Journal of Informetrics 11 (2017) 1201–1212

Contents lists available at ScienceDirect

Journal of Informetrics

journal homepage: www.elsevier.com/locate/joi

Regular article

# Microsoft Academic: A multidisciplinary comparison of citation counts with Scopus and Mendeley for 29 journals

Mike Thelwall

*Statistical Cybermetrics Research Group, University of Wolverhampton, UK*

ARTICLE INFO

*Article history:*
Received 11 August 2017
Received in revised form 6 October 2017
Accepted 28 October 2017

*Keywords:*
Microsoft Academic
Citation analysis
Early citations
Google Scholar
Mendeley

ABSTRACT

Microsoft Academic is a free citation index that allows large scale data collection. This combination makes it useful for scientometric research. Previous studies have found that its citation counts tend to be slightly larger than those of Scopus but smaller than Google Scholar, with disciplinary variations. This study reports the largest and most systematic analysis so far, of 172,752 articles in 29 large journals chosen from different specialisms. From Scopus citation counts, Microsoft Academic citation counts and Mendeley reader counts for articles published 2007–2017, Microsoft Academic found a slightly more (6%) citations than Scopus overall and especially for the current year (51%). It found fewer citations than Mendeley readers overall (59%), and only 7% as many for the current year. Differences between journals were probably due to field preprint sharing cultures or journal policies rather than broad disciplinary differences.



## 1. Introduction

Microsoft Academic is the replacement for Microsoft Academic Search, generated by Microsoft Asia (Sinha et al., 2015). Like Google Scholar, it is a free search engine for academic research and includes a citation index. Unlike Google Scholar (Halevi, Moed, & Bar-Ilan, 2017), it allows automatic data collection via an API (Chen, 2017) and so has the potential to be used for scientometric applications that require large amounts of data, such as calculating field normalised indicators (Thelwall, 2017b; Waltman, van Eck, van Leeuwen, Visser, & van Raan, 2011) (for example, see: Hug, Ochsner, & Brändle, 2017) and analyses of large groups of researchers (Science-Metrix, 2015). Given this potential it is important to compare the coverage of Microsoft Academic with that of other citation indexes, such as Google Scholar, Scopus and the Web of Science (WoS) to evaluate its suitability as a scientometric data source (Harzing & Alakangas, 2016). Its potential advantages over existing sources are (Hug & Brändle, 2017): allowing automatic data collection (compared to Google Scholar); collecting more citations through access to web crawler data (compared to Scopus and WoS); collecting non-academic citations, such as from newspapers (Hug et al., 2017); collecting early citations to recently-published papers, also through access to a web crawler (compared to Scopus and WoS); and for data triangulation to check for database-induced biases in indicator values.

Although Microsoft Academic was only formally released in July 2017, several papers have investigated its key properties. Like Google Scholar, it has indexing errors, such as incorrect publication dates (Harzing & Alakangas, 2016; Harzing & Alakangas, 2017a; Hug et al., 2017), as could be expected from an index that incorporates web crawler data. Its metadata accuracy seems to have improved since its early version, however (Harzing & Alakangas, 2017b).

*E-mail address:* m.thelwall@wlv.ac.uk

https://doi.org/10.1016/j.joi.2017.10.006

### 1.1. Citation counts in Microsoft Academic, Google Scholar, Scopus and Web of Science

Microsoft Academic calculates two citation count for each article. It reports a standard count of (algorithmically verified) citations (its CC field) but also an Estimated Citation Count (ECC) that adds the number of citations that it estimates exist but have not been found using statistical techniques (Harzing & Alakangas, 2017a). One study found ECC values for individual academics to sometimes be identical to CC values for all their publications but in other cases totalling their ECC values gave more than twice the total of their CC values (Harzing & Alakangas, 2017a). The difference seems to be field-based, with Microsoft Academic's algorithm apparently estimating that it has covered some fields comprehensively but that it has missed most articles in other fields. The current paper uses CC values and the discussion below refers to these traditional citation counts unless ECC is specified.

An October 2016 study collected data from Microsoft Academic, Scopus and WoS on the publications (of all types) of 145 University of Melbourne professors and associate professors from five broad areas. At the level of academics, Google Scholar found the most citations in all areas; Microsoft Academic found more citations than Scopus and WoS in Social Sciences, Engineering and Humanities, but slightly fewer than Scopus and about the same as WoS in Life Sciences and Sciences (Harzing & Alakangas, 2017a; see also: Harzing, 2014).

A June 2017 comparison between Google Scholar and Microsoft Academic for the publications in the profiles of the same 145 academics from five areas found that average estimated citation counts (i.e., ECC values) were about the same as Google Scholar in Life Sciences, but lower in Sciences, Social Sciences, Engineering, and Humanities (Harzing & Alakangas, 2017b). Given that ECC values tend to be higher than CC values, by June 2017 Microsoft Academic still had not found as many citations as Google Scholar. Since the relationship between ECC and CC varies by field, it is not possible to estimate citation counts from the ECC values to compare with WoS or Scopus. For the same data, there were high Spearman correlations between Microsoft Academic and both Scopus and WoS citation counts, varying between 0.73/0.74 (Scopus/WoS) for Humanities and 0.86/0.85 for Natural Sciences. These values include articles from multiple years (2008–2015) and multiple fields and are therefore likely to overestimate the underlying relationship strength. This is because the correlation coefficients will be inflated by year differences (older articles tending to have higher citation counts than younger articles) and field differences (articles in high citation fields tending to have higher citation counts than articles in low citation fields).

One study compared the journal-normalised citation scores of three academics based on their publications in Scientometrics 2010–2014, finding that the choice of citation source changed the results (e.g., from 0.55 to 0.69 for researcher B) and altered the rank order of the researchers (Hug et al., 2017).

An institution-level analysis of mandated publications in the University of Zurich digital repository found that Microsoft Academic indexed fewer journal articles from 2008 to 2015 than Scopus but slightly more than WoS (Hug & Brändle, 2017). For a combined data set of journal articles, conference papers, monographs, book chapters and edited volumes 2008–2014 from the same repository, Microsoft Academic (19.5 citations per publication) had lower citation counts than Scopus (25.8) and WoS (26.3) in Natural Sciences. The three sources gave broadly similar results in Engineering & Technology, Medical & Health Sciences and Agricultural Sciences, Social Sciences and Humanities, with WoS having the lowest values in these areas.

At the journal level and taking into account the time dimension, average citation counts between Scopus and Microsoft Academic have been compared in one field (library and information science) and for the general scientific journals *Science* and *Nature*, analysing the years 1996–2017 separately (Thelwall, in press). Scopus reported higher citation counts for *Science* and *Nature* whereas Microsoft Academic reported higher values for several library and information science journals. The differences were small – typically about 5–10% and did not vary much over time (1996–2016) in percentage terms.

### 1.2. Early citations

Early citations are important in scientometrics because citation analyses typically need a few years of publication data in Scopus or WoS to get high enough citation counts for a reasonable analysis (Campanario, 2011; Wang, 2013). Any data source that finds earlier impact evidence may help analysts to conduct more timely analyses. Google Scholar and Microsoft Academic have, in theory, an advantage for early citations because they can index informally published preprints from the web in advance of their official publication or early view dates.

Microsoft Academic reports (although does not necessary extract citations from) journal articles, conference papers, books, book chapters (but see below), white papers and newsletters (Harzing, 2016). It also covers working papers (essentially the same as white papers) and has a low but non-zero coverage of habilitations and book chapters, but no coverage of newspaper articles (Hug et al., 2017). It is important to distinguish between coverage (having a metadata record within Microsoft Academic) and indexing (extracting the citations from the document). It seems reasonable to assume that Microsoft Academic would have records for documents that it cannot index but that it would attempt to index any document (including white papers) for which it can access the full text.

Given that Microsoft Academic, like Google Scholar, finds and presumably indexes working/white papers, including preprints, it seems likely that it would tend to report higher citation counts for recently published articles than Scopus and WoS. Nevertheless, Microsoft Academic does not seem to be able to find enough early citations to make a noticeable difference for articles in their publication year, although this has only been checked for library and information science journals, *Science* and *Nature* (Thelwall, in press).