Ready for BioData.pt Management?

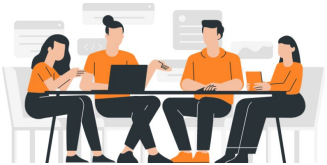
Intensive Course

Introduction to Research Data Management

Daniel Faria, Jorge Oliveira, Gil Poiares-Oliveira

BioData.pt | elixir PORTUGAL

Ready for BioData.pt Management?



Intensive Course

Introduction to Research Data Management

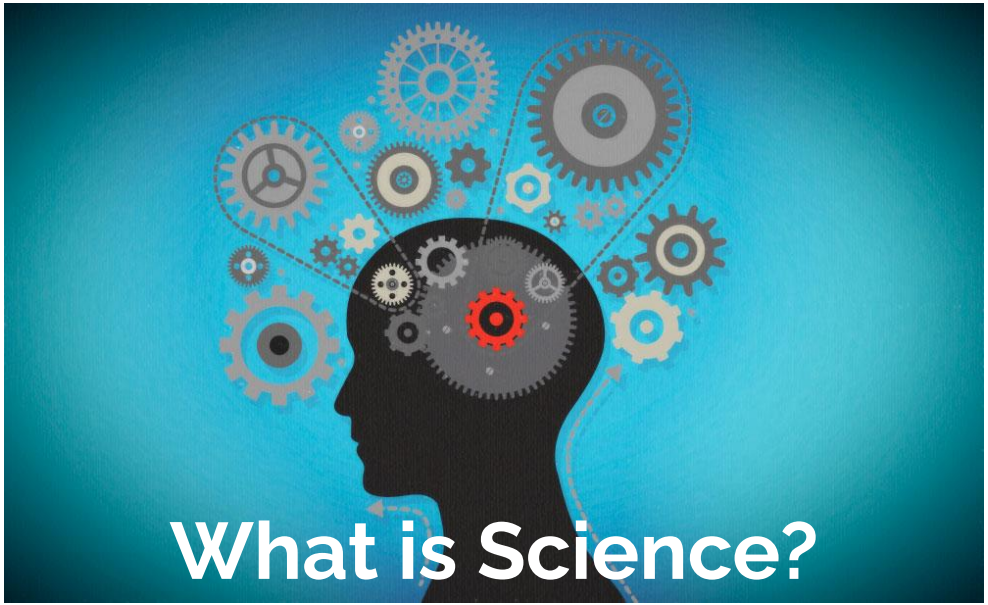
I – The Core Concepts

Learning Outcomes:

- Understand the basic requirements for science
- Distinguish data, information and knowledge

BioData.pt | elixir PORTUGAL

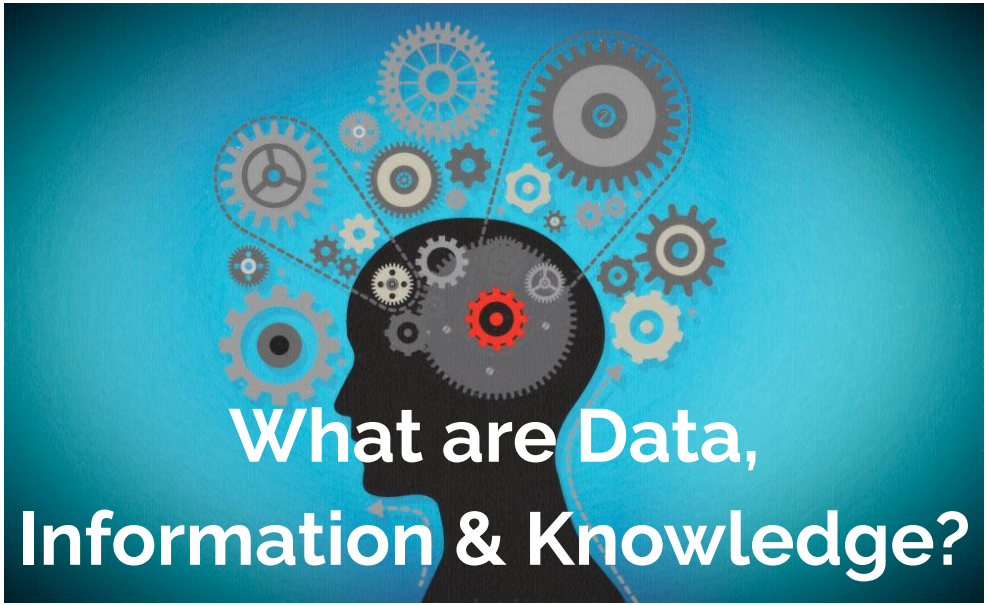
Brainstorming Moment



What is Science?

- A **knowledge** discovery paradigm
- Predicated on hypothesis testing through experimentation
- Usually implies **data** acquisition and analysis
- Requires testability and **reproducibility**

Brainstorming Moment

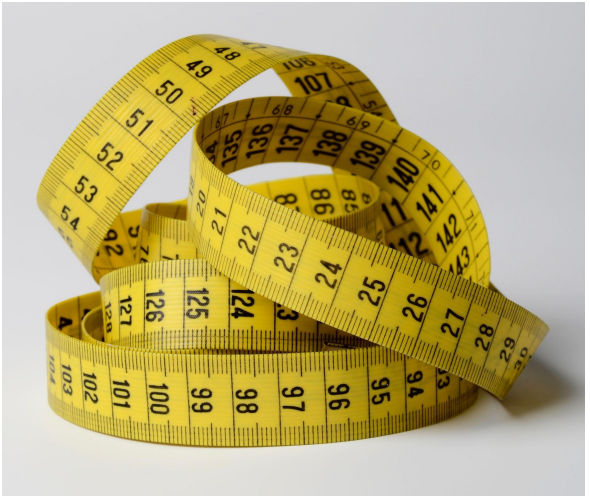


What is Data?

- **Datum:** an atomic fact or piece of "information"
  - e.g. a triple asserting a predicate about a subject:

    <water> <boiling point> <100℃>

- **Dataset:** a collection of data that share a subject (sensu latu) or scope



By Marta Longas from Pexels

What is Information?

- **Information:** data + context (metadata)
  - enabling the interpretation of the data:

    purity: <500 mg/L dissolved solids

    pressure: ~1 bar

What is Metadata?

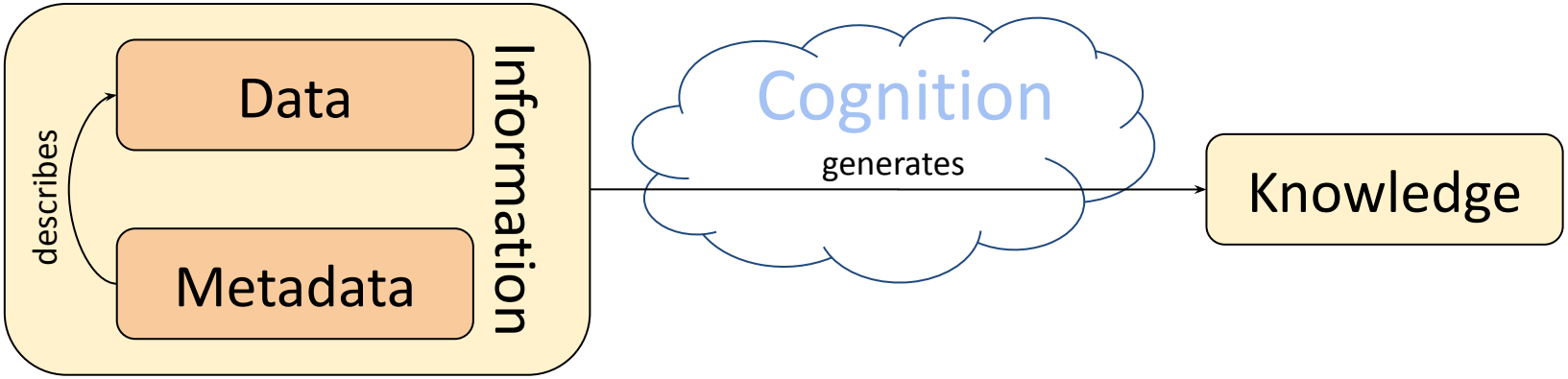
- **Metadata:** data about data, providing context
  - Who produced the data?
  - When was the data produced?
  - Why was the data produced?
  - What is the data about?
  - How can the data be used?

By Dr. Marcus Gossler - Own work, CC BY-SA 3.0

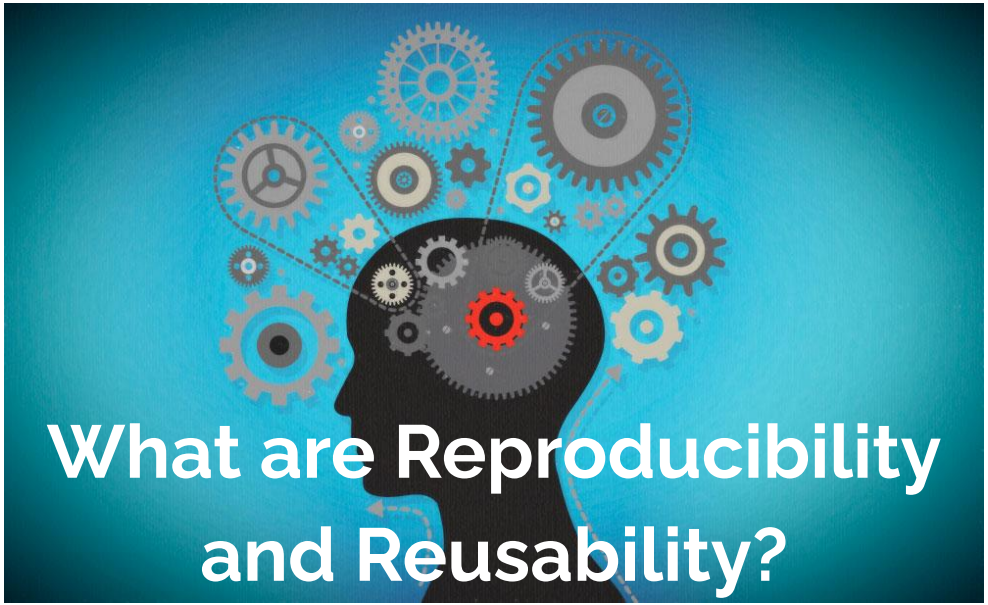
What is Knowledge?

- **Knowledge:** (actionable) understanding of information
  - If I heat tap water until it starts boiling, I can cook food at 100℃
  - If I see water boiling, I shouldn't put my hand in it

Data, Information & Knowledge



Brainstorming Moment



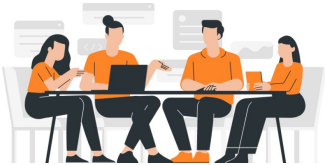
What is Reproducibility?

- **Reproducibility:** the same analysis performed on the same data produces the same results
  - If an experiment is not reproducible, it is not science
  - If it is not generalisable, then it cannot lead to knowledge

		Data	
		Same	Different
Analysis	Same	Reproducible	Replicable
	Different	Robust	Generalisable

What is Reusability?

- **Reusability**: data generated to test a scientific hypothesis (or for exploratory science) can be readily adapted (and combined with other data) to test other hypotheses
  - Critical in data-intensive research fields, where data acquisition is expensive



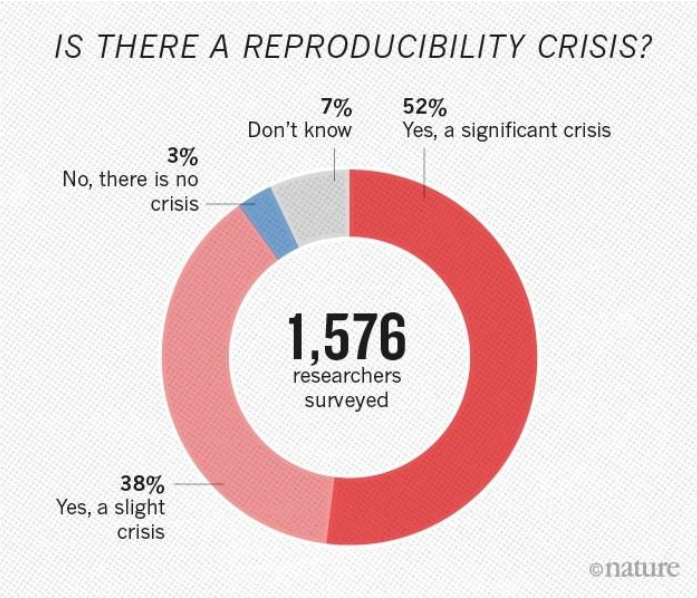
II – The Research Landscape

Learning Outcomes:

- Discuss the merits and demerits of the Open Science movement
- Explain how to comply with the FAIR data

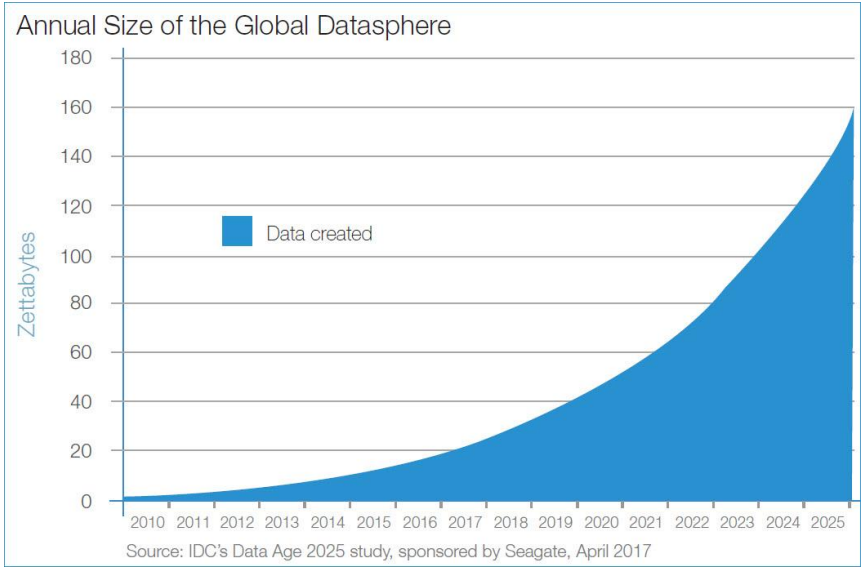
The Reproducibility Crisis

- Researchers are pressured to publish as they are evaluated on their publication record
- The prevalent culture is that only research with positive results merits publication
- These lead to fake science (e.g. p-hacking)



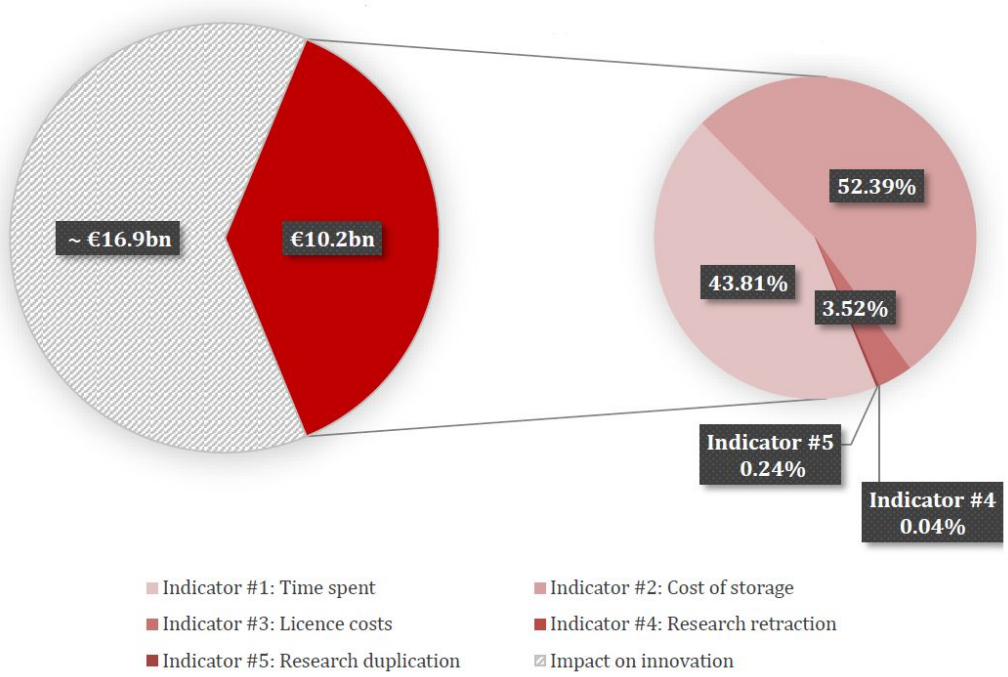
Ready for BioData.pt Management?

Exponential Data Production



- The same pattern emerges if we look at just research data or just research articles

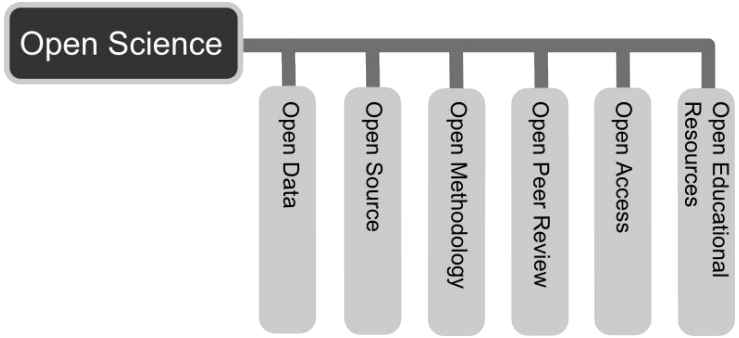
The Cost of Low Reusability



From EC report "Cost of not having FAIR research data", March 2018

Open Science

- **Vision:** scientific research and its dissemination accessible to all levels of society
  - publications
  - data
  - samples
  - software

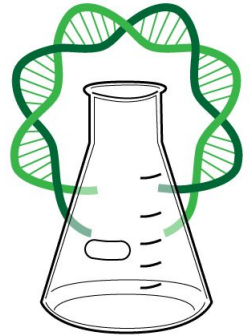


- **Goal:** transparent and accessible knowledge shared and developed through collaborative networks

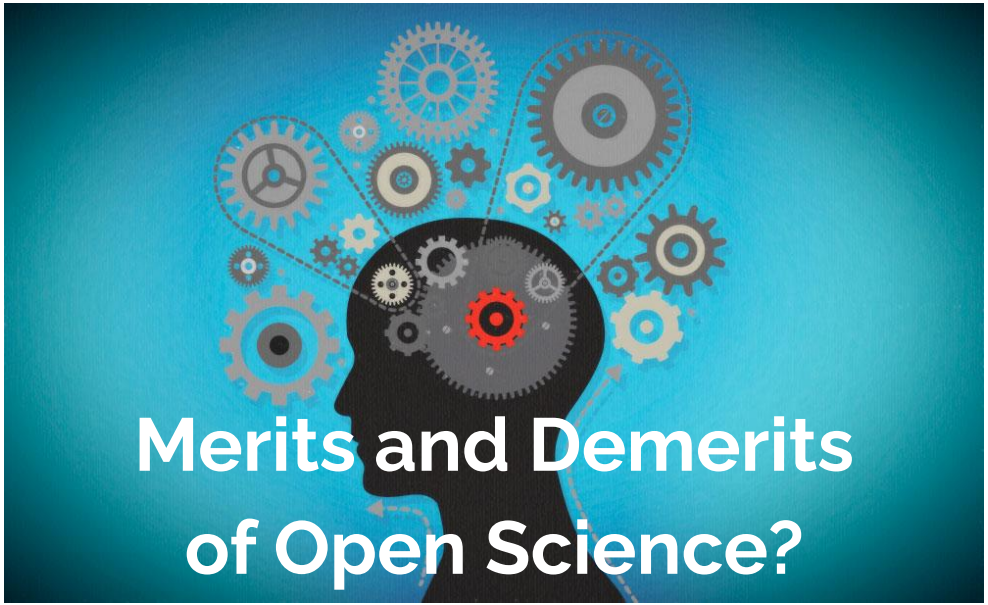
Open Science

Layers:

- **Open Access:** research outputs distributed online, free of cost or access barriers
- **Open Research:** data, result and methodology clearly documented and freely available online
- **Open-Notebook Science:** primary record of a research project publicly available online as it is recorded—no insider information



Brainstorming Moment



Merits of Open Science

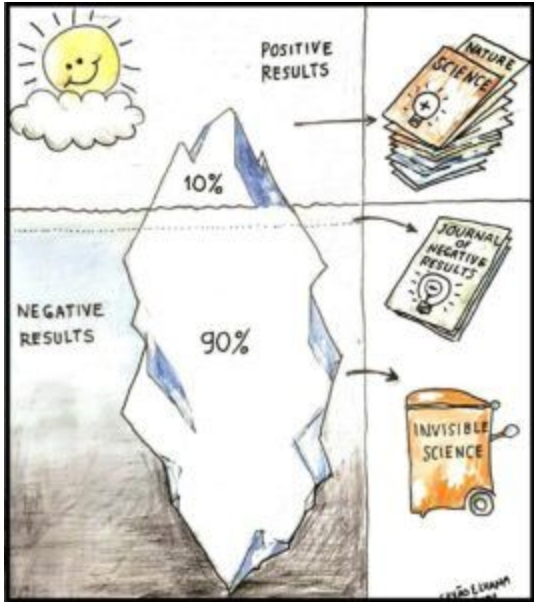
- Sharing the data and detailed methods is essential for **reproducibility**
  - There must be **very** strong arguments to justify publishing an article without sharing the data (e.g. sensitive and/or personal data)
  - Open Science prevents fake science

Merits of Open Science

- Sharing the data and detailed methods is essential for **reusability**
  - If the research was publicly funded, the funder rightfully expects a return to society, and data is often the most valuable output
  - Open Science increases the value of science to society

Merits of Open Science

- Open notebook science **prevents publication bias** (i.e. only positive results get shared)
  - Negative or inconclusive results may be of value if the science is sound
  - The data is likely of value



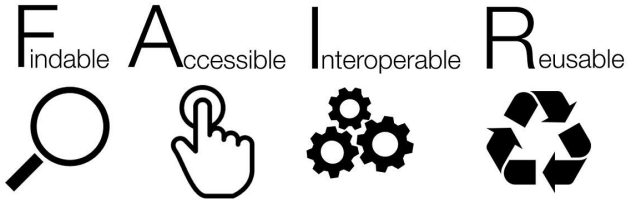
Demerits of Open Science

- Does not explicitly contemplate:
  - Forms of direct (economic) value from research such as the creation of patents or the commercialization of products/services, which are usually welcomed by funders
  - Personal and/or sensitive data
  - Criteria to ensure reproducibility and reusability

Demerits of Open Science

- Open notebook science overlooks the fact that there are research domains where there is strong competition, in which sharing ongoing research outcomes would expose the authors to theft
  - Demands for researchers to share data openly must be matched with measures to safeguard their right to publish their findings from their own data

The FAIR Data Principles



By SangyaPundir - Own work, CC BY-SA 4.0

- Build upon the Open Science vision by:
  - Specifying the criteria that research data should comply with to be reusable (and reproducible)
  - Explicitly contemplating the need for data privacy "*as open as possible, as closed as necessary*"

Findability

- Describe data with precise metadata (keywords) for searching
    - Use ontologies for metadata fields and values
- Put data in a repository that uses persistent unique identifiers, indexes metadata and allows searches

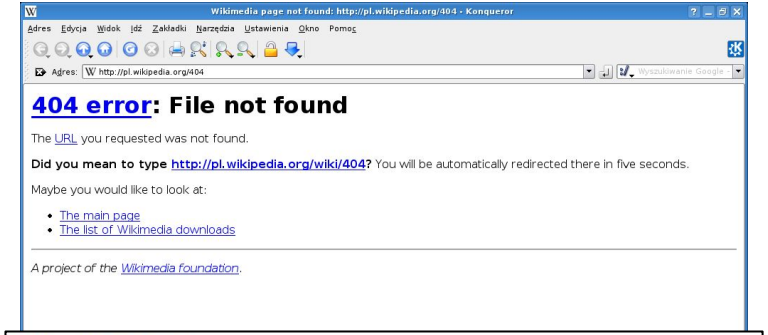
Things get lost in a sea of things!

By Martin Handford, retrieved from: https://exploringyourmind.com/how-does-our-brain-find-waldo/

Accessibility

- Put data in a repository that:
  - Uses persistent unique identifiers
  - Has a standard access protocol
  - Preferably supports both human and computer access
  - Has authentication and authorization protocols, if the data require it

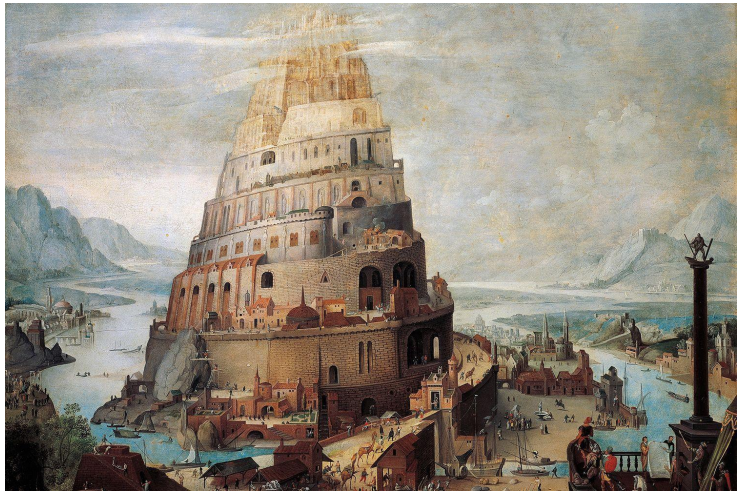
Define who can access and how



Interoperability

- Use standard (open) file formats
- Use ontologies for metadata fields and values
- Follow applicable metadata standards
- Include cross-references to external data objects when suitable

Reconcile vocabulary & file formats among research communities

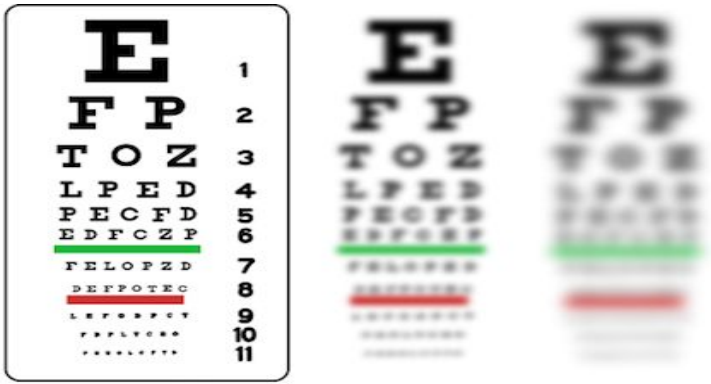


By Abel Grimmer, retrieved from: http://cbcpnews.net/cbcpnews/the-tower-of-babel/

Reusability ♻

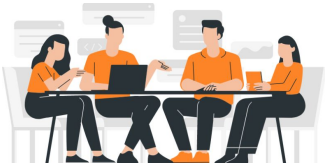
- Hinges on **interpretability**
- Describe data with sufficient metadata for interpreting it and understanding the experimental context (datasets should be self-contained)
- Use ontologies for metadata fields and values

We can't afford to read a research paper to interpret every dataset!



Key Points

- Publishing data only as an appendix in a scientific papers is not enough
  - Papers are not efficient vehicles for knowledge transfer!!!
- Data must be published in a FAIR manner
  - It need not be published before the scientific publication
  - It need not be fully public

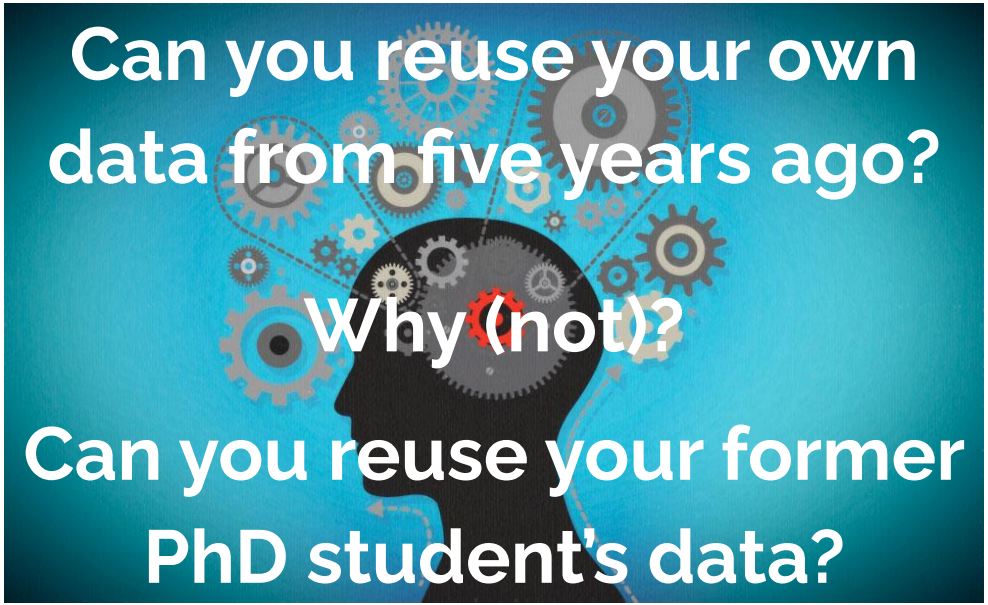


III – Research Data Management

Learning Outcomes:

- Explain the need for Research Data Management
- Understand the demands of the funders

Brainstorming Moment



Self-Reuse Challenge

Can you reuse your own data from five years ago?

Can you reuse your former PhD student's data?

- Both require good data management practices across the data lifecycle:
  - Data organization, documentation, storage, versioning
    - FAIR data by design

Research Data Management (RDM)

- Encompasses all the processes involving research data across all the stages of the data lifecycle, e.g.: collecting, organizing, documenting, processing, analysing, storing, sharing
- It is a research field in its own right, but core RDM knowledge is critical to all researchers (like knowing how to use a computer or write an article)

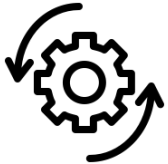
https://rdmkit.elixir-europe.org/

Research Data Management (RDM)

Improves research…

By LAFS, CC BY 2.0

Effectiveness: better handling of data leads to better science



By Youmena, CC BY 2.0

Efficiency: less time wasted in routine tasks improves productivity

By ROZMOWA, CC BY 2.0

Security: control over data access and less risk of data loss

By Nithinan Tatah, CC BY 2.0

Impact: easier to produce FAIR data, more likely to be reused

Documentation, Documentation, Documentation

- Researchers often only think about documentation when forced to (for writing an article or publishing a dataset)
  - Risk of losing data for irrelevant experiments made
  - Tendency to do the minimum required
  - Struggle with using applicable metadata standards and ontologies

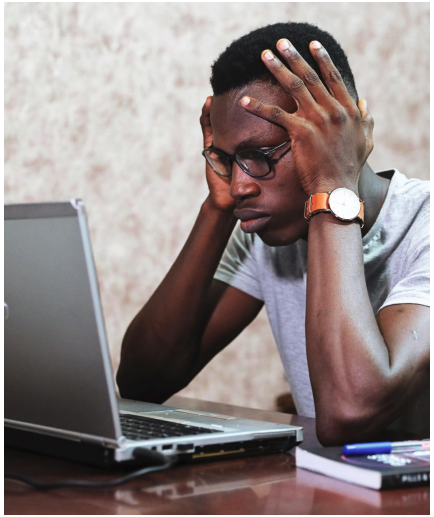


Photo by Oladimeji Ajegbile from Pexels

Documentation, Documentation, Documentation

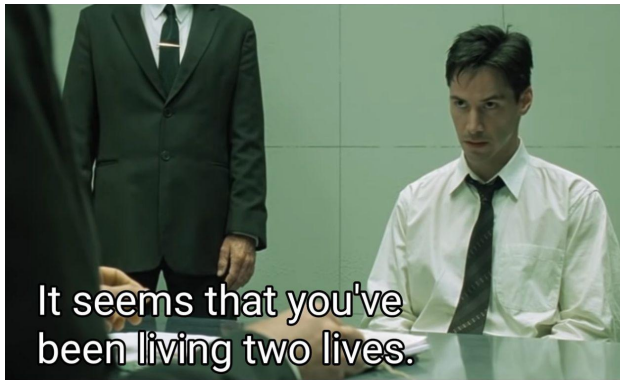
- **Continuous** documentation across the data lifecycle is the key for good data management
- Everything that is relevant for data reuse, from the experimental design (plan) to the usage license (share) including all processes the data undergo, must be clearly documented

https://rdmkit.elixir-europe.org/

Data Management Plan (DMP)

A. A formal document used to plan and support data management activities by anticipating needs and requirements in a (research) project, facility or institution

B. A bureaucratic demand of funding agencies that we draft because we have to and then archive in the basement



Homework

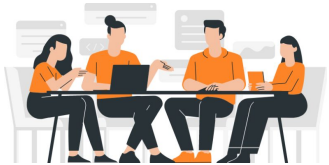
- Is there a default public database or repository for your research domain?
  - What are its metadata requirements?
- Is there a community metadata standard?
  - Does it cover your use case?
- Are there adequate ontologies?
- Are there default data (open) file formats

Homework

- Is there a data steward in your institution?
- Are there RDM guidelines or an institutional DMP?
- Is there IT infrastructure for data storage and computing?
  - What are the access conditions and costs?

Getting Started

- RDMkit a knowledge hub on research data management that approaches topics from multiple vantage points and links to relevant tools and training
- FAIRCOOKBOOK a compendium of specific recipes for producing FAIR data
- FAIRsharing.org standards, databases, policies an index of metadata standards, ontologies, databases, and data policies that should be considered when publishing FAIR data



Thank You!

Questions?