



MINISTÈRE  
DE L'ENSEIGNEMENT  
SUPÉRIEUR  
ET DE LA RECHERCHE

*Liberté  
Égalité  
Fraternité*



UNIVERSITÉ  
DE LORRAINE

*Inria*



# 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



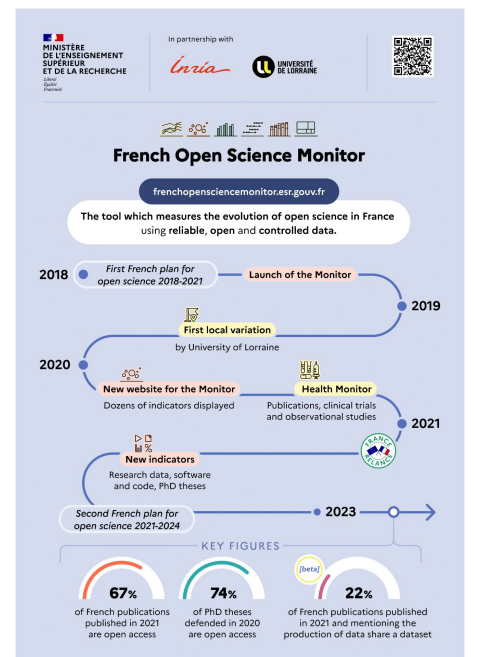
## MINISTÈRE DE L'ENSEIGNEMENT SUPÉRIEUR ET DE LA RECHERCHE

*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](http://www.ouvrirlascience.fr)

[frenchopensciencemonitor.esr.gouv.fr/](http://frenchopensciencemonitor.esr.gouv.fr/)

[marin.dacos@recherche.gouv.fr](mailto: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

A dark blue diagonal graphic that starts from the bottom left corner and extends towards the top right corner, covering the lower half of the slide.

# 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 31<sup>st</sup>, 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).<sup>1</sup> Issued by the White House Office of Science and Technology Policy (OSTP), the 2013 Memorandum

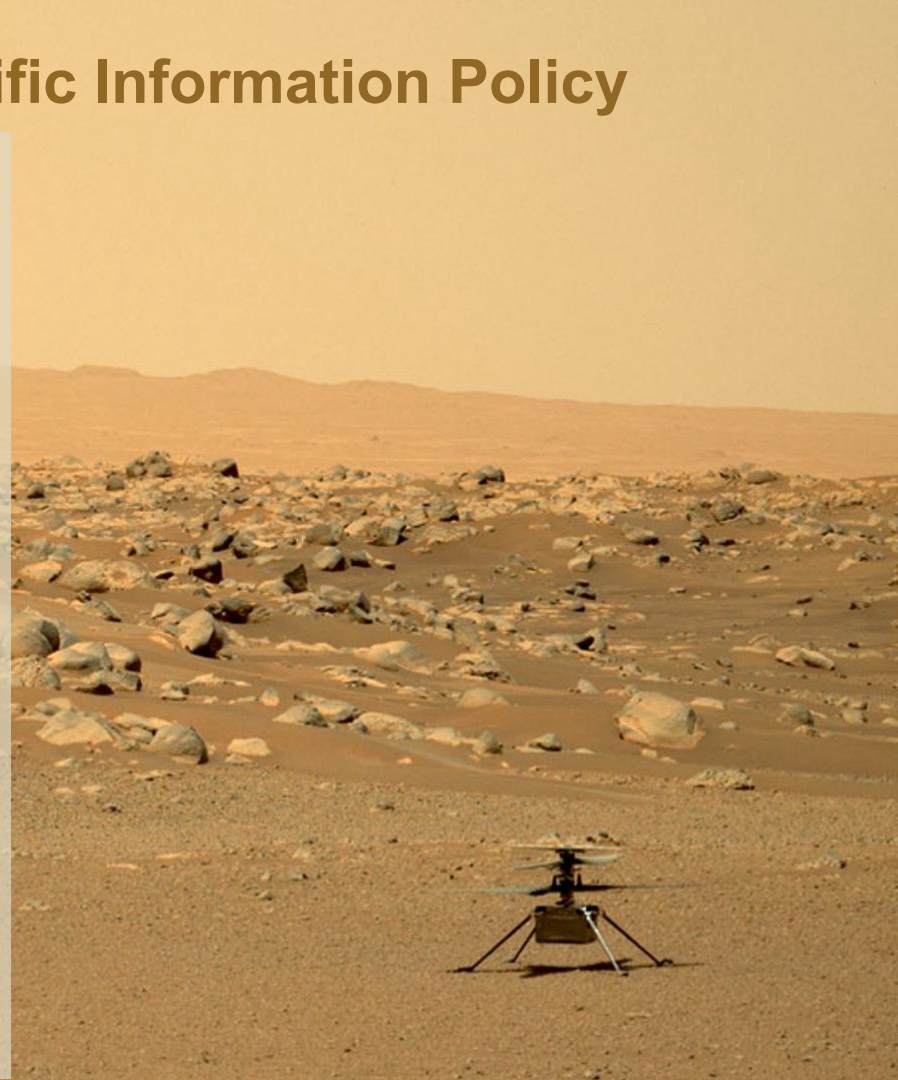
Released in August 2022 with the requirements that agencies update their Research Access plans to include **immediate and free access to publications and data** and to ensure research integrity and equity.

- 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

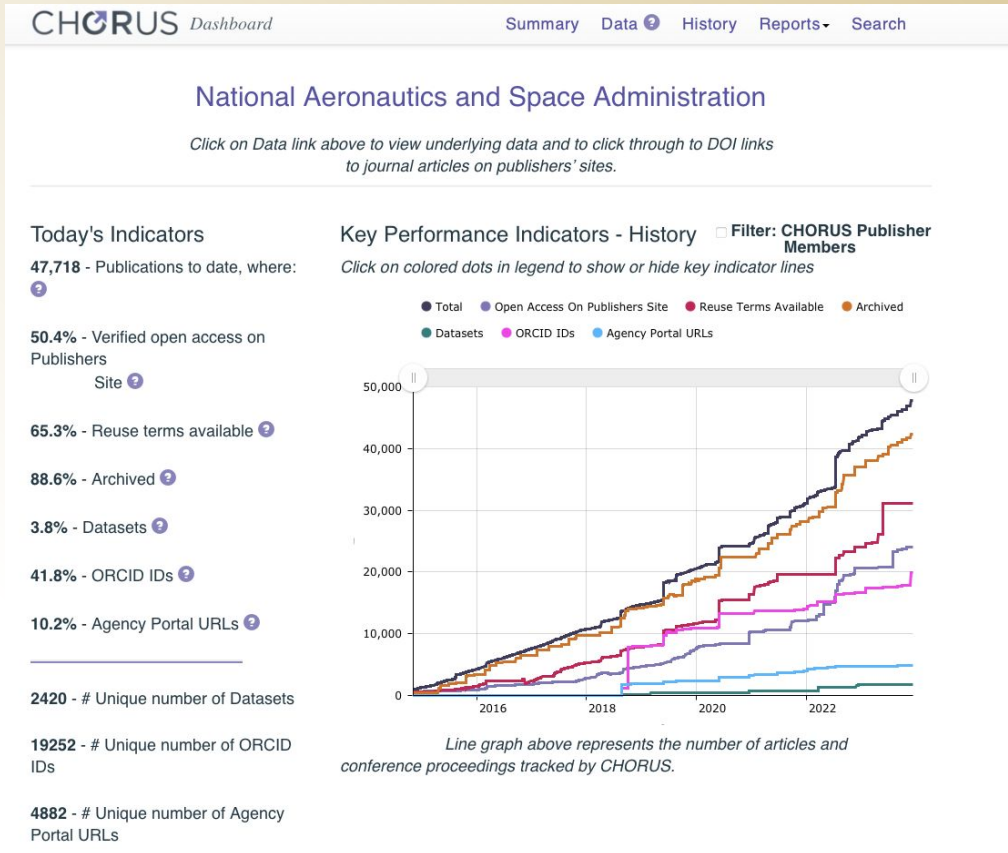
Use open science infrastructure to aid tracking with DOI's on awards: [Dept. of Energy example.](#)

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>



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/>

The screenshot shows the NASA ADS search interface. The search bar contains the query "ack:NASA\* year:2010-2023" and shows "Your search returned 105,937 results". The left sidebar lists various filters such as "AUTHORS", "COLLECTIONS", "REFEREED", "INSTITUTIONS", "KEYWORDS", "PUBLICATIONS", "BIB GROUPS", "SIMBAD OBJECTS", and "NED OBJECTS". The main content area displays a list of search results, including titles like "Possible connection between solar activity and local seismicity" and "PKGui: A GUI software for Polubarinova-Kochina's solutions of steady unconfined groundwater flow". On the right, there is a bar chart showing the number of "refereed" (blue) and "non-refereed" (green) articles from 2010 to 2020.

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](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

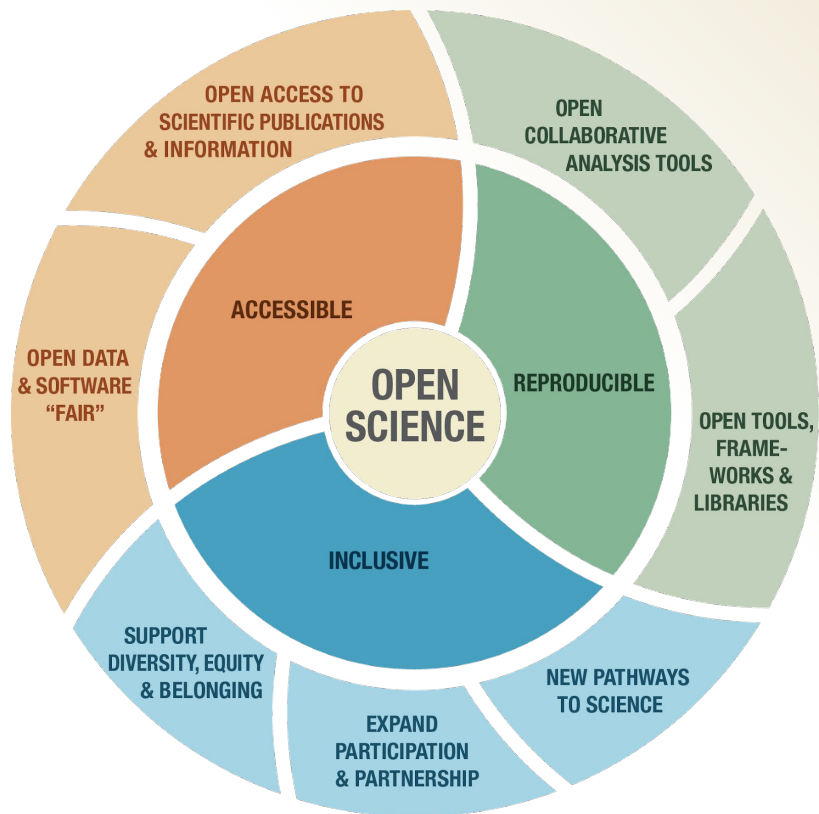
Along with data and publications, NASA also mandates the release of scientific software developed as part of a publication. Credit and recognition for scientific software is relatively new.

Possible metrics and measurement methods:

- Downloads, installations, usage, imports/dependencies, contributors, services
- Citations, Altmetrics, ImpactStory, libraries.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](https://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 be transparent and reproducible, and thus **completely independent from any proprietary data source**.

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

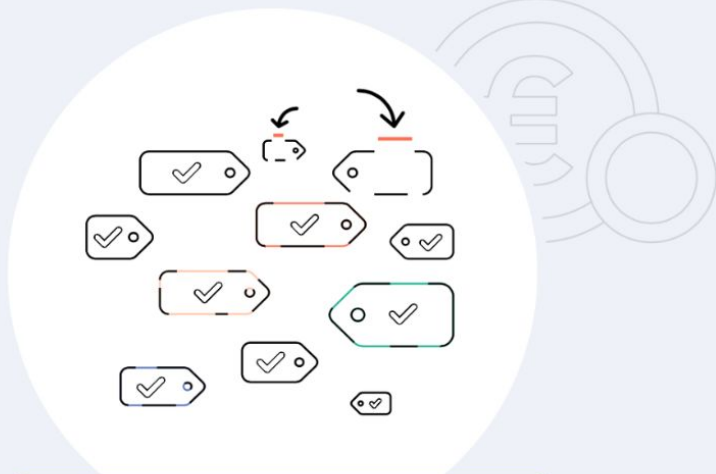
## Second French Plan for Open Science



# As of 2018: the metadata gap between proprietary and open

## Proprietary bibliographic databases

remedy these defects  
by enriching these metadata

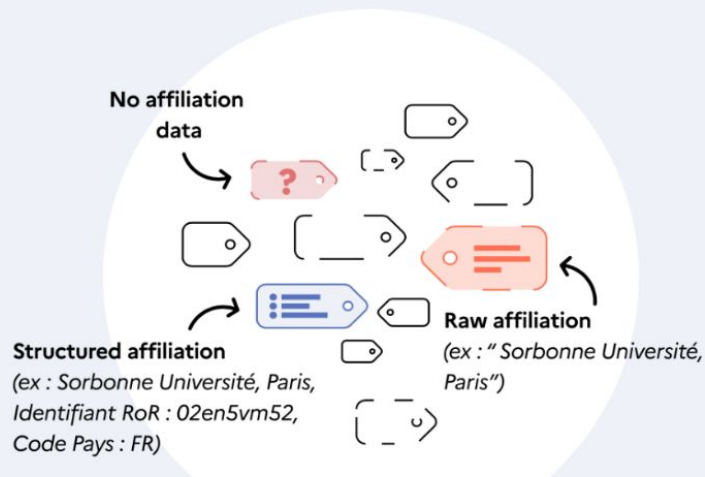


Proprietary bibliographic databases:

- 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

# As of 2018: the metadata gap between proprietary and open

## Open metadata does not exist?

Proprietary bibliographic databases  
remedy these defects  
by enriching these metadata

Open bibliographic databases  
offer a low amount of affiliation metadata  
and of disparate quality

### Let's try to create it!

### How ?

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

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

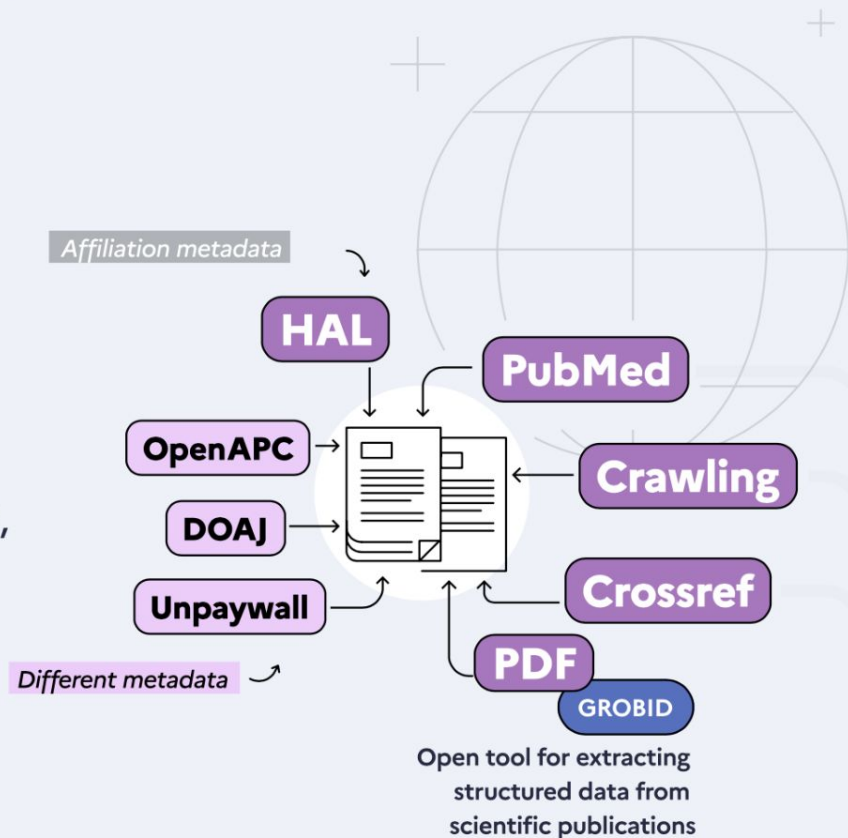
# Our open methodology

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.

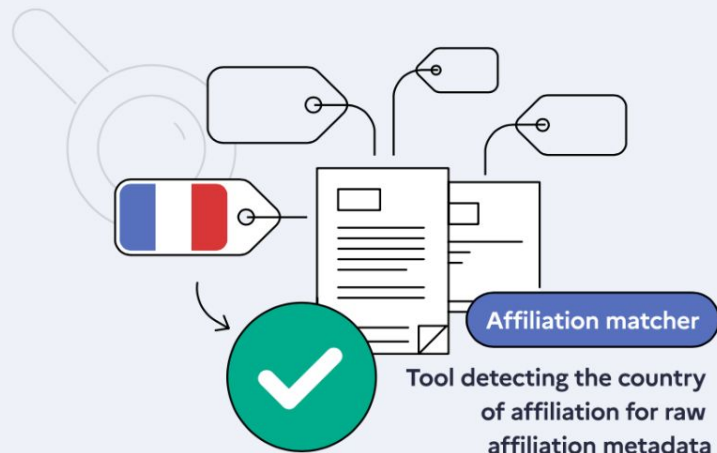
#### Detection rate of french scientific publications



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).



"Sorbonne Université, Paris" → France ✓

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

### 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](https://doi.org/10.1162/qss_a_00179)

### #3 Enhance...

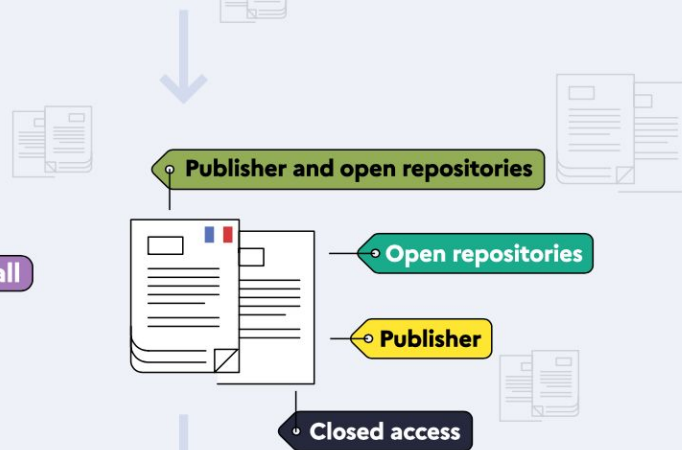
#### ... the opening status

For crossref DOI:

the information stems from **Unpaywall**

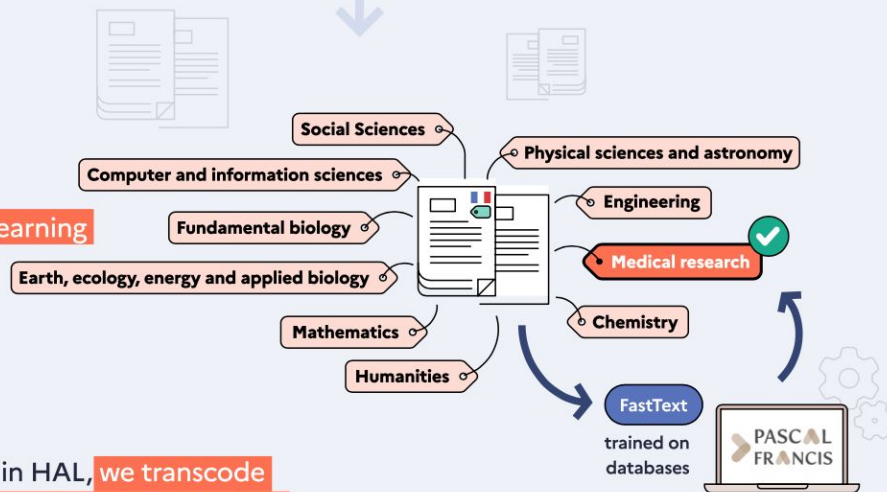
For publications in HAL (no DOI):

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 initio* 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 PROCRA (Wiederstein and Sippl, 2007).

#### Southern blot analysis

Genomic DNA of foxtail millet was extracted from leaves using the cetyltrimethylammonium bromide (CTAB) method (Saghai-Maroo *et al.*, 1984), digested with *Pvu*II and *Hind*III (New England Biolabs), fractionated in a 1.0% agarose gel, and blotted on a Hybond N<sup>+</sup> membrane (Amersham). The blots were hybridized to a 705 bp *SiDREB2* probe radioactively labelled with [ $\alpha$ -<sup>32</sup>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.

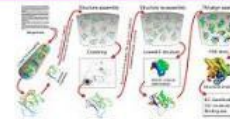
#### Subcellular localization of the *SiDREB2* protein

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 structure 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	<a href="http://europepmc.org/articles/pmc2782770?pdf=render">http://europepmc.org/articles/pmc2782770?pdf=render</a>
publisher	Wiley

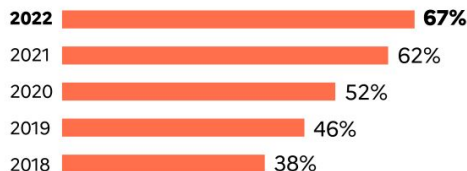
I-TASSER (Iterative Threading ASSEMBly Refinement) is a bioinformatics method for predicting three-dimensional structure model of protein molecules from amino acid sequences. It detects structure templates from the Protein Data Bank by a technique called

# Main results of the French Open Science Monitor

[FrenchOpenScienceMonitor.esr.gouv.fr](https://FrenchOpenScienceMonitor.esr.gouv.fr)

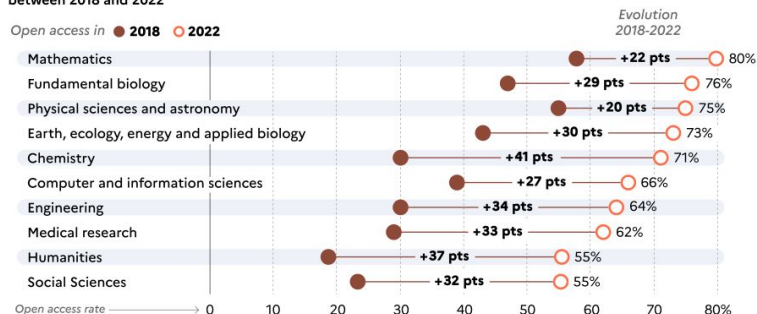
## Publications

Open access rate of scientific publications in France, with a Crossref DOI, published during the previous year, by observation year



Growth  
(all fields)  
2018-2022  
**+29 points**

Rate of open access publications in France, for each discipline between 2018 and 2022

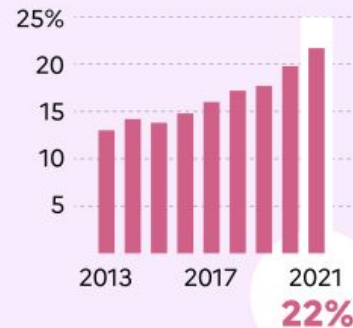


## Research data and software

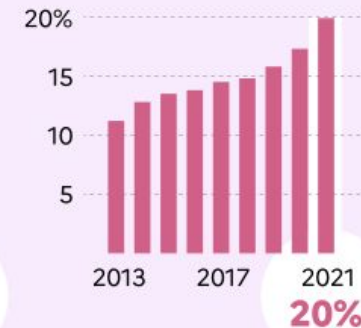
[beta]

Proportion of publications that share:

A dataset



A software or code



Share of clinical trials registered and completed in France in the past 10 years that have posted or published results

All types of lead sponsor\*:



Industrial lead sponsor:



Academic lead sponsor:




\* Individual or legal entity in charge of research conducted on human beings who initiates, finances and supervises the conduct of the clinical trial.

# Lessons learnt

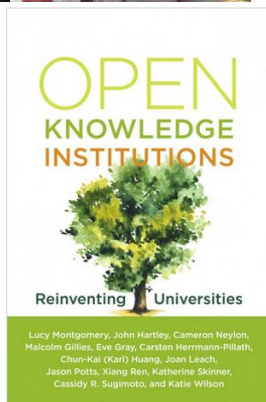
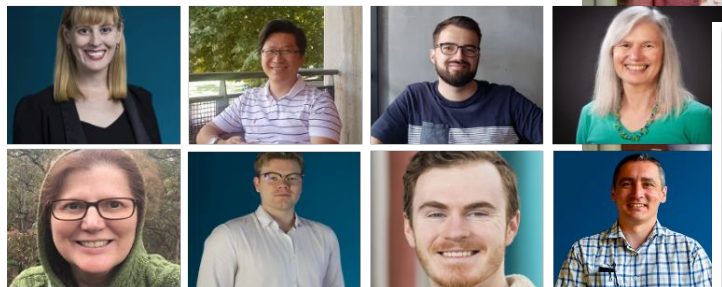
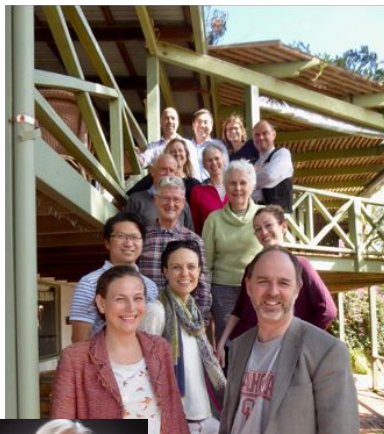
- It is possible to build an open science monitor at the national level without any proprietary data, taking advantage of the progress in **machine learning** and **cloud computing**.
- An iterative process is needed to improve and extend the results
- Collaborating at different scales is key
  - Necessity 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)

A dark blue diagonal graphic that starts from the bottom left corner and extends towards the top right corner, covering the bottom half of the slide.

# 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

# THE AUSTRALIAN CONTEXT

---

Wednesday, 15 November 2023

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

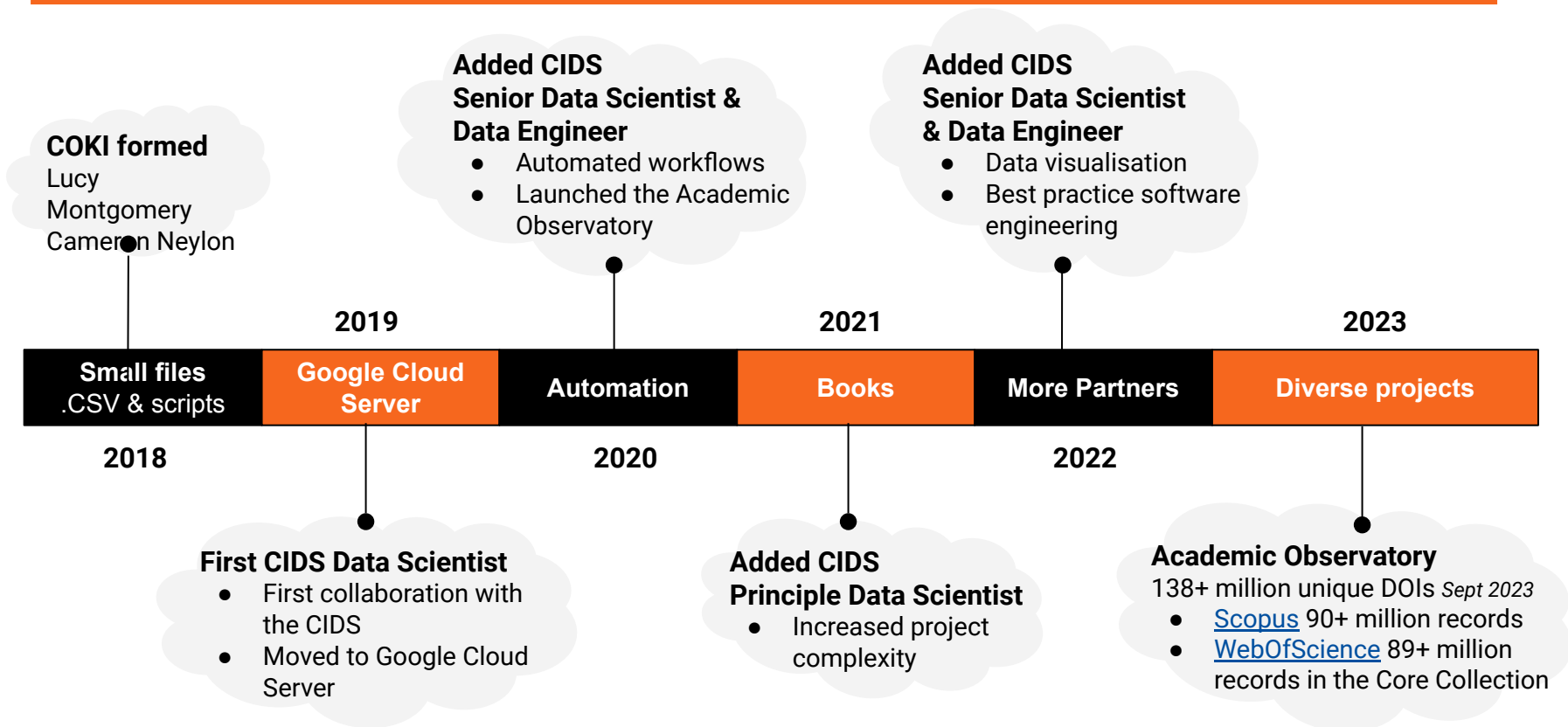
*“The current system for assessing research careers for hiring, promotion and funding is not fit for purpose.”*

*“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

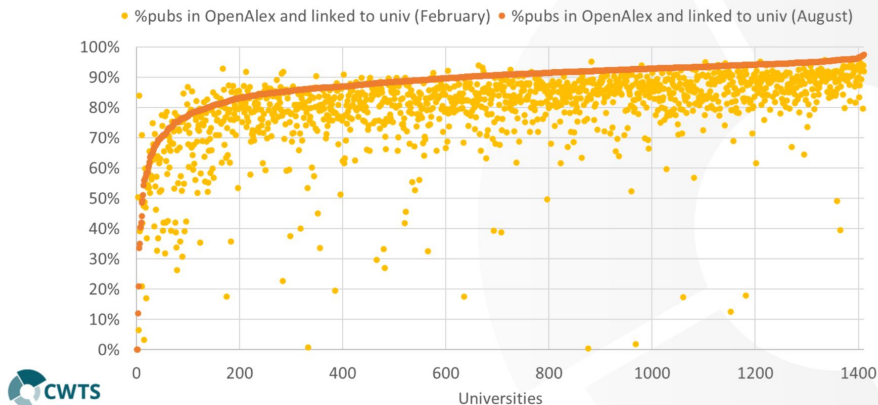


# COKI & Curtin Institute for Data Science



# MEANWHILE...NEW OPPORTUNITIES

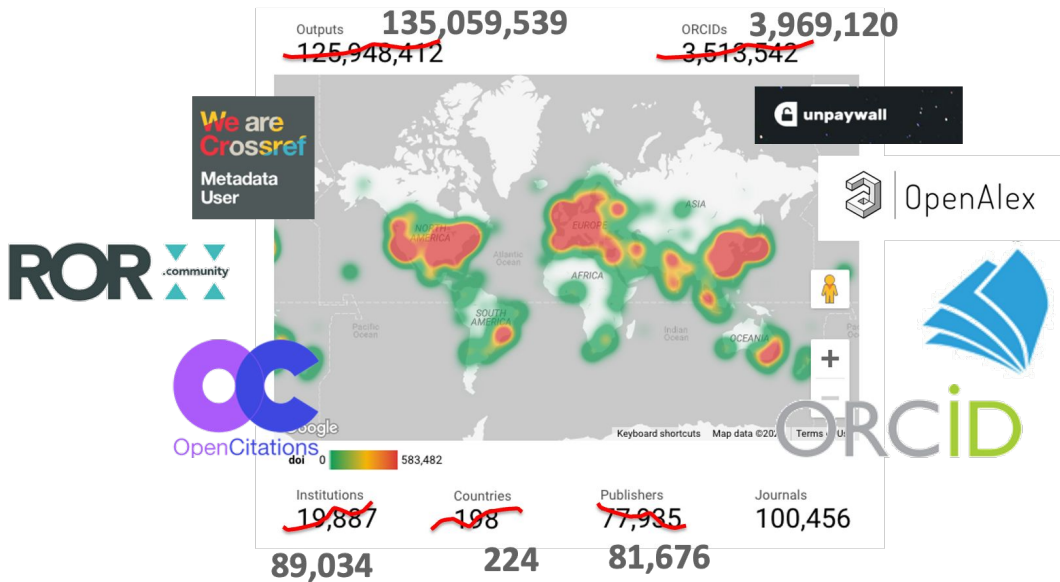
## Publication coverage of LR2023 universities: Improvement of false negatives in OpenAlex



- 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>

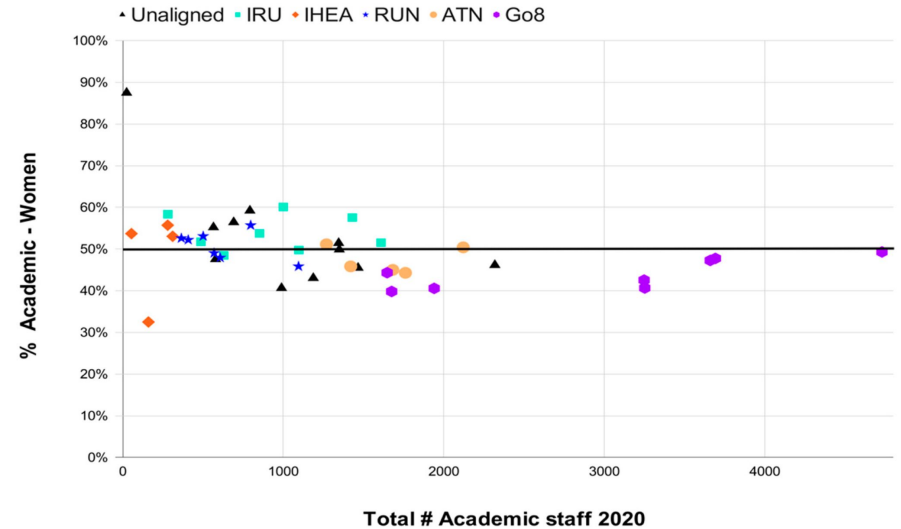
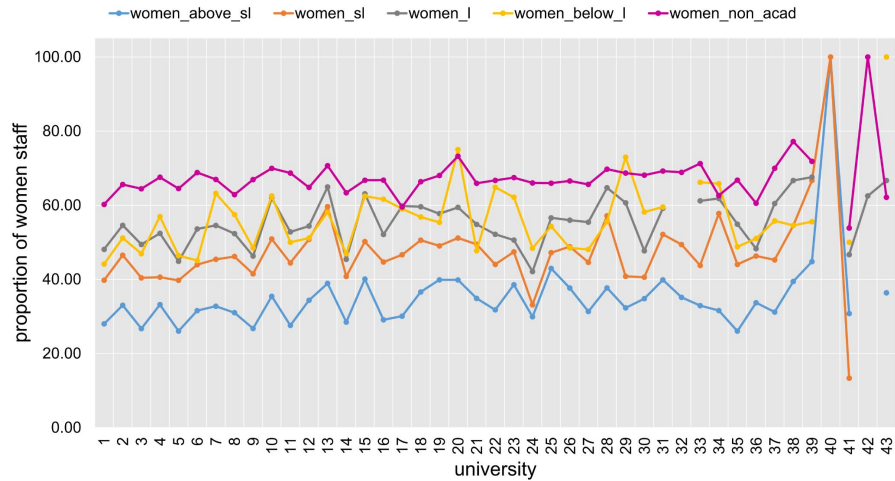
## MEANWHILE...NEW OPPORTUNITIES



- 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

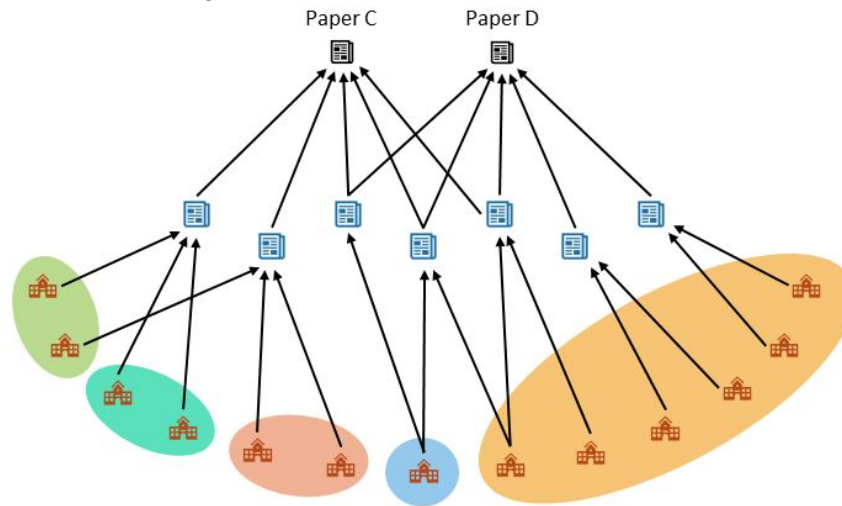
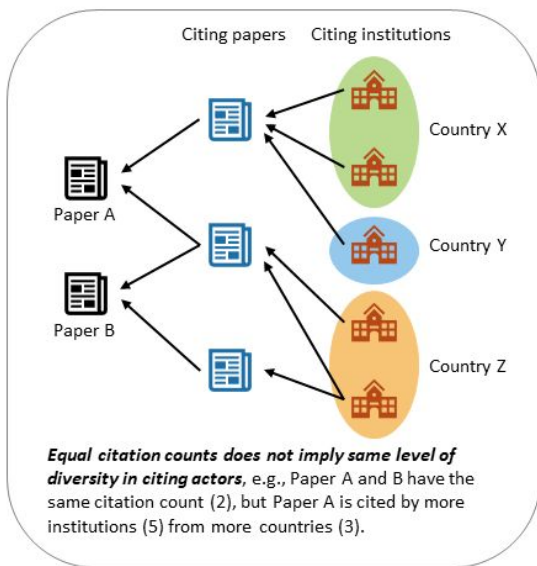
- Diversity in knowledge production and dissemination:



(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:



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.

HOME

## 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			
COUNTRY	OPEN ↓	BREAKDOWN PUBLISHER OPEN BOTH OTHER PLATFORM OPEN CLOSED	TOTAL PUBLICATIONS	OPEN PUBLICATIONS	
SÃO TOMÉ AND PRÍNCIPE	93%		1,633	1,519	<a href="#">LEARN MORE &gt;</a>
INDONESIA	89%		1,077,670	959,400	<a href="#">LEARN MORE &gt;</a>
NICARAGUA	82%		5,554	4,574	<a href="#">LEARN MORE &gt;</a>
EQUATORIAL GUINEA	82%		153	125	<a href="#">LEARN MORE &gt;</a>
PERU	80%		90,707	72,589	<a href="#">LEARN MORE &gt;</a>

FILTERS

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



redalyc AmeliCA

Building an Open Science  
monitoring framework with  
open technologies



# Monitoring Open Science in the South

## Monitoring Open Science from the essential 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)



# When open is the default, what should be monitored?

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

Distributed investment  
Universal benefit

12 country-level networks of institutional repositories



AmeliCA latindex



Natural  
Open  
Access

Academic journal publishing

~12.000 online journals

~2.700 quality-certified journals

63 OA mandates

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



>3.000 OJS  
installations

Institutional journal portals  
and repositories

1.460  
diamond OA  
journals  
published by  
700 academic  
institutions

No fees neither for  
authors nor for readers

Owned by academic  
sector

Nonprofit

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?

# Distributed investment

Not expenditure

## Investment

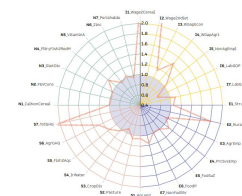
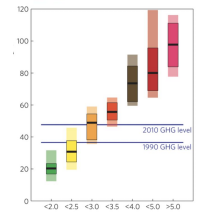
- Implications from the commodification
- Non-commercial ecosystem degradation
  - Total number of journals sold to commercial publishers
  - Total number of journals flipped to APC
- Investment in Open Science infrastructure
  - Repositories
  - Journals' sustainability
  - Open data
- R&D Personnel (FTE)
- Labour force (FTE)

## Infrastructure indicators

- Open infrastructures sustained by the country
- Open infrastructures sustained by the institution
- Open infrastructures services
- Open infrastructures contribution to O

## Business models

- 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



Openness 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.

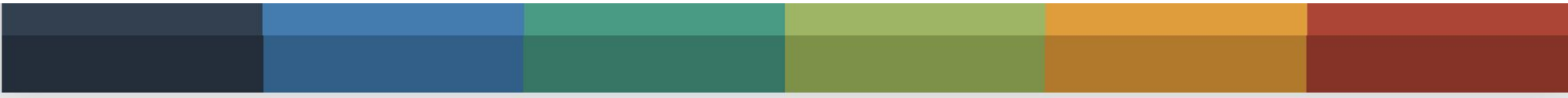
*First adopted in Europe, the wave of TAs has now reached libraries in Asia, Africa, the Americas and Australia. With more than half a million new research articles published openly through TAs negotiated by institutions in 67 countries to date, there can be no doubt that TAs increase global access to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

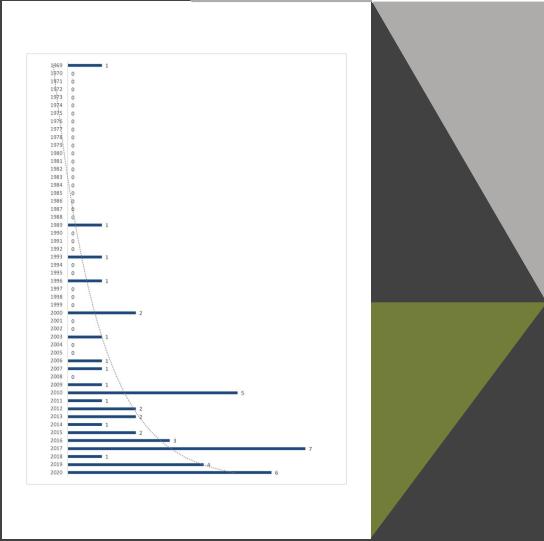
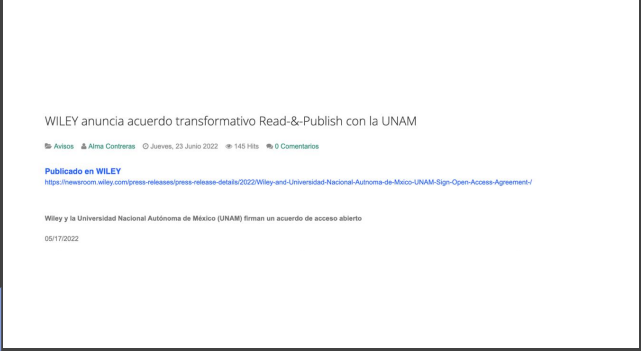
Colleen Campbell, Ádám Dér, Kai Geschuhn and Ana Valente (2022). How are transformative agreements transforming libraries? IFLA, WLIC, Dublin

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

d a t O S

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.



- 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, multilingualism, diversity



# Unveil the structure of knowledge

The image shows a screenshot of the Redalyc.org website. At the top, the Redalyc.org logo is visible, along with navigation links for 'Acerca de Redalyc', 'Acceso Abierto Diamante', 'Principios y Valores', 'Tecnología de Publicación Digital (XML JATS)', 'Indexación de Revistas', 'Servicios', and 'ESP ENG'. Below the navigation bar, there is a search bar and several menu options: 'Artículos por palabra clave', 'Representación del conocimiento', 'Artículos por país', 'Base de conocimiento en SPARQL', and 'Acerca de'. The main content area features a network visualization of terms related to COVID-19. The central node is 'COVID', which is connected to numerous other nodes, including 'epidemic', 'SARS', 'contagion', 'epidemics', 'infectious diseases', 'coronavirus', 'diseases', 'epidemiology', 'pandemia', 'pandemics', 'infectious', 'epidemiologic', 'coronavirus', 'epidemiology', 'contagion', 'epidemic', 'epidemiological', 'pandemia', 'infectious diseases', 'pneumonia', 'flu', 'tuberculosis', 'Influenza', 'bubonic plague', 'disease', 'infection', 'virus', 'disaster', 'outbreaks', 'disease outbreaks', 'outbreak', 'endemic', 'epidemics', 'disease', 'outbreak', 'Disaster', 'infestation', and 'pest'. A tooltip is displayed over the 'COVID' node, listing related terms: 'COVID Neighbors: epidemic, SARS, contagion, epidemics, infectious diseases, coronavirus, diseases, epidemiology, pandemia, pandemics'. To the right of the network visualization, there is a list of recommended articles, each with a title and a brief description of the journal and issue.

redalyc.org UNAM

Acerca de Redalyc Acceso Abierto Diamante Principios y Valores Tecnología de Publicación Digital (XML JATS) Indexación de Revistas Servicios ESP ENG

Buscador Artículos por palabra clave Representación del conocimiento Artículos por país Base de conocimiento en SPARQL Acerca de

COVID Neighbors:  
epidemic  
SARS  
contagion  
epidemics  
infectious diseases  
coronavirus  
diseases  
epidemiology  
pandemia  
pandemics

COVID

epidemic  
SARS  
contagion  
epidemics  
infectious diseases  
coronavirus  
diseases  
epidemiology  
pandemia  
pandemics

epidemiologic  
coronavirus  
epidemiology  
contagion  
epidemic  
epidemiological  
pandemia  
infectious diseases  
pneumonia  
flu  
tuberculosis  
Influenza  
bubonic plague  
disease  
infection  
virus  
disaster  
outbreaks  
disease outbreaks  
outbreak  
endemic  
epidemics  
disease  
outbreak

Disaster  
infestation  
pest

Recomendaciones para realización de traqueostomías y atención de los pacientes traqueostomizados en Colombia durante la pandemia COVID-19  
Revista Colombiana de Cirugía, 2020 2(35)

COVID-19 y su relación con poblaciones vulnerables  
Revista Habanera de Ciencias Médicas, 2020 1(19)

Vitamina D, sus acciones "no clásicas" y su utilidad en la pandemia del COVID-19  
Revista de nefrología, diálisis y transplante, 2020 4(40)

Políticas implementadas por el gobierno mexicano frente al COVID-19. El caso de la educación básica  
Revista Latinoamericana de Estudios Educativos (México), 2020 Esp.-(L)

Labor preventiva e implementación de estrategias docentes durante la COVID-19 en la Universidad de Ciencias Médicas de Santiago de Cuba  
MEDISAN, 2020 8(24)

Desigualdades en Salud Bucal para Personas Mayores en Tiempos del COVID-19. La Teleodontología y la Odontología de Mínima Intervención como Caminos de Solución.  
International Journal of Interdisciplinary Dentistry, 2020 3(13)

Habitar el Valparaíso neoliberal: vivienda, hacinamiento y pobreza como marco de la pandemia  
O Social em Questão, 2020 48(23)

Materiales didácticos digitales y coronavirus en tiempos de confinamiento en el contexto español  
Práxis Educativa (Brasil), 2020 (15)

# 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

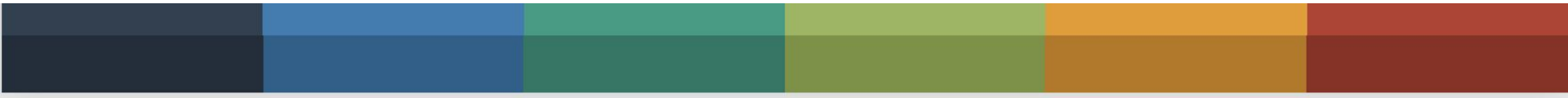


The harmful transition from communication  
to commodification



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](mailto:arianna.becerril@redalyc.org)

<http://www.redalyc.org/autor.oa?id=25>

<http://orcid.org/0000-0003-0278-8295>

@ariannabec



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)

A dark blue diagonal gradient bar that starts from the bottom left corner and extends towards the top right corner, covering the lower half of the slide.

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

 OpenAlex

Count Sort Column View Export Help

-  the work  open access 
-  the year is  2012- 
-  the institution country is **France**   
-  the SDG is **Life below water**   
-  the citation count is  1- 
-  the language is **NOT en**   
-  the work  from Global South 
-  the work  indexed by DOAJ 

[+ add filter](#) 

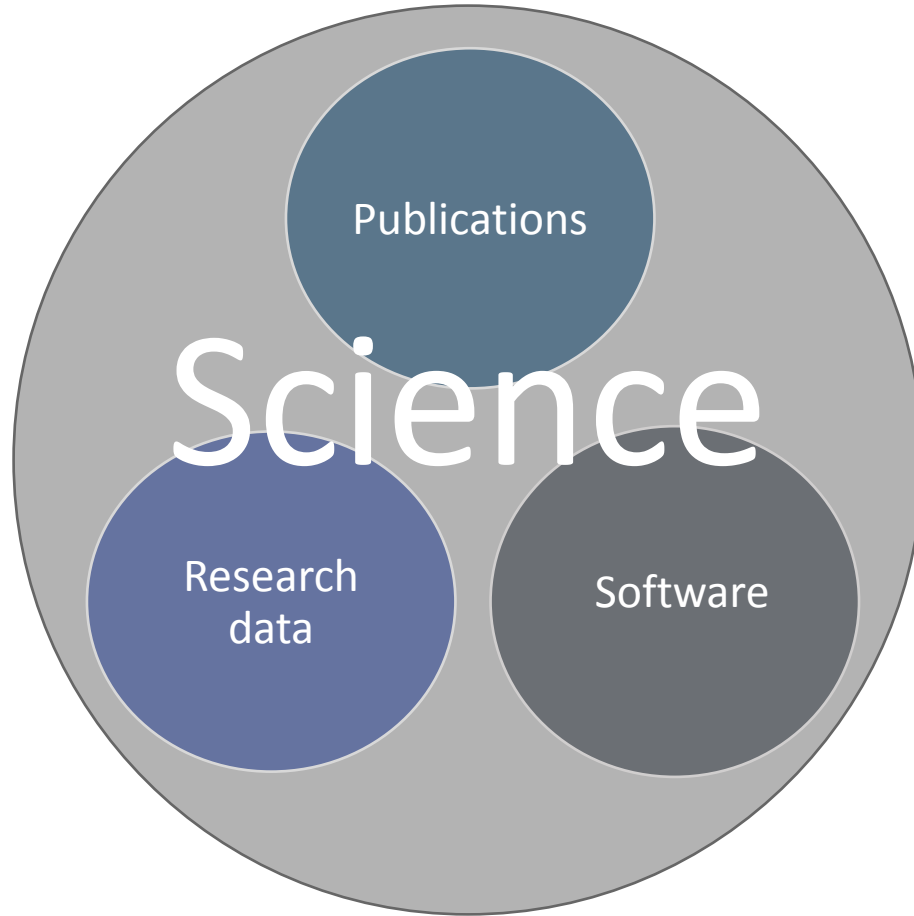
About 126 results

title	year	type	open access	citation count	Citation percentile (year)	language
Genus-level taxonomic changes implied by the mitochondrial phylogeny of grey mullets (Teleostei: Mugilidae)	2012	article		53	96	fr
Acoustic metamaterials for sound mitigation	2016	article		47	97	fr

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

Patrice Lopez (science-miner)

A dark blue diagonal gradient bar that starts from the bottom left corner and extends towards the top right corner, covering the bottom half of the slide.



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

## Citation

edpomacedo. (2023). edpomacedo/bdij-lexemes: v (wikibaseintegrator). Zenodo. <https://doi.org/10.5281/zenodo.10395844>

Style

APA



# 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

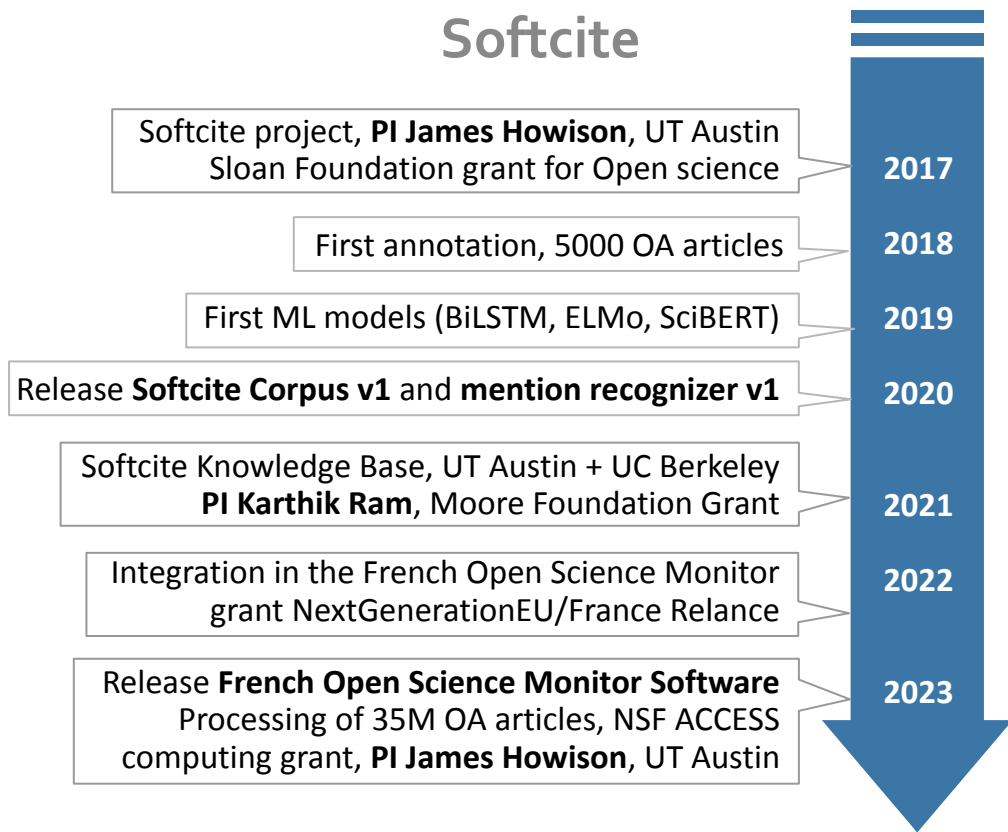
used



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

Softcite

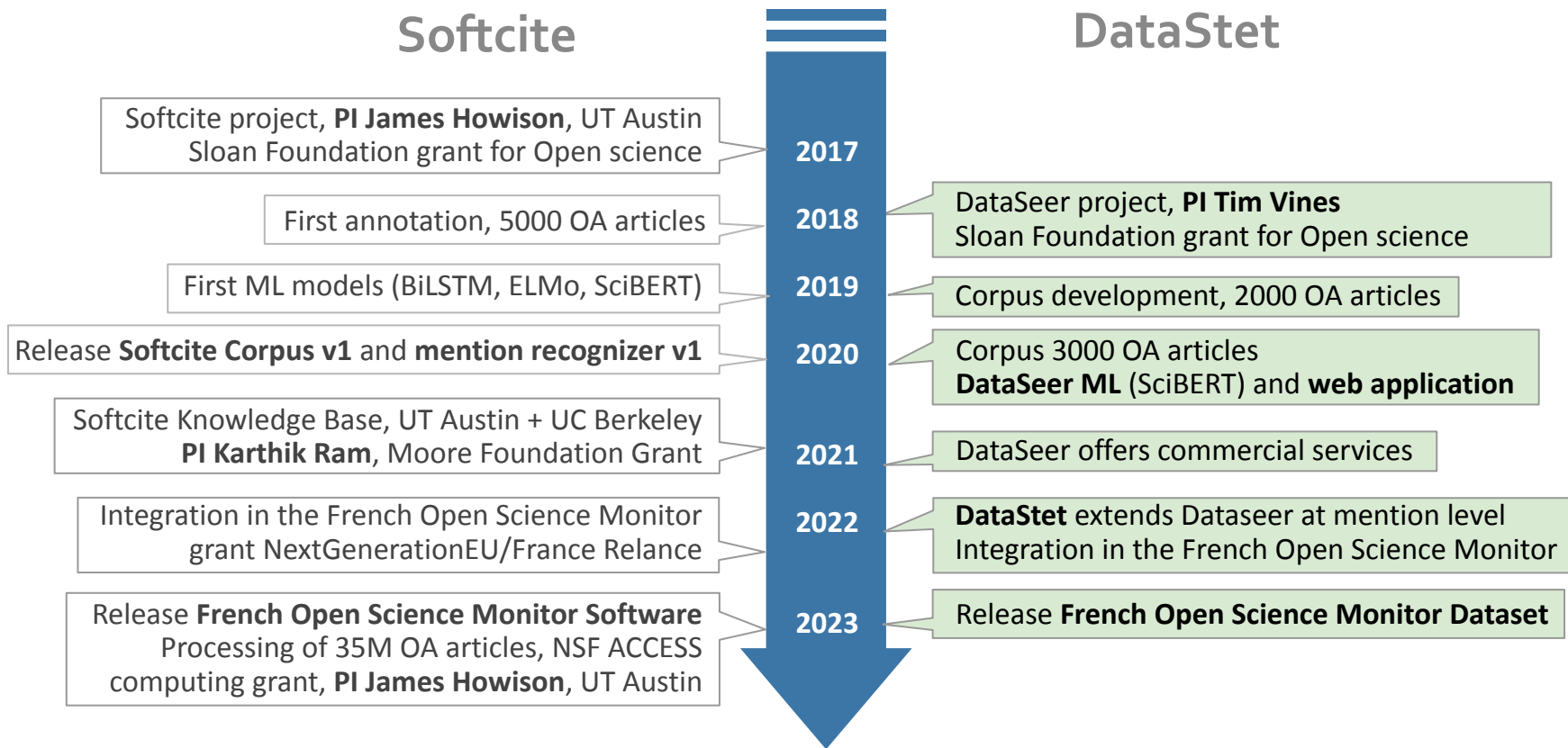
DataStet



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

## Softcite

## DataStet



## Softcite Mention recognizer

set work out, so we need to run many experiments carefully and quickly.

Solving this task has both computational aspects, namely comparing per-text, and domain-specific aspects, namely distributing the work of processing a corpus across a cluster. Most existing tools for data-oriented compilation require a multiplicity of domain-specific pipelines in a general-purpose language (e.g. [PINA](#), [ROSEFINKA](#), [Pilot](#), [Squeak](#) [4]), or vice-versa (e.g. [Pilot](#), [ROSEFINKA](#)) and [SQL](#)-based approaches [16, 17]. There are few languages where both compilation and data-partitions are first-class in readable expressions [\[18\]](#).

In practice, programmers use an assortment of tools, each with different strengths. A programmer might use a batch pipeline system, like [Pilot](#) or [Squeak](#) to iterate over the output of documents, and use [Hadoop](#) or [Cron](#) to schedule the pipeline. If the output is large, then a programmer will make a sequence of ad hoc queries to understand the result using a distributed SQL-like system, such as [Hive](#) or [Spark](#) [19], or even scripts running locally on a workstation.

Switching among languages and tools has both cognitive and practical costs. For our example, the programmer must write many different kinds of code (e.g. [Pilot](#) pipelines, [Hive](#) SQL, and [Hadoop](#) UDFs), having to complete code that is difficult to write, extend, or change. Because each task is in a different language, it is difficult to try out a task on small data as a local machine, or to re-do a single job on a high-level language if the intermediate data turns out to be messy enough. The programmer must duplicate code, and write code that is hard to read with similarities between the data models of each language and the on-disk data.

In this paper we describe Volok (Knowledge Logic), a declarative programming language based on [Datalog](#). Volok has two primary goals:

- Make computations over semi-structured data easy and succinct (Section 3). Data problems require lots of experimentation, and succinct code allows faster iteration. A declarative language makes programmers solve problems quickly by focusing on the high-level logic of a task instead of low-level details. Working with semi-structured data is often laborious and error-prone. Volok's data model is based around patterns.

**DRYADLINGO**

Type: software

Raw name: DRYADLINGO

References:

[4] Yu et al (2008)

authors: Yuan Yu, Michael Isard, Dennis Fetterly, Mihai Butiu, Ulter Erlingsson, Pradeep Kumar Gupta, Jim Conry

title: Dryad-ING: A system for general-purpose distributed data-parallel computing using a high-level language

date: 2008

book/ISS: Proceedings of Operating Systems Design and Implementation (OSDI)

conference: Operating Systems Design and Implementation (OSDI)

volume: 8

first page: 1

last page: 14

<https://github.com/softcite/software-mentions>  
[https://github.com/softcite/software\\_mentions\\_client](https://github.com/softcite/software_mentions_client)

## DataStet

TCGA gene expression dataset

Normalized [gene-level expression data](#), assayed by RNA-sequencing, for 817 primary breast cancers analyzed as part of the [TCGA](#) program was obtained from the [TCGA](#) data portal website: [ftp://cge-data.nci.nih.gov/TCGA](#). Details of the data processing can be found in [Ciriello et al.](#)

Association between AR primary tumor expression, clinical and tumor characteristics, chemotherapy response, and outcome

Associations between AR expression and clinical and tumor characteristics were assessed using the Wilcoxon rank-sum test (for two-level factors) or the Kruskal-Wallis test (for multi-level factors). The [clinical characteristics](#)

rgg Breast Cancer (2019) 47

page 6/7

**TCGA**

Type: dataset-name

Raw name: TCGA

URL: <http://cge-data.nci.nih.gov/TCGA>

References:

8 Ciriello et al (2015)

authors: Giovanni Ciriello, Michael Gatz, Katherine A Hoadley, Hsiki Zhang, Sukruth Pilla, Renee Beatty, Matthew Wilerson, Cyril Aksoy, Michael Menden, Andrew Chinnai, Peter W. Laird, Chris Sander, Tomi Rigg, Christofel Petrucci

title: Abstract S2-04. Comprehensive molecular characterization of invasive breast breast tumors

date: 2015-04-30

book title: General Session Abstracts

volume: 153

first page: 506

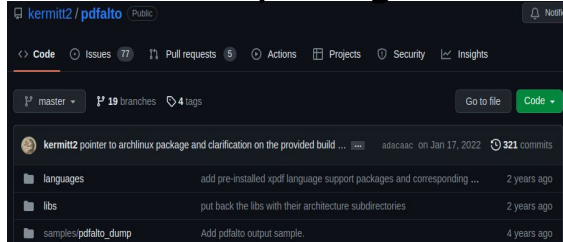
last page: 519

DOI: 10.1158/1538-7445.sbs014-02-04

publisher: American Association for Cancer Research

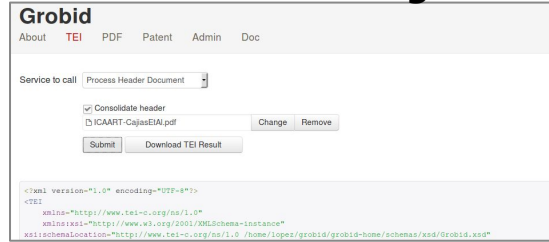
<https://github.com/kermitt2/datatset>

# PDF parsing



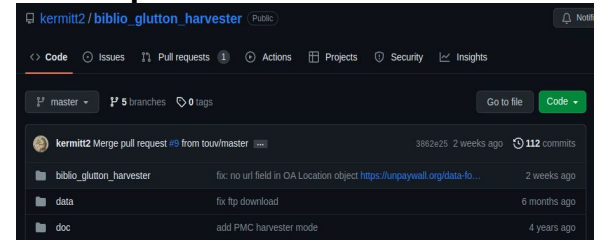
<https://github.com/kermitt2/pdfalto>

# PDF structuring



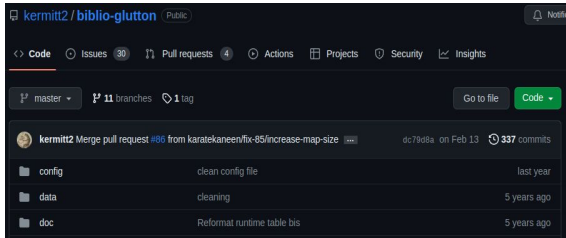
<https://github.com/kermitt2/grobid>

# Open Access harvester



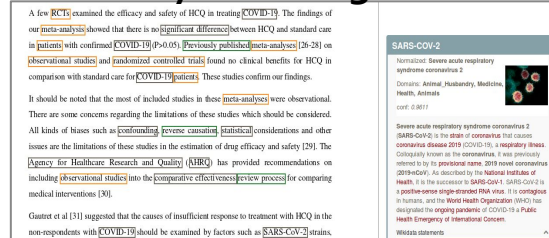
<https://github.com/kermitt2/biblio-glutton-harvester>

# Biblio. reference resolution



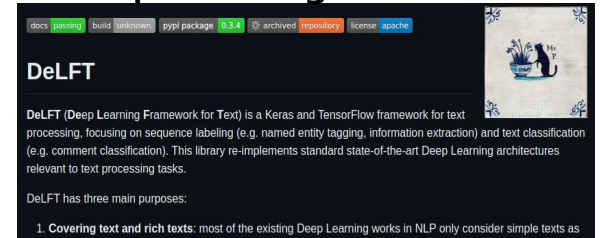
<https://github.com/kermitt2/biblio-glutton>

# Entity disambiguation



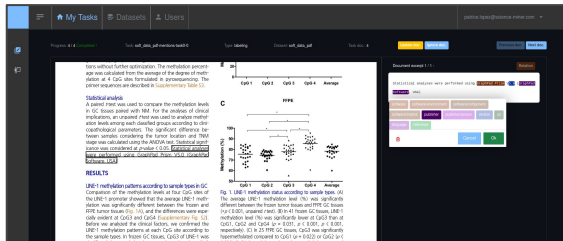
<https://github.com/kermitt2/entity-fishing>

# Deep Learning for rich text



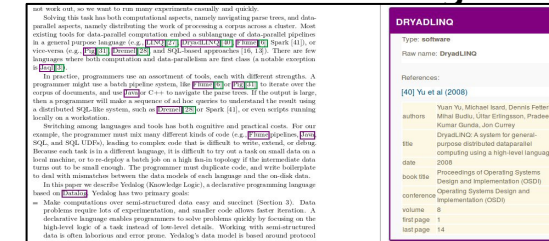
<https://github.com/kermitt2/delft>

# Manual annotations



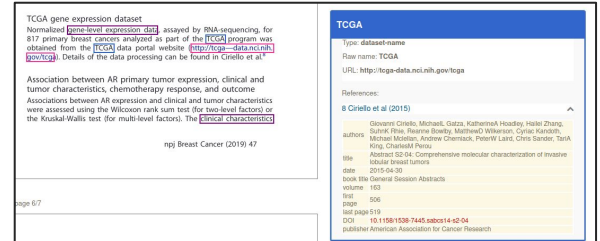
<https://github.com/kermitt2/kish>

# Softcite Mention recognizer



<https://github.com/kermitt2/software-mentions>  
[https://github.com/kermitt2/software\\_mentions\\_client](https://github.com/kermitt2/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	<b>94.94</b>	200
not data sentence	97.56	95.92	<b>96.73</b>	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	<b>89.24</b>	466
unnamed mentioned dataset	71.85	67.15	<b>69.38</b>	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%)
<b>software name</b>	74.01	88.98	<b>80.81</b>	989
version	83.99	90.81	<b>87.27</b>	283
publisher	75.51	88.80	<b>81.62</b>	250
url	53.97	82.93	<b>65.38</b>	41
<b>all (micro avg.)</b>	75.22	89.12	<b>81.58</b>	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 ?	<b>used</b>	96.83	94.18	<b>95.49</b>	292
	<b>not used</b>	84.40	91.09	<b>87.62</b>	101
mentioned dataset/software created/extended ?	<b>created</b>	81.08	83.33	<b>82.19</b>	31
	<b>not created</b>	98.31	98.04	<b>98.18</b>	362
mentioned created dataset/software shared ?	<b>shared</b>	81.82	90.00	<b>85.71</b>	26
	<b>not shared</b>	99.35	98.71	<b>99.03</b>	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	<b>63.7 %</b>	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

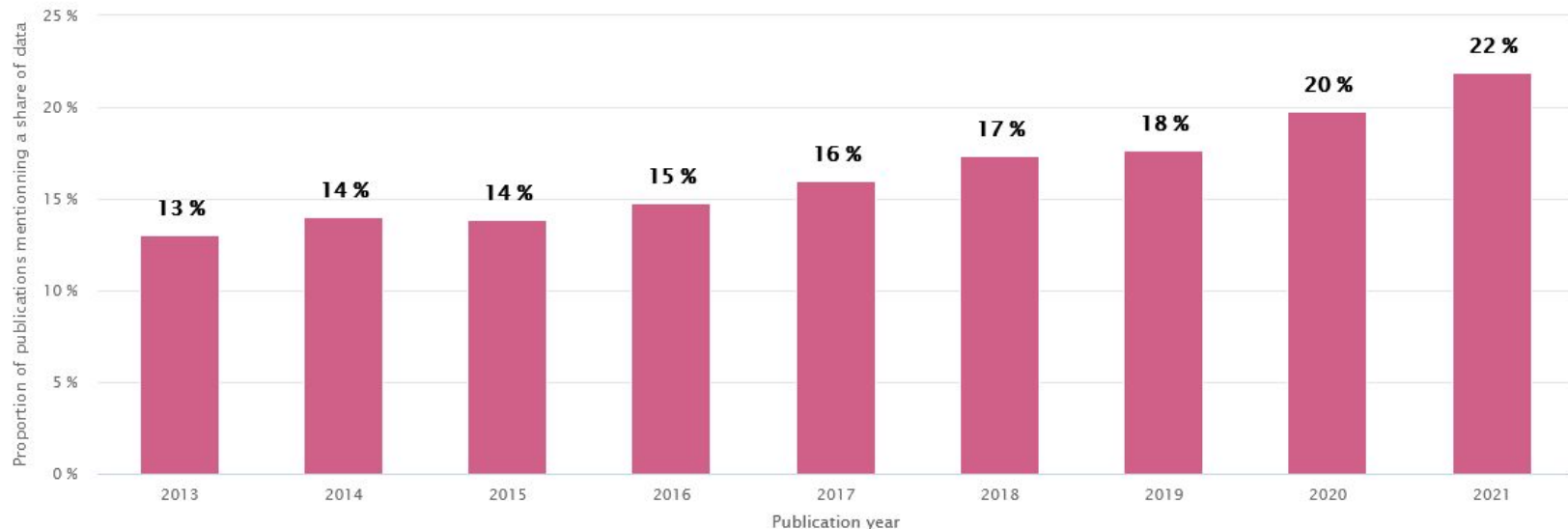
among **those mentioning the production of data**,

share of publications mentioning the sharing of data

# Publications mentioning sharing their produced data

Version [bêta]

Proportion of publications in France that mention the sharing of their data



French Open Science Monitor

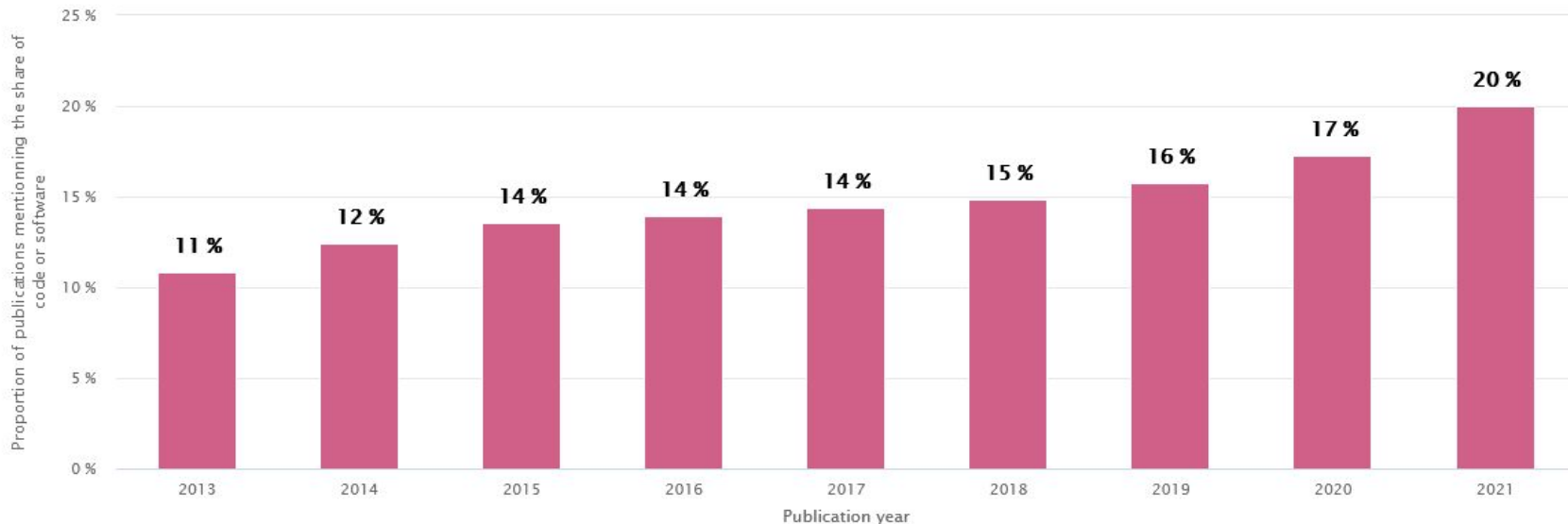
## Comment

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]

Proportion of publications in France that mention the sharing of their code or software



French Open Science Monitor

## 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

This is a demonstrator based on the processing of French publications (at least one author with a French affiliation) in the period 2013-2021, approx. 700K full text articles

Entity ▼  
software (347,876)  
persons (24,820)  
organizations (1,625)  
licenses (155)








Author ▼  
RStudio (144)  
Hadley Wickham (141)  
Scott Chamberlain (109)  
Jeroen Ooms (80)  
Kurt Hornik (65)  
Dirk Eddelbuettel (62)  
Gábor Csárdi (55)  
Achim Zeileis (52)  
Bob Rudis (47)  
Kirill Müller (46)

Languages ▼  
R (17,153)  
C (1,402)  
C++ (1,366)  
Java (796)  
Python (445)  
PHP (256)  
JavaScript (250)  
C# (156)  
Perl (156)  
assembly language (141)

+ add new facet    📄 copy as url

software ▼    must ▼    search term    ⚙️ ▼    +

374,476 results - in 4047 ms (server time)

	<b>Matlab</b> - Numerical computing environment and programming language	51493 mentions in 28768 documents
	<b>ImageJ</b> - image processing software	36435 mentions in 21149 documents
	<b>GraphPad Prism</b> - 2D graphing and statistics software	24422 mentions in 19624 documents
	<b>Statistical Package for the Social Sciences</b> - software Abstract <span>▼</span>	19968 mentions in 17206 documents
	<b>SAS</b> - statistical software	21395 mentions in 15616 documents
	<b>Excel</b> - spreadsheet editor, part of Microsoft Office	15811 mentions in 11766 documents
	<b>Stata</b> - statistical software package Abstract <span>▼</span>	11902 mentions in 9939 documents

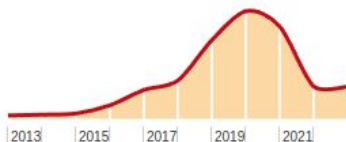
[https://cloud.science-miner.com/software\\_kb\\_bso/frontend/index.html](https://cloud.science-miner.com/software_kb_bso/frontend/index.html)

# From publication-level indicators to dataset and software Knowledge Base

**STAR** - Spike Train Analysis with R wikidata-simplified wikidata codemeta

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

**Mentions** 2146 mentions in 1056 documents (click to view mentions) 238 documents in 2021



Year	Mentions
2013	~10
2015	~50
2017	~150
2019	~350
2021	~238

**Publisher** University of Michigan HPC (Wikidata)

**Excel** - spreadsheet editor, part of Microsoft Office 15811 mentions in 11766 documents

**Stata** - statistical software package 11902 mentions in 9939 documents

Abstract ▼

**Entity**

- software (347,876)
- persons (24,820)
- organizations (1,625)
- licenses (155)

**Author**

- RStudio (144)
- Hadley Wickham (141)
- Scott Chamberlain (10)
- Jeroen Ooms (80)
- Kurt Hornik (65)
- Dirk Eddelbuettel (62)
- Gábor Csárdi (55)
- Achim Zeileis (52)
- Bob Rudis (47)
- Kirill Müller (46)

**Languages**

- R (17,153)
- C (1,402)
- C++ (1,366)
- Java (796)
- Python (445)
- PHP (256)
- JavaScript (250)
- C# (156)
- Perl (156)
- assembly language (141)


+ add new facet    📄 copy as url

[https://cloud.science-miner.com/software\\_kb\\_bso/frontend/index.html](https://cloud.science-miner.com/software_kb_bso/frontend/index.html)

# From public Knowledge



# software


- Entity
  - software (347)
  - persons (24.8)
  - organizations
  - licenses (155)
- Author
  - RStudio (144)
  - Hadley Wickh
  - Scott Chambe
  - Jeroen Ooms
  - Kurt Hornik (6)
  - Dirk Eddelbue
  - Gábor Csárdi
  - Achim Zeileis
  - Bob Rudis (47)
  - Kirill Müller (4)
- Languages
  - R (17,153)
  - C (1,402)
  - C++ (1,366)
  - Java (796)
  - Python (445)
  - PHP (256)
  - JavaScript (25)
  - C# (156)
  - Perl (156)
  - assembly lang
- + add new f


**STAR** - Spike Train Analysis with R 


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



0 – 10 of 1,056 next »


*Evaluation of STAR and Kallisto on Single Cell RNA-Seq Data Alignment*, Du et al., G3 Genes|Genomes|Genetics, DOI: 10.1534/g3.120.401160, PMID: 32220951, PMC ID: PMC7202009  View mentions in PDF 


**STAR** alignment: STAR alignment was performed for Drop-seq data using STAR version 2.5.2 a 


Data processing and analysis on 10x mouse cortex 1 single nuclei RNA-seq data **STAR**solo: STAR version 2.7.3 a with -solo command was used. 

[show other 69 mentions](#) 

*The STAR MAPS-based PiXeL detector*, Contin et al., Nuclear Instruments and Methods in Physics Research Section A: Accelerators, Spectrometers, Detectors and Associated Equipment, DOI: 10.1016/j.nima.2018.03.003  View mentions in PDF 

The **STAR** PXL detector collected more than 3 Billion minimum-bias Au+Au Following the successful experience in STAR, the next-generation MAPS sen- sors will be used for the ALICE Inner Tracking System (ITS) Upgrade [21] at LHC and the CBM Micro-Vertex Detector (MVD) [22] at FAIR, and have been proposed for the MAPS-based VerTeX detector (MVTX) of sPHENIX [23], a next-generation nuclear physics experiment for multiscale studies of the strongly coupled quark-gluon plasma planned for the year 2022 and beyond at RHIC. 

The **STAR** PXL detector collected more than 3 Billion minimum-bias Au+Au Following the successful experience in STAR, the next-generation MAPS sen- sors will be used for the ALICE Inner Tracking System (ITS) Upgrade [21] at LHC and the CBM Micro-Vertex Detector (MVD) [22] at FAIR, and have been proposed for the MAPS-based VerTeX detector (MVTX) of sPHENIX [23], a next-generation nuclear physics experiment for multiscale studies of the strongly coupled quark-gluon plasma planned for the year 2022 and beyond at RHIC. 

[show other 35 mentions](#) 

codemeta 

ents in 2021

downloaded from <https://github.com/BUSStools/bustools>. The matrix was generated following the python code available at URL: [https://github.com/BUSStools/BUS\\_notebooks\\_python/blob/master/dataset-notebooks/10x\\_hgmm\\_100\\_python/10x\\_hgmm\\_100.ipynb](https://github.com/BUSStools/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 **bustool** 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](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](https://www.kallistobus.tools/velocity_tutorial.html). The spliced and unspliced matrices were processed following the instruction from [https://github.com/BUSStools/getting\\_started/blob/master/velocity\\_tutorial.ipynb](https://github.com/BUSStools/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 **bustool** 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/MBGRLHGP\\_2019/releases](https://github.com/pachterlab/MBGRLHGP_2019/releases). The pseudo

[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) ([Torr 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** **genombar** command ([Yi et al. 2018 preprint](#)) followed by **featureCount**. **Kallisto** **genombar** 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	15
page	last page 21
ISSN	1367-4803
e ISSN	1460-2059
DOI	10.1093/bioinformatics/bts635
publisher	Oxford University Press (OUP)
url	<a href="https://doi.org/10.1093/bioinformatics/bts635">https://doi.org/10.1093/bioinformatics/bts635</a>

# 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 CORP-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)

A dark blue diagonal graphic that starts from the bottom left corner and extends towards the top right corner, covering the lower half of the slide.



# 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!**



## Open Science to Improve the Reproducibility of Science

### Evidence Based Reproducibility





- Protocols?
- What to monitor
- How to monitor
  
- Behaviour



# Protocols?



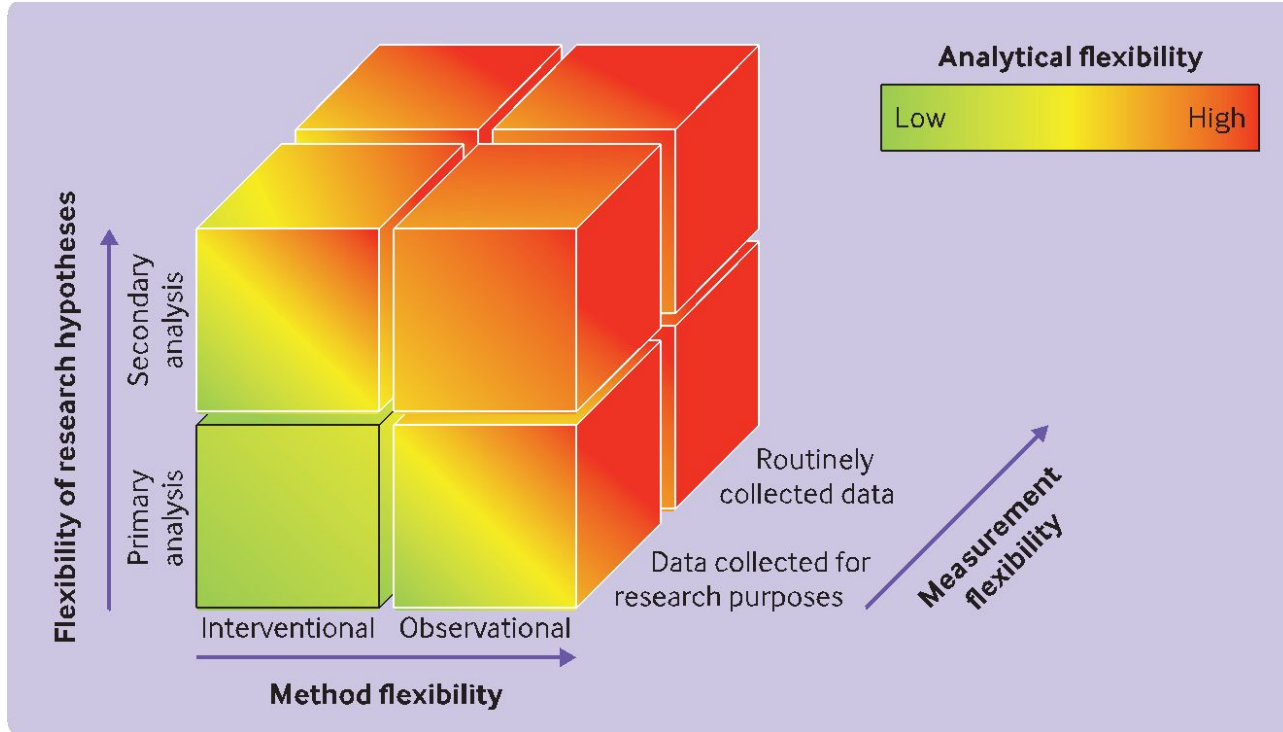
[www.osiris4r.eu](http://www.osiris4r.eu)



UMC Utrecht







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

[www.osiris4r.eu](http://www.osiris4r.eu)



# Protocols?

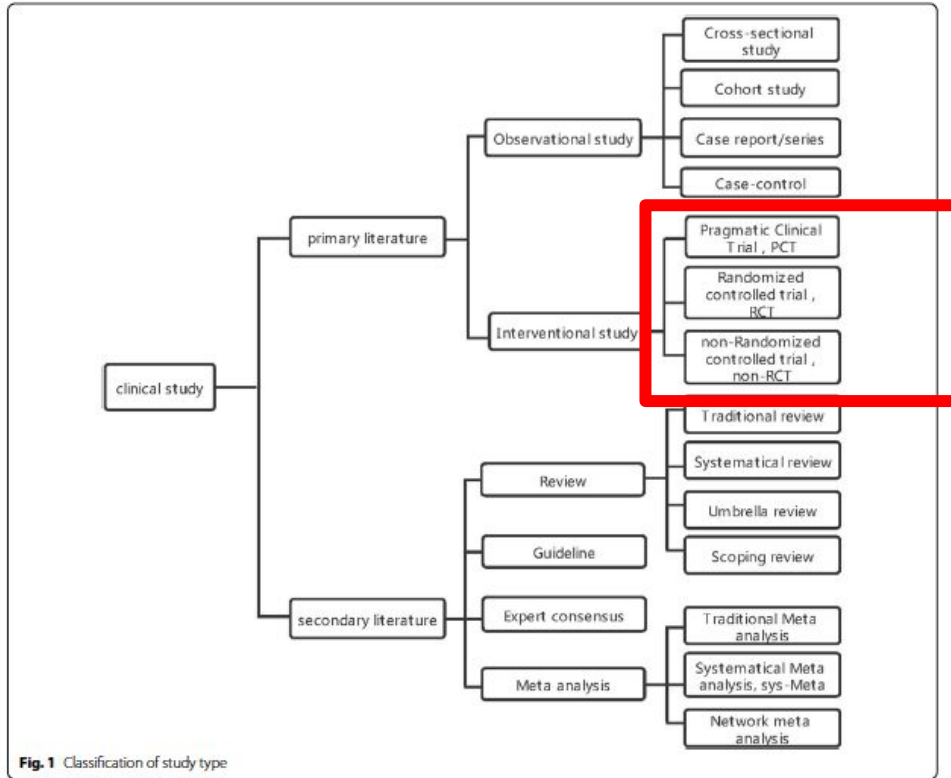
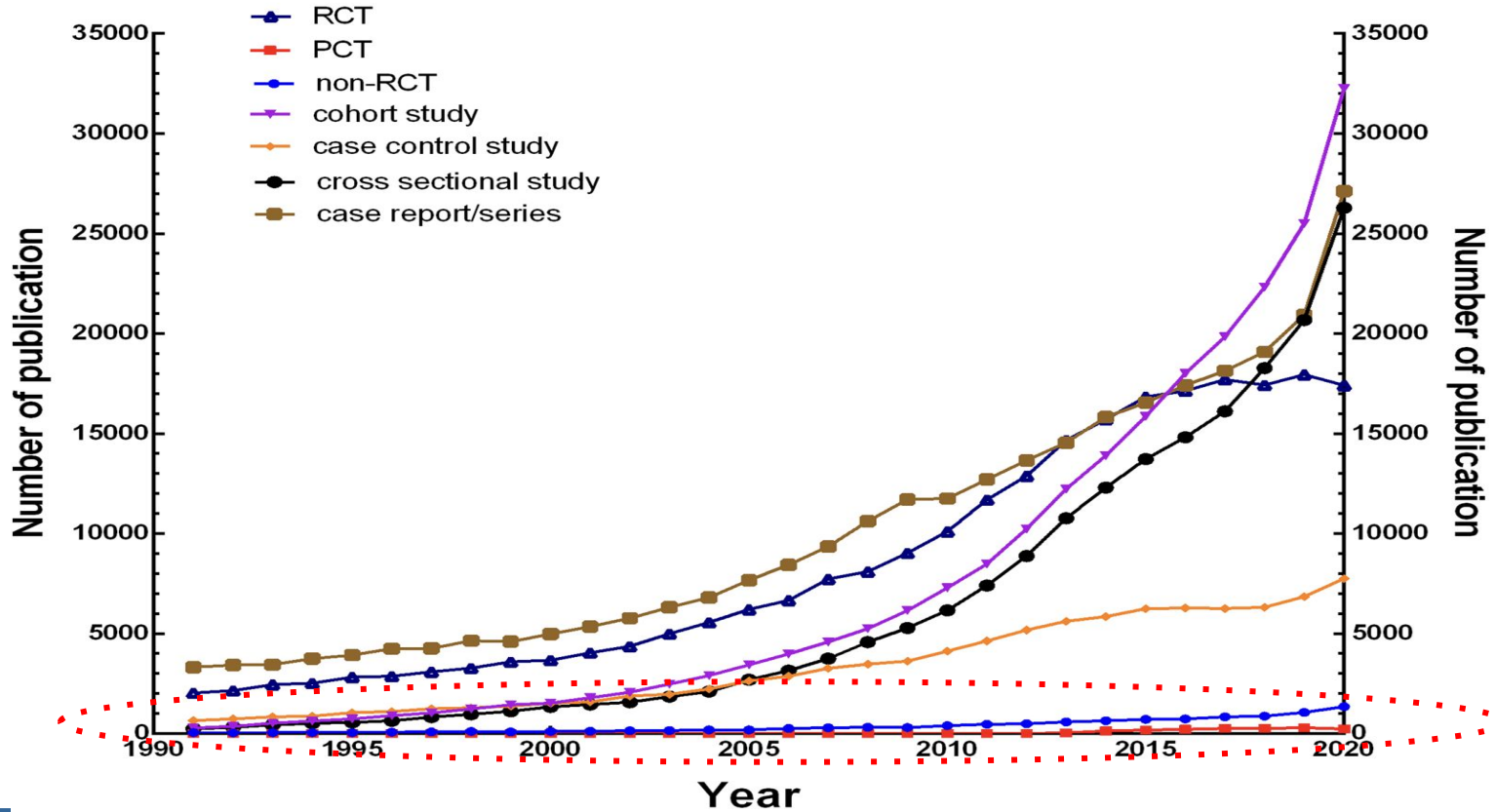


Fig. 1 Classification of study type

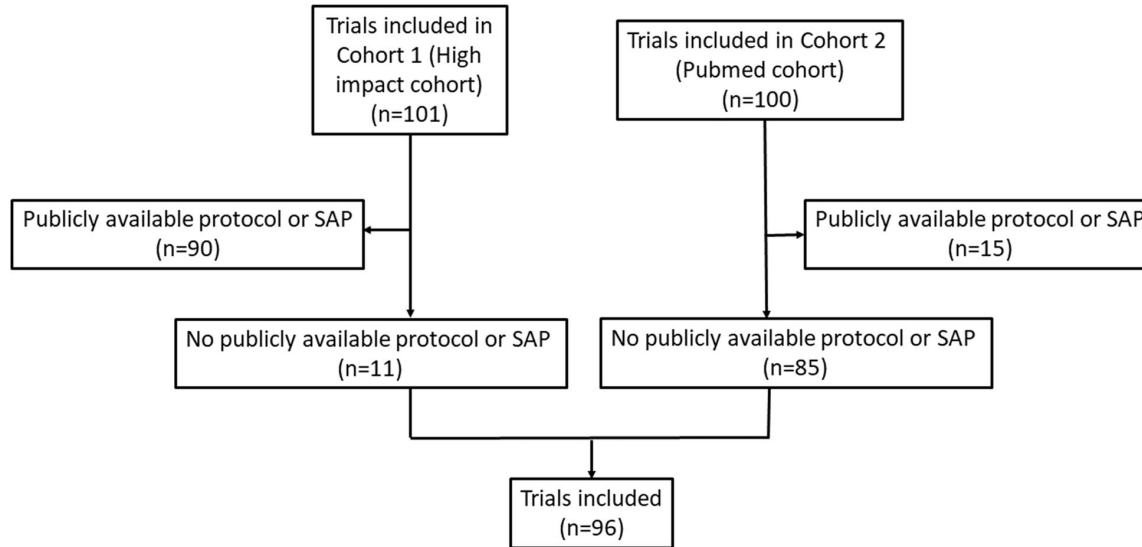
Zhao et al.  
*European Journal of Medical Research* (2022) 27:95



# Protocols?



# Protocols



Campbell et al. *Trials* (2022) 23:674



# Protocols

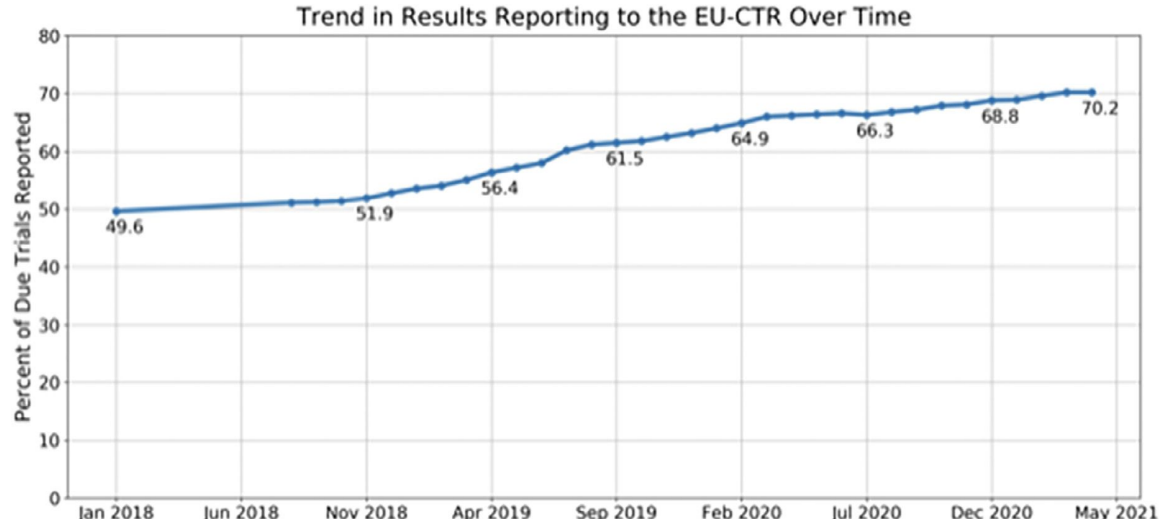
**Table 2** Study results

	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 <sup>a</sup>			
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)

<sup>a</sup> Partial excerpt from a protocol (n=1), research ethics application form (n=1)



# Protocols



**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.

Expand protocol requirement to all types of biomedical studies





# What and how to monitor



[www.osiris4r.eu](http://www.osiris4r.eu)



UMC Utrecht



CONSENSUS VIEW

### Community consensus on core open science practices to monitor in biomedicine

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

**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,



*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
<b>Traditional open science practices</b>		
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
<b>Broader transparency practices</b>		
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.

[www.osiris4r.eu](http://www.osiris4r.eu)



# 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>

[www.osiris4r.eu](http://www.osiris4r.eu)



# Plos tools to monitor Open Science

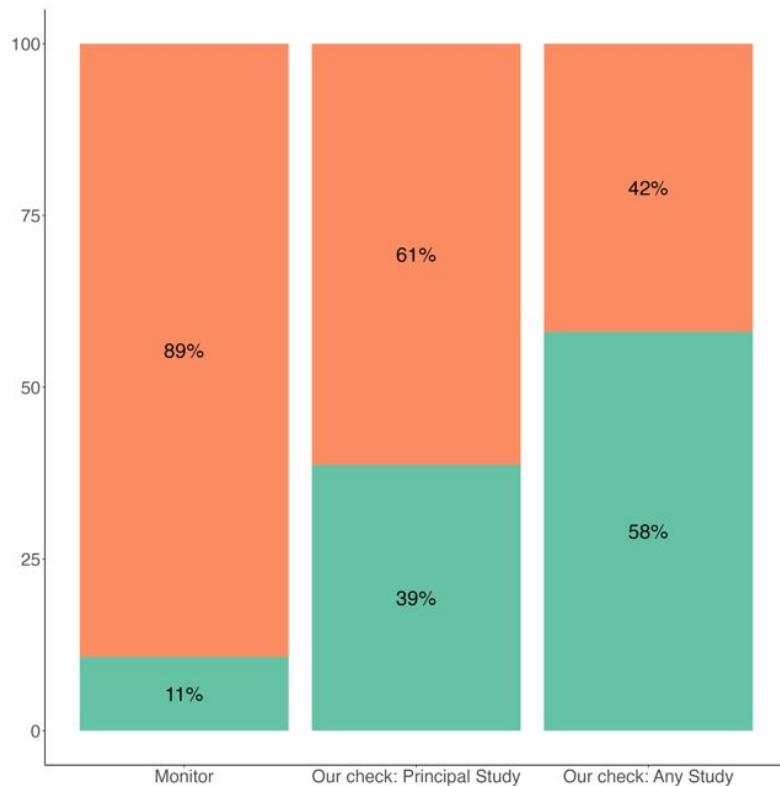
- Sharing of research data, in particular data shared in data repositories
- Sharing of code
- Posting of preprints



# 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.**

[www.osiris4r.eu](http://www.osiris4r.eu)





# 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

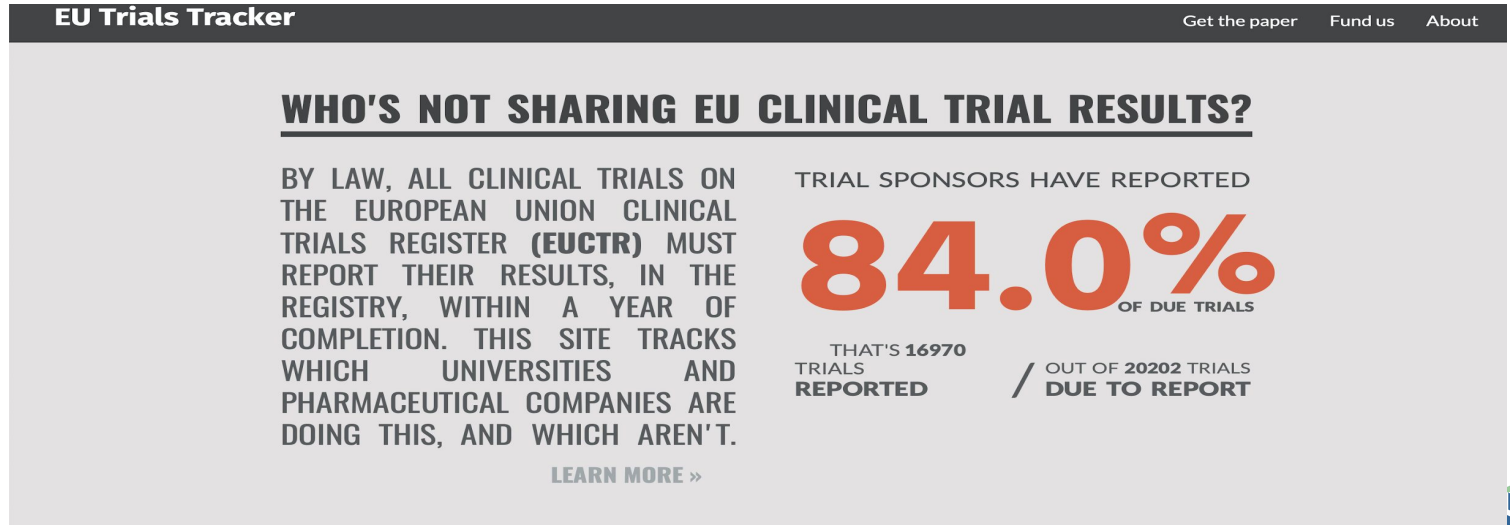
- Data citation
- Data transparency
- Analytical code transparency
- Materials transparency
- Design and analysis transparency
- Study preregistration
- Replication
- Publication bias
- Open science badges

<https://osf.io/t2yu5>



# Other examples

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



The screenshot shows the 'EU Trials Tracker' website. At the top, there is a navigation bar with links for 'Get the paper', 'Fund us', and 'About'. The main heading is 'WHO'S NOT SHARING EU CLINICAL TRIAL RESULTS?'. Below this, a text block explains that by law, all clinical trials on the European Union Clinical Trials Register (EUCTR) must report their results within a year of completion. The site tracks which universities and pharmaceutical companies are doing this and which aren't. A large orange percentage '84.0%' is displayed, indicating that 84.0% of due trials have been reported. This corresponds to 16970 reported trials out of a total of 20202 trials due to report. A 'LEARN MORE >>' link is provided at the bottom of the text block.

**EU Trials Tracker** Get the paper Fund us About

## WHO'S NOT SHARING EU CLINICAL TRIAL RESULTS?

BY LAW, ALL CLINICAL TRIALS ON THE EUROPEAN UNION CLINICAL TRIALS REGISTER (**EUCTR**) MUST REPORT THEIR RESULTS, IN THE REGISTRY, WITHIN A YEAR OF COMPLETION. THIS SITE TRACKS WHICH UNIVERSITIES AND PHARMACEUTICAL COMPANIES ARE DOING THIS, AND WHICH AREN'T.

TRIAL SPONSORS HAVE REPORTED

# 84.0%

OF DUE TRIALS

THAT'S **16970** TRIALS REPORTED / OUT OF **20202** TRIALS DUE TO REPORT

[LEARN MORE >>](#)



# Conclusion

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



# Most important!!



[www.osiris4r.eu](http://www.osiris4r.eu)

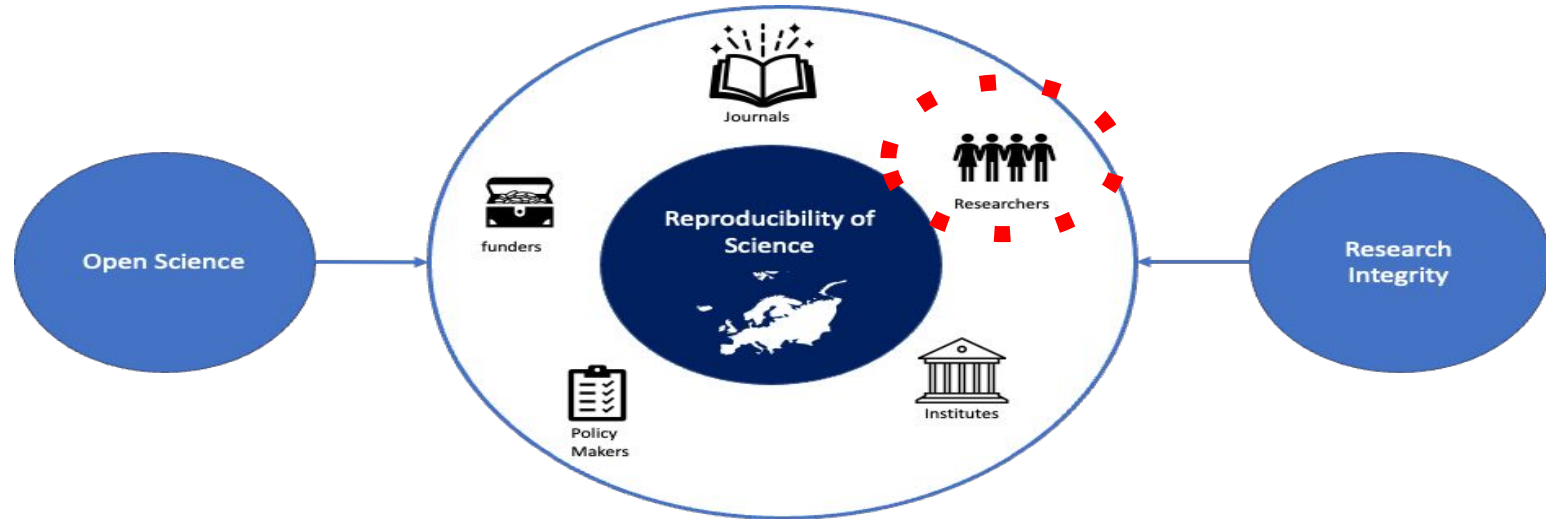


# Most important

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



# Behavioural change





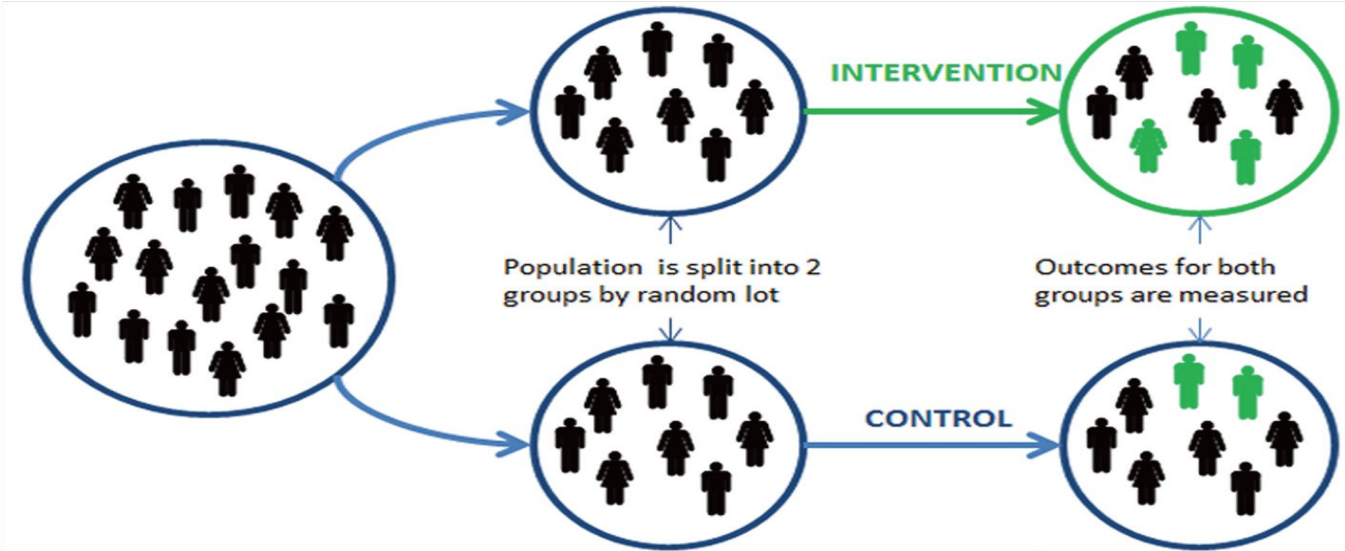
# Evidence Based

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.





### 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

A dark blue diagonal graphic that starts from the bottom left corner and extends towards the top right corner, covering the lower half of the slide.

# 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.**

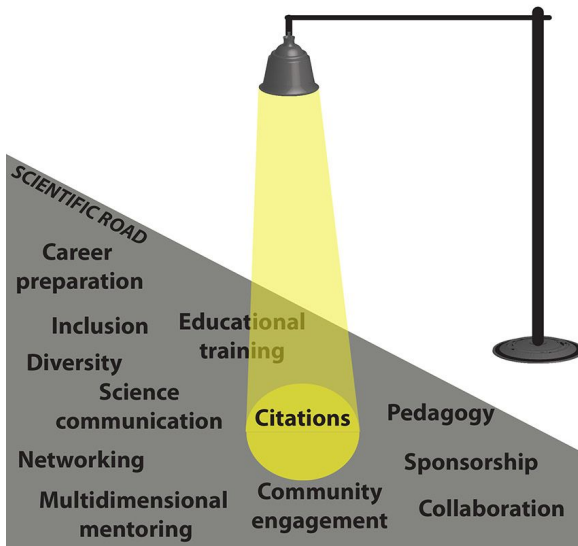
We propose three principles for monitoring OS:

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)**

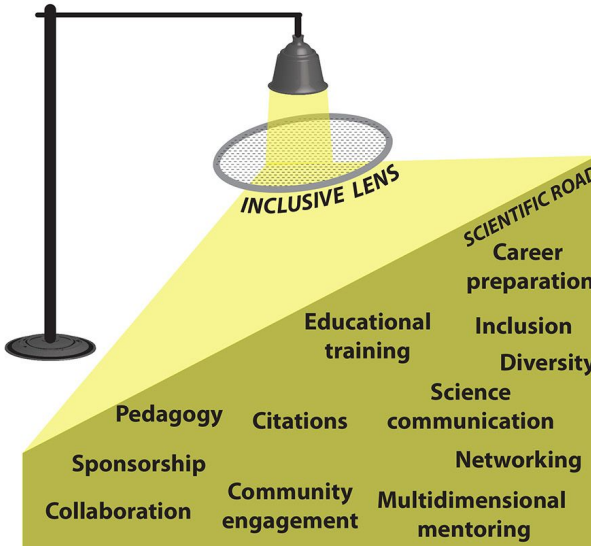
# The streetlight effect of indicators

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

A) Narrow View of Scientific Impact



B) Inclusive View of Scientific Impact



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

# 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.”



# 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

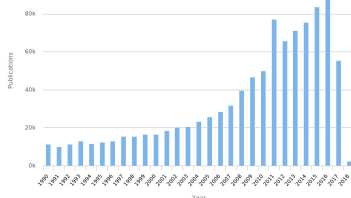
## 2. Opening up OS: showing multiple trajectories within OS

### effect of appraisal 'outputs' on decision-making

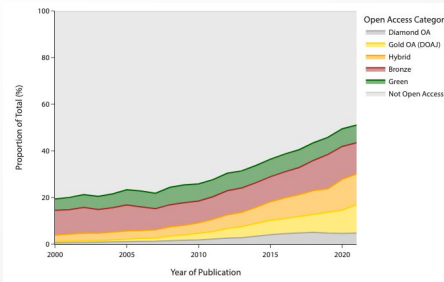
closing-down

opening-up

narrow

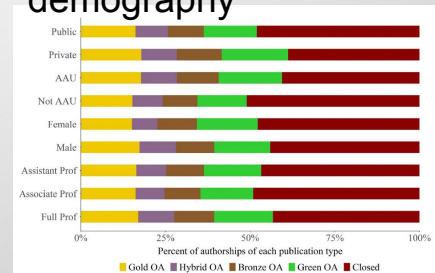


Trends in OA



Plural dimensions  
Trends in OA by  
type

Outcomes:  
OA pubs by  
demography



Opening Up versus Closing Down (Stirling, 2008)

range of  
appraisals  
inputs

(issues, perspectives,  
scenarios, methods)

broad

# 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 govts.)
- 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/>

# 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)

A dark blue diagonal graphic element that starts from the bottom left corner and extends towards the top right corner, creating a triangular shape in the bottom right of the slide.







## UNESCO Recommendation on 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	<ul style="list-style-type: none"><li>• Compilation/index of the existing open science training modules and materials</li><li>• Creation and delivery of new and additional necessary training modules on open science for different open science actors</li></ul>
OS policies and strategies	<ul style="list-style-type: none"><li>• Global Repository of Open Science Policy Instruments</li><li>• Development of Open Science Policy Guide</li></ul>
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

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

*To be kept under public oversight*



# 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:

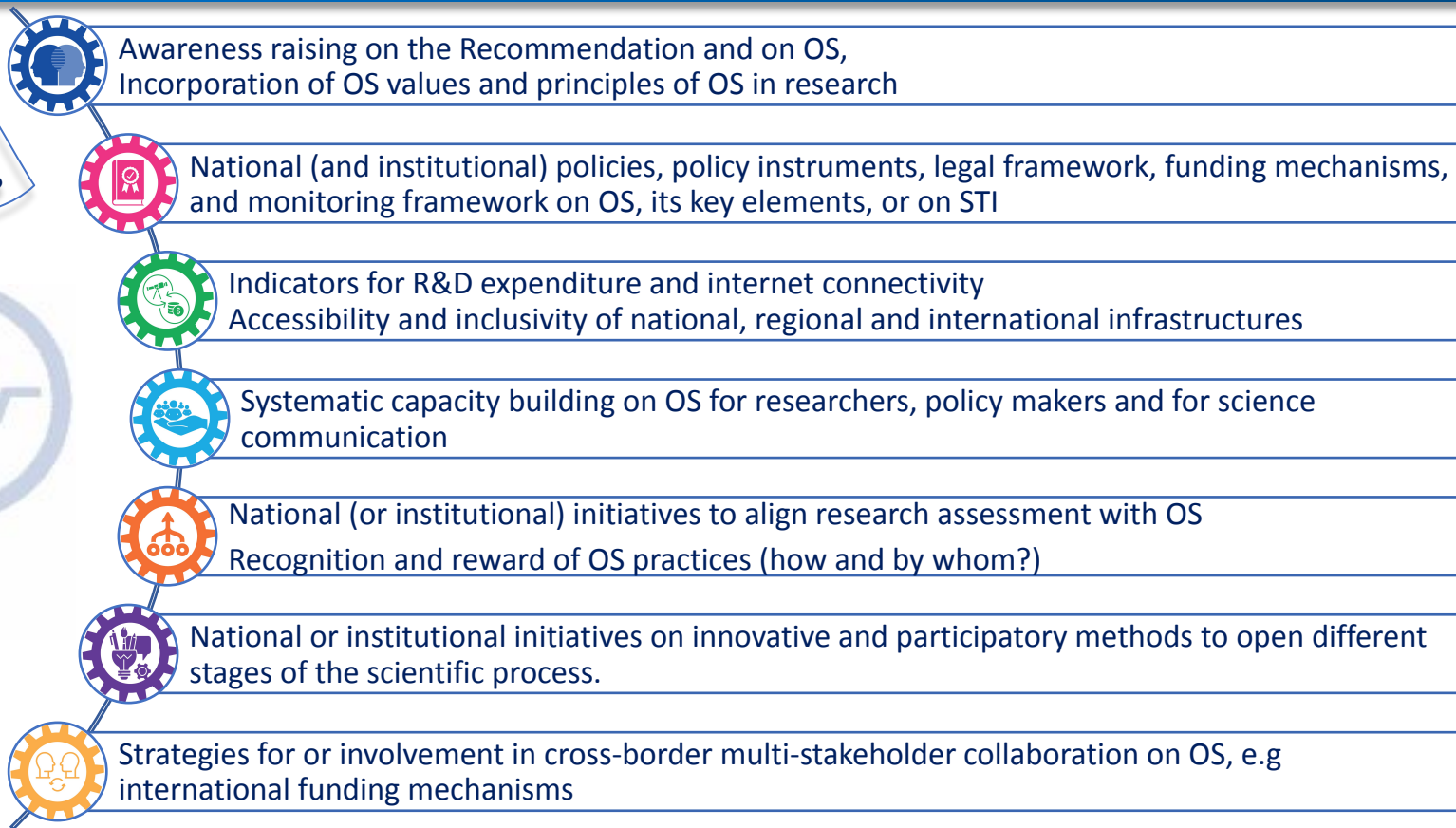
- ❑ survey for Member States to report on policies and actions promoting open science in line with the Recommendation on Open Science
- ❑ an output analysis based on inclusive ‘Global Scientific Scholarly Database(s)’
- ❑ surveys to research organizations
- ❑ opinion surveys to actors regarding the values 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  
1<sup>st</sup> round: 2024-2025





# Output analysis based on open inclusive global databases

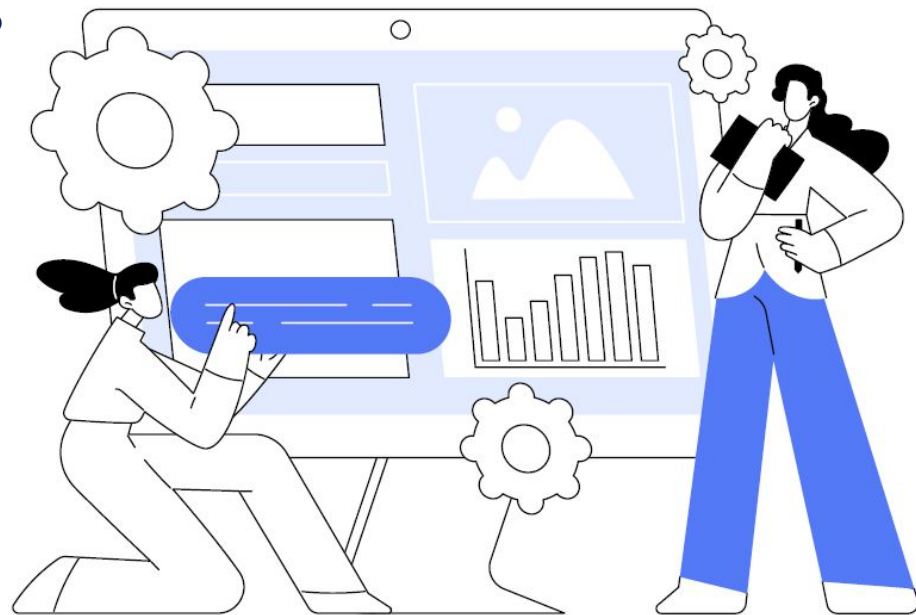
What aspects of open science **should** be measured?

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?

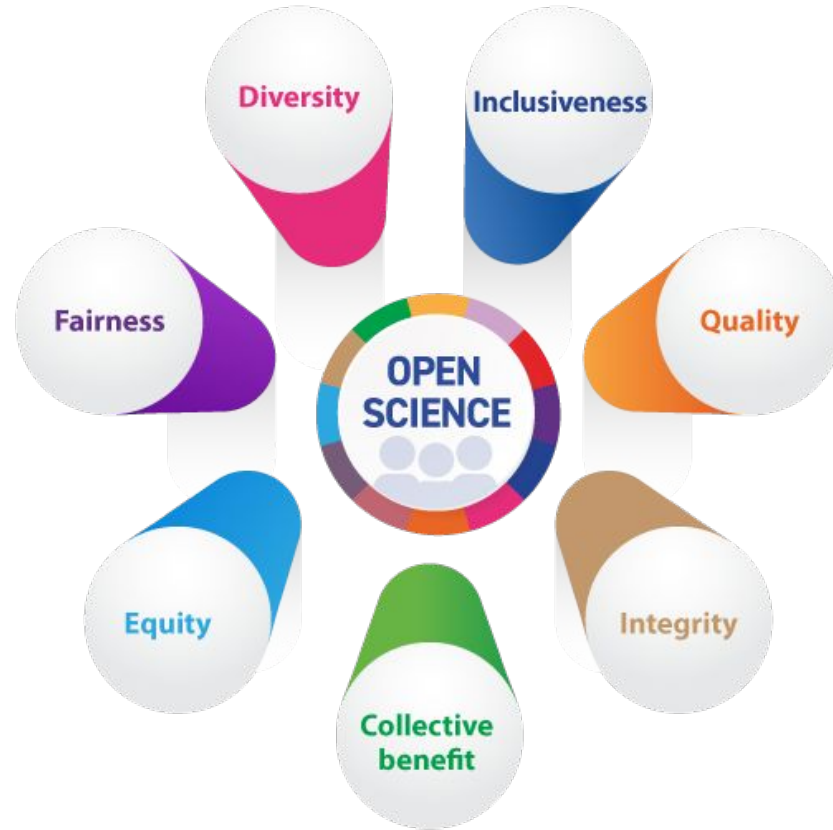


[Graphic: Shutterstock/Visual Generation](#)

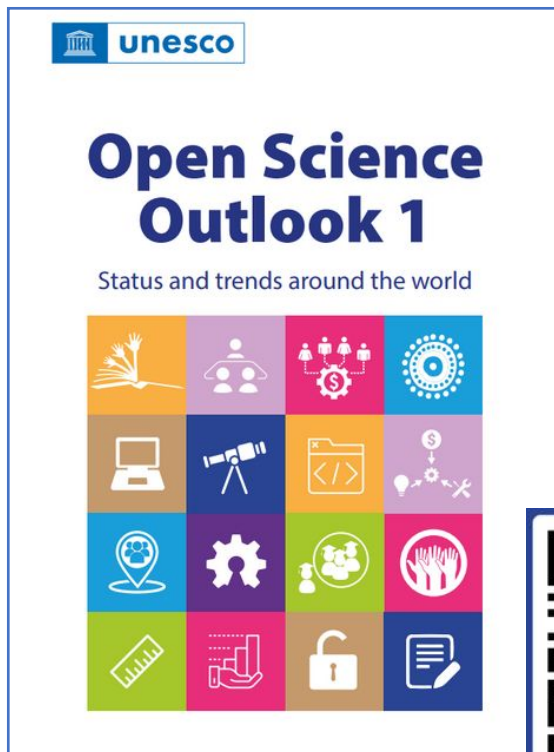


unesco

# Working Group: Open science values provide a shared framework



# Open Science Outlook



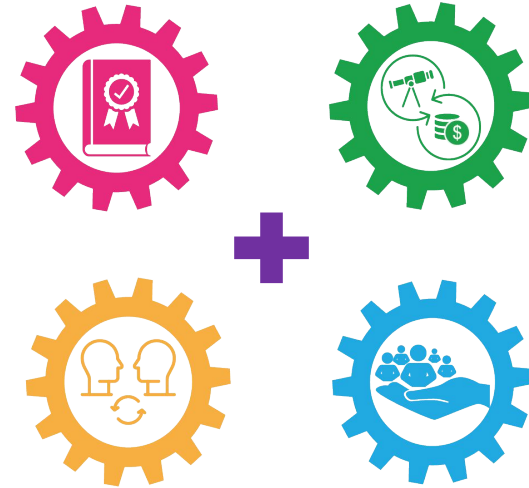
Counting is not enough  
Current system of rankings do not  
promote inclusion, equity and  
openness

- 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

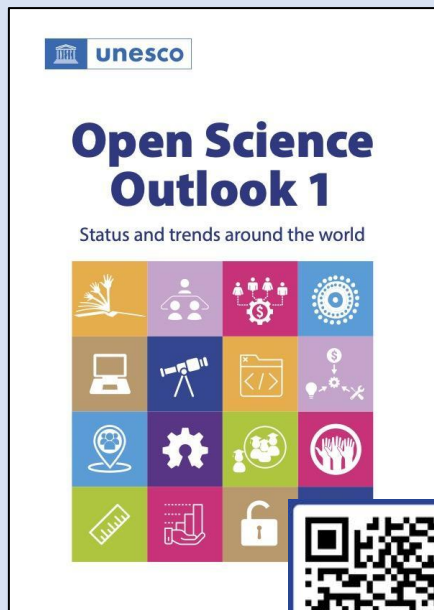
# Open Science Outlook



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](mailto:openscience@unesco.org)

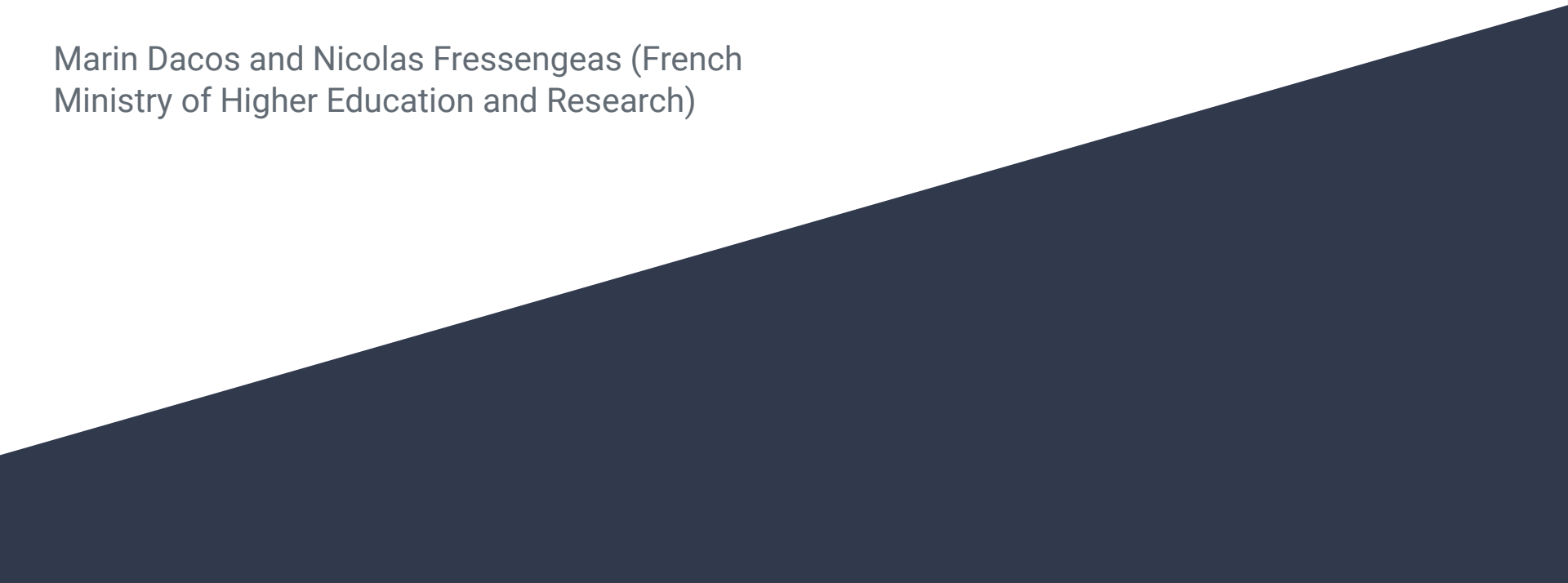
UNESCO Open science website:  
<https://www.unesco.org/open-science>



unesco

# Proposal for an Open Science Monitoring Framework

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

A dark blue diagonal graphic that starts from the bottom left corner and extends towards the top right corner, covering the lower half of the page.

# 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

POSM

## 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 ?

POSM

## 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

POSM

## 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

An Open Science monitor should be open

POSM

## 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

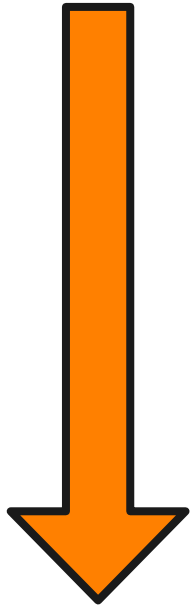
POSM

## 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 ?

