STI 2018 Leiden

23rd International Conference on Science and Technology Indicators
"Science, Technology and Innovation Indicators in Transition"

STI 2018 Conference Proceedings

Proceedings of the 23rd International Conference on Science and Technology Indicators

All papers published in this conference proceedings have been peer reviewed through a peer review process administered by the proceedings Editors. Reviews were conducted by expert referees to the professional and scientific standards expected of a conference proceedings.

Chair of the Conference

Paul Wouters

Scientific Editors

Rodrigo Costas
Thomas Franssen
Alfredo Yegros-Yegros

Layout

Andrea Reyes Elizondo
Suze van der Luijt-Jansen

The articles of this collection can be accessed at https://hdl.handle.net/1887/64521

ISBN: 978-90-9031204-0

23rd International Conference on Science and Technology Indicators (STI 2018)
"Science, Technology and Innovation indicators in transition"
12 - 14 September 2018 | Leiden, The Netherlands
#STI18LDN

Accuracy of affiliation information in Microsoft Academic: Implications for institutional level research evaluation

Bijan Ranjbar-Sahraei[*], Nees Jan van Eck[*] and Rutger de Jong[**]

[*] *b.ranjbarsahraei@cwts.leidenuniv.nl; ecknjpvan@cwts.leidenuniv.nl*
Centre for Science and Technology Studies, Leiden University, PO box 905, Leiden, 2300 AX (The Netherlands)

[**] *r.m.de.jong@library.leidenuniv.nl*
Science Library, Leiden University, PO box 9502, Leiden, 2300 RA (The Netherlands)

Introduction

Since the introduction of Microsoft Academic (MA), the successor of Microsoft Academic Search, different researchers have studied its coverage and data quality. Harzing (2016) suggested that MA might be a "Phoenix arisen from the ashes". Later, Harzing and Alakangaz (2017) suggested that "MA Phoenix is undeniably growing wings" and it might provide an excellent alternative for citation analysis. Hug and Brändle (2017) also suggested that "MA is on the verge of becoming a bibliometric superpower."

In this work, we study the accuracy of affiliation information in MA. To conduct this study, we have considered the full set of publications assigned to Leiden University (LU)[1] as provided by two different data sources: MA and Web of Science (WoS). The results of this study suggest that a considerable number of publications in MA have missing or wrong affiliation information.

Data

We used the Academic Knowledge API[2] to collect 131,868 publications from MA that are published in the period 1980-2017 and authored by LU researchers according to MA. We refer to this dataset as MA(LU). Additionally, a set of 110,133 publications indexed in WoS was used. This set contains publications published in the period 1980-2017 and authored by researchers affiliated with LU according to the 2017 edition of the CWTS Leiden Ranking[3] (Waltman et al., 2012). We refer to this dataset as WoS(LU).

Method

To study the accuracy of affiliation information in MA, we matched the publications in the MA(LU) and WOS(LU) datasets. The matching was performed in two steps:

1. Publications with the same title (after removing non-alpha-numeric characters) and a publication year difference of at most one year were matched. This rule has been reported by Thelwall (2018) to have high accuracy.

[1] https://www.universiteitleiden.nl
[2] https://docs.microsoft.com/en-us/azure/cognitive-services/academic-knowledge/home
[3] http://www.leidenranking.com

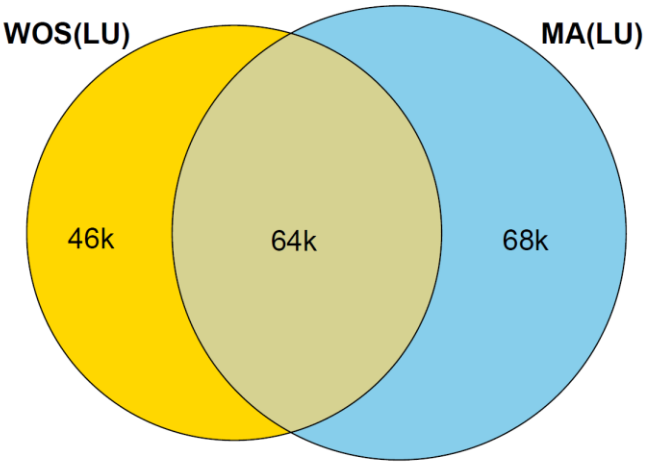
2. For the publications that were not matched in the first step, publications with identical meta data for the volume number, the first page number, and the last name of first author plus a publication year difference of at most one year were considered as a match. Van Eck and Waltman (2017) used a similar rule for matching publications.

To avoid ambiguous matches, we ignored the matches that were responsible for linking publications from one dataset to multiple publications in the other dataset.

Results

The sets of publications that are assigned to LU by MA and WoS are clearly different: despite an overlap of 64,239 publications, 67,629 publications in MA(LU) do not have a match in WoS(LU), and 45,894 publications in WoS(LU) do not have a match in MA(LU) (Figure 1). Next, we will discuss each of these three subsets in more detail.

Figure 1. Comparison of the publications in MA(LU) and WoS(LU).



Publications with an MA-WoS match

The majority of the common publications are journal articles: 84% are of document type ‘article’ and published in WoS indexed journals. Since both data sources indicate that these publications contain an LU affiliation, we did not examine this subset further.

Publications in MA without a match in WoS

Table 1 shows the results of a manual validation of 100 randomly selected MA(LU) publications without a match in WoS(LU). The manual inspection of the PDF of the publications showed that in 29 cases the LU affiliation information was not mentioned anywhere in the PDF. For the majority of these cases, affiliation information of other Dutch universities such as Utrecht University or University of Amsterdam was found in the PDF. It is unclear why MA provides wrong affiliation information for these publications.

Table 1. Results of the manual validation of a sample of 100 MA(LU) publications without a match in WoS(LU).

MA affiliation information	**No. of pub.**	**Observation**
incorrect	29	LU affiliation is not mentioned in the PDF.
correct	41	LU affiliation is mentioned in the PDF and the publication does not exist in WoS.
	23	LU affiliation is mentioned in the PDF and the publication exists in WoS(LU), but the match is not detected.
unknown	7	PDF was not accessible

Publications in WoS without a match in MA

Table 2 shows the results of a manual validation of 100 randomly selected WoS(LU) publications without a match in MA(LU). In 40 cases the LU affiliation information was found in the PDF and the corresponding publication could be retrieved using the MA web interface[4]. However, none of the authors were affiliated to LU according to MA. In some cases, the affiliation of the first non-LU author was assigned to all other authors. In some other cases, the second affiliation of the author(s) was ignored by MA. We have also seen cases were the affiliation information of all authors was simply missing. No fixed pattern, however, was found that explains the source of these errors.

Table 2. Results of the manual validation of a sample of 100 WoS(LU) publications without a match in MA(LU).

MA affiliation information	**No. of pub.**	**Observation**
incorrect	40	LU affiliation is mentioned in the PDF and the publication exists in MA, but LU affiliation is not provided by MA.
correct	16	LU affiliation is mentioned in the PDF and the publication exists in MA(LU) with correct affiliation, but the match is not detected.
-	44	LU affiliation is mentioned in the PDF but the publication does not exist in MA.

Conclusion

Out of the 200 publications that were manually checked, 69 publications had missing or wrong affiliation information in MA. This is an indication that considerable number of publications in MA have unreliable affiliation information. Therefore, institutional level research evaluation based on MA data should be performed only when proper attention is paid to this limitation.

References

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

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

Hug, S. E., & Brändle, M. P. (2017). The coverage of Microsoft Academic: Analyzing the publication output of a university. *Scientometrics*, *113*(3), 1551-1571.

Thelwall, M. (2018). Microsoft Academic automatic document searches: accuracy for journal articles and suitability for citation analysis. *Journal of Informetrics*, *12*(1), 1-9.

Van Eck, N. J. & Waltman, L. (2017). Accuracy of citation data in Web of Science and Scopus. *Proceedings of the 16th International Conference of the International Society for Scientometrics and Informetrics*, 1087-1092.

Waltman, L., Calero-Medina, C., Kosten, J., Noyons, E., Tijssen, R. J., Van Eck, N. J., ... & Wouters, P. (2012). The Leiden Ranking 2011/2012: Data collection, indicators, and interpretation. *Journal of the Association for Information Science and Technology*, *63*(12), 2419-2432.

[4] https://academic.microsoft.com (accessed April 2018)