A new framework for the French Open Science Monitor

Anne L'Hôte[1] and Eric Jeangirard[1]

[1]French Ministry of Higher Education, Research and Innovation,
Paris, France

November 2021

Keywords: open science, open access, unpaywall, clinical trials

1. Introduction

2. Method

2.1 Publications

2.1.1 Perimeter definition

2.1.1.1 French Open Science Monitor The French Open Science Monitor is a tool that aims at steering the Open Science policy in France. As such, it produces statistics that are analyzed over time and it has to focus on "French" productions. Also, as stated in (COSO 2018), we want to use only public or open datasources. Two constraints of perimeters thus appear naturally :

- **only publications with at least an author who has a French affiliation** are considered. The nationality of the authors does not come into play. Still, this raises the issue of access to affiliation information. Affiliation metadata are present in specific sources, like PubMed, but very rarely in the whole Crossref data. To fill in the gaps, we propose to crawl the affiliation information displayed publicly from the publications webpages. On top of that, identifying a country from an affiliation text is not that straightforward. If you are not convinced, think about an affiliation stating "Hôtel Dieu de France, Beirut, Lebanon": this does not refer to a French affiliation even though the word "France" is present. We use an automatic detection algorithm, based on Elasticsearch, described in (L'Hôte and Jeangirard 2021), to infer the countries from the affiliations texts.
- **only the publications with a Crossref DOI** are considered. Dupli-

> cates have to be avoided, in order not to count twice (or more) a publication and thus add a bias to the statistics tha are produced. It is then key to use a Persistent IDentifier. Also, we choose to use Unpaywall data for Open Access (OA) discovery. This service produces open data and offers the possibility to snapshot the whole database, which is an asset to analyse the OA dynamics. For now, Unpaywall focuses only on Crossref DOI, which leads us to adopt the same perimeter. We are aware this is a bias against some disciplines, Humanities and Social Sciences for example.

All genres of publications are considered (journal articles, proceedings, books …) as long as the publication is associated to a Crossref DOI. Many types appear in the metadata, but for clarity, we group them in categories, namely journal articles, proceedings, preprints, book chapters, books, the rest being grouped in a category 'Others'. It is important to note that the 'preprint' type does not appear as such directly in the available metadata (it is generaly declared as journal article). So preprint detection is based on the dissemination platform information. At the time of writing, only the Cold Spring Harbor Laboratory (BioRxiv, MedRxiv) case is covered, but it can be extended as soon as another preprint dissemination platform would start using Crossref DOIs.

2.1.1.2 French Open Science Monitor in Health The French Open Science Monitor also introduces a focus on the Health domain. Delimiting a clear perimeter for Health is not very easy. For now, we simply have chosen to consider **in the scope all PubMed publications, and only these**. The publications data used in the French Open Science Monitor in Health is then a subset of the publications descried above, adding the PubMed presence criterion. Note that "Health" is seen more as a domain than a discipline. In fact, publications from several disciplines are taken into account in the French Open Science Monitor in Health. A domain-specific set of disciplines is used in the French Open Science Monitor in Health, as described below.

2.1.2 Open access dynamic

From the first edition of the French Open Science Monitor, it was clear that the estimate of the open access rate should not be static but should try to capture the dynamics of the opening (Jeangirard 2019). Indeed, the 0-day open access exists but we cannot assume it represents the totality of the open access. Therefore, for a given set of publications, say the publications published during the year Y, it makes sense to measure the open access rate at different point in time, for example at some moment in year Y+1, Y+2 …

To do so, it becomes necessary to historicize the database containing the open access information. So, instead of maintaining a database that keeps track of the opening of each publication, like Unpaywall is doing, we have to make regular snapshots of the whole Unpaywall database. Each snapshot is used as an observation date to measure the open access rate. It is important to note that this method natively embeds the potentials open access discovery errors from the underlying Unpaywall database, that can be false negative (a publication is actually open at this point in time but it not detected) or false postive (wronly seen as open whereas it is closed).

This method of analysis therefore reveals two temporal dimensions: publication dates and observation dates. Obviously, the observation date must be after the publication date. To avoid that the proliferation of possible analyses blurs the message, we propose to look mainly at two elements :

- A main statistics that is the **1Y Open Access rate**: it represents the open access rate of the publications published during year Y and measured (observed from the snapshot of the OA discovery database) at point during year Y+1 (generally in December if the data is available).
- Also the **shape of open access curve** (open access rate function of the publication year). From an observation date to another, the evolution of the shape gives an insight of the speed of opening. An inverted U curve means the open access rate is lower for recent publications. Flat curves means the open access rate is the barely the same, whatever the age of the publication. Increasing curve instead would mean recent papers are more and more open.

2.1.3 Discipline and language impact

All discplines and publication languages are covered. Again, however, no metadata exists for describing the discipline or the publication language. To enrich the metadata, we then rely on machine learning approches, that try to infer discipline and language from the available metadata.

For the language detection, only the title, and the abstract if available are used, with the lid.176.bin fasttext word embedding machine learning algorithm (Joulin et al. 2016).

Discipline detection also uses journal and keywords metadata if available. A general classifier is implemented for all domains, it classifies the publications into 10 macro disciplines: Mathematics, Chemistry, Physics & astronomy, Fondamental biology, Medical research, Computer sciences, Earth science ecology energy & applied biology, Humanities, Social sciences, Engineering. It is trained on data from the Pascal & Francis database and uses a Fasttext classifier. More details are discussed in the previous paper (Jeangirard 2019).

A domain-specific classifier is implemented for the Health domain. It classifies the publications into 17 disciplines, built from the Fields of Research taxonomy. The full methodology is detailed in (Jeangirard 2021).

The main purpose of these metadata enrichments is to be able to analyse the open access rate in function of languages and disciplines. We expect to observe differences not only in the global OA rate (which discipline is the most open ?), but also in the dynamics trends (which discipline show the strongest increase over time ?) or in the opening uses (relying on publihser hosted open access versus open repositories).

2.1.4 Publishers and dissemination platforms strategies

2.1.4.1 Identification of the dissemination platforms

2.1.4.2 Business models Diamond DOAJ, Gold, Hybrid

2.1.4.3 Licences licence info from unpaywall normalized into cc-by …

2.1.4.3 Article Processing Charges (APC) estimation estimation from openAPC and DOAJ for hybrid and gold (by definition 0 for diamond).

2.1.5 The role of the open repositories

normalization of the oa_locations data from unpaywall. Beware, mixes with preprint servers.

2.1.6 Other impacts on open access

2.1.6.1 Funding PubMed gives info on grant declaration. When this info is available, can we observe an impact on measured OA ?

2.1.6.2 Main authors affiliation country Info from PubMed (only health domain). Main authors : first and last Correlation between OA rate and affiliation country of main authors ? -> link to funding and OA mandates

2.2 Clinical trials and observational studies

2.2.1 Perimeter

clinicaltrials.org and EUCTR reconciliation based on id one of the location in France

2.2.1 Main opening indicators

Results and publication declaration Time to register the study Time to register the results

2.2.2 Lead sponsor impact

2.3 Data collection system and architecture

3. Results

4. Discussion and conclusion

4.1 Findings

4.2 Limitations and future research

Software and code availability

https://github.com/dataesr/bso-publications https://github.com/dataesr/bso-clinical-trials

Data availability

portail MESRI

Acknowledgements

WeDoData Florian Naudet (https://orcid.org/0000-0003-3760-3801) University of Rennes 1, Rennes, France

References

COSO, French Open Science Committee. 2018. “Feedback on EC Open Science Monitor Methodological Note.” https://www.ouvrirlascience.fr/feedback-ec-science-monitor/.

Jeangirard, Eric. 2019. “Monitoring Open Access at a National Level: French Case Study.” In *ELPUB 2019 23d International Conference on Electronic Publishing*. OpenEdition Press. https://doi.org/10.4000/proceedings.elpub.2019.20.

———. 2021. “Content-Based Subject Classification at Article Level in Biomedical Context.” *arXiv:2104.14800 [Cs]*, May. http://arxiv.org/abs/2104.14800.

Joulin, Armand, Edouard Grave, Piotr Bojanowski, and Tomas Mikolov. 2016. “Bag of Tricks for Efficient Text Classification.” *arXiv:1607.01759 [Cs]*, August. http://arxiv.org/abs/1607.01759.

L'Hôte, Anne, and Eric Jeangirard. 2021. "Using Elasticsearch for Entity Recognition in Affiliation Disambiguation." *arXiv:2110.01958 [Cs]*, October. http://arxiv.org/abs/2110.01958.