SciELO as an open scientometric research infrastructure: general discussion of coverage in OpenAlex, WoS, Scopus and Dimensions

João de Melo Maricato[*], Alysson Mazoni[**], Rogério Mugnaini[***], Abel L. Packer[***], Rodrigo Costas[****]

[*] *jmmaricato@unb.br*
ORCID: https://orcid.org/0000-0001-9162-6866
University of Brasília, Faculty of Information Science, Brazil

[**] *afmazoni@unicamp.br*
ORCID: https://orcid.org/0000-0001-5265-6894
University of Campinas, Department of Scientific and Technological Policy, Institute of Geosciences, InSySPo, Brazil

[***] *mugnaini@usp.br*
ORCID: https://orcid.org/0000-0001-9334-3448
University of São Paulo, School of Communication and Arts, Brazil

[****] *abel.packer@scielo.org*
ORCID: https://orcid.org/0000-0001-9610-5728
Director of SciELO / FAPESP, Brazil

[*****] *rcostas@cwts.leidenuniv.nl*
ORCID: https://orcid.org/0000-0002-7465-6462
Centre for Science and Technology Studies, Leiden University, Leiden, The Netherlands
DSI-NRF Centre of Excellence in Scientometrics and Science, Technology and Innovation Policy (SciSTIP), Stellenbosch University, Stellenbosch, South Africa

Abstract. This research comparatively analyses the coverage of publications indexed in SciELO in OpenAlex, WoS, Scopus and Dimensions databases. DOIs from SciELO publications were used to study their coverage in the selected databases. We found that 68% of publications indexed in SciELO have DOI, with progressive growth over the years. There are discrepancies in the number of DOIs per SciELO collection, ranging from 99% in SciELO Brazil to less than 5% in other countries. Coverage of SciELO publications with DOIs is high in Dimensions (95%) and OpenAlex (93%) and low in Scopus (57%) and WoS (33%). There is a need to look for ways to improve the large-scale availability of SciELO data. These improvements could represent the tipping point for the SciELO database to become an open research scientometric infrastructure in its own right.

1. Introduction

Many different scientometric databases have emerged over the past years. The possibility of extracting large volumes of data has been facilitated by means of Application Programming Interfaces (APIs) and cloud computing technologies. The capacity of producing and collecting data through large-scale scientometric infrastructures have increased enormously through digital publishing and automated data processing. There is still, however, an important lack of understanding of the criteria and coverage of the science and research practice that are captured in these infrastructures (Krüger, 2020).

Classical scientometric databases like Web of Science (WoS) or Scopus have historically received numerous criticisms. One of the main problems of these databases is that users must pay to access them. Moreover, WoS and Scopus are selective in choosing journals to be indexed. Their selection process tends to introduce disciplinary, geographic and language biases, often excluding journals outside the dominant commercial oligopoly, as well as those that communicate research from the social sciences and humanities or are published in non-English languages (Loprieno et al, 2015; Larivière & Sugimoto, 2018).

The lack of comprehensive coverage is the most common and important criticism of the WoS and Scopus databases. A country like Brazil has been estimated to have a coverage as low as 53% of its scientific journals, and 52% of its scientific publications is WoS (Melo, Trinca & Maricato, 2021). Publications in a language other than English (common practice in Iberoamerica and in other non-English regions) are particularly poorly covered in these selective databases (Simons, 2008), including citation flow between languages (Santos et al., 2021).

The underrepresentation of many countries in selective databases does not allow for the proper scientific evaluation of countries that are not part of the scientific mainstream, causing biases in the research evaluations elaborated from their data (Demachki & Maricato, 2022; Krüger, 2020; Mongeon & Paul-Hus, 2016; CLACSO, 2020). This low representation of publications has been considered as one of the main reasons for creating specific indexes and databases in Latin America as well as other peripheral countries (Santin & Caregnato, 2018).

Some countries have developed regional databases and citation indexes, considered more appropriate for analysing and evaluating their more local scientific production. Examples are SciELO Citation Index, Chinese Science Citation Database (CSCD), China National Knowledge Infrastructure (CNKI), Russian Science Citation Index, Indian Citation Index, Taiwan Citation Index, or the Russian Science Citation Index (Santin & Caregnato, 2018).

Other experiences have been carried out by different types of organisations. Examples include commercial databases, such as Dimensions (from the company Digital Science) or Lens.org (Cambia), and non-profit organisations, such as Unpaywall and OpenAlex from OurResearch, or the Crossref databases. The existence of such alternatives comes with new advantages: reduction of costs of databases; possibility of comparison and verification between them; and possibilities of analysing different levels of coverage, thus making it possible to identify the most pertinent database for evaluation in a given context (Thelwall, 2018).

The coverage of a database can be assessed from various perspectives, such as a coverage of indexed sources, publications, citations, disciplines and subject areas, document types, regional and non-English literature, content overlap, quality, etc. (Pranckut, 2021; Rafols, Ciarli, & Chavarro, 2020; Visser, van Eck & Waltman, 2021). The WoS and Scopus potentialities and limitations are widely known and discussed in the literature (Melo, Trinca & Maricato, 2021; Mugnaini, Digiampietri & Mena-Chalco, 2014; Moed, 2002; Mongeon & Paul-Hus,2016; Pranckute, 2021). Therefore, some characteristics of the three lesser-known databases (SciELO, Dimensions and OpenAlex) deserve to be briefly presented. In this research we will comparatively analyse the coverage of publications indexed in the SciELO Network database in relation to the WoS, Scopus, Dimensions and OpenAlex databases.

Although, there are relevant works that have analysed different aspects and discussions related to the coverage of SciELO (Packer, 2014; Minniti, Santoro & Belli, 2018; Beigel et al, 2024),

Dimensions (Martín-Martín, et al 2021) and OpenAlex (Scheidsteger & Haunschild, 2022), there has not yet been a large-scale comparison of these databases, particularly with the specific question of *how well covered are SciELO publications in these new databases?*

1.1. SciELO in the context of scientometric research infrastructures

The Scientific Electronic Library Online (SciELO) is a database started in 1998, predominantly funded (between 85 and 90%) by the São Paulo State Research Support Foundation (FAPESP). It fulfils, according to Packer et al (2019), indexing, storage, publication and dissemination functions of open access journals. It emerged as a program to support the Brazilian research infrastructure, which selects journals published by the country's institutions. SciELO is not a publisher or editor of journals, but a meta-publisher digital library, performing the functions of collection development, including permanence and withdrawal of journals using indexing criteria with bibliometric methods (Packer et al, 2019; Packer, 2020). SciELO plays an extremely relevant role for the internationalisation of scientific production, and has become one of the best options for monitoring the evaluation of the science developed in the participating countries.

SciELO has expanded its collections over time, spreading out to Latin America and Caribbean (LAC), South Africa, Portugal, and Spain, as well as having thematic collections. It also publishes other types of information sources, such as books, preprints and data sets. According to the library's website, currently (April 2023), 1,905 journals, more than 1 million documents and 27.5 million references are indexed, from which, SciELO Brazil contributes with 399 journals, 483,541 documents and 12,749,068 references.

Over the years, both Scopus and WoS have taken some steps to expand their regional coverage, for example, both databases partly incorporating SciELO (Scopus in 2007 and WoS in 2014), with the aim of covering more research from Latin America and the Caribbean (Pranckut, 2021). These databases, however, do not not allow the extraction or export of all SciELO data from its collections in an intuitive way or through tables or a robust API. Moreover it is unclear how well covered are SciELO publications in their main indexes.

The Dimensions database was launched in 2018 by Digital Science. Dimensions indexes publications (papers, book chapters, proceedings, monographs, preprints), datasets, grants, patents and clinical trials. Dimensions is a prominent new player for bibliometric work, covering a wider range of sources than WoS and Scopus (Krüger, 2020). Dimensions data is presumably sourced from publishers and has the advantage that their data has less spans and errors, seemingly filling a gap for large-scale analytics (Thelwal, 2018). According to the company's website and 2019 report, currently (April 2023) the database covers 135 million publications from more than 50,000 source titles and more than 1 billion citations. Dimensions' data is freely accessible via its search interface, although the full version is a paid service, providing a greater number of analyses and API access. Despite this, the company grants no-cost access to the full version of the database for non-commercial purposes (Herzog, Hook & Konkiel, 2020).

Finally, the OpenAlex database emerges from the discontinued Microsoft Academic Graph (MAG). The non-profit organisation OurResearch announced that they would preserve and incorporate the last complete MAG corpus (excluding only patent data) and would aggregate data from Crossref. In January 2022 OurResearch launched OpenAlex, providing access via API and the ability to export all their data for free (Scheidsteger & Haunschild, 2022).

According to information from the website (April 2023), OpenAlex has approximately 250 million publications and 1.9 billion citations. The database does not yet offer a user-friendly interface for searches, but makes all its data available through an API and data dumps.

1.2. Objectives of this study

In this research we will comparatively analyse the coverage of publications indexed in the SciELO Network database in relation to the WoS, Scopus, Dimensions and OpenAlex databases. With this we expect to contribute with the first large-scale analysis of the coverage a potential of each of them for the production of scientometric indicators for SciELO publications.

More specifically, this research targets the following questions:

- What is the total number of publications and publications with DOIs in the SciELO database and their temporal trends?
- How is the coverage of publications with DOIs from the SciELO database by country collections?
- How many SciELO publications with DOIs are covered in the WoS, Scopus, Dimensions and OpenAlex databases? What is the temporal trend of their coverage?

2. Data and methods

Despite that large amounts of data are available, it is often very difficult to use them given file formats, different ways data forms are filed, different storage and searching tools and lack of standardised documentation. For this work we have used a cloud computing infrastructure – using Google BigQuery technology- developed in the context of the project InSySPo (University of Campinas) [https://www.ige.unicamp.br/insyspo/], as well as the SQL data infrastructure available at CWTS (Centre for Science and Technology Studies, Leiden University).

Through a data dump provided by SciELO (October 2022), including publications' metadata, we constructed different tables to collect bibliographic information (see scripts associated with this manuscript in section Open science practices). The data dump is made available as a folder with XML files, which were converted into JSON format and uploaded to the Google Big Query (Krishnan et al, 2015) data warehouse. Inside this platform, it is possible to collect the fields with relevant information and export them into tables specific for our queries.

We collected information about the number of publications, DOIs and publication years. Also, we collected the data on the same publications with DOIs from OpenAlex, Scopus, Dimensions and WoS. The commercial databases were available only at CWTS and OpenAlex was uploaded according to their documented instructions. Using the DOI as publication identifier we studied the coverage of SciELO publications in all the different databases. Analyses and figures were produced using the Python language that directly query the tables we have collected the comparisons here presented (Bisong & Bisong, 2019).

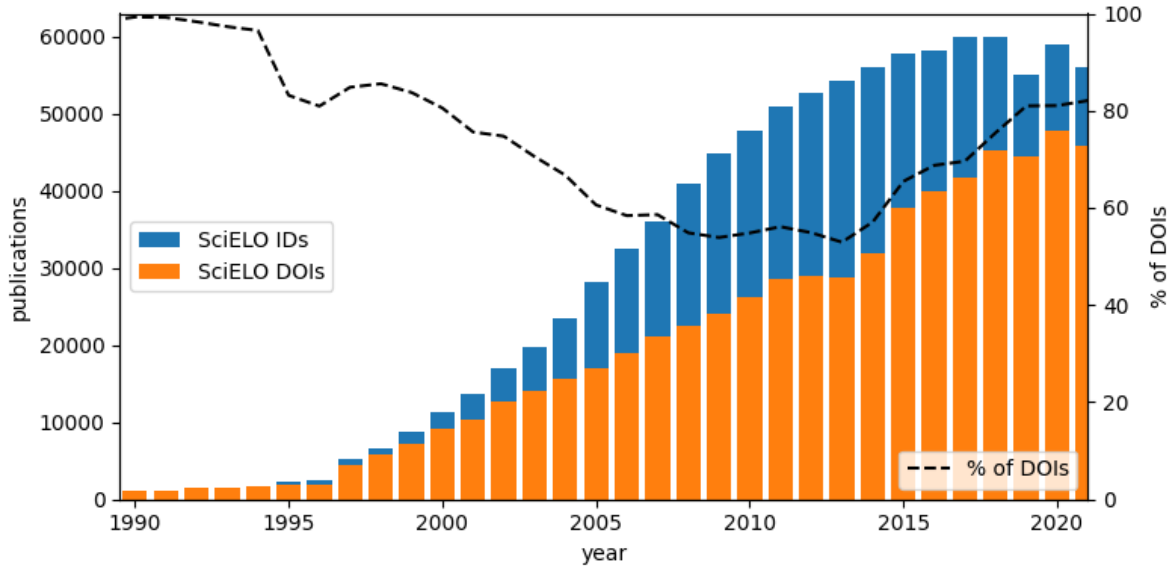
All the steps and links are made public in a Github repository accessible in https://github.com/alyssonmazoni/scielo_scientometrics.

3. Results and discussion

The total number of publications indexed in the SciELO dump is 1,017,103, of which 687,470 (68%) have a DOI. The database has publications from 1909 to 2024. The number of articles indexed up to 1995 was proportionally small (3% of the total). Since 1997 when SciELO was created, we observed that the number of indexed articles grew progressively: 94% of all publications between 1997 and 2022.

The average annual growth of publications indexed in the database during the entire period was 22%. The average annual growth between 1997 and 2013 was maintained in 2013 at 22%. After this period, the growth of the number of publications decreases, with an average annual growth between 2014 and 2021 of 1% (Figure 1). In parallel, the number of records with DOIs shows a gap that has evolved over time and seems to be decreasing in the past few years. Between 2017 and 2021, 80% of all articles indexed in SciELO have a DOI (Figure 1).

Figure 1: Publications with and without DOI indexed in the SciELO database by year (1990-2021).



SciELO currently has 19 collections, whose publications are highly concentrated (Table 1). Most of them are country collections, while there is a thematic collection (Public Health) and Ciência e Cultura. The largest collections are SciELO Brazil, with approximately 45% of all publications, followed by SciELO Colombia (9%) and SciELO Mexico and SciELO Chile (8% of publications each). The other collections together account for approximately 30% of publications. This concentration in some countries represents the scientific apparatus, population size and publication policies adopted by each of the countries, among others.

In Table 1 we can also observe the total publications and those with DOIs indexed in SciELO, by collections. There are important disparities between the number of publications with and without DOIs depending on the collection. The SciELO Brazil collection has DOIs for practically all the publications (99%). Other countries have collections with relatively good DOIs coverage, such as SciELO Ecuador (88%), SciELO Chile (79%), and SciELO Paraguay (72%). Other collections have a proportion of publications with a DOI that leaves much to be desired, with emphasis on SciELO Argentina, SciELO West Indians, SciELO Bolivia, SciELO Cuba and SciELO Venezuela, which appear at the bottom of Table 1. Collections with as many

DOIs as possible, would facilitate the production and analysis of more robust bibliometric indicators, facilitating the planning of scientific policies for all countries/collections that are part of SciELO.

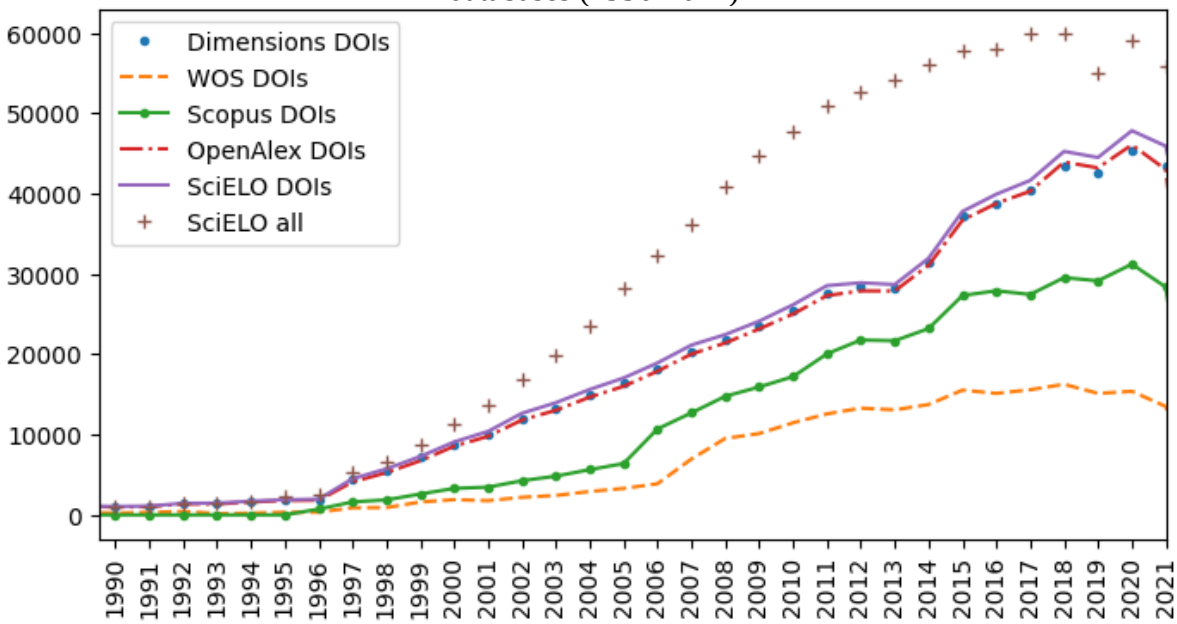
Table 1 - publications with and without DOI indexed in the SciELO database by collection (1909-2024)

Collection	Total of publications	Publications with DOIs	%
SciELO Brazil	477971	472003	99%
SciELO Public Health	50383	44247	88%
SciELO Ecuador	2219	1901	86%
SciELO Chile	81311	64101	79%
SciELO Paraguay	3629	2629	72%
SciELO South Africa	37078	21318	57%
SciELO Colombia	94325	44291	47%
SciELO Peru	14450	6082	42%
SciELO Spain	46480	19170	41%
SciELO Portugal	25813	10582	41%
SciELO Uruguay	6211	2315	37%
SciELO Mexico	87625	29647	34%
SciELO Costa Rica	10810	3353	31%
SciELO Ciência e Cultura	1975	508	26%
SciELO Argentina	51875	6905	13%
SciELO West Indians	1391	84	6%
SciELO Bolivia	5342	63	1,2%
SciELO Cuba	43759	111	0,3%
SciELO Venezuela	18971	0	0%

We observe that an increasing number of SciELO publications with DOIs are being indexed by the analysed databases. Dimensions and OpenAlex cover virtually all SciELO DOIs over time. Considering the entire period (1909 to 2024), the greatest coverage of SciELO publications with DOIs is done by Dimensions (95%), closely followed by OpenAlex (93%). With lower coverage are Scopus (57%) and WoS databases (33%).

The WoS and Scopus databases started to cover a larger part of the current SciELO production only from 2018 onwards. Before that, the average annual coverage of WoS was approximately 5% in both databases. Considering the period between 2010 and 2021, WoS now covers an average of 40% and Scopus 70% of the production indexed in SciELO with DOIs. It should be noted that approximately 30% of the total SciELO publications do not have DOIs, which if they would be included in the analysis, the results could be significantly different (see the “SciELO all” row in Figure 2).

Figure 2: Coverage of SciELO DOIs indexed in WoS, Scopus, Dimensions and OpenAlex databases (1990-2021)



4. Preliminary conclusions and outlook

Although publications with DOIs in SciELO have apparently stabilised at high levels (80%), the ideal would be to reach 100%. There is apparent stagnation in the growth of the number of articles indexed in SciELO, which suggests that the database is approaching the core collection of journals that complies with SciELO indexing criteria and therefore is not indexing new journals at the same speed as before. We understand that both points should be the focus of SciELO's future attention.

DOIs are a great facilitator for collection and analysis of bibliometric (and altmetric) data, its absence brings numerous difficulties and limitations to the production of reliable indicators. Although approximately 70% of the articles indexed by SciELO have DOIs, we observe important disparities between the collections/countries. It is recommended that SciELO seek ways to expand the insertion of DOIs, supporting countries with greater technical difficulties and creating mechanisms for editors to include this identifier in all publications.

As the Dimensions and OpenAlex databases cover practically all of the SciELO DOIs, these databases represent relevant sources for studying SciELO-covered production. However, the ideal would be for SciELO to promote better ways of making its data available, as there is a lot of relevant information from SciELO that is not captured by other databases (for example, details of citation data, funding, authorship, affiliations, date stamps from submission to publication, rankings of publications by collections, number of accesses and downloads, etc.). Future research should also focus on the metadata completeness of SciELO records in both Dimensions and OpenAlex, which would be an additional indication of the potential of these databases to study SciELO publications. Moreover, SciELO publications without DOIs should also be scrutinised for coverage in these and other databases (for example, Crossref) in future research.

Regarding the WoS and Scopus databases, they show a relatively lower coverage of SciELO publications. This raises the following three points of reflection. The first refers to the lack of

retrospective coverage of SciELO oldest publications, probably due to the more rigid WoS and Scopus indexing policies. The second refers to the decline in the coverage of more recent (2013-2022) SciELO publications in WoS and Scopus (approximately 34% and 62%, respectively), which may call for investigations on the consistency, integrity and transparency of the indexing policies and criteria of these two selective databases. Third, our results indicate a levelling off and even decreasing trend in the number of SciELO publications in most recent years, which might be due to a decrease in research that is communicated by SciELO journals, or an increase in submissions to non-SciELO journals.

In this research, we only analysed aspects of coverage of articles with DOIs from the SciELO database in WoS, Sopus, Dimensions and OpenAlex databases. Thus, we assume the limitation of not having analysed the full coverage of SciELO articles in these databases. It is our intention in future research to analyse the coverage in its entirety, using other techniques to find similarities between other elements of the article (such as title, author, journal, etc.). It is possible that the provided SciELO data dump may still miss some of its DOIs, therefore using a matching algorithm on titles and authors could alleviate this problem.

There are other analytical possibilities for future research around the SciELO database. Some examples are coverage in terms of journals, citations, countries, languages of publication, downloads, publication access, presence in social networks, etc. We also intend to carry out a detailed analysis of SciELO publications as a future work alongside a knowledge graph encompassing all (or nearly all) metadata presented as interconnected tables. In our research, we did not analyse the coverage of DOIs per collection in relation to each of the databases. As the coverage of the Dimensions and OpenAlex databases is high, we intend to do future analysis of coverage by collection in WoS and Scopus. We believe that there may be important disparities, related to the low occurrence of DOIs in some collections.

The SciELO database has unique analytical characteristics that need to be valued. It has a relevant source for open access scientific publication in the regions that it covers. It is also a platform targeted to increase the visibility and internationalisation of the countries' scientific output; and, as shown in this study, it also provides a unique database for the analysis of the scientific dynamics of the regions and researchers publishing there. Even though the Dimensions and OpenAlex databases have a good coverage of SciELO, SciELO on its own remains relevant for scientometric studies, as it contains metadata elements that may not exist in other databases (e.g. acknowledgements, counts on views and downloads, submission to publication dates, etc.). Finally, we would like to remark the need to seek ways to improve the large-scale availability of SciELO data for researchers, policy makers, librarians, and any other interested stakeholders. Such large-scale availability may represent the tipping point for the SciELO database to become an open research scientometric infrastructure on its own.

5. Bibliographic references

Beigel, F., Packer, A. L., Gallardo, O., & Salatino, M. (2024). OLIVA: La Producción Científica Indexada en América Latina. Diversidad Disciplinar, Colaboración Institucional y Multilingüismo en SciELO y Redalyc (1995-2018). *Dados*, 67(1), e20210174. DOI: https://doi.org/10.1590/dados.2024.67.1.307

Bisong, E., & Bisong, E. (2019). Google colaboratory. Building machine learning and deep learning models on google cloud platform: a comprehensive guide for beginners, 59-64.

CLACSO. Consejo Latinoamericano de Ciencias Sociales. (2020). Diagnosis and proposals for a regional initiative: Towards a Transformation of Scientific Research Assessment in Latin America and the Caribbean Series from The Latin American Forum for Research Assessment (FOLEC). Buenos Aires. Available from: http://biblioteca.clacso.edu.ar/Argentina/folec/20210528060657/FOLEC-Diagnostico-IN.pdf

Demachki, É., & Maricato, J. D. M. (2022). Coverage of Data Sources and Correlations Between Altmetrics and Citation Indicators: The Case of a Brazilian Portal of Open Access Journals. *Serials Review*, *48*(1-2), 151-166. DOI: https://doi.org/10.1080/00987913.2022.2066967

Herzog, C., Hook, D., & Konkiel, S. (2020). Dimensions: Bringing down barriers between scientometricians and data. *Quantitative Science Studies*, *1*(1), 387-395. DOI: https://doi.org/10.1162/qss_a_00020

Krishnan, S. P. T., Gonzalez, J. L. U., Krishnan, S. P. T., & Gonzalez, J. L. U. (2015). Google BigQuery. Building Your Next Big Thing with Google Cloud Platform: A Guide for Developers and Enterprise Architects, 235-253. DOI: https://doi.org/10.1007/978-1-4842-1004-8

Krüger, A. K. (2020). Quantification 2.0? Bibliometric infrastructures in academic evaluation. *Politics and Governance*, *8*(2), 58-67. https://doi.org/10.17645/pag.v8i2.2575

Larivière, V., & Sugimoto, C. R. (2018). *Mesurer la science*. Les Presses de l'Université de Montréal.

Loprieno, A., Werlen, R., Hasgall, A., & Bregy, J. (2016). The 'Mesurer les Performances de la Recherche'Project of the Rectors' Conference of the Swiss Universities (CRUS) and Its Further Development. *Research Assessment in the Humanities: Towards Criteria and Procedures*, 13-21. DOI: https://doi.org/10.1007/978-3-319-29016-4_2

Martín-Martín, A., Thelwall, M., Orduna-Malea, E., & Delgado López-Cózar, E. (2021). Google Scholar, Microsoft Academic, Scopus, Dimensions, Web of Science, and OpenCitations' COCI: a multidisciplinary comparison of coverage via citations. *Scientometrics*, *126*(1), 871-906. DOI: https://doi.org/10.1007/s11192-020-03690-4

Melo, J. H. N. D., Trinca, T. P., & Maricato, J. D. M. (2021). Limites dos indicadores bibliométricos de bases de dados internacionais para avaliação da Pós-Graduação brasileira: a cobertura da Web of Science nas diferentes áreas do conhecimento. *Transinformação*, *33*. https://doi.org/10.1590/2318-0889202133e200071

Minniti, S., Santoro, V., & Belli, S. (2018). Mapping the development of open access in Latin America and Caribbean countries. An analysis of web of science core collection and SciELO citation index (2005–2017). *Scientometrics*, *117*(3), 1905-1930. DOI: https://doi.org/10.1007/s11192-018-2950-0

Moed, H. (2002). Measuring China" s research performance using the Science Citation Index. *Scientometrics*, *53*(3), 281-296. DOI: https://doi.org/10.1023/A:1014812810602

Mugnaini, R., Digiampetri, L. A., & Mena-Chalco, J. P. (2014). Comunicação científica no Brasil (1998-2012): indexação, crescimento, fluxo e dispersão. *Transinformação*, *26*, 239-252. DOI: https://doi.org/10.1590/0103-3786201400030002

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

Packer, A. L. (2014). The emergence of journals of Brazil and scenarios for their future. *Educação e Pesquisa*, *40*, 301-323. DOI: https://doi.org/10.1590/S1517-97022014061860

Packer, A. L., Meneghini, R., Santos, S., Mendonça, A., Ramalho, A., Gesseff, E. & Saad, R. (2019). A coleção SciELO Brasil aos 20 anos. *Recuperado do https://www. scielo20. org/redescielo/wp-content/uploads/sites/2/2018/09/Informe-SciELO-Brasil-atualizada-1. pdf em*, *15*(09).

Packer, A. L., (2020). The Pasts, Presents, and Futures of SciELO. In: EVE, M. P, GRAY, J. eds. Reassembling Scholarly Communications: Histories, Infrastructures, and Global Politics of Open Access. (Cambridge): The MIT Press, 2020. pp.297-313. DOI: https://doi.org/10.7551/mitpress/11885.003.0030

Pranckutė, R. (2021). Web of Science (WoS) and Scopus: The titans of bibliographic information in today's academic world. *Publications*, *9*(1), 12. https://doi.org/10.3390/publications9010012

Rafols, I., Ciarli, T., & Chavarro, D. (2020, October 4). Under-reporting research relevant to local needs in the global south. Database biases in the representation of knowledge on rice. https://doi.org/10.31235/osf.io/3kf9d

Santin, D. M., & Caregnato, S. E. (2018). ÍNDICES DE CITAÇÃO NACIONAIS E REGIONAIS: importância, experiências e perspectivas para a América Latina. *Encontro Brasileiro de Bibliometria e Cientometria (6.: 2018 jul. 17-20: Rio de Janeiro, RJ). Anais [recurso eletrônico]. Rio de Janeiro, RJ: UFRJ, 2018*. https://www.lume.ufrgs.br/bitstream/handle/10183/183984/001075906.pdf?sequence=1

Santos, S. M., Fraumann G., Belli, S., & Mugnaini, R (2021). The relationship between the language of scientific publication and its impact in the field of Public and Collective Health, *Journal of Scientometric Research,* 10(1). DOI: https://doi.org/10.5530/JSCIRES.10.1S.24

Scheidsteger, T., & Haunschild, R. (2022). Comparison of metadata with relevance for bibliometrics between Microsoft Academic Graph and OpenAlex until 2020. *arXiv preprint arXiv:2206.14168*. DOI: https://doi.org/10.48550/arXiv.2206.14168

Simons, K. The Misused Impact Factor. Science. 2008;322(5899):165-5. DOI: https://doi.org/10.1126/science.1165316

Thelwall, M. (2018). Dimensions: A competitor to Scopus and the Web of Science?. *Journal of informetrics*, *12*(2), 430-435. DOI: https://doi.org/10.1016/j.joi.2018.03.006

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. DOI: https://doi.org/10.1162/qss_a_00112

Open science practices
The authors took care in making all the results public. All the steps in the analysis can be reproduced using the codes provided in the Github repository mentioned in the text: https://github.com/alyssonmazoni/scielo_scientometrics. We ask that credit be given to the authors. The raw tables for open data are available as public datasets inside the Google Big Query platform. The only steps that are not executed there are the queries to the commercial databases which cannot be made public. However, the nature of comparison demands their use. The aggregated results as presented here are made available in tables that are described and linked in the same repository. The tools used here were commercial tools applied to open data. The provider of these tools (Google) allows free access to some functionalities up to a certain use amount. We have used them to a larger capacity thanks to a research budget. The repository also includes some instructions for reproducibility in a local environment.

Acknowledgments
The authors thank the SciELO team for insightful conversations and feedback on the more technical aspects of the database and its metadata.

Author contributions
João de Melo Maricato - Conceptualization; Data curation; Methodology; Investigation; Validation; Writing – original draft.
Alysson Mazoni - Conceptualization; Data curation; Methodology; Investigation; Writing – original draft.
Rogério Mugnaini - Validation; Writing – review & editing.
Abel Packer - Validation; Writing – review & editing.
Rodrigo Costas - Conceptualization; Data curation; Methodology; Investigation; Validation; Writing – original draft; Writing – review & editing; Supervision.

Competing interests
ALP and RM are part of the SciELO project.

Funding information
Rodrigo Costas is partially funded by the South African DSI-NRF Centre of Excellence in Scientometrics and Science, Technology and Innovation Policy (SciSTIP) and supported by the InSySPo project.
João de Melo Maricato is funded by the Coordenação de Aperfeiçoamento de Pessoal de Nível Superior (CAPES).
Alysson Mazoni is funded by Fundação de Amparo à Pesquisa do estado de São Paulo (Fapesp) (process number 2021/05823-1).