Promoting Open Science through bibliometrics

Laetitia Bracco

Promoting Open Science through bibliometrics: a practical guide to building an open access monitor

Abstract

In order to assess the progress of Open Science in France, the French Ministry of Higher Education, Research and Innovation published the French Open Science Monitor in 2019. Even if this tool has bias, for only the publications with a DOI can be considered, thus promoting article-dominant research communities, its indicators are trustworthy and reliable. The University of Lorraine was the very first institution to reuse the National Monitor in order to create a new version at the scale of one university in 2020. Since its release, the Lorraine Open Science Monitor has been reused by many other institutions. In 2022, the French Open Science Monitor further evolved, enabling new insights on open science. The Lorraine Open Science Monitor has also evolved since it began. This paper details how the initial code for the Lorraine Open Science Monitor was developed and disseminated. It then outlines plans for development in the next few years.

Key Words: bibliometrics; Open Science; monitoring; indicators

1. Introduction

The tools used to evaluate research are still often oriented towards the impact factor or h-index, as in the Web of Science. However, public policy and funding requirements are leading researchers to open their publications and data more and more. But the efforts made by researchers to open their results are still little considered in their evaluation. Bibliometrics and open science seem, therefore, to be irreconcilable enemies. On the one hand, we have a controversial set of traditional research performance indicators, which are regularly criticised by scholars (Gingras, 2014, Alstete, Beutell & Meyer 2018). On the other hand, we have Open Science, which incorporates principles of open access, sharing, and accessibility of knowledge opposed to the traditional evaluation processes, which are yet mainly bibliometrics-oriented (Cagan, 2013). If we accept that a strategic shift towards open, inclusive, and transparent research performance indicators is needed, it is useful to consider the current status of open knowledge in a nation-level case study.

In France, the reflection on monitoring Open Science started in 2018, with the publication of the first National Open Science Plan[i]. Following this strong commitment, indicators had to be set, in order to assess the effectiveness of the Plan. In 2019, the French Ministry of Higher Education, Research and Innovation published the French Open Science Monitor. This first version (Jeangirard, 2019) presented general indicators about Open Science in France: the global rate of opening of publications for the considered year, an evolution by year and the rate of opening by discipline and by publisher.

This was the first time such a tool was developed. To our knowledge, no other Open Science dashboard had been made available to institutions in France.

As part of an open data strategy, the Ministry made available all the datasets used to create the graphs, as well as the code and documentation to understand how the project had been realised. As the scripts were freely reusable, it seemed relevant to try to apply these indicators at the scale of one university, in order to assess its own Open Science policy and present an institutional-level case

study exploring how open science can be measured, understood and implemented. This paper details how the Lorraine Open Science Monitor was designed, implemented and disseminated.

2. Creating a local Open Science Monitor

2.1. A local context in favour of Open Science: The University of Lorraine

The University of Lorraine is multidisciplinary, with 60,000 students and nearly 4,000 researchers in 60 laboratories. As an institution, it is strongly committed to Open Science. Indeed, an institutional archive, HAL-UL[ii], was made openly available in 2016. In 2018, the steering committee of the University made the deposit in HAL-UL of all publications' references, with the full text for articles, mandatory.

Since 2019, an Open Science steering committee has been in place, whose objectives are to promote and develop the Open Science policy among researchers. From this steering committee, two operational committees have been created, one dedicated to research data and the other one to open publications, in which the university libraries are deeply involved.

At the University of Lorraine, the numerous university libraries are part of a large transversal service, the Documentation Department. The Documentation Department manages library policy at all levels, whether it be documentary policy, users training, research support. As part of its research support policy, the libraries have set up a network of fifteen librarians dedicated to helping researchers in the use of the open archive.

Finally, symposia dedicated to Open Science have been organised every year since 2018.

A strong political commitment and human resources have therefore been brought together in favour of Open Science, however, the effects of this policy and the progress of Open Science at the University had yet to be measured, quantified, or explained. The National Open Science Monitor was the ideal tool for it.

2.2. How to isolate the publications of a single university

The Documentation Department, strongly engaged in all matters regarding Open Science, was appointed to take a lead in adapting the National Monitor datasets so they could be applied at the University of Lorraine.

There were no preexisting skills on programming among the library's staff; these were nevertheless essential to adapt the National Monitor. The project was entrusted to a data librarian who, therefore, learned the Python programming language and was trained more specifically in data analysis to carry out this project.

The first challenge was to isolate the University of Lorraine's production. In order to build the French Open Science Monitor (hereafter the National Monitor), the first step had been to create a database gathering the metadata of all French publications from 2013 to 2019. Being an Open Science oriented project, the National Monitor could not use commercial databases, such as the Web of Science. Therefore, the Ministry's team of developers created parsers - in other words, detection tools on online scientific publications - to identify French affiliations. An equivalent work could not be done for a single university, as there are so many name variants in the affiliations, so an alternative strategy was needed.

To build the corpus of the institution, the data librarian had to choose databases in which metadata extractions were possible, to which the institution subscribed and in which the affiliations were reliable. After several attempts, the choice fell on a combination of five different sources: the Web of Science, PubMed, the open archive of the University HAL-UL, Lens.org and data from article processing charges.

Why were these databases chosen? The Web of Science is an international reference regarding scientific publications. The library regularly controls the university publications indexed in the Web of Science; therefore we can consider that this source is reliable. But the Web of Science is not designed to be comprehensive, that is why this corpus had to be completed with other sources.

PubMed is also a worldwide reference, in which the affiliations were also deemed reliable after several tests. The open archive of the University, HAL-UL, was of course an important source too, for the library staff controls and monitors every deposit. Even though some publications, that can be found on the Web of Science for instance, are not deposited in HAL, it is still an unavoidable source, as well as the list of publications for which the university paid article processing charges. Finally, Lens.org was also considered a reliable source after some tests.

These data sources do not all allow extractions in the same format. Moreover, many publications are available in several sources at the same time, for example in the Web of Science and in the university's open archive. It was, therefore, necessary to cross-reference these data, to remove duplicates, to harmonise them in order to obtain, in the end, a unified list.

The Python language is particularly adapted to this type of automated file processing. The first step consisted in writing a code allowing to read and process the data and to remove the duplicates from very heterogeneous sources, to obtain the DOI of all the relevant publications. This first step allowed us to build a corpus of about 20,500 publications for the period 2016-2020, or about 4,100 publications per year for 4000 teachers-researchers (of which only 2,400 are researchers).

This first step is already significant. Indeed, the evaluation of research in France relies heavily on publications reported in the Web of Science, which is far from being exhaustive. However, about 20% of the publications in the corpus are not in the Web of Science, which is a significant proportion.

2.3. Replicating the national graphs in Python

The second step of the project was to write the code to generate the graphs. The graphs of the National Monitor were written in JavaScript, a language for which we didn't have any in-house skills. That is why these graphs were completely rewritten in Python.

The script that identifies the open access status for each publication was written by the Ministry. Its purpose is to use the Unpaywall[iii] application which, from a DOI, is able to determine the title, the author, the date, the type of document and of course its open access status. This script was reused as is in the Lorraine Monitor. The Ministry's script that assigns a discipline to each publication was also reused. Finally, all of these processes had to be brought together into a single coherent entity.

The creation of an Open Science Monitor at the institutional level could potentially interest many other universities and research organisations in France. This is why it seemed very important to think, from the beginning of the project, about how it could be reused. Consequently, free tools to develop the code were chosen, namely Gitlab to store the data and the scripts, and Jupyter notebooks for the code itself.

The idea was to offer a ready-to-use tool that other institutions could download and reuse by simply changing the extractions from the bibliographic databases and replace these with a local dataset.

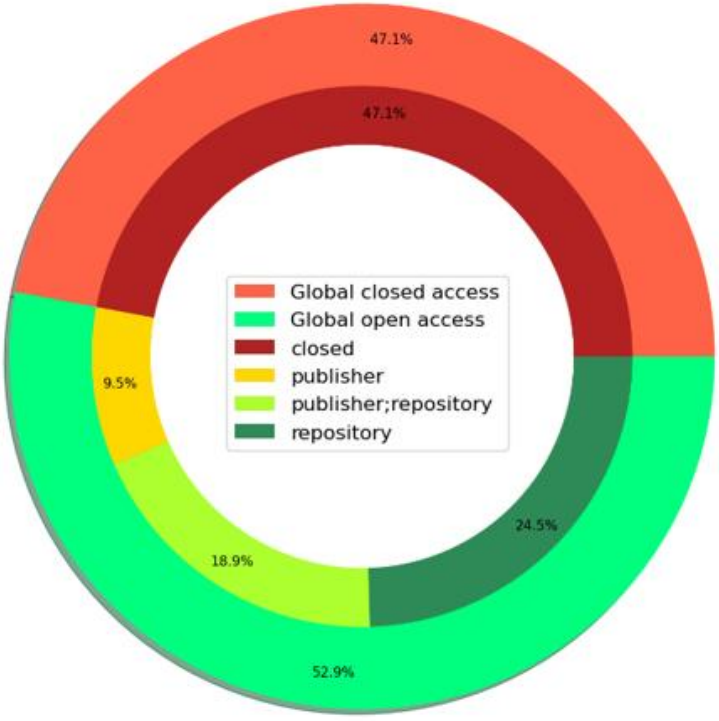
Jupyter notebooks are interfaces in which text and code can be mixed. It's very practical for beginners in programming because it allows a check at each step to test the code is working as expected. It is also very practical to explain the code to beginners. Indeed, text and images can be added to the notebooks to make the code more easily understandable.

Each line of code is annotated so that anyone, even without any programming knowledge, can understand it and act on the code, for example, by changing the name of the institution.

2.4. Which results?

Let's take a closer look at the results obtained in 2021. The Lorraine Monitor was then composed of 8 graphs, each of which had a different purpose. These graphs are no longer available on our website, since the new version is displayed. However, they are still available online on HAL-UL and on the Software Heritage Archive (Bracco, 2020).

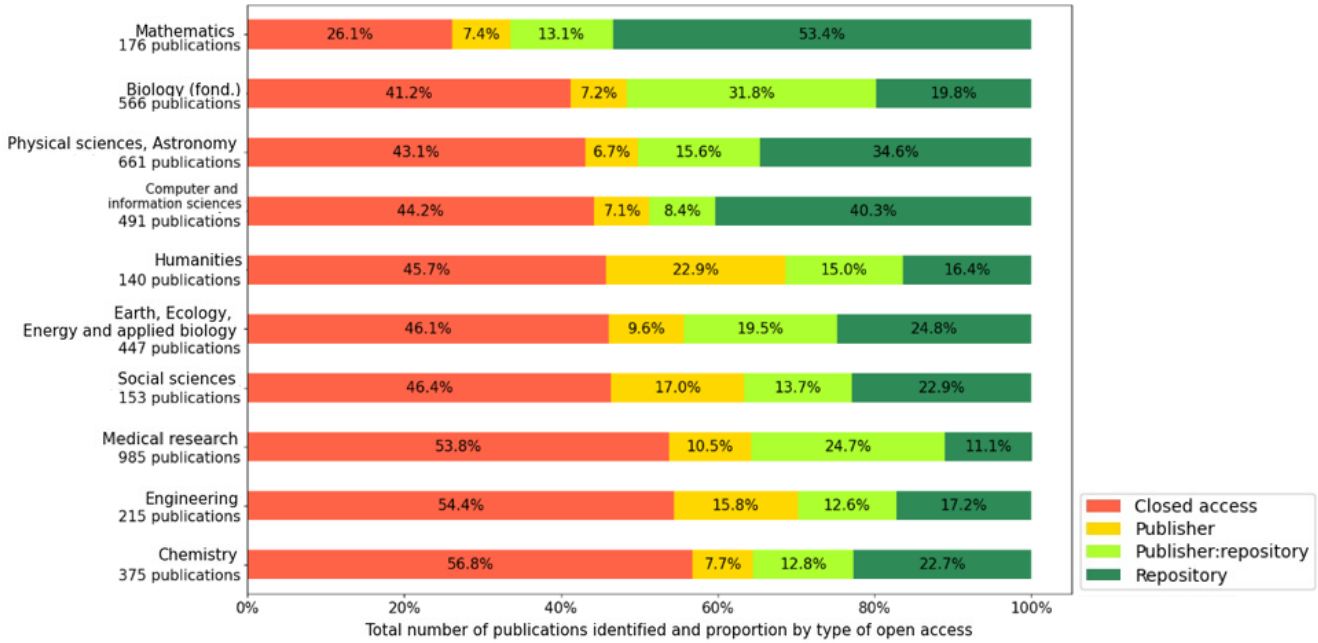
Fig. 1: Proportion of open access publications in 2019



The first graph (fig.1) gives a global view over one year. This version of the Monitor of the University of Lorraine was generated in January 2021, and yet it is the year 2019 that is highlighted. Indeed, in January 2021, many publications from the year 2020 could not legally be in open access. The French law for a digital Republic only allows the deposit of articles in their post-print version in an open archive after 6 or 12 months of embargo depending on the discipline, unless the publisher authorises it before. It would, therefore, not have been relevant to consider the year 2020 as early as January 2021. The same choice was made at the national level.

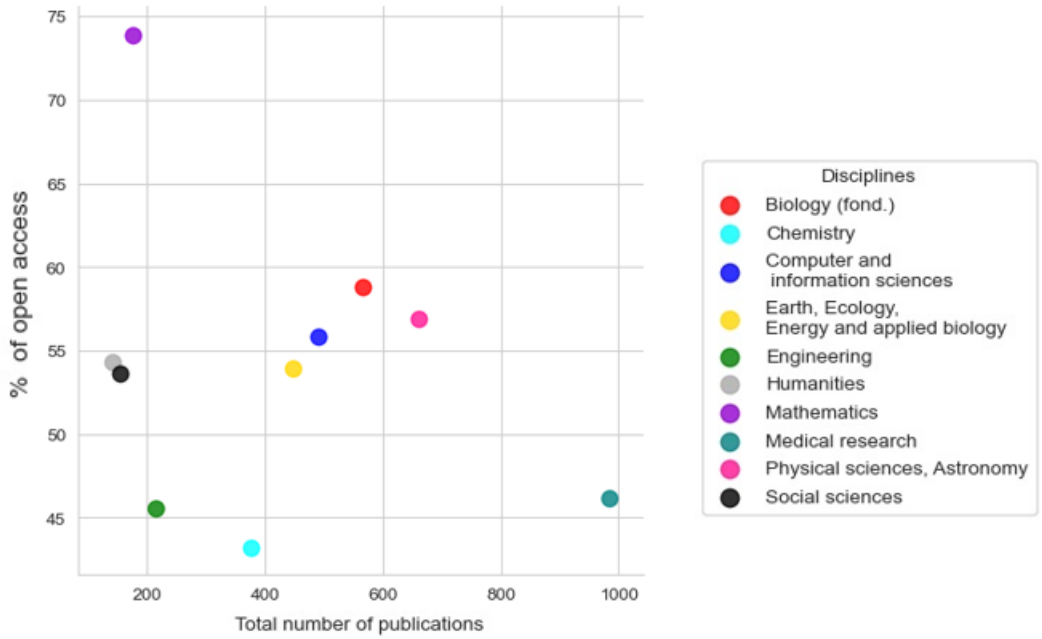
The University of Lorraine was, at the time, in line with the French dynamic, with a rate of open access that was fairly close (nearly 53% compared to 56% at the national level). The open access rate has increased to 66% in 2022, which is beyond the national level for the same period (62%). The Fig. 1 shows the different types of open access: open access in an open archive only (dark green), open access in an open archive and from a publisher at the same time (light green), open access from a publisher only (yellow). Obtaining the list of publications that are in open access only from the publisher allows, for example, to warn the authors that their publications could also be archived in the open archive, which would ensure that appropriate digital preservation policies were applied.

Fig. 2: Open access to publications by discipline in 2019



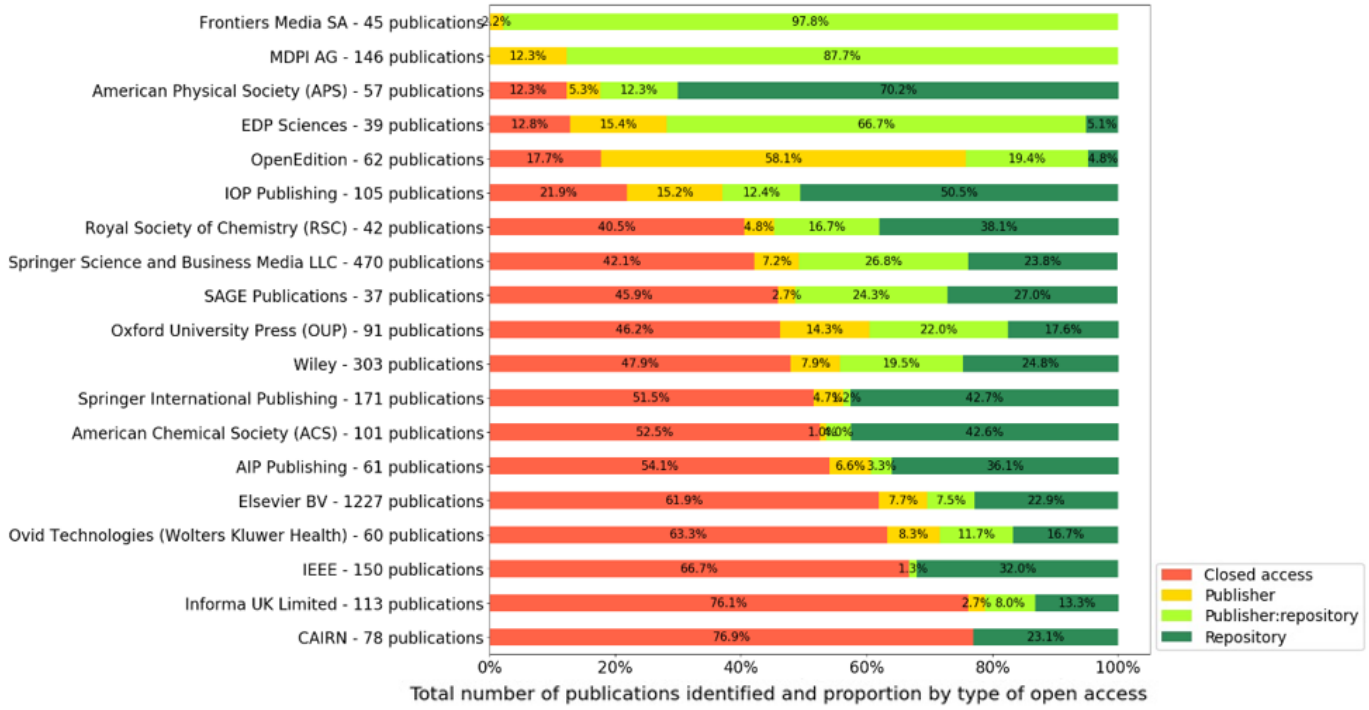
The graph by discipline (Fig. 2) is particularly illuminating because it shows how much habits can change depending on the scientific community. Mathematics led the disciplines being the most Open Science-friendly, with a 73.9% open rate for 2019 publications. On the contrary, chemistry seemed to be a relatively weak performer, with only 43.2% of publications in open access. However, it is important to measure the progress made by this discipline, and this is one of the advantages of this tool: in 2020, when the measurement was made on the 2018 publications, the open access rate was much lower, at only 29.7%.

Fig. 3: Open access to publications by discipline related to the number of publications in 2019



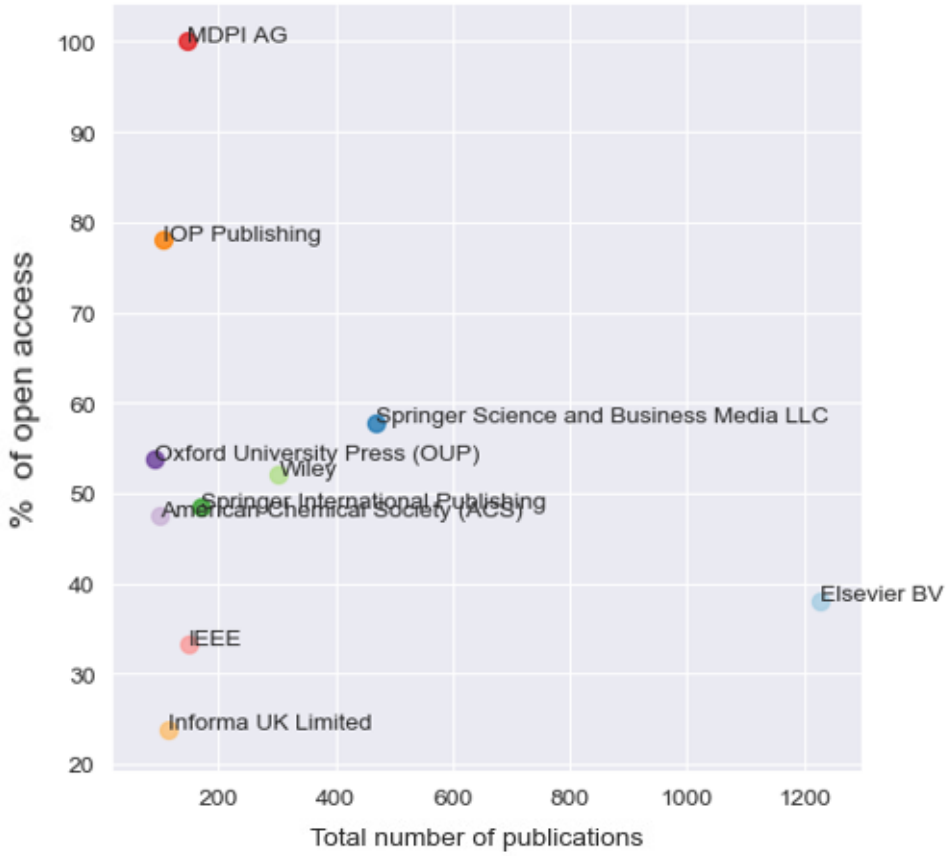
A second representation by discipline (Fig. 3) allows us to visualise these disparities even more while also underlining the lack of representativeness of the databases. Indeed, the humanities are poorly referenced. This second visualisation of the same data, which was also carried out for the publishers/platforms, was not present in the first National Monitor. It provided a more immediate vision of the important differences that can exist between scientific communities in terms of Open Science, and underlines the lack of visibility of certain disciplines (Humanities in particular) in the most used databases. This under-representation of humanities and social sciences in the monitor is due to several causes. Firstly, the databases most used in France in these fields assign few DOIs, and rarely have reliable affiliation data. Therefore, they could not be used to build the corpus. Also, publications in humanities and social sciences, especially French-speaking, are not very prominent in bibliographic and bibliometric databases such as the Web of Science.

Fig. 4: Open access to publications by publisher or platform in 2019



The graph by publisher or platform (Fig. 4) allowed us to see where our researchers chose to publish. Here again, there are strong disparities, for example, between MDPI, a publisher of scholarly open access journals, whose publications are 100% open access, and the Institute of Electrical and Electronics Engineers, whose publications are overwhelmingly in closed access.

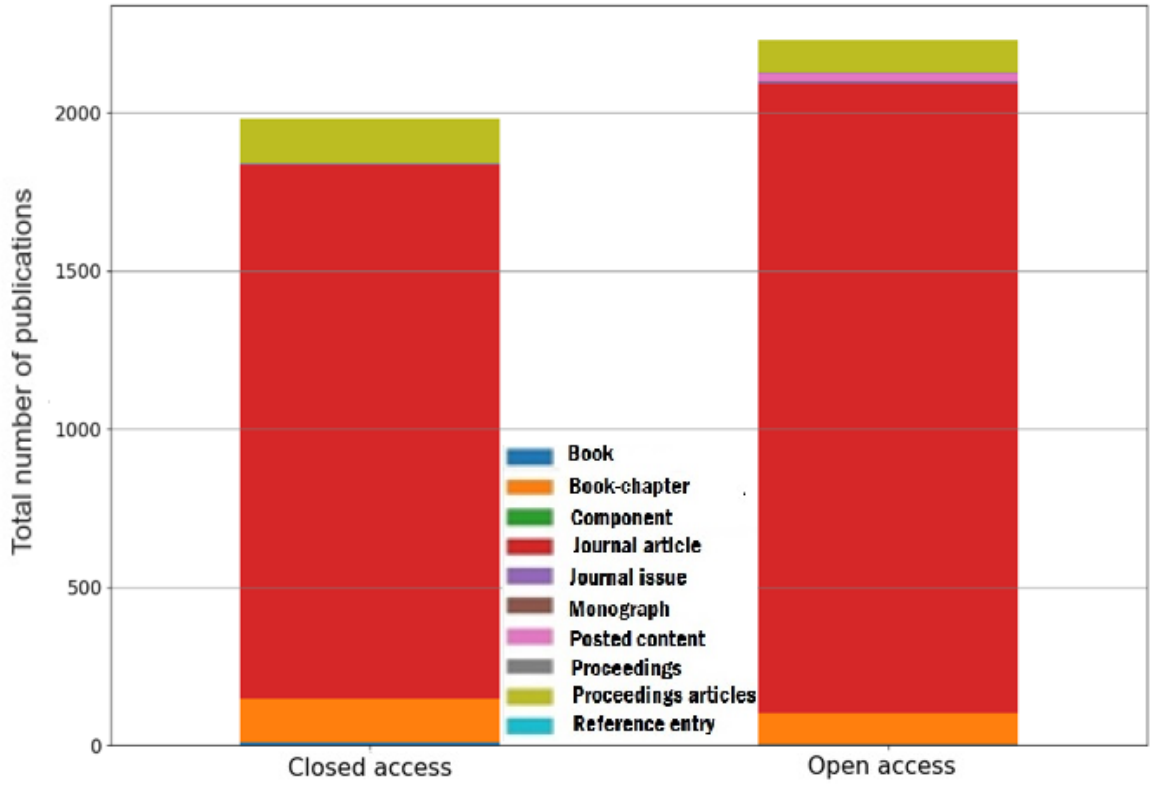
Fig. 5: Open access to publications by publisher or platform related to the number of publications in 2019



As for the graph by discipline, this graph by publisher/platform (Fig. 5) has been enriched with a second visualisation that highlights the crushing domination of Elsevier in the scientific production of Lorraine, since it alone represents one-third of the publications of the year.

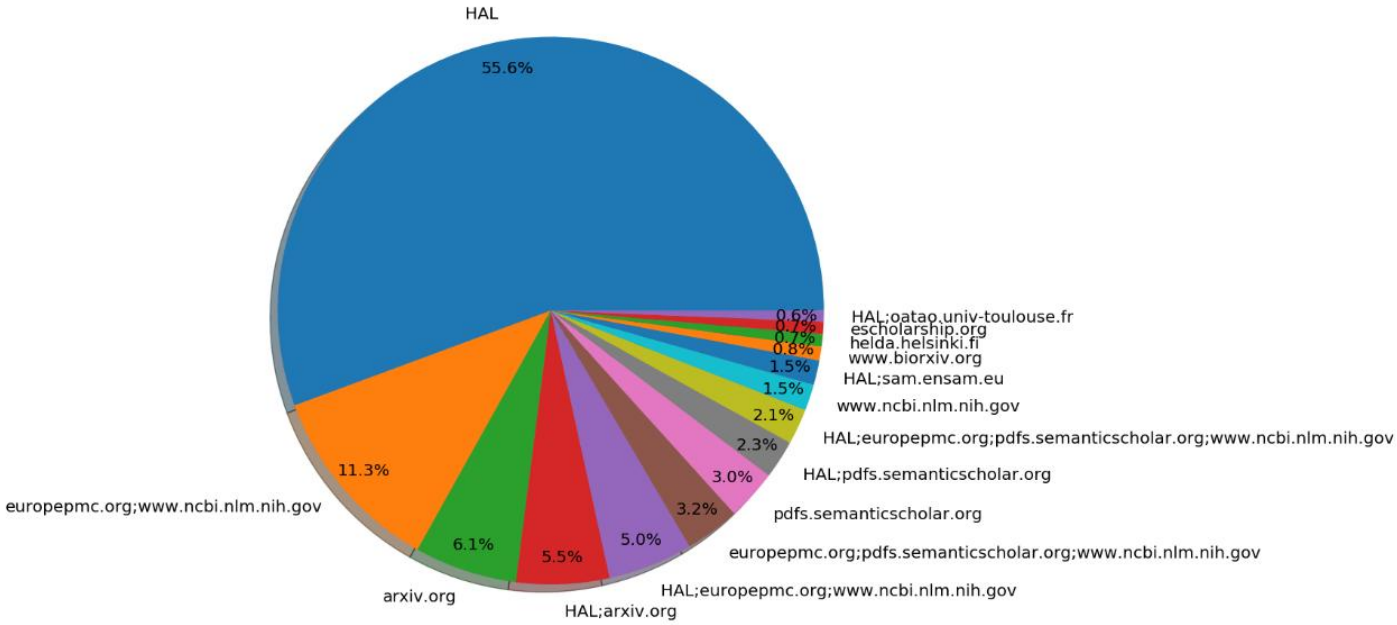
In addition to the comparison between publishers, this graph allows us to know where the researchers of the University of Lorraine publish the most (from right to left). This graph can help us to make a choice when analysing subscriptions: to end the subscription to Elsevier, for example, would not have neutral consequences in terms of access to our own scientific production. The necessity to deposit full-text publications in our open archive is only more obvious. However, even though it became mandatory in 2018, lots of our researchers' articles are only reported in HAL-UL, without the full text.

Fig. 6: Open access to publications by type of document in 2019



The approach by type of publication (Fig. 6) shows with force the domination of the journal article within the databases, whereas other types of research products (books, for example) are less prominent, and this can be explained by several factors. We have seen that the humanities and social sciences are poorly represented in the corpus: traditionally, researchers in hard sciences produce mainly articles. In addition, other types of publications (posters, books, theses...) are less often provided with a DOI. Finally, we can also conclude that the domination of the journal article is a result among many others of the policy of quantitative evaluation of research: it is more rewarding to have many publications, even short ones.

Fig. 7: Ranking of the most represented open archives in 2019



A final visualisation (Fig. 7) allows us to identify the self-archiving platforms preferred by our researchers. As at the national level, HAL is predominant, but there is also a practice of multi-platform deposit.

In order to understand the data generated by the Lorraine Open Science Monitor, a few points should be considered. Only publications with a DOI are usable because it is on this criterion that Unpaywall works. There are, therefore many publications that are not considered. It is difficult to determine precisely the percentage of missing publications in the corpus. We can nevertheless give an approximation from the HAL-UL open archive. For the year 2016, 2951 publications with a DOI are integrated in the corpus. However, there are 6312 publications from 2016 in HAL-UL. We can therefore consider that the Barometer gives us a trend on the progression of open access, but it cannot establish a precise quantitative evaluation of the scientific production of the institution. The other bias is that humanities and social sciences disciplines are less well represented in the databases. These two biases are also present in the National Monitor, the comparison between the University and the National Monitor, therefore, remains relevant.

3. Disseminating the code and drawing conclusions

3.1. An open and regularly improved code

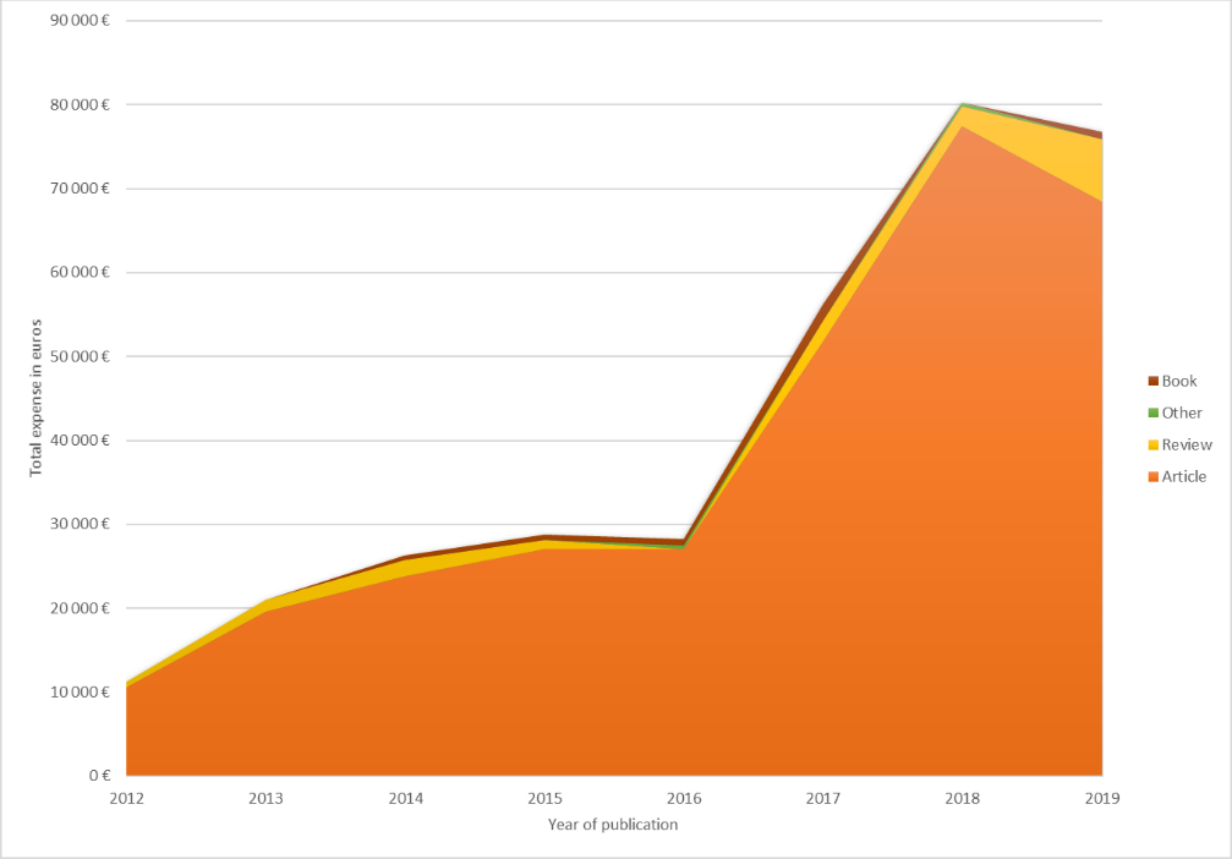
For its commitment to Open Science, it was very important for the University of Lorraine to open this code as much as possible. Immediately after the release of the code, several universities, research organisations and colleges contacted us in order to reuse it with their own data.

The code being exposed with a readme file, and the data librarian being available to answer questions, many institutions were able to reuse the code and to share it. As at early 2022, more than

fifteen French institutions have already reused the code and most of them have published their own version. For now, there has been no reuse outside of France; but since the code is applicable to international databases, it could also be reused by other nations' institutions. It has not been done so far, but two Swiss institutions have formulated the wish to know more about the code so that they could reproduce it.

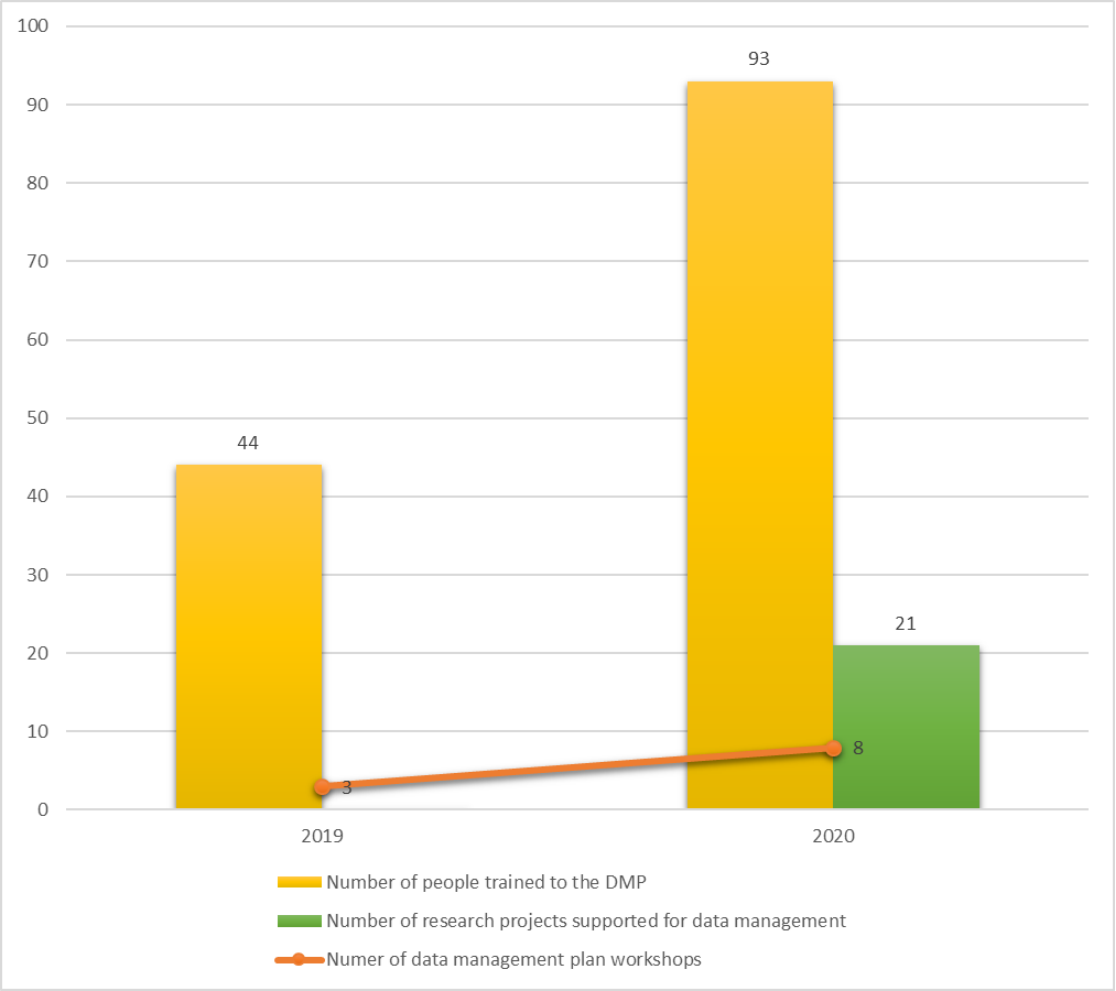
Additional graphs were made with Excel and added to the Lorraine Open Science Monitor in spring 2021 to further enhance the availability of relevant indicators. These new graphs focus on the costs of article processing charges and on Open Science training indicators, data that are necessarily very specific to the institution and cannot be reused by others.

Fig. 8: Volume of real expenses by year and type of publication 2012-2019



With Fig. 8, we can note a strong increase in costs of open access over time.

Fig. 9: Training and support for researchers on research data management 2019-2020



Finally, in order to measure our Open Science training activities, three graphs were added (Fig. 9): training of doctoral students, training, and support for the data management plan, training and support for the use of the open archive.

3.2. What to do with these figures?

From these graphs, some lessons can be outlined for the future.

The continuous increase in the rate of open access to publications is the product of a constant effort that must be maintained: the awareness and training of researchers in Open Science require a sustained effort and human resources. Openness is still far from being automatic in scientific production.

Some disciplines could be the object of particular attention, notably through more recurrent training: although certain scientific communities, such as mathematics, have a long practice of publishing in open access, others are still extremely reluctant.

The majority of publishers at the University of Lorraine are consistent with a worldwide trend of editorial concentration, which must be diversified: a third of the articles published by researchers at the University of Lorraine are published by Elsevier.

Unsubscribing from the more dominant publishers, especially where this involves disciplines with a lower number of deposits in the institutional archive, may result in the end of access to our own researchers' publications. Before the Monitor was published, the University of Lorraine has chosen to unsubscribe from Springer and the Institute of Electrical and Electronics Engineers. With the Monitor, we can see that these two publishers are widely represented in our scientific production. It is thus essential to encourage the researchers to deposit the full text of their publications in HAL-UL to allow everybody to have access to it now that we do not have a subscription to these journals anymore.

Non-journal publications must be more widely referenced through the systematic use of a DOI. This can only be done gradually by changing the practices of book publishers in particular. It would be interesting to have a national or even international strategy in this field.

The Lorraine Open Science Monitor is a tool that allows the University to measure the progress of open access within its scientific production. In this respect, it constitutes a set of original indicators complementary to traditional bibliometrics. Designed from the outset as a tool that can be reused by other institutions, it should be enriched as it is reused.

It is not intended to evaluate research, but to become a tool to help an institution establish an effective action plan to advance Open Science at institutional level.

3.3. New skills in the library

This work represents an important achievement for our university. But to reach it, many skills had to be developed within the library. The project involved only one person from the University of Lorraine, the data librarian, who had no pre-existing programming skills. Therefore, it was a challenge to learn Python language, to use Jupyter Notebooks or to get familiar with the Git environment at the same time. She trained herself online and had the opportunity to follow a 12-hour training regarding data science; but it is not really common among librarian skills. The IT staff of the university were not involved in the project, for it was at the beginning a mere experiment. Thanks to regular exchanges with the National Monitor staff, the project came to life after eight months of hard work.

The difficulties encountered were not only technical, but also intellectual: which databases are the most suitable? How to create an understandable and reusable code? How to present the results which were, at the beginning, not at their best?

The first version of the Lorraine Monitor was of most interest to the library and the Research Department. It was broadly disseminated among scientific communities. However, the comments and reactions came mostly from the University’s executive, and far less from scientists themselves.

Conclusion: What future for the Lorraine Open Science Monitor?

The code was broadly reused, and its success can be measured in the increased number of institutions that publish their own version.

However, it costs to the University of Lorraine a great deal in terms of resource allocation. Indeed, even with a readme file and precise instructions, and because most of the time, librarians in charge of this topic do not have programming skills, there are still many questions to answer. Many training sessions had to be scheduled, as well as improvements in the code itself for institutions with different databases.

The next version of the National Open Science Monitor, thanks to a very important work from the team in charge of its development at the Ministry and ongoing exchanges with the University of Lorraine, provides a built-in function to obtain institution level graphs natively. This new version will simplify the generation of graphs and significantly reduce the workload required to set up a local monitor.

The first version of the Lorraine Monitor was not connected to the National Monitor, even if some parts of the National Monitor's code were reused. The second version is different regarding this matter. Indeed, institutions now simply have to gather their lists of DOI (using the Lorraine Monitor code) and send it to the Ministry. The Ministry can then update its database in order to attribute each DOI to the right institution; and then the graphs can be generated directly on the National Monitor's website by each institution. This methodology is precisely detailed (Bracco, L'Hôte, Jeangirard, Torny, 2022). It therefore represents less work for the libraries, for the graphs are easily displayed on a website, and an improvement of the metadata contained in the National Monitor.

The Lorraine Open Science Monitor therefore paved the way for a wide use of these new indicators on Open Science.

As the first institution reusing the national code, the University of Lorraine has been asked by the Ministry to manage the next National Monitor, along with its team and Inria (French Institute for Research in Computer Science and Automation). This new Monitor, which will be developed between 2021 and 2023, will also include indicators about research data and softwares.

Laetitia Bracco

References

Alstete, Jeffrey W., Beutell, Nicholas J. & P. Meyer, John P. (2018). *Evaluating Scholarship and Research Impact: History, Practices, and Policy Development*. Emerald. https://doi.org/10.1108/9781787563872

Bracco, Laetitia (2020). Baromètre lorrain de la Science Ouverte. (2020). Université de Lorraine. https://hal.univ-lorraine.fr/hal-03450104v1

Bracco, Laetitia, L'Hôte, Anne, Jeangirard, Eric, Torny, Didier. *Extending the open monitoring of open science: A new framework for the French Open Science Monitor (BSO)*. (2022). Ministère de l'Enseignement Supérieur, de la Recherche et de l'Innovation, Université de Lorraine. https://hal.archives-ouvertes.fr/hal-03651518

Cagan, Ross. (2013). The San Francisco Declaration on Research Assessment. *Disease Models & Mechanisms, 6 (4)*, 869-870. https://doi.org/10.1242/dmm.012955

Gingras, Yves. (2014). *Les dérives de l'évaluation de la recherche. Du bon usage de la bibliométrie*, Raisons d'agir.

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

[i] Ouvrir la science. (2018). *National Plan for Open Science*. https://www.enseignementsup-recherche.gouv.fr/sites/default/files/content_migration/document/SO_A4_2018_EN_01_leger_982501.pdf

[ii] Université de Lorraine. (2021). *Archive ouverte de l'Université de Lorraine*. https://hal.univ-lorraine.fr/

[iii] Unpaywall is an online open-source software. Its goal is to build a database that distributes open access publications from more than 50,000 publishers and open archives. In Unpaywall, you can find open access versions of scientific publications that are usually paywalled. Unpaywall offers extensions on Google Chrome or Firefox.