Accepted for publication in Quantitative Science Studies (MIT Press – ISSN 2641-3337) on January 18, 2022

Identify scientific publications country-wide and measure their open access: the case of the French Open Science Barometer (BSO)

Lauranne Chaignon
ORCID 0000-0003-4055-8431
Université PSL, 75006 Paris, France

Daniel Egret *
ORCID 0000-0003-1605-7047
Observatoire de Paris, Université PSL, 75006 Paris, France
daniel.egret@psl.eu
* Corresponding author

Abstract:
We use several sources to collect and evaluate academic scientific publication on a country scale, and we apply it to the case of France for the years 2015-2020, while presenting a more detailed analysis focused on the reference year 2019. These sources are diverse: databases available by subscription (Scopus, Web of Science) or open to the scientific community (Microsoft Academic Graph), the national open archive HAL, and databases serving thematic communities (ADS and PUBMED). We show the contribution of the different sources to the final corpus. These results are then compared to those obtained with another approach, that of the French Open Science Barometer (Jeangirard, 2019) for monitoring open access at the national level. We show that both approaches provide a convergent estimate of the open access rate.
We also present and discuss the definitions of the concepts used, and list the main difficulties encountered in processing the data.
The results of this study contribute to a better understanding of the respective contributions of the main databases and their complementarity in the broad framework of a country-wide corpus. They also shed light on the calculation of open access rates and thus contribute to a better understanding of current developments in the field of open science.

Keywords : publications - databases - open science

1. Introduction

Open access to publications (see e.g. Laakso & Björk, 2012; Piwowar et al., 2018) within the general framework of Open Science is now an issue shared by many institutions, universities and research organizations, or funders. France is no exception: two national plans for Open Science have been successively launched, in 2018 and 2021, by the Ministry of Higher Education, Research and Innovation (MESRI). Generalizing open access to publications is the first axis of these two plans, with a goal of 100% of French scientific publications in open access by 2030[1], either through a publication natively in open access or through a deposit in an open archive. This national plan is in line with the European Plan S[2].

[1] National Plan for Open Science: https://www.ouvrirlascience.fr/national-plan-for-open-science-4th-july-2018/; https://www.ouvrirlascience.fr/second-national-plan-for-open-science/
[2] Plan S: https://www.coalition-s.org/

To support the policies thus deployed, a good knowledge of the state of publications and their open access rate seems necessary and many measurement tools have been developed for this purpose, in different contexts, such as the European Open Science Monitor (OSM), the German Open Access Monitor (OAM), the Danish Open Access Indicator, or the COKI Open Access Dashboard. Other countries have also adopted national strategies for monitoring Open Access (Carvalho et al., 2017).

In its guide to assisting research organisations and funders in setting up a tool for monitoring Open Access publications (Philipp et al., 2021), the organisation Science Europe considers the constitution of the corpus of publications to be analysed as one of the key stages in the process. We could add that it is even one of the major challenges of this exercise. Indeed, no database provides an easy and complete answer to this question. The large databases such as the Web of Science (WoS) and Scopus have the advantage of systematically listing a large part of the millions of scientific publications published each year in the world. The metadata are standardized and allow for efficient searching. However, the coverage of science, technology and medicine (STM) and of English-language publications in international journals is privileged, while other disciplinary fields, other languages of publication, other sources or document types are less fully surveyed (Van Leeuwen et al., 2001; Mongeon & Paul-Hus, 2016; Vera-Baceta, Thelwall & Kousha, 2019). Moreover, these databases are accessible only by subscription, so their data are not open or reusable. If we consider thematic databases such as PubMed or NASA/ADS, their metadata are both high quality and open. On the other hand, they cover a very specific disciplinary field: an exhaustive census of publications in a multidisciplinary context will therefore require multiple sources.

As for open archives, while they have the advantage of listing types of publications, languages and sources that are often absent from large databases, they offer insufficiently standardized metadata, which complicates their collection and processing. Thus, no single database offers comprehensiveness, standardized metadata and openness. As Huang et al (2020) conclude in a recent article: "Any institutional evaluation framework that is serious about coverage should consider incorporating multiple bibliographic sources."

Current Research Information Systems (CRIS) can be a way around this difficulty, provided that they are not fed solely by the large commercial databases mentioned above. They are increasingly being used in universities to help manage, understand and evaluate research activities. However, most CRIS are, today, still used only at an institutional level (Sivertsen, 2019). Although their aggregation at the country level in order to constitute a national base is progressing, it is still most often correlated with the implementation of a public funding policy based on scientific publication performance, as is the case in Denmark, Finland, Hungary, Italy, Norway or Poland (Puuska et al., 2020). If the motivation is primarily financial, a national database is an opportunity to set up an effective monitoring of open access policies at the country level, as Finland has experimented with (Pölönen et al., 2020).

For countries that do not have such a pool of data, the implementation of a monitoring tool on this scale implies selecting from among the existing databases, whether commercial or not, those that will best meet the objective set. The German Ministry of Education and Research has thus chosen to use the *Dimensions* and *Web of Science* databases to establish its corpus[3]. *Universities UK*, the association of 140 UK universities, has chosen to use *Scopus to* produce its latest report on the effects of new policies to promote open access[4].

In the case of France, the objective of the MESRI was to set up a tool that would enable the steering of the national policy on open science, by measuring, on an annual basis, the level of open access of all publications with at least one French affiliation. This request was

[3] https://jugit.fz-juelich.de/synoa/oam-dokumentation/-/wikis/Quelldatenbanken/Quelldatenbanken
[4] https://www.universitiesuk.ac.uk/sites/default/files/field/downloads/2021-09/monitoring-transition-open-access-2017-annexe-1-methodology.pdf

accompanied by a very specific requirement: "a transparent methodology and reproducible results". It is with this in mind that the French Open Science Barometer (BSO) was carried out[5], as described by Eric Jeangirard (2019). For the BSO, the constitutive choice is to use only open sources. The methodology used consists in scanning all the papers references in Unpaywall and in the national open archive HAL (see below), in order to identify either the French authors or the presence of the mention of France in the affiliation. The publications thus identified were then enriched with information on their scientific discipline, using natural language processing (NLP), also based on an open source, to determine, from the title, the discipline to which a document belongs. Finally, the open access status was determined using the Unpaywall database. The corpus obtained by this strategy is available in open access from the MESRI OpenData portal[6]. In accordance with the recommendations made at the European level (Open Access Monitoring: Philipp et al. 2021), the French National Open Science Barometer is published on an annual basis.

About 150,000 publications are thus identified each year by the BSO. The purpose of this study is to consider an alternative approach, this time based on the use of the main open or non-open bibliographic databases, and to analyse the extent to which this new corpus differs from that of the BSO. Our approach is based on the use of six complementary sources, namely *WoS*, *Scopus*, *Microsoft Academic Graph*, *PubMed*, *NASA/ADS* and the HAL open archive, to identify and assess academic scientific publication at the scale of a country, in this case France, for publications released during the six years 2015-2020. As the year scale seemed to us more relevant to characterize scientific production, we chose to highlight, in the context of this article, the data related to the year 2019[7]. We then compare the corpus obtained with that of the BSO, and we show to what extent the diversity of the sources used makes it possible to refine the identification and characterization of French scientific production, as well as the estimation of the open access rate.

While there is an abundant literature on the comparison between Scopus, WoS and other generalist databases (see, for example, in a national production context: Bartol et al., 2014; Moed, Markusova & Akoev, 2018; Archambault et al., 2009; or for a statistical comparison of large reference databases: Mongeon & Paul-Hus, 2016; Pranckuté, 2021; Visser et al., 2021), our study provides a detailed quantitative view in the specific context of French research. Far from identifying a source that would be optimal, our study shows the importance of diversifying the sources used to provide complementary views on a country's publication.

2. Constitution of the France 2015-2020 corpus: data and methods

2.1 Definitions

Before describing in detail the methodology used to establish our corpus, we present and discuss here the main concepts used.

DOI (Digital Object Identifier): The DOI[8] is a persistent identifier that can be assigned to any type of content, be it text, software, datasets, etc. (Simmonds, 1999). It will be used as a common metadata for the entire study.

Scientific publications: We consider here scientific publications indexed in databases (private or public) and accessible in open archives. All types of documents are taken into account. This primarily concerns articles, generally published in international peer-reviewed journals, but also conference proceedings, book chapters, or any other publication, provided

[5] http://bso.esr.gouv.fr
[6] https://data.enseignementsup-recherche.gouv.fr/explore/dataset/open-access-monitor-france/
[7] The counts for each of the six years are available in the supplementary data file.
[8] DOIs are managed by the non-profit association CrossRef (Hendricks et al., 2020).

that it has a DOI. However, the restriction to only documents with a DOI is an important restriction, which we must explain here.

In order to facilitate the aggregation of results, and to avoid duplication, we have chosen, as does the BSO (French Open Access Monitoring), to restrict the cross-referencing of data to publications identified by a DOI number. This step is necessary to allow the efficient cross-referencing of documents identified in each database by their DOI identifier, common to all databases. In addition, the Unpaywall database, which will inform us about open access in the next step, only lists publications with a DOI.

Let us note that the requirement of the presence of a DOI immediately rules out a certain number of journals that do not adhere to this very general technology of persistent identifiers (Gorraiz et al., 2016); some of these journals may be, as Wang et al. (2020) point out, key journals in their discipline with the example, for the field of Artificial Intelligence of the *Journal of Machine Learning Research.*

Moreover, grey literature, under which we can group preprints, reports, theses and in some cases conference proceedings (Schöpfel & Prost, 2019), is often ignored by open access measurement tools, mainly for two reasons: the first corresponds to a concern to discard literature whose scientific relevance cannot be sufficiently controlled (lack of peer review); the second is rather related to technical considerations, in particular a difficulty in identifying these publications in the absence of complete and standardized metadata, and in particular persistent identifiers. In practice, this leads to ignoring a large proportion of the work published in certain disciplines where the thematic field, the regional vocation or the applicative nature of the publications take precedence over international referencing.

Our methodology, based on the use of the DOI, therefore effectively excludes some of the documents that might be of interest to us. This is why we will come back to publications without DOIs at the end of our study, by proposing an estimate of the share of grey literature in French national production (part 5.2).

Finally, it should be noted that the publications taken into account to establish our corpus are exclusively those that have a digital version: it is this digital version that we will try to measure the degree of accessibility. Thus, peer-reviewed research published in books/monographs is only covered when it is in digital format and has a DOI. For this reason, non-academic publishing generally falls outside the scope of our study.

Open access: A scientific article that is only available on payment of a subscription or a toll (price per article) is considered closed. In contrast, a scientific article that is freely available, either on a publisher's website or after the deposit of the full text (in its final layout or not) on an open archive, is deemed open.

Our source of information for the open access status of an article will be the Unpaywall database (Piwowar et al., 2018), specifically the data in the "is_oa" field. If the value returned for a given publication is equal to "True", the publication will be considered open. If this value is "False", the publication will be considered closed. The so-called “bronze” status is considered open.

Note that the open access status may vary over time, since a closed publication may have its embargo lifted or be subsequently deposited in an open archive. Thus, in our study, it will be the status observed in February 2021, as recorded in the Unpaywall database snapshot for that date.

Let us recall that for France, the Law for a Digital Republic of October 7, 2016[9] establishes the possibility of deposit on an open archive of the postprint of any scientific article resulting from research funded at least for half by the State or public authorities, at the expiration of a period of 6 months to 12 months depending on the scientific field (respectively, STM or Humanities & Social Sciences).

[9] Law for a Digital Republic: see in particular its article 30
https://www.legifrance.gouv.fr/dossierlegislatif/JORFDOLE000031589829/

2.2 Sources used to constitute the FR-2015-2020 corpus

The collection of metadata related to a large set of publications is facilitated by the use of databases that systematically, if not exhaustively, collect a large part of the millions of scientific publications published each year worldwide.

In this article, we have privileged the databases providing a search capability for the mention of the country in the affiliation, and we have collected the publications whose affiliation mentions the country considered in our study, France, using the corresponding query modes of six databases that, to our knowledge, effectively cover the French scientific production.

We did not use the Dimensions database, as it is not considered to be a reliable source for establishing a corpus on a country scale (Guerrero-Bote et al., 2021).

We use the following databases in our study:

- Scopus (Baas et al., 2020) references more than 25,000 journals and is considered one of the most comprehensive databases for international peer-reviewed journals. Query by country is possible. Metadata extraction is limited to batches of 20,000 documents. This database is available by subscription from Elsevier.
- Web of Science (Birkle et al., 2020) has been the reference database for scientometrics since the pioneering work of Garfield (1964). The query by country is provided in the advanced query mode. This database is available by subscription from Clarivate Analytics. In this study, we use all the indexes (including ESCI: *Emerging Sources*) except for the *Book Citation Index* which was not available to us.
- The HAL open archive https://hal.archives-ouvertes.fr/ (Charnay & Michau, 2007) is a national multidisciplinary open archive intended for the deposit and dissemination of research-level scientific articles (published or not), theses and other objects emanating from French or foreign teaching and research establishments, public or private laboratories. Created in 2001 with ArXiv as a model, this platform has gradually become one of the main tools for reporting French research. A partnership agreement in favour of this archive was signed in 2013 by the Conference of University Presidents (CPU) and 22 institutions. In July 2021, the MESRI also committed to supporting the development of this archive, both in terms of technical aspects and governance, as part of its second national plan for open science 2021-2024.

  French researchers are invited to deposit on this platform the products of their research, whether they are publications (article in a journal, communication in a conference, chapter of a book, book, poster, file, patent), unpublished documents (pre-publication, working document, report), academic works (thesis, HDR, course) or research data (image, video, software, map or sound). The recorded documents are either in the form of a notice only, or accompanied by the full text of the article. This production can be grouped within different collections or portals relating to a theme (SHS for example), a medium (images and videos) or a research structure (university, laboratory or research team), but it remains possible to carry out queries covering all portals and collections. After 20 years of use (Berthaud, Charnay, & Fargier, 2021), more than 2,700,000 works are now recorded in this archive.

  HAL data can be queried using an advanced query or the API. The latter, which is available free of charge, allows the identification of the country of affiliation.
- The NASA/ADS database (Kurtz et al., 2000) is one of the most recognized examples of a bibliographic database covering a research field: astrophysics and physics. Its query mode allows the query by country. Access is free.
- The PUBMED database is one of the preferred and free access points for metadata related to biomedical science research. A query by affiliation is possible (Ibarra et al., 2018).
- The Microsoft Academic Graph database (Wang et al., 2019; Herrmannova & Knoth, 2016), one of the three products of the Microsoft Research project, is one of the largest open publication and citation data sets. It is populated automatically, using bibliographic data from web pages crawled by the Bing search engine, also a Microsoft product. The data can be accessed using the Academic Knowledge API. It should be noted that MAG does not contain structured data on affiliation country. Identification of French outputs (provided by the Curtin Open Knowledge Initiative team) was by applying a query to the affiliation string

(OriginalAffiliation data element from the MAG PaperAuthorAffiliations table, linked via the PaperID to the DOI) that sought to determine whether the affiliation string finished with "France" (or one of a small set of non-English names). This number may not match that in the online COKI country dashboard, which maps affiliation country from GRIDs in MAG to the country of organisation in the GRID database[10].

Some of the characteristics of these databases as well as the number of documents obtained for one year (the year 2019), in the framework of the query "France 2015-2020" carried out in October 2021 are presented in Table 1.

Base	Sample Query (France, year 2019)	Number of documents France 2019	Types of documents	Domains	Practical limitations
Scopus	AFFILCOUNTRY (france) and PUBYEAR = 2019	123,181	All	All areas	Export in batches of 20,000
Web of Science	CU = FRANCE AND PY =2019	124,790	All	All areas	Export in batches of 5000
HAL (Open Archive, France)	Via API: producedDateY_i:2019 structCountry_s:fr	158,937	Open archive of French laboratories	All areas	Export in batches of 10,000
NASA/ADS	aff: "France" AND year:2019-2019	19,997	All	Physics and Astrophysics	Export in batches of 500
PUBMED	(France[Affiliation]) AND ("2019"[Date - Publication])	56,038	All	Medicine, Biology, Health	Export in batches of 10,000
MAG	mag.Year = 2019 AND ((SELECT COUNT(1) FROM UNNEST(mag.authors) as auth WHERE REGEXP_EXTRACT (auth.OriginalAffiliation, r'Fran(ce|kreich|cia)(?:\W|\s+|$)') is not null) > 0	101,885	All (with DOI)	All areas	(COKI, private communication)

Table 1 - Sources used: queries, number of records returned for the year 2019

2.3 Aggregation of results for publications identified by a DOI

As mentioned above, in order to facilitate the aggregation of results and to avoid duplication, we have chosen, as does the BSO (French Open Access Monitoring), to restrict the cross-matching of data to publications identified by a DOI number.

Table 2 shows the counts obtained for the year 2019: DOIs are available for 94% of the documents indexed in *Scopus* and 85% of those in the *Web of Science*. We can notice, in addition, that a major part of the documents without DOI corresponds to communications to conferences (for France and the year 2019: 54% of the documents without DOI in Scopus are communications; 78% in the Web of Science). For ADS the documents without DOI are mainly conference abstracts, while documents without DOI represent only 1% in PubMed.

For the HAL archive, the situation is different: the fact is that the DOI identifier is not systematically filled in because it is not a compulsory metadata during the deposit. While only 2 to 3% of the documents characterized as articles in WoS or Scopus do not have a DOI recorded, this proportion rises to 22% for the documents characterized as articles in HAL. In addition, the open archive contains many unpublished documents, preprints, reports or theses

[10] https://openknowledge.community/dashboards/coki-open-access-dashboard/

that do not have (or not yet) a DOI: with the book chapters, these documents represent half of the publications without DOI, which will not be considered for the rest of the study.

However, we will return to HAL in Section 5 for a discussion of the grey literature.

Note that for MAG, we had direct access to the DOI lists through the COKI team, whom we thank for their help.

Query France 2019	Number of documents	Documents with DOI	%DOI	Category: Articles No DOI
Scopus	123,181	115,273	94%	1,709
WoS	124,790	101,377	85%	2,763
HAL	158,937	66,836	42%	*16,992*
ADS	19,997	15,731	79%	56
PUBMED	56,038	55,516	99%	522
MAG		101,885	-	-

Table 2 - DOI counts in the 6 sources for the year 2019.
The last column shows the numbers of documents without DOIs in the Article category alone.

2.4 Open access and external validation: using ***Unpaywall***

One of the objectives of this study is the measurement of the share of open access to publications. For this we use the *Unpaywall* database[11] which is the leading database in this field (Piwowar et al., 2018; Holly, 2018).

This database offers a simplified access mode (by batches of 1,000 DOIs) which allows to easily obtain the status of a publication (open or closed access, with the publisher and/or in an open archive) at the time of the query. It is also possible to download a complete version of the database (called a Snapshot). For this study, we used the version dated February 2021. For the year 2019, this version lists more than 6 million publications.

Querying the *Unpaywall* database also allows us to validate the DOIs identified in the previous step: we consider that DOIs not found in *Unpaywall* generally correspond to identifiers that have not been confirmed by Crossref, the agency that certifies their quality and continuity.

Moreover, it is not uncommon to find differences in the date of publication from one database to another (often due to the time lag between the version published online (early access) and the "final" publication). We have chosen to use the year of publication provided in the Unpaywall database as the reference year (see Table 3), whether or not it is consistent with the year of publication mentioned in the source database. This choice is also the one adopted by the BSO (French Open Access Monitoring).

	Total with DOI 2019	DOI confirmed Unpaywall 2019
Scopus	115,273	111,422

[11] Unpaywall. http://www.unpaywall.org

WoS	101,377	96,712
HAL	66,836	63,413
ADS	15,731	15,410
PUBMED	55,516	48,047
MAG	101,885	102,338
Total Corpus FR-2019		**139,514**

Table 3 - Unpaywall Cross-Reference: DOI and Year of Publication

Table 3 presents the results of the cross-matching between the six sources, and their validation with Unpaywall.

The first column recalls the number of DOIs obtained from each source, already presented in Table 2. The second column presents the numbers of DOIs found in Unpaywall and recorded in this database as published in 2019.

Note that to obtain the counts in Table 3 we cross-referenced the results of queries covering for the six sources the whole of the years 2015-2020, with the year 2019 from Unpaywall. Discrepancies in publication dates affect about 8% of the documents. Because of the reassignment of publication dates, the number of DOIs with confirmed output (second column of Table 3) for a given year, may be larger than the original number of DOIs for this year (case of MAG), despite a small loss of unidentified DOIs.

In the following section, the 139,514 records described in column 2 will be cross-referenced with the BSO.

3. Comparison of the FR-2019 and BSO datasets

3.1 Overlap of the two sets

The corpus thus constituted (FR-2019) can now be compared with that of the French Open Science Barometer (BSO), which also aims to cover all French production, for several years including 2019[12].

Since the BSO data are also restricted to publications with a DOI and have benefited from the *Unpaywall* query, it is easy to cross-reference the two sets of DOIs. The result is summarized in Table 4.

	France 2019 (# DOI)	Contribution to the global corpus
Corpus FR-2019	139,514	83%
BSO 2019	153,705	92%
In common	125,807	75%

[12] The BSO data have been produced in December 2020 and are made available on the Open Data portal of the Ministry of Higher Education (MESRI) : https://data.enseignementsup-recherche.gouv.fr/explore/dataset/open-access-monitor-france/

BSO only	27,898	17%
FR-2019 only	13,707	8%
Global corpus FR-2019 + BSO (without duplicates)	**167,412**	100%

Table 4 - Cross-referencing of FR-2019 sources with BSO data
(Source BSO: Jeangirard, 2019)

Table 4 shows that, if we restrict ourselves to the data validated after querying *Unpaywall*, 8% of the total data set (i.e., 13,707 DOIs) are not identified in the BSO, while conversely 17% of the documents (i.e., 27,898 DOIs) had not been identified in our FR-2019 corpus

3.2 Data from our FR-2019 corpus that are not part of the BSO corpus

The data from our sources not included in the BSO corpus seem to correspond mainly to a failure to identify the France affiliation in the algorithm developed by Jeangirard (2019). This was expected and corresponds to what Jeangirard calls false negatives – which he says he cannot estimate and which we estimate here at 9% of the BSO corpus.

In our study, the main sources contributing to this subset not identified by the BSO are Scopus (63%), WoS (41%) and MAG (23%). We believe that these documents come from the less represented publishers, for whom it is likely that specific algorithms for extracting the country of affiliation have not been developed for BSO.

3.3. Data from the BSO corpus absent from the FR-2019 corpus

The data from the BSO corpus not included in our sources come mainly from humanities and social sciences journals (44%), biomedical journals (24%) and basic biology journals (12%). We note a significantly higher proportion of articles in French in this BSO-only subset: 31% compared to the average of 15% for the global corpus (the language analysis methodology will be presented below, in section 4.4).

These are mainly journals or resources not covered by the databases we have used, in particular documentary resources and journals with a national scope, in French or English: for example, the most represented sources in this set are:

-Case Medical Research: international database of clinical trials
-Faculty Opinions - Post-Publication Peer Review of the Biomedical Literature
-SSRN Electronic Journal: database of social science preprints.

This set of documents also includes the "false positives" reported by Jeangirard (2019), i.e., documents that their algorithm wrongly identified as publications from the France set. These are publications, for which none of the authors has an affiliation in France, but which the BSO algorithm nevertheless retained. Jeangirard estimates the false positive rate at 4% (which would correspond to about 6,000 publications for the year 2019).

We can try to estimate more precisely this share of "false positives": the search in Scopus of DOIs corresponding to publications collected for the BSO but not confirmed by our other sources sheds light on this subject:

Search in Scopus	Number	Comment
BSO only	27,898	
Not found	23,706	Journals not indexed by Scopus

Found in other years	576	Year assignment discrepancy
Found same year	3,616	Probable false positives from the BSO

Table 5 - Search in Scopus for false positives of BSO

This search allows us to identify 3,616 probable false positives: the Scopus database recognizes the DOI, the year is indeed 2019, but the article does not include, according to Scopus, an affiliation in France. This corresponds to 3.5% of the DOIs common to BSO and Scopus: this count thus seems compatible with the 4% estimated by Jeangirard (2019). Let us note once again that the cross-referencing of the different sources highlights divergent assessments of the publication date of the articles.

3.4 Contribution of the different sources to the overall aggregated corpus

Table 6 presents the contributions of each source to the overall corpus (aggregating the two approaches: our FR-2019 corpus and the one collected for the BSO):

	Scopus	WoS	HAL	ADS	PUBMED	MAG	BSO
Share of Total	67%	58%	38%	9%	29%	61%	92%
In one source	7,211	4,009	6,335	155	230	11,665	27,898

Table 6 - Share of each source in the overall aggregated corpus (FR-2019 + BSO)
The second line gives the number of documents found in only one source (year 2019)

	Scopus	WoS	HAL	ADS	PUBMED	MAG	BSO
Scopus	**111,422**	88,327	54,611	14,851	46,503	85,873	102,736
WoS	88,327	**96,712**	49,664	14,507	44,493	76,286	91,159
HAL	54,611	49,664	**63,413**	10,521	22,934	45,608	61,440
ADS	14,851	14,507	10,521	**15,410**	3,243	11,270	14,780
PUBMED	46,503	44,493	22,934	3,243	**48,047**	44,071	47,696
MAG	85,873	76,286	45,608	11,270	44,071	**102,338**	98,604
BSO	102,736	91,159	61,440	14,780	47,696	98,604	**153,705**

Table 7 - Cross contributions from each source to the overall France 2019 corpus

Table 7 presents the cross-referenced contributions of the sources to the overall corpus. It should be noted that the fact that a publication is identified in database A and is not identified in database B as being part of the corpus does not necessarily mean that it is absent from database B: it may be present in database B, but with a DOI that has not been filled in or is incorrect, or a failure to identify the country (no affiliation with France).

4. Estimated rate of Open Access publications

4.1 Unpaywall results: Share of open access publications (year 2019)

Table 8 presents the main results of the Open Access (OA) rate estimate observed in February 2021, based on Unpaywall.org, for each of the sources.

Note that we do not use here the original BSO open access observations, which were made at a different date, and thus could not be directly compared to ours. We have chosen to report all the calculations to the same observation date: that of the production of the Unpaywall *snapshot* in February 2021.

Publications France 2019	# DOI	Total OA	% OA	OA Articles	%OA Articles
Scopus	111,422	61,854	56%	56,538	59%
WoS	96,712	56,975	59%	54,473	60%
HAL	63,413	42,316	67%	38,513	69%
ADS	15,410	11,981	78%	11,608	80%
PUBMED	48,047	29,907	62%	29,818	63%
MAG	102,338	53,392	52%	48,647	55%
FR-2019	**128,344**	**75,070**	**54%**	67,285	**57%**
BSO	153,953	82,267	54%	70,197	57%
FR-2019 + BSO	167,412	88,365	53%	75,413	56%

Table 8 - Share of open access for each source (OA calculation: Unpaywall)
For all sources, including the BSO: open access as of February 2021

Table 8 illustrates the results obtained, depending on the sources used, to determine the Open Access rate (%OA) observed in February 2021: overall we find 54% both for the BSO corpus, and for our FR-2019 corpus. The aggregation of the two results gives a slightly lower overall rate of 53% for all 167,412 publications.

The reader is referred to Aliakbar & Stahlschmidt (2019) for a discussion of the merits and limitations of these rate calculations. In their conclusions the authors recommend the use of multiple sources to reduce errors and gaps, and this is clearly a view we share. Cross-matching all these datasets allowed us to correct, at least in part, the problem of false negatives and to obtain a refined estimate of the open access rate.

4.2 Variation in open access rate by document type

The calculation for the articles alone, using the *journal-article* nomenclature proposed by Unpaywall, shows, as expected, a significantly higher rate of opening: 57% for the BSO corpus, and for our corpus, and 56% for the corpus resulting from the aggregation of the two sets.

This category is interesting insofar as the national policy enacted by Article 30 of the 2016 law mentioned above concerns a "scientific writing [...] published in a periodical appearing at least once a year", i.e., in our terminology, a scientific journal article.

In this context, it is worth mentioning that the approaches presented here do not distinguish between publicly funded research articles and other articles from private and industrial research, for which the open science commitments do not apply.

The details of the types of documents identified for both approaches are given in Table 9. The percentages observed are very similar in the two datasets (FR-2019 and BSO) for articles and conference proceedings. The differences are more noticeable for book chapters and can be explained by a significantly wider coverage in the case of the BSO. The 'other' category covers too many different situations for the differences in the observed rate to be significant.

Type of document	Number of DOIs	Share	% OA	%OA FR2019	%OA BSO
journal-article	133,638	80%	56%	57%	57%
book-chapter	13,268	8%	25%	24%	27%
proceedings-article	12,987	8%	40%	40%	41%
other	7,519	4%	60%	54%	64%
Total FR-2019+BSO	167,412	100%	53%		

Table 9 - Share of Open Access by document type
(Overall dataset FR-2019 + BSO)

4.3 Observation of annual trends (2015-2020)

In order to detect the ability to measure annual changes, we extracted the data —and present the annual counts in Table 10a— for each of the years 2015 to 2020, following the same methodology as outlined for 2019. For 2019 the counts are identical to those in Tables 3 and 7, above. Table 10b provides for the years 2015 to 2019 the data from Table 4, above (the BSO does not cover the year 2020).

Year	HAL	PUBMED	ADS	WoS	Scopus	MAG
2015	51,734	41,287	15,387	91,028	108,195	92,722
2016	57,851	44,785	16,396	96,186	112,486	96,850
2017	59,451	46,057	16,806	95,731	113,077	95,808
2018	61,997	46,490	16,254	97,012	114,069	99,356
2019	63,413	48,047	15,410	96,712	111,422	102,338
2020	59,796	55,293	16,077	94,237	104,533	100,608

Table 10a - Counts obtained for publications from France,
for the years 2015 to 2020, using the same methodology

	FR-2015-20	BSO	Global corpus	%FR15-20	%BSO
2015	133,817	140,493	157,053	85%	89%
2016	138,885	148,476	164,772	84%	90%
2017	138,845	146,179	162,179	86%	90%

2018	141,059	159,380	171,987	82%	93%
2019	139,514	153,705	167,412	83%	92%

Table 10b – Results of the two approaches for the years 2015 to 2019.
See Table 4 above.

For the observation of open access, the reference remains Unpaywall (snapshot of February 2021). The results are shown in Table 11. As expected, they show a steady increase in the open access rate from 2015 to 2019.

The year 2020, observed in February 2021, has a different character since the observation is made before the 6-month, 1-year, or in some cases longer embargoes have expired.

Open access rate / Year of publication	FR2015-2020	BSO	Global corpus
2015	45,4%	45,5%	44,5%
2016	47,8%	47,6%	46,6%
2017	50,0%	50,0%	48,9%
2018	51,7%	50,6%	49,9%
2019	53,8%	53,5%	52,8%
2020	52,6%	-	52,6%

Table 11 - Change in open access rate, observed in February 2021
for publications dated from 2015 to 2020
(The global corpus is the aggregation of the two datasets FR-2015-20 and BSO)

In Table 12, we give examples of observations of the open access status (Gold, Green, etc.) as provided by Unpaywall for two distinct years. These few examples allow us to affirm the absence of significant bias between the two datasets: the two strategies lead to quite similar estimates.

Open access status	FR-2015	BSO 2015	FR-2019	BSO 2019
Gold	12%	13%	18%	18%
Hybrid	12%	12%	9%	10%
Bronze	4%	4%	7%	6%
Green	18%	16%	20%	20%
Closed	55%	54%	46%	46%

Table 12 - Open access status, observed in February 2021
for publications dated 2015 and 2019.

A comparison of the rates obtained for the French corpus with those obtained on an international scale would go beyond the limits of this article: the interested reader may refer to the recent study by Robinson-Garcia, Costas, & van Leeuwen (2020), which also presents a discussion of the different modes of open access mentioned here (Gold, Bronze, Hybrid, Green).

4.4 Are French articles more often in open access?

It is possible to cross-reference the observations presented above with information on the language in which the article is written: are articles in French, for example, more often, or less often, in open access? To examine this, as this information is not systematically provided by all databases, we analyzed the title of the article as provided by Unpaywall, by applying a simple language detection software *langdetect*[13]. Only detections assigned with a displayed probability greater than 0.99 were retained.

In the framework of our study of French national scientific production, for the year 2019, the two main languages concerned are English (83% of the detected documents) and French (15%), the rest of the detected languages not exceeding 3% in total. The distribution is not identical according to the document type, in particular the communications to –mostly international— conferences (labelled proceedings-article in Unpaywall) are almost always in English.

	% English	% French
journal-article	82%	16%
book-chapter	77%	14%
proceedings-article	97%	1%
other	87%	4%

Table 13 - France 2019: language by document type

Table 14 shows that the rates of open access observed vary greatly according to the disciplines (extracted here from the BSO). As a general rule, documents detected as being written in French are much less frequently in Open Access.

	Number of documents	% documents in French	% OA documents in English	% OA documents in French
Total with language and discipline detected	153,272	15%	58%	26%
Chemistry	7,050	5%	53%	50%
Computer and information sciences	10,225	8%	55%	37%
Mathematics	3,914	11%	73%	55%
Medical research	48,191	24%	57%	8%
Biology (fond.)	21,535	12%	69%	57%
Social sciences	8,020	69%	40%	37%
Physical sciences, Astronomy	15,701	7%	64%	73%
Earth, Ecology, Energy and applied biology	12,222	16%	59%	42%
Engineering	4,402	24%	40%	40%

[13] Langdetect (https://pypi.org/project/langdetect/) is a python-port of Nakatani Shuyo's language-detection library (https://github.com/shuyo/language-detection). When published (in 2010), it claimed to reach 99%+ accuracy on 49 supported languages.

Humanities	9,388	66%	41%	43%

Table 14 - France 2019: Open access rate by language and discipline.
Calculations are restricted to documents for which the language can be determined
and whose discipline is assessed in the BSO.

Most of the French language material without Open Access comes from three areas: medical research, including journals for practitioners; and the humanities and social sciences.

5. Results and Discussion

5.1 Discussion of the sources used

The six sources we have chosen to use actually provide three different insights:

(1) Scopus and Web of Science provide extensive coverage of the literature in peer-reviewed journals and international conference proceedings; while Scopus has a slightly wider coverage, the use of the two databases together provides a 10 to 20% improvement over what would be obtained with a single database. The MAG database, which will soon be discontinued, brings, as a complement, a set of documents not indexed by WoS and Scopus, contributing to a further increase of about 10% of the corpus identified in our study.

(2) The HAL open archive is filled at the initiative of the authors who deposit the bibliographic record (metadata) and, if applicable, the full text in its preprint or editor version. Part of the archive contains grey literature (Schöpfel, Prost, & Ndiaye, 2019) and moreover the DOI is filled in irregularly and not systematically. The metadata and DOI do not seem to be thoroughly quality controlled: for this reason, this source should be considered with caution for bibliometric studies. However, it is a reference source for French research and a cornerstone of the national open science policy.

(3) The ADS and PUBMED databases are thematic databases and are therefore only intended to cover parts of the research field. On the other hand, both databases are deep in their field and cover grey literature and sources not indexed by the large generalist databases.

This study sheds new light on the coverage of French scientific production by the various databases. While the Web of Science and Scopus voluntarily restrict themselves to the perimeter of peer-reviewed publications appearing in referenced journals or books (Birkle et al., 2020; Baas et al., 2020), the use of complementary databases, whether thematic or not, allows us to have a more complete view of the share of literature that is not or poorly referenced, and that may be less general in scope, geographically, linguistically, or thematically. We observe that the strategy adopted by the BSO allows for the systematic collection of data on a significant quantity of these publications –often neglected in bibliometric studies. Far from identifying an optimal source, our study shows the importance of diversifying the sources used to provide complementary views on a country's publication.

5.2 Characteristics of excluded national production without DOI

Publications without DOI form a heterogeneous group of peer-review and grey literature. The share of unreferenced grey literature can be approached in particular through the HAL open archive, by considering documents without DOI, which were not taken into account in our study. However, it is advisable to make sure beforehand that the absence of DOI is not due to a lack of information, but corresponds to articles from journals that do not use this identification mode. Since the open archive, which is mainly fed by author deposits, is not fed in a complete and systematic way, this approach can only be qualitative.

We note, first of all, without surprise, a very strong disciplinary variation: only 15% of the documents in the field of humanities and social sciences (SSH) deposited in HAL have a DOI, while the proportion is 70% in Chemistry or Physics, the global average being 42% for the year 2019 considered here (see Table 2). This rate reaches 50% in the field of Computer Science.

Among the records without DOI the share of records from the SSH fields is 52%, compared to an SSH share of 12% of publications with DOI.

We also note that the full text is deposited significantly less frequently for documents without a DOI: 39%, whereas the average is 44%.

We can also note, for HAL (year 2019) a strong differentiation according to the language (we use here the language informed in the archive):

- Among the documents without DOI, the proportion of articles in French is 57% (49% for articles in English), while for articles with DOI it is only 8%.
- 91% of the documents in French have no DOI (or no DOI indicated).

In total, we found nearly 90,000 records without a DOI in HAL (Table 2). If we restrict ourselves to documents classified as Articles, book chapters or conference papers, nearly 56,000 records without a DOI (or without a DOI indicated) listed in HAL had to be excluded from this study.

- For the journal articles (category ART in HAL) we tried to estimate the part which corresponds to not informed DOI: if we consider the articles without DOI published in a journal for which other articles have DOI, we note that it concerns 31% of the articles without DOI (in HAL in 2019). We therefore estimate that at least 30% of DOIs are missing in HAL due to DOIs that are not filled in. Most of this 30% can be expected to be covered by the other sources. If this assumption is correct, it would mean that out of the 56,000 records without DOI entered in HAL, we can estimate that there are around 40,000 articles or communications without DOI, which were therefore not taken into account. This point will be the subject of further study.

5.3 Validation of the open strategy used for the BSO

The comparison between the result obtained with our sources and the open strategy of the BSO validates the use of the latter: this strategy, if we summarize it in a few words, consists in scanning all the DOIs available by Unpaywall, and also by HAL, to identify either the French authors, or the presence of the mention of France in the address.

We observe that this strategy makes it possible to identify more than 20,000 records (if we exclude the false positives) not found by our approach, i.e., about 17% of the total: these are mainly journals that are not indexed in the major international databases, and more particularly in the biomedical and social science fields.

Our approach also identified approximately 13,000 DOIs not included in the BSO and thus estimated the false negative rate in the BSO strategy to be close to 9% (see Table 4 above).

Recurrent sources of error include conflicting approaches to publication date (with the usual confusions between the first online publication and the final date of the reference; see for example Liu, 2021).

6. Conclusions

The main results of our Study are as follows:

- Our study validates a strategy of determining a collection of scientific publications with an affiliation in France, for a given year. This corpus is deliberately restricted by the use of DOI. We present the details of the counts for the year 2019. We estimate that the corpus of outputs with DOI covers around 80% of the French national scholarly production in 2019, with an additional set of 40,000 articles or communications without DOI not taken into account here.
- Our determination of cross-coverage by the various databases provides useful insight for users of these databases. We believe that these counts can help users of these databases to identify overlaps and complementarities, in a context comparable to that of our study.
- The use of multiple sources ensures validation at a sufficiently fine level to shed light on the geographical, thematic, linguistic, etc. disparities that affect bibliometric studies. Our study confirms the relevance of adopting a multi-source approach.

- The open-source strategy used by the BSO effectively identifies the vast majority of publications with a persistent identifier (DOI) for Open Science monitoring.
- The determination of the open access rate has been refined. It should be remembered that this rate depends on the date of observation and may differ depending on the type of documents we wish to consider. Our objective is not to comment here on the 54% or 53% rate reached for the opening of publications in 2019 (observed in February 2021), but to note the convergence of two different methodologies that allow us to accurately draw the shifting landscape of open science at the country level.

The question of the place of the national open archive HAL, and of other open archives, in the strategy of Open Science deserves a specific development which should be the subject of a further study. The objective of such a study would be to examine the possibilities of convergence between on the one hand the specific challenges of open archives, allowing an easy deposit at the disposal of the authors, and on the other hand the requirements of referencing and query environment which should not only provide open access to scientific knowledge produced by French research, but also support the most diverse possible readership in their consultation process.

Acknowledgements

We thank the two reviewers for their stimulating comments, which we believe have significantly helped to improve our work.

Author Contributions

Lauranne Chaignon: Writing—review & editing, Validation. Daniel Egret: Supervision, Data curation, Writing—original draft, Writing—review & editing.

Competing Interests

The authors have no competing interests.

Funding Information

No specific funding has been received for this research.

Data Availability

Data tables providing the detailed number of records for each year, as well as a notebook describing the whole procedure, are available as supplementary data files on HAL Open Archive: https://hal.archives-ouvertes.fr/hal-03537679. Subscriptions to Scopus and Web of Science are required to replicate the research, with the methods described above.

References

Aliakbar, A., & Stahlschmidt, S. (2019). Merits and Limits: Applying Open Data to Monitor Open Access Publications in Bibliometric Databases. SocArXiv. https://doi.org/10.31235/osf.io/npj4h.

Archambault, É., Campbell, D., Gingras, Y. and Larivière, V. (2009). Comparing bibliometric statistics obtained from the Web of Science and Scopus. Journal of the American Society for Information Science and Technology, 60: 1320-1326. https://doi.org/10.1002/asi.21062

Baas, J., Schotten, M., Plume, A., Côté, G., & Karimi, R. (2020). Scopus as a curated, high-quality bibliometric data source for academic research in quantitative science studies. Quantitative Science Studies 2020; 1 (1): 377–386. https://doi.org/10.1162/qss_a_00019

Bartol, T., Budimir, G., Dekleva-Smrekar, D., Pusnik, M., & Juznic, P. (2014). Assessment of research fields in Scopus and Web of Science in the view of national research evaluation in Slovenia. Scientometrics, 98(2), 1491-1504. https://doi.org/10.1007/s11192-013-1148-8

Berthaud, C., Charnay, D., & Fargier, N. (2021). Diffuser et pérenniser le savoir scientifique : 20 ans d'histoire de HAL. Histoire de la Recherche Contemporaine, 10 (2) | 2021. https://doi.org/10.4000/hrc.6330

Birkle, C., Pendlebury, D. A.,Schnell, J., & Adams, J. (2020). Web of Science as a data source for research on scientific and scholarly activity. Quantitative Science Studies,1(1),363–376. https://doi.org/10.1162/qss_a_00018

Carvalho, J., Laranjeira, C., Vaz, V., & Mendes Moreira, J. (2017). Monitoring a National Open Access Funder Mandate. Procedia Computer Science, 106, pp. 283-290. https://doi.org/10.1016/j.procs.2017.03.027.

Charnay, D., Michau. C. (2007). L'archive ouverte HAL. JRES 2007, Nov 2007, Strasbourg, France. ⟨hal-01101888⟩

Garfield, E. (1964). Science Citation Index — A new dimension in indexing science. Science,144(361), 649–654. https://doi.org/10.1126/science.144.3619.649

Gorraiz, J., Melero-Fuentes, D., Gumpenberger, C., & Valderrama-Zurián, J.-C. (2016). Availability of digital object identifiers (DOIs) in Web of Science and Scopus. Journal of Informetrics, 10 (1), pp. 98-109. https://doi.org/10.1016/j.joi.2015.11.008

Guerrero-Bote, V. P., Chinchilla-Rodríguez, Z., Mendoza, A., & de Moya-Anegón, F. (2021). Comparative Analysis of the Bibliographic Data Sources Dimensions and Scopus: An Approach at the Country and Institutional Levels. Frontiers in research metrics and analytics, 5, 593494. https://doi.org/10.3389/frma.2020.593494

Hendricks, G., Tkaczyk, D., Lin, J., & Feeney, P. (2020). Crossref: The sustainable source of community-owned scholarly metadata. Quantitative Science Studies,1(1), 414–427. https://doi.org/10.1162/qss_a_00022

Herrmannova, D., & Knoth, P. (2016). An analysis of the Microsoft Academic Graph. D-Lib Magazine, 22(7), 9–10. https://doi.org/10.1045/september2016-herrmannova

Holly, E. (2018). The rise and rise of Unpaywall. Nature, 560, 7718, 290-291. https://doi.org/10.1038/d41586-018-05968-3

Huang, C.-K., Neylon, C., Brookes-Kenworthy, C., Hosking, R., Montgomery, L., Wilson, K., & Ozaygen, A. (2020). Comparison of bibliographic data sources: Implications for the robustness of university rankings. Quantitative Science Studies 2020; 1 (2): 445–478. https://doi.org/10.1162/qss_a_00031

Ibarra, M.E., Ferreira, J.P., Torrents, M. et al. (2018). Changes in PubMed affiliation indexing improved publication identification by country. Scientometrics 115, 1365–1370. https://doi.org/10.1007/s11192-018-2714-x

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

Kurtz, M.J., Eichhorn, G., Accomazzi, A., Grant, C.S., Murray, S. S., & Watson, J. M. (2000). The NASA Astrophysics Data System: Overview. Astron. Astrophys. Suppl. Ser., 143, 41. http://dx.doi.org/10.1051/aas:2000170

Laakso, M., Björk, BC. (2012). Anatomy of open access publishing: a study of longitudinal development and internal structure. BMC Med 10, 124. https://doi.org/10.1186/1741-7015-10-124

Liu, W. (2021). A matter of time: publication dates in Web of Science Core Collection. Scientometrics 126, 849–857. https://doi.org/10.1007/s11192-020-03697-x

Moed, H. F., Markusova, V., & Akoev, M. (2018). Trends in Russian research output indexed in Scopus and Web of Science. Scientometrics, 116(2), 1153–1180. https://doi.org/10.1007/s11192-018-2769-8.

Mongeon, P., & Paul-Hus, A. (2016). The journal coverage of Web of Science and Scopus: A comparative analysis. Scientometrics,106(1), 213–228. https://doi.org/10.1007/s11192-015-1765-5.

Philipp, T., Botz, G., Kita, J.-C., Richards, P., Sänger, A., Siegert, O., & Reumaux, M. (2021). Open Access Monitoring: Guidelines and Recommendations for Research Organisations and Funders. Science Europe, Briefing Paper, May 2021. https://doi.org/10.5281/zenodo.4905553

Piwowar, H., Priem, J. Larivière, V., Alperin, J.P., Matthias, L., Norlander, B., Farley, A., West, J., & and Haustein, S. (2018). The State of OA: A Large-Scale Analysis of the Prevalence and Impact of Open Access Articles. PeerJ 6:e4375. https://doi.org/10.7717/peerj.4375.

Pölönen, J., Laakso, M., Guns, R., Kulczycki, E., & Sivertsen, G. (2020). Open access at the national level: A comprehensive analysis of publications by Finnish researchers. Quantitative Science Studies, 1 (4): 1396–1428. https://doi.org/10.1162/qss_a_00084

Pranckutė, R. (2021). Web of Science (WoS) and Scopus: The Titans of Bibliographic Information in Today's Academic World. *Publications 2021*, 9(1), 12. https://doi.org/10.3390/publications9010012.

Puuska, H.-M., Nikkanen, J., Engels, T., Guns, R., Ivanović, D., & Pölönen, J. (2020). Integration of national publication databases – towards a high-quality and comprehensive information base on scholarly publications in Europe. In ITM Web Conf. 33 (2020) 02001. https://doi.org/10.1051/itmconf/20203302001

Robinson-Garcia, N., Costas, R., van Leeuwen, T.N. (2020). Open Access uptake by universities worldwide. PeerJ 8:e9410. https://doi.org/10.7717/peerj.9410

Schöpfel, J., & Prost, H. (2019). The scope of open science monitoring and grey literature. *12th Conference on Grey Literature and Repositories, National Library of Technology (NTK),* Oct 2019, Prague, Czech Republic. ⟨hal-02300020⟩

Schöpfel, J., Prost, H., & Ndiaye, E. (2019). Going Green. Publishing Academic Grey Literature in Laboratory Collections on HAL. GL21 International Conference on Grey Literature, 22-23 October 2019, Hannover, Germany. ⟨hal-02300017⟩

Simmonds, A.W. (1999). The Digital Object Identifier (DOI). Publishing Research Quaterly, 15, 10–13. https://doi.org/10.1007/s12109-999-0022-2

Sivertsen, G. (2019). Developing Current Research Information Systems as Data Sources for Studies of Research. In Wolfgang Glänzel, Henk F. Moed, Ulrich Schmoch & Mike Thelwall

(eds.), Springer Handbook of Science and Technology Indicators. Springer Verlag, 667-683. https://doi.org/10.1007/978-3-030-02511-3_25

Van Leeuwen, T.N., Moed, H.F., Tijssen, R.J.W. *et al.* (2001). Language biases in the coverage of the Science Citation Index and its consequences for international comparisons of national research performance. *Scientometrics* **51,** 335–346. https://doi.org/10.1023/A:1010549719484

Vera-Baceta, MA., Thelwall, M. & Kousha, K. (2019). Web of Science and Scopus language coverage. *Scientometrics* **121,** 1803–1813. https://doi.org/10.1007/s11192-019-03264-z

Visser, M., van Eck, N. J., Waltman, L. (2021). Large-scale comparison of bibliographic data sources: Scopus, Web of Science, Dimensions, Crossref, and Microsoft Academic. Quantitative Science Studies 2 (1): 20–41. https://doi.org/10.1162/qss_a_00112

Wang, K., et al. (2019). A Review of Microsoft Academic Services for Science of Science Studies. Frontiers in Big Data, 2, 45. 10.3389/fdata.2019.00045

Wang, K., Shen, Z., Huang,C., Wu, C.-H., Dong, Y., & Kanakia, A.(2020). Microsoft Academic Graph: When experts are not enough. Quantitative Science Studies, 1(1), 396–413. https://doi.org/10.1162/qss_a_00021