MONITORING OPEN SCIENCE BEYOND PUBLICATIONS

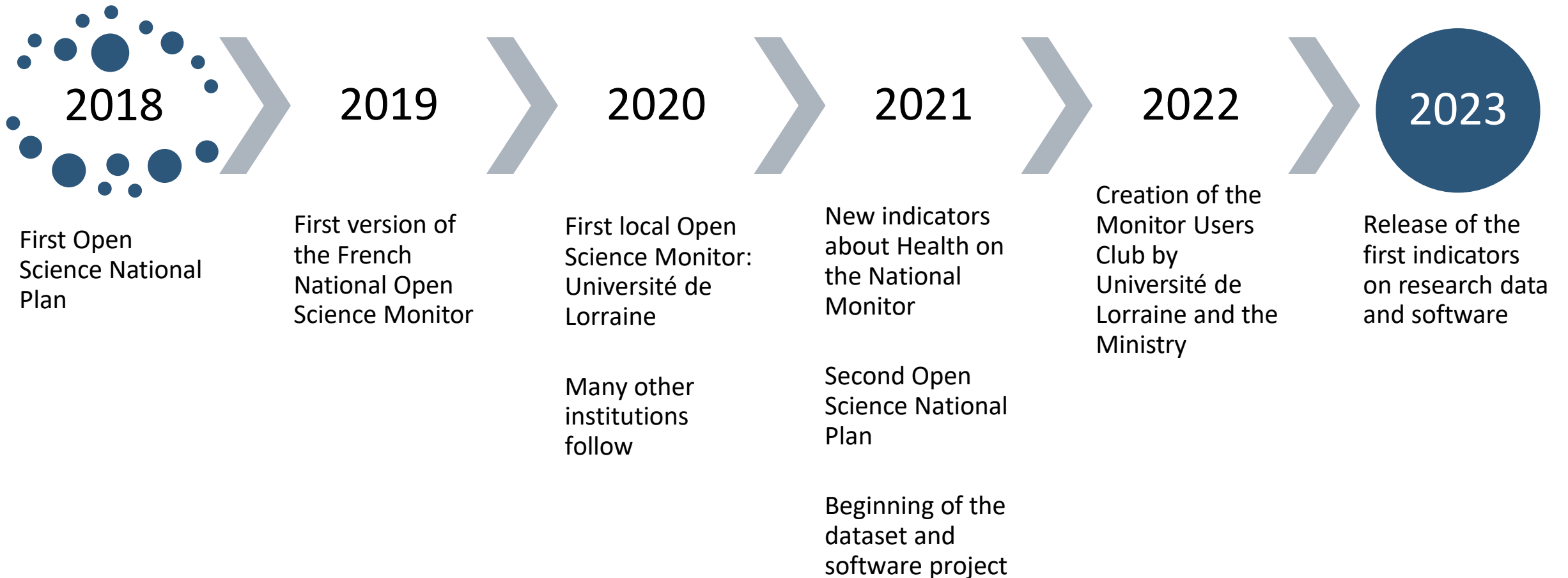**DATASETS AND SOFTWARE AS RESEARCH PRODUCTS TO BE SHARED**

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 dataset 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 where the requirements?
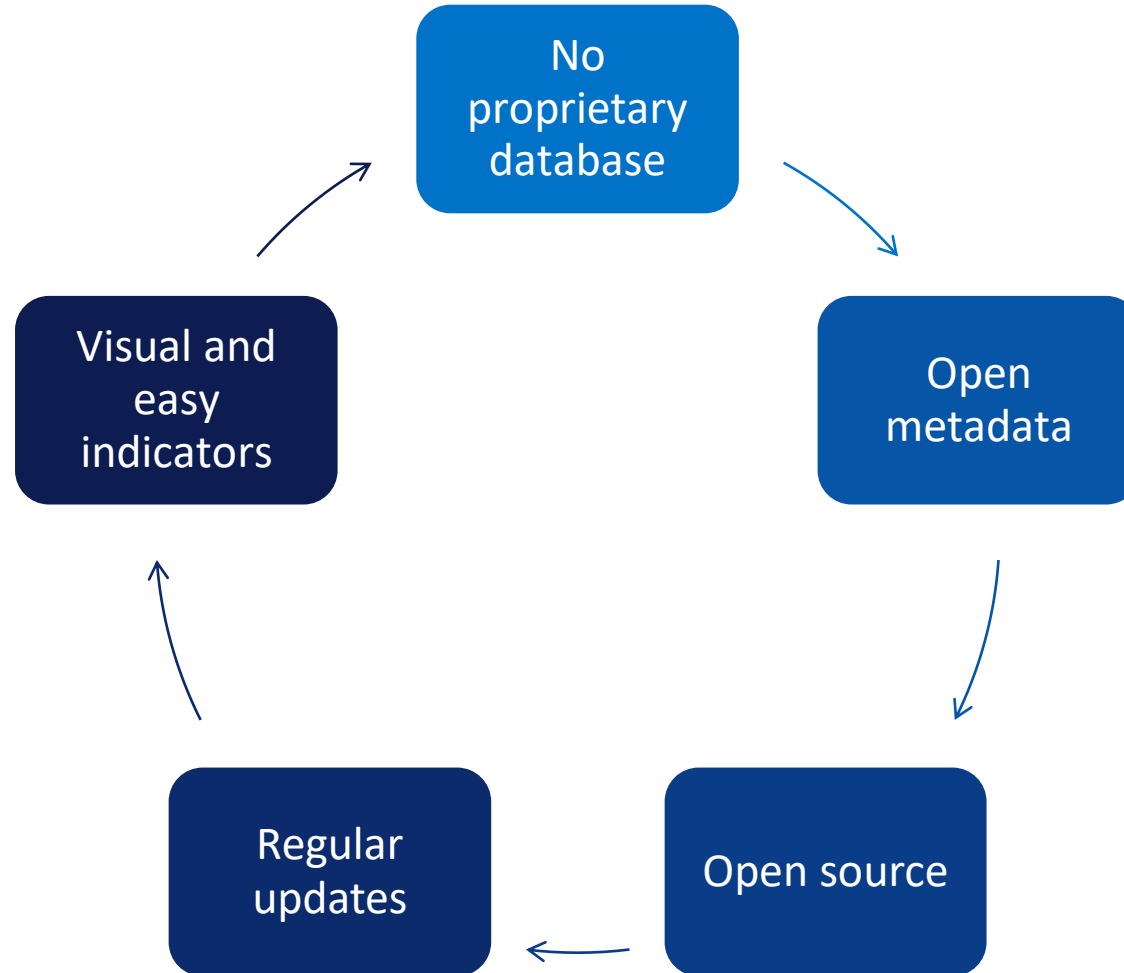
THE BUILDING BLOCK OF THE FRENCH OPEN SCIENCE MONITOR

🏛 **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

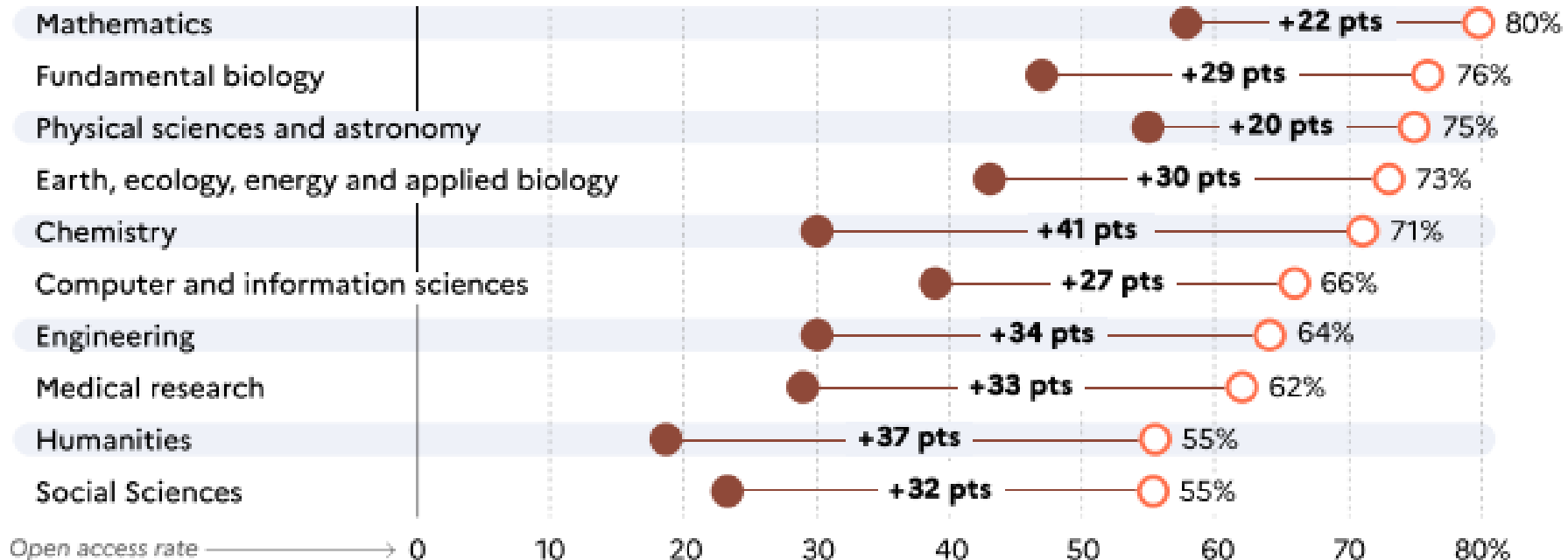| Year	Rate
2022	**67%**
2021	62%
2020	52%
2019	46%
2018	38%
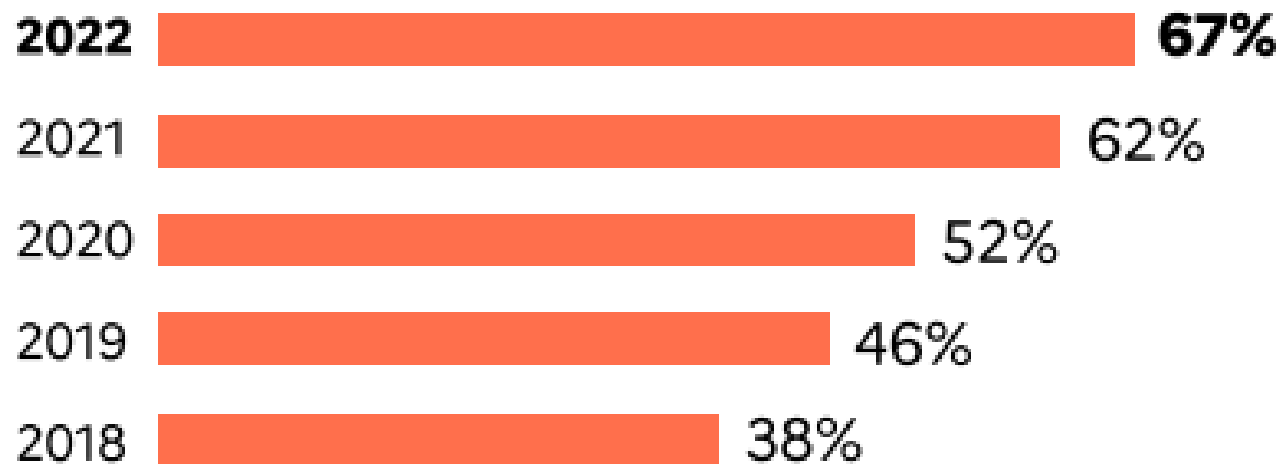
Growth
(all fields)
2018-2022

+29 points

THE RESULTS OF THE LATEST RELEASE: BY DISCIPLINE

Rate of open access publications in France, for each discipline between 2018 and 2022

Open access in ● 2018 ○ 2022

Evolution 2018-2022

Discipline	2018-2022	2022
Mathematics	+22 pts	80%
Fundamental biology	+29 pts	76%
Physical sciences and astronomy	+20 pts	75%
Earth, ecology, energy and applied biology	+30 pts	73%
Chemistry	+41 pts	71%
Computer and information sciences	+27 pts	66%
Engineering	+34 pts	64%
Medical research	+33 pts	62%
Humanities	+37 pts	55%
Social Sciences	+32 pts	55%

Open access rate ⟶ 0   10   20   30   40   50   60   70   80%

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

All types of lead sponsor*:
57%

Industrial lead sponsor:
77%

Academic lead sponsor:
31%

* 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

Opening rate of doctoral theses in France by year
of defense (observational year 2022)

80%

60

40

20

2011   2013   2015   2017   2019   2021*

71%
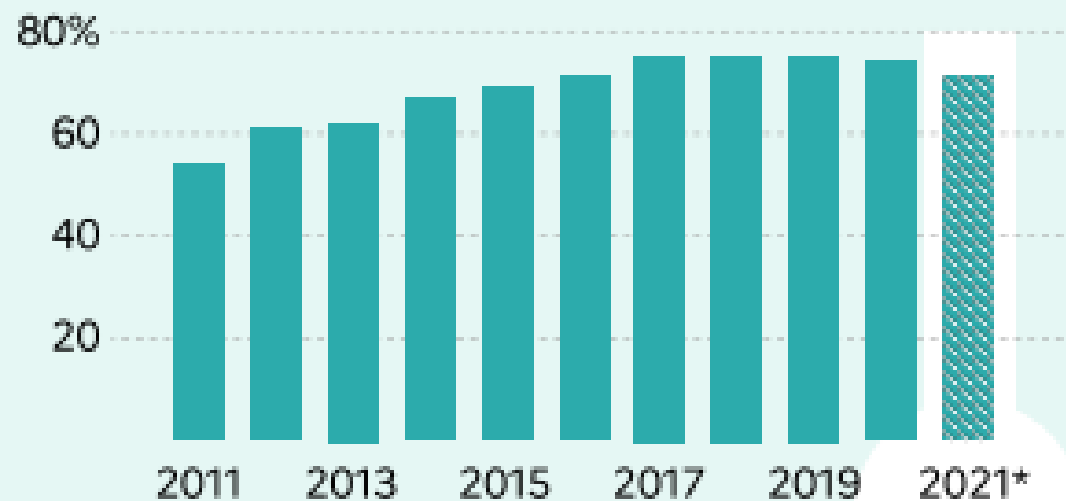
* The slight decline shown for 2021 reflects
a number of theses under ongoing embargo

…TO MONITORING OPEN SCIENCE

WHAT SHOULD WE CONSIDER AS A SCIENTIFIC PRODUCTION?
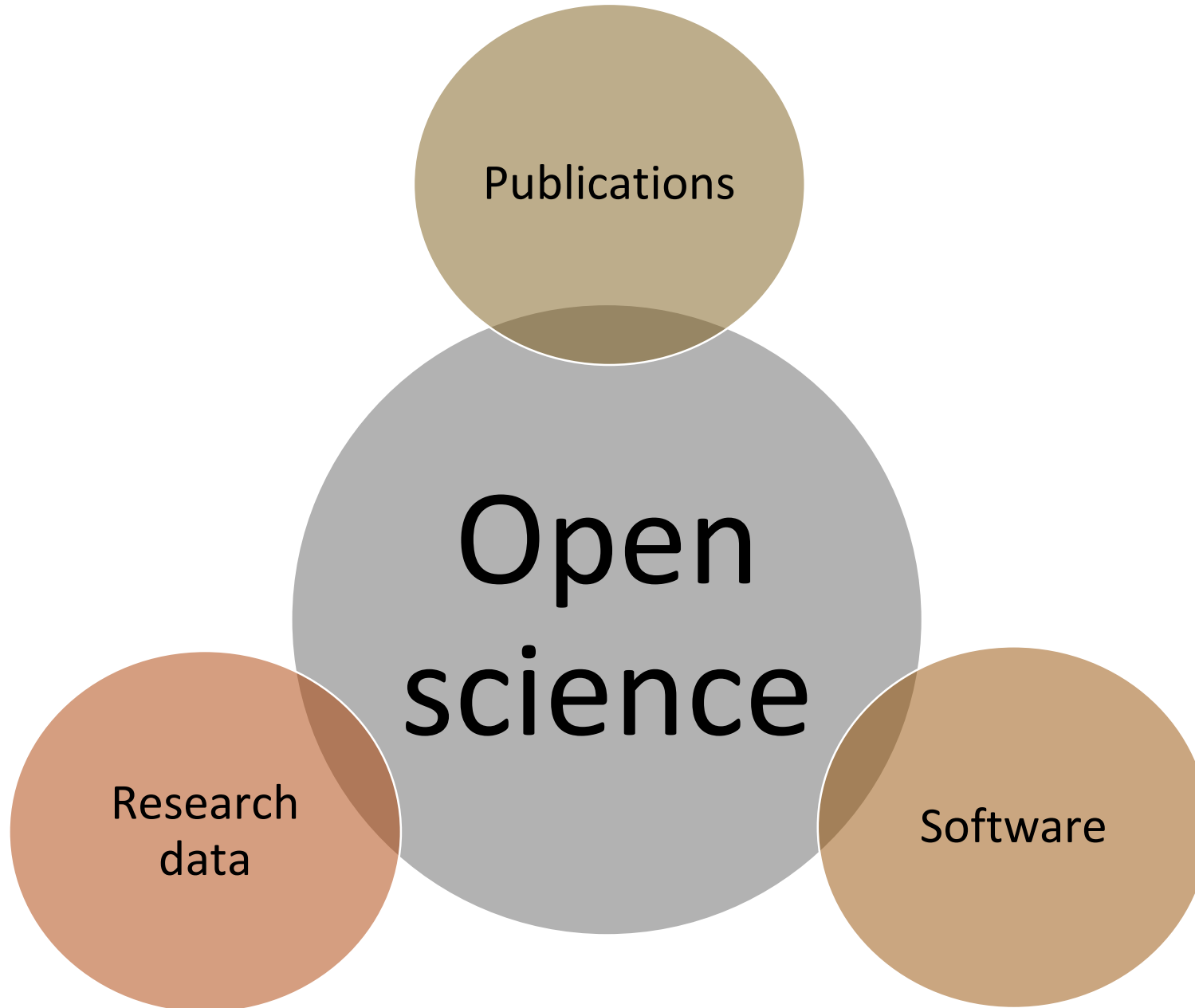
THE NEW FRENCH OPEN SCIENCE MONITOR: DATASETS AND SOFTWARE

- Gathering of a threefold and complementary team:

- Winner of a funding from the European recovery plan:

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

WHAT ARE THE MAIN CHALLENGES?

Technical	Factual
• 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	• 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**

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
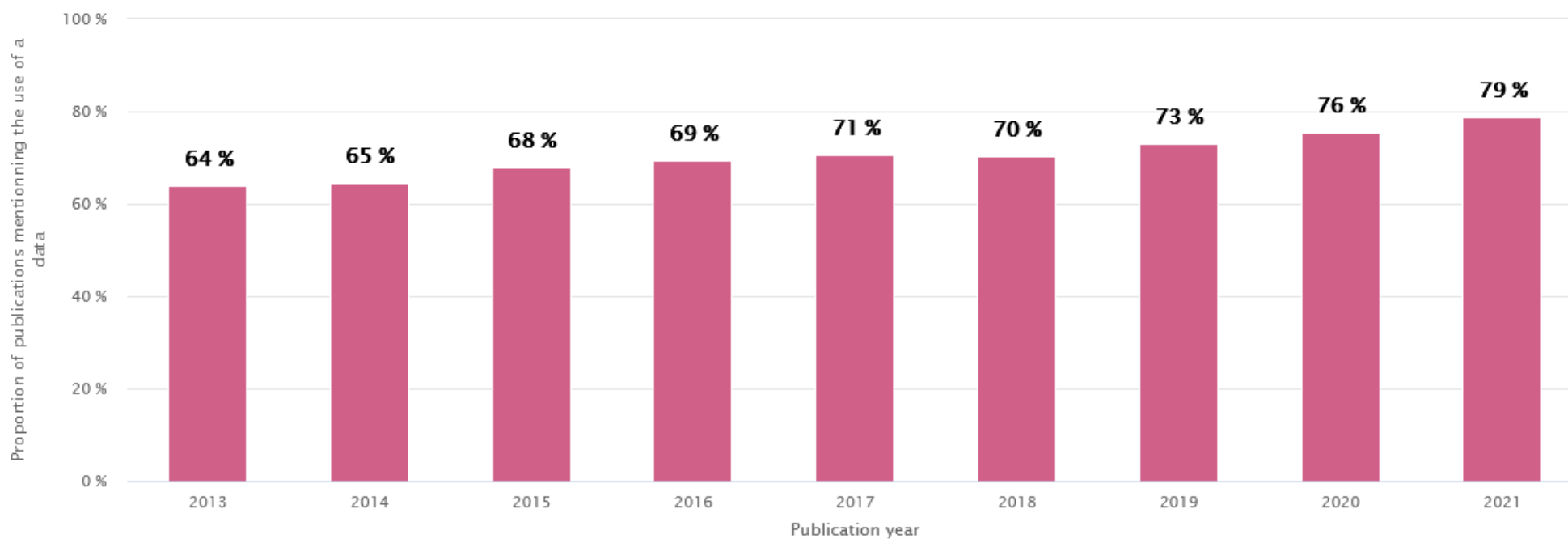
Comment

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

French Open Science Monitor

Comment

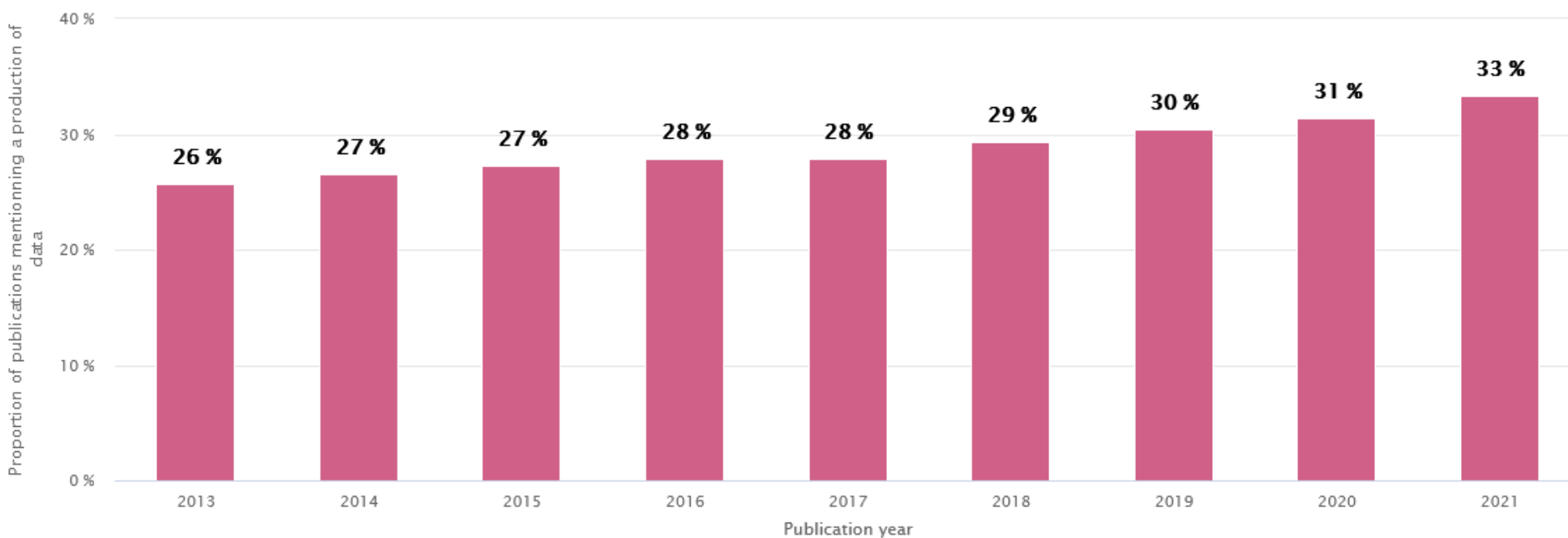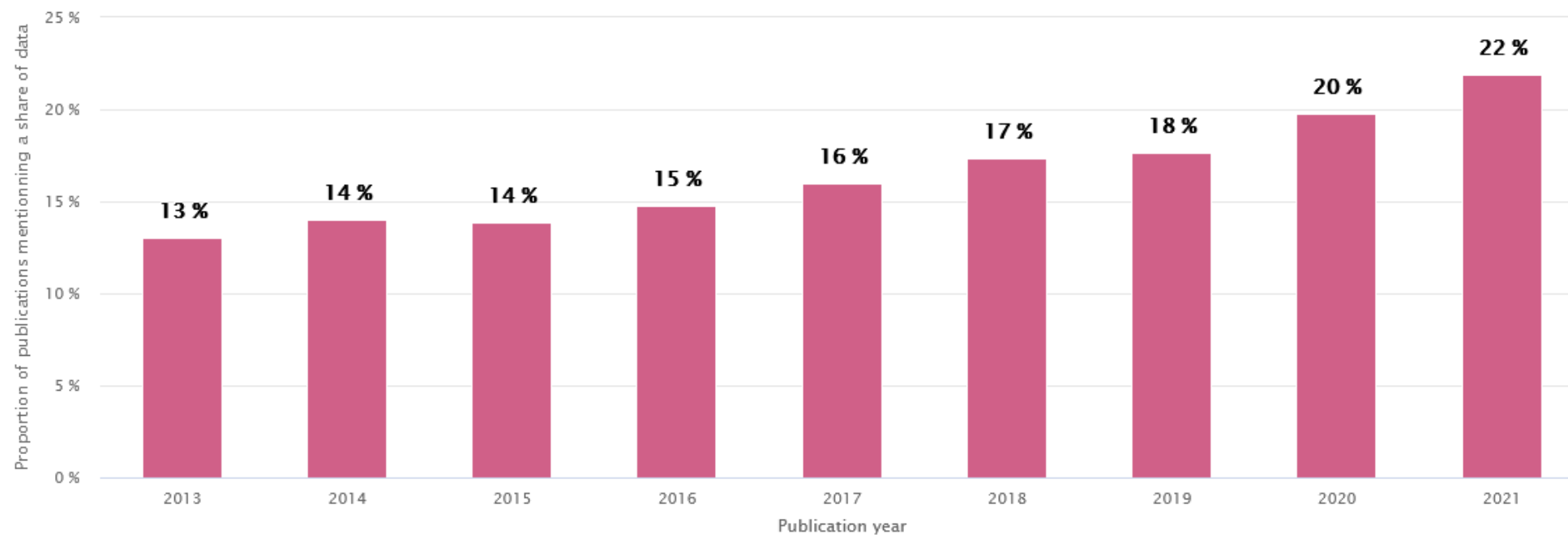This graph shows, by publication year, the proportion of publications for which a mention of data production has been detected, among the publications that use data. This detection is achieved through an automatic analysis of the full text by the DataStet tool.

FIRST RESULTS: SHARING DATASETS

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

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
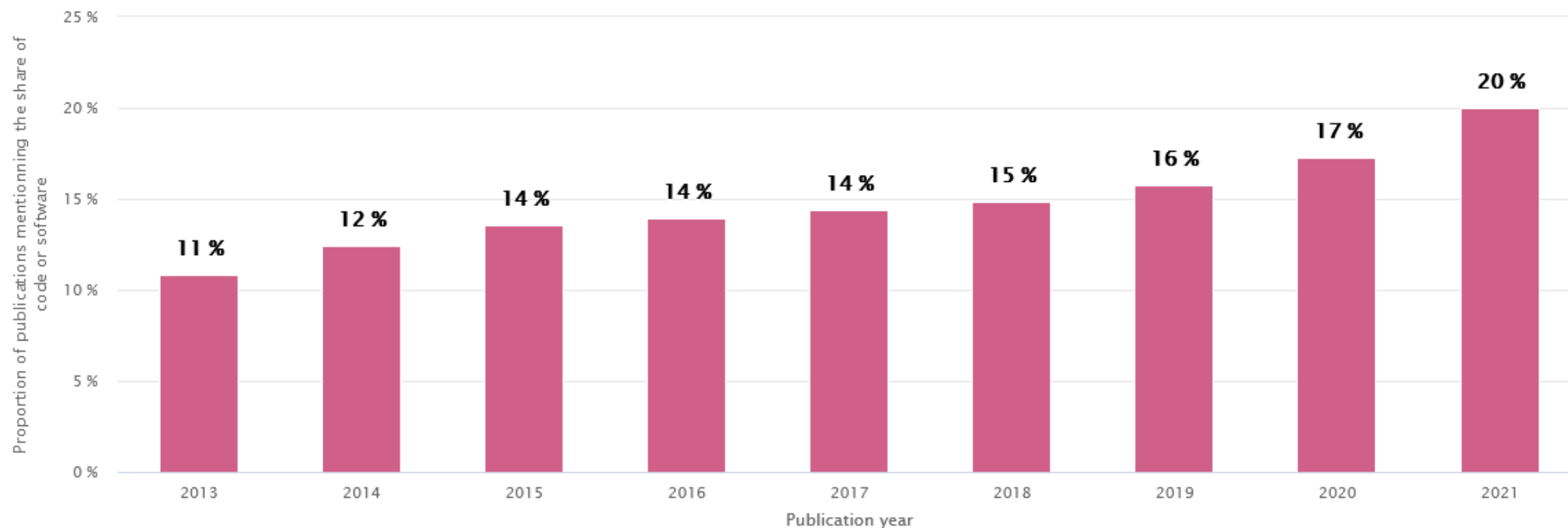
French Open Science Monitor

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