Science and Technology Observatory

Broadening Data Sources for Bibliometric Analyses

Recent Results and Further Developments

Seminar

Paris, Hcéres – Wednesday 28 February 2024

April 2024

Table

Introduction

Frédérique Sachwald, OST - Hcéres

Broadening data sources for bibliometric analyses may have two objectives. The first one is to use open access sources in order to be able to be transparent about the raw data and ease replication. The second one is to be able to work on larger data bases that may include more diversified types of publications and offer a better coverage of countries around the world. Such broader data sources may also offer a better coverage of scientific publications across disciplines.

The first three presentations focus on the first issue by exploring the use of OpenAlex for some bibliometric analyses. The fourth presentation deals with the second issue and explores the impact of using different corpora within the Web of Science on bibliometric indicators.

The general discussion is then introduced by Vincent Larivière and Peter van den Besselaar.

Learning from the first Leiden Ranking Open edition

Nees Jan van Eck, Centre for Science and Technology Studies (CWTS), Leiden University

- *Link to the presentation: https://doi.org/10.5281/zenodo.10806619*
- *Link to Leiden Ranking Open Edition: https://open.leidenranking.com/*

The Leiden Ranking of Universities

Open data sources such as OpenCitations, OpenAIRE, and OpenAlex offer unrestricted access for using and reusing the data, leading to initiatives that utilize open data, such as the COKI (Curtin Open Knowledge Initiative) Open Access Dashboard[1] or the French Open Science Monitor[2]. The idea of a fully transparent and reproducible ranking system has resulted in the development of the Leiden Ranking Open Edition, which is a part of a broader movement. In order to allow for direct comparisons, this new edition maintains the same methodology as the traditional Leiden Ranking while providing transparency of the data.

The inaugural version of the Leiden Ranking, a multidimensional assessment of universities, was released in 2007 and it now includes 1,411 universities from 72 countries. It provides bibliometric indicators based on the Clarivate's Web of Science database, focusing on the number of publications, scientific impact, collaboration, open access, and gender statistics. Each indicator is provided separately and the Leiden ranking does not rely on composite indicators or surveys. It acknowledges that the overall performance of a university can vary depending on the specific aspect of interest. However, the traditional Leiden Ranking was not completely transparent, particularly in making underlying data available to users, due to restrictions imposed by the proprietary nature of the Web of Science data, which prevents access to the actual publications contributing to a university's indicators.

Preparing the Leiden Ranking Open Edition

To address this issue, the Leiden Ranking team has produced an open edition based on OpenAlex data, in collaboration with the OpenAlex team. This open edition follows the traditional ranking in methodology and the number of universities included in order to facilitate comparisons and maintain the trust of the academic community. The open edition aims to explore the possibilities of using open data, including by identifying data quality issues. The Leiden team provided feedback to OpenAlex in order to contribute to data improvement. The Leiden Ranking Open Edition was launched in January 2024.

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

[2] https://frenchopensciencemonitor.esr.gouv.fr/

The production process began with the extraction of data from OpenAlex, focusing on identifying “core publications”, i.e., publications with specific metadata in international scientific journals[3]. A publication classification system, based on whole OpenAlex, was developed to calculate field-normalized indicators, and publications were assigned to universities using an organization registry, similar to the one used in the traditional Leiden Ranking. The OpenAlex team has improved its institution parsing and linking system, allowing for better affiliation matching. This has resulted in a significant increase in the percentage of publications correctly linked to universities. Three types of affiliated organizations were identified: component organizations that are fully controlled by a university, joint organizations that are controlled by multiple universities, and associated organizations that are related to but not fully controlled by a university. ROR identifiers were used to map these affiliations. The ranking ensures transparency by listing all affiliated organizations considered for each university. Publications were classified into 4,521 fields of science using clustering algorithms.

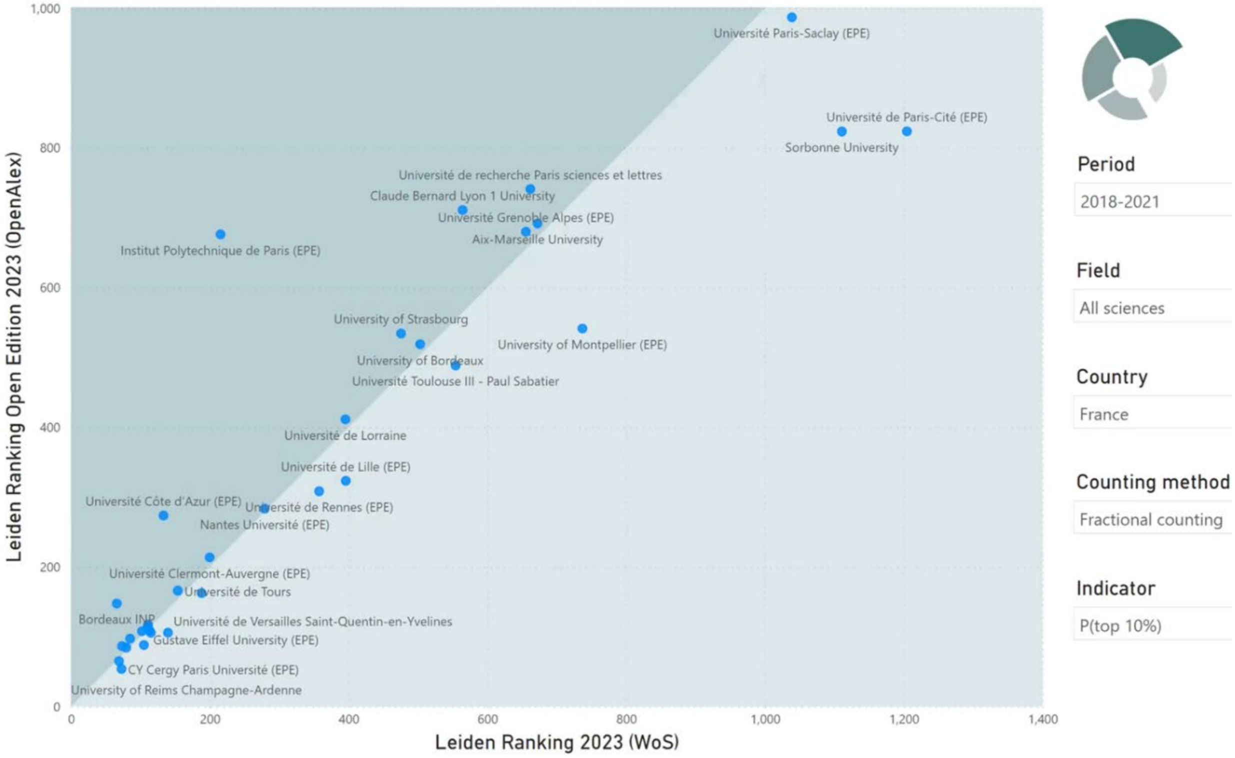
The methodology of the Leiden Ranking includes four main components: volume of publications, scientific impact, open access, and collaboration, each of which is crucial to the ranking process. This Leiden Ranking Open Edition can be accessed online. It includes both size-dependent and size-independent indicators, with detailed pages for each university displaying data in tables and figures.

Comparing results of the Leiden Ranking Open Edition

A comparison between the traditional Leiden Ranking and the Open Edition reveals differences in how publications are assigned to universities. A manual analysis showed that errors in assignment were due to missing or incorrect affiliation data, with 25% of errors attributed to the traditional ranking and 75% to OpenAlex. Improvements have been made, and further enhancements are expected.

Scatterplots comparing bibliometric indicators from the two editions show that most universities' results align closely, with some deviations, particularly among Chinese universities, which warrants further investigation. French universities generally align well, with some exceptions like the *Institut Polytechnique de Paris* (IPP).

Open Edition vs. traditional Leiden Ranking, number of top10% publications, France, 2018-21



Period
2018-2021

Field
All sciences

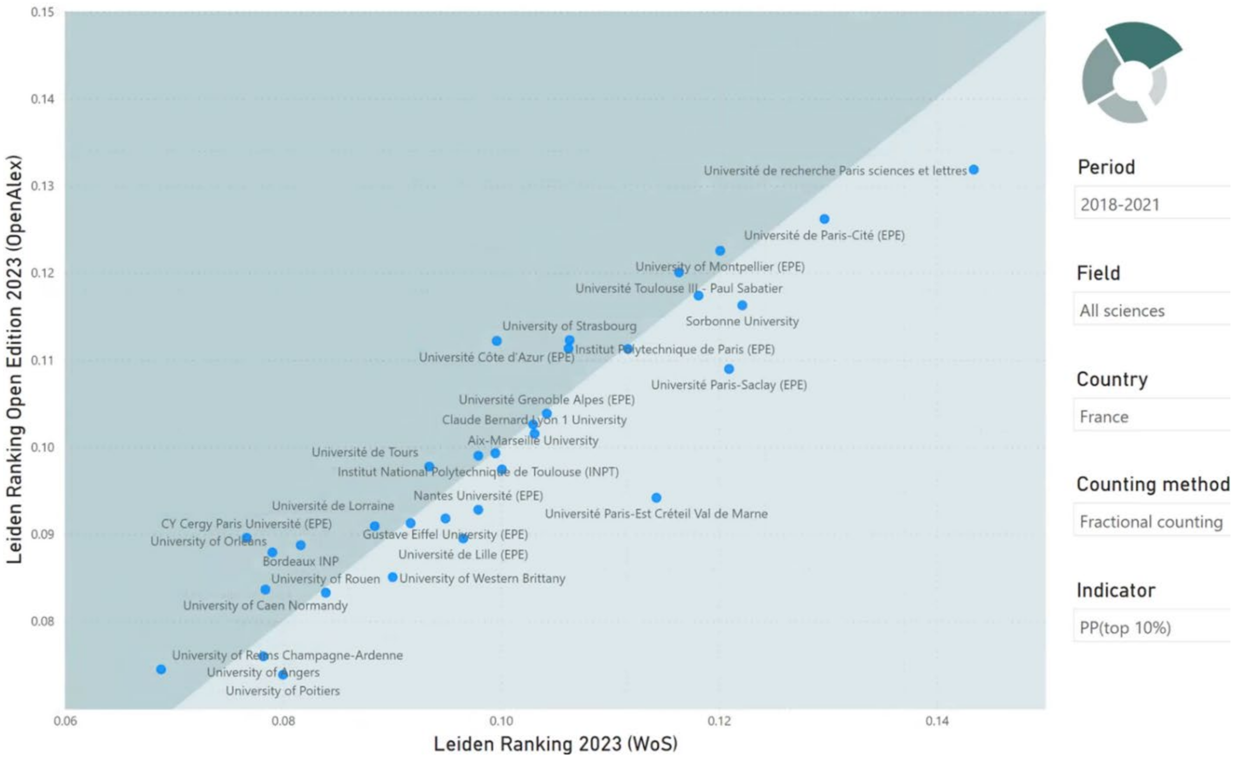
Country
France

Counting method
Fractional counting

Indicator
P(top 10%)

[3] For the complete list of the criteria defining a core publication, see: https://open.leidenranking.com/information/indicators

Open Edition vs. traditional Leiden Ranking, Proportion of publications of top 10%, France, 2018-21



Period
2018-2021

Field
All sciences

Country
France

Counting method
Fractional counting

Indicator
PP(top 10%)

Conclusion

In conclusion, open data sources offer significant advantages over traditional proprietary data, democratizing access to research information and enhancing transparency. While there are still quality concerns with open data on publications, adopting a stance of waiting for perfection is not the right approach. Actively using open data sources is essential for identifying and resolving data and technical issues, as demonstrated by the collaboration with the OpenAlex team.

Comparison of field normalized scores of German universities calculated on different databases

Thomas Scheidsteger, Max Planck Institute for Solid State Research

- *Link to the presentation:* https://doi.org/10.5281/zenodo.10942822
- *Link to the contribution presented at STI 2023:* https://doi.org/10.55835/6441118c643beb0d90fc543f

Research evaluation using bibliometric methods is frequently based on commercial bibliographic databases such as Web of Science, Scopus or Dimensions but there are now free alternatives such as OpenAlex. With the many databases available, one question is how similar are field-normalized citation scores using the same indicator definition but different underlying databases.

Calculation of citation scores for 48 German universities in four publication data bases

A preliminary study had analyzed the publications of a computer science institute with a small number of papers (442).[4] The comparison between the Web of Science and Microsoft Academic Graph showed some promising results and motivated the present study to do the same on more than three hundred thousand papers from 48 German universities.[5]

German universities were chosen for this study because of their high-quality disambiguated and unified address information developed by the I2SoS Bibliometrics Team at the University of Bielefeld[6] and provided by the German "Competence Network for Bibliometrics"[7]. The publication set contains 334,511 publications (articles or reviews) from 48 German universities that published more than 3,000 papers between 2013 and 2017. All these papers are indexed in the four databases of interest: WoS, Scopus, Dimensions and OpenAlex. They are identified by a unique DOI. Only papers in the Top 4 OECD fields were considered: natural sciences, medicine, engineering and social sciences.

The normalized citation scores were calculated in the traditional way. The citations count for each paper is divided by the reference value, which is the mean of similar papers (same publication year, same document type and same subject classification). The expected citation rates for the NCS were calculated based on different field categorization schemes in the four databases: 252 subject categories for the Web of Science, 335 journal classification codes for Scopus, the second level 154 categories for Dimensions and the 284 OpenAlex sub-categories.

Comparison between the four data bases

The comparison between the data bases uses several statistical measures: coefficients of a linear regression, Spearman rank correlation coefficient (r_s), Lin concordance correlation coefficient (r_ccc) and Mean normalized citation score (MNCS).

Considering Scopus vs WoS or OpenAlex vs WoS, scatterplots and linear regressions show that outliers have an effect on the linear regression and Lin's r_ccc - but not on Spearman's r_s that is higher than 0.88 for all cases. To quantify the effect of outliers, the papers in the Top 1% NCS values in each database were excluded. This results in a strong increase in Lin's concordance coefficient for the WoS vs. Scopus comparison and the smallest absolute change for the Dimension vs. OpenAlex comparison. The other comparisons show similar values for Lin's r_ccc with or without outliers.

Looking at the MNCS in the four databases for the 48 German universities, the order of the universities is very similar: NCS_WoS < NCS_Scopus < NCS_Dimensions < NCS_OpenAlex. Looking at Lin's concordance, when outliers are removed, the overall spread is strongly reduced and in most cases Lin's r_ccc is reduced (except in Scopus vs WoS). Comparisons show a broader spread with Top 1% papers included and no more extreme r_ccc values without Top 1% papers.

Comparison of OpenAlex vs. WoS in the top 4 OECD categories shows strong to almost complete agreement for Natural Sciences and Engineering, and less strong agreement for Medicine and Social Sciences. Looking only the natural sciences and medicine for the 48 universities separately, outliers appear for 3 universities when comparing OpenAlex vs 3 commercial databases: Looking at 3 outliers in detail for a single university like University of Mainz shows many more citations in OpenAlex than in WoS. The slope of the linear regression hardly changes without the 3 outliers, but Lin's r_ccc is well within the range of good agreement.

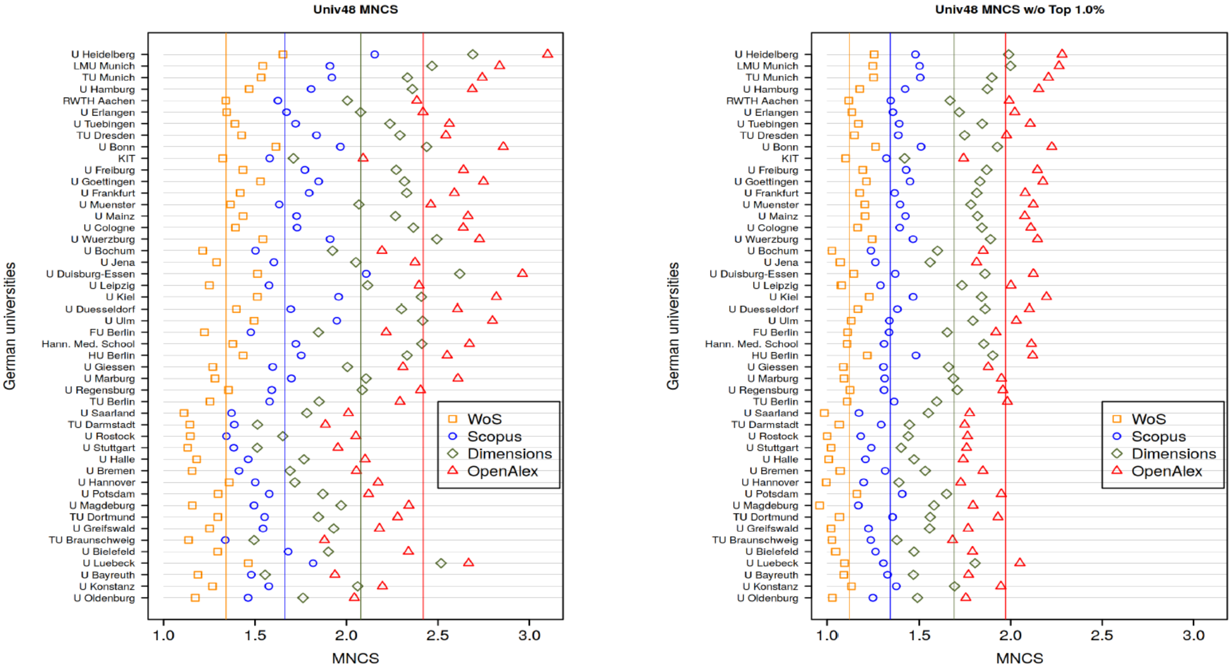
[4] Scheidsteger, T., Haunschild, R., Hug, S., & Bornmann, L. (2018). ***The concordance of field-normalized scores based on Web of Science and Microsoft Academic data: A case study in computer sciences***, 23th International Conference on Science, Technology and Innovation Indicators, Leiden. https://hdl.handle.net/1887/65358.

[5] It is based on: Scheidsteger, T., Haunschild, R. & Bornmann, L. (2023). How similar are field-normalized scores from different free or commercial databases calculated for large German universities? 27th International Conference on Science, Technology and Innovation Indicators (STI 2023). https://doi.org/10.55835/6441118c643beb0d90fc543f

[6] https://www.uni-bielefeld.de/einrichtungen/i2sos/bibliometrie/index.xml

[7] http://www.bibliometrie.info

MNCS across the 48 German universities (ordered by publication output) in the four databases



Conclusion

In conclusion, when all publications are taken into account together, all comparisons show almost complete, or at least strong, agreement. Moreover, the removal of the Top 1% most cited publications leads to small decreases in most cases. Considering the 48 universities separately, there is strong to almost complete agreement between the three commercial databases. Besides there are several cases of very low r_ccc values in the comparisons between OpenAlex and the three commercial databases. Removing the Top 1% papers resulted in strong or almost complete agreement between data bases in most cases.

The three commercial databases are fairly similar to each other. OpenAlex is a bit different but still seems to be similarly suited for bibliometric evaluations like the established commercial databases.

However, there are two limitations. First, generalization to other countries is only possible if institutional data with a very high quality of disambiguation are available. Second, NCS and MNCS are susceptible to outliers.

Reference Coverage Analysis of OpenAlex compared to Web of Science and Scopus

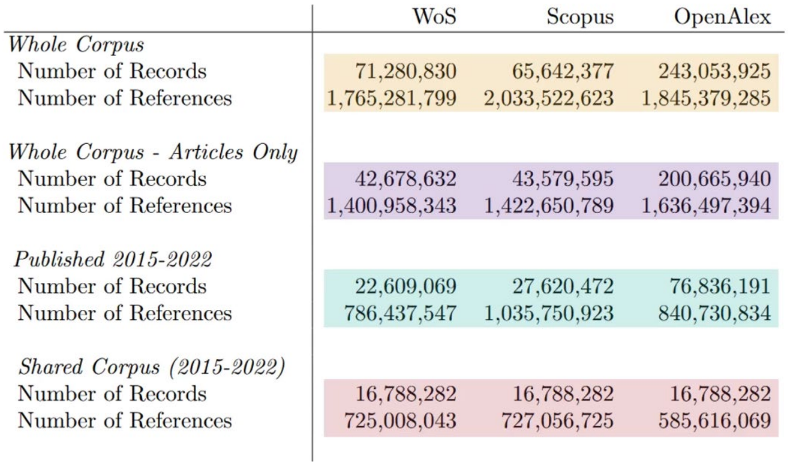
Jack Culbert, GESIS, Leibniz Institute for the Social Sciences

- *Link to the presentation: https://zenodo.org/records/10777335*
- *Link to the related paper: https://arxiv.org/abs/2401.16359*

OpenAlex is a new open source of scholarly metadata, which can become a competitor to established commercial sources. OpenAlex was released by OurResearch in 2022 as a replacement for the discontinued Microsoft Academic Graph (MAG). As this promising alternative source is rapidly evolving and the data contained within is expanding and changing, the question of its trustworthiness arises. In other words, considering its free and open nature in contrast to the expensive and closed-access models of Web of Science and Scopus, is OpenAlex ready for bibliometrics?

This research is carried out as part of the project Comparative Analysis and Curation of German Metadata in open Bibliometric Data (OPENBIB)[8]. The purpose is to establish an open bibliometrics database within the Kompetenznetzwerk Bibliometrie (KB).

Construction of the Shared Corpus

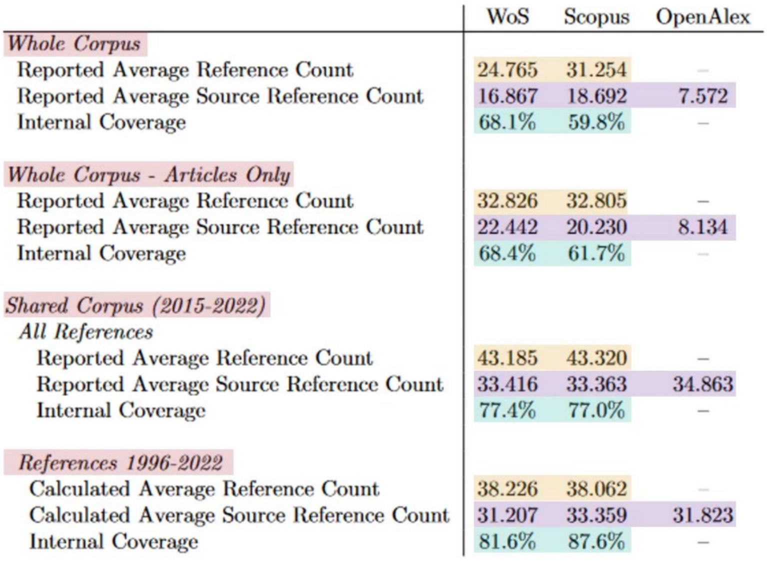


	WoS	Scopus	OpenAlex
Whole Corpus			
Number of Records	71,280,830	65,642,377	243,053,925
Number of References	1,765,281,799	2,033,522,623	1,845,379,285
Whole Corpus - Articles Only			
Number of Records	42,678,632	43,579,595	200,665,940
Number of References	1,400,958,343	1,422,650,789	1,636,497,394
Published 2015-2022			
Number of Records	22,609,069	27,620,472	76,836,191
Number of References	786,437,547	1,035,750,923	840,730,834
Shared Corpus (2015-2022)			
Number of Records	16,788,282	16,788,282	16,788,282
Number of References	725,008,043	727,056,725	585,616,069

Comparing metadata coverage

With 243 million records, OpenAlex is much larger than WoS and Scopus, which include 71 and 66 million records respectively. However, the total number of references of OpenAlex is comparable to that of the two other databases, which leads to a very low average reference count: 7.5 references on average for OpenAlex versus more than twice as much for WoS and Scopus. When the databases are limited to articles only, the pattern is similar. However, when restricted to the common corpus, OpenAlex was found to have a similar reference coverage as the Web of Science or Scopus. That is also confirmed when references are limited to the recent literature, that is to say to the references published between 1996 and 2022.

Comparison of Reference Counts



	WoS	Scopus	OpenAlex
Whole Corpus			
Reported Average Reference Count	24.765	31.254	–
Reported Average Source Reference Count	16.867	18.692	7.572
Internal Coverage	68.1%	59.8%	–
Whole Corpus - Articles Only			
Reported Average Reference Count	32.826	32.805	–
Reported Average Source Reference Count	22.442	20.230	8.134
Internal Coverage	68.4%	61.7%	–
Shared Corpus (2015-2022)			
All References			
Reported Average Reference Count	43.185	43.320	–
Reported Average Source Reference Count	33.416	33.363	34.863
Internal Coverage	77.4%	77.0%	–
References 1996-2022			
Calculated Average Reference Count	38.226	38.062	–
Calculated Average Source Reference Count	31.207	33.359	31.823
Internal Coverage	81.6%	87.6%	–

[8] https://bibliometrie.info/en/

In terms of metadata coverage observed at the journal level, OpenAlex has a much better coverage of ORCID compared to Scopus and Web of Science. The proportion of articles with a piece of abstract information in OpenAlex is less important (87%) than in WoS or Scopus (92%). One explanation could be that abstracts were not shared openly by large publishers (Elsevier, Taylor & Francis, IEEE, etc.) via Crossref. As for metadata on open access (OA) status information, coverage between OpenAlex and the two proprietary databases is highly correlated but slightly in favour of OpenAlex – possibly due to the time lag for the proprietary databases to integrate Unpaywall open access status information. The proportion of open access information in all three datasets is around 49%.

OpenAlex needs to improve metadata coverage, volatility, and potential over-allocation of articles to ORCID identifiers. For instance, in a few cases, ORCIDs were attributed to more than 10,000 records in the corpus, indicating issues with author name disambiguation in OpenAlex.

The challenge in extracting references for OpenAlex is due to its reliance on web scraping, as opposed to the manual indexing procedures of Web of Science and Scopus.

Conclusion

Overall, the paper provides valuable insights into the strengths and weaknesses of OpenAlex compared to established databases like the Web of Science and Scopus. The discussion underscored the need to understand the limitations of OpenAlex's reference coverage and its implications for bibliometric analyses. Potential avenues for further exploration include the examination of patterns based on publishers, journal types, and scientific fields. Addressing these issues is essential for leveraging OpenAlex's potential for academic research.

Comparison of France scientific profile measured on various publication perimeters

Agénor Lahatte, Frédérique Sachwald, OST-Hcéres

- *Link to the presentation:* https://www.hceres.fr/sites/default/files/media/downloads/seminar-broadening-data-sources-ost-on-the-case-of-france_hceres.pdf

The objective is to analyze the influence of the coverage of publication data bases on a set of bibliometric indicators for the main publishing countries. The analysis does not compare two different data bases, but successively compares different perimeters within the Web of Science: first a standard perimeter with a corpus that includes the Emerging Sources Citation Index and second a total world corpus with one including only articles in English.

Comparison of the Standard perimeter with a Large perimeter

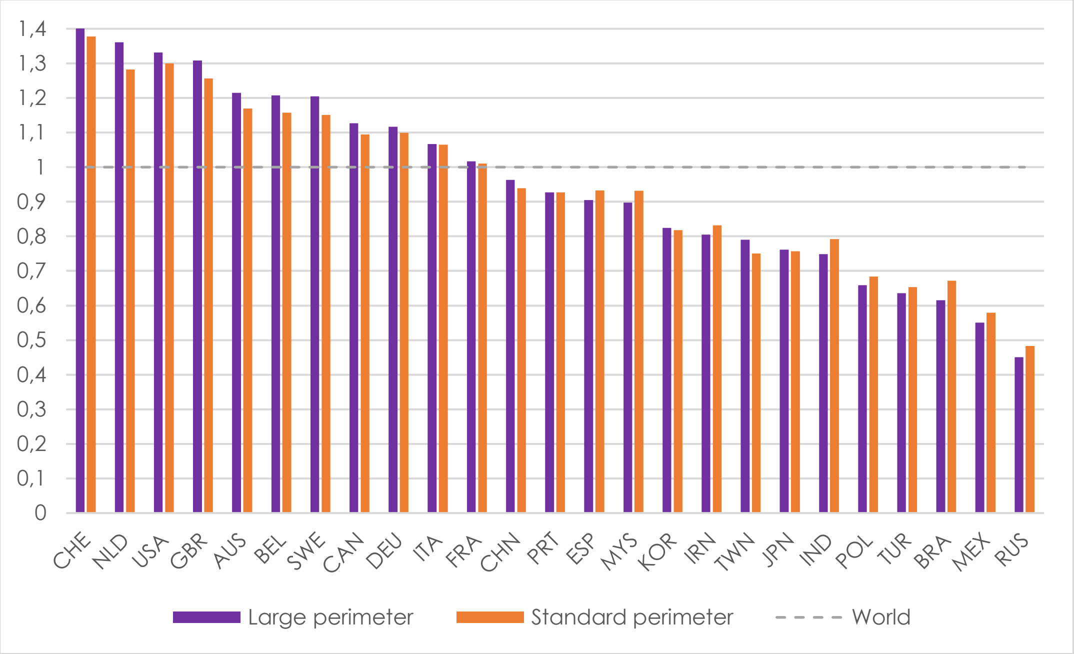
The standard perimeter of OST publication data base, a home version of WoS, includes the following indexes: SCIE, SSCI, AHCI, CPCI-S, CPCI-SSH. In 2023, OST has added the Emerging Sources Citation Index (ESCI), resulting in the large perimeter. In both cases, only articles and reviews in journals and proceedings are included. In 2021, the World Standard perimeter, or corpus, includes 2.7 million publications and the large corpus 3 million – a 9% increase.

The large corpus includes the same share of publications in Life sciences, a slightly smaller share in Physical sciences & Engineering (-6%) and a larger share in Social sciences and Humanities (+31%).

Among the largest publishing countries, some gain very few publications in the large corpus: less than 6% for China, Switzerland, the Netherlands or Germany. France gains 6% more publications. On the contrary, Russia, India, Turkey and Brazil gain more than 18%. Spain or Poland are in between at 12-13%.

These same countries that gain the most in terms of publications tend to lose the most in terms of impact as measured by the Mean normalized citation score (Figure below), ordered by decreasing MNCS in the large perimeter). Symmetrically, the countries having a similar number of publications in both corpora experience a slight increase in their MNCS in the larger corpus. France has 6% more publications in the large perimeter and its MNCS is about equal, around 1.0.

MNCS in the large vs standard perimeter, 2010-21



Source: OST publication data base, WoS, treatment by OST

Focus on publications in English

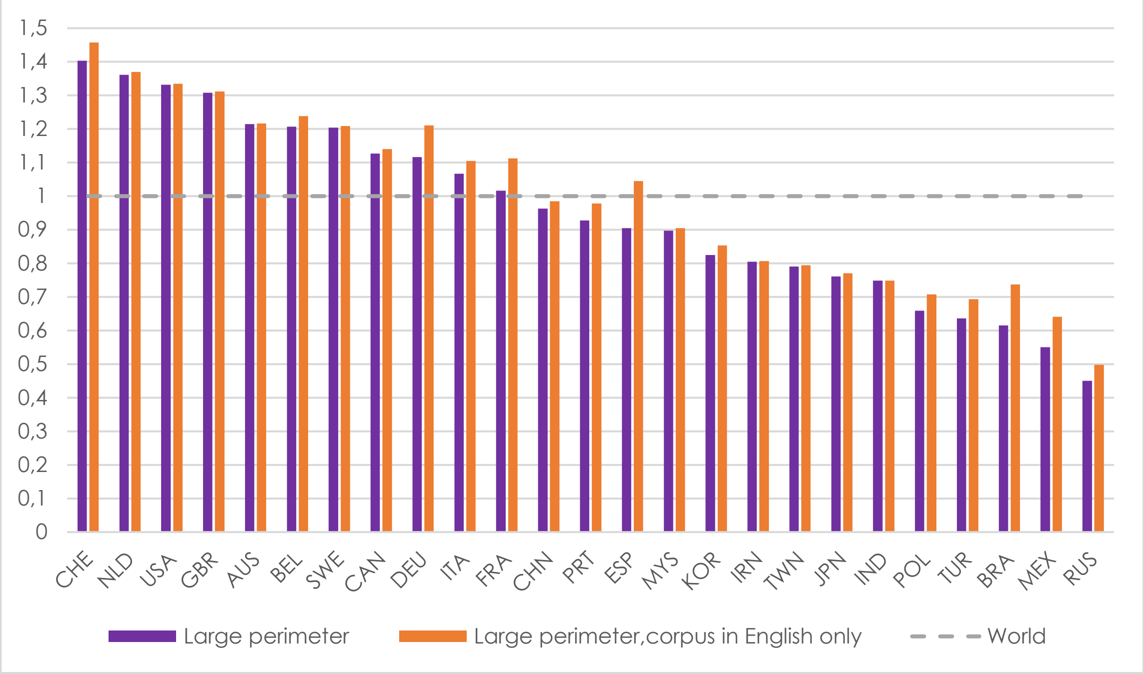
At the world level, the large corpus includes a higher share of publications that are not written in English.[9] Among the main publishing countries, it is highest for Brazil (17%), Mexico, Russia (both 15%) and Spain (14%). Both France and Germany have a little less than 7% of their publications not in English, Italy 3%, China 2% and Canada 1%.

The publications that are not written in English tend to have a smaller or much smaller potential audience. As a result, the countries with the highest share of non-English publications tend to have substantially higher impact indicators when the corpus is restricted to English publications. This is the case of Brazil, Mexico and Russia in particular (figure below).

In the case of France, the share of non-English publications is highest in Social sciences and Humanities (32%) and the impact indicator (MNCS) is also much higher in the English corpus of these disciplines (+48% as compared to the total large corpus). The impact indicator increases the most in some domains of the humanities.

[9] This refers to the text with a specific WoS field – not to the abstract (that may be both in English and other languages).

MNCS: large perimeter compared to only publications in English, 2010-21



Source: OST publication data base, WoS, treatment by OST

Conclusion

This presentation has focused on two sets of changes in the perimeter of world publications, and conclusions are similar. The countries that gain most publications in a broader perimeter – either by enlarging the number of journals or by including non-English publications – loose most in terms of indicators of impact. There is a trade-off between the breadth of the corpus and the impact of included publications as measured by the MNCS. In the two cases that have been explored, the changes are larger for Social Sciences and Humanities.

The analysis has been conducted within OST publication data base, which is a home version of the WoS. The hypothesis is that this exploration can be an illustration of what will be observed when comparing selective data bases to larger ones. It is by the way consistent with the results of the second presentation of this seminar on German universities.

Will broader data sources enable more relevant bibliometric analyses?

Introduction

Vincent Larivière, Université de Montréal

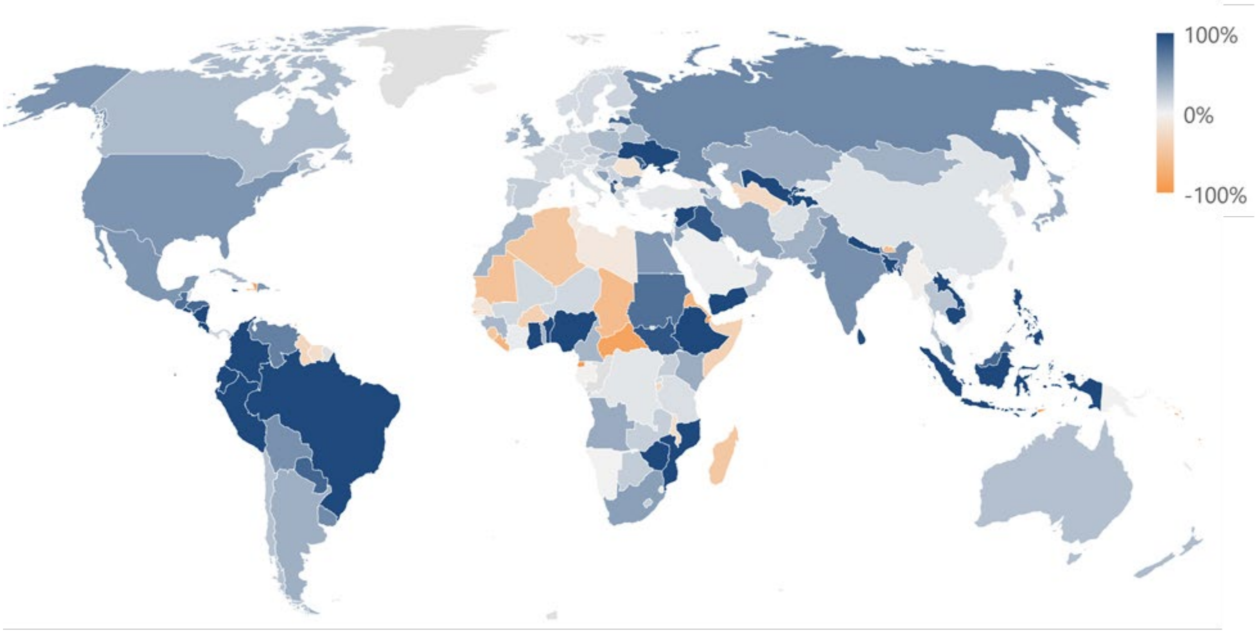
There is a quite clear conclusion emerging from recent research, including some of those presented in this seminar. It is the fact that if Open Alex is used with the same cut than other data bases like Web of Science or Scopus, results are quite similar. It is a good point and shows a certain level of robustness of OpenAlex, which is open and free.

But the ambition of OpenAlex is to be able to cover a much broader perimeter of publications, including in national languages. Limits of Web of Science and Scopus are well known and Open Alex offers a broader coverage with papers not in English, papers from the global south, more proceedings, and books. This second advantage is crucial but when we expand datasets, there are many issues about the quality of metadata. This is the case in particular for new types of documents that are not included in the selective commercial data bases. For example affiliations are totally missing on books metadata.

Another issue is the coverage of different countries. The additional coverage of OpenAlex is not evenly distributed across countries. For example, Brazil or Indonesia benefit a lot from the broader coverage of OpenAlex (see the map below). The United States and India also substantially increase their number of publications. European countries and China are stable, while some countries in Africa have actually less publications in OpenAlex.

There are a number of ways to improve the situation. More bibliometricians and librarians should go on working on OpenAlex, publishing and disseminating knowledge. In parallel, national policies should promote open access publications that would be easy to integrate in OpenAlex. And, having in mind the team of OpenAlex is quite small, the scientific community should find a way to improve the metadata by for example creating a shared space to fill the missing metadata.

Country difference in the number of articles, OpenAlex as compared to WoS, in %, 2015-22



Sources: data from WoS and OpenAlex, treatment by Vincent Larivière

General discussion

Major characteristics of OpenAlex

A first set of questions and comments related to the need to further clarifying to what extent and how OpenAlex can be used presently. Concerning the broader coverage, one potential advantage is a better coverage of emerging countries, but the map showed by Vincent Larivière suggests that for the time being this advantage is quite unequal among countries. The coverage advantage should also be documented across scientific fields.

One advantage of OpenAlex is its better coverage of publications in national languages (for example in French) of publications in Social sciences and Humanities. However, given the difficulties to compare these publications across countries and the fact that OpenAlex coverage is geographically uneven, this broader coverage by OpenAlex may not be adapted to international comparisons.

One major asset of OpenAlex is open access; it means that it is accessible to everyone in the world free of charge. But on the other hand, it requires adequate skills and knowledge of scientific publications to use OpenAlex properly, in particular for bibliometric analyses.

Remaining technical challenges

One issue is that of duplicate publications and of the impact on citation count. Duplicates are mainly due to different types of documents and doesn't impact counting if restricted to articles. Yet another issue is the fact that the identification of document types is not reliable.

More generally, it is important to document OpenAlex challenges, as it has been done for the WoS over the years, resulting in the bibliometric literature. One example is the fact that journals have to be available online to be indexed in OpenAlex. Currently, China appears to be rather absent of the debate while it is the first producer of scientific articles. This is due to the fact that Chinese affiliations remain rather unavailable in the database. It is important not to reproduce or create biases in OpenAlex. Besides, some problems are not new; again about China, CNKI (China National Knowledge Infrastructure) is not available in the WoS nor in OpenAlex.

Improving the quality and reliability of OpenAlex

Several participants discussed the way to contribute to the improvement of the quality of OpenAlex metadata. The issue of the attribution of scientific articles to researchers was specifically discussed. It involves a good disambiguation of researchers' names. At the individual level, providing feedback to the OpenAlex team may help, but will not be sufficient. A centralized infrastructure could be designed to help researchers exchange information with OpenAlex. An automated process could be developed if the ORCID was used more systematically and reliably by researchers, which would be more economical than creating a new infrastructure. Some researchers are reluctant to use ORCID, but open access to publications and to quality data on publications requires some efforts, like more accessible information from researchers. Otherwise, costs may be too high, leading to compromises on some of the initial objectives.

Another improvement could be for OpenAlex to implement better reference extractions, including full text.

The two avenues for broadening publication data sources

On the basis of this seminar, provisional conclusions can be set forth on each of the two avenues for broadening data sources for bibliometric analyses.

OpenAlex represents a formidable opportunity to have a both open access and very large source of data on productions from the scientific communities around the world. Bibliometricians have already succeeded in replicating previous analyses on OpenAlex. However, at this stage, such studies partially rely on external information, including from the historical data bases. New analyses on the sole basis of OpenAlex require further investment to increase the quality and reliability of the data base.

The second objective involves the use of broader data sources to work on diversified types of publications and have a better coverage of scientific publications across disciplines. This can be done by exploring various corpora to analyze the impact of each broadening of the data sources on indicators. OpenAlex could certainly be considered as a source for such exploration. Given the variety of document types in OpenAlex, such explorations would require to precisely define the perimeter of the documents of interest, for example articles in scientific journals or broader sets of scholarly literature. Specific perimeters could also be designed on the basis of the language of publications.

Evaluation · Research · Higher Education
Hcéres

2 rue Albert Einstein
75013 Paris, France
T. 33 (0)1 55 55 60 10
hceres.fr

Haut Conseil de l'évaluation de la recherche et de l'enseignement supérieur
High Council for the evaluation of research and higher education