A Perspective on Merging databases for bibliometric analysis

Namrata Dagli [1], Jestina Rachel Kurian [2], Mainul Haque [3]

ABSTRACT

Keywords

Bibliometrics, Database-merging, Database, Integration, Scopus, WoS, Dimensions, Scientometrics, Techno-bibliometric

Bibliometric analysis refers to the quantitative evaluation of academic literature, including the counting of citations, tracking of publication trends, mapping of co-authorship networks, and identification of emerging themes. Traditionally, scholars have relied on one major database such as Scopus or Web of Science (WoS). In recent years, the trend of merging data from multiple sources has been on the rise, driven by the perceived benefits of combining bibliographic databases. This editorial provides a critical perspective on the rationale and challenges of cross-database integration for bibliometric analysis.

An analysis by Singh et al. 2021 found that 99.11% of the journals indexed in WoS are also indexed in Scopus and 96.61% in Dimensions. In addition, Scopus has 96.42% of its indexed journals also covered by Dimensions, showing a high degree of overlap [1]. Another study also concluded that while all databases retain some unique content, Scopus shares a substantial overlap with Dimensions and Crossref [2]. These findings suggest that merging can broaden coverage, but the benefit is mostly marginal.

In addition, the marginally broadened coverage achieved through database merging presents significant challenges. A major concern involves the comparability of h-index values, which vary notably between databases due to differences in coverage. A study found that among 350 Monash University researchers, only 31% had identical h-indices in both Scopus and WoS; 55% scored higher in Scopus, while 9% scored higher in WoS [3]. This clearly shows that h-indices from different platforms are not comparable, and merging studies must always report the source of the metric.

Other methodological issues include duplicate records, which are caused by minor differences in titles, author name formats, or missing DOIs, and can distort counts. Author names and institutional affiliations often appear in varying formats across databases, requiring manual harmonization. Additionally, different authors with the same acronyms might appear as a single author, distorting productivity. However, recent advancements have introduced tools to address these challenges in merging bibliometric data from databases like Scopus, WoS, Dimensions,

1. Adjunct Research Faculty, Center for Global Health Research, Saveetha Medical College, Saveetha Institute of Medical and Technical Sciences (SIMATS), Chennai, India. Email: dr.namrata.dagli@gmail.com.
2. Jestina Rachel Kurian, Department of Data Science, Prasanna School of Public Health, Manipal Academy of Higher Education, Manipal, India. Email: jestina.rachel@learner.manipal.edu
3. Independent Researcher. Former Professor of Malaysia and Bangladesh. Block C, Road 10, House 266, Khilgaon, Dhaka 1219, Bangladesh. Email: runurono@gmail.com.

Correspondence

Mainul Haque, Independent Researcher. Former Professor of Malaysia and Bangladesh. Block C, Road 10, House 266, Khilgaon, Dhaka 1219, Bangladesh. Email: runurono@gmail.com. Cell Phone: +8801703605918. WhatApp: +60109265543

DOI: https://doi.org/10.3329/bjms.v24i4.84667

and OpenAlex. Tools such as BibexPy [4] and ASySD [5] improve deduplication and metadata harmonization. Other open-source solutions, including a high-accuracy preprocessing framework [6] and KKU-BiblioMerge [7], support seamless integration and cleaning for more reliable analysis. While the use of the software might make the merging process more standardized by making it reproducible and less subjective, several issues still persist, affecting the accuracy of analysis, such as definitions of article types are also inconsistent; what one database labels "conference paper" might be listed as "journal article" in another [2,8].

Citation counts here, too, differ, reflecting the varied indexing strategies employed. Citation network studies (e.g., co-citation analysis or bibliographic coupling) are also compromised because identical references may be formatted differently across sources, resulting in fragmented or artificial links [2]. Moreover, field-normalized metrics become invalid without consistent subject categorization. Addressing these inconsistencies and category mismatches between databases remains an obstacle, requiring sophisticated preprocessing and normalization strategies that are currently underdeveloped. Current technology for merging and mapping bibliometric data requires significant upgrades to effectively address existing limitations. While various tools aim to automate deduplication, their performance is still contingent on the quality of the data input and robust preprocessing strategies.

Additionally, the effectiveness of these tools in merging very large quantities of data is limited by factors such as system memory [6]. Evidence suggests that utilizing merged data can lead to different outcomes compared to analyses based solely on a single database [6]. As merging bibliometric data involves multiple steps, such as deduplication, metadata harmonization, and format conversion, the results can vary depending on the tools, settings, and techniques used. Due to this technique-sensitivity, reproducibility becomes a challenge, making it harder to compare findings across bibliometric studies. Furthermore, widely used mapping tools like VOSviewer encounter specific structural and technical limitations when encountering merged data and therefore fail to generate maps according to the tool's full functionality. Similarly, since PubMed does not include cited references, maps for co-citation or co-authorship cannot be generated. As a result, the retrieved information does not contribute to citation-related analysis. All these issues complicate merging and demand intensive quality control.

These challenges don't just affect traditional bibliometric studies; they also hinder our ability to connect research papers with technological developments. For instance, Techno-bibliometric approaches attempt to map the interface between scientific literature and technological innovation by analyzing paper and patent metadata, including citations. When performing techno-bibliometric analysis, particularly with patent databases like Lens.org or PatCite, it is essential that the scientific datasets being merged preserve clean, consistent metadata to ensure valid mapping. Thus, while merging may enhance coverage, it also introduces risks of disrupting citation structures that underpin assessments of science–technology interactions.[9]

Given these issues, merging is most appropriate only for metadata that remains consistent across databases, such as DOIs, titles, publication years, and journal identifiers. These fields are reliably matched and less prone to variation across sources. In contrast, author names, affiliations, and citation count often vary and require manual refinement. Despite advancements in deduplication and integration tools, rigorous data cleaning remains essential to ensure accuracy. Thus, manual efforts in data reconciliation are time-intensive and prone to human error, especially when dealing with large datasets from multiple sources [10].

Apart from the inherent limitations of incompatible databases, variations in data quality, limitations of tools and methods for cross-database merging, and constraints in mapping tools for reading the merged data, researchers face additional challenges. Many lack the extensive skills in coding and data management required to successfully and accurately harmonize and analyze the merged data [11]. In countries with limited resources, researchers may not have access to multiple scientific databases, raising questions of inclusivity and equity in research opportunities. On a different note, while merged datasets often limit the feasibility of network-based analyses such as co-authorship mapping, citation metrics, or institutional performance tracking due to inconsistencies in author names, affiliations, and

citation formats, they can still be valuable for thematic analyses. Specifically, merged data can support topic clustering, keyword-based visualizations, and other forms of content-driven exploration where exact metadata alignment is less critical.

A key question then emerges: When is it appropriate to use cross-database inputs for bibliometric analysis? A high overlap suggests that merging may add marginal value while considerably increasing effort, particularly when the objectives are exploratory rather than confirmatory. In addition, if the overlap is high, merged datasets may disproportionately capture low-impact or duplicated records, diluting the informative value rather than enriching it. This suggests the benefits of merging vary by field and research purpose. The criteria for deciding the appropriateness of combining or merging the databases are mentioned in Figure 1.

Merging databases: Decision matrix

Consideration	Merge if...	Do Not Merge if...
Unique coverage	Significant non-overlapping records	Content is almost identical across databases
Metadata compatibility	Fields like DOI, author, title align across sources	Inconsistent formats and missing identifiers
Tool availability	Access deduplication tools (e.g., DOI matching)	No reliable deduplication method available
Objective of Analysis	You need broader coverage (e.g., thematic mapping)	Analyzing citation networks/ impact indicators
Naming conventions	Affiliations and country fields are standardized across sources	There are many variations in names, affiliations.
Time and expertise	You have resources and time for cleaning and validation	Timeline or resources are tight

Figure 1- Decision Matrix: The criteria to decide the appropriateness of merging the databases

In conclusion, while merging databases can enhance bibliometric analysis to some extent, it incurs significant costs in terms of time and effort. If the goal is a broad, exploratory analysis, using a single well-chosen database can be sufficient. However, if completeness, discipline-specific coverage, or cross-validation is essential to meet specific research goals, merging may be justified if conducted with a carefully planned methodology that includes consistent metadata, careful deduplication using reliable software, clear h-index source attribution, and transparency at every stage. The focus should be on ensuring data integrity, reproducibility, and clarity of methods rather than simply layering more databases onto a study.

Looking ahead, there is a need for more integrated and standardized bibliometric practices. Unified APIs and standard metadata export formats would make it easier to combine datasets cleanly. Bibliometric tools should be enhanced to support multi-database import while automatically identifying and merging duplicates, identifying and harmonizing author and institution names, and labeling each metric with its source. There is also a growing need to establish bibliometric indices that maintain validity and comparability across databases, helping ensure consistent evaluations regardless of the source used. This can minimize bias caused by differences in coverage or the definition of metrics across platforms. Standard reporting guidelines, like a checklist specifying data sources, coverage dates, deduplication strategies, and source attribution for metrics, would further enhance transparency. Although merging databases can enhance study robustness, it demands careful decision-making, prioritizing depth over breadth and quality over quantity.

Consent for Publication

The author has reviewed and approved the final version and agrees to be accountable for all aspects of the work, including any accuracy or integrity issues.

Disclosure

Mainul Haque works as an editorial team member of the Journal of Applied Pharmaceutical Science, India. The remaining authors declare that they do not have any financial involvement or affiliations with any organization, association, or entity directly or indirectly related to the subject matter or materials presented in this review paper.

Data Availability

Information for this review paper is taken from freely available sources.

Authorship Contribution

All authors contributed significantly to the work, whether in the conception, design, utilization, collection, analysis, or interpretation of data, or all these areas. They also participated in the paper's drafting, revision, or critical review, gave their final approval for the version that would be published, decided on the journal to which the article would be submitted, and made the responsible decision to be held accountable for all aspects of the work.

REFERENCES

1. Singh VK, Singh P, Karmakar M, Leta J, Mayr P. The journal coverage of Web of Science, Scopus, and Dimensions: A comparative analysis. *Scientometrics*. 2021;**126**(6):5113–5142. Doi: 10.1007/s11192-021-03948-5
2. Visser M, van Eck NJ, Waltman L. Large-scale comparison of bibliographic data sources: Scopus, Web of Science, Dimensions, Crossref, and Microsoft Academic. *Quant Sci Stud*. 2021;**2**(1):20–41. DOI: 10.1162/qss_a_00112
3. Khurana, Parul, and Kiran Sharma. Impact of H-index on Authors Ranking: A Comparative Analysis of Scopus and WoS. *ArXiv*, 2021 Available from https://arxiv.org/abs/2102.06964. [Accessed on 24 July, 2025]
4. Kara BC, Şahin A, Dirsehan T. BibexPy: Harmonizing the bibliometric symphony of Scopus and Web of Science. *SoftwareX*. 2025;**30**:102098. Doi: 10.1016/j.softx.2025.102098
5. Hair K, Bahor Z, Macleod M, Liao J, Sena ES. The Automated Systematic Search Deduplicator (ASySD): a rapid, open-source, interoperable tool to remove duplicate citations in biomedical systematic reviews. *BMC Biol*. 2023;**21**(1):189. doi:10.1186/s12915-023-01686-z.
6. Nikolić D, Ivanović D, Ivanović L. An open-source tool for merging data from multiple citation databases. *Scientometrics*. 2024;**129**(7):4573-95. Doi:10.1007/s11192-024-05076-2
7. Chansanam W, Lai C. KKU-BiblioMerge: A novel tool for multi-database integration in bibliometric analysis. *Iberam J Sci Meas Commun*. 2025;**5**(1):4. Available from https://dialnet.unirioja.es/servlet/articulo?codigo=10103160 [Accessed on 20 Jul 2025]
8. Jiao C, Li K, Fang Z. How are exclusively data journals indexed in major scholarly databases? An examination of four databases. *Sci Data*. 2023;**10**(1):737. Doi:10.1038/s41597-023-02625-x.
9. Konu Kadirhanogullari M, Özay Köse E. Bibliometric Analysis: Technology Studies in Science Education. *Intl J Tech Educ and Sci*. 2023;**7(**2):167-91. doi:10.46328/ijtes.469
10. Echchakoui S. Why and how to merge Scopus and Web of Science during bibliometric analysis: the case of sales force literature from 1912 to 2019. *J Market Anal*. 2020;**8**(3):165-184. doi: 10.1016/j.softx.2025.102098
11. Caputo A, Kargina M. A user-friendly method to merge Scopus and Web of Science data during bibliometric analysis. *J Market Anal*. 2022;**10**(1):82-8. Doi: 10.1016/j.softx.2025.102098