Citation Analysis in Open Access World: A Case Study of Health Science Open Access Repositories

Zahid Ashraf Wani
Assistant Professor, Department of Library and Information Science, University of Kashmir, Hazratbal, Srinagar, Jammu & Kashmir, India

Tariq Shafi
Assistant Librarian, J&K Academy of Art, Culture & Languages, Lalmandi, Srinagar, Jammu & Kashmir, India

DOI: 10.18329/09757597/2015/8104

World Digital Libraries 8(1): 49–58 (2015)

Abstract

The study made an attempt to carry out the citation analysis of primary literature archived in select Open Access Institutional repositories in the field of Health Science listed by Directory of Open Access Repositories (OpenDOAR). A total of 99 articles harvested from three select Health Science repositories were tested in two primary literature indexing databases, viz., Web of Science and Google Scholar for citation analysis. Among these, Google Scholar could not retrieve three (10.34 per cent) articles while for Web of Science this number goes to 26 (89.66 per cent). The number of citations received by the deposited contents is higher in Google Scholar (59.57 per cent) as compared to Web of Science (40.42 per cent). However, the average citation per article is higher in Web of Science (3.24) in comparison to Google Scholar (2.53). There are 27 articles in the resource corpus which have not received any citation. Number of such articles in Google Scholar and Web of Sciences are 20 (20.83 per cent) and 7 (9.59 per cent), respectively.

Keywords: Citation analysis, Open access, Open access repositories, Health science

1. Introduction

Since the dawn of the Open Access movement some major changes have been witnessed in scholarly publishing world especially emergence of Open Access Journals and the setting up of Open Access Institutional Repositories that let the authors to self-archive their scholarly works and to make their current research results freely accessible to the potential users, even before appearing in print format. The combination of institutional repositories and open access journals is increasing, giving libraries and researchers their first chance to change fundamentally the way scientific information is communicated. They hold out the promise of providing a fairer, more equitable, and more efficient system of scholarly communication, and one that can better serve the international research community (Prosser 2004).

The purpose with which open access institutional repositories are created is to encourage scholarly communication outside traditional publishing models, demonstrate the prestige of institutions by highlighting their scholarly output, and to make this output accessible to the wider academic community (Crow 2002). However, while institutional repositories have been adopted across the academic spectrum, the quality of the materials maintained within them is not often representative of the institution's academic stature. In order to increase access to quality materials and create real alternatives to journal publication, open access repositories must contain materials of value, both to serve the needs of academic institution as well as the larger scholarly community (Wacha and Wisner 2011).

The setting up of an institutional repository represents significant institutional investment as it hosts the intellectual assets of an institution and, thus, there should be some sound policies underlying the depositing of rich scholarly content and every possible effort should be made to evaluate and assess the usage statistics of the deposited content so as to get a clear view of the benefits which an institutional repository can possibly bring to the affiliated institutions in general and individual researchers in particular. Citation analysis of the deposited content in the open access institutional repositories can be used as one of the measures to justify this significant institutional investment.

Citing is the process by which scholars give recognition to research used by another academic researcher. Citation resources are tools used by academic scholars for keeping track of who did what research and the impact of the research within the discipline. Citation analysis is therefore an attempt to measure the impact and contribution of a study to the body of knowledge and research (Adriaanse 2011).

Citation analysis is an important tool used to trace scholarly research, measure impact, and justify tenure and funding decisions (Bauer and Bakkalbasi 2005). The number of citations received by a particular publication is seen as a quantitative measure of the resonance and impact created by that publication in the scientific community (Neuhaus and Danie 2006).

The citation resource by The Institute for Scientific Information (ISI), Web of Science (WoS), was traditionally the citation tool of choice of academics for more than 40 years. Changes in scholarly communication, including preprint/postprint servers, technical reports available via the Internet, and open access e-journals are developing rapidly, and traditional citation tracking using WoS may miss much of this new activity (Bauer and Bakkalbasi 2005). The arrival of Scopus in 2004—a fee-based citation resource, and Google Scholar (GS) — a citation resource available for free and accessible via the Web, presented WoS with competition (Adriaanse 2011).

The emergence of GS as citation tracking database has been warmly welcomed by the scholarly community as it has widened the scope of citation based metrics. But at the same time this new entrant has been put to the rigorous

tests to evaluate and assess its usefulness as citation database and to establish whether GS is a substitute for or complementary to the traditional tools.

2. Literature Review

The contents deposited in Open Access Institutional Repositories can be used without any restriction from any part of the world at any time which in turn can lead to increased readership, download hits, and ultimately higher citation impact. There are many studies which have attributed open access availability of scholarly content to increased citation rates.

The first seminal work to establish whether papers available for free on the web have higher citation impact was done by Lawrence (2001). He analysed the difference in citation rates between articles freely available on the web and those only available through either toll-access services, or paper-only. He examined a total of 119,924 articles published from 1990 to 2000 in computer science and estimated citation counts and online availability using Research Index, excluding self-citations. He found 4.5 times more citations to the articles that were freely available than the articles which were put behind the subscription barriers. However, Lawrence did not mention the number of open access articles in the data set, he used. Based on the findings, he suggests that in order to maximize impact, minimize redundancy and speed scientific progress. Authors and publishers should aim to make research easy to access.

A study by Harnad and Broody (2004) shows the results of an analysis of 95,012 journal articles and conference papers in Physics indexed by the ISI between the years 1992–2001. They compared the citation count of those articles that had been self-archived (making them available freely) by their authors to the citation counts to those authors who had not. They were able to show that there was a significant advantage in terms of the number of citations received by self-archived articles ranging from 2.5 per cent for restricted access articles to 5.8 per cent for self-archived articles on an average.

Antelman (2004) undertook a study to see whether research articles in four disciplines (Philosophy, Political Science, Electrical and Electronic Engineering, and Mathematics) at varying stages of adoption of Open Access (OA) have a greater impact as measured by citations in the ISI Web of Science database when their authors make them freely available on the Internet. Out of 602 articles, 17 per cent were OA in Philosophy; 29 per cent of articles among 299 were OA in Political Science; 37 per cent of articles out of 502 were OA in Electrical and Electronic Engineering; and 69 per cent of articles were OA in Mathematics out of 610 articles. She found a significant difference in the mean citation rates of OA articles and those that are not freely available online in all the four disciplines. The relative increase in citations for OA articles ranged from a low of 45 per cent in Philosophy to 51 per cent in Electrical and Electronic Engineering, 86 per cent in Political Science, and 91 per cent in Mathematics.

Xia, Myers and Wilhoite (2011) examined the relationship between multiple open access availability of journal articles and the citation advantage by collecting data of OA copies and citation numbers in 20 top library and information science journals. They discovered a correlation between the two variables; namely, multiple OA availability of an article has a positive impact on its citation count. The results of the study reveal that for every increase in the availability of OA articles, citation numbers increase by 2.348.

There are evidences in the literature that the papers posted as preprints prior to their publication have an added advantage in terms of their citation rates. Schwarz and Kennicutt (2004) investigated this 'preprint publishing culture' by using data from the Astrophysics Data System (ADS), the American Astronomical Society (AAS), and the arXiv electronic preprint server (astro-ph), to study the publishing,

preprint posting, and citation patterns for papers published in Astrophysical Journal (ApJ) in 1999 and 2002. Results of the study reveal that the ApJ papers posted prior to publication as astro-ph preprints are cited more than twice as often as papers that are not posted on astro-ph. The citation analysis in the fields of high-energy Physics and Astrophysics performed by Youngen (1998) indicates that the number of citations to traditional preprints has gradually declined over the past 10 years, and that citations to electronic preprints nearly double every year since 1992.

An analysis of 2,765 articles published in four maths journals from 1997 to 2005 by Davis and Fromerth (2006) indicates that articles deposited in the arXiv received 35 per cent more citations on an average than non-deposited articles (an advantage of about 1.1 citations per article), and that this difference was most pronounced for highly-cited articles. In a similar kind of study to test whether open access increases citation impact Brody (2006) used the arXiv—a collection of author self-archived Physics, Maths, and Computer Science e-prints. Comparing the number of citations to journal papers with and without an e-print in arXiv, he found that the papers with an arXiv e-print receive about twice as many citations as the papers without an e-print in arXiv. Henneken *et al.* (2006) also support the view that there is a major difference between the normalized citation rate for papers from the pre-arXiv era and papers that have been offered as e-prints in the arXiv repository.

In an another interesting study, Kim (2012) examined the relationship between free access to research articles and the diffusion of their ideas as measured by citation counts by using a dataset from the Social Science Research Network (SSRN), an open repository of research articles, by employing a natural experiment (select group of published articles posted on SSRN at a time chosen by their authors' affiliated organizations or SSRN, not by their authors) that allowed the estimation of the value of free access separate from confounding factors such as early viewership and quality differential. Using a difference-in-difference method and comparing the citation profiles of the articles before and after the posting time on SSRN against a group of control articles with similar characteristics, he estimated the effect of the SSRN posting on citation counts. The articles posted on SSRN receive more citations even prior to being posted on SSRN, suggesting that they are of higher quality. Their citation counts further increase after being posted, gaining an additional 10–20 per cent of citations. This gain is likely to be caused by the free access that SSRN provides.

Metcalfe (2005) in his study compared citations to articles in 13 major astrophysics journals with citations to articles in those journals that had also been made OA by posting in the arXiv and found a two-fold increase in citations for OA articles. He has clearly stated the benefits of OA by revealing that higher impact journal articles not posted to arXiv are cited less often than those from lower impact journals posted to arXiv. In an another study, wherein Metcalfe (2006) used wider data sources and compared OA (articles posted in the arXiv and Montana State University's Solar Physics Open Access Archive) to Non-OA articles in Solar Physics and confirmed that the articles posted to MSU's archive gained 1.7 times as many citations as non-OA articles and those posted to ArXiv received 2.6 times as many citations.

3. Scope

The scope of the study is limited to the open access institutional repositories in Health Science. The study is also limited to two citation databases Web of Science and Google Scholar which have been employed to obtain citation metrics acquired by the resource corpus from 2008 to April 15, 2013.

4. Objectives

The following objectives are laid down for the study:

- To measure the citation counts received by the contents deposited in select repositories during the study period.
- To identify the contents that have not received any citation.
- To determine the comprehensiveness of citation tracking tools.
- To determine the effect of authorship on citations.

5. Methodology

The study is carried out in the following three stages:

1. Selection of Health Science Open Access Institutional Repositories

A list of Open Access Institutional Repositories pertaining to Health Science was obtained from the Directory of Open Access Repositories (OpenDOAR). A total of 17 repositories having English Language database were identified and only 15 per cent of the repositories (3) were selected by purposive or judgement sampling. The following repositories were selected for the study:

- Digital Commons@Becker (DC@Becker)
- Digital Knowledge Repository of Central Drug Research Institute (DKR@CDRI)
- ECNIS Repository (Environmental Cancer Risk, Nutrition and Individual Susceptibility).

2. Harvesting of Resource Corpus from Select Repositories

In this stage, 20 per cent of the resource corpus was harvested from the select repositories by quasi-random sampling.

3. Determining the Citation Metrics of Harvested Resource Corpus

The harvested resource corpus was run on the Web of Science and Google Scholar for collecting the necessary data in accordance with the set objectives of the study for analysis and interpretation.

6. Analysis

6.1 Comprehensiveness of citation tracking tools

Amongst 99 articles harvested from the select repositories, 29 articles are not retrieved by Google Scholar (GS) and Web of Science (WoS). Among these, 3 (10.34 per cent) articles are not retrieved by GS and 26 (89.66 per cent) are not retrieved by WoS and thus reducing the resource corps to 96 and 73 articles in GS and WoS respectively. The maximum number of non-retrieved articles is observed in 'Digital Commons@Becker' (65.52 per cent) and the least in 'Environmental Cancer Risk, Nutrition and Individual Susceptibility Repository' (13.79 per cent). Table 1 offers a lucid picture.

The results clearly reveal the comprehensiveness of GS in tracking down the contents from Open Access Institutional Repositories owing to the fact that it indexes the resources from multiple locations where authors self archive their research results besides OA institutional repositories, thus giving the bibliometricians and the like, something to ponder upon, as it is emerging as a tough competitor to the subscription based Elsevier's 'Scopus' and Thomson Reuters 'Web of Science' which just index the resources from journals that are registered/indexed by these databases.

6.2 Articles with zero citations

A total of 27 articles have not received any citation. The maximum number of articles that are devoid of any citations in the select citation databases belong to GS with 20.83 per cent of the articles while the least articles that have not accumulated any citation during the study period belong to WoS (9.59 per cent) as is indicated in Table 2.

Table 1: Comprehensiveness of citation tracking tools

S.no.	Repository	Total no. of articles	No. of articles not retrieved		Total
			GS	WoS	
1.	CDRI	21	1	5	6 (20.69)
2.	DC@Becker	61	1	18	19 (65.52)
3.	ECNIS	17	1	3	4 (13.79)
Total		**99**	**3 (10.34)**	**26 (89.66)**	**29**

(Figures in parentheses indicate percentage)

Table 2: Articles with zero citations

S.no.	Repository	GS		WoS	
		No. of retrieved articles	Articles with 0 citations	No. of retrieved articles	Articles with 0 citations
1.	CDRI	20	3 (15.0)	16	1 (14.28)
2.	DC@Becker	60	16 (80.0)	43	5 (71.43)
3.	ECNIS	16	1 (5.0)	14	1 (14.28)
Total		**96**	**20 (20.83)**	**73**	**7 (9.59)**

(Figures in parentheses indicate percentage)

The maximum number of articles that have not received any citation in GS belong to 'Digital Commons@Becker' (80.0 per cent) followed by the articles deposited in 'Central Drug Research Institute' (15.0 per cent) and the least in 'Environmental Cancer Risk, Nutrition and Individual Susceptibility Repository (ECNIS)' with just 5.0 per cent articles.

The highest number of articles that are devoid of any citations in WoS comes from 'Digital Commons@Becker' (71.43 per cent) followed by CDRI and ECNIS with 14.28 per cent articles each.

6.3 Citation metrics in select repositories

The citation tracking tools employed for the study have a varying degree of strength in terms of their citation tracking metrics. It is pertinent to mention here that only those articles which have received citations (=1 or >1) have been taken into account. The titles which were neither retrieved nor have received any citations were not considered.

It is evident from Table 3 that GS is leading the deck with 59.57 per cent of the total citations followed by WoS (40.42 per cent). However, the average citations received per article is higher in WoS (3.24) as compared to GS (2.53).

The 'Digital Commons@Texas Medical Centre' has received maximum number of citations to its content by both the citation tracking tools (GS: 52.38 per cent and WoS: 48.85 per cent) followed by 'ECNIS' (WoS: 26.63 per cent and GS: 26.38 per cent). The least number of citations have been received by the contents deposited in 'Central Drug Research Institute' (WoS: 24.52 per cent and GS: 21.30 per cent) as is depicted in Table 4.

Table 3: Citation metrics

S.no.	Citation tracking tool	Total no. of articles	Total no. of citations	Average citation per article
1.	Google Scholar	76	3,005 (59.57)	2.53
2.	Web of Science	66	2,039 (40.42)	3.24
Total			**5,044**	

(Figures in parentheses indicate percentage)

6.4 Effect of authorship on citations

Google Scholar

In order to understand the impact of solo and collaborative endeavours on citation count in GS data was analysed and it was found that with an average of 39.54 mean citations, 76 papers have received a total of 3,005 citations (Table 5). Works which are produced in collaboration received more citations as compared to works that are solo efforts. Sixty-eight collaborative works have received mean citations of 43.31 while eight solo works have received only 7.5 mean citations. Amongst collaborative works, papers that are produced by a team of five individuals received maximum mean citations of 49.69 followed respectively by a group comprising more than five authors and four authors with 48.71 mean citations and 39.53 mean citations respectively.

Table 5: Effect of authorship on citations in Google Scholar

No. of authors	No. of papers	Total no. of citations	Mean no. of citations
1	8	60	7.5
2	7	127	18.14
3	9	407	45.22
4	15	593	39.53
5	16	795	49.69
>5	21	1,023	48.71
Total	**76**	**3,005**	**39.54**

(Figures in parentheses indicate percentage)

Web of Science

The WoS has fished out a total of 66 articles which have received 2,039 citations with a mean of 30.89 citations (Table 6). Among 66 papers, 64 papers have been written in collaboration and have received 2,020 citations with a mean of

Table 4: Citation metrics in select repositories

S.no.	Repository	GS		WoS	
		No. of articles	No. of citations	No. of articles	No. of citations
1.	CDRI	17	640 (21.30)	15	500 (24.52)
2.	DC@Becker	44	1,574 (52.38)	38	996 (48.85)
3.	ECNIS	15	791 (26.32)	13	543 (26.63)
Total		**76**	**3,005**	**66**	**2,039**

(Figures in parentheses indicate percentage)

Table 6: Effect of authorship on citations in Web of Science

No. of authors	No. of papers	Total no. of citations	Mean no. of citations
1	2	19	9.50
2	5	209	41.80
3	12	270	22.50
4	8	315	39.37
5	13	405	31.15
>5	26	821	31.58
Total	**66**	**2,039**	**30.89**

(Figures in parentheses indicate percentage)

45.0 citations. There are only two papers which bear just a single author and have received 19 citations with a mean of 9.50 citations. Mean number of citations for collaborative works is higher for the papers written in the team of two authors (41.80) followed by the team of four authors (39.37), and the group comprising more than five authors (31.58).

7. Conclusion

Open Access Institutional repositories are set up with the aim to maximize the use of the resources of a particular institution which in turn can bring laurels to the organization. The present study has revealed this fact by employing the citation analysis method to the contents deposited in Open Access Institutional repositories. It is clearly seen in the results that how well the contents in these repositories are put to use. A few number of articles have acquired such a huge number of citations which clearly reveal the benefits of OA on increased readership. There are many citation databases available in the market both free (e.g., Google Scholar) and proprietary (e.g., Web of Science and Scopus) with varying degree of strengths and weaknesses which can be used in carrying out citation based studies. The emergence of Google Scholar as a free citation tracking tool, has given something to cheer-upon, to the scholars interested in carrying out the citation based studies, belonging to the institutions who cannot afford to subscribe to the proprietary based citation indexing tools.

References

Adriaanse L S. 2011. **A comparison of the fee-based citation resources Web of Science and Scopus with the free citation resource Google scholar.** Available at <http://hdl.handle.net/10210/4938>

Antelman K. 2004. **Do open-access articles have a greater research impact?** *College & Research Libraries* **65**(5): 372–382.

Bauer K and Bakkalbasi N. 2005. **An examination of citation counts in a new scholarly communication environment**. *D-Lib Magazine* **11**(9). Available at <http://www.dlib.org/dlib/september05/bauer/09bauer.html>

Brody T D. 2006. **Evaluating research impact through open access to scholarly communications**. University of Southampton PhD thesis. (Retrieved January 25, 2013) Available at <http://www.erevistas.csic.es/descargas/brody%5B1%5D.pdf>

Crow R. 2002. **The case for institutional repositories: A SPARC position paper**. Available at <http://www.sparc.arl.org/bm~doc/ir_final_release_102.pdf>

Davis P M and Fromerth M J. 2006. **Does the arXiv lead to higher citations and reduced publisher downloads for mathematics articles?** (Retrieved March 23, 2013) Available at <http://arxiv.org/ftp/cs/papers/0603/0603056.pdf>

Harnad S and Broody T. 2004. **Comparing the impact of Open Access (OA) vs. Non-OA articles in the same journals**. *D-Lib Magazine* **10**(6). Available at <http://www.dlib.org/dlib/june04/hamad/06hamad.html>

Henneken, *et al.* 2006. **Effect of E-printing on citation rates in astronomy and physics**. (Retrieved March 11, 2013) Available at <http://arxiv.org/ftp/cs/papers/0604/0604061.pdf>

Kim H H. 2012. **The effect of free access on the diffusion of scholarly ideas**. Available at <http://mis.eller.arizona.edu/docs/events/2012/MIS_speakers_series_effect_of_free_access.pdf >

Lawrence S. 2001. **Free online availability substantially increases a paper's impact**. (Retrieved January 18, 2013) Available at <http://www.nature.com/nature/debates/e-access/Articles/lawrence.html>

Metcalfe T S. 2005. **The rise and citation impact of astro-ph in major journals**. (Retrieved January 20, 2012) Available at <http://arxiv.org/pdf/astro-ph/0503519.pdf>

Metcalfe T S. 2006. **The citation impact of digital preprint archives for solar physics papers**. (Retrieved January 20, 2012) Available at <http://arxiv.org/pdf/astro-ph/0607079.pdf>

Neuhaus C and Danie H D. 2006. **Data sources for performing citation analysis: An overview**. *Journal of Documentation* **64**(2): 193–210.

Prosser D C. 2004. **The next information revolution—How open access repositories and journals will transform scholarly communications**. *Liber Quarterly* **14**(1): 23-36.

Schwarz G J and Kennicutt R C. 2004. **Demographic and citation trends in astrophysical journal papers and preprints**. (Retrieved January 05, 2012) Available at <http://arxiv.org/pdf/astro-ph/0411275.pdf>

Xia J, Myers R L, and Wilhoite S K. 2011. **Multiple open access availability and citation impact**. *Journal of Information Science* **37**(1):19–28.

Wacha M and Wisner M. 2011. **Measuring value in open access repositories**. *The Serials Librarian* **61** (3-4): 377-388.

Youngen G K. 1998. **Citation patterns to electronic preprints in the astronomy and astrophysics literature**. *Library and Information Services in Astronomy*. Available at <www.stsci.edu/stsci/meetings/lisa3/youngeng.html>