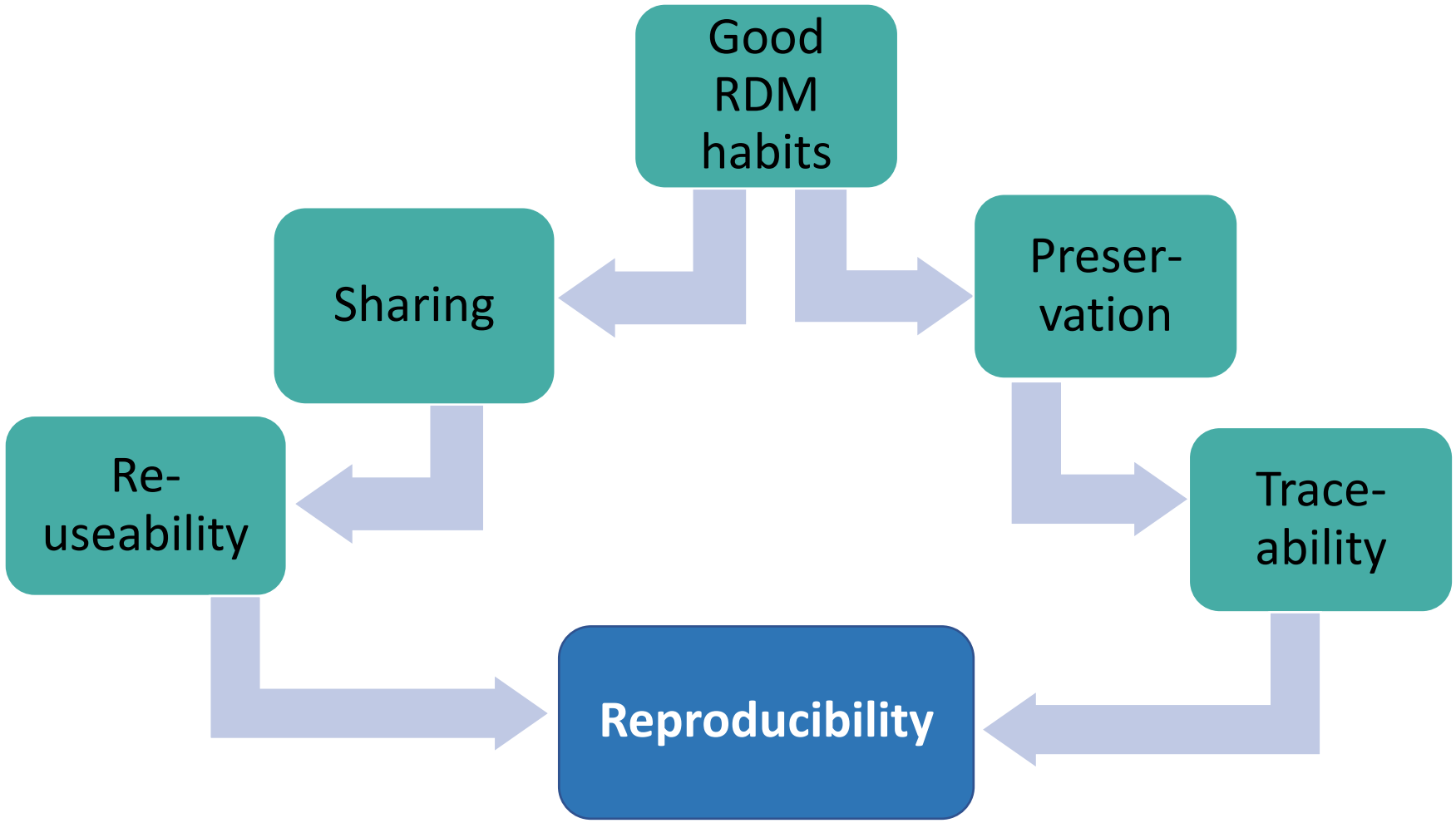
Why is it so important?

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 or even fraud
- 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

Why is it so important?



Reproducible research

We have the « why » pretty much covered, let us get to the « **how** »

- Data FAIRness
- Data storage
- Data security
- Data documentation
- Metadata

- Data sharing
  - Regulations
  - Data repositories
  - Licenses
  - Data protection

Data planning: rules and regulations

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

For many, after-work drinks are a common way of relaxing after a busy week.

But one worker in Japan could be nursing a protracted hangover after he lost a USB memory stick following a night out with colleagues.

Why? It contained the personal details of nearly half a million people.

The unnamed man placed the memory stick in his bag before an evening of drinking in the city of Amagasaki, north-west of Osaka.

He spent several hours drinking in a local restaurant before eventually passing out on the street, local media reported.

When he eventually came around, he realised that both his bag and the memory stick were missing.

BBC NEWS

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.

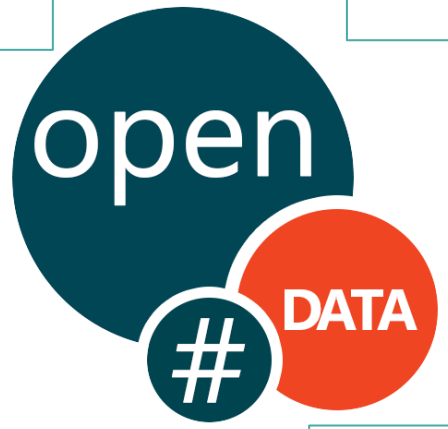
The data FAIRness spectrum

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

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?)**

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**
- Digital Object Identifier
- Other standard identifier

Accessible

Interoperable

Reusable

FAIR data

The data FAIRn

Paper metadata

The Location of Young Pulsar PSR J0837−2454: Galactic Halo or Local Supernova Remnant?

Show affiliations

Pol, Nihan; Burke-Spolaor, Sarah; Hurley-Walker, Natasha; Blumer, Harsha; Johnston, Simon; Keith, Michael; Keane, Evan F.; Burgay, Marta; Possenti, Andrea; Petroff, Emily; Bhat, N. D. Ramesh

We present the discovery and timing of the young (age ~ 28.6 kyr) pulsar PSR J0837−2454. Based on its high latitude ($b = 9.8^{o}$) and dispersion measure (DM $= 143$~pc~cm^{-3}), the pulsar appears to be at a z-height of >1 kpc above the Galactic plane, but near the edge of our Galaxy. This is many times the observed scale height of the canonical pulsar population, which suggests this pulsar may have been born far out of the plane. If accurate, the young age and high z-height imply that this is the first pulsar known to be born from a runaway O/B star. In follow-up imaging with the Australia Telescope Compact Array (ATCA), we detect the pulsar with a flux density $S_{1400} = 0.18 \pm 0.05$ mJy. We do not detect an obvious supernova remnant around the pulsar in our ATCA data, but we detect a co-located, low-surface-brightness region of $\sim 1.5^{o}$ extent in archival Galactic and Extragalactic All-sky MWA Survey data. We also detect co-located Hα emission from the Southern Hα Sky Survey Atlas. Distance estimates based on these two detections come out to ~0.9 kpc and ~0.2 kpc respectively, both of which are much smaller than the distance predicted by the NE2001 model (6.3 kpc) and YMW model (> 25 kpc) and place the pulsar much closer to the plane of the Galaxy. If the pulsar/remnant association holds, this result also highlights the inherent difficulty in the classification of transients as "Galactic" (pulsar) or "extragalactic" (fast radio burst) toward the Galactic anti-center based solely on the modeled Galactic electron contribution to a detection.

Publication: eprint arXiv:2104.11680

Pub Date: April 2021

arXiv: arXiv:2104.11680

Bibcode: 2021arXiv210411680P

Keywords: Astrophysics - High Energy Astrophysical Phenomena

E-Print Comments: Published in ApJ. 12 pages, 9 figures, 2 tables; doi:10.3847/1538-4357/abe70d

The data FAIRnes

Dataset metadata

Files | Metadata | Terms | Versions

Export Metadata

Citation Metadata

Dataset Persistent ID	doi:10.14428/DVN/FT2TX8
Publication Date	2021-02-05
Title	Replication Data for: Expectancy-value-cost motivational theory to explore final year medical students' research intentions and past research experience: a multicentre cross-sectional questionnaire study
Author	Van Maele, Louis (IRSS, CAMG, Université catholique de Louvain, Belgium) - ORCID: 0000-0003-1683-1207 Devos, Christelle (IPSY, Université catholique de Louvain, Belgium) Guisset, Séverine (SMCS, LIDAM, Université catholique de Louvain, Belgium) Leconte, Sophie (IRSS, CAMG, Université catholique de Louvain, Belgium) Macq, Jean (IRSS, Université catholique de Louvain)
Contact	Use email button above to contact. Van Maele, Louis (IRSS, CAMG, Université catholique de Louvain)
Description	The purpose of this dataset was to study final-year medical students' research intentions and motivation based on the Expectancy-Value-Cost motivational theory. The data comes from an online questionnaire sent in February 2017 to final-year medical students in three French-speaking Belgian universities (ULB, UCLouvain and ULg).
Subject	Medicine, Health and Life Sciences; Other
Keyword	Motivation (MeSH) https://www.ncbi.nlm.nih.gov/mesh/?term=motivation Medical Students (MeSH) https://www.ncbi.nlm.nih.gov/mesh/?term=medical+student Research (MeSH) https://www.ncbi.nlm.nih.gov/mesh/?term=activities%2C+research Questionnaire (MeSH) https://www.ncbi.nlm.nih.gov/mesh/?term=design%2C+questionnaire Belgium (MeSH) https://www.ncbi.nlm.nih.gov/mesh/?term=belgium
Production Date	2017-03-30
Production Place	Belgium
Depositor	Van Maele, Louis
Deposit Date	2020-12-02

The data FAIRness spectrum

Findable

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

- **Metadata**
- Digital Object Identifier
- Other standard identifier

In most cases, at least the metadata can be shared

Accessible

Interoperable

Reusable

FAIR data

The data FAIRness spectrum

Findable

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

- **Metadata**
- Digital Object Identifier
- Other standard identifier

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

Using a data repository usually checks most boxes

Interoperable

Reusable

FAIR data

The data FAIRness spectrum

Findable

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

- **Metadata**
- Digital Object Identifier
- Other standard identifier

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

Using a data repository usually checks most boxes

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

FAIR data

The data FAIRness spectrum

Findable

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

- **Metadata**
- Digital Object Identifier
- Other standard identifier

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

Using a data repository usually checks most boxes

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.

- The **licenses** are as open as possible.
- The format is as universal as possible
- The data is well documented

FAIR data

FAIR data sharing

As a scientist who wants to publish data:

Two possibilities:

- Deposit data and metadata in an online data repository
- Publish data as annex files to a paper

Should be anticipated as early as possible!

Attention points:

- **IPR** regulations and law
- **Patenting** regulations and laws
- **GDPR**
- Contracts with **third parties**
- Using a **license**

FAIR data sharing

A license defines how to **reuse** the content:

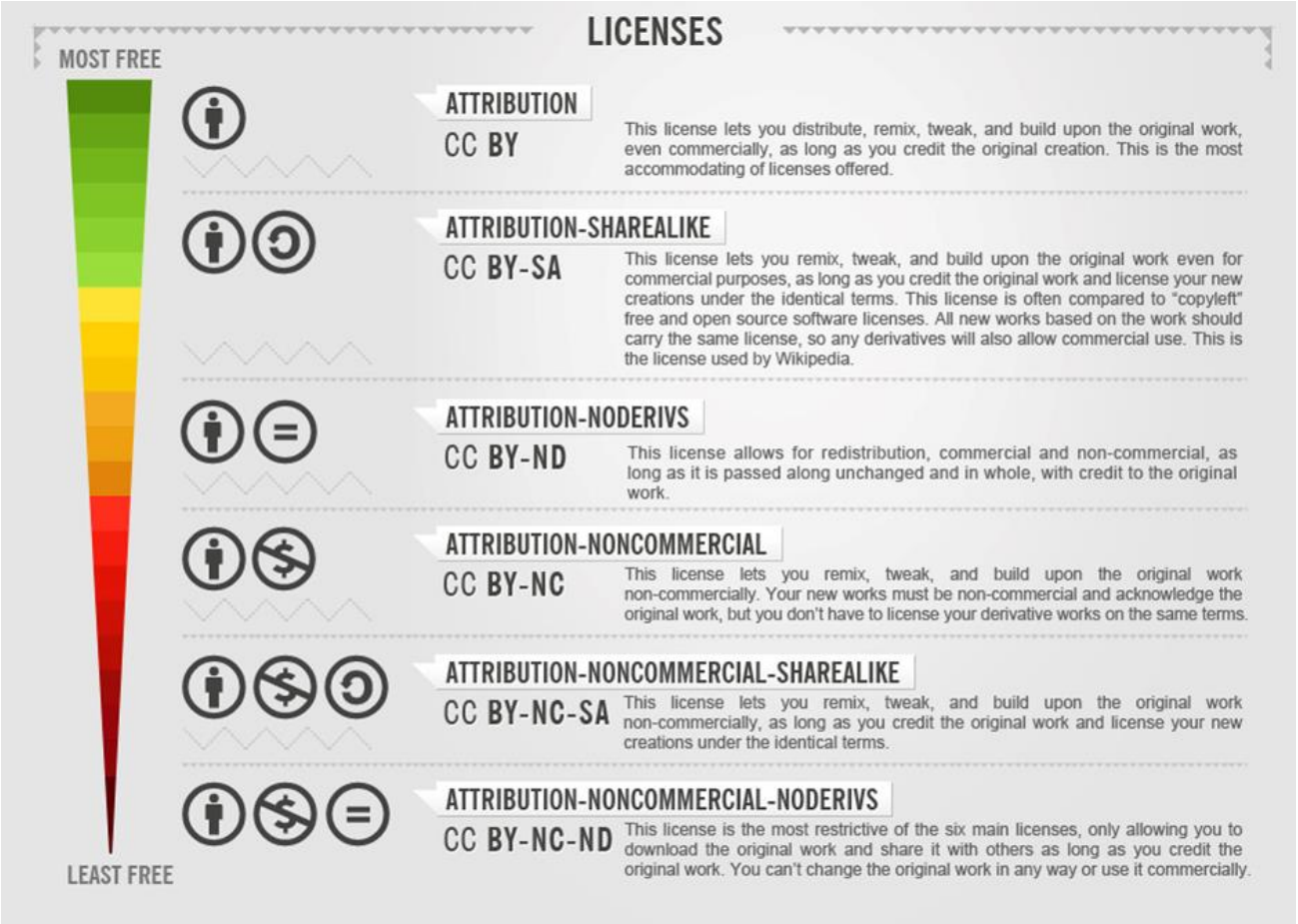
Rights to **reuse**, to **modification**, to **commercial** use, **obligation** to mention the **attribution** and to **share alike**

Sometimes, choosing a license can be restricted:

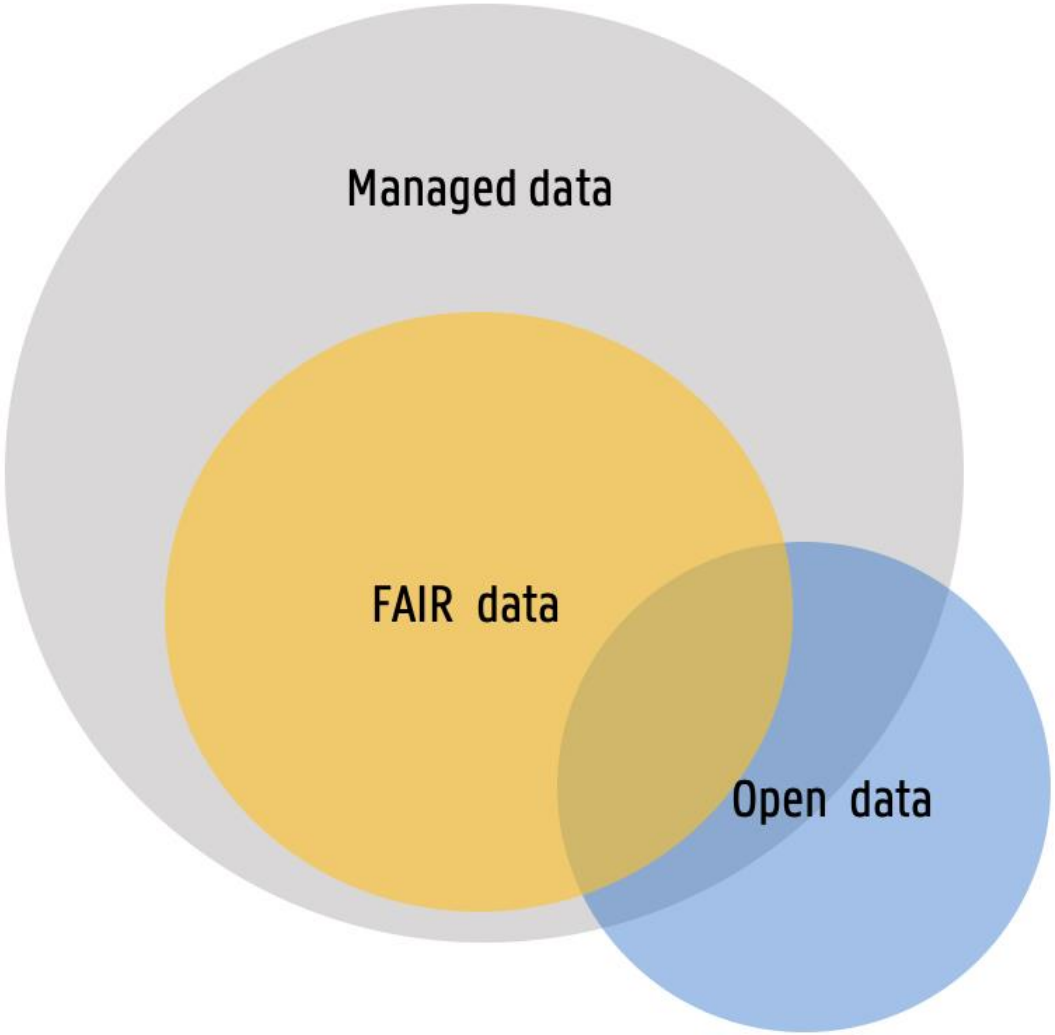
- The license may come with the use of a **repository**
- The license may come with the publication through an **editor** (journal)

Who can **help** me?

- Interface ULiège
- Libraries



FAIR data sharing

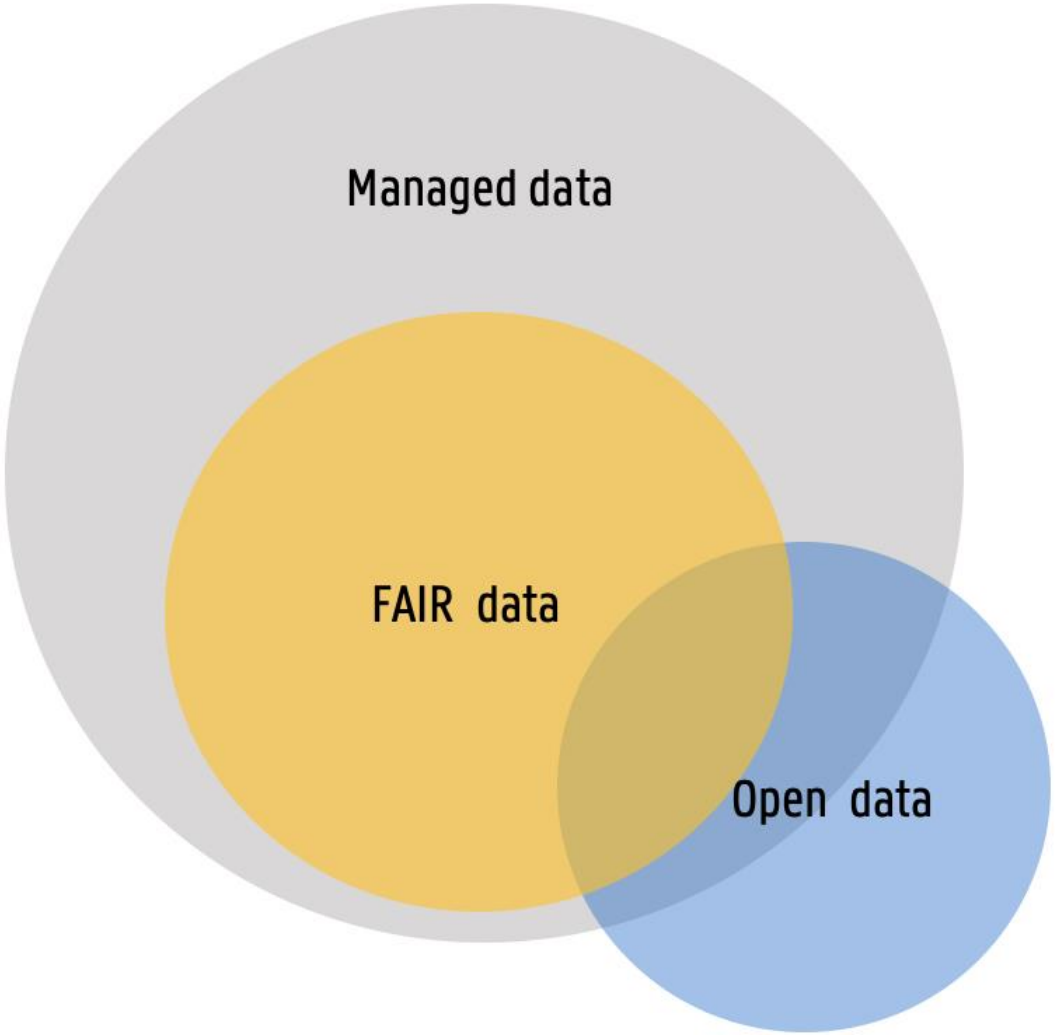


To sum up :
FAIR data is a **bridge** between individual good RDM habits and open data.

Making data FAIR is not only about the sharing step of the project, it starts at data creation:

- Storage
- Documentation
- Protection
- Traceability
- ...

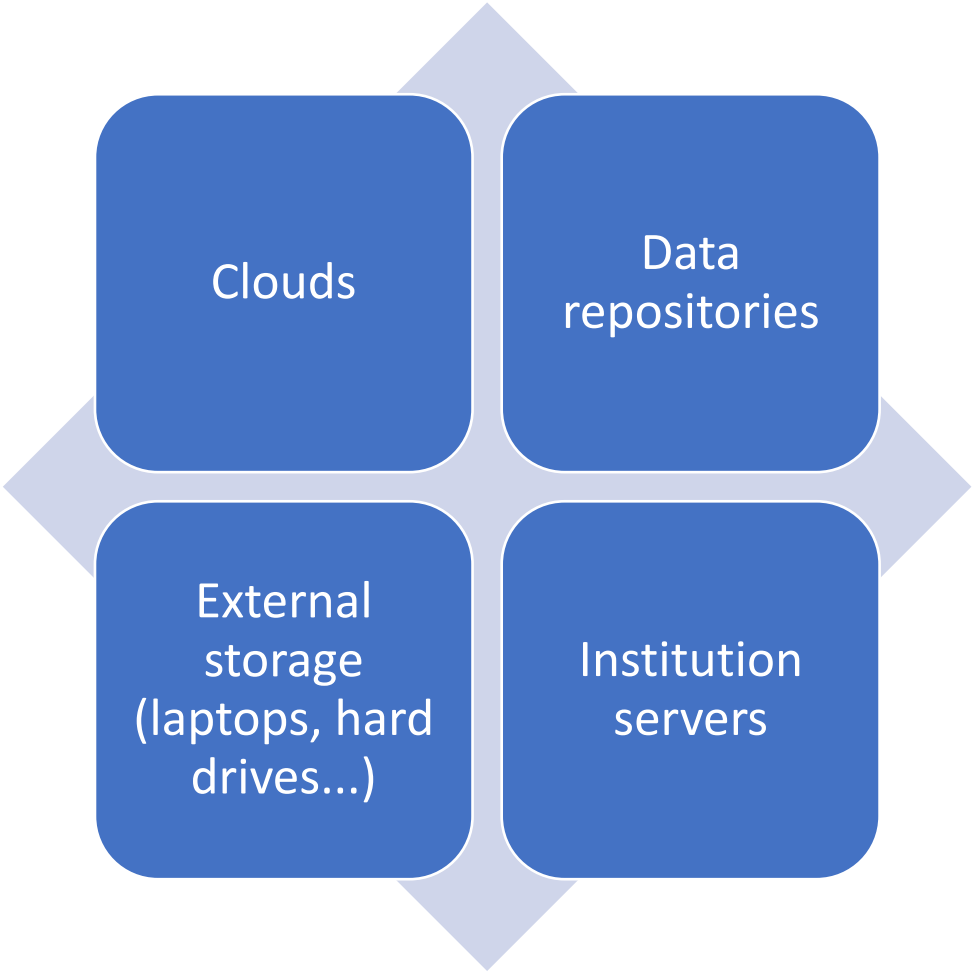
FAIR data sharing



What **practical habits** can you take up early on to facilitate FAIR data sharing at the end of your PhD?

Data storage

There are four main families of storage solutions:



Data storage

How do I choose?

Documentation

Organisation

Security

Sustainability

Data storage

How do I choose?

Documentation

- How / Why / By whom was the data created ?
- The difference b/w data dredging and reproducibility is telling what you did
- Keep track as much as possible between raw data and results, even **inconclusive**
- **Never erase anything**

Documentation makes it possible for an independent user to re-use the data, it makes it meaningful and supports **reproducibility**. It provides the context of the data acquisition, its (pre-)processing, its history.
The idea “what would a future user of the data need to know ?”

It is diverse : notes, codebooks, …

It can be as simple as describing what each column of an Excel file contains

But there is one **standardized** way of documenting data and that is **metadata**
Metadata is machine and human readable
There are standards applicable to scientific communities and types of data

RDA standards catalog : https://rdamsc.bath.ac.uk/
Examples such as DDI for human and med sciences, FITS for astrophysics, ISO 19115 for geography, …
The standards AND the metadata generation **usually comes with the chosen directory** as an XML file (or other computing language) and looks something like this
Controlled vocabulary, i.e. agreeing on terms, spelling, formats, … help visibility (e.g. keywords such as “galaxies” instead of “galaxy”, “Galaxy”, “Galactic objects”, …)

Note: data papers

Data storage

How do I choose?

Documentation

- How / Why / By whom was the data created ?
- The difference b/w data dredging and reproducibility is telling what you did
- Keep track as much as possible between raw data and results, even **inconclusive**
- **Never erase anything**

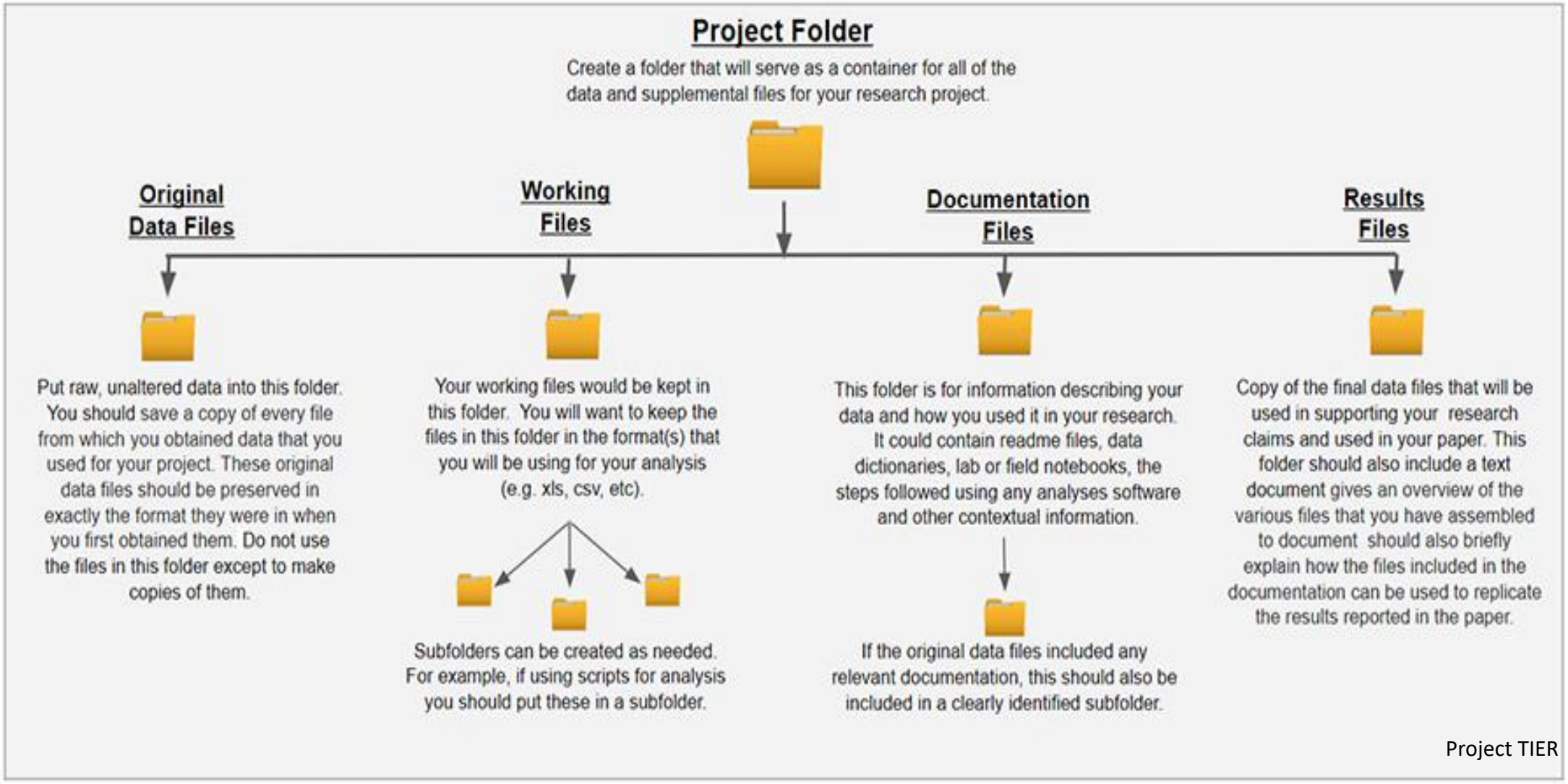
Organisation

- Knowing that, how much volume do I need?
- **Tree structure?**
- Explicit filenames
- Important for traceability, for yourself as much as for the next user

Security

Sustainability

Data storage



Project TIER

Data storage

How do I choose?

Documentation

- How / Why / By whom was the data created ?
- The difference b/w data dredging and reproducibility is telling what you did
- Keep track as much as possible between raw data and results, even **inconclusive**
- **Never erase anything**

Organisation

- Knowing that, how much volume do I need?
- **Tree structure?**
- Explicit filenames
- Important for traceability, for yourself as much as for the next user

Security

- Some risks associated with storage : hardware failure, theft, unauthorized access, natural disasters
- This can be consequential to your research further research, or your subjects
- **Backup ? 3** copies, **2** different storage solutions, **1** off-site
- Keep your data as in-house as possible
- Encryption or passwords are always a good idea

Sustainability

Content : D. Green - DCC

Data storage

How do I choose?

Documentation

- How / Why / By whom was the data created ?
- The difference b/w data dredging and reproducibility is telling what you did
- Keep track as much as possible between raw data and results, even **inconclusive**
- **Never erase anything**

Organisation

- Knowing that, how much volume do I need?
- **Tree structure?**
- Explicit filenames
- Important for traceability, for yourself as much as for the next user

Security

- Risks & consequences
- **Backup ? 3** copies, **2** different storage solutions, **1** off-site
- Keep your data as in-house as possible
- Encryption or passwords are always a good idea

Sustainability

- Could someone reuse this easily in ten years?
  - Availability, location
  - Documentation, metadata
  - Format
- A guide to decide what to archive

Rule of thumb: anything that underpins an article must be kept unless regulations force you to erase

Data storage

How do I select a data repository?

General	Discipline-specific
Zenodo OSF Figshare Dataverse Institutional repositories (« data ORBi »)	Some examples : The QDR (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

Data storage

How do I select a data repository?

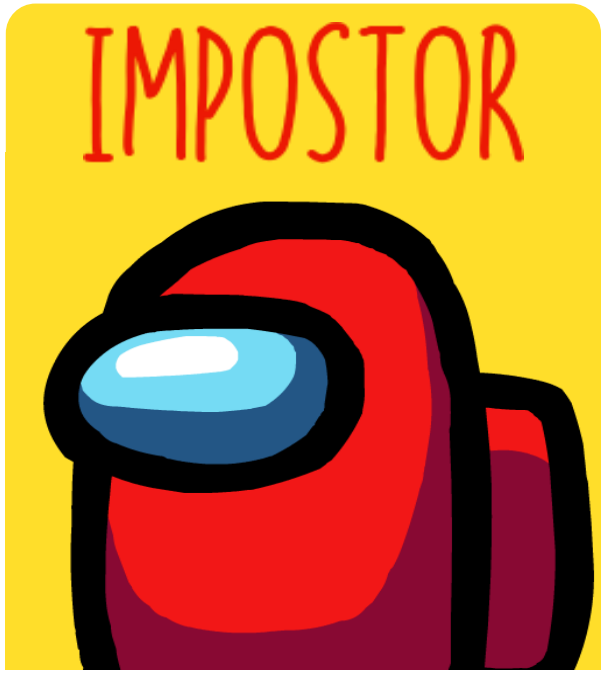
General	Discipline-specific
Zenodo OSF Figshare Dataverse Institutional repositories (« data ORBi »)	Some examples : The QDR (HSS), CDS (astro), NCBI (genomics), .. Catalogs of directories : Re3data, FAIRsharing Ask your peers and supervisor

A good repository:

- Is recognized by your peers
- Provides a persistent [illegible]fier such as a DOI
- Comes [illegible]
- [illegible] is
- Le[illegible] all your rights
- Has [illegible]certification such as CoreTrustSeal

If external repository, keep a sustainable local copy

Data management and ethics



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

https://retractionwatch.com/

Research lives in a paradoxical context that may push us, even unconsciously, towards questionable practice

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

The academic journal

The online news story

- **Pressure to publish** with tenure and funding on the line
- Pressure to find results that seem **new and striking**
- Numerous ways to **tweak** your study, **consciously**

or not, until you get a result that counts as **statistically significant** even though it is probably meaningless:

→ Altering how long it lasts
→ Play with the sample size
→ P-hacking (collecting lots of variables and playing with data until finding counts as statistically significant)

- As a result: many studies that get media coverage seem to contradict each other, impeding the **trust** of society in (good) science

Rule of thumb : it is okay to play around with your data, but the difference b/w data dredging and exploring a dataset is telling about it in your publications

What is data dredging, why does it happen, and what are its consequences?

Data management and ethics

P-hacking

Chopping up, testing, arranging, filtering, tweaking and/or tuning your dataset to obtain a **statistically significant result**

Even if it is random

Data management and ethics

I am testing **a hypothesis H**

ex: these diet pills do work

ex: this dice is loaded

I collect **relevant data**

ex: weight of a group of people before and after taking diet pills for a month

ex : number of times each face comes up after 50 dice rolls

Data management and ethics

I am testing **a hypothesis H**

ex: these diet pills do work

ex: this dice is loaded

I collect **relevant data**

ex: weight of a group of people before and after taking diet pills for a month

ex : number of times each face comes up after 50 dice rolls

I compute the **probability to obtain this same data even if my hypothesis H is wrong**

ex: if these pills do not work, what is the probability that these people would have lost weight anyway?

ex: if the dice is not loaded, what is the probability that face 6 only comes up 5 times out of 50?

Data management and ethics

I am testing **a hypothesis H** ex: these diet pills do work ex: this dice is loaded	I collect **relevant data** ex: weight of a group of people before and after taking diet pills for a month ex : number of times each face comes up after 50 dice rolls	I compute the **probability to obtain this same data even if my hypothesis H is wrong** ex: if these pills do not work, what is the probability that these people would have lost weight anyway? ex: if the dice is not loaded, what is the probability that face 6 only comes up 5 times out of 50?

The data drives the conclusion, not the opposite

Playing around with the p-value is fine, but **boiling down a complex scientific result to only one p-value** is not.

A small p-value is a **good indicator** that your hypothesis is correct, but is not enough:

→ **It does not prove H is true** (it only proves the opposite of H is improbable given this particular dataset)

→ It **does not prove** that the dataset is suitable for the test, or that the model is suitable for the hypothesis.

→ It **does not prove** the quality of the dataset (completeness, sample size, accuracy, ...)

Data management and ethics

No panic:

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

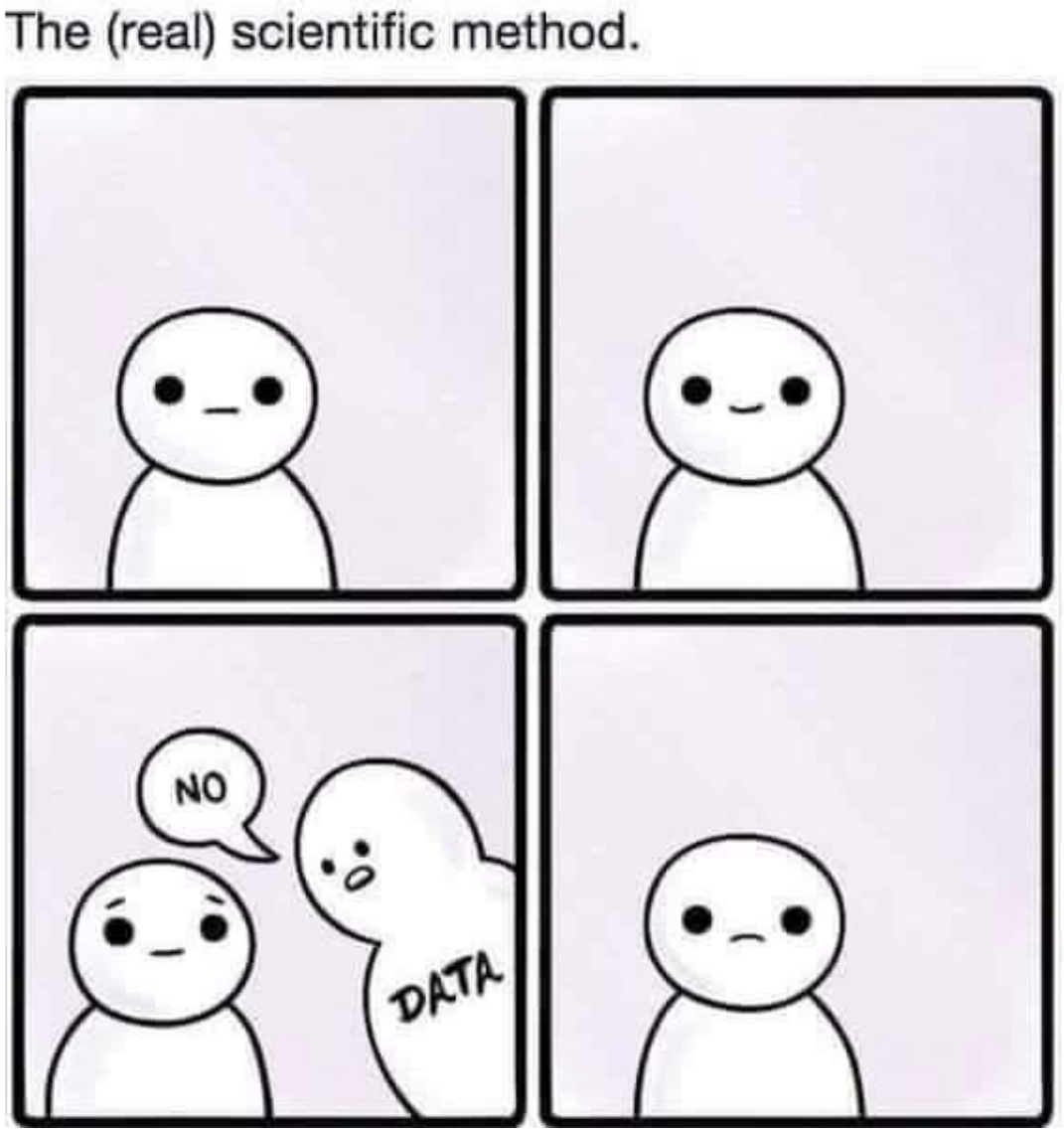
Making raw data, protocols, methodologies...

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

Your doctoral school, your supervisor, your lab, your stats teacher...

Training sessions in catalog:

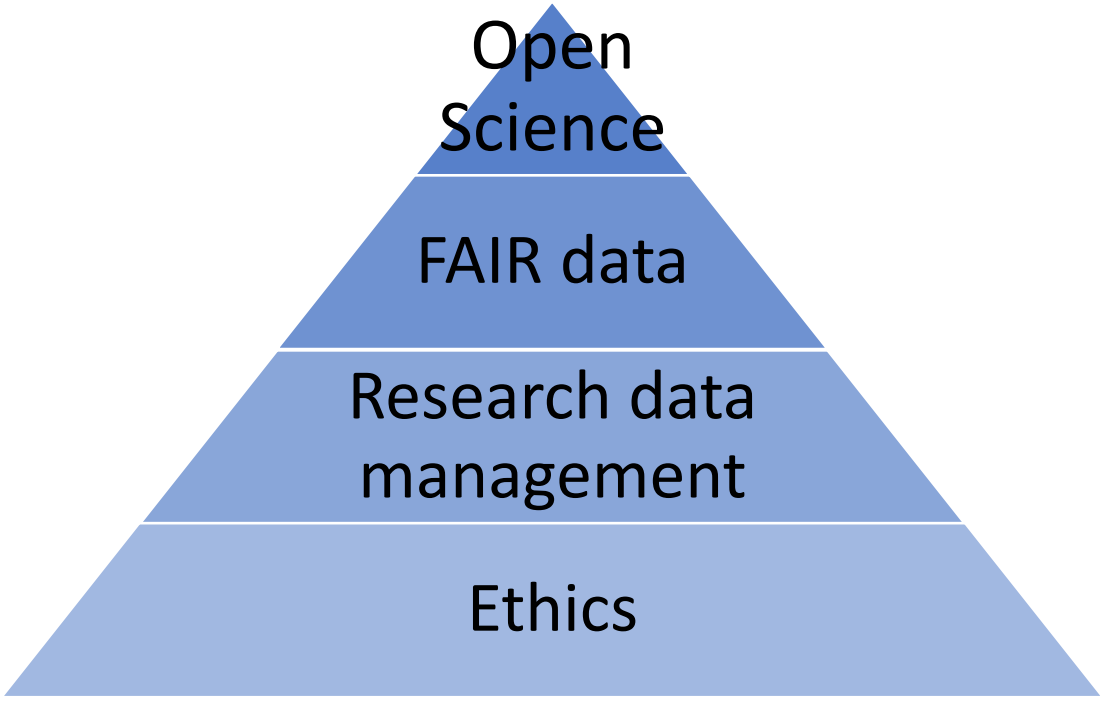
- Probabilités et statistiques de base (A2-7)
- Statistique multivariée (A2-8)



The bigger picture

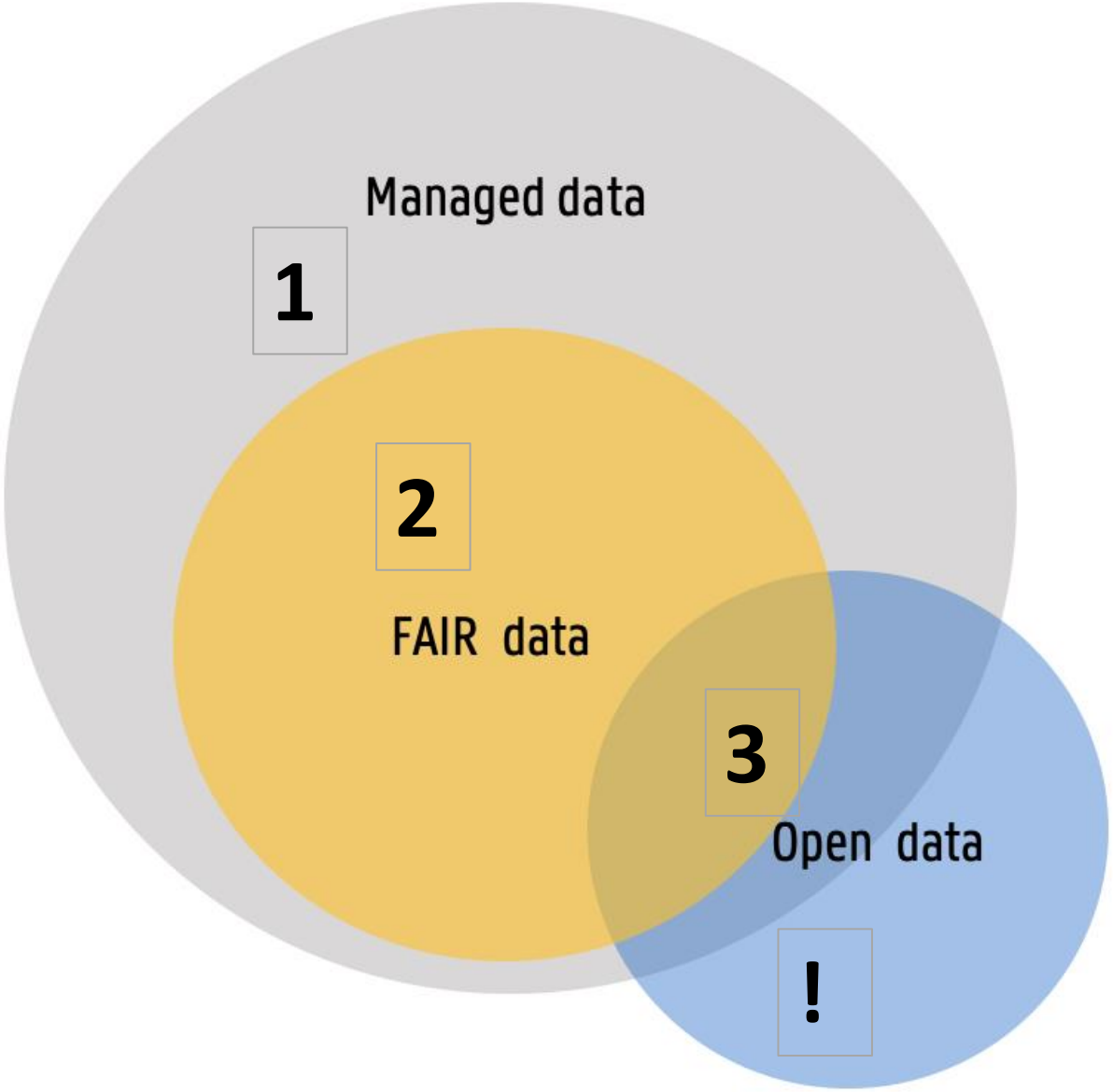
RDM is first and foremost a set of **best practices** that support scientific **reproducibility**

- Keeping track of every step of your data analysis and being transparent about it prevents any suspicion of **data dredging**
- Good habits in **data storage** enables this traceability and accountability
- It also ensures compliance **to rules and regulations**

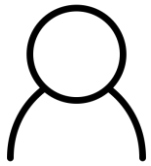


Good RDM habits facilitate **FAIR data sharing**, with **open science** being the cherry on top of the cake

The bigger picture



Find help and resources



Talk to your supervisor
Research office:
Judith Biernaux jbiernaux@uliege.be + Jérome Eeckhout : jeeckhout@uliege.be
GDPR: Pierre-François Pirlet pfpirlet@uliege.be
Ethics & scientific integrity: https://www.recherche.uliege.be/cms/c_9022717/en/ethics-and-scientific-integrity
ULiège Ethics Board: ceis@uliege.be
Legal Affairs: +32 (0)4 366 52 38 – Via Pages Web
Dual use: https://www.recherche.uliege.be/cms/c_11374378/fr/dual-use
Interface: Pages Web (patenting, licences, ...)

Data Management Plan (DMP) -> http://dmponline.be + online tutorial
Checklist Grey et al 2020
6 directories for sharing your data
Data Storage and Organization tips from Macalester College MN
Suggestions**: Doc Fetcher**, **Obsidian**, **Jupyter** or **Gitlab**

Charte Européenne du Chercheur
Ethics in research and international cooperation (EU)
Ethical aspects of new ICT technologies (EU)
Guidelines on Enhancing the QUAlity and Transparency Of health Research (EQUATOR)

Sources

[*]
J. Ioannidis, 2005, Contradicted and Initially Stronger Effects in Highly Cited Clinical Research, JAMA. 2005;294(2):218-228. doi:10.1001/jama.294.2.218

Mark Otto Baerlocher et al., 2010, Data integrity, reliability and fraud in medical research, Elsevier European Journal of Internal Medicine 21 (2010) 40–45

Monya Baker, 1,500 scientists lift the lid on reproducibility, Nature 533, 452–454 (26 May 2016) doi:10.1038/533452a

Sources

“Three Camps, One Destination: The Intersections of Research Data Management, FAIR and Open”.
Higman, Rosie, Daniel Bangert, and Sarah Jones. 2019. *Insights* 32 (1): 18.
DOI: http://doi.org/10.1629/uksg.468

Formations courtes en ligne sur la gestion responsable des données, Macalester College, Minnesota, consulté le 18/09/20
https://libguides.macalester.edu/c.php?g=527786&p=3608657

What Is Data Quality and Why Is It Important? Aaron Moss, PhD, consulté le 18/09/20
https://www.cloudresearch.com/resources/guides/ultimate-guide-to-survey-data-quality/guide-data-quality-what-is-data-quality-why-important/

Research data management explained, University of Leeds, consulté le 13/09/20
https://library.leeds.ac.uk/info/14062/research_data_management/61/research_data_management_explained

Fostering the practical implementation of Open Science in Horizon 2020 and beyond, consulté le 14/08/20
https://www.fosteropenscience.eu/node/1420

Ice Cream Sales Lead to Higher Homicide Rates: How Correlation Doesn’t Always Equal Causation, consulté le 18/09/20
https://www.egenerationmarketing.com/blog/causation-and-correlation-for-a-law-firm

How researchers dupe the public with a sneaky practice called "outcome switching“, consulté le 17/09/20
https://www.vox.com/2015/12/29/10654056/ben-goldacre-compare-trials

Quick Data Lessons: Data Dredging, consulté le 03/09/20
https://www.geckoboard.com/blog/quick-data-lessons-data-dredging/

Sources

You Can’t Trust What You Read About Nutrition , consulté le 16/09/20
https://fivethirtyeight.com/features/you-cant-trust-what-you-read-about-nutrition/

Science Isn’t Broken, consulté le 16/09/20
https://fivethirtyeight.com/features/science-isnt-broken/#part1

Page « PACE trial » on me-pedia, consulté le 16/09/20
https://me-pedia.org/wiki/PACE_trial

For my next trick... Consulté le 17/09/20
https://www.economist.com/science-and-technology/2016/03/26/for-my-next-trick

Simmons, Nelson & Simonsohn, 2011
https://journals.sagepub.com/doi/pdf/10.1177/0956797611417632

American Statistical Association (ASA) Statement on Statistical Significance and P-Values, 2010
https://amstat.tandfonline.com/doi/full/10.1080/00031305.2016.1154108

Baerlocher et al., 2010
https://www.sciencedirect.com/science/article/pii/S0953620509002337

Sources

Manon Knockart et Thomas Tombal, « Quels droits sur les données », in *Actualités en droit du numérique*, Limal, Anthémis, 2019, p. 53 et suiv.

Thierry Léonard, Bojana Salovic, Olivia Guerguinov, « Protection des données : quel cadre juridique pour la recherche scientifique en Belgique ? », blog Droit et Technologies, 1er avril 2019 https://www.droit-technologie.org/wp-content/uploads/2019/04/v2.pdf (consulté le 24 février 2021)

Lionel Maurel, « A qui appartiennent les données de la recherche ? », Webinaire Tuto@Mate organisé par le Réseau Méthodes Analyses Terrains Enquêtes en SHS le14 septembre 2020 https://mate-shs.cnrs.fr/wp-content/uploads/2020/09/tuto25-mate-Données-de-recherche.pdf (consulté le 24 février 2021)

Anne-Laure Stérin, Camille Noûs, « Ouverture des données de la recherche : les mutations juridiques récentes », *Tracés. Revue de Sciences humaines* [En ligne], #19 | 2019, mis en ligne le 22 juillet 2020 http://journals.openedition.org/traces/10603 (consulté le 24 février 2021)

Questions juridiques liées aux données de recherche, interview de Lionel Maurel réalisée à l'occasion de la séquence de com' : La licence ouverte, à l'Inist-CNRS (Nancy) le 02 juillet 2019 https://doranum.fr/aspects-juridiques-ethiques/questions-juridiques-liees-aux-donnees-de-la-recherche/ (consulté le 24 février 2021)

Sources

Herbert Gruttemeier, Thérèse Hameau, « Accès aux données scientifiques et contraintes juridiques – une question d'équilibre », *I2D - Information, données & documents*, 2016/2 (Volume 53), p. 20-22 https://www.cairn.info/revue-i2d-information-donnees-et-documents-2016-2-page-20.htm (consulté le 24 février 2021)

Lionel Maurel, Données de la recherche et questions juridiques au sein des plans de gestion de données https://www.hisoma.mom.fr/sites/hisoma.mom.fr/files/docs/Recherche/quinquenal-2016-2020/axe-t/emmanuelle-morlock/seminaire_pgd_juridique_maurel.pdf (consulté le 24 février 2021)

Extra : Why is it so difficult?

https://www.youtube.com/watch?v=FpCrY7x5nEE