Microsoft Academic is one year old: the Phoenix is ready to leave the nest

ANNE-WIL HARZING
Middlesex University
The Burroughs, Hendon, London NW4 4BT
Email: anne@harzing.com

SATU ALAKANGAS
University of Melbourne
Parkville Campus, Parkville VIC 3010, Australia

Abstract

We investigate the coverage of Microsoft Academic (MA) just over a year after its re-launch. First, we provide a detailed comparison for the first author's record across the four major data sources: Google Scholar (GS), MA, Scopus and Web of Science (WoS) and show that for the most important academic publications, journal articles and books, GS and MA display very similar publication and citation coverage, leaving both Scopus and WoS far behind, especially in terms of citation counts.

A second, large scale, comparison for 145 academics across the five main disciplinary areas confirms that citation coverage for GS and MA is quite similar for four of the five disciplines. MA citation coverage in the Humanities is still substantially lower than GS coverage, reflecting MA's lower coverage of non-journal publications. However, we shouldn't forget that MA coverage for the Humanities still dwarfs coverage for this discipline in Scopus and WoS.

It would be desirable for other researchers to verify our findings with different samples before drawing a definitive conclusion about MA coverage. However, based on our current findings we suggest that, only one year after its re-launch, MA is rapidly become the data source of choice; it appears to be combining the comprehensive coverage across disciplines, displayed by GS, with the more structured approach to data presentation, typical of Scopus and WoS. The phoenix seems to be ready to leave the nest, all set to start its life into an adulthood of research evaluation.

Microsoft Academic is one year old: the Phoenix is ready to leave the nest

Introduction

Just over a year ago, we conducted the first study of Microsoft Academic (MA) coverage (Harzing, 2016). We showed that, for the first author's publication record, MA outperformed both Web of Science (WoS) and Scopus in terms of publication coverage and citation counts. Like Google Scholar (GS), MA found all of the academic's journal articles and books; it did not, however, match GS's coverage for book chapters, conference papers and other publications. MA's citation counts were also lower than GS citation counts. Just over half a year ago, we expanded this analysis (Harzing & Alakangas, 2017) and showed that this general conclusion was also valid for a sample of 145 academics across five disciplines. The only other study to date on MA coverage (Hug & Brändle, 2017), based on title searches for the 2008-2015 publications of an entire university, found Scopus coverage for journal articles to be marginally better than MA coverage, with both data sources outperforming the WoS. MA, however, showed the highest proportion of unique coverage for journal articles. It also significantly outperformed the two other data sources in all other document types.

In this short letter, we investigate whether MA has sustained its advantage over the commercial data sources and whether it has made any further headway in comparison to its non-commercial rival. We do so by combining the approach of our two earlier articles. First, we provide a detailed comparison of the first author's publication and citation record across the four data sources. Second, we compare MA and GS for a sample of 145 academics across five main disciplines. For full details and a justification of these two samples, please refer to Harzing (2016) and Harzing and Alakangas (2016, 2017). All data were collected in the first week of June 2017. Searches for MA and GS were run in Publish or Perish (PoP) (Harzing, 2007). We used a Google Scholar Profile search, newly available in PoP version 5, for those academics that had established such a profile[1] and a regular GS search for all other academics. Searches for WoS and Scopus were conducted in their native interfaces, exported and subsequently imported into PoP.[2]

Results

Before turning to our data source comparisons, we first verify whether the MA's teething problems, as highlighted in Harzing (2016), had been resolved. First, although MA still reports so-called stray publications or sometimes attributes chapters in an edited book to the editor, with 30 stray/inaccurate publications, largely without citations, out of a total of 131, this problem is substantially less prominent than in GS. In the regular GS search, we find around 350 results for the first author, nearly two thirds of which were stray or inaccurate publications. As such, especially for academics without a well-maintained GS Profile, a MA search is likely to provide a much "cleaner" result than a GS search.

Second, in 2016 MA had considerable problems in parsing titles with a main and sub-title separated by a semi-colon; it reported two versions – one with and one without subtitle – with citations split between them. This problem has all but disappeared. In the rare cases where it still occurs, it concerns articles where the split version has no citations (and hence can simply be disregarded as a stray publication). A third, quite serious, problem reported in 2016 concerned incorrect year allocations. No less than 18 out of the first author's 89 publications in MA carried the wrong publication year, in some cases the year was "way out" (sometimes more than ten years). Currently, there are only two (out of 100) publications with the "wrong" publication year. In both cases, MA parsed the earlier online-first year rather than the print publication year[3], something that still happens on a very regular basis in GS.

[1] Half of the academics in our sample had created a GS profile. This varied from one third for the Life Sciences, to half for the Humanities and the Sciences, and two thirds for the Social Sciences and Engineering.

[2] We used the basic/general search option for WoS and Scopus rather than the "cited reference search". The five key reasons for this choice are detailed in Harzing (2013). Briefly, both WoS and Scopus have more stray citations than either GS or MA, their cited reference search is very time-consuming and unwieldy to use, and it doesn't allow merging, sorting, exporting, or any further analysis of the data.

[3] Obviously one could argue that this is in fact the correct publication year; it is the year the publication first became available.

Detailed comparison of publications across four data sources

The first author's publication record includes 78 journal articles, three books, seventeen book chapters, a software program, and an online compilation of journal rankings. It also includes more than 100 conference papers and more than 100 other publications, such as white papers, newsletter/magazine articles, and blog posts. The conference papers are by and large not available online, however, and the other publications are not generally recognised as academic publications. Hence we would not expect substantive coverage of these two publication categories in any of the data sources.

As Table 1 shows, both GS and MA record all of the author's journal articles and books. Scopus does not record any of the books, but, in contrast to our earlier study (Harzing, 2016), *does* record nearly all (96%) of the journal articles. Whereas a year ago, Scopus missed 13 articles, this number has now been reduced to only three, one of which is a very recent article, not yet available in online first, but captured by GS and MA through the Middlesex University Research Repository. The ten newly covered articles are most likely a result of the Scopus expansion program, which finished late 2016 and included adding back volumes from 36 major publishers (Elsevier, 2016). The WoS performs much more poorly, recording none of the books and only 55 out of the 78 journal articles.

Table 1: Publication coverage across four data sources

Data source/ Document type	Journal Articles	Books	Chapters	Conference papers	Other publications	Software / Data	Total (excl. conf. & other)
All publications	78	3	17	100+	100+	2	300+ (100)
Google Scholar	78	3	17	14	13	2	127 (100)
Microsoft Academic	78	3	5	10	4	1	101 (87)
Scopus	75	0	1	2	0	0	78 (76)
Web of Science	55	0	1	0	0	0	57 (57)

In terms of the remaining publications, GS still has an edge over MA, covering far more book chapters and other publications, although MA does cover 10 out of the 14 conference papers listed in GS. Further, even though it is still missing the Journal Quality List, MA does record the Publish or Perish software. Scopus and the WoS have negligible coverage of non-journal publications; both report one book chapter and Scopus lists two conference papers.

Detailed comparison of citations across four data sources

PoP version 5 uses MA's estimated citation counts, reported as default in the MA web interface since July/August 2016, rather than the previously reported linked citation counts. For a detailed discussion of how MA estimates these counts, see Harzing and Alakangas (2017). Table 2 shows that for the combined 78 journal articles MA citations are only 2% lower than GS citations. However, even for books and software/data MA estimated citation counts are quite close to GS counts.

Table 2: Citation coverage across four data sources

Data source/ Document type	Journal Articles	Books	Chapters	Conference papers	Other publications	Software / Data	Total
Google Scholar	9842	1118	528	127	121	709	12445
Microsoft Academic	9600	984	35	31	17	620	11287
Scopus	3805	0	1	2	0	0	3808
Web of Science	2323	0	0	0	0	0	2323

Reflecting its much lower coverage of book chapters and other publications, MA citations in these categories are only a fraction of GS citations. Finally, despite its relatively good coverage of conference papers, citations for this publication type in MA are low, largely caused by the fact that the conference papers that were most cited in GS are not included in MA.

Citation counts in Scopus and WoS are substantially lower than in both GS and MA, in total Scopus records around a third of GS/MA citations, whereas WoS only records around a fifth. Even if we compare only journals articles, Scopus only records around 40% of GS/MA citations, whereas WoS records less than a quarter of GS/MA citations. It is clear that for this particular Social Science academic, the non-commercial data sources display a substantially better coverage.

Detailed comparison of metrics across four data sources

Table 3 illustrates how key metrics are affected by the use of different data sources. It shows again that MA and GS provide very similar metrics, with the largest difference occurring for the number of publications with more than 10 citations per year (39 vs. 45). In contrast, both Scopus and WoS provide much lower metrics than MA and GS, especially in the area of yearly citations and the number of articles with more than 10 citations per year. The hI,annual - an annual individual h-index (see Harzing, Alakangas & Adams, 2014) - shows the lowest variance across data sources, partly because WoS misses coverage of the academic's older articles, thus reducing the number of years since first publication, the denominator in this metric. Even so, this metric is substantially higher in both GS and MA.

Table 3: Key metrics across four data sources

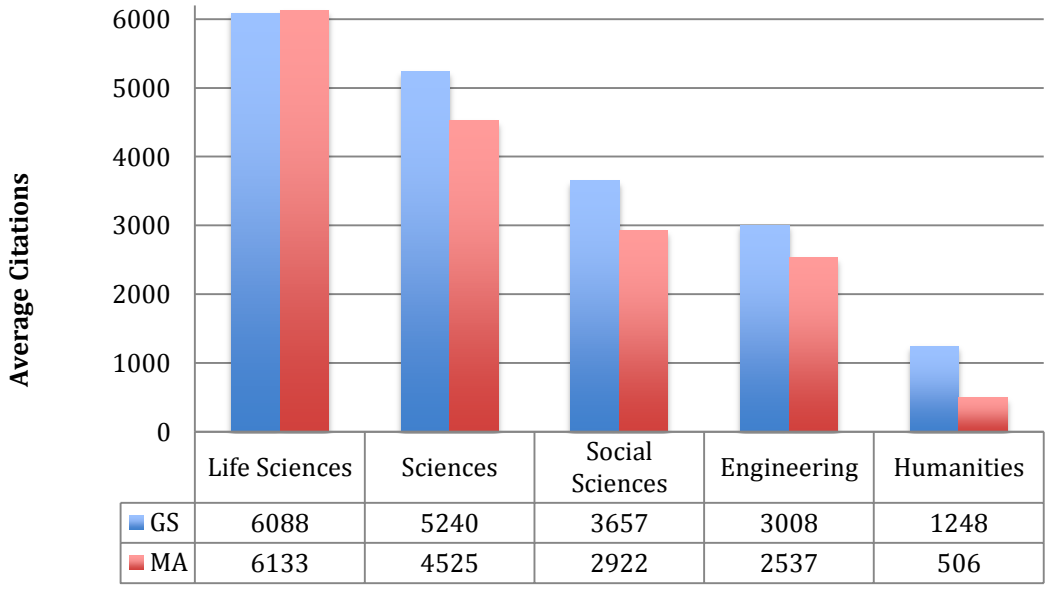
Data source/ Metric	Cites/ year	h-index	g-index	hI,norm	hI,annual	>10 cites/ per year
Google Scholar	566	52	111	42	1.91	45
Microsoft Academic	513	48	104	39	1.77	39
Scopus	173	31	61	24	1.09	17
Web of Science	137	25	48	18	1.06	10

Disciplinary comparison between GS and MA

As the result for an individual academic's publication record might be idiosyncratic, we also provide a high-level overview for a sample of 145 academics across five main disciplines: Life Sciences, Sciences, Social Sciences, Engineering, and Humanities. As Harzing and Alakangas (2017) showed that MA outperformed both WoS and Scopus and that data collection for these commercial data source is infinitely more time-consuming than for MA and GS, we focus our comparison on MA and GS only. MA and GS publication counts include many stray publications and hence necessitate time-consuming merging and data cleaning to achieve full accuracy. We therefore focus our comparison on total citations and one of the metrics - the hI,annual – only as these are not influenced by stray publications.

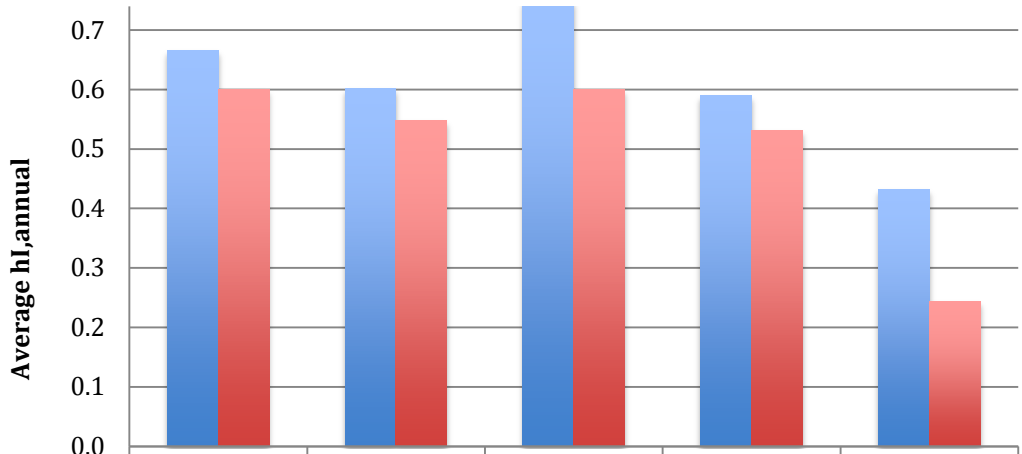
Figure 1 shows that citation counts for MA are roughly identical to GS citation counts for the Life Sciences and are 14-20% lower for the Sciences, Engineering and Social Sciences. In the Humanities they are nearly 60% lower. Comparing our results with those of Harzing and Alakangas (2017), collected 7 months earlier, we find that MA citation counts have declined slightly for the Life Sciences (-/- 0.6%) and Sciences (-/- 4.6%), whereas GS citation counts for these disciplines increased by 8-10%. The decline for the Sciences was largely caused by two authors that were conflated with namesakes (resulting in inflated counts) in the 2016 data; with these corrected the decline was only 1.6%.

Figure 1: Comparison of citation counts across disciplines for Google Scholar and Microsoft Academic



Further investigation showed that MA citation counts for around half of the academics in the Life Sciences and Sciences had declined[4], whereas citation counts for the remaining academics in these disciplines had increased by 7-11%, i.e. at a level similar to GS. MA estimated citations counts for Engineering and the Social Sciences increased by 13-17%. For Engineering, this was quite similar to its increase in GS citations (15%), for the Social Sciences the MA increase exceeded the GS increase (12%). MA's biggest relative gain for was made for the Humanities, which – although outperforming the discipline's abysmal record in Scopus and WoS – still shows the poorest relative performance in MA; citation counts for this discipline increased by 50% since our 2016 data collection.

Figure 2: Comparison of the hI,annual across disciplines for Google Scholar and Microsoft Academic



	Life Sciences	Sciences	Social Sciences	Engineering	Humanities
GS	0.67	0.60	0.75	0.59	0.43
MA	0.60	0.55	0.60	0.53	0.24

Figure 2 shows that a comparison of the hI,annual displays a similar picture. MA metrics are 9-10% lower than GS metrics for the Life Sciences, Sciences and Engineering, 20% for the Social Sciences and 44% for the Humanities. As we found in our comparison using the 2013 Scopus hI,annual for the same data set (Harzing, Alakangas & Adams, 2014), the 2017 GS and MA hI,annual are not significantly different across four of the five disciplines, thus again confirming the metric's relevance for cross-disciplinary comparisons. It should be noted that, especially for the Life Sciences, the lower hI,annual for MA is partly based on its more complete reporting of authors. Because of problems with author truncation in GS searches, the actual number of authors is likely to be underestimated.[5] When comparing the average number of authors for MA and GS, we found them to be 10-14% higher in the Humanities, Social Sciences and Engineering, 20% higher in the Sciences and 54% higher in the Life Sciences.

Conclusions

The aim of this letter was to investigate the coverage of MA just over a year after its re-launch. We first provided a detailed comparison of the publication and citation coverage for an individual academic across the four major data sources: GS, MA, Scopus and WoS. We showed that for the most important academic publications, journal articles and books, GS and MA displayed very similar publication and citation coverage, leaving both Scopus and WoS far behind, especially in terms of citation counts.

[4] Consultation with the MA team suggests that the earlier (November 2016) citation counts for the Life Sciences and Sciences were likely to have been inflated, because of parsing errors for titles with chemical compounds, DNA, protein names or non-Roman characters. These errors, now largely addressed with improved paper-matching algorithms, meant that these papers and their citations were double-counted. Our subsequent detailed analysis of citation levels for academics in (Life) Sciences indeed showed that those sub-disciplines likely to contain problematic titles, such as Biochemistry, Genetics, Microbiology, Neuroscience, and Pathology showed a decline in citations, whereas those in Audiology, Mathematics, Physics, Population Health, Veterinary Science, and Zoology showed increases. In general though, citation counts in the Life Sciences, and to a lesser extent the Sciences, are likely to be closer to GS counts as these disciplines have fewer non-journal publications.

[5] The PoP GS Profile search doesn't suffer from this same problem, but as indicated above only half of the academics in our sample and only one third of the Life Science academics had created a profile.

A second, large scale, comparison for 145 academics across the five main disciplinary areas confirmed that citation coverage for GS and MA is quite similar for four of the five disciplines, resulting in comparable hI,annual metrics for these disciplines across the two data sources. MA citation coverage in the Humanities is still substantially lower than GS coverage, illustrating MA's lower coverage of non-journal publications. However, we shouldn't forget that MA coverage for the Humanities still dwarfs coverage for this discipline in Scopus and the WoS.

We found that the teething problems in MA with regard to title splits and incorrect year allocations had been resolved. As indicated above, MA also shows much cleaner search findings than GS and allows API access to its data, allowing for easier and quicker searches. We were able to conduct the MA searches for 145 academics in less than 10 minutes, whereas – due to the necessary delays between queries – this took several hours for GS. Finally, MA doesn't suffer from the author and journal title truncation problems that are experienced for regular GS searches in PoP and thus provides more reliable authors counts and more complete bibliographic details. There are still occasional problems with conflated authors and missing publications; however, this issue is likely to improve as MA now offers author profiles that can be maintained by the academic in question.

It would be desirable for other researchers to verify our findings with different samples before drawing a definitive conclusion about MA coverage. However, based on our current findings we suggest that, only one year after its re-launch, MA is rapidly become the data source of choice; it appears to be combining the comprehensive coverage across disciplines, displayed by GS, with the more structured approach to data presentation, typical of Scopus and WoS. The phoenix seems to be ready to leave the nest, all set to start its life into an adulthood of research evaluation.

References

Elsevier (2016) *Scopus: Content Coverage Guide*. https://www.elsevier.com/__data/assets/pdf_file/0007/69451/scopus_content_coverage_guide.pdf. Accessed 6 June 2017.

Harzing, A. W. (2007). *Publish or Perish*. https://harzing.com/resources/publish-or-perish

Harzing, A.W. (2013). A preliminary test of Google Scholar as a source for citation data: A longitudinal study of Nobel Prize winners, *Scientometrics*, vol. 93, no. 3, pp. 1057-1075.

Harzing, A.W. (2016). Microsoft Academic (Search): A Phoenix arisen from the ashes?, *Scientometrics,* vol. 108, no. 3, pp. 1637-1647.

Harzing, A.W., & Alakangas, S. (2016). Google Scholar, Scopus and the Web of Science: A longitudinal and cross-disciplinary comparison, *Scientometrics*, vol. 106, no. 2, pp. 787-804.

Harzing, A.W., & Alakangas, S. (2017). Microsoft Academic: Is the Phoenix getting wings?, *Scientometrics,* vol. 110, no. 1, pp. 371-383.

Harzing, A.W., Alakangas, S., & Adams, D. (2014). hIa: An individual annual h-index to accommodate disciplinary and career length differences, *Scientometrics*, vol. 99, no. 3, pp. 811-821.

Hug, S.E., & Brändle, M.P. (2017). The coverage of Microsoft Academic: Analyzing the publication output of a university. *arXiv preprint arXiv:1703.05539*.