Online Information Review

Open bibliometrics and undiscovered public knowledge
David Stuart,

Open bibliometrics and undiscovered public knowledge

David Stuart
School of Mathematics and Computer Science, University of Wolverhampton, Wolverhampton, UK

Received 13 July 2017
Revised 21 August 2017
Accepted 23 August 2017

Abstract

Purpose – The purpose of this paper is to raise awareness of the potential of open bibliometrics, especially for the discovery of previously undiscovered public knowledge.
Design/methodology/approach – The viewpoint considers the limitations of the most popular current bibliometric tools and the possibilities offered from more open tools. It is supported by analysis of the openness of keywords associated with bibliometric studies in 2016.
Findings – The paper finds that although tools are emerging that offer more open bibliometrics, bibliometric research nonetheless continues to make use of restricted services.
Originality/value – This viewpoint on the potential of open bibliometrics is supported by an analysis of the current openness of bibliometric keywords.

Keywords Bibliometrics, Open access, Altmetrics, Webometrics, Open bibliometrics
Paper type Viewpoint

Introduction

There are many different threads to open science, and some have captured the scientific community's attention more than others. Unfortunately, open bibliometrics probably comes a long way down on a list of open science priorities for most researchers, and calls for open bibliometrics and citations (e.g. Shotton, 2013) have not captured the academic community's attention as much as calls for open access or open data. This, however, may be a mistake, as the importance of open bibliometrics grows with the importance of both bibliometrics and open science, and neither show signs of slowing down.

Bibliometrics is multifaceted, and therefore so is the potential impact of open bibliometrics. Bibliometric studies may be broadly categorized as either relational or evaluative, either offering insights into the relationship between units of analysis or aiding in the evaluation of units of analysis. This viewpoint considers open bibliometrics from the perspective of relational bibliometric analysis. More specifically, bibliometric analysis for the discovery of undiscovered public knowledge. Whilst evaluative bibliometrics often gets the most attention (albeit often for the wrong reasons), relational bibliometrics offers some of the more exciting avenues of research.

Undiscovered public knowledge is knowledge that whilst in the public domain is undiscovered due to its fragmented nature (Swanson, 1986). Relational bibliometrics can help to bring these fragments together, although it requires open bibliometric tools and resources. To help understand the openness of current bibliometric studies, this viewpoint is supported by an analysis of keywords associated with current bibliometric studies.

Background

Bibliometrics is an increasingly important and distinct branch of library and information science (Milojevic´ *et al.*, 2011), and both "bibliometric" and "citation" are growth terms relative to information and library science as a whole (Larivière *et al.*, 2012). Defined by Pritchard (1969) as "the application of mathematics and statistical methods to books and other media of communication" (p. 349), bibliometrics is now one of many terms available for applying "mathematical and statistical tools to an increasingly elusive set of objects" (De Bellis, 2014, p. 23). Increasingly these objects are found online, and the shift to the

Online Information Review
Vol. 42 No. 3, 2018
pp. 412-418

1468-4527
DOI 10.1108/OIR-07-2017-0209

networked age is significant for bibliometrics in that it has been said that the web offers the potential to "expand and democratize the tools and techniques for communicating, evaluating, and counting science" (De Bellis, 2014, p. 410). However, whilst there has been an increase in the number of citation services available, as well as the emergence of webometrics and altmetrics, we are a long way from an open bibliometrics.

Citation services

Citations are the most explicit form of an intellectual debt that is generally made between two papers, and Shotton (2013) has described the lack of free and easy access to citation data as a "scandal," pointing to the difficulty researchers can have in accessing the major citation indexes, the limited usability of the data, and the restrictions in republishing data. In recent years, the Web of Science has been joined by three other major services for citation analysis, Scopus, Google Scholar, and, most recently, Microsoft Academic, but whilst competition in the citation marketplace is to be welcomed, we are still a long way from free and easy access to quality citation data. The fact that Google Scholar and Microsoft Academic are free-to-access services is an undoubtedly important step in the right direction, but there are significant differences between free-to-reuse and free-to-access, and even between the ways data can be accessed.

Of the two major free-to-access services, Google Scholar is the longest established and continues to have greater coverage than Microsoft Academic (Harzing and Alakangas, 2017), however Microsoft Academic has an application programming interface (API) enabling access to a greater amount of data and a wider variety of analysis. In fact, the whole of the underlying Microsoft Academic Graph of publication records was available for download as part of the 22nd ACM SIGKDD Conference on Knowledge Discovery and Data Mining (Microsoft, 2016), although as the graph has grown in size, it is now only available via the API. In comparison, Google Scholar does not provide an API due to publisher restrictions (van Noorden, 2014). Nevertheless, there are still limitations with Microsoft Academic: questions have been raised about the completeness of the affiliation and citation data (Herrmannova and Knoth, 2016), and researchers are still limited by Microsoft's terms and conditions. The terms and conditions associated with the Microsoft Academic Graph seem quite generous at first: 10,000 API calls a month available for free, and an invitation to develop your own web services with an appropriate acknowledgment. However, such limits can quickly be used up, and search engine functionality has been known to disappear from search engine APIs when they are no longer in their commercial interest (Ortega *et al.*, 2014).

Other services have been established to provide open access to citations with fewer restrictions, for example, the Open Citations Corpus (http://opencitations.net/corpus) data has been made available under a Creative Commons Zero license, and CiteSeerX (citeseerx.ist.psu.edu) data are available under a Creative Commons Attribution-NonCommercial-ShareAlike license. Such services, however, are often subject specific and are far smaller than the commercial offerings.

Altmetric and webometric services

The web has not only provided access to new citation services, but also opportunities for insights into less formal discourse, which has led to the emergence of altmetrics and webometrics. Webometrics is the "study of quantitative aspects of the construction and use of information resources, structures and technologies on the Web drawing on bibliometric and informetric approaches" (Björneborn and Ingwersen, 2004, p. 1217), whilst altmetrics focuses on the structured nature of social network technologies to establish alternative filters and research indicators (Priem *et al.*, 2010). Webometrics and altmetrics offer the opportunity for new and fast insights into the impact of science and the relationship between fields and ideas. Importantly, it also allows a wider range of outputs to be

measured, and the development of metrics can theoretically reflect the needs of the community rather than being dictated by what citation databases choose to index and provide access to. The new opportunities have not gone unnoticed by commercial providers with the most recent altmetric acquisition being Plum Analytics being bought by Elsevier (2017) who also own Scopus.

Whilst ostensibly altmetrics and webometrics are "open" as they make use of data on the public web, in reality researchers must make use of third party resources through which to view the web. This may be directly through the use of an API associated with a particular site or service (e.g. Twitter or Mendeley), or indirectly through a third party that provides access to the aggregated data (e.g. a search engine or Altmetric.com). As with citation services, social network sites and data aggregators impose conditions on how the data may be accessed and used, with an investigation of a dozen different social network sites potentially having a dozen different sets of terms and conditions to be accommodated. Whether altmetrics is really any more open than traditional citation analysis is a matter of debate, although services such as Common Crawl (http://commoncrawl.org), an open repository of web crawl data, provides the opportunity for more open webometrics, at least for those with the requisite technical skills.

How open was bibliometrics in 2016?

An indication of the current openness of bibliometrics can be gained through an analysis of the terms associated with the topic: bibliometrics, altmetrics, and webometrics were searched for in the Web of Science to give an indication of the relative size of each of those areas with a more detailed analysis of the author keywords associated with bibliometrics.

The growth of bibliometrics, altmetrics, and webometrics

"bibliometric*," "altmetric*," and "webometric*" were searched for in the Topic Field in the Web of Science. This returns those records that include each of the terms, along with their plural and any inflected forms, in the Title, Abstract, Author Keywords, or Keywords Plus fields.

As can be seen in Figure 1, bibliometrics continues to be the most popular term by far, webometrics fell significantly in 2016, and altmetrics is on the rise, albeit still nowhere near the same levels as bibliometrics.

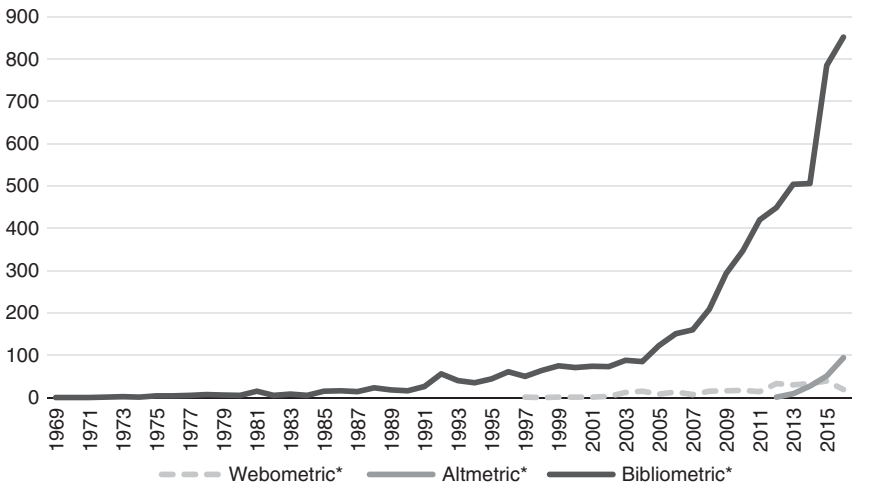

Figure 1.
Number of papers indexed by the Web of Science with bibliometric*/ altmetric*/ webometric* in the Topic Field

Of the 94 papers that are retrieved by a topic search for "altmetric*" in 2016, only 33 (35 percent) of these are also retrieved by "bibliometric*." Comparatively, whilst 19 papers were retrieved by a topic search for "webometric*," only 3 (15 percent) of these are also retrieved by "bibliometric*."

The openness of bibliometric author keywords

A search for "bibliometric*" in the Topic Field of the Web of Science was also used to identify bibliometric papers for a more detailed analysis into the openness of the topic. Title, Abstract, Author Keywords, and Keywords Plus fields, may each be used as surrogates to provide insights into the contents of a paper. Titles are designed for capturing attention and are overloaded with information (Milojević *et al.*, 2011), but not all relevant key terms necessarily fit easily into a title, and all free text fields necessitate additional steps in transforming the text into distinct terms. Keywords Plus are created automatically from "significant, frequently occurring words in the titles of an article's cited references" (Web of Science, 2008), and have been found to be effective in analyzing the knowledge structure of science, but they are less representative of an article's content (Zhang *et al.*, 2016). Author Keywords are used for the concept analysis to ensure only those terms deemed relevant by the author are included. Of the 852 papers retrieved from the Web of Science with a topic search of "bibliometric*," 769 had author keywords. Each keyword was analyzed to determine whether it indicated a more open type of bibliometrics.

In total there were 2,206 distinct keywords, 2,111 after the application of the Porter2 stemming algorithm in the stemming 1.0 Python package (https://pypi.python.org/pypi/stemming/1.0). For those papers that had author keywords, the mean number of author keywords was 4.9, although the most author keywords associated with one paper was 14.

The top 20 author keywords associated with bibliometrics are provided in Table I.

The composition of the top 20 keywords broadly reflects that of the author keywords as a whole, primarily consisting of traditional bibliometric methods and sources, and the topics and countries that were the focus of bibliometric analysis.

Analysis of the 2,111 stemmed author keywords showed that few reflected an increasingly open bibliometrics. The number of mentions of a citation service in fact increases with a lack of openness: Web of Science (40), Scopus (24), Google Scholar (7), Microsoft Academic Search (1), Open Citations Corpus (0), and CiteSeerX (0).

Additional terms that did indicate a degree of openness primarily reflect the rise in altmetrics and online content: Altmetrics (16), Social Media (7), Social Networking (4), Mendeley (3), Twitter (1), Twitter Counts (1), Moocs (1), Social Media Metrics (1), Social Networking Sites (1) Web 2.0 (1), Library 2.0 (1).

An interest in openness can also be seen at the fringes of more closed bibliometrics: Open access (5), Open Access Concept (1), and Open Innovation (1) have all been the focus of

Keyword	Frequency (769 papers with Author Keywords)		
Bibliometrics	338	Scopus	24
Bibliometric Analysis	111	Research	23
Citation Analysis	66	Research Evaluation	22
Scientometrics	48	Social Network Analysis	19
Web of Science	40	Impact Factor	18
Citation	37	Publications	18
Bibliometric Index	28	Altmetrics	16
H-Index	28	Innovation	14
Scientific Production	25	Research Trendstrends	13
Bibliometric Studies	24	China	13

Table I. Most frequently used author keywords for WoS records retrieved with a "bibliometric*" Topic Field search

bibliometric studies, and free bibliographic tools are being used that have been created specifically for the bibliometric community: CiteSpace (3), BibExcel(2), Cited References Explorer (1), VOSviewer (1), VOSviewer Map (1).

Discussion

Bibliometrics is still closed

There is little doubt that bibliometrics continues to be dominated by the traditional form of citation analysis, and that citation analysis is dominated by paid to access services. Whilst altmetrics has seemingly made the breakthrough that webometrics never managed, it nonetheless still only accounts for 10.3 percent of the combined "altmetric*" and "bibliometric*" bibliographic set, and even then it should not be overlooked that altmetrics and webometrics are heavily reliant on services provided by third parties.

The lack of a central open citation service around which bibliometricians can coalesce has undoubtedly been a limitation for the development of open bibliometrics, although this may be beginning to change. Whilst the potential of Microsoft Academic Search has not yet been reflected in the bibliometric studies of 2016, the functionality should encourage wider use as bibliometric tools are built on top of it; it has already been incorporated into Publish or Perish (https://harzing.com/resources/publish-or-perish) and is being incorporated into Webometric Analyst (http://lexiurl.wlv.ac.uk). There is also the potential for new providers of citation services to emerge as the barriers to the establishment of such services falls; over 45 percent of scholarly literature from 2015 was found to be available in an open access format (Piwowar *et al.*, 2017), and open source libraries are being developed that enable the extraction of citation data from unstructured documents (e.g. GROBID, https://github.com/kermitt2/grobid). There are still issues, however, with the quality of the data, and the rights of access, that will need to be overcome before we can expect the position of Web of Science, or even Scopus, to be usurped. It should also be noted that the rise in open access, as well as lowering the barriers for new services, could also compound data quality problems as multiple versions are made available online.

Although the rapid rise of altmetrics shows a rising interest in less formal types of publication, and potentially a willingness for less robust indicators, this is not necessarily unadulterated good news for an increasingly open bibliometrics. The structured nature of altmetric data that has driven much of the interest comes at the cost of the data being owned by the dominant social network sites. Webometrics does not necessarily have the same power imbalance between researcher and data provider, but unfortunately this is in decline.

It is also important that we do not overlook the rapidly changing and fragmented nature of social network sites, and focus on those services that can be easily measured rather than those that should be measured. For example, the traditional openness of Twitter and the extensiveness of its APIs meant that it was the focus of more studies than Facebook, whilst WhatsApp is a black box, despite having over one billion users.

Bibliometrics for undiscovered public knowledge

Open bibliometrics are considered in this viewpoint in connection with relational bibliometrics and the identification of undiscovered public knowledge, because of the particular requirements such a bibliometric study places on a bibliometric service. Whereas evaluative bibliometrics are typically accommodated by the simplest of bibliometric services, relational bibliometrics often require far greater functionality.

Consider a typical evaluative bibliometric use, analyzing the citation impact of a set of documents as part of a research assessment exercise, people may be cautioned against such use, but they continue to do it anyway (Sayer, 2015). The bibliometric investigator merely needs to go to the citation service of their choice, enter the title of the paper (or papers) they are interested in, and they will quickly be presented with the number of citations that paper

has received. The conscientious investigator may wish to go a step further and consider the nature of those citations (e.g. are citing papers highly cited, are they self-citations), but one degree of separation would probably be considered more than sufficient.

In comparison, a service that wishes to use bibliometrics in the identification of undiscovered public knowledge is likely to be interested in two degrees of separation. The seminal undiscovered public knowledge example is that of the relationship between dietary fish oils and Raynaud's disease; that dietary fish oils could lower blood viscosity was separately known to the fact that those with Raynaud's disease had abnormally high bloody viscosity. For Swanson and Smallheiser (1996), complementary literatures were found through title words in bibliographic records in MEDLINE. Citations provide an additional method of identifying complementary literature.

The skewed nature of citation distribution means that the difference between one and two degrees of separation are equally skewed. But take, for example, a highly cited OIR article, Jasco's Google Scholar: the pros and cons. The 108 papers that cite the article (i.e. at one degree of separation), mushroom to 1,481 papers at two degrees of separation. If each bibliographic record has to be downloaded separately, a month's 10,000 API requests would quickly disappear.

It may also be argued that such a relational study requires a more integrated approach of bibliometric analysis, combining citations with altmetrics and webometrics to get a more nuanced understanding of the relationship between different sets of complementary literature. The ability of the web to offer transversal links, short cuts between different web clusters, has long been recognized in webometrics (Björneborn and Ingwersen, 2001), and altmetrics may provide a richer data set for such transversal links.

Conclusion

There has been a rapid rise in interest in bibliometrics and altmetrics in recent years, and that shows no signs of slowing, but there is a huge gap in the openness of current bibliometrics and the openness that is necessary for the most robust and insightful evaluative and relational bibliometrics possible. Undoubtedly the status quo is probably helped in part by the current emphasis on evaluative bibliometrics, where researchers find ways of working within the limitations of the bibliometric services.

Tibor Braun described 1992, the year that Eugene Garfield sold the Web of Science, as the end of the romantic period of bibliometrics (van Raan, 2013), and whilst the web may offer the potential for the democratization of citation tools, there are few signs of a pre-lapsarian, less commercial garden on the horizon. If there is hope for open bibliometrics, it is in the "moral ballast," to use Sayer (2015, p. 91) phrase, that accompanies the "openness" that is increasingly a norm of science.

References

Björneborn, L. and Ingwersen, P. (2001), "Perspectives of webometrics", *Scientometrics*, Vol. 50 No. 1, pp. 65-82.

Björneborn, L. and Ingwersen, P. (2004), "Toward a basic framework for webometrics", *Journal of the American Society for Information Science and Technology*, Vol. 55 No. 14, pp. 1216-1227.

De Bellis, N. (2014), "History and evolution of (biblio)metrics", in Cronin, B. and Sugimoto, C.R. (Eds), *Beyond Bibliometrics: Harnessing Multidimensional Indicators of Scholarly Impact*, MIT Press, Cambridge, MA, pp. 23-44.

Elsevier (2017), *Elsevier Acquires Leading 'Altmetrics' Provider Plum Analytics*, Elsevier, available at: www.elsevier.com/about/press-releases/corporate/elsevier-acquires-leading-altmetrics-provider-plum-analytics

Harzing, A.-W. and Alakangas, S. (2017), "Microsoft academic: is the phoenix getting wings?", *Scientometrics*, Vol. 110 No. 1, pp. 371-381.

Herrmannova, D. and Knoth, P. (2016), "An analysis of the microsoft academic graph", *D-Lib Magazine*, Vol. 22 Nos 9/10, available at: www.dlib.org/dlib/september16/herrmannova/09herrmannova.html

Larivière, V., Sugimoto, C.R. and Cronin, B. (2012), "A bibliometric chronicling of library and information science's first hundred years", *Journal of the American Society for Information Science and Technology*, Vol. 63 No. 5, pp. 997-1016.

Milojević, S., Sugimoto, C.R., Yan, E. and Ding, Y. (2011), "The cognitive structure of library and information science: analysis of article title words", *Journal of the American Society for Information Science and Technology*, Vol. 62 No. 10, pp. 1933-1953.

Ortega, J.L., Orduña-Malea, E. and Aguillo, I.F. (2014), "Are web mentions accurate substitutes for inlinks for Spanish universities?", *Online Information Review*, Vol. 38 No. 1, pp. 59-77.

Piwowar, H., Priem, J., Larivière, V., Alperin, J.P., Matthias, L., Norlander, B., Farley, A., West, J. and Haustein, S. (2017), "The state of OA: a large-scale analysis of the prevalence and impact of Open Access articles", *PeerJ Preprints*, Vol. 5, p. e3119v1, available at: https://peerj.com/preprints/3119/

Priem, J., Taraborelli, D., Groth, P. and Neylon, C. (2010), "Altmetrics: a manifesto", altmetrics, available at: http://altmetrics.org/manifesto (accessed August 21, 2017).

Pritchard, A. (1969), "Statistical bibliography or bibliometrics?", *Journal of Documentation*, Vol. 25 No. 4, pp. 348-349.

Sayer, D. (2015), *Rank Hypocrisies: The Insult of the REF*, SAGE Publications, London.

Shotton, D. (2013), "Publishing: open citations", Nature, available at: www.nature.com/news/publishing-open-citations-1.13937 (accessed August 21, 2017).

Swanson, D.R. (1986), "Undiscovered public knowledge", *Library Quarterly*, Vol. 56 No. 2, pp. 103-118.

Swanson, D.R. and Smallheiser, N.R. (1996), "Undiscovered public knowledge: a ten-year update", in Simoudis, E., Han, J. and Fayadd, U. (Eds), *Proceedings of the 2nd International Conference on Knowledge Discovery and Data Mining*, AAI PRESS, Menlo Park, CA, pp. 295-298.

van Noorden, R. (2014), "Google scholar pioneer on search engine's future", *Nature,* available at: www.nature.com/news/google-scholar-pioneer-on-search-engine-s-future-1.16269 (accessed August 21, 2017).

van Raan, T.F.J. (2013), "Citations, h-index, journal impact and rankings: not all sorrow and misery. CWTS: a short history of measuring science", in van Holsteyn, J., Mom, R., Smit, I., Tromp, H. and Wolters, G. (Eds), *Perspectives on the Past: 50 years of FSW*, Biblioscope, Utrecht, pp. 86-103, available at: www.cwts.nl/TvR/documents/AvR-2013-FSW50-ENG.pdf

Web of Science (2008), ISI Proceedings, available at: https://images.webofknowledge.com/WOK48B3/help/ISIP/h_fullrec.html (accessed August 21, 2017).

Zhang, J., Yu, Q., Zheng, F., Long, C., Lu, Z. and Duan, Z. (2016), "Comparing keywords plus of WOS and author keywords: a case study of patient adherence research", *Journal of the Association for Information Science and Technology*, Vol. 67 No. 4, pp. 967-972.

Corresponding author
David Stuart can be contacted at: dp_stuart@hotmail.com