◈**pro metrics**

KKU-BiblioMerge: A novel tool for multi-database integration in bibliometric analysis

Wirapong Chansanam[1], Chunqiu Li[2]

[1] Khon Kaen University, Thailand.
[2] Beijing Normal University, China.
   Corresponding author
   Email: lichunqiu@bnu.edu.cn. **ORCID:** https://orcid.org/0000-0003-1520-1297.

ABSTRACT

Objective. The objective of this study was to develop and validate KKU-BiblioMerge V.1.0, a bibliometric tool designed to address the limitations of single-source data in bibliometric analysis by integrating data from multiple databases, specifically Scopus and Web of Science (WoS).

Design/Methodology/Approach. The tool was developed using the R Shiny framework and incorporated key functions for data deduplication, field mapping, and integrity checks to ensure effective dataset merging. The performance of KKU-BiblioMerge was assessed by testing its ability to import, merge, and export bibliometric data, focusing on the efficiency and accuracy of consolidating records from Scopus and WoS.

Findings. The KKU-BiblioMerge application effectively processed and integrated 686 initial documents, eliminating 24.49% duplicate records to produce a final dataset of 518 unique entries. The tool demonstrated strong data consistency and high accuracy in field mapping, offering reliable cross-platform integration of bibliometric data compared to tools such as VOSviewer and Biblioshiny.

Originality/Value. KKU-BiblioMerge V.1.0 was a user-friendly, robust solution for multi-database bibliometric analysis. It enabled a more comprehensive and unbiased understanding of research landscapes. Its capability to integrate diverse datasets laid a foundation for advancing bibliometric software, broadening the scope and accuracy of analyses across scientific domains

Keywords: multi-database integration; bibliometric tool; scientometric analysis; data deduplication; cross-platform analysis; KKU-BiblioMerge.

1. INTRODUCTION

Bɪʙʟɪᴏᴍᴇᴛʀɪᴄ analysis is imperative in comprehending prevailing research trends, the impact of scholarly work, and the interconnections between disparate scientific domains. A multitude of studies have examined the evolution, capabilities, and applications of various bibliometric tools, particularly within the purview of scientific research and development. Moral-Muñoz *et al.* (2020) conducted a thorough review of the available software

tools utilized for bibliometric and scientometric analysis. Their study identifies key tools for data acquisition, performance analysis, and visualization, providing a structured guide on selecting suitable software based on user requirements. This review is especially valuable in understanding the existing landscape of bibliometric software and the criteria for choosing appropriate tools for specific types of analyses. Pessin *et al.* (2022) introduced an integrated method known as "smart bibliometrics," which combines science mapping and bibliometric analysis. This approach enhances the accuracy and depth of bibliometric studies by enabling a more detailed understanding of research trends and knowledge structures. By combining the two methodologies, the authors offer a framework that is particularly beneficial for uncovering hidden patterns and emerging trends within scientific disciplines.

Meanwhile, the application of scientometric analysis in the evaluation of research trends, the mapping of knowledge domains, and the assessment of scholarly impact across diverse fields has become imperative. The versatility of this technique is illustrated by recent studies, which demonstrate its depth in exploring specific research domains and guiding future inquiry. For instance, Fathi *et al.* (2024) conducted a global scientometric review of energy and building research, underscoring a growing emphasis on sustainable development. Their analysis mapped key trends and created a foundation for future scholarship. Similarly, Kastrin and Hristovski (2021) provided a comprehensive review of literature-based discovery (LBD), tracing four decades of evolution from co-occurrence to semantic-based methods, enhanced by network science. Additionally, Chen and Song (2019) proposed a systematic approach to scientometric reviews, emphasizing citation expansion to connect local and global research perspectives, thereby improving review accuracy. In specific sectors, scientometric methods offer practical insights. Ghaleb *et al.* (2022) explored project complexity in construction, identifying themes such as schedule performance and risk assessment, which are pivotal for managing complex projects. In the health research domain, Liu *et al.* (2020) used scientometric analysis to examine the link between pesticide exposure and

Parkinson's disease, visualizing trends that underscore the need for deeper investigation into environmental health risks. Furthermore, scientometric evaluations have also proven beneficial in academic administration. Győrffy *et al.* (2023) assessed the impact of scientometric rankings on grant application evaluations, while Munkácsy *et al.* (2022) studied PhD-level productivity, contributing valuable insights into career development and research policy. Moreover, scientometric analysis is shaping research on global issues. For instance, Sianes *et al.* (2023) underscored the influence of sustainable development goals (SDGs) on academic research, signifying a mounting commitment to global challenges. Additionally, Waqas *et al.* (2019) mapped research on online, unveiling evolving trends and emerging focus areas. These studies highlight the utility of scientometric analysis across various domains, advancing knowledge and shaping future research directions.

In a similar vein, Pereira *et al.* (2023) presented pyBibX, a Python library specifically designed for bibliometric and scientometric analysis. This tool stands out due to its integration of artificial intelligence (AI) capabilities, which allows for more sophisticated and automated analysis processes. The library offers a modular approach, supporting both seasoned bibliometric researchers and newcomers, as it is flexible and capable of handling a wide array of bibliometric tasks. Van Eck and Waltman (2017) discussed the use of two specific tools, CitNetExplorer (van Eck & Waltman, 2014) and VOSviewer, which are widely recognized for their clustering capabilities. By clustering scientific publications, these tools enable researchers to visualize and analyze the structural connections among different research topics, a capability that is critical for understanding knowledge domains and identifying the intellectual structure of scientific fields. This structural understanding can serve as a foundation for further research in bibliometric clustering techniques. Delgado López-Cózar *et al.* (2018) explored the role of Google Scholar as a bibliographic tool, emphasizing its status as a "big data" resource. The authors analyzed the features and document typologies covered by Google Scholar, revealing its advantages and limitations as a bibliometric database. This

study provides a critical evaluation of Google Scholar's potential and constraints, informing researchers about how this widely used tool can complement more specialized bibliometric databases. The field of bibliometric analysis has also greatly benefited from Bibliometrix, an R-based tool, developed by Aria and Cuccurullo (2017). This tool facilitates comprehensive science mapping and bibliometric analysis, offering an accessible platform for researchers with a variety of functionalities, including data import, analysis, and visualization. Aria and Cuccurullo's work is particularly noteworthy for its flexibility and adaptability, allowing researchers to conduct sophisticated analyses directly in the R environment. Bibliometrix has become widely adopted, demonstrating its utility in facilitating detailed bibliometric studies across various disciplines.

In the rapidly evolving domain of bibliometric analysis, a range of tools has emerged to support the evaluation and visualization of scientific literature. These tools, including ScientoPy, HistCite, Biblioshiny, CitNetExplorer, VOSviewer, CiteSpace, BibExcel, and BiblioMagika, offer diverse functionalities to enhance the analysis of scientific literature. ScientoPy, a Python-based tool, is particularly adept at customizable trend analysis, while HistCite maps citation networks over time, providing historical insights (Garfield *et al.,* 2006; Silva & Ramos, 2021). Biblioshiny, as an accessible interface for the Bibliometrix R-package, supports comprehensive analyses with interactive visuals for non-coders (Aria & Cuccurullo, 2017). CitNetExplorer and VOSviewer facilitate the visualization of citation networks, with VOSviewer being particularly distinguished for its bibliometric mapping capabilities (Van Eck & Waltman, 2010). CiteSpace, adept at detecting emerging trends, supports research on field evolution (Chen, 2006). Conversely, BibExcel and BiblioMagika prioritize data preparation, ensuring the generation of accurate and standardized datasets for analysis (Persson *et al.,* 2009; Sharma & Das, 2023). These tools empower researchers to conduct targeted bibliometric analyses, facilitating the tracking of scientific trends, the mapping of networks, and the exploration of disciplinary developments across diverse fields (Chen *et al.,* 2021). In sum, the reviewed studies cover a spectrum of bibliometric tools and methodologies, from traditional bibliometric software to advanced AI-enhanced libraries and widely used search engines. These tools collectively contribute to enhancing the efficiency and depth of bibliometric analysis, enabling researchers to explore scientific trends and the interconnections within and across fields.

Recent bibliometric research has increasingly highlighted the challenges of relying on a single database for comprehensive analysis, noting limitations and suggesting improvements for future studies. Romanelli *et al.* (2021) conducted a bibliometric analysis in the field of environmental sciences using the Web of Science (WoS) database. They identified four main challenges related to conducting bibliometric reviews, including limited data coverage and representation when relying on a single source. The authors recommended that future research integrate multiple databases to address these limitations, enhancing the inclusivity and robustness of the findings. In a similar vein, Guerrero-Bote *et al.* (2021) analyzed the comparative strengths of Dimensions and Scopus databases for university-level bibliometric analyses. They found that relying solely on one database, such as Scopus, could result in incomplete assessments of research output, particularly when analyzing broad academic landscapes like university-level research production. In light of these findings, the authors proposed that subsequent studies should employ data from multiple databases to ensure a more comprehensive understanding of research output and impact. Passas (2024) also addressed the limitations of single-database bibliometric analysis, which utilized Scopus for bibliometric insights. It is acknowledged that using a single database may introduce biases and data gaps that could skew the results of bibliometric analyses. Passas recommended a multi-database approach for future studies to reduce these limitations and provide a more balanced view of the research landscape. This recommendation aligns with the views of Romanelli *et al.* (2021) and Guerrero-Bote *et al.* (2021), collectively suggesting that leveraging multiple sources can address data inconsistencies and improve bibliometric research quality. Studies by Chansanam and Li (2022, 2023) offer valuable bibliometric

insights into poverty research, emphasizing a shift toward sustainability and policy-oriented themes aligned with SDGs. The 2022 analysis of Scopus data (1964-2022) emphasized an interdisciplinary focus, with rising keywords such as "sustainable development" and "inequality." In contrast, their 2023 study, which employed the WoS, mapped themes including "income inequality" and "social justice," underscoring a growing interest in applied, policy-relevant research for poverty alleviation. Despite the presence of commonalities in the examined topics and methodologies, the two studies were conducted independently, lacking a comprehensive integration of data from both databases. This discrepancy highlights the inherent limitations of single-database bibliometric analyses and underscores the value of incorporating multiple databases. This approach not only addresses the deficiencies in data coverage but also mitigates potential biases, thereby facilitating more accurate and reliable bibliometric insights across diverse research domains.

Bibliometric databases, including WoS, Scopus, and OpenAlex, are tools that facilitate research trend analysis and impact assessment. While WoS offers archival data, it is limited in scope and cost-prohibitive for many users. Scopus provides extensive journal coverage but requires a subscription. OpenAlex, a free, inclusive database, exceeds others in scope but lacks advanced analytics and visualization capabilities. Bibliometric analysis necessitates advanced bibliometric merging tools to effectively integrate data from multiple sources, enhancing coverage, inclusivity, and accuracy through deduplication, integrity checks, and precise field mapping. These tools empower researchers to derive cross-disciplinary insights and comprehensively analyze global research trends. However, the effectiveness of these tools hinges on sophisticated integration algorithms, which can introduce complexity. Compatibility challenges with proprietary database formats and the absence of built-in visualization capabilities often limit the utility of these tools, necessitating supplementary tools for a complete analysis. This study focuses on developing and validating a framework for combining at least two databases (Scopus and WoS) to address the limitations of single-source data

in bibliometric analysis. The selection of these databases was deliberate, aiming to showcase the viability and efficacy of the proposed KKU-BiblioMerge framework in seamlessly integrating data from widely recognized sources with varied formats and structures. While OpenAlex is extensive and comprehensive, it was not the primary focus of this research, which was to establish a proof of concept for the data integration process. The integration of Scopus and WoS data serves as a substantial test case, demonstrating the framework's aptitude for handling heterogeneous data sources. Once the framework has been validated, future work will explore its scalability and adaptability to incorporate OpenAlex and other large-scale databases, thereby expanding its utility further.

A critical limitation of bibliometric tools is their inability to integrate multiple databases for simultaneous analysis. Although bibliometric analysis is increasingly recognized as a critical tool for understanding research trends, impact, and interconnections across scientific domains, most tools are constrained to data from a single source. As indicated by numerous studies, this limitation can introduce biases, restrict data coverage, and impede a comprehensive understanding of the research landscape (Guerrero-Bote *et al.,* 2021; Passas, 2024; Romanelli *et al.,* 2021). To enhance the depth and reliability of bibliometric insights, researchers have increasingly called for methods that enable the integration of multiple databases, thus addressing data inconsistencies and enabling a broader understanding of academic outputs. This study builds on this need by proposing a framework enabling multi-database integration, which is crucial for advancing bibliometric analysis capabilities and addressing existing tool limitations in future research. The objective of this study is to develop a robust bibliometric tool, KKU-BiblioMerge V.1.0, to integrate data from two primary and globally renowned research databases. This tool will address the limitations of single-source bibliometric analyses by incorporating advanced features for data deduplication, field mapping, and integrity checks. By offering an efficient solution for multi-database integration, this tool seeks to enhance the capabilities and accuracy of bibliometric analysis.

2. METHODOLOGY

The Bibliometrix R-package provides a comprehensive toolkit for bibliometric and scientometric analysis, which is ideal for quantitative research. Developed in R, this package leverages robust statistical algorithms, advanced numerical routines, and integrated data visualization capabilities, making R a preferred language for scientific computation. The bibliometrix workflow facilitates science mapping across stages, including data collection, analysis, and visualization (Cobo *et al.,* 2018). Initially, it enables data collection by converting bibliographic data into an R data frame. The analysis phase includes descriptive analyses of bibliographic data, network creation for various connections such as bibliographic coupling and co-citation, and normalization. Finally, data visualization is achieved through conceptual structure and network mapping, providing an insightful representation of the bibliometric data.

As illustrated in Figure 1, the development of KKU-BiblioMerge V.1.0, a Shiny-based application designed for integrating bibliometric data, has been underway. The tool, once developed, will facilitate the merging and analysis of bibliometric information from multiple sources, enhancing research efficiency and data management.
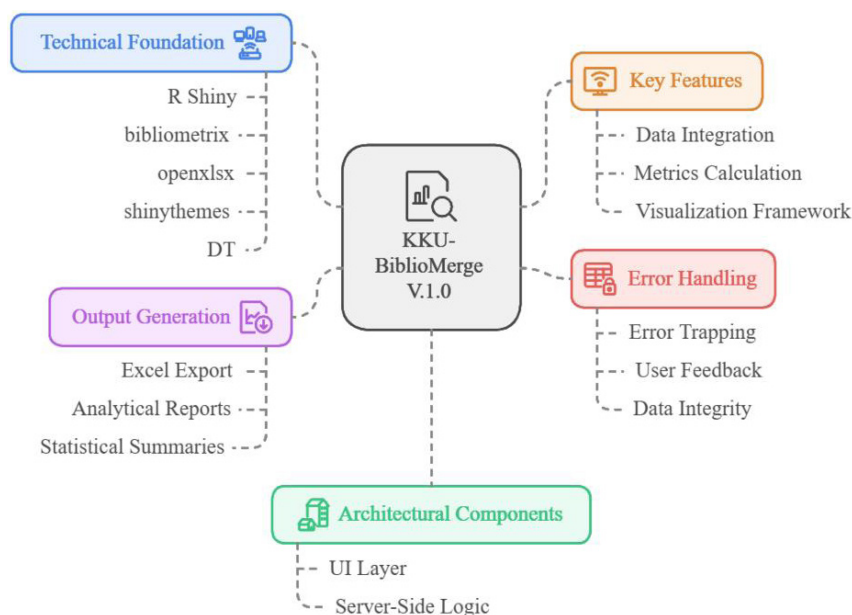

Figure 1. Application architecture and development framework.

The details on KKU-BiblioMerge V.1.0 working steps:

1. Data input stage:
   - Users can upload bibliographic data from Scopus (".bib" format) and WoS (".txt" format) through an intuitive interface.
   - Input files are validated to ensure compatibility and proper formatting before processing begins.
2. Data preprocessing:
   - Includes file format conversion to standardize data structures.
   - Implements data normalization techniques, such as harmonizing author names, standardizing journal titles, and ensuring consistent citation formats.
3. Duplicate detection and removal:
   - Advanced algorithms identify and eliminate duplicate entries using metadata fields such as title, DOI, and publication year.
   - Deduplication ensures the final dataset retains unique records without redundancies.
4. Data integration:
   - Merges records from multiple databases while preserving bibliometric integrity.
   - Ensures consistent field mapping (e.g., aligning author affiliations and keywords across databases).

5. Metrics calculation:
   - Calculates key bibliometric metrics, including citation counts, co-authorship networks, document type distributions, and publication trends.
   - Offers insights into the scholarly impact and collaboration patterns.
6. Data visualization and export:
   - Provides interactive dashboards to visualize bibliometric trends, including temporal publication patterns and network maps.
   - Export functionality lets users download the integrated dataset for further analysis using tools such as Biblioshiny.
7. Error handling and feedback:
   - Real-time feedback during data processing highlights errors or inconsistencies, ensuring user awareness and data accuracy.

KKU-BiblioMerge is a robust solution for bibliometric data integration. It combines data from multiple significant databases, thereby overcoming the limitations of single-source analyses. Its advanced deduplication and field mapping mechanisms ensure the merged dataset is accurate, consistent, and complete. The system is highly efficient, enabling the quick processing of large datasets while minimizing manual effort. By consolidating diverse sources, it provides a comprehensive view of research landscapes, uncovering trends that might remain hidden with individual databases. The Shiny-based interface is characterized by its user-friendliness and accessibility to researchers with varying technical expertise, offering clear instructions, interactive visualizations, and intuitive controls. Additionally, the system is scalable and flexible, designed to integrate new databases in the future, and it includes robust validation mechanisms to prevent processing errors and maintain data quality. The analyze_overlap() function employs a robust method for identifying and resolving duplicate bibliometric records by creating unique identifiers for each entry. The process entails title normalization (removal of non-alphanumeric characters and conversion to lowercase), matching of publication years, and normalization of author strings (removal of special characters and conversion to lowercase). Deduplication is executed in two stages: initially, overlapping records are identified using the intersect() function on unique identifiers; subsequently, merging and deduplication are executed via the mergeDbSources() function of the bibliometrix package. The system meticulously tracks key statistics, including the total number of records per database, the number of overlapping entries, and the unique records for each source. The results are saved in an MS Excel format to ensure reliability. This method handles diverse text formats, accounts for multiple fields to minimize false matches, and utilizes standardized identifiers for consistent record integration. This application exemplifies a sophisticated implementation of bibliometric data integration and analysis, suitable for academic research and bibliometric studies. It effectively manages the complexity of merging disparate bibliometric data sources while providing comprehensive analytical capabilities. The development adheres to contemporary software engineering principles, emphasizing modularity, error handling, and user experience. The application's architecture permits future expansions and modifications while maintaining robust functionality for current bibliometric analysis needs.

3. RESULTS

In this query example, data were extracted from the Scopus and WoS databases using specific selection criteria. The study focused on (1) the broad keyword "public financial management" as the main topic, (2) articles in English, (3) inclusion of all subject categories, and (4) a publication range spanning from 1977 to 2024. Following the acquisition of the files from both Scopus and WoS, the next step involved the utilization of KKU-BiblioMerge V.1.0 (https://wirapongchansanam.shinyapps.io/KKU-BiblioMerge-V-1-0/) to merge these bibliography datasets. In accordance with the detailed processing information as illustrated in Figure 2, step-by-step instructions are displayed on the page to ensure accurate integration of the dataset files. This methodological approach enables users to effectively combine and organize bibliometric data for subsequent analysis.
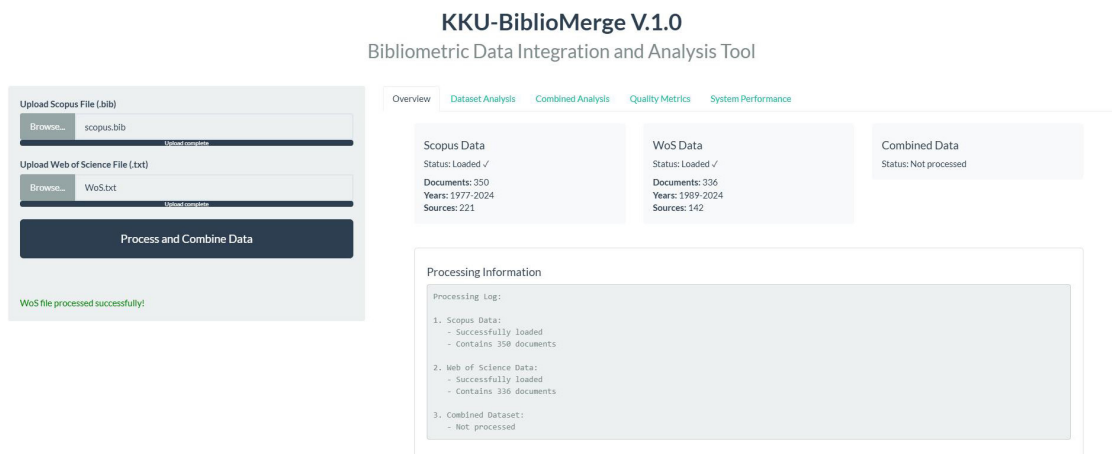
KKU-BiblioMerge V.1.0
Bibliometric Data Integration and Analysis Tool

Figure 2. Main interface and data processing dashboard
of KKU-BiblioMerge V.1.0 for bibliometric data integration.

The development of KKU-BiblioMerge V.1.0 has resulted in a sophisticated bibliometric data integration and analysis tool featuring a comprehensive interface with five analytical modules: overview, dataset analysis, combined analysis, quality metrics, and system performance. As illustrated in Figure 2, the application effectively processes and integrates bibliometric data from two major databases: Scopus (comprising 350 documents from 1977 to 2024 across 221 sources) and WoS (encompassing 336 documents from 1989 to 2024 from 142 sources). The system's "Processing Information" log indicates successful data loading and validation stages for both databases, with clear status indicators reflecting the current state of data processing. The user interface facilitates seamless navigation and real-time processing feedback, exemplified by the "Loaded ✓" status for individual databases while maintaining a "Not processed" status for the combined dataset until user initiation. This implementation exemplifies the sophisticated capacity of KKU-BiblioMerge V.1.0 to manage disparate bibliometric data sources while upholding data integrity and providing lucid processing information, thereby accomplishing its objective of delivering a user-friendly platform for comprehensive bibliometric analysis.

As illustrated in Figure 3, the "Dataset Analysis" interface of KKU-BiblioMerge V.1.0 reveals comprehensive bibliometric patterns across both the Scopus and WoS databases. The Scopus dataset encompasses 350 documents (1977-2024) from 221 unique sources, with 768 distinct authors, and an average citation rate of 7.42 per document across 11 document types. Concurrently, the WoS dataset comprises 336 documents (1989-2024) from 142 unique sources, featuring 681 unique authors and averaging 4.56 citations per document across 13 document types. The temporal publication trends, depicted through dual time-series graphs, manifest noteworthy growth patterns: Scopus publications show a marked increase from 2010 onwards, peaking at approximately 40 publications per year in recent years, while WoS publications exhibit a similar upward trend from 2005, reaching peaks of around 35 publications annually. This comparative analysis illustrates the complementary nature of both databases and validates the system's capability to process and analyze distinct bibliometric data sources while maintaining their unique characteristics and temporal distributions.

The metrics generated by KKU-BiblioMerge V.1.0, including citation rates, publication trends, and document type distributions, provide a multifaceted understanding of bibliometric landscapes. By integrating data from Scopus and WoS, the tool enables comparative and holistic analyses that are unattainable with single-database approaches. For instance, combining datasets revealed broader temporal trends and disciplinary coverage, validating the system's ability to address data fragmentation. This comprehensive integration facilitates a more inclusive exploration of scholarly outputs, uncovering collaborative networks, thematic patterns, and emerging research areas. These
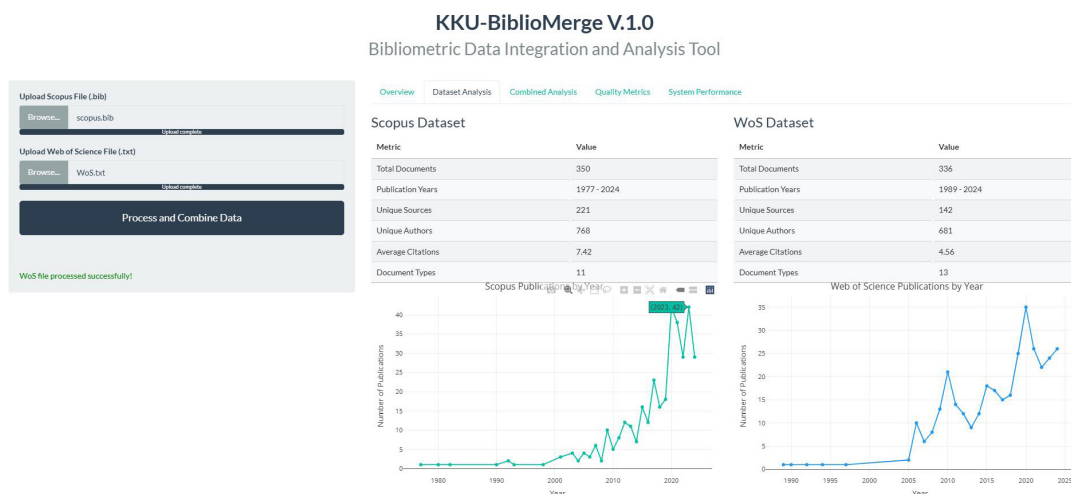
Figure 3. Dataset analysis dashboard showing comparative
bibliometric patterns in KKU-BiblioMerge V.1.0.

insights significantly enhance bibliometric understanding, offering researchers a richer and more accurate view of research dynamics.

As illustrated in Figure 4, the "Combined Analysis" of dataset interface of KKU-BiblioMerge V.1.0 exemplifies the successful integration and deduplication of bibliometric data from the Scopus and WoS databases. The system efficiently processed 686 initial documents, yielding 518 unique records after the removal of 168 duplicates, thereby exhibiting a 24.49% duplication rate. The integrated dataset spans from 1977 to 2024, encompassing 273 unique sources and 986 unique authors, with an average citation impact of 4.72 per document. The publication trend visualization reveals a significant growth pattern, particularly accelerating from 2010 onwards, with peak productivity reaching 50 publications annually in recent years. Document type distribution analysis demonstrates a diverse scholarly output, with research articles (235) and dissertations/theses (102) dominating the corpus, followed by book chapters (65) and conference papers (9). The system's tabular presentation of the first 10 records exemplifies the standardized integration of metadata fields, including authors, titles, sources, publication years, document types, and citation counts. This demonstration illustrates the tool's capability to maintain bibliographic integrity while harmonizing data from different sources.
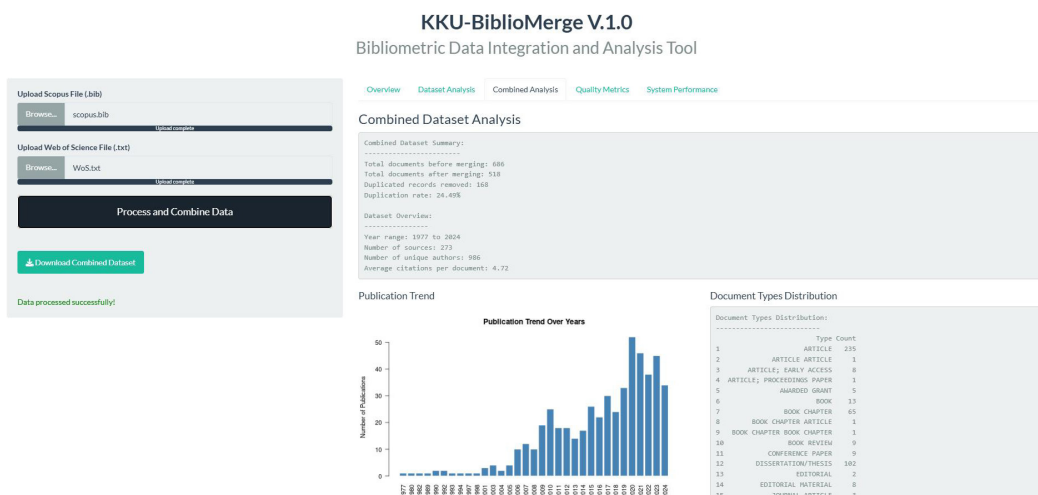

Figure 4. Bibliometric data integration outcomes
and analytical visualization in KKU-BiblioMerge V.1.0.

As illustrated in Figure 5, the KKU-BiblioMerge V.1.0 software facilitates the export of data. Following the successful processing of data, which is indicated by the "Data processed successfully!" message, users can download the combined bibliometric dataset by selecting the "Download Combined Dataset" button. The system then generates an MS Excel file (".xlsx" format) with a standardized naming convention, "combined-bibliometric-data-," followed by the current date. This export feature enables users to preserve their merged and deduplicated bibliometric data for subsequent analysis in MS Excel. The red arrows in the image delineate the workflow from successful processing to file download.
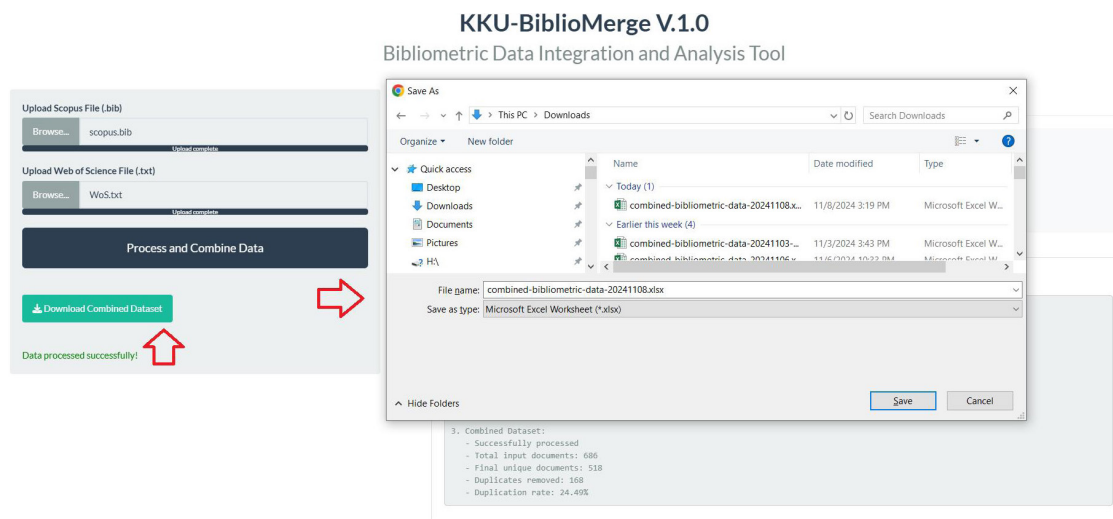

Figure 5. The data export functionality of KKU-BiblioMerge V.1.0.

Once exported to an MS Excel file format, the data file can be seamlessly imported into the Biblioshiny application, a web-based interface of the Bibliometrix R-package, to conduct advanced bibliometric analysis (Aria & Cuccurullo, 2017) as shown in Figure 6. This process enables researchers to leverage Biblioshiny's comprehensive suite of tools, allowing for an in-depth exploration and visualization of scholarly data. The utilization of MS Excel for data importation enables the seamless integration of Biblioshiny's sophisticated functionalities, empowering researchers to undertake comprehensive analyses of citation patterns, co-authorship networks, thematic trends, and other pivotal bibliometric indicators. This approach facilitates a systematic examination of research landscapes, yielding critical insights into publication trends, influential authors (Hou *et al.*, 2020), prominent journals, and emerging topics. Furthermore, the Excel-to-Biblioshiny workflow streamlines data management, ensuring accessibility for researchers with varying technical proficiencies. Consequently, this approach enhances the accuracy and depth of bibliometric investigations, thereby supporting evidence-based decision-making in research planning, academic collaboration, and knowledge discovery. In summary, exporting bibliographic data to MS Excel for subsequent import into Biblioshiny is an effective method for conducting rigorous bibliometric analysis, which ultimately contributes to the advancement of academic research and scientific knowledge. As illustrated in Figure 6, the "Main Information" dashboard of the bibliometrix web interface presents a comprehensive array of bibliometric statistics for the merged dataset. The analysis encompasses a period from 1977 to 2024, encompassing 518 documents from 273 sources, exhibiting an annual growth rate of 7.79%. The collaboration metrics reveal a total of 986 authors, including 227 single-authored documents, with an international co-authorship rate of 0.1931% and an average of 2.18 co-authors per document. Additional metrics demonstrate 441 author keywords, an average document age of 7.78 years, and a mean citation rate of 4.724 citations per document. This dashboard offers a concise overview of the dataset's primary bibliometric indicators.
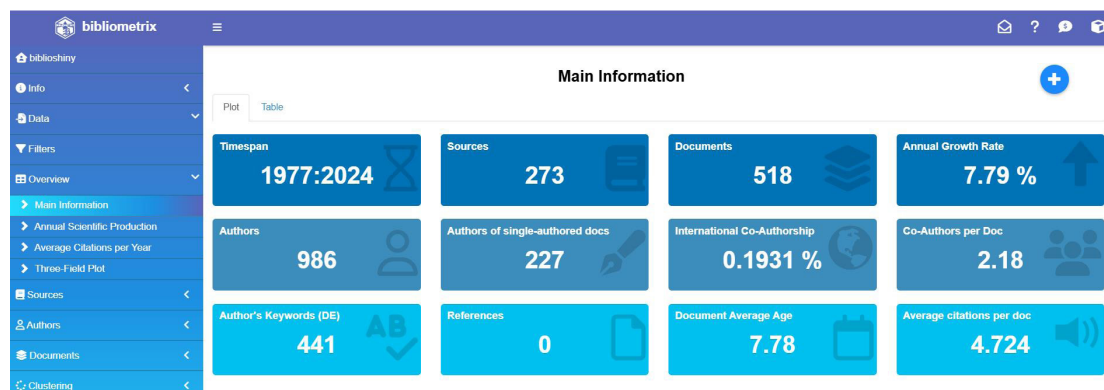
Figure 6. The main interface of Biblioshiny-R application.

As illustrated in Figure 7, the KKU-Biblio-Merge V.1.0 workflow is methodically structured in three phases: data collection, data analysis, and data visualization. The process initiates with "data query and loading," which supplies the KKU-BiblioMerge V.1.0 system with the necessary foundation for "descriptive bibliometric analysis." The "data analysis" phase employs a dual-track approach: the first track involves "document × attribute matrix creation" with normalization, leading to "data reduction" through PCA, MDS, MCA, and Clustering techniques; the second track focuses on "network matrix creation." Both tracks converge in the "mapping" phase, producing various visualization outputs, including factorial maps, dendrograms, semantic maps, network maps, and historiography. The system's capabilities are further demonstrated by its ability to support advanced bibliometric analyses, including bibliographic coupling, co-citation analysis, collaboration networks, co-occurrence analysis, and historiographic mapping.
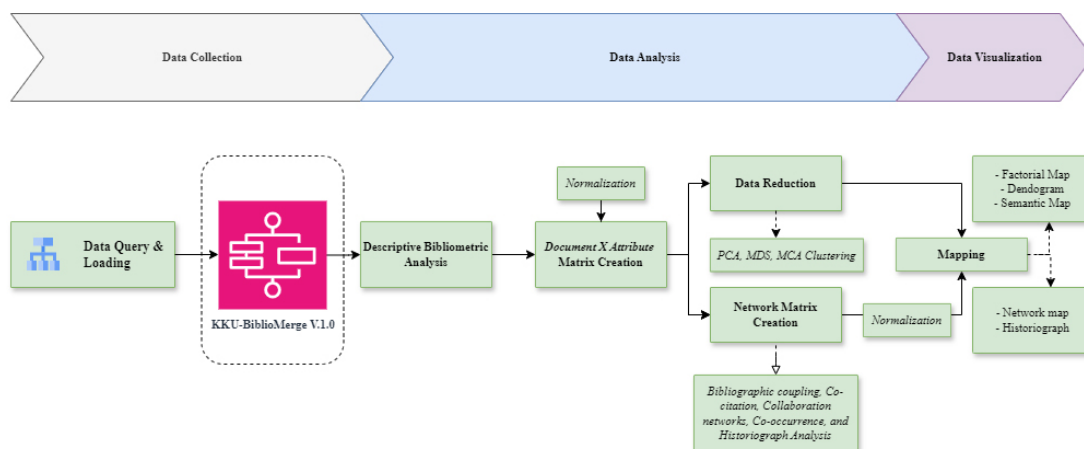

Figure 7. The KKU-BiblioMerge V.1.0 comprehensive integration with the bibliometrix package workflow architecture. Source. Based on Aria and Cuccurullo (2017).

4. EVALUATION FOR KKU-BIBLIOMERGE V.1.0

The evaluation of KKU-BiblioMerge V.1.0 is centered on two fundamental dimensions: data integration quality and system performance efficiency. This targeted evaluation approach is poised to yield the most valuable insights for assessing and enhancing the application's effectiveness.

4.1. Data integration quality dimension

The tool's capacity to precisely integrate bibliographic data from Scopus and WoS is of the utmost importance. Key metrics encompass the accuracy of duplicate detection, verified against the anticipated duplication rate of 24.49%, and the thoroughness of merged

records, ensuring the retention of no critical information during the integration process. Furthermore, the precision of field mapping —comprising author names, titles, sources, and publication years— is imperative for upholding data integrity, in conjunction with ensuring uniformity in the final dataset. Validation methods entail manual inspections, comparisons with original data sources, and cross-verification of merged fields.

As illustrated in Figure 8, the "Quality Metrics" (data quality assessment) interface of KKU-BiblioMerge V.1.0 demonstrates robust data validation across three critical dimensions. The "completeness metrics" reveal perfect (100%) completeness for most bibliometric fields, including authors (AU), titles (TI),

publication years (PY), document types (DT), and citation counts (TC), with only source (SO) information showing a slightly lower completeness rate of 79.15%. The "consistency metrics" indicate full validity (100%) for both year formats and author name standardization, though three title duplicates were identified in the dataset. The "mapping accuracy" assessment confirms perfect (100%) alignment across all key fields (titles, sources, and years) between the Scopus and WoS databases, validating the system's robust data integration capabilities. These metrics underscore the tool's efficacy in upholding stringent data quality standards while integrating bibliometric data from disparate sources, exhibiting particular strengths in field completeness and mapping accuracy.
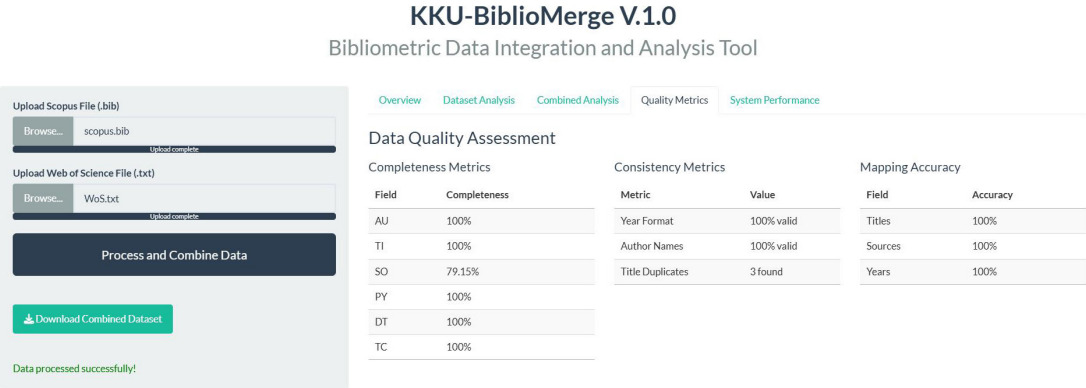

Figure 8. Data quality assessment dashboard of KKU-BiblioMerge V.1.0: analysis of bibliometric field completeness, consistency, and mapping accuracy.

4.2. System performance efficiency dimension

The effectiveness of this measure is gauged by several operational metrics, with particular emphasis on the tool's responsiveness and speed. These metrics encompass the time required to upload and process files, merge data, and generate analyses. Additionally, resource utilization is a focal point, with CPU and memory usage monitoring to ensure efficient performance across a range of file sizes. The duration of each operation is meticulously tracked to identify areas that could be improved, and performance testing is conducted across diverse dataset sizes to assess scalability and reliability under varied loads. This comprehensive evaluation framework ensures the accuracy, operational speed, and resource efficiency of KKU-BiblioMerge

V.1.0, thereby establishing a foundation for a robust and reliable bibliographic analysis tool tailored to research needs.

As illustrated in Figure 9, the "System Performance" metrics interface of KKU-BiblioMerge V.1.0 demonstrates efficient processing capabilities across three key operational phases. In the "file processing performance" phase, the processing speed of Scopus data (0.84 MB) resulted in 262.49 records/second, completing 350 records in 1.33 seconds, while the processing speed of WoS data (0.49 MB) achieved higher efficiency at 555.91 records/second, processing 336 records in 0.60 seconds. The "data merging performance" phase demonstrated rapid integration, with a processing time of 0.06 seconds and the generation of 518 merged records. The "analysis performance" phase exhibited optimal efficiency,

analyzing all 518 records in 0.00 seconds. These metrics illustrate the system's high-performance capabilities in handling bibliometric data integration tasks, with impressive processing speeds for database formats and efficient merging operations.
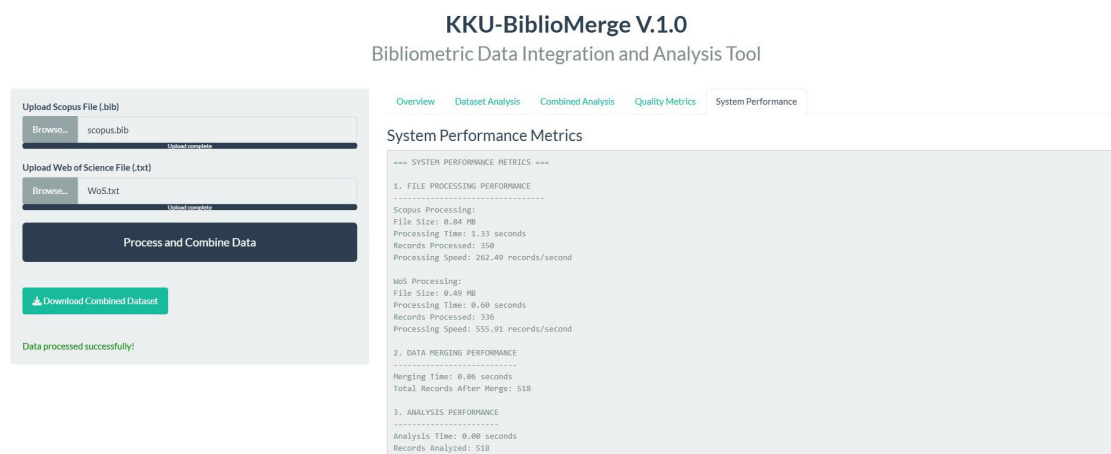
KKU-BiblioMerge V.1.0
Bibliometric Data Integration and Analysis Tool

Figure 9. System performance analysis dashboard of KKU-BiblioMerge V.1.0: processing, merging, and analysis metrics.

The KKU-BiblioMerge V.1.0 successfully processed 686 documents, eliminating 24.49% of duplicate records to produce a consolidated dataset of 518 unique entries, ensuring data integrity and comprehensiveness. The KKU-BiblioMerge demonstrated high accuracy in field mapping, offering a reliable cross-platform integration of bibliometric data compared to tools like VOSviewer and Biblioshiny. The KKU-BiblioMerge V.1.0 offers a variety of visual outputs to facilitate the interpretation of bibliometric data integration and analysis. Beyond basic frequency calculations, the tool incorporates visualization panels that include temporal publication trends, illustrating growth patterns in research output over time, and document-type distributions, highlighting various scholarly contributions such as research articles, dissertations, and book chapters. Additionally, it showcases author collaboration networks and citation impact visualizations, offering insights into the interconnectedness and scholarly influence within the dataset. The integration outcomes, such as deduplication rates and consolidated datasets, are presented using tabular formats and graphical trends to ensure transparency and accuracy in data handling. Furthermore, the combined dataset's bibliometric patterns are displayed via comparative temporal graphs, revealing complementary trends across databases. These visual capabilities support descriptive analysis and provide a foundation for advanced exploration using tools such as Biblioshiny, leveraging the exported datasets for deeper bibliometric insights.

5. DISCUSSION

This study presents the development and implementation of the KKU-BiblioMerge V.1.0 framework, an innovative tool and concept designed to enhance bibliometric analysis through multi-database integration. KKU-BiblioMerge V.1.0 is conceived to facilitate the merging of bibliometric data from multiple sources, particularly Scopus and WoS, thereby addressing a significant gap in current bibliometric practices. This necessity arises from the recognition that single-source bibliometric analyses are subject to limitations, including the potential for incomplete or skewed insights, particularly in interdisciplinary or emerging fields (Guerrero-Bote *et al.,* 2021; Passas, 2024; Romanelli *et al.,* 2021). KKU-BiblioMerge addresses these limitations by consolidating data from multiple databases, ensuring a more comprehensive representation of scholarly outputs and trends. The necessity for multi-database integration has been recognized in bibliometric studies, with researchers underscoring its role in achieving accurate, representative, and robust findings. As Romanelli *et al.* (2021) have observed,

analyses confined to a single database may suffer from a lack of inclusivity, potentially leading to the exclusion of significant publications, particularly when databases exhibit varying levels of field-specific coverage. Similarly, Guerrero-Bote *et al.* (2021) have underscored that studies based on a single source, such as those confined to Scopus, may not adequately represent the breadth of research output, potentially leading to the oversight of key studies from other fields or regions. The integration capabilities of KKU-BiblioMerge address these limitations by allowing data merging across major bibliometric databases, thus supporting a more holistic understanding of research landscapes.

The KKU-BiblioMerge V.1.0 multi-database approach aligns with ongoing efforts to enhance bibliometric accuracy and reliability, offering researchers a tool that enables more inclusive and cross-disciplinary insights. This development marks a significant advancement in bibliometric methods, providing a solution to the fragmented nature of traditional single-database analyses and responding directly to the challenges identified in the literature. In comparison to existing bibliometric tools such as VOSviewer and CitNetExplorer, which predominantly excel in mapping and clustering scientific data but lack capabilities for integrating data across multiple sources (Van Eck & Waltman, 2017), KKU-BiblioMerge V.1.0 offers distinct value by emphasizing comprehensive data consolidation. The KKU-BiblioMerge V.1.0 tool is notable for its ability to facilitate high-accuracy data integration, incorporating features such as deduplication and rigorous data integrity checks. These features are essential for reliable bibliometric analysis. In contrast to the capabilities of VOSviewer and CitNetExplorer, which are powerful tools for visualizing citation networks and knowledge domains, KKU-BiblioMerge V.1.0 offers distinct advantages due to its ability to integrate data from multiple sources. The KKU-BiblioMerge stands out due to its capacity to seamlessly integrate bibliometric data from multiple major databases, including Scopus and WoS, while preserving data integrity through advanced deduplication and field mapping processes. Its user-friendly R Shiny framework ensures efficient processing, robust analysis, and adaptability for future expansions, distinguishing it from traditional tools.

Figures 1-9 collectively illustrate the potential and uniqueness of KKU-BiblioMerge V.1.0 by showcasing its robust design, functionality, and integration capabilities. They highlight the tool's architecture and development framework (Figure 1), user interface and data processing features (Figure 2), and its ability to analyze and compare bibliometric patterns between databases (Figure 3). The effectiveness of the tool in data integration, deduplication, and consolidation is demonstrated by bibliometric outcomes (Figure 4), while its export functionality underscores its interoperability (Figure 5). The systematic workflow architecture (Figure 7) also emphasizes the comprehensive approach to data collection, analysis, and visualization. The evaluation of data quality (Figure 8) and system performance (Figure 9) further reinforce its accuracy, reliability, and operational efficiency, positioning KKU-BiblioMerge V.1.0 as a significant advancement in bibliometric analysis.

Furthermore, while Biblioshiny offers a user-friendly analysis interface as part of the Bibliometrix R-package, it is similarly limited to single datasets (Aria & Cuccurullo, 2017). However, KKU-BiblioMerge expands beyond this constraint by enabling a unified data input from multiple databases, allowing researchers to analyze bibliometric data from a broader, cross-platform perspective. This capability is particularly advantageous for studies that require a comprehensive view of scholarly outputs across disciplines and sources, thus enhancing the depth and breadth of bibliometric analysis. KKU-BiblioMerge meets a critical need for more integrated, inclusive bibliometric tools to provide richer insights into research trends, influential publications, and collaborative networks across scientific domains by enabling a multi-database approach. The results of this study also highlight significant processing efficiency and data quality in KKU-BiblioMerge. This capability is analogous to recent advancements in bibliometric tool development, such as the integration of AI in pyBibX for sophisticated data management (Pereira *et al.,* 2023). Consequently, the current tool signifies an advancement by addressing the demand for multi-source data integration, thereby offering a more comprehensive bibliometric framework than previously available.

6. CONCLUSION

This study hypothesizes that integrating multiple bibliometric databases using KKU-BiblioMerge V.1.0 will address the limitations of single-source analyses, enhancing the comprehensiveness, accuracy, and inclusivity of bibliometric research. The findings strongly support this hypothesis, demonstrating the system's ability to integrate data from Scopus and WoS with high accuracy, robust deduplication, and reliable field mapping, thereby mitigating biases and data gaps inherent in single-database analyses. KKU-BiblioMerge facilitates researchers in conducting broader, cross-disciplinary studies with enhanced data coverage by offering a scalable and user-friendly platform. While the tool signifies a substantial advancement in bibliometric analysis, there are potential areas for enhancement, including the incorporation of visualization features, the expansion of integration with other databases such as OpenAlex, and the mitigation of limitations such as the absence of real-time updates and the reliance on external visualization tools. The KKU-BiblioMerge V.1.0 release contributes substantially to the advancement of bibliometric methodologies and establishes the foundation for further innovations in multi-database integration.

Conflict of interest

The authors declare that there is no conflict of interest.

Contribution statement

Wirapong Chansanam: Conceptualization, literature search, writing and editing the manuscript, data curation, formal analysis, software programming and development, final approval, and submission.

Chunqiu Li: Conceptualization, literature search, data curation, formal analysis, writing and editing the manuscript, and final approval.

Wirapong Chansanam and Chunqiu Li: Collaboratively contributed to conceptualization, literature search, data curation, formal analysis, writing, and editing the manuscript, with both authors providing final approval.

Data availability statement

The data generated during the research process have been included in the article. ●

REFERENCES

Aria, M., & Cuccurullo, C. (2017). Bibliometrix: An R-tool for comprehensive science mapping analysis. *Journal of Informetrics, 11*(4), 959-975. https://doi.org/10.1016/j.joi.2017.08.007

Chansanam, W., & Li, C. (2022). Scientometrics of poverty research for sustainability development: Trend analysis of the 1964-2022 data through Scopus. *Sustainability, 14*(9), 5339. https://doi.org/10.3390/su14095339

Chansanam, W., & Li, C. (2023). Knowledge structure and trends in poverty research on the Web of Science database: A bibliometric analysis. *Journal of Mekong Societies, 19*(2), 120-152.

Chen, C. (2006). CiteSpace II: Detecting and visualizing emerging trends and transient patterns in scientific literature. Journal of the American Society for Information *Science and Technology, 57*(3), 359-377. https://doi.org/10.1002/asi.20317

Chen, C., Ibekwe-SanJuan, F., & Hou, J. (2021). The structure and dynamics of cocitation clusters: A multiple-perspective cocitation analysis. *Journal of the American Society for Information Science and Technology, 62*(7), 1386-1402. https://doi.org/10.1002/asi.21557

Chen, C., & Song, M. (2019). Visualizing a field of research: A methodology of systematic scientometric reviews. *PLoS One, 14*(10), Article e0223994. https://doi.org/10.1371/journal.pone.0223994

Cobo, M. J., López-Herrera, A. G., Herrera-Viedma, E., & Herrera, F. (2018). SciMAT: A new science mapping analysis software tool. *Journal of the American Society for Information Science and Technology, 62*(8), 1609-1630. https://doi.org/10.1002/asi.21525

Delgado López-Cózar, E., Orduna-Malea, E., Martin-Martin, A., Ayllon, J. M. (2018). Google Scholar: The "big data" bibliographic tool. *ArXiv Preprint,* arXiv:1806.06351. https://arxiv.org/abs/1806.06351

Fathi, S. J., Habibi, A., & Vakilinezhad, R. (2024). Scientometric literature review and visualization of global research on energy and building. *Environment, Development and Sustainability,* 1-47. https://doi.org/10.1007/s10668-024-04912-y

Garfield, E., Pudovkin, A. I., & Istomin, V. S. (2006). HistCite: A software tool for informetric analysis of citation linkage. *Information and Computation, 42*(3), 334-339. https://doi.org/10.1016/j.ic.2005.10.011

Ghaleb, H., Alhajlah, H. H., Bin Abdullah, A. A., Kassem, M. A., & Al-Sharafi, M. A. (2022). A scientometric analysis and systematic literature review for construction project complexity. *Buildings, 12*(4), 482. https://doi.org/10.3390/buildings12040482

Guerrero-Bote, V. P., Chinchilla-Rodríguez, Z., Mendoza, A., & de Moya-Anegón, F. (2021). Comparative analysis of the bibliographic data sources Dimensions and Scopus: An approach at the level of universities. *Frontiers in Research Metrics and Analytics, 5,* Article 593494. https://doi.org/10.3389/frma.2020.593494

Győrffy, B., Weltz, B., & Szabó, I. (2023). Supporting grant reviewers through the scientometric ranking of applicants. *PLoS One, 18*(1), Article e0280480. https://doi.org/10.1371/journal.pone.0280480

Hou, J., Yang, X., & Chen, C. (2020). Measuring researchers' potential scholarly impact with structural variations: Four types of researchers in information science (1979-2018). *PLoS One, 15*(6), Article e0234347. https://doi.org/10.1371/journal.pone.0234347

Kastrin, A., & Hristovski, D. (2021). Scientometric analysis and knowledge mapping of literature-based discovery (1986-2020). *Scientometrics, 126*(2), 1415-1451. https://doi.org/10.1007/s11192-020-03811-z

Liu, C., Liu, Z., Zhang, Z., Li, Y., Fang, R., Li, F., & Zhang, J. (2020). A scientometric analysis and visualization of research on Parkinson's disease associated with pesticide exposure. *Frontiers in Public Health, 8,* 91. https://doi.org/10.3389/fpubh.2020.00091

Moral-Muñoz, J. A., Herrera-Viedma, E., Santisteban-Espejo, A., & Cobo, M. J. (2020). Software tools for conducting bibliometric analysis in science: An up-to-date review. *Science Mapping and Knowledge Discovery Journal. 29*(1). https://doi.org/10.3145/epi.2020.ene.03

Munkácsy, G., Herman, P., & Győrffy, B. (2022). Comparison of scientometric achievements at PhD and scientific output ten years later for 4,790 academic researchers. *PLoS One, 17*(7), Article e0271218. https://doi.org/10.1371/journal.pone.0271218

Passas, N. (2024). Bibliometric analysis: The main steps. *Research Metrics Journal, 4*(2), 65. https://doi.org/10.3390/rmj4020065

Pereira, V., Basilio, M. P., Tarjano Santos, C. H. (2023). pyBibX — A Python library for bibliometric and scientometric analysis powered with artificial intelligence tools. *ArXiv Preprint,* arXiv:2304.14516. https://arxiv.org/abs/2304.14516

Persson, O., Danell, R., & Schneider, J. W. (2009). How to use BibExcel for various types of bibliometric analysis. In *Proceedings of the International Society for Scientometrics and Informetrics Conference,* Rio de Janeiro, Brazil, 9-13.

Pessin, V. Z., Yamane, L. H., & Siman, R. R. (2022). Smart bibliometrics: An integrated method of science mapping and bibliometric analysis. *Scientometrics.* https://doi.org/10.1007/s11192-022-04406-6

Romanelli, M., Gonçalves, M. C. P., de Abreu Pestana, L. F., Hitaka Soares, J. A., Boschi, R. S., & Andrade, D. F.. (2021). Four challenges when conducting bibliometric reviews and how to deal with them. *Environmental Science and Pollution Research, 28*(33), 45692-45706. https://doi.org/10.1007/s11356-021-16420-x

Sharma, R., & Das, S. (2023). BiblioMagika: A bibliometric and network analysis tool for early-stage researchers. *Journal of Scientometric Research, 12*(1), 45-56. https://doi.org/10.5530/jscires.2023.12.7

Sianes, A., Vega-Muñoz, A., Tirado-Valencia, P., & Ariza-Montes, A. (2022). Impact of the sustainable development goals on the academic research agenda. A scientometric analysis. *PLoS One, 17*(3), Article e0265409. https://doi.org/10.1371/journal.pone.0265409

Silva, J., & Ramos, S. (2021). ScientoPy: A python-based tool for the analysis of scientific trends in publications. *SoftwareX, 14,* Article 100691. https://doi.org/10.1016/j.softx.2021.100691

van Eck, N. J., & Waltman, L. (2010). Software survey: VOSviewer, a computer program for bibliometric mapping. *Scientometrics, 84*(2), 523-538. https://doi.org/10.1007/s11192-009-0146-3

van Eck, N. J., & Waltman, L. (2014). CitNetExplorer: A new software tool for analyzing and visualizing citation networks. *Journal of Informetrics, 8*(4), 802-823. https://doi.org/10.1016/j.joi.2014.07.006

van Eck, N. J., & Waltman, L. (2017). Citation-based clustering of publications using CitNetExplorer and VOSviewer. *ArXiv Preprint,* arXiv:1702.03411. https://arxiv.org/abs/1702.03411

Waqas, A., Salminen, J., Jung, S. G., Almerekhi, H., & Jansen, B. J. (2019). Mapping online hate: A scientometric analysis on research trends and hotspots in research on online hate. *PLoS One, 14*(9), Article e0222194. https://doi.org/10.1371/journal.pone.0222194