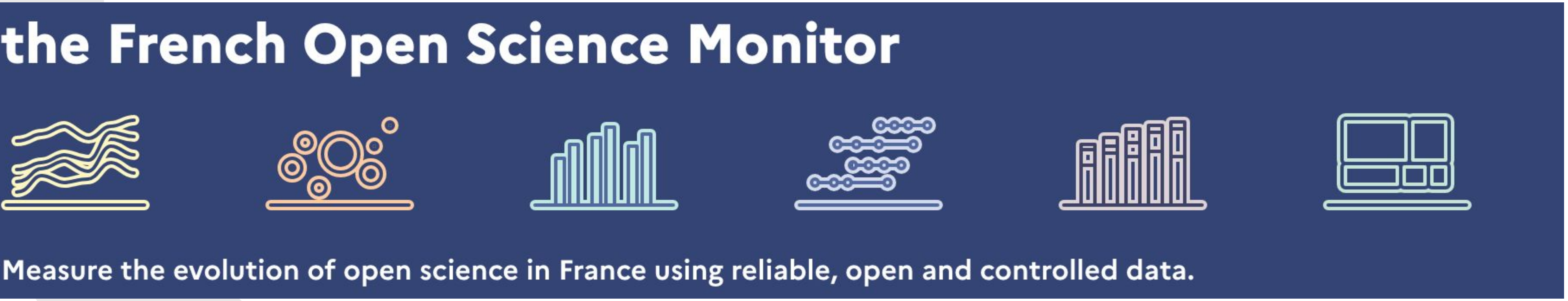
MINISTÈRE
DE L'ENSEIGNEMENT
SUPÉRIEUR
ET DE LA RECHERCHE
Liberté
Égalité
Fraternité

UNIVERSITÉ
DE LORRAINE

Inria

THE FRENCH OPEN SCIENCE MONITOR



the French Open Science Monitor

Measure the evolution of open science in France using reliable, open and controlled data.

EOSC Tripartite event, 2024/09/12

Eric JEANGIRARD, French Ministry of Higher Education and Research

FROM MONITORING OPEN ACCESS …

Since 2018, the French OSM is a monitoring tool for the Open Science public policy

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

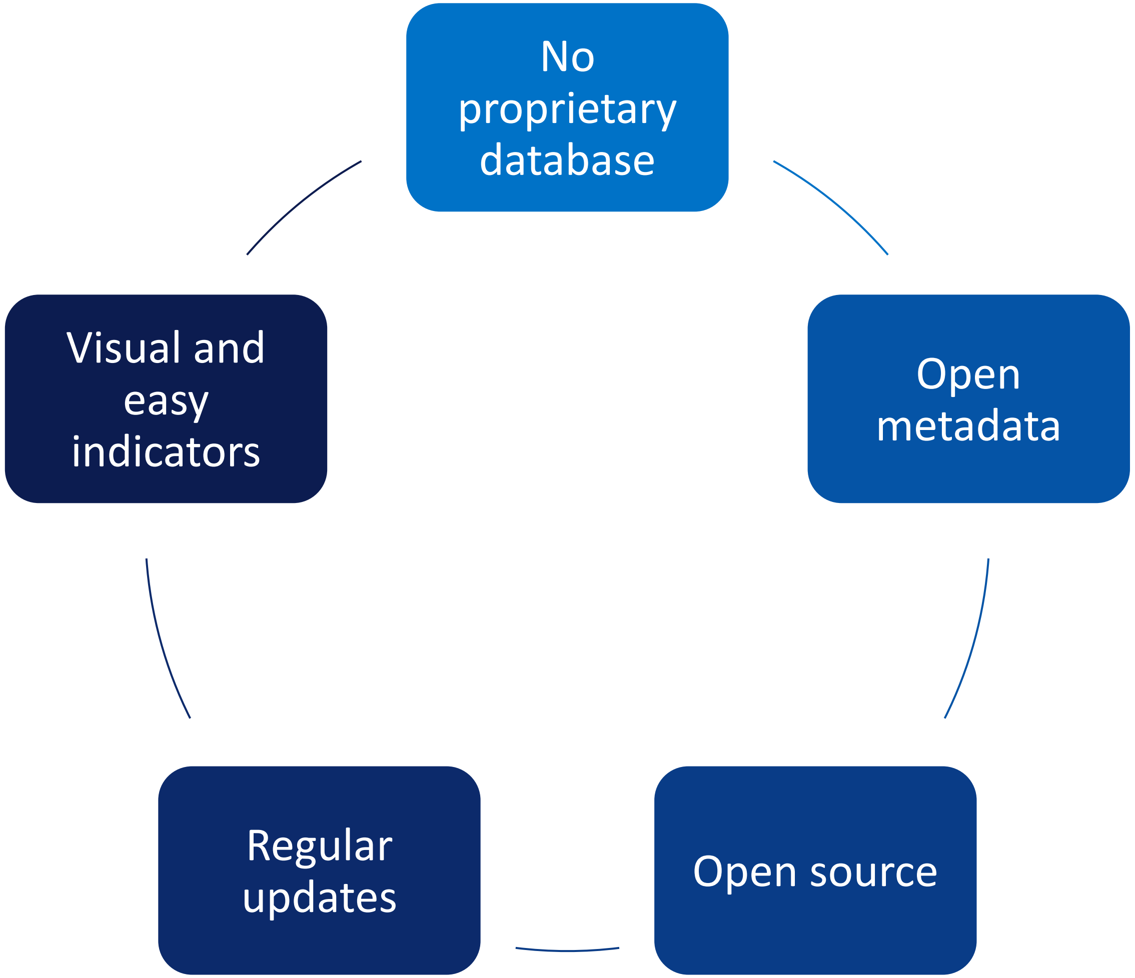
2023

Release of the first indicators on research data and software

Two principles: usefulness to the community and openness of data and processing to create a virtuous circle

- As a monitoring tool, useful for decision makers at different levels
  - national
  - institutions
  - laboratories
  - libraries
  - disciplines
  - …

- **Its openness enables the creation of new services based on the French OSM data**, making the tool even more relevant and useful for the community



No proprietary database

Open metadata

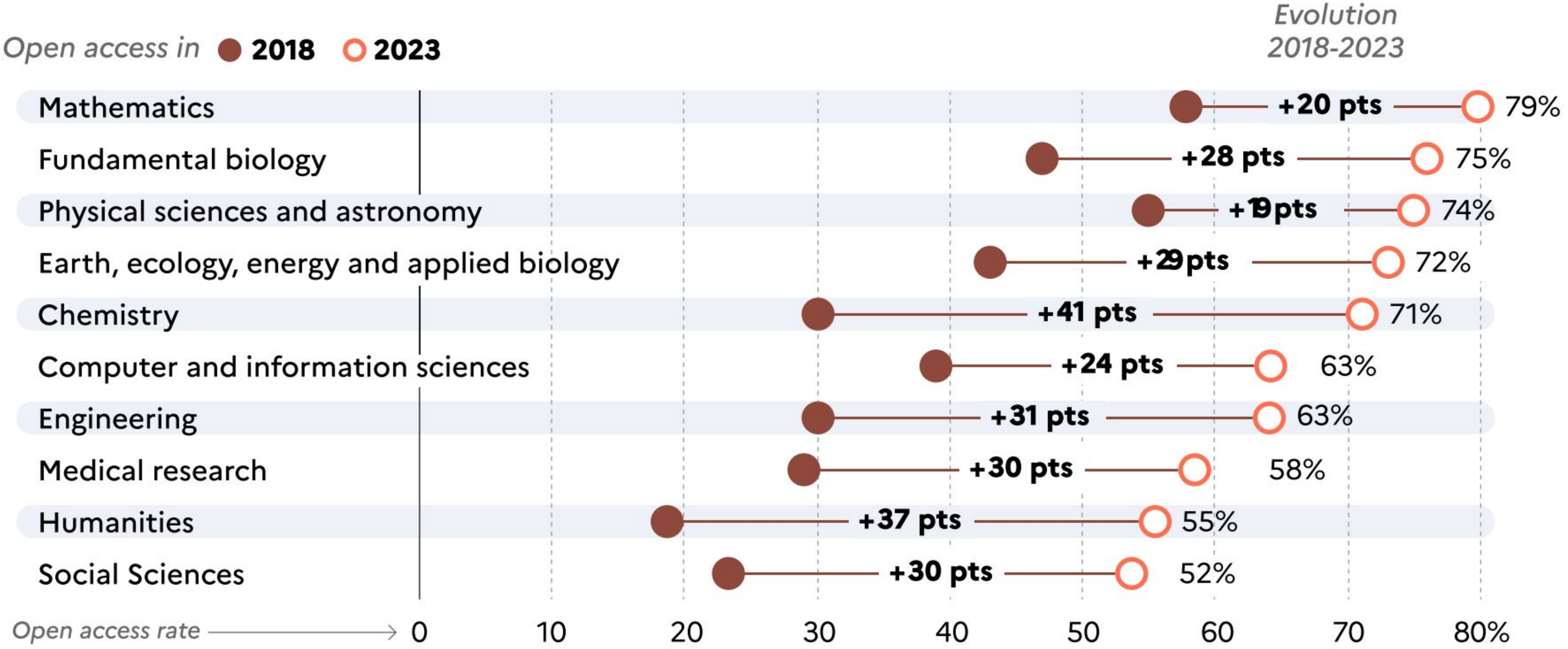
Open source

Regular updates

Visual and easy indicators

Quick overview: the open access rate in France has strongly improved since 2018

Rate of open access publications in France, for each discipline between 2018 and 2023



Changing scale: from national to local monitoring (institution level)

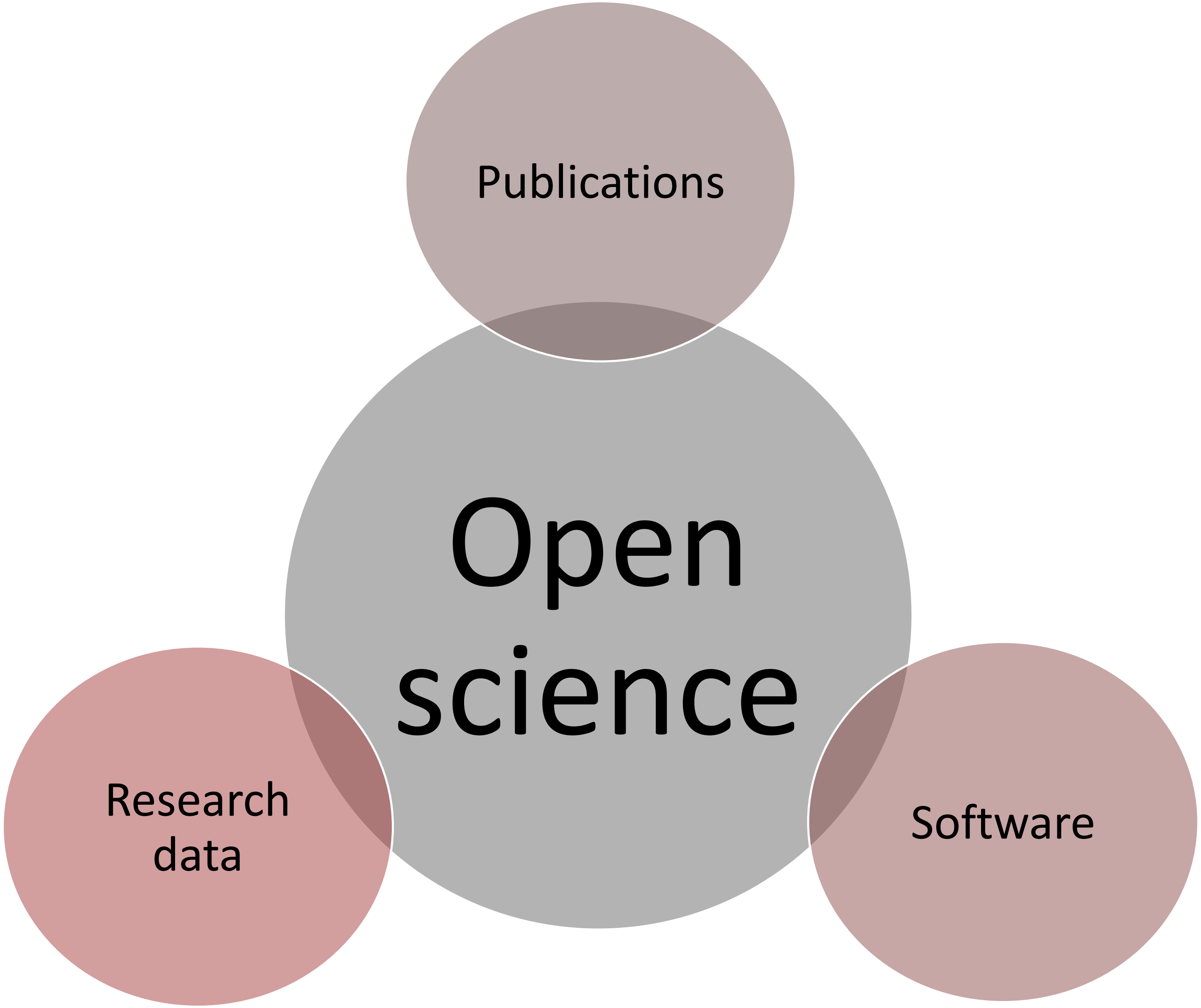
- The national monitoring tool can be derived to monitor local OS progress in any French institutions

- More than 80 French institution use this tool now, including large research organisms and universities, funders, but also labs

- The infrastructure and development is handled at the national level and benefit to the whole French ecosystem, providing open data and open services

…TO MONITORING OPEN SCIENCE

Monitoring the pillars of Open Science?



Publications

Open science

Research data

Software

Monitoring research outputs like software or data is harder than publications

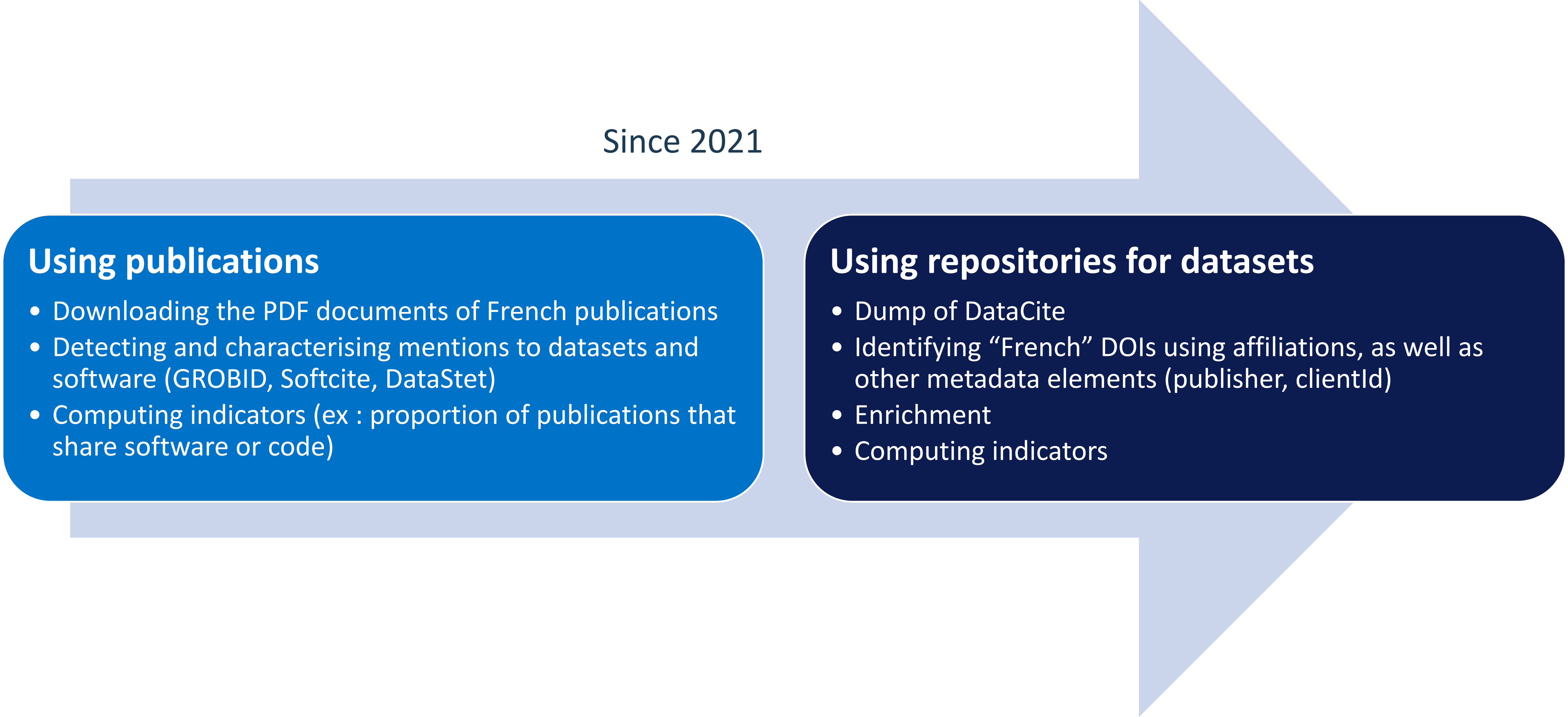
Technical

- No global database for research data and software
- Poor (if any) metadata for monitoring: affiliations, topics…
- Variety of PIDs

Factual

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

Multiples approaches to monitor monitor multiple practices



Since 2021

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 for datasets

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

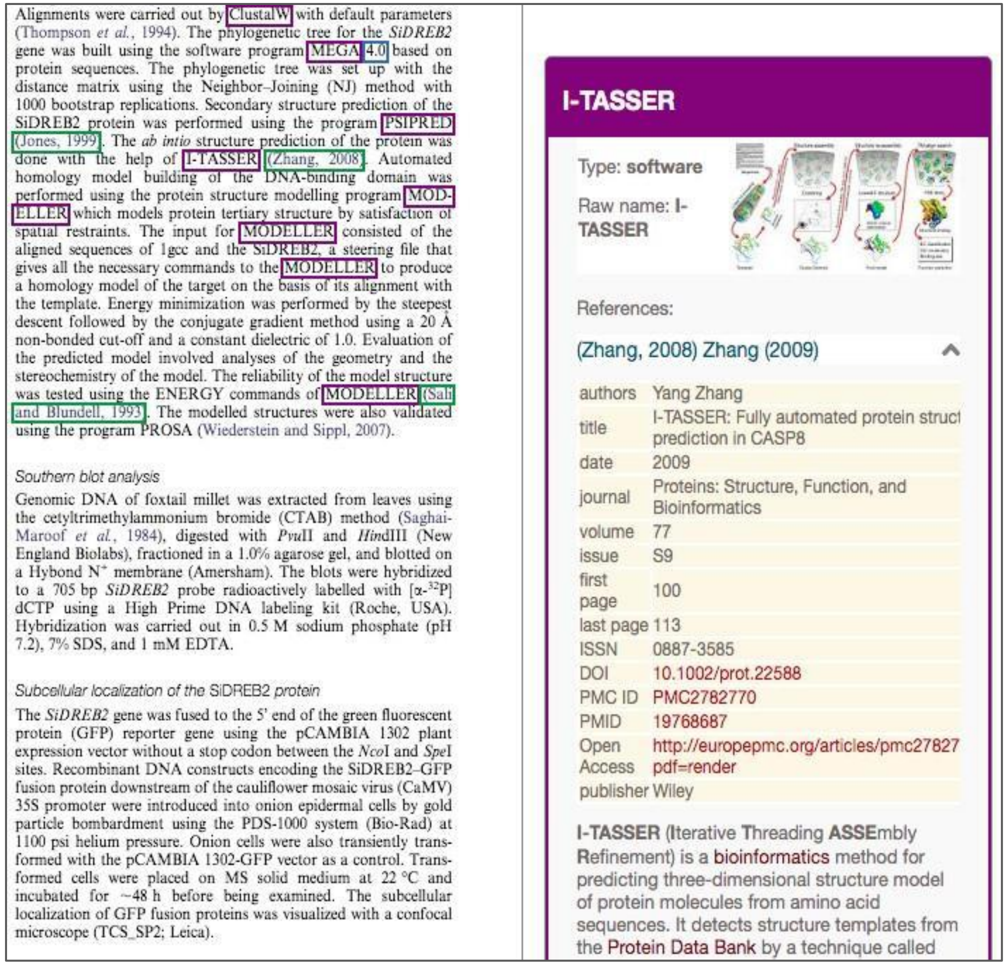
Mining full-text to detect software mentions

- **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:

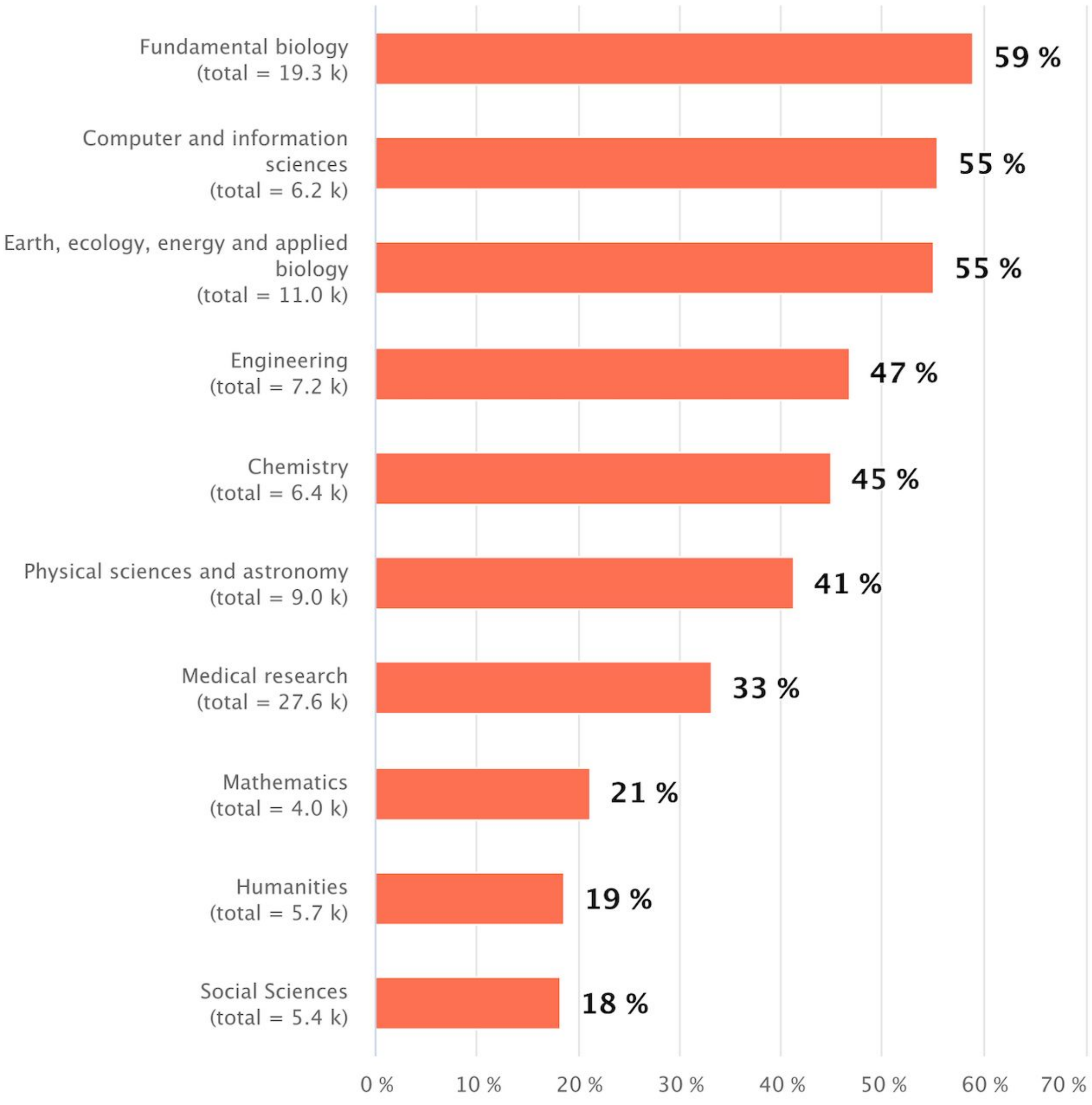
(Zhang, 2008) Zhang (2009)

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 (**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 Protein Data Bank by a technique called

Software is used across all the disciplines

Proportion of publications in France that mention the use of code or software by discipline



- **Almost half** of French publication **mention the use of software**
- **All disciplines are involved**
- Mining PDFs is costly and difficult to scale because of
  - PDF accessibility
  - computation costs

  ⇒ **There is a need for global cooperation on this topic**

International initiatives to scale up

Open Science Monitoring Initiative

- The French OSM is the first national Open Science Monitor to tackle software monitoring using software mention detections on a large scale (more than 1 millions PDFs analyzed)

- Working on a large-scale infrastructure to monitor software use through the scholarly publications remains a challenge. International coordination is key
  - to **build a consensus** on the detection techniques
  - to push for **open source software to be used for these detections** (relying on proprietary tools would create new dependances that we could avoid)
  - to build a large scale infrastructure

- 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

PERSPECTIVES

Upcoming challenges

- **International initiative** on open science monitoring
- Dedicated **infrastructure** to analyse software through publications
- A complementary approach, based on software metadata directly (not the mentions in the publications) is also important to **build software catalog with a high quality of metadata**

THANK YOU!

✉ ERIC.JEANGIRARD@RECHERCHE.GOUV.FR

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

CREDITS

Caution: Image by memyselfaneye from Pixabay
Green statistics: Storyset by Freepik
Telescope: Everypixel by Arnaud Papa
Thank you: Image by Ryan McGuire from Pixabay