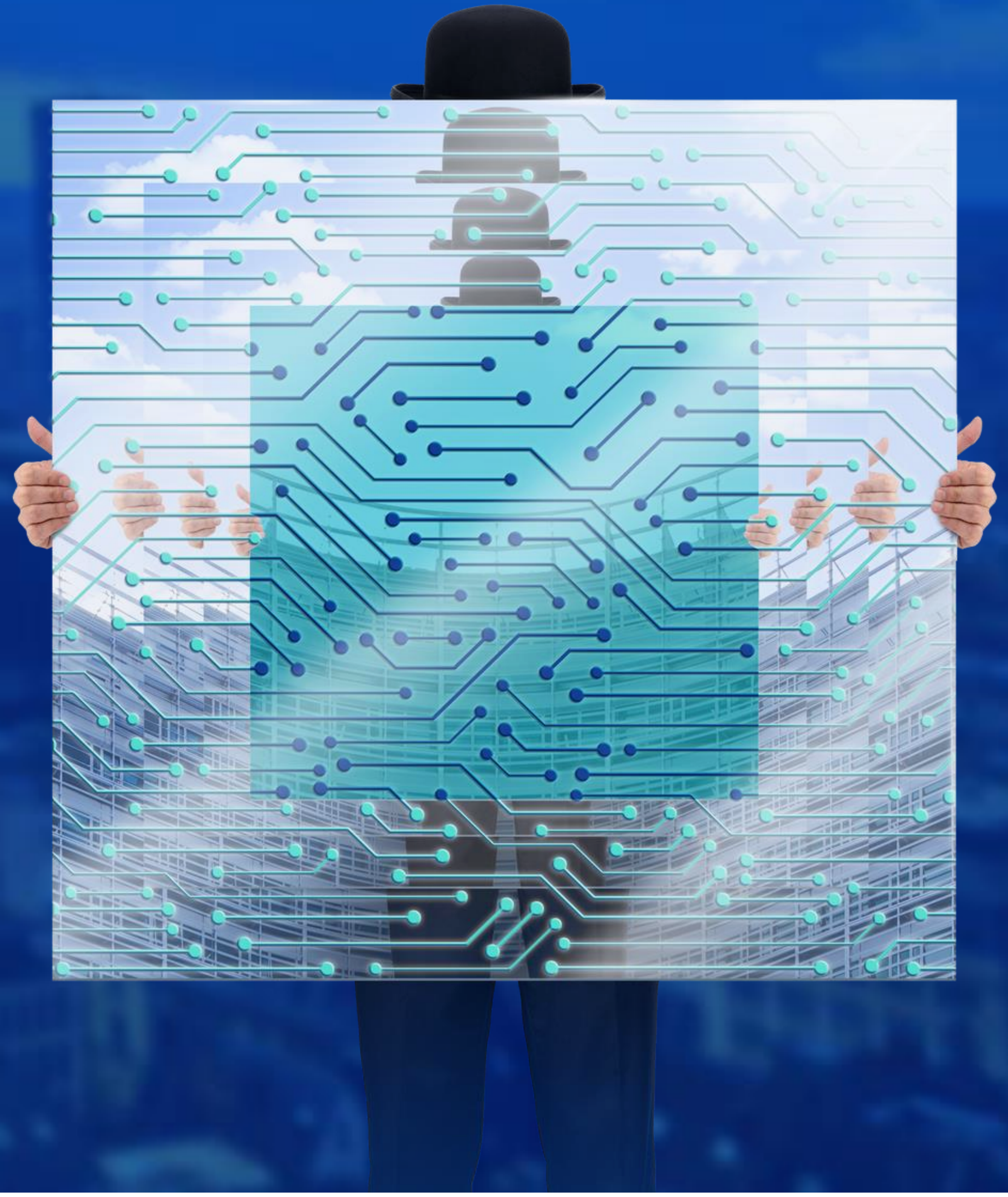
FARI BRUSSELS CONFERENCE

LIÈGE université

Can AI, Data and Robotics Research be Open and Accessible to All? Exploring the challenges

From the perspective of open research data management



FARI AI FOR THE COMMON GOOD INSTITUTE BRUSSELS
UNIVERSITÉ LIBRE DE BRUXELLES ULB VUB VRIJE UNIVERSITEIT BRUSSEL
innoviris.brussels we fund your future
paradigm.brussels
RÉGION DE BRUXELLES-CAPITALE
Funded by the European Union NextGenerationEU

What does a research data officier do?

Data Management Helpdesk ☺
Boosting compliance with:

- FAIR data principles
- Open Science guidelines
- Stakeholders demands
- Applicable regulations or standards

All along the data life cycle of their projects

LIÈGE université

FARI AI FOR THE COMMON GOOD INSTITUTE BRUSSELS · UNIVERSITÉ LIBRE DE BRUXELLES ULB VUB VRIJE UNIVERSITEIT BRUSSEL · innoviris.brussels · paradigm.brussels · RÉGION DE BRUXELLES-CAPITALE · Funded by the European Union NextGenerationEU

Open data, open source, and AI

Open science means more research outputs online…
… and more training material for machine-learning models

-> Boosting open science **supports** innovation capabilities in AI, even outside academia

Open data, open source, and AI

Open science means more research outputs online…
… and more training material for machine-learning models

-> Is OS enough?
-> Will research be ever really open?

Technical challenges

Financial challenges

Systemic challenges

Technical challenges

Just because a dataset is open does not mean it is:

Easy to find

Directly reusable

Formats & metadata, provenance

Of good quality

Technical debt

Compliant

True anonymity vs big data

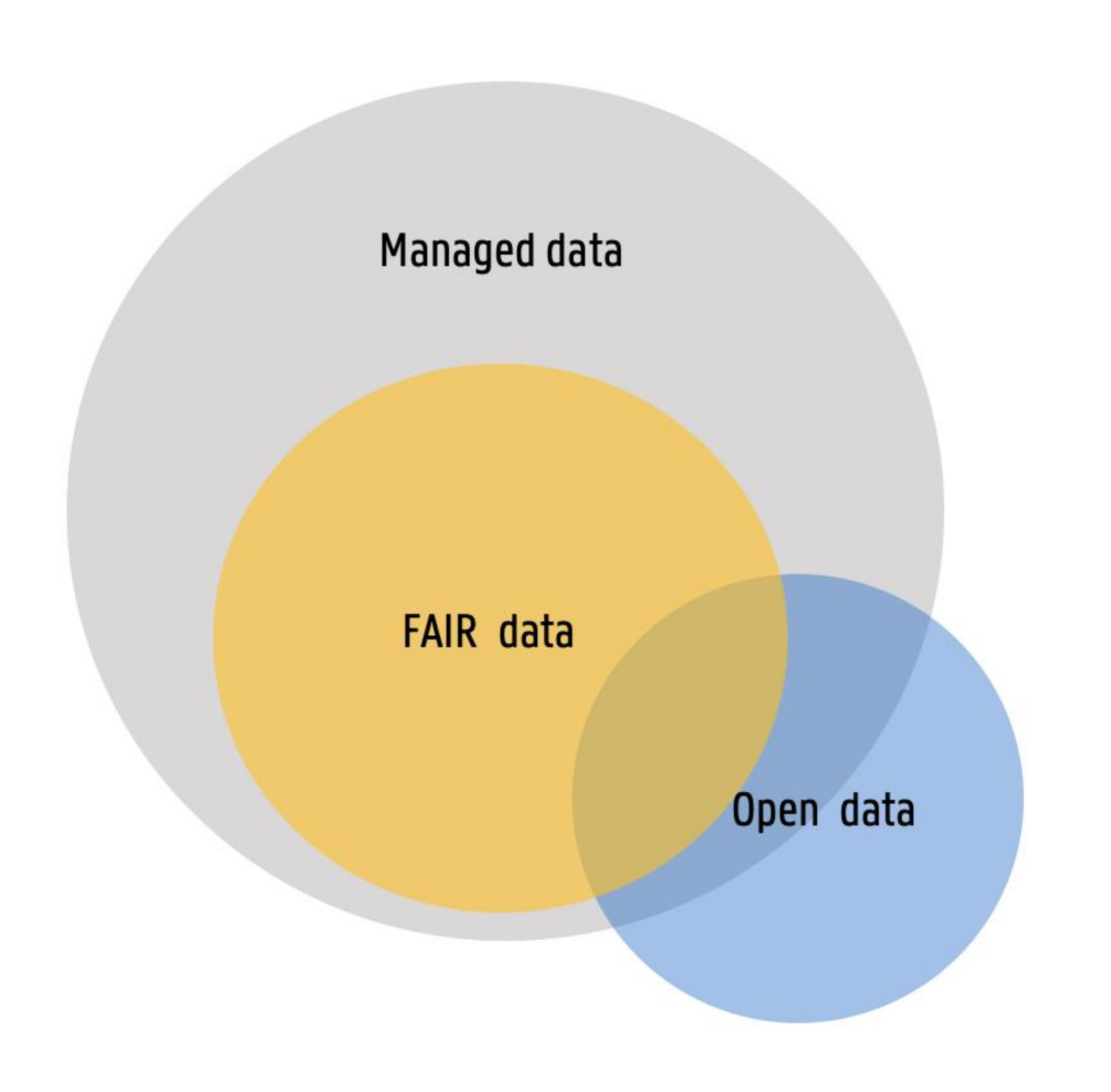
Non-fraudulent

LancetGate

Technical challenges

Shifting the norm to the FAIR data principles helps addressing some of these challenges
->**Awareness** in research communities
->**Tools and support** available in RPOs

Technical challenges



Findable

Data repositories with machine-readable metadata, including a DOI

Accessible

Standardized retrieval protocol (such as « log in and download »)

Interoperable

Non-proprietary format, standard vocabularies, references, languages

Reusable

Sufficient documentation for autonomous reuse, including non-experts, open licence

Technical challenges

Just because a piece of code is open does not mean it is:

Portable

Well-documented

Years and layers of decision-making

Model-data blur

Of good quality

Technical debt

Generalisable

Verification vs hope for re-use and co-dev

Lawfully reusable

What really is an open source licence?

Technical challenges

Some of the FAIR principles can support open source too

Journal policies and good practice guides

Producing truly open code takes **skills and time**

Technical challenges

Case study: re-using social network data for AI-based research

How to collect?

Tools, method, selection/rejection criteria? Quality control? Need for inter-disciplinarity (IT, law, numerical humanities, …)

How to reuse?

Ethical considerations: consent, sensitivity, illegal content, … Technical considerations: storage security, volume vs cost, duration …

How to publish?

Legal considerations: freedom of speech, copyright, privacy protection, true anonymity, social network policies…

LIÈGE université

Financial challenges

Indirect, non-negligible costs

Human ressources for maintenance and user support
- Often poorly recognised tasks in research careers

Cumulative **data storage** vs digital sobriety

Article Processing Charges for open science papers are a major barrier to OS culture in general

In 2022, cumulated ULiège expenses in APC = 460k€ for 238 articles

LIÈGE université

Systemic challenges

Can science ever be truly open?

Terrorists will use it	We'll get spam	It's too big	It's not very interesting
Thieves will use it	I don't mind, but someone else might	We will get too many enquiries	Lawyers want a custom License
There's no API	Poor Quality	There's already a project to...	We might want to use it in a paper
It's too complicated	Data Protection	People may misinterpret the data	What if we want to sell it later

#opendataexcuses

Systemic challenges

Biases in open science culture impact AI innovation

Positivity bias
Negative results are seldom published,
painting **a too positive picture**

-> Skewed training datasets
-> Skewed ML model

e.g. enhancing chemical properties of material

Systemic challenges

Biases in open science culture impact AI innovation

Lack of bibliodiversity
Metrics culture effectively silences a whole facet of scientific productions

-> Global North over-represented
-> Gender, ethnic, linguistic biases

-> Vendor lock-in, limited publication models

Systemic challenges

Research communication remains a market, research outputs are managed **more as commercial goods** than as a public good.

Its biases are carried out in machine-learning models that are trained on this improper representation.

A **deeper shift** of publication strategies needs to be incentivized towards truly reusable research.

So what do we do ? A culture shift

AI in and of itself might just help detect and discuss such biases ☺

*“By analysing big data, **researchers have confirmed** deeply rooted intersectional inequalities in science production and publishing. And they have a fairly good understanding of how various dimensions of diversity, from gender and racial or ethnic composition to interdisciplinarity and geography, relate to outcomes”*

Broader scope is key to the future of ‘science of science’. Nat Hum Behav 6, 899–900 (2022). https://doi.org/10.1038/s41562-022-01424-5

LIÈGE université

FARI AI FOR THE COMMON GOOD INSTITUTE BRUSSELS · UNIVERSITÉ LIBRE DE BRUXELLES ULB VUB VRIJE UNIVERSITEIT BRUSSEL · innoviris.brussels we fund your future · paradigm.brussels · RÉGION DE BRUXELLES-CAPITALE · Funded by the European Union NextGenerationEU

So what do we do ? A culture shift

Top-down initiatives from big stakeholders
Since the Jussieu Call and the DORA manifesto

EOSC: a diversity of services and infrastructures, supporting OS in all its forms

Open and FAIR as norm for € grants, FNRS/ FWO, …

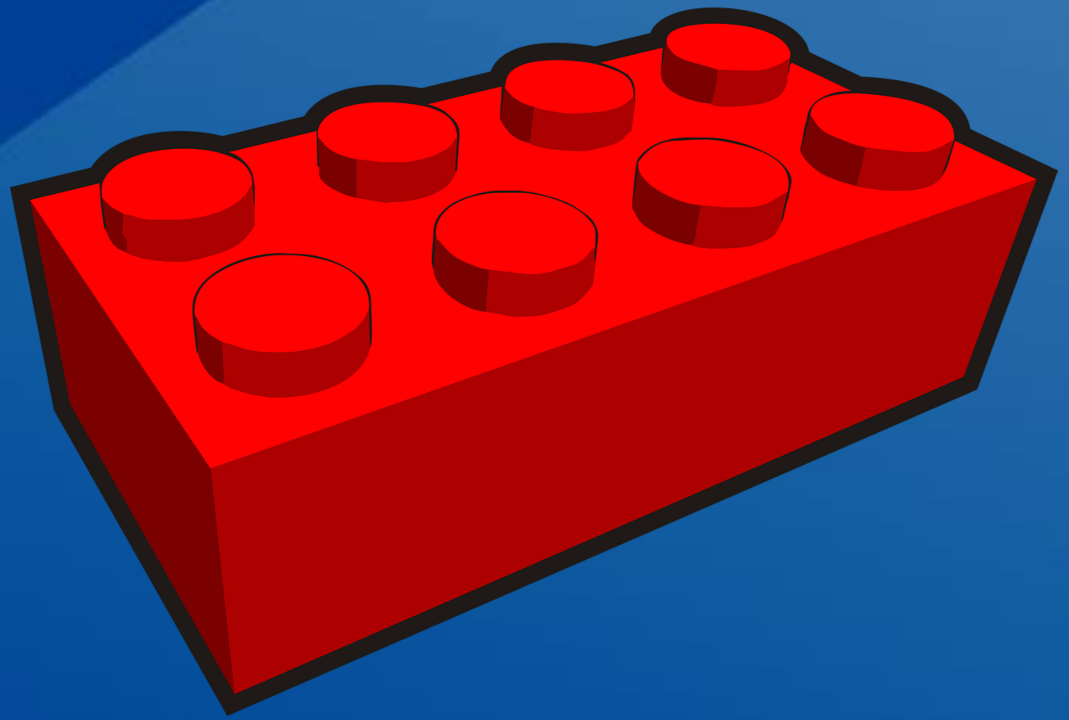
CoARA: RPOs are commiting to steer away from metrics-based evaluations

So what do we do ? A culture shift

Bottom-up initiatives:

Sector-level guidelines
Such as TOP or EQUATOR guidelines

Belgian Reproducibility Network
Flemish Research Data Network
FWB Data Ambassadors



Hold your institution CoARA accountable
Challenge publication culture towards truly open science

Ackowledgements

Additionnal bibliography (most resources linked in text)

Birhane, A., Kasirzadeh, A., Leslie, D. et al. Science in the age of large language models. Nat Rev Phys 5, 277–280 (2023). https://doi.org/10.1038/s42254-023-00581-4

Illuminating 'the ugly side of science': fresh incentives for reporting negative results, Rachel Brazil, Nature – Career Feature, 8 May 2024, https://www.nature.com/articles/d41586-024-01389-7 (accessed on Aug 6, 2024)

Shearer, K., Chan, L., Kuchma, I., & Mounier, P. (2020). Fostering Bibliodiversity in Scholarly Communications: A Call for Action. Zenodo. https://doi.org/10.5281/zenodo.3752923

Bardiau, M., & Dony, C. (2024). Measuring back: bibliodiversity and the Journal Impact Factor™ brand, a case study of IF-journals included in the 2021 Journal Citations Report™. Insights: the UKSG Journal, 37. doi:10.1629/uksg.633

Curry, Stephen. « The Intersections Between DORA, Open Scholarship and Equity ». Proceedings of the Paris Open Science European Conference, OpenEdition Press, 2022, https://doi.org/10.4000/books.oep.16151.

Eglen, S., Marwick, B., Halchenko, Y. et al. Toward standard practices for sharing computer code and programs in neuroscience. Nat Neurosci 20, 770–773 (2017). https://doi.org/10.1038/nn.4550

Easterbrook, S. Open code for open science?. Nature Geosci 7, 779–781 (2014). https://doi.org/10.1038/ngeo2283

Open Science Isn't Always Open to All Scientists, Christie Bahlai, Lewis J. Bartlett, Kevin R. Burgio, Auriel M. V. Fournier, Carl N. Keiser, Timothée Poisot, Kaitlin Stack Whitney, American Scientist March-April 2019, Volume 107, Number 2, Page 78 DOI: 10.1511/2019.107.2.78 (accessed on Aug 6, 2024)

Thanks to Lyne Van de Vyvere and Jérémie Fays for their support

LIÈGE université