Large-scale Machine-Learning analysis of scientific PDF for monitoring the production and the openness of research data and software in France

Aricia Bassinet, Laetitia Bracco, Anne L'Hôte, Eric Jeangirard, Patrice Lopez, Laurent Romary

Large-scale Machine-Learning analysis of scientific PDF for monitoring the production and the openness of research data and software in France

Aricia Bassinet [1], Laetitia Bracco [1], Anne L'Hôte [2], Eric Jeangirard [2], Patrice Lopez [3], Laurent Romary [4]

1. University of Lorraine, France
2. French Ministry of Higher Education and Research, France
3. science-miner, France
4. Inria, France

Abstract. There is today no standard way to reference research datasets and software in scientific communication. Emerging editorial workflows and supporting infrastructures dedicated to research datasets and software are still poorly adopted in current publishing practices and are highly fragmented. To better follow the production of research datasets and software, we present a text mining method applied to scientific publications at scale and implemented at the French national level. Our approach relies on state-of-the-art Machine Learning and document engineering techniques to ensure reliable accuracy across multiple research areas and document types. The annotations produced by our system are used by the French Open Science Monitor (BSO) platform to follow the production and the openness of research datasets and software, in the context of the second National Plan for Open Science. The source code and the data of the French Open Science Monitor, as well as all the associated tools and training data, are all available under open licences.

Keywords: research data, research software, open access, open science, scientometrics

1 Introduction

1.1 Motivations

Research datasets and software are today core elements of the research activities. 90-95% of researchers in the US and the UK rely upon software, and more than 60% would be unable to continue working if such software stopped functioning (Philippe et al., 2019). Nearly half of the researchers commonly use data generated by other scientists (Science staff, 2011) and the vast majority of researchers support data sharing (Tenopir et al., 2015). The critical role of research data and software is today broadly acknowledged, in particular for improving the reuse, the reproducibility and the transparency of research results (Laurinavichyute et al., 2022).

Following several recommendations such as the San Francisco Declaration on Research Assessment, pro-active policies to enforce higher standards of openness and visibility for all research results have been introduced in the last years. Significant national open science policies are currently implemented. Recently, the U.S. OSTP (White House Office of Science and Technology Policy) launched the Year of Open Science to advance national open science policies across the federal US government in 2023 (House, 2023; Nelson, 2022). This new policy introduces in particular a mandate for free, immediate public access to government-funded research to take effect by the end of 2025, including research datasets, and the development of key performance indicators.

Another example is the French National Plan for Open Science, a long-term, forerunner effort supported by 5 million euros annual budget in 2018-2021 to promote every aspects of Open Science in France. Under this framework, a public dashboard to follow the Openness of scientific publication with key performance indicators has been developed and deployed already since 2019, called the French Open Science Monitor (BSO). This plan was extended in a second phase in 2022-2024 with an increased budget to 15 million euros per year,

giving additional attention to the openness of research datasets and software.

To evaluate, adapt and maximize the adoption of these policies, their effects must be measured. Monitoring tools and dashboards are crucial to follow the evolution of practice regarding openness and sharing of research datasets and software. However, in contrast with the well-established practice of citing scholarly publications, the visibility of these research products is considered largely insufficient and challenging.

(Park et al., 2018) shows that most data citations are informal (mentions without cross-reference with a bibliographical section and without identifiers) and found mostly in footnotes, acknowledgements or supplementary materials sections of research articles. Similarly, software are not cited in scholarly publications in a consistent and easily readable manner (Howison & Bullard, 2016). Multiple initiatives took place in the last decade to address this issue, in particular focusing on improving research datasets and software cataloging (Elger et al., 2016; Garijo et al., 2019), standards for data citation and software citation (Katz et al., 2021; Silvello, 2018) and advocacy efforts. The main advocacy effort regarding dataset and software identification is currently focusing on using PID similar to the successful Crossref DOI for the research publications.

The usage of PID, combined with PID authority agencies and infrastructures, offers indeed a comprehensive technical solution for the identification and citation of any research entities in scholar communication (Cousijn et al., 2019; Juty et al., 2020). However, (Du et al., 2022) failed to observe any usages of PID associated to software when they exist in random samples of software mentions from recent articles on COVID-19. Regarding dataset citations, (He & Han, 2017) report that less than 10% of publications that include data mentions contain any PID. According to the study of (Macgregor et al., 2022), PID associated to data have a much more limited awareness among researchers as compared to publication PID. With their current adoption and impact, PID cannot provide today realistic measurements of the usage, creation and sharing of research data and software.

One immediately available alternative to PID for datasets and software is to rely on a text mining approach to automatically detect their mentions in scholar full texts and their role in the described research work. If the corpus of scientific full text is comprehensive enough, this approach can provide a realistic snapshot of the actual practices regarding research datasets and software at a given time. Such continuous source of data could then lead to trustful indicators for monitoring the impact of open science policies with limited latency.

For extending the French Open Science Monitor (BSO) platform (Bracco et al., 2022) to measure the openness of research datasets and software, we implemented a large text mining pipeline applied to a comprehensive corpus of scientific publications including at least one author with a French affiliation. After describing how we created this corpus and harvested full texts, we will show that recent advances in scientific text mining make fine-grained extraction approaches effective. Our models capture reliably research datasets and software mentions directly from the full text publications, but also characterize these mentions in term of usage, creation and sharing. Finally we will present the indicators developed from these text mining results and how they are integrated in the French Open Science Monitor infrastructure and dashboards, offering at the same time country-wide measurements and custom monitors at the level of an institution or a scientific field.

1.2 The French Open Science Monitor

The French Open Science Monitor[1], also called BSO for *Baromètre de la Science Ouverte*, is a tool for steering the public policy introduced by the first French National Plan for Open Science (MESR, 2018). A first version was released in 2019 providing a collection of measurements and dashboards on the rate of Open Access publications produced by all public French research entities (Jeangirard, 2019). In 2021, the new version of BSO was extended to better encompass the field of health, with additional information on clinical trials and observational studies openness.

[1] https://frenchopensciencemonitor.esr.gouv.fr

Following efforts initiated by the University of Lorraine in 2020, the national monitor also offers tools and learning materials to support the publication by individual institutions of their own Open Science monitor from the national data.

While opening scientific publications is a crucial aspect of Open Science, other research outputs need to be considered. The follow-up second Plan for Open Science (MESR, 2021), which started in 2022, includes a particular focus on research datasets and software, supporting numerous projects and initiatives to broaden their openness. Although the French Open Science Monitor is updated every year (Bracco et al., 2022), measurements related to research datasets and software were not covered yet and require specific indicators.

1.3 Quality criteria for Open Science indicators

Before discussing existing works and describing our approach, we introduce the main quality criteria that we think should be considered when designing open science indicators. We also highlight some of the main challenges associated to these quality criteria when applied to research datasets and research software.

- **Coverage**: Ideally the indicators should cover all the research outputs of interest in the context of Open Science policies. This is very challenging for research datasets and software, because, to a large extend, they are not identified nor indexed as traditional scholar articles. A high coverage is however mandatory to create reliable estimates that are inclusive for all scientific fields and all types of institutions.
- **Accuracy**: Indicator should be reliable, in particular avoiding false positive and duplicates. This criteria supposes thorough and reproducible evaluations in term of standard accuracy metrics (e.g. precision, recall, F1-score) and to rely on reliable authoritative sources.
- **Freshness**: Policy indicators are developed to capture recent changes in publishing practices. The data acquisition underlying these indicators must reduce as much as

possible delays between actual publication dates and measurements.

- **Adaptability to different geographical and organizational levels**: To exploit indicators, we expect that further analyses are possible at lower scale than only national level. Deriving indicators at the level of geographical areas and at the level of individual organizations (universities, research institutes, research laboratories) are requirements for proper study and adaptation of an Open Science policy.
- **Adaptability to different research domains**: Practices vary significantly from one research domain to another. The volume of scientific production is also specific to research areas, and one field could be entirely diluted and invisible within global indicators. Following the evolution of indicators by scientific and technical domains is a key requirement.
- **Fairness**: Indicators should maintain consistency in terms of domains and languages. As much as possible, we want to avoid exclusion of some research areas and languages. This aspect is challenging for example with Social Sciences and Humanities, where publications are more incompletely referenced by large bibliographical index, and frequently published in languages other than English.
- **Understandability and interpretability**: Indicators should present measurements easy to understand for researchers and for the public. For example, if expressed as a percentage, the indicator maximum value (100%) should be clear and correspond directly to a goal of the evaluated public policy.
- **Consistency maintained over time**: Indicators produced for a given year must be directly comparable with the indicators from previous years, in order to follow correctly the evaluation of research activity over time. The consistency should be valid in terms of measurement methodology, corpus, and presentation.
- **Independence and trustfulness for the researchers**: to maximize trust by re-

searchers, public indicators should preferably be independent from proprietary resources and interests. Open source, open data and open access documentation are therefore to be prioritized. In addition, we think methods and data sources for creating the indicators should be fully documented to meet this criteria.

1.4 Existing Open Science indicators for research datasets and software

To our knowledge, there are currently only two examples of deployed Open Science monitors related to research data and software, one part of the OpenAire infrastructure and the PLOS Open Science indicators. We discuss these two examples based on the quality criteria presented in the previous section.

1.4.1 OpenAire

OpenAire is an infrastructure dedicated to open science and mainly funded by EU Horizon 2020 program. Among other services, OpenAire developed the Open Science Observatory[2] to better understand the European open research landscape. The service includes dashboards related to publications, datasets and software.

Relatively to datasets, if we focus on France, this observatory references only 958 datasets affiliated to an organization in the country for the period 2014-2023, with 373 reported as Open Access (as of May 2023). It also reports 16,863 Open Access datasets in the country's repositories (for reference, there are a total of 42M item registered on DataCite). Similarly, 533 software are affiliated to an organization in France, with 111 identified as Open source. As visible, the coverage is extremely limited. Coverage of research domains is similarly incomplete and not consistent.

The method for generating the indicators is currenly not documented, beyond an indication

[2] https://osobservatory.openaire.eu

that it relies on the OpenAire Research graph. This includes a “PID graph” aggregating open metadata databases such as CrossRef, PubMed or arXiv, but we were not able to find additional information about the involved ”intelligent linking”. We assume dataset and software metadata sources mainly correspond to those implementing manual references via PID with sufficient metadata, such as parsed affiliations. As such, we think the OpenAire Observatory illustrates the limit of an approach relaying on PID and on manual referencing for research datasets and software. However, it overall lacks transparency and might be more a proof of concept than a usable service.

Such fragmentary coverage leads indeed to biased and unreliable indicators. Under these conditions, we think that dashboards are counter-productive. Publishing aggregated dashboard on so few and non-representative data can lead to disengagement of the public for the tool, false interpretation, wrong public policy decisions and inability to assess public the application of Open Science policies.

1.4.2 PLOS Open Science indicators

In late 2022, and with an update in March 2023, PLOS published a series of open science indicators (Public Library of Science, 2023). These indicators are evaluated on the 71,109 PLOS research articles published during the 4 year period from January 2019 to December 2022. Three main indicators are provided: the rates of research data sharing per article, the rate of code sharing per article and rate of preprints associated to published article. The PLOS method relies on DataSeer text mining tools (“DataSeer project”, 2019–2023).

As compared to the OpenAire dashboards, the PLOS Open Science indicators correspond to a significantly higher quality considering the criteria we introduced in section 1.3. The corpus is well delimited and the indicators are regularly updated. The text mining process is comprehensive because applied to all the publications. The indicators can be adapted by geographical areas and institutions. The relative ratios over time are easy to interpret

and capture directly the relevant output for PLOS Open Science policy.

In term of scale, diversity and complexity, the PLOS collection is however narrow in comparison with the national ambition of the BSO. These indicators cover only one publisher and are limited mostly to the Life Science domain. In addition, the extraction process can take advantage of the Data Availability statements exceptionally present in all the PLOS publications over the considered period, and of the JATS XML format, which is easier to process than PDF.

Finally, although the text mining tools from DataSeer are open source and documented, their training data and Machine Learning models are currently not open. The trustfulness criteria for the researchers is impacted by the limitation for reproducing the annotations used to derived the indicators.

2 Identification of countrywide research publications

We emphasized in section 1.3 the importance of the coverage and freshness criteria to produced meaningful indicators. The examples of the OpenAire and PLOS indicators illustrate the challenge of addressing these criteria.

In this section, we explain how the corpus of French research publication is created. Identifying in a reliable and comprehensive manner the global set of French research publications is the mandatory basis for developing targeted indicators. A detailed presentation of the approach is given in (Bracco et al., 2022).

In 2018 and 2019, the following two principles have been established when launching the French Open Science Monitor project:

1. The French research publication corpus is defined as all the publications with at least one author with a French affiliation.

2. The French Open Science Monitor is a sovereign tool to steer public policies, it is independent from proprietary database and providers, therefore the project only uses open data for creating the reference set of French research publication.

The first principle supposes to be able to access the affiliation information of every authors for all the publications. However, open bibliographic databases such as Crossref and PubMed lack metadata related to publication country and affiliations. To solve this limitation without relying on proprietary data such as Web of Science (WoS) and Scopus, the French Open Science Monitor has developed an original data ingestion and extraction pipeline running every year (Bracco et al., 2022):

1. Start with all CrossRef DOI for a given publication time range.
2. Collect additional open metadata relative to each DOI, from several sources:
    - harvest at scale and parse every HTML landing pages associated to a DOI to extract additional affiliation raw strings,
    - harvest repositories metadata (PubMed, HAL),
    - when the full-text is available, use GROBID ("GROBID", 2008–2023) to extract affiliations metadata.
3. Use an in-house open source affiliation matcher (L'Hôte & Jeangirard, 2021) to detect affiliation countries from raw affiliation strings.

This approach appears to be highly competitive and reliable when compared to proprietary data such as WoS and Scopus, while relying only on transparent open data sources. Studying French publications in 2019, (Chaignon & Egret, 2022) shows that the open-data method implemented by the French Open Science Monitor (BSO) is able to identify the largest share of French publications as compared to the combined aggregated corpus of Scopus, Web of Science (WoS), HAL archive, the Astrophysics Data System Abstract Service (ADS), PubMed and Microsoft Academic Graph (MAG).

Table 1: Share of the different sources in the overall French publication aggregated corpus (total of 167,412 publications) for the year 2019, as reported by (Chaignon & Egret, 2022).

	Scopus	WoS	HAL	ADS	PubMed	MAG	**BSO**
Share of total (%)	67	58	38	9	29	61	**92**

(Chaignon & Egret, 2022) further indicates that the approach used by the French Open Science Monitor effectively identifies the vast majority of publications with a persistent identifier (DOI) for Open Science monitoring. Table 1 reports the coverage estimated by this study.

3 Full text harvesting

After the creation of the bibliographical set of French publications, the next step is to harvest the largest possible amount of full texts to apply text mining.

In the case of Open Access publications, we identify one or more possible download URL using Unpaywall, which is broadly acknowledged as the most advanced database for this purpose (Else, 2018). Table 2 presents the proportion of Open and Closed Access publication for the complete BSO collection (2013-2021) as of year-end 2022. We also indicate the number for the first considered publication year (2013) and last one (2021) to illustrate the current Open Access dynamic.

Despite the presence of one or several direct full text access URL, around 15% of download fail. Main failure reasons include broken link, IP blocking after a threshold of downloads, or Cloudflare challenges to block machine-based download. The usage of robust web harvesting techniques (rotating HTTP header, support of simple Cloudflare challenges, random delays, bulk download of the full arXiv collection, etc.) only partially mitigates the risk of failed download.

Table 2: Number of publications in the French Open Science Monitor with Open Access (OA) and Closed Access (CA) shares with associated download rates of full texts (PDF), as of year-end 2022.

Publication periods	**2013-2021**	**2013**	**2021**
Overall # publications in BSO	1,426,140	143,095	160,217
- # succesful dowload	908,567	85,914	103,211
- % succesful dowload rate	63.7	60.0	64.4
# OA publications	773,753	61,849	107,722
- # succesful downloads	660,501	52,549	85, 073
- % succesful download rate	85.4	85.0	79.0
Number of CA publications	652,387	81,246	52,495
- # succesful dowmloads	248,066	33,365	18,138
- % succesful dowmload rate	38.0	41.1	34.6

For closed access publications, programmatic download of PDF depends on the nature of the contractual agreement with individual publisher for Text and Data Mining (TDM) application, leading to complex and time-consuming contractual negotiations. Even when such a contractual agreement exists, implementing it can be costly. Most publishers have set procedures and technical constraints to control or even discourage this type of project. Moreover, such a contract may not be renewed, thus not guaranteeing the sustainability of the approach over time for closed publications. Our full text harvesting currently includes the Elsevier and Wiley TDM web services.

Table 2 further summarizes the success rate of the harvesting of full text documents realized in 2022.

4 Machine Learning for mention detection and characterization

As discussed in the section 1.1, PID and metadata driven approach related to research datasets and software cannot lead currently to realistic evaluations and indicators due to

low adoptions and lack of awareness. In contrast, automatic recognition of these mentions from the full texts offers potentially a factual and comprehensive approach, directly usable to estimate quantitatively these research outputs.

4.1 Advantages of text mining publications

We think that mining research datasets and software mentions in scientific publications can provide solutions for most of the quality criteria for indicators introduced in section 1.3:

- In term of **coverage and freshness**, a corpus of scientific publications can offer a trustful snapshot of the scientific production if the text mining is applied to a very significant amount of scientific publications. First the ratio of Open Access publication is today of more than 50% of the whole publications. Second, copyright exception for text mining for subscription-based publications makes possible to harvesting legally a large corpus of closed-access publications. Accessing a comprehensive corpus close to completeness is thus realistic.
- When enough training data is available, the **accuracy** of modern machine learning techniques, in particular based on Deep Learning architecture, has improved significantly. We will see that large scale datasets of manually annotated research datasets and software mentions have been released recently, and we can expect reaching a satisfactory accuracy.;
- The **adaptability** to different geographical and organizational levels and different scientific and technical domains have already been addressed with high reliability at the level of publication in the previous version of the French Open Access Monitor (Bracco et al., 2022).
- With respect to **fairness**, the systematic application of text mining on a comprehensive corpus can cover research domains where awareness of metadata and PID

referencing is very low, because online access to full-texts publications is today a universal practice, avoiding their exclusions.

- **Consistency**: an automated text mining solution can be re-applied to a full corpus regularly, including back files, and produces consistent indicators over any periods. The process is independent from a manual referencing of research datasets and software that could happen to already published articles, certain research fields, from certain publishers and is applied to any types of publications.
- **Independence and trustfulness**: we think that this criteria can be fulfilled if the software used to obtain these indicators are open source, publicly documented and transparently evaluated.

The first text-mining approaches for capturing information on data and software have been rule-based techniques. We discuss in the next section their different limitations.

4.2 Limitation of rule-based tools

In the context of Open Science and bibliometrics, some pattern matching tools have been developed to capture statements about data and software openness in a given corpus. For example, (Larregue et al., 2020) uses some keyword matching applied to the data availability statements and obtained an approximation of the data sharing status for 7,394 COVID-19 articles deposited on medRxiv. However, data sharing statements are, to a very large extend, not limited to data availability statements (Park et al., 2018), COVID-19 articles are not representative of general data sharing practices (Sevryugina & Dicks, 2022) and keyword matching is far from a state of the art automation technique. Such work can be seen as one shot exercise for reporting interesting results in a limited scope, but not as a method that can be generalized and re-used for mining data usage in full texts.

As a more general purpose tool, ODDPub (Riedel et al., 2020) has been developed to detect open data statements in full texts and applied at scale to 2.75 million articles on

PubMed Central XML publications (Serghiou et al., 2021). It implements a rule-based approach for capturing open data statement patterns. ODDPub gives then global information about openness of data and software used in a document. The rules have been developed for biomedical literature. However, we think ODDPub has underlying limitations for the production of data and software Open Science indicators and these limitation can be generalized to the other similar pattern matching tools:

- ODDPub produces information about *openness statements* related to data and software used in the work described in a publication. The tool does not produce more fine-grained information about their re-use, production and sharing. The global volume of novel data and code is unknown. This makes impossible to estimate the key ratio of created data and software openly shared, which is the measurement that Open Science policies need to monitor and maximize.

- Rule-based approaches are useful when no training data is available or when transparency for auditing error is relevant, but they are technically outdated. They perform with lower accuracy and portability to other domains that modern machine learning when quality training data is available (Chiticariu et al., 2010; Trienes et al., 2020). We will see that such quality training data exist today for software and dataset mentions.

- Reported ODDPub F1-score are 0.73 for data openness recognition and 0.64 for code openness recognition at document level. As indicated in (Lafia et al., 2022), performance metrics for the open code detection are unreliable because of a lack of data for significant evaluation (open code in the evaluation data was found in only 11 publications[3]). ODDPub was designed to be part of a workflow including a manual validation step. In this context, accuracy of the automated step can be balanced to ensure an expected quality level. However, for a fully automated text mining at scale, the reported accuracies are more challenging.

[3]It means for example that one additional example of open code will increase or decrease the reported F1-score by around 10 points.

- ODDPub has often no clear distinction between code and data - including in the evaluation data of the tool. Data present on a GitHub repository for example will often be classified as open software.
- ODDPub produces global screening information about openness at document level, but the tool does not have the technical capacity to identify mentions (dataset and software names), nor to provide usable information at software and dataset levels (dataset and software attributes and mention contexts). We think that the level of complexity of the rules required to extract mentions and mention-level characterizations would make the approach very hard and time-consuming to extend to such recognition, if doable.

4.3 Machine Learning for mention detection and characterization

The automatic recognition approach raises crucial challenges that are necessary to tackle and evaluate rigorously for a valid application:

- processing PDF as input for ensuring coverage, freshness and fairness,
- accuracy and sparsity to address accuracy and trustfulness,
- robustness, speed and production-level technical capacities for supporting freshness, consistency and accuracy requirements.

We discuss these three points in the next sections.

4.3.1 Processing PDF as input

To perform text mining on scientific articles, we cannot assume the availability of clean and structured text. The most widespread and easily available scientific publication format is raw PDF, a presentation-oriented format that destroys the semantics and the original structure of data, introducing noise in text encoding and text order stream. This format

raises issues for text mining applications, both in terms of significant source of errors and technical feasibility (Westergaard et al., 2017).[4]

Often presented as an alternative, publisher structured XML including the text body is text-mining friendly, but they have limited availability. First, XML are costly to produce, many small, medium and academic publishers do not have the resources today to produce formats beyond PDF. Second, recent publications are critical for developing Open Science indicators, but the availability of XML is delayed and uncertain as compared to PDF (e.g. for conference proceedings or preprint initially in PDF). Third, accessing possible XML often require additional subscription-based API services even when the Open Source PDF version is available. Last, processing only XML would create a strong dependency on publishers. Even when available, XML full-texts are in a variety of different native publisher XML formats, often incomplete and inconsistent from one to another, involving the development of many custom parsers when considered at scale.

Thus, supporting PDF by using a layout-aware parsing and conversion tool appeared very early as a key requirement for any scalable scientific text mining task task.

4.3.2 Accuracy and sparsity

Accuracy of the recognition of software and dataset mentions from within scholar full texts is challenging first due to the high sparsity of these mentions. This work requires the application of state-of-the-art ML methods to millions of published PDFs across different scientific domains, where dataset and software mentions represent only a few relevant tokens out of several thousands in every document.

Considering the Softcite dataset version 1.0 (Du et al., 2021), the 4,971 full-texts contain a total of around 46 million tokens, but only 15,280 tokens are relevant to a software mention.

[4] From one of the author of (Westergaard et al., 2017), an effort to apply text mining to 15 million scientific articles, “We probably spent more computational resources teasing the text out of PDFs and beating it into shape than we spent on the actual text mining.” (McKenzie, 2017)

Around one token is positively labeled for each 3,000 “negative” tokens, with a ratio as low as one token per 17,500 tokens for *publishers* and *URL* fields. An Imbalance Ratio value above 500 is usually already considered to be extreme (Lee & Deleris, 2020). With the higher observed Imbalance Ratio here, from 1:3,000 to 1:17,500, an ML approach to finding new unseen software mentions is very challenging.

4.3.3 Robustness, speed and production-level capacities

Given the technical environment of the French Ministry of Higher Education and Research, which is operating the French Open Science Monitor, and the volume of full texts downloaded reported in section 3, the following technical constraints need to be covered:

- scaling to at least 1 million PDF in manageable time and standard computing infrastructure,
- secure storage for harvested full texts,
- repeatable processing every year for the complete corpus,
- high robustness to avoid service failure, in particular in the common case of ill-formed and corrupted PDF,
- fully automated process for deployment on a Kubernetes cloud.

4.4 Research software

Automatic recognition of software mentions has attracted a lot of interest in parallel with the development of research software citation advocacy in the last decade. (Krüger & Schindler, 2020) presents these first approaches, based mainly on gazetteers and rules. Machine Learning promises significantly higher accuracy and coverage, but appeared however first limited by the lack of manually annotated training data necessary for reliable models - the largest public dataset until 2020 being limited to only 85 annotated documents

(Duck et al., 2015). With the development of large annotated gold corpus, the Softcite dataset (Du et al., 2021), 4,971 articles in Life Sciences and Economics (4,093 software mentions), and the SoMeSci dataset (Schindler et al., 2021), 1,367 articles in Life Sciences (3,756 software mentions), Deep Learning approaches have been recently possible for this task.

4.4.1 Prior work

To our knowledge, three systems have used modern Deep Learning techniques and the recent large annotated gold corpus for software mention recognition. (Lopez et al., 2021), (David et al., 2022) and (Istrate et al., 2022). All are using fine-tuned SciBERT models (Beltagy et al., 2019).

(Lopez et al., 2021) is the earliest published system and was developed in parallel with the creation of the Softcite dataset. The system includes a variety of models trained with the Softcite dataset. The training uses sampling techniques to increase the number of negative examples for mitigating the problem of mention sparsity. The best model is a SciBERT model, with a CRF activation layer, fine-tuned with the positive examples of the Softcite dataset and additional negative examples selected with a technique called active sampling. The training set uses a 1:20 ratio (20 paragraphs without annotations for one paragraph with at least one annotation). With the Softcite dataset, it corresponds to around 2000 positive paragraphs and 40K negative paragraphs. The model has been evaluated on full paper content for 20% of the annotated articles. Additional entity disambiguation is used to filter out false positives and document-level mention propagation is used to increase recall. Input of the service can be PDF or various publisher XML formats.

(David et al., 2022) is trained with the SoMeSci dataset (Schindler et al., 2021), which is smaller and limited to Life Sciences, but has a more comprehensive set of annotations than (Du et al., 2021), including relationships between mentioned software. The model uses a

more complex architecture to also predict these relationships. Training is realized with the annotated corpus which include only the positive sentences (sentence with at least one software mention) for 787 PLoS articles, method sections for 480 PLoS articles (a total of around 3,200 paragraphs) and complete content for 100 articles. The distribution of software mentions is thus significantly oversampled as compared to the actual distribution. The system supports JATS XML documents as input.

(Istrate et al., 2022) is the latest published system, but also the simplest, with a more limited scope. It is a SciBERT model fine-tuned with the Softcite dataset (Du et al., 2021). The trained model used software name and version information and does not cover the other annotated attributes available in the dataset (url, publisher). The model is trained only on the training examples of the dataset, which are all positive paragraph examples (the paragraphs with at least one software mention). The model can then be used in a pipeline starting from XML documents as input, relying on *GNU parallel* command lines.

4.4.2 Comparing model predictions with real mention destributions

The accuracy and sparsity aspect is important to stress for comparing these tools. (David et al., 2022) and (Istrate et al., 2022) both trained and evaluated their models on the annotated corpus. While software mentions are extremely sparse in actual scholar full texts, they are represented in almost every sentences/paragraphs in these corpus, which lead to model unrealistic ratio of software mentions if used as such. The models tend then to predict software in many sentences/paragraphs, leading to a very high amount of false positives. Because these models also report evaluation scores produced on partitions of the annotated data, the reported evaluations present similarly unrealistic figures, over-representing expected software mentions. The invalid evaluation of a model for an unbalanced classification problem after sampling the training data is a known issue discussed in (Vandewiele et al., 2021).

Table 3: Reported scores, reproduced scores and benchmark scores against an holdout set of full papers for three Deep Learning software mention recognizers.

Software mention recognizer	F1-score on annotated corpus as reported in original publication	F-1 score on annotated corpus, reproduced	**F1-score on holdout set** (full article content)	Note
CZI recognizer	92.0	85.5[a]	**56.3**[a]	(software name and version only)
SoMeSci recognizer	88.3	84.0[b]	**62.4**[b]	("Application" name only)
Softcite recognizer (using around 40K negative sampling examples)	-	82.3	**79.1**	(software name, version, publisher, url)

[a] Fork for reproduced cross-evaluation and Softcite holdout evaluations available at https://github.com/kermitt2/software-mention-extraction-czi
[b] Fork for reproduced SoMeSci cross-evaluation and Softcite holdout evaluations available at https://github.com/kermitt2/SoMeNLP

Table 3 illustrates the issue of training and predicting with very different mention distributions. We indicate reported scores based on cross-evaluation on the annotated corpus (all over-representing mentions) in the first two columns. In the third column, we report realistic evaluation scores on the full content of a set of annotated documents (the softcite holdout set, 20% of the Softcite articles).

This shows that a parser like (Istrate et al., 2022) trained on unrealistic over-sampled distributions will perform with much lower accuracy when applied to real full documents, producing a high rate of false positive. For this reason, in (Istrate et al., 2022), manual corrections have been realized at very large scale on the extracted results.

Similar observation applies to (David et al., 2022). The recognizer is trained on the SoMeSci corpus. This corpus is over-representing software mentions due its construction. However,

some differences in the annotation guidelines of “software names” between the SoMeSci and Softcite datasets might also artificially impact the matching accuracy.

To address false positives, (Lopez et al., 2021) has explored several sampling ratios and techniques to adapt the model to the actual extremely sparse distribution of mentions. With F1-score around 80% on actual mention distribution and full article content, the results do not require manual corrections for tasks which are robust enough to cope with some uncertainty, like statistical analysis, indicators or knowledge base construction.

4.4.3 Production considerations

The system of (Lopez et al., 2021) has the advantage of being integrated with GROBID to allow processing of PDF, while the other tools work with XML full text inputs and would require further development to support PDF parsing and structuring.

GROBID is used for parsing, extracting, and structuring the content of scientific articles in PDF, but also to drive the entity recognition in relevant sections. This tool provides in particular clean text paragraphs, solving issues like character encoding, character composition, hyphenation, reading order, identification of reference markers, footnotes, headnotes, tables and figures, which improve any subsequent text mining process. In addition, GROBID parses bibliographical references, match them to CrossRef DOI and identifies citation contexts in the text body associated to dataset and citation mentions.

The Softcite software mention recognizer includes production-ready releases, with docker images, REST API service and a python client for parallel processing. It is thus ready to deploy in the existing cloud-based BSO infrastructure and service pipeline. The system also includes a web GUI console for the service that display annotations on PDF for easy visual inspection.

We have therefore chosen to extend the system of (Lopez et al., 2021) for the production of the French Open Science indicators, considering its accuracy, the support of PDF and

its engineering maturity. Improvements to the system developed in this work includes: a new extended and enriched Softcite corpus, refinement of software types and relationships, automatic characterization of mention context in term of usage/creation/sharing and some improvement of models using LinkBERT fine-tuning (Yasunaga et al., 2022).

4.4.4 Data model and training data

Definition A "software" entity can be defined as *a collection of computer programs that provides the instructions for telling a computer what to do and how to do it*. In the present work, we mostly reused existing annotation guidelines and tried to follow consensus on the definition of "software". We consider as software: programs, packages, scripts, plug-ins, workflow scripts, API, OS, programming environments, macros and embedded software. We exclude databases, models (simulation, machine learning), algorithms/methods, programming languages and file formats.

Mentions are however often ambiguous. For example, it is quite frequent that the name of an algorithm and its implementation (as software) are used in papers in an interchangeable manner. The mention to all these entities should be interpreted in context, checking if the statement refers to the software implementation of an algorithm. Similarly, mentions to devices, databases, models, etc. require to check in context if the mention actually refers to the software part of these entities or not.

For a more detailed discussion and more examples, see the *Annotation Guideline*s coming with the public dataset (Howison et al., 2023).

A software mention contains at least a software substantive, which could be a proper name or an implicit name like *script*, *program*, etc., and optional attributes. In our framework, we considered as attributes: URL, publisher (organization or individual which publishes the software), version, programming language and one or several bibliographical references.

In order to characterize software contributions in research papers, we need to identify

precisely the created software parts. For example when a script is developed and should be run in a particular environment, we distinguish the script as one sharable software component distinct from its software environment. For this, we introduce a typing for software mention:

- **standalone software**: a software expressed in the mention context without dependency to another software and which does not require code.
- **software environment**: a software requiring some code/scripts to realize the research task, expressed in the mention context with or without dependency to another software.
- **named software component**: a named software depending on another software environment to run, the software environment being expressed in the mention context.
- **implicit software component**: an unnamed software depending on another software environment to run, the software environment being expressed in the mention context, and where the software refering expression is a generic term for program. such as *program*, *code*, *script*, *macro*, *package*, *library*, etc.

Annotations The Softcite dataset is encoded in TEI XML. In its first version, software mentions were all encoded as standalone software. To produce indicators on software and code sharing associated to research works, we extended the annotations for refining the type of the software mentions and to encode the dependencies between mentioned software. With the additional dependency annotations between software, the new Softcite dataset version is comparable to SoMeSci in term of annotation comprehensiveness, with the benefit of a larger amount of annotated documents and a coverage not limited to Life Science. Like the previous versions of the Softcite dataset, the additional annotations have been produced in parallel by 2 annotators and a reconciliation phase have been realized by a curator for all disagreements.

Table 4 present some statistics about this new revised version (2.0) of the Softcite dataset. A significant amount of software mentions have been broken down into software parts to encoded their relationships.

Table 4: Annotation overview for the Softcite corpus version 1.0 and 2.0.

Softcite dataset version	v1.0 (2020)	v2.0 (2023)
number of documents	4,971	4,971
software name (total)	4,093	5,134
- environment	-	1,089
- component	-	88
- implicit	-	106
version	1,258	1,478
publisher	1,111	1,311
URL	172	231
programming language	-	71

The latest version of the Softcite dataset is available on Zenodo (Howison et al., 2023).

4.4.5 Evaluation

The following tables present the evaluation of the two main models involved in the identification of software mentions, based on version 0.7.2 of the Softcite software mention recognizer and the updated version 2.0 of the Softcite dataset.

Software mentions and attributes

Similarly to the work in (Lopez et al., 2021), to evaluate the model on the actual distribution of software mentions in scientific articles, we have reused an *holdout set* containing the complete full text of 20% the Softcite Dataset documents (994 articles). This set reproduces the overall distribution of documents with annotation (29.0%), the distribution between Biomedicine and Economics fields, and the overall distribution of mentions per document. The *holdout set* captures therefore a realistic distribution of software mentions and can be used to produce stable evaluation using different version of the training data.

We used the remaining 80% of documents (3,977 articles), divided at paragraph-level into positive (1,886 paragraphs with at least one manual annotation) and negative (612,597 paragraphs without manual annotations). Optimizing the balance between positive and negative sampling, the final model was trained on the 1,886 positive paragraphs of the Softcite dataset and a set a of 50,000 negative paragraphs selected with active sampling as described in (Lopez et al., 2021).

Our best model is a fine-tuned SciBERT base model (Beltagy et al., 2019) with an additional CRF activation layer, which performs slightly but consistently better for this task than a fine-tuned LinkBERT model (Yasunaga et al., 2022).

Table 5: Evaluation scores of fine-tuned SciBERT model for software mention recognition.

	precision	recall	**F1-score**	support
publisher	75.51	88.80	**81.62**	250
software	74.01	88.98	**80.81**	989
url	53.97	82.93	**65.38**	41
version	83.99	90.81	**87.27**	283
all (micro avg.)	75.22	89.12	**81.58**	1563

Software type refinement

We report the evaluation of our best model for the typing of software mentions, once again a fine-tuned SciBERT model with an additional CRF activation layer. This model is applied only to sequences where at least one software mention has been identified previously. The following scores are thus produced using 10-fold cross-validation on a training set based on all the paragraph with at least one annotated software.

Table 6: Evaluation scores of fine-tuned SciBERT model for predicting software mention sub-types.

	precision	recall	**F1-score**	support
component	71.43	71.43	**71.43**	7
environment	83.33	83.33	**83.33**	102
implicit	72.73	57.14	**64.00**	14
language	100.00	61.54	**76.19**	13
all (micro avg.)	82.81	77.94	**80.30**	136

4.5 Research datasets

4.5.1 Existing work

Similarly to software mentions, data citations are inconsistent, vague, incomplete and indirect (Mooney, 2011). Beyond rule-based method such as ODDPub (Riedel et al., 2020), for which we have discussed the limitations in section 4.2, several works have explored machine learning techniques for the automatic recognition of datasets in technical and scientific documents.

Not going to the level of dataset labeling, (Zhao et al., 2019) relied on fine-tuning BERT and achieved a F1-score of 78.7 for dataset screening on a sentence-level classification task. (Hou et al., 2019) developed automatic extraction of task, dataset, metric and score also by fine-tuning BERT, reaching 67.8 F1-score. In the domain of Machine Learning, NLP and Information Retrieval, (Heddes et al., 2021) investigated BERT and SciBERT models too, reporting F1-score of 0.78 on reused named dataset mention recognition on positive sentences, and 0.73 on a more realistic ratio of positive/negative sentences artificially augmented. (Lafia et al., 2022) experimented on the Social Science domain, identifying automatically 68% of all sentences coded as having data references based on a spaCy NER model. The manually annotated dataset (not made public) was initially obtained by filtering sentences with patterns in a very large corpus of articles.

The Coleridge Initiative held a Kaggle competition on dataset recognition in 2021 in the

field of public health and social policies (“Coleridge Initiative - Show US the Data”, 2021). However, we found very hard to rely on the dataset or methods developed for this competition, as further discussed in section 4.5.2.

(Jain et al., 2020) developed SciREX, a dataset of 438 annotated arXiv documents on the Machine Learning domain, with identification of named datasets, among other entities. The reported global F1-score for mention recognition with a BERT-based model is 71.2. Finally, EneRex (Yousuf et al., 2022) is system for extraction various including technical facets from scientific publications, again on the Machine Learning field, including *used datasets*. They report F1-scores of 48.6, 63.4 and 66.2 for dataset mention recognition on the EneRex, SciREX and Papers With Code datasets respectively.

We also mention the DataSeer project (“DataSeer project”, 2019–2023), funded by the Sloan Foundation, which focused on the a sentence-level classification task related to data and data type prediction. Although the dataset and models of the project was not made available publicly, the development was entirely realized in Open Source.

To our knowledge there is currently no dataset and work with a broad coverage in term of scientific domains. The exiting work usually ignore unnamed mention of data (with the exception of ODDPub and DataSeer). Although almost all based on fine-tuned BERT, each recognizer depends greatly on domains, definition of datasets and on the distribution of dataset mentions in the training data. These differences make the reported accuracies impossible to compare, as the distribution of mentions in the evaluation data varies considerably, being in general not realistic.

4.5.2 Data model and training data

Dataset mentions are usually less complex that software mentions, but the exact scope of “research datasets” varies significantly from one existing work to another and needs some clarifications.

Implicit versus explicit research datasets Although the production of various data is very common is a scientific work, a large amount of the data discussed in scholar articles are actually not named, not curated and not shared. For example:

```
The 47 tilapia DNA samples were sent to Beijing Genome Institute (BGI) for 101
bp paired-end sequencing using Illumina Hiseq 2500. (10.1038/srep14168)
```

The Illumina Hiseq 2500 device produces sequencing data, which are part of the described study. These sequencing data are however not named and not shared, they remain implicit. Knowledge about the field and the involved devices are necessary to infer that data are produced. We define this sort of research dataset as *implicit dataset*. These data are generally not considered as valuable by the researchers and often not in a sharable state (proprietary format, embedded as project resources). We consider fully automatic recognition of this kind of implicit datasets currently beyond what is feasible given the current existing annotated corpus and machine learning technique accuracy.

In contrast, in this work we focus on *explicitly mentioned datasets*, which we define as:

1. all named datasets,
2. unnamed datasets explicitly mentioned as existing data.

Explicitly mentioned datasets are research data identified as such by the authors, discussed and considered valuable with respect to the scientific claims of the publication. They are usually available in a form that makes sharing possible. Datasets related to 2) are unnamed data mentioned as used, produced or shared. They are usually referred to with generic substantive *data* or *dataset*, which can be annotated:

```
The data has been collected by the UN Comtrade organization
(https://comtrade.un.org/), and cleaned by CEPII. (10.1371/journal.pone.0203915)
```

```
All the data were recorded after each sub-culture (at 3, 6 and 9 weeks).
(10.1038/s41598-018-37335-7)
```

```
The dataset used in this study consist of natural products that have been tested
for in-vitro antiplasmodial activities (NAA) compiled in-house from literature,
PhD Theses and public chemical databases. (10.1371/journal.pone.0204644)
```

Similarly to software, a reference to a named dataset can be ambiguous and should be examined in context. For example, is a mention to a **database** the same as a mention to a dataset? If research data have been loaded and shared via this database, we can consider in general that this data are packaged as a dataset. On the other hand, a "dataset" can exists independently from a database management system and is not ambiguous with respect to its storage and software access environment. So for deciding if a mention of a database can be considered as a valid dataset mention, one has to clarify if the research work is referring to the data stored in the database or more generally to the software or the service associated to the database.

In addition, we have defined the following main guidelines:

- *Data sharing initiative/project*: we currently exclude the mentions to initiatives, collaborations (HEP or Astronomy) and projects from our dataset extraction. Only mentions to individual research dataset or collection are considered as a "dataset".
- *Accession number*: the references to an entry in a database, for instance via a unique identifier such as an accession number, is considered as a valid mention to a dataset.

Training Developing an annotated corpus of dataset mentions entirely from scratch would not have been possible for this project. The Softcite dataset for example took several years of development and involved a total of 38 different human annotators. To train our dataset mention recognizer, we reused existing annotated corpus, re-annotated them to follow our guidelines and produce some additional annotations for a limited subset.

As noted in section 4.5.1, the datasets of many existing work covering dataset mention recognition have not been made publicly available for reuse. Relatively the Coleridge Kag-

gle competition ("Coleridge Initiative - Show US the Data", 2021), we found that the annotations were unreliable, inconsistent and highly incomplete (less than 70 distinct "datasets" annotated only in 8,000 documents). In addition, this effort was realized without clear definition of a "dataset". For example, a research initiative name such as ADNI could be labeled as "dataset" even in context not related to the datasets produced by this initiative. Unfortunately, the Coleridge dataset makes impossible to apply state-of-the-art NER techniques without a considerable additional labeling effort and we could use it in a very limited manner explained below.

One additional challenge is that, to our knowledge, there is no annotation on a complete full text versions of the annotated articles, well adapted to tackle the sparsity problem on real distribution. The existing training data are always sets of positive sentences with at least one dataset annotation. In general, there is no guarantee that the rest of these articles do not include other datasets and checking such a large amount of content is highly time-consuming. To mitigate this issue, we divided the task in two steps, the first one to identify the data sentences in a complete article, and the second one to spot dataset mentions in selected data sentences.

The first step is implemented as data sentence classification model. We used 2,000 positive sentences containing at least one dataset mention and 20,000 negative data sentences without dataset mention from PubMed Central full-texts and from the Coleridge dataset. This combined dataset was used to fine-tune a LinkBERT-base binary classifier.

The second step is the dataset mention recognition in the identified data sentences. To train this model, we use the following resources:

- a re-annotated version of the dataset[5] developed by (Heddes et al., 2021), a set of 6,000 sentences in the IR/ML/NLP domain with 3,684 dataset mentions. We fully reviewed the dataset and re-annotated to follow our dataset annotation principles: it

[5] https://github.com/xjaeh/ner_dataset_recognition

covers now new datasets (not just reused ones) with annotations at individual dataset level (avoid one annotation for a conjunction expression of datasets).

- a set of approx. 1000 sentences from PubMed Central full-texts with at least one dataset mention (explicit and implicit datasets) and partial data acquisition device annotations.

These resources were used to train a LinkBERT base model (Yasunaga et al., 2022) with an additional CRF activation layer.

4.5.3 Accuracy

As explained above, the first model is classifying if a sentence introduces a dataset or not. A 10-fold cross-validation on our dataset gives the accuracy presented in Table 7.

Table 7: Evaluation scores for first-stage data sentence recognition.

	precision	recall	**F1-score**	support (10%)
data sentence	93.70	96.21	94.94	200
not data sentence	97.56	95.92	96.73	2000

Similarly, we evaluate the mention recognition using 10-fold cross-validation on the annotated dataset, Table 8. In the actual application of the models, the second model is applied only to sentences detected as “data sentence”, so the error of the previous model detecting “data sentences” will be propagated to the second model. However, even combining the two error rates, the recognition of mentions of explicit datasets (dataset with names or without name but expressed as dataset or data) appears very reliable with the current amount of training data. This result supports the relevance of these extracted mentions for building indicators.

At this stage, unnamed datasets are harder to recognize and will require additional training data and modeling efforts. We also present for reference the current recognition scores

Table 8: Evaluation scores for dataset mention recognition in predicted data sentences.

	precision	recall	**F1-score**	support (10%)
named dataset	89.04	89.46	**89.24**	466
unnamed mentionned dataset	71.85	67.15	**69.38**	927
data device	51.91	37.94	42.61	97

for data device mentions, but their manual annotations are still work-in-progress and very limited. We think that the automatic identification of data acquisition devices or data processing devices could help in the future to spot implicit data in a more reliable way.

4.6 Characterization of mention contexts

Whether or not a research dataset or software mentioned in an article was used, created and shared is important to monitor the compliance with Open Science policies. We hypothesize here that the wording used to introduce and describe a dataset or software mention can characterize its possible usage, creation and sharing. The sentences containing the mentions are used as classifier input, without additional features. As we observed that the wording used to describe the role of these mentions is very similar for dataset and software, we use the same classifiers for characterizing both research dataset and software mentions.

4.6.1 Training data

The annotated data for training the classifiers are a combination of existing training data and additional manual annotation realized during the project:

- l'attribut *used* of the Softcite dataset
- the corresponding annotations in the SoMeSci dataset
- a new additional set of 500 contexts manually annotated focusing more on datasets

and the minoroty classes (*created* and *shared*)

Table 9 presents the distribution of classes in this assembled training dataset.

Table 9: Distribution in manually annotated training data of the classes for dataset and software mention characterization.

total contexts	3,643
used	2,774
not used	869
created	338
not created	3,305
shared	266
not shared	3,377

4.6.2 Accuracy

The evaluation scores presented in Table 10 are produced using 10-fold cross-validation based on three binary classifiers, one per class used/created/shared. The classifiers are fine-tuned LinkBERT base model (Yasunaga et al., 2022). Binary classifiers for each class perform significantly better than a single multiclass classifier (up to 4 F1-score points).

Table 10: Evaluation scores for dataset and software mention characterization.

	precision	recall	**F1-score**	support
used	96.83	94.18	**95.49**	292
not used	84.40	91.09	**87.62**	101
created	81.08	83.33	**82.19**	31
not created	98.31	98.04	**98.18**	362
shared	81.82	90.00	**85.71**	26
not shared	99.35	98.71	**99.03**	385

4.7 Recognition of availability statements

We have extended GROBID to identify automatically data and code availability statements in research publications. We define a data and/or code availability statements as a standalone section of a research publication (with a section title and one or several paragraphs) describing how the data and code involved in the research work can or cannot be accessed. Availability statements appear usually in the front page of an article or at the end as a annex, but we also considered positions inside the main body, not rare with preprints. We do not put any constraints on the section title associated to "data availability statements".

95 articles out of the 520 training articles of the GROBID `segmentation` model have been further annotated with data availability section markups. This GROBID model is used to segment the main zone of a scientific article, such as header, body, bibliographical section, acknowledgement or funding. The data availability section is then structured and identified as such in the file TEI result file.

To cover the whole spectrum of difficulty, we evaluated the reliability of recognition on two set of high quality publisher publications and on one set of preprints:

- a set of 1000 random PLOS articles in PDF and JATS XML (2003-2022) containing 779 data and code availability statement markup information
- a set of 984 random eLife articles in PDF and JATS XML (from the complete eLife collection) containing 585 data availability statements markup information
- a set of 2000 random bioRxiv articles, which have been reviewed and completed manually, containing 473 data availability statements

Table 11 presents an evaluation of the accuracy of the data and code availability statement recognition in a end-to-end scenario with PDF as input.

On recent publications (2020 and after), we observed that the presence of data availability sections is correctly recognized in nearly all PLOS and eLife. On the other hand, preprint

Table 11: Evaluation scores for the automatic identification of Data Availability Statements. Datasets are available at https://zenodo.org/record/7708580.

collection	precision	recall	**F1-score**	support
PLOS 1000 articles	99.57	89.73	**94.4**	779
eLife 984 articles	96.62	92.82	**94.68**	585
bioRxiv 2000 articles	82.7	78.25	**80.41**	473

data availability statements can be challenging to identify because they can appear in non-usual positions in the article. They can also be introduced by a large variety of section titles depending on the described data. In peer-reviewed publisher versions, we generally observe a high regularity and very high precision similar to PLOS/eLife articles.

4.8 Architecture of the mention recognizers

The dataset and software mention extraction process are very similar and shares a common software architecture summarized by Figure 1.

We expect input full texts to be in PDF formats, but a large range of publisher XML formats are also supported. PDF are first parsed by a PDF parser based on the Open source library Xpdf called *pdfalto*. Complementary to the support of ALTO output, a modern format for OCR output, pdfalto implements additional features relevant to scientific documents, in particular better handling of multi-column documents, the recognition of superscript/subscript style and the robust recognition of line numbers for review manuscripts.

GROBID is then applied to structure the raw PDF stream into header, sections, paragraphs, footnotes, etc. Grobid applies in cascade a set of machine learning models to create hierarchical structures, combining text content and layout features (e.g. font and style information, relative positions, SVG objects, etc.). Bibliographical references are also extracted and parsed. If the input is encoded in a publisher XML format, we transform the XML into the same TEI format as produced by GROBID thanks to the Open Source Pub2TEI tool,

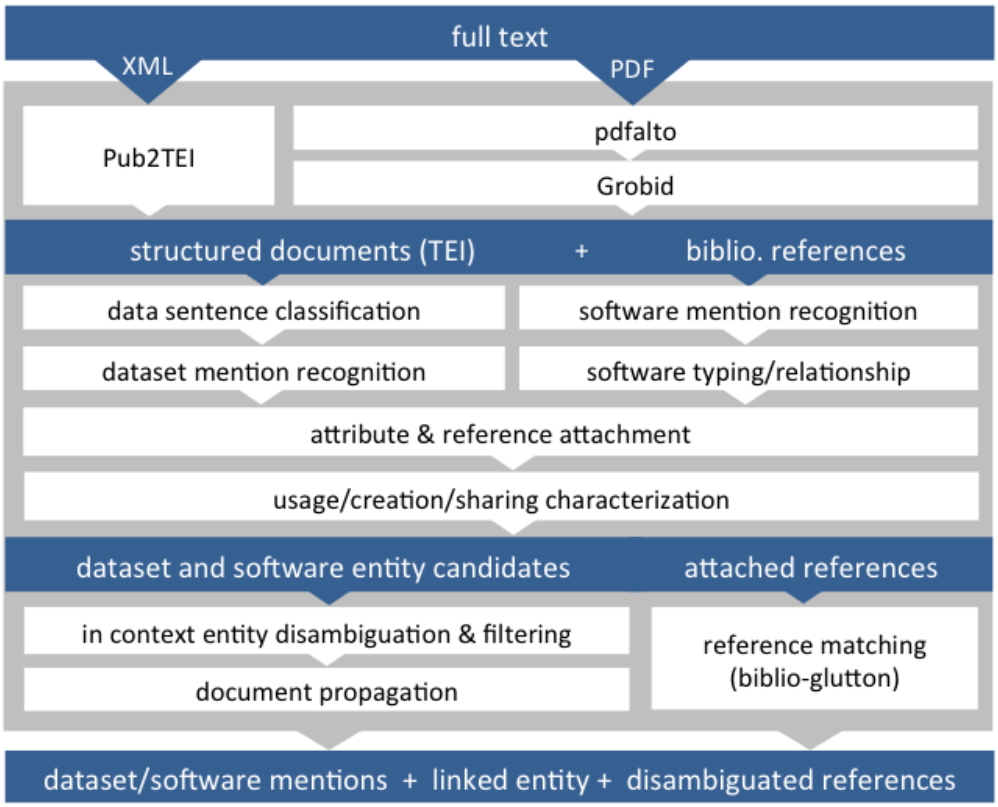


Figure 1: Overview of the dataset and software mention recognition process.

which support the main XML publisher formats such as JATS, Elsevier, Wiley, etc. We can therefore centralize our document processing across PDF and XML sources.

Mining for specific entities is only relevant to certain textual structures of a scientific document, such as paragraphs, abstracts, figure captions, etc. In the Softcite corpus, we calculated that, on average, 28% of publication content should be filtered out. This includes metadata (author, affiliations), bibliographical sections, table and figure content, formulas, headnotes, page numbers, reference markers, or some editorial annexes (like conflicts of interest). Relying on GROBID not only improve the quality of textual content, but it also makes possible to apply the text mining process only to relevant structures, avoiding possible sources of false recognition.

To identify software names and related attributes, a first mention recognition is applied to

relevant textual structures. For identified software names, a second model is then applied to refine the software mention types. For recognizing datasets, two models are applied. A first classification model is applied to every sentences of relevant textual structures to determine if the sentence is describing or not a dataset. A second model is applied on positive data sentences to identify dataset mention spans and attributes like URL and bibliographical references. This division of the recognition task into two steps is introduced to mitigate the problem of lack of dataset annotations for complete documents, mentioned in the section 4.5.2.

Attachment of the bibliographical references in the context of the mentions are evaluated and if successful these bibliographical references are fully resolved against CrossRef DOI via the `biblio-glutton` service ("biblio-glutton", 2018–2023). An entity disambiguation is realized in context using `entity-fishing` ("entity-fishing", 2016–2023) for every software mention candidates. If the candidates is significantly more likely scientific entities different from a software, the candidates are discarded to avoid likely false positives. Otherwise, if the software entity is a known disambiguated software present in Wikidata, the entity is linked via the resulting Q Wikidata identifier. Finally a document-level propagation is realized to identified possible overlooked dataset and software name matches in the same article in order to improve recall.

5 Application to the French Open Science Monitor

5.1 Monitoring indicators

To help steer French public policy, we have seen in section 1.3 that our indicators must meet several minimum quality conditions. In particular, indicators must have floor and ceiling easy to understand and interpret by the public, with ceiling values as objectives of a policy.

For research publications, the French Open Science Monitor reports the percentage of open access publications of the previous years. Each year the new indicators are recalculated and do not depend on the indicator of the previous years. The openness indicators vary between 0% and 100%, and 100% is the French National Open Science Plan objective for 2030.

For consistency and readability, we consider that the same requirements apply to indicators for research data and software. The mention detection uses the full-text of the publications as raw material. It is therefore straightforward to also propose indicators relating to the proportion of publications. As the methodology is similar for both datasets and software, the proposed indicators will be equivalent, replacing dataset by software.

Research dataset and software notations and indicators

$n_{publication}$: number of publications

$n_{text-mined}$: number of text-mined publications

n_{use}: number of text-mined publications that mention the usage of at least one dataset

n_{create}: number of text-mined publications that mention the usage and the production of at least one of their datasets

n_{share}: number of text-mined publications that mention the usage, the production and the sharing of at least one of their dataset

$$P_{text-mined} = \frac{n_{text-mined}}{n_{publication}} \quad (1)$$

$$I_{naive} = \frac{n_{share}}{n_{publication}} \quad (2)$$

$$I_{use} = \frac{n_{use}}{n_{text-mined}} \quad (3)$$

$$I_{create} = \frac{n_{create}}{n_{use}} \quad (4)$$

$$I_{share} = \frac{n_{share}}{n_{create}} \quad (5)$$

A direct indicator would be, for example, the proportion of publications that share a dataset

as given by I_{naive} (2). This indicator is simple to understand, but has several shortcomings. First, concerning the denominator, not all publications can be analysed by Softcite and DataStet, because the PDF could not systematically be downloaded. It should therefore be replaced by the number of publications that have been processed. Second, we want to have indicators whose upper bound has interpretations matching the objectives of the public policy. For example, if the described research work does not require the use of any datasets or software, the article is not relevant for sharing ratio. Thus, the notion of a publication that shares a dataset should be clarified as a publication that uses, creates and shares a dataset.

We therefore propose to monitor the modified key indicator I_{share} (5). Publications that do not create a dataset are however also meaningful to understand which proportion of the publication is relevant for data sharing. Therefore, we propose to complete the analysis with the two additional indicators I_{use} (3) and I_{create} (4).

These three indicators I_{use}, I_{create} and I_{share}, correspond to a funnel analysis. They give a global and separated view of all relevant cases: the bottom of the funnel correspond to the shareable data, which is actionable and which we want to increase. The other two are more descriptive, but will provide insights on understanding the practices regarding data, particularly by breaking them down by subject area.

The three indicators can be further broken down by metadata facets of the publications: year of publication, disciplines, publishers, etc.

We supplement these indicators by providing also the ratio of publication where an explicit Data Availability Statement section has been recognized in the full text. Although Data Availability Statements offer no guarantee of actual sharing for various reasons (Gabelica et al., 2022), it permits to track the evolution of adoption of this practice.

5.2 Infrastructure and runtime

Figure 2 outlines the overall automated processing of the full texts part of the BSO platform. After the harvesting of PDF, several machine learning modules are applied in parallel to extract structured information from the full text: GROBID, Softcite and DataStet. The different extracted mention annotations and metadata are analyzed and aggregated by a Python module to produce the indicators at document level. This module feeds an ElasticSearch instance, used to export a public data dump and as back-end for the different BSO related dashboards.

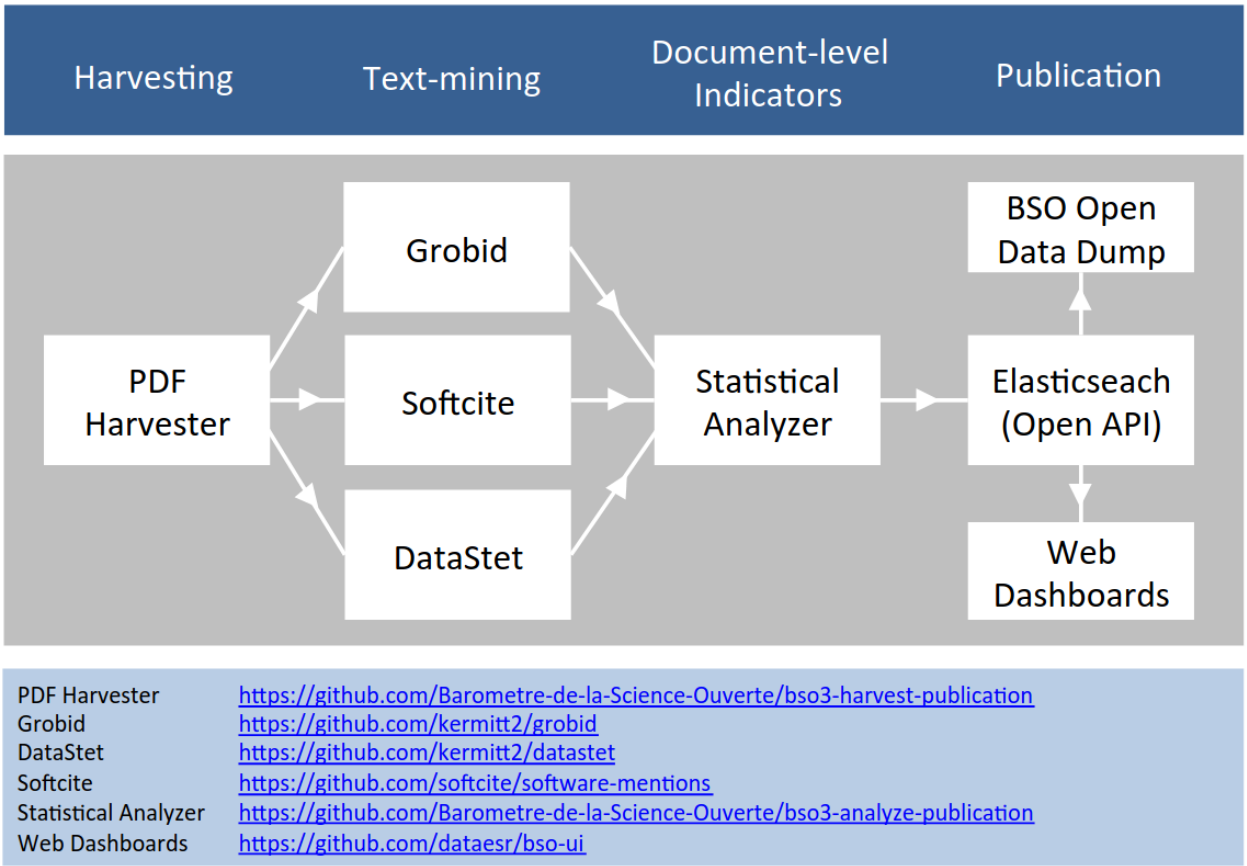


Figure 2: Global overview of the text-mining data flows

All the modules are deployed on a managed Kubernetes cluster provided by the public cloud OVHcloud. Unfortunately, the cloud provider did not make available GPU at the time of the run. We use 5 servers with 16 CPU 240GB RAM and 5 servers with 32 CPU

120GB RAM to deploy the GROBID, Softcite and DataStet services. An extra server is used to orchestrate all the services calls. Processing around 700k publications took about one month, longer that initially expected. To avoid delaying the scheduled release date of the new Monitor version, we did not process 100% of the downloaded publications as shown by Table 12 - something we expect to solve in the next year version by adjusting our computing capacities. For this reason, we currently present these new indicators on research datasets and software as *beta*. All the data resources are stored on secured cold storage, representing a total of 1.86TB.

5.3 Dataset and software mention extraction

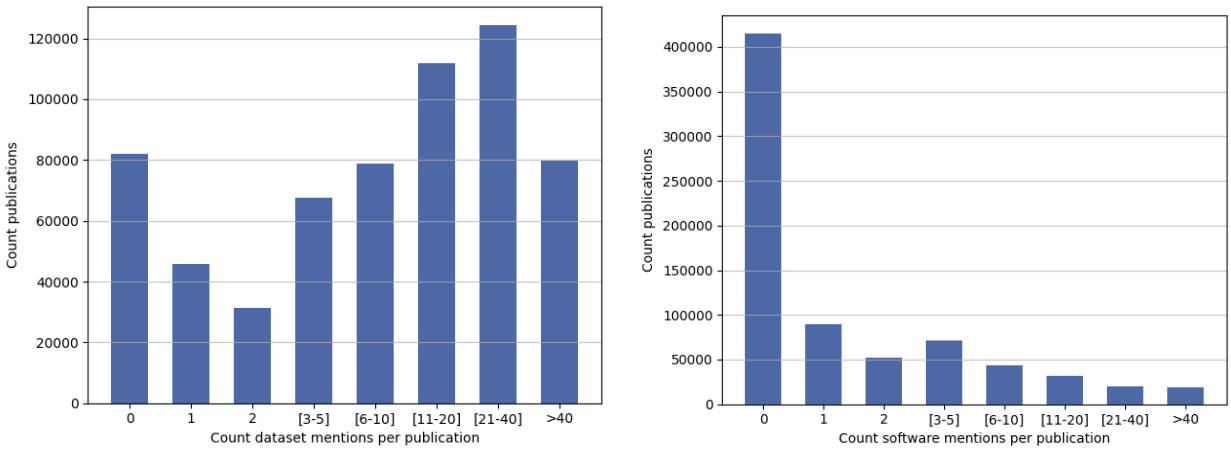
Table 12 presents statistics on the number of documents processed by the text mining modules and the total of extracted mentions in the BSO published in early 2023 (period 2013-2021).

Table 12: Statistics on the proportion of text-mined publications in the French Open Science Monitor corpus covering 2013-2021.

	Publications 2013-2021		
	# documents	share (%)	# mentions
Full corpus	1,426,140	100.0	
Full texts downloaded	908,567	63.7	
Processed with GROBID	743,700	52.1	
Processed with Softcite	742,289	52.0	3,567,547
Processed with DataStet	621,306	43.6	5,607,080

Producing document-level indicators as defined in 5.1 is interesting in term of interpretability, but also for reliability. As the number of extracted mentions increases for one document, the number of observations for producing the indicators also increases.

Figure 3 shows the distribution of documents by number of extracted mentions. We see that for datasets, 462,486 documents have three or more extracted mentions (85% of all documents having at least one dataset mention), and for software 185,710 documents (56%



(a) (b)

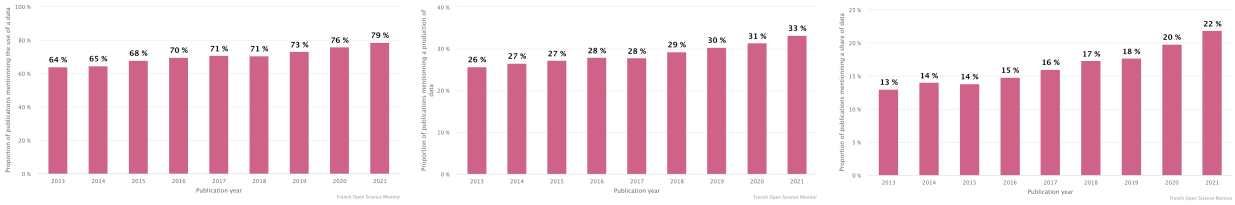
Figure 3: Distribution of documents by number of extracted mentions in the French Open Science Monitor corpus covering 2013-2021, for (a) datasets and (b) software.

of all documents having at least one software mention). Considering for instance an error rate at mention level at 15-20%, when several mentions are extracted this error rate follows a polynomial reduction. Several mentions similarly classified (as used/created/shared) consolidates the document-level indicator at higher certainty levels.

5.4 French Open Science Monitor indicators and dashboards

The first step of the funnel analysis described in section 5.1 focuses on the use of data in research works. Figure 4(a) shows that the majority of the publications is mentioning the use of data. It also reveals a positive trend over time, with almost 80% of processed publications mentioning the use of data for the publication year 2021.

The second step of the funnel analysis focuses on the production of research data 4(b). It analyses, among the text-mined publications that mention the use of data, the proportion of the publications that also mention the production of data. Around one third of the text-mined publication mention having produced their data, and again the trends goes upwards over time.

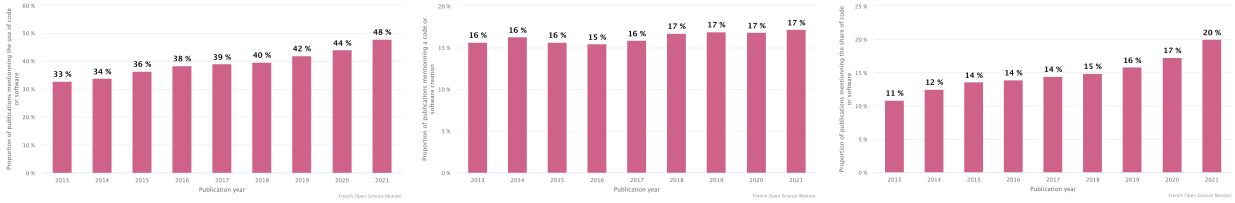


(a) (b) (c)

Figure 4: (a) Proportion of publications in France that mention the use of data. (b) Proportion of publications in France that mention having produced data. (c) Proportion of publications in France that mention the sharing of their data.

The last step of the funnel analysis focuses on the sharing of the produced data 4(c). This indicator goes from 13% for the publications published in 2013 to 22% for the publications published in 2021.

The same funnel analysis conducted for research data is applied for software and code, present on Figure 5, showing lower ratio in all categories.



(a) (b) (c)

Figure 5: (a) Proportion of publications in France that mention the use of code or software. (b) Proportion of publications in France that mention having created code or software. (c) Proportion of publications in France that mention the sharing of their code or software.

As explained in section 4.7, we also track the proportion of publications that include a data and code availability statement section. As show in Figure 6, the upward trend is very strong, as this indicators goes from only 1% in 2013 to 21% in 2021.

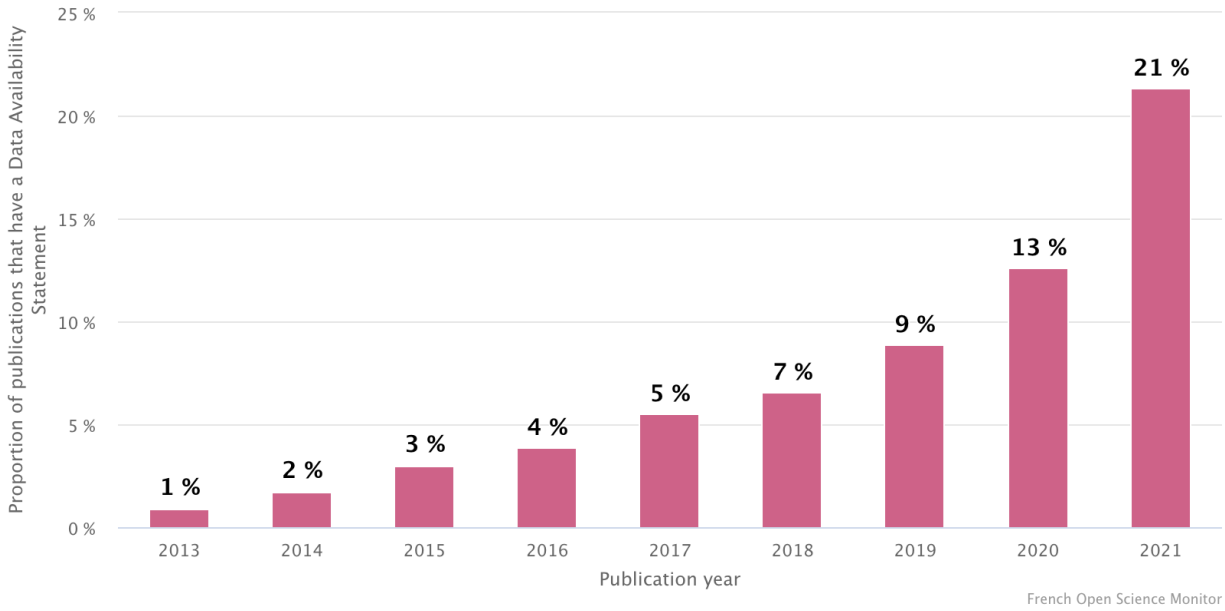


Figure 6: Proportion of publications in France including a Data Availability Statement.

5.5 Local Open Science Monitor

One of the objectives of the French Open Science Monitor is to provide all research institutions with a local version of the indicators. Similarly as for measuring the openness of research publications on the existing platform, the new indicators can be bounded to datasets and software produced by the researchers only affiliated with a particular organization.

Not requiring any coding skills, more than 40 French research institutions have published their local monitor with the dataset and software indicators, as of May 2023. This capacity of simple customization has led to the creation of a French Open Science Monitor user group with more than 200 members.

6 Limitations and future work

Corpus completeness Several factors limit currently the completeness of our text mining processing:

- The creation of the country-wide corpus of publications is only covering articles with DOI.
- In the current first version of the Monitor, around two thirds of full texts have been successfully harvested.
- Only English are currently supported by Softcite and DataStet.

We are aware that these factors in combination impact more particularly certain scholar fields like Law, Social Science and Humanities. These fields have a lower adoption of DOI, Open Access and publish more frequently in French language.

We expect to increase both the coverage of publication by improving our affiliation matching and by extending further the collection to publications without DOI. The French monitor corpus already contains PhD thesis and all the documents indexed on the French open repository HAL, regardless of the presence of a DOI. These documents could potentially be analyzed to detect datasets and software mentions.

The effort required to the support of the French language by Softcite and DataStet will be also evaluated. Accordingly, a plan for producing additional training data in French language will be planed for the next year.

Performance across domains Research dataset and software mention recognizers are currently limited by the relatively poor multidisciplinary coverage of the training data. (Lopez et al., 2021) shows that the recognition performance for software mention falls by around 15 points F1-score on an entirely new scientific domain.

To mitigate this issue, we have started new round of manual annotations on multiple do-

mains with systematic model retraining and evaluation. We already observed that, as the global corpus grows, the required number of annotated documents to cover a new domain decreases. We think that we can realistically expect a relatively uniform accuracy of the recognizer across all scholar fields at medium term.

Large-scale research entity disambiguation Recognizing dataset and software entities could make possible indicators centered not only on documents, but also on global production of research datasets and software and their reuse. By recognizing and reporting data and software reuse among a large corpus of publications over time, researchers could receive credit and acknowledgement for the impact of their datasets and software. Such credit can be strong incentive for researchers to open their data and software, as well as producing better datasets and software over time.

However dataset and software-level indicators suppose to disambiguate different mentions of the same datasets and software entity. This task is complicated because software names tend to be very ambiguous and research datasets are mostly unnamed. In addition, no reference database of research software with uniform metadata exists today. We have started to explore mention disambiguation at scale, including referencing when possible datasets with DataCite entries and software with Software Heritage archive and will assess the feasibility of dataset and software-level indicators in the future BSO.

Local Open Science Monitor Automatic affiliation-level identification is harder than country identification. Moreover, smaller institutions might not gather enough publications for the minority categories *production* and *sharing* on datasets and software for reliable trends over time. To partially mitigate these issues, the BSO includes the possibility to create local monitors based on list of DOI, which is relevant for the institutions collecting the list of all their researcher publications for general reporting and evaluation purposes.

Covering dataset repositories While datasets associated to an existing publication are particularly important in the context of Open Science and scientific quality, other types of datasets might be developed without mentions in published research works. Some of them can simply be deposited on research dataset repositories. We think that an additional indicator following the global production of deposited datasets would complement the BSO and help to understand and drive research practices. However, poor affiliation metadata associated to datasets in repositories is an additional challenge. We are currently working on such an additional indicator, which is expected in the next BSO version.

7 Conclusion

The French Open Science Monitor shows first that quality indicators can rely on open bibliographical data only. Being independent from proprietary databases is an advantage in term of public trust, sovereignty and sustainability. Beyond research publications, our platform also demonstrates that reliable indicators on datasets and software openness can be developed thanks to modern text mining techniques at a large scale. This approach does not depend on the adoption of PID for these research products, which is currently too limited.

Even if indicators are estimates and not exact absolute values, we think that the relative trends over several years are reliable. First as the indicators are produced under the same conditions over time, consistent global trends are captured. Second, several mentions identified in the same document lead to very reliable document-level measurements because the global error rate margin is significantly reduced. Rather than perfect absolute measurements, reliable trends are enough to start following the impact of Open Science policies and adapt effort accordingly.

We expect that these techniques will also help to explore other research measurements in the next version of the Monitor. Dataset and software-level citations and impact in

particular can pave the way to new incentive to recognize and improve the production of datasets and software by scientists.

References

Beltagy, I., Lo, K., & Cohan, A. (2019). Scibert: A pretrained language model for scientific text.

Biblio-glutton. (2018–2023). https://github.com/kermitt2/biblio-glutton

Bracco, L., L'Hôte, Anne, Jeangirard, Eric, & Torny, Didier. (2022). Extending the open monitoring of open science: A new framework for the French Open Science Monitor (BSO). https://hal.archives-ouvertes.fr/hal-03651518

Chaignon, L., & Egret, D. (2022). Identifying scientific publications countrywide and measuring their open access: The case of the French Open Science Barometer (BSO). *Quantitative Science Studies*, *3*(1), 18–36. https://doi.org/10.1162/qss_a_00179

Chiticariu, L., Krishnamurthy, R., Li, Y., Reiss, F., & Vaithyanathan, S. (2010). Domain adaptation of rule-based annotators for named-entity recognition tasks. *Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing*, 1002–1012.

Coleridge initiative - show us the data. (2021). https://www.kaggle.com/competitions/coleridgeinitiative-show-us-the-data

Cousijn, H., Feeney, P., Lowenberg, D., Presani, E., & Simons, N. (2019). Bringing citations and usage metrics together to make data count. *Data Science Journal*, *18*(1), 9. https://doi.org/http://doi.org/10.5334/dsj-2019-009

Dataseer project. (2019–2023). https://github.com/dataseer

David, S., Felix, B., Stefan, D., & Frank, K. (2022). The role of software in science: A knowledge graph-based analysis of software mentions in PubMed Central. *PeerJ Computer Science*, *8*, e835. https://doi.org/10.7717/peerj-cs.835

Du, C., Cohoon, J., Lopez, P., & Howison, J. (2021). Softcite dataset: A dataset of software mentions in biomedical and economic research publications. *Journal of the Association for Information Science and Technology*, *72*(7), 870–884. https://doi.org/10.1002/asi.24454

Du, C., Cohoon, J., Lopez, P., & Howison, J. (2022). Understanding progress in software citation: A study of software citation in the CORD-19 corpus. *PeerJ Computer Science*, *8*, e1022. https://doi.org/10.7717/peerj-cs.1022

Duck, G., Kovacevic, A., Robertson, D. L., Stevens, R., & Nenadic, G. (2015). Ambiguity and variability of database and software names in bioinformatics. *Journal of biomedical semantics*, *6*(1), 29.

Elger, K., Biskaborn, B. K., Pampel, H., & Lantuit, H. (2016). Open research data, data portals and data publication–an introduction to the data curation landscape. *Polarforschung*, *85*(2), 119–133.

Else, H. (2018). The rise and rise of unpaywall. *Nature*, *560*(7718), 290–291.

Entity-fishing. (2016–2023). https://github.com/kermitt2/entity-fishing

Gabelica, M., Bojfçifá, R., & Puljak, L. (2022). Many researchers were not compliant with their published data sharing statement: A mixed-methods study. *Journal of Clinical Epidemiology*, *150*, 33–41. https://doi.org/https://doi.org/10.1016/j.jclinepi.2022.05.019

Garijo, D., Osorio, M., Khider, D., Ratnakar, V., & Gil, Y. (2019). Okg-soft: An open knowledge graph with machine readable scientific software metadata. *2019 15th International Conference on eScience (eScience)*, 349–358. https://doi.org/10.1109/eScience.2019.00046

Grobid. (2008–2023). https://github.com/kermitt2/grobid

He, L., & Han, Z. (2017). Do usage counts of scientific data make sense? an investigation of the dryad repository. *Library Hi Tech*, *35*(2), 332–342. https://doi.org/10.1108/LHT-12-2016-0158

Heddes, J., Meerdink, P., Pieters, M., & Marx, M. (2021). The automatic detection of dataset names in scientific articles. *Data*, *6*(8). https://doi.org/10.3390/data6080084

Hou, Y., Jochim, C., Gleize, M., Bonin, F., & Ganguly, D. (2019). Identification of tasks, datasets, evaluation metrics, and numeric scores for scientific leaderboards construction. *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, 5203–5213. https://doi.org/10.18653/v1/P19-1513

House, T. W. (2023). Fact sheet: Biden-harris administration announces new actions to advance open and equitable research [Accessed on May 17th, 2023]. https://www.whitehouse.gov/ostp/news-updates/2023/01/11/fact-sheet-biden-harris-administration-announces-new-actions-to-advance-open-and-equitable-research/

Howison, J., & Bullard, J. (2016). Software in the Scientific Literature: Problems with Seeing, Finding, and Using Software Mentioned in the Biology Literature. *Journal of the Association for Information Science and Technology*, *67*(9), 2137–2155. https://doi.org/10.1002/asi.23538

Howison, J., Lopez, P., Du, C., & Cohoon, H. (2023). Softcite dataset version 2. https://doi.org/10.5281/zenodo.7995565

Istrate, A.-M., Li, D., Taraborelli, D., Torkar, M., Veytsman, B., & Williams, I. (2022). A large dataset of software mentions in the biomedical literature. https://arxiv.org/abs/2209.00693

Jain, S., van Zuylen, M., Hajishirzi, H., & Beltagy, I. (2020). Scirex: A challenge dataset for document-level information extraction. *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*.

Jeangirard, E. (2019). Monitoring Open Access at a national level: French case study. *ELPUB 2019 23rd edition of the International Conference on Electronic Publishing*, *Academic publishing and digital bibliodiversity*. https://doi.org/10.4000/proceedings.elpub.2019.20

Juty, N., Wimalaratne, S. M., Soiland-Reyes, S., Kunze, J., Goble, C. A., & Clark, T. (2020). Unique, persistent, resolvable: Identifiers as the foundation of fair. *Data Intelligence*, *2*(1-2), 30–39.

Katz, D., Chue Hong, N., Clark, T., Muench, A., Stall, S., Bouquin, D., Cannon, M., Edmunds, S., Faez, T., Feeney, P., Fenner, M., Friedman, M., Grenier, G., Harrison, M., Heber, J., Leary, A., MacCallum, C., Murray, H., Pastrana, E., … Yeston, J. (2021). Recognizing the value of software: A software citation guide [version 2; peer review: 2 approved]. *F1000Research*, *9*(1257). https://doi.org/10.12688/f1000research.26932.2

Krüger, F., & Schindler, D. (2020). A Literature Review on Methods for the Extraction of Usage Statements of Software and Data. *Computing in Science Engineering*, *22*(1), 26–38. https://doi.org/10.1109/MCSE.2019.2943847

Lafia, S., Fan, L., & Hemphill, L. (2022). A natural language processing pipeline for detecting informal data references in academic literature. *Proceedings of the Association for Information Science and Technology*, *59*(1), 169–178. https://doi.org/https://doi.org/10.1002/pra2.614

Larregue, J., Vincent-Lamarre, P., Lebaron, F., & Larivière, V. (2020). Covid-19: where is the data? https://doi.org/10.5281/zenodo.2585783

Laurinavichyute, A., Yadav, H., & Vasishth, S. (2022). Share the code, not just the data: A case study of the reproducibility of articles published in the journal of memory and language under the open data policy. *Journal of Memory and Language*, *125*, 104332. https://doi.org/https://doi.org/10.1016/j.jml.2022.104332

Lee, J., & Deleris, L. (2020). Sequencing, combining and sampling classifiers to help find needles in haystacks. *24th European Conference on Artificial Intelligence, ECAI*.

L'Hôte, A., & Jeangirard, E. (2021). Using elasticsearch for entity recognition in affiliation disambiguation. https://arxiv.org/abs/2110.01958

Lopez, P., Du, C., Cohoon, J., Ram, K., & Howison, J. (2021). Mining software entities in scientific literature: Document-level NER for an extremely imbalance and large-scale task.

Proceedings of the 30th ACM International Conference on Information & Knowledge Management, 3986–3995. https://doi.org/10.1145/3459637.3481936

Macgregor, G., Lancho-Barrantes, B. S., & Pennington, D. R. (2022). Exploring the concept of pid literacy: User perceptions and understanding of persistent identifiers in support of open scholarly infrastructure. *arXiv preprint arXiv:2211.07367*.

McKenzie, L. (2017). Want to analyze millions of scientific papers all at once? here's the best way to do it. *Science*. https://doi.org/10.1126/science.aan7139

MESR. (2018). National Plan for Open Science. https://cache.media.enseignementsup-recherche.gouv.fr/file/Recherche/50/1/SO_A4_2018_EN_01_leger_982501.pdf

MESR. (2021). 2nd National Plan for Open Science. https://cache.media.enseignementsup-recherche.gouv.fr/file/science_ouverte/20/9/MEN_brochure_PNSO_web_1415209.pdf

Mooney, H. (2011). Citing data sources in the social sciences: Do authors do it? *Learned Publishing*, *24*(2), 99–108. https://doi.org/https://doi.org/10.1087/20110204

Nelson, A. (2022). Memorandum for the heads of executive departments and agencies - ensuring free, immediate, and equitable access to federally funded research [Accessed on May 17th, 2023]. https://www.whitehouse.gov/wp-content/uploads/2022/08/08-2022-OSTP-Public-Access-Memo.pdf

Park, H., You, S., & Wolfram, D. (2018). Informal data citation for data sharing and reuse is more common than formal data citation in biomedical fields. *Journal of the Association for Information Science and Technology*, *69*(11), 1346–1354. https://doi.org/https://doi.org/10.1002/asi.24049

Philippe, O., Hammitzsch, M., Janosch, S., van der Walt, A., van Werkhoven, B., Hettrick, S., Katz, D. S., Leinweber, K., Gesing, S., Druskat, S., Henwood, S., May, N. R., Lohani, N. P., & Sinha, M. (2019). softwaresaved/international-survey: Public release for 2018 results. https://doi.org/10.5281/zenodo.2585783

Public Library of Science. (2023). PLOS Open Science Indicators. https://doi.org/10.6084/m9.figshare.21687686.v2

Riedel, N., Kip, M., & Bobrov, E. (2020). Oddpub - a text-mining algorithm to detect data sharing in biomedical publications. *Data Science Journal*, *19*(1), 42. https://doi.org/http://doi.org/10.5334/dsj-2020-042

Schindler, D., Bensmann, F., Dietze, S., & Krüger, F. (2021). Somesci- a 5 star open data gold standard knowledge graph of software mentions in scientific articles. *Proceedings of the 30th ACM International Conference on Information & Knowledge Management*, 4574–4583. https://doi.org/10.1145/3459637.3482017

Science staff. (2011). Challenges and opportunities. *Science*, *331*(6018), 692–693. https://doi.org/10.1126/science.331.6018.692

Serghiou, S., Contopoulos-Ioannidis, D., Boyack, K., Riedel, N., Wallach, J., & Ioannidis, J. (2021). Assessment of transparency indicators across the biomedical literature: How open is open? *PLoS Biol*, *19*(3), e3001107. https://doi.org/https://doi.org/10.1371/journal.pbio.3001107

Sevryugina, Y. V., & Dicks, A. J. (2022). Publication practices during the covid-19 pandemic: Expedited publishing or simply an early bird effect? *Learned Publishing*, *35*(4), 563–573. https://doi.org/https://doi.org/10.1002/leap.1483

Silvello, G. (2018). Theory and practice of data citation. *Journal of the Association for Information Science and Technology*, *69*(1), 6–20. https://doi.org/https://doi.org/10.1002/asi.23917

Tenopir, C., Dalton, E. D., Allard, S., Frame, M., Pjesivac, I., Birch, B., Pollock, D., & Dorsett, K. (2015). Changes in data sharing and data reuse practices and perceptions among scientists worldwide. *PLOS ONE*, *10*(8), 1–24. https://doi.org/10.1371/journal.pone.0134826

Trienes, J., Trieschnigg, D., Seifert, C., & Hiemstra, D. (2020). Comparing Rule-based, Feature-based and Deep Neural Methods for De-identification of Dutch Medical Records. *Proceedings of the 1st ACM WSDM Health Search and Data Mining Workshop (HSDM2020)*.

Vandewiele, G., Dehaene, I., Kovács, G., Sterckx, L., Janssens, O., Ongenae, F., De Backere, F., De Turck, F., Roelens, K., Decruyenaere, J., Van Hoecke, S., & Demeester, T. (2021). Overly optimistic prediction results on imbalanced data: A case study of flaws and benefits when applying over-sampling. *Artificial Intelligence in Medicine*, *111*, 101987. https://doi.org/https://doi.org/10.1016/j.artmed.2020.101987

Westergaard, D., Stærfeldt, H.-H., Tønsberg, C., Jensen, L. J., & Brunak, S. (2017). Text mining of 15 million full-text scientific articles. https://doi.org/10.1101/162099

Yasunaga, M., Leskovec, J., & Liang, P. (2022). Linkbert: Pretraining language models with document links. *Association for Computational Linguistics (ACL)*.

Yousuf, R. B., Biswas, S., Kaushal, K. K., Dunham, J. W., Gelles, R., Muthiah, S., Self, N., Butler, P., & Ramakrishnan, N. (2022). Lessons from deep learning applied to scholarly information extraction: What works, what doesn't, and future directions. *ArXiv*, *abs/2207.04029*.

Zhao, H., Luo, Z., Feng, C., Zheng, A., & Liu, X. (2019). A context-based framework for modeling the role and function of on-line resource citations in scientific literature. *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, 5206–5215.

Funding information The extension of the French Open Science Monitor (BSO) presented in this document was supported by France Relance/NextGenerationEU funding.

Previous work on the Softcite dataset and the Softcite Software mention recognizer was supported by the Alfred P. Sloan Foundation, Grant/Award Number: 2016-7209 (2018-2020) and the Gordon and Betty Moore Foundation, Grant/Award Number 8622 (2021).

DataStet is a continuation of dataseer-ml, developed by one of the co-author in the context of the DataSeer project (2019-2020), supported by the Alfred P. Sloan Foundation.

Author contributions All authors have contributed equally.