Goodbye, Microsoft Academic – Hello, open research infrastructure?

blogs.lse.ac.uk/impactofsocialsciences/2021/05/27/goodbye-microsoft-academic-hello-open-research-infrastructure/

May 27, 2021

The announcement of the closure of Microsoft Academic later this year, may have left the research community largely unmoved, although its demise has significant implications for those working with the service's substantial database. Here, **Aaron Tay, Alberto Martín-Martín,** *and* **Sven E. Hug¸** *discuss what set Microsoft Academic apart from its competitors and the potential consequences of Microsoft's withdrawal from scholarly metadata for the development of open research infrastructures.*

Recently, Microsoft announced that it will shut down Microsoft Academic, the second largest academic search engine after Google Scholar. Although the global scientific community took little notice of this announcement, many computer scientists, meta-researchers, librarians, and start-ups were shocked, because they had been building an ecosystem of information services around the database.

Microsoft Academic is not the company's first attempt to build a literature search tool. An earlier project, Microsoft Academic Search, ran from 2009 to 2012 and fell into shabby disrepair before being officially relaunched as Microsoft Academic in 2016. This is indicative of how Microsoft had never intended to enter into the business of scholarly metadata. Instead, the tech giant has been using data on scholarly communication as testing ground for big data and artificial intelligence (AI) technologies, as a recent article by Redmond researchers suggests. It is rumoured that Microsoft may offer the tested technologies to harvest knowledge from documents in Office 365.

A sophisticated search engine

While traditional citation indexes, such as Web of Science and Scopus, are mainly based on selected journals, Microsoft Academic's strength has been the way it crawls the web and it's use of AI technologies to populate its database. It is thus not surprising that Microsoft Academic has been faster at indexing new publications and contains significantly more records (194 million, without patents) than the Web of Science Core Collection (79 million) and Scopus (75 million). Microsoft Academic also encompasses a much broader range of publication types (preprints, working papers, dissertations, etc.) and shines in research fields that traditional citation databases often do not cover well, such as computer science, social sciences, and humanities.

> Microsoft Academic's strength has been the way it crawls the web and it's use of AI technologies to populate its database.

A major advantage of Microsoft Academic over Google Scholar is the search interface, which for now still offers ample filtering and sorting options and provides various rankings (topics, journals, institutions, etc.) as well as visualizations of summary statistics. Although the search engine is free of charge and features an integrated social network for academics, it has never been popular with researchers, as can be seen from web traffic statistics:

Total visits in April 2021 according to SimilarWeb (in million)	
scholar.google.com	137.5
semanticscholar.org	8.9
scopus.com	5.2
webofknowledge.com	4.4
academic.microsoft.com	0.7

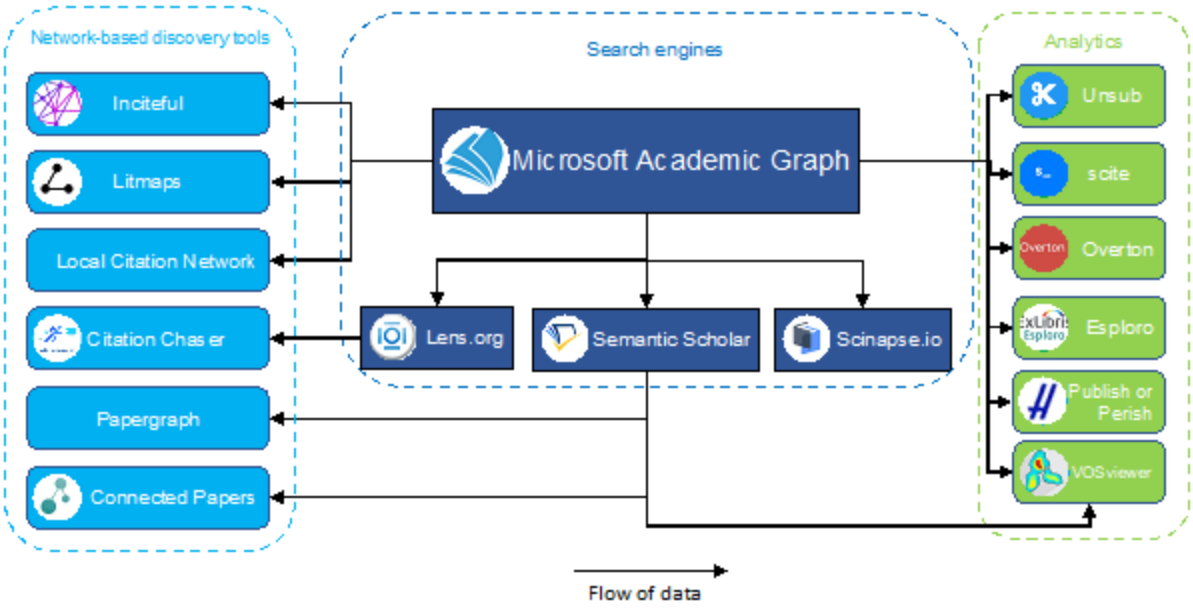
The main reason for this low usage is likely the search interface itself. It differs fundamentally from traditional academic search systems as it is driven by AI technologies. Specifically, the interface offers true semantic search instead of the usual keyword search with Boolean operators. Or as Microsoft once explained: 'Microsoft Academic understands the meaning of words, it doesn't just match keywords to content. For example, when you type "Microsoft", it knows you mean the institution, and shows you publications authored by researchers affiliated with Microsoft.' In addition, the search engine is based on more than 700,000 'fields of study' (i.e., topics or concepts) that were created and are continuously expanded by algorithms, whereas other search systems use fixed, human curated, and less complex classifications. Furthermore, the search engine employs two unique metrics, saliency and estimated citation counts, which are difficult to understand and interpret for most users. Overall, these AI-driven features create a search experience that is very different from what users are accustomed to. It thus seems that the AI technologies employed are either too avant-garde for users or not mature enough.

A wealth of data for free

While the search engine has not been embraced by the scientific community, the underlying data, the Microsoft Academic Graph, has attracted many users. There are several reasons for this. The data set is huge, well structured, and detailed. Its use is free of charge, and access is convenient (API or full data-dump). In contrast, direct access to Google Scholar data is impossible, and data can only be scraped from Google Scholar to a very limited extent. Although Microsoft exclusively employs AI technologies to collect and curate data, data quality is reasonably accurate and suitable for the large-scale analysis of some aspects of scholarly communication.

> Microsoft Academic has enabled researchers and commercial enterprises alike to work with comprehensive metadata at little cost

In this way, Microsoft Academic has enabled researchers and commercial enterprises alike to work with comprehensive metadata at little cost. Before Microsoft made its database available, only researchers at a few institutes (in rich countries) had access to large data sets, and the companies owning such data mostly used it for their own products. The paper introducing the Microsoft Academic Graph has been cited more than 500 times since 2015, which indicates how useful the database has been in research. The graph is also used in many commercial and non-commercial tools and services (e.g., VOSviewer, Unsub, Litmaps, scite). And there are even some bibliographic databases and search engines that tap into the wealth of Microsoft Academic (e.g., Semantic Scholar, The Lens, Scinapse).



Although the closure of Microsoft Academic will not impact the performance of these tools and services in the same way, it is obvious that a valuable resource will be lost at the end of this year. It remains to be seen whether and how it can and will be replaced. The least expensive solution would be to pay Microsoft for continuing the database, which of course would require Microsoft to be willing to keep it running. The annual cloud computing cost for updating the content of the Microsoft Academic Graph is roughly equal to the salary of a single experienced data scientist. A developer of the database recently estimated that maintaining Microsoft Academic at the current technical level would cost about one-third of the amount a medium-sized university would pay for data from a traditional citation index.

Towards open research infrastructures?

Microsoft Academic has demonstrated the value of openly available metadata that has been collected and curated by AI technologies. It has provided a fertile ground for both researchers and commercial enterprises. There are, of course, other open metadata sources. For example, Crossref contains more than 125 million records, 48 million of which have open references thanks to the Initiative for Open Citations and the collaborating publishers. However, Crossref is smaller, contains less detailed data, is less consistently curated, and only indexes publications with a DOI (digital object identifier).

In the end, Microsoft's project has demonstrated that it is not enough to make a database publicly available – a database must also be sustainable. If we want open and sustainable databases, it would probably be a good idea to invest more time and resources in building them. And to begin with, we could support, for example, those who plan to build an open-source and free to use replacement for Microsoft Academic.

Note: This article gives the views of the author, and not the position of the Impact of Social Science blog, nor of the London School of Economics. Please review our Comments Policy if you have any concerns on posting a comment below.

Image Credit: Alexandre Debiève via Unsplash.
