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

UNIVERSITÉ
DE LORRAINE

Inria

MONITORING OPEN SCIENCE BEYOND PUBLICATIONS

DATASETS AND SOFTWARE AS RESEARCH PRODUCTS TO BE SHARED

Laetitia BRACCO, Université de Lorraine

FROM MONITORING OPEN ACCESS TO PUBLICATIONS...

A LITTLE BIT OF CONTEXT IN FRANCE…

2018

First Open Science National Plan

2019

First version of the French National Open Science Monitor

2020

First local Open Science Monitor: Université de Lorraine

Many other institutions follow

2021

New indicators about Health on the National Monitor

Second Open Science National Plan

Beginning of the datasets and software project

2022

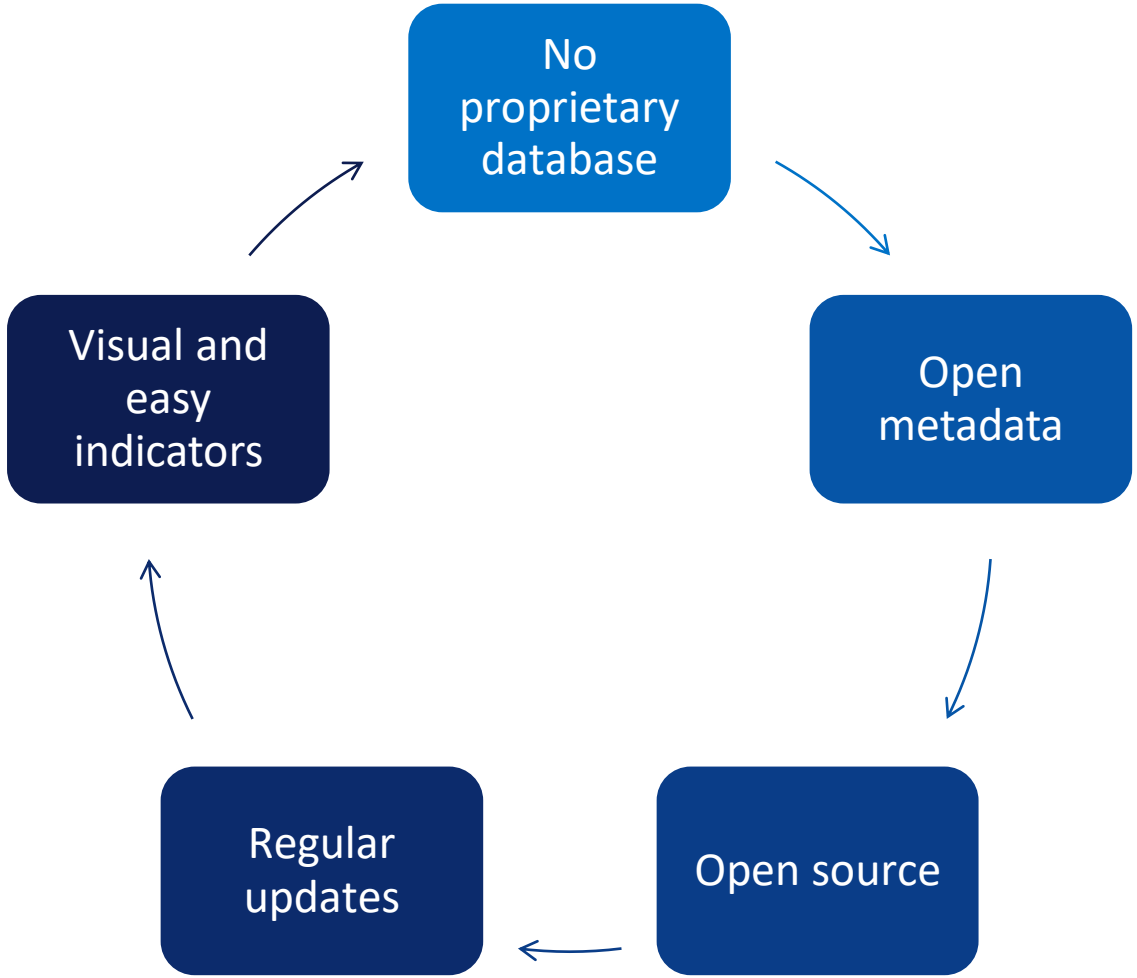
Creation of the Monitor Users Club by Université de Lorraine and the Ministry

2023

Release of the first indicators on research data and software

FOCUS ON THE NATIONAL OPEN SCIENCE MONITOR

- A need for a national open science monitor with open indicators
- What were the requirements?



THE BUILDING BLOCK OF THE FRENCH OPEN SCIENCE MONITOR

🏛 Affiliation metadata

- PubMed, Crossref, HAL
- 🏗 Crawling web pages
- 🏗 Automatic detection of countries

🔍 Characterising openness

- Detecting if the article is open access or not : Unpaywall
- 🏗 Qualifying the type of open access

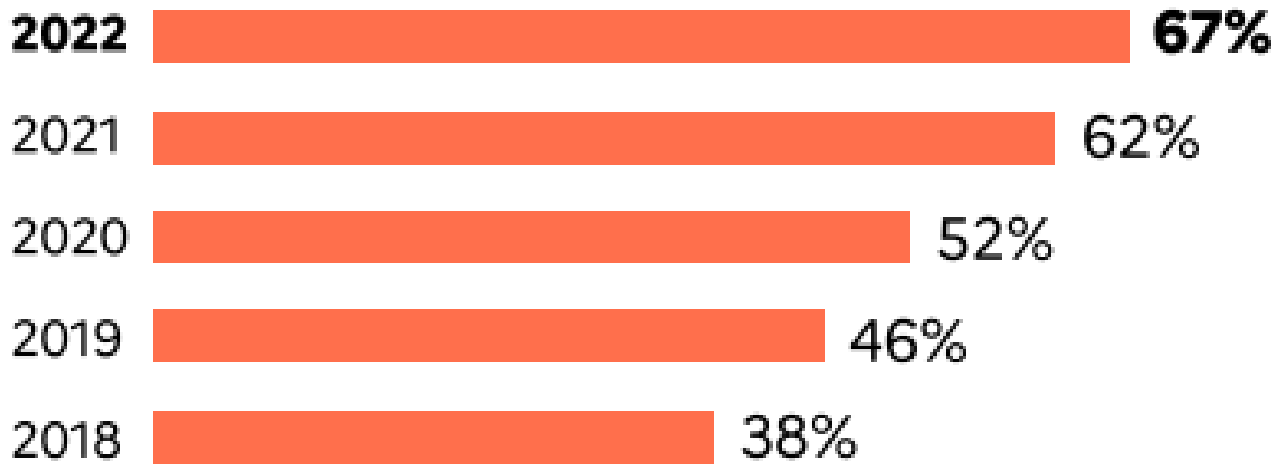
🤘 Thematic classification

- Training data : Pascal and Francis databases, Field of Research (FoR)
- 🏗 Automatic classification model (fastText)

🏗 : built-in by the Ministry for the project

THE RESULTS OF THE LATEST RELEASE: GLOBAL PUBLICATIONS

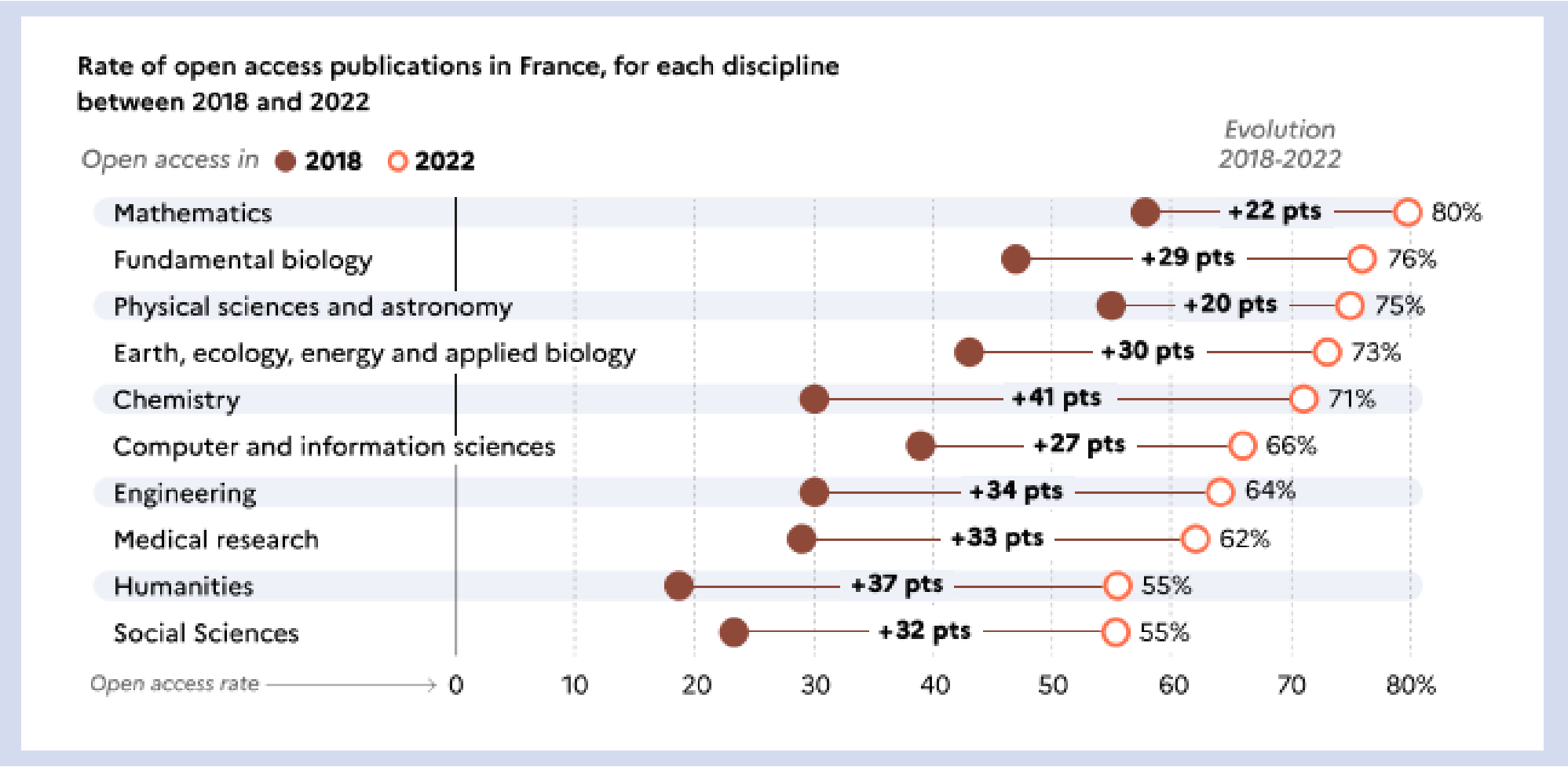
Open access rate of scientific publications in France, with a Crossref DOI, published during the previous year, by observation year



Growth
(all fields)
2018-2022

+29 points

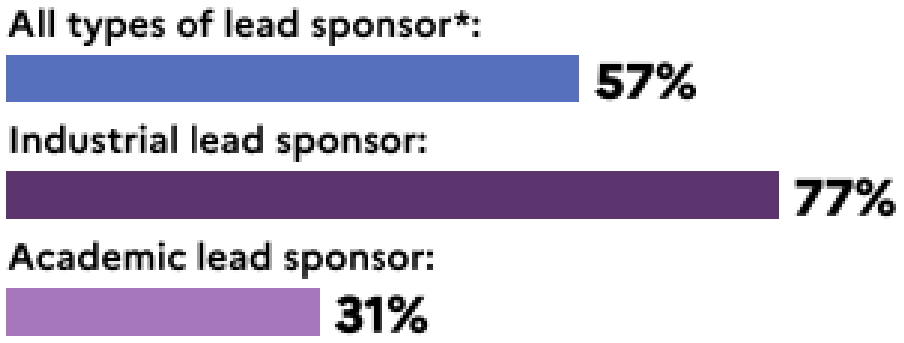
THE RESULTS OF THE LATEST RELEASE: BY DISCIPLINE



THE RESULTS OF THE LATEST RELEASE: CLINICAL TRIALS

Clinical trials: 57% share their results

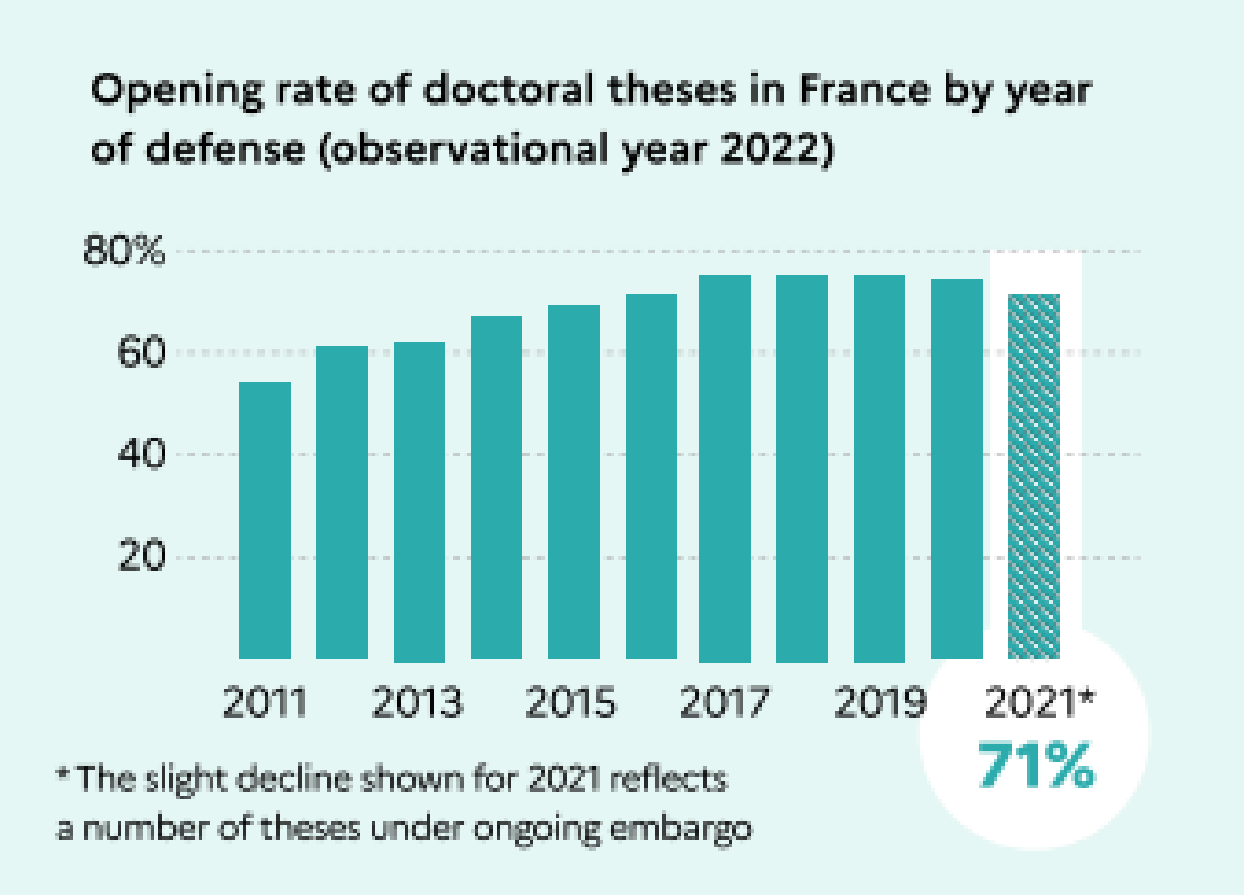
Share of clinical trials registered and completed in France in the past 10 years that have posted or published results



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

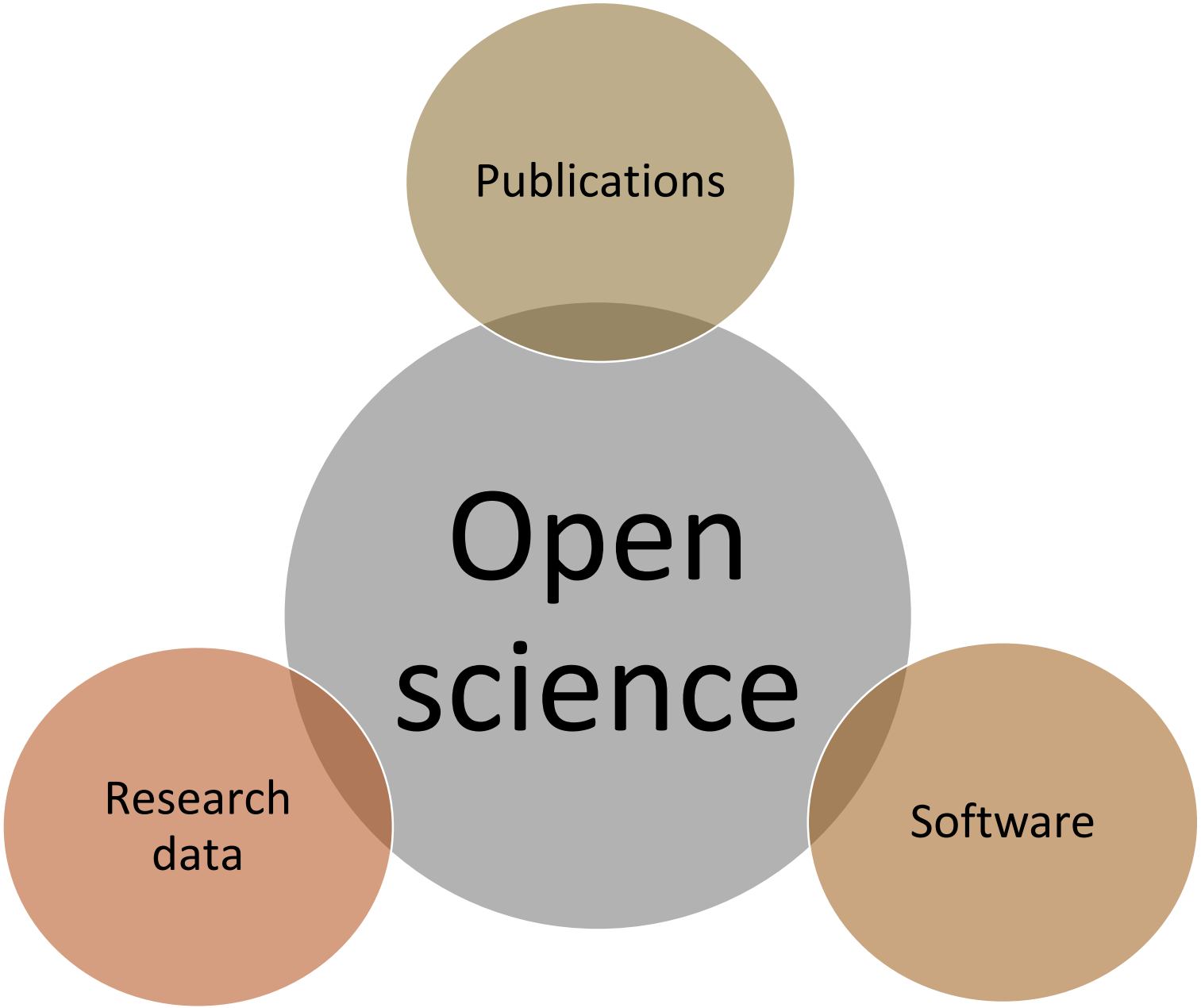
Openness of results of clinical trials has not moved since the later edition, with a sharing ratio of 57%. The registration of clinical trials and their results in public databases allows a rapid circulation of results, even when these have been unsuccessful and do not lead to a scientific publication.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

THE RESULTS OF THE LATEST RELEASE: THESES



…TO MONITORING OPEN SCIENCE

WHAT SHOULD WE CONSIDER AS A SCIENTIFIC PRODUCTION?



THE NEW FRENCH OPEN SCIENCE MONITOR: DATASETS AND SOFTWARE

- Gathering of a threefold and complementary team:

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

UNIVERSITÉ DE LORRAINE

Inria

- Winner of a funding from the European recovery plan:

Funded by the European Union
NextGenerationEU

FRANCE RELANCE

- Total cost of the project so far: **572 000 €**

WHAT ARE THE MAIN CHALLENGES?

Technical

- No global database for research data and software
- Too many identifiers for research data: DOI, accession number, entry number…
- And too few identifiers for both

Factual

- Low awareness from researchers on the value of these research products
- Low recognition in the individual assessment process

A DUAL METHODOLOGICAL APPROACH

2021/2023

Using publications

- Downloading the PDF documents of French publications
- Detecting and characterising mentions to datasets and software (GROBID, Softcite, DataStet)
- Computing indicators (ex : proportion of publications that share software or code)

2023/2024

Using repositories

- Dump of DataCite
- Identifying “French” DOIs using affiliations, as well as other metadata elements (publisher, clientId)
- Thematic enrichment
- Computing indicators

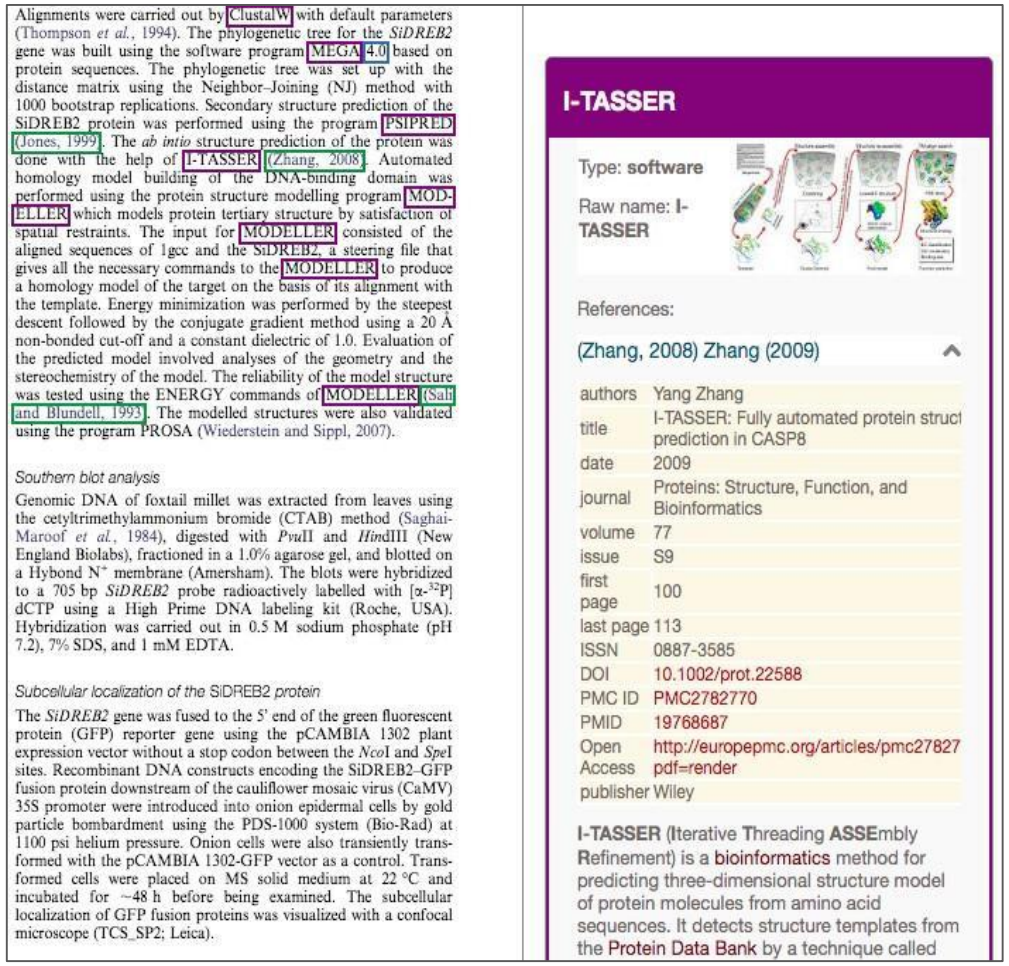
MINING FULL-TEXTS TO DETECT MENTIONS TO DATASETS AND SOFTWARE

- **Innovative approach** based upon the use and development of machine learning tools
  - GROBID: full-text structuring
  - Softcite: **software mention detection**
  - DataStet: **data set mention detection**

- Automatic characterisation of mentions: **usage / production or creation / sharing**

- Another challenge: **downloading massive amounts of full-texts**

science-miner



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

Southern blot analysis

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

Subcellular localization of the SiDREB2 protein

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

I-TASSER

Type: **software**

Raw name: **I-TASSER**

References:

(Zhang, 2008) Zhang (2009)

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

I-TASSER (Iterative Threading **ASSE**mbly 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

CHARACTERISING MENTIONS TO DATASETS AND SOFTWARE

Automatic characterization of mentions to **software** by means of Softcite, to **datasets** via DataStet:

- **used:** is the software/dataset mentioned in the article used in the research work?
- **created:** does the software/dataset results (created or being contributed to) from the research project?
- **shared:** is the software/dataset shared?

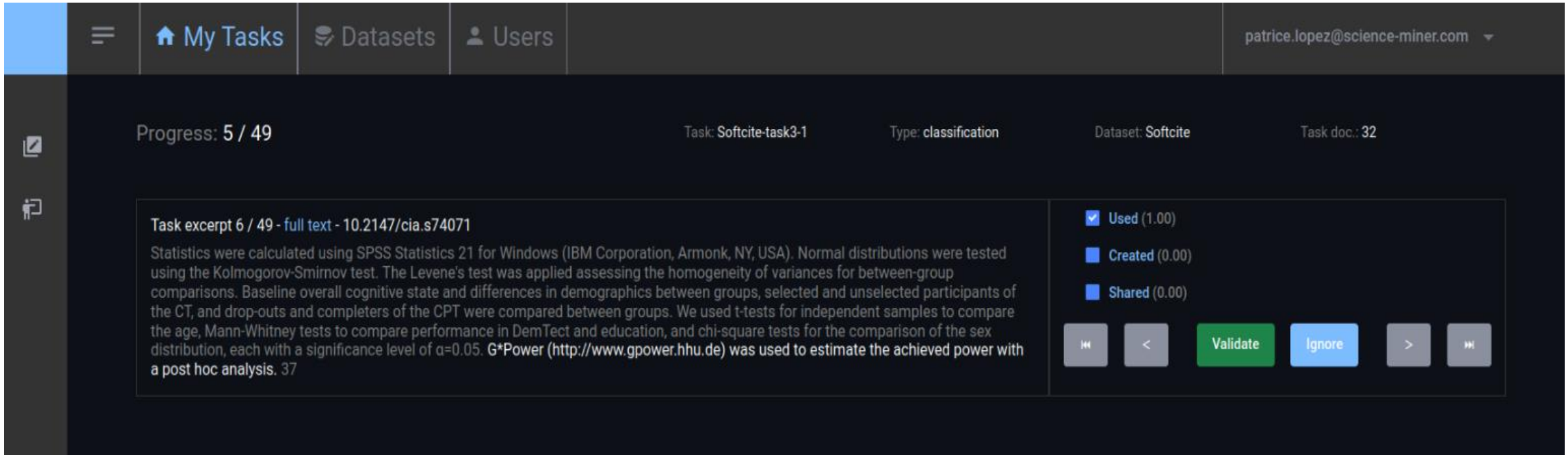
The classification models, based upon LinkBERT, have been trained on the basis of:

- the **Softcite** corpus (UT Austin/science-miner) : 4971 articles
- the **SoMeSci** corpus (GESIS Koeln/Uni Rostock) : 1367 articles

→ https://cloud.science-miner.com/software/
→ https://github.com/softcite/software-mentions

SUBSEQUENT MANUAL VERIFICATION

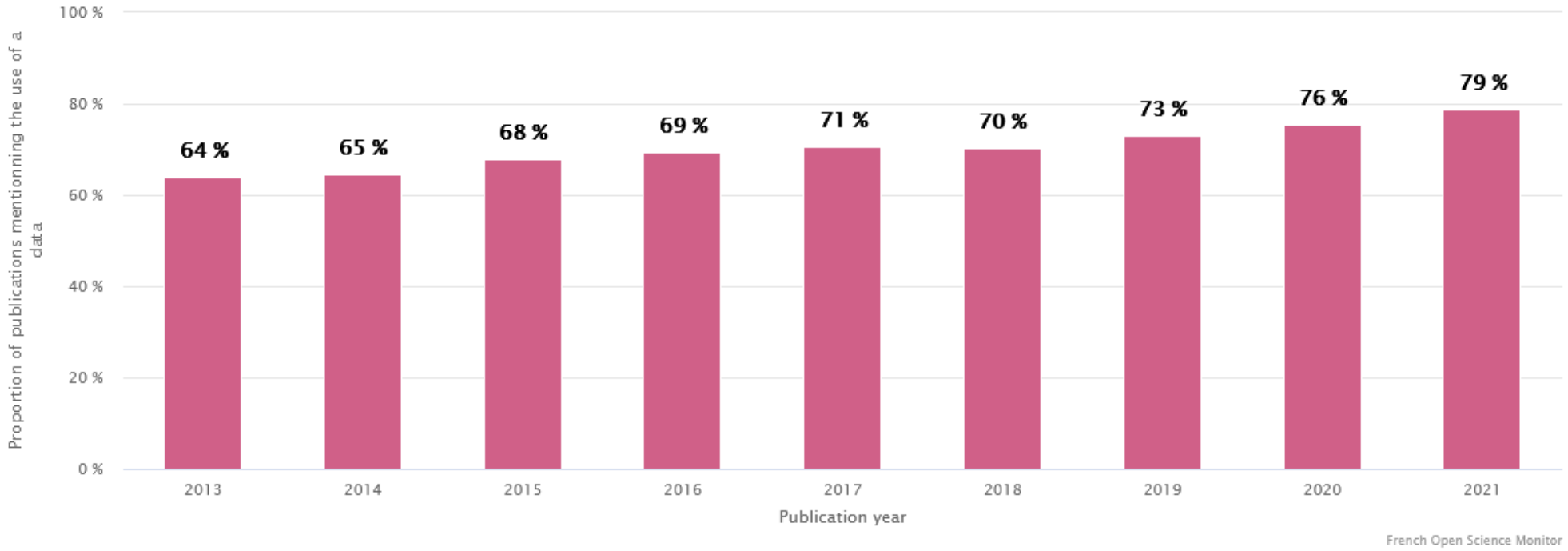
→ Manual annotation to improve the learning corpus



FIRST RESULTS: USING DATASETS

Version [bêta]

Proportion of publications in France that mention the use of data



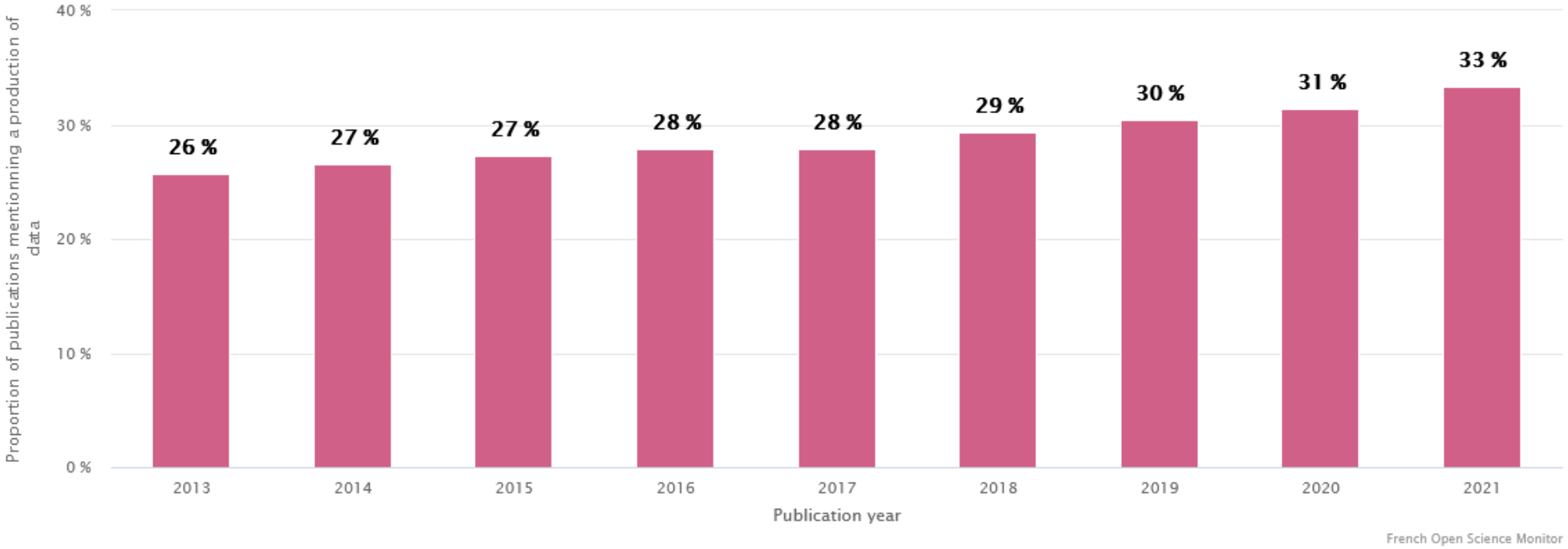
Comment

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

FIRST RESULTS: CREATING DATASETS

Version [bêta]

Proportion of publications in France that mention having produced their data



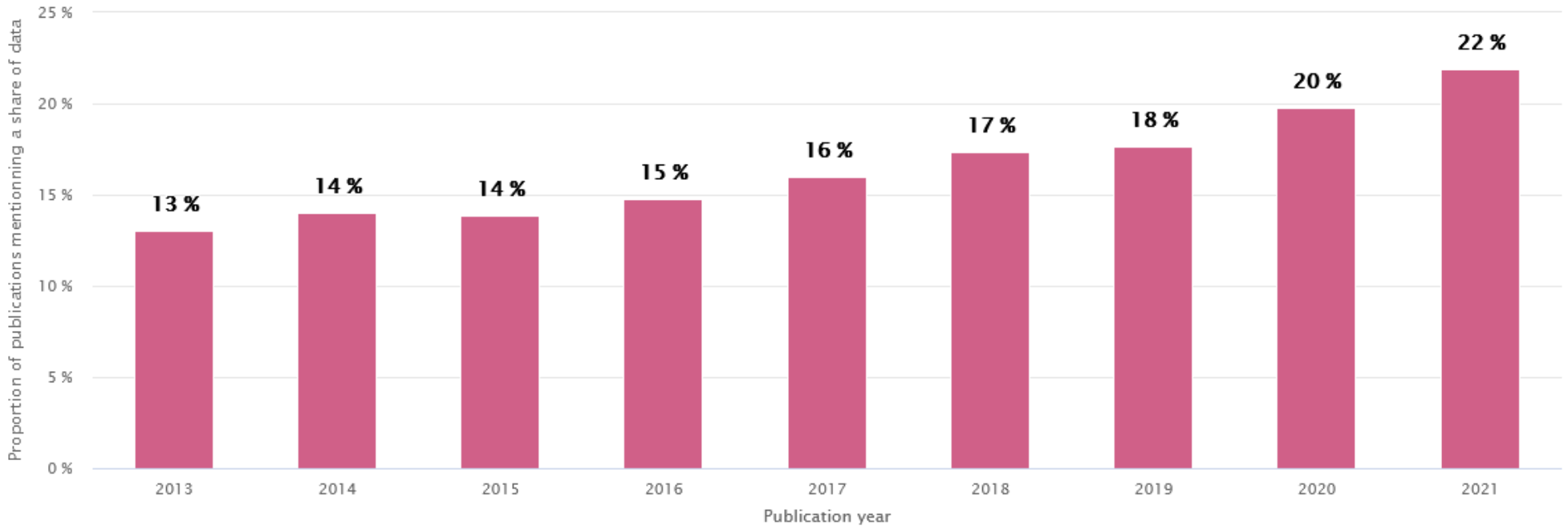
Comment

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.

FIRST RESULTS: SHARING DATASETS

Version [bêta]

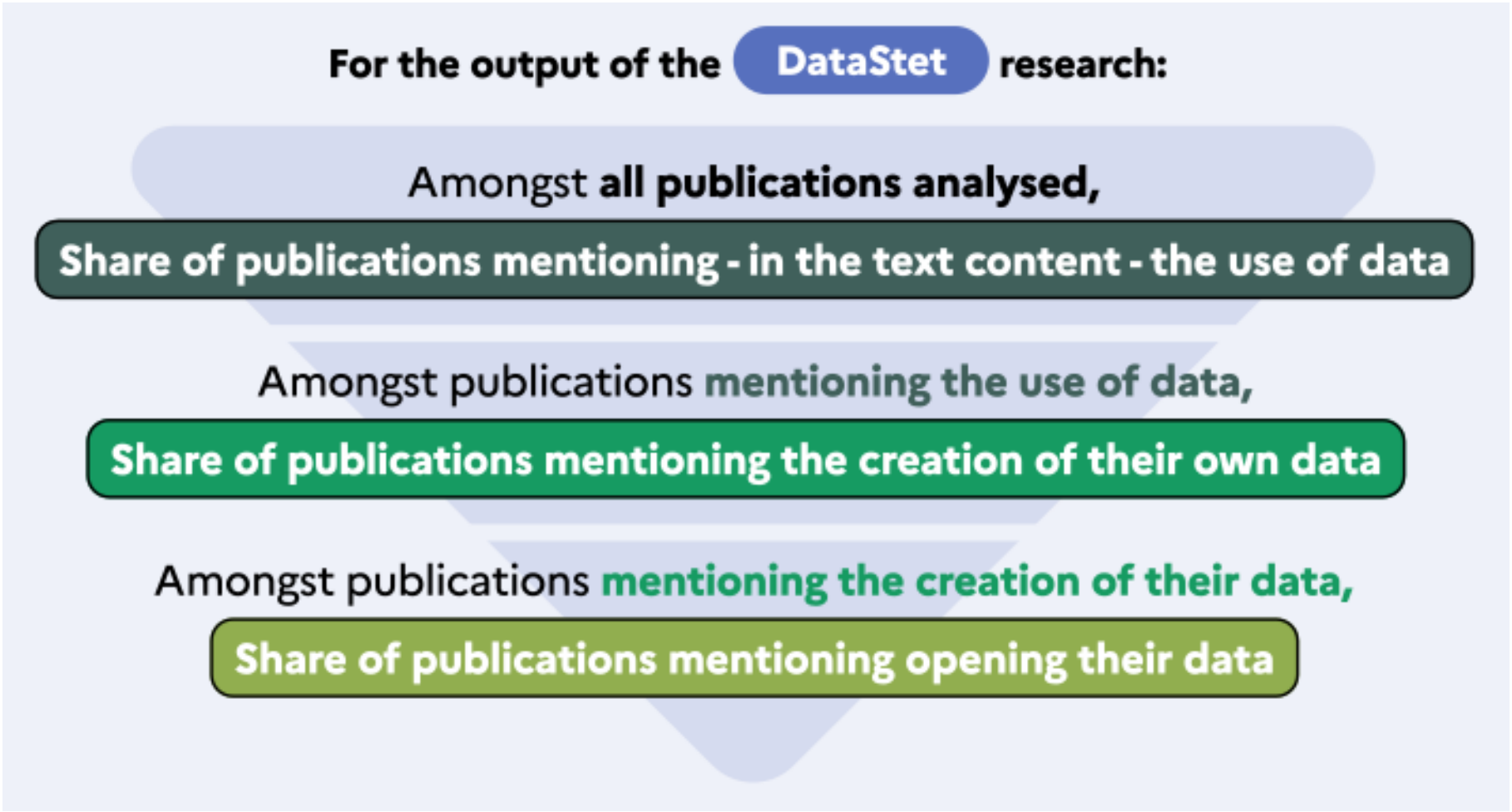
Proportion of publications in France that mention the sharing of their data



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.

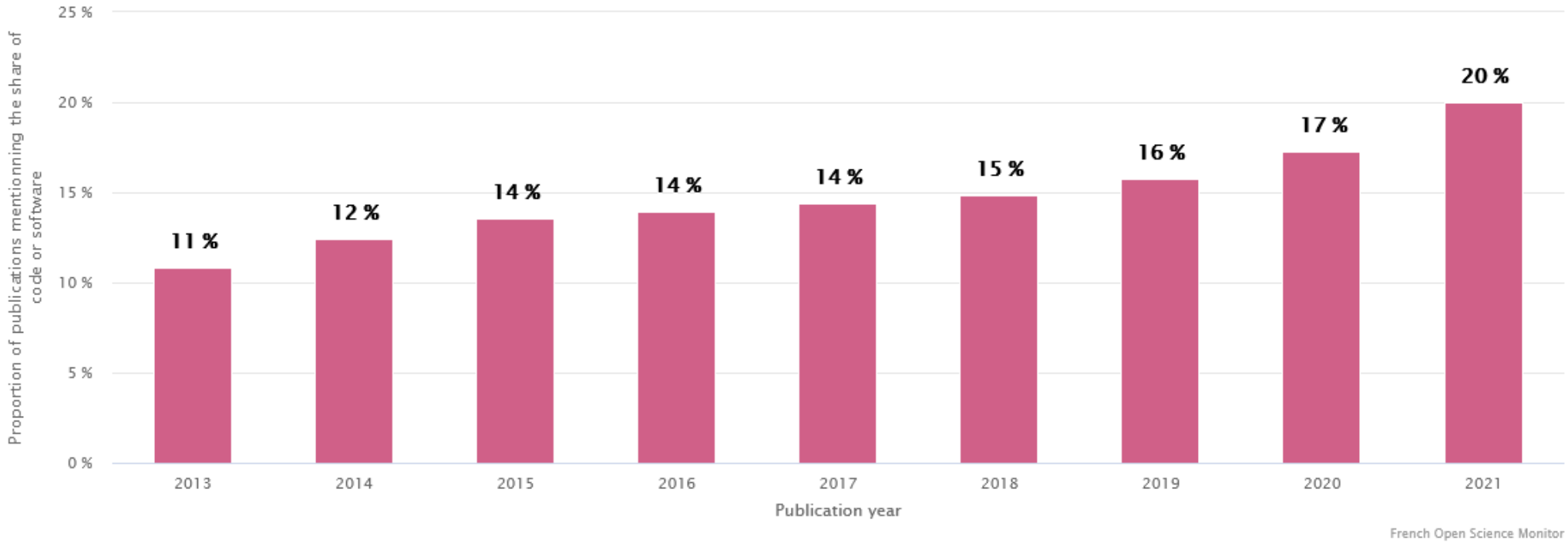
IN BRIEF


For the output of the DataStet research:
Amongst all publications analysed,
Share of publications mentioning - in the text content - the use of data
Amongst publications mentioning the use of data,
Share of publications mentioning the creation of their own data
Amongst publications mentioning the creation of their data,
Share of publications mentioning opening their data

FIRST RESULTS: SHARING 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.

METHODOLOGY FOR DATASET AND SOFTWARE SHARING

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

- Only for English publications

LOCAL MONITORS

APPLYING THE MONITOR TO AN INSTITUTION

- 177 universities, research organizations or research units have started an Open Science Monitor at their scale
- A strong local dynamics with an ever-growing community
- More than 200 individuals have subscribed to the Open Science Monitor Users Club

PERSPECTIVES

WHAT'S NEXT ?

- **Repository approach** (DataCite harvesting and enrichment), synergy identified with the Recherche Data Gouv repository
- **Improved full-text mining models** for data and code/software
- Processing of **French-language publications**
- New **ORCID** monitoring indicators
- **International development**
  - UNESCO
  - OpenAlex
  - COKI
  - CWTS

THANK YOU!

LAETITIA.BRACCO@UNIV-LORRAINE.FR

HTTPS://FRENCHOPENSCIENCEMONITOR.ESR.GOUV.FR/

CREDITS

Road: Image by Larisa Koshkina from Pixabay
Aurora: Image by Noel Bauza from Pixabay
Caution: Image by memyselfaneye from Pixabay
Green statistics: Storyset by Freepik
Telescope: Everypixel by Arnaud Papa