Does Microsoft Academic Find Early Citations?[1]

Mike Thelwall, Statistical Cybermetrics Research Group, University of Wolverhampton, UK. m.thelwall@wlv.ac.uk

This article investigates whether Microsoft Academic can use its web search component to identify early citations to recently published articles to help solve the problem of delays in research evaluations caused by the need to wait for citation counts to accrue. The results for 44,398 articles in Nature, Science and seven library and information science journals 1996-2017 show that Microsoft Academic and Scopus citation counts are similar for all years, with no early citation advantage for either. In contrast, Mendeley reader counts are substantially higher for more recent articles. Thus, Microsoft Academic appears to be broadly like Scopus for citation count data, and is apparently not more able to take advantage of online preprints to find early citations.
Keywords: Microsoft Academic; Google Scholar; Mendeley; citation analysis; early citations.

Introduction

When citation counts are used for research evaluation, it is usual to employ a citation window of 2-5 years to ensure that citation count data is stable enough to make impact comparisons (Wang, 2013). This imposes a delay between publication and evaluation, and any source of early citation data might be able to shorten this delay. Google Scholar is a potential source of early citations due to its indexing of conference papers and preprints on the web, which allows it to find more citations than traditional citation indexes (Falagas, Pitsouni, Malietzis, & Pappas, 2008; Harzing, 2014; Moed, Bar-Ilan, & Halevi, 2016). It must be manually searched (except for Publish or Perish: Harzing, 2007), which is impractical for many scientometric studies. This paper assesses whether Microsoft Academic, which supports large scale data collection, shares the wider coverage advantage of Google Scholar so that it can be a useful automated source of early citations.

Microsoft Academic Search was an early competitor to Google Scholar as a free citation index of academic publications culled from the web (Carlson, 2006). It was problematic (Jacsó, 2011; Ortega & Aguillo, 2014) and far less successful, with the project supporting it ceasing in 2012 (Microsoft, 2017a). It continued to be available after this without being updated (Orduña-Malea, Martín-Martín, Ayllon, & Delgado Lopez-Cozar, 2014). Its replacement, Microsoft Academic, was formally launched in July 2017, although a preliminary version had previously been available. Its main goal was to be semantic academic search engine (Sinha, Shen, Song, Ma, Eide, Hsu, & Wang, 2015) that would help users to find relevant research documents even if they did not match the query terms (Microsoft, 2017b). Like Google Scholar, Microsoft Academic allows users to have their own profile pages in the site. An important difference is that Microsoft Academic allows automated searches (Chen, 2017). This is valuable for scientometric purposes when large sets of publications are analysed.

Analyses of the beta version of Microsoft Academic have described its core properties. As might be expected from a tool that extracts references from web texts, it includes indexing errors such as incorrect publication years. An early investigation of Harzing's academic publications of all types found that Google Scholar indexed more of her

---

publications and citations than Microsoft Academic (both indexed all her journal articles) and Microsoft Academic found more publications and citations than either Scopus or the Web of Science (WoS), although Scopus and WoS indexed some non-article publications (e.g., book chapters) not in Microsoft Academic (Harzing, 2016). A study of University of Zurich publications from its open access repository found Microsoft Academic to index fewer journal articles than Scopus but more than WoS; it indexed more books, book chapters and conference papers than WoS and Scopus (Hug & Brändle, in press). A larger investigation studied 145 academics from five broad disciplines. The citation counts for their publications were higher than those from Scopus and WoS in Engineering, Social Sciences, and Humanities, but similar in Science and Life Sciences. In all areas, Google Scholar found the most citations (Harzing & Alakangas, 2017a). A more recent study of the citations to the works of 145 authors found Microsoft Academic to have similar coverage to Google Scholar except that in the humanities it indexed fewer non-journal publications than Google Scholar but more than Scopus and the WoS in the humanities (Harzing & Alakangas, 2017b). Finally, one investigation identified the lack of coherent subject categories in Microsoft Academic and the lack of a document type field. Based on the publications of three researchers in Scientometrics 2010-2014, if found that field normalised indicators gave reasonable values, despite the limitations discussed (Hug, Ochsner, & Brändle, 2017).

Although, as discussed above, prior research suggests that the beta version of Microsoft Academic has similar coverage to Scopus and WoS in some areas and higher coverage in others (Harzing & Alakangas, 2017a), it is not known whether it finds substantial numbers of early citations from web preprints of articles. The best current source of early impact evidence is the social reference manager Mendeley because articles published in the most recent few years tend to have much higher Mendeley reader counts than citation counts (Maflahi & Thelwall, in press; Thelwall & Sud, 2016; Thelwall, 2017a). Thus, any early citation advantage for Microsoft Academic must be considered in the context of the potential to use Mendeley Reader counts instead. The goal of this article is to assess whether Microsoft Academic obtains citation counts that are comparable to major citation indexes, and whether it is particularly useful for early citations. The following research questions drive the study. Scopus was chosen for comparison purposes since it has greater coverage of academic research than the WoS (Mongeon & Paul-Hus, 2016) and therefore its citation counts should be higher. Correlation tests were used to help assess the extent to which the different sources reflect the same type of impact (Sud & Thelwall, 2014). High correlations have been found between citations from Microsoft Academic and both Scopus and the WoS for a single university (Hug & Brändle, in press).

1. Are citation counts reported by Microsoft Academic greater in magnitude to those reported by Scopus for recently-published articles?
2. Do citation counts reported by Microsoft Academic have a significant positive correlation with the citation counts provided by Scopus for recently-published articles?
3. How do Microsoft Academic citation counts compare to Mendeley reader counts for the above questions?

Methods

The research questions were addressed with nine journals. Journals were chosen rather than fields because both Scopus and Microsoft Academic allow journal-based searching but Microsoft Academic does not allow field-based searching. Its equivalent is keyword-based

searching, which is not comparable. *Science* and *Nature* were selected as large and important science journals. Seven journals from library and information science were also selected for a more in depth analysis of a single field. A field known to the author was chosen to aid interpretation of the results. The years 1996-2017 were selected for analysis to give a long period, starting with a year in which Scopus expanded its coverage. Although the focus of this article is on recent years, a long period is used to help identify any long term trends.

Data

All documents of type journal article in the nine journals were downloaded from Scopus on 17 July 2017 using ISSN-based journal queries, such as the following for *Science* in 2001.

- ISSN(0036-8075) AND PUBYEAR IS 2001 AND DOCTYPE(ar))

This query excludes reviews, editorials and other non-research document types. The *Journal of the Association for Information Science and Technology* (JASIST) used two ISSNs during this period so both were used. The data were checked for incorrectly categorised articles, resulting in some *Nature* news articles being removed and some articles with incorrect DOIs being removed from both *Science* and *Nature*.

Documents are not categorised by type in Microsoft Academic and so documents of all types from the nine journals were downloaded from it. For this, the standardised names used for the journals by Microsoft Academic were identified by ad-hoc searches in the site (academic.research.microsoft.com) for articles and then the queries below were submitted to it via its API, using a function added to the free software Webometric Analyst (*Citations* menu, *Microsoft Academic* menu item: http://lexiurl.wlv.ac.uk). The following queries were used for the seven information science journals, including two for JASIST due to its different names.

Composite(J.JN=='journal of informetrics')
Composite(J.JN=='j amer soc inform sci')
Composite(J.JN=='jasis')
Composite(J.JN=='scientometrics')
Composite(J.JN=='ipm')
Composite(J.JN=='journal of information science')
Composite(J.JN=='j doc')
Composite(J.JN=='libr inform sci res')

Experiments showed that not all *Science* and *Nature* documents were returned by these queries because there were too many and so two date specific queries were submitted for these instead. All Microsoft Academic queries were submitted on 16 or 17 July 2017 and the citation counts (CC field) recorded for each match.

And(Composite(J.JN=='nature'),Y>2006)
And(Composite(J.JN=='science'),Y>2006)
And(Composite(J.JN=='nature'),And(Y>1995,Y<2007))
And(Composite(J.JN=='science'),And(Y>1995,Y<2007))

The Scopus results were submitted to Mendeley to obtain reader counts in 17 July 2017 using its Applications Programming Interface (API) in Webometric Analyst, matching documents by DOI and by title/author/years and combining the results for the greatest coverage (Zahedi, Haustein, & Bowman, 2014).

A combined dataset was built from the above three sets of results for each journal by matching documents based on their DOIs and discarding any that were absent from at

least one of the three sources. DOI-based matching ensures near-perfect matching, in comparison to metadata matching. Manual checks did not find any incorrect matches. Articles missing from any of the three sources were discarded because they may be present but with an incorrect or missing DOI. Incorrect DOIs were a problem for JASIST because some older article had DOIs for a journal issue (e.g., 10.1002/asi.v60:6) rather than the individual article. Some of the excluded articles would presumably be absent from some of the services rather than not found. This issue is discussed again after the results.

The above steps produced a final dataset of 44,398 documents that were classified as journal articles in Scopus, with a DOI, and with a record with a matching DOI in both Microsoft Academic and Mendeley.

Analyses

The average citation or reader count of articles were calculated for each year and journal to compare the magnitudes between the three sources. Geometric means were used because citation counts are highly skewed and arithmetic means can therefore be unduly influenced by individual outliers (Thelwall & Fairclough, 2015; Zitt, 2012). Averages were calculated separately for each year because the time needed to accrue citations and changes in database coverage give citation and readership data a temporal dimension (Larivière, Archambault, & Gingras, 2008).

Spearman correlations were calculated between the three data sources separately for each journal and year. Spearman rather than Pearson correlations were used because the data sources are skewed. The Spearman correlation assess the extent to which documents are in the same rank order between the different data sets. This broadly suggests whether changing data source could affect the results of a scientometric impact study using the different data sources.

Results

For all journals and most years, the average citation counts per document are similar for Scopus and Microsoft Academic (Figure 1). The average citation counts tend to be slightly higher for Scopus for *Nature* and *Science* (Figure 1) and slightly higher for Microsoft Academic for the seven library and information science journals. The averages tend to be most similar for recent articles and there is no tendency for Microsoft Academic to identify early citations.

The Spearman correlations between citation counts for Microsoft Academic and Scopus are almost uniformly close to 1 except for years with few documents that give unreliable results (Figure 2). Thus, the two are essentially interchangeable for citation count-based ranking purposes.

For most journals, there are fewer Mendeley readers than citations for older articles but all journals have more readers than citations for more recent articles. Mendeley is therefore a better source of data for new documents. Correlations between Mendeley reader counts and both sources of citation counts tend to be very high, suggesting that they reflect similar types of impact. These correlations are in line with previous correlations for the most recent year of the information science journals (Figure 8 of: Maflahi & Thelwall, in press) and for overall correlations for the library and Information Science Scopus category (Figure 5 of: Thelwall & Sud, 2016). The *Nature* and *Science* correlations are higher than for a previous study of them (Table 4, 5 of: Li, Thelwall, & Giustini, 2011), probably reflecting a greater current uptake and maturing of Mendeley. The lower correlations for years where

Mendeley has a different average could be due to the increased amount of data allowing a more refined rank order (e.g., if there are less uncited articles that would be tied in last place; see: Thelwall, 2016). This is not a possible explanation for the years when the Mendeley reader count average is approximately the same as the citation count averages (e.g., 2008 for *Nature*: Figure 1) and so it is clear that there are genuine, but small, underlying differences between Mendeley readers and academic citations.

The jagged patterns in the graphs for older years are mostly caused by low numbers of matching documents (Table 1). Nature (Figure 1) is an exception, with jaggedness despite large numbers of articles. This seemed to be due to changes in the types of article categorised as journal articles in some years.
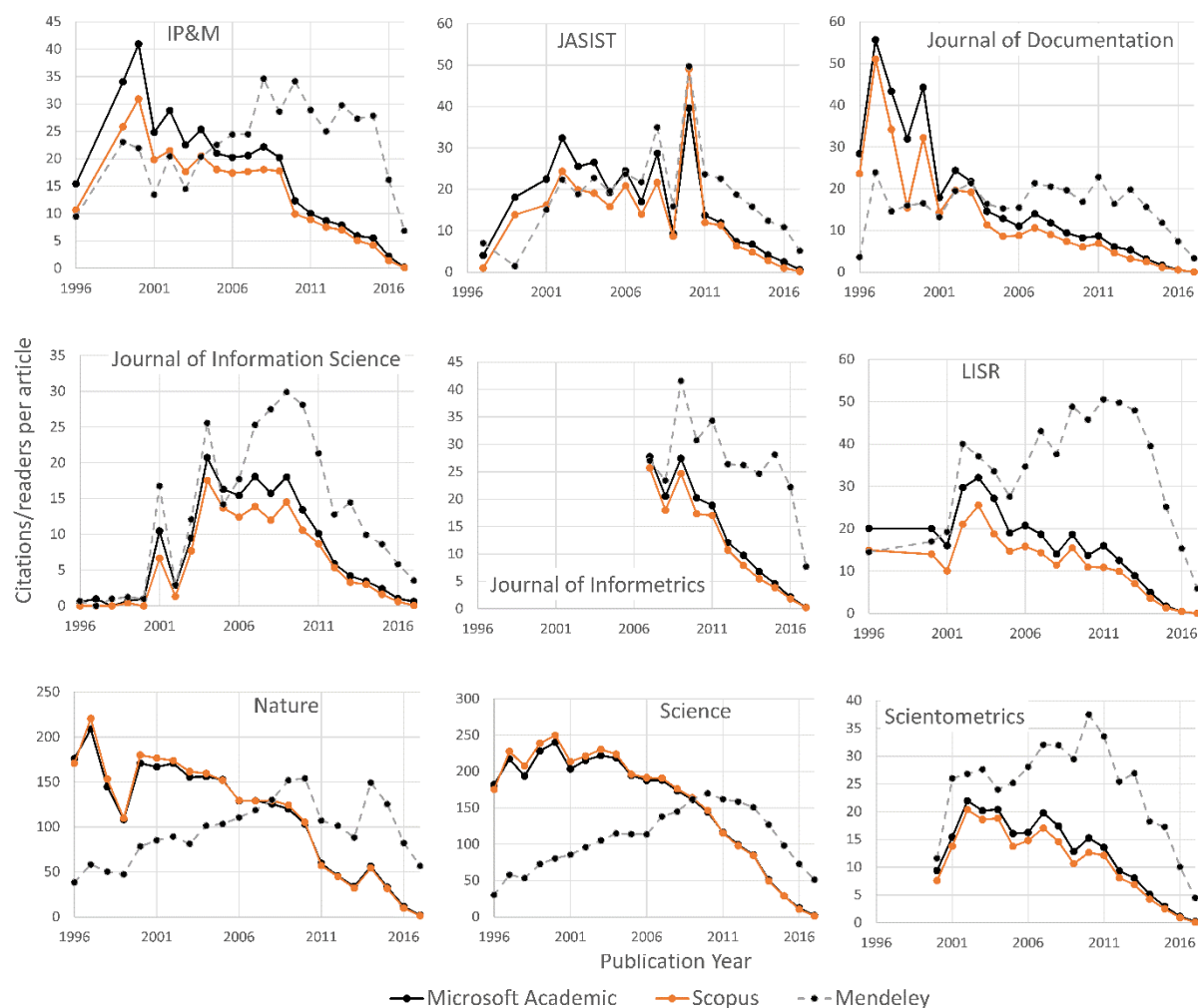

Figure 1. Geometric mean Microsoft Academic citations, Scopus citations and Mendeley readers for the nine journals analysed. Documents are only included when found and matched in all three sources; fewer documents have matching DOIs in earlier years.
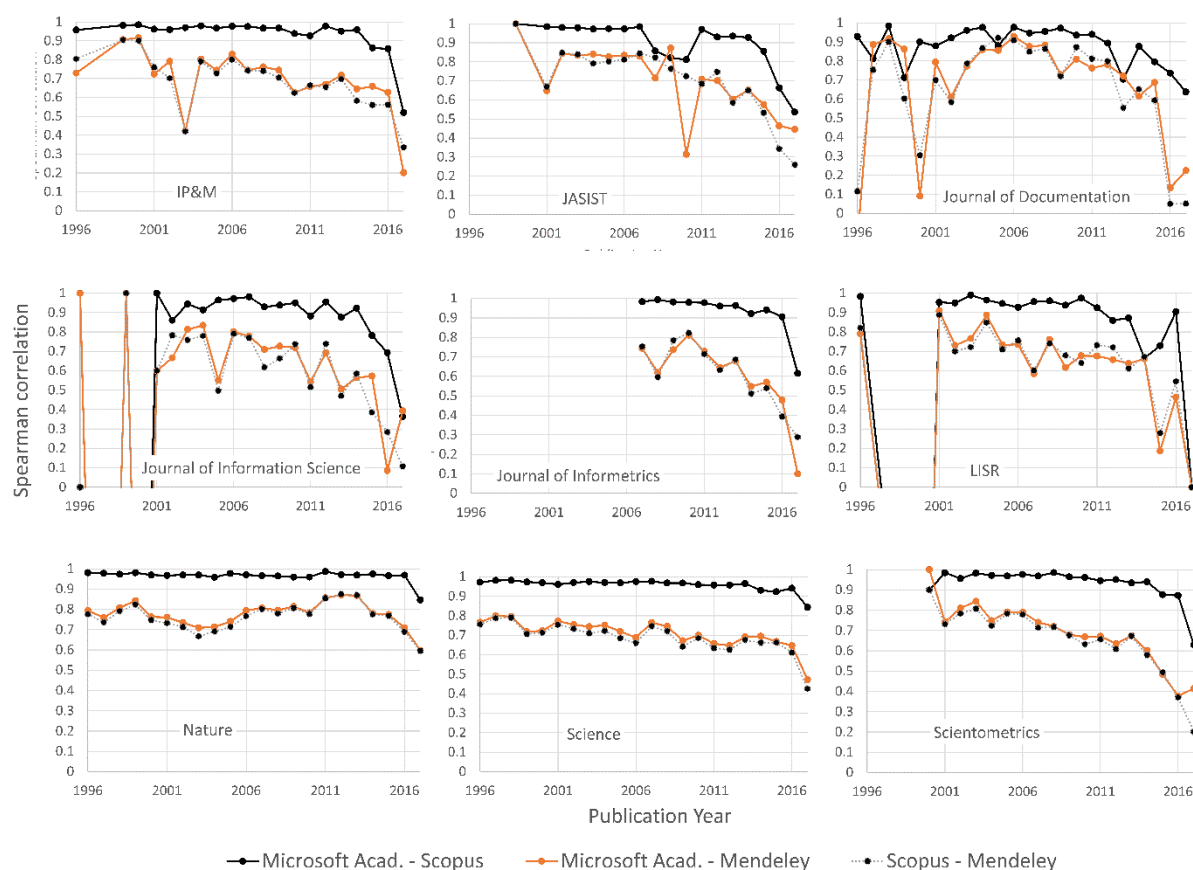
Figure 2. Spearman correlations between Microsoft Academic citations, Scopus citations and Mendeley readers; all against publication year for the nine journals analysed. Documents are only included when found and matched in all three sources; fewer documents have matching DOIs in earlier years.

Table 1. Sample statistics about the journals used in Figures 1 and 2. Years with no data are ignored in the calculations.

Journal	Min. in a year	Max in a year	Average	Total
IP&M	30	112	61	1219
JASIST	1	212	80	1524
J. Documentation	6	60	33	720
J. Information Science	1	71	30	654
J. Informetrics	31	92	61	667
Lib. & Info. Sci. Res.	1	38	25	474
Nature	440	1145	903	19860
Science	242	856	742	16323
Scientometrics	5	370	164	2957
Overall				**44398**

The slightly higher citation counts for Microsoft Academic were investigated for the journal *Scientometrics* by checking the citation counts for articles with many Microsoft Academic citations compared to Scopus citations. Scientometrics was chosen as a journal familiar to the readers of this article. As illustrated below, the results suggest that marginally higher

citation counts for Microsoft Academic may be partly due to mistakes, indexing grey literature sources and allocating citations to preprints and conference papers to the subsequent journal article.

- *What drives the relevance and reputation of economics journals? An update from a survey among economists (2015)*: This had 0 Scopus citations and 73 Microsoft Academic citations. The Microsoft Academic citations checked were incorrect or to a preprint with the same name and authors. Thus, the higher value in this case is from a mixture of mistakes and (correct) document merging. Some of the citations in this case originated from social science and economics preprint sharing archives (RePEc and SSRN).

- *Coverage and adoption of altmetrics sources in the bibliometric community (2014)*: This had 14 Scopus citations and 81 Microsoft Academic citations. The additional citations in this case mainly originated from citations to the conference paper from which this *Scientometrics* article was derived.

- *Open access scientometrics and the UK Research Assessment Exercise* (2009). This had 22 Scopus citations and 59 Microsoft Academic citations. The additional citations in this case again mainly originated from citations to the conference paper from which this *Scientometrics* article was derived.

Discussion

The most important limitation of this study is that it only covers two large general science journals and one field, library and information science. There may be substantial differences between the two citation sources in some other fields. This could include conference-based fields with much online publishing (e.g., computing) that could be found by Microsoft Academic, areas with preprint repositories (e.g., physics, economics) or subject areas with publishers that have decided not to cooperate with one of Elsevier and Microsoft for citation indexing. This paper also does not analyse the influence of occasional errors in citation indexing, which has been a problem for Microsoft Academic (Harzing & Alakangas, 2017a; Hug, Ochsner, & Brändle, 2017), but is much reduced now (Harzing & Alakangas, 2017b). The research design decision to exclude articles not found by at least one of the three sources assumes that only query matching problems cause articles to be missing (e.g., for articles that are in Mendeley but that were not found by the search process). It is likely that some of the articles that had not been found were completely missing (e.g., articles that are not in Mendeley at all) and therefore should be registered with a citation or reader count of 0. This may result in higher average citation and reader counts for all three sources, especially for Mendeley due to its reliance upon user inputting of publication records.

The slightly higher citation counts for the seven library and information science journals agree with a previous study based on the publications of 145 academics across multiple years that Microsoft Academic finds more citations than Scopus (and WoS) for Social Sciences (based on 24 academics' publications) and Humanities (19 academics) categories (Figure 3 of: Harzing & Alakangas, 2017a). The marginally lower citation counts found here for *Science* and *Nature* also agree with previous results based on individual academics that Microsoft Academic has slightly lower citation counts than Scopus (and WoS) for the Life Sciences (43 academics) and Sciences (39 academics) categories (Figure 3 of: Harzing & Alakangas, 2017a). The results also extend previous analyses by showing that they hold true when data is systematically collected from journals and over a long period. The results also show, for the first time, that there is a very high positive correlation

between Scopus citation and Microsoft Academic citations at the journal level and that Mendeley gets much higher counts for recent articles. Google Scholar finds more citations than Scopus (Falagas, Pitsouni, Malietzis, & Pappas, 2008; Harzing & Alakangas, 2016) and slightly more citations than Microsoft Academic overall (Harzing, 2016; Harzing & Alakangas, 2017ab; Hug, Ochsner, & Brändle, 2017). Nevertheless, even Google Scholar is probably inferior to Mendeley for early impact evidence for recent articles (e.g., Maflahi & Thelwall, in press). Mendeley can be spammed and so is not recommended for formal evaluations when the data source is known in advance.

The lack of early citations in Microsoft Academic is strange, given that it finds citations from academic preprints and working papers online and so should, in theory, deliver much higher citation counts than Scopus for recently-published articles. The results suggest that Microsoft Academic takes a more conservative approach than Google Scholar when identifying citations by including only a limited set of preprint sources (e.g., only recognised repositories and not preprints elsewhere) or by having more stringent criteria for judging that a document is academic so that its citations should be indexed.

Conclusions

The magnitude (Figure 1) and correlation (Figure 2) results give clear evidence that Microsoft Academic is essentially equivalent to Scopus in terms of the magnitudes of its citation counts, agreeing with a previous analysis of a university (Hug & Brändle, in press). Because it is free and, unlike Google Scholar, allows automated searching, it is an important new data source for scientometrics. Surprisingly, however, it does not give an early citation advantage, and Mendeley remains the best source of early impact evidence. Microsoft Academic is therefore not recommended for researchers that already have convenient access to Scopus or the Web of Science, except if there are subject areas in which it has greater coverage. Researchers can run automatic searches by either writing their own code to access it or using the automated search facility in Webometric Analyst (http://lexiurl.wlv.ac.uk), or Publish or Perish (version 5+) for individual scholars (Harzing, 2007).

There are several practical issues with using Microsoft Academic for scientometric purposes. If the impact of the articles produced by a department, university or country is to be assessed with field normalised indicators (e.g., Science-Metrix, 2015; Thelwall, 2017b; Waltman, van Eck, van Leeuwen, Visser, & van Raan, 2011) then a reference set is needed of documents produced in the fields and years investigated for the rest of the world. This is straightforward with Scopus and the WoS. For example, Scopus has a Library and Information Sciences category and it is simple and quick to download from Scopus the citation counts of all journal articles within this category from a given year. There is currently no way to download a reference set from Microsoft Academic because it does not include useful field categories (Hug, Ochsner, & Brändle, 2017). Whilst it would be possible to collate a set of journal articles from any given category (e.g., Library and Information Science) and download the citation counts from these from Microsoft Academic, additional work would be needed to filter out unwanted document types, such as review articles or editorials. If these two categorisation issues can be fixed in the future, and Microsoft Academic's former indexing errors (e.g., for title and publication years) are confirmed to be minor and disappearing (Harzing & Alakangas, 2017b), then it would greatly simplify the task of using Microsoft Academic for scientometric research and evaluations. It also seems technically possible for Microsoft Academic to identify more early citations from online

preprints and if its designers can do this then its value for scientometrics would be greatly enhanced.

Although Microsoft Academic does not seem to find substantially more early citations than Scopus, the current study adds to the growing body of evidence showing Microsoft Academic to be a robust source of free citation data.

References

Carlson, S. (2006). Challenging Google, Microsoft unveils a search tool for scholarly articles. Chronicle of Higher Education. 52 (33), A43. https://eric.ed.gov/?id=EJ773667

Chen, A. (2017). Academic Knowledge API. https://docs.microsoft.com/en-us/azure/cognitive-services/academic-knowledge/home

Falagas, M. E., Pitsouni, E. I., Malietzis, G. A., & Pappas, G. (2008). Comparison of PubMed, Scopus, Web of Science, and Google Scholar: strengths and weaknesses. The FASEB journal, 22(2), 338-342.

Harzing, A. W., & Alakangas, S. (2016). Google Scholar, Scopus and the Web of Science: a longitudinal and cross-disciplinary comparison. Scientometrics, 106(2), 787-804.

Harzing, A. W., & Alakangas, S. (2017a). Microsoft Academic: Is the phoenix getting wings? Scientometrics, 110(1), 371-383.

Harzing, A. W., & Alakangas, S. (2017b). Microsoft Academic is one year old: The Phoenix is ready to leave the nest. Scientometrics, 112(3), 1887-1894.

Harzing, A.W. (2007). Publish or Perish. http://www.harzing.com/pop.htm

Harzing, A. W. (2014). A longitudinal study of Google Scholar coverage between 2012 and 2013. Scientometrics, 98(1), 565-575.

Harzing, A. W. (2016). Microsoft Academic (Search): A Phoenix arisen from the ashes? Scientometrics, 108(3), 1637-1647.

Hug, S. E., & Brändle, M. P. (in press). The coverage of Microsoft Academic: Analyzing the publication output of a university. Scientometrics. doi:10.1007/s11192-017-2535-3

Hug, S. E., Ochsner, M., & Brändle, M. P. (2017). Citation analysis with Microsoft Academic. Scientometrics, 111(1), 371-378.

Jacsó, P. (2011). The pros and cons of Microsoft Academic Search from a bibliometric perspective. Online Information Review, 35(6), 983-997.

Larivière, V., Archambault, É., & Gingras, Y. (2008). Long-term variations in the aging of scientific literature: From exponential growth to steady-state science (1900–2004). Journal of the Association for Information Science and Technology, 59(2), 288-296.

Li, X., Thelwall, M., & Giustini, D. (2011). Validating online reference managers for scholarly impact measurement. Scientometrics, 91(2), 461-471.

Maflahi, N, & Thelwall, M. (in press). How quickly do publications get read? The evolution of Mendeley reader counts for new articles. Journal of the Association for Information Science and Technology. doi:10.1002/asi.23909

Microsoft (2017a). Why a new site? https://web.archive.org/web/20170105184616/https://academic.microsoft.com/FAQ

Microsoft (2017b). What's new in version 2.0? https://academic.microsoft.com/#/faq (accessed 1 August 2017)

Moed, H. F., Bar-Ilan, J., & Halevi, G. (2016). A new methodology for comparing Google Scholar and Scopus. Journal of Informetrics, 10(2), 533-551.

Mongeon, P., & Paul-Hus, A. (2016). The journal coverage of Web of Science and Scopus: a comparative analysis. Scientometrics, 106(1), 213-228.

Orduña-Malea, E., Martín-Martín, A., M. Ayllon, J., & Delgado Lopez-Cozar, E. (2014). The silent fading of an academic search engine: the case of Microsoft Academic Search. Online Information Review, 38(7), 936-953.

Ortega, J. L., & Aguillo, I. F. (2014). Microsoft Academic Search and Google Scholar citations: Comparative analysis of author profiles. Journal of the Association for Information Science and Technology, 65(6), 1149-1156.

Science-Metrix (2015). Analysis of Bibliometric indicators for European policies 2000–2013. http://ec.europa.eu/research/innovationunion/pdf/bibliometric_indicators_for_europ ean_policies.pdf

Sinha, A., Shen, Z., Song, Y., Ma, H., Eide, D., Hsu, B. J. P., & Wang, K. (2015). An overview of Microsoft Academic service (mas) and applications. In Proceedings of the 24th international conference on world wide web (pp. 243-246). New York, NY: ACM Press.

Sud, P. & Thelwall, M. (2014). Evaluating altmetrics. Scientometrics, 98(2), 1131-1143. 10.1007/s11192-013-1117-2.

Thelwall, M. & Fairclough, R. (2015). Geometric journal impact factors correcting for individual highly cited articles. Journal of Informetrics, 9(2),263–272.

Thelwall, M. & Sud, P. (2016). Mendeley readership counts: An investigation of temporal and disciplinary differences. Journal of the Association for Information Science and Technology, 57(6), 3036-3050. doi:10.1002/asi.2355

Thelwall, M. (2016). Interpreting correlations between citation counts and other indicators. Scientometrics, 108(1), 337-347.

Thelwall, M. (2017a). Are Mendeley reader counts high enough for research evaluations when articles are published? Aslib Journal of Information Management, 69(2), 174-183. doi:10.1108/AJIM-01-2017-0028

Thelwall, M. (2017b). Three practical field normalised alternative indicator formulae for research evaluation. Journal of Informetrics, 11(1), 128–151. 10.1016/j.joi.2016.12.002

Waltman, L., van Eck, N. J., van Leeuwen, T. N., Visser, M. S., & van Raan, A. F. (2011). Towards a new crown indicator: An empirical analysis. Scientometrics, 87(3), 467-481.

Wang, J. (2013). Citation time window choice for research impact evaluation. Scientometrics, 94(3), 851-872.

Zahedi, Z., Haustein, S. & Bowman, T (2014). Exploring data quality and retrieval strategies for Mendeley reader counts. Presentation at SIGMET Metrics 2014 workshop, 5 November 2014. Available: http://www.slideshare.net/StefanieHaustein/sigmetworkshop-asist2014

Zitt, M. (2012). The journal impact factor: Angel, devil, or scapegoat? A comment on JK Vanclay's article 2011. Scientometrics, 92(2), 485-503.