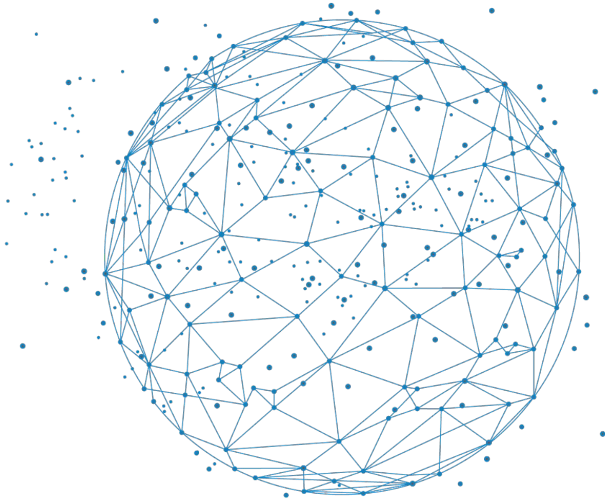




<https://moodle.france-bioinformatique.fr/course/view.php?id=38>



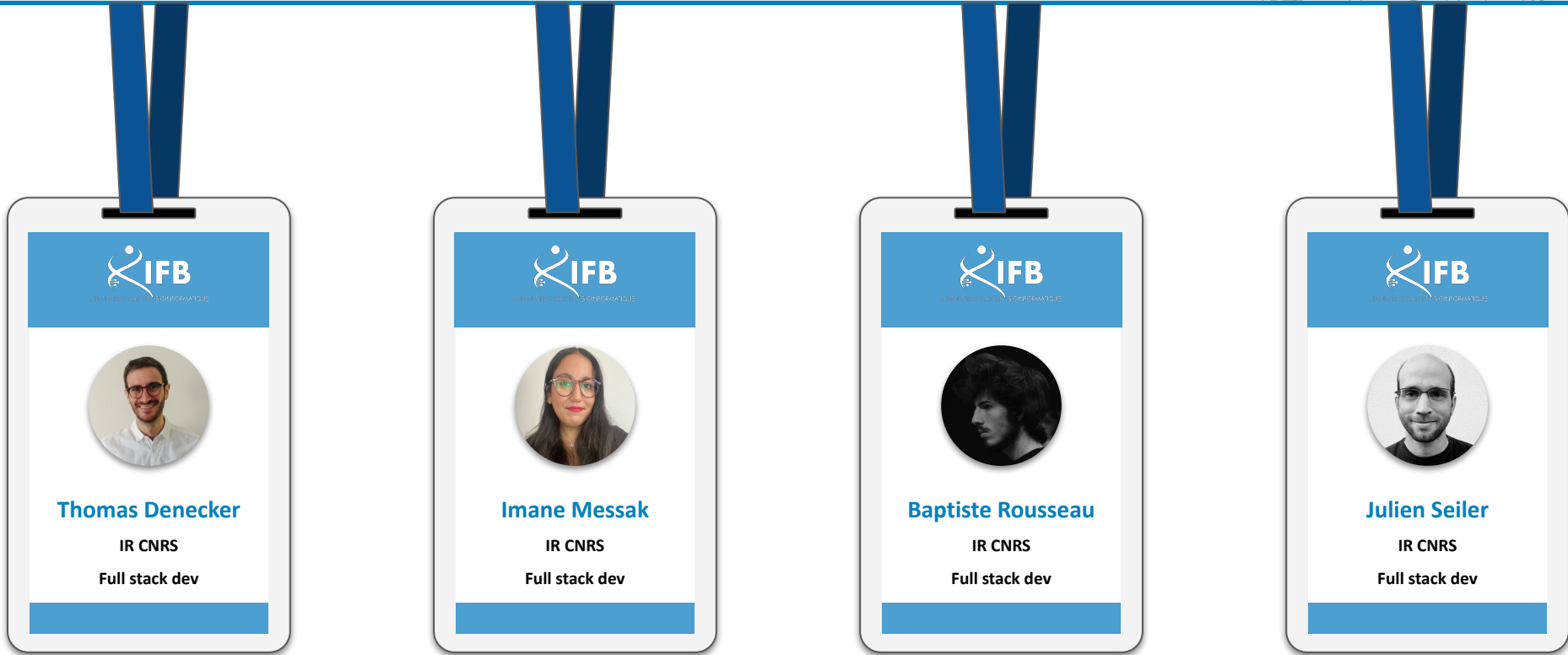
2025 June 5th



FAIR principles applied to bioinformatics

MESSAK Imane

Hello !



Part 1 : Context

- Open Science
- The crisis of reproducibility

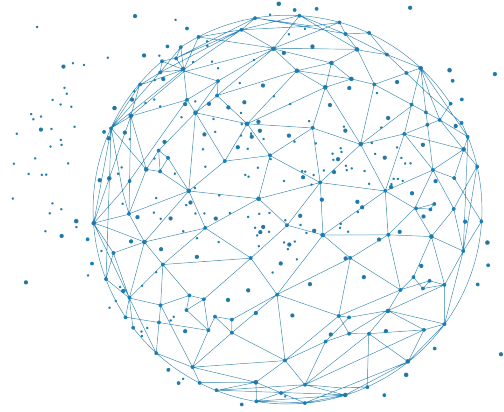
Part 2 : Basic concepts

- The definitions of reproducibility
- FAIR Principles

Part 3 : A solution proposal

- A solution proposal
- FAIR bioinfo training

Open Science, the crisis of reproducibility

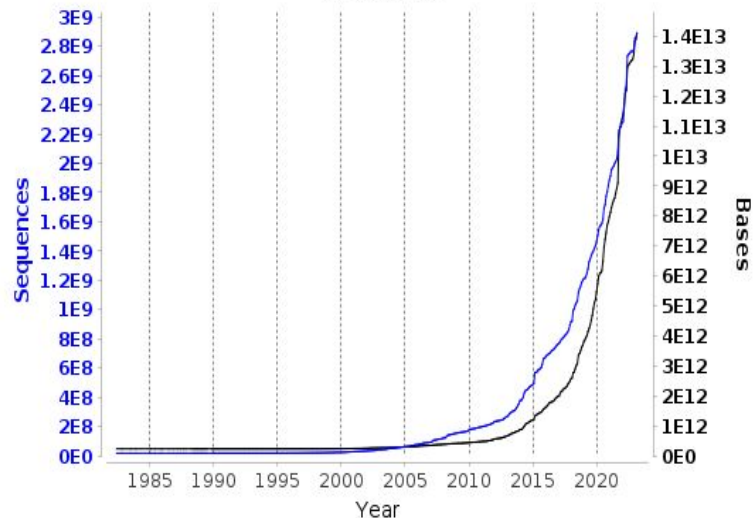


More and more data...



Assembled/annotated sequence growth

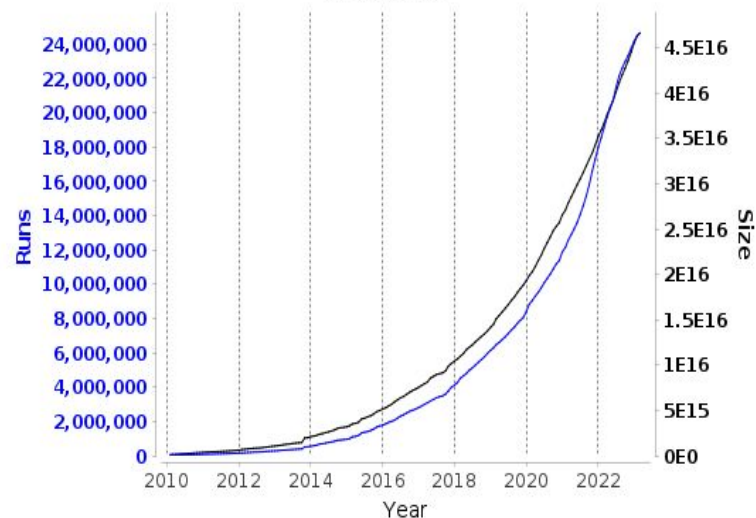
20-Mar-2023



— Sequences (2.9 billions) — Bases (14.1 trillions)

Runs growth

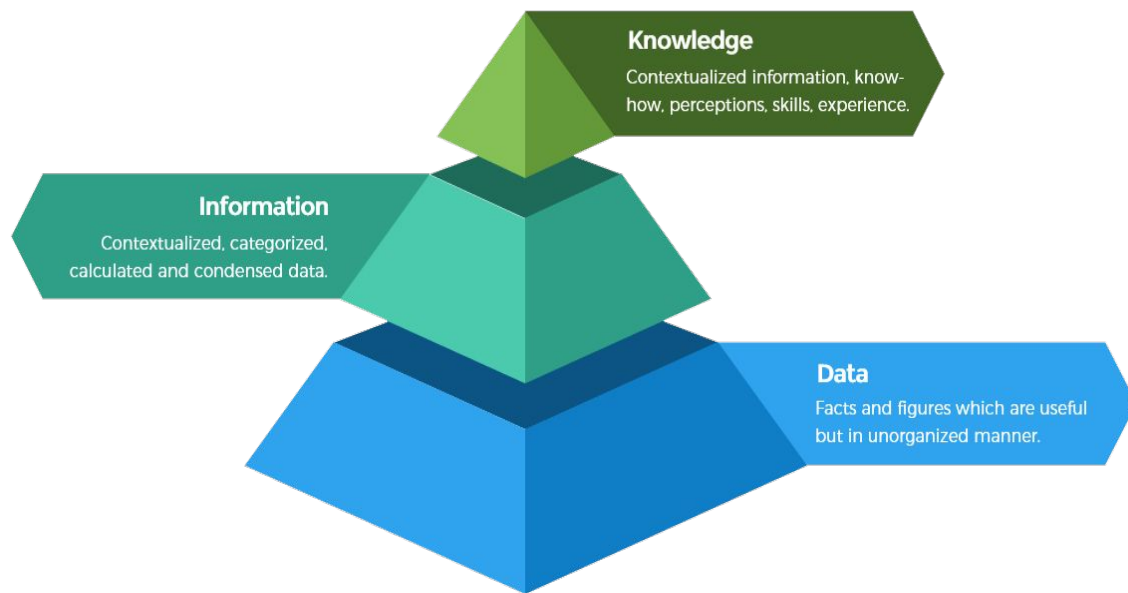
20-Mar-2023



— Runs (24.6 millions) — Size (46.6 petabytes)

<https://www.ebi.ac.uk/ena/browser/about/statistics>

... to create more knowledge



PHPKB, Defining Knowledge, Information, Data
phpkb.com/kb/article/defining-knowledge-information-data-239

Knowledge

- Is subjective
- Has meaning for a specific purpose
- Is processed and understood
- Is not quantifiable, there is no knowledge overload

Information

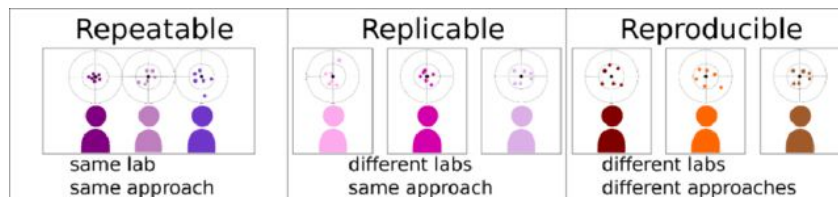
- Should be objective
- Has a meaning
- Is processed
- Is quantifiable, there can be information overload

Data

- Is objective
- Has no meaning
- Is unprocessed
- Is quantifiable, there can be data overload

But how to produce good data?

Repeatable
eplicable
eproductible

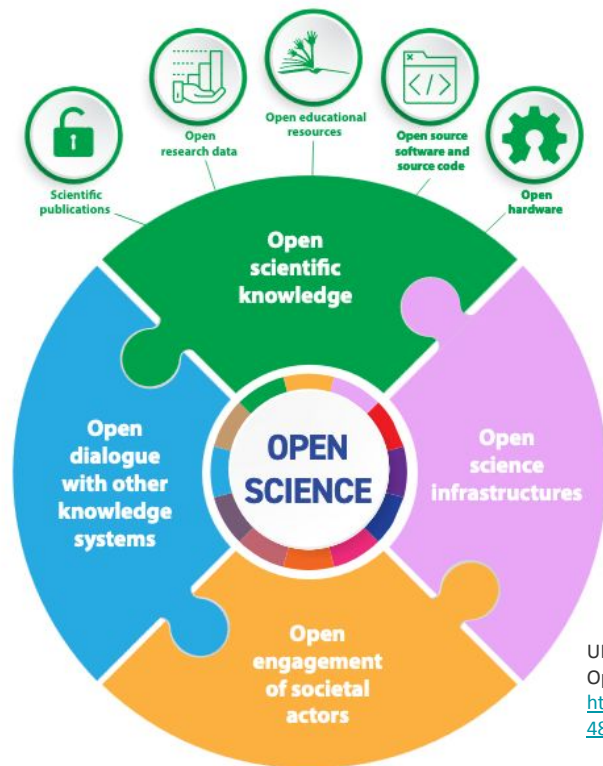


doi.org/10.1038/s41597-020-0486-7

→ From collecting to analyzing data

→ From preserving to re-using data

UNESCO Recommendations



UNESCO Recommendation on
Open Science, nov 2021
<https://unesdoc.unesco.org/ark:/48223/pf0000379949.locale=en>

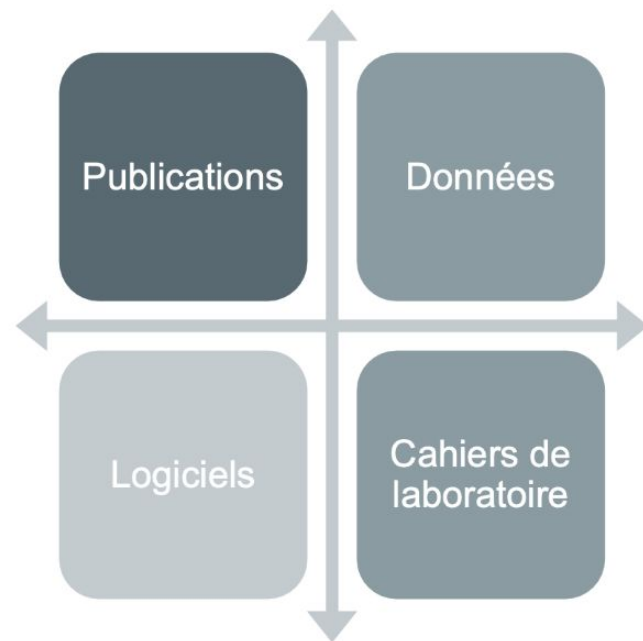
Objective: Make research accessible to everyone

- Not just access to the **knowledge** itself
- The entire process of its **creation** and **dissemination**
- The possibility of **reuse**
- Open **dialogue** with all stakeholders, interdisciplinarity
- Commitment to and from **society**



Ambitions

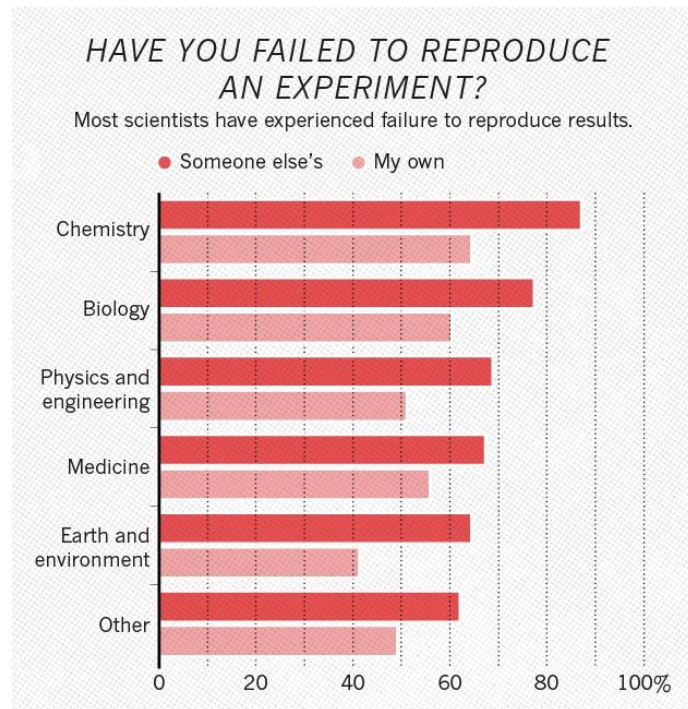
- **Democratize access to knowledge**
- **Make science more cumulative**, strongly supported by data, and more transparent
- **Increase research efficiency** by avoiding duplicated efforts and reusing data or scientific material
- **Promote scientific advancements and innovation**
- Foster public **trust** in science



In biology,

76 %

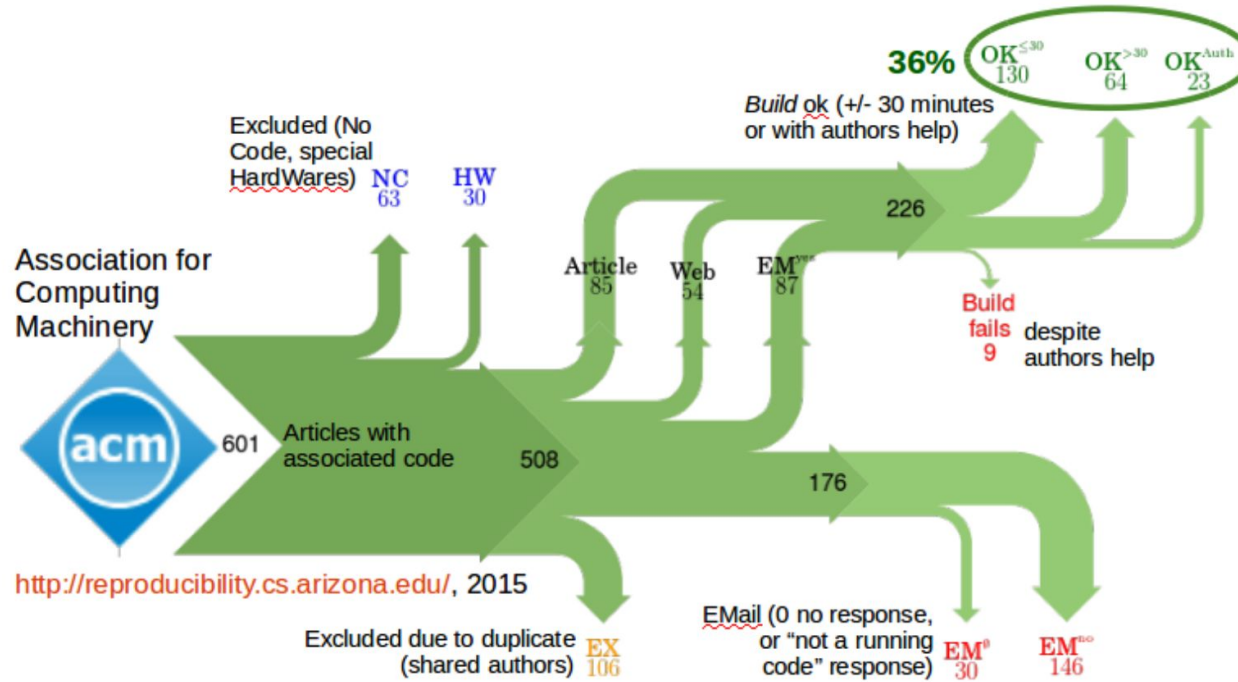
of the researchers surveyed
failed to reproduce results.



Monya Baker, 2016

<https://doi.org/10.1038/533452a>

And in reality? A reproducibility crisis



Collberg et al. 2015

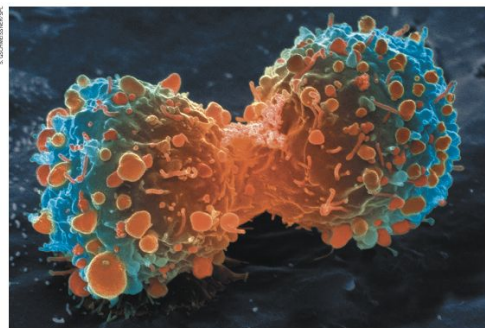
COMMENT

ARAB INFLUENCE Shift expertise to track mutations where they emerge **p.534**

EARTH SYSTEMS Past climates give valuable clues to future warming **p.537**

HISTORY OF SCIENCE Descartes' lost letter tracked using Google **p.540**

OUTRAGE Wylie Vale and an elusive stress hormone **p.542**



Many landmark findings in preclinical oncology research are not reproducible, in part because of inadequate cell lines and animal models.

Raise standards for preclinical cancer research

C. Glenn Begley and Lee M. Ellis propose how methods, publications and incentives must change if patients are to benefit.

Efforts over the past decade to characterize the genetic alterations in human cancers have led to a better understanding of molecular drivers of this complex set of diseases. Although we in the cancer field hoped that this would lead to more effective drugs, historically, our ability to translate cancer research to clinical success has been remarkably low. Sadly, clinical

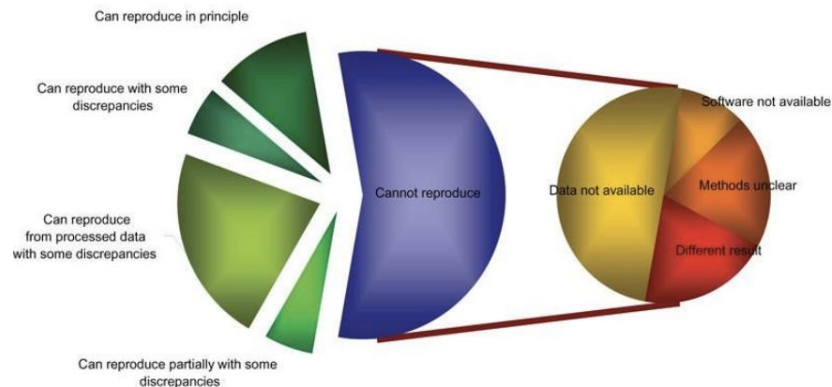
trials in oncology have the highest failure rate compared with other therapeutic areas. Given the high unmet need in oncology, it is understandable that barriers to clinical development may be lower than for other disease areas, and a larger number of drugs with suboptimal preclinical validation will enter oncology trials. However, this low success rate is not sustainable or acceptable, and

investigators must reassess their approach to translating discovery research into greater clinical success and impact. Many factors are responsible for the high failure rate, notwithstanding the inherently difficult nature of this disease. Certainly, the limitations of preclinical tools such as inadequate cancer-cell-line and mouse models make it difficult for even

© 2012 Macmillan Publishers Limited. All rights reserved. 29 MARCH 2012 | VOL 483 | NATURE | 533

Repeatability of published microarray gene expression analyses

John P A Ioannidis¹⁻³, David B Allison⁴, Catherine A Ball⁵, Issa Coulibaly⁴, Xiangqin Cui⁴, Aedin C Culhane^{6,7}, Mario Falchi^{8,9}, Cesare Furlanello¹⁰, Laurence Game¹¹, Giuseppe Jurman¹⁰, Jon Mangion¹¹, Tapan Mehta⁴, Michael Nitzberg⁵, Grier P Page^{4,12}, Enrico Petretto^{11,13} & Vera van Noort¹⁴



Alsheikh-Ali et al. PLoS ONE (2011)
Nekrutenko & Taylor, Nature Genetics (2012)
Begley & Ellis Nature (2012)



An interesting article ...

... but a deceptive M&M

PLOS ONE Publish About Browse Search advanced search

OPEN ACCESS PEER-REVIEWED
RESEARCH ARTICLE

The *Arthrobacter arilaitensis* Re117 Genome Sequence Reveals Its Genetic Adaptation to the Surface of Cheese

Christophe Monnet , Valentin Loux, Jean-François Gibrat, Eric Spinnler, Valérie Barbe, Benoît Vacherie, Frederick Gavory, Edith Goubeyre, Patricia Siguier, Michaël Chandler, Rayda Elleuch, Françoise Irlinger , Tatiana Vallaeys

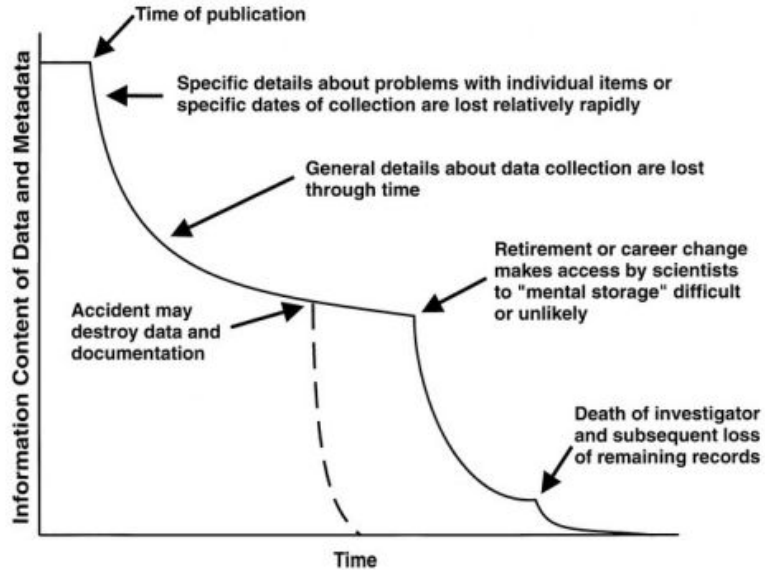
Published: November 24, 2010 • <https://doi.org/10.1371/journal.pone.0015489>

91 Save	84 Citation
8,442 View	2 Share

The genome sequences of *Arthrobacter aureus* TC1 [24] (accession number CP000474), *Arthrobacter* sp. strain FB24 (accession number CP000454), a strain isolated from a xylene and chromate enriched soil microcosm [25], and *A. chlorophenolicus* A6 (accession number CP001341), a strain capable of degrading high concentrations of 4-chlorophenol [26], were used for comparative genomic analyses. The sequence data from the strains FB24 and A6 were produced by the US Department of Energy Joint Genome Institute (<http://www.jgi.doe.gov/>) in collaboration with the user community. Genome comparisons were performed using Origami, an in-house tool developed for microbial genome comparison. Orthologs were defined as reciprocal best hits with an e-value lower than 10^{-3} . Transposases were excluded from the analysis. Core genes were defined as orthologs shared between the four *Arthrobacter* strains. Synteny was studied using an in-house developed tool, Align, using dynamic programming to search conserved gene trains allowing gaps and “mismatches” (homology relation instead of orthology). Circular representation of the genome was produced using the Circos software [27].



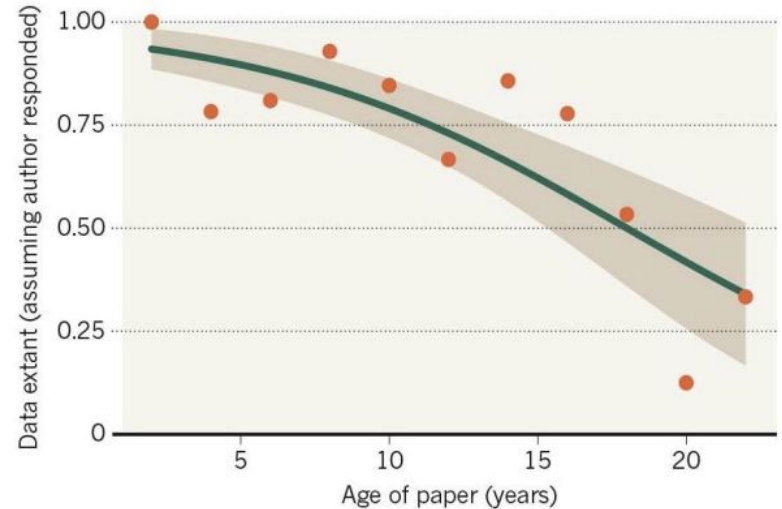
Scriberia



[Michener \(1997\)](#)

MISSING DATA

As research articles age, the odds of their raw data being extant drop dramatically.



[Gibney \(2013\)](#)



Cost of not having FAIR research data

Cost-Benefit analysis for FAIR research data

Written by PwC EU Services
March – 2018

Research and
Innovation

Likely cost of not having FAIR research data

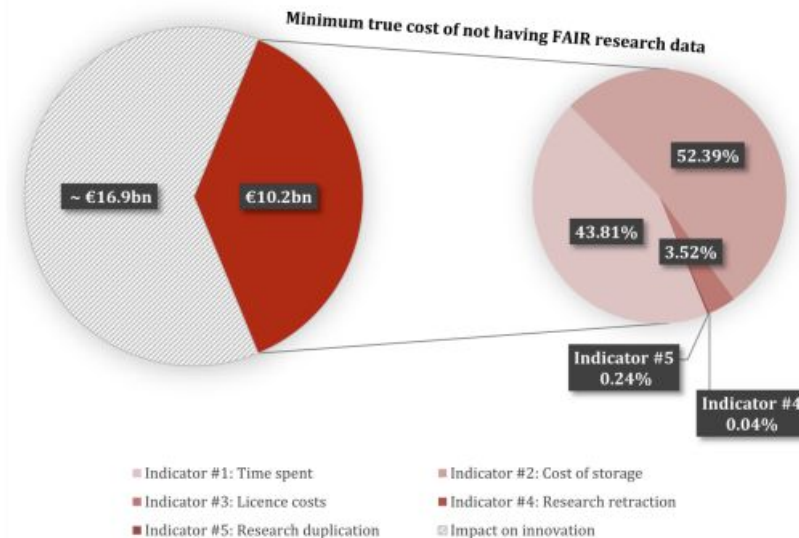
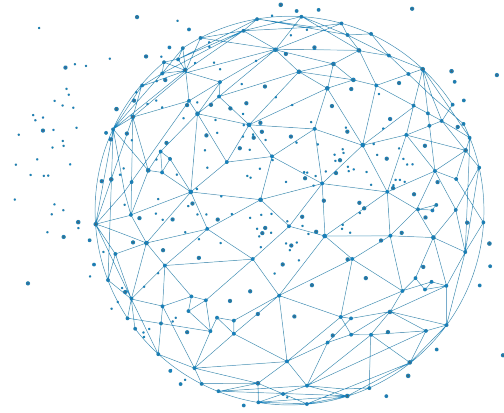


Figure 5: Cost breakdown

€10.2bn every year !!!

The basic concepts of reproducibility





*Reproducible research, Repeatability,
Replicability, Reproducibility,
Replication...*

Overlapping concepts

⇒ many definitions!

Definitions from the



(2016) :

Repeatability: same team, same experimental design

Reproducibility: different teams, same experimental design

Replicability: different teams, different experimental designs

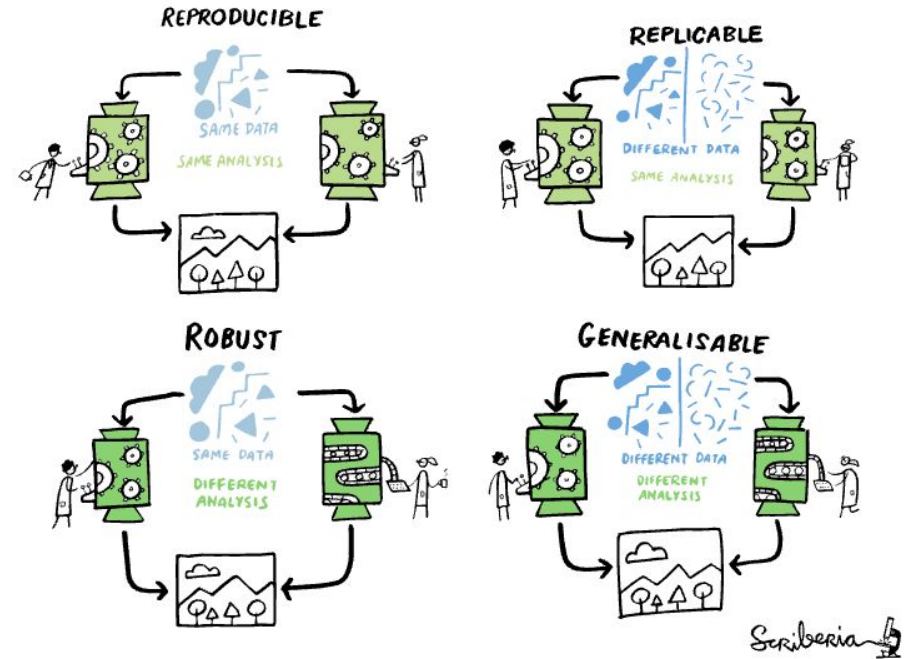
https://www.researchgate.net/publication/323118701_Terminologies_for_Reproducible_Research

<https://www.acm.org/publications/policies/artifact-review-and-badging-current>

Matrix for the Reproducibility of Whitaker (2017):

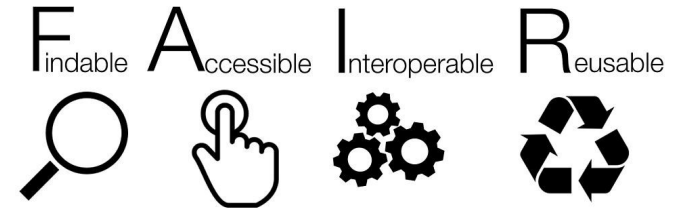
		Data	
		Same	Different
Code	Same	Reproducible	Replicable
	Different	Robust	Generalisable

<https://doi.org/10.6084/m9.figshare.5443201.v1>, Slide 7



The Turing Way project illustration by Scriberia. Used under a CC-BY 4.0 licence. DOI: [The Turing Way Community & Scriberia \(2024\)](#).

		Data	
		Same	Different
Code	Same	Reproducible	Replicable
	Different	Robust	Generalisable



In brief, FAIR data should be:

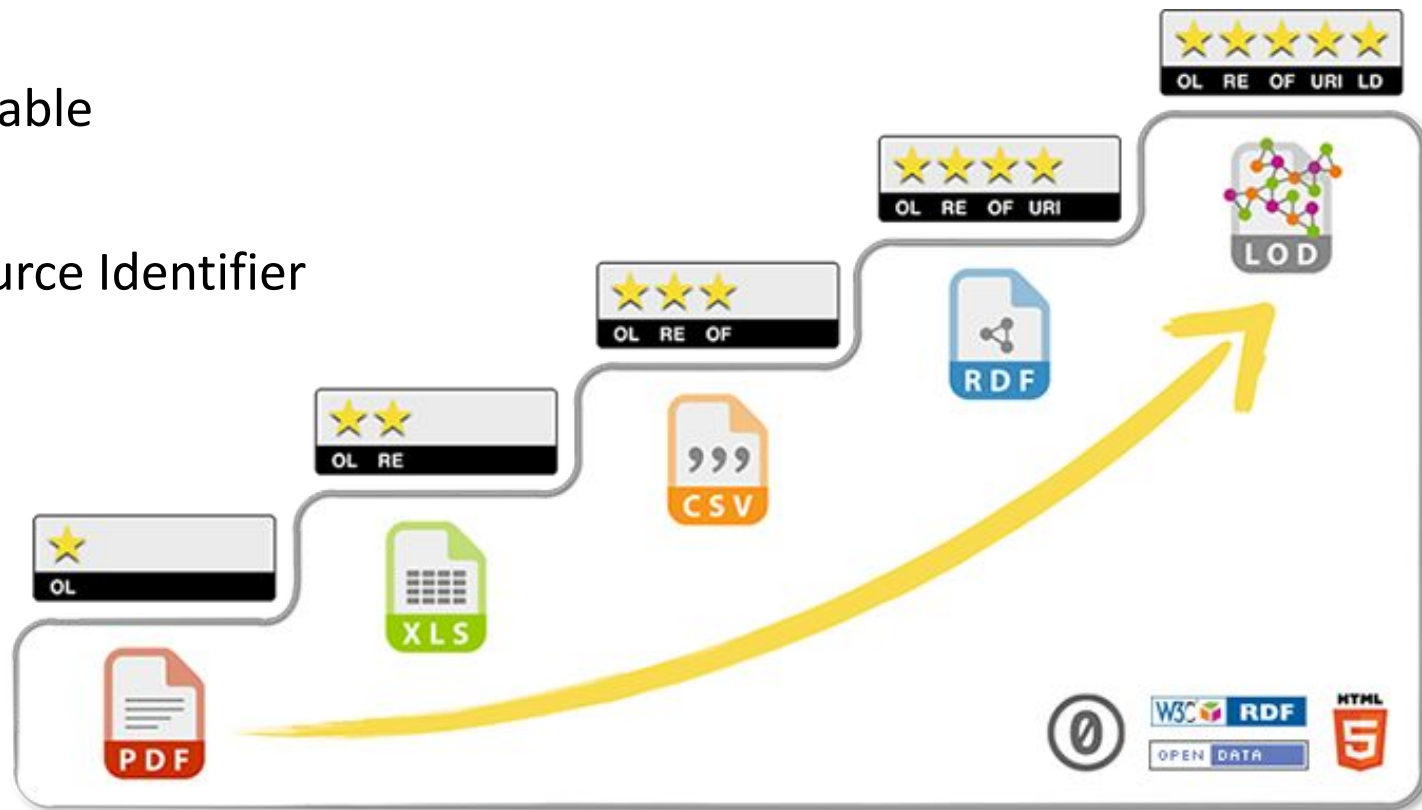
-  **Findable:** The first step in (re)using data is to find it! Descriptive metadata (information about the data such as keywords) is essential.
-  **Accessible:** Once the user finds the data and software they need to know how to access it. Data could be openly available but it is also possible that authentication and authorisation procedures are necessary.
-  **Interoperable:** Data needs to be integrated with other data and interoperate with applications or workflows.
-  **Reusable:** Data should be well-described so that they can be used, combined, and extended in different settings.



The Turing Way project illustration by Scriberia. Used under a CC-BY 4.0 licence. DOI: [The Turing Way Community & Scriberia \(2024\)](#).

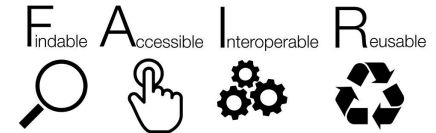
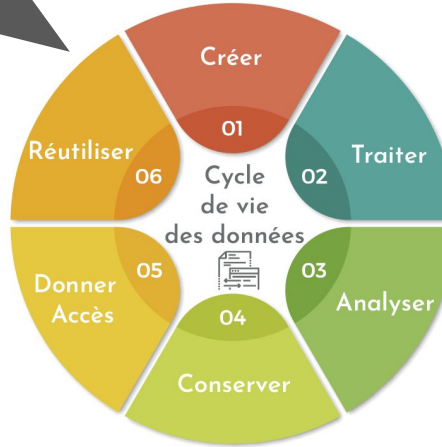
- ★ Open License
- ★ machine REadable
- ★ Open Format
- ★ Uniform Resource Identifier
- ★ Linked Data

Bernes-Lee





		Data	
		Same	Different
Code	Same	Reproducible	Replicable
	Different	Robust	Generalisable



<https://science-ouverte.univ-artois.fr/es-donnees-de-la-recherche/>

?



Like data, software is backed up, but it's not "just" data. They are **living** and **complex** entities.

	Data	Software
are	facts, observations	creations
produce	evidence	an executable tool
change?	no (unless underlying hardware changes)	yes: continuous maintenance & updates
lifespan	long-term necessary when the experiment cannot be reproduced (cost of collection or validation)	short: often rebuilt to use other software (complex dependencies), replaceable (another performs the task better)

Katz DS, Niemeyer KE, Smith AM, Anderson WL, Boettiger C, Hinsén K, Hooft R, Hucka M, Lee A, Löffler F, Pollard T, Rios F. 2016. Software vs. data in the context of citation. PeerJ Preprints 4:e2630v1 <https://doi.org/10.7287/peerj.preprints.2630v1>



Findable (for humans and machines)

- Unique identifier (e.g., a DOI with **Zenodo**)
- Metadata describing the analysis and tools; they are FAIR, searchable, and indexable (e.g., **README**)
- Available on a forge (**GitHub**, **GitLab**) and in an archive repository (**SWH**)



Accessible

- **Retrievable** using a standardized protocol
- **Open**, free and universally implementable
- Metadata accessible, even when the software is no longer available (**SWH**)



Interoperable

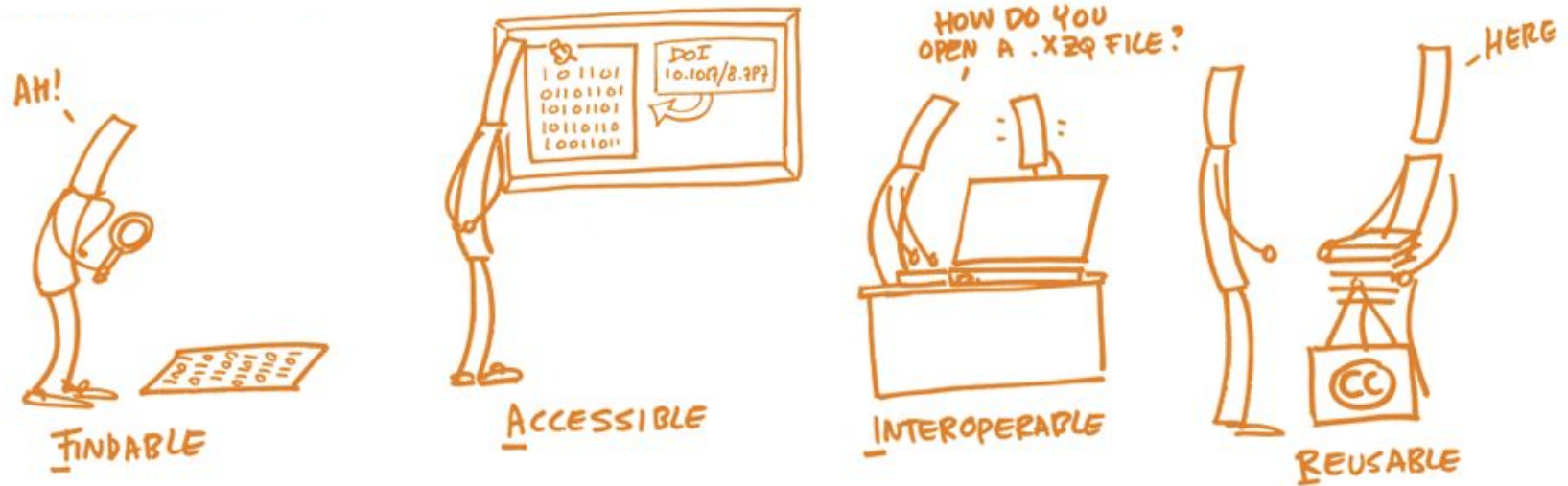
- Use a **formal, accessible, shared** and broadly **applicable language**
- Tool cooperation (**snakemake, conda, and docker**) locally or on a server (cloud or cluster)



Reusable

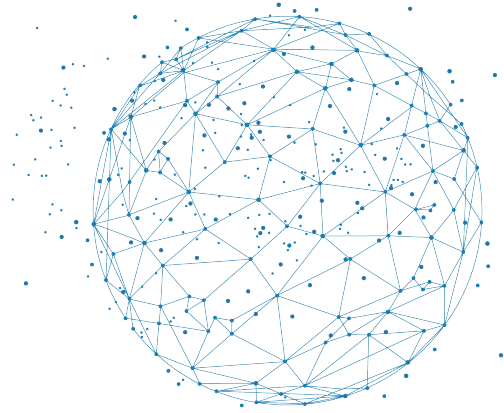
- Clear **licensing** and **access rights** (according to employer, DPO)
- Have a **detailed provenance**
- Follows **community standards**

In short, it's very close!



[OpenAire](#)

A suggested solution to make your project reproducible



Reminder : the road to our goal



The Turing Way project illustration by Scriberia. Used under a CC-BY 4.0 licence. DOI: [The Turing Way Community & Scriberia \(2024\)](#).



The pillars of reproducibility

Reproducibility

Versioning

Sharing

Archiving

**Create a
virtual
environment**

**Installing
tools using
a package
manager**

**Create
analysis
scripts**

**Offloading
analysis to a
server**

**Portability
of results
exploration**

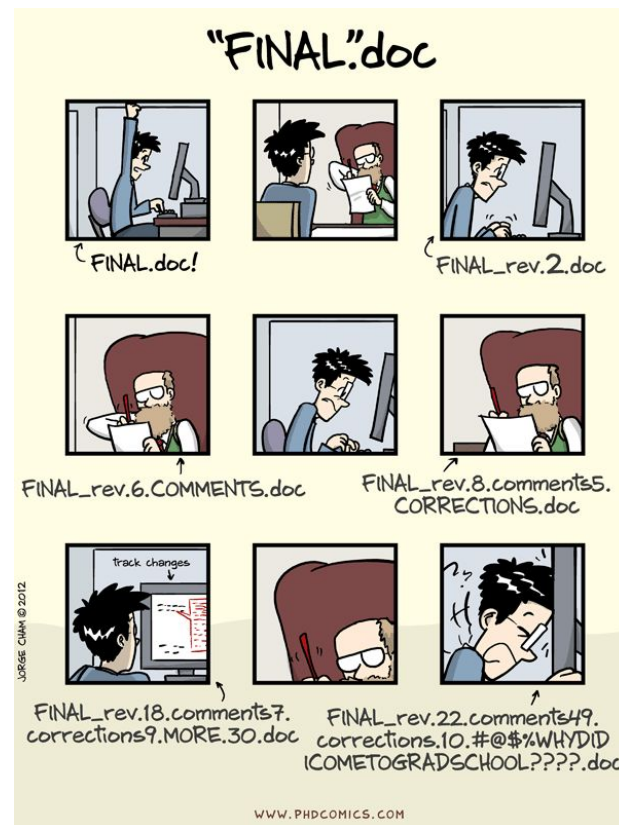
**Editing the
analyses
reports**

based on

The five pillars of computational reproducibility: Bioinformatics and beyond
Mark Ziemann, Pierre Poulain, Anusuiya Bora, OSF preprint, 2023

Why ?

- Having the right version of the code
- Temporal insight
- Openness to the community



Pillar 1 - Versioning, sharing, and archiving code



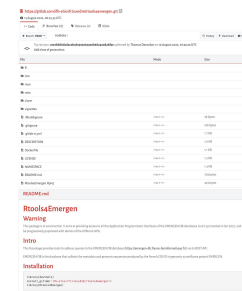
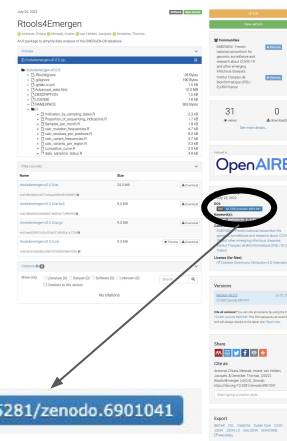
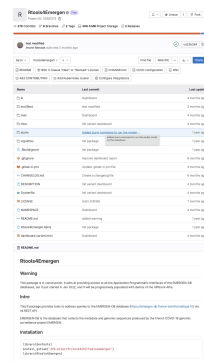
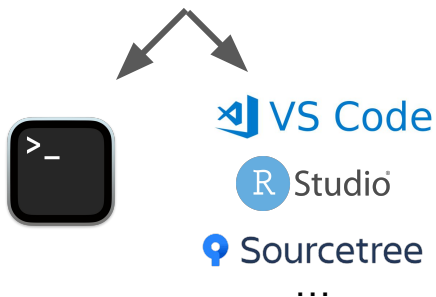
Versioning

Versioning /
sharing

Sharing

Archiving

how ?

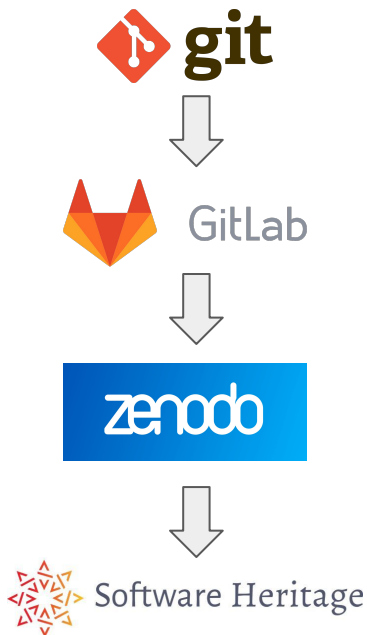




Why ?

- Freezing the environment
- Sharing the environment

How ?



Advantages / Disadvantages

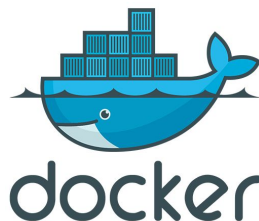
- + Code backup
- + Easy for sharing
- + Automatic version management
- Not easy for novices



Why ?

- Freezing the environment
- Sharing the environment

How ?



Advantages / Disadvantages

- + Fast and lightweight
- + Portable
- + Easy to share and deploy
- With an up-to-date system
- [docker] Accepted in your structure?



Why ?

- Having the right version of the tools used
- Installing them simply

How ?

CONDA



PIXI

Advantages / Disadvantages

- + Simple manager to install
- + Simple package installation
- + Version management
- Can be heavy (miniconda solution)
- Missing packages (R)



Why ?

- Having a reproducible analysis script
- Not redoing what has already been done
- Parallelizing

How ?



nextflow

Advantages / Disadvantages

- + Job management
- + Powerful and fast
- + Capable of using Conda environments
- + Parallelizable
- A logic to learn (syntax less simple than shell script)



Why ?

- Controlled environment
- Offloading analysis

How ?



Advantages / Disadvantages

- + Easy to set up
- + Increased power (cloud or cluster)
- + For everyone
- Not easy for novices
- Attention to sensitive data





Why ?

- Making exploration simple
- Easy to share

How ?



Advantages / Disadvantages

- + Portable (HTML)
- + Accessible everywhere
- + Interactive (configurable, dynamic graphs, etc.)
- Mix of languages



Why ?

- Having a record of the analysis (date, time, parameters, etc.)
- Storing tool versions

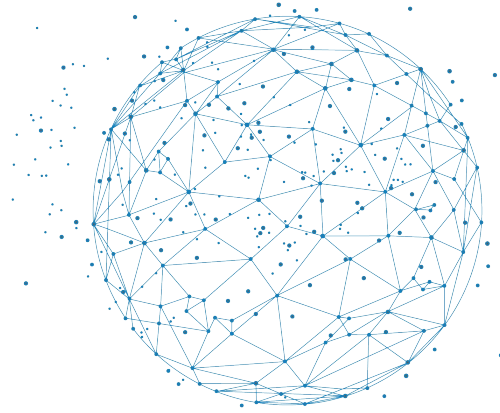
How ?



Advantages / Disadvantages

- + Simple syntax (Markdown)
- + Sharing (PDF, HTML, etc.)
- Rare visualization problems using $\text{\LaTeX}$

Conclusion





1. Data

- a. Do I do even it if the data are not 100% FAIR?
- b. How to manage large volumes of data?
- c. How to manage data and metadata updates?

2. Code

- a. How to ensure that the code will always be accessible?
- b. Is it acceptable to make adaptations? To what extent?
- c. Provide all the code? (valorization, creation of a start-up, etc.)

3. Computation time (days, months, years)

4. Skill and sensitivity

- a. Willingness but technical inability to do so
- b. "Why bother, it's useless"
- c. too long

5. The coverage

- a. Should everything be made reproducible?

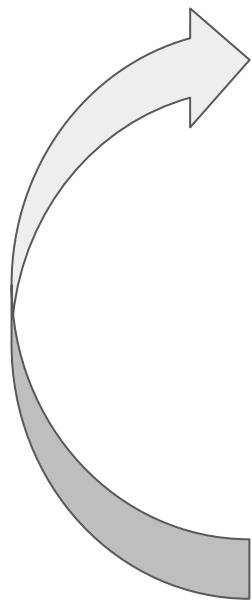
6. When to do it?

- a. At the beginning? But what if it doesn't work?
- b. At the end?

A genuine reflection on the reproducibility of analyses
(balance, time, scope, etc.)

Proposal of a solution that helps make any analysis protocol reproducible.

Reproducibility is an added value for your work !



FAIR raw data

+

FAIR scripts/protocols

=

FAIR processed data

And you? How many pillars support your reproducibility?

Reproducibility

**Versioning
Sharing
Archiving**



**Create a
virtual
environment**



**Installing
tools using
a package
manager**



**Create
analysis
scripts**



**Offloading
analysis to a
server**



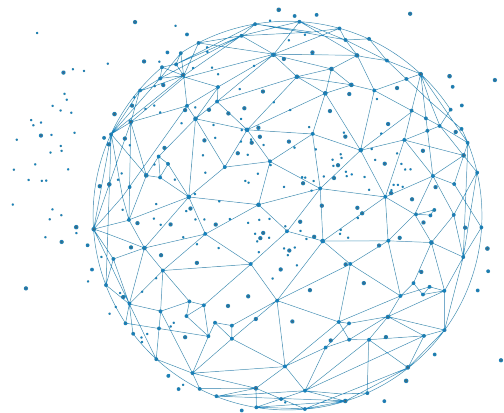
**Portability
of results
exploration**



**Editing the
analyses
reports**

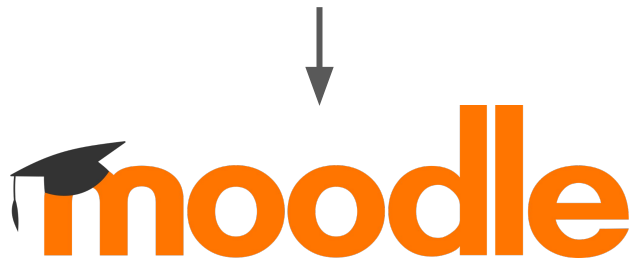


Trainings by the IFB



IFB trainings on the theme of FAIR

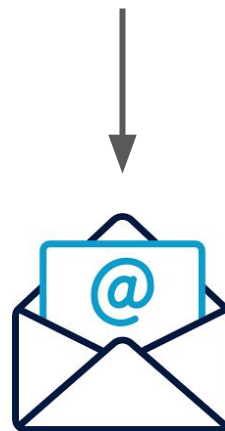
1. FAIR principles for research data management in life sciences (**FAIRdata**)
2. The FAIR principles in a Bioinformatics project (**FAIRbioinfo**)



<https://moodle.france-bioinformatique.fr/>

More training sessions to come

Stay Tuned!





Thank you for your
attention !



INSTITUT FRANÇAIS DE BIOINFORMATIQUE

