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

. . .

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

December 2022

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

Datasets and software are today a core element of the research activities, whose increasing role is broadly acknowledged []. With the objective of better supporting the reuse and the reproducibility of research results, many initiatives to improve the visibility of research dataset and software took place in the last decade, in particular focusing on improving dataset and software cataloging [] and standards for citation [].

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

To address these pitfalls, and as a way to enforce higer 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. . .

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 (MESRI 2018). 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 (MESRI 2021) 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 (Bracco et al. 2022), 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.

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 and code

(Du et al. 2021)

2.3 Research datasets

2.4 Matching and aggregation

3. Implementation

Architecture diagram

Workflow description

Infrastructure and runtime indications

4. Results

4.1 Full text harvesting and affiliation extraction

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, Research and Innovation 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

Bracco, Laetitia, L'Hôte, Anne, Jeangirard, Eric, and 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.

Du, Caifan, Johanna Cohoon, Patrice Lopez, and James Howison. 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): 870–84. https://doi.org/10.1002/asi.24454.

MESRI. 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.

———. 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.