




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



UNIVERSITÉ
DE LORRAINE

Inria

MONITORING OPEN SCIENCE BEYOND PUBLICATIONS

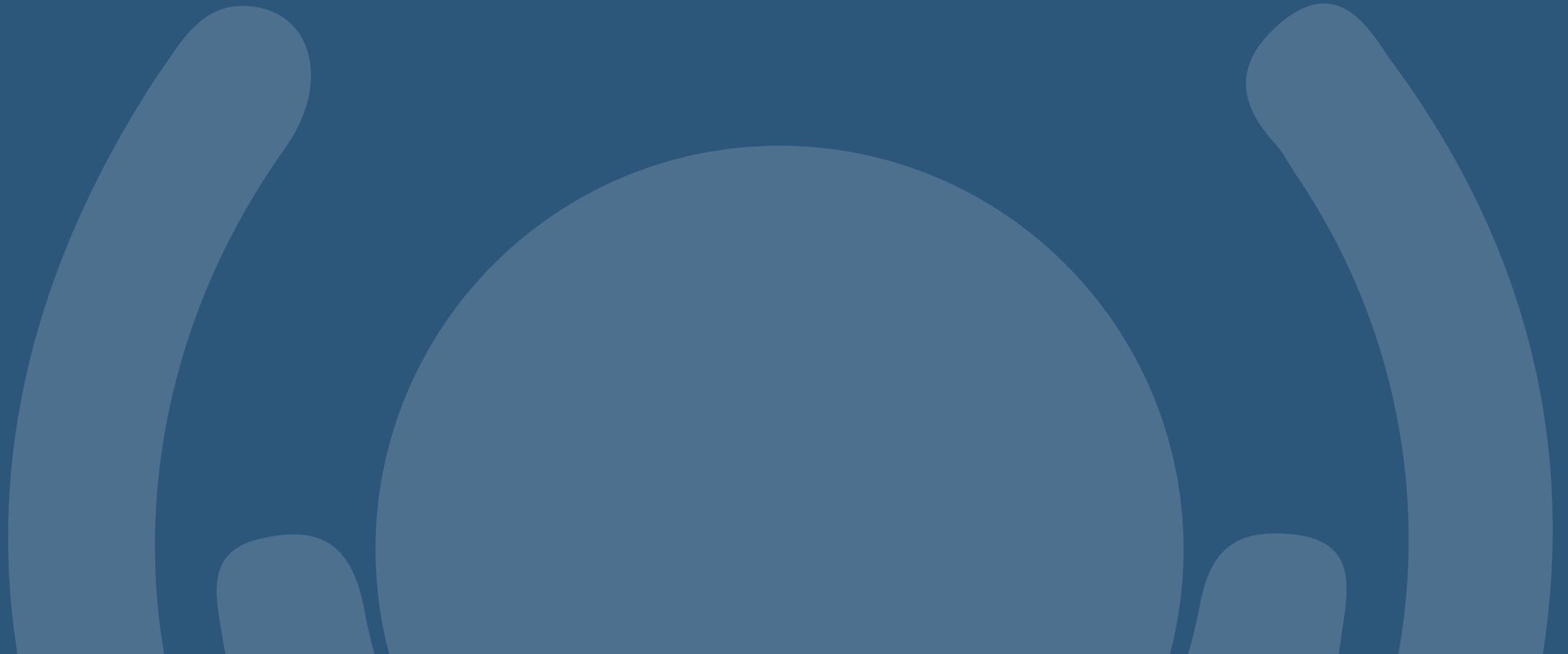
**THE FRENCH OPEN SCIENCE MONITOR ON RESEARCH
DATA AND SOFTWARE**

Slovak Open Science Forum, 14/11/24

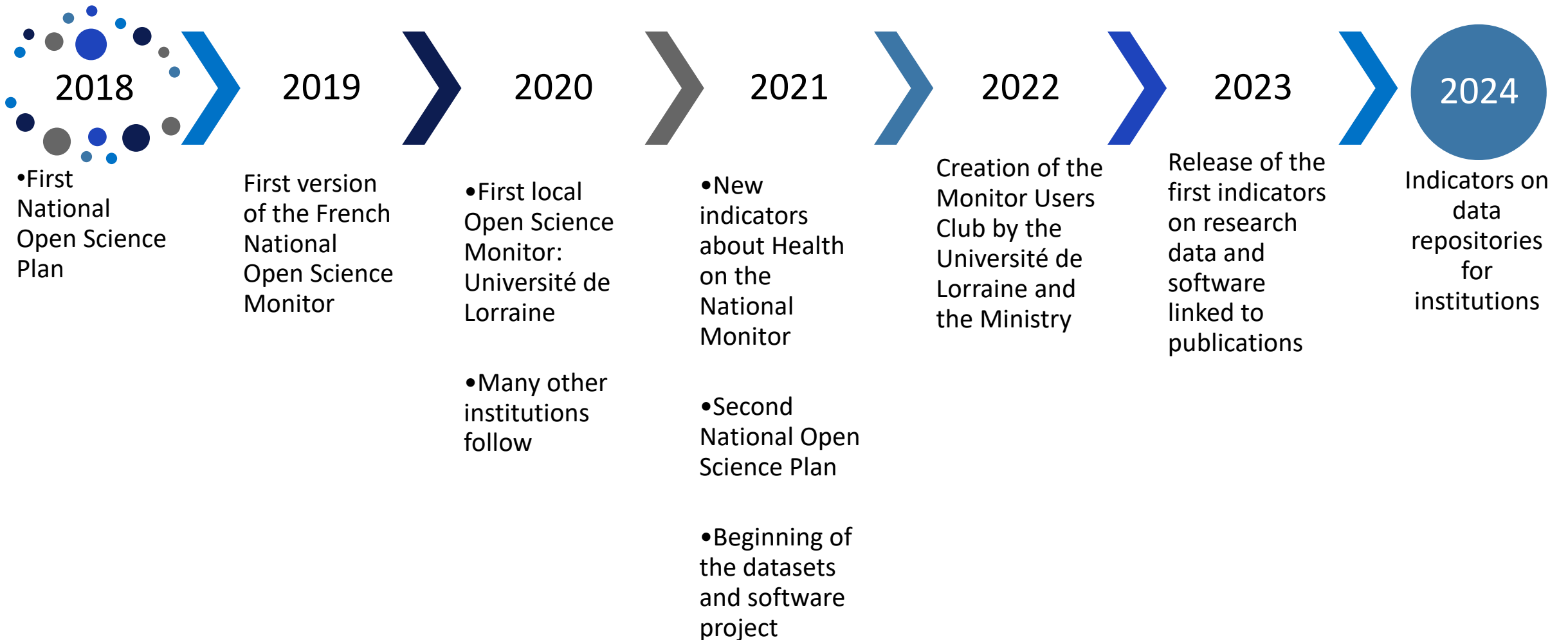
Laetitia BRACCO, Université de Lorraine



FROM MONITORING OPEN ACCESS TO PUBLICATIONS...

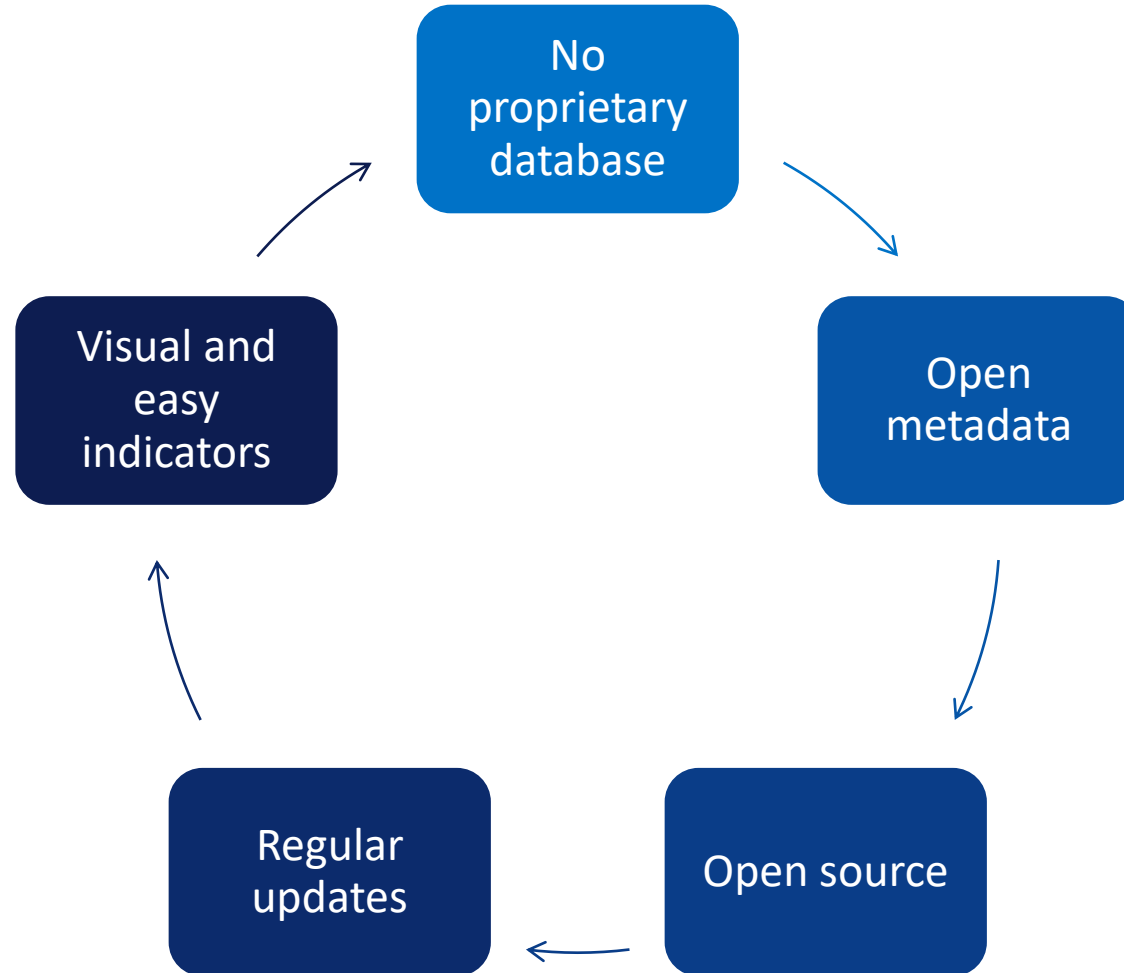


A LITTLE BIT OF CONTEXT IN FRANCE...



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

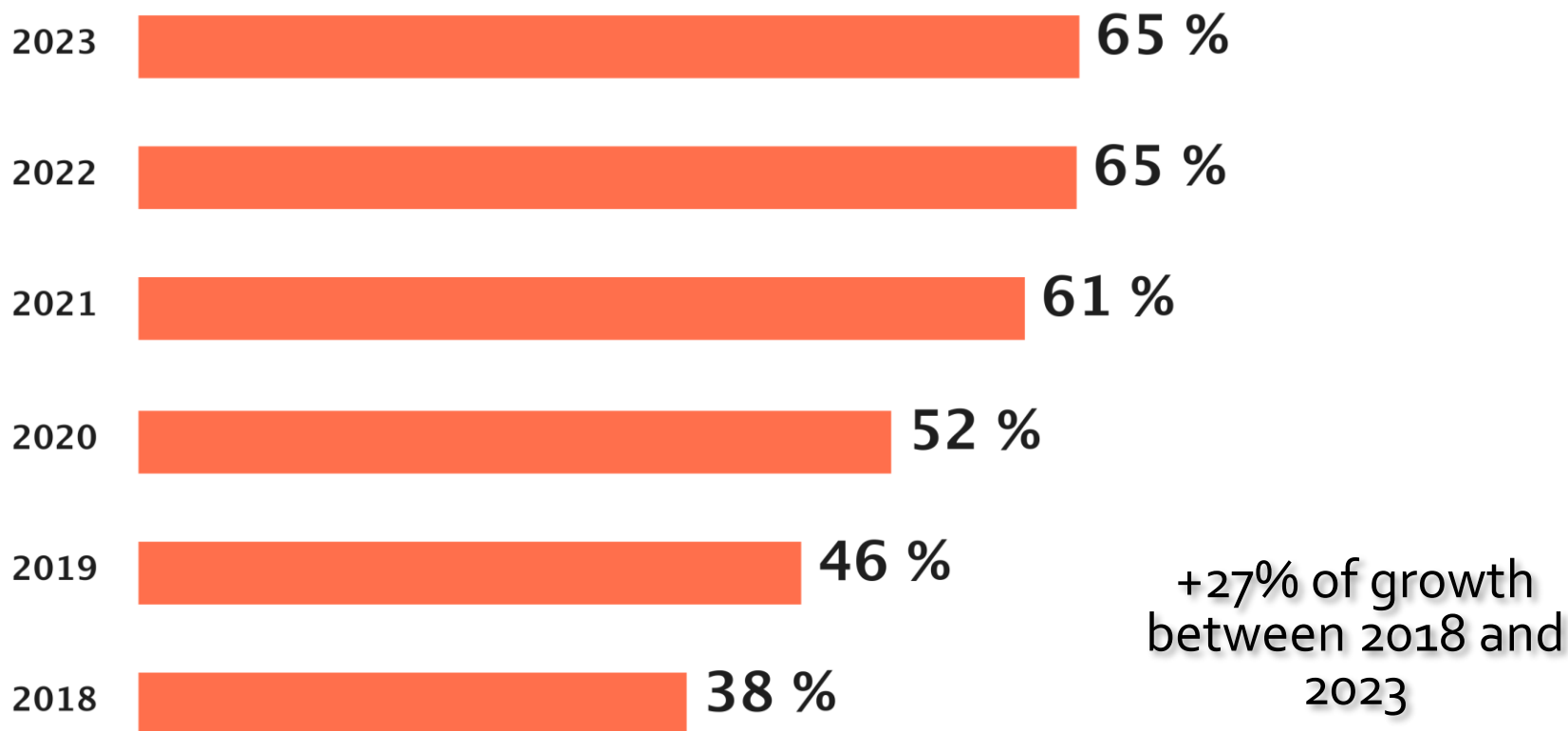
🔍 Thematic classification

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

🔍 : built-in by the Ministry for the project

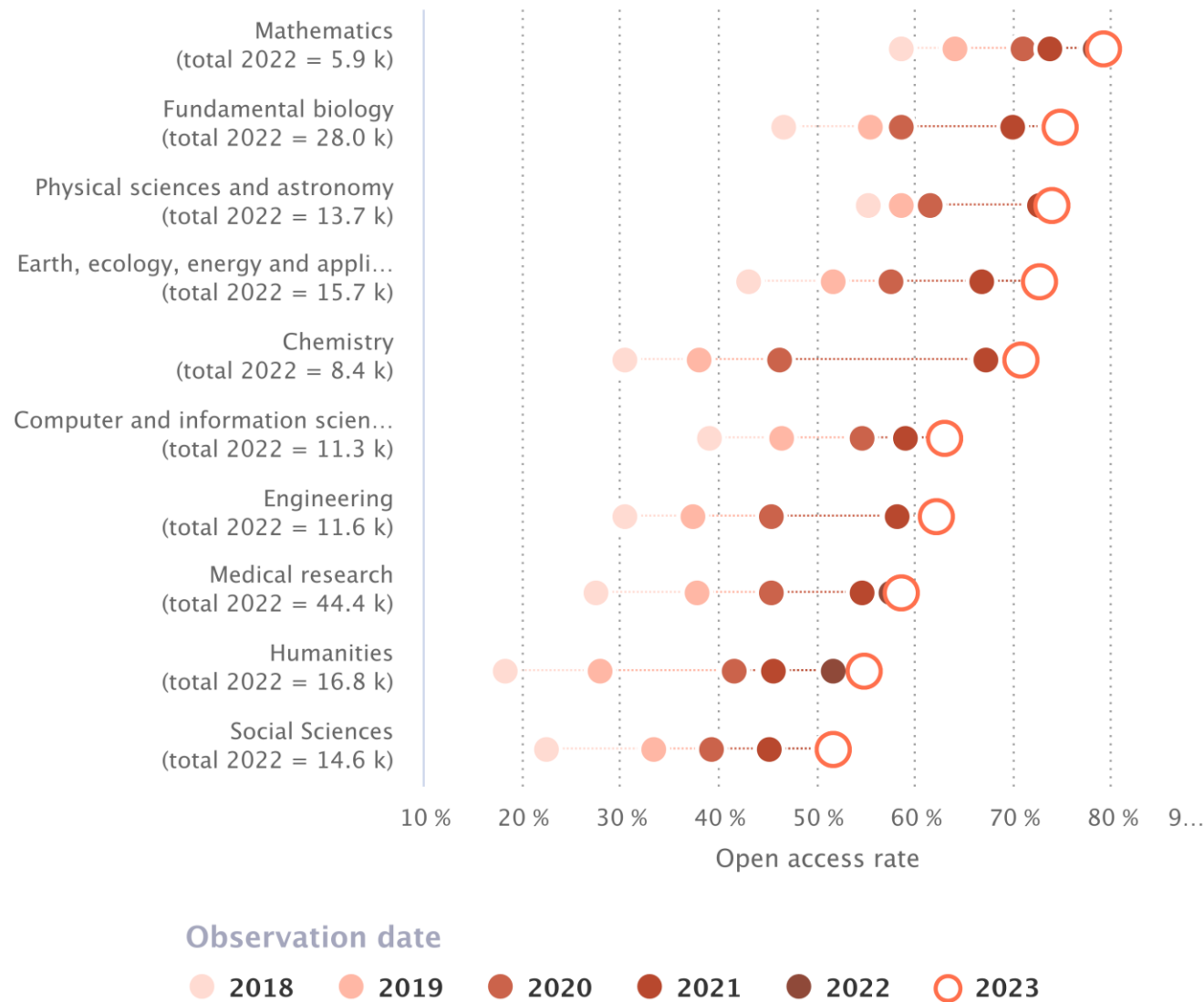
THE RESULTS OF THE LATEST RELEASE: GLOBAL PUBLICATIONS

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



THE RESULTS OF THE LATEST RELEASE: BY DISCIPLINE

Evolution of the rate of scientific publications in open access in France, with a Crossref DOI, for each discipline by year of observation

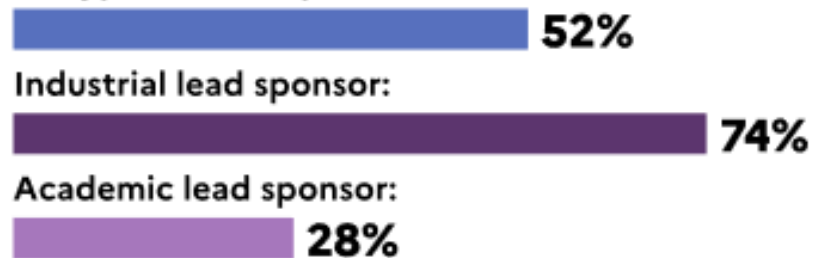


THE RESULTS OF THE LATEST RELEASE: CLINICAL TRIALS

Clinical trials: 52% share their results within 3 years

Percentage of registered clinical trials completed in 2020 that have posted a result and/or declared a scientific publication within 3 years of the end of the trial

All types of lead sponsor*:

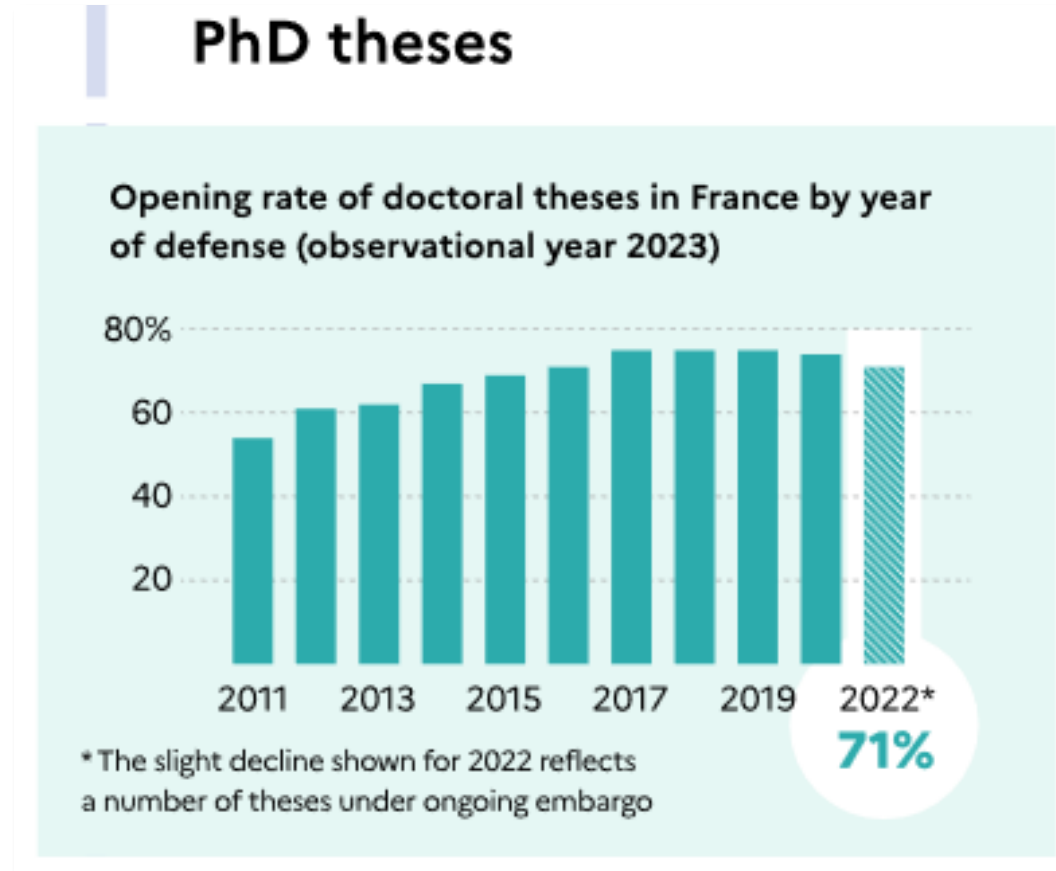


* Individual or legal entity in charge of research conducted on human beings who initiates, finances and supervises the conduct of the clinical trial.

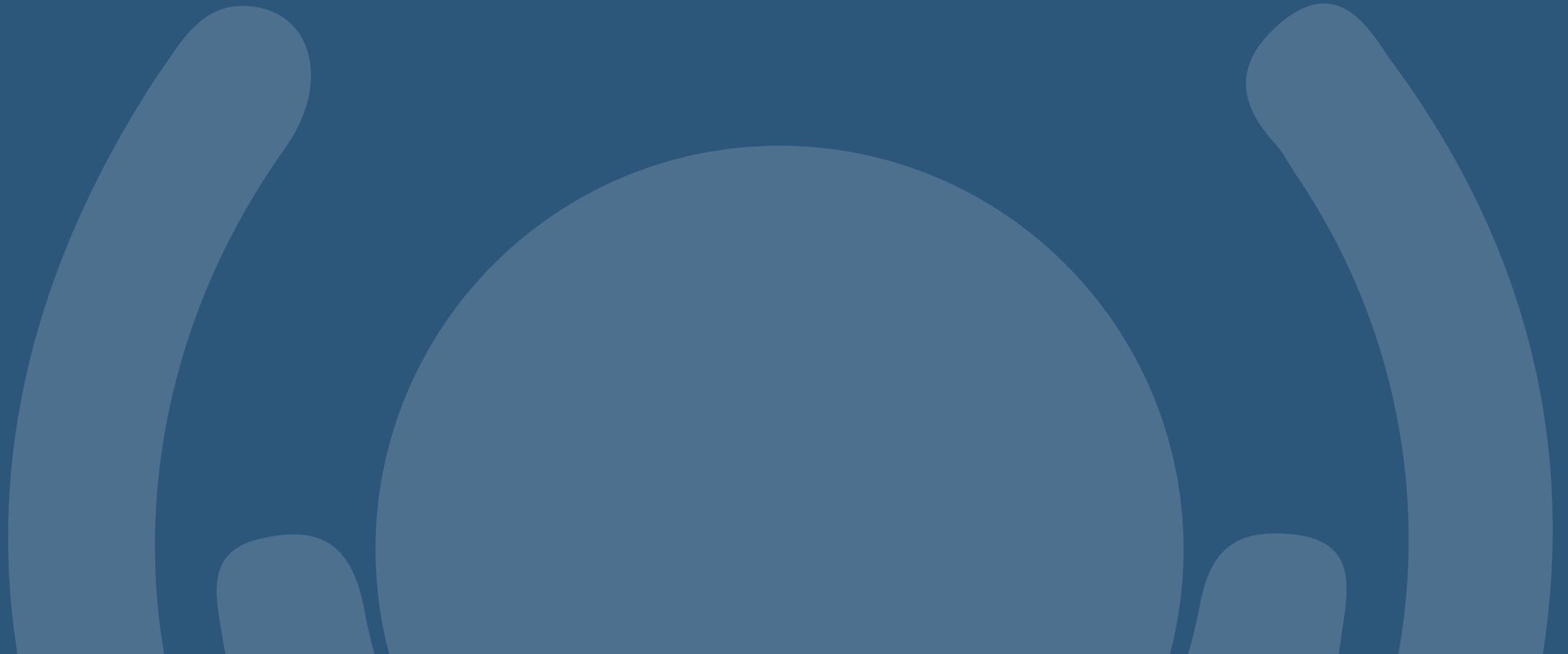
The rate of opening clinical trial results within 3 years is 52%. This is higher than in 2014 (46%) but still low.

The declaration of clinical trials and their results in public databases allows rapid circulation of results, including those that have been unsuccessful and are not published scientifically. There is a very wide disparity between industrial and academic promoters.

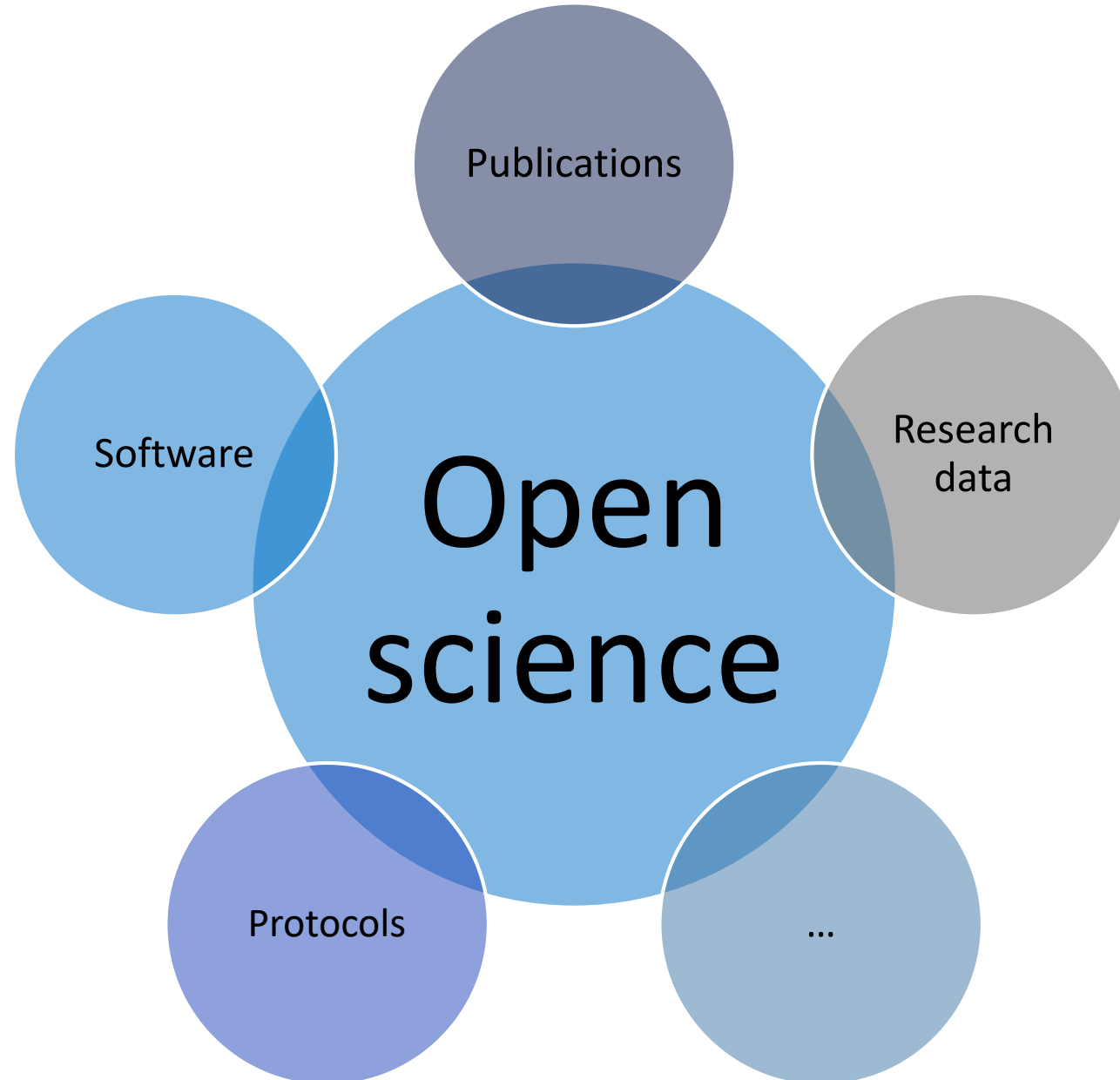
THE RESULTS OF THE LATEST RELEASE: THESES



...TO MONITORING OPEN SCIENCE



WHAT SHOULD WE CONSIDER AS A SCIENTIFIC PRODUCTION?



THE NEW FRENCH OPEN SCIENCE MONITOR: DATASETS AND SOFTWARE

- In 2021, gathering of a threefold and complementary team: French Ministry of Higher Education and Research, Université de Lorraine, Inria



- Winner of a 357k € funding from the European recovery plan



- Goals:
 - producing indicators on datasets and software related to French publications, using AI to detect in full-texts: datasets (with [DataStet](#)) and software (with [Softcite](#))
 - and on data repositories
- Part one of the mission accomplished in 2023, part two currently developed

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

2023/2024

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)

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**

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** (Salvi 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 *PvuII* and *HindIII* (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 -³²P] 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 *NcoI* and *SpeI* 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:

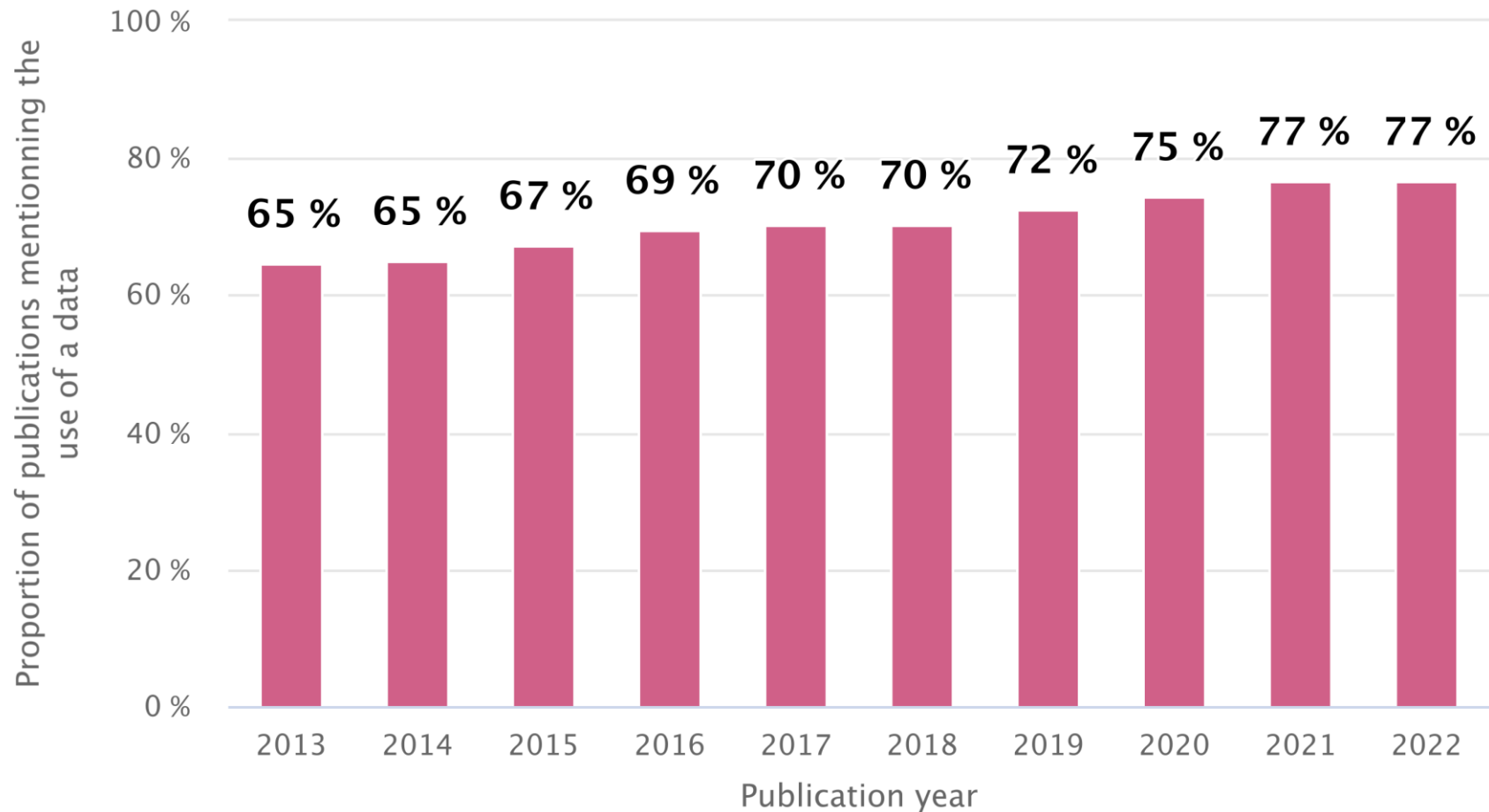
(Zhang, 2008) Zhang (2009) ^

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	http://europepmc.org/articles/pmc2782770
Access	pdf=render
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

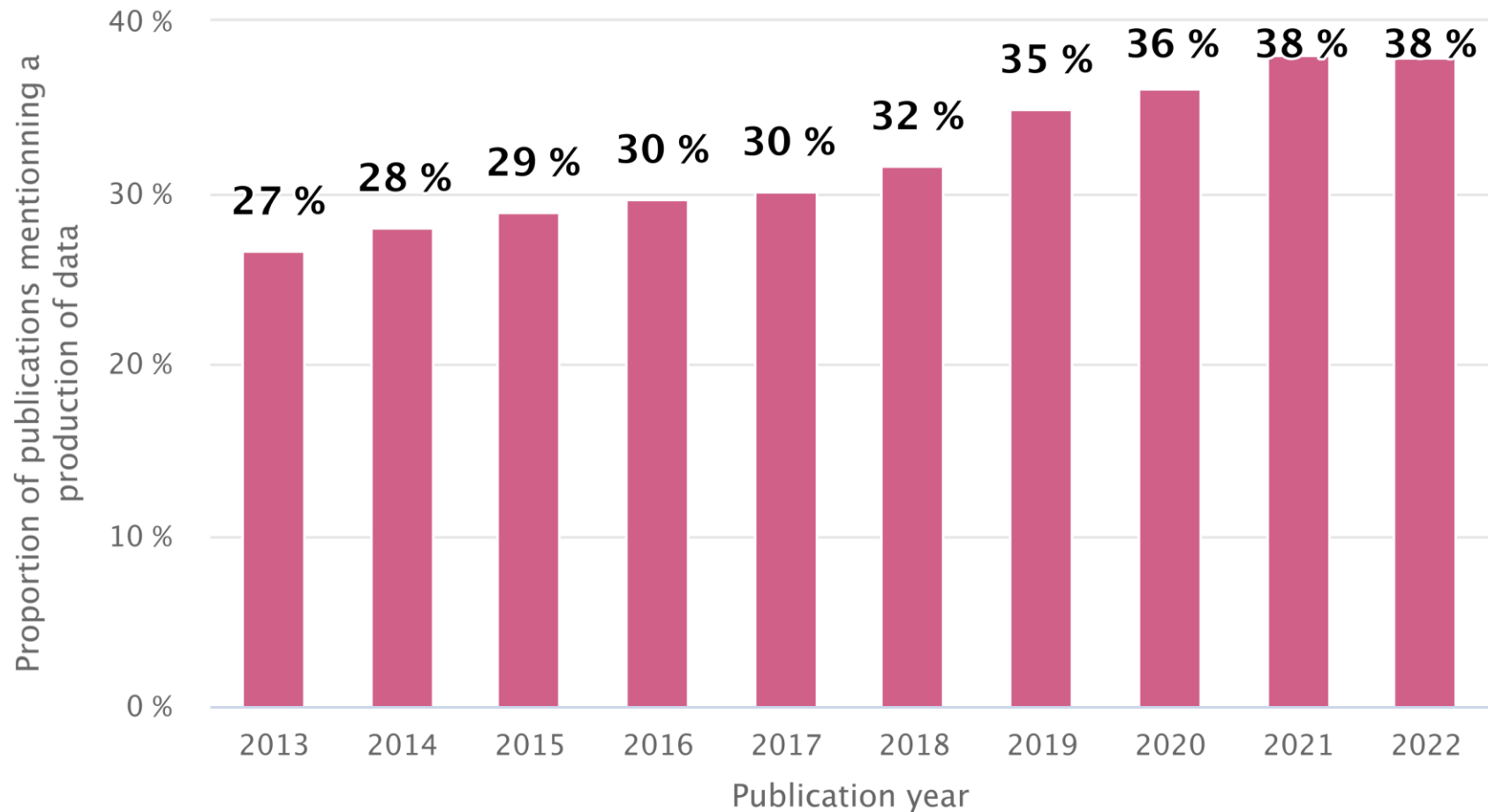
FIRST RESULTS: USING DATASETS

Proportion of publications in France that mention the use of data by publication year



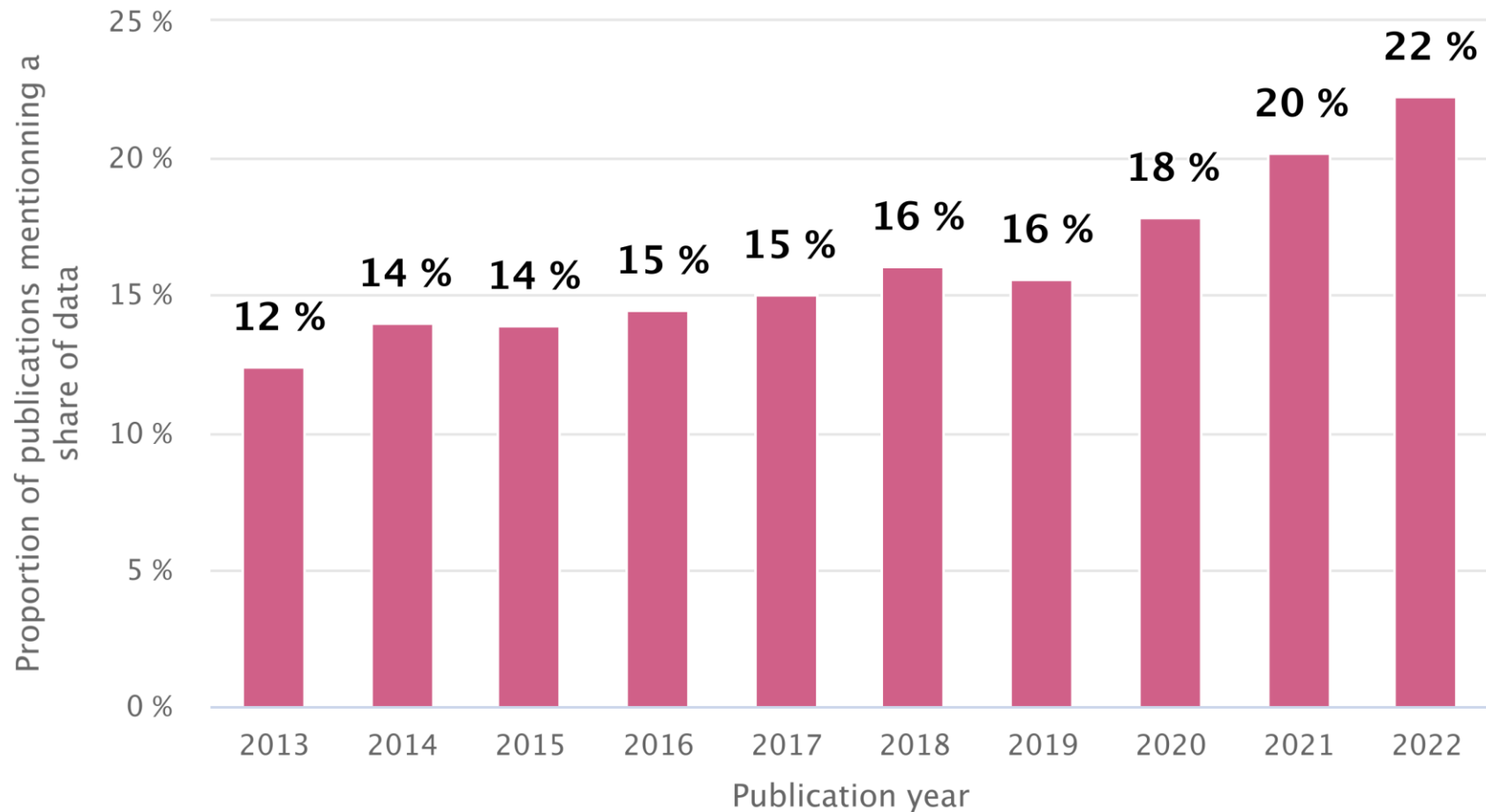
FIRST RESULTS: CREATING DATASETS

Proportion of publications in France that mention having produced their data by publication year



FIRST RESULTS: SHARING DATASETS

Proportion of publications in France that mention sharing a dataset by publication year



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

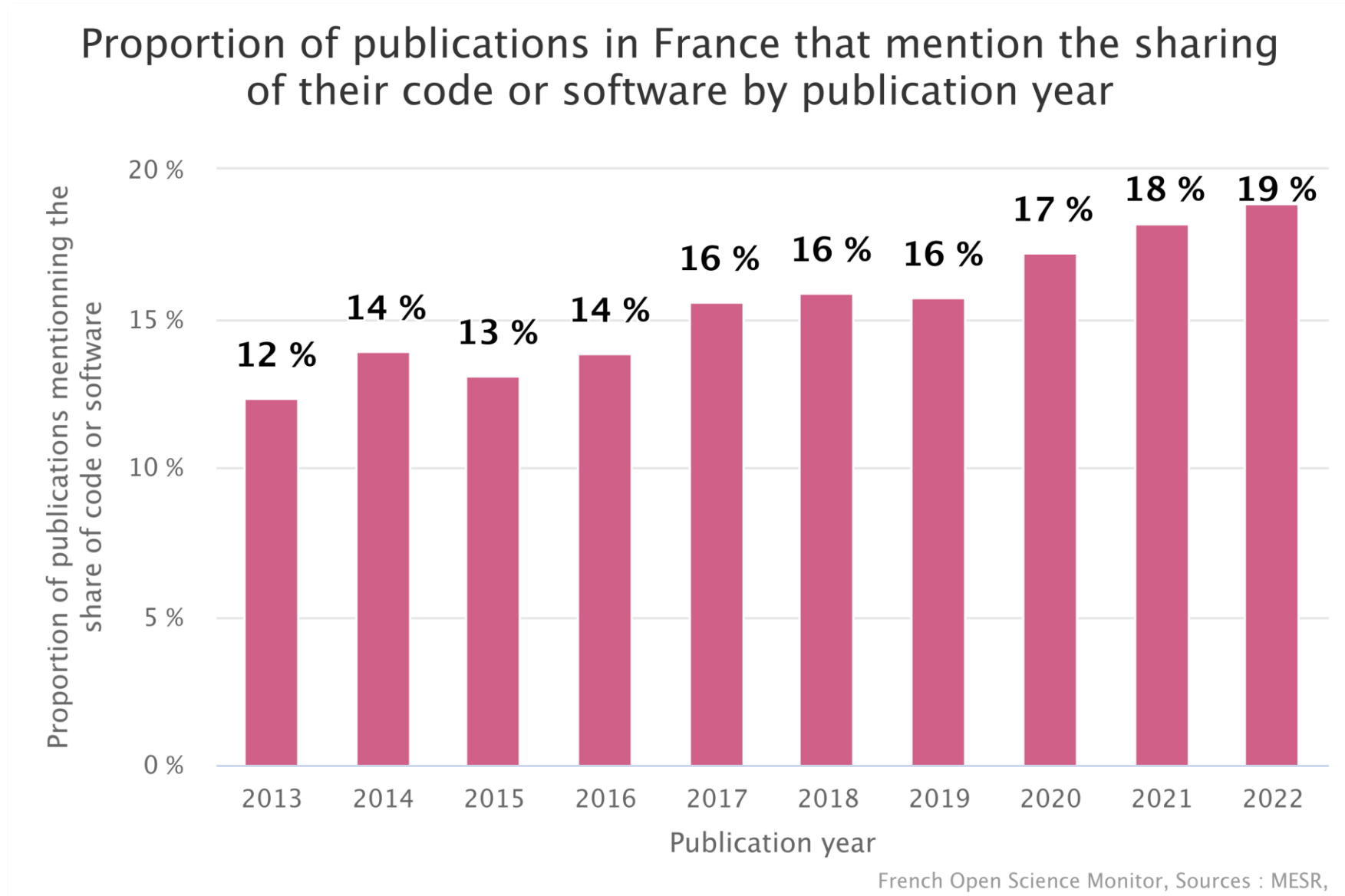
Amongst publications **mentioning the use of data**,

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



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

- More than 70 research performing organisations, universities, grandes écoles and research centres have their own tailored monitor
- A strong local dynamics with an ever-growing community
- More than 250 individuals have subscribed to the Open Science Monitor Users Club



REUSING THE FRENCH OPEN SCIENCE MONITOR OUTPUTS

- All outputs are openly available:
<https://frenchopensciencemonitor.esr.gouv.fr/about/opendata>
- The machine learning algorithms as well:
 - [DataStet](#)
 - [Softcite](#)
- Current collaboration between France and CERN to reuse the OSM code for CERN publications



PERSPECTIVES



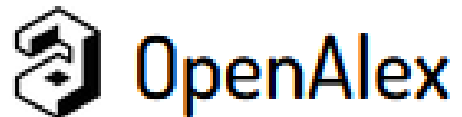
WHAT'S NEXT ?

- [works-magnet](#): a tool developed by the French Ministry of Higher Education and Research to retrieve the scholarly works of an institution (publications and datasets)
- New indicators on datasets for the institutions on the next version of the Monitor
- Partnership with [OpenAlex](#): the works-magnet can also be used to pinpoint and report ROR errors to OpenAlex



Works magnet 

Retrieve the scholarly works of your institution



WHAT'S NEXT ?



Open Science
Monitoring
Initiative

- Throughout the project, many exchanges with other countries
- Many similar initiatives exist, following more or less the same general Open Science guidelines
- But no common understanding on what should be monitored, how, or for whom
- The French Ministry of Higher Education and Research, the Université de Lorraine, Inria and Unesco organized a workshop on the subject in December 2023
- It led to a first draft for [Principles of Open Science Monitoring](#), currently reviewed internationally
- **This is the starting point for OSMI, the [Open Science Monitoring Initiative](#)**





THANK YOU!



LAETITIA.BRACCO@UNIV-LORRAINE.FR



[HTTPS://FRENCHOPENSOURCEMONITOR.ESR.GOUV.FR/](https://frenchopensciencemonitor.esr.gouv.fr/)

CREDITS

Road: Image by [Larisa Koshkina](#) from [Pixabay](#)

Caution: Image by [memyselfaneyeye](#) from [Pixabay](#)

Green statistics: Storyset by Freepik

Green indicators: Storyset by Freepik

Telescope: Everypixel by Arnaud Papa

Thank you: Image by [Ryan McGuire](#) from [Pixabay](#)