How to build an Open Science Monitor based on publications? A French perspective

Laetitia Bracco[2], Eric Jeangirard[1], Anne L'Hôte[1], and Laurent Romary[3]

[1]French Ministry of Higher Education and Research, Paris, France
[2]University of Lorraine, France
[3]INRIA, France

December 2024

Abstract

Many countries and institutions are striving to develop tools to monitor their open science policies. Since 2018, with the launch of its National Plan for Open Science, France has been progressively implementing a monitoring framework for its public policy, relying exclusively on reliable, open, and controlled data.

Currently, this monitoring focuses on research outputs, particularly publications, as well as theses and clinical trials. Publications serve as a basis for analyzing other dimensions, including research data, code, and software. The metadata associated with publications is therefore particularly valuable, but the methodology for leveraging it raises several challenges.

Here, we briefly outline how we have used this metadata to construct the French Open Science Monitor.

Keywords: open access, open science, open data, open source

1. Requirements

1.1 Data to gather

The starting point for these analyses is a corpus of publications. Defining the appropriate target scope is essential to provide relevant insights. Details about the metadata required for this corpus are provided in Section 2. In summary, describing this corpus with a PID (Persistent Identifier) and associated metadata is crucial. The default primary PID should be the Crossref DOI. Other PIDs can be used, but the methodology and code must be adapted accordingly, particularly for Open Access (OA) status discovery.

Additional metadata are also required. A normalized scientific field classification enables the creation of KPIs by scientific domain. If such metadata are unavailable, they can be inferred using machine learning models. OpenAlex also provides computed metadata that can be leveraged. Metadata such as publishers, repositories, journals, affiliations, and publication types can support further analyses, but they must be normalized to ensure that the insights derived are meaningful.

Depending on the context, the corpus can be extracted from CRIS systems, global databases (such as OpenAlex), or even custom-built. The French Open Science Monitor (OSM) opted for the latter approach, combining open data from Crossref, PubMed, open repositories, web crawling, and bottom-up data contributions from French institutions that wish to develop their own OSM. For global data sources, country affiliation is determined based on harvested (or crawled) raw affiliation strings using

the affiliation matcher detailed in (L'Hôte and Jeangirard 2021). The code and a Docker image are available here: https://github.com/dataesr/affiliation-matcher.

For Open Access publication KPIs, an OA status discovery tool is required. By default, Unpaywall provides information for Crossref DOIs.

For indicators related to datasets and software, in addition to the metadata corpus, the full texts of publications in PDF format are also required. Text and Data Mining (TDM) techniques can then be applied to compute the KPIs.

Warning: To achieve the best possible results, it is essential to download as many full-text publications as possible. In the European context, this is feasible under the framework of the European directive allowing text and data mining for research purposes1. It is necessary to have lawful access to the downloaded content if it is not already openly accessible, for instance, via a subscription. Outside the European Union, the legal framework must be carefully reviewed. This note does not address the context beyond Europe.

1.2 Open source software used

The French Monitor code is freely available under open license (MIT License). It is modular as detailed in the infrastructure section in (Bracco et al. 2022) (for OA to publications) and in (Bassinet et al. 2023) (for datasets and software). However, it is closely linked to our data acquisition pipeline (web crawling, bottom data collection from French institutions, extension to cover French specific PID (HAL)) and to the data architecture we built on the OVH public cloud - S3 Object Storage used by the Ministry. Also, Python has been used. If it is to be implemented in another country, parts of the code will have to be rewritten to match the local requirements. The core monitor code also relies on other free and open source services / software:

- Open access discovery tool (as explained above): Unpaywall The premium service of Unpaywall is used to get a quarterly full snapshot of the database. These snapshots are used to historicize the OA status of each publication, useful to analyse the OA dynamics.
- A Text and Data Mining (TDM) tool to detect research datasets mentions from the full-text. We use DataStet from the Docker image 0.8.0
- A Text and Data Mining (TDM) tool to detect code and software mentions from the full-text. We use Softcite from the Docker image 0.8.0
- A smart scholarly PDF parsing tool to structure metadata and the full-text content from a PDF. We use GROBID from the Docker image 0.8.0. Later versions (from 0.8.1) include fixes on grant ids detection that can be very relevant.

We also developed extra modules that glues together and orchestrates all the previous tools:

- The module bso3-analyse-publications implements in Python the whole TDM pipeline, and also the analysis funnel analysis at the document level of the resulting outputs to get publication-wise KPIs.
- The module bso-publications implements in python an extract-transform-load process and stores the final results, at the publication level, in an Elasticsearch index.

1.3 Computation power consumed

We deploy our code on a public cloud infrastructure, by OVH. We use their managed Kubernetes service, using multiple nodes (servers):

- 1 large server (16 CPU 240Go RAM) is used to host the metadata Mongo databases (Unpaywall snapshots, French corpus metadata) and run most of the OA KPIs calculations.

- 6 smaller servers are used for data acquisition (crawling, parsing, harvesting) and enrichment (language detection, discipline inference, tasks monitoring ...) (4 CPU 15Go RAM and 2 CPU 7Go RAM).
- 1 medium server (4 CPU 60Go RAM) used to harvest the PDF.
- 5 servers (32 CPU 120Go RAM) to run the TDM analysis.
- 3 servers (with redundancy) (16 CPU 60Go RAM) to host the Elasticsearch resulting indices.
- 6 servers (with redundancy + staging/prod) (2 CPU 7Go) to host the website.

This infrastructure is also used for other projects, so 100% cannot be affected to the OS monitoring (in particular the Elasticsearch and websites hosting part). However, the PDF harvesting and TDM analysis are a very specific need. For this specific need, around 20k euros were spent to analyse 700k PDFs in one month. The rest of the infrastructure is around 70k a year, but is not specific to OS monitoring. Relying on other services (like OpenAlex) could help reduce the costs.

1.4 Team and human resources

A deep understanding of the Open science / scholarly communications area is key to make it happen. Random software engineers do not have this knowledge and it may take time for them to understand what is at stake for the monitoring to be relevant. Building the whole pipeline can be implemented with about 2 FTE for 6 months. An extra FTE (project manager or so) can be needed to make sure software developments are inline with the project goals. Maintenance costs are lower, about 0.5 FTE a year. However, things evoles fast and new features (new objects to monitor, new types of analysis ...) are generally necessary so a maintenance only scenario is not very likely to happen.

2. A few methodological considerations

2.1 Corpus creation

2.1.1 Defining a perimeter

Research outputs can be indexed in large databases (Crossref, Datacite, OpenAlex), or not. It can be necessary to put in place specific harvesters to get metadata from other places (disciplinary or institutional repositories for example).

Making sure these extra data are correctly ingested with the data coming from the large database is a challenge (data format, no duplicates ...). An option could be to make those extra research outputs fit into one of these databases: adding DOI for example, or asking OpenAlex to harvest extra repositories.

Also, about the distinction between types of publication, for example, the distinction with professional articles: it all depends on the available data, of course, but also the main point to define the perimeter is to know upfront what is the goal of the monitoring. If it is to steer and analyze the impact of a public policy, then the perimeter has to be in line with the public policy itself. That may be different from one national/institutional situation to another. That is also why, in France, we propose a national monitoring, but also "local" monitoring in which the perimeter can be customed by the users.

2.1.2 Handling duplicates

Detecting duplicates can be challenging. For publications with a Crossref DOI, this PID is in general enough, even though, for some cases, preprints and published versions sometimes have a Crossref DOI each and could be considered as only one research output.

For works from Datacite, that can be much more complicated. The Datacite API has a field "relatedIdentifiers" that can be used to detect duplicates.

In the example from OpenAlex one can see from the datacite API https://api.datacite.org/dois/10.5281/zenodo.8042997 that this DOI has several versions. However that may not be enough and some heuristics based on title, authors ... or even more complicated processes (machine learning etc) could be used to detect duplicates. The same goes for publications from other systems (open repositories for example) that may contain a lot of duplicates.

Also, during the deduplication phase, duplicate records may have to be merged into one single record. The list of affiliations, keywords, and authors can be different. Merging strategies have to be settled. In the French case, for key elements, Crossref data have the top priority, followed by the other sources (web scraping, PubMed, HAL, ...). For some metadata that are of type list, it is possible to enrich the final metadata with all the metadata coming from each source (affiliations, keywords for example).

2.1.3 Author disambiguation and affiliations

Grobid can help to get authors' PID (like ORCID) or raw affiliation strings and PID (like ROR) when they are present in the full text. However other sources have to be considered to get better coverage.

Affiliations can be harvested from the landing pages in many cases (from HTML Highwire header or from HTML parsing). That is one the techniques used by OpenAlex. Aligning those raw affiliation strings to ROR is another step where heuristics or machine learning can be helpful (see (L'Hôte and Jeangirard 2021)). The Works-magnet (see (Jeangirard, Bracco, and L'Hôte 2024)) is a tool designed to help improve this matching done automatically in OpenAlex. Author disambiguation is not an easy task either as the majority of authors have no ORCID at all. OpenAlex uses clustering techniques https://github.com/ourresearch/openalex-name-disambiguation/tree/main/V3

In the French case, we benefit from a large registry of persons (https://www.idref.fr/) maintained by ABES. This registry has quite a good coverage of the researchers working in France. Again, clustering techniques can be used to disambiguate author name with a PID, but a good person registry helps. These techniques are used in the French research portal scanR (https://scanr.enseignementsup-recherche.gouv.fr/) to add PID (namely idref ID) to the authors of publications.

2.2 Corpus enrichment

Once the publication corpus is defined, many enrichment are necessary to add dimensions to analyse. The use of machine-learning techniques can help a lot, but additional care is required to verify the results (see (Jeangirard 2022)).

2.2.1 Discipline classification

Different algorithms based on title, ISSN, authors etc ... can be used to classify publications. OpenAlex implements some. The French Open Science Monitoring uses its own to classify publications into 10 macro-categories (cf (Jeangirard 2019)).

2.2.2 Open Access discovery

The Open Access status information is generally not retrieved with Grobid (publication parser), but instead using information from the landing page (HTML parser) or from repositories. Unpaywall (and OpenAlex now) already implements that logic, at least for all Crossref DOI. For other publications (with no Crossref DOI), open access discovery is not that easy and depends on local specificities. In the future, we could expect that, if those publications are in OpenAlex, they also benefit from a better open access discovery service just like with Unpaywall. Also, it is important to note that the open

access status is not a fixed metadata (contrary to the title or the list of authors for instance). It can evolve over time. More details on this aspect are given in (Bracco et al. 2022).

2.2.3 Open Access types

In the French Open Science Monitor, we analyze in different ways the type of open access. In particular,

- is it opened via the publisher or via a repository (or both?) - regardless of the license
- if it is opened via the publisher, what kind of business model is it used?
- if it is opened via the publisher, is there any proper license, and which one?

More details are described in (Bracco et al. 2022).

3. A few advices and impact

Having an objective and quantitative tool for monitoring, easy to plug, which makes it easier for funders and institutions to communicate and steer their open science policy. Having a tool that can be easily adapted to different contexts like funders or institutions is then a key component. However, a tool remains a tool and is only complementary to the local policies and mandates.

Some tips also can help in building a reliable and effective Open Science Monitoring.

- Clean up OpenAlex affiliations to obtain a reliable corpus of publications (the works-magnet can help).
- Maintaining its own infrastructure is costly (money and HR). A call to open infrastructure could help reduce the costs and invest in a shared infrastructure / methodology.
- If a national dashboard cannot be created, it is already possible to obtain useful indicators on open access publications directly via COKI (see (Diprose 2023)), as it is based on OpenAlex data.
- In the event of legal and/or economic difficulties in accessing non-open access content, a downgraded version of the result based solely on open access full text and under CC licence is possible.

References

Bassinet, Aricia, Laetitia Bracco, Anne L'Hôte, Eric Jeangirard, Patrice Lopez, and Laurent Romary. 2023. "Large-scale Machine-Learning analysis of scientific PDF for monitoring the production and the openness of research data and software in France." https://hal.science/hal-04121339.

Bracco, Laetitia, Anne L'Hôte, Eric Jeangirard, and Didier Torny. 2022. "Extending the open monitoring of open science." https://hal.science/hal-03651518.

Diprose, Hosking, J. P. 2023. "User-Friendly Dashboard for Tracking Global Open Access Performance." *The Journal of Electronic Publishing*. https://doi.org/10.3998/jep.3398.

Jeangirard, E. 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.

Jeangirard, Eric, Laetitia Bracco, and Anne L'Hôte. 2024. "Works-magnet : aucune de perdue, 10 000 de retrouvées." Abes; Journées Abes 2024. https://doi.org/10.5281/zenodo.11471247.

Jeangirard, Éric. 2022. "L'utilisation de l'apprentissage automatique dans le Baromètre de la science ouverte : une façon de réconcilier bibliométrie et science ouverte ?" *Arabesques*, no. 107 (September): 10–11. https://doi.org/10.35562/arabesques.3084.

L'Hôte, Anne, and Eric Jeangirard. 2021. "Using Elasticsearch for entity recognition in affiliation disambiguation." https://hal.science/hal-03365806.