MINISTÈRE
DE L'ENSEIGNEMENT
SUPÉRIEUR
ET DE LA RECHERCHE

Liberté
Égalité
Fraternité

French Open Science Monitor

FrenchOpenScienceMonitor.esr.gouv.fr

7th June 2023

Eric Jeangirard

Agenda

- **What is the French Open Science Monitor?**
  - Objectives
  - Strengths
  - Main results
- **Methodology**
  - Open access to publications
  - Dataset sharing, software sharing
  - “Local” monitor for any French institutions
- **Lessons learnt**
  - Not using proprietary sources actually is possible, and beneficial !
  - An iterative process is needed to improve and extend the results
  - Collaborations at different scale are key
- **Perspectives**

What is the French Open Science Monitor?

Objectives of the French Open Science Monitor

Since the launch of the **National Plan for Open Science** in July 2018, the monitor has been designed as:

- a **sovereign** and **evolving** tool for **assessing the impacts** of the open science **public policy**
- a strategic tool to refine and adjust open science **public policies**
- a **lever for improving knowledge of French scientific production**, beyond the Open Science aspects
- the monitor now covers **publicatications**, **PhD thesis**, **clinical trials**, **research data**, **software and code**

Strengths of the French Open Science Monitor

The monitor has been developed with the values of sharing and openness to **promote transparency** and **facilitate reuse**.

- it is the **most comprehensive tool** to describe the French scientific production (+30 pts of coverage w.r.t proprietary sources like WoS)
- **Open source** and **open data** with open licences
- **Transparent** and **documented** methodology
- A visual data **exploration website** : **https://FrenchOpenScienceMonitor.esr.gouv.fr**
- A tool that can **easily** be **adapted by institutions and laboratories** (80+ reuse as of today)

Key assets built for the French Open Science Monitor

- **Affiliation-matcher tool**
  - From a raw affiliation string (e.g “Sorbonne University, Paris”) to a country code (FR) and RoR code (02en5vm52)
  - Combined with scraping, it enables to build the **most comprehensive source of French publications**

- **Research data and software mention detection within the publications full-text**
  - DataStet and Softcite modules built on top of GROBID use cutting-edge deep learning models to detect and characterize mentions of research data and software in the publications

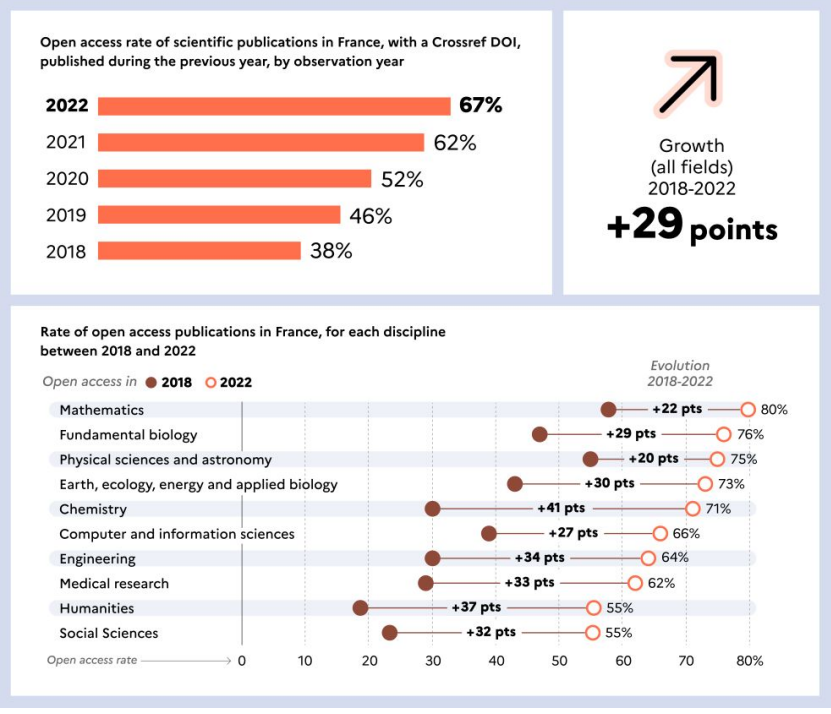
- **The community of users in the French institutions**
  - 150+ local adaptations, a community of more 200 persons
  - Regular feedback on local needs and specificities

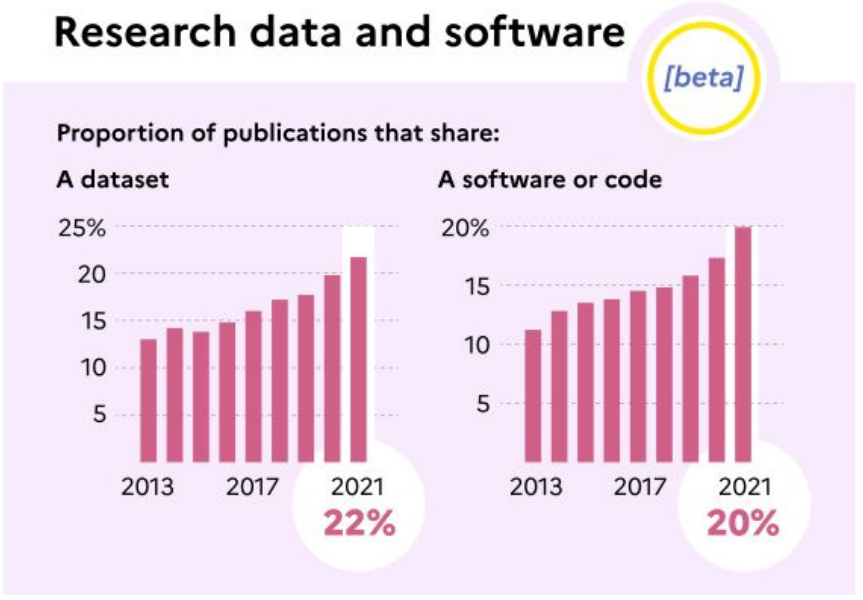
MINISTÈRE
DE L'ENSEIGNEMENT
SUPÉRIEUR
ET DE LA RECHERCHE
Liberté
Égalité
Fraternité

Main results of the French Open Science Monitor

FrenchOpenScienceMonitor.esr.gouv.fr

Publications



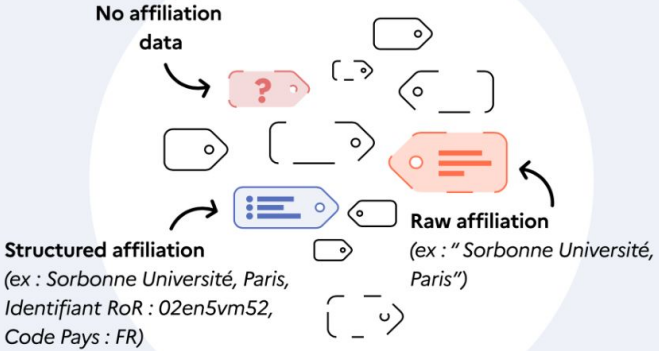


Methodology for publications

Methodology for publications openness: our observation

Open bibliographic databases

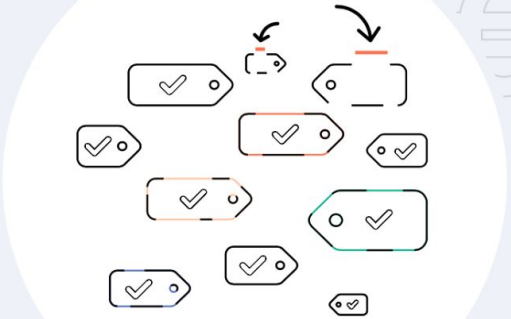
offer a low amount of affiliation metadata
and of disparate quality



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

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

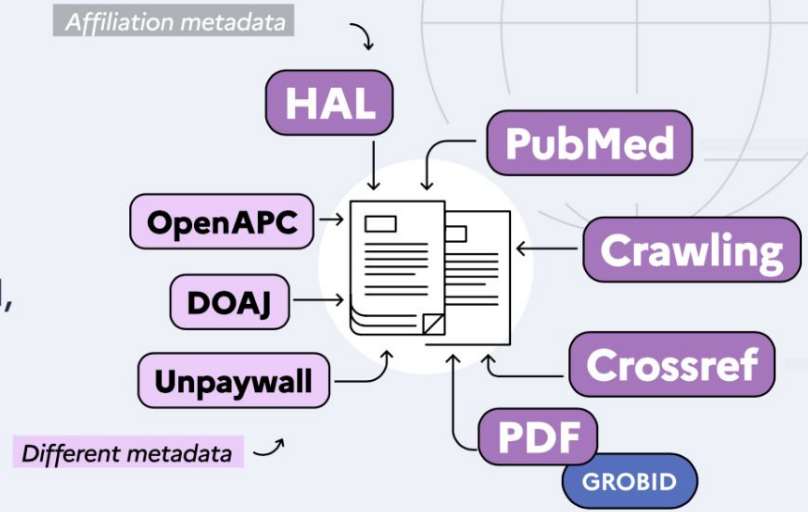
Our open methodology

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

#1 Collect

as much metadata as possible

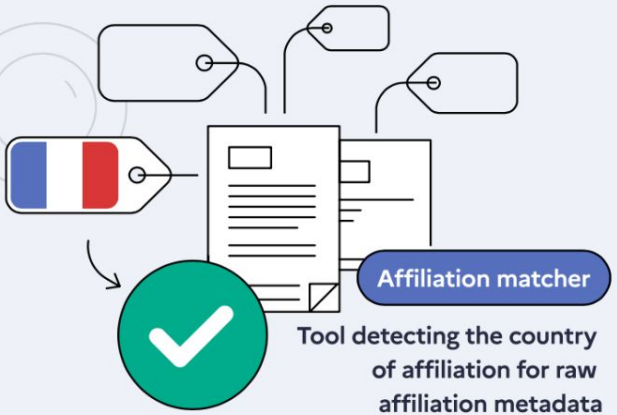
For each individual publication in the world, a variety of sources aggregated.



#2 Detect

the country of affiliation

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



"Sorbonne Université, Paris" → *France*

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

Detection rate of french scientific publications

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

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

Database of French scientific publications

170 000/year

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

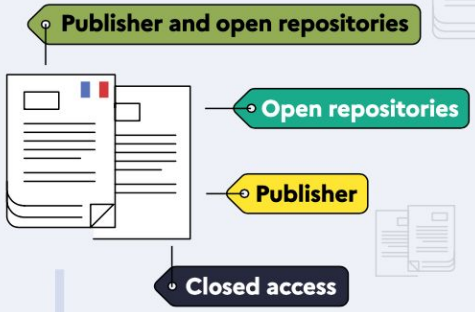
MINISTÈRE
DE L'ENSEIGNEMENT
SUPÉRIEUR
ET DE LA RECHERCHE
Liberté
Égalité
Fraternité

#3 Enhance...

... the opening status

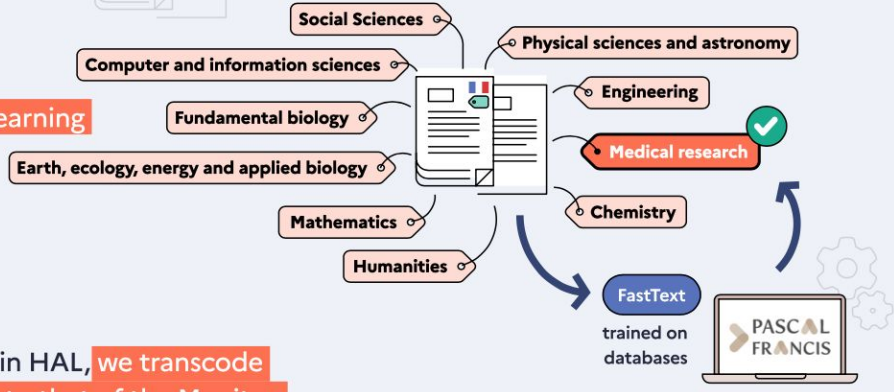
For crossref DOI:
the information stems from **Unpaywall**

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.

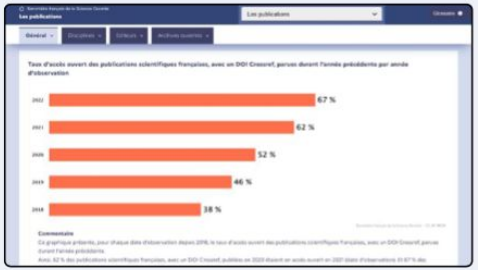
MINISTÈRE DE L'ENSEIGN SUPÉRIEUR ET DE LA REC
Liberté
Égalité
Fraternité

#4 Share

with the community all these aggregated and computed data

Datavizualisations on the Monitor's website...

frenchopensciencemonitor.esr.gouv.fr/



... and available on the open data portal of MESR

data.enseignementsup-recherche.gouv.fr

But also...

Local variations with Local monitors

barometredelascienceouverte.esr.gouv.fr/declinaisons/howto

Our tools' code are under open license

github.com/dataesr

bso-parser-html
bso-publications
bso-ui
scientific-tagger
harvest-pubmed
affiliation-matcher
harvest-hal

Methodology for dataset and software

Methodology for dataset and software sharing

- Main idea is to leverage on the publications full-text, which is the most generic approach (does not depend on disciplinary, repositories, etc)

- **Step 1**: from the list of publications gathered with the previous methodology, download as much full-texts as possible
  - 85% success for OA
  - 38% success for closed OA (using Elsevier and Wiley paying subscription)
  - 64% overall (900k PDF for 1.4m publications, from 2013 to 2021)
- **Step 2**: Apply 3 deep learning modules
  - GROBID ⇒ structures full-text to TEI-XML format, identifies Data Sharing statements
  - DataStet ⇒ identifies and classifies dataset mentions [used, created, shared]
  - Softcite ⇒ identifies and classifies code and software mentions [used, created, shared]
- **Step 3**: Produce aggregate KPI, with a funnel analysis

Methodology for dataset and software sharing [FR]

Methodology for dataset and software sharing

- Methodology is costly in terms of budget and time
  - Access to PDF can be difficult
  - NLP techniques are compute-intensive

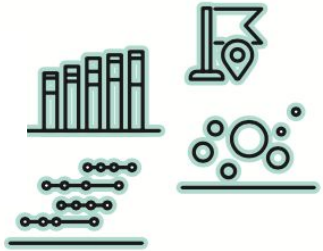
From national to local monitoring

At the service of institutions

Une méthodologie simplifiée

Send a publications list via the dedicated web page

Graphics generated in 1 click



More than local variations **85**

instituts · universités · écoles · organismes de recherche · unités de recherche

A user club with more than **200** membres

- Regular webinars
- Designed for sharing tips and tricks between institutions
- Designed to find help

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

SCIENCE OUVERTE
à l'Université de Lorraine

À L'UL | PUBLICATIONS OUVERTES | DONNÉES DE LA RECHERCHE | BIBLIOMÉTRIE | AGENDA | CONTACT

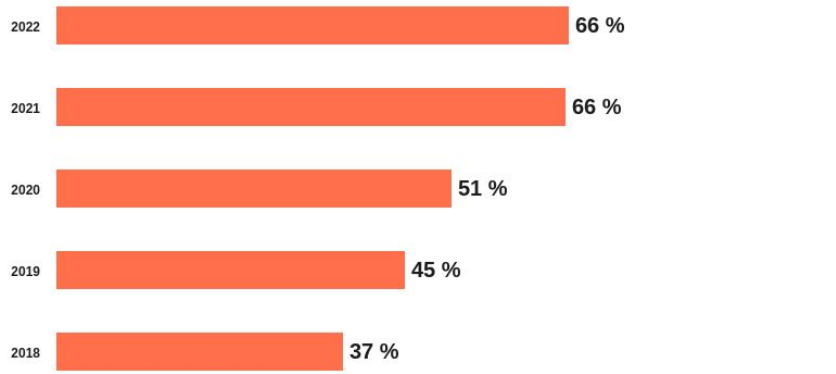
UNIVERSITÉ DE LORRAINE

BAROMÈTRE LORRAIN DE LA SCIENCE OUVERTE

Home » Bibliométrie » Baromètre lorrain de la Science Ouverte

LA PROGRESSION DE LA SCIENCE OUVERTE À L'UNIVERSITÉ DE LORRAINE

Université de Lorraine (UL) : Taux d'accès ouvert des publications scientifiques de l'université de Lorraine, avec un DOI Crossref, parues durant l'année précédente par année d'observation



Baromètre français de la Science Ouverte - CC-BY MESR

> VOIR TOUS LES INDICATEURS

DÉPÔT DANS HAL
hal-contact@univ-lorraine.fr

GESTION DE VOS DONNEES DE RECHERCHE
donnees-recherche@univ-lorraine.fr

PUBLIER EN OPEN ACCESS
copo-contact@univ-lorraine.fr

BIBLIOMETRIE
bibliometrie-contact@univ-lorraine.fr

EDITER UNE REVUE
ddoc-edition-contact@univ-lorraine.fr

Boîte à outils

Nos Prochains Événements

1 juin | 14 h 00 min

MON IDENTITÉ DANS HAL

5 juin | 14 h 00 min - 15 h 00 min

ATELIER PLAN DE GESTION DE DONNÉES

Lessons Learnt (for now)

No use of proprietary sources brings benefits!

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

- That requires more **work on the data quality**, that has to be checked with local knowledge of the Higher Education and Research ecosystem

- That allows to **openly share the results** (indicators, code, data) but also services can be built on top of the resulting data and API

An iterative process is needed to improve and extend the results

- The French monitor has been built step by step, layer after layer : publications, then clinical trials, and now adding research data and code. And we even plan more layers.
- **The machine learning** models constructions are themselves iteratives processes as **they improve with more feedbacks**
- Monitoring hidden things is complicated, and the **development of PID policy or implementation** enables a more accurate monitoring

Collaborations at different scales are key

- Many initiatives exist
  - necessity to be **complementary** and not to reinvent the wheel
  - facilitate re-use (open services, code, data, doc) in a **logic of community creation**
  - The French monitor is now supported by 3 French institutions : the **French HER ministry**, **Lorraine University** and **INRIA**.

- **From national to local**
  - the French national monitor provides an open service to the French institutions so they can benefit from the same tool within their perimeter
  - each institution contributes with their local data (list of publications) and feedback to improve the national monitoring

- **From national to international**
  - Open initiatives exist, like OpenAlex or COKI. There is room to coordinate so that (open) data quality improves globally and cutting-edge detection methods are shared

Perspectives

Perspectives

- **Keep on developing our key assets**
  - improving **AI models for research data and software detections** (DataStet and Softcite on top of GROBID)
  - affiliation-matcher tool
  - engage with our local users **community**

- **Improving and extending** the French monitor scope
  - More details on clinical trials at the institution level and work together with the academic lead sponsors to **increase the sharing of clinical trials results**
  - Start following the **ORCID use in France**

- **Coordinate to scale up**
  - Coordinate and align at the international level on cutting-edge tools for research data and software sharing monitoring
  - Leverage on open and global datasources like OpenAlex, CORE, COKI