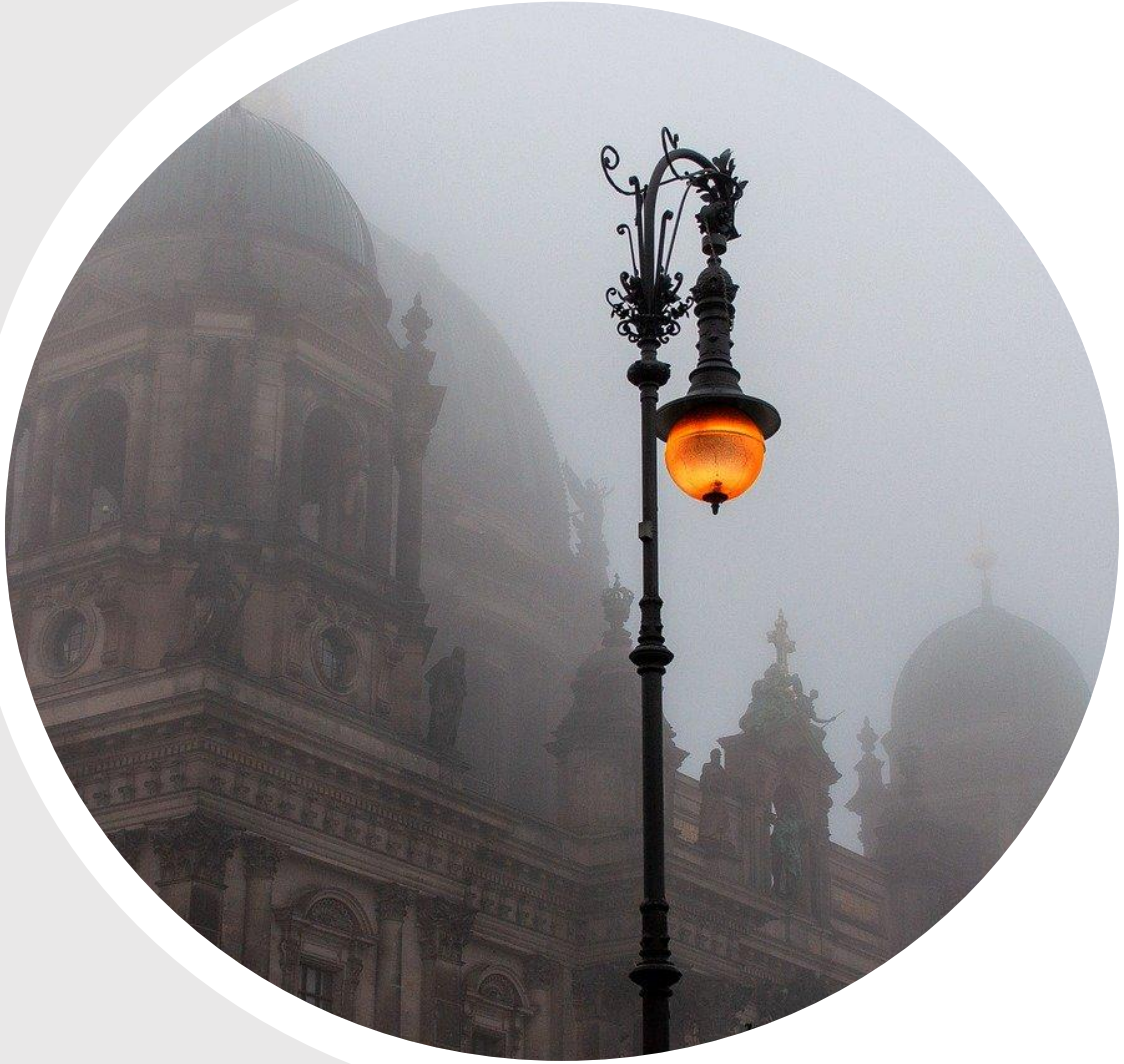


MINISTÈRE
DE L'ENSEIGNEMENT
SUPÉRIEUR
ET DE LA RECHERCHE
Liberté
Égalité
Fraternité

UNIVERSITÉ
DE LORRAINE

Inria

The French Open Science Monitor

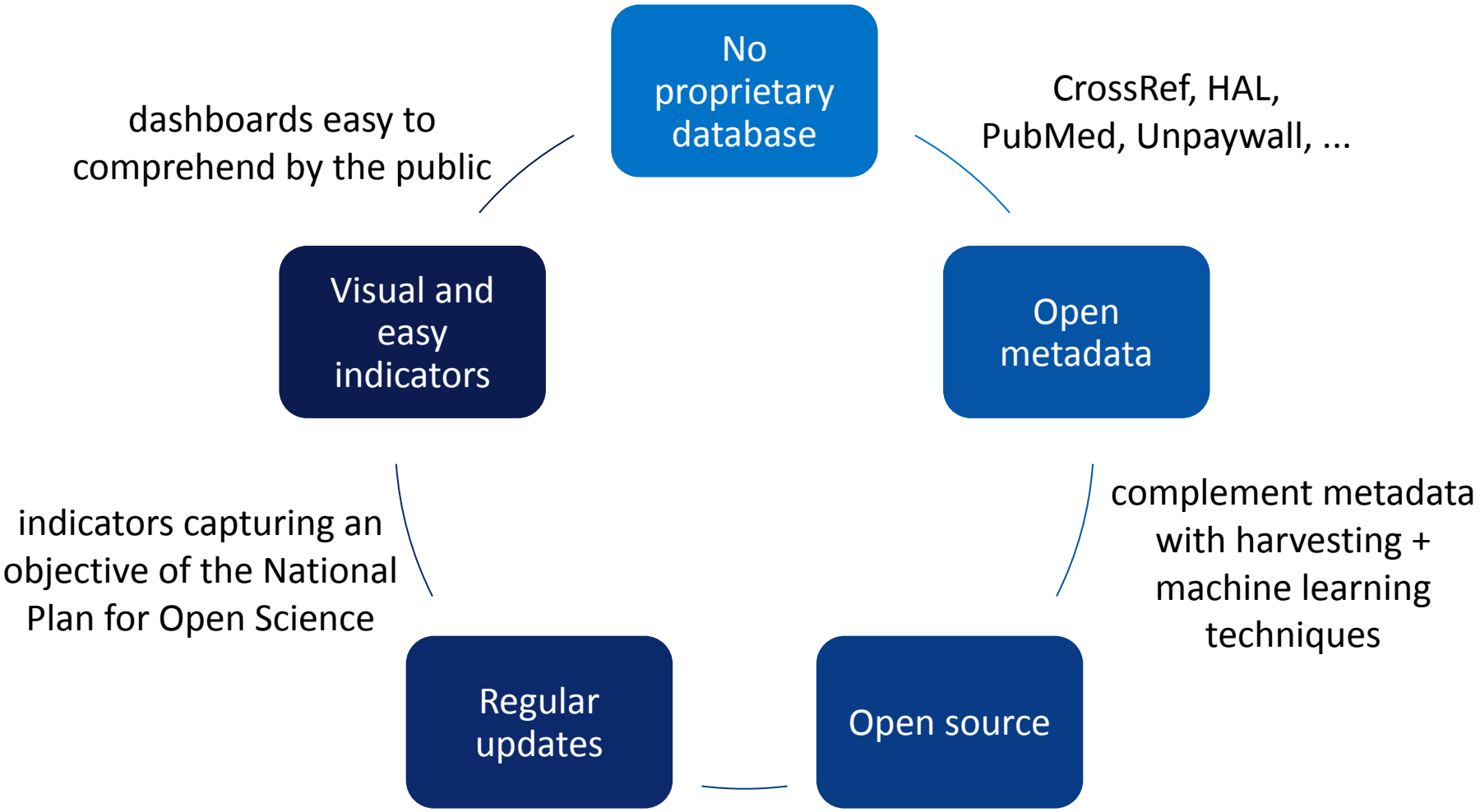
Monitoring Open Science Beyond publications

Patrice Lopez, science-miner

FOCUS ON THE FRENCH OPEN SCIENCE MONITOR

2018

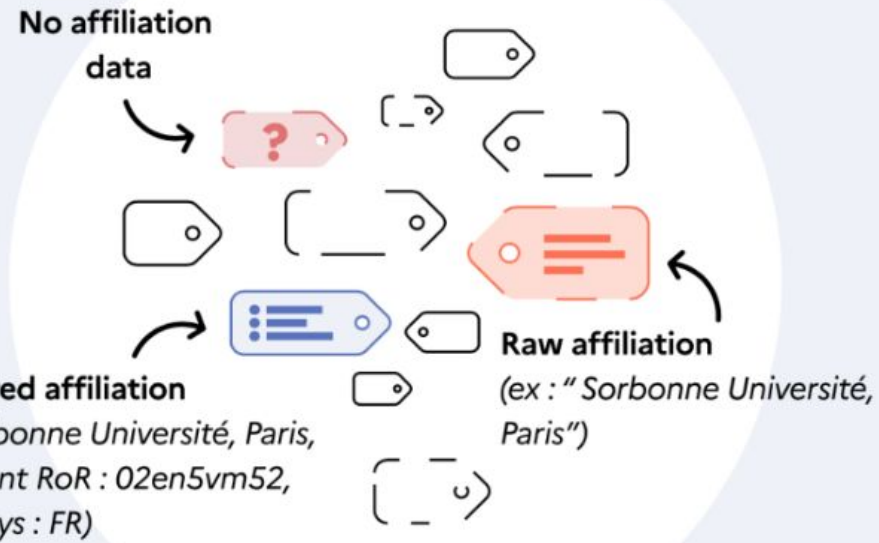
French National Plan for Open Science



HTTPS://FRENCHOPENSCIENCEMONITOR.ESR.GOUV.FR

Open bibliographic databases

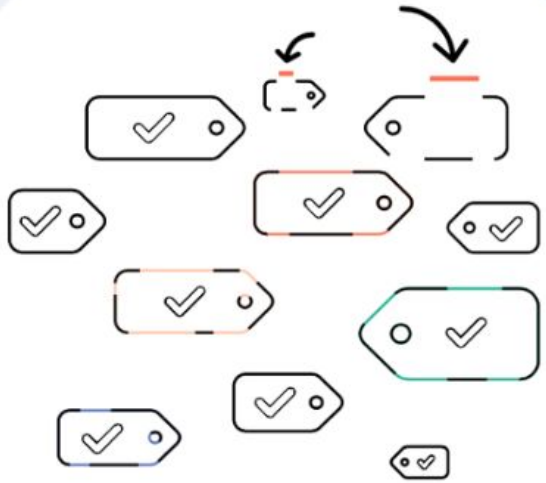
offer a low amount of affiliation metadata
and of disparate quality



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

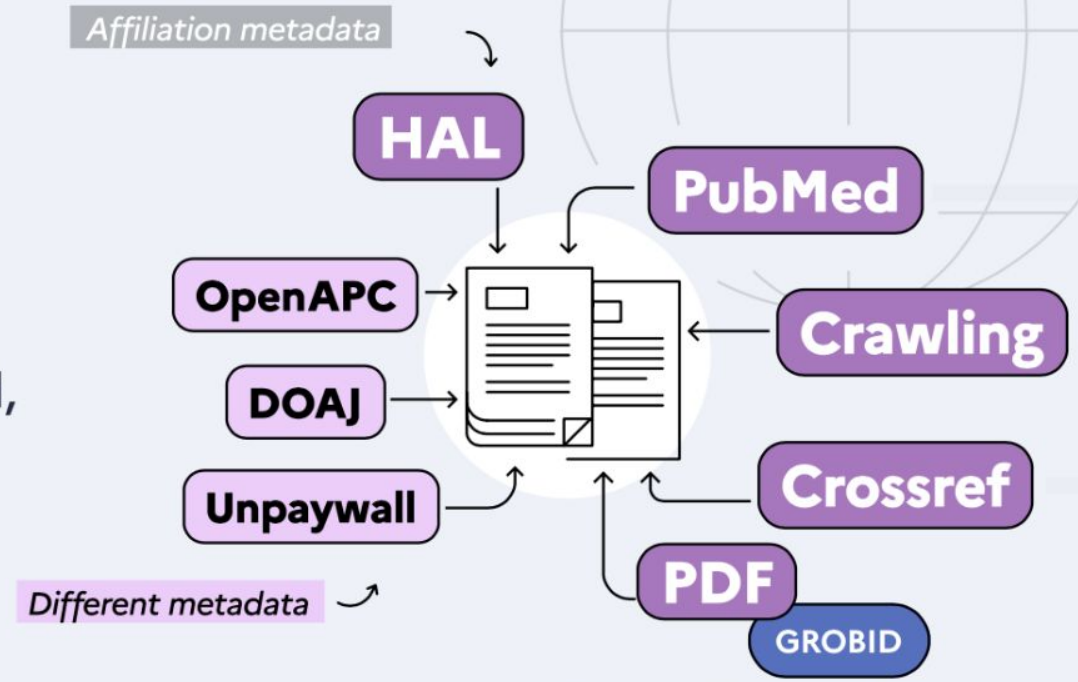
Our open methodology

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

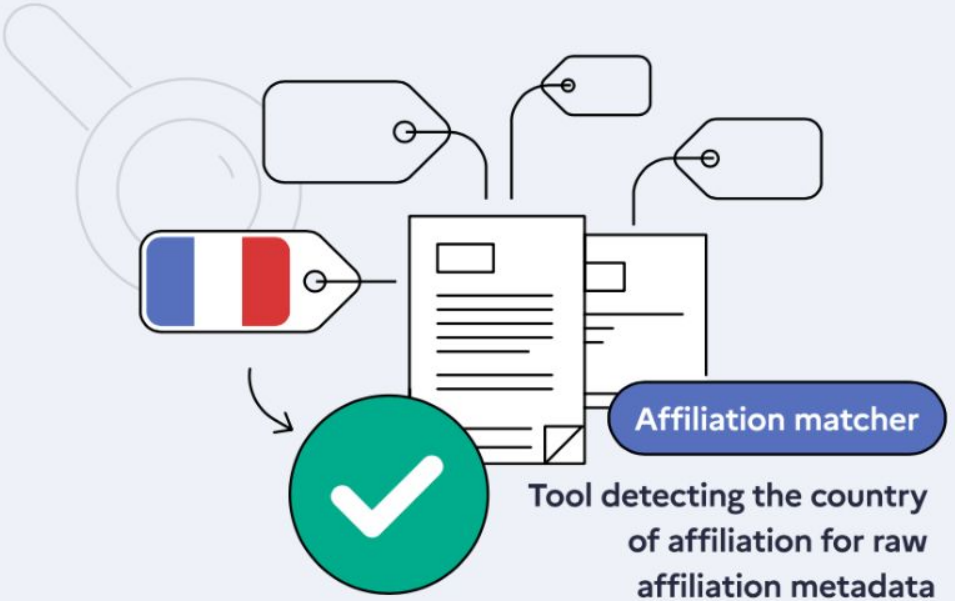
For each individual publication in the world, a variety of sources aggregated.



#2 Detect

the country of affiliation

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



Tool detecting the country of affiliation for raw affiliation metadata

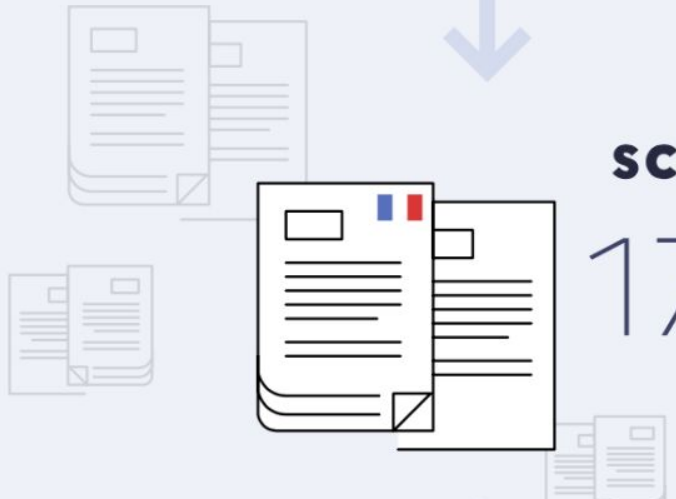
"Sorbonne Université, Paris" → *France*

"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

* Comparing sources and French Open Science Monitor corpus

➡ The approach used by the French Open Science Monitor effectively identifies the vast majority of publications with a DOI for Open Science monitoring.

	Scopus	**WoS**	**HAL**	**ADS**	**PubMed**	**MAG**	**BSO**
Share of total (%)	67	58	38	9	29	61	**92**

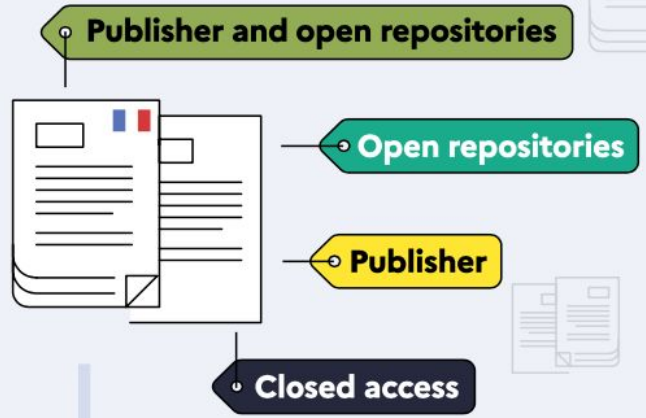
*Share of the different sources in the overall French publication aggregated corpus (total of 167,412 publications) for the year 2019, as reported by **[2]***

#3 Enhance...

... the opening status

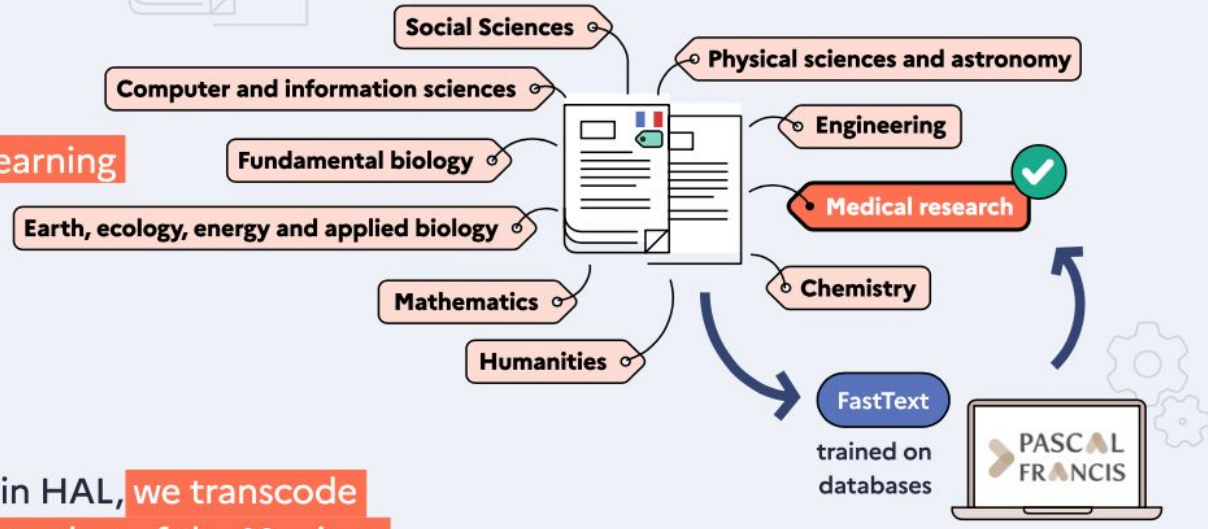
For crossref DOI:
the information stems from **Unpaywall**

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.



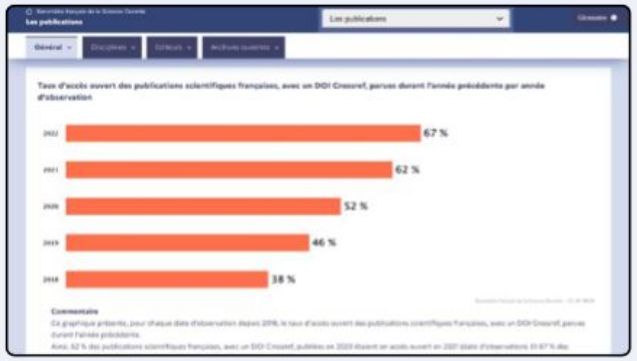
If metadata is available in HAL, we transcode the HAL classification into that of the Monitor.

#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

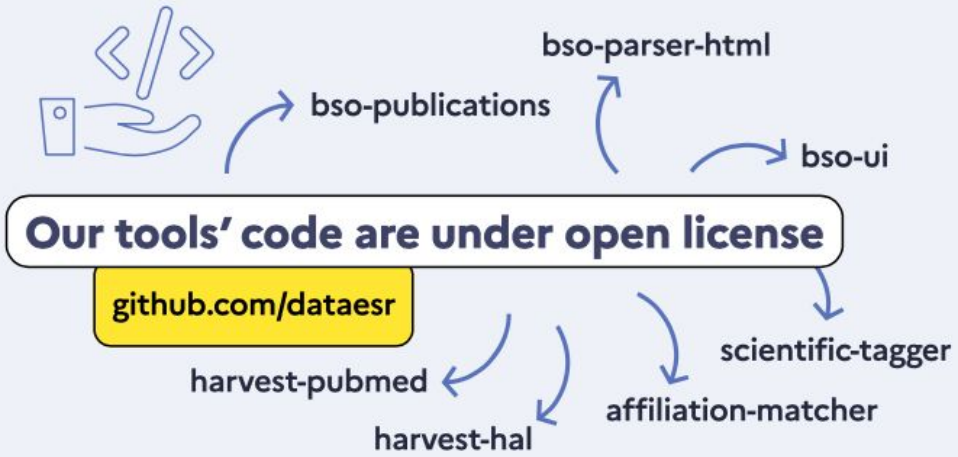
But also...

Local variations with Local monitors

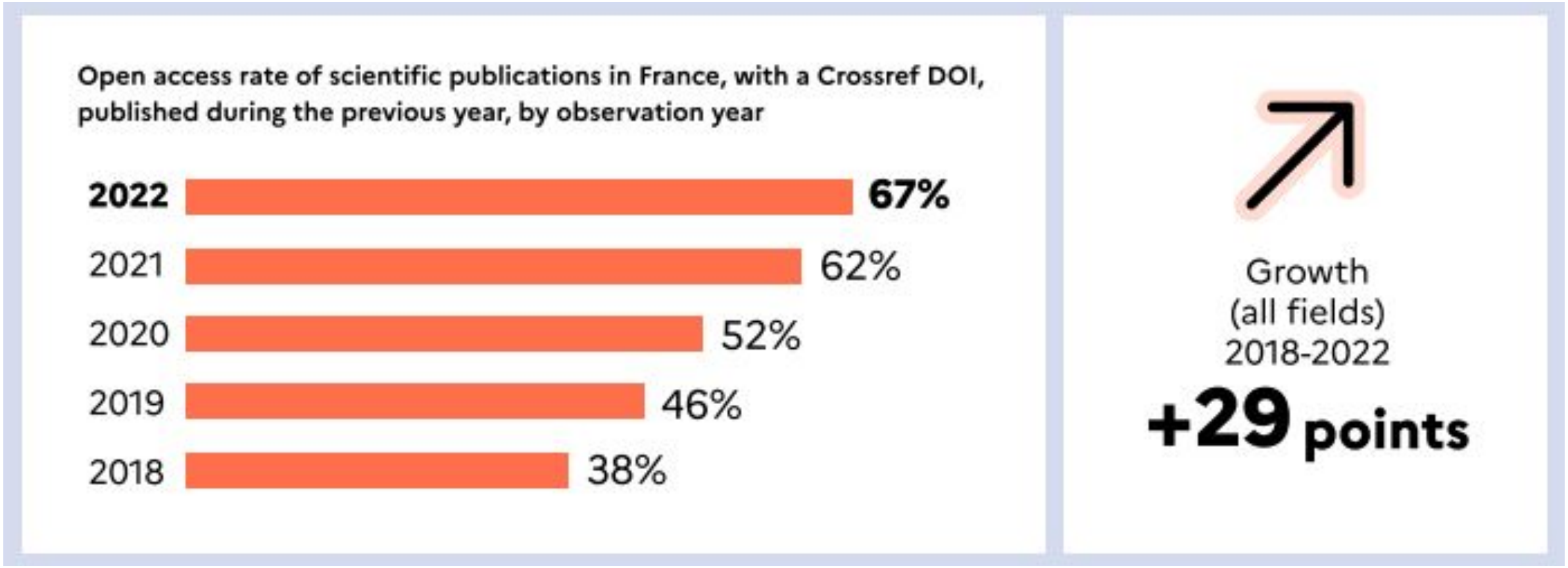
barometredelascienceouverte.esr.gouv.fr/declinaisons/howto

Our tools' code are under open license

github.com/dataesr



OPEN ACCESS RATE OF FRENCH PUBLICATIONS

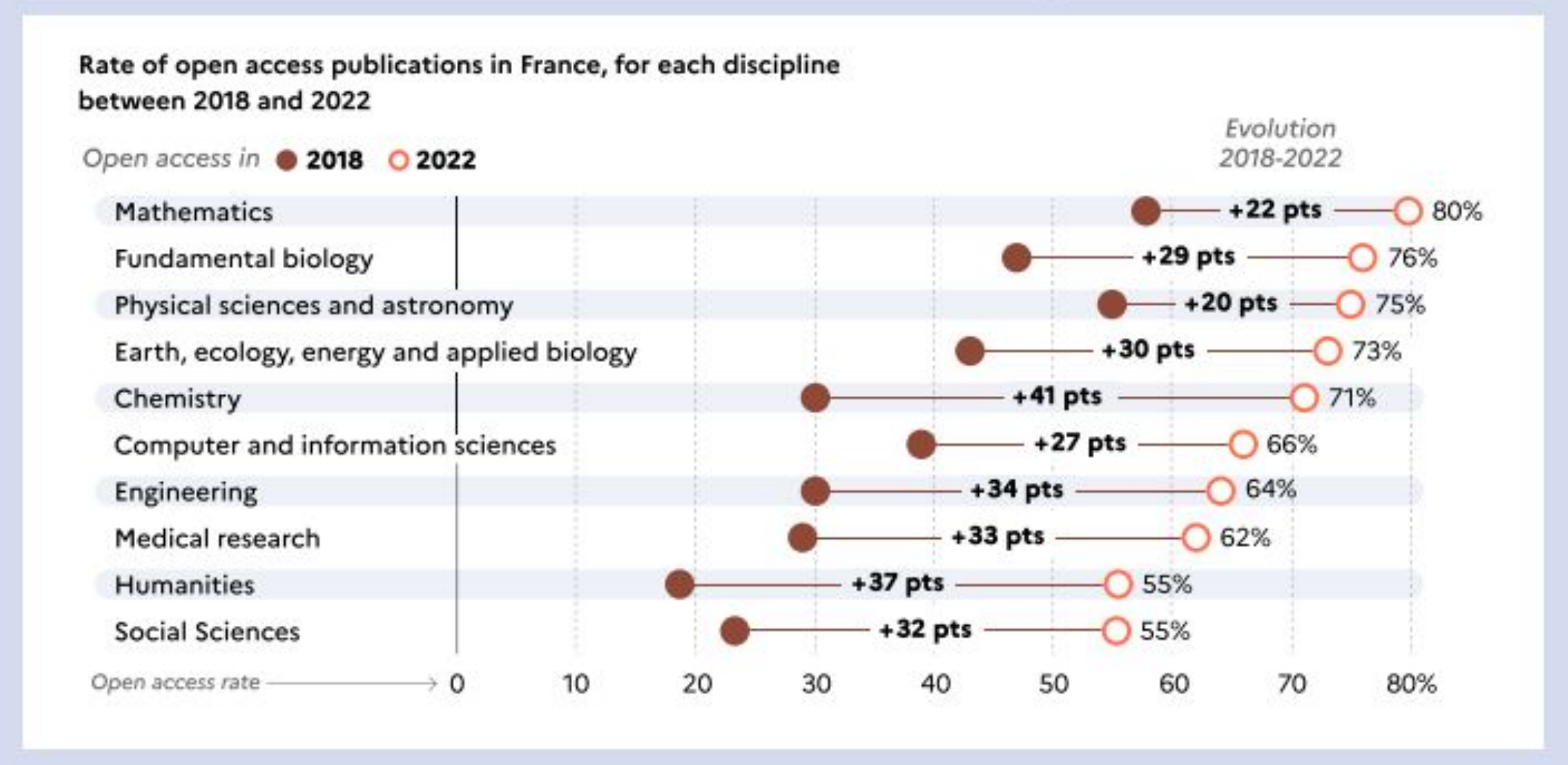


2021: 160,217 publications

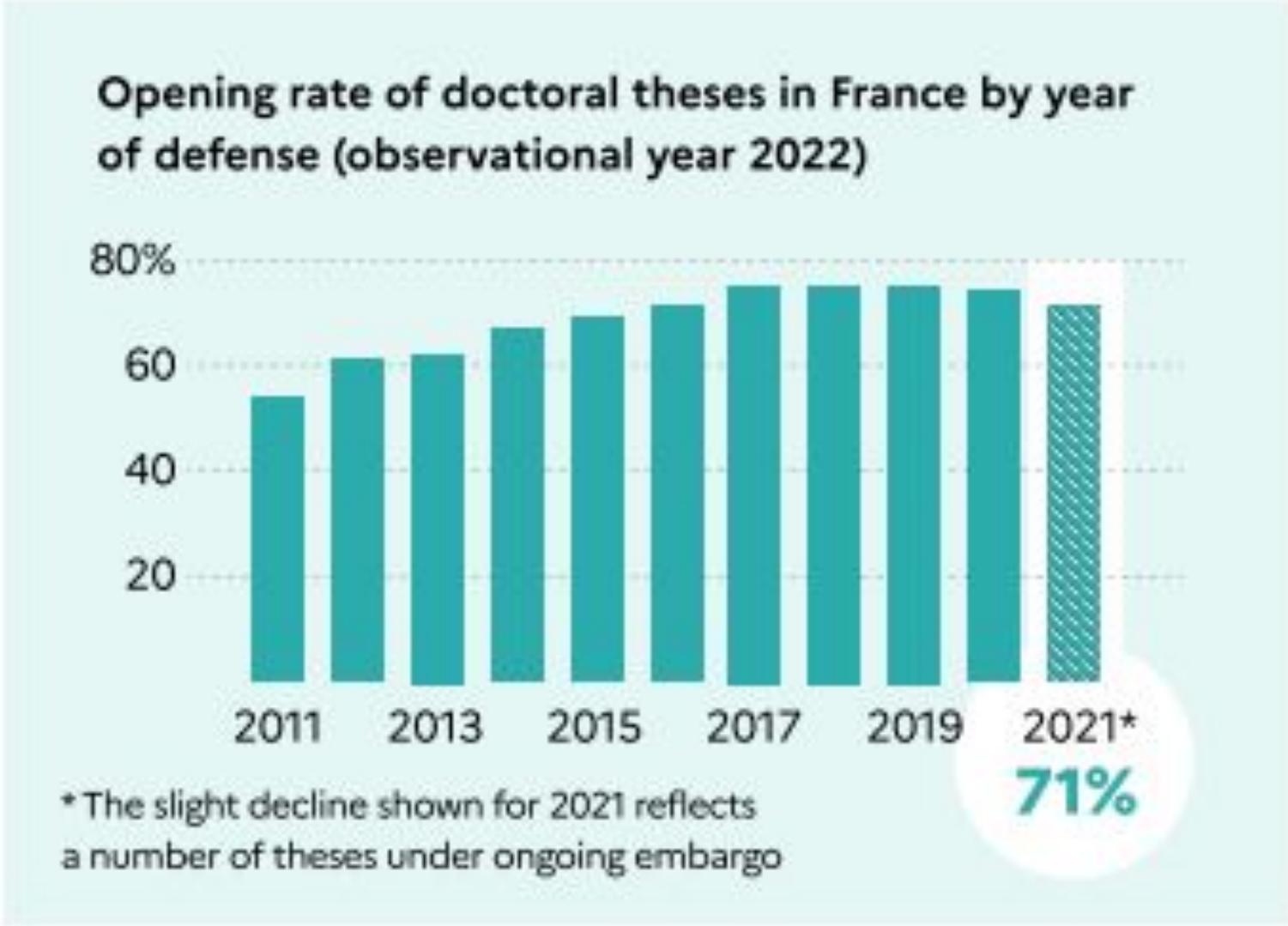
2013-2021: 1,426,140 publications

OPEN ACCESS RATE OF FRENCH PUBLICATIONS

OPEN ACCESS RATE OF PUBLICATIONS: BY DISCIPLINE



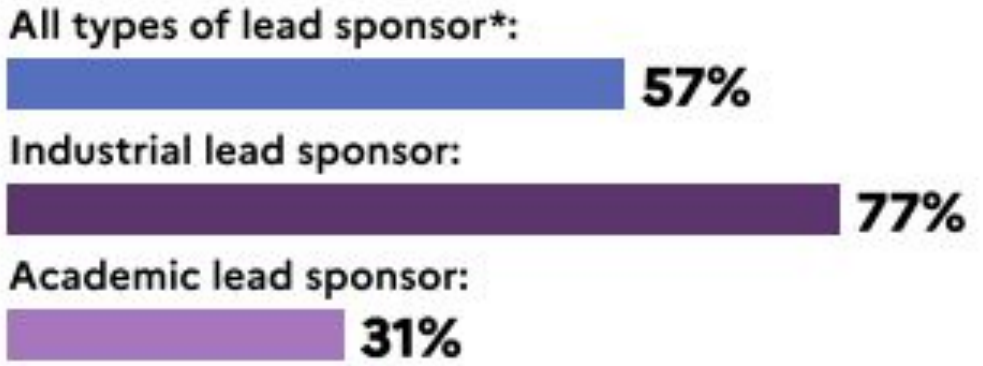
THE RESULTS OF THE LATEST RELEASE: PH.D. THESES



THE RESULTS OF THE LATEST RELEASE: CLINICAL TRIALS

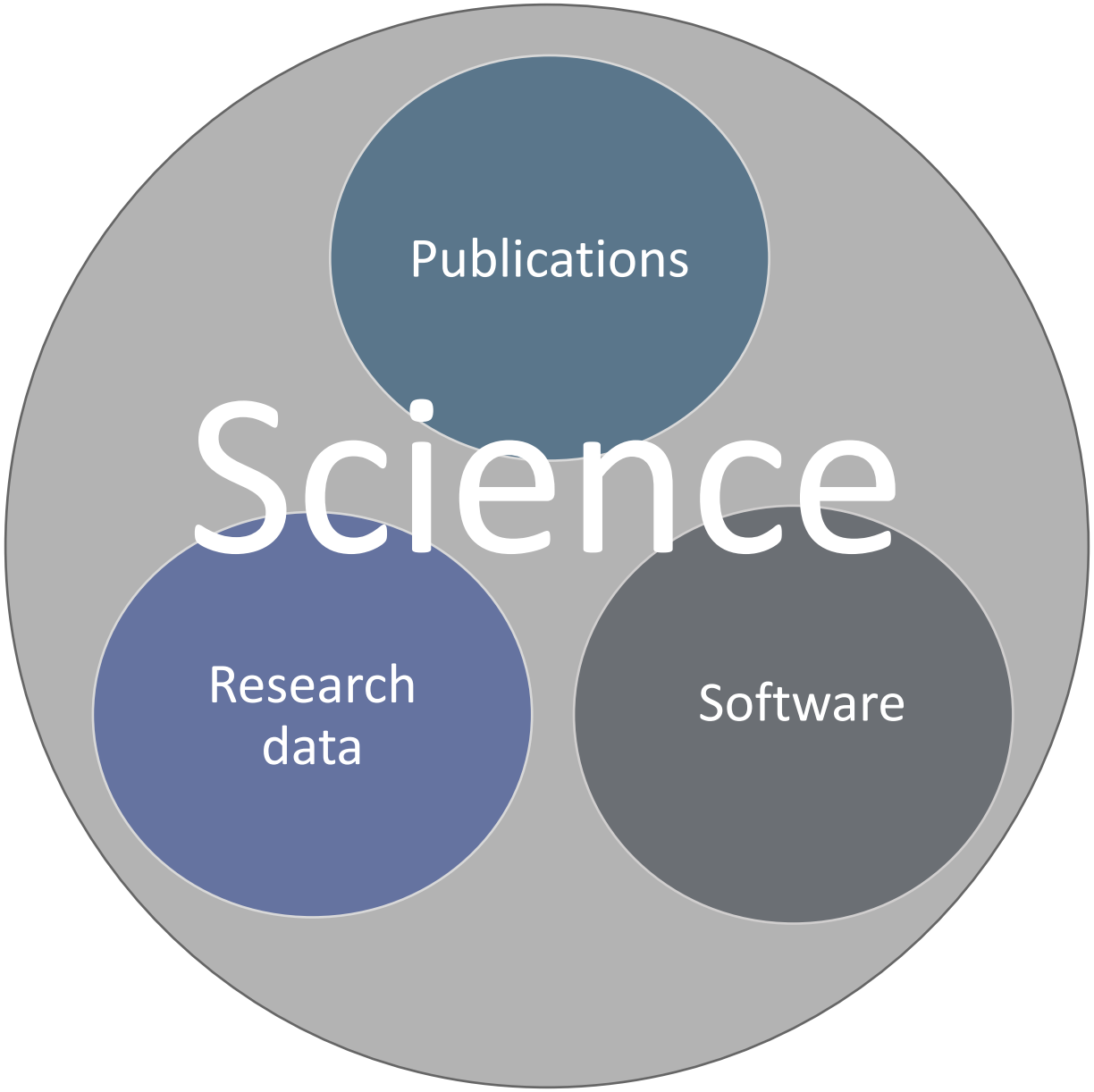
Clinical trials: 57% share their results

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



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

Openness of results of clinical trials has not moved since the later edition, with a sharing ratio of 57%. The registration of clinical trials and their results in public databases allows a rapid circulation of results, even when these have been unsuccessful and do not lead to a scientific publication. The significant variation between industrial and academic sponsors should be noticed.


Publications
Science
Research data
Software

Understanding research datasets

Research data repositories ?

Data repositories via DataCite suffer from many limitations:

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

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

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

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

Following research software activities

Software development in research is collaborative and distributed:

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

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

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

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

Mining data and software activities in scholarly full texts

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

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

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

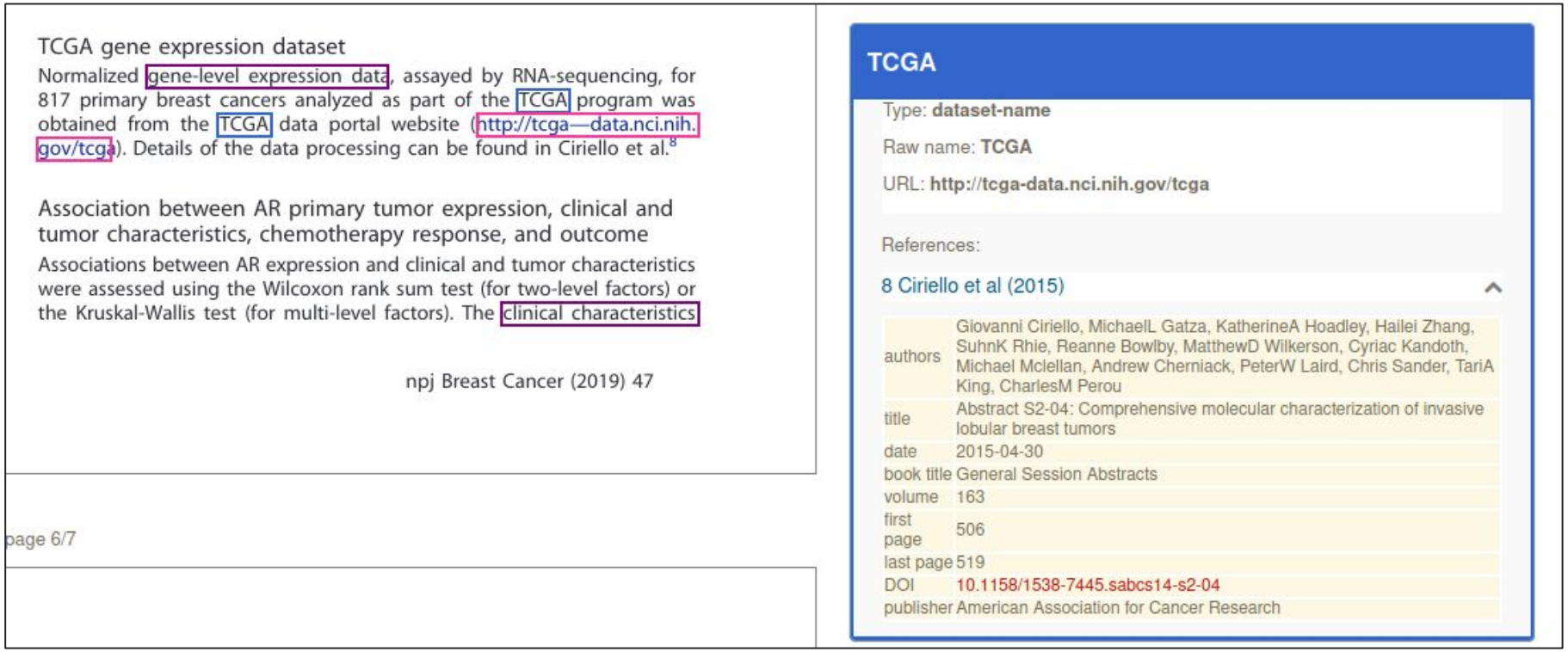
2) **Automatic characterization of the mention context:** is a mentioned dataset or software **used/created/shared** ?

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

MINING FULL-TEXTS FOR DATASETS MENTIONS

- **Approach** based on machine learning tools
  - **GROBID**: full-text structuring of PDF
  - detection of Data Availability sections, Materials and Methods, etc.

- **DataStet**: **dataset mention detection:**
  - based on DataSeer (2018-20, Sloan Found.)
  - trained on 22,000 manually annotated sentences
  - https://github.com/kermitt2/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 (http://tcga—data.nci.nih.gov/tcga). Details of the data processing can be found in Ciriello et al.[8]

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

npj Breast Cancer (2019) 47

page 6/7

TCGA

Type: **dataset-name**

Raw name: **TCGA**

URL: **http://tcga-data.nci.nih.gov/tcga**

References:

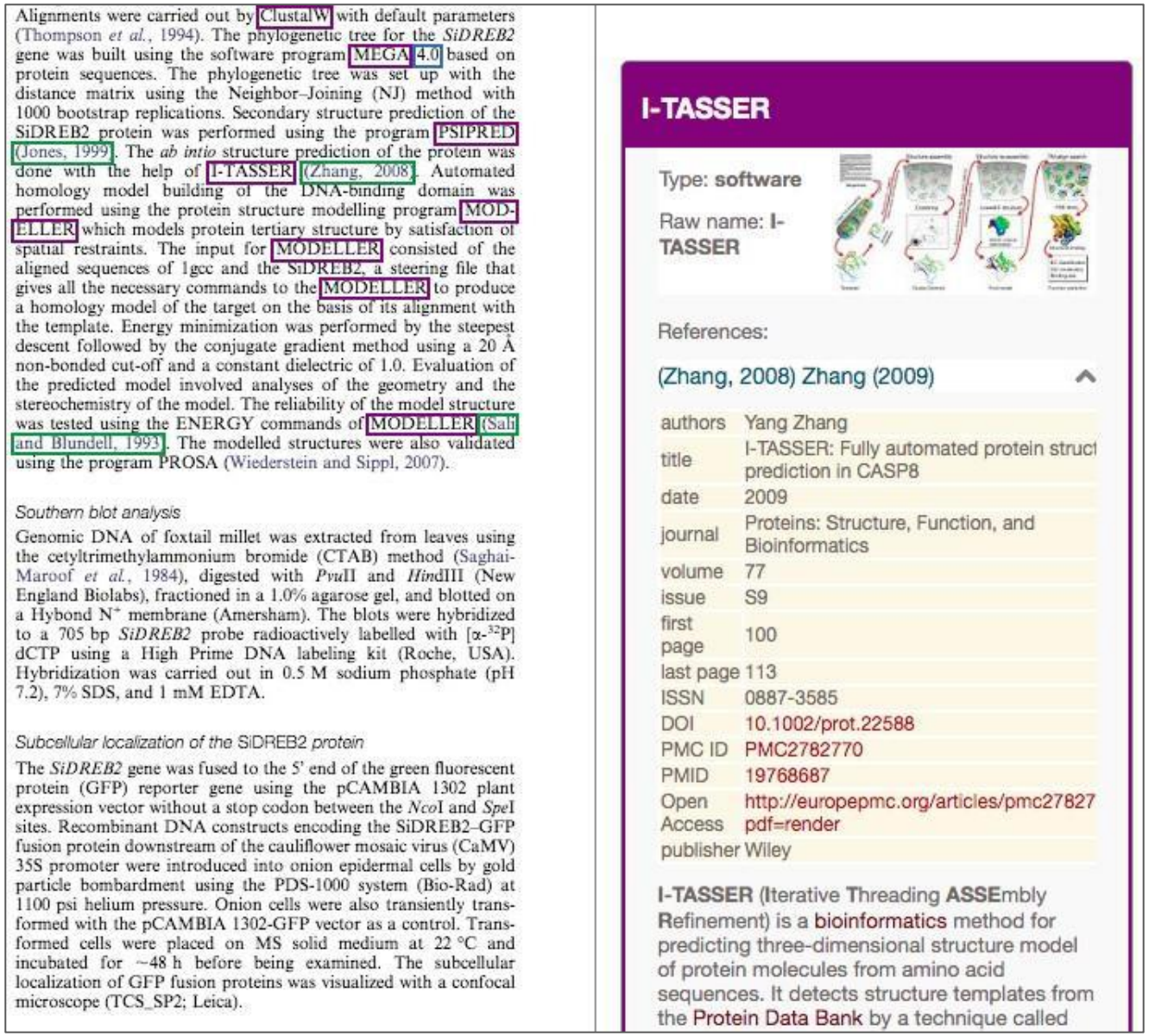
8 Ciriello et al (2015)

authors	Giovanni Ciriello, MichaelL Gatza, KatherineA Hoadley, Hailei Zhang, SuhnK Rhie, Reanne Bowlby, MatthewD Wilkerson, Cyriac Kandoth, Michael Mclellan, Andrew Cherniack, PeterW Laird, Chris Sander, TariA King, CharlesM Perou
title	Abstract S2-04: Comprehensive molecular characterization of invasive lobular breast tumors
date	2015-04-30
book title	General Session Abstracts
volume	163
first page	506
last page	519
DOI	10.1158/1538-7445.sabcs14-s2-04
publisher	American Association for Cancer Research

MINING FULL-TEXTS FOR SOFTWARE MENTIONS

- **Softcite**: **software mention detection**
  - funding Sloan & Moore Foundations, and French Open Science Plan
  - trained on 4,971 manually annotated documents (37 annotators)
  - https://github.com/softcite

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



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

Southern blot analysis

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

Subcellular localization of the SiDREB2 protein

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

I-TASSER

Type: **software**

Raw name: **I-TASSER**

References:

(Zhang, 2008) Zhang (2009)

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

I-TASSER (**I**terative **T**hreading **ASSE**mbly **R**efinement) is a **bioinformatics** method for predicting three-dimensional structure model of protein molecules from amino acid sequences. It detects structure templates from the **P**rotein **D**ata **B**ank by a technique called

MENTIONS TO DATASETS AND SOFTWARE

	# documents	share	sucessful download rate
Full corpus (2012-2021)	1,426,140	100 %	
Full text downloaded	908,567	63.7 %	63.7 %
→ open access	→ 660,501	46.3%	85.4%
→ closed access	→ 248,066	17.4%	38.0%

	# full text documents	# mentions
processed with Softcite	742,289	3,567,547
processed with DataStet	621,306	5,607,080

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

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

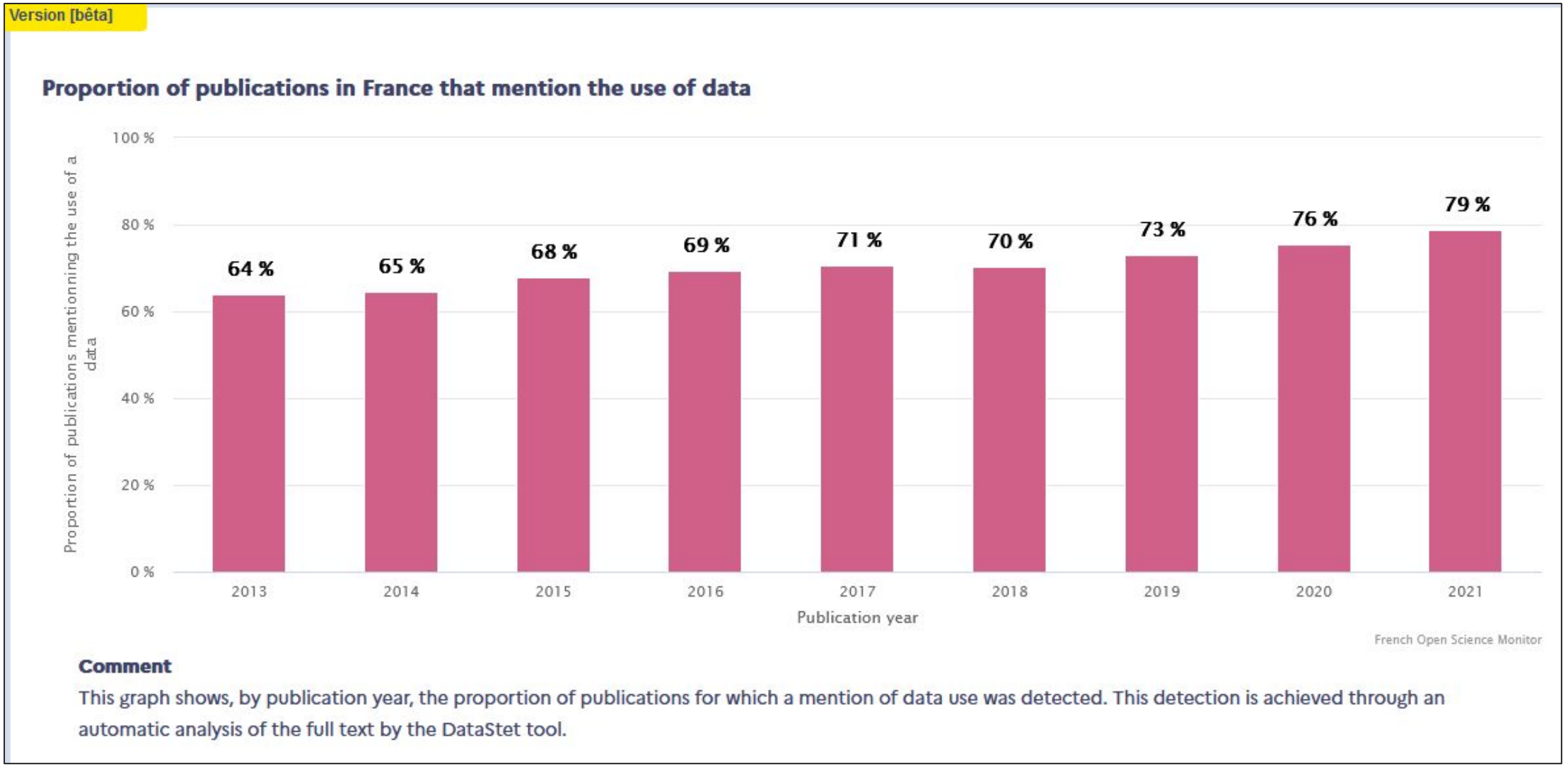
Monitoring dataset and software production

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

among all **processed publications**,

share of publications mentioning the use of data

Publications mentioning the use of data



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

Monitoring dataset and software production

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

among all **processed publications**,

share of publications mentioning the use of data

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

Monitoring dataset and software production

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

among all **processed publications**,

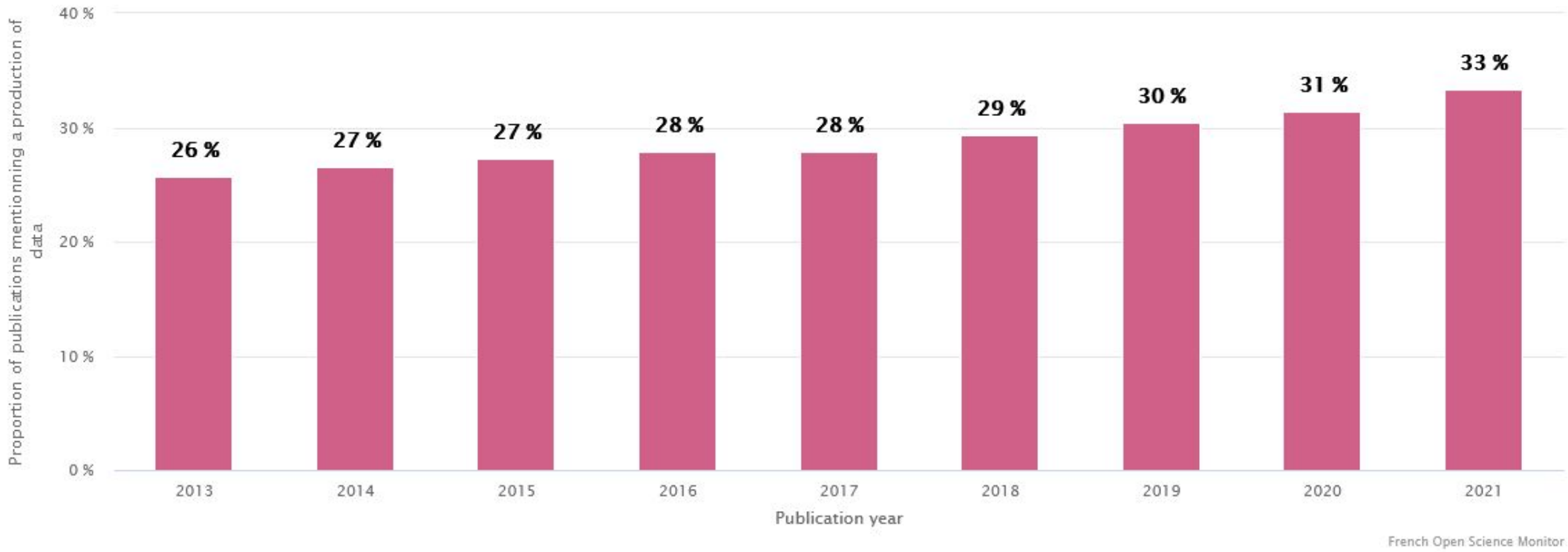
share of publications mentioning the use of data

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

share of publications mentioning the production of data

Publications mentioning the production of data

Version [bêta]



Comment

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

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

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

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

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**,

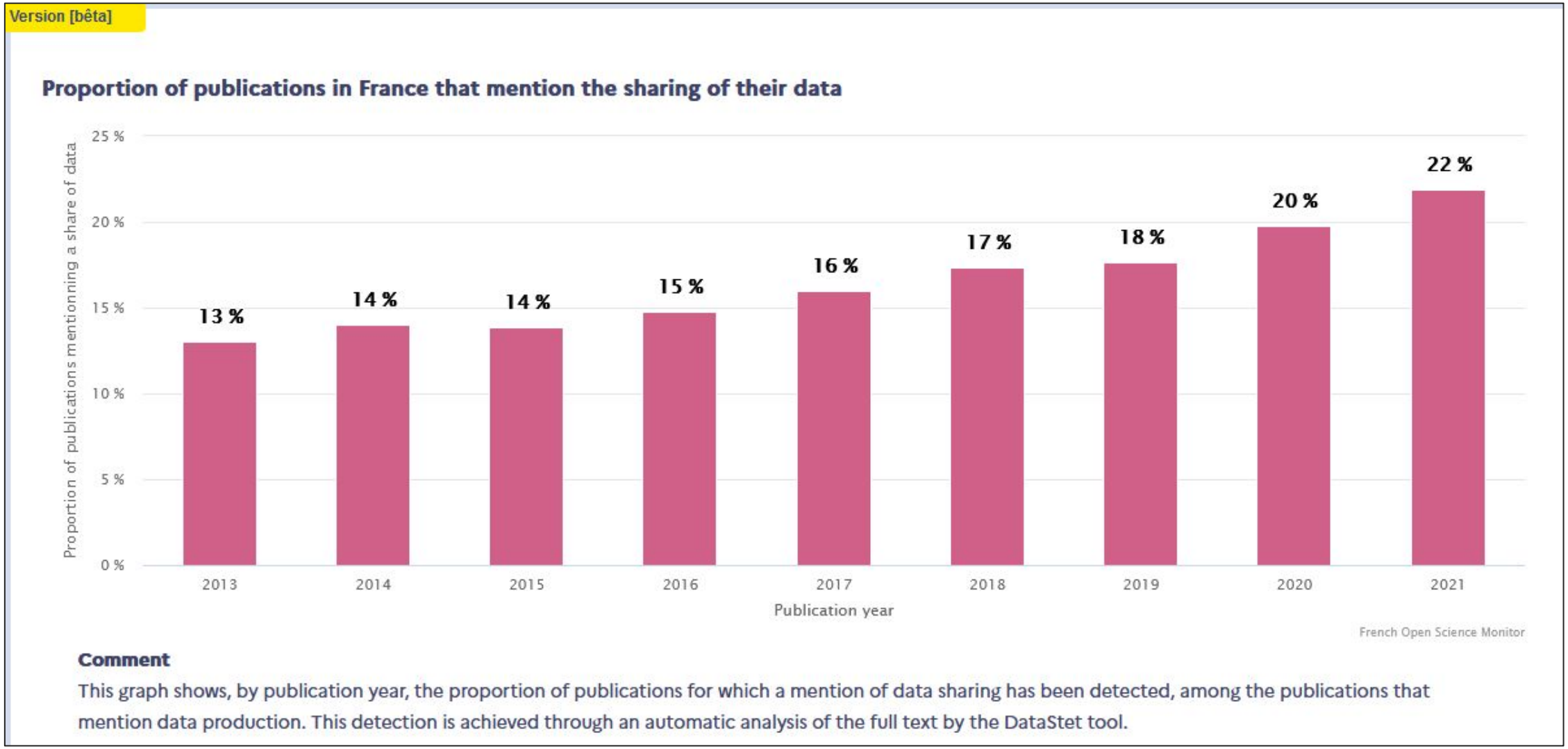
share of publications mentioning the production of data

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

share of publications mentioning the sharing of data

Publications mentioning sharing their created data

Version [bêta]

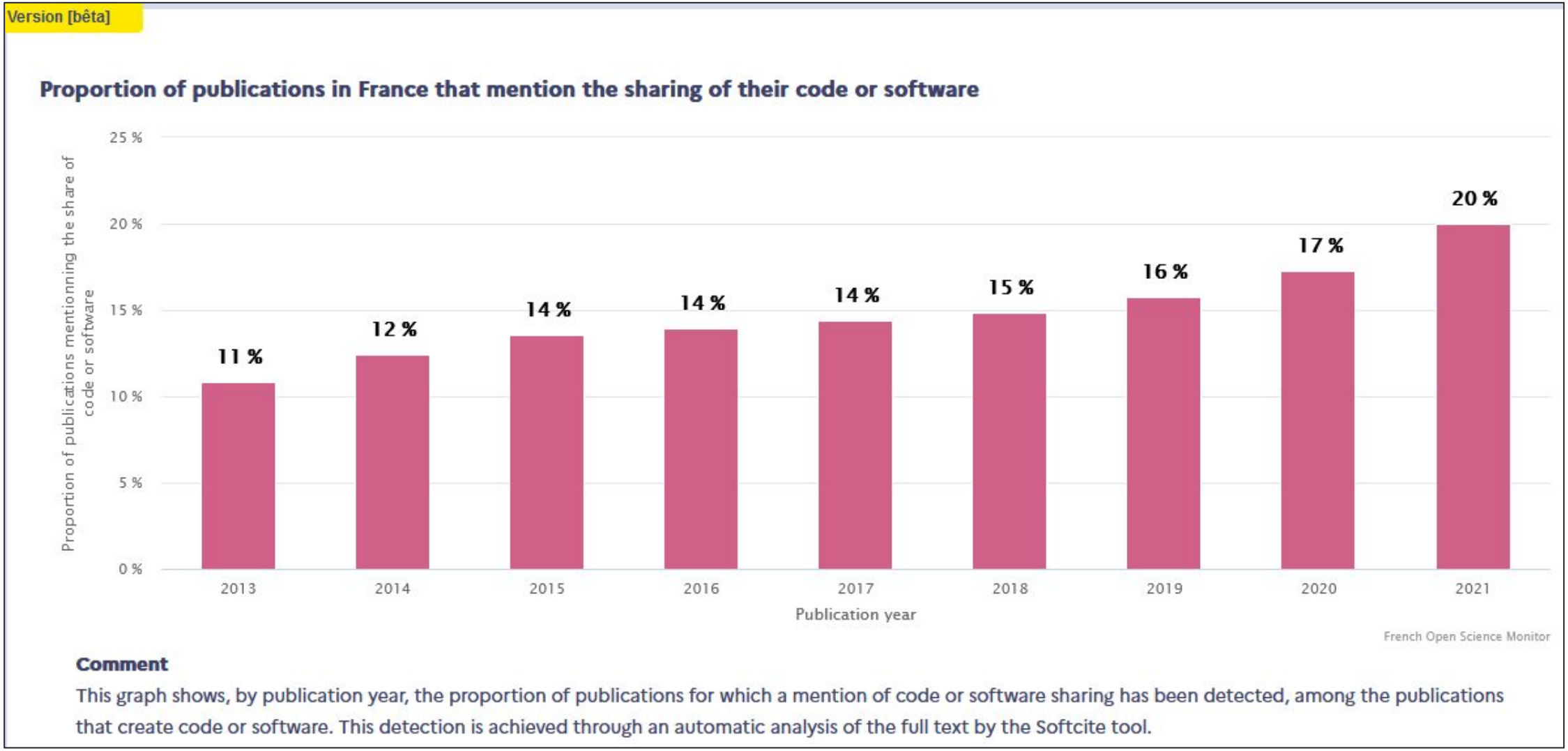


Comment

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

Publications mentioning sharing of their software

Version [bêta]



Comment

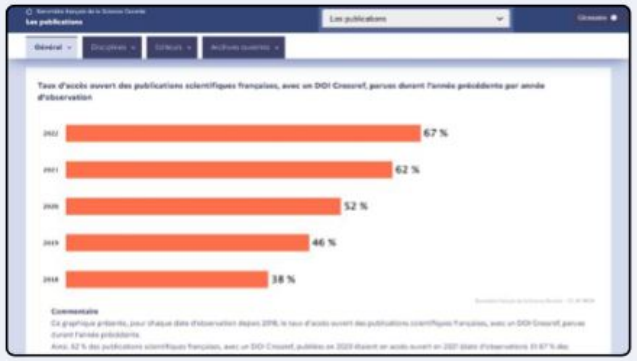
This graph shows, by publication year, the proportion of publications for which a mention of code or software sharing has been detected, among the publications that create code or software. This detection is achieved through an automatic analysis of the full text by the Softcite tool.

#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

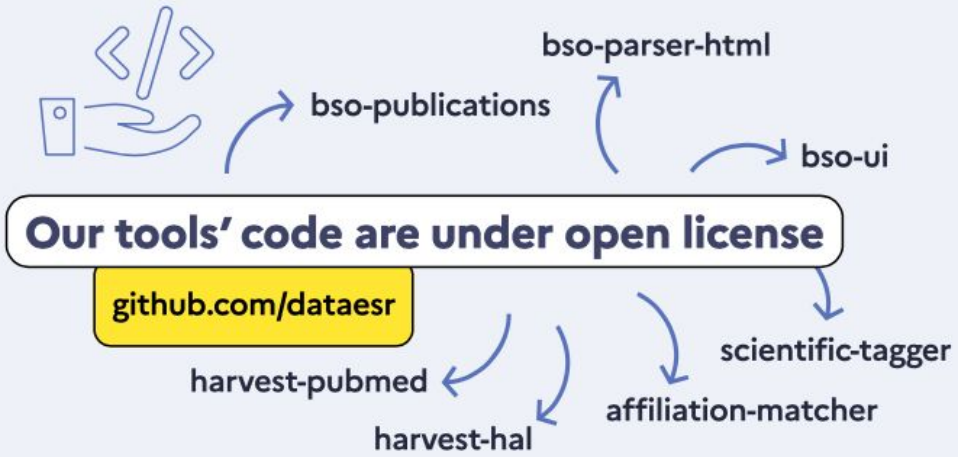
But also...

Local variations with Local monitors

barometredelascienceouverte.esr.gouv.fr/declinaisons/howto

Our tools' code are under open license

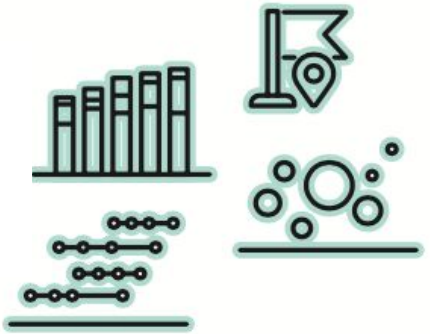
github.com/dataesr



At the service of institutions

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**

APPLYING THE MONITOR LOCALLY TO AN INSTITUTION

THANK YOU!

patrice.lopez@science-miner.com

HTTPS://FRENCHOPENSCIENCEMONITOR.ESR.GOUV.FR

CREDITS

Berlin cathedral dom: Image by user 12138562O from Pixabay
Parliament glass dom: Image by Thibaud Frere from Pixabay

REFERENCES

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

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

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

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