Data retrieval, data cleaning and data merging: creating a database for bibliometric analyses

ROMANA JADRIJEVIĆ[1 2], ANTONIJA MIJATOVIĆ[1], IVANA BABIĆ[2] & ANA MARUŠIĆ[1 2]

[1] University of Split School of Medicine
[2] University Hospital of Split

ABSTRACT

Introduction: Bibliometric analyses have become fairly frequent in today's science (Donthu et al., 2021; Klarin, 2024; Öztürk et al., 2024). They help map out research areas and evaluate the quality of scientific research. Most bibliometric analyses today gather their data from Web of Science or Scopus. Yet, the methodology sections on obtaining that data often leave us wanting, especially in the few cases where datasets from both databases were merged (Echchakoui, 2020).

The aim of our research was to create a comprehensive database of all publications by Croatian clinicians, regardless of the research area, so we could map their areas of interest during the period 2005-2022, as well as study potential effects of Croatia's European Union (EU) membership. Our focus on publications by clinicians rather than publications in clinical medicine led us to create a more complex data retrieval strategy, using affiliations instead of research areas.

Methods:

Identification of publications

Using the government-provided list of registered healthcare legal subjects, a list of all possible names for all types of medical subjects was created, excluding medical faculties. Using truncation, the terms were grouped where possible. Operators OR, AND, and SAME were used to create the search strategy in Web of Science: Core Collection (WoS:CC) and Scopus.

The search was conducted separately for each database. The results in both databases were limited to publication years 2005 through 2022, in order to obtain data for a proportional number of years prior and after Croatia joined the EU.

Data retrieval

The results were downloaded via built-in export functions in both databases. The built-in option was used due to API keys either being too expensive or limited in scope for this part of the process.

The data from WOS:CC had to be downloaded in groups of 1,000 due to the

limitations of our national licenses. The first 20,000 results from Scopus, sorted by date (newest first), were exported simultaneously. As Scopus only allows the first 20,000 results to be downloaded, we changed the sorting to date (oldest first), and downloaded the remaining 8,159 results. All exported data were deposited as raw data in the form in which they were downloaded. Copies were made and then combined into a single Excel document for each database separately.

We used Python to automatize parts of our cleaning process. We implemented random controls of data after any step including Python, to ensure the different rows and columns did not shuffle.

Deduplication

The most distinguishable identifier common to both datasets was the DOI number. We considered the title as well, but given how a portion of the publications had the same title, e.g. Reply or Letter to the Editor, we realized the most trustworthy identifier was DOI.

Using Excel's built-in sorting function, we were able to easily remove all records with the empty DOI field. We used Python to identify records that appeared in both datasets and removed the duplicates found in Scopus, leaving the WoS records as they had less missing data than Scopus.

Merging and data cleaning

To merge the two datasets, we had to decide which data columns we needed to keep, and, of those, which ones could be combined into one column. We used Excel's Merge Tables function to create the unique dataset.

During the previous steps, we noticed that some of our data probably did not meet the inclusion criteria. Through trial and error, we realized that the most thorough way to ensure the eligibility of all records in our dataset was to check each entry manually. All in all, we checked 21,651 records, 3,531 of which were excluded for not meeting the inclusion criteria.

Missing data

The most important type of missing data was the publication month, as our interrupted time-series analysis would use them as observation points. We automatized this process for 989 records using DOIs and PubMed export, and manually extracted the 1,530 remaining.

The majority of missing publication month data originated from Scopus' records for publications from Croatian journals.

Results: In total, 26,873 records from WoS and 28,159 records from Scopus were identified. Of those, 9,045 from WoS and 10,600 from Scopus were filtered out for not having a DOI number and are awaiting manual deduplication and data cleaning.

A total of 13,730 duplicates were removed.

A total of 21,657 records with DOI were assessed for eligibility. We excluded 3,530 records for not having a required affiliation and 2 for not being able to find any publication month-related data.

All in all, 18,125 records with DOI were included in the database for our study.

Discussion: While being aware that no database can be perfect, we believe we will have the most comprehensive bibliographic database of the research Croatian clinicians published in the period from 2005 to 2022 after including the records without DOI.

Sometimes limiting our bibliometric analyses to a database's predetermined research areas and document types may not be enough to gain a full overview of a body of research. We believe affiliations are the only currently viable way to gain insight into the research that specific types of institutions conduct.

KEYWORDS

bibliometrics; data cleaning; data merging; data retrieval; database merging

REFERENCES

1. Donthu, N., Kumar, S., Mukherjee, D., Pandey, N., & Lim, W. M. (2021). How to conduct a bibliometric analysis: An overview and guidelines. Journal of Business Research, 133, 285– 296. https://doi.org/10.1016/j.jbusres.2021.04.070

2. Echchakoui, S. (2020). Why and how to merge Scopus and Web of Science during bibliometric analysis: The case of sales force literature from 1912 to 2019. Journal of Marketing Analytics, 8(3), 165–184. https://doi.org/10.1057/s41270-020-00081-9

3. Klarin, A. (2024). How to conduct a bibliometric content analysis: Guidelines and contributions of content co-occurrence or co-word literature reviews. International Journal of Consumer Studies, 48(2), e13031. https://doi.org/10.1111/ijcs.13031

4. Öztürk, O., Kocaman, R., & Kanbach, D. K. (2024). How to design bibliometric research: An overview and a framework proposal. Review of Managerial Science. https://doi.org/10.1007/s11846-024-00738-0