Research Data Management and Reproducibility

Good habits for good research

Introduction to scientific computing

Oct 18, 2021 - (GIGA)

LIÈGE université

Judith Biernaux, Dr.
Administration R&D
Place du XX Août, 7 (Bât. A1)
B-4000 Liège
Tél : +32 (0)4 366 55 14
Jbiernaux@uliege.be

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**

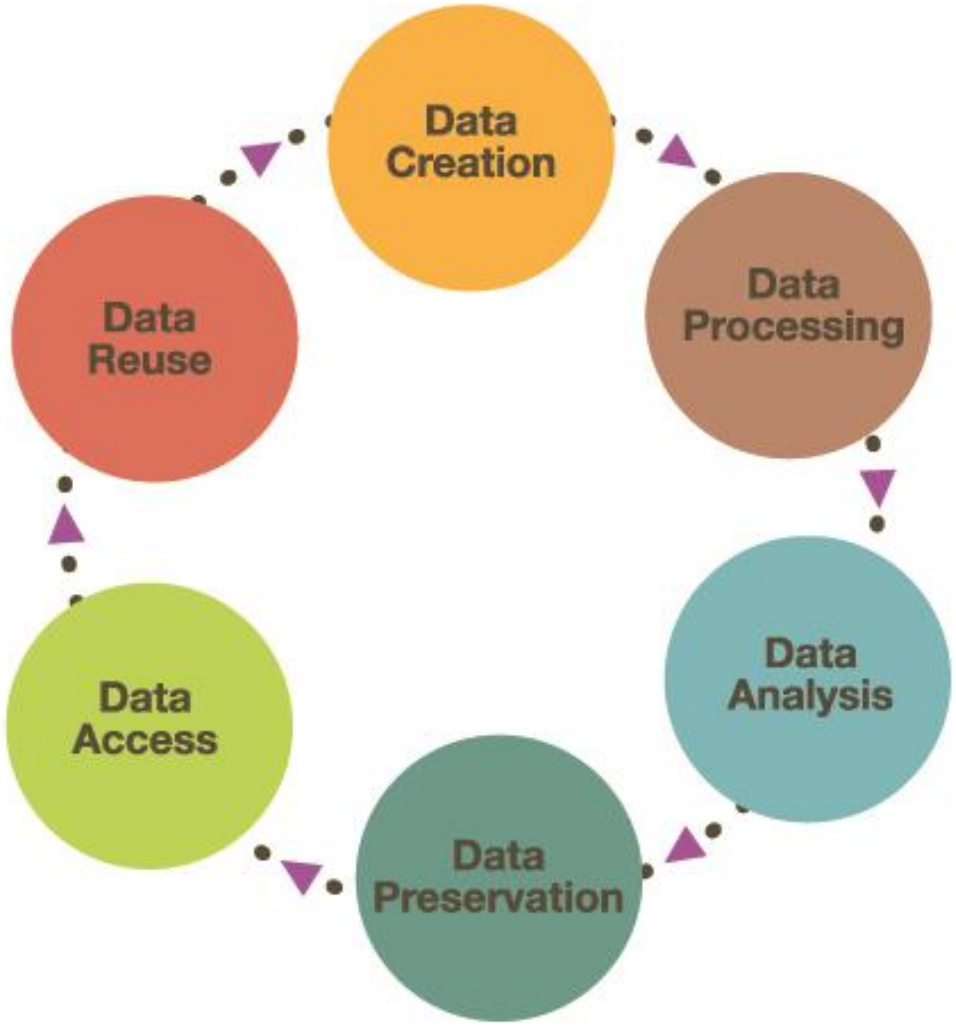
-> They may have an impact on the **quality** of your results

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

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

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

What is the use of research data?



What usually happens to data:

They are **created**
- in a lab, through fieldwork, measurement, on a computer, ...

They are **processed**
- cleaned up, sampled, converted, ...

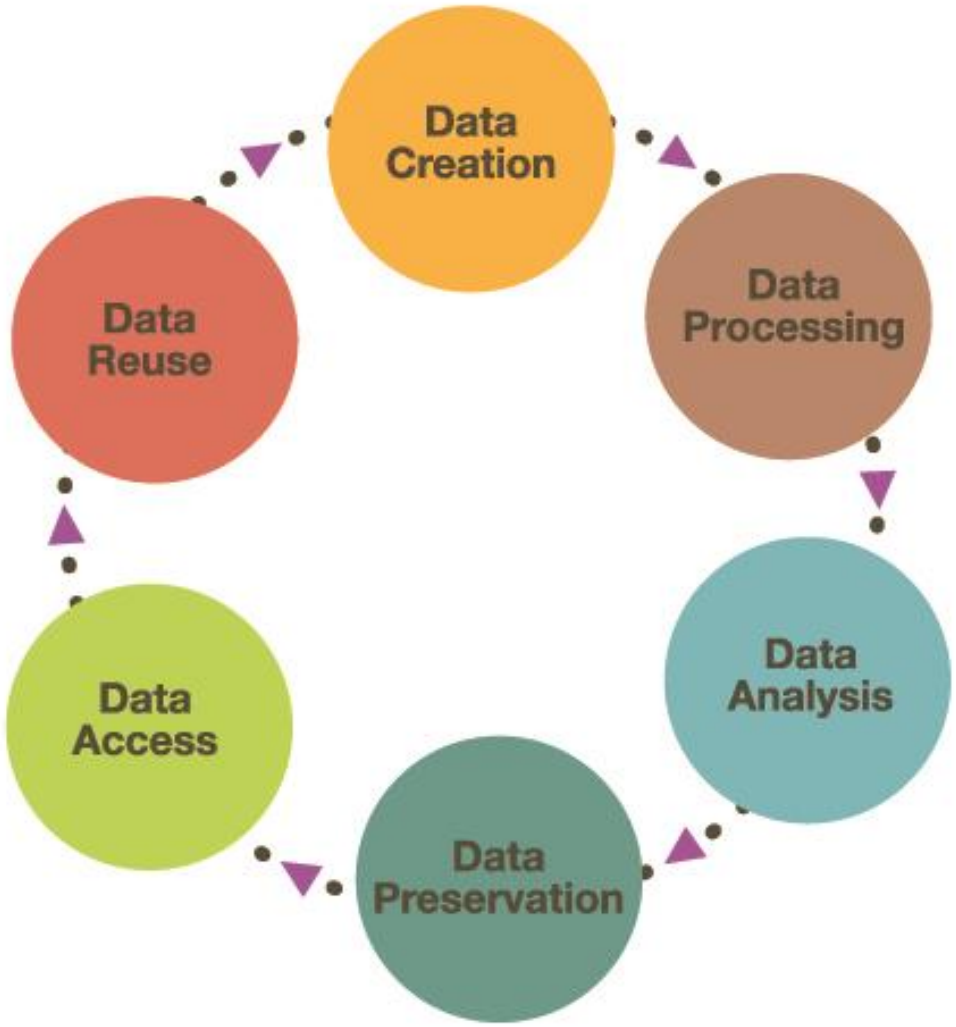
They are **analysed**
- statistics, fitting, study, comparison, interpretation, ...

They are **stored**
- for long-term preservation

They are **shared**
- as open as possible, as closed as necessary

Someone else **re-uses** them

What can I do?



Before each stage, **step back and ask :**

What **sort** of data do I need?

Nature, format, volume, source

Who do they **belong** to? Are they subject to applicable **regulations**?

Access? Ownership? Conditions?

What am I putting them through? Are they fit for my **methodology**?

Quality? Sample size?

How and where should I **store** them?

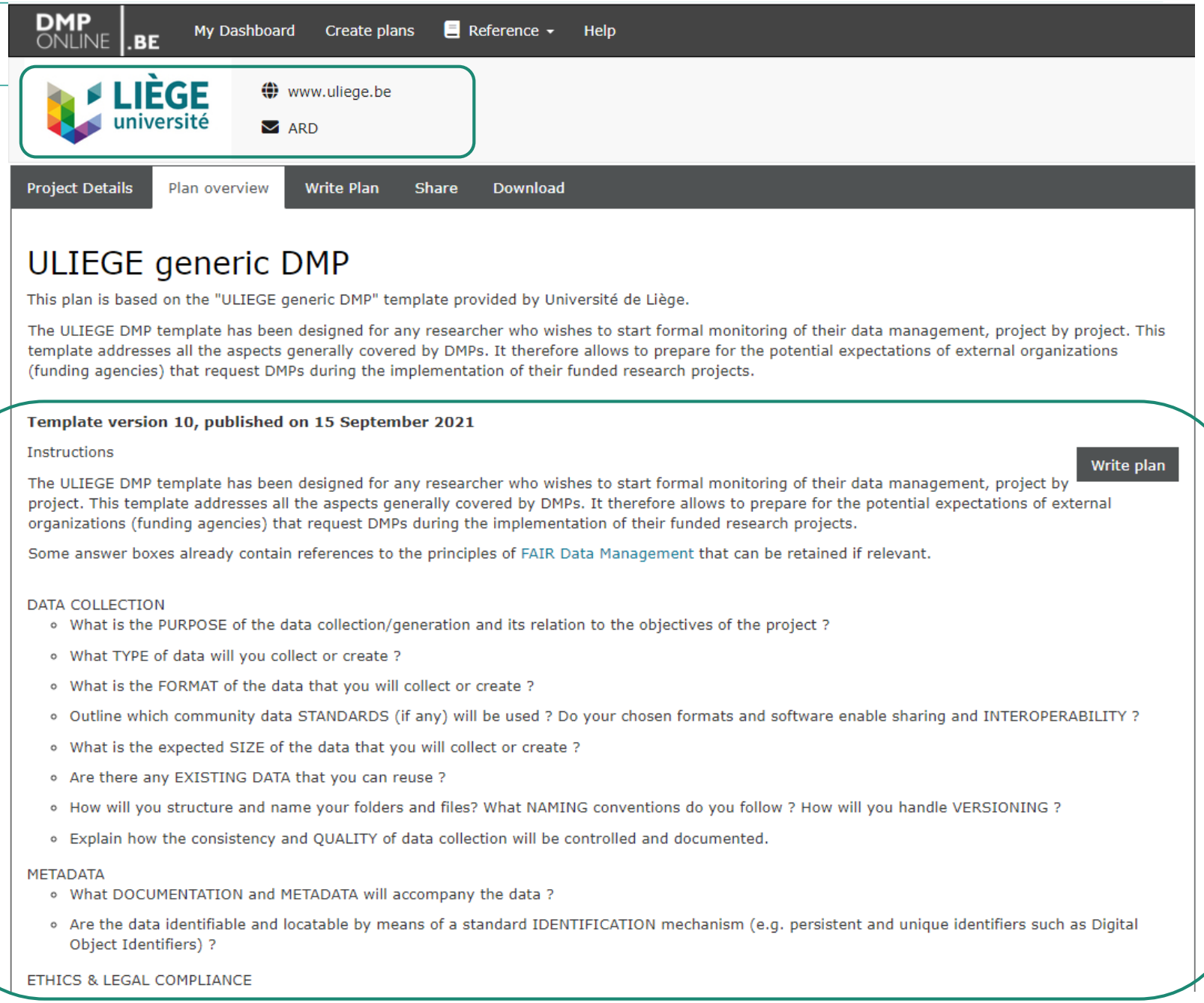
Volume? Safety and security? Long-term?

How should I **share** them?

Open Data, FAIR data, licenses

Data Management Plan

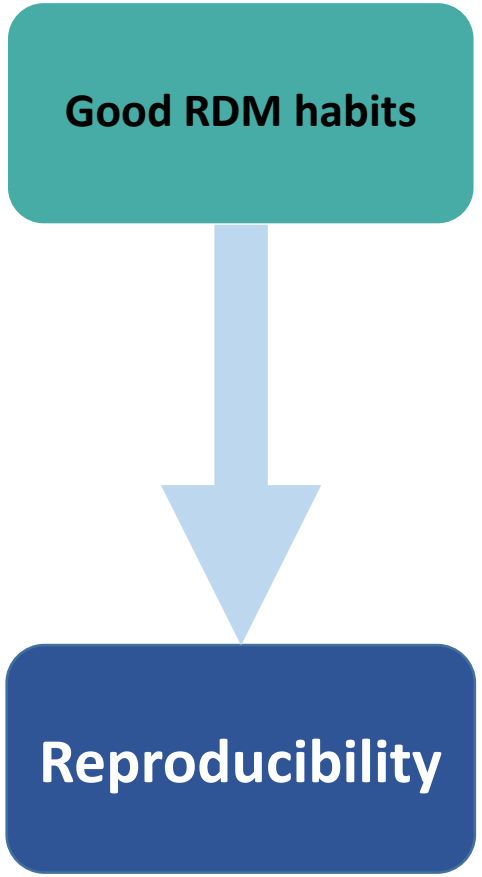
DMP training for PhD students



Why is it so important?

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**
- 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.**

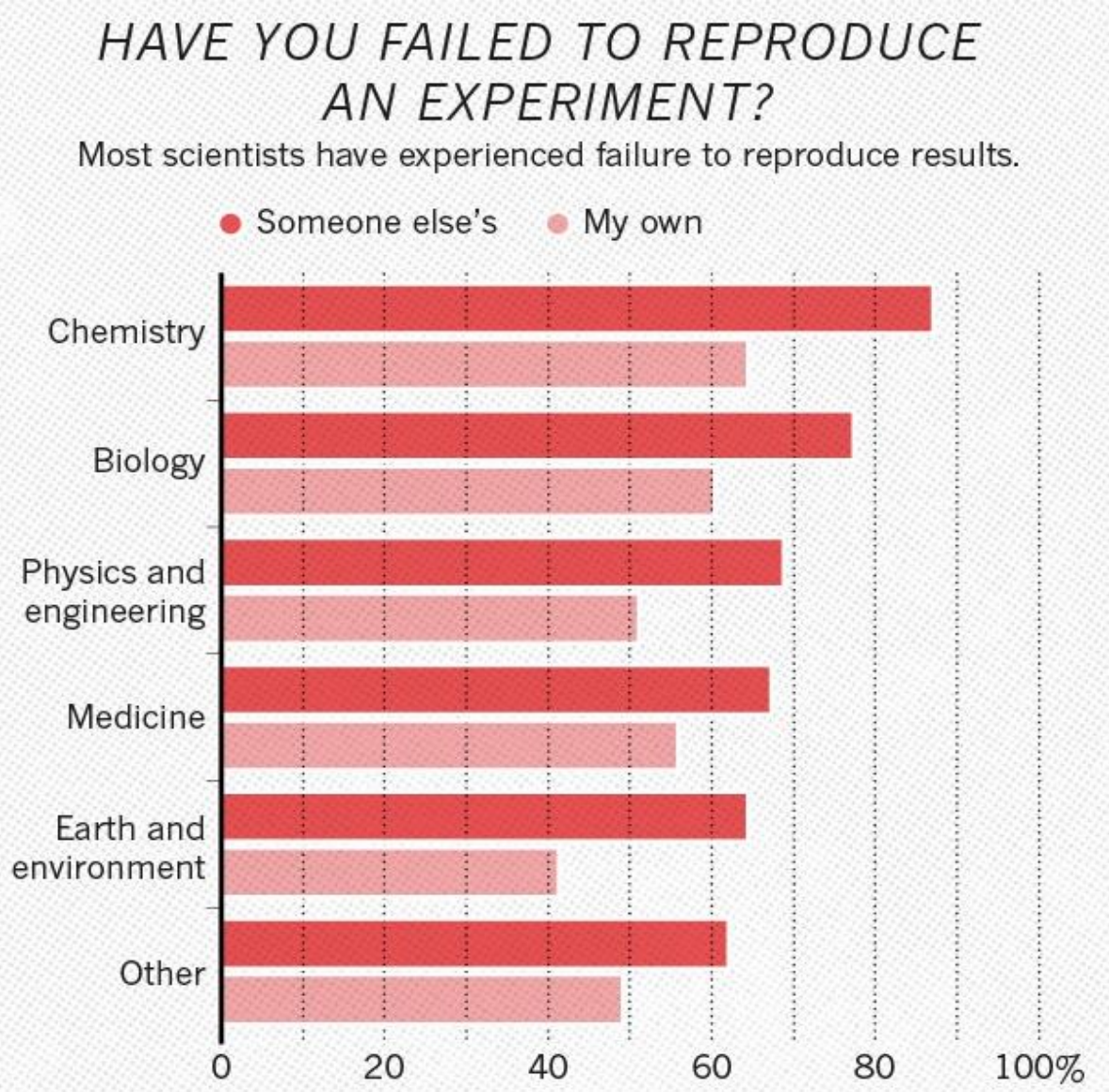
It is what makes your research **alive**
It makes it **useful** and **trustworthy**

Why is it so important?

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

Why is it so important?

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

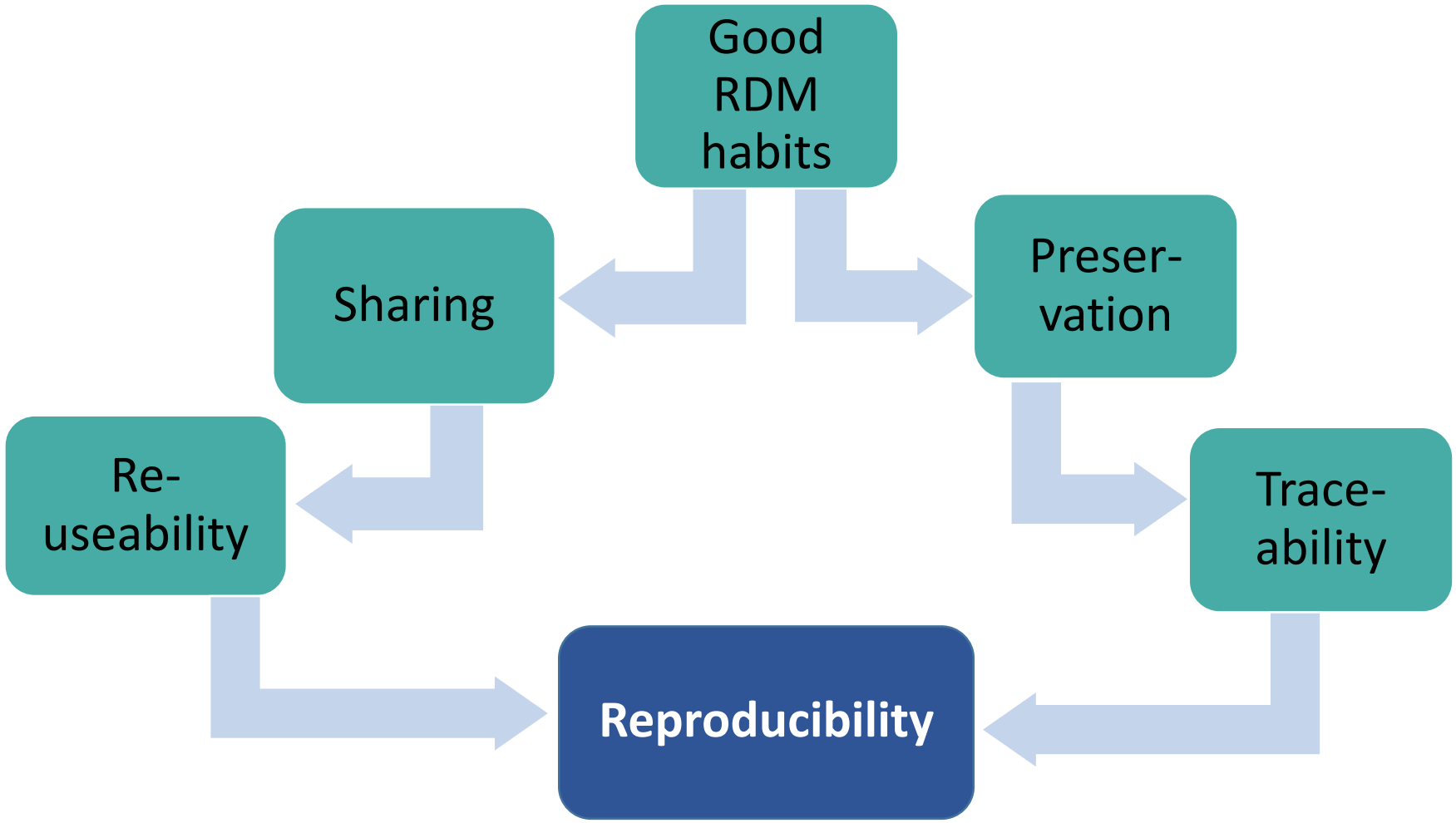
Why is it so important?

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?



Are they subject to specific applicable regulations?

A piece of data may be subject to different **regulations**

- to be used validly
- to be transmitted to third parties

Content : Catherine Thiry

Are they subject to specific applicable regulations?

Data Type / Sector	Applicable regulations / Risks
Personal data (be mindful of anonymization!)	
Works of art	
Data from other sources	
Genetic resources (plant, animal, human)	
Industrial or economical information from private companies	
Data that can be used for hostile purpose	

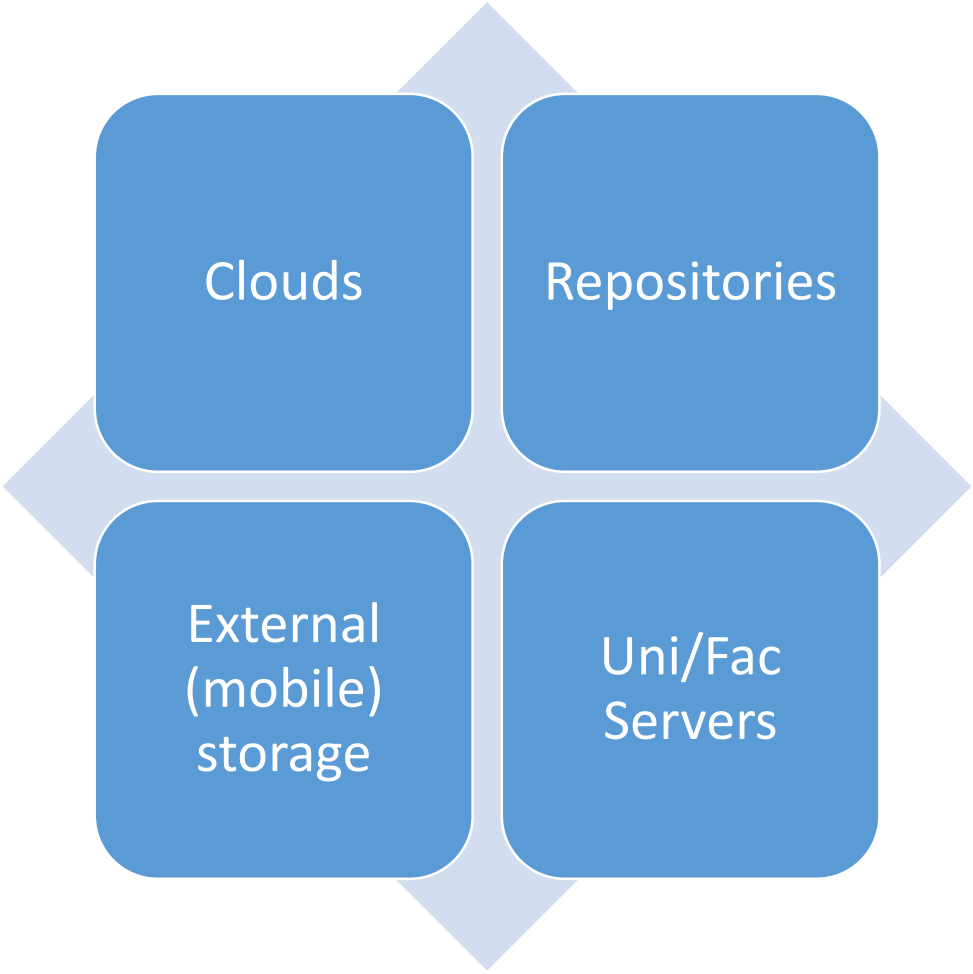
Content : Catherine Thiry

Are they subject to specific applicable regulations?

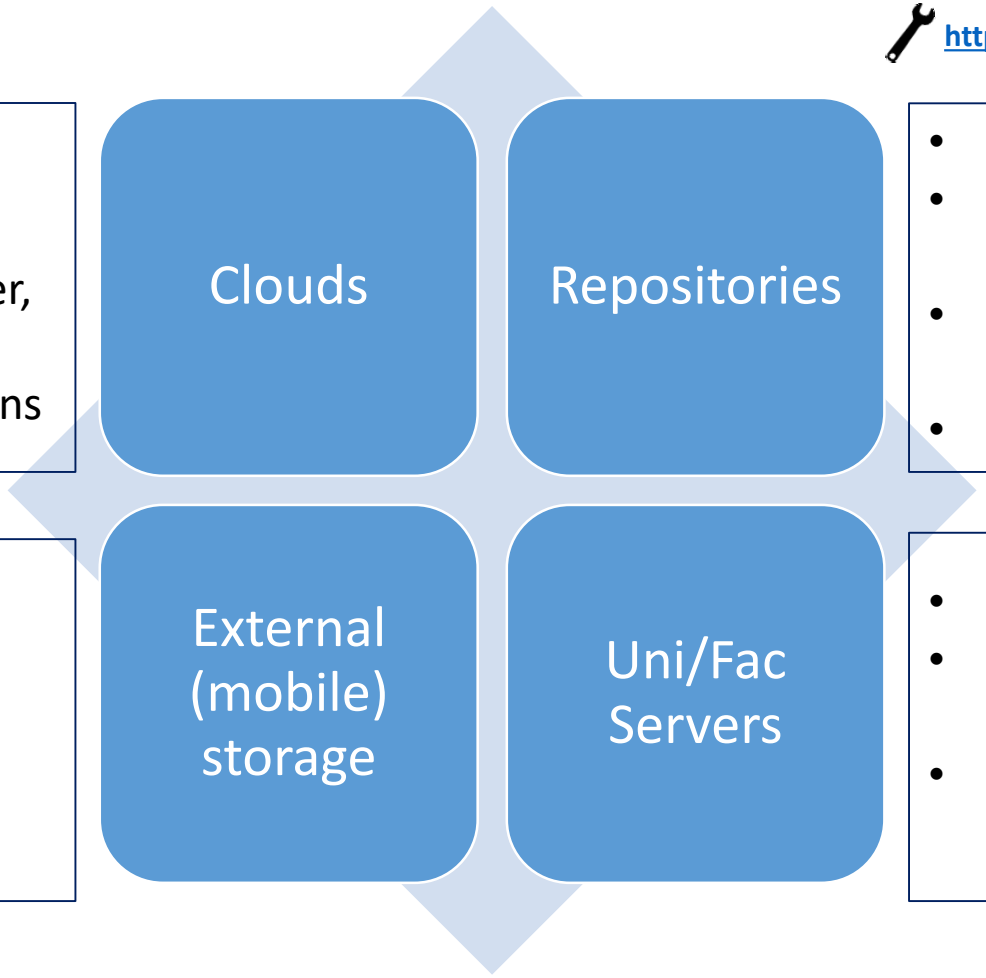
Data Type / Sector	Applicable regulations / Risks
Personal data (be mindful of anonymization!)	GDPR
Works of art	Copyright, intellectual property regulations, licenses
Data from other sources	
Genetic resources (plant, animal, human)	Nagoya protocol / GDPR if human
Industrial or economical information from private companies	Confidentiality agreement / NDA
Data that can be used for hostile purpose	Dual Use (protection license)

How should I store my data?

How should I store my data?



How should I store my data?



- Google Drive, DropBox, ... Be mindful of **outsourcing**
- DoX or OneDrive: internal so safer, **security** control
- Pay attention to volume limitations

https://www.teamscopeapp.com/blog/6-repositories-to-share-your-research-data

- Zenodo, Figshare, Dataverse, ...
- Excellent encoding structure with **metadata**
- Mostly made for **sharing**, not preserving
- Pay attention to volume limitations

- Practicality but **mobility risks**
- Careful **protection** needed (password or other)

- Internal so safer, **security** control
- **Remotely** accessible -> fewer mobility risks
- Easily shared **within** Uni but not with external users

How should I store my data?

Documentation

Organisation

Security

Preservation

How should I store my data?

https://libguides.macalester.edu/data1

Documentation

- **Never erase anything**
- Keep track of any data processing or analysis (even inconcluding)
- Using metadata or « readme » files
- Make sure you guarantee **traceability**

Organisation

- What volume do I need? GB? TB?
- **Tree view** (added volume?)
- Naming convention?

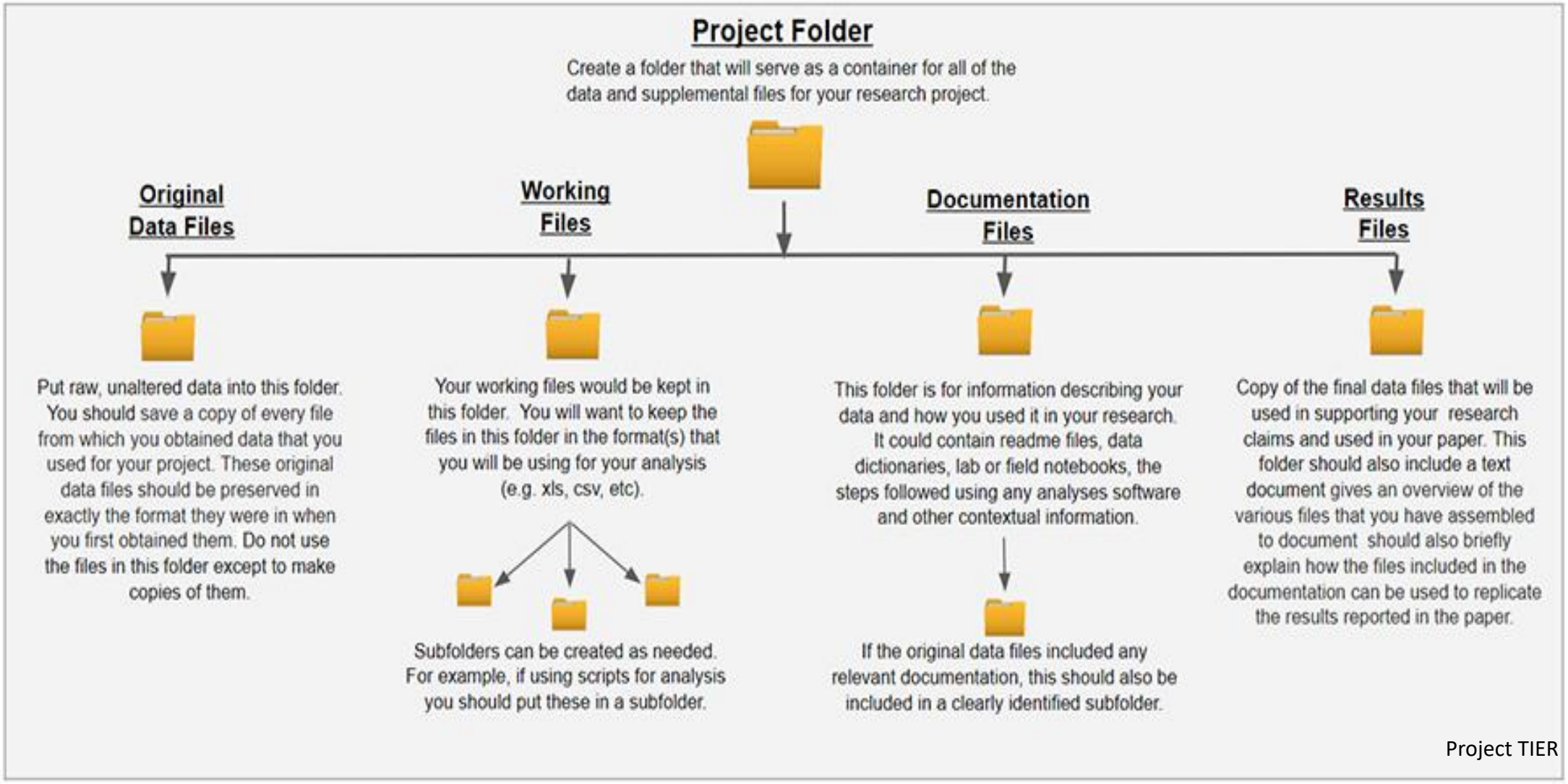
Security

- Level of confidentiality? (See applicable regulations)
- **Backup solutions?**
  3 copies
  2 different storage solutions
  1 off site (if possible)

Preservation

- Format (open vs proprietary)
- Long-term storage solution?
- What about after your thesis?

How should I store my data?



Project TIER

What does my data go through?

Research context sometimes pushes us towards **bad habits** or **questionable** practice

What does my data go through?

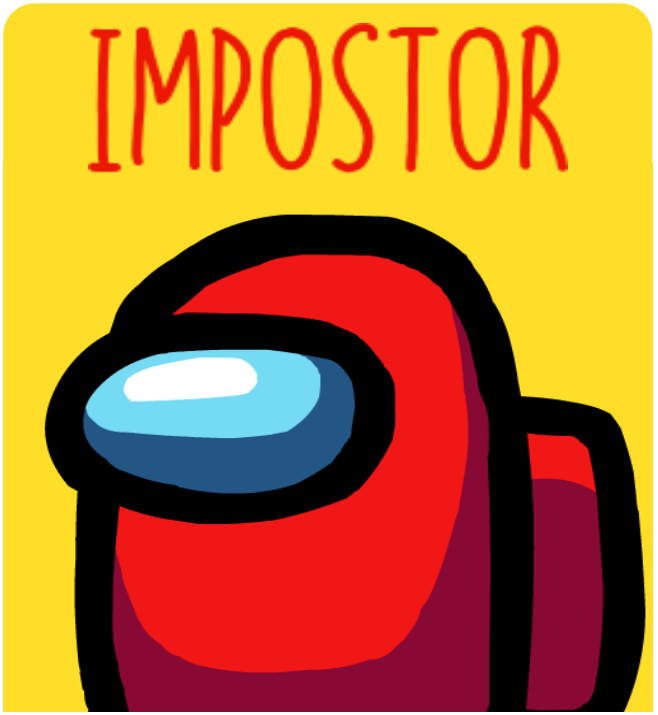
Fraud

Falsification, fabrication, plagiarism – zero tolerance

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/



What does my data go through?

What does my data go through?

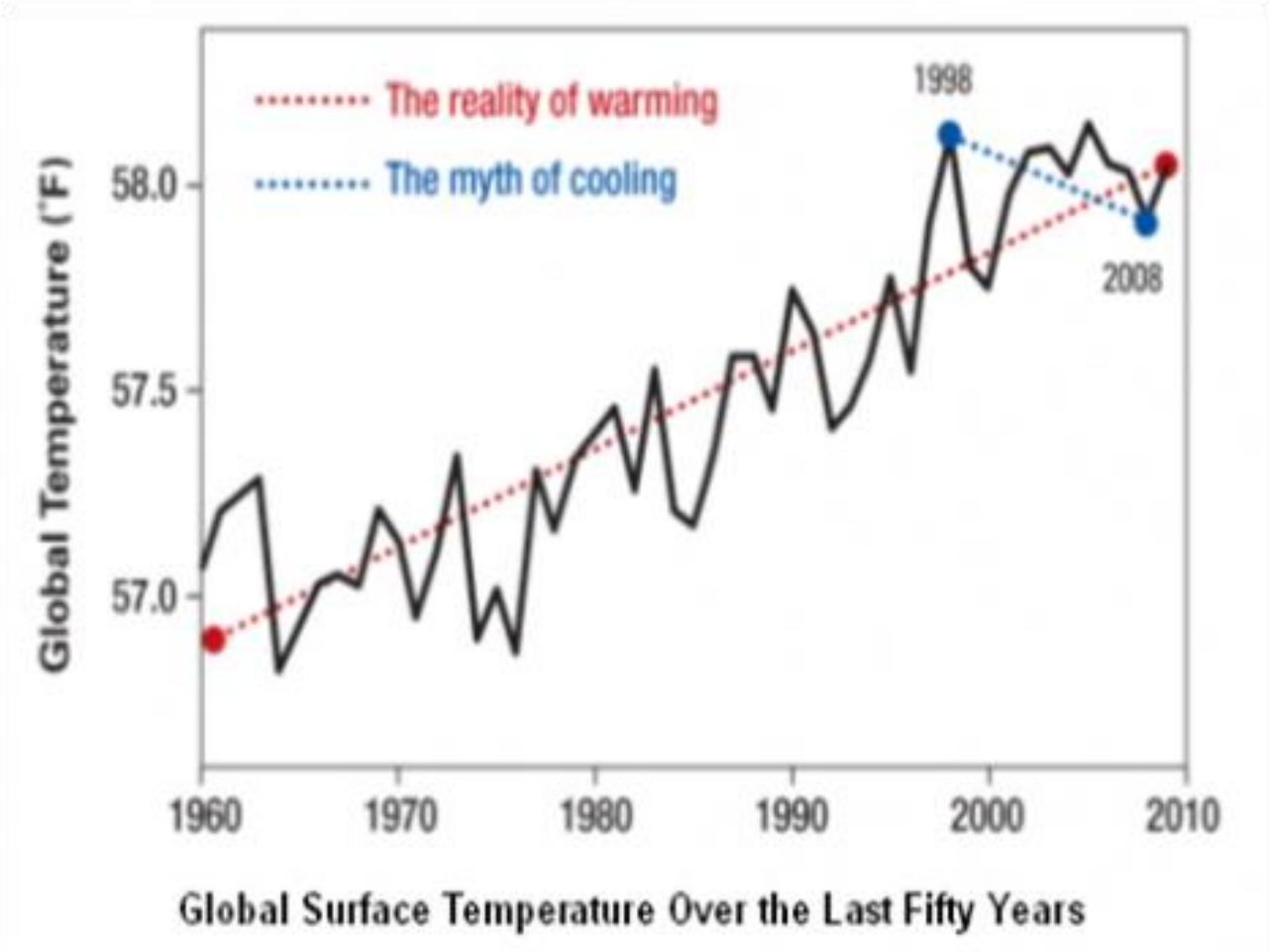
Shortcuts

- **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)

What does my data go through?

Cherry-picking

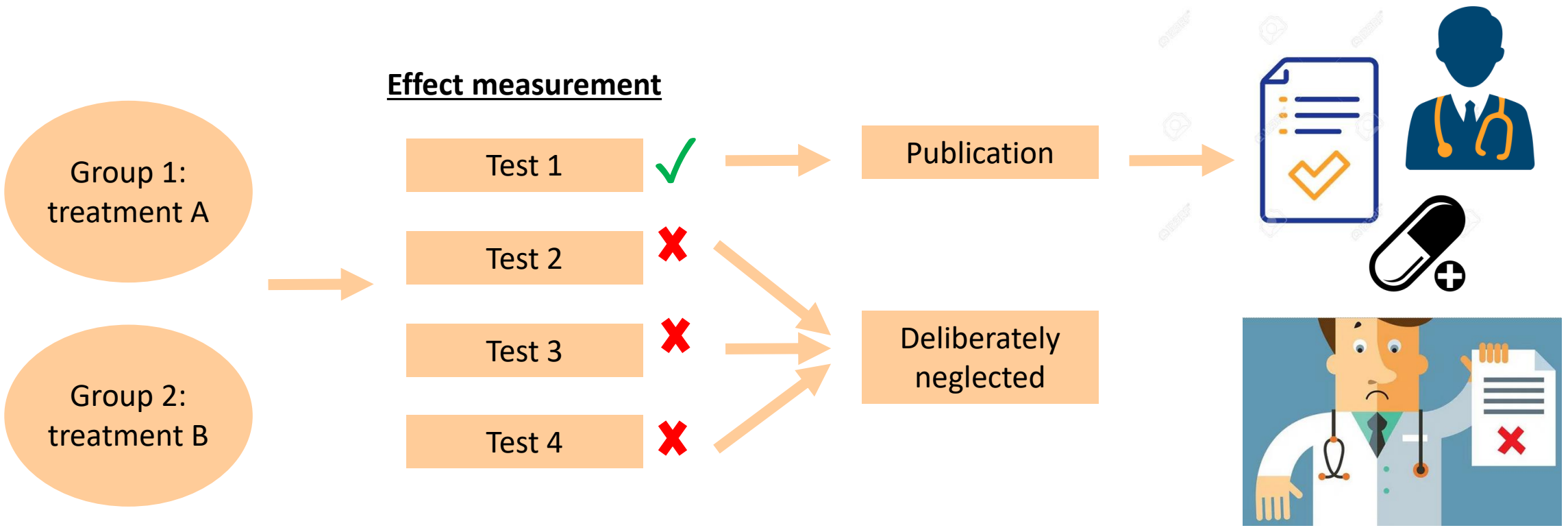
Intentionally filtering out data that does not support a pet hypothesis



What does my data go through?

Outcome switching

Changing the course of a protocol to **eliminate** inconclusive or negative results
Usually encountered in clinical trials



What does my data go through?

P-hacking

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

Even if it is random

What does my data go through?

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?

Bad habit = to boil down a complex scientific result to only one p-value!

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, ...)

Keep in mind that the data drives the conclusion, not the opposite

What does my data go through?

No panic:

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

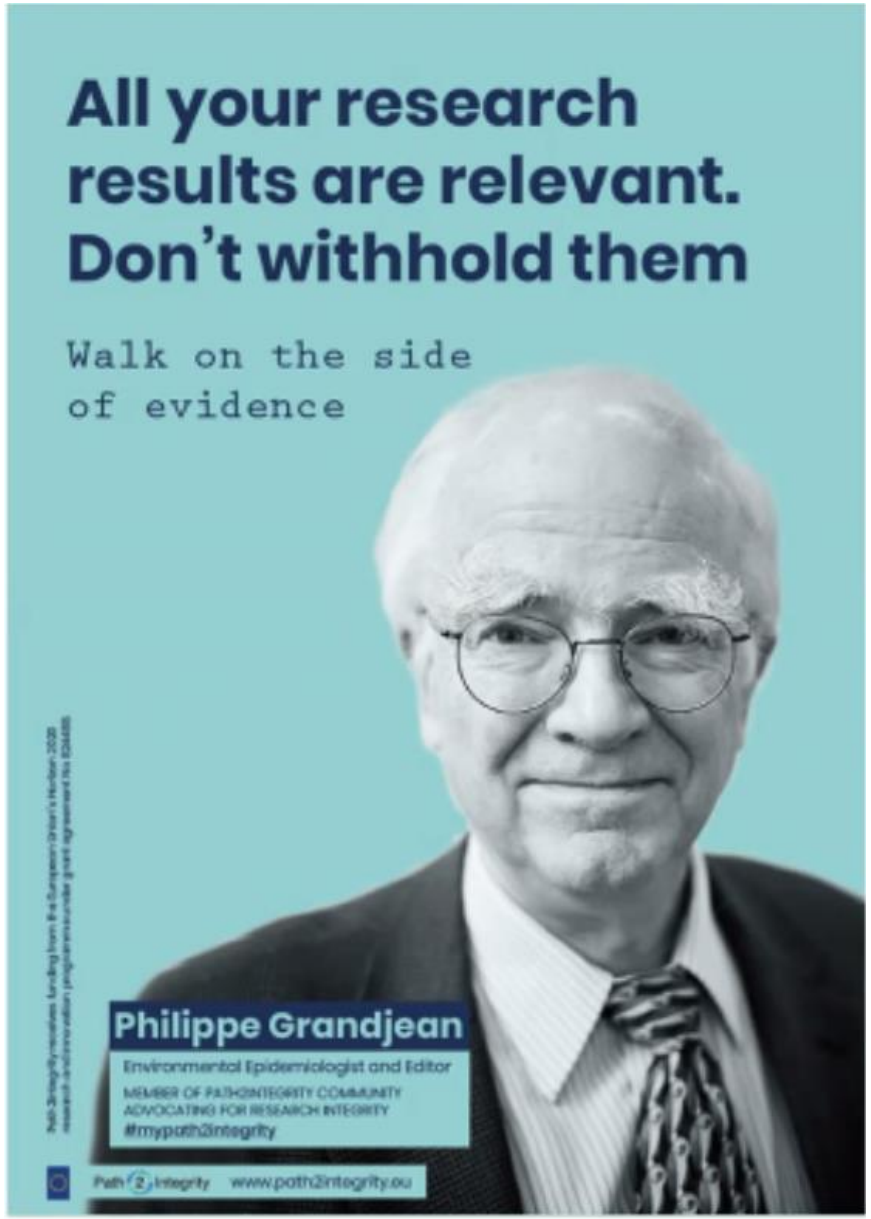
Making raw data, protocols, methodologies...

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

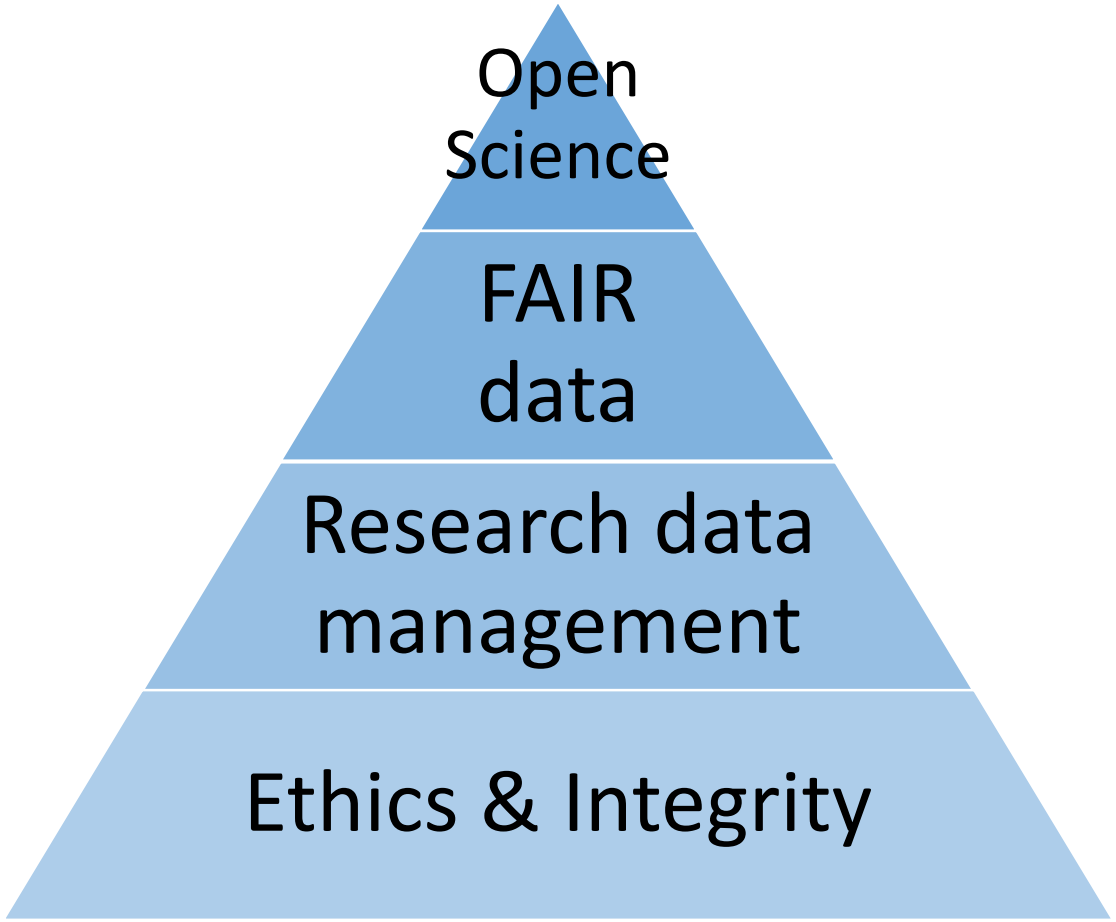
Your doctoral school, your supervisor, your lab, your stats teacher...

Training sessions in catalog:

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



How should I share my data?



- More and more **funders** require FAIR/Open data
- Increased **visibility** for your research: more access means more citations (+reputation)
- **Re-usability**-> better **collaboration** and **productivity**
- Possibility for the community to **replicate** and validate your result
- Continuity and **sustainability** of the data, even after staff change for example

How should I share my data?

As open as possible, as closed as necessary

Cannot be shared

GDPR,
Confidentiality

There is still hope...!

RDM habits are also for yourself ☺

Open Data

Zero obstacles

Not a token of quality...!

Should be the general aim

How should I share my data?

As open as possible, as closed as necessary

Cannot be shared

Open Data

FAIR data

FAIR principles

Findable Data are discoverable and easy to find, either by humans or computer - **Metadata** - *Digital Object Identifier* - Or other standard identifier	

FAIR principles

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

FAIR principles

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

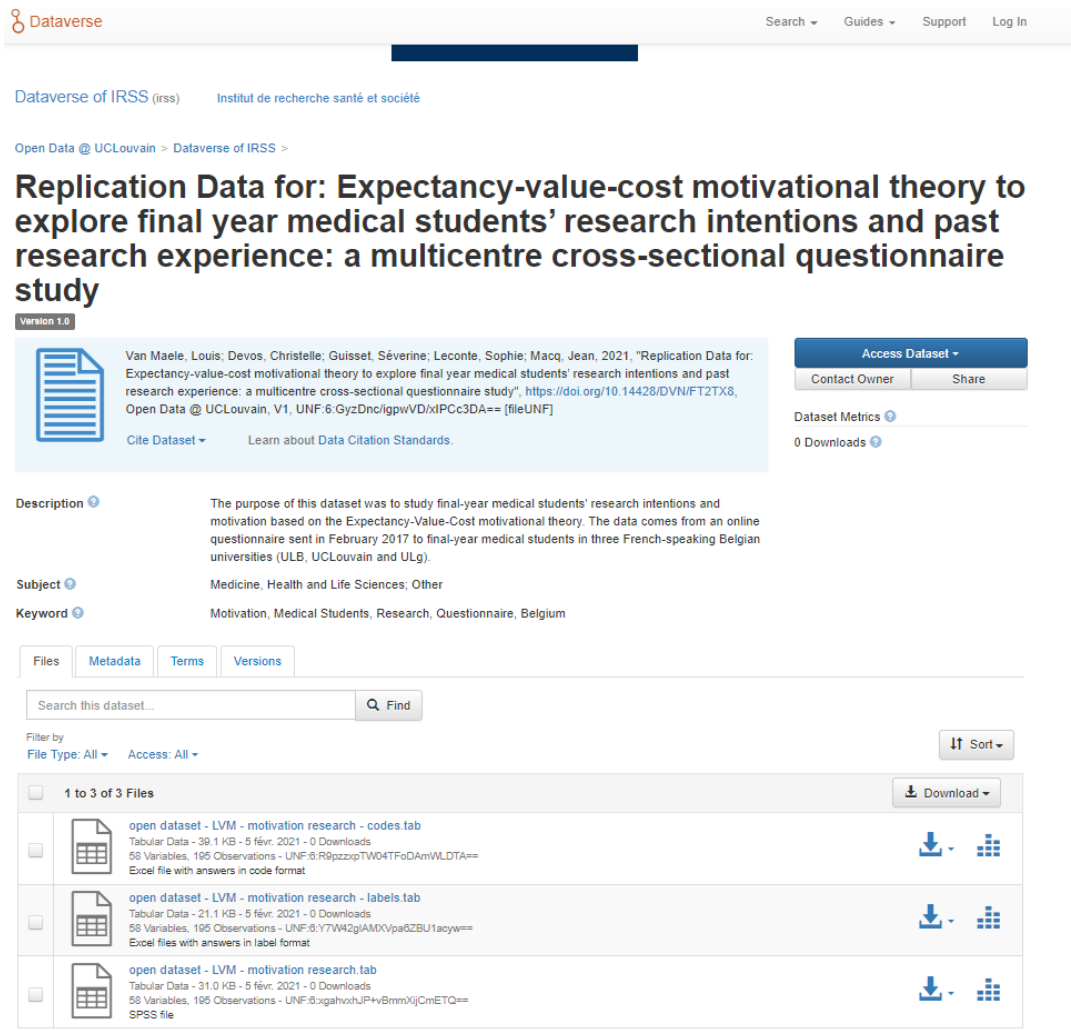
FAIR principles

How to **generate** metadata:

- When using a pre-existing data repository: **text fields** to be filled out

- Many file formats use **headers** (but not always easy to edit)

- If all else fails: companion text file !



FAIR principles

Findable

Data are discoverable and easy to find, either by humans or computer

- **Metadata**
- *Digital Object Identifier*
- Or other standard identifier

Accessible

Data are made available in a sustainable way, even after the end of the project

- The data (or **at least** metadata) are **retrievable** in a relatively well-known **directory** (not only on your website)
- If the data cannot be shared, it has to be **justified**

FAIR principles

Findable

Data are discoverable and easy to find, either by humans or computer

- **Metadata**
- *Digital Object Identifier*
- Or other standard identifier

Accessible

Data are made available in a sustainable way, even after the end of the project

- The data (or **at least** metadata) are **retrievable** in a relatively well-known **directory** (not only on your website)
- If the data cannot be shared, it has to be **justified**

Interoperable

Data can be operated, exchanged, compared or re-used, in a **variety** of contexts, institutions, workflows, software, systems...

FAIR principles

Findable

Data are discoverable and easy to find, either by humans or computer

- **Metadata**
- *Digital Object Identifier*
- Or other standard identifier

Accessible

Data are made available in a sustainable way, even after the end of the project

- The data (or **at least** metadata) are **retrievable** in a relatively well-known **directory** (not only on your website)
- If the data cannot be shared, it has to be **justified**

Interoperable

Data can be operated, exchanged, compared or re-used, in a **variety** of contexts, institutions, workflows, software, systems...

Reusable

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 and sustainable as possible

How should I share my data?

How to implement FAIR / open 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**

How should I share my data?

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

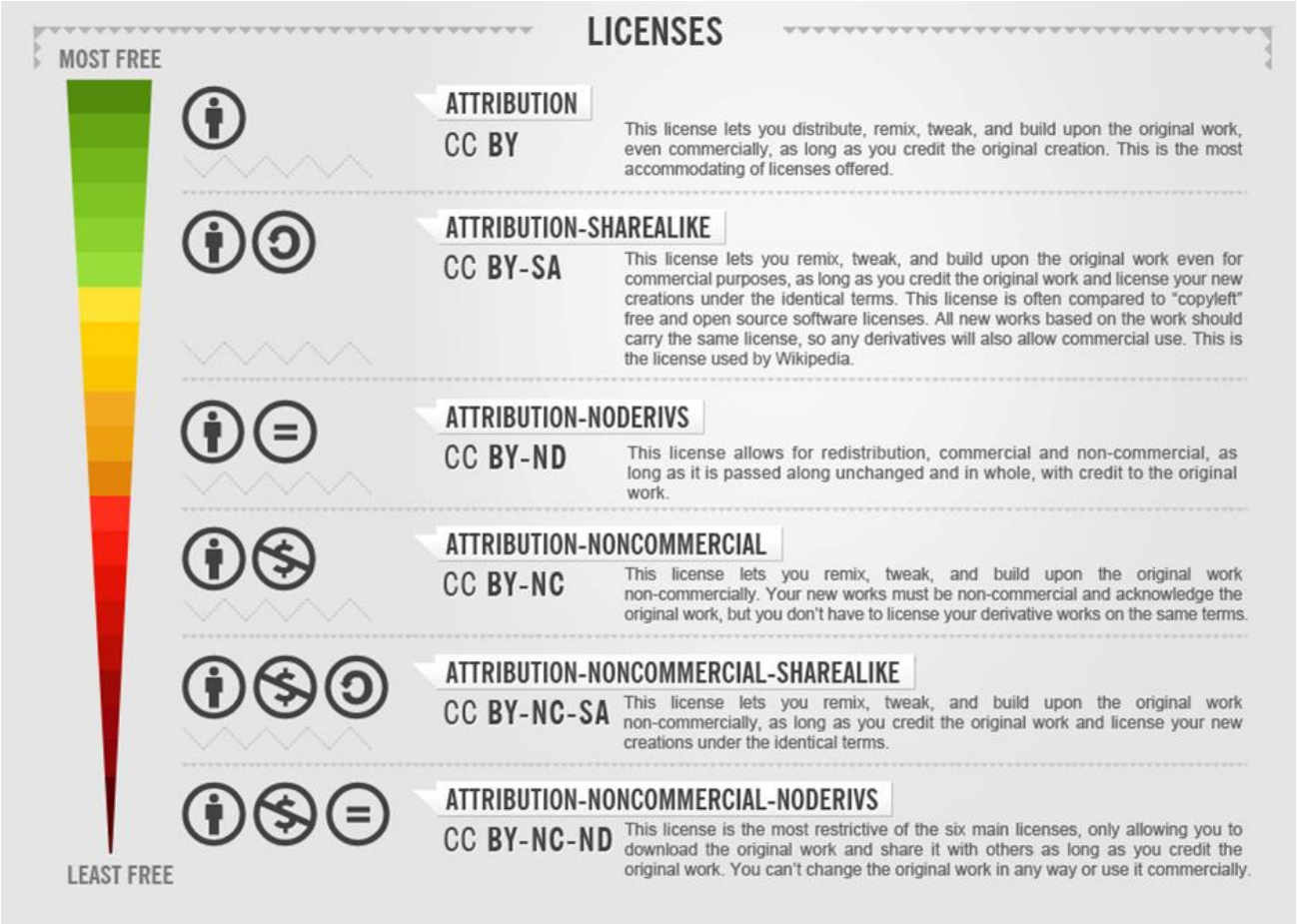
Rights to **reuse**, to **modification**, to **commercial** use, **obligation** to mention the **attribution** and to **share alike**

Sometimes, choosing a license can be restricted:

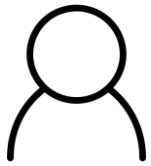
- The license may come with the use of a **repository**
- The license may come with the publication through an **editor** (journal)

Who can **help** me?

- Interface ULiège
- Libraries



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)