Online Information Review

Michael Norris, Charles Oppenheim and Fytton Rowland
Department of Information Science, Loughborough University, UK

Finding open access articles using Google, Google Scholar, OAIster and OpenDOAR

Abstract

Purpose – The paper seeks to demonstrate the relative effectiveness of a range of search tools in finding open access (OA) versions of peer reviewed academic articles on the WWW.

Design/methodology/approach – Some background is given to why and how academics may make their articles OA and how they may be found by others searching for them. Google, Google Scholar, OAIster and OpenDOAR were used to try to locate OA versions of peer reviewed journal articles drawn from three subjects (ecology, economics, and sociology).

Findings – Of the 2519 articles 967 were found to have OA versions on the WWW. Google and Google Scholar found 76.84% of them. The results from OpenDOAR and OAIster were disappointing, but some improvements are noted. Only in economics could OAIster and OpenDOAR be considered a relative success.

Originality/value The paper shows the relative effectiveness of the search tools in these three subjects. The results indicate that those wanting to find OA articles in these subjects, for the moment at least, should use the general search engines Google and Google Scholar first rather than OpenDOAR or OAIster.

Keywords: Open access; Google; Google Scholar; OpenDOAR; OAIster; Retrieval Performance

Paper type Research paper

Introduction

Academics can make their articles open access (OA) and thus freely available to anyone with Internet access by self-archiving electronic versions of their articles on their own personal web page, their department's web page, a subject repository or by

depositing them in an institutional repository as well as submitting them to an OA journal, a means of access not considered here. Articles deposited in repositories which use the Open Archives Initiative Protocol for Metadata Harvesting (OAI-PMH) can have their metadata harvested, by for example, OAIster which currently holds the metadata of about 10 million articles. Anyone with Internet access may then search OAIster or use any of the general search engines to try and locate a particular author's work on the World Wide Web. It has been argued that articles, which are OA, accrue more citations than articles that remain behind subscription barriers (Harnad & Brody, 2004). Attempts to quantify this citation advantage have generally involved finding those articles that are OA on the World Wide Web and comparing their citation count to articles from the same journal issue, which remain accessible only by subscription. The mean citation counts of the two sets of articles are then compared (Antelman, 2004). In the first of two studies, similar in method to that conducted by Antelman the authors determined whether a particular set of OA articles did in fact have a citation advantage over their toll access (TA) counterparts (Norris, Oppenheim & Rowland in press). As part of the two studies, the authors used OAIster, OpenDOAR, Google and Google Scholar to try to locate as many OA versions of the articles as possible from the different subjects. The second study extended the first, by taking a further set of articles and used the same search tools to try to locate OA versions of them. This paper reports, primarily, the relative success of these search tools using article records from the second of the two studies.

Background

There are a growing number of institutional repositories that are OAI-PMH-compliant and consequently harvestable by service providers. Currently, the Registry of Open Access Repositories ROAR (2008) has over 1000 repositories registered worldwide, of which 536 are based at research institutions. These 536 archives hold a total of 2,309,512 records, averaging 5087 records each with a median figure of 938. In terms of the two million or so peer reviewed research articles published on a yearly basis, this represents a small but growing part of the total output. The graph shown in Figure 1 shows the rapid growth of institutional archives to March 2008 (Registry of Open Access Repositories (ROAR) 2008).

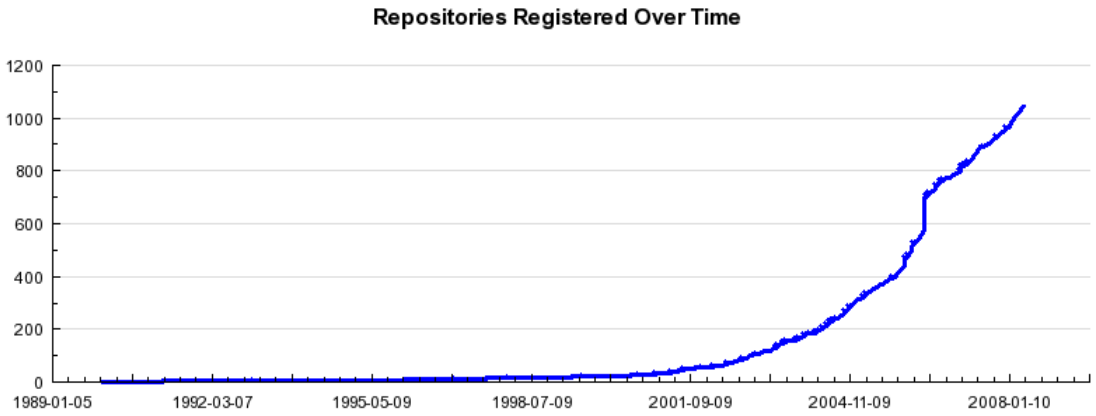

Figure 1. Repository numbers (ROAR 2008)

These repositories may be subject based like RePEC (Research Papers in Economics) or the physics archive arXiv, or may be more general in nature like the DEPOT or may be found at an institutional level. Authors may, of course self-archive their work to a home or departmental web page. OAIster, which harvests metadata is a union catalogue of digital sources hosted at the University of Michigan; “to establish a broad, generic retrieval service for information about publicly available digital library resources provided by the research library community” (About OAIster, 2007). Repositories make their records available to OAIster, where they harvest “their descriptive metadata (records) using OAI-PMH” (About OAIster, 2008). This service currently harvests from about 900 contributors and contains over 15 million records, which can be searched by author, title, language or subject and by resource type. Similarly OpenDOAR (About OpenDOAR, 2007), hosted by the University of Nottingham, facilitates access, worldwide, to institutional repositories. It is part of SHERPA (Securing a Hybrid Environment for Research Preservation and Access). Initially a directory of OA repositories, it now offers a trial service to search the contents of the repositories that it lists (OpenDOAR, 2006). Unlike OAIster, OpenDOAR does not search the repositories’ metadata even if they are OAI-PMH compliant, but “relies on Google's indexes, which in turn rely on repositories being suitably structured and configured for the Googlebot web crawler” (SHERPA news, 2006). In contrast general search engines like Google and Google Scholar will search non-compliant home and departmental web pages as well as compliant OAI repositories.

Those undertaking research to see if there is an OA citation advantage have used differing strategies to find OA articles. Much research has focussed on the citation advantage of articles deposited in arXiv that, on deposit, immediately become OA, compared to articles in the same subject that are not deposited and so remain TA (Kurtz *et al*. 2005: Moed 2007). Other authors have used various general search engines, either manually or by using robots to search the web for OA articles (Antelman 2004: Hajjem, Harnad, & Gringras (n.d.)). Whilst no one search engine can say that it indexes and searches all of the web, Google has established itself as the most frequently used general web search tool, with Google Scholar appearing as an addition to the Google stable of products focussing on scholarly materials. To a certain extent, Google and Google Scholar return many exclusive hits, returning differing results even when the same search terms are used.

In the case of Google and Google Scholar there have been many articles that have reviewed their performance and coverage. Google Scholar, in particular has had its coverage and citation structure reviewed by many authors. Jacsó has scrutinised closely the performance of Google Scholar, whilst in general highly critical, he does concede that "GS is good for locating relevant items, leading users some of the time to an open access version of a document, but it is not appropriate for bibliographic studies" (Jacso 2006, p. 307). Markland (2006) examined the effectiveness of both Google and Google Scholar at retrieving a defined set of items using keywords and title searches taken from 26 institutional repositories in the UK. Between them, using a title search, they found 25 out of the 26 items, with Google itself being the more successful finding all 25. Google Scholar found 17 items from within the repositories and found a further three items outside of the repositories. In contrast, when the repositories were searched directly using key words or titles taken from their own records, three items were not found. Walters compared the performance of Google Scholar to seven other databases (Academic Search Elite, AgeLine, AricleFirst, GEOBASE, POPLINE, Social Sciences Abstracts and Social Sciences Citation Index). He used a reference set of 155 articles on later life migration and found that Google Scholar found 93% of them, covering 27% more than Social Sciences Citation Index, its nearest rival.

Earlier work by the authors using Google Scholar in a pilot study carried out in late 2005, were disappointing, but subsequent work showed it to be more successful. It is assumed that in the intervening period between the pilot study and this research that the search capabilities of Google Scholar have been enhanced. This seems to be borne out by recent comments from Jacsó (2008) who notes the increase in coverage of Google Scholar whilst still, however, deploring its software. On this basis, Google, Google Scholar, OAIster and the OpenDOAR service were used in combination as the search tools for finding OA versions of journal articles.

Methodology

A random sample of 628 articles was taken from the 10,119 that appeared in the 112 ecology journals listed in the 2005 Journal Citation Reports that were published in 2003. A purposive sample of 966 articles was taken from 21 mid-impact economics journals appearing in 2003 and 925 articles were taken from high impact sociology journals that appeared in 2004. The bibliographic details of each of the articles were taken. The four search tools, OAIster, OpenDOAR, Google and Google Scholar were identified as being likely to find as many OA articles as possible. The search for OA articles was conducted by entering the article's title as a phrase in each search tool. As the primary purpose of the research was to locate OA articles, the search sequence was designed to be progressive rather than exhaustive of each search tool, starting with OAIster, and then OpenDOAR, followed by Google Scholar, and finally Google. OAIster and OpenDOAR were always searched. If no hits were found using these two, then Google Scholar was searched; if Google Scholar also did not yield a result, then Google was also interrogated.

Results and discussion

Of the 2519 articles selected, 967 (38.39%) were found to have OA versions on the World Wide Web. Table 1 shows how these 967 articles are broken down by their OA status and subject.

Table 1. OA status by subject

Subject	Total Articles	% OA	% TA
Ecology	628	34.39	65.61

Economics	966	54.45	45.44
Sociology	925	24.32	75.68

Given the search protocol adopted, the results for Google and Google Scholar cannot be said to reflect the absolute potential of either of them. However, taken together, they jointly found 76.84% of the articles. The percentage of records found for each search tool was; Google 8.79%, Google Scholar 68.04%, OAIster 2.38%, OpenDOAR 11.17% and where OAIster and OpenDOAR retrieved the same article, their combined score was 9.62%. Table 2 gives a more detailed breakdown of hits by subject and the search tool which found them. The hits for OAIster and OpenDOAR appear in columns four, five and six, columns four and five give exclusive hits and column six gives hits where both search tools have found the same record.

Table 2. Break down of OA hits by subject and search tool

Subject	Google	Google Scholar	OAIster	OpenDOAR	OAIster & OpenDOAR	Total
Ecology	20	194	2	0	0	216
Economics	32	287	13	108	86	526
Sociology	33	177	8	0	7	225
Total	85	658	23	108	93	967

Google Scholar was much more successful than OAIster and OpenDOAR, whose overall success was relatively poor. OAIster and OpenDOAR, however, could be considered useful search tools, finding 39.35% of the hits for economics. It is notable that in sociology, which had the smallest percentage of OA articles overall, that the majority of them were found using Google and Google Scholar, suggesting that those who do self-archive their work, in this subject at least, are not using repositories that can be found by using OAIster and OpenDOAR.

When the OA articles were broken down by first author affiliation, North America was found to be the region from which most articles originated. By subject, the percentage article counts from North America were, ecology 61.97%, economics 70.37% and for sociology 77.53% (data not shown).

There are major variations in OA hits when broken down by the search tool which found them and by first author affiliation. Table 3 shows the percentage of hits by each search tool. North America had the highest percentage of hits using Google and Google Scholar with the UK having the lowest percentage.

Table 3. Percentage OA hits by region and search tool.

Search tool	N America	Europe*	UK	Rest of World	Total
Google	4.34	2.59	0.93	0.93	8.79
Google Scholar	39.09	14.37	5.89	8.69	68.04
OAIster	1.24	0.52	0.41	0.21	2.38
OpenDOAR	4.76	3.62	1.65	1.14	11.17
OAIster & OpenDOAR	4.76	2.59	0.83	1.45	9.62
Total	54.19	23.68	9.72	12.41	100.00

* Does not include the UK

When OA hits were further examined by subject, their ranking by combined Google and Google Scholar were ecology 99.07%, economics 60.65%, and for sociology 93.33%. Given that both OAIster and OpenDOAR list the economics database RePEc among the sources from which they collect articles records, it is not surprising that there was a reasonable number of hits in this subject when using these two search tools. It is notable that OpenDOAR is overall more successful than OAIster in finding OA economics articles, presumably because it is searching RePEc, other repositories which allow Google's robots access.

As part of the first of the two studies undertaken by Norris, Oppenheim and Rowland (in press) they also examined, and briefly reported the success of the four search tools to find OA articles for the same subjects (including mathematics). Articles for this first study were taken from high impact journals from 2003. When the first study data is compared to the second, there are some notable differences. The second sample of high impact sociology articles were taken from 2004 and the percentage of hits dropped for *Google Scholar*, but rose for *Google* between the first and second study. Overall, their combined share of the hits for sociology drops from 98.37% to 93.33% to the benefit of OAIster. This is in contrast to the hits in ecology, where the combined *Google* score was 96.27% rising to 99.07% for the second study. Given that institutional repositories are more likely to be found at the more successful institutions (Directory of World Repositories 2008) and that the sample for the second round ecology data was randomly taken and hence more likely to come from a range of different institutions, it could be argued that it is more likely that the authors would self-archive to their own websites if repositories were not available at their own institution. However, for economics, where OpenDOAR was particularly successful, the combined scores for *Google* drops from 78.76% to 60.65% giving 39.35% of the

share of the hits to OAIster and OpenDOAR in the second study, perhaps mirroring the growing success of these harvesters.

The relative success of OAIster and OpenDOAR is attributed to their harvesting the metadata from RePEc and the need for academics to share informal research results in general symposia and in working paper series. Antelman (2006, p.89) examined self-archiving practices within the social sciences, taking approximately 2000 articles from 22 high impact journals from 11 different publishers with varying self-archiving policies, including economics and sociology. For economics, she found an overall rate of self-archiving of 59% and for sociology 24%. For the two samples taken here, the rate for the economics' first study data was in the order of 65% and for the second round data 54.45% and for sociology 21% and 24.32% respectively, a noticeably similar result. Antelman goes on to explain the overall level of self-archiving as characteristic of the discipline, for the social sciences this is one where authors are less reliant on a culture of sharing information for example in the exchange of preprints. Economics, however, is characterised as a discipline with a higher degree of mutual dependence where working papers are shared through repositories with other authors. Apart from the RePEc there are few disciplinary depositories for the social sciences. Hence, there is little difference between the results between the first and second round studies for sociology, with the OA hits being found almost exclusively by *Google* and *Google Scholar* and with few academics archiving to any sort of repository.

Bergstrom and Lavaty (2007) used Google, Google Scholar and OAIster equally to help them find OA articles in economics and political science. From a sample of 703 economics articles, they could find most OA articles using Google, with Google Scholar finding some ten-percentage points less than Google. They found, using OAIster about 25% of their sample articles. This is a similar result to those found here in the second study, where 18.82% of the articles were located by searching OAIster. RePEc provided 27% of the articles, which is in itself, an interesting result given that OAIster lists RePEc as one of the sources it trawls. When the holdings of the two sources are compared, it is clear that not all the records available from RePEc are reported by OAIster and presumably, this explains the difference, although it is very unlikely that there were any items discovered in RePEc that could not be found by using Google or Google Scholar.

Conclusion

Despite the increasing number of institutional repositories and their harvesting by such services as OAIster, it is apparent that finding OA articles in the four subjects selected here was greatly facilitated by the use Google and Google Scholar. What is clear is that whilst OAIster and OpenDOAR are reliant for the majority of their content from institutional repositories, it appears that the majority of authors in this sample at least are not self-archiving their work to them or if they do, it is to non-compliant or unregistered repositories or to locations not accessible to these search tools. Alternatively, there may be of lack coverage by OAIster and OpenDOAR for other as yet unidentified reasons. Authors prefer, it seems, when they do self-archive their work, to do so to their personal or departmental web page where metadata harvesters such as OAIster cannot readily find them, but where Google and Google Scholar can. Those wanting to find OA articles, it is suggested, are more likely to find them using Google or Google Scholar rather than OpenDOAR or OAIster.

References

About OAIster. (2008), Available at: http://*OAIster*.umdl.umich.edu/o/*OAIster*/about.html

About OpenDOAR. (2006), Available at: http://*OpenDOAR*.org/about.html

Antelman, K. (2004), "Do open-access articles have a greater research impact", *College and Research Libraries*, Vol. 65 No 5, pp. 372-382.

Antelman, K. (2006) "Self-archiving practice and the influence of publisher policies in the social sciences", *Learned Publishing*, Vol. 19 No 2, pp. 85-95.

Bergstrom, T. and Lavaty, R. (2007) "How often do economists self-archive?", available at: http://repositories.cdlib.org/ucsbecon/bergstrom/2007a/

Directory of World Repositories. (2008), available at: http://www.webometrics.info/premierleague.asp

Hajjem, C., Harnad, S. and Gringras, Y., (n.d.), "Ten-year cross-disciplinary comparison of the growth of OA and how it increases citation impact", available at: http://eprints.ecs.soton.ac.uk/12906/

Harnad, S. and Brody, T. (2004), "Comparing the impact of OA (OA) vs. non-OA articles in the same journals", *D-Lib Magazine*, Vol. 10 No 6, pp.1-5. available at: http://mirrored.ukoln.ac.uk/lis-journals/dlib/dlib/dlib/june04/harnad/06harnad.html

Jacso, P. (2006), "Deflated, inflated and phantom citation counts", *Online Information Review*, Vol. 30 No 3, pp. 297-309.

Jacso, P. (2008), "Savvy searching Google Scholar revisited", *Online Information Review*, Vol. 32 No 1, pp. 102-114.

Kurtz, M., Eichhorn, G., Accomazzi, A., Grant, C., Demleitner, M., Henneken, E and Murray, S. (2005), "The effect of use and access on citations", *Information Processing and Management*, Vol. 41 No 6, pp. 1395-1402.

Markland M., (2006), "Institutional repositories in the UK: what can the Google user find there?, *Journal of Librarianship and Information Science*, Vol.38 No 4. pp. 221-228.

Moed, H., (2007), "The effect of 'Open Access' upon citation impact: An analysis of ArXiv's Condensed Matter Section", *Journal of the American Society for Information Science and Technolog*, Vol. 58 No 13, pp. 2047-2054.

Norris, M., Oppenheim, C. and Rowland F. (In Press), "The Citation Advantage of Open Access Articles" *Journal of the American Society for Information Science and Technology*.

Registry of open access repositories (ROAR), (2008), available at: http://roar.eprints.org/index.php?action=generate_chart

SHERPA News. (2006), available at: http://www.sherpa.ac.uk/news/opendoaroct06.html

Walters, W. (2006), "Google Scholar coverage of a multidisciplinary field" *Information Processing and Management*, Vol. 43 No 4 pp. 1121-1132.