Introduction to data management in hard science

Damien Belvèze

29th of april 2022

Research Data: What are we talking about?

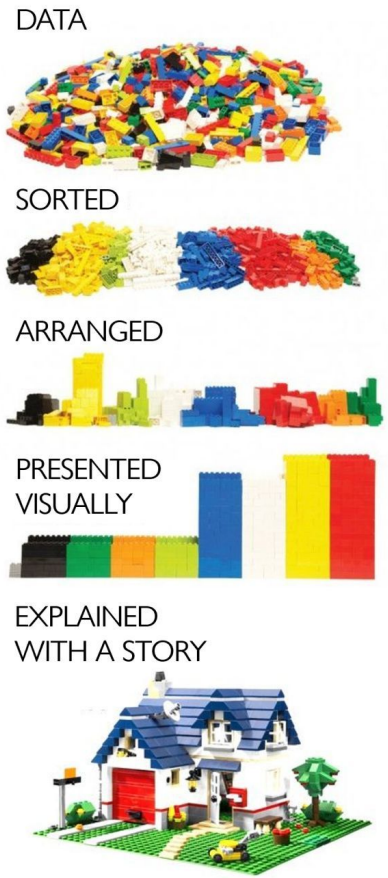
some definitions on data management

- **research data**: "recorded objects (figures, texts, images or sounds), which are used as main sources for scientific inquiry and are generally considered by the scientific community as compulsory to valid scientific results" (OECD, 2007)
- **Datasets**« Aggregation (...) of raw data or intermediary data with some unity, gathered in order to form a consistent whole » (Gaillard, 2014)
- **Open Science** :

> Open Science is the practice of science in such a way that others can collaborate and contribute, where research data, lab notes and other research processes are freely available, under terms that enable reuse, redistribution and reproduction of the research and its underlying data and methods

Open Science Definition.

Research Data: what are we talking about?

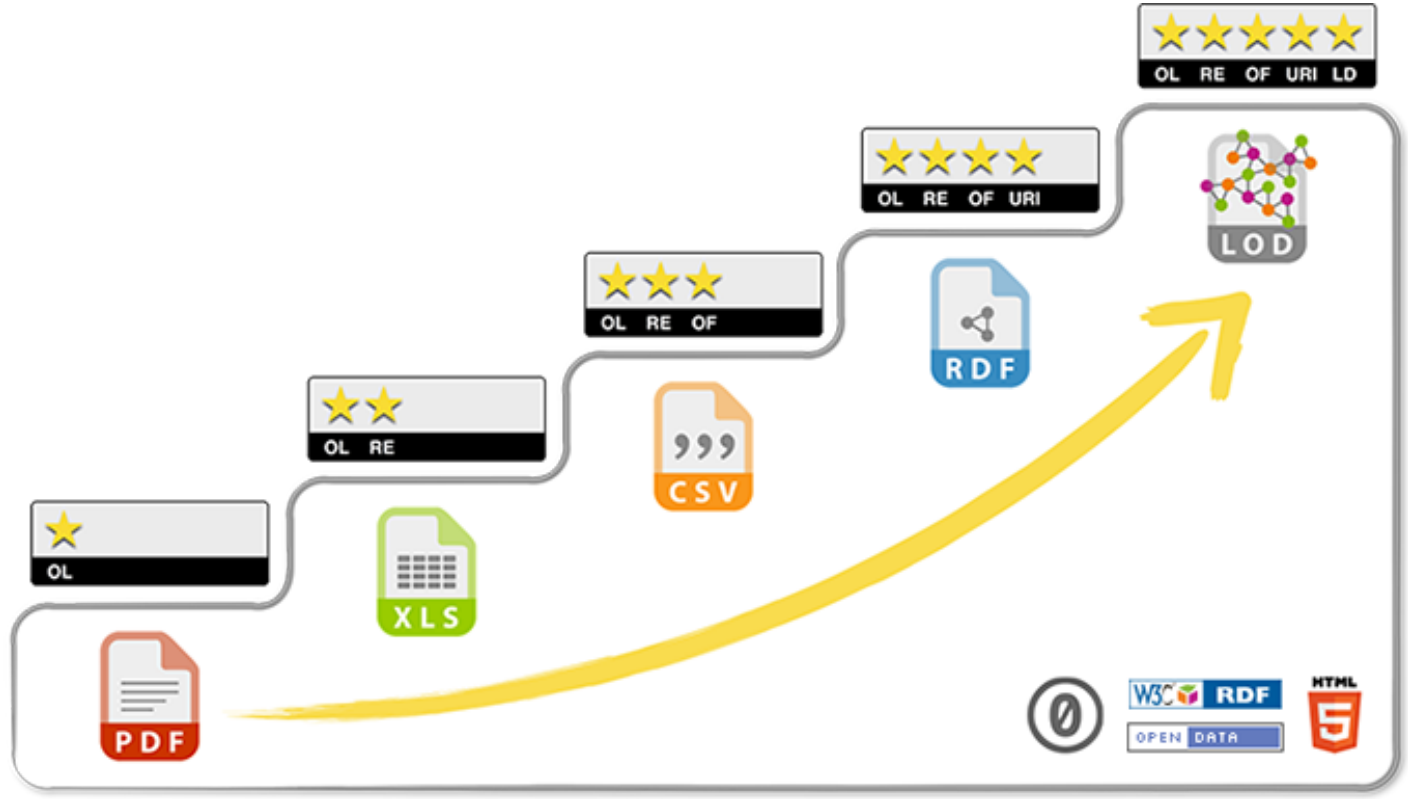


- Raw data
- derivative data
- analysed data (or interpreted data)

> Data doesn’t say anything. Humans say things.
> (Andreo Jones-Rooy)

Data accessibility : an old issue

As of year 1997, pioneers of the web have been wondering how data could be more structured in order to be more usable.



What about you? What kind of data do you produce?

Data storage, curation and publication: what is at stake?

Economic aspects

> Research data in digital form is increasingly used in works beyond the project for which it was originally collected, as well as in other research areas and in industry

OECD report 2007

Health issues

Sharing medical data helps cure people

collecting individual cases data for future meta-analyses has become strategic for public health

> no other patient was recruited for a study ; individual cases data, once gathered, were enough for the Gustave-Roussy Institute to make a meta-analysis that showed that concomitant chemotherapy is more efficient than sequential chemotherapy (Florian Naudet, Claude Pellen)

Source : The conversation

Democracy at stake

data as a Common

Obépine (research project) = measures the prevalence of Sars-Cov-2 in French waste waters - Request for access to data produced by a public entity

The answer refers to the "Law for a Digital Republic" aka "Lemaire's Law" (2016) whose goal is to improve circulation of data and knowledge from public institutions to the public.

demande Dada

data or *capta* ?

personal consent (GDPR)

If you use conversation transcripts or forms, do not forget to collect and confidentially store the consent forms for each transaction (form submission, interview)

Issues related to data retention and sharing

Scientific Issues

Science **reproducibility** and **reliability** are at stake

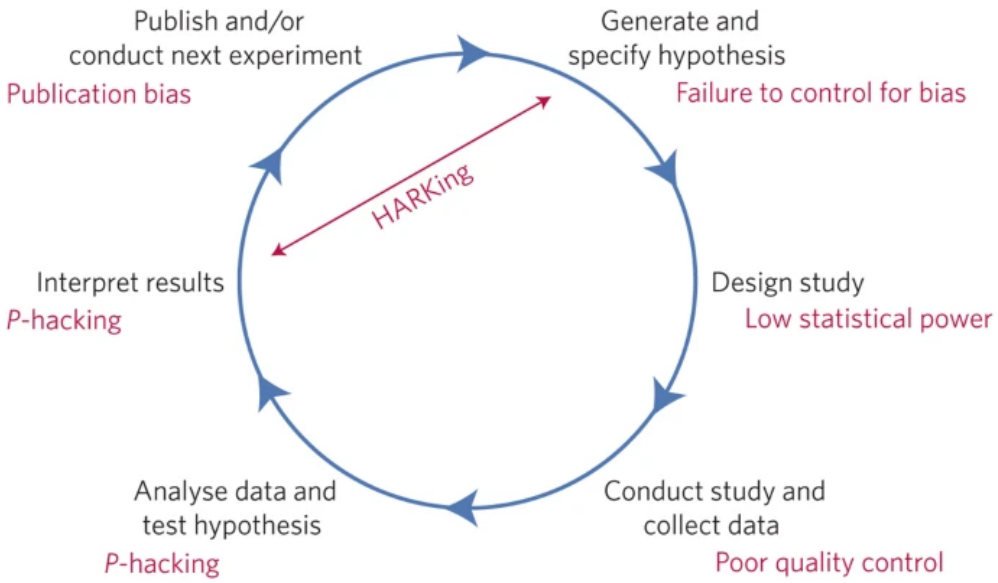
ie. LancetGate The Lancet retracts a paper on HCQ based on data that Surgisphere refused to provide to reviewers.

Lancet stated on June 5th 2020 that...

> Our independent peer reviewers informed us that Surgisphere would not transfer the full dataset [...] As such, our reviewers were not able to conduct an independent and private peer review

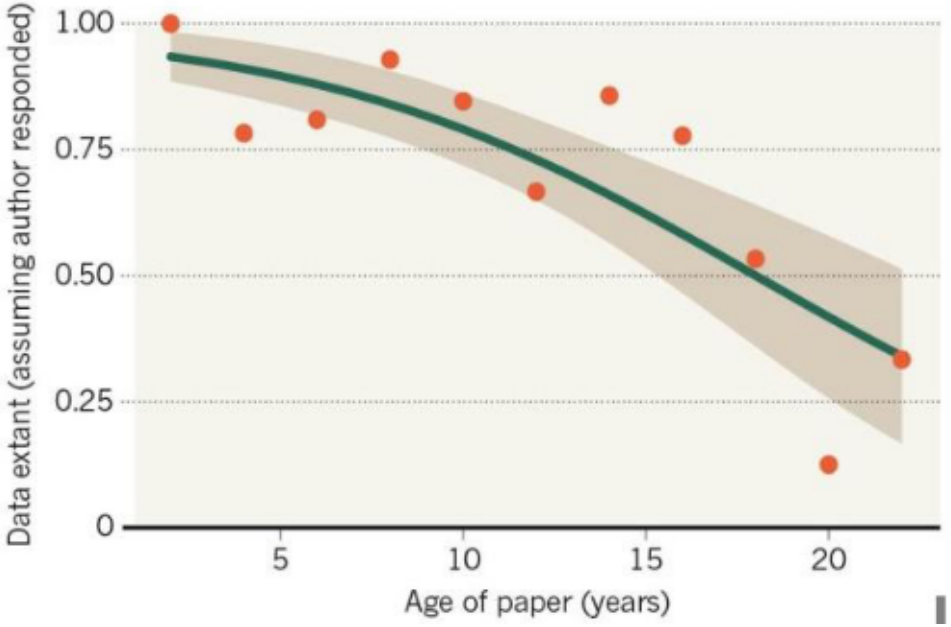
Le Monde, 4 juin 2020

Accessing the data is a condition to reproducibility



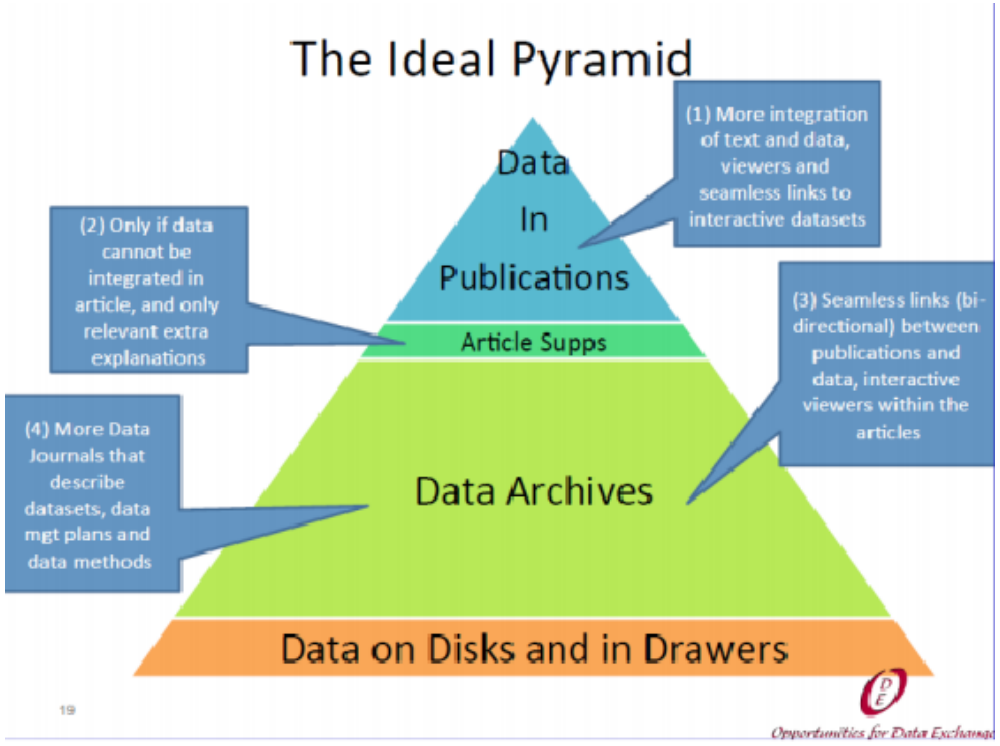
poor quality control on data Manifesto for reproducible science

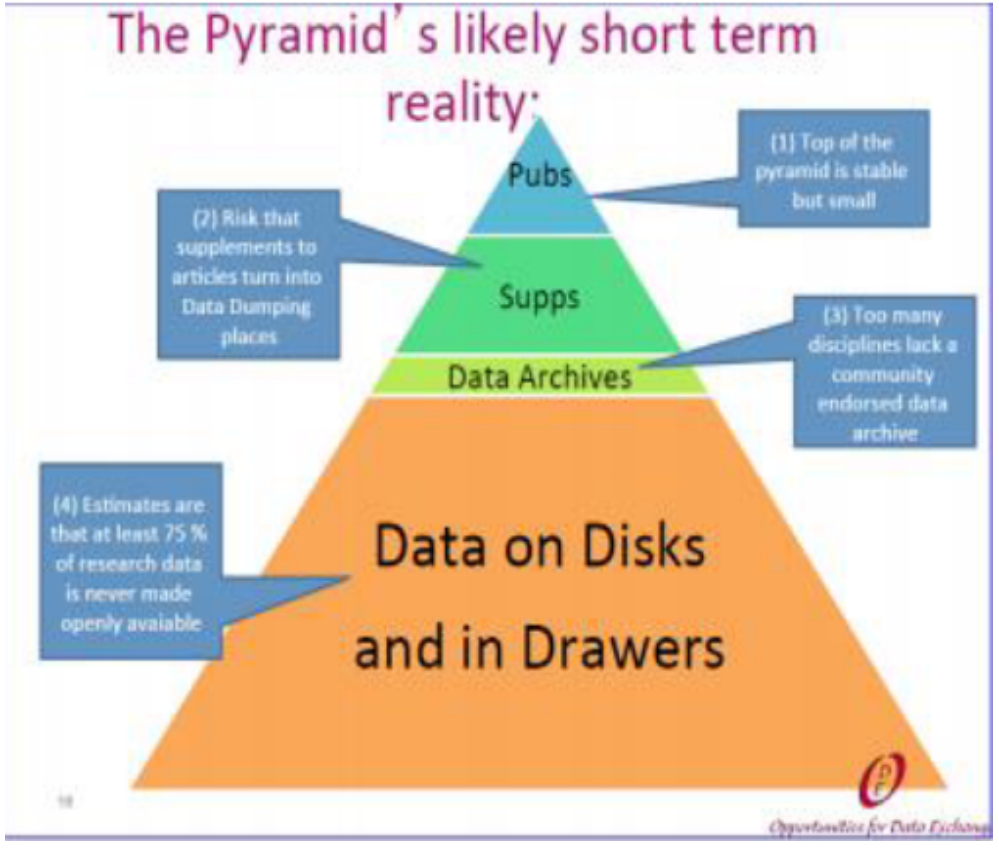
ephemeral data



80% of the data produced these last twenty years are lost.

How it should be and how it is





(Marie Puren, 2021)

National Plan for Open Science (S Plan)

- make open access by default for publications
- **Plan S strongly encourages that research data and other research outputs are made as open as possible and as closed as necessary**
- code that underlies the publication should be made available in external repositories

Plan S: principles and implementations

FAIR principles

Findable / Accessible / Interoperable / Reusable

How to make your data *findable*?

- dataset -> DOI (exemple : https://doi.org/10.5281/zenodo.5903186)
- data repositories should be well indexed by search engines like Google data set search (SKOS implementation improves this referencing)

How to make my data accessible ?

> as open as possible, as close as necessary

Be explicit on how the dataset could be reused and shared with others. Choose a licence adapted to these conditions

Citation advantage

> We also find an association between articles that include statements that link to data in a repository and **up to 25.36% (± 1.07%) higher citation impact** on average (Colavizza, Hrynaszkiewicz, 2020)

Category	Definition	Example
0	Not available	*No additional data available* (common).
1	Data available on request or similar	*Supporting information is available in the additional files and further supporting data is available from the authors on request* (DOI: 10.1186/1471-2164-14-876).
2	Data available with the paper and its supplementary files	*The authors confirm that all data underlying the findings are fully available without restriction. All data are included within the manuscript* (DOI: 10.1371/journal.pone.0098191).
3	Data available in a repository	*The authors confirm that all data underlying the findings are fully available without restriction. The transcriptome data is deposited at NCBI/Gene Bank as the TSA accession SRR1151079 and SRR1151080* (DOI: 10.1371/journal.pone.0106370).

https://doi.org/10.1371/journal.pone.0230416.t001

How to make my data interoperable?

- Use open formats for your data
- use ontologies and controlled languages (thesaurus) that are standard in your field

And why I would share MY data?

- This is a lot of work to collect them!
- These data are a competitive asset for my laboratory
- Loi Lemaire : the public pre-empts the data to prevent it from becoming the property of private actors (avoiding enclosures)
- Unlike her works, the data actually belong to the researcher's institution, they are not hers.

How to write a data management plan (DMP)

DMP is compulsory for any research contract with National Research Agency (ANR)

must be submitted 6 months after the signature of the contract

The DMP states what data will be collected and how they will be processed to be made FAIR compliant

where should you store your data?

Not on the publisher's website

- data repositories give you more place and freedom to describe your data
- publisher's policy may impede free reusability or sharing of your data (according to Open Science principles)
- publishers may have interest to make money with your data or use them to build products for sale (see for instance Elsevier and Scival)

Where should we store our data then?

- re3data.org (waiting for *recherche data gouv* next summer)
- wiser to load them on a repository **before the publication of the related article** (once the paper is published, no possibility to make a link from the publisher website to the dataset on the repository) it it's too late, you can still make a link from the open access repository (HAL) where you have deposited your AAM to the dataset.
- if your data are not ready to be deposited, you may nevertheless reserve a place on a data repository to get a permanent ID to the dataset by just loading a file (data description for instance)
- if you use your own data, provide a space before the bibliography entitled "data availability" for instance (and put here the DOI of the datasets)
- if you reuse date from another project, cite datasets in the bibliography

consider the cost of data preservation

- The cost of data preservation should be balanced with the cost of production of a new dataset
- if your data are expensive to collect, you should archive them for longer, if the cost of preservation are higher than the cost of production, consider collecting new data

Store enough data to make reproducibility possible but no more :

- If your processes are well documented, you will be able to reproduce intermediary data from raw data (in this case, only raw data have to be stored in order to minimize the environmental footprint)
- computational data : virtualization (Docker) and versioning (github)

> every state of data that can be reproduced may not be stored. If we need intermediary data, and if every process is properly documented, it will only cost some computing time & resources to get them from the raw data and thus you will have spared the cost of storage for a large amount of refined data (Olivier Collin)

Why it is so important to describe your data

Write and make available at the root of your folder a readme file, with these informations:

- file organisation and localisation in the folder
- definition of the variables in the data and explanations about how the observations were conducted
- details on the experimentation settings
- details on the data processing (which operations, with what tools...)

description : ontologies and controlled languages.

Ontologies :

- genomic ontologies
- biological ontologies

thesaurus, controlled vocabularies :

- MeSH (médecine)
- INRAE thesaurus

Documenting your data make them more reusable (and prevent bad interpretation)

Lack of context = no reusability

data : from a privacy perspective

anonymisation # pseudonymisation

Tools: Amnesia et Arx

The more pseudonomysed personal data you have in a dataset, higher is the risk of a reidentification

credits

This presentation was widely inspired by :

- Thierry Fournier, introduction aux données de la recherche en sciences exactes, 2022
- Cécile Arènes, Rédiger un plan de gestion des données, 2021

bibliography

Blanc, I. (2020, novembre 20). Données relatives à la concentration de SARS-CoV-2 dans les eaux usées—Une demande d'accès à l'information à Ministère de l'enseignement supérieur, de la recherche et de l'innovation. Consulté 14 avril 2022, à l'adresse Ma Dada website: https://madada.fr/demande/donnees_relatives_a_la_concentra

Brunori, G. (2020). Data Management Plan. https://doi.org/10.5281/zenodo.3664215 Colavizza, G., Hrynaszkiewicz, I., Staden, I., Whitaker, K., & McGillivray, B. (2020). The citation advantage of linking publications to research data. PLOS ONE, 15(4), e0230416. https://doi.org/10.1371/journal.pone.0230416

Munafò, M. R., Nosek, B. A., Bishop, D. V. M., Button, K. S., Chambers, C. D., Percie du Sert, N., … Ioannidis, J. P. A. (2017). A manifesto for reproducible science. Nature Human Behaviour, 1(1), 1-9. https://doi.org/10.1038/s41562-016-0021

OCDE. (2007). Principes et lignes directrices de l'OCDE pour l'accès aux données de la recherche financée sur fonds publics. Consulté à l'adresse https://www.oecd.org/fr/science/inno/38500823.pdf

Sandve, G. K., Nekrutenko, A., Taylor, J., & Hovig, E. (2013). Ten Simple Rules for Reproducible Computational Research. PLOS Computational Biology, 9(10), e1003285. https://doi.org/10.1371/journal.pcbi.1003285

« The Lancet » annonce le retrait de son étude sur l'hydroxychloroquine. (2020, juin 4). Le Monde.fr. Consulté à l'adresse https://www.lemonde.fr/sciences/article/2020/06/04/hydroxychloroquine-trois-auteurs-de-l-etude-du-lancet-se-retractent_6041803_1650684.html

Vines, T. H., Albert, A. Y. K., Andrew, R. L., Débarre, F., Bock, D. G., Franklin, M. T., … Rennison, D. J. (2014). The Availability of Research Data Declines Rapidly with Article Age. Current Biology, 24(1), 94-97. https://doi.org/10.1016/j.cub.2013.11.014