kompetenzzentrum
bibliometrie

Summary of Forschungspool 2020 reports comparing Dimensions bibliometric database with Web of Science and Scopus

In 2020, the Kompetenzzentrum Bibliometrie (KB)[1] commissioned three reports to assess the new Dimensions bibliometric database. The purpose of the reports was to evaluate the Dimensions database regarding its potential inclusion in the KB infrastructure. Dimensions, released by Digital Science in 2018, differs from Scopus and WoS[2] in its larger size - due to its index of a broader range of document types and less rigorous curation processes - and potentially offers the chance to address known issues with WoS and Scopus, such as under-representation of the social sciences and humanities as well as content from outside North America and Western Europe. A small number of previous studies of these databases concluded that Dimensions performs similarly to Scopus and both have higher coverage and citation counts than WoS. The three KB reports sought to determine whether Dimensions is suitable for the KB's purposes and could provide additional value beyond that currently obtainable via WoS and Scopus.

The reports were prepared by Forschungszentrum Jülich, Universität Bielefeld, and the Deutsches Zentrum für Hochschul- und Wissenschaftsforschung (DZHW). FZ Jülich quantitatively analysed Dimensions' content between 1996 and 2018 compared to WoS and Scopus, with a particular focus on typical variables necessary for bibliometric analyses. The DZHW examined the outcomes of such bibliometric analyses by comparing the differences in normalised citation impact of German publications between the databases and the databases' citation networks. Universität Bielefeld addressed the KB-specific requirement for reliable coding of German institutions by examining the feasibility of applying the existing KB institutional address coding system to Dimensions data and an evaluation of the address coding results compared to WoS. The three reports' results are summarised here.

FZ Jülich's analysis confirmed that, overall, Dimensions constitutes a larger corpus than Scopus and WoS. Both FZ Jülich and DZHW reported that Scopus held the largest number of German publications. However all 3 reports indicate that Dimensions was missing GRID-IDs - the data field from which the affiliation and corresponding country can be obtained indirectly - for up to 50% of publications. This result suggests that Dimensions would hold the largest collection of German articles should this issue be resolved. Indeed, Dimensions had twice as many German-language publications as Scopus and 3 times more than WoS, and an overall more diverse array of languages (FZ Jülich) indicating that Dimensions' larger scope may provide better coverage of regional content. However, the two largest sources of Dimensions' content were preprint and abstract repositories (FZ Jülich), highlighting Dimensions' broader interest in the entire spectrum of scientific publishing, including documents not typically applied in bibliometric analyses. Moreover, 20% of items in Dimensions did not have a source, which is problematic for bibliometric analyses that normalise indicators against the publishing journal.

In terms of disciplinary coverage, after mapping each database's native discipline classifications to the Fields of Research, for publications from 2016, Dimensions followed the same trend as WoS and Scopus in having much stronger coverage of the natural and medical sciences and engineering fields than the social sciences and humanities, although Dimensions' coverage in these latter fields was

[1] The KB is funded by the German Federal Ministry for Education and Research (Grant number 01PQ17001) and maintains a quality controlled in-house data infrastructure for bibliometric research and services. Further information: http://bibliometrie.info/

[2] Defined as the indices SCIE, SSCI, AHCI and CPCI

improved over WoS and Scopus (DZHW). Scopus held the largest coverage in most other disciplines and WoS the least: However, Scopus has been noted to somewhat “over-attribute” publications to multiple disciplines based on only loose associations, that may probably explain this observation. Dimensions’ article-level classification via AI and natural language processing might be more accurate than the journal-based approaches by WoS and Scopus, but the coverage is insufficient. Nearly 30% of publications in Dimensions were unclassified compared to less than 1% in WoS and Scopus (FZ Jülich). This poses a challenge for normalisation of indicators by discipline.

FZ Jülich also reported greater diversity in the document types used in WoS and Scopus over Dimensions. Only around half of the items of a ~500 sample classified as “articles” by Dimensions were also classified as articles in WoS or Scopus and were otherwise notes, book reviews, editorial material, and so on. This misattribution of document types in Dimension is a result of a simplistic allocation procedure,that classifies all items in journals to document type “article”. In other words Dimensions does not offer a classification system of types of documents in journals to allow distinctions between substantial original research in research articles or review papers and less substantial or original content in other document types like, editorials, etc. DZHW also identified this processing as problematic. Moreover, as bibliometric methods compute statistics across all relevant documents by discipline, the extensive inclusion of uncharacteristic material in a particular document type due to inaccurate classification can thus potentially lower the signal-to-noise ratio and impact these statistics, as observed by the DZHW.

Regarding other data fields, FZ Jülich and DZHW also found that Dimensions had more complete DOIs (<1% missing) than WoS or Scopus (4-27%), but Dimensions was missing author names for 8% of items, compared to 2% in Scopus and 0% in WoS (FZ Jülich). WoS also held more information about funding organisations than Dimensions, and both more than Scopus.

The completeness of data was also a key issue for Bielefeld University when applying institutional coding to Dimensions data. They reported good coverage of data fields but that the full institutional address data was missing for 35% of authors, as were 50% of GRID-IDs, both of which are critical for institution coding. However, if these issues were resolved, Bielefeld University states that institutional coding could be applied to Dimensions with a similar success rate to the ~95% currently achieved in WoS.

Comparing the relative citation impact of the same German publications in each database, DZHW noted discipline-level differences in publications’ observed and expected citations due to the databases’ exclusive content. Once aggregated to the institutional sector level, the largest difference was between WoS and Scopus, with all sectors performing better in Scopus, particularly those oriented toward applied science. Slightly higher values were also seen in Dimensions over WoS, but performance in Dimensions and Scopus was very similar.

Further, an analysis of the citation networks of publications jointly indexed in all three databases and publications exclusive to each database identified that each database maintains the same core of highly cited publications and each its own periphery of exclusive publications that draw on the core, but are themselves less well-cited documents. In Scopus and Dimensions, there is a greater degree of knowledge transfer to the periphery visible. However, in none of the databases are these exclusive publications particularly influential to the core. These findings suggest that, at this current time, the exclusive content of Dimensions (and Scopus) offer a varying perspective of the science system which does not differ structurally. Rather, there is a large and clearly identifiable core of publications that all three data sources cover to a similar degree.

The three reports indicate that incomplete data is the most critical issue in Dimensions. However, FZ Jülich’s analysis of two snapshots of Dimensions taken one year apart showcased how Digital Science rapidly addresses weaknesses of the database, such as missing discipline and language information.

Alongside its continuously growing content, recent developments indicate that Dimensions will overcome current limitations in future. However, its rapid growth and transformation in these early years should also be considered in the challenges it could pose for bibliometric analyses involving a time-series or performed annually using this database.

Appraisal

Based on the outcomes of these reports, the following conclusions are made to the KB:

1. Dimensions constitutes a comprehensive database which already meets several requirements for the bibliometric work in the KB.
2. The relatively young database is constantly improving and shows great potential to match the established databases.
3. The data quality of Dimensions does not currently live up to its own standards. The coverage and precision of metadata is missing for a substantial share of the indexed publications, impacting bibliometric analyses.
4. Given the currently achieved state and the constant improvements, the data quality in the Dimensions database might be closely monitored by the KB to observe if the current gap in quality diminishes.
5. The apparent resemblance with Scopus motivates a future analysis of the wider Dimensions universe, as the links of publications to other data types might define the unique trait of Dimensions beyond the current offering in WoS or Scopus.