MINISTÈRE
DE L'ENSEIGNEMENT
SUPÉRIEUR
ET DE LA RECHERCHE
Liberté
Égalité
Fraternité

UNIVERSITÉ
DE LORRAINE

Inria

unesco

Building an Open Science Monitoring Framework with open technologies

19th of December
UNESCO headquarters

Opening Ceremony

Host: Laetitia Bracco (Université de Lorraine)

- Jaco du Toit, UNESCO, Chief of the Universal Access to Information and Digital Inclusion Section
- Shaofeng Hu, UNESCO, Director Division for Science Technology and Innovation Policy
- Marin Dacos, French Ministry of Higher Education and Research

MINISTÈRE
DE L'ENSEIGNEMENT
SUPÉRIEUR
ET DE LA RECHERCHE

Liberté
Égalité
Fraternité

Welcome words from France

Marin Dacos
French National Open Science Coordinator

A national public policy

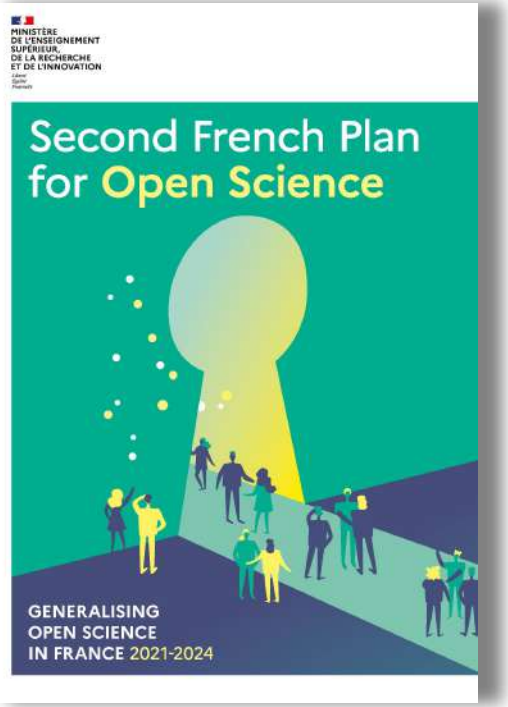
MINISTÈRE
DE L'ENSEIGNEMENT
SUPÉRIEUR
ET DE LA RECHERCHE

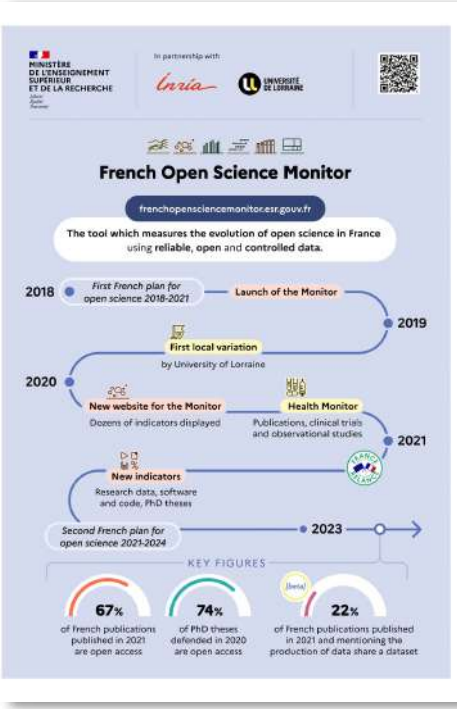
Liberté
Égalité
Fraternité

And the tools :

- French Open Science Committee
- French Open Science Fund
- French Open Science Monitor

https://www.ouvrirlascience.fr/home/





We need a global approach to monitor open science progress

In all domains,

NOT ONLY open access to publications

ALSO :

- research data,
- open source software,
- clinical trials,
- open science impacts,
- open science costs,
- etc.

The possible Open Science Monitoring Framework structure

In line with the future
Declaration on Open Research Information

1 - PRINCIPLES

2 - CORE INDICATORS

3- TECHNICAL SPECIFICATIONS

4- A GROUP OF STAKEHOLDERS BUILDING THE FRAMEWORK

We need you !

- You are the best experts
- We need your ideas, your imagination and your comments
- We need your contributions to build a consensual global open science monitoring framework

Thank you !

MINISTÈRE
DE L'ENSEIGNEMENT
SUPÉRIEUR
ET DE LA RECHERCHE

Liberté
Égalité
Fraternité

www.ouvrirlascience.fr

frenchopensciencemonitor.esr.gouv.fr/

marin.dacos@recherche.gouv.fr

Panel Discussion I - Large-scale Open Science monitoring initiatives

- Open science monitoring at NASA: Steve Crawford (NASA)
- Methodology of the French Open Science Monitor: Eric Jeangirard (French Ministry of Higher Education and Research)
- COKI- Curtin University's Open Knowledge Initiative: Cameron Neylon and Lucy Montgomery (Curtin University)
- Monitoring Open Science in the South, Arianna Becerril (Redalyc)

Open science monitoring at NASA

Steve Crawford, Science Data Officer (NASA)

Rachel Paseka, Chelle Gentemann, Jeamay Palo

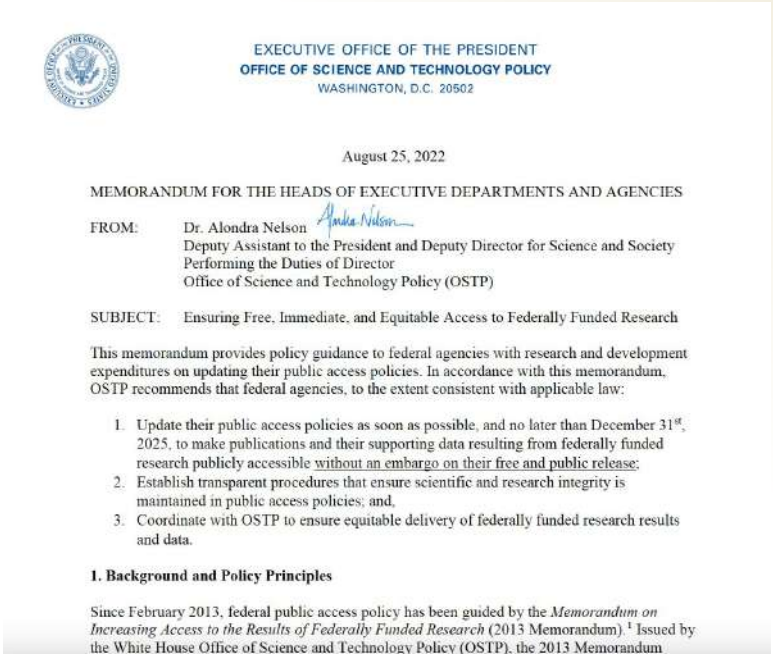
The White House announces

2023 A Year of Open Science

CDC ✦ DOA ✦ DOC ✦ DOE ✦ DOS ✦ DOT ✦ NASA ✦ NEH ✦ NIH ✦ NIST ✦ NOAA ✦ NSF ✦ SI ✦ USDA ✦ USGS

Open Science is the principle and practice of making research products and processes available to all, while respecting diverse cultures, maintaining security and privacy, and fostering collaborations, reproducibility and equity.

Ensuring Free, Immediate, and Equitable Access to Federal Funded Research



EXECUTIVE OFFICE OF THE PRESIDENT
OFFICE OF SCIENCE AND TECHNOLOGY POLICY
WASHINGTON, D.C. 20502

August 25, 2022

MEMORANDUM FOR THE HEADS OF EXECUTIVE DEPARTMENTS AND AGENCIES

FROM: Dr. Alondra Nelson
Deputy Assistant to the President and Deputy Director for Science and Society
Performing the Duties of Director
Office of Science and Technology Policy (OSTP)

SUBJECT: Ensuring Free, Immediate, and Equitable Access to Federally Funded Research

This memorandum provides policy guidance to federal agencies with research and development expenditures on updating their public access policies. In accordance with this memorandum, OSTP recommends that federal agencies, to the extent consistent with applicable law:

1. Update their public access policies as soon as possible, and no later than December 31st, 2025, to make publications and their supporting data resulting from federally funded research publicly accessible without an embargo on their free and public release;
2. Establish transparent procedures that ensure scientific and research integrity is maintained in public access policies; and,
3. Coordinate with OSTP to ensure equitable delivery of federally funded research results and data.

1. Background and Policy Principles

Since February 2013, federal public access policy has been guided by the *Memorandum on Increasing Access to the Results of Federally Funded Research* (2013 Memorandum).[1] Issued by the White House Office of Science and Technology Policy (OSTP), the 2013 Memorandum

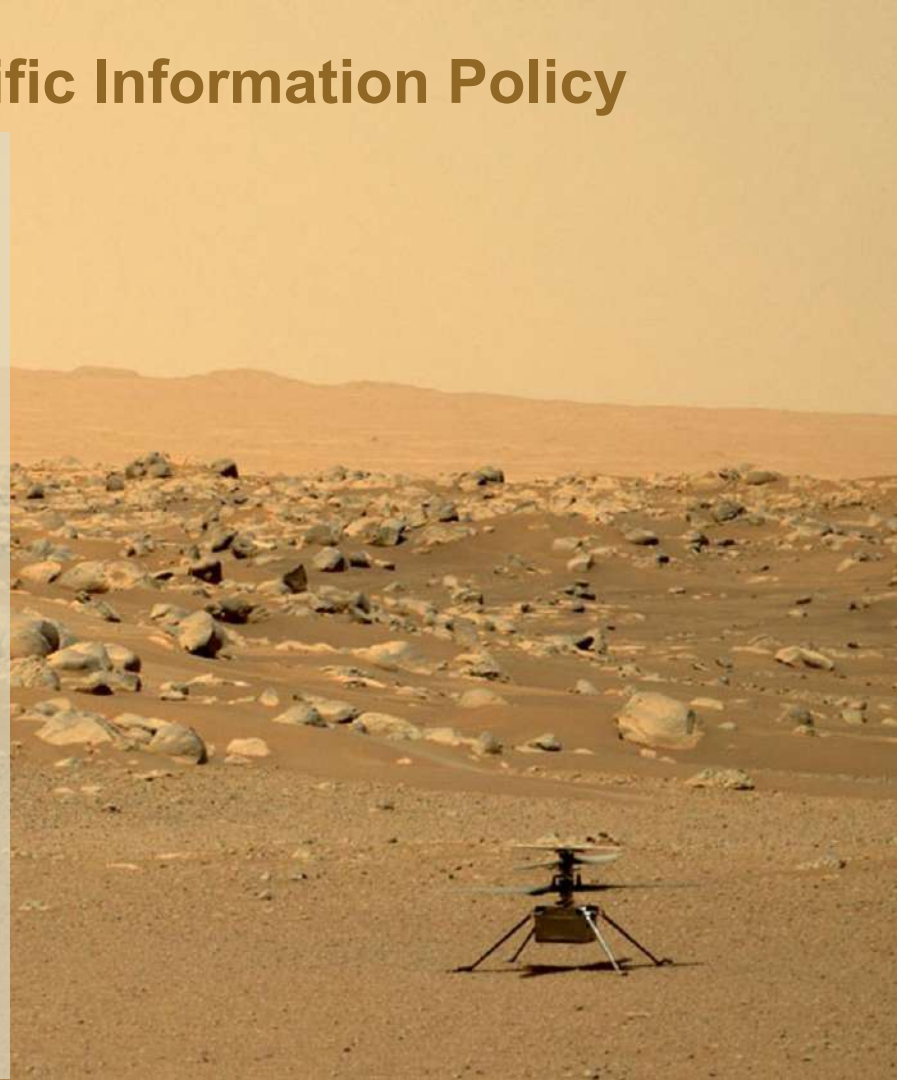
Released in August 2022 with the requirements that agencies update their Research Access plans to include **immediate and free access to publications and data** and to ensure research integrity and equity.

- Recommend standard consistent **benchmarks and metrics** to monitor implementation
- Regularly reporting back **statistics** on publications and implementation

NASA SMD's updated Scientific Information Policy

Major Policy Updates

- Peer-reviewed **publications are made openly available** with no embargo period.
- Research **data and software are shared** at the time of publication or the end of the funding award.
- Mission **data are released as soon as possible,** and unrestricted mission software is developed openly.
- Science **workshops and meetings are held openly** to enable broad participation.



Measuring Open Science Products

Quantitative measurements help assess and monitoring progress on the Open Science requirements.

- Publications
- Data
- Software
- Events

Challenges

How to determine the “denominator”?

How to assess derivative products?

How to track over the long term?

Opportunities

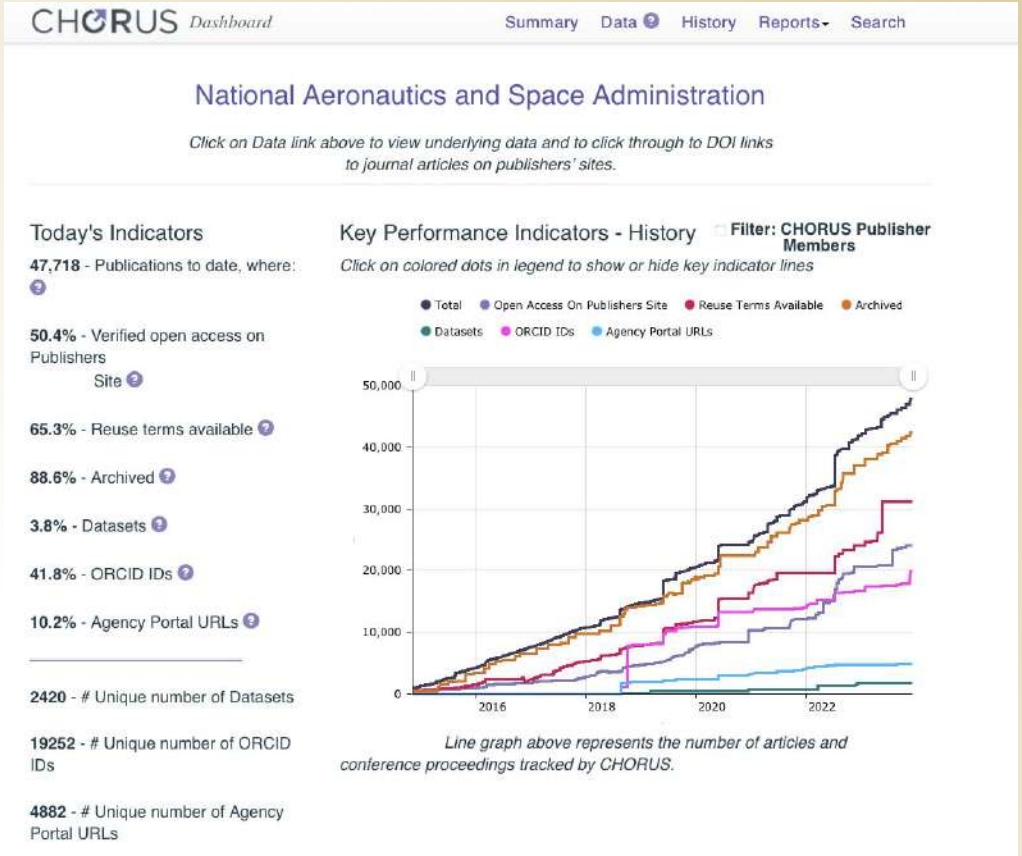
Use open science infrastructure to aid tracking with DOI’s on awards: Dept. of Energy example.

Use “living” Data Management Plans: NASA Task Book.

CHORUS

CHORUS helps funders, institutions, publishers, societies, and the public see, find, and understand the **status of outputs of funded research.**

CHORUS provides NASA with a list of publications from CHORUS partner publishers that are funded by NASA and publications are ingested into NASA's PubSpace repository.



https://dashboard.chorusaccess.org/nasa#/summary

SciX

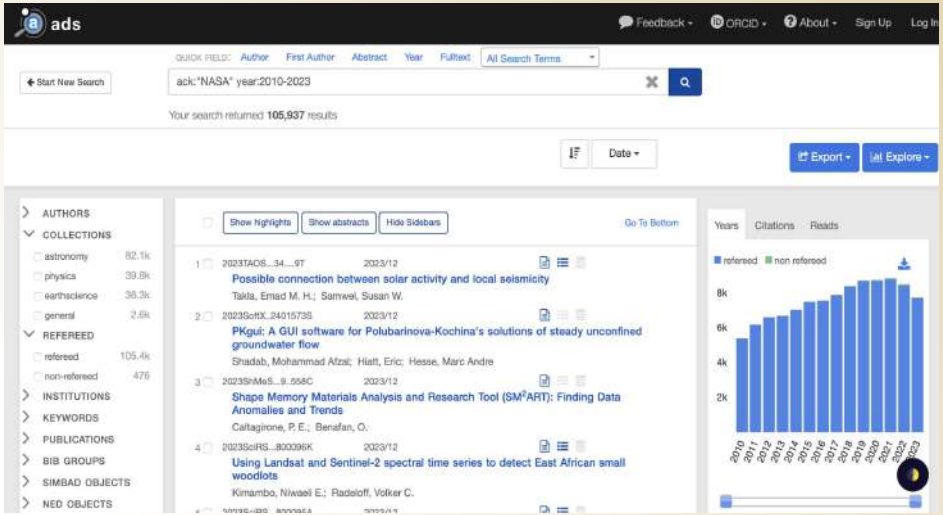
[SciXplorer.org]

NASA SciX is a literature-based, **open digital information system** covering the fields of Astrophysics, Planetary Science, Heliophysics, Earth Science, and NASA space-based experiments.

It can be used to identify NASA funded research in Earth and Space Science.

Beta version is now available.

https://scixplorer.org/



Example search based on acknowledgements from the ADS, from which SciX is developed: https://ui.adsabs.harvard.edu/search/q=ack%3A%22NASA%22%20year%3A2010-2023&sort=date%20desc%2C%20bibcode%20desc&p_=0

Measuring Software Impact

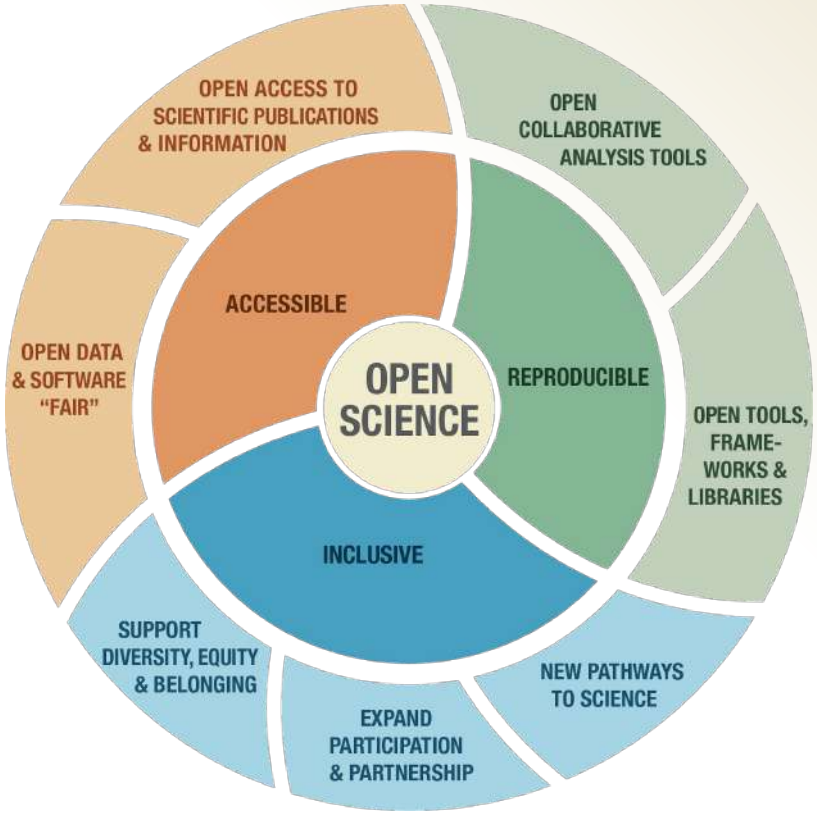
Along with data and publications, NASA also mandates the release of scientific software developed as part of a publication. Credit and recognition for scientific software is relatively new.

Possible metrics and measurement methods:

- Downloads, installations, usage, imports/dependencies, contributors, services
- Citations, Altmetrics, ImpactStory, libaries.io, Software Heritage
- Usage by Missions, observatories, research centers, or data repositories
- Funding, awards, or other incentives or recognition.
- Recognizing contributors and level of contribution.

Adopted from D. Katz https://zenodo.org/records/4058718

Examples: Mars Perseverance Rover,. Astropy, Journal of Open Source Software



Measuring the Impact of Open Science

How do we make science more **accessible, inclusive, and reproducible**? How do we assess if we are successful?

Example: NASA ROSES Yearbook shares proposers statistics.

Activities:
Listening sessions, impact assessments, community dialogue

NASA's Transform to Open Science (TOPS)

A 5-year mission to accelerate adoption of open science

Goals:

- Increase understanding and adoption of open science principles and techniques
- Broaden participation by historically excluded communities
- Accelerate scientific discovery

Open Science 101

A community-developed introduction to **core open science skills** released on Dec 6!

https://nasa.github.io/Transform-to-Open-Science/

Methodology of the French open science monitor

FrenchOpenScienceMonitor.esr.gouv.fr

Eric Jeangirard (French Ministry of Higher Education and Research)

Objectives of the French Open Science Monitor

Open science monitoring has become instrumental in France since the launch of the **National Plan for Open Science** in July 2018.

A **sovereign** and **evolving** tool was needed for **assessing the impacts** of the open science **public policy.**

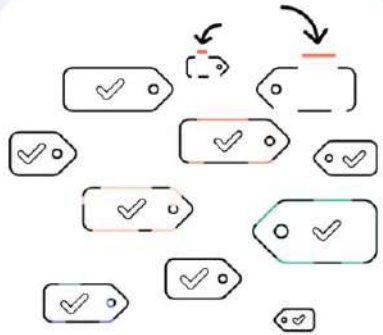
This tool had to transparent and reproducible, and thus **completely independent from any proprietary datasource**.

First focus on open access to publications, but now also covers PhD thesis, dataset, research software and clinical trials.

As of 2018: the metadata gap between proprietary and open

Proprietary bibliographic databases

remedy these defects
by enriching these metadata

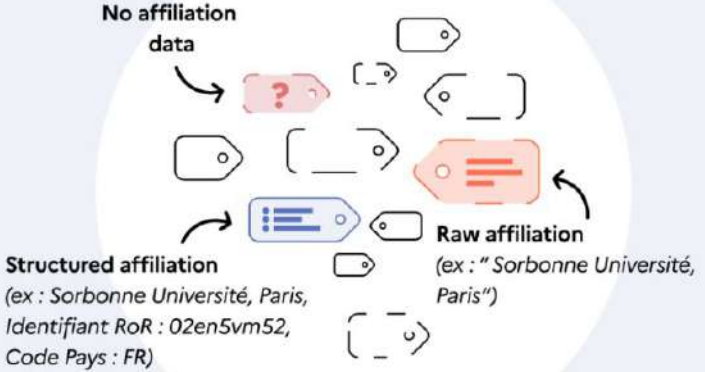


Proprietary bibliographic databases:

- are not shareable under an open license
- are biased and do not allow the bibliodiversity of the production to be taken into account

Open bibliographic databases

offer a low amount of affiliation metadata
and of disparate quality



Open bibliographic databases make it possible to share and reuse data, even to build new services on shared data

MINISTÈRE
DE L'ENSEIGNEMENT
SUPÉRIEUR
ET DE LA RECHERCHE

As of 2018: the metadata gap between proprietary and open

Open metadata does not exist ?

Let's try to create it!

How ?

- machine learning
- cloud computing
- … and some common sense

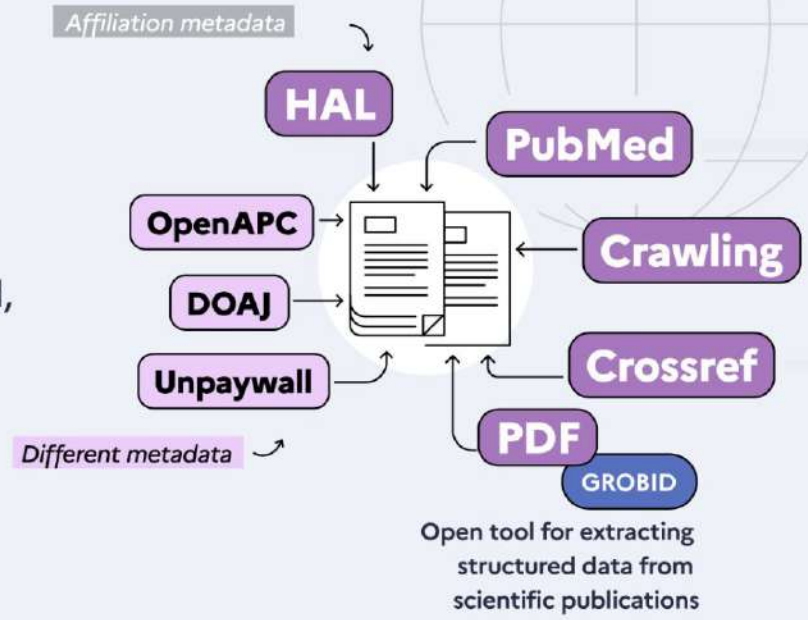
Our open methodology

For each publication in the world, we have chosen to collect as much affiliation metadata as possible, using **a variety of open sources**. Our idiosyncrasy: no use of proprietary databases.

#1 Collect

as much metadata as possible

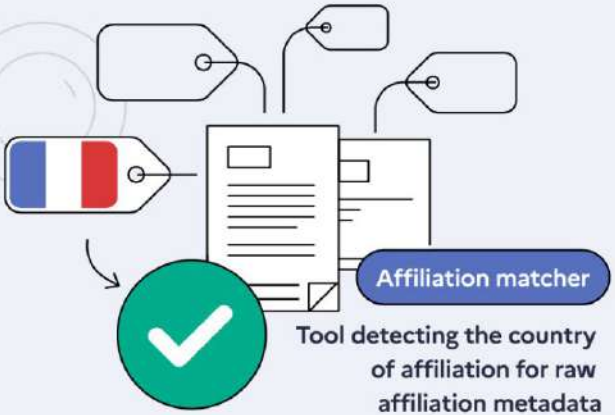
For each individual publication in the world, a variety of sources aggregated.



#2 Detect

the country of affiliation

Publications are filtered to exclusively retain those with at least one French affiliation.



"Sorbonne Université, Paris" → *France* ✓

"Hotel Dieu de France, Beirut, Lebanon" → *Liban* ✗

Detection rate of french scientific publications

90% The Monitor's methodology has enabled to establish to this day **the most comprehensive database for French publications in the world***.

60%
for a worldwide standard tool, the Web of Science (WoS).

Database of French scientific publications

170 000/year

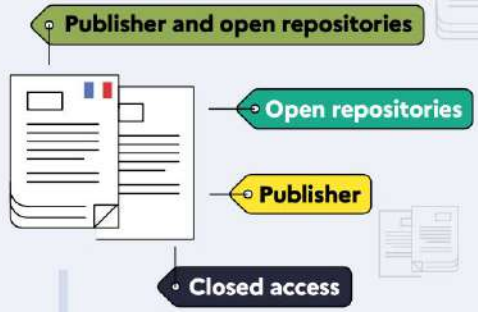
* Lauranne Chaignon, Daniel Egret; Identifying scientific publications countrywide and measuring their open access: The case of the French Open Science Barometer (BSO). Quantitative Science Studies 2022; 3 (1): 18–36. doi: doi.org/10.1162/qss_a_00179

#3 Enhance...

... the opening status

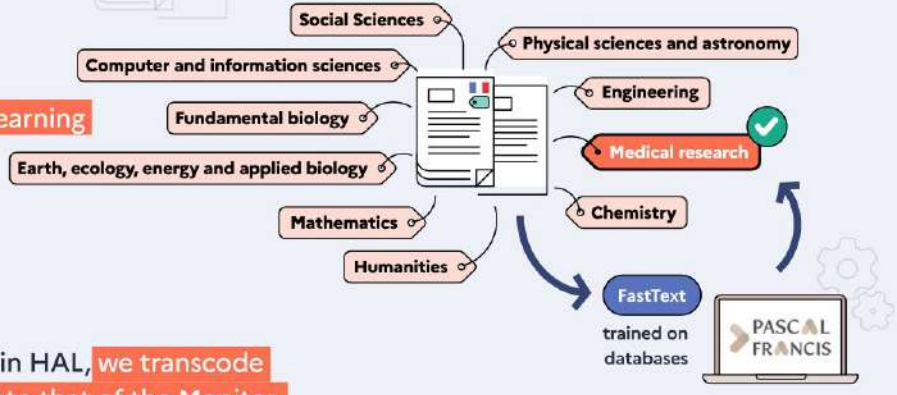
For crossref DOI:
the information stems from **Unpaywall**

For publications in HAL (no DOI):
the information stems from **HAL**



... the disciplinary classification

Via an automatic classification machine learning algorithm (fastText) using titles, summary and name of journal.



If metadata is available in HAL, we transcode the HAL classification into that of the Monitor.

Custom KPI designed to steer our public policy

⇒ Mixed OA route (publisher and open repository) highlighted

⇒ Focus on Diamond

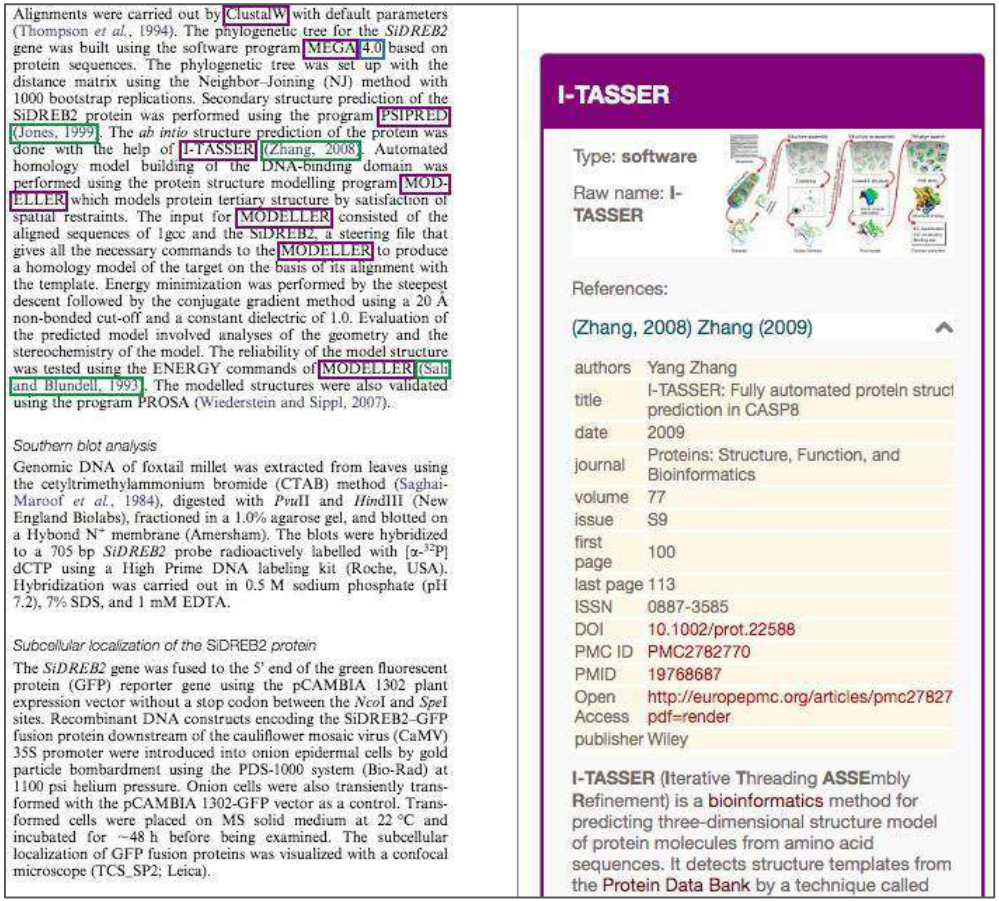
⇒ National APC expenditures estimates

Beyond publications: monitoring clinical trials, datasets and software

- **Clinical trials transparency using public registries (european and american)**

- **Dataset and software**
  - Trained on 4,971 manually annotated documents (37 annotators)
  - https://github.com/softcite

  - Automatic characterization of mentions: **used / created / shared**
    - Trained on 3,643 manually annotated sentences



Alignments were carried out by ClustalW with default parameters (Thompson *et al.*, 1994). The phylogenetic tree for the *SiDREB2* gene was built using the software program MEGA 4.0 based on protein sequences. The phylogenetic tree was set up with the distance matrix using the Neighbor–Joining (NJ) method with 1000 bootstrap replications. Secondary structure prediction of the SiDREB2 protein was performed using the program PSIPRED (Jones, 1999). The *ab intio* structure prediction of the protein was done with the help of I-TASSER (Zhang, 2008). Automated homology model building of the DNA-binding domain was performed using the protein structure modelling program MODELLER which models protein tertiary structure by satisfaction of spatial restraints. The input for MODELLER consisted of the aligned sequences of 1gcc and the SiDREB2, a steering file that gives all the necessary commands to the MODELLER to produce a homology model of the target on the basis of its alignment with the template. Energy minimization was performed by the steepest descent followed by the conjugate gradient method using a 20 Å non-bonded cut-off and a constant dielectric of 1.0. Evaluation of the predicted model involved analyses of the geometry and the stereochemistry of the model. The reliability of the model structure was tested using the ENERGY commands of MODELLER (Sali and Blundell, 1993). The modelled structures were also validated using the program PROSA (Wiederstein and Sippl, 2007).

Southern blot analysis

Genomic DNA of foxtail millet was extracted from leaves using the cetyltrimethylammonium bromide (CTAB) method (Saghai-Maroof *et al.*, 1984), digested with *Pvu*II and *Hin*dIII (New England Biolabs), fractioned in a 1.0% agarose gel, and blotted on a Hybond N$^+$ membrane (Amersham). The blots were hybridized to a 705 bp *SiDREB2* probe radioactively labelled with [α-^{32}P] dCTP using a High Prime DNA labeling kit (Roche, USA). Hybridization was carried out in 0.5 M sodium phosphate (pH 7.2), 7% SDS, and 1 mM EDTA.

Subcellular localization of the SiDREB2 *protein*

The *SiDREB2* gene was fused to the 5′ end of the green fluorescent protein (GFP) reporter gene using the pCAMBIA 1302 plant expression vector without a stop codon between the *Nco*I and *Spe*I sites. Recombinant DNA constructs encoding the SiDREB2–GFP fusion protein downstream of the cauliflower mosaic virus (CaMV) 35S promoter were introduced into onion epidermal cells by gold particle bombardment using the PDS-1000 system (Bio-Rad) at 1100 psi helium pressure. Onion cells were also transiently transformed with the pCAMBIA 1302-GFP vector as a control. Transformed cells were placed on MS solid medium at 22 °C and incubated for ~48 h before being examined. The subcellular localization of GFP fusion proteins was visualized with a confocal microscope (TCS_SP2; Leica).

I-TASSER

Type: **software**

Raw name: **I-TASSER**

References:

(Zhang, 2008) Zhang (2009)

authors	Yang Zhang
title	I-TASSER: Fully automated protein struct prediction in CASP8
date	2009
journal	Proteins: Structure, Function, and Bioinformatics
volume	77
issue	S9
first page	100
last page	113
ISSN	0887-3585
DOI	10.1002/prot.22588
PMC ID	PMC2782770
PMID	19768687
Open Access	http://europepmc.org/articles/pmc27827 pdf=render
publisher	Wiley

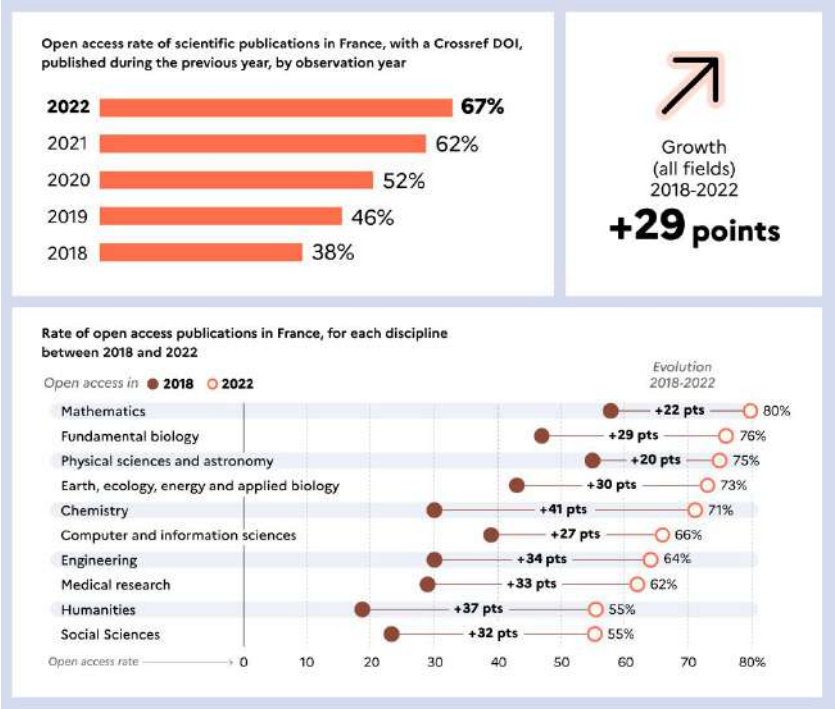
I-TASSER (Iterative Threading **ASSE**mbly **R**efinement) is a bioinformatics method for predicting three-dimensional structure model of protein molecules from amino acid sequences. It detects structure templates from the Protein Data Bank by a technique called

MINISTÈRE DE L'ENSEIGNEMENT SUPÉRIEUR ET DE LA RECHERCHE

Main results of the French Open Science Monitor

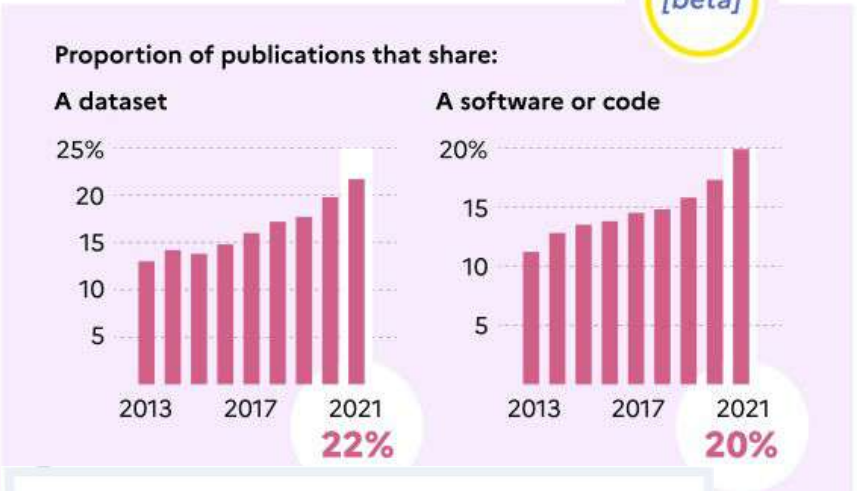
FrenchOpenScienceMonitor.esr.gouv.fr

Publications



Research data and software

[beta]



Lessons learnt

- It is possible to build an open science monitor at the national level without any proprietary data, taking advantage of the progress in **machine learning** and **cloud computing**.

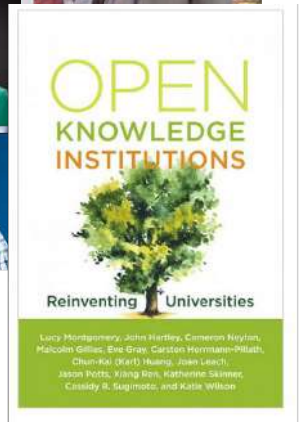
- An iterative process is needed to improve and extend the results

- Collaborating at different scales is key
  - Necessity to be complementary and not to reinvent the wheel
  - From national to local
  - From national to international: Open initiatives exist, like OpenAlex or COKI and others. There is room to coordinate so that (open) data quality improves globally and cutting-edge detection methods are shared

COKI – Curtin University's Open Knowledge Initiative

Lucy Montgomery and Cameron Neylon
(Curtin University)

THE COKI PROJECT



- Curtin Open Knowledge Initiative
- Commenced in 2018
- Curtin strategic initiative
- ~$10M in funding
- Founded in the Centre for Culture and Technology
- Collaboration with Curtin Institute for Data Science

COKI CURTIN OPEN KNOWLEDGE INITIATIVE

Curtin University

THE AUSTRALIAN CONTEXT

Wednesday, 15 November 2023

Media statement by Australia's Chief Scientist, Dr Cathy Foley

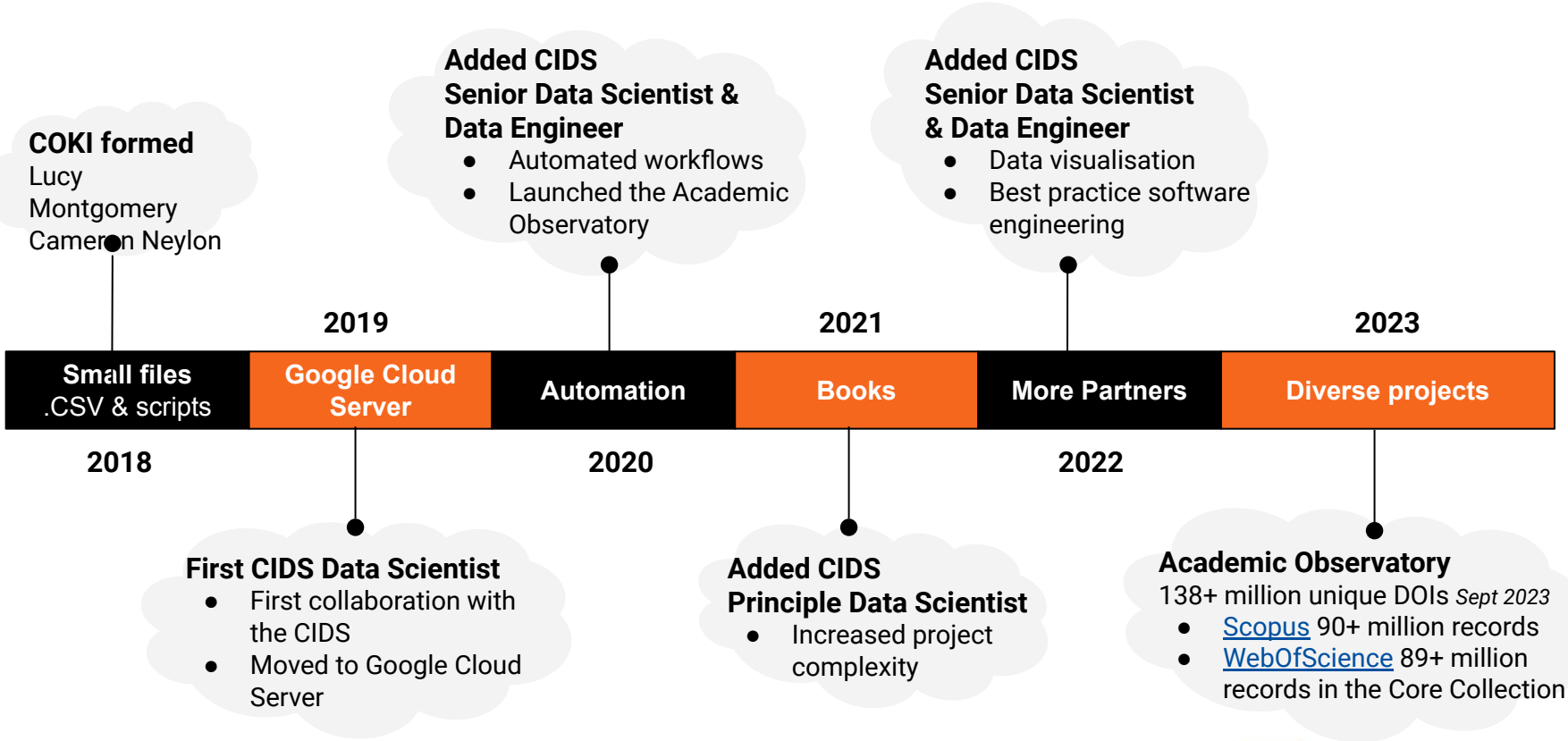
"The current system for assessing research careers for hiring, promotion and funding is not fit for purpose."

"The current practices do not incentivise innovation or multidisciplinary research, nor recognise the breadth of roles in a healthy science and research system."

- 2021: Australia's Chief Scientist declares open science a priority
- 2022 - 24: Major review(s) of research sector following change of government
- 2023: Critical review of research evaluation
- 2024: New "accord" between government and universities to be announced

https://www.chiefscientist.gov.au/news-and-media/australias-systems-assessing-research-careers-not-fit-purpose

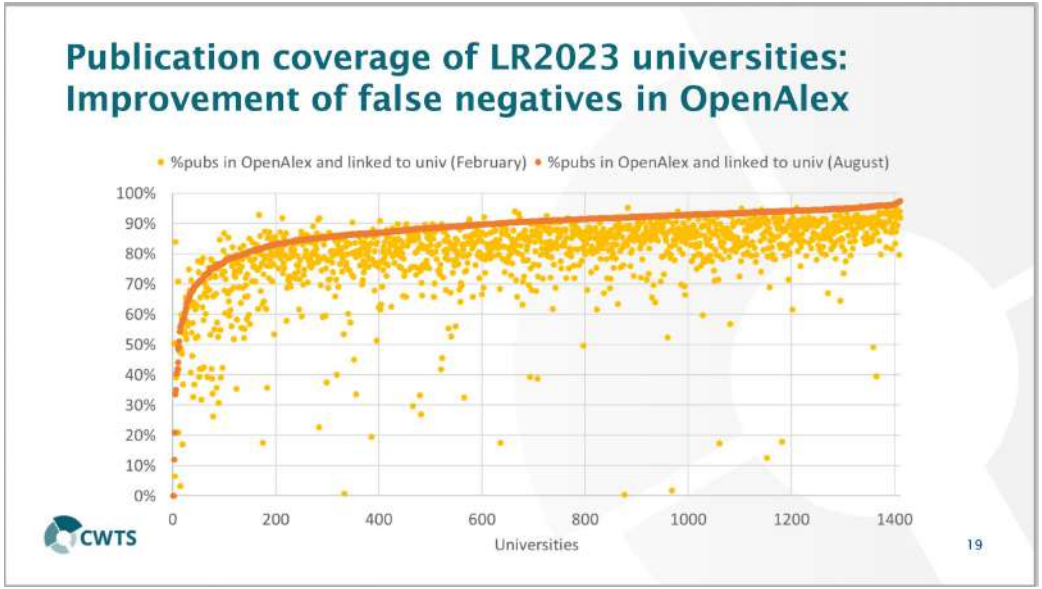
COKI & Curtin Institute for Data Science



CURTIN OPEN KNOWLEDGE INITIATIVE

Curtin University

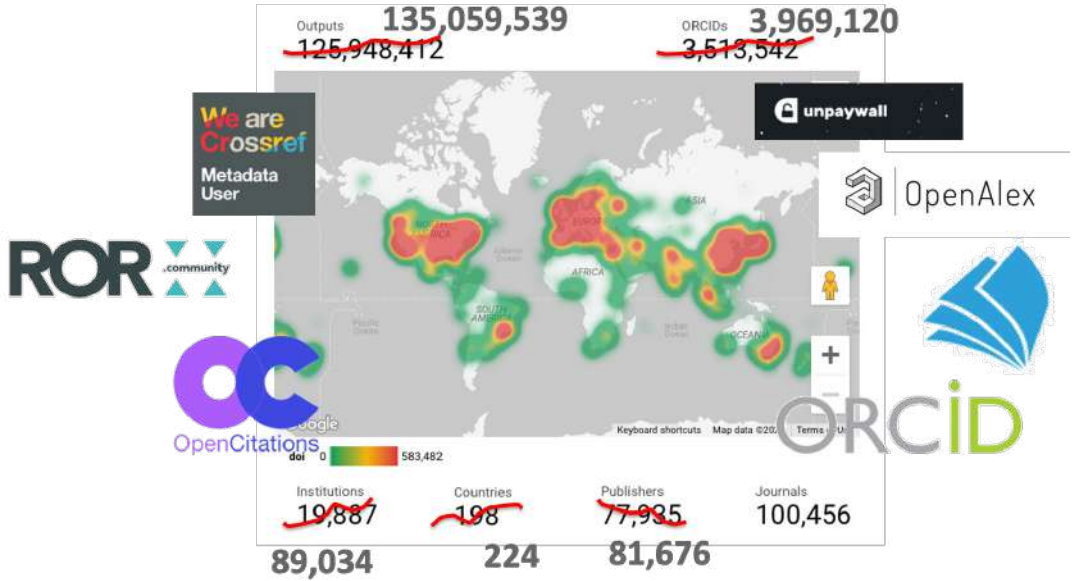
MEANWHILE…NEW OPPORTUNITIES



- Opportunities for community curation of open data sets
- Working with the open research information community to understand and improve open data sets and explore their possibilities

Van Eck, N. J. (2023, November 9). Leiden Ranking Open Edition. Zenodo. https://doi.org/10.5281/zenodo.10107263

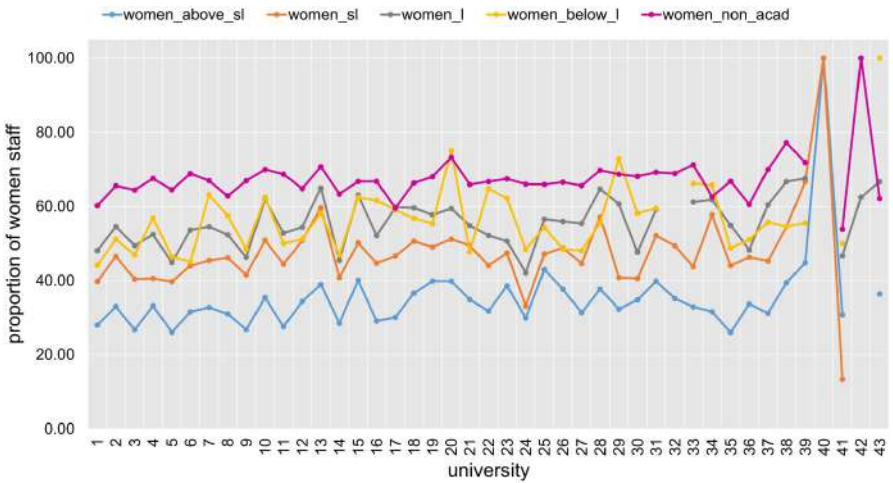
MEANWHILE...NEW OPPORTUNITIES

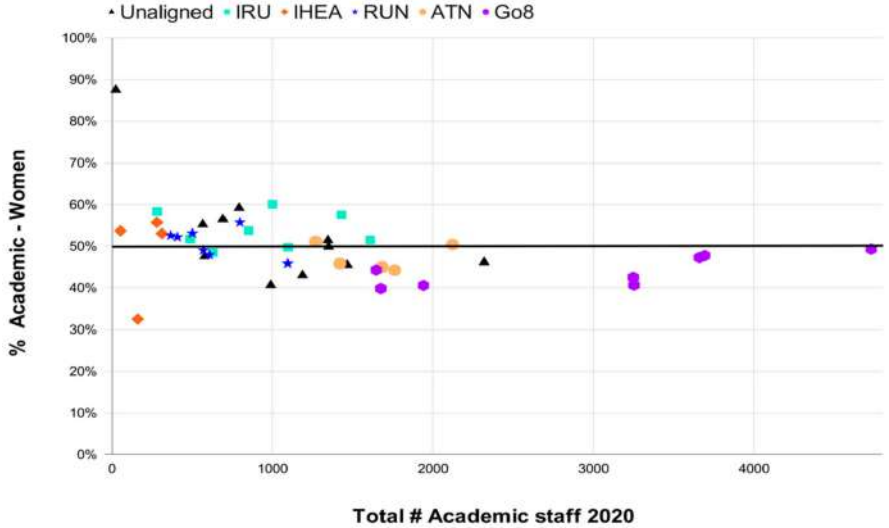


- Open Data landscape is now more comprehensive than proprietary
- We can do “open science on open science” with open data, open source, open systems

RESEARCH

- *Diversity* in knowledge production and dissemination:

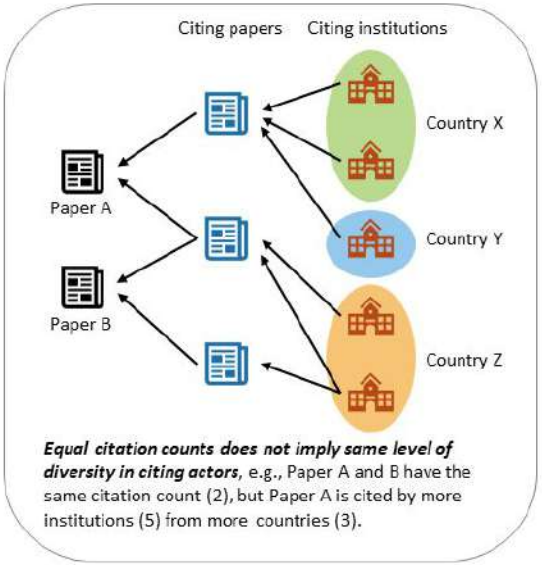




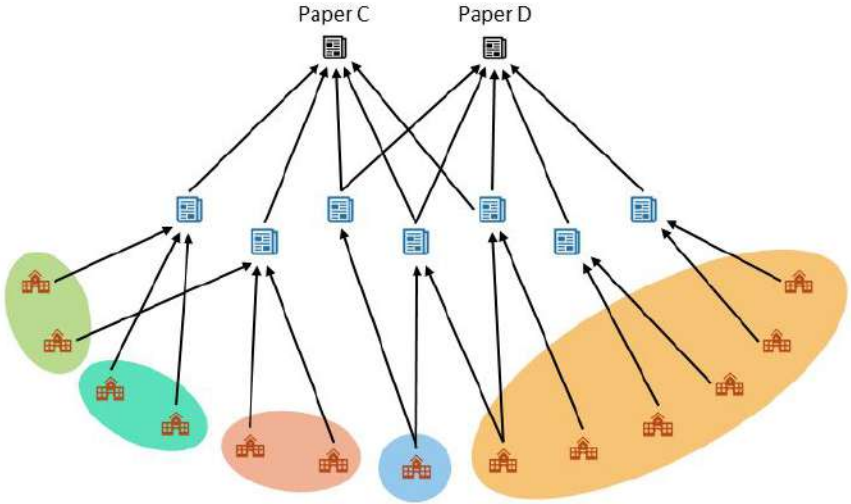
(LEFT) Huang et al. (2021). Mapping open knowledge institutions: an exploratory analysis of Australian universities. PeerJ 9:e11391 https://doi.org/10.7717/peerj.11391
(RIGHT) Wilson et al. (2022). Changing the Academic Gender Narrative through Open Access. Publications, 10, 22. https://doi.org/10.3390/publications10030022

RESEARCH

- Thinking about ***citations*** in a different way:



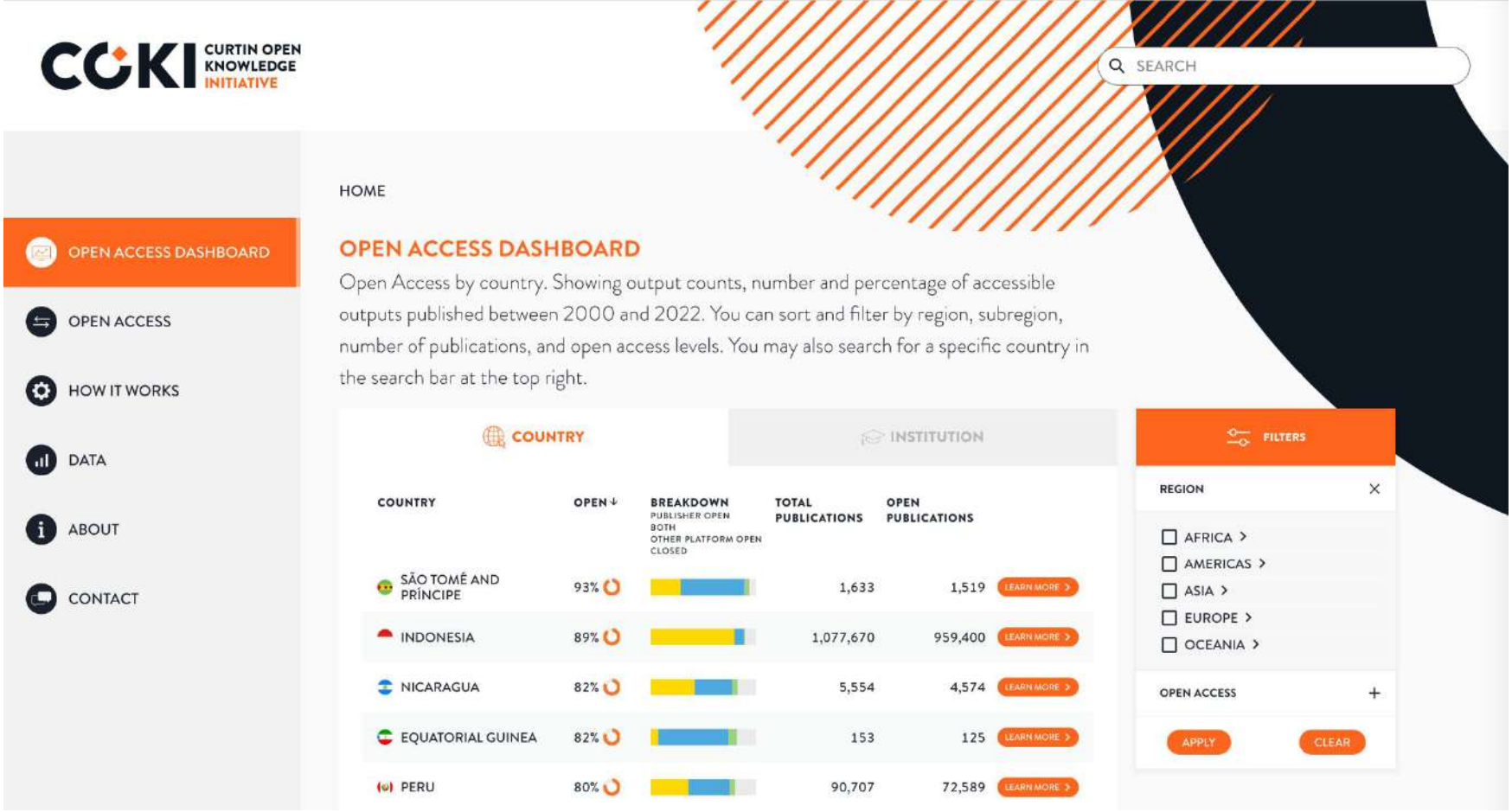
Equal citation counts does not imply same level of diversity in citing actors, e.g., Paper A and B have the same citation count (2), but Paper A is cited by more institutions (5) from more countries (3).



Paper	# Citations	# Citing institutions	# Citing countries	Gini-Simpson*	Shannon*
C	5	9	5	0.79	1.59
D	5	7	2	0.55	0.53

* Calculated based on grouping all "citing paper to citing institution" pairs by countries.

Huang et al. (2022). Open Access Research Outputs Receive More Diverse Citations (Version 2). Zenodo. https://doi.org/10.5281/zenodo.7099438


CURTIN OPEN KNOWLEDGE INITIATIVE
SEARCH
HOME
OPEN ACCESS DASHBOARD
OPEN ACCESS
HOW IT WORKS
DATA
ABOUT
CONTACT
OPEN ACCESS DASHBOARD
Open Access by country. Showing output counts, number and percentage of accessible outputs published between 2000 and 2022. You can sort and filter by region, subregion, number of publications, and open access levels. You may also search for a specific country in the search bar at the top right.
COUNTRY
INSTITUTION
FILTERS
COUNTRY
OPEN
BREAKDOWN
PUBLISHER OPEN
BOTH
OTHER PLATFORM OPEN
CLOSED
TOTAL PUBLICATIONS
OPEN PUBLICATIONS
SÃO TOMÉ AND PRÍNCIPE 93% 1,633 1,519 LEARN MORE
INDONESIA 89% 1,077,670 959,400 LEARN MORE
NICARAGUA 82% 5,554 4,574 LEARN MORE
EQUATORIAL GUINEA 82% 153 125 LEARN MORE
PERU 80% 90,707 72,589 LEARN MORE
REGION
AFRICA
AMERICAS
ASIA
EUROPE
OCEANIA
OPEN ACCESS
APPLY
CLEAR

Monitoring Open Science in the South

Arianna Becerril García

Autonomous University of the State of Mexico

Building an Open Science monitoring framework with open technologies

Monitoring Open Science in the South

Monitoring Open Science from the essencial values of science

Difference between South to North?

Mertonian norms (Universalism, communality, ...)

Monitoring when open is the default

(exclusion, inequity, losses, distortions)

Monitoring Open Science from the global public good

(non-excludable, non-rivalrous)

AmeliCA

When open is the default,

what should be monitored?

A different starting point in Latin America to reflect on Open Science monitoring

AmeliCA

Distributed investment **Universal benefit**

12 country-level networks of institutional repositories

AmeliCA latindex

1.460 diamond OA journals published by 700 academic institutions

Natural Open Access

>3.000 OJS installations

No fees neither for authors nor for readers

Owned by academic sector

Nonprofit

Academic journal publishing

~12.000 online journals

~2.700 quality-certified journals

63 OA mandates

4 national mandates (AR, MX, BR, PE)

Institutional journal portals and repositories

Nonprofit platforms and infrastructures for capacity building and sustainability

A different starting point in Latin America to Openness

Distributed investment

Universal benefit

What it should be monitored?

Where it should be monitored?

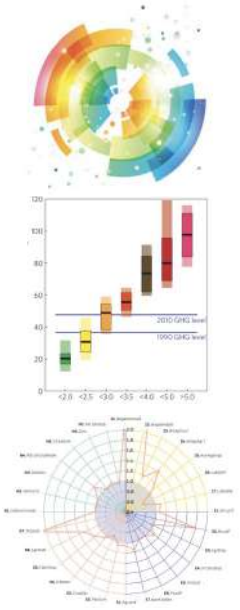
AmeliCA

Distributed investment

Not expenditure

Investment	Infrastructure indicators	Business models

• Implications from the commodification
• Non-commercial ecosystem degradation
  • Total number of journals sold to commercial publishers
  • Total number of journals flipped to APC
• Investment in Open Science infrastructure
  • Repositories
  • Journals' sustainability
  • Open data
• R&D Personnel (FTE)
• Labour force (FTE)
	• Open infrastructures sustained by the country
	• Open infrastructures sustained by the institution
	• Open infrastructures services
	• Open infrastructures contribution to O
		• Sustainability
		• Evolution
		• Market forces



Universal benefit

Exclusion

- Total number of beneficiaries
- Beneficiaries by country
- Beneficiaries by language
- Beneficiaries by gender
- Beneficiaries by race
- Beneficiaries by age
- Beneficiaries by ethnical condition

Gaps

- Access gap
- Author gap
- Gender gap
- Digital gap
- Intellectual property monitoring

Opennes should be means to an end.
Openness will grow in indicators always.
Growth indicators hide or make invisible gaps and the ones left behind.
Growth indicators hide or make invisible the consolidation/decline of models.

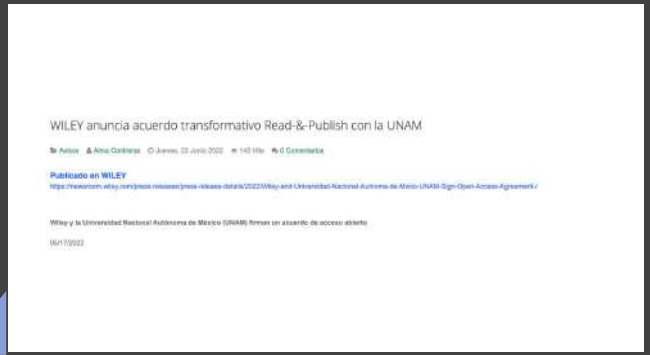
First adopted in Europe, the wave of TAs has now reached libraries in Asia, Africa, the Americas and Australia. With more than half a million new research articles published openly through TAs negotiated by institutions in 67 countries to date, there can be no doubt that TAs increase global access to research.

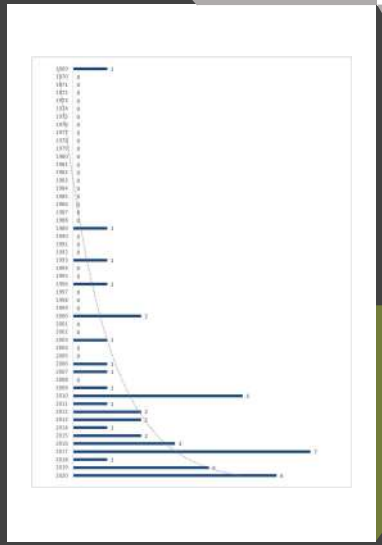
Colleen Campbell, Ádám Dér, Kai Geschuhn and Ana Valente (2022). How are transformative agreements transforming libraries? IFLA, WLIC, Dublin

Exclusion is not monitored

Through key "transformation drivers", characteristic of TAs, libraries, globally, are advancing toward a fully open paradigm in scholarly communication

The lack of investment in OPEN infrastructure is not monitored





Commodification of science in Latin America

- Alarming weakening of Diamond OA.
- In the last 2 years more than 100 journals have started to charge authors.
- Journals being acquired by commercial publishers, clear acceleration in the last decade.

AmeliCA datOS

A framework for data sharing and monitoring of OpenScience Interoperability for an inclusive Open Science and to move forward research assessment, a map that represent the knowledge that is being generated and circulated in non-commercial channels.

Platforms, data providers

Data availability:
APIs / WS / OAI-PMH

Interoperability guidelines and a technical committee

Data collection and integration

Evolution of research assessment, innovation, discoverability improved, CRIS …

- Open infrastructures
- Institutional repositories
- National networks of repositories
- National Systems

- Decentralization
- Independence sharing a goal: sharing data for research evaluation

- Syntactic layer
- Semantic layer
- Shared vocabularies
- Metadata standards
- Persistent identifiers

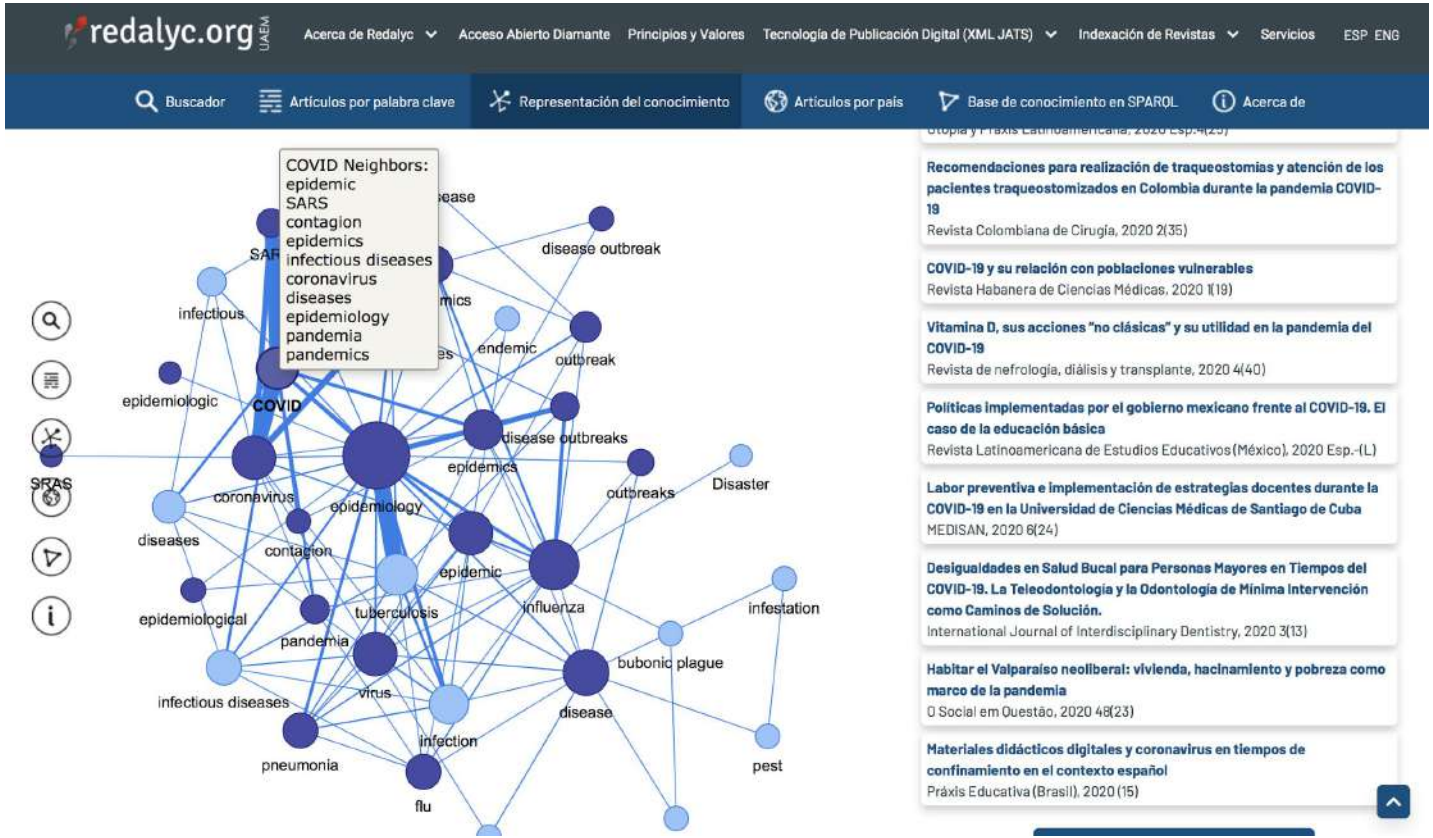
Transparent and effective integration
- By using international standards, the integration is optimized
- No normalization processes are necessary since common vocabularies are used
- Redundant or duplicate production is identified on the various platforms

Sustainability
- Update in real time
- Distributed system that respects the independence of the initiatives, however they share a common language
- Compatible with the different routes of Open Access
- Extensible to other branches of Open Science

Compatibility
- Compatible with initiatives from other regions and other Open Access pathways.
- Compatible with traditional bibliometrics, which allows the integration of Latin American production in international bibliometric studies, rankings, and developments that impact the evaluation of research.

Principles & Values: Inclusion, non-commercial, scholar-owned, multilingüalism, diversity

Unveil the structure of knowledge



Evolution of Openness?

Let’s monitor the evolution, the (harmful) effects of different strategies, the OS models consolidation/decrease,

Access to knowledge, processability, deposit, text mining, processes, …

From knowledge to ‘solve’ to knowledge to publish

The harmful transition from communication to commodification

Control of scientific circuit Ownership

Open Science should be monitored as a mean to an end

Thank you

Arianna Becerril García
arianna.becerril@redalyc.org
http://www.redalyc.org/autor.oa?id=25
http://orcid.org/0000-0003-0278-8295
@ariannabec

redalyc UAEMéx AmeliCA

Panel Discussion II - Issues and opportunities for monitoring Open Science

- Scientific Knowledge Graph beyond proprietary data with OpenAlex, Jason Priem (OpenAlex)
- Mining software and research dataset creations, sharing and citations in scientific literature, Patrice Lopez (science-miner)
- Monitoring the opening of protocols and clinical study reports, Inge Stegeman (UMC Utrecht)

Scientific Knowledge Graph beyond proprietary data with OpenAlex

Jason Priem (OpenAlex)

OpenAlex is…

An open index of the world’s research system

Why OpenAlex?

- bigger (240M works)
- easier to use (webapp, API, database)
- **OPEN** (code, data)

OpenAlex has:

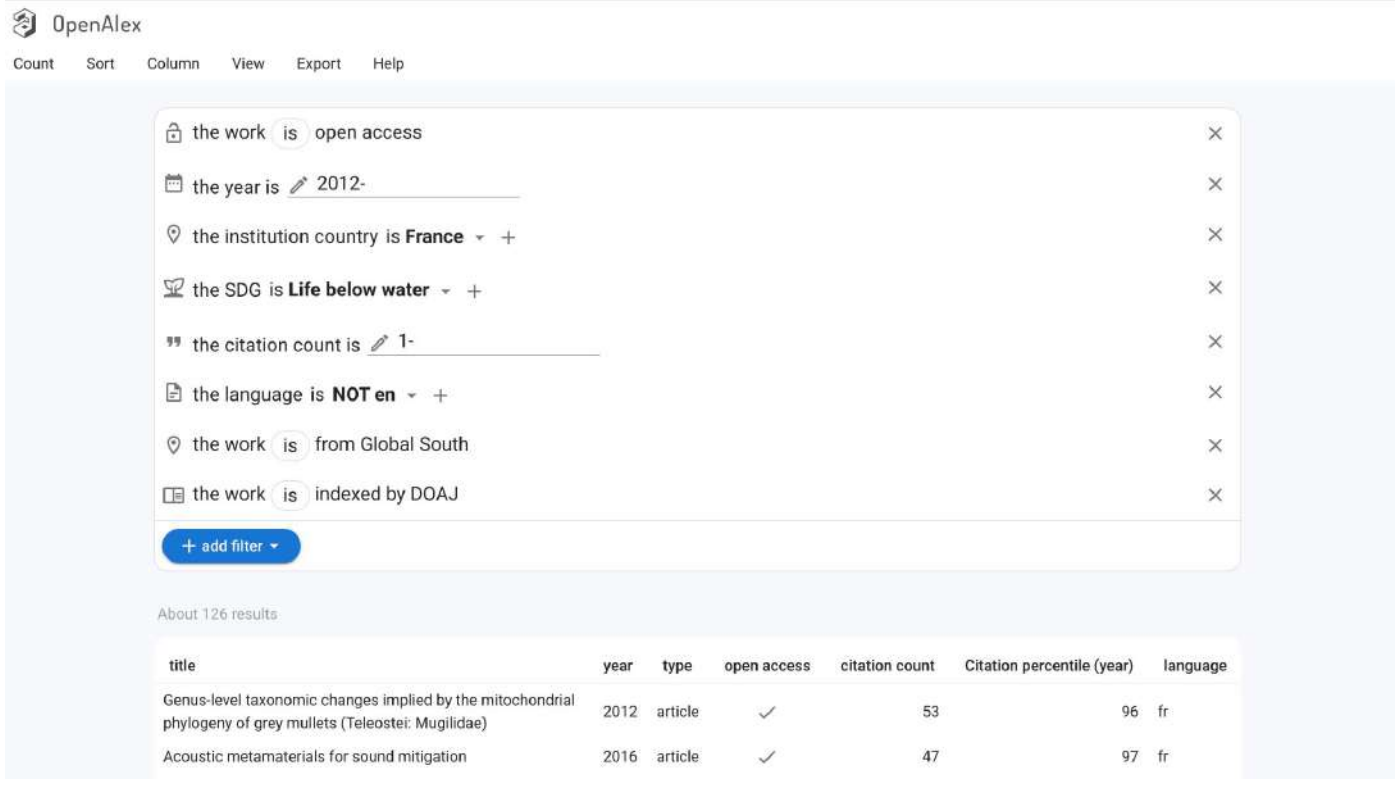
- Works (articles, books, datasets)
- Authors
- Sources (journals, repositories)
- Publishers
- Funders
- Institutions (universities, centres)
- Concepts (fields, topics, keywords)

OpenAlex data comes from:

- Crossref
- PubMed
- ORCID
- ROR
- OpenAPC
- DOAJ

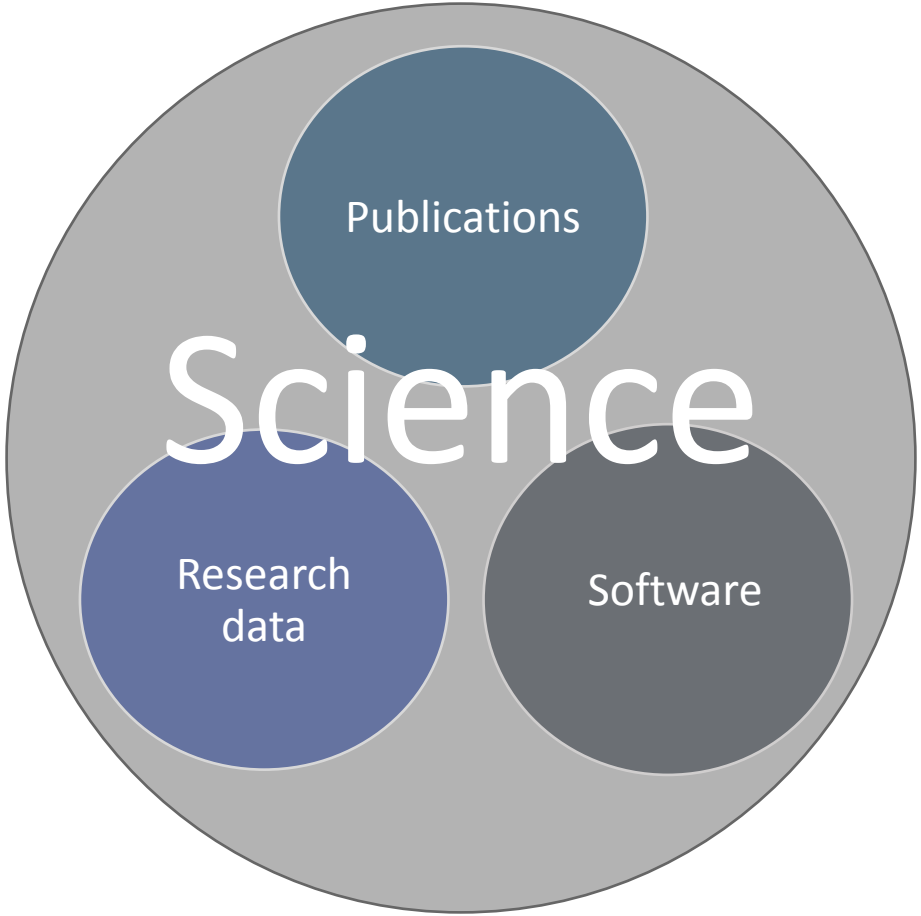
- Repositories:
  - institutional
  - national (HAL)
  - disciplinary (ArXiv)

Lots of filters



Mining software and research datasets creations, sharing and citations in scientific literature

Patrice Lopez (science-miner)


Publications
Science
Research data
Software

Understanding research datasets

Research data repositories ?

Data repositories via DataCite suffer from many limitations:

- Data repositories only inform about shared datasets
  - They do not cover mainstream databases & accession numbers, e.g. GenBank, PDB, PubChem
- Metadata debt: lack of affiliation and domain information for meaningful indicators
  - Granularity issues: 1 dataset with 10,000 images can give 10,000 DOI of type “dataset”
  - Deposits of datasets in repositories are often not correlated with actual data production

Only around 10% of dataset mentions in articles had PID in 2017 **[4]**

… and most datasets are mostly unnamed and not shared, e.g.:

*“**data** were recorded using an MR-compatible 32-channel BrainAmp MR plus amplifier.”*

Following research software activities

Software development in research is collaborative and distributed:

- Many platforms and catalogs/registries, no central metadata repository
  - Software are not data. Open Source software are made to evolve: pull request, versions, fork, etc.
  - How to identify software relevant to research?

Software citations are mostly informal, only 1-8% of mentions as bibliographic references **[2,3]**

PID are still not taking off: 0-0.6% of mentions with PID in 2022-2023 **[2,3]**

118,403 software entries on Zenodo, mostly via GitHub integration - but a large number without usable metadata

Mining data and software activities in scholarly full texts

Publications can be used as **proxies** to the dataset and software usage, creation and sharing:

1) **Text mining of dataset and software mentions in the full texts**

   ➡ Ensures data and software are related to actual research works
   ➡ Make possible to rely on document metadata to produce meaningful indicators
   ➡ Scalable and representative

2) **Automatic characterization of the mention context:** is a mentioned dataset or software **used/created/shared** ?

   ➡ Insights on the role the mentioned dataset or software wrt. the research work

Text mining data and software mentions in scholarly publications is complicated

- **PDF format** is mandatory, but hard to support
- **Sparsity**: a few mentions in average per articles (5,000-10,000 words)
- **Document-level**: more relevant parts of the document, multiple mentions of same product
- Mentions are heterogeneous and mostly **informal**

- Datasets are mostly unnamed, e.g.:

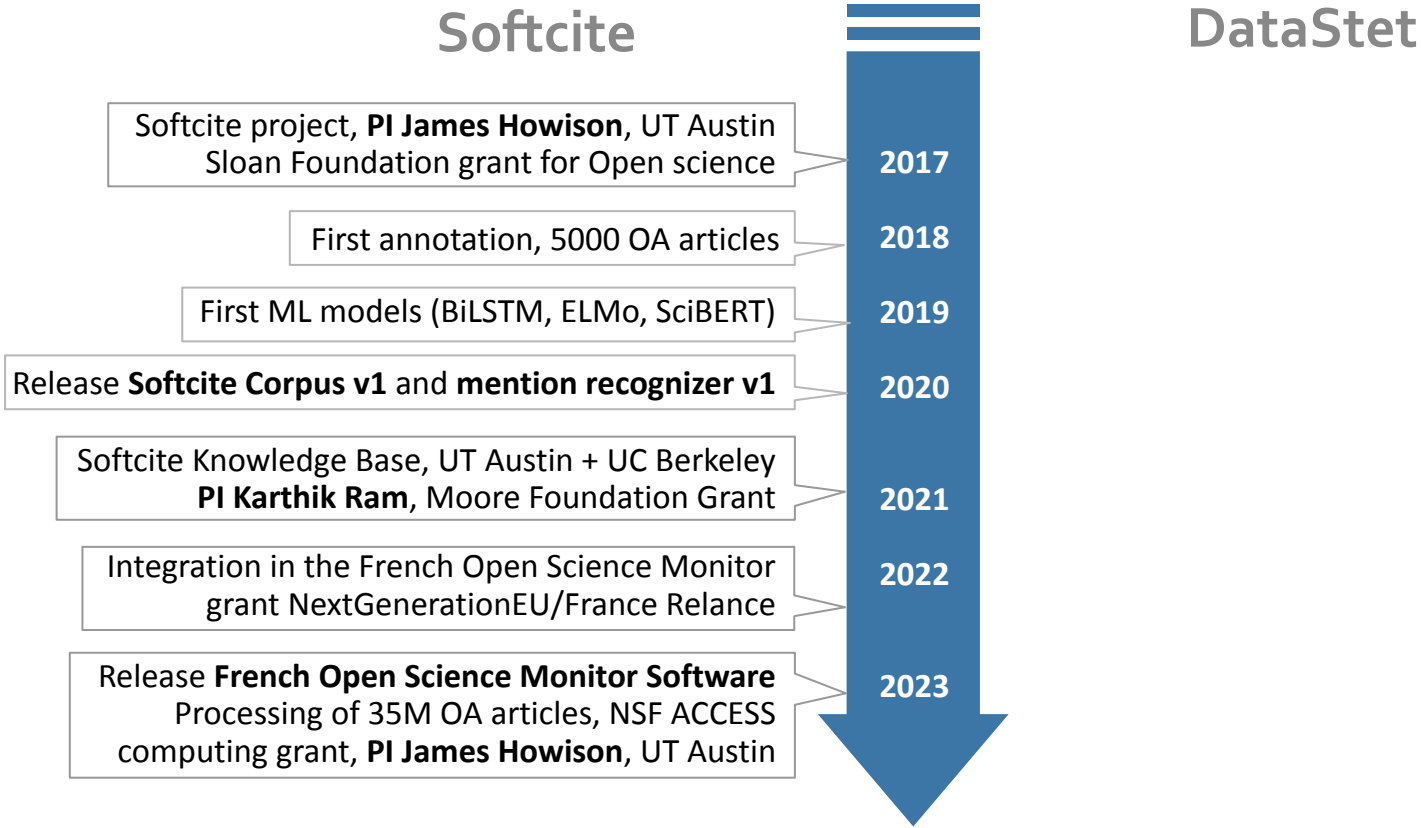
  *"The **data** has been collected by the UN Comtrade organization, and cleaned by CEPII."*

- Software relations can be complex:

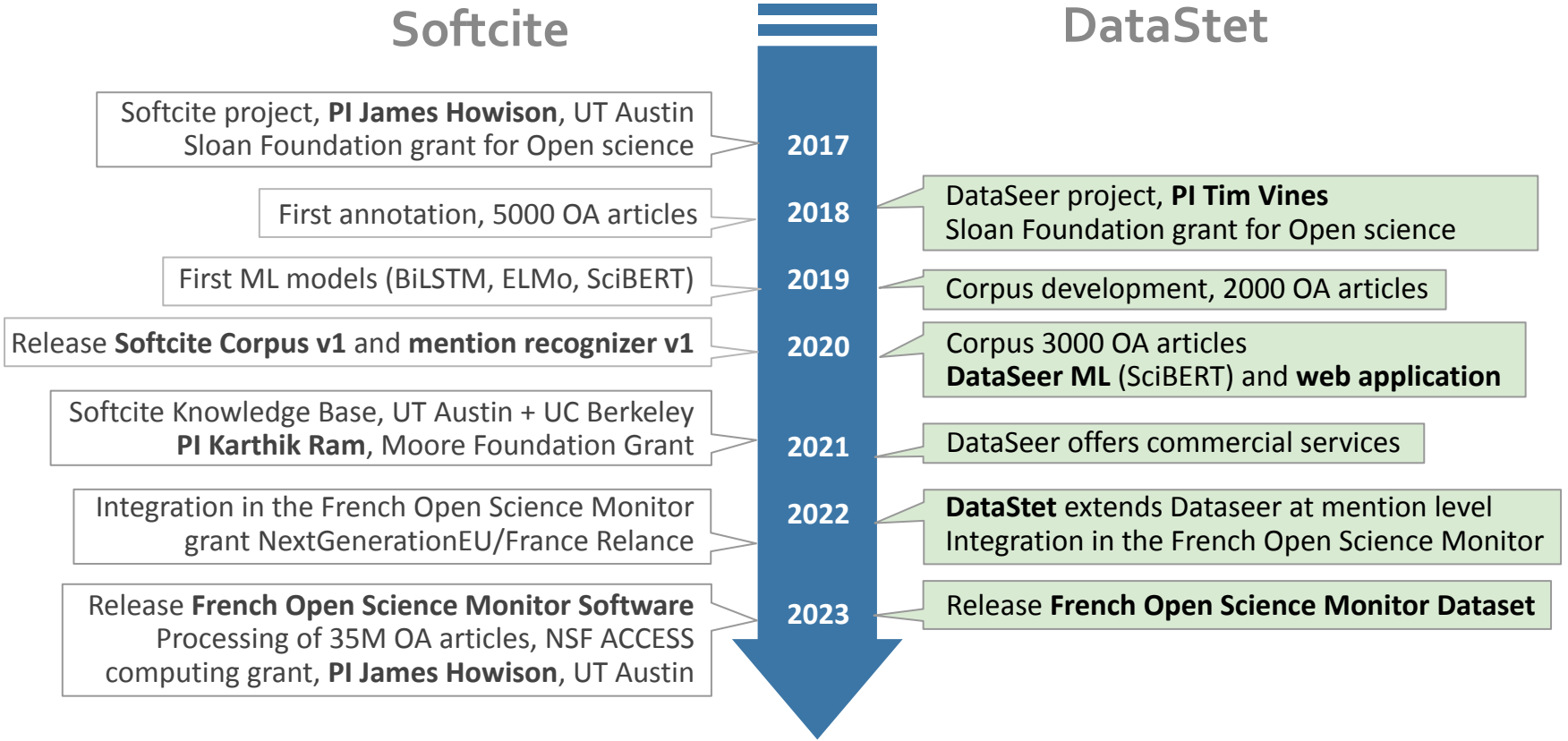
  *"All the **methods** were implemented in the **Scikit Learn** package of **Python 3.1**."*

created used used

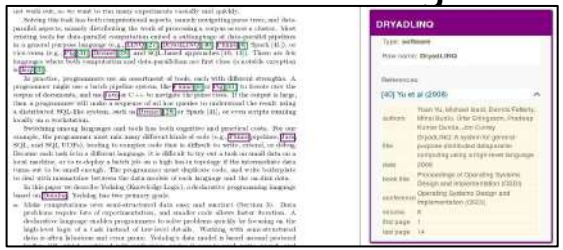
Two 5-years R&D effort for reliable mention extractions with Deep Learning techniques



Two 5-years R&D effort for reliable mention extractions with Deep Learning techniques

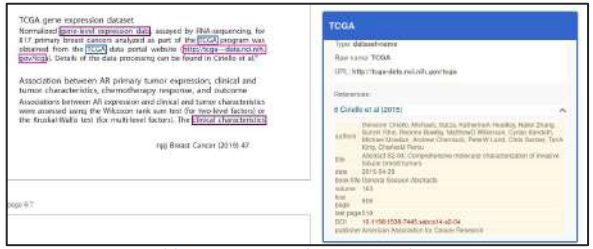


Softcite Mention recognizer



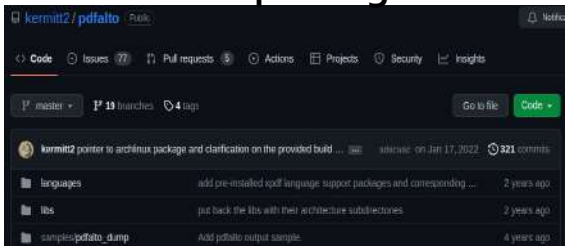
https://github.com/softcite/software-mentions
https://github.com/softcite/software_mentions_client

DataStet



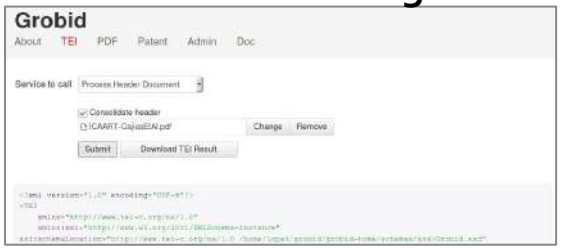
https://github.com/kermitt2/datastet

PDF parsing



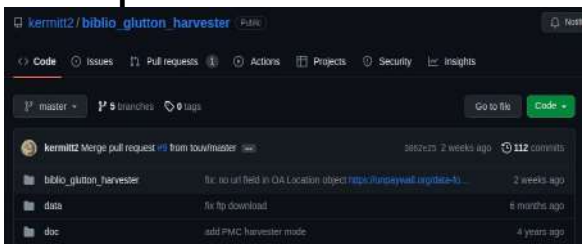
https://github.com/kermitt2/pdfalto

PDF structuring



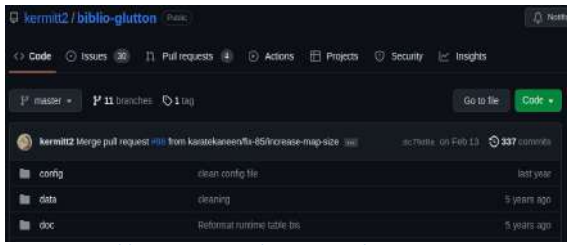
https://github.com/kermitt2/grobid

Open Access harvester



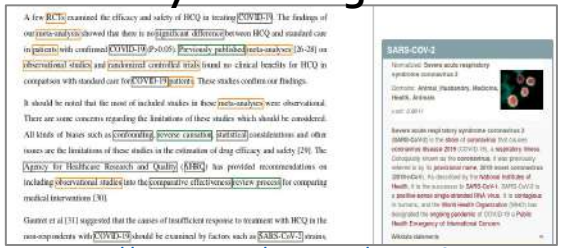
https://github.com/kermitt2/biblio-glutton-harvester

Biblio. reference resolution



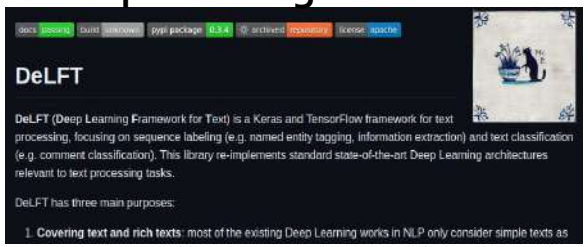
https://github.com/kermitt2/biblio-glutton

Entity disambiguation



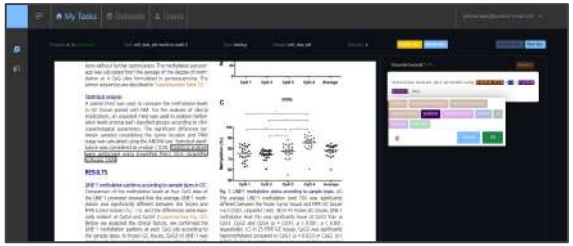
https://github.com/kermitt2/entity-fishing

Deep Learning for rich text



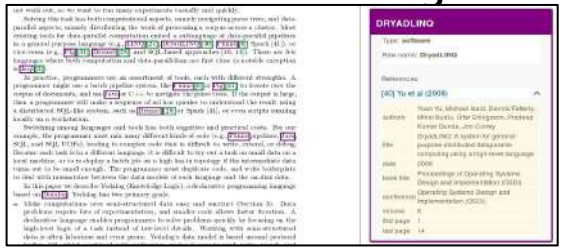
https://github.com/kermitt2/delft

Manual annotations



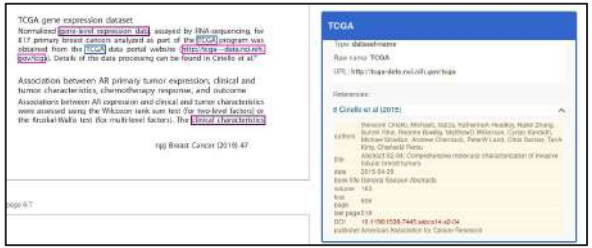
https://github.com/kermitt2/kish

Softcite Mention recognizer



https://github.com/softcite/software-mentions
https://github.com/softcite/software_mentions_client

DataStet



https://github.com/kermitt2/datastet

Deep Learning techniques are reliable for mention detections

- First step, **sentence classification** in relevant document structures, trained on 22,000 sent.

SciBERT	precision	recall	F1-score	support (10%)
data sentence	93.70	96.21	**94.94**	200
not data sentence	97.56	95.92	**96.73**	2000

- Second step, **entity recognition at sentence level,** trained on 6,000 annotated sentences

LinkBERT+CRF	precision	recall	F1-score	support (10%)
named dataset	89.04	89.46	**89.24**	466
unnamed mentioned dataset	71.85	67.15	**69.38**	927
data device	51.91	37.94	42.61	97

For more information and evaluations, see our preprint https://hal.science/hal-04121339 **[1]**

Deep Learning techniques are reliable for mention detections

- Recognition in **one pass**, in relevant document structures, trained on Softcite corpus 4,971 manually annotated documents

SciBERT+CRF	precision	recall	F1-score	support (10%)
software name	74.01	88.98	**80.81**	989
version	83.99	90.81	**87.27**	283
publisher	75.51	88.80	**81.62**	250
url	53.97	82.93	**65.38**	41
all (micro avg.)	75.22	89.12	**81.58**	1563

For more information and evaluations, see our preprint https://hal.science/hal-04121339 **[1]**

Mention context characterization is also reliable

- 3 binary classification, trained on 3,643 manually annotated sentences, from **Softcite** corpus (4971 articles) and **SoMeSci** corpus (GESIS Cologne/Uni Rostock, 1367 mostly partial articles)

	LinkBERT	precision	recall	F1-score	support (10%)
mentioned dataset/software used in the described research work ?	**used**	96.83	94.18	**95.49**	292
	not used	84.40	91.09	**87.62**	101
mentioned dataset/software created/extended ?	**created**	81.08	83.33	**82.19**	31
	not created	98.31	98.04	**98.18**	362
mentioned created dataset/software shared ?	**shared**	81.82	90.00	**85.71**	26
	not shared	99.35	98.71	**99.03**	385

For more information and evaluations, see our preprint https://hal.science/hal-04121339 **[1]**

French Open Science Monitor:
Mentions to datasets and software: 2013-2021

	# documents	share	successful download rate
Full corpus (2013-2021)	1,426,140	100 %	
Full text downloaded	908,567	**63.7 %**	63.7 %
→ open access	→ 660,501	46.3%	85.4%
→ closed access	→ 248,066	17.4%	38.0%

	# full text documents	# mentions	Runtime 2023 (1 instance with GPU)
processed with Softcite	742,289	3,567,547	1.34 PDF/s
processed with DataStet	621,306	5,607,080	0.65 PDF/s

For more information and evaluations, see our preprint https://hal.science/hal-04121339 **[1]**

Monitoring dataset and software production

For **research datasets** extracted with **DataStet**

among all **processed publications**,

share of publications mentioning the use of data

among **those mentioning the use of data**,

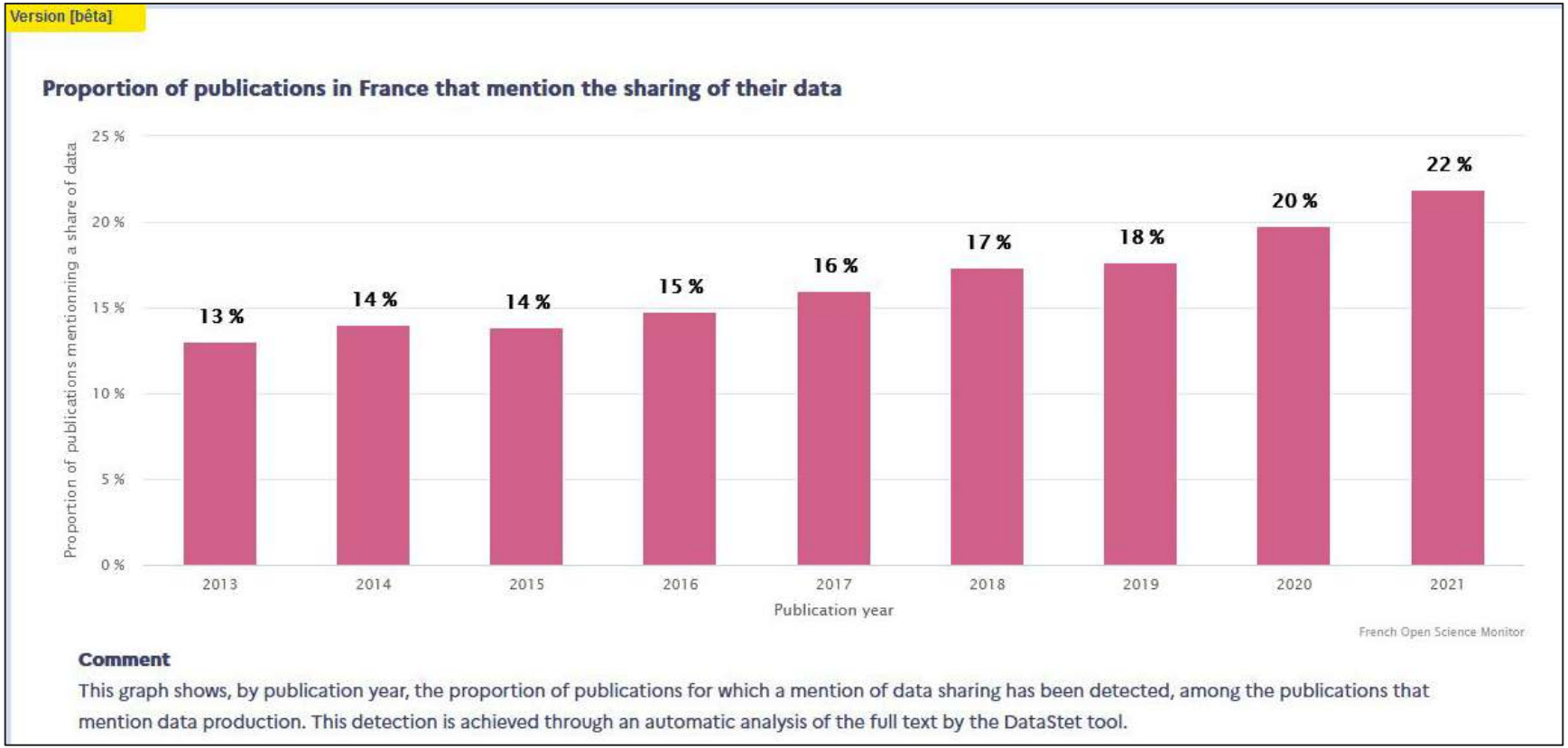
share of publications mentioning the production of data

among **those mentioning the production of data**,

share of publications mentioning the sharing of data

Publications mentioning sharing their produced data

Version [bêta]

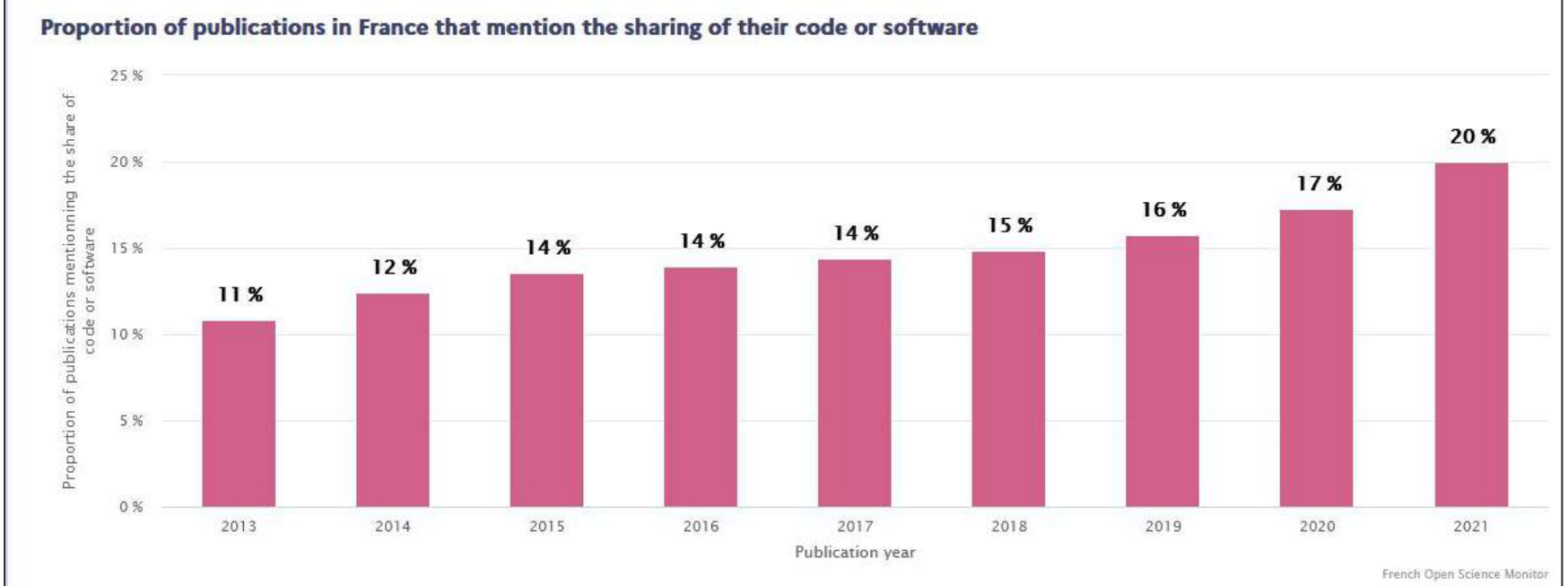


Comment

This graph shows, by publication year, the proportion of publications for which a mention of data sharing has been detected, among the publications that mention data production. This detection is achieved through an automatic analysis of the full text by the DataStet tool.

Publications mentioning sharing of their created software

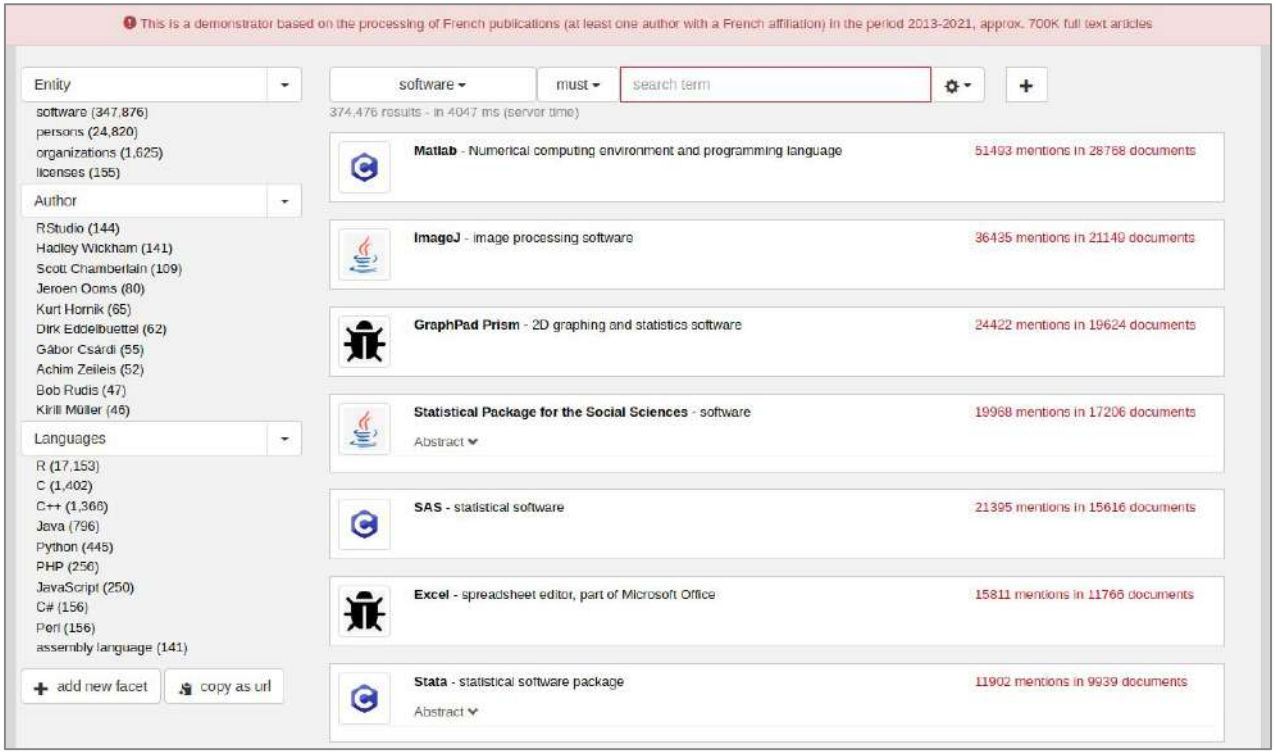
Version [bêta]



Comment

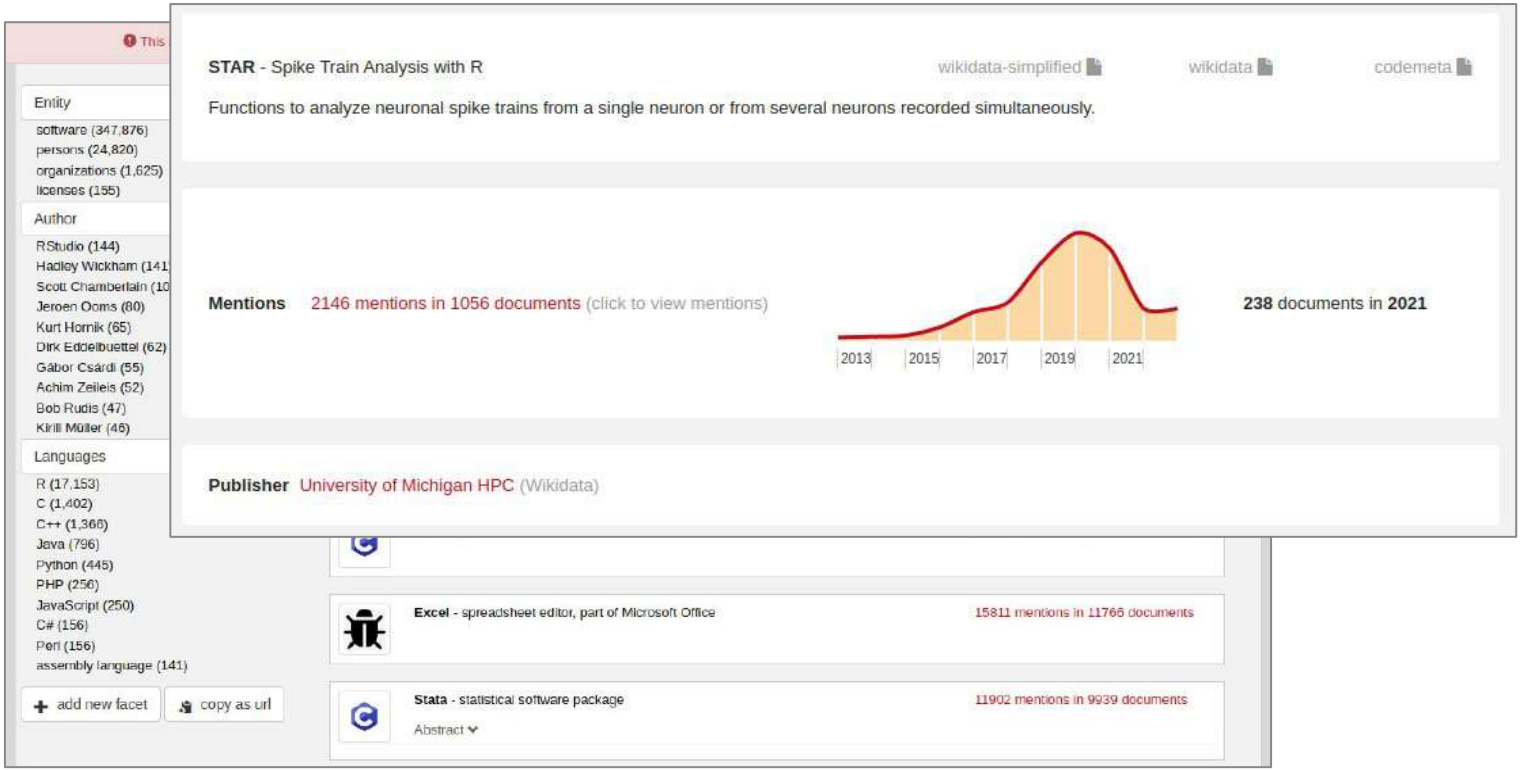
This graph shows, by publication year, the proportion of publications for which a mention of code or software sharing has been detected, among the publications that create code or software. This detection is achieved through an automatic analysis of the full text by the Softcite tool.

From publication-level indicators to dataset and software Knowledge Base



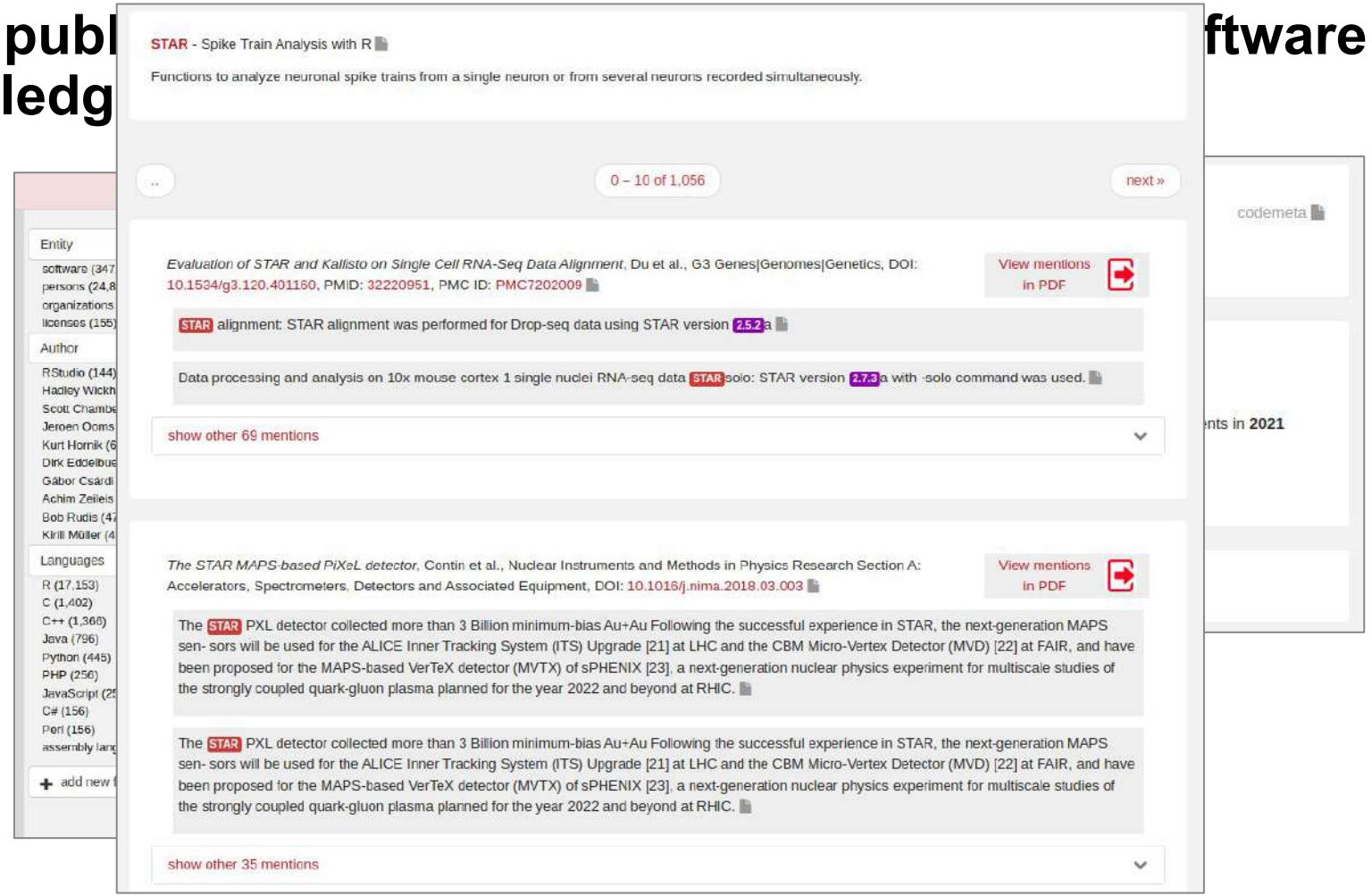
https://cloud.science-miner.com/software_kb_bso/frontend/index.html

From publication-level indicators to dataset and software Knowledge Base



https://cloud.science-miner.com/software_kb_bso/frontend/index.html

From publ ftware Knowledg



From publ ftware Knowledg

STAR - Spike Train Analysis with R

Functions to analyze neuronal spike trains from a single neuron or from several neurons recorded simultaneously.

downloaded from https://github.com/BUStools/bustools. The matrix was generated following the python code available at URL: https://github.com/BUStools/BUS_notebooks_python/blob/master/dataset-notebooks/10x_hgmm_100_python/10x_hgmm_100.ipynb.

Kallisto With Human cDNA and intron index: Kallisto version 0.46.1 and bustools version 0.39.3 were used. The cDNA+intron index and relevant files were downloaded from the github page: https://www.kallistobus.tools/velocity_tutorial.html. The pseudoalignment and sequential correction and counting processes were done following the instruction from https://www.kallistobus.tools/velocity_tutorial.html. The spliced and unspliced matrices were processed following the instruction from https://github.com/BUStools/getting_started/blob/master/velocity_tutorial.ipynb. The cells with less than 3 expressed genes, and genes expressed in less than 200 cells were removed.

Downstream clustering analysis: Seurat version 3.2.0 was used for downstream analysis. The filtration criteria include min.cells = 3, min features = 200. Data were then log normalized with a scale factor of 10000 in Seurat. The cell types were annotated manually based on the FeaturePlot of each marker gene.

Data processing and analysis on 10x mouse cortex 1 single nuclei RNA-seq data

STARsolo: STAR version 2.7.3a with *–solo* command was used. For single nuclei RNA-seq data, command "–soloFeatures Gene SJ GeneFull" was used for generating counts for both exonic RNA and pre-mRNA. The STAR index was built with a read length of 50. STARsolo was configured for 16bp GemCode barcode, 10bp UMI, and 50bp transcript.

Kallisto with mouse cDNA and intron index: Kallisto version 0.46.1 and bustools version 0.39.3 was used. The mouse ensembl 86 cDNA+ intron index and relevant files were downloaded from the github page: https://github.com/pachterlab/MBGBLHGP_2019/releases. The pseu-

show other 35 mentions

Data availability

The authors state that all data necessary for confirming the conclusions presented in the article are represented fully within the article, figures and tables. Supplemental material available at figshare: https://doi.org/10.25387/g3.11866281.

RESULTS AND DISCUSSION

Comparisons of STAR vs. Kallisto alignment results on Drop-Seq and Fluidigm data

STAR and Kallisto are based on different concepts. STAR is a conventional aligner that aligns to the reference genome, whereas Kallisto uses transcriptome quantification for pseudoalignment. To compare these two methods, we downloaded the raw sequencing reads from a previously published GEO data set (GSE99330) (Torre *et al.* 2018). Briefly, this dataset is composed of 8640 single cells generated by Drop-seq platform and 800 single cells generated from Fluidigm (C1 mRNA Seq HT IFC) platforms, using WM989-A6-G3 cell line as the biological material. The RNA-FISH validation data on 26 genes serve as the standard that could help validate the expression level as the result of different alignment methods. We used GRCh38 as the reference genome for STAR and GRCh38 as the reference transcriptome for Kallisto, per recommendation of the authors.

For the scRNA-seq reads from Drop-seq platform, STAR has 62.40% alignment rate, compared to 35.11% pseudoalignment rate from Kallisto; for the reads from Fluidigm platform, STAR has 66.57% alignment rate, compared to 34.03% from Kallisto (Table S1). To generate the count matrix, we used STAR and Kallisto genombam command (Yi *et al.* 2018 *preprint*) followed by featureCount. Kallisto genombam command projects the pseudoalignments to genomic space using a model of transcriptome consisting of genes, transcripts and exon coordinates, which allows the interchange between pseudoalignment and genome alignment possible. We then evaluated the aligners on the count matrix output (Figure 1).

Specifically, we first checked the overall correlation of alignments from STAR and Kallisto workflows. We added a pseudo-count of

STAR

Type: software

Raw name: STAR

Version: 2.5.2

Publisher: University of Michigan HPC

References:

(Dobin et al. 2013) Dobin et al (2013)

authors	Alexander Dobin, CarrieA Davis, Felix Schlesinger, Jorg Drenkow, Chris Zaleski, Sonali Jha, Philippe Batut, Mark Chaisson, ThomasR Gingeras
title	STAR: ultrafast universal RNA-seq aligner
date	2013
journal	Bioinformatics
volume	29
issue	1
first page	15
last page	21
ISSN	1367-4803
e ISSN	1460-2059
DOI	10.1093/bioinformatics/bts635
publisher	Oxford University Press (OUP)
url	https://doi.org/10.1093/bioinformatics/bts635

Challenges in monitoring dataset and software production

- Publication corpus **completeness**
  - Access to full-texts often difficult
  - Limited coverage of documents without DOI
  - Current dataset & software extraction supports only English

- Budget and time **cost**
  - Modern NLP techniques are computing-intensive
  - New requirements like GPU and cloud-based solution for scaling

- Performance across **domains**
  - Example of Softcite: Currently good coverage/accuracy in Life Sciences and Economics…
  - … but estimate of 15 points F1-score loss on an entirely new scientific domain

- Software is more than what is visible from publications: **library/package dependencies**

References

[1] Aricia Bassinet, Laetitia Bracco, Anne L'Hôte, Eric Jeangirard, Patrice Lopez, et Laurent Romary. 2023. Large-scale Machine-Learning analysis of scientific PDF for monitoring the production and the openness of research data and software in France. 2023. https://hal.science/hal-04121339

[2] Du, C., Cohoon, J., Lopez, P., & Howison, J. 2022. Understanding progress in software citation: A study of software citation in the CORD-19 corpus. PeerJ Computer Science, 8, e1022. https://doi.org/10.7717/peerj-cs.1022

[3] David Schindler, Tazin Hossain, Sascha Spors, Frank Krüger. 2023. A multi-level analysis of data quality for formal software citation. arXiv:2306.17535v1, https://arxiv.org/abs/2306.17535

[4] He, L., & Han, Z. 2017. Do usage counts of scientific data make sense? An investigation of the dryad repository. Library Hi Tech, 35(2), 332–342. https://doi.org/10.1108/LHT-12-2016-0158

Monitoring the opening of protocols and clinical study reports

Inge Stegeman (UMC Utrecht)

siris

Funded by
the European Union

UMC Utrec

Summary

- Expand protocol requirement to all types of biomedical studies
- Monitor what is done instead of what is written that ‘will be done’
- Improve the writing of papers in order for automated monitoring
- Behavioural change

AND

- Evidence Based! Do the research! Test test test!

OSIRIS

Open Science to Improve the Reproducibility of Science

Evidence Based Reproducibility

UMC Utrecht

KU LEUVEN

Amsterdam UMC
University Medical Centers

UNIVERSITY OF OXFORD

Czech University of Life Sciences Prague

ISTITUTO DI RICERCHE FARMACOLOGICHE MARIO NEGRI · IRCCS

SENSE about SCIENCE

lorier
L'organisation pour une recherche Inserm éthique et responsable

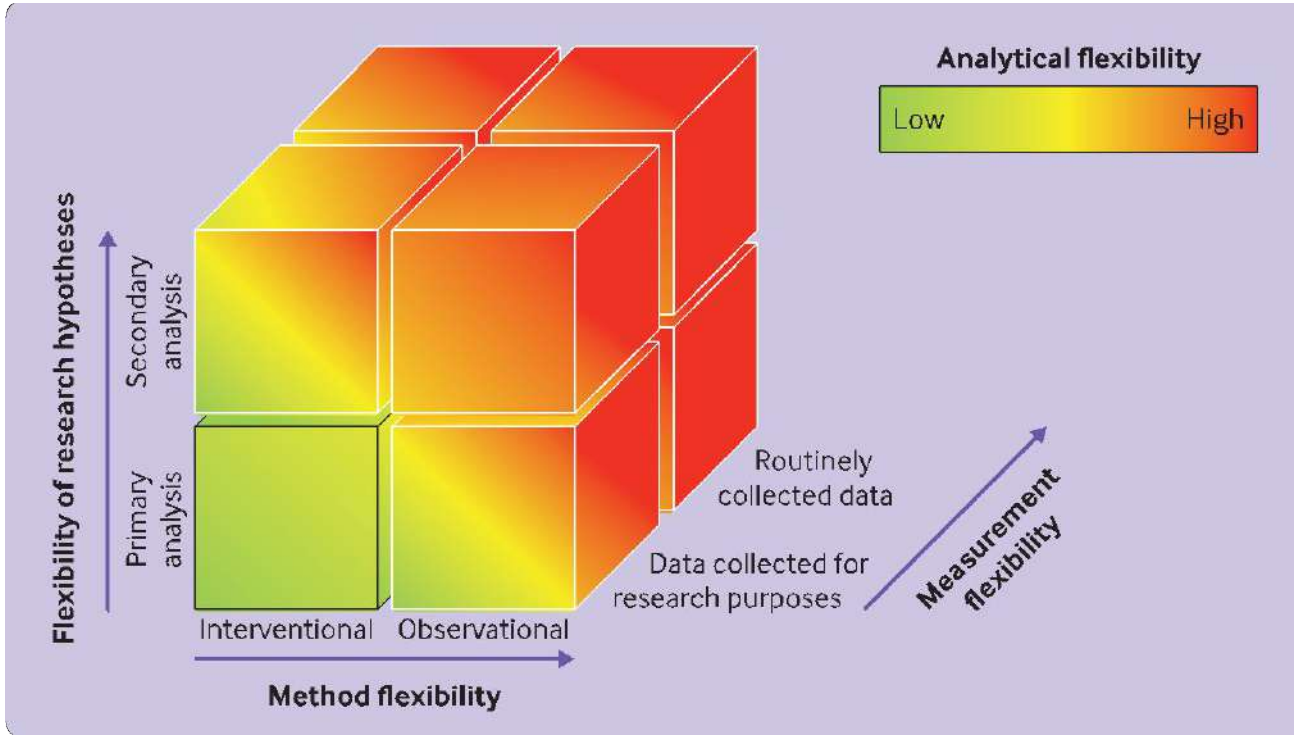
MATE
HUNGARIAN UNIVERSITY OF AGRICULTURE AND LIFE SCIENCES

ypard

- Protocols?
- What to monitor
- How to monitior

- Behaviour

Protocols?



Naudet et al. BMJ. Improving the transparency and reliability of observational studies through registration

Protocols?

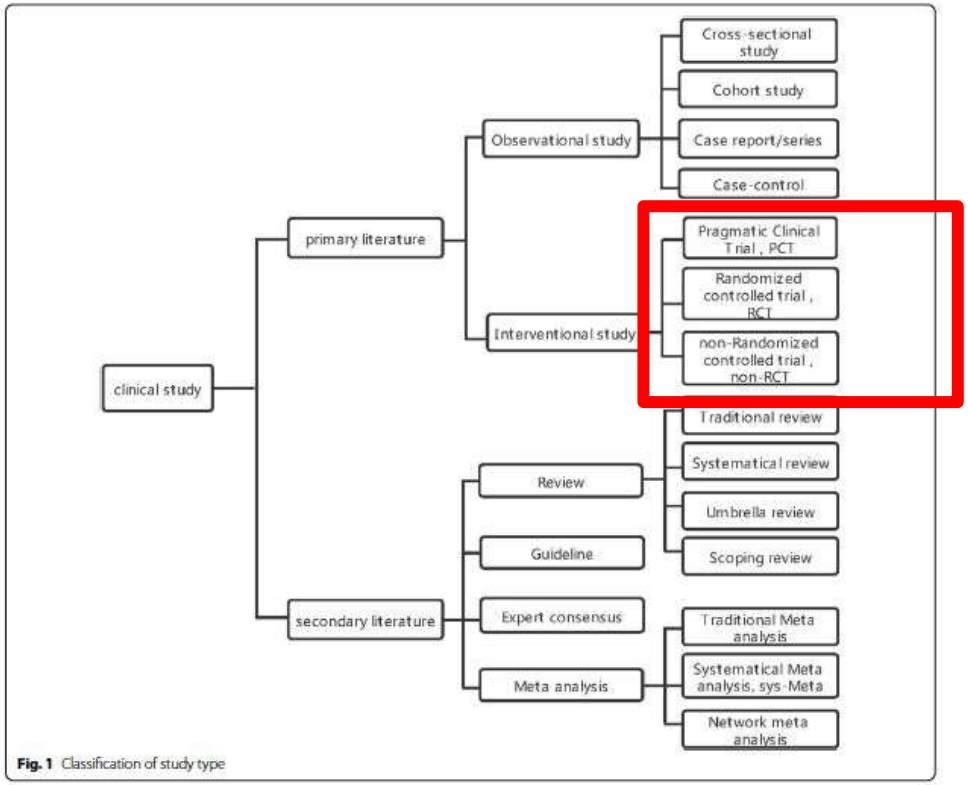
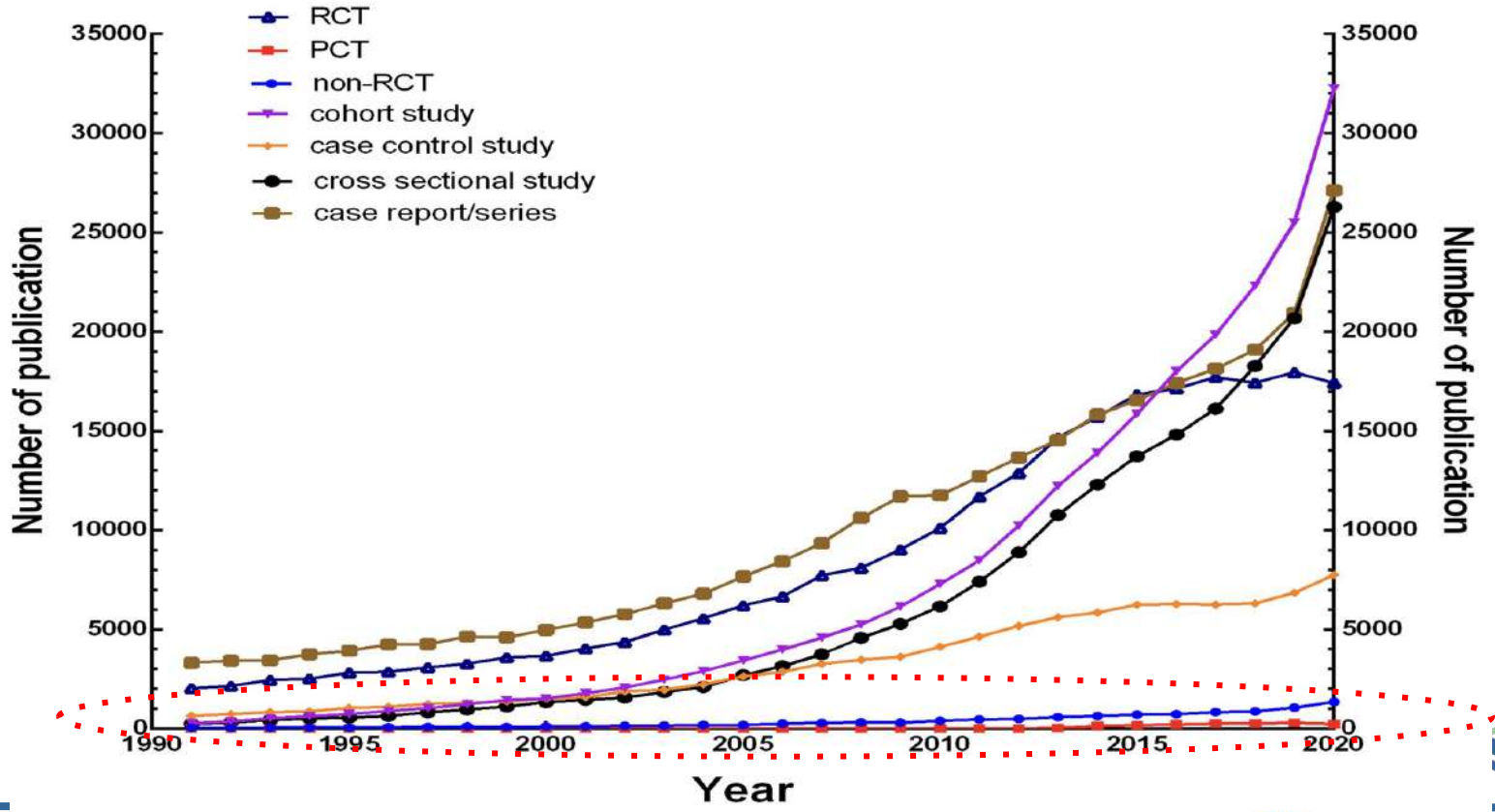


Fig. 1 Classification of study type

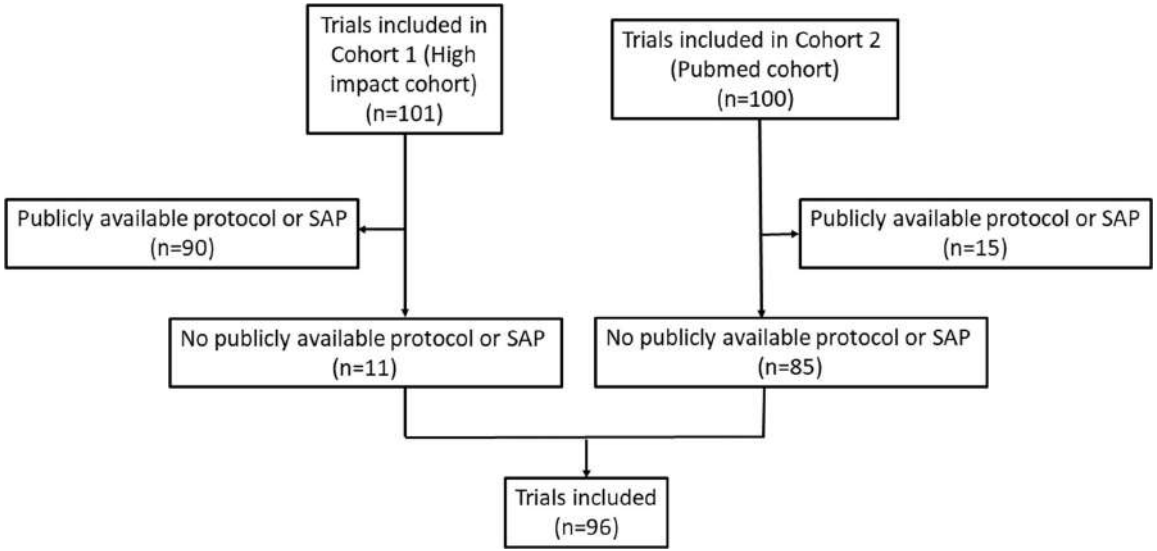
Zhao et al.
European Journal of Medical Research (2022) 27:95

Protocols?



Zhao et al. European Journal of Medical Research (2022) 27:95

Protocols



Campbell et al. Trials (2022) 23:674

Protocols

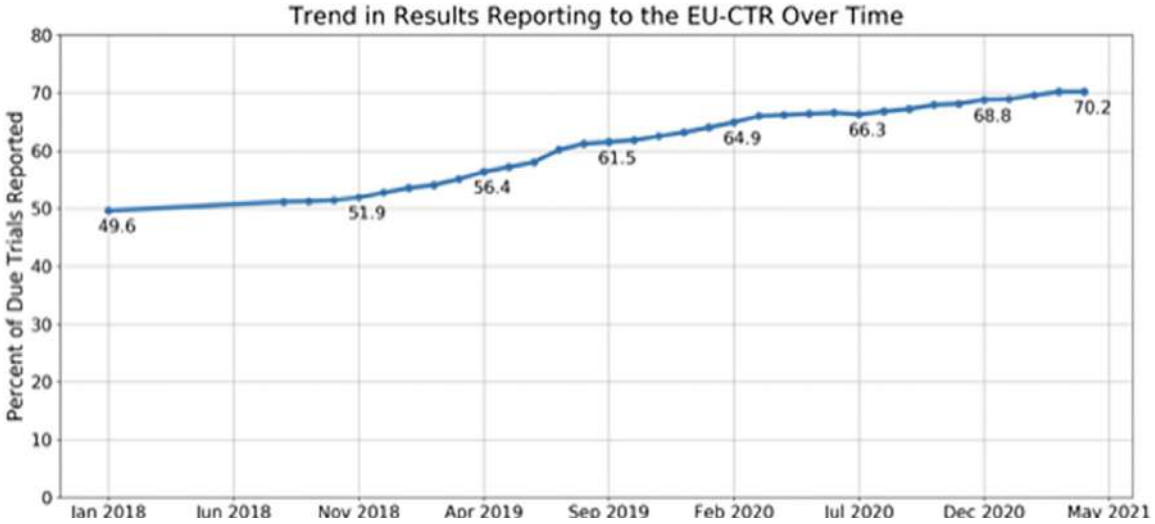
Table 2 Study results

	High impact cohort	PubMed cohort	Total
Publicly available protocol in stage 1			
Yes (excluded from stage 2)	90/101 (89)	15/100 (15)	105/201 (52)
No (included in stage 2)	11/101 (11)	85/100 (85)	96/201 (48)
Replied to email requesting protocol/SAP			
Yes	6/11 (55)	9/85 (11)	15/96 (16)
No	5/11 (45)	76/85 (89)	81/96 (84)
Days until first response (among those responding)			
Median (IQR)	23 (4, 35)	3 (2, 21)	10 (2, 35)
Shared some study documents			
Yes	4/11 (36)	4/85 (5)	8/96 (8)
No	7/11 (64)	81/85 (95)	88/96 (92)
Days until documents shared (among those sharing)			
Median (IQR)	31 (14, 42)	21 (11, 22)	22 (11, 31)
Shared protocol			
Yes	4/11 (36)	2/85 (2)	6/96 (6)
No	7/11 (64)	83/85 (98)	90/96 (94)
Shared statistical analysis plan			
Yes	1/11 (9)	0/85 (0)	1/96 (1)
No	10/11 (91)	85/85 (0)	95/96 (99)
Shared other study documents[a]			
Yes	0/11 (0)	2/85 (2)	2/96 (2)
No	11/11 (100)	83/85 (98)	94/96 (98)
Study protocol available at end of stage 2 (either available publicly in stage 1 or shared upon request in stage 2)			
Yes	94/101 (93)	17/100 (17)	111/201 (55)
No	7/101 (7)	83/100 (83)	90/201 (45)

[a] Partial excerpt from a protocol ($n = 1$), research ethics application form ($n = 1$)

Campbell et al. Trials (2022) 23:674

Protocols



Journal of Clinical Epidemiology 142 (2022) 161–170

Table 2| **Proportion of trials with results posted on ClinicalTrials.gov at three and six months. Values are numbers (percentages) unless stated otherwise**

Analysis	Intervention	Control	Risk difference (95% CI)*	Relative risk (95% CI)	P value
Primary analysis:	n=190	n=189			
3 months	36 (19)	24 (13)	6.2 (−1.1 to 13.6)	1.5 (0.9 to 2.4)	0.096
6 months	46 (24)	27 (14)	9.9 (2.1 to 17.8)	1.7 (1.1 to 2.6)	0.014
Sensitivity analysis†:	n=164	n=167			
3 months	10 (6)	2 (1)	4.9 (0.9 to 8.9)	5.1 (1.1 to 22.9)	0.02
6 months	20 (12)	5 (3)	9.2 (3.6 to 14.8)	4.1 (1.6 to 10.6)	0.001

Three month assessment corresponds to posting results on 1 December 2012.

Six month assessment corresponds to posting results on 1 March 2013.

*Asymptotic 95% confidence interval.

†Excluding 48 trials not meeting inclusion criteria because "results first received date" were before randomization.

Ravoud et al. BMJ 2014;349:g5579 doi: 10.1136/bmj.g5579

Expand protocol requirement to all types of biomedical studies

What and how to monitor

What to monitor

PLOS BIOLOGY

CONSENSUS VIEW

Community consensus on core open science practices to monitor in biomedicine

Kelly D. Cobey[1,2*], Stefanie Haustein[3,4], Jamie Brehaut[2,5], Ulrich Dirnagl[6,7], Delwen L. Franzen[7], Lars G. Hemkens[7,8,9], Justin Presseau[2,5,10], Nico Riedel[6], Daniel Strech[6,11], Juan Pablo Alperin[4,12], Rodrigo Costas[13], Emily S. Sena[14], Thed van Leeuwen[13], Clare L. Ardern[15,16], Isabel O. L. Bacellar[17], Nancy Camack[5], Marcos Britto Correa[18], Roberto Buccione[19], Maximiliano Sergio Cenci[18], Dean A. Fergusson[2,5], Cassandra Gould van Praag[20], Michael M. Hoffman[21,22,23,24], Renata Moraes Bielemann[25], Ugo Moschini[26], Mauro Paschetta[27], Valentina Pasquale[26], Valeria E. Rac[28,29,30], Dylan Roskams-Edris[31,32], Hermann M. Schatzl[33], Jo Anne Stratton[31], David Moher[2,5]

1 University of Ottawa Heart Institute, Ottawa, Ontario, Canada, **2** School of Epidemiology and Public Health, Faculty of Medicine, University of Ottawa, Ottawa, Ontario, Canada, **3** School of Information Studies, Faculty of Arts, University of Ottawa, Ottawa, Ontario, Canada, **4** Scholarly Communications Lab, Ottawa and Vancouver, Canada, **5** Clinical Epidemiology Program, Ottawa Hospital Research Institute, Ottawa, Ontario, Canada, **6** Department of Experimental Neurology, Charité-Universitätsmedizin Berlin, Berlin, Germany, **7** QUEST Center for Responsible Research, Berlin Institute of Health at Charité-Universitätsmedizin Berlin, Berlin, Germany, **8** Department of Clinical Research, University of Basel and University Hospital Basel,

Check for updates

To reach consensus on what open science practices to monitor at biomedical research institutions, we conducted a modified 3-round Delphi study. Participants were research administrators, researchers,

Table 3. Prioritization of traditional open science practices and broader transparency practices.

No.	Practice	Score
	Traditional open science practices	
1	Reporting whether clinical trials were registered before they started recruitment	9.71
2	Reporting whether study data were shared openly at the time of publication (with limited exceptions)	9.18
3	Reporting what proportion of articles are published open access with a breakdown of time delay	8.12
4	Reporting whether study code was shared openly at the time of publication (with limited exceptions)	7.94
5	Reporting whether systematic reviews have been registered before data collection began	6.76
6	Reporting whether clinical trials results appeared in the registry from 1 year after study completion	6.76
7	Reporting whether there was a statement about study materials sharing with publications	6
8	Reporting whether a reporting guideline checklist was used	5.88
9	Reporting citations to data	5.53
10	Reporting trial results in a manuscript-style publication (peer reviewed or preprint)	4.82
11	Reporting the number of preprints	4.35
12	Reporting systematic review results in a manuscript-style publication (peer reviewed or preprint)	2.94
	Broader transparency practices	
1	Reporting whether author contributions were described	5.12
2	Reporting whether author conflicts of interest were described	4.71
3	Reporting the use of persistent identifiers when sharing data/code/materials	4.65
4	Reporting whether ORCID identifiers were used	4.47
5	Reporting whether data/code/materials are shared with a clear license	3.47
6	Reporting whether research articles include funding statements	3
7	Reporting whether the data/code/materials license is open or not	2.59

Cobey KD, et al. (2023) Community consensus on core open science practices to monitor in biomedicine. PLoS Biol 21(1): e3001949.

Consensus Core Open Science characteristics to monitor

19 outcomes amongst:

1. **Reporting whether clinical trials were registered before they started recruitment.** *This practice is required by several organizations and funders internationally. Despite clear mandates for registration, we know this practice is not optimal. Standardized reporting of trial registration will allow for linkage of trial outputs to the registry and help contribute to the reduction of selective outcome reporting and non-reporting.*

Cobey KD, et al. (2023) Community consensus on core open science practices to monitor in biomedicine. PLoS Biol 21(1): e3001949.

Monitor what is done instead of what is written

Summary points

- Efficient sharing and reuse of data from clinical trials are critical in advancing medical knowledge and developing improved treatments.
- We believe that the International Committee of Medical Journal Editors (ICMJE) clinical trial data sharing policy is currently inadequate.
- Although data sharing plans help increase transparency, they do not ensure that data are shared, and they are often inadequately implemented.
- We believe that the ICMJE should adapt a stronger policy on data sharing that is enforced rigorously in all ICMJE members and affiliated journals.
- The policy should include a strong evaluation component to ensure that all clinical trial data are shared, their value maximized, and data producers incentivized.

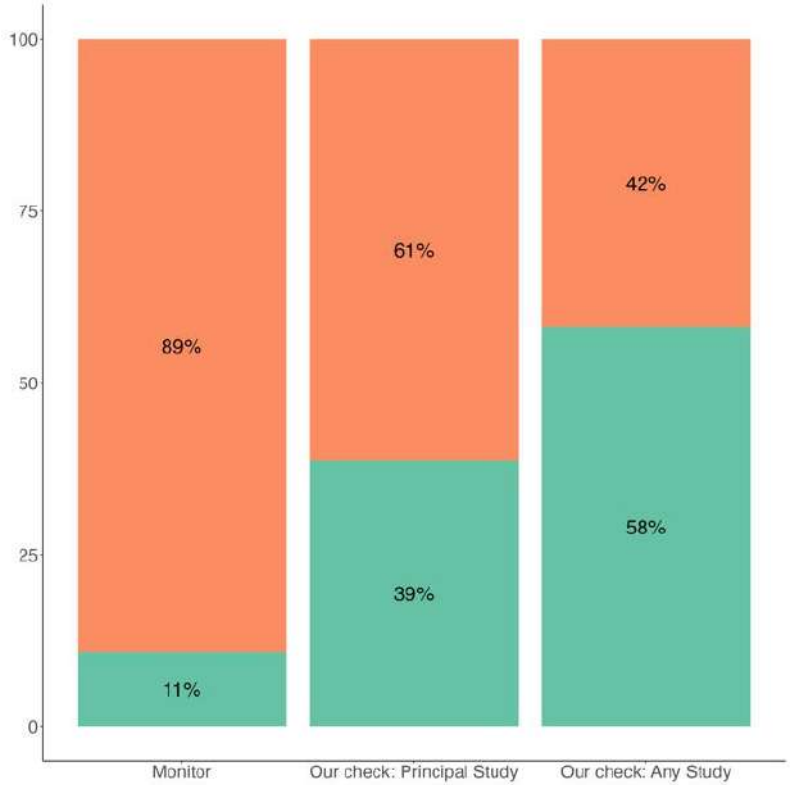
Naudet F, Siebert M, Pellen C, Gaba J, Axfors C, Cristea I, et al. (2021) Medical journal requirements for clinical trial data sharing: Ripe for improvement. PLoS Med 18(10): e1003844.
https://doi.org/10.1371/journal.pmed.1003844

Plos tools to monitor Open Science

- Sharing of research data, in particular data shared in data repositories
- Sharing of code
- Posting of preprints

How to monitor

- Automated
- Non-automated



Alix-Doucet et al. Reporting of interventional clinical trial results in a French academic center: a survey of completed studies to be submitted.

Non automated

Editors and community members can complete a journal evaluation form on the TOP Factor website to accelerate the process

Top Factor

- reports the steps that a journal is taking to implement open science practices, practices that are based on the core principles of the scientific community.

- It is an alternative way to assess journal qualities, and is an improvement over traditional metrics that measure mean citation rates.

- The TOP Factor is transparent (see underlying data and the evaluation rubric) and will be responsive to community feedback.

Rubrik of top factor

- Data citation
- Data transparancy
- Analytical code transparancy
- Materials transparancy
- Design and analysis transparancy
- Study preregistration
- Replication
- Publication bias
- Open science badges

https://osf.io/t2yu5

Other examples

- Charité Dashboard on Responsible Research
- https://eu.trialstracker.net

Conclusion

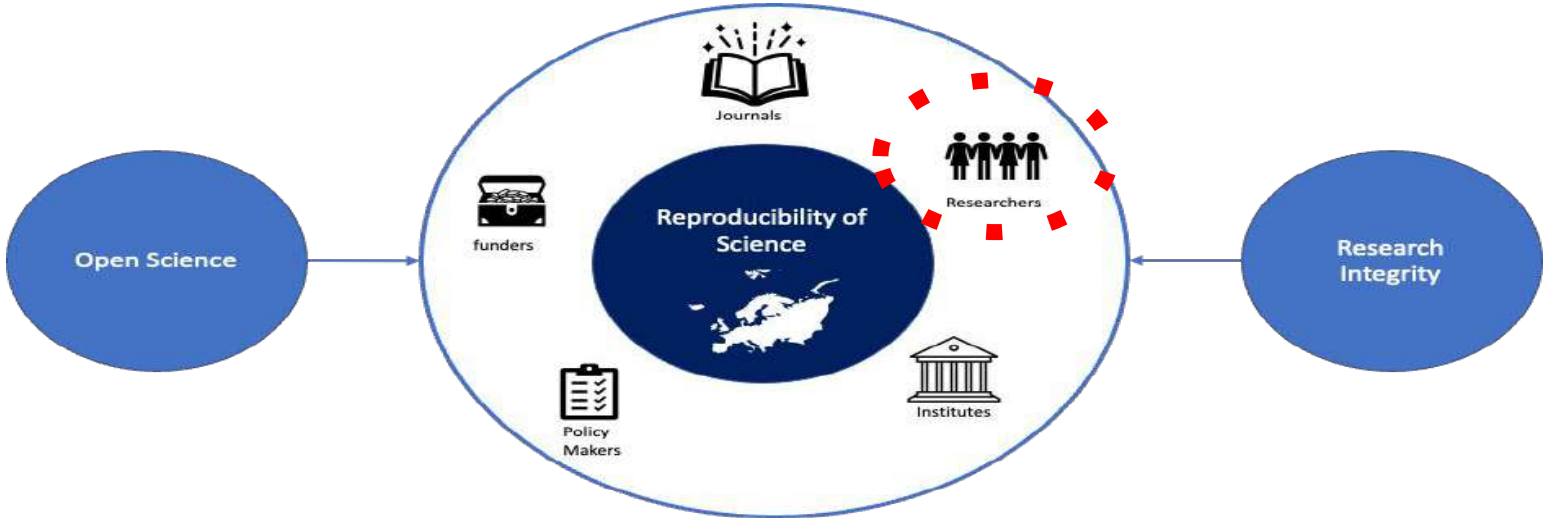
Automated monitoring would be ideal, but currently reporting is the challenge

Most imporant!!

Most important

- Monitoring is important, but behavioural change is the key
  - Recognition and reward system

Behavioural change

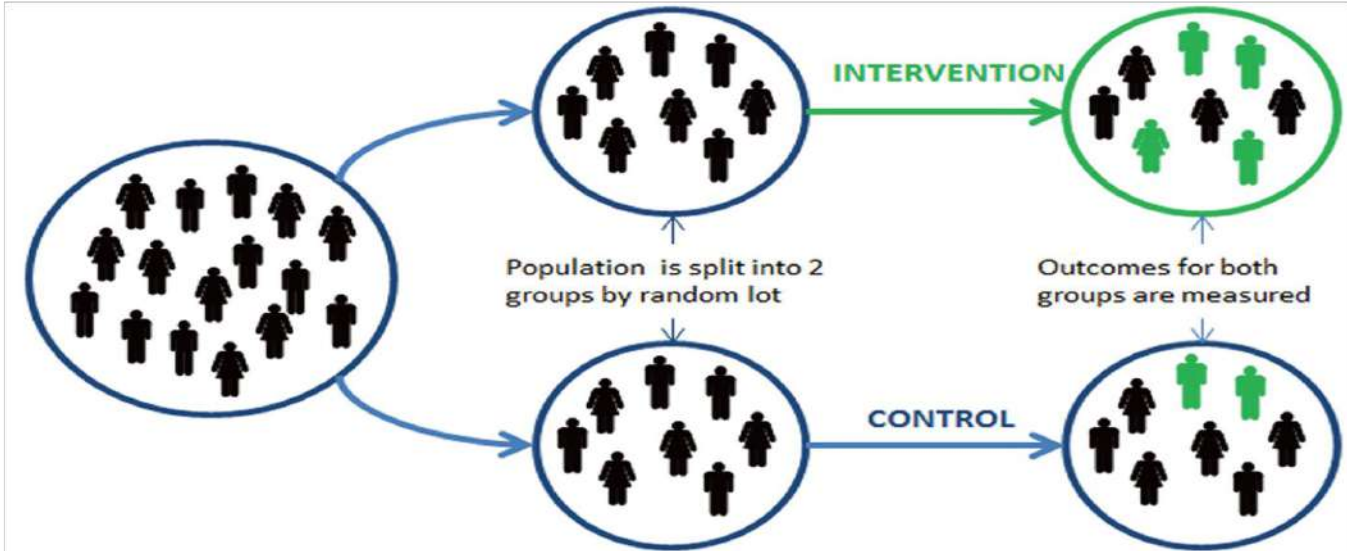


Evidence Based

Monitoring tools have the potential to improve impact of protocols and Open Science practices…….

But

First assess the effectiveness, harms and benefits of the measure before monitoring.


INTERVENTION
Population is split into 2 groups by random lot
Outcomes for both groups are measured
CONTROL

OSIRIS

INVITATION TO COLLABORATE

GOALS

- Identify effective (open science) interventions that increase reproducibility of research
  - Understand drivers & barriers
- Develop and test interventions to improve reproducibility for researchers & institutions
  - Dashboards of indicators
- Develop and evaluate systems for reproducibility compliance for publishers & funders
- Co-create, design and user-test training resources

Summary

- Expand protocol requirement to all types of biomedical studies
- Monitor what is done instead of what is written that ‘will be done’
- Improve the writing of papers in order for automated monitoring
- Behavioural change

AND

- Evidence Based! Do the research! Test test test!

Towards a monitoring framework that reflects values and outcomes

Ismael Rafols, CWTS, Leiden University

UNESCO Chair in Diversity and Inclusion in Global Science

Argument: A framework of OS trajectories

- Open Science is(part of) a **change of the model** of how science works
  - New model of science → **new monitoring framework**
- Open Science has **multiple dimensions to include (not only outputs)**
- Its monitoring is about mapping **directionality related to values under highly uncertain conditions.**

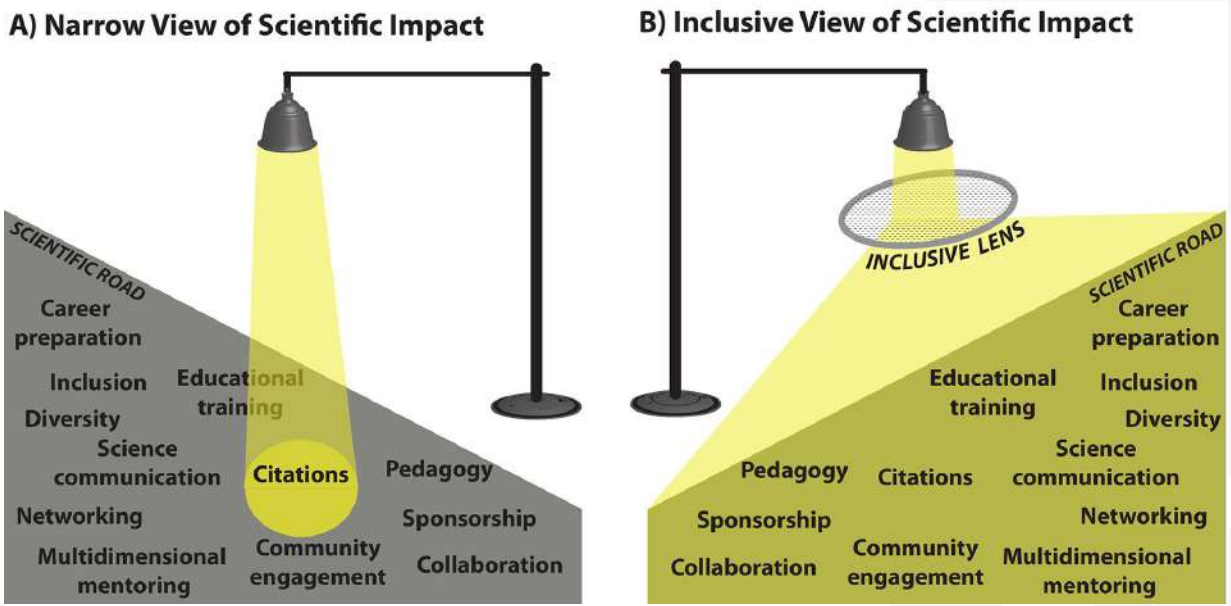
We propose three principles for monitoring OS:

1. Monitoring needs to **include values & normative** commitments (expected impacts)
2. **Formative monitoring & opening up**: monitoring should foster reflection on alternative transformations
3. **Focus on outcomes, not outputs** – given uncertainties and ambiguities of transformation
   - Need to **survey practices of subjects (orgs and people),** not only objects (outputs)

CWTS Meaningful metrics

The streetlight effect of indicators

- Incentives: indicators signal to stakeholders what is important. consequences on research system
  - Goal displacement: instead of mission, follow indicators



EC Expert Group on Indicators for OS (2018) for assessment:

Indicator frameworks (for each context different sets of indicators)

- List of tentative 150 indicators
- No CORE set of indicators

Davies et al. (2023) **Promoting inclusive metrics of success and impact**

The benefits of open science are not inevitable: Problems with epistemic diversity and injustice in current OS

Sabina Leonelli (2023):

"…the interpretation of openness as the sharing of resources, so often encountered in OS initiatives and policies, may have the unwanted effect of **constraining epistemic diversity and worsening epistemic injustice**, resulting in **unreliable and unethical scientific knowledge**. "

"…some OS policies – despite their good intentions and progressive slant – [are] acting as a **reactionary force which reinforces conservatism, discrimination, commodification and inequality in research**, thus ultimately closing down opportunities for inquiry in a disastrous reversal of what they set out to achieve."

CWTS Meaningful metrics

1. Values and normative commitments need to be included in the monitoring

For a given transformative innovation, there are normative commitments associated to values. Monitoring should help in visualising how these commitments fare:

“More is not better” □ we need to discuss the directions, the trajectories

For example: Open Access publications in Gold/Hybrid OA increasibility of pubs…BUT

also creates:

- barriers **to equity and fairness** (as seen in demography)
- problems **of quality and integrity** (lack of rigorous reviewing: MDPI & Frontiers)
- challenge **to collective benefit** (more visibility to topics of the rich countries?)
- lack of **transparency** (‘soft’ peer review given incentives to publish in some journals)

2. Formative Monitoring: a map of trajectories supporting strategic decision-making

- There is not one single transformation to monitor (from A to B: perhaps from A to C or to D)
- A transformation involves multiple aspects, and for each aspects, there are multiple trajectories
- Not about more or less Open Science but **what type of Open Science**

- The main purpose of monitoring is **fostering reflection about the trajectory taken in a transformation** and its implications in relation with the normative aims
- Open Access is increasing… but what type of OA? What are the broader consequences?

- Formative & Learning component in monitoring:
  - Avoiding the street-light effect.
  - Facilitate decision-making and navigation between alternative trajectories
  - Identifying views, interests, choices

CWTS Meaningful metrics

2. Opening up OS: showing multiple trajectories within OS

effect of appraisal 'outputs' on decision-making

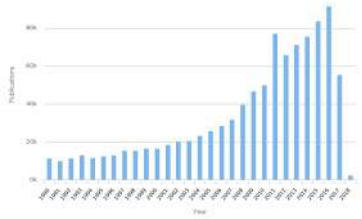
closing-down ⟷ opening-up

narrow

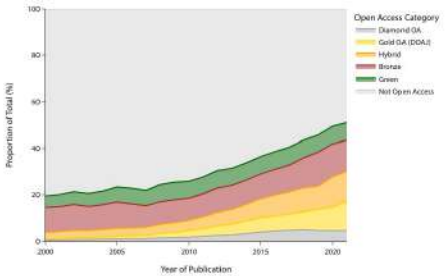
range of appraisals inputs

(issues, perspectives, scenarios, methods)

broad

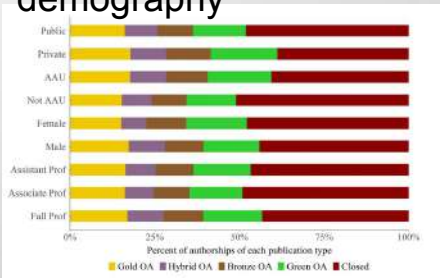


Trends in OA



Plural dimensions
Trends in OA by type

Outcomes:
OA pubs by demography



Opening Up versus Closing Down (Stirling, 2008)

3. From *Outputs to Processes and Outcomes*

In OS, the focus of monitoring is currently on outputs (science supply).

- % of papers in OA, % with OD, # open software scripts, # “citizen science” projects

Little on processes (participation, dialogue) or outcomes (changes in beneficiaries)

But what are the uses? Assumptions driving OS policies might be wrong

- Is OA leading to broader readership outside of academia? Not large, but not negligible
- What is the evidence of re-use of OD sets by other researchers? Perhaps low? Is it worth the effort?
- Evidence of re-use of open software? Yes. perceived as very high and impactful.

Re-using Open Data
EC report (2020)

3. Need of surveys: subject-based monitoring of processes rather than object-based of outputs

To monitor ‘outcomes’ (i.e. effects of policies in practices related to OS)

we cannot rely on counting objects/products (though these may be valuable)

We need to survey:

- organisations: which policies? (e.g.UNESCO survey to govs.)
- researchers: which practices? which engagements? (e.g. Berlin survey, SuperMoRRI)
- non-academics: which engagement? which benefits?

So that they explain their (potential changes) in behaviour.

SuperMoRRI project interviews/surveys to universities and researchers on RRI

Existing surveys to citizens on their relationship with science

Summary

Given that transformative change is **about mapping directionality related to values under highly uncertain conditions…**

…we propose three principles for monitoring:

1. Monitoring need to **include values & normative commitments**
2. **‘Formative’ monitoring**: monitoring should foster learning of alternatives OS transformations (including values and impacts)
3. **Focus on processes and outcomes (changes induced), not only outputs** – given uncertainties and ambiguities of transformation
   - Need to survey practices of subjects (orgs and people), not only objects

See blog at: https://blogs.lse.ac.uk/impactofsocialsciences/2023/08/14/the-benefits-of-open-science-are-not-inevitable-monitoring-its-development-should-be-value-led/

CWTS Meaningful metrics

Panel Discussion III – Towards an Open Science Monitoring Framework

- Monitoring Open Science beyond open access to scientific knowledge: Reflections of the UNESCO Working Group on a Global Open Science Monitoring Framework. Ana Persic, Programme Specialist, UNESCO

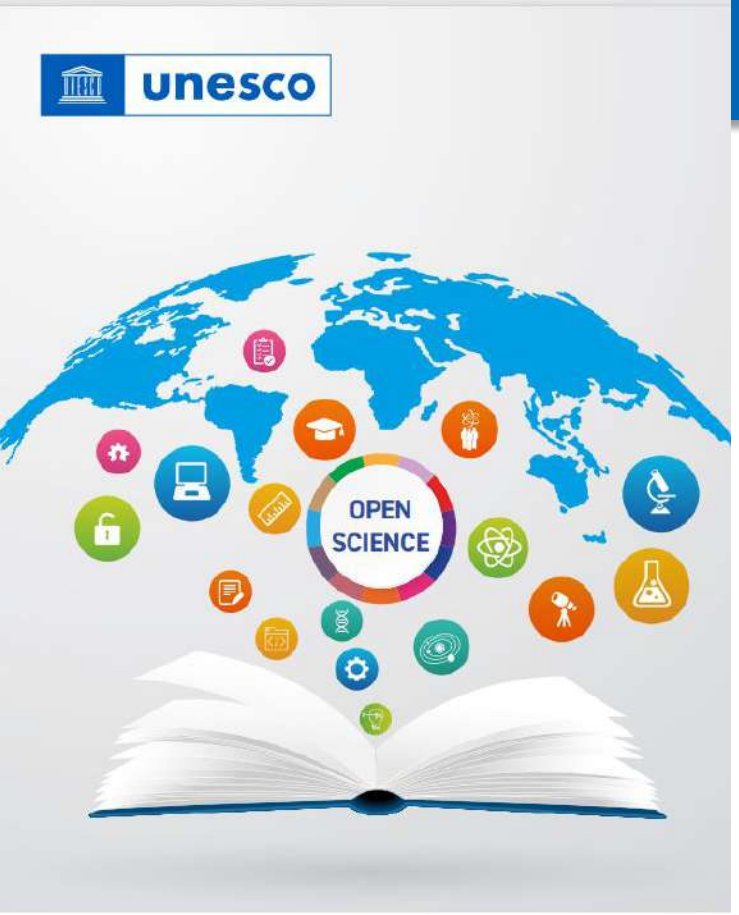
- Proposal for an Open Science Monitoring Framework: Marin Dacos and Nicolas Fressengeas (French Ministry of Higher Education and Research)

Monitoring Open Science beyond open access to scientific knowledge: Reflections of the UNESCO Working Group on a Global Open Science Monitoring Framework

Ana Persic (UNESCO)

Monitoring Open Science beyond open access to scientific knowledge: Reflections of the UNESCO Working Group on a Global Open Science Monitoring Framework

Ana Persic, Programme Specialist, Science Technology Innovation Policy and Open Science



2021 UNESCO Recommendation on Open Science

- It is the first **international normative instrument** on open science;
- it contains the first **internationally agreed definition** of open science;
- it spells out the common **core values and guiding principles** of open science;
- it addresses **multiple actors and stakeholders** of open science;
- it recommends **actions on different levels** to operationalize the principles of open science;
- it proposes **innovative approaches for open science** at different stages of the **scientific cycle**;
- it calls for development of a **comprehensive open science monitoring framework**.

Key pillars of open science

Open Scientific Knowledge: scientific publications, research data, software, source code, hardware and educational resources available in the public domain or under copyright with open license

Open Science infrastructures: scientific equipment or sets of instruments, knowledge-based resources such as collections, repositories, archives and scientific data, open computational and digital infrastructures

Open engagement of societal actors: collaboration between scientists and societal actors beyond the scientific community, opening up practices and tools that are part of the research cycle by making the scientific process more inclusive and accessible to the broader inquiring society

Open dialogue with other knowledge systems: recognition of richness and complementarities between diverse epistemologies, including indigenous knowledge systems

Key Objectives – Key Areas of Action

Promoting a **common understanding** of OS and its associated benefits and challenges, as well as the **diverse paths** to OS

Developing an **enabling policy environment** for OS

Investing in **infrastructure** and services for OS

Investing in training, education, digital literacy and **capacity-building,** for researchers and other stakeholders

Fostering a culture of OS and **aligning incentives** for OS

Promoting **innovative approaches** to OS at different stages of the scientific process

Promoting **international and multistakeholder co-operation** in the context of OS with a view to reducing digital, technological and knowledge gaps.

Addressing the challenges for OSR Implementation

Working Groups	Deliverables
OS capacity building	• Compilation/index of the existing open science training modules and materials • Creation and delivery of new and additional necessary training modules on open science for different open science actors
OS policies and strategies	• Global Repository of Open Science Policy Instruments • Development of Open Science Policy Guide
OS financing and incentives	Proposals for regional and thematic open science funding mechanisms and recommendations for revision of the current research careers assessments and evaluation criteria
OS infrastructures	Support for /development of international, regional and thematic open science platforms for sharing of knowledge and best practices. Specific focus will be on thematic platforms in UNESCO's priority areas, including biodiversity, water, disaster risk reduction, geosciences, ocean sciences, climate change…
OS monitoring framework	Global monitoring framework for open science

Working Group: Monitoring the implementation of the ROS

To be kept under public oversight

A complex and multilayered **process**, that requires:

- ❑ inputs from different groups of stakeholders
- ❑ both qualitative and quantitative indicators
- ❑ responsible design of indicators
- ❑ use of open non-proprietary and transparent infrastructures, when possible
- ❑ use of available relevant indicator and data sources
- ❑ consideration of synergies and overlaps with existing monitoring frameworks
- ❑ identification of unintended consequences and potential negative effects
- ❑ multi-stakeholder participatory approach, including scientific community

Working Group: Monitoring the implementation of the ROS

- **Open Science monitoring is a complex and multilayered exercise** which might require a "**pluralistic monitoring framework**" including:

  - survey for Member States to report on policies and actions promoting open science in line with the Recommendation on Open Science
  - an output analysis based on inclusive 'Global Scientific Scholarly Database(s)'
  - surveys to research organizations
  - opinion surveys to actors regarding the values and practice of open science

Shared expertise and exploration is essential as open science evolves

Survey for Member States reporting on the implementation of the Recommendation

Every four years
1st round: 2024-2025

- Awareness raising on the Recommendation and on OS,
  Incorporation of OS values and principles of OS in research
- National (and institutional) policies, policy instruments, legal framework, funding mechanisms, and monitoring framework on OS, its key elements, or on STI
- Indicators for R&D expenditure and internet connectivity
  Accessibility and inclusivity of national, regional and international infrastructures
- Systematic capacity building on OS for researchers, policy makers and for science communication
- National (or institutional) initiatives to align research assessment with OS
  Recognition and reward of OS practices (how and by whom?)
- National or institutional initiatives on innovative and participatory methods to open different stages of the scientific process.
- Strategies for or involvement in cross-border multi-stakeholder collaboration on OS, e.g international funding mechanisms

Output analysis based on open inclusive global databases

What aspects of open science **should** be measured?

What aspects of open science **can** be measured?

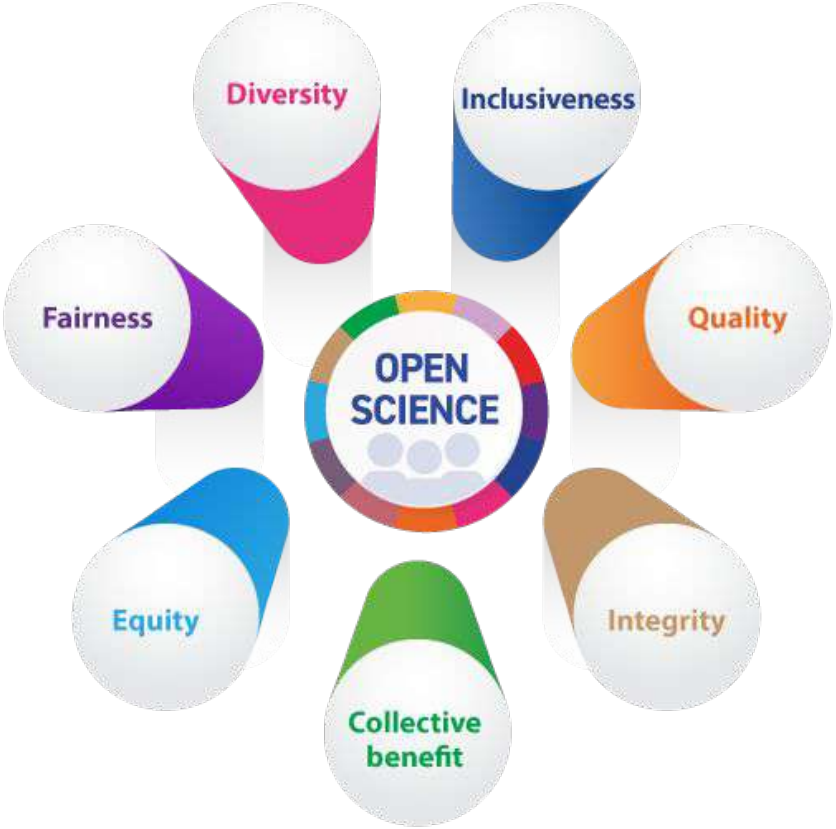
What **indicators** should be used?

Which **data sources** are the most relevant/reliable/inclusive/comparable?

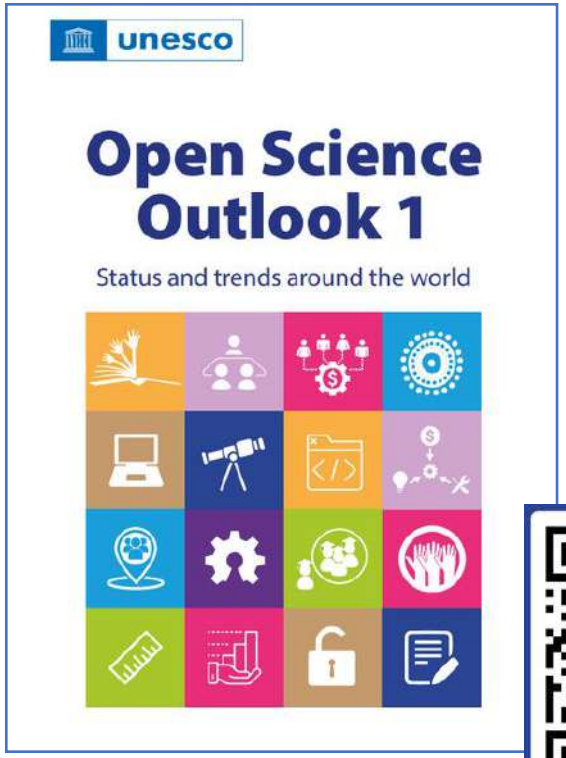
What are the **gaps** in available data sources?

Graphic: Shutterstock/Visual Generation

Working Group: Open science values provide a shared framework



Open Science Outlook



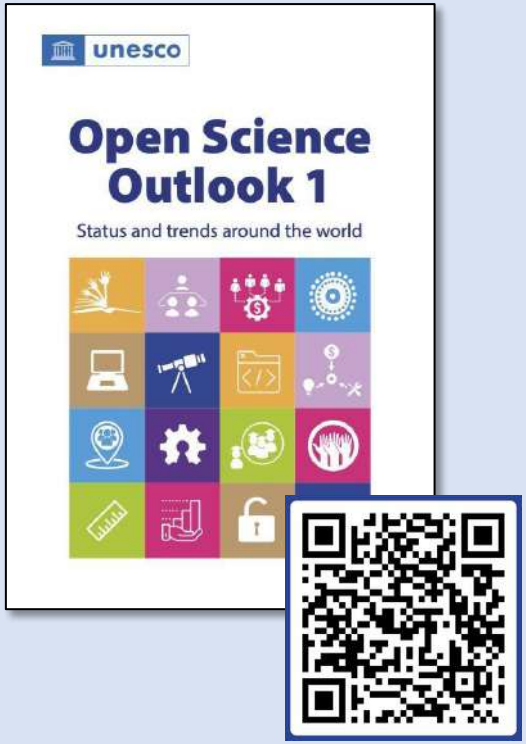
Counting is not enough
Current system of rankings do not promote inclusion, equity and openness

- standard approaches & **existing indicators are insufficient** to monitor openness across the scientific cycle & all pillars of open science
- innovation needed in **open qualitative & quantitative assessments** to monitor change & align with the values & principles of open science
- **overall need to monitor a comprehensive transformation** to open science & its impacts on STI systems and on society

Open Science Outlook

Opportunity to strengthening the focus on values and people, not just products

Thank you!



Join the UNESCO Open Science Working Groups

Contribute to global open science calls

Be in touch openscience@unesco.org

UNESCO Open science website:
https://www.unesco.org/open-science

Proposal for an Open Science Monitoring Framework

Marin Dacos and Nicolas Fressengeas (French Ministry of Higher Education and Research)

May 2023 G7 Communique (excerpt)

G7 cooperation on open science is set to continue, in particular to encourage a framework for monitoring the progress and obstacles of open science.

The need for a global coordination

Initiatives are flourishing worldwide

- Diversity and multiplicity is good news
- Distinct or incompatible ways may not be

Could we agree on principles for monitoring ?

- Preserve initiatives, diversity and local needs
- Towards a common shared goal

Declaration on Open Research Information

Barcelona

Another initiative on research monitoring

- Agreement on opening the present Research Information
- Focus on opening already available metadata

Link with our Principles proposal ?

- Our Principles would be guidelines for the building of new research information, on research openness
- Which would then need to be opened

The Principles for Open Science Monitoring

POSM

An output for today's meeting ?

- Draft proposed and shared
- Many comments already there
- Further input welcome

Two afternoon breakout sessions

- Further discussions and input
- To get to a common agreement

Workflow proposal

Today’s meeting for discussions and inputs

- Share your views using all canals
  - Google Doc
  - Breakout Sessions
  - ...

For a final set of Principles in 2024

- Shared for a final round after the meeting
- Before going public

A set on Principles in 2024… what for ?

Public dissemination ?

- How ?
  - Open Archive ?
  - A specific website ?
  - ...

What next ?

- Public endorsement ?
- On a specific website ?
- … other ideas ?

Principles : Part 1. *Relevance* of the monitoring

Everything that can be counted does not necessarily count

Monitoring should…

- Be meaningful for public policy
- Be consensual
- Be comprehensive
- Include a core set of indicators
- Foster comparison at international level
- Be mature
- Favor quality over quantity

Principles : Part 2. *Transparency and reproducibility*

Monitoring should…

- Document processes and methodology
- Be transparent and explicit about indicator quality
- Provide pragmatic indicators
- Allow its input data to be reused
- Produce open and FAIR output data
- Be open source
- Explicit data lineage and licenses
- Allow accountability by third parties

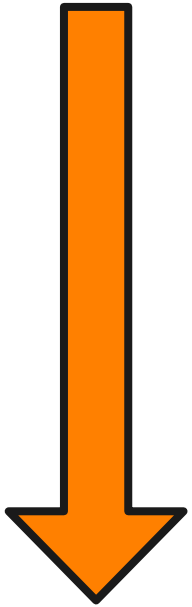
Principles : Part 3. *Governance*

Monitoring should

- Be reusable through API
- Be able to self assess against the Principles
- Be standardized, at least for core indicators
- Undergo continuous assessment

From principles to actual monitoring

Expected outcome for today



Principles of Open Science Monitoring

Core set of indicators

Technical Specifications

Breakout rooms

After Lunch

Principles

- 2 breakout rooms

Community building

- 1 breakout room

Technical Specifications

- 2 breakout rooms

Principles breakout rooms

Room 1

- Review principles starting with part 1
- Define the core set of indicators
- Try to shorten the document

Room 2

- Review principles starting with part 2
- Define the core set of indicators
- Try to shorten the document

Community breakout room

Towards a community for the monitoring of Open Science

Which initiatives are we missing ?

What form should take this community ?

At what frequency should it meet ?

What is its scope ?

…

Specifications breakout rooms

- Publications
- Clinical trials
- Research dataset
- Software and code
- Costs
- Usages
- Impact
- …

Shared views ?