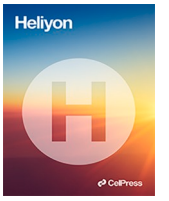
Heliyon 9 (2023) e21527

Contents lists available at ScienceDirect

Heliyon

journal homepage: www.cell.com/heliyon



A revisit to the specification of sub-datasets and corresponding coverage timespans when using Web of Science Core Collection

Andy Wai Kan Yeung [a]

[a] *Oral and Maxillofacial Radiology, Applied Oral Sciences and Community Dental Care, Faculty of Dentistry, University of Hong Kong, Hong Kong, China*

ARTICLE INFO

Keywords:
Web of Science Core collection
Reproducibility
Responsible research
Bibliographic database
Bibliometric analysis
Replication study

ABSTRACT

Many papers used the Web of Science Core Collection (WOSCC) as the data source. The work by Liu (2019, Scientometrics) revealed that only 48 % of such papers published in information science and library science journals during 2017–2018 specified which sub-datasets they used, and subsequently urged researchers to provide such information together with the corresponding coverage timespans to improve transparency and reproducibility. This work revisited this issue to reveal if the current condition has improved following Liu's recommendations. Using WOSCC, 934 bibliometric open-access papers published during 2020–2022 in non-information science and library science journals were evaluated. Of these 934 papers, 45.0 % specified the sub-datasets of WOSCC they used for data collection, and 4.8 % specified the coverage timespan(s) of corresponding sub-datasets or the overall dataset. There seemed to be no improvement in the specification of data source using WOSCC.

1. Introduction

Nowadays many bibliometric studies would extract and analyze data from the Web of Science Core Collection (WOSCC) database. In its official website, Clarivate Analytics stated that WOSCC has indexed "85.9 million records going back to 1900" published in more than "21,000 peer-reviewed journals" [1]. Indeed, the entire database is enormous, but academic institutions usually have customized subscription plans, meaning that only part of the entire database can be accessed [2]. Liu has highlighted this issue by showing the differences in the Citation Indexes (sub-datasets) and their corresponding coverage timespans according to the subscriptions from three different universities [2]. Liu illustrated that, at the time of the writing (July 18, 2019), University of Manchester could access to 8 Citation Indexes and 2 Chemical Indexes by its WOSCC subscription. Simultaneously, Xi'an Jiaotong University could access to 6 Citation Indexes and 2 Chemical Indexes, and Zhejiang University of Finance and Economics could only access to 4 Citation Indexes and 2 Chemical Indexes. Unlike the former 2 universities that could access some of the Citation Indexes as early as from 1900, the last university could only retrieve indexed records published since 2015 or 2016. For comparison, the subscription by the University of Hong Kong enabled access to 6 Citation Indexes and no Chemical Index, and records published as early as from 1956 indexed in the Social Sciences Citation Index (Fig. 1).

In short, the point to re-iterate here is that basically each institution could have a different subscription plan, so that the bibliometric results will be very different according to the Citation Indexes available and selected. Liu [2] evaluated 243 papers that used WOSCC as the data source. They were published in 34 journals in the category of Information Science and Library Science during

E-mail address: ndyeung@hku.hk.

https://doi.org/10.1016/j.heliyon.2023.e21527
Received 11 September 2023; Received in revised form 22 October 2023; Accepted 23 October 2023
Available online 2 November 2023

2017–2018. Indeed, it was found that 52 % of these papers did not specify which sub-datasets they used for data collection. No evaluation was done regarding the specification of the corresponding coverage timespans of the sub-datasets. The reported figure was surprising, as it was assumed that library science researchers would have a deeper understanding on this issue of heterogeneous datasets under the umbrella term of WOSCC. More than 3 years has passed since Liu published these findings and the recommendation to specify sub-datasets and corresponding coverage timespans, this report aimed to be a follow-up study to reveal if there have been any signs of improvement in the reporting details of bibliometric studies in this aspect. In particular, papers published in non-Information Science and Library Science journals were considered, to reveal if the recommended reporting practice was followed by the general research community. This was highly relevant as many bibliometric papers were published in the general literature.

2. Methods

On 1st May 2023, the WOSCC was queried (https://www.webofscience.com/wos/woscc/basic-search). To clarify, the following sub-datasets were included in the search:

- Science Citation Index Expanded (SCI-EXPANDED): 1970-present
- Social Sciences Citation Index (SSCI): 1956-present
- Arts & Humanities Citation Index (AHCI): 1975-present
- Conference Proceedings Citation Index - Science (CPCI–S): 2009-present
- Conference Proceedings Citation Index – Social Science & Humanities (CPCI-SSH): 2009-present
- Emerging Sources Citation Index (ESCI): 2005-present

The following search string was used:
#1: TS = "WOSCC" OR "Web of Science Core Collection" OR "WOS Core Collection"
#2: TS = bibliometric* OR scientometric*
#3: TS = Scopus.
#4: #1 AND #2 NOT #3.

Papers mentioning Scopus in their title, abstract or keywords were intentionally excluded, because many of such papers compared the strengths and weaknesses between WOSCC and Scopus in a qualitative way without data collection. Other databases were not mentioned in the search strategy as they were less frequently involved in database comparisons. Meanwhile, the terms bibliometric* or scientometric* were added to remove systematic reviews and meta-analysis papers without bibliometric components. Afterwards, the following filters were applied:

- Document Types = Article or Review Article
- Exclude – Web of Science Categories = Information Science Library Science
- Publication Years = 2020–2022
- Languages = English
- Open Access = All Open Access

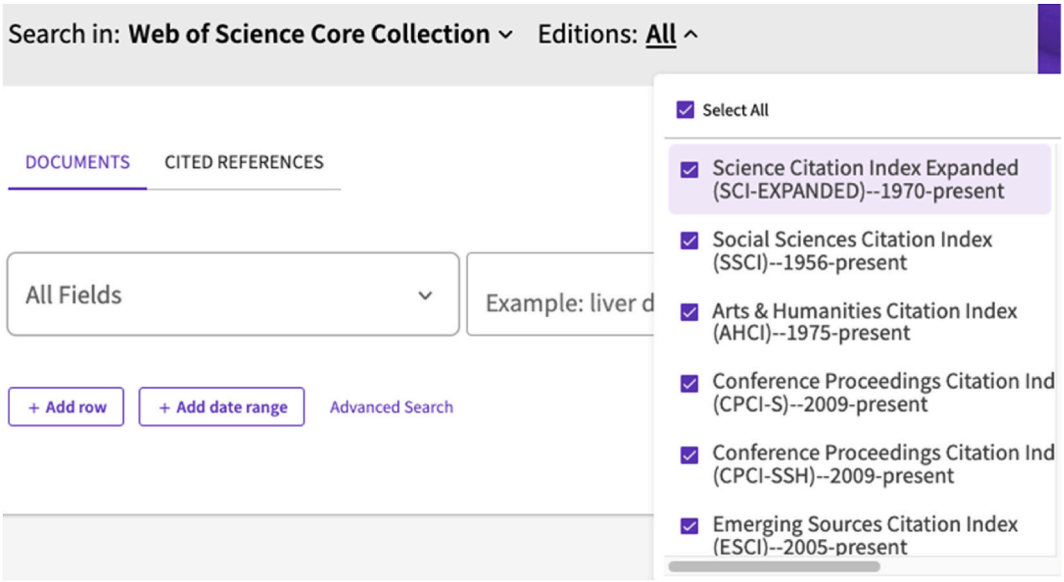


Fig. 1. Web of Science Core Collection subscribed by the University of Hong Kong. Date accessed on 11 May 2023.

The inclusion of open access papers only could allow readers to verify the results more easily. A total of 962 papers were resulted from the search. 28 of them were excluded with the following reasons: no open access to its main text (n = 19, implying potential mislabeling by WOS); published in 2023 (n = 6); duplicated papers (n = 2); and not using WOSCC for the literature search (n = 1). It might be counter-intuitive to see that 6 papers published in 2023 were returned from the search that limited publication years to 2020–2022. However, the inbuilt filters in the WOS platform, at the time of this writing, recognized the Publication Year as the year when the paper was published as an "online ahead of print" or "early access". Contradictorily, as the data was exported, the year when the paper was officially published in a journal volume with pagination details was recognized. This filtering issue for publication year has also been documented previously [3].

Hence, 934 papers were analyzed with the following evaluation items:

- Specification of WOSCC sub-datasets (Yes/No)
- Specification of the corresponding coverage timespans of the sub-datasets (Yes/No): Applicable only if it specified WOSCC sub-datasets
- Specification of overall WOSCC coverage timespan (Yes/No): Applicable only if it did not specify WOSCC sub-datasets

In addition, the timespan of literature search and its justification were also recorded, as the latter might sometimes implicitly provide information regarding the coverage timespan:

- Specification of timespan of literature search (Number of years)
- Justification of such timespan (Yes/No)

3. Results

The coded data sheet for the 934 papers was uploaded as Supplementary File 1. Overall, 420 of the 934 papers (45.0 %) specified the sub-datasets of WOSCC they used for data collection. This ratio remained relatively stable during the surveyed period (2020: 44.8 %, 2021: 48.7 %, 2022: 43.8 %; Fig. 2).

Among papers that specified the sub-datasets of WOSCC, only 25 (6.0 % of 420) papers specified the corresponding coverage timespans of the sub-datasets. Again, this ratio remained stable during the surveyed period (2020: 5.1 %, 2021: 5.2 %, 2022: 6.3 %; Fig. 3).

Among papers that did not specify the sub-datasets of WOSCC, 1 paper (0.2 % of 514) specified the overall coverage timespan:

> *"Our data are retrieved from the Web of Science Core Collection (from 1985 to 2020) … and 1087 articles were selected from the last ten years (2011–2020)"* [4].

In other words, regardless of whether the sub-datasets of WOSCC were specified or not, most of the papers did not disclose the coverage timespan(s) of their WOSCC subscription. Next, the justifications of the timespan of the literature search provided by the papers were examined. For the 908 papers that did not specify the coverage timespans of either the sub-datasets or the entire dataset, 19 of them (20.9 % of 908) provided justifications to their literature search timespan, together with the exact search timespan, that might implicitly provide information regarding the overall coverage timespan:

- From the start of the database coverage to the last completed year at the time of data collection, or to the data extraction date (n = 17)
- Used the default timeframe setting (n = 1)
- Due to the restrictions of the affiliation's subscription (n = 1)

By combining the numbers together (25 + 1 + 19), one could compute that information about the coverage timespan(s) of

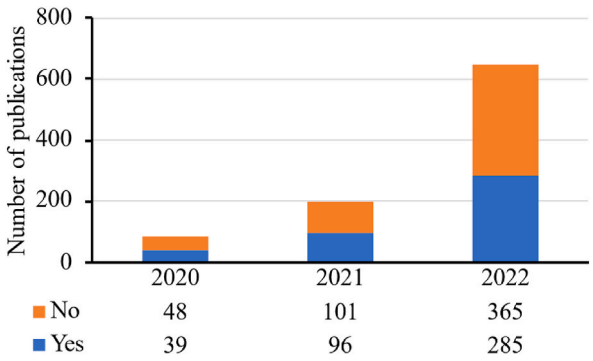


Fig. 2. Number of papers specifying the WOSCC sub-datasets during 2020–2022.

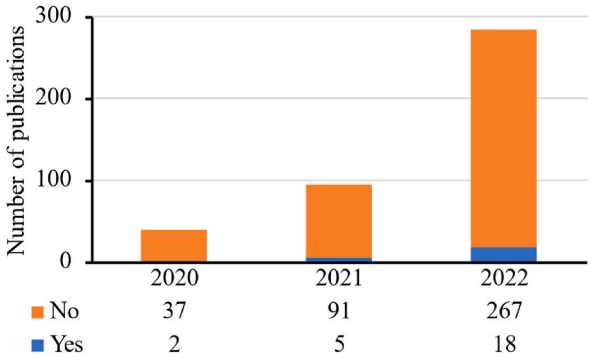


Fig. 3. Number of papers specifying the corresponding coverage timespans of the used WOSCC sub-datasets during 2020–2022.

corresponding sub-datasets or overall database was available from 45 papers (4.8 % of 934). This ratio is much lower than that of specifying the sub-datasets of WOSCC they used for data collection (45.0 % as reported above).

4. Discussion

This study found that 45.0 % of the 934 bibliometric papers published during 2020–2022 have specified the sub-datasets of WOSCC they used for data collection. This number was slightly smaller than the 48 % reported by Liu [2] who analyzed 243 papers published during 2017–2018 in journals specialized in information science and library science. The difference seemed to be very small. This implied that the situation has not improved since Liu [2] has published the findings and urged for better reporting details by specifying the WOSCC sub-datasets and corresponding coverage timespans.

Worse than that, only 4.8 % of the 934 papers specified the coverage timespan(s) of each sub-dataset or the entire dataset used for the bibliometric analysis. At times, the literature search timespan could be based on when the research topic has accumulated a specific number of publications, or when a seminal/landmark document was published [5]. Hence, researchers might mislead readers if the data interpretation and discussion did not account for the coverage timespans of the datasets. For instance, if a research team had subscription to the WOSCC that enabled access to publications since 2010, the discussion part of the bibliometric results should not suggest that the analyzed research field had been largely stagnant but grown rapidly since 2010.

Among the 934 papers analyzed here, only one study [6] has cited [2] to substantiate the line that "Web of Science is a commonly used database in many disciplines, particularly in the medical and health domain", which was not the main point raised by the latter. Unfortunately, this study did not specify WOSCC sub-datasets, or the coverage timespan(s) of sub-datasets/entire dataset. Of course, researchers may not choose to cite any references given that a methodology or a concept has become well-known in a universal way, a citation behavior called "obliteration by incorporation" previously observed for various research concepts [7,8]. In this case, however, the widespread specification of WOSCC sub-datasets and coverage timespans could not be observed. Actually, there have been other independently published recommendations suggested for bibliometric studies. Besides advocating for the specification of WOSCC sub-datasets and their corresponding coverage timespans, Liu also suggested that, whenever a literature search via WOSCC had its timespan covering 1990 or before, additional caution should be taken during data interpretation [9]. It is because the abstract, author keywords, and KeyWords Plus information of records published on 1990 or before were mostly unavailable on the sub-datasets of SCI-Expanded, SSCI, and AHCI. Meanwhile, the research team led by Ho suggested that literature searching via WOSCC should be done with the "front page" search technique, meaning that the search terms should only be applied to the title, abstract, and author keywords fields, excluding the KeyWords Plus field, to improve the relevance of the search results [10,11]. Moreover, Donthu et al. [12] published a step-by-step walkthrough on how to conduct and report bibliometric studies. For instance, they suggested that, during data network visualization, definitions should be provided to the node size, edge (link) thickness, and color of the components. They also suggested that a dataset of >500 publications is more suitable for a bibliometric analysis, whereas data cleaning (such as removing duplicates and wrong/irrelevant entries) may still be important. Interestingly, this walkthrough was published in Journal of Business Research instead of an information science or library science journal. Moreover, researchers who queried WOSCC with a collection of digital object identifiers (DOIs) should be aware that WOSCC might record wrong DOI names for the indexed papers, such as replacing "O" by "o", and "b" by "6" [13]. A recent study has introduced an algorithm to systematically "clean"/correct DOI names in WOSCC with some success [14]. Readers should also be aware that the data filters of WOS are not perfect, as papers could be tagged/labelled wrongly, such as document type assignment [15,16] and funding information [17]. To safeguard data accuracy/quality, researchers should perform manual inspection for a subset of the entire dataset, e.g. randomly select 10 %, exported from WOSCC to be analyzed.

Considering the need to standardize the reporting details of bibliometric studies, a reporting guideline for bibliometric studies, called Guidance List for repOrting Bibliometric AnaLyses (GLOBAL), is currently under preparation by an international research team with its research protocol available at the Open Science Framework website [18]. This is a great initiative, as bibliometric studies are specialized documents that cannot be universally classified as either original article or review paper [19]. It is hoped that this GLOBAL guideline could integrate the individual recommendations, be well disseminated and followed by the research community resembling guidelines for other study designs, such as the Preferred Reporting Items for Systematic Reviews and Meta-Analyses (PRISMA) [20], or

the Strengthening the Reporting of Observational Studies in Epidemiology (STROBE) guidelines [21].

There were some limitations of this study. First, the examined dataset did not cover the entire relevant literature. For instance, the inclusion of open access papers could allow readers to verify the results without the barrier from the paywall, but inevitably some relevant papers were subsequently excluded that could compromise the generalizability of the results. At the same time, readers should also note that some open access journals might have financial considerations to maximize publishing revenue by having low selectivity (high acceptance rate) [22], which might again affect the generalizability of the results from this study. Moreover, few relevant papers that did not mention bibliometric* OR scientometric* in their title, abstract, and keywords would be excluded. Besides, readers should be aware of the differences in WOS, (Web of Knowledge) WOK, and WOSCC [23]. Basically, the platform to access the literature databases was called WOK prior to 2014, and renamed as WOS in 2014. Within the WOS (platform), users can access to various citation indexes (one of which is called WOSCC), as well as other product databases and the Derwent Innovations Index. Unfortunately, at times researchers might use these terms interchangeably, rendering some confusion [23].

5. Conclusion

In conclusion, this study found that only 45.0 % of the 934 bibliometric papers published during 2020–2022 have specified the sub-datasets of WOSCC they used for data collection. Worse than that, only 4.8 % of the 934 papers specified the coverage timespan(s) of each sub-dataset or the entire dataset used for the bibliometric analysis. Detailed and correct information regarding the use of WOSCC should be reported in future studies. If word limit is an issue, authors should consider reporting the details as supplementary materials.

Data availability statement

Data is provided as Supplementary File 1.

CRediT authorship contribution statement

Andy Wai Kan Yeung: Writing – review & editing, Writing – original draft, Data curation, Conceptualization.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgments

This work was supported by departmental funds only.

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.heliyon.2023.e21527.

References

[1] Clarivate Analytics, Web of Science Core Collection, 2023. https://clarivate.com/products/scientific-and-academic-research/research-discovery-and-workflow-solutions/web-of-science/web-of-science-core-collection/.
[2] W. Liu, The data source of this study is Web of Science Core Collection? Not enough, Scientometrics 121 (3) (2019) 1815–1824.
[3] W. Liu, A matter of time: publication dates in Web of Science Core Collection, Scientometrics 126 (1) (2021) 849–857.
[4] Y. Jia, Y. Chen, P. Yan, Q. Huang, Bibliometric analysis on global research trends of airborne microorganisms in recent ten years (2011-2020), Aerosol Air Qual. Res. 21 (2) (2021), 200497.
[5] J.P. Romanelli, M.C.P. Gonçalves, L.F. de Abreu Pestana, J.A.H. Soares, R.S. Boschi, D.F. Andrade, Four challenges when conducting bibliometric reviews and how to deal with them, Environ. Sci. Pollut. Control Ser. 28 (2021) 60448–60458.
[6] Z. Liu, L. Ren, C. Xiao, K. Zhang, P. Demian, Virtual reality aided therapy towards health 4.0: a two-decade bibliometric analysis, Int. J. Environ. Res. Publ. Health 19 (3) (2022) 1525.
[7] W. Marx, L. Bornmann, How accurately does Thomas Kuhn's model of paradigm change describe the transition from the static view of the universe to the big bang theory in cosmology? A historical reconstruction and citation analysis, Scientometrics 84 (2) (2010) 441–464.
[8] A.W.K. Yeung, Is the influence of freud declining in psychology and psychiatry? A bibliometric analysis, Front. Psychol. 12 (2021), 631516.
[9] W. Liu, Caveats for the use of Web of Science Core Collection in old literature retrieval and historical bibliometric analysis, Technol. Forecast. Soc. Change 172 (2021), 121023.
[10] H.-Z. Fu, Y.-S. Ho, Top cited articles in adsorption research using Y-index, Res. Eval. 23 (1) (2014) 12–20.
[11] H.-Z. Fu, M.-H. Wang, Y.-S. Ho, The most frequently cited adsorption research articles in the Science Citation Index (Expanded), J. Colloid Interface Sci. 379 (1) (2012) 148–156.
[12] N. Donthu, S. Kumar, D. Mukherjee, N. Pandey, W.M. Lim, How to conduct a bibliometric analysis: an overview and guidelines, J. Bus. Res. 133 (2021) 285–296.
[13] J. Zhu, G. Hu, W. Liu, DOI errors and possible solutions for Web of Science, Scientometrics 118 (2019) 709–718.

[14] S. Xu, L. Hao, X. An, D. Zhai, H. Pang, Types of DOI errors of cited references in Web of Science with a cleaning method, Scientometrics 120 (2019) 1427–1437.
[15] P. Donner, Document type assignment accuracy in the journal citation index data of Web of Science, Scientometrics 113 (1) (2017) 219–236.
[16] A.W.K. Yeung, Comparison between Scopus, Web of science, PubMed and publishers for mislabelled review papers, Curr. Sci. 116 (11) (2019) 1909–1914.
[17] B. Álvarez-Bornstein, F. Morillo, M. Bordons, Funding acknowledgments in the Web of Science: completeness and accuracy of collected data, Scientometrics 112 (2017) 1793–1812.
[18] J.Y. Ng, S. Haustein, S. Ebrahimzadeh, C. Chen, M. Sabe, M. Solmi, D. Moher, Guidance List for repOrting Bibliometric AnaLyses (GLOBAL), 2023, https://doi.org/10.17605/OSF.IO/MTXBF.
[19] A.W.K. Yeung, Document type assignment by Web of science, Scopus, PubMed, and publishers to "top 100" papers, Malaysian, J. Libr. Inf. Sci. 26 (3) (2021) 97–103.
[20] M.J. Page, J.E. McKenzie, P.M. Bossuyt, I. Boutron, T.C. Hoffmann, C.D. Mulrow, et al., The PRISMA 2020 statement: an updated guideline for reporting systematic reviews, Int. J. Surg. 88 (2021), 105906.
[21] E. Von Elm, D.G. Altman, M. Egger, S.J. Pocock, P.C. Gøtzsche, J.P. Vandenbroucke, The Strengthening the Reporting of Observational Studies in Epidemiology (STROBE) statement: guidelines for reporting observational studies, Lancet 370 (9596) (2007) 1453–1457.
[22] K. Siler, Demarcating spectrums of predatory publishing: economic and institutional sources of academic legitimacy, J Assoc Inf Sci Technol 71 (11) (2020) 1386–1401.
[23] G. Hu, L. Wang, R. Ni, W. Liu, Which h-index? An exploration within the Web of science, Scientometrics 123 (2020) 1225–1233.