Microsoft Academic is on the verge of becoming a bibliometric superpower


*Last year, the new Microsoft Academic service was launched. **Sven E. Hug** and **Martin P. Brändle** look at how it compares with more established competitors such as Google Scholar, Scopus, and Web of Science. While there are reservations about the availability of instructions for novice users, Microsoft Academic has impressive semantic search functionality, broad coverage, structured and rich metadata, and solid citation analysis features. Moreover, accessing raw data is relatively cheap. Given these benefits and its fast pace of development, Microsoft Academic is on the verge of becoming a bibliometric superpower.*

In 2016, Microsoft released a new academic search engine. This happened quietly, as if the company was afraid of embarrassing itself again, as it did years ago in its loss to Google Scholar in the search engine race. However, there is absolutely no need to hide the new database, Microsoft Academic, since it has the potential to outduel Google Scholar, Web of Science, and Scopus. In fact, with 168 million records as of early 2017, the database has already outstripped Web of Science (59 million records) and Scopus (66 million records) in terms of coverage. Nothing can be reliably said with regard to Google Scholar, as its size has not been declared (estimations range from 160-200 million records). To access the new database, one can use the Microsoft Academic search interface or the Academic Knowledge API.

Search less, research more?

First trials show that the search interface of Microsoft Academic returns relatively few but very accurate results. This is due to its semantic search engine, which leverages entities associated with a paper (e.g. fields of study, journal, author, affiliation). In contrast, most other scholarly databases rely on search terms, which are also employed by Microsoft Academic but only if semantic search fails. Much like library databases, Microsoft Academic offers a range of filtering and sorting options to refine search results. This is very convenient and a plus compared to Google Scholar, which provides only very limited refinement options.

However, the search interface in its current stage is not without pitfalls and drawbacks. Above all, tutorials or instructions are virtually missing, leaving any first-time user puzzled. For example, who would have guessed that the small symbols showing up in the search slot represent the entities that constitute the database (e.g. a laboratory flask for “field of study”)? And who would have known that natural language queries such as “papers about bibliometrics after 1977 citing Eugene Garfield” can be performed? Also, one has to be aware that queries need a bit of patience and can be choppy at times, as it takes the engine a while to suggest supplementary search terms and to eventually display the results.

Microsoft recognises that semantic query is still not popular and that users need time to adapt. Hence, the new database may not yet live up to its slogan: “research more, search less”. However, Microsoft Academic is being developed at a relentless pace. Just recently, a social networking site for academics has been integrated. Hopefully, the performance of and the instructions for the search interface will be further improved soon.

Beyond searching: citation analysis

To fully tap the wealth of Microsoft Academic, one has to employ the Academic Knowledge API, which comes at relatively low cost ($0.25 per 1,000 queries). We have examined the API from the perspective of bibliometrics (i.e. the quantitative study of scholarly communication) and found that the metadata is structured and rich and can easily be retrieved, handled, and processed. The API allows retrieving aggregated citation counts and frequency distributions of citations. These features enable the calculation of a wide range of indicators and are a major advantage of Microsoft Academic over Google Scholar. First studies have shown that citation analyses with Microsoft Academic, Scopus, and Web of Science yield similar results with respect to the h-index, average-based and distribution-based indicators, and rank correlations of citation counts.

However, there are some limitations regarding the available metadata. First, the database does not provide the document type of a publication, which is often used for normalising indicators. Second, the fields of study – there are more than 50,000 of them! – cannot readily be employed for bibliometric analyses as they represent the semantics of a paper rather than traditional research fields.

Testing the coverage

To learn how well an actual publication list is represented in Microsoft Academic, we examined the coverage of the publication output of the University of Zurich in the database. We used Scopus and Web of Science as benchmarks and the university's publication repository to complete missing information (e.g. document type). When focusing on main document types we found that, overall, the coverage of Microsoft Academic (56.6%) and Scopus (57.9%) is almost equal, with Web of Science (52.6%) trailing slightly. However, Microsoft Academic clearly surpasses Scopus and Web of Science with respect to book-related document types and conference items, but falls somewhat behind Scopus with regard to journal articles (see Figure 1). All three databases show similar biases with regard to the coverage of the social sciences and humanities, non-English publications, and open access publications.

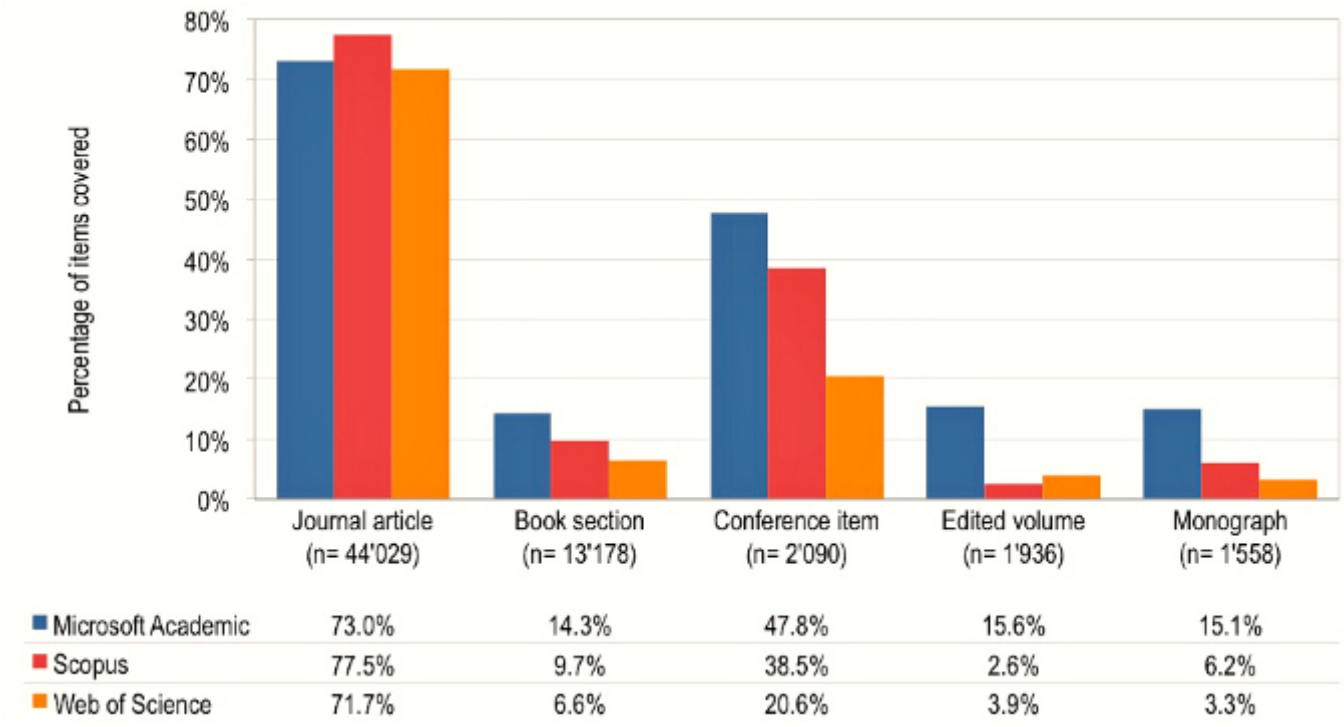

Figure 1: Coverage of the publication output of the University of Zurich (2008-2015) by main document types. Source: Hug and Brändle (2017). The coverage of Microsoft Academic: Analyzing the publication output of a university

These findings suggest that Microsoft Academic performs similarly to Scopus and Web of Science when focusing on main document types. On the one hand, this speaks for the quality of Microsoft Academic, while on the other hand it is somewhat deflating as one expects a more comprehensive coverage of a database that gets the majority of its data from web pages. We assume that Microsoft first filled the database with metadata feeds from the largest journal publishers and then went on with web indexing. This could explain why the coverage of Microsoft Academic is similar to the benchmark databases. Since the size of Microsoft Academic is growing rapidly, we expect future studies to find a much broader coverage.

Data quality

Not much is known about the data quality yet. We found that 89.5% of the publication years are correct, 7.0% differ by ±1 year, and 3.5% feature larger differences. In addition, we found that 95.1% of the journal articles list the correct number of authors.

A new bibliometric superpower?

Microsoft Academic combines the following features of existing databases: a broad coverage (Google Scholar), structured and rich metadata as well as functionality (Scopus and Web of Science), and a social network for academics (ResearchGate). In addition, accessing raw data is relatively cheap. Given these features and the ongoing development, we believe Microsoft Academic is on the verge of becoming a bibliometric superpower.

This blog post is based on the authors' articles, "Citation analysis with microsoft academic", published in Scientometrics (DOI: 10.1007/s11192-017-2247-8); and "The coverage of Microsoft Academic: Analyzing the publication output of a university", a preprint currently available on arXiv.

Featured image credit: Sketch of Twitter Data Visualization by Patrick Dinnen (licensed under a CC BY-SA 2.0 license).

Note: This article gives the views of the authors, and not the position of the LSE Impact Blog, nor of the London School of Economics. Please review our comments policy if you have any concerns on posting a comment below.

About the authors

Sven E. Hug *is working as a research associate at the Professorship for Social Psychology and Research on Higher Education (ETH Zurich) and as a project manager at the Evaluation Office of the University of Zurich. His interest is in scientometrics, peer review, and research evaluation. His ORCID iD is 0000-0002-7624-9529.*

Martin P. Brändle *is working as an information scientist and software engineer at the IT services of the University of Zurich. His interest is in information system development, repository software and open access initiatives, long term archiving of research data, and information visualisation. His ORCID iD is 0000-0002-7752-6567.*

June 19th, 2017 | Citations, Metrics | 2 Comments

Share This Story, Choose Your Platform!