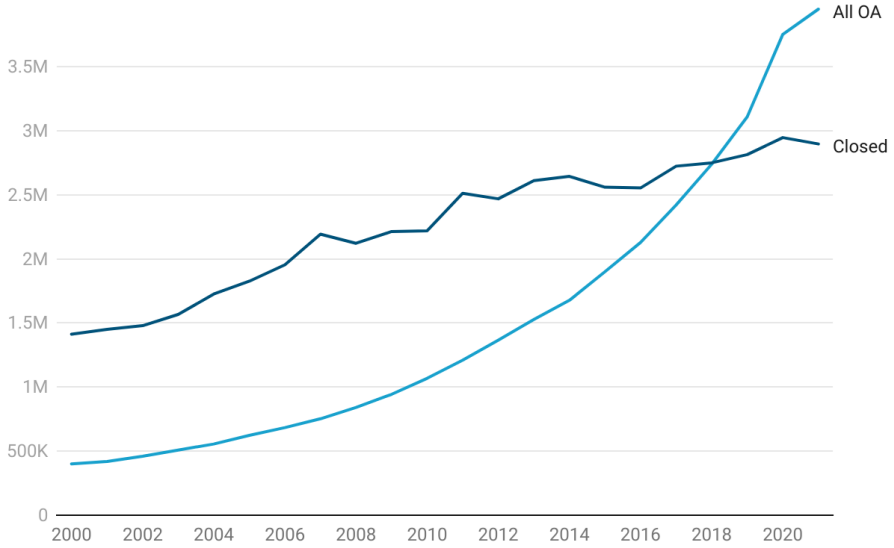
Academic Data Citation and Reuse

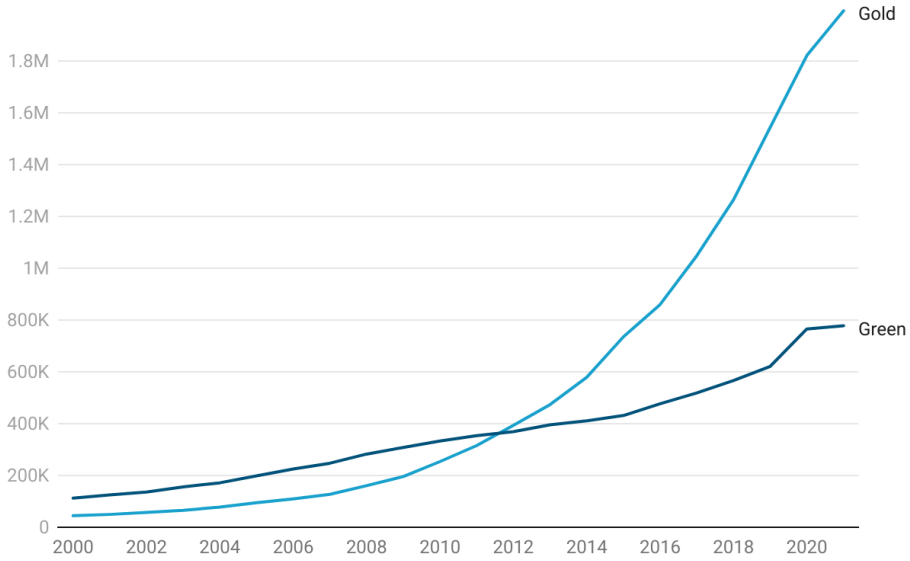
Mark Hahnel
VP Open Research & Founder of Figshare

Academic Publishing Outputs

Open vs Closed Access Publishing



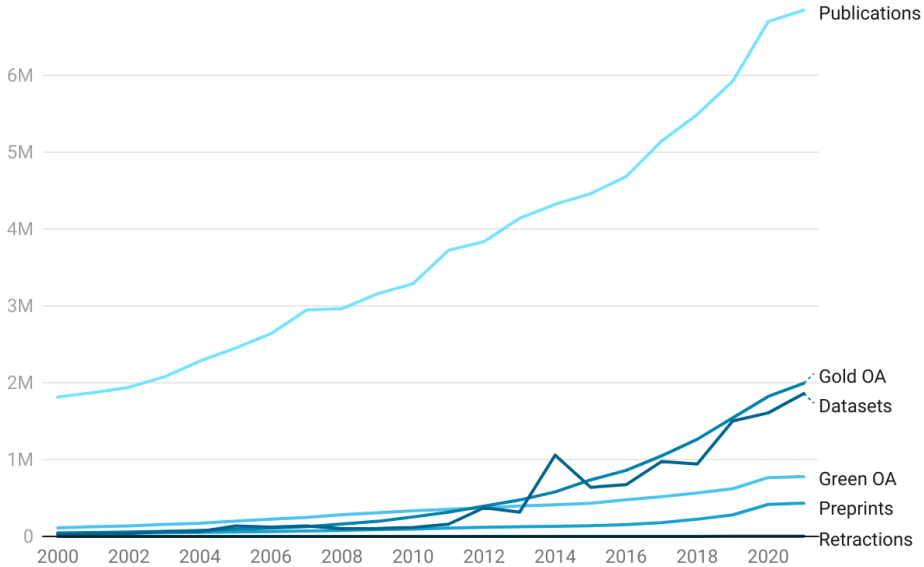
Gold vs Green Open Access Publishing



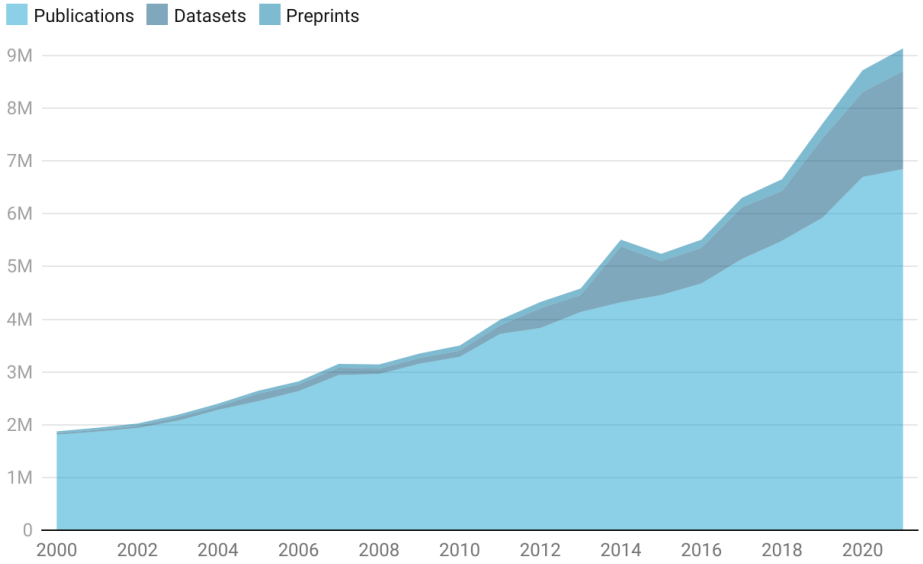
Academic Publishing Outputs

DIGITAL science

Academic Publishing Outputs



Cumulative Academic Publishing Outputs



Why?

DIGITAL science

The State of Open Data

Has been running for 9 years

Had over 30,000 respondents from 192 countries over that time

2023 had first partner publication from CAS

DIGITAL science

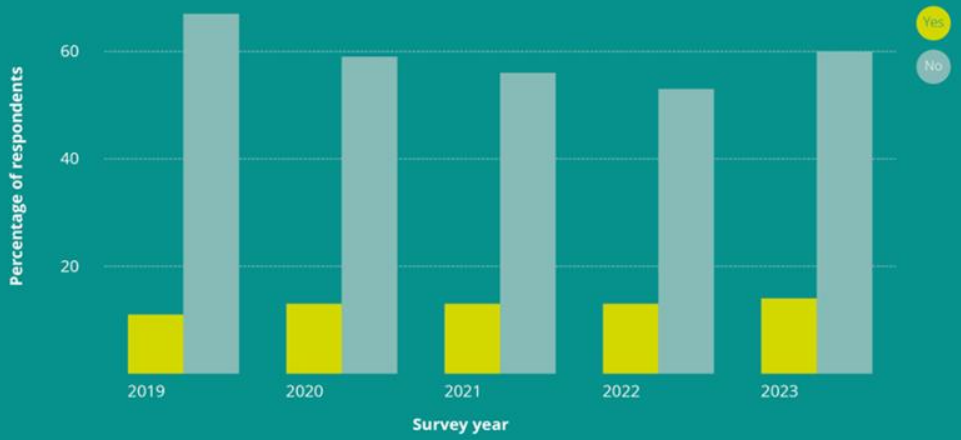
SPRINGER NATURE

figshare

Credit is an ongoing issue

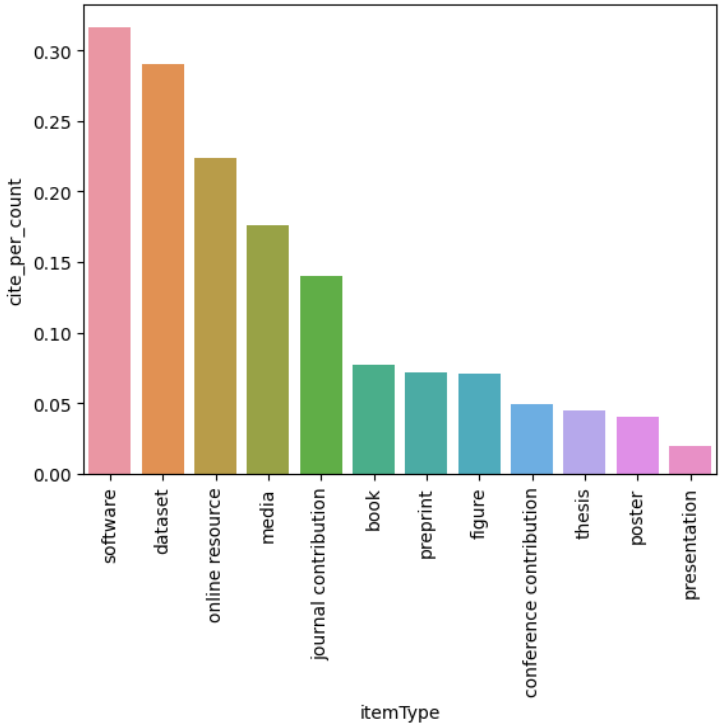
For eight years running, our survey has revealed a recurring concern among researchers: the perception that they don't receive sufficient recognition for openly sharing their data.

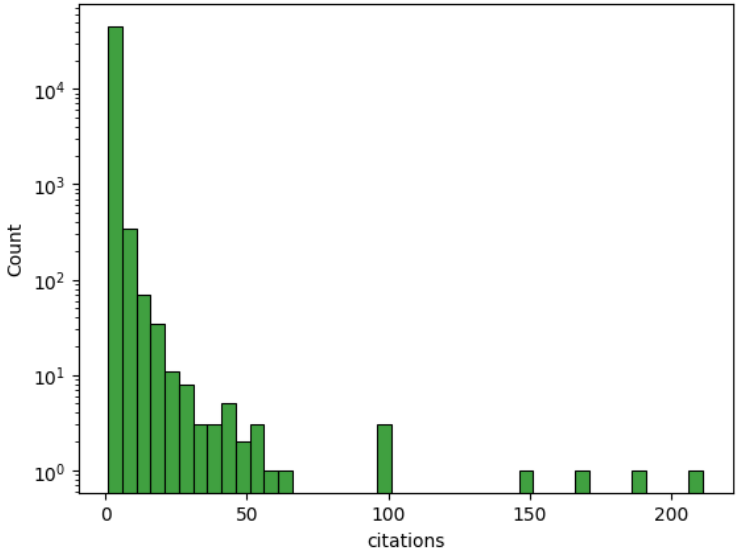
Do you think researchers currently get sufficient credit for sharing data?



Longitudinal survey data from 2019-2023 for the question 'Do you think researchers currently get sufficient credit for sharing data?'

Citation Counts on Figshare.com





DIGITAL science

SPRINGER NATURE

figshare

Support is not making its way to those who need it

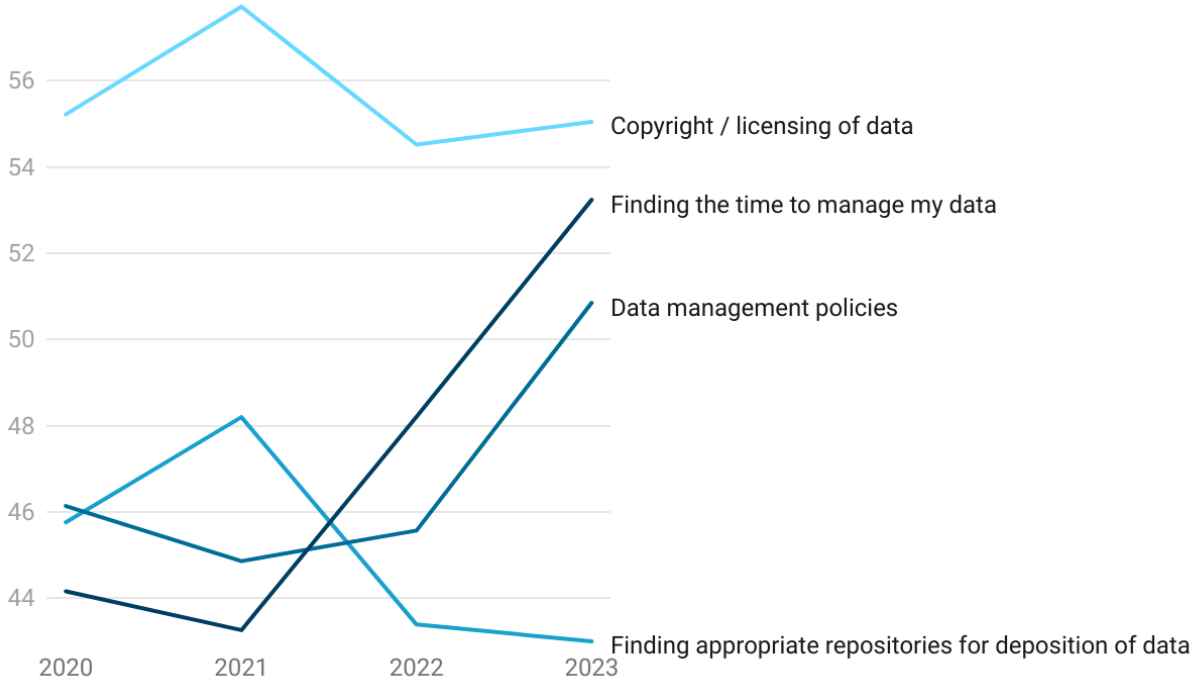
Almost three-quarters of respondents had never received support with making their data openly available.

What percentage of researchers "strongly agree" that the following should be made open, by job title?

The State of Open Data

DIGITAL science

What areas, if any, do you feel you need help with in regard to making your research data openly available?



DIGITAL science

SPRINGER NATURE

figshare

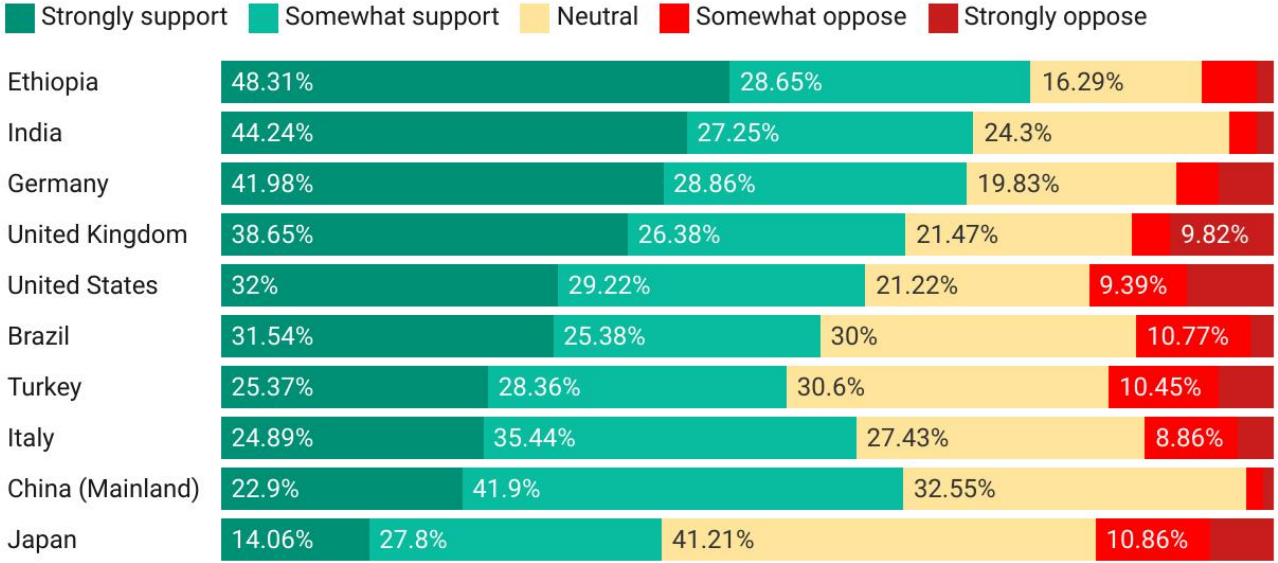
The State of Open Data 2023

One size does not fit all

Variations in responses from different subject expertise and geographies highlight a need for a more nuanced approach.

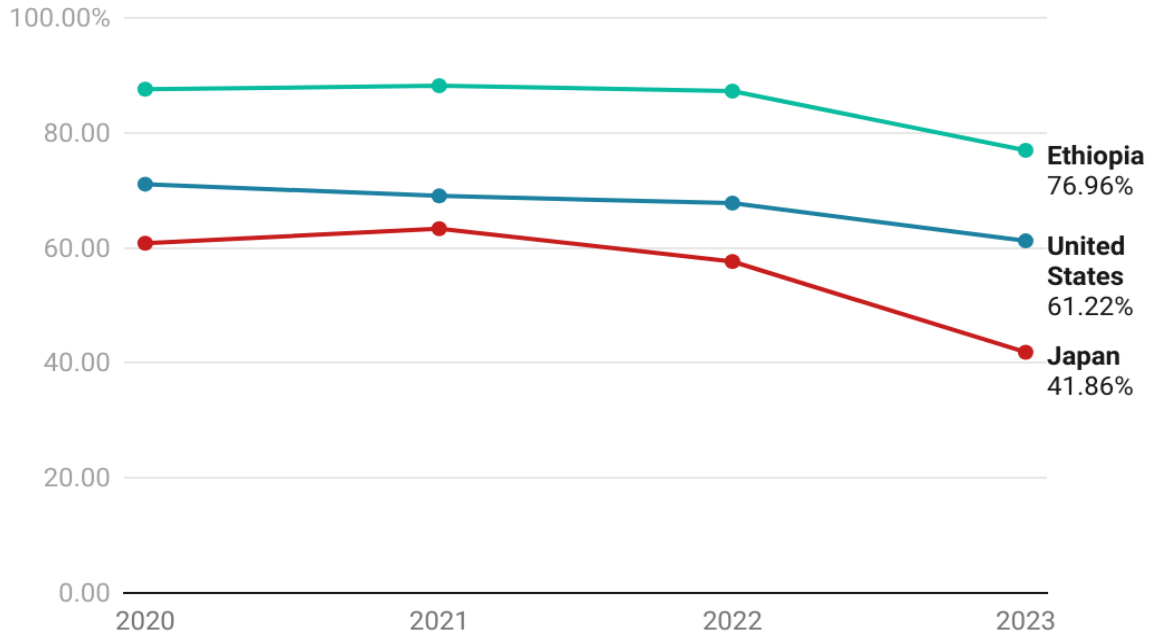
#StateOfOpenData

Thinking about the country in which you are currently working, how supportive are you of the idea of a national mandate for making research data openly available?

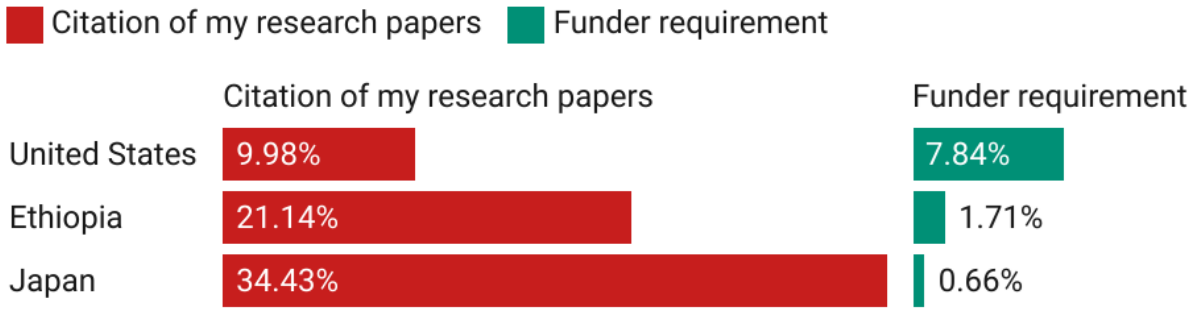


Created with Datawrapper

How supportive would you be of a national mandate for making research data openly available?



Which one of these circumstances would motivate you the most to share your data?

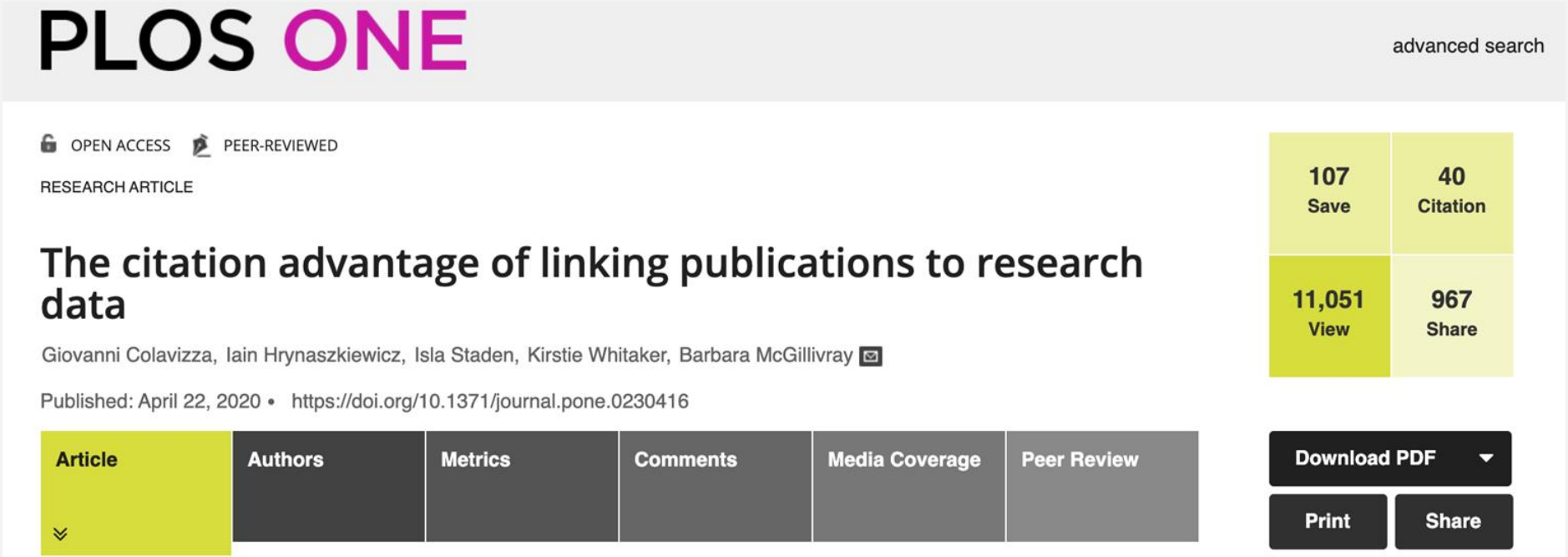


as a proportion of total responses to survey question

Created with Datawrapper

*"Linking papers to their supporting data in a repository was associated with on average a **25% increase** in citations"*

https://doi.org/10.1371/journal.pone.0230416



PLOS ONE

advanced search

OPEN ACCESS PEER-REVIEWED

RESEARCH ARTICLE

The citation advantage of linking publications to research data

Giovanni Colavizza, Iain Hrynaszkiewicz, Isla Staden, Kirstie Whitaker, Barbara McGillivray

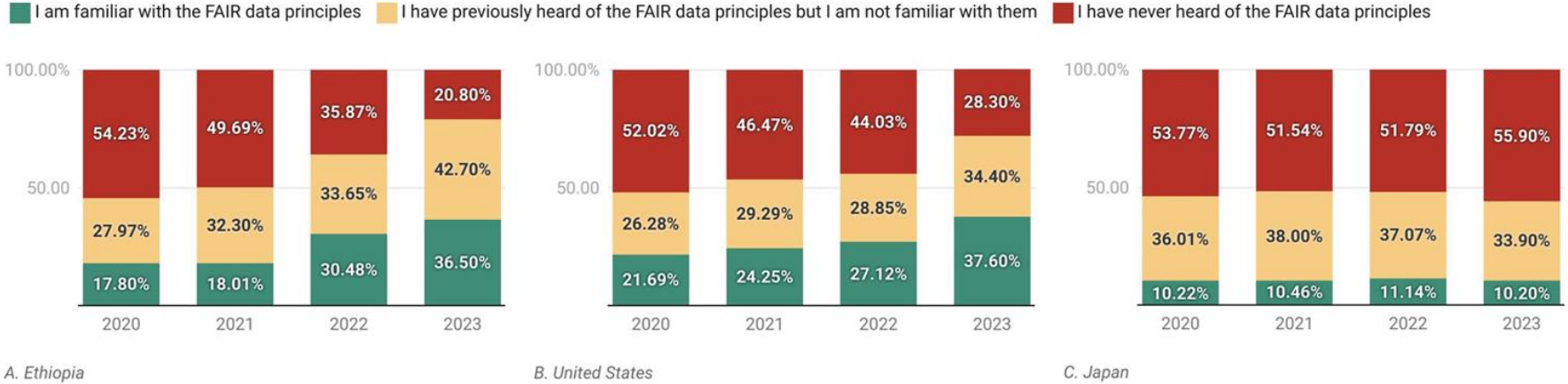
Published: April 22, 2020 • https://doi.org/10.1371/journal.pone.0230416

Article | Authors | Metrics | Comments | Media Coverage | Peer Review

107 Save | 40 Citation | 11,051 View | 967 Share

Download PDF | Print | Share

How familiar are you with the FAIR data principles in relation to Open Data?



A. Ethiopia

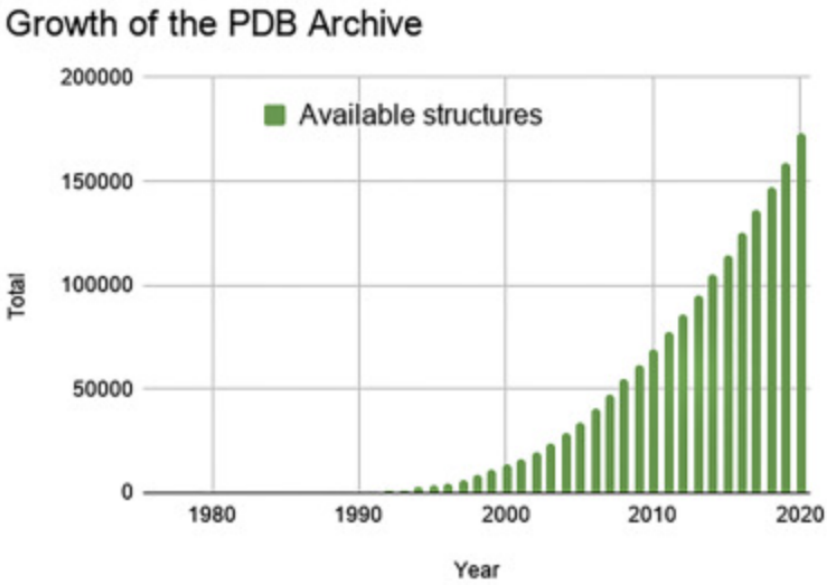
B. United States

C. Japan

But why?

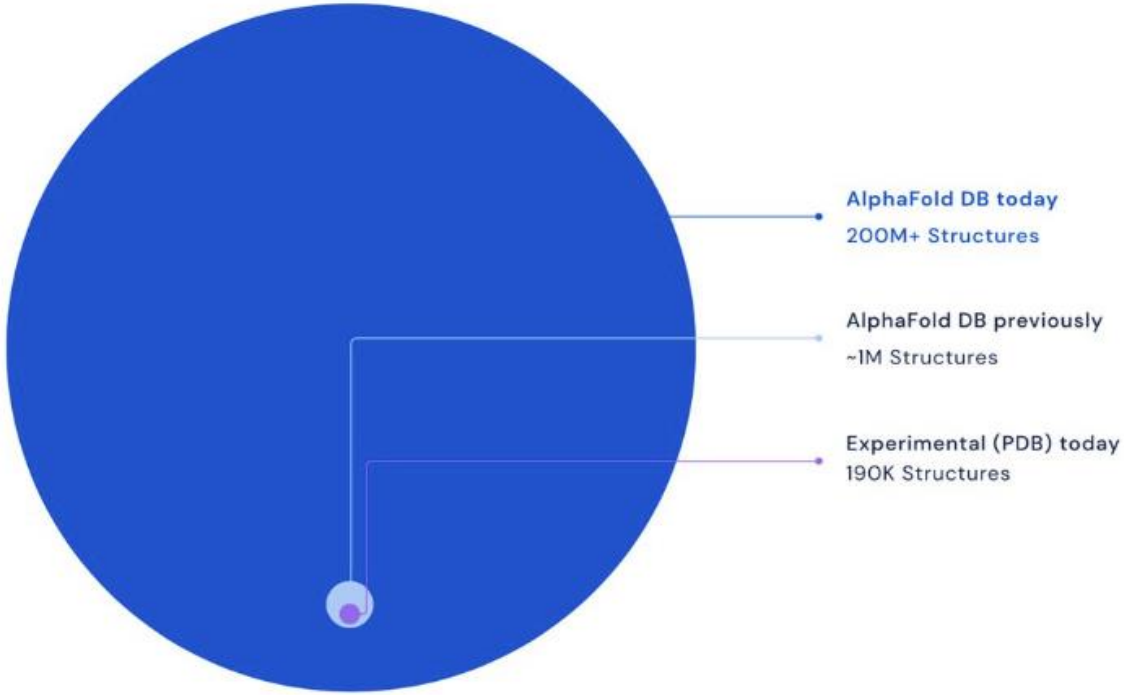
DIGITAL science

“We trained this system on publicly available data consisting of ~170,000 protein structures from the protein data bank”

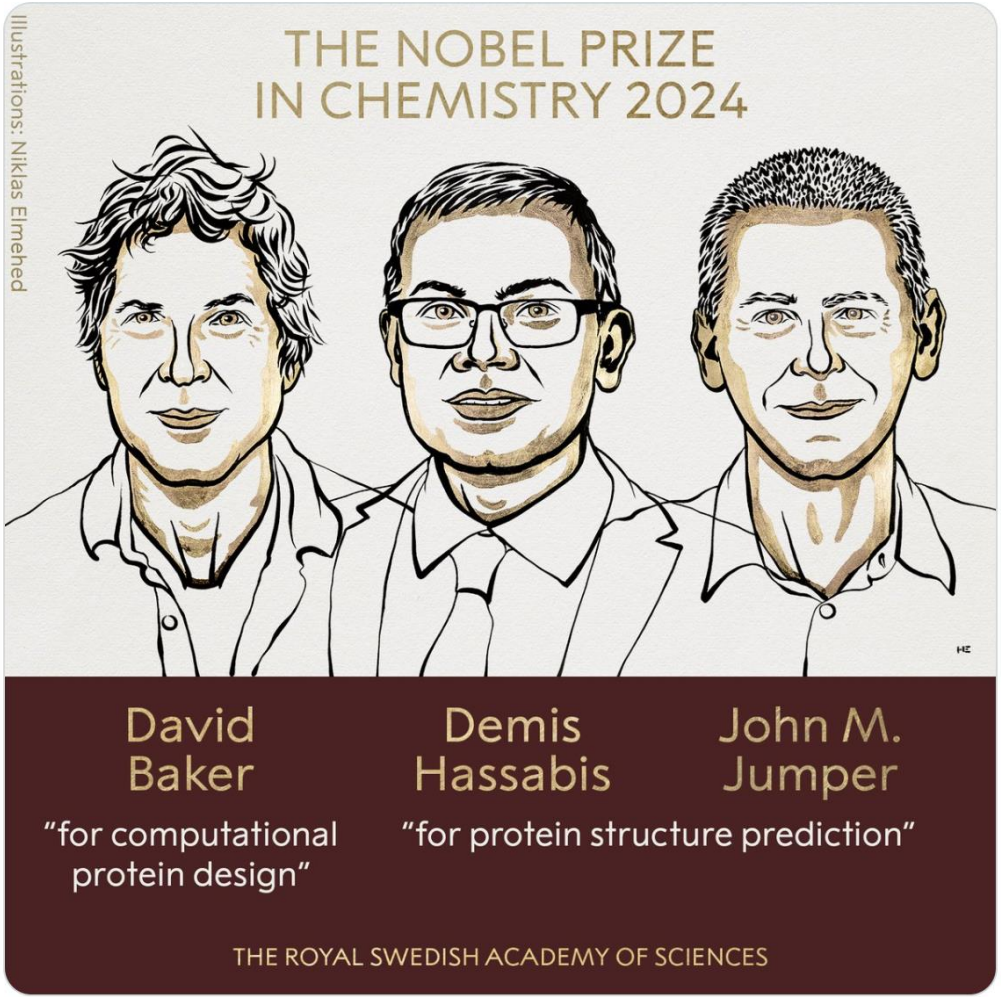


The estimated replacement cost of current PDB archival contents exceeds US$20 billion (assuming an average cost of US$100,000 for regenerating each of the >200,000 experimental structures).

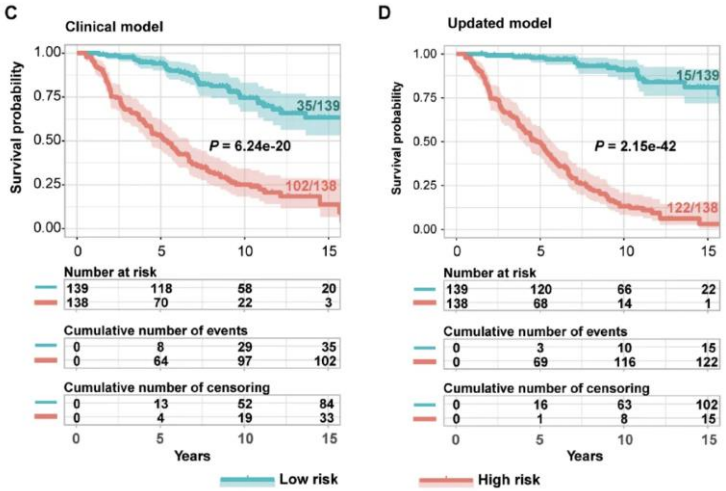
https://www.rcsb.org/pages/about-us/economic-impact



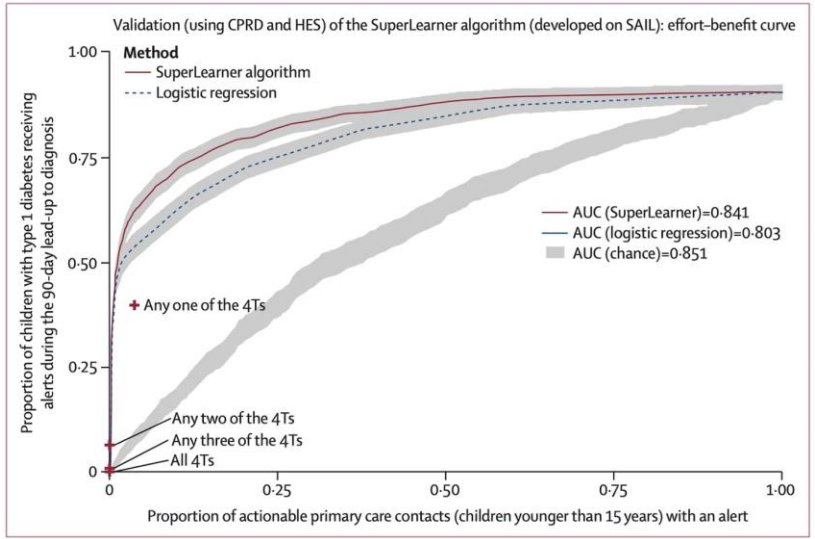
At launch	$100,000 * 1,000,000 = $100,000,000,000
After 1 year	$100,000 * 200,000,000 = $20,000,000,000,000


THE NOBEL PRIZE
IN CHEMISTRY 2024
Illustrations: Niklas Elmehed
David Baker
"for computational protein design"
Demis Hassabis
John M. Jumper
"for protein structure prediction"
THE ROYAL SWEDISH ACADEMY OF SCIENCES

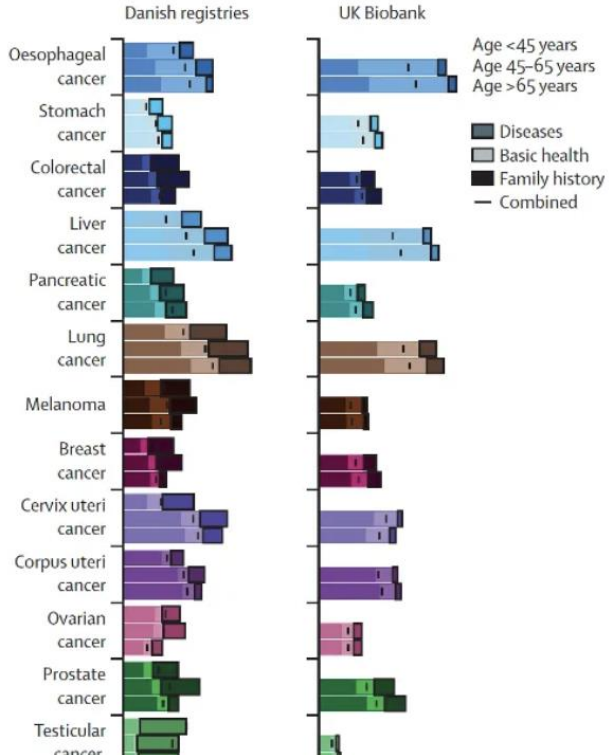
Multiple omics (DNA methylations, proteins, micro-RNAs, gene variants) for predicting kidney disease and failure in people with Type 1 diabetes compared with the standard clinical model for risk



Predicting Type 1 diabetes and timing in children



Predicting multi-cancer risk over 3-years using a national health resource (Denmark)



Nature | Vol 621 | 21 September 2023 | 467

DIGITAL science

CHANGING DRUG-DISCOVERY PIPELINES

Advances in artificial-intelligence (AI) tools and innovations in protein science and laboratory techniques mean that protein drugs can be developed faster and more efficiently today than was possible using older methods.

The process starts with millions of naturally occurring proteins, such as antibodies found in mice.

Conventional drug discovery

High-throughput screening identifies a handful of proteins that bind to a desired target at an appropriate strength.

6 months

Many cycles of engineering are needed to modify the proteins so that they have the properties a drug needs.

18 months

Modified proteins are rigorously tested to see whether they are safe, efficient and suitable for

AI and automation

Machine-learning models will soon be used to filter for protein sequences that bind to the desired target.

Less than 3 months

Such AI tools find likely binding sequences in an initial pool of millions.

FGIPCCPVHL...

Other machine-learning models are used to predict the properties that these protein sequences will have and to design improved candidates.

Feedback improves the predictions.

Automated assays test candidates.

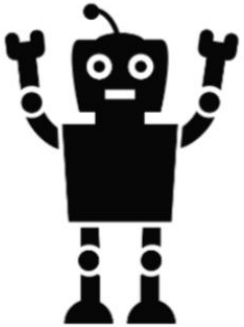
6 months

Higher confidence in the methods used means fewer candidates need to be rigorously

Leaps in technology are supporting AI-guided drug design, such as this fully robotic workstation that can purify proteins and move liquids.

AI can help to speed up drug discovery — but only if we give it the right data

FAIR



FAIR

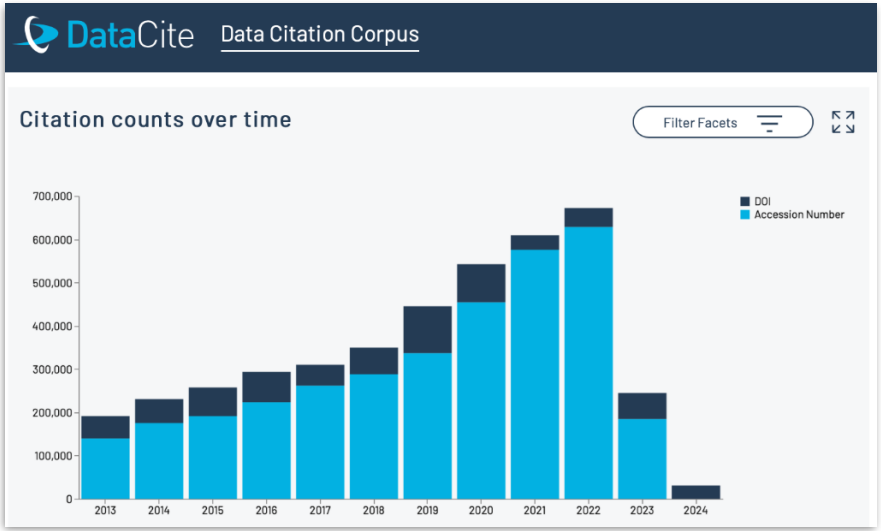
But why me as a researcher?

DIGITAL science

The Data Citation Corpus

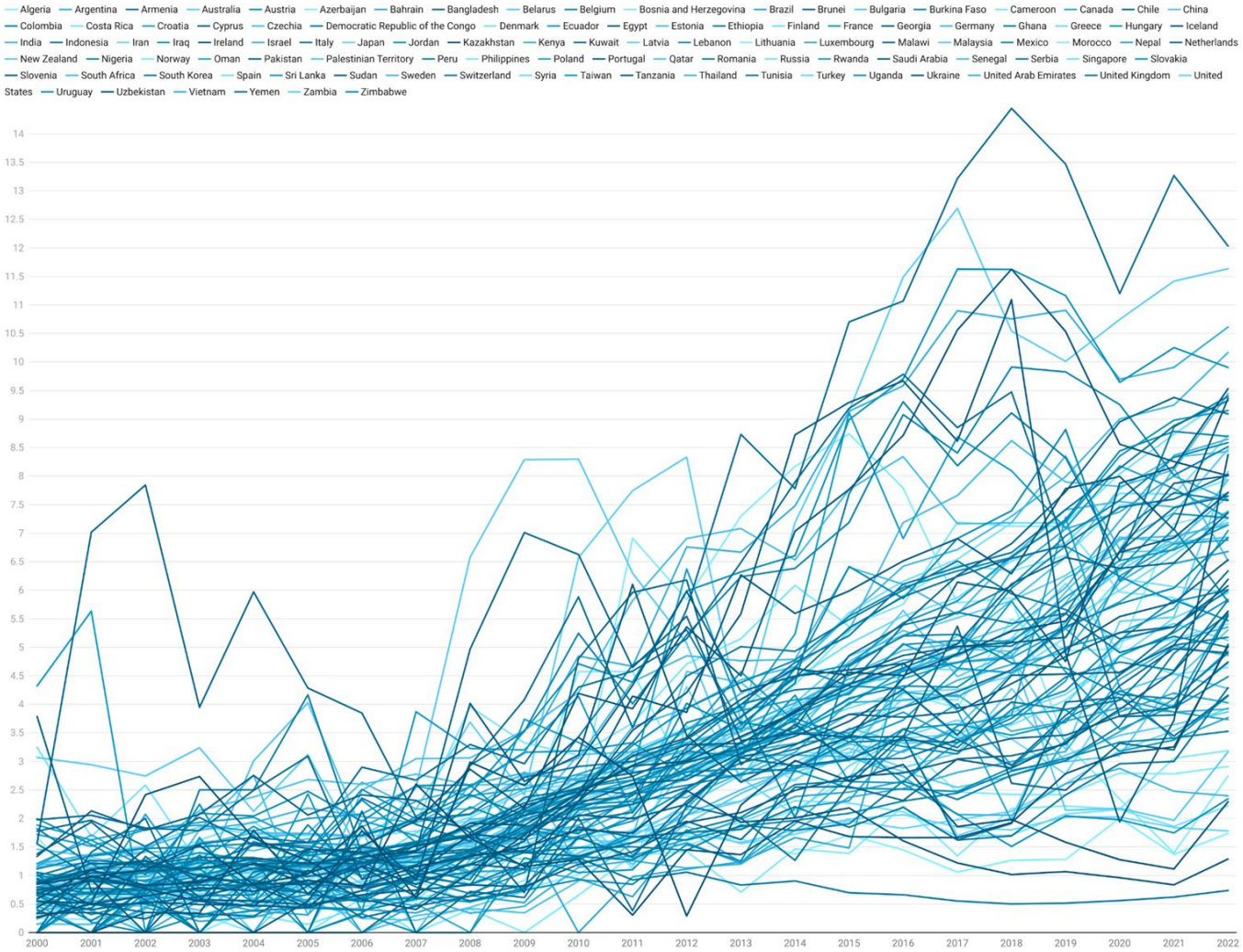
- Open aggregate of data citations
- 5 million citations
- Citations from diverse sources, including DOI metadata and machine learning

Check the latest data file for the Data Citation Corpus:
10.5281/zenodo.11196858

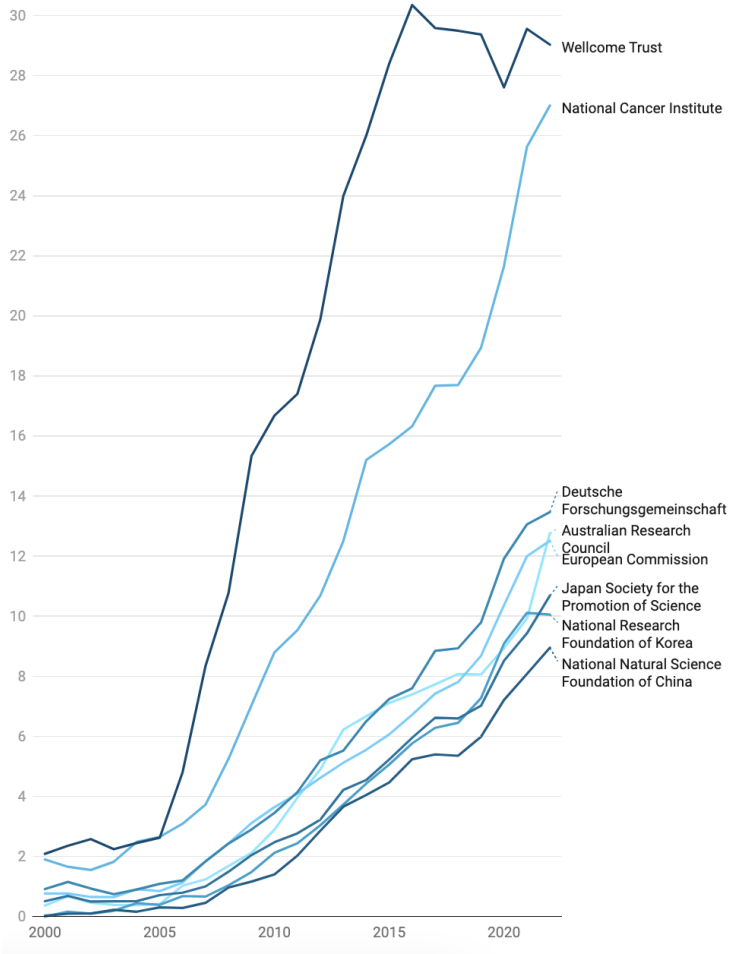


Data Citation Corpus dashboard:
https://corpus.stage.datacite.org/dashboard

Links to datasets in CZI corpus from countries with more than 1000 publications per year

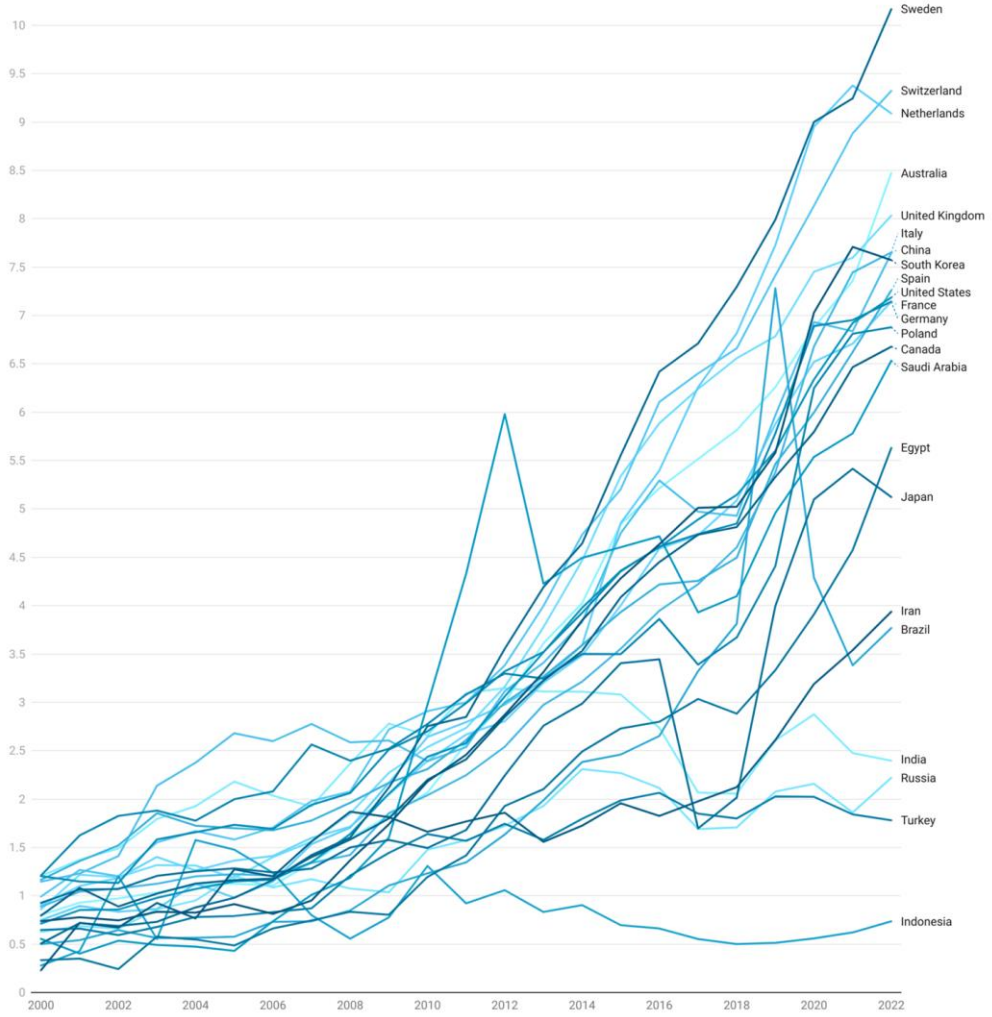


Percentage of Papers Citing a Dataset in CZI Corpus by Funder



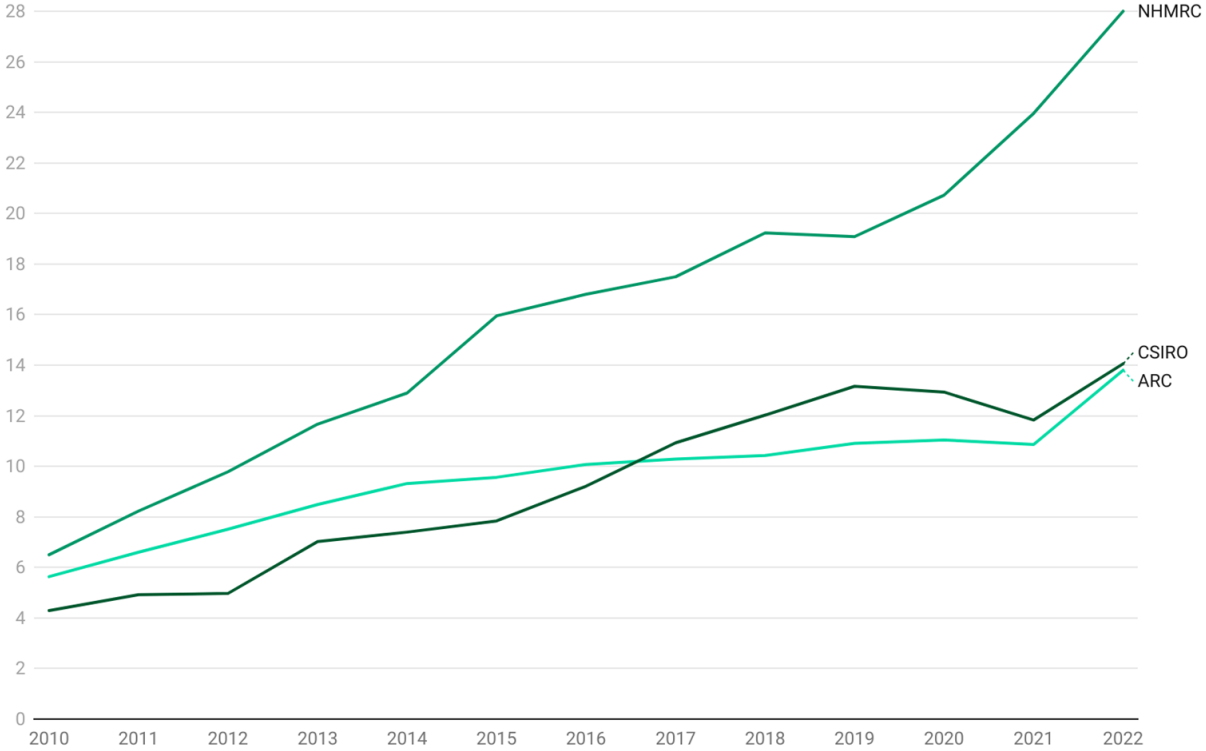
figshare
A **Digital Science** Solution

Links to datasets in CZI corpus from countries with more than 50,000 publications per year



figshare
A **Digital Science** Solution

Percentage of papers linking to datasets in the CZI corpus funded by Australian Funders that fund >1000 publications a year

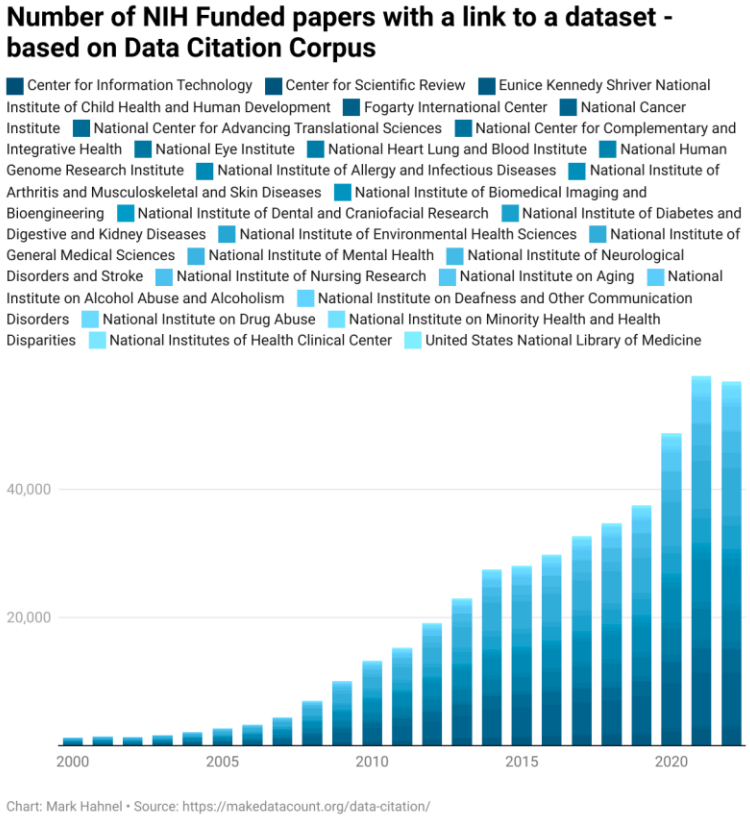


ARC: OA policy (2013), encourages open data.
CSIRO: OA policy (2016), open data policy (2016).
NHMRC: OA policy (2012), open data policy (2020).

figshare

A **Digital Science** Solution

Policy > Mandate > Compliance > **Measurement**



NIH Data Sharing Index (S-index) Challenge

Promoting data sharing and developing a robust metric to reward exemplary data sharers.

This Challenge aims to incentivize and reward data sharing excellence, promoting a new metric for assessing how effectively researchers share valuable data, driving a culture of openness in science.

Apply starting 04/21/25

Follow challenge (51)

Share

Submission period: Phase 1 opens on 04/21/25 01:00 PM GMT+1

Challenge type: Scientific

Total cash prizes: $1,000,000

Print challenge

DIGITAL science

Research Transformation

Change in the era of AI, open and impact:

voices from the academic community

Scan to download

Thank you

mark@figshare.com

@figshare

figshare.com