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

Widening the French Open Science Monitor scope

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

[1]French Ministry of Higher Education, Research, Paris, France
[2]University of Lorraine, France
[3]science-miner, France
[4]Inria, France

January 2023

Abstract

There is today no standard way for referencing research datasets and research software in scientific communication. Emerging editorial workflows and supporting infrastructures dedicated to dataset and software are still poorly adopted by 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 satisfactory accuracy across multiple research areas and document types, combining full-text harvesting, mention extraction, context characterization and corpus-level analysis.

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 data 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 datasets, are available under open licences.

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

1. Introduction

Motivations

(note: to be reviewed, it is too circular)

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 [8]. Nearly half of the researchers commonly use data generated by other scientists [9] and the vast majority of researchers support data sharing [10]. The critical role of research data and software is today broadly acknowledged, in particular for better supporting the reuse and the reproducibility of research results [4].

However, in contrast with the well established practice of citing publication, the visibility of research datasets and software is considered largely insufficient. Software is not cited in scholarly publications in a consistent and easily readable manner [3]. When they exist, the PID associated to software are not used (Du et al. 2022). Much of the published data is still essentially unavailable for integration into secondary data analysis and evaluation of reproducibility (). The deposited data is also be incomplete, sometimes intentionally and fragmented (). Similarly as for software, PID associated to data are almost not used to reference datasets in publications ().

Multiple initiatives took place in last decade to address this issue, in particular focusing on improving dataset and software cataloging [], advocacy efforts and standards for data and software citation [], with limited impact until now (Du et al. 2022).

Among the main reasons for this limited impact, we can mention the lack of incentive for researchers to invest time on work not credited and not considered for career and promotion, the fragmentation of policies, standards, infrastructures and workflows, and the absence of investment by most of the scientific publishers to implement standard for referencing datasets and software.

To address these pitfalls, it is considered that more pro-active and voluntarist policies are necessary to enforce higher standards of openness and visibility for all research results. National Open Science policies are currently rapidly developing... To measure the effect of these policies and adapt them to maximize their adoption, monitoring tool and dashboards are crucial...

However, the usage, creation, and sharing of information for the large majority of research data and software are only available in narative forms inside the content of scientific publications...

The French Open Science Monitor

The French Open Science Monitor, also called BSO for 'Baromètre de la Science Ouverte', is a tool for monitoring and steering the public policy linked to the first French National Plan for Open Science [7]. A first version of the tool was launched in 2019 providing measurements of the rate of Open Access publications produced by all public French research entities. A follow-up second Plan for Open Science [6] has started to further promote and develop the French open science policy, including a focus on research data and research software. The French Open Science Monitor is updated every year [1], however measurements related to research datasets and software were not covered yet.

The extension of the French Open Science Monitor to research datasets and software was funded following a call for projects within the framework of the French Recovery Plan. The University of Lorraine has been asked by the Ministry of Higher Education and Research (MESR) to lead this project alongside the MESR's Department of Decision Support Tools and Inria.

Quality criteria for Open Science indicators

What are the quality criteria for useful and reliable indicators on the production of research datasets and research software ?

- coverage
- accuracy
- freshness
- adaptibility to different geographical and organizational levels
- adaptibility to different scientific and technical domains
- fair: maintain consistency in term of domains and languages

Why the existing Open Science monitors related to data and software fail regarding these criteria? e.g. OpenAire. Extremely limited coverage leads to extremely biased and unreliable indicators, leading to counter-productive dashboards (for instance indicating that a handful of datasets is produced at the scale of a whole University during one year). Publishing aggregated dashboard on unreliable and non-representative data can lead to disengagement of the public for the tool, false interpretation, wrong public policy decisions and unability to assess public the application of Open Science policies.

Why do we think an automatic data-centric approach to capture the actual scientific activity and production could lead to higher quality indicators than the manually referencing approach? Because it can offer a more trustful snapshot of the scientific production, but it needs to be reliable enough (precision and coverage) and applied to a very significant amount of scientific publications.

Challenges

In view of the slow adoption of more comprehensive formal and manual reference workflows, mining dataset and software mentions in scientific publications appeared early as a possible solution to increase at scale the visibility of these new key research products.

Description of prior works... dictionary/term search, regular expressions, domain-specific rules, restriction to XML full texts availability. Lack of annotated data and reliable evaluations.

Characterizing the data and software usage and sharing: Often too much "publication centric" (e.g. is the publication uses/shares some data and code? versus what are the data/code product and how these data and software are they shared and reused?)

Lack of quality metadata and considerable fragmentation of source code repositories.

Which indicators should be selected to provide meaningful information on data and software production and openness: how the reliability and coverage of automatic extraction can influence these indicators? Which minimal information should be extracted to ensure clear and trustworthy indicators for visitors of the BSO platform?

2. Method

2.1 Machine Learning for mention detection and characterization

2.2 Research software

[5] [11] [2]

2.3 Research datasets

2.4 Matching and aggregation

2.5 Monitoring indicators construction

To help steer French public policy, the BSO must have indicators that meet several conditions:

- indicators that evolve "quickly" according to changes in uses and practices (indicators with little temporal inertia)
- indicators that are easily understandable and interpretable
- indicators whose floor and ceiling are known in advance and which are achievable.

For example, for the publications component, the BSO looks at the percentage of open access publications for the previous year: each year the new indicator does not depend on previous years. It varies between 0% and 100%, and 100% is the target for 2030.

Although the subject is more complex, it is important that the same applies to indicators for research data and software. The detection work uses the full-text PDFs of the publications as raw material. It is therefore natural to 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 each time dataset by software. A "naive" indicator would be, for example, the proportion of publications that share a dataset :

$$I_{naive} = \frac{\text{number of publications that 'shares' a dataset}}{\text{total number of publications}}$$

This indicator is simple, but has several shortcomings:

- first, concerning the denominator, not all publications can be analysed by Softcite / DataStet, because the PDF could not systematically be downloaded (closed access for which we do not have a subscription, missing PDF, etc ...). It should therefore be replaced at least by the number of publications that have been analysed.
- secondly, we want to have an indicator whose upper bound has a meaning linked to the objectives of the public policy. For example, an article for which the research work did not require the use of any dataset cannot share it.

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 a modified key indicator:

$$I_{share} = \frac{\text{number of publications that use, create and share a dataset}}{\text{number of publications analyzed that use and create a dataset}}$$

The disadvantage of reasoning in this way is that we lose some of the lessons that could be learned from the non-actionable part of the scope (publications that do not create a dataset). This is why we propose to complete the analysis with two additional indicators:

$$I_{create} = \frac{\text{number of publications that use and create a dataset}}{\text{number of publications analyzed that use a dataset}}$$

and

$$I_{use} = \frac{\text{number of publications that use a dataset}}{\text{number of publications analyzed}}$$

These three indicators therefore make it possible to have a global and decomposable view of all the cases: the top of the pyramid, which is actionable, and which we want to increase. The other two are more descriptive, but will provide a better understanding of practices and uses, particularly by breaking them down by subject area.

We note in particular that we find the naive indicator initially proposed as follows:

$$I_{naive} = I_{share} \times I_{create} \times I_{use} \times P_{analyzed}$$

where

$$P_{analyzed} = \frac{\text{number of publications analyzed}}{\text{total number of publications}}$$

All these indicators can be broken down by the components linked to the publications themselves: year of publication, disciplines, publishers etc. Monitoring by year of publication (rather than the full stock of publications since the 2010s) allows for a responsive indicator.

We supplement these indicators with a simpler one, linked simply to the presence, in the structure of the full text, of an explicit Data Availibility Statement paragraph (which in no way guarantees sharing) but nevertheless makes it possible to observe the evolution of this practice (which is very much linked to the publication's editor).

3. Implementation

Architecture diagram

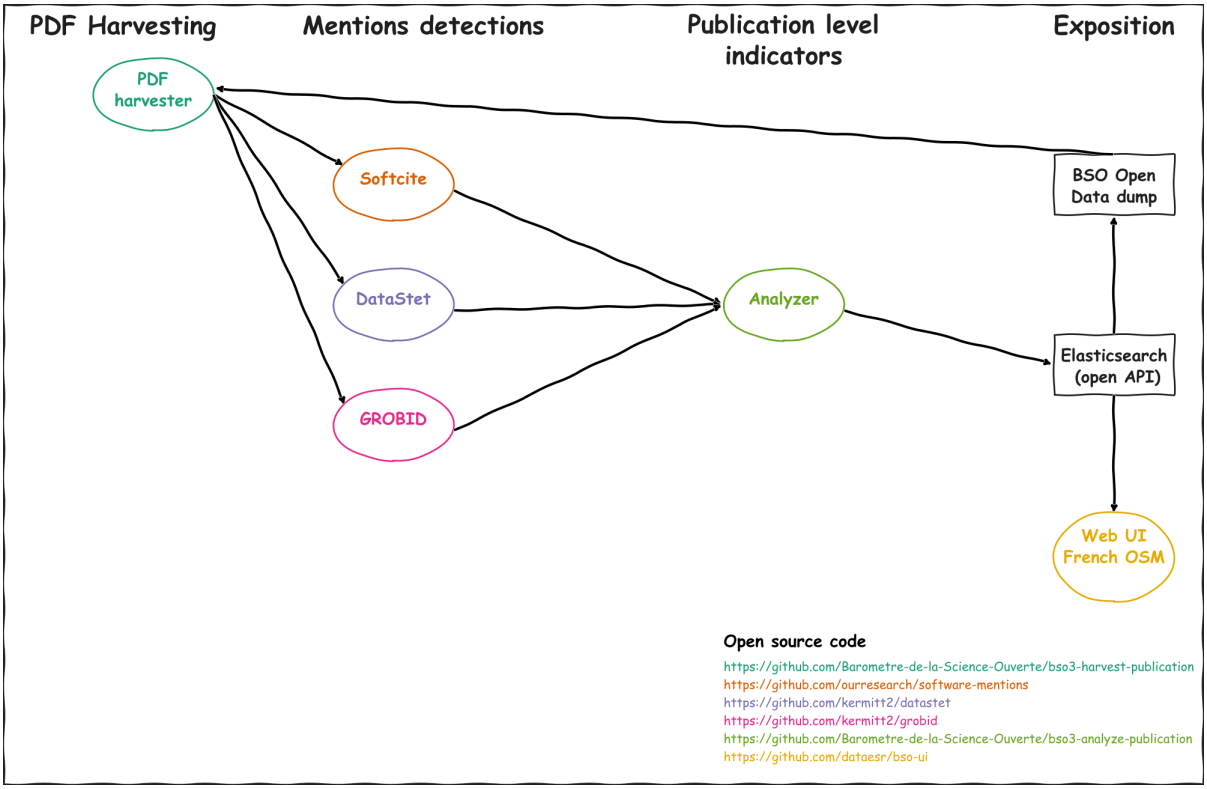

Figure 1: Global overview of the publications data flows

Workflow description

Infrastructure and runtime indications

4. Results

4.1 Full text harvesting

is_oa	number of publications	number of PDF downloaded	% download success
True	698,610	616,733	88 %
False	659,584	272,070	41 %
Total	1,358,194	888,803	65 %

harvester_type	number of PDF downloaded	% of total
standard	559,332	63 %
Elsevier TDM API	200,155	22 %
arXiv	78,078	9 %
Wiley TDM API	51,238	6 %

4.2 Dataset and software mention extraction

4.3 Research product deduplication

4.4 French Open Science Monitor indicators and dashboards

...

One of the ojectives 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 to a particular organization. The ability to measure the research activity at different scale, from a given organization to the national level has led to the creation of a BSO user group (https://groupes.renater.fr/sympa/info/bso-etablissements).

5 Limitations and future work

5.1 Limitations

5.2 Future work

5.2.1 Domain coverage

5.2.2 Large-scale research entity disambiguation

5.2.3 Local Open Science Monitor

Data and software availability

The data resulting from this work are publicly available on the French Ministry of Higher Education and Research open data portal:

The source code used for the French Open Science Monitor is available on GitHub, and shared with open source licences.

Acknowledgements

Thank you to the readers.

References

[1] Bracco, L. et al. 2022. Extending the open monitoring of open science: A new framework for the French Open Science Monitor (BSO). (2022).

[2] Du, C. et al. 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 (Jul. 2021), 870–884. DOI:https://doi.org/10.1002/asi.24454.

[3] Howison, J. and 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 (2016), 2137–2155. DOI:https://doi.org/10.1002/asi.23538.

[4] Laurinavichyute, A. et al. 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, (2022), 104332. DOI:https://doi.org/https://doi.org/10.1016/j.jml.2022.104332.

[5] Lopez, P. et al. 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* (New York, NY, USA, 2021), 3986–3995.

[6] MESR 2021. 2nd National Plan for Open Science.

[7] MESR 2018. National Plan for Open Science.

[8] Philippe, O. et al. 2019. softwaresaved/international-survey: Public release for 2018 results. Zenodo.

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

[10] Tenopir, E.D.A.A., Carol AND Dalton 2015. Changes in data sharing and data reuse practices and perceptions among scientists worldwide. *PLOS ONE*. 10, 8 (Aug. 2015), 1–24. DOI:https://doi.org/10.1371/journal.pone.0134826.

[11] GROBID. GitHub.