BRIEF COMMUNICATION

Call for Caution in the Use of Bibliometric Data

Erwin Krauskopf
Facultad de Ciencias Biologicas, Universidad Andres Bello, Santiago, Chile and Fundacion Ciencia & Vida, Zañartu 1482, Santiago, Chile. E-mail: erwin.krauskopf@unab.cl

Introduction

Seventy years ago, Eugene Garfield published an article proposing the creation of a novel bibliographic metric, the citation index, based on the use of a numerical code that identified each journal and article (Garfield, 1955). One of the main advantages of the citation index was the possibility of assessing a particular work and its impact on a field, allowing users to track the influence of their own work. This index preceded the creation of a multidisciplinary citation index by the Institute for Scientific Information (later acquired by Thomson Reuters) known as the *Science Citation Index* during the early 1960s. Thomson Reuters manages the Web of Science database (WoS), which additionally includes the *Social Sciences Citation Index*, the *Arts & Humanities Citation Index*, and the *Conference Proceedings Citation Index*, among others. It is important to note that Thomson Reuters recently announced a definite agreement to sell its intellectual property and science business to private equity funds (Marketwatch, 2016). Subsequently, the Elsevier Publishing Company introduced Scopus in 2004, a database that also offered citation searching through a wide range of fields such as science, technology, medicine, social sciences, and the arts & humanities, in addition to patent search.

Both databases serve multiple purposes, as their browsing and searching options are not only useful to researchers interested in a specific topic, but to university administrators and government authorities, who constantly need to be making decisions such as the allocation of research funds or the evaluation of university faculties and departments. An unforeseen consequence of these databases was the development of several ranking systems that have increased institutional visibility worldwide. Indeed, most of the existing global ranking systems, such as the Times Higher Education World University Rankings, QS World University Rankings, Scimago Institutions Rankings, and the CWTS Leiden Ranking, use bibliometric data from one or other of these databases. All these ranking systems provide a set of indicators, some of which are publication-based (Claasen, 2015).

While studying the information published by a Chilean university ranking, I came across some data that seemed flawed, and needed to be confirmed. Thus, I began collecting bibliometric data from Scopus (from 2012–2014) for all Chilean universities using the country as the search query and then generating a list of institutions with the number of documents published by each of them for every year. To most readers, the task just described should be very common and straightforward. However, and to my surprise, when the search strategy was modified by querying directly for the Chilean institution one-by-one, the output was different. The same process was repeated using the WoS with a similar outcome.

One expects that the information provided is accurate, but it seems that depending on the methodology applied to retrieve data, the output of some bibliometric indicators used to estimate institutional publication performance might be different. Since 3 years ago a study reported the lack of standardization in the institutional address of Chilean universities (Krauskopf, 2013), the objective of this communication is to verify if these discrepancies are also present in queries for US institutions.

Methods

Bibliometric information was extracted on the week of December 7 2015 for US institutions from Scopus and the WoS for a 3-year period (2012–2014). The first approach (country approach) consisted of searching for the target

Received February 15, 2016; revised August 16, 2016; accepted August 26, 2016

© 2017 ASIS&T • Published online 3 May 2017 in Wiley Online Library (wileyonlinelibrary.com). DOI: 10.1002/asi.23809

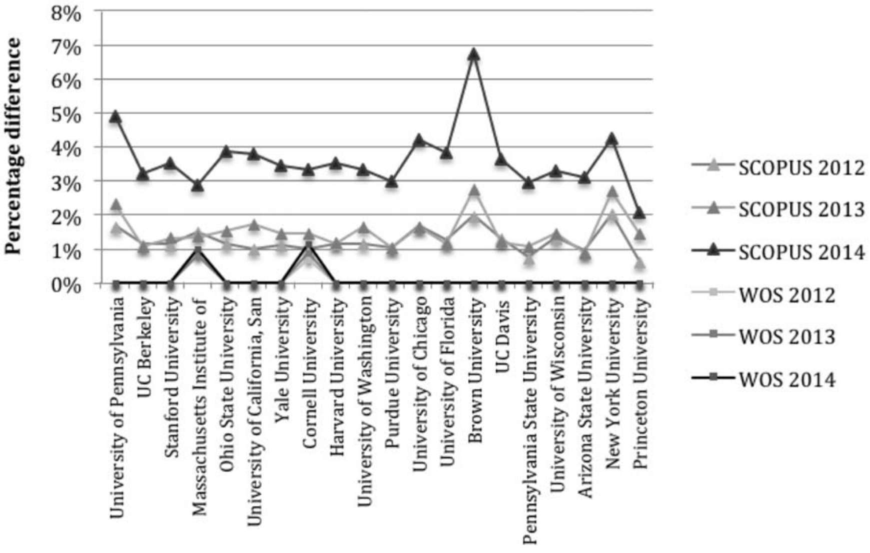

FIG. 1. Percentage difference in the total number of documents registered to an institution utilizing two different search queries using Scopus or the Web of Science as the source of information.

country (using AFFILCOUNTRY [US] for Scopus, while CU = USA was used for the WoS) through an advanced search for each year and then querying the database for a list of all the institutions that had published documents within that time period. From this list, 20 institutions were chosen randomly among the top-100 universities (in terms of documents published) to assess the search consistency within each database.

The second approach (affiliation) consisted of using an advanced query to search specifically for each of the 20 institutions chosen previously, again for each of the 3 years. The "affiliation search" option was used for Scopus, while the "organization-enhanced" (code OG) was used for the WoS.

For each database we established the difference in the number of publications for each university using both approaches and expressed it in terms of percentage. All types of documents published were considered for this analysis.

Results

As Figure 1 depicts, data retrieved using Scopus showed noteworthy differences for the 20 institutions, ranging up to 6.7%. In fact, these differences were larger for queries for more recent years. It was interesting to notice that the highest differences throughout the period examined were always associated with the same three institutions of the 20 chosen: Brown University, New York University, and the University of Pennsylvania.

A closer look at the data from these institutions revealed minor problems with misspelled addresses, although most of the cases did not show any problem that could be highlighted. For example, Table 1 displays documents that are registered to each of the 20 universities, but can only be found using the second approach (the affiliation search instead of the country-based search). To discard the likelihood that this difference is attributed to international collaborations, we analyzed the affiliations of each author. On all of the documents we found only authors affiliated with institutions with a US address.

As a user, one expects that these discrepancies would be nonexistent or reduced to a minimum, since Scopus has incorporated an "Affiliation Identifier," a unique number that distinguishes institutions with similar names. Nevertheless, it does not seem to be working properly, as the affiliation ID was included during the Boolean combination query, as shown below:

AF-ID (("Princeton University" 60003269)) AND NOT (AFFILCOUNTRY (united AND states)) AND (LIMIT-TO (PUBYEAR, 2014)) AND (LIMIT-TO (AF-ID, "Princeton University" 60003269))

Conversely, it is worth noting that the WoS showed significantly lower differences between both search strategies. Only Cornell University and the Massachusetts Institute of Technology (MIT) showed differences during the 3 years assessed, reaching up to 1.2% during 2014 (Figure 1). The other 18 institutions did not show any differences between both search methods in the WoS. Further analysis of these differences revealed two interesting details. In the case of Cornell University, all of the documents that make up the 1.2% difference are from the Weill Cornell Medical College Qatar, which was established by Cornell University in partnership with Qatar Foundation for Education, Science, and Community Development in 2001.

In the case of MIT, the 1% difference corresponded to documents that were authored by researchers that did not register MIT as their affiliation, even though MIT appears in all cases as the "Organization-Enhanced Name." These researchers belonged to foreign institutions that had collaboration agreements with MIT: the Skolkovo Institute of

TABLE 1. Examples of documents from each institution, retrieved using the affiliation search option from Scopus.

University	Journal	Year	Vol	Pag	Total Citations
University of Pennsylvania	*Nature Immunology*	2014	15	929	50
UC Berkeley	*Nature Communications*	2014	5	3,526	59
Stanford University	*Nature Medicine*	2014	20	552	82
Massachusetts Institute of Technology	*Genome Biology*	2014	15	409	50
Ohio State University	*Annual Review of Pathology*	2014	9	287	77
University of California, San Diego	*Chemical Reviews*	2014	114	8,662	35
Yale University	*Clinical infectious diseases: an official publication of the Infectious Diseases Society of America*	2014	59	1,066	34
Cornell University	*Nature Genetics*	2014	46	1,034	23
Harvard University	*Science*	2014	346	1,335	25
University of Washington Seattle	*Clinical infectious diseases: an official publication of the Infectious Diseases Society of America*	2014	59	147	42
Purdue University	*Science*	2014	344	263	32
University of Chicago	*Nature Chemical Biology*	2014	10	93	74
University of Florida	*Science*	2014	346	1,078	37
Brown University	*Nature Communications*	2014	5	3,256	15
UC Davis	*American Journal of Psychiatry*	2014	171	627	44
Pennsylvania State University	*Journal of Virology*	2014	88	10,056	29
University of Wisconsin Madison	*Cell*	2014	158	1,389	26
Arizona State University	*PLoS Pathogens*	2014	10	e1004202	17
New York University	*Nature Neuroscience*	2014	17	1,198	9
Princeton University	*Nature Neuroscience*	2014	17	1,816	17

Note. None of these documents were discovered using the global approach. Examples of documents retrieved from Scopus using only the affiliation search option.

Science and Technology, founded in 2011 by nine Russian institutions and organizations; and the Singapore-MIT alliance, founded in 1998 by two universities from Singapore and MIT. The differences found for the WoS with the two search strategies are not, in the end, erroneous, since the papers that were not found using the country search are not from the US per se.

Conclusions

Although previous studies have compared Scopus and the WoS in terms of journal coverage (Falagas, Kouranos, Arancibia-Jorge, & Karageorgopoulos, 2008; Lopez-Illescas, Moya-Anegon, & Moed, 2008; Mongeon & Paul-Hus, 2016) and citation analysis (Li, Burnham, Lemley, & Britton, 2010; Meho & Sugimoto, 2009), inaccuracies have been recently reported for these databases (Franceschini, Maisano, & Mastrogiacomo, 2015, 2016; Valderrama-Zurian, Aguilar-Moya, Melero-Fuentes, & Aleixandre-Benavent, 2015).

Then, how can the accuracy and consistency of the information contained within these databases be optimized? Manual editing of these databases is unrealistic due to their large size; therefore, new algorithms should be created and implemented to find inconsistencies such as the ones described in Scopus. Nonetheless, institutions also need to assume responsibility for their lack of awareness.

Most of the university rankings weigh factors that measure research productivity and performance based on globally accessible bibliometric indicators such as total amount of publications and citations. Since these institutional rankings have been developed by various types of organizations (news media, universities, nongovernmental organizations, etc.) (Cakir, Acarturk, Alasehir, & Cilingir, 2015) their database search strategy may differ. Thus, in the interest of their users, each ranking system should indicate the search strategy used in their methodology section. For instance, are international branch campuses considered in the final tally? Currently, there are over 200 international branch campuses from 73 different countries distributed worldwide (Cross-Border Education Research Team, 2016; Healey, 2016). Without the complete information users (prospective students, government agencies, etc.) might make wrong assumptions. A similar issue has been raised by Mongeon and Paul-Hus (2016) when defining the geographical origin of a journal: should it be by the country of its publisher or its editor?

At the end of the day, the better knowledge we have about the method by which data are collected, the smaller the risk of misuse or misinterpretation.

Acknowledgment

This work was supported by the grant Proyecto Financiamiento Basal (PFB-016).

References

Cakir, M.P., Acarturk, C., Alasehir, O., & Cilingir, C. (2015). A comparative analysis of global and national university ranking systems. Scientometrics, 103(3), 813–848.

Claasen, C. (2015). Measuring university quality. Scientometrics, 104(3), 793–807.

Cross-Border Education Research Team. (2016). C-BERT branch campus listing. [Data originally collected by Kevin Kinser and Jason E. Lane]. Retrieved from http://globalhighered.org/branchcampuses.php.

Falagas, M.E., Kouranos, V.D., Arancibia-Jorge, R., & Karageorgopoulos, D.E. (2008). Comparison of SCImago journal rank indicator with journal impact factor. The FASEB Journal, 22(8), 2623–2628.

Franceschini, F., Maisano, D., & Mastrogiacomo, L. (2015). Errors in DOI indexing by bibliometric databases. Scientometrics, 102(3), 2181–2186.

Franceschini, F., Maisano, D., & Mastrogiacomo, L. (2016). The museum of errors/horrors in Scopus. Journal of Informetrics, 10(1), 174–182.

Garfield, E. (1955). Citation indexes for science. Science, 122(3159), 108–111.

Healey, N.M. (2016). The challenges of leading an international branch campus: The “lived experience” of in-country senior managers. Journal of Studies in International Education, 20(1), 61–78.

Krauskopf, E. (2013). Standardization of the institutional address. Scientometrics, 94(3), 1313–1315.

Li, J., Burnham, J.F., Lemley, T., & Britton, R.M. (2010). Citation analysis: Comparison of Web of Science, Scopus, SciFinder, and Google Scholar. Journal of Electronic Resources in Medical Libraries, 7(3), 196–217.

Lopez-Illescas, C., Moya-Anegon, F., & Moed, H.F. (2008). Coverage and citation impact of oncological journals in the Web of Science and Scopus. Journal of Informetrics, 2(4), 304–316.

Marketwatch. (2016) Thomson Reuters announces definitive agreement to sell its intellectual property & science business to Onex and Baring Asia for $3.55 billion. Retrieved from http://www.marketwatch.com/story/thomson-reuters-announces-definitive-agreement-to-sell-its-intellectual-property-science-business-to-onex-and-baring-asia-for-355-billion-2016-07-11

Meho, L.I., & Sugimoto, C.R. (2009). Assessing the scholarly impact of information studies: A tale of two citation databases — Scopus and Web of Science. Journal of the American Society for Information Science and Technology, 60(12), 2499–2508.

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

Valderrama-Zurian, J.C., Aguilar-Moya, R., Melero-Fuentes, D., & Aleixandre-Benavent, R. (2015). A systematic analysis of duplicate records in Scopus. Journal of Informetrics, 9(3), 570–576.

DOI: 10.1002/asi