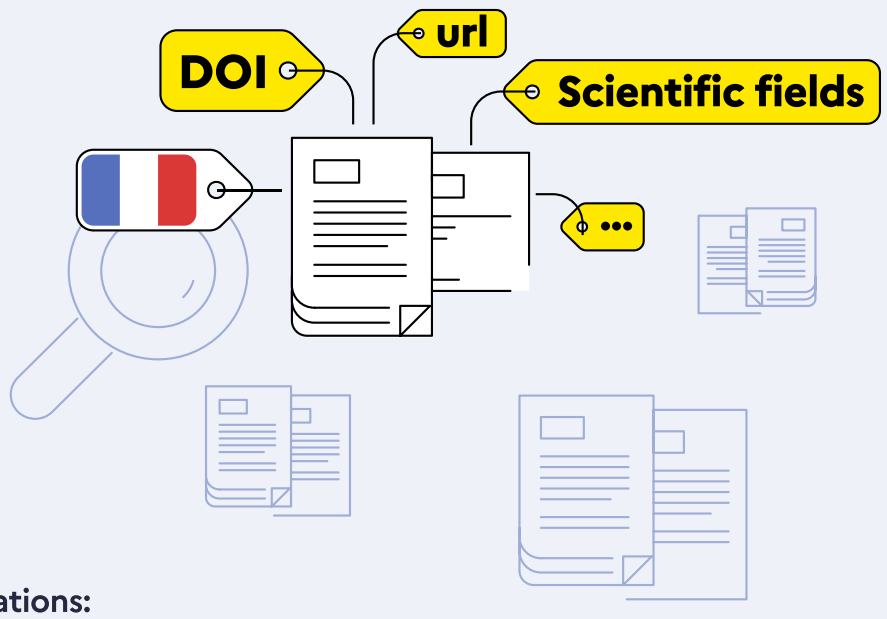
French Open Science Monitor

Data, code and software: An innovative methodology

A generic text analysis process is applied to the publication content for detecting the use, production and openness of datasets and software.



Step one

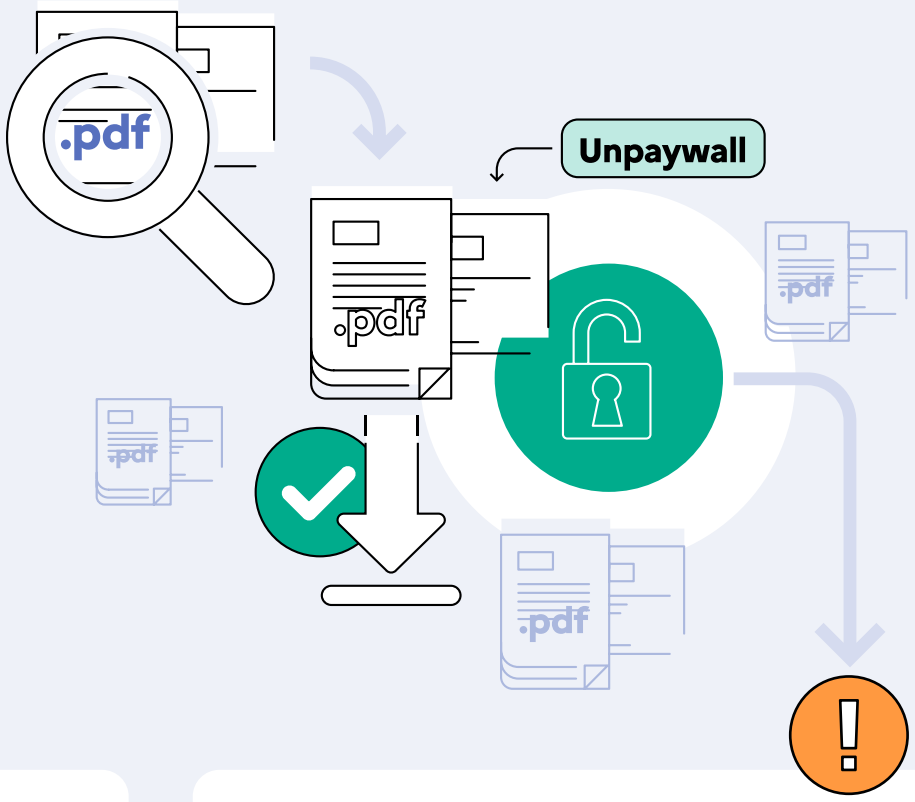
Reference and describe country-wide French research publications

Collecting the open metadata enriched with the Monitor's information (DOI, URL, scientific field classification, affiliations...)

Learn more about the Monitor's methodology regarding publications:

frenchopensciencemonitor.esr.gouv.fr/about/methodology

and the associated preprint hal.science/hal-03651518



Step two

Download the text content of all publications in the corpus

Publications in open access

For each publication, using its DOI, Unpaywall provides a URL that enables to download the full texts into a PDF format.

of publications in open access downloaded by the Monitor

Some technical hindrances encountered: interactive challenge, non-existing page, temporarily unavailable website, broken link, bot-blocking, publication in open access in browser but automatic download unavailable...

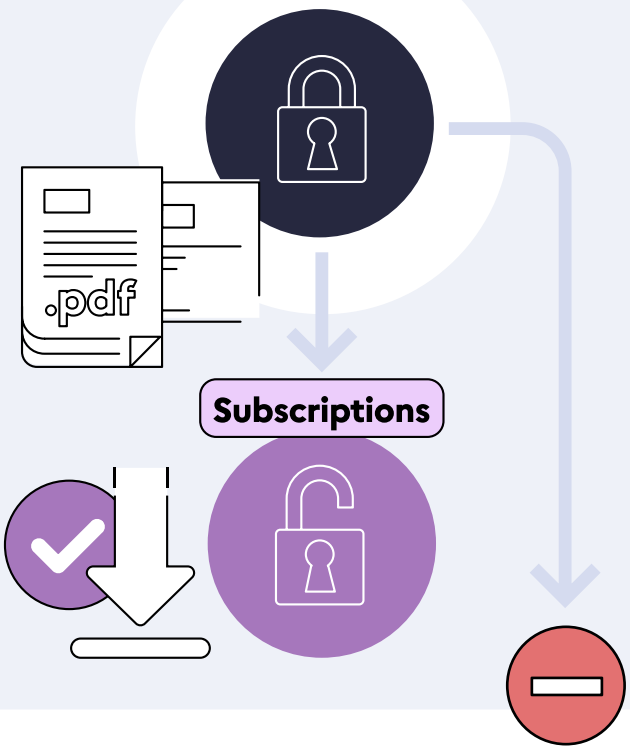
MINISTÈRE
DE L'ENSEIGNEMENT
SUPÉRIEUR
ET DE LA RECHERCHE

Liberté
Égalité
Fraternité

frenchopensciencemonitor.esr.gouv.fr/about/methodology

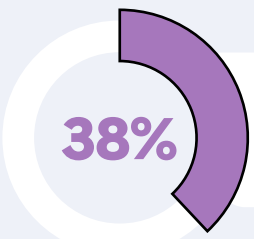
Publications with restricted access

We use the subscriptions of our project partners
(national subscription, University of Lorraine's subscriptions).



Allowed by the transposition of **2019/790* European Union on copyright** and related rights in the Digital Single Market

*https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=celex%3A32019L0790



of closed access publications downloaded by the Monitor

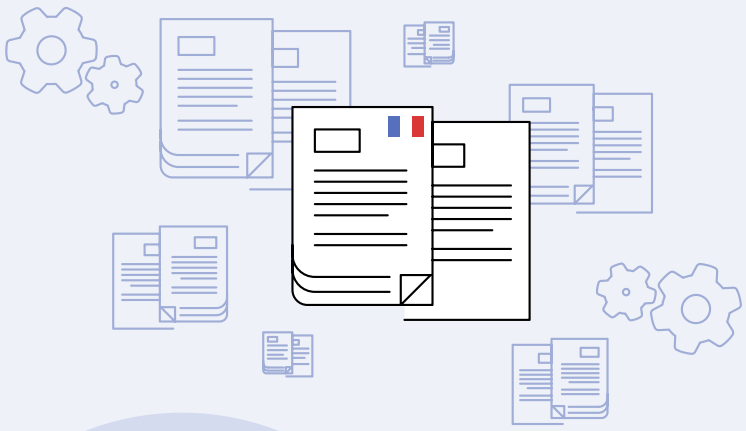
However some publishers with whom we had subscription contracts have introduced mechanisms to control such downloading (API, downloading tokens...).

For publishers with whom we had no subscription, downloading PDF of closed access publications was not possible.

Step three

Enrich through machine learning

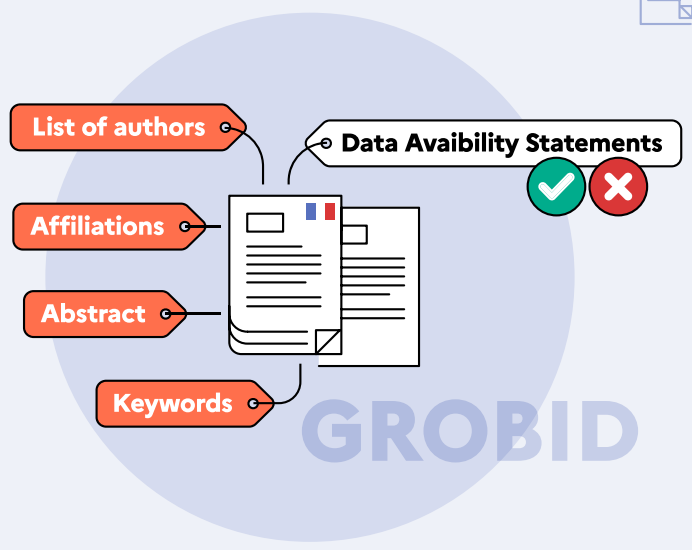
3 tools – including some deep learning
- are applied to these downloaded PDF:



GROBID

The tool extracts metadata and structures PDF into a standard XML-TEI format.

For this project, the capacity to detect "Data Availability Statements" or statements regarding the opening of data has been added to GROBID.



DataStet

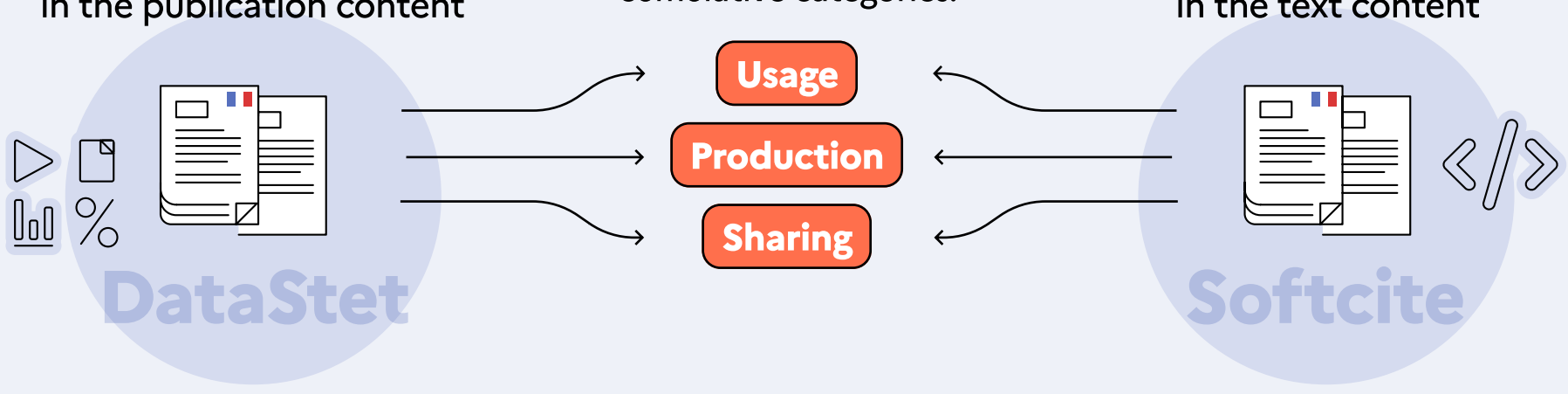
identifies
all mentions of code and software
in the publication content

Each is characterized along three potentially cumulative categories:

Usage
Production
Sharing

Softcite

Identifies
all mentions of code and software
in the text content



Step four

Use these observations and enrichments to create indicators

For datasets as well as code and software, several indicators have been created via **a funnel approach:**

From the DataStet enrichments

Amongst **all publications analysed,**

Share of publications mentioning - in the publication 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

Softcite

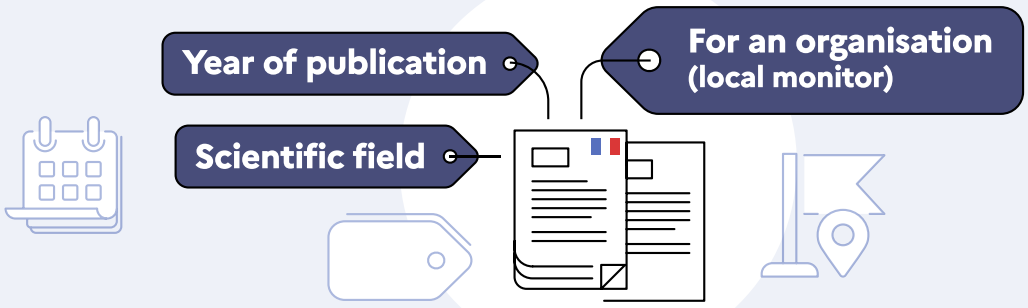
A similar analysis is run for mentions of code and software

GROBID

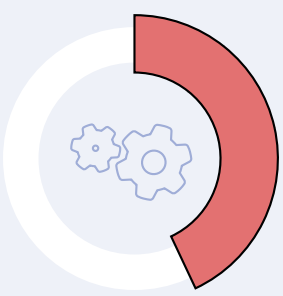
Share of publications that include a "Data Availability Statement"

(statement regarding the opening of data)

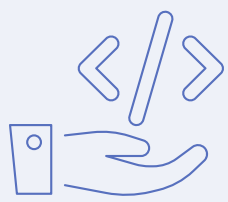
Such overall indicators are rolled out according to the publication facets:



Unreachable publications and processing costs have limited the analysis to 43% of all publications.



So far completed



Methodology published on HAL

hal.science/hal-04121339

This novel methodology procedure was created with:

UNIVERSITÉ DE LORRAINE · Inria · science-miner

With the support of:

FRANCE RELANCE · Funded by the European Union NextGenerationEU