

Essay

Why Most Published Research Findings Are False

John P.A. Ioannidis

Summary

There is increasing concern that most current published research findings are false. The probability that a research claim is true may depend on study power and bias, the number of other studies on the same question, and, importantly, the ratio of true to no relationships among the relationships probed in each scientific field. In this framework, a research finding is less likely to be true when the studies conducted in a field are smaller; when effect sizes are smaller; when there is a greater number and lesser preselection of tested relationships; where there is greater flexibility in designs, definitions, outcomes, and analytical modes; when there is greater financial and other interest and prejudice; and when more teams are involved in a scientific field in chase of statistical significance. Simulations show that for most study designs and settings, it is more likely for a research claim to be false than true. Moreover, for many current scientific fields, claimed research findings may often be simply accurate measures of the prevailing bias. In this essay, I discuss the implications of these problems for the conduct and interpretation of research.

factors that influence this problem and some corollaries thereof.

Modeling the Framework for False Positive Findings

Several methodologists have pointed out [9–11] that the high rate of nonreplication (lack of confirmation) of research discoveries is a consequence of the convenient, yet ill-founded strategy of claiming conclusive research findings solely on the basis of a single study assessed by formal statistical significance, typically for a *p*-value less than 0.05. Research is not most appropriately represented and summarized by *p*-values, but, unfortunately, there is a widespread notion that medical research articles

It can be proven that most claimed research findings are false.

should be interpreted based only on *p*-values. Research findings are defined here as any relationship reaching formal statistical significance, e.g., effective interventions, informative predictors, risk factors, or associations. “Negative” research is also very useful.

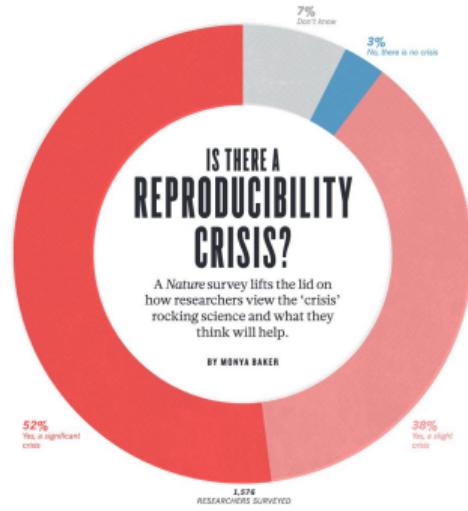
is characteristic of the field and can vary a lot depending on whether the field targets highly likely relationships or searches for only one or a few true relationships among thousands and millions of hypotheses that may be postulated. Let us also consider, for computational simplicity, circumscribed fields where either there is only one true relationship (among many that can be hypothesized) or the power is similar to find any of the several existing true relationships. The pre-study probability of a relationship being true is $R/(R+1)$. The probability of a study finding a true relationship reflects the power $1-\beta$ (one minus the Type II error rate). The probability of claiming a relationship when none truly exists reflects the Type I error rate, α . Assuming that *c* relationships are being probed in the field, the expected values of the 2×2 table are given in Table 1. After a research finding has been claimed based on achieving formal statistical significance, the post-study probability that it is true is the positive predictive value, PPV. The PPV is also the complementary probability of what Wacholder et al. have called the false positive report probability [10]. According to the 2×2 table, one gets $PPV = (1-\beta)R/(1-\beta)R + \alpha(1-R)$.

Crisis elements

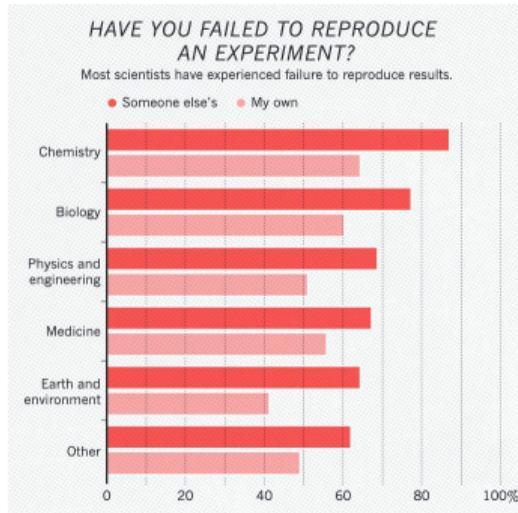
- Highlighted around 2005
- Since 2010 more articles related to the non reproducibility
- Medicine is one of the most impacted discipline

Reproducibility crisis

2016

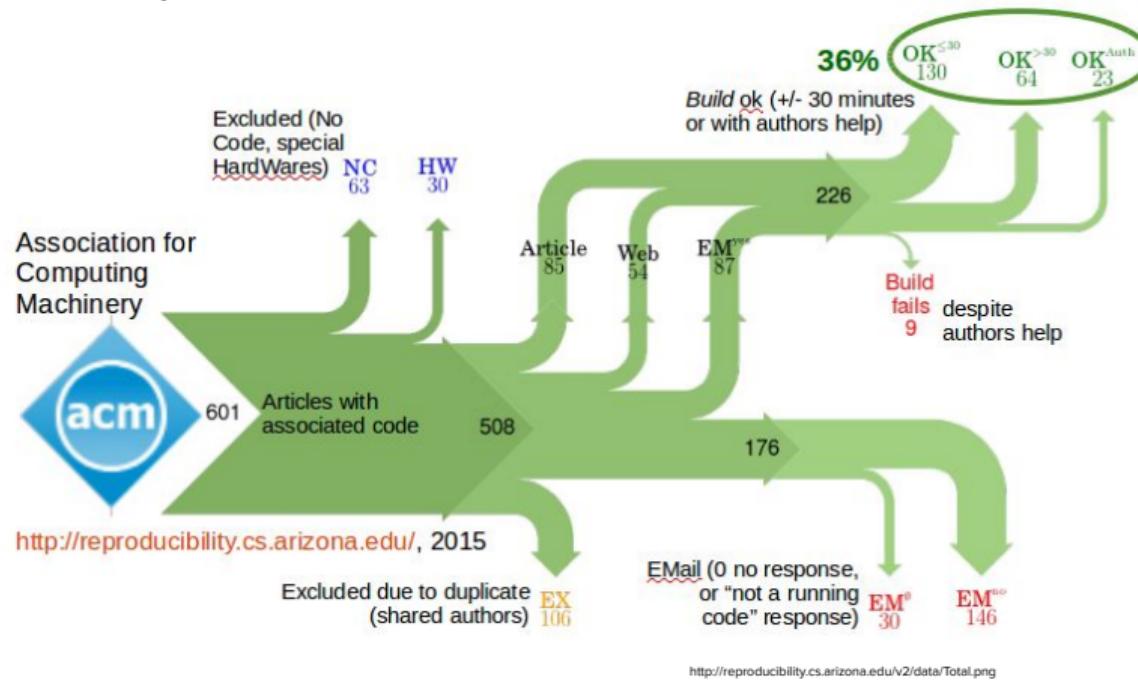


Baker, M. 1,500 scientists lift the lid on reproducibility. *Nature* 533, 452–454 (2016). <https://doi.org/10.1038/533452a>



4

Also in computer sciences



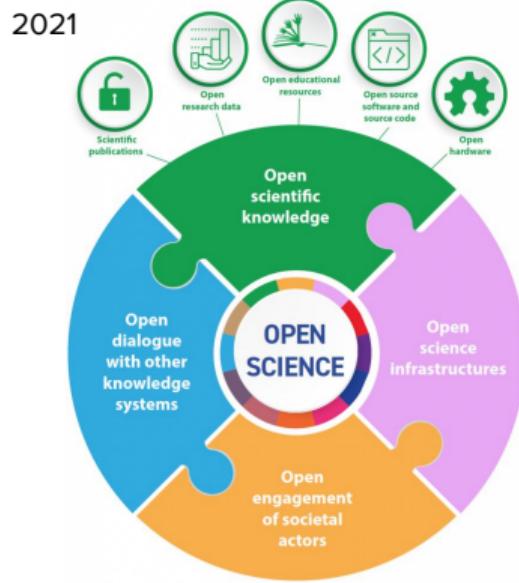
Long term negative impact of retracted papers

Article	Year of retraction	Citing Articles before retraction	Citing Articles after retraction	Total cites (journals indexed by Web of Science)
1. Primary Prevention of Cardiovascular Disease with a Mediterranean Diet. N ENGL J MED; APR 2013 . Estruch R, et al.	2018	1919	816	2735
2. Ileal-lymphoid-nodular hyperplasia, non-specific colitis, and pervasive developmental disorder in children. LANCET; FEB 28 1998 . Wakefield AJ, et al.	2010	642	867	1509
3. Visfatin: A protein secreted by visceral fat that mimics the effects of insulin. SCIENCE; JAN 2005 . Fukuhara A, et al.	2007	232	1192	1424
4. An enhanced transient expression system in plants based on suppression of gene silencing by the p19 protein of tomato bushy stunt virus. PLANT J; MAR 2003 . Voinnet O, et al.	2015	896	375	1271
5. Lysyl oxidase is essential for hypoxia-induced metastasis. NATURE; APR 2006 . Erler JT, et al.	2020	977	81	1058

Retraction Watch : Top 10 most highly cited retracted papers
<https://retractionwatch.com/the-retraction-watch-leaderboard/top-10-most-highly-cited-retracted-papers/>

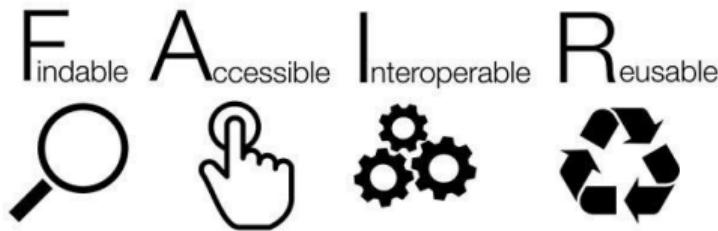
6

A way out: Open science and FAIR principles



Graphic on page 11. [UNESCO Recommendation on Open Science](#). CC BY IGO 3.0 C. Green

2016



Wilkinson, M., Dumontier, M., Aalbersberg, I. et al. The FAIR Guiding Principles for scientific data management and stewardship. *Sci Data* **3**, 160018 (2016).
<https://doi.org/10.1038/sdata.2016.18>

7

FAIR history

- Born in 2016 with *The FAIR Guiding Principles for scientific data management and stewardship*
- How to build, stock, share, use and publish data
- Make criteria to better use our data

. <https://doi.org/10.1038/sdata.2016.18>

SCIENTIFIC DATA

Amended: Addendum

OPEN

SUBJECT CATEGORIES

- » Research data
- » Publication characteristics

Received: 10 December 2015

Accepted: 12 February 2016

Published: 15 March 2016

Comment: The FAIR Guiding Principles for scientific data management and stewardship

Mark D. Wilkinson et al.[#]

There is an urgent need to improve the infrastructure supporting the reuse of scholarly data. A diverse set of stakeholders—representing academia, industry, funding agencies, and scholarly publishers—have come together to design and jointly endorse a concise and measurable set of principles that we refer to as the FAIR Data Principles. The intent is that these may act as a guideline for those wishing to enhance the reusability of their data holdings. Distinct from peer initiatives that focus on the human scholar, the FAIR Principles put specific emphasis on enhancing the ability of machines to automatically find and use the data, in addition to supporting its reuse by individuals. This Comment is the first formal publication of the FAIR Principles, and includes the rationale behind them, and some exemplar implementations in the community.

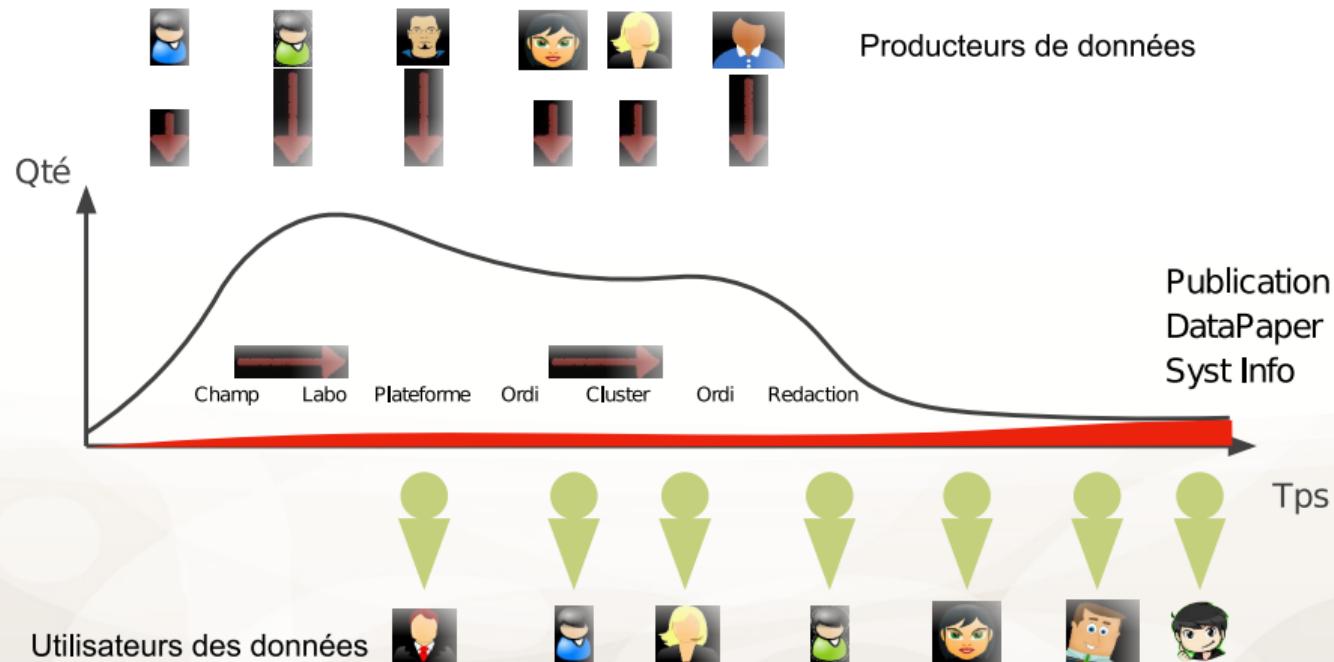
. <https://doi.org/10.1038/sdata.2016.18>

Apply FAIR TO

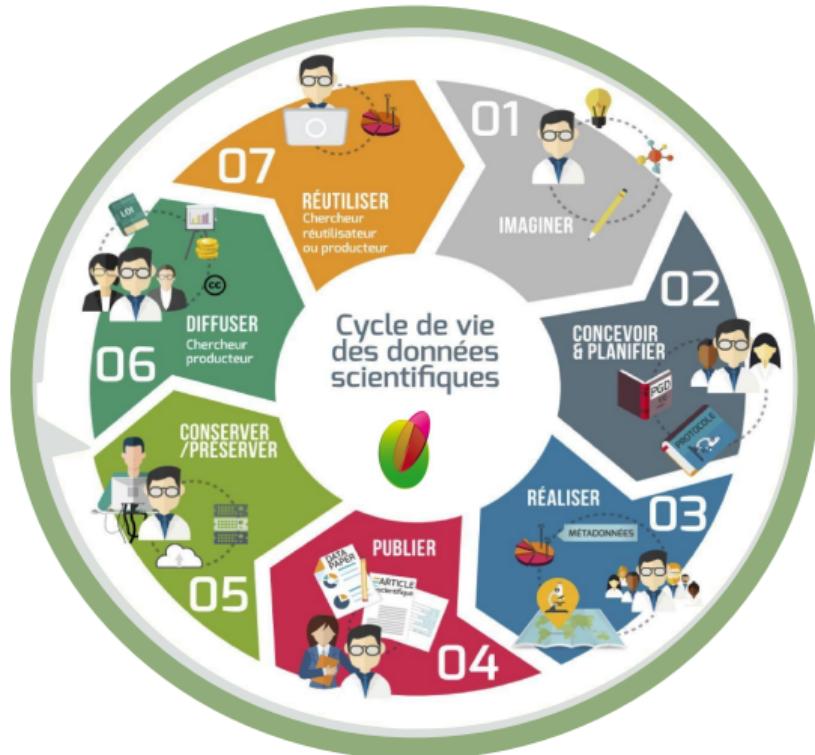
■ Your DATA

- Data lifecycle
- Data Management Plan (DMP)
- Metadata
- Data storage

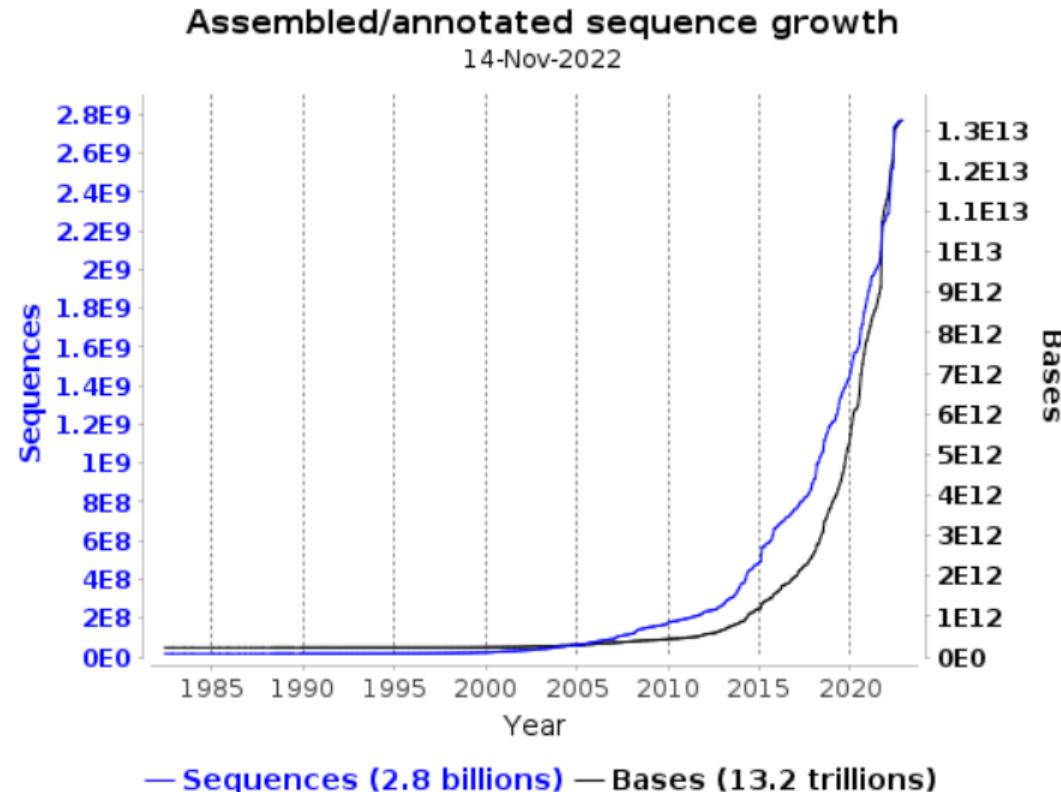
Focus on the data lifecycle



Focus on the data lifecycle



Focus on the data lifecycle



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

Focus on the data exchange

Transfert de vos données de recherche



Comment transmettre vos données ?





Cost of not having FAIR research data

Cost-Benefit analysis for FAIR research data

Apply FAIR TO

- Your DATA

- Data lifecycle
- Data Management Plan (DMP)
- Metadata
- Data storage

- Your scripts, environment...

- Objective of this training

FAIR principles

F
indable



By 維基小霸王 - Own work, CC BY-SA 4.0.
<https://commons.wikimedia.org/w/index.php?curid=88894774>

PID
Repository

8

<https://doi.org/10.1038/sdata.2016.18>

404

This is not the
web page you
are looking for.



To be Findable

- (meta)data are assigned a globally unique and persistent identifier
- data are described with rich metadata
- metadata clearly and explicitly include the identifier of the data it describes
- (meta)data are registered or indexed in a searchable resource

FAIR principles

F
indable



By 維基小霸王 - Own work, CC BY-SA 4.0,
<https://commons.wikimedia.org/w/index.php?curid=88894774>

A
ccessible



<https://nitsfirstworldproblems.tumblr.com/post/147555650875/i-can-t-reach-the-top-shelves-of-the-kitchen>

PID
Repository

Protocols
(free, open, auth.)

9

<https://doi.org/10.1038/sdata.2016.18>

To be Accessible

- (meta)data are retrievable by their identifier using a standardized communication protocol
- the protocol is open, free, and universally implementable
- the protocol allows for an authentication and authorization procedure, where necessary
- metadata are accessible, even when the data are no longer available

FAIR principles

Findable



By 維基小霸王 - Own work, CC BY-SA 4.0,
<https://commons.wikimedia.org/w/index.php?curid=88894774>

Accessible



<https://nitsfirstworldproblems.tumblr.com/post/147555650875/i-can-t-reach-the-top-shelves-of-the-kitchen>

Interoperable



By Unknown author - Popular Science Monthly Volume 88, Public Domain
<https://commons.wikimedia.org/w/index.php?curid=22614407>

PID
Repository

Protocols
(free, open, auth.)

Standards
(format, vocabulary)

10
<https://doi.org/10.1038/sdata.2016.18>

To be Interoperable

- (meta)data use a formal, accessible, shared, and broadly applicable language for knowledge representation.
- (meta)data use vocabularies that follow FAIR principles
- (meta)data include qualified references to other (meta)data

FAIR principles

Findable



By 維基小霸王 - Own work, CC BY-SA 4.0,
<https://commons.wikimedia.org/w/index.php?curid=88894774>

Accessible



<https://nillsfirstworldproblems.tumblr.com/post/147555650875/i-can-t-reach-the-top-shelves-of-the-kitchen>

Interoperable



By Unknown author - Popular Science Monthly Volume 88, Public Domain
<https://commons.wikimedia.org/w/index.php?curid=22614407>

Reusable



By Sun Ladder - Own work, CC BY-SA 3.0,
<https://commons.wikimedia.org/w/index.php?curid=5746428>

PID
Repository

Protocols
(free, open, auth.)

Standards
(format, vocabulary)

Metadata
License
Origin

11
<https://doi.org/10.1038/sdata.2016.18>

To be Reusable

- meta(data) are richly described with a plurality of accurate and relevant attributes
- (meta)data are released with a clear and accessible data usage license
- (meta)data are associated with detailed provenance
- (meta)data meet domain-relevant community standard

To be Reusable

A point on the licences

- number different types of licences from proprietary to totally free AND open
- more than 50 different kinds
- No licence = authors' rights
- France = support for free and open data

To be Reusable

A point on the licences

- number different types of licences from proprietary to totally free AND open
- more than 50 different kinds
- No licence = authors' rights
- France = support for free and open data

Recherche dans GitHub par type de licence

Vous pouvez filtrer les référentiels en fonction de leur licence ou famille de licences à l'aide du qualificateur `license` et du mot clé de licence exact :

Licence	Mot clé de licence
Academic Free License v3.0	afl-3.0
Licence Apache 2.0	apache-2.0
Licence Artistic 2.0	artistic-2.0
Licence logicielle Boost 1.0	bsl-1.0
Licence BSD « simplifiée » à 2 clauses	bsd-2-clause
Licence BSD « nouvelle » ou « révisée » à 3 clauses	bsd-3-clause

To be Reusable

A point on the licences

- number different types of licences from proprietary to totally free AND open
- more than 50 different kinds
- No licence = authors' rights
- France = support for free and open data

The screenshot shows the homepage of the Joinup Licensing Assistant. At the top left is the European Commission logo. The top navigation bar includes links for "Interoperable Europe", "Interoperability Solutions", "Sign in", and "Get started". A search icon is also present. The main header features the text "Joinup Licensing Assistant" over a background of binary code. Below the header, there's a diagram illustrating the licensing process: a circular icon with a person head, a circular icon with a document and a star, a circular icon with a checkmark, and a circular icon with a folder labeled "Joinup Licensing Assistant". To the right of the diagram is a button labeled "SUBSCRIBE TO THIS SOLUTION". A small box indicates the topic is "EUPL". The bottom right corner features the logo of Université Clermont Auvergne.

To be Reusable

A point on the licences

- number different types of licences from proprietary to totally free AND open
- more than 50 different kinds
- No licence = authors' rights
- France = support for free and open data

[Overview](#)

[Members](#)

[About](#)

[JLA - Compatibility Checker](#)

[JLA - Find and compare software licenses](#)

[REPORT ABUSIVE CONTENT](#)

A unique tool allowing everyone to compare and select open licences based on their content.

[Select licence terms below](#)

	Can	Must	Cannot	Compatible	Law	Support
Use/reproduce	Incl. Copyright	Hold liable	None N/A	EU/MS law	Strong Community	
Distribute	Royalty free	Use trademark	Permissive	US law	Governments/EU	
Modify/merge	State changes	Commerce	GPL	Licensor's law	OSI approved	
Sublicense	Disclose source	Modify	Other copyleft	Other law	FSF Free/Libre	

Pierre MARIN (Université Clermont Auvergne, AuBi, Mésocentre)

FAIR Bioinfo 2022

3 avril 2023

24 / 30

To be Reusable

A point on software archive

- Git-like (github, gitlab) are web services not archive
- Software are fragile

To be Reusable

A point on software archive

- Git-like (github, gitlab) are web services not archive
- Software are fragile

The screenshot shows the Software Heritage website. At the top, there is a navigation bar with links for Mission, Archive, Communauté, Grants, Soutien, A propos, and a search icon. The main header features the Software Heritage logo (a stylized orange and yellow starburst) and the text "Software Heritage". Below this, a sub-header reads "préserve le code source des logiciels, pour les générations actuelles et futures". To the right, there is a photograph of a modern library or archive interior with white bookshelves and a staircase. The central part of the page has a large red banner with the text "Nous construisons l'archive universelle des logiciels". At the bottom, there is a footer section with the Software Heritage logo and text about collecting and preserving software source code, followed by a "Collect Preserve Share" call-to-action and a note about conserving accessible software.

Software Heritage

Mission Archive Communauté Grants Soutien A propos

Software Heritage
préserve le code source des logiciels, pour les générations actuelles et futures

Nous construisons l'archive universelle des logiciels

Collect Preserve Share

Nous collectons et préservons les logiciels sous forme de code source parce qu'ils sont le support indissociable des connaissances techniques et scientifiques de l'humanité tout entière et que nous ne pouvons pas prendre le risque de les perdre.

Nous conservons et rendons accessible tous les logiciels que nous collectons car c'est uniquement en les partageant que nous pouvons

UNIVERSITÉ Clermont Auvergne

FAIR tools

Findable



Accessible



Interoperable



Reusable



Data



Software
and
analyses



[FAIRsharing.org](https://fairsharing.org)

standards, databases, policies



CeCILL

A complete integrated FAIR environment

The Galaxy project



A complete integrated FAIR environment

The Galaxy project

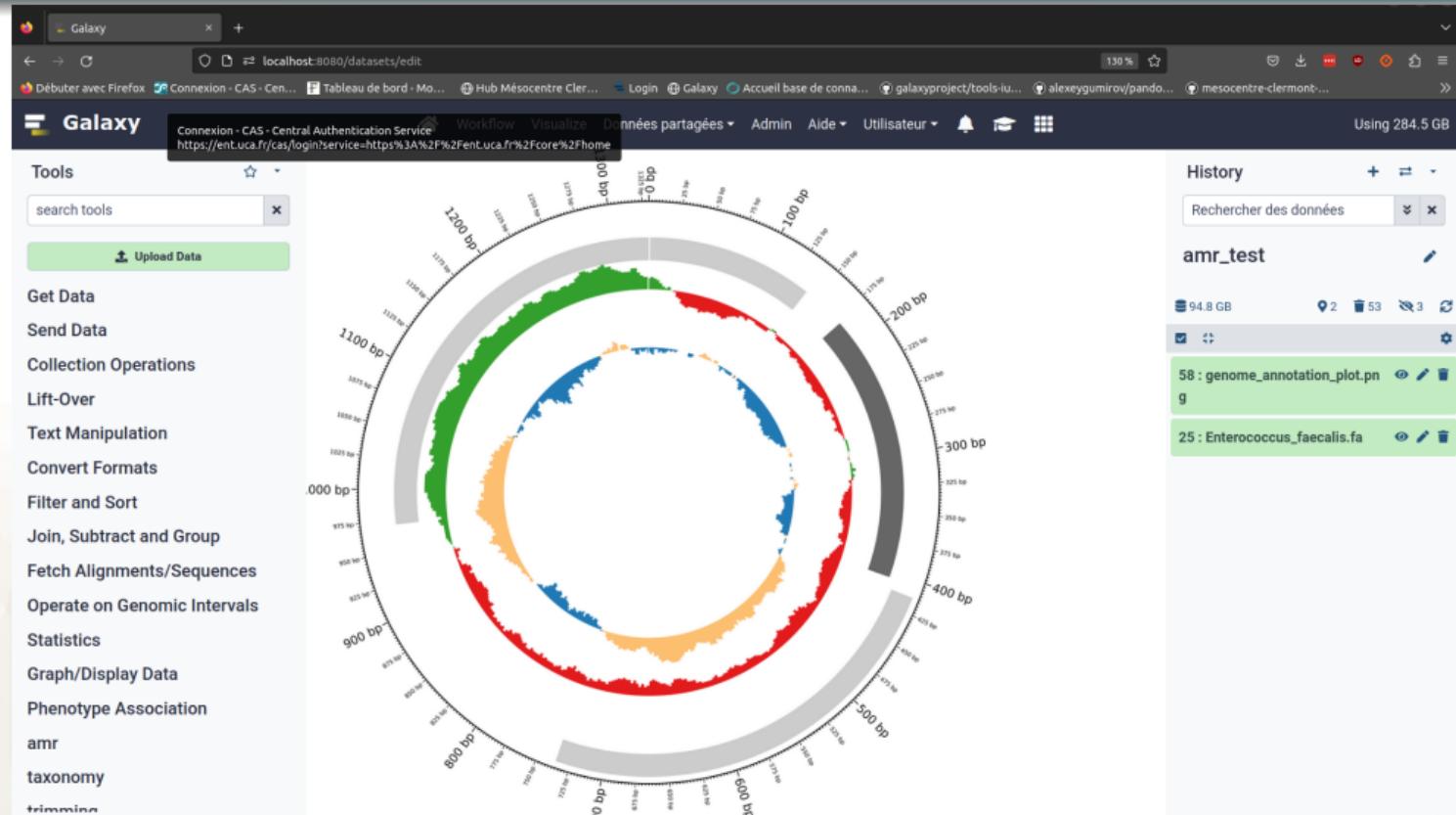
Galaxy

Galaxy is an open-source platform for FAIR data analysis that enables users to :

- Use tools from various domains (that can be plugged into workflows) through its graphical web interface.
- Run code in interactive environments (RStudio, Jupyter...) along with other tools or workflows.
- Manage data by sharing and publishing results, workflows, and visualizations.
- Ensure reproducibility by capturing the necessary information to repeat and understand data analyses.

A complete integrated FAIR environment

The Galaxy project



A complete integrated FAIR environment

The Galaxy project

The screenshot shows the Galaxy web interface running on localhost:8080. The main panel displays a workflow titled "staramr" which scans genome assemblies against three databases: ResFinder, PlasmidFinder, and PointFinder. The workflow has a single input file, "25 : Enterococcus_faecalis.fa", listed under the "genomes" section. Below the workflow, there are several configuration sliders for BLAST parameters:

- Percent identity threshold for BLAST: 98.0
- Percent length overlap of BLAST hit for ResFinder database: 60.0
- Percent length overlap of BLAST hit for PointFinder database: 95.0
- Percent length overlap of BLAST hit for PlasmidFinder database: 60.0

On the right side of the interface, there is a "History" panel titled "amr_test" containing a file named "58 : genome_annotation_plot.png". The top right corner of the interface indicates "Using 284.5 GB".

A complete integrated FAIR environment

The Galaxy project

The screenshot shows the Galaxy web interface running a workflow titled "Workflow: abromics_SR_PE_workflow". The interface includes a sidebar with various tool categories like Tools, Get Data, Send Data, etc. The main area displays the workflow steps and their outputs. A "Run Workflow" button is visible. To the right, a "History" panel shows a list of data files and their details.

Workflow: abromics_SR_PE_workflow

History Options

Send results to a new history
 No

1: R1_fastq

2: R2_fastq

3: abromics_SR_PE_trimming

4: abromics_SR_PE_taxonomy

5: abromics_SR_PE_assembly

6: abromics_assembly_antimicrobial_detection

7: abromics_assembly_annotation

8: abromics_SR_PE_QC_metrics

9: abromics_depth_amr_gene_workflow

History

Rechercher des données

amr_test

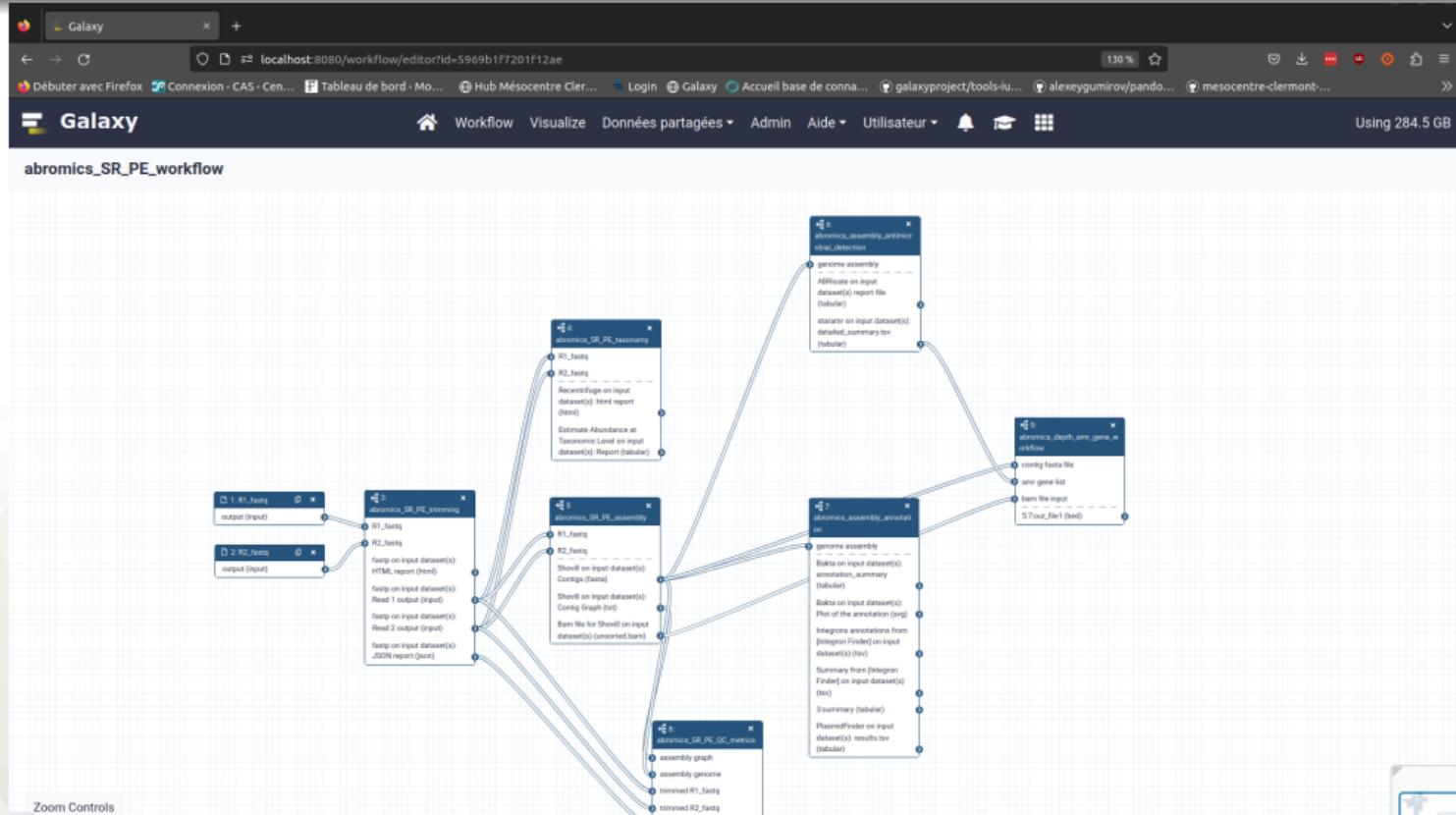
94.8 GB

58 : genome_annotation_plot.png

25 : Enterococcus_faecalis.fa

A complete integrated FAIR environment

The Galaxy project



A FAIR approach at any level



"One practice to rule them all, One practice to find them, One practice to bring them all and in the FAIR bind them."

A FAIR approach at any level



"One practice to rule them all, One practice to find them, One practice to bring them all and in the FAIR bind them."

Improvement for who ?

At each level a good practice help you

- Long term efficiency to a bioinformatician
- stop wasting time for a beginner