BioData.pt elixir PORTUGAL

Ready for BioData Management?

DAY 2 DAY DATA MANAGEMENT COURSE

FOR LIFE SCIENCES

Funding:

- ELIXIR Implementation Study "Impact evaluation at Node-level - getting it done"
- ELIXIR-CONVERGE Project “Connect and align ELIXIR Nodes to deliver sustainable FAIR life-science data management services”

Agenda

9h30-10h30: Introduction to Research Data Management

10h30-11h00: Resources for Research Data Management (RDMkit, DMPortal, DSW, ELN)

11h00-11h15: Coffee break

11h15-12h00: The IGC use case

12h00-13h00: Hands-On Exercise with eLab Notebook, in pairs

13h00-14h00: Lunch break

14h00-15h00: Hands-On Exercise with eLab Notebook (cont.)

15h00-15h10: Ice breaker

15h10-16h30: Hands-On Exercise with DMPortal, in pairs

16h30-17h00: Wrap-up and closing

Trainers

Daniel Faria
Assistant Professor at Instituto Superior Técnico & Researcher at INESC-ID
Scientific Coordinator of the Ready for BioData Management program

Jorge Oliveira
CTO of BioData.pt
Invited Assistant Professor at Instituto Superior Técnico

Carolina Ventura
Data Steward at Instituto Gulbenkian de Ciência

BioData.pt elixir PORTUGAL

Ready for BioData Management?

DAY 2 DAY DATA MANAGEMENT COURSE

FOR LIFE SCIENCES

Introduction to Research Data Management

Daniel Faria

Part I – The Core Concepts

Learning Outcomes:

- Distinguish Data, Information and Knowledge
- Understand Data Management and DMPs
- Define Reproducibility and Reusability

What is Science?

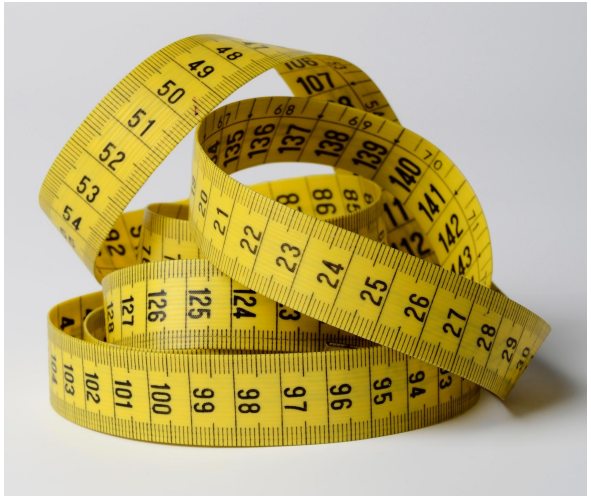
What is Science?

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

What are Data, Information and Knowledge?

What is Data?

- Datum: an atomic fact or piece of “information”
  - Melting Point: 0℃
  - Boiling Point: 100℃
- Dataset: a collection of data that share an object or scope



By Marta Longas from Pexels

What is Information?

- Information: data + context (metadata)
  - Object: Water w/ < 500 mg/L dissolved solids
  - Experimental conditions: 1 atm
  - …

What is Metadata?

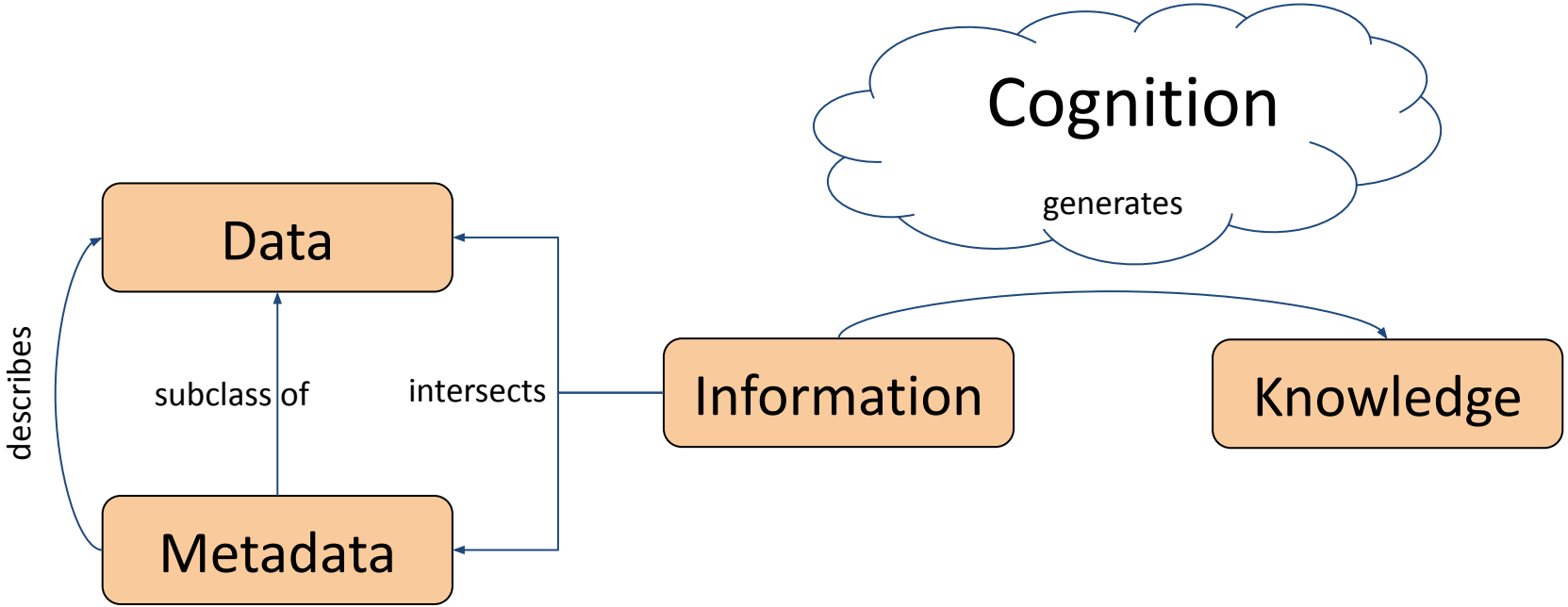
- Metadata is data about data, providing context:
  - **Who** produced the data?
  - **When** was the data produced?
  - **What** is the data about?
  - **Why** was the data produced?
  - **How** can the data be used? (license)

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

What is Knowledge?

- Knowledge: information + (actionable) understanding
  - 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

What are Data, Information and Knowledge?



What is Research Data Management?

What is Research Data Management?

- Research data management (RDM) encompasses all the processes involving research data across all the stages of the data lifecycle: collecting, organizing, documenting, processing, analysing, storing, and sharing

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

What is Research Data Management?

- Good RDM aims at:
  - Improving research effectiveness and efficiency
  - Improving data security (avoiding (meta)data loss)
  - Preventing errors
  - Increasing research impact
  - Ensuring **reproducibility** and promoting **data reusability**

What is a Data Management Plan?

- A DMP is a formal document used to plan and support data management activities by anticipating needs and requirements in a (research) project, facility or institution
- It is to data management what a blueprint is to construction

What are Reproducibility and Reusability?

What is Reproducibility?

- Research is reproducible if the same analysis performed on the same dataset consistently produces the same result
- Reproducibility is necessary but not sufficient for science
- Lack of reproducibility means poor scientific practices or fake science

		Data	
		Same	Different
Analysis	Same	Reproducible	Replicable
	Different	Robust	Generalisable

What is Reusability?

- Research data is reusable if it can be readily adapted and combined with other data to test research hypotheses other than the one for which it was collected
- Reusability is critical in data-intensive research fields, especially where data acquisition is expensive

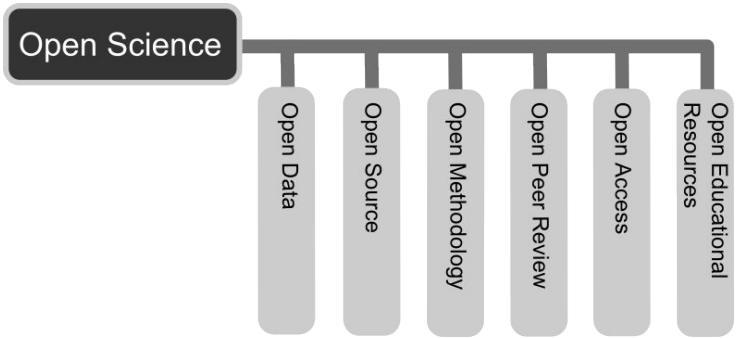
Part II – The Demands & Motivations

Learning Outcomes:

- Explain the Open Science movement and the FAIR data principles
- Understand the demands of funders

Open Science

- Scientific research and its dissemination accessible to all levels of society
  - publications
  - data
  - physical samples
  - software
  - ...



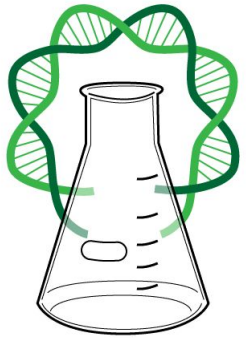
By Andreas E. Neuhold, CC BY 3.0

- 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



By Greg Emmerich, CC BY-SA 3.0

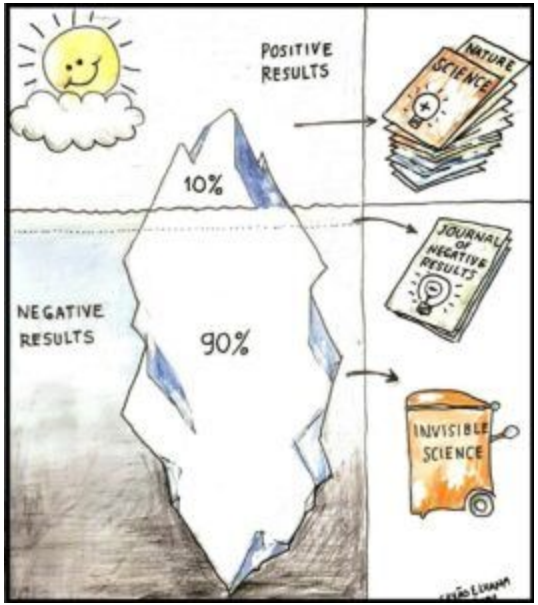
To Share or Not to Share Research Data?

To Share

- If the research was funded through public funds, the funder expects some return to society, and data is often the most valuable output of the project (**reusability**)
- Sharing the data and experimental protocol is essential for **reproducibility**, and should be demanded by journals and/or during peer review

To Share

- Negative results should also be published and even data that was not conducive to scientific publications can and should be shared
    - It may be useless to you but it can be useful for someone else

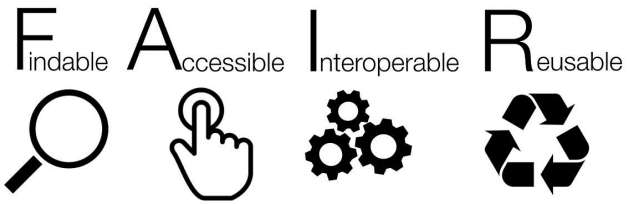


Not To Share

- Data supporting a pending patent or that can be leveraged into a start-up company
    - Both returns to society as far as funders are concerned
- Personal and/or sensitive data

The FAIR Data Principles

Findability: (Meta)data are easy to find for both humans and computers



By SangyaPundir - Own work, CC BY-SA 4.0

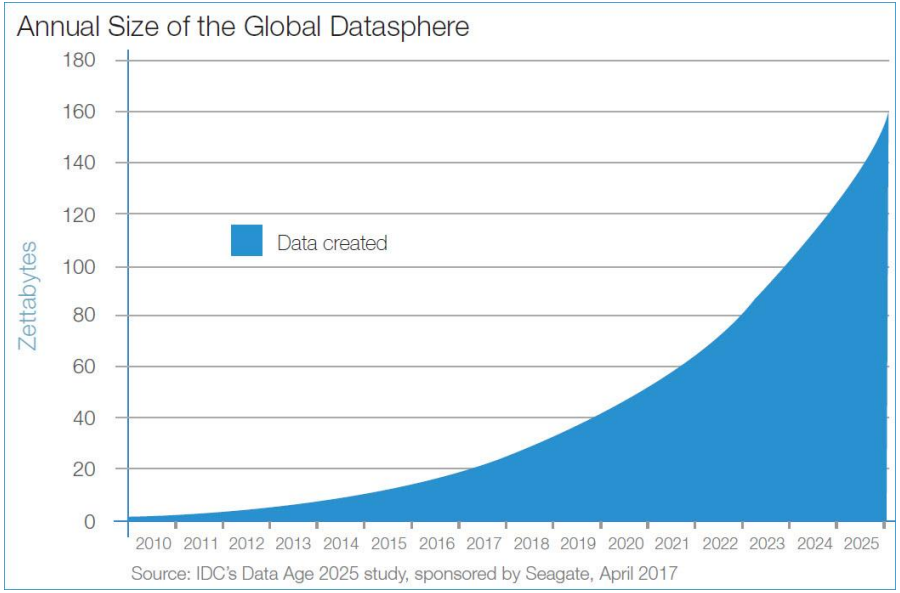
Accessibility: (Meta)data have a defined access protocol with authentication and authorization rules

Interoperability: (Meta)data are integratable with other similar datasets and interpretable by applications or workflows for analysis, storage and processing

Reusability – (Meta)data should be well described so that it can be interpreted and reused

Findability

Our data production is growing exponentially



Findability

- More data ⇒ harder search
- Things can get lost amid a sea of things
- If we can’t find it, we can’t reuse it

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

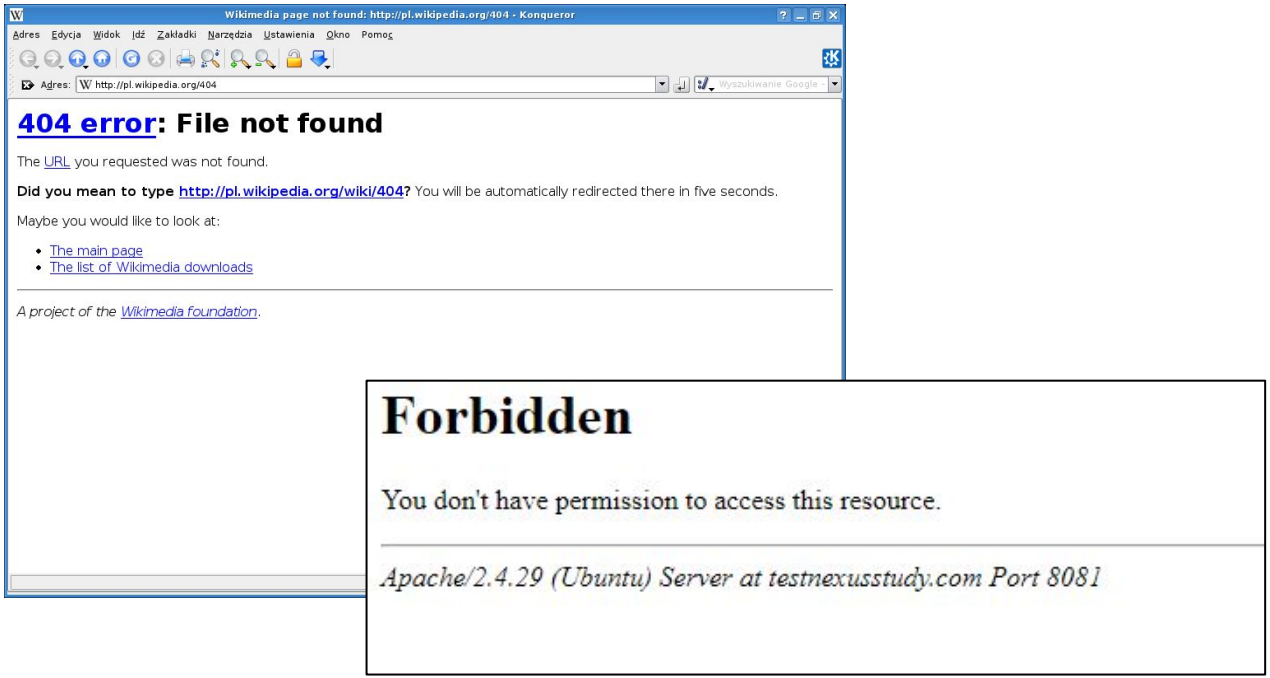
How to Make Research Data Findable?

How to Make Research Data Findable?

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

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

Accessibility

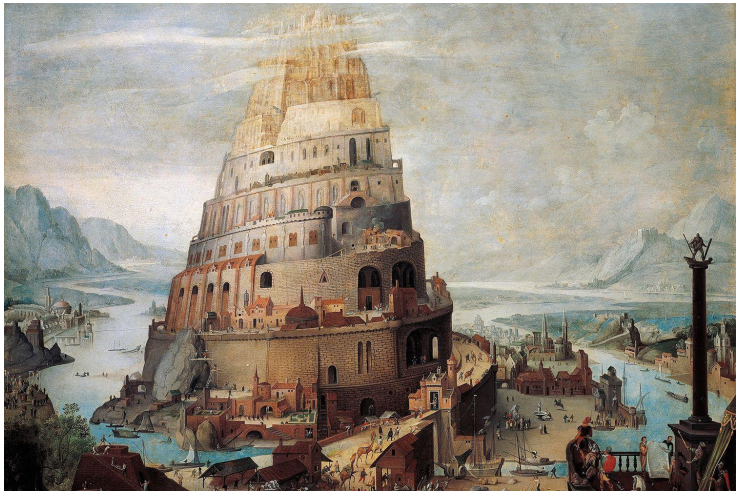


How to Make Research Data Accessible?

- 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 requires it

Interoperability

- Specialization of science ⇒ divergence of vocabulary, file formats, data organization and viewpoints
- All lead to sundered data that cannot be combined with other data and/or used by particular applications or workflows

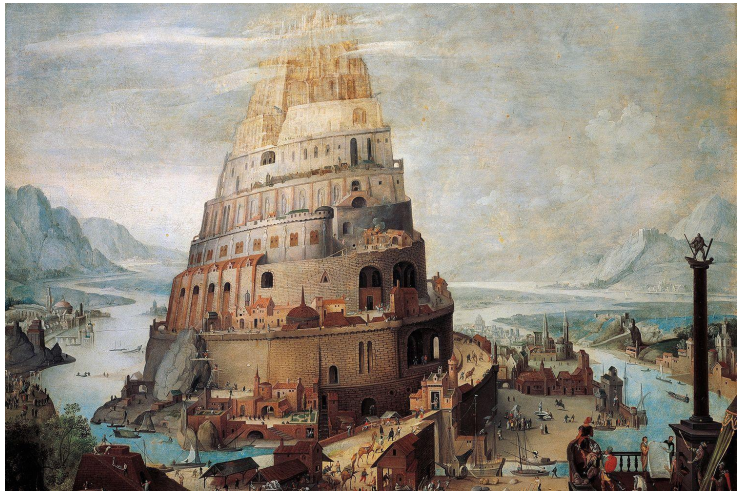


By Abel Grimmer, retrieved from: http://cbcpnews.net/cbcpnews/the-tower-of-babel/

How to Make Research Data Interoperable?

How to Make Research Data Interoperable?

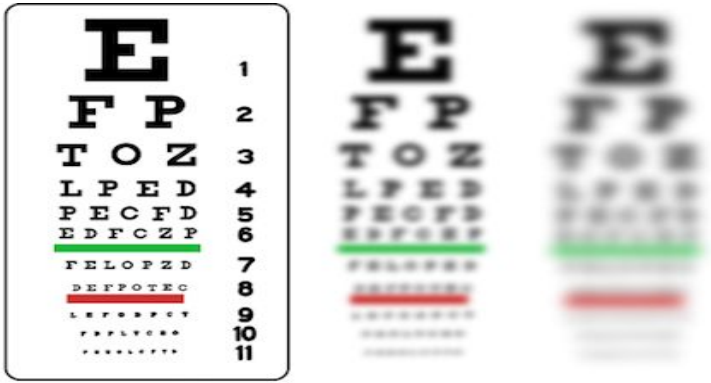
- Use standard (open) file formats
- Use a (structured) controlled vocabulary for metadata fields and values
- Include cross-references to external data objects whenever suitable



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

Reusability

- Reusability hinges on findability and interoperability, but also on **interpretability**
- We need to be able to readily interpret the data to figure out if and how we can reuse it
- We can’t afford to read a whole research paper to interpret every dataset

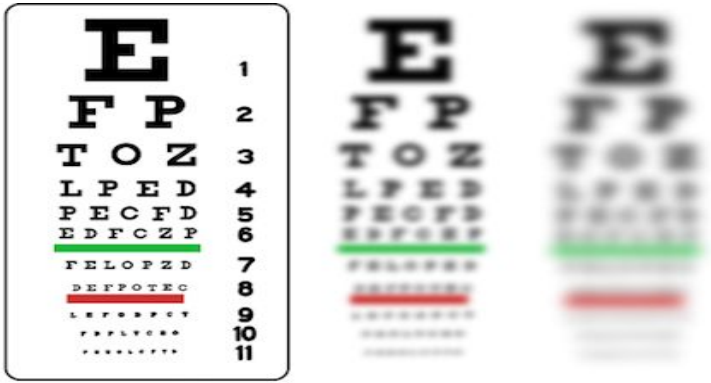


By Daniel P. B. Smith, CC BY-SA 3.0

How to Make Research Data Interpretable?

How to Make Research Data Interpretable?

- Describe data with sufficient metadata for interpreting it and understanding the experimental context—each dataset should be fully self-contained
- Use a (structured) controlled vocabulary for metadata fields and values



By Daniel P. B. Smith, CC BY-SA 3.0

The Funders' Demands

- Research funders are increasingly demanding that you comply with the FAIR data principles and that you include DMPs in your research projects that demonstrate how you will comply
- Very soon it will be impossible to get European or national funding without a DMP
- And validation of FAIRness is coming in the near future

Wrap Up

- Publishing data only in scientific papers is not enough
  - Papers are not efficient vehicles for knowledge transfer!!!
- Data must be published in a FAIR manner
- Data need not be published before the scientific publication
- Data need not be fully public
  - “As open as possible, as closed as necessary”

Part III – Day-to-Day Data Management

Learning Outcomes:

- Describe the challenges of producing FAIR data
- Explain the benefits of good RDM

Producing FAIR Data

- Consider national policies, institutional policies, funder's policies
- Learn specific recipes in the FAIR Cookbook
- Consult information hubs about existing standards, such as FAIRsharing.org
- Search for key concepts through ontology lookup services, such as BioPortal

Producing FAIR Data

- Is there a default public database or repository for your research domain?
  - Does it have a metadata schema?
- Are there community metadata standards?
  - Do they cover your use case?
- Are there adequate ontologies?
- Are there default data (open) file formats?

Producing FAIR Data

- Organize, Document & Annotate:
  - Your code / scripts / workflows,
  - Your protocols
  - Your data & metadata
- According to the applicable guidelines / standards or the repository where you're depositing your data / materials
- Using domain ontologies, recommended file formats

Producing FAIR Data

- **Is a lot of work**
  - Especially if you only think about it when publishing your data
    - You have to trace all the data and experimental details—risk of (meta)data loss
    - Inertia and rush lead to poor job

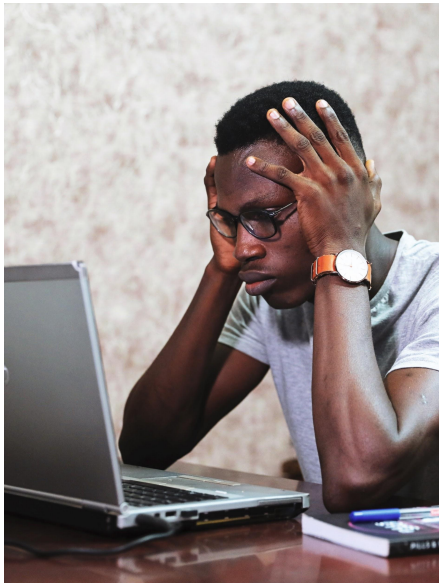


Photo by Oladimeji Ajegbile from Pexels

Producing FAIR Data

- **Is much easier with good data management across the data lifecycle**

Plan

- Do your homework
  - Legal and ethical issues
  - National / funder's / institutional policies
  - Best practices in your domain
  - Standards and ontologies
- Allocate resources to subsequent stages
  - People, equipment, time
- **Make a DMP!**

Collect

- Capture **provenance** and **methodological metadata** according to established standards
- Use standard (open) file formats
- Ensure data quality
- Ensure (temporary) data storage: persistent, secure, redundant
- Use an **electronic lab notebook** (ELN) to keep a digital record of your research!

Process & Analyse

- Document all processing and analysis steps, according to established **metadata** standards
- Use/create workflows to ensure **reproducibility**
  - Document your code!
- Ensure link between storage and computing
- Name and organize your files sensibly and document file structure if not self-evident
- Use version control software (e.g. git)

Preserve

- Ensure long term data storage (for the duration determined during planning): persistent, secure, redundant
- Determine accessibility to the data: authentication and authorization protocols
- Ensure data is documented with sufficient metadata to enable its verification and **reuse**

Share

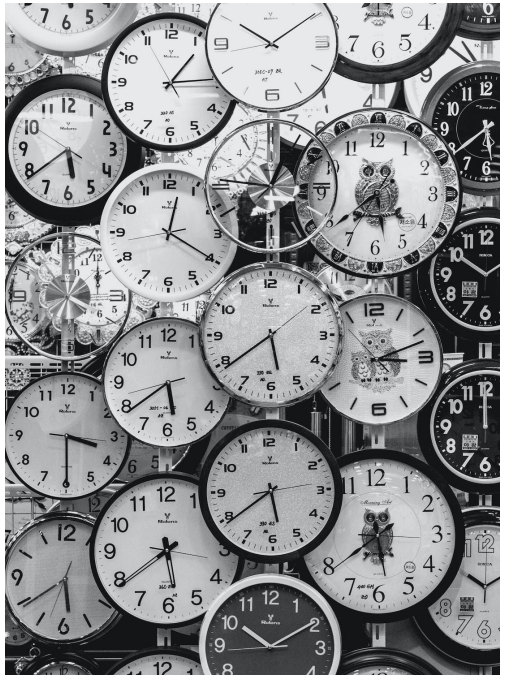
- Determine if the data can be publicly shared (in its entirety or partially) or how access to it can be gained
- Deposit data in a suitable repository, with sufficient metadata to enable **reuse**
  - Indexes metadata and enables searches
  - Has machine accessibility
  - Has authentication and authorization protocols if necessary

Reuse

- Not really a stage
- Data that you reuse should be contemplated at the plan and collect stages
- Making your data reusable is a concern that should permeate the whole data lifecycle

Day-to-Day Data Management

- Take some time each day:
  - To document what you did
    - In an ELN
    - Following metadata standards
    - Using controlled vocabularies
  - To organize your data
  - To document your code
  - To check and if necessary update your DMP



DAY 2 DAY DATA MANAGEMENT COURSE
FOR LIFE SCIENCES

Questions?

DAY 2 DAY DATA MANAGEMENT COURSE
FOR LIFE SCIENCES

Acknowledgements

- ELIXIR Implementation Study "Impact evaluation at Node-level - getting it done"
- ELIXIR-CONVERGE Project “Connect and align ELIXIR Nodes to deliver sustainable FAIR life-science data management services”

BioData.pt elixir PORTUGAL

Ready for BioData Management?

DAY 2 DAY DATA MANAGEMENT COURSE

FOR LIFE SCIENCES

Organization:

Ana Portugal Melo - INESC-ID / BioData.pt

Carolina Costa - IGC / BioData.pt

Daniel Faria - INESC-ID / IST / BioData.pt

Jorge Oliveira - INESC-ID / IST / BioData.pt

@BioData_PT

@BioData.PT

@BioData-PT

www.BioData.pt

inesc id lisboa | GULBENKIAN CIÊNCIA | TÉCNICO LISBOA | BioData.pt

elixir CONVERGE