The Citation Advantage of Open Access Articles

Michael Norris, Charles Oppenheim and Fytton Rowland

Department of Information Science, Loughborough University, Loughborough, LE11 3TU, UK. E-mail: {M.Norris2, C.Oppenheim, J.F.Rowland}@lboro.ac.uk

Tel +44 (0) 1509 223065 Fax +44 (0) 1509 223053

Corresponding author Charles Oppenheim

Abstract

Four subjects, ecology, applied mathematics, sociology and economics, were selected to assess whether there is a citation advantage between journal articles that have an open access (OA) version on the Internet compared to those articles that are exclusively toll access (TA). Citations were counted using the Web of Science and the OA status of articles was determined by searching OAIster, OpenDOAR, Google and Google Scholar. Of a sample of 4633 articles examined, 2280 (49%) were OA and had a mean citation count of 9.04, whereas the mean for TA articles was 5.76. There appears to be a clear citation advantage for those articles that are OA as opposed to those that are TA. This advantage, however, varies between disciplines, with sociology having the highest citation advantage but the lowest number of OA articles from the sample taken and ecology having the highest individual citation count for OA articles but the smallest citation advantage. Tests of correlation or association between OA status and a number of variables were generally found to be weak or inconsistent. The cause of this citation advantage has not been determined.

Introduction

Academics are frequently judged, in part at least, on the quality of their published research. The greater the impact of that research as counted by, for example, the number of citations it receives, the better, it is believed, is the quality of the work (van Leeuwen et al. 2003, pp. 262-263). Receiving many citations for academic research generally correlates strongly with academic success; an analysis of Nobel laureates and their citation counts by Garfield (1979, pp. 63-64) and Opthof (1997, p. 2), although tenuous, gives some credibility to the idea that the two are linked. Likewise a similar ranking by Hirsch (2007, pp. 16569-16572) using his *h-index*, which uses

article citation counts, has been successfully used to identify and rank prominent physicists.

In recent years, it has become possible for authors to self-archive an electronic version of their work in a variety of locations from personal web pages, to a disciplinary archive or to an institutional repository. In so doing, authors make their work open access (OA) and freely available to anyone who has Internet access. Toll access (TA) articles (often known as closed access articles) remain behind subscription barriers and are only accessible by a personal or institutional subscription. If it can be shown that self-archived OA research often receives more citations than closed access research, then a convincing argument can be made to persuade researchers to self-archive their work. A number of studies (Antelman, 2004; Eysenbach, 2006; Harnad & Brody, 2004; Lawrence, 2001) have shown that those authors who make their work OA will receive more citations and hence achieve greater impact than those authors whose articles remain behind subscription barriers. Despite this advantage, few authors, however, self-archive their work (Swan & Brown, 2005, pp. 62-68); the self-archiving rate of authors seems to be at best, around 15% (Hajjem et al. 2005; Sale, 2005). However, increasing citation impact is not the only benefit of self-archiving; work that is freely available for anyone to read should increase research access and impact, and allow those that fund research through their taxation access as well (Harnad, 2006, p. 73).

There has been much discussion on the causes of this citation advantage. Kurtz et al. (2005), Davis and Fromerth (2006) and Moed (2006) are not convinced that simply making an article open access is sufficient cause for any increased citation counts. Rather, they suggest that authors may self-archive their better quality work, and because some articles are made available as preprints before publication, they have a longer period in which to attract citations. Metcalfe (2006, p. 549), however, thinks that, in solar physics at least, higher citation rates are not the result of authors archiving their higher quality papers, or necessarily that better authors more readily archive their work. What is evident, however, despite what might be the cause of any OA citation advantage, is that the evidence accumulated so far indicates that those authors who make their peer-reviewed work more visible by self-archiving their articles receive more citations than those who do not.

Previous research

Open Access Citation Advantage

Lawrence (2001) was the first to show that conference articles that were OA and freely available on the World Wide Web were more frequently cited than articles that were offline. Since Lawrence's pioneering work, there have been a number of studies that have demonstrated a similar citation advantage (Antelman, 2004; Eysenbach, 2006; Hajjem et al. 2005; Harnad, 2004). Harnad and his colleagues (Hajjem et al. 2005; Harnad & Brody, 2004; Harnad et al. 2004) have carried out large-scale trials where they examined the citation counts of OA and TA articles from the same journals from a database of 14 million articles. In physics and in a range of other subjects, they have found a significant citation advantage for those articles that were OA. In these studies, they identified OA versions of articles either by trawling the web using a computer algorithm or by taking self-archived versions from a disciplinary archive and then compared the citation counts of both OA and TA versions. In contrast to this approach, Antelman (2004) selected four subjects and a relatively small number of articles and manually identified OA versions of articles and their respective citation counts. Again, there was a significant citation advantage for those articles that were OA, but with noticeable variations between subjects.

These two approaches counted the citations from work that was made available by authors by self-archiving their work where it could be accessed. Eysenbach (2006) took a selection of articles that appeared in a single journal (*Proceedings of the National Academy of Sciences*), some of which were TA and others which were OA by virtue of their authors paying for their publication, even though after a six months moratorium all articles appearing in the journal become OA. The OA articles were available from the publisher's web site. Overall, Eysenbach found, that even when taking into account factors such as the number of authors, country of origin and discipline, that OA articles were still twice as likely to be cited as the non-OA articles appearing in the same journal. Eysenbach (2006, p. 697) also suggested, that those OA articles that were hosted on the publisher's website were more heavily cited than some of the original TA articles which were subsequently made OA by being self-

archived by their authors elsewhere. Given the status of the *Proceedings of the National Academy of Sciences* as a prestigious journal with a high rejection rate and high impact, the results found by Eysenbach are not necessarily applicable to journals containing articles of a more variable quality.

Causation

Lawrence (2001) found a significant correlation between conference papers that were available online and their greater citation counts as compared to offline papers, he was unable to identify the cause of this correlation, although he did suggest in his analysis of papers from the same conferences that "online articles are more highly cited because of their easier availability" (p.521). This uncertainty as to the cause of any OA citation advantage has led to speculation that there are other reasons for this advantage other than simply that the article is OA. Several possible reasons have been suggested, including article age, number of authors, the quality of articles, and the status of authors or of their institutions. Kurtz et al. (2005) looked at three possible reasons for this advantage in astronomy. They found evidence for an early access (EA) effect caused by an article preprint being made freely available prior to journal publication, and a self-selection bias (SB) where the author has self-archived their better work; but were unable to find a specific open access (OA) effect. They concluded, in astronomy at least, that this lack of an OA effect was probably caused because authors in astronomy must already have access to the literature in order for them to carry out and report their research. Wren (2005) found that articles from high-impact biomedical journals are more likely to be found at non-journal websites, suggesting, possibly, that these are better quality papers which are made more readily available by their authors.

Although working with a small sample and a single journal, Eysenbach (2006, p. 697) thought that "…publishing papers as OA articles on journal sites facilitates knowledge dissemination to a greater degree than self-archiving…". This view that self-archiving is less efficient in terms of accruing citations is contentious. Harnad (2007a) has reported the preliminary findings in which his team have quantified four components of a citation advantage from biomedical articles that had been self-archived. It was shown "that each of the four factors contributes an independent,

statistically significant increment to the citation counts" (Harnad, 2007b). The largest increment to any OA citation advantage was the number of years since publication, followed by the impact factor of the journal in which article appeared, the number of authors of the article and that, although the smallest contributor, the fact that the article had been self-archived. Davis and Fromerth (2006), taking article-level data from four mathematics journals, 18.5% of which had been deposited in the arXiv archive, could only find reasonable evidence to support a quality differential where more highly citable articles had been deposited in arXiv. Using a similar approach, Moed (2006) looked to estimate the early view and quality bias effect on the citation impact of preprint articles found in the condensed matter section of arXiv. Taking a large sample from 24 journals of deposited and non-deposited articles, Moed found a strong early view and quality bias, but was unable to find a general open access citation advantage. In a recent review of the literature Craig et al. (2007) could find no evidence of an OA effect, rather they suggested that an article's OA "status alone had little or no effect on citations". The authors supported the work of Moed (2006) which they regarded as the most rigorous to date and if replicated, they argued, this might help determine the generality of the results found by Moed.

Metcalfe (2006) compared the citation rates of solar physics articles made freely available in arXiv or in the Montana State University archive found a citation advantage compared to those articles that had not been deposited. More interestingly, Metcalfe (2006, p. 551) suggested that this effect is due to improved visibility rather than authors selecting their better papers to archive. Metcalfe noted the results of Schwarz and Kennicutt (2004) who found that astrophysics conference papers posted to arXiv were cited twice as frequently as those that were not. Metcalfe sampled a set of conference proceedings from solar physics and found a comparable boost in citation rates for those that had been self-archived to arXiv. Metcalfe (2006, p. 551) suggests that conferences in astronomy and astrophysics are not affected by a quality bias because they are the place to publish work in progress or details that are not significant enough by themselves to merit publication in a peer-reviewed journal, and so he concludes are of lesser quality.

Research background

The present study extends the range of subjects examined by Antelman (2004) and on a smaller scale supplements the work of Hajjem et al. (2005). Four subjects are examined to see if there is an OA citation advantage from articles published in a range of high impact journals. Subject differences are investigated in their level of OA and citation advantage, and the sources of the citations are broken down into, for example journal self-citations and author self-citations to examine any effect on OA advantage. Some measure of causation of the OA effect is made by examining correlations between the number of authors and their articles, by examining the country of origin of OA authors and their particular subjects.

Methods

Harnad and Brody (2004) argued that the best way to test for a citation advantage for OA articles is to compare the citation counts of individual OA and non-OA articles appearing in the same non-OA journal. This process is dependent on these articles accruing citations which can be counted and compared. Given that as many as 50% (Garfield, 2005) of articles are not cited at all, choosing high impact factor (IF) journals from which to take a sample of articles should increase the likelihood that there is a substantial citation count for both TA and OA articles. The impact factor of a journal is calculated by taking the number citations to all documents published in a journal over a consecutive two year period and then dividing this count by the number of citable items from that journal during the same period (Garfield 1979). This metric is calculated annually by Thomson ISI which then ranks the journals it indexes on this basis within subject categories. This data is made available by Thomson ISI in its *Journal Citation Reports* (JCR) and the bibliographic data and accruing citation counts associated with the articles within the indexed journals appears in its citations indexes found in its Web of Science (WoS) database. Moed (2005, pp. 113-114) describes the advantages of using the WoS citation indexes, not least of which is the frequency with which they have been used by other researchers. Coverage of subjects within WoS varies between disciplines, with the sciences predominating. The database is, however, sufficiently broad to enable records to be collected from a range of subjects. Disciplines vary in their level of citedness and the coverage of the subject

by journals as opposed to coverage by monographs; sociology is a particular example where monographs play a significant part in scholarly communication (Nederhof, 2006, pp. 83-86).

Four subjects were selected for examination; these were: applied mathematics; ecology; economics; and sociology. The subjects represent a selection from the sciences and the social sciences. Moed (2005, pp. 126-131) ranks the ISI coverage of these subjects by the number of references made to articles published in a sample of up to eight ISI source journals relative to the total number of references which appear in those source journals. Ecology has the highest coverage at 64%, followed by applied mathematics at 54%, economics at 47% and sociology at 27%. Sociology emerges typically, as Hicks (2004, pp. 480-484) describes, as a discipline which is biased towards publishing a significant amount of material in monographs, leading to a lower number of citations to core journals; that is, authors cite significantly fewer journal articles than, say, in ecology. Both sociology and economics have a relatively high national orientation, as defined by the share of the papers from the country most frequently publishing in a journal, relative to the total number of papers published in the journal (Moed, 2005, p. 131). This suggests that articles published in these subjects are of interest locally to the country rather than being of international significance. Generally speaking, the 'harder' the science, the more likely that scholarly communication will be through journals that are more international in scope. For example, chemistry has an ISI coverage of 84% and a low national orientation (Moed, 2005, pp. 129-131).

A deliberately purposive sampling approach was adopted in the selection of journal titles, since the aim of the work was to assess whether there was an OA citation advantage, and not to determine whether the distribution of OA articles was random or otherwise. By their very nature, high impact journals attract a greater number of citations than their lower impact counterparts and are more likely to have articles from leading academics and their institutions. Whilst the sample was clearly biased in favor of high impact journals, the citation counts attracted by articles from them, would reflect the citing behavior of that particular discipline and hence allow comparisons between them, and a measure of any OA citation advantage if present. A sample of high impact journals from the four disciplines as defined by subject

categories in the 2005 edition of the JCR was taken. Appendix A gives the journal titles, their subject category and impact factor. Checking the publisher's websites of the 65 chosen journals showed that they were all available electronically and with the exception of three, they were also available in print format. The status of each journal was checked to ensure that it was completely TA and that it was not available in OA form after any embargo period.

The bibliographic details and citation counts of all the articles published in 2003 for the journals selected were taken from the citation indexes on WoS. In the case of sociology, because of the high number of book reviews in the journals, a small number of article records and their citation counts were taken from the latter part of 2002 to increase the sample size. This approach was adopted to give rough parity in terms of journal impact factor between the subjects selected without having to take article records from mid-range impact journals and to give a similar sample size to the other subjects. Letters, editorial material and corrections were excluded. In this process, 4633 articles were identified. Moed (2005, p. 95) demonstrates that, in general, the peak in citation frequency for journal articles is usually achieved by about the third year after publication, but there is some variation in this between disciplines. The citation counts from these records were broken down into journal and author self-citations and other author citations. Finding OA versions of TA articles on the World Wide Web can sometimes be difficult and misleading. Searches using computer algorithms, although manageable, can give some false drops (Hajjem et al. 2005). Similarly, manual searches using search engines can lead unwittingly to publisher's websites, especially if searches are made from subscribing institutions where the IP address is recognised by the publisher's server. Additionally, given the many spurious hits that a search engine can give, finding OA versions of articles may prove difficult, even if in fact they are there.

The majority of earlier studies looking for an OA citation advantage have searched the web either manually or by trawling using a computer algorithm. Carr (2006) reports that, of those making searches on the WWW for articles, over 96% of these searchers get to the Eprints repository at Southampton University by using Google (76.05%), Google Scholar (15.25%) and Yahoo (4.93%). The use of Google Scholar to find OA articles has not as yet, we believe, been reported in the literature. In a

number of papers, Jacso (2005a, pp. 208-214; 2005b, pp. 1537-1547) has reviewed Google Scholar, from which it is clear that it is not currently an adequate tool for citation counting as such, but it is useful in locating OA versions of journal articles. The view that Google Scholar has limitations is also shared, generally, by others who have also found significant omissions in the coverage and recall from this database (Myhill, 2005; Notess, 2005). In recent research conducted by Meho and Yang (2007) they found that Google Scholar was particularly strong in its coverage of conference proceedings and international non-English language journals, despite its evident limitations. Whilst these criticisms of Google Scholar are fair, a recent test of its recall by Norris and Oppenheim (2007) found that in terms of finding links to individual articles taken from a common database of articles from the social sciences, Google Scholar had a hit rate of 87%, compared to 88% for WoS and 95% for Scopus. Additionally, in a small pilot study, a hundred article records were taken from different subjects and used as a sample in the search engines Yahoo, Google and Google Scholar. Each article title was entered as a phrase in each of the search engines. Yahoo was not as successful at finding hits as was Google or Google Scholar, nor did Yahoo find any hits in addition to those found by Google or Google Scholar. However, Google and Google Scholar had little overlap and did return unique hits that found OA article records, suggesting that in combination they would yield more OA results than if used singly.

There has been a significant growth in the number of institutional repositories into which authors can self-archive their journal output and make it freely available, thus broadening the availability of OA material. These repositories can have their records harvested by service providers such as OAIster. OAIster is a union catalogue of digital sources hosted at the University of Michigan (OAIster 2007). Repositories make their records available to OAIster, who then harvest their descriptive metadata using OAI-PMH (the Open Archives Initiative Protocol for Metadata Harvesting). OAIster currently harvests records from over 700 repositories and contains over 10 million records, which are searchable from a single access point. OpenDOAR, (OpenDOAR 2007) hosted by the University of Nottingham, is a similar centralised access point to worldwide institutional repositories. OpenDOAR, initially a directory of open access repositories, now offers a trial service to search the contents of the repositories that it lists. Unlike OAIster, OpenDOAR does not search the repositories'

metadata even if they are OAI-PMH compliant, but relies on Google's indexes, and repositories being suitably structured for the Googlebot web crawler. Both of these service providers enable access to well known repositories, including the major subject repositories such as arXiv and RePEc. RePEc (http://repec.org/docs/RePEcIntro.html) is “the world’s largest collection of scholarly information for economics” and its database holds details of 12,700 professionals and 10,250 institutions associated with economics”. For our research, Google, Google Scholar, OAIster and the OpenDOAR service were used in combination as the search tools to maximise the findings of OA versions of journal articles.

To determine the OA status of the 4633 articles identified, article titles were entered as a phrase search in OAIster, OpenDOAR, Google Scholar and Google. The search sequence started with OAIster,and proceeded through OpenDOAR, and if necessary Google Scholar, and finally Google. OAIster and OpenDOAR were always searched; if these two failed to produce a hit, then Google Scholar was searched, and finally, if necessary, Google was used. OA articles were those articles that could be identified as being completely freely available from an individual’s website, a departmental site, subject repository or an institutional repository. Such finds included preprints, postprints and drafts, and were counted as OA if both the title of the article and the article’s authors were the same as that found in the journal in which the article was published. Hits that led to a publisher’s website were discounted, as in general a subscription is needed to access the full text.

Findings

Of the 4633 articles tested, 2280 were OA and in total, the 4633 articles accrued 34,156 citations between them; 489 of the articles did not receive any citations at all. Of the 489 articles that did not receive any citations 309 (63.2%) were TA and the remaining 180 (36.8%) articles were OA. Overall, including zero citation count records, the gross mean citation count for those articles that were OA was 9.04 compared to 5.76 for the TA articles. This represents an OA citation advantage of 57% (OA-TA/TA citation counts). When journal and author self-citations were excluded, the mean citation counts for the two article sets were OA 6.47 and TA 3.93, an OA citation advantage of 64%. Figure 1 shows the proportion of OA/TA articles found by subject.

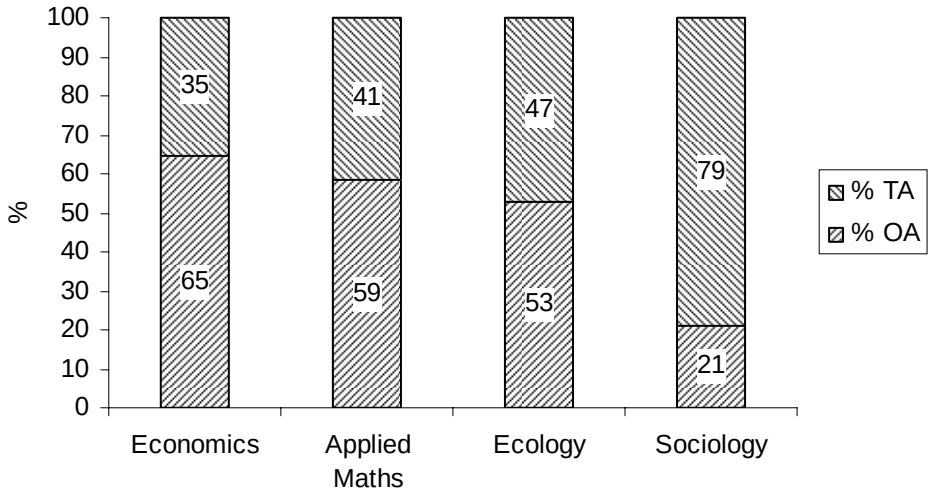

FIG. 1. Proportion of OA/TA articles by subject

The mean citation counts of the two populations, both gross and net of journal and author self-citation counts, were compared using the independent 2 sample *t*-test; the result showed them to be from populations with different means ($p<0.001$). Similarly when the test was conducted for each of the four subjects, the same result was found. Although the frequency distributions of citation counts are usually skewed, such distributions can, when the sample size is sufficiently large, be considered to have means normally distributed in accordance with the central limit theorem (Hinton, 2004, p. 55). The non-parametric Kolmogorov-Smirnov Z test was also used to confirm that the two groups, OA and TA articles, were not drawn from the same populations, and in every instance, the test confirmed that this was the case.

Table 1 gives the gross citation counts for the four subjects; the OA advantage ranges from 88% for sociology to 44% for ecology.

Table 1. Gross citation counts

	TA citations	TA articles	Avg citations TA article	OA citation	OA articles	Avg citations OA article	OA % advantage ±
Applied maths	1627	480	3.39	3518	678	5.19	53
Ecology	6240	553	11.28	10012	618	16.20	44
Economics	1716	402	4.27	5099	739	6.90	62
Sociology	3961	918	4.31	1983	245	8.09	88
Total	13544	2353	5.76	20612	2280	9.04	57

This advantage is maintained when journal and author self-citations are removed, leaving just the citations from other authors writing in journals other than the cited article journals; this is shown in Table 2.

Table 2. Citation count net of author and journal self-citations

	TA citations	TA articles	Avg citations TA article	OA citation	OA articles	Avg citations OA article	OA % advantage ±
Applied maths	854	480	1.78	2065	678	3.05	71
Ecology	4246	553	7.68	7058	618	11.42	49
Economics	1245	402	3.10	4056	739	5.49	77
Sociology	2891	918	3.15	1568	245	6.40	103
Total	9236	2353	3.93	14747	2280	6.47	65

Sociology has the highest citation advantage for those articles that are OA, but overall as shown in Figure 1 its authors make the smallest number of their articles OA. Ecology, with the second lowest rate of open access, has the highest citation count for its articles. Economics has the highest rate of OA adoption and is second to sociology in citation advantage.

Figure 2 shows a breakdown of the OA citation count by the four types identified. Journal author self-citations (JASC) are where the cited author is citing themselves and writing in the same journal as the original cited article. Journal self-citations (JSC) are citations where authors other than the original article author have cited the article within the same journal. Author self-citations (ASC) are where the authors are citing themselves but are writing in a journal other than the journal in which their original article appeared. Finally, other citations (OC) are from authors unrelated to the original cited journal or any of its authors.

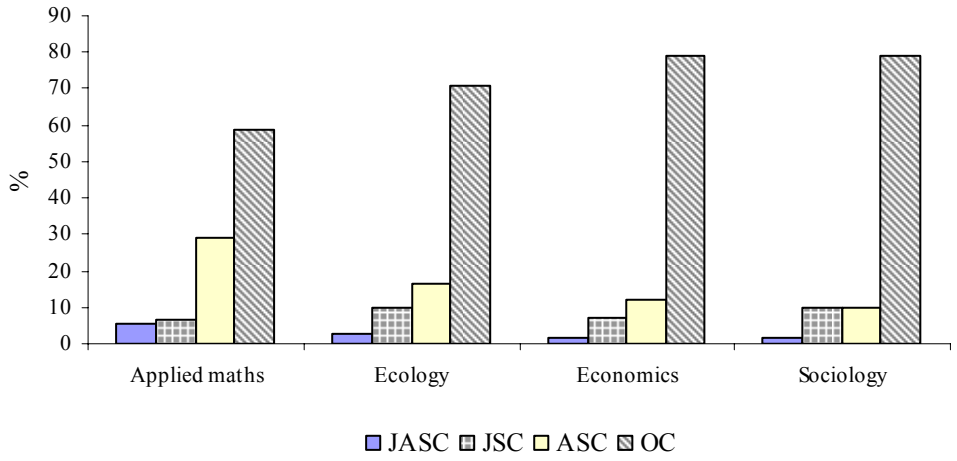

FIG. 2. Breakdown of OA citations by subject

Other author citations form the largest single category for all the subjects and author self-citation rate is highest in applied mathematics and lowest in sociology. The combined self-citation rates are 41% for maths, 29% ecology, and 20% for both economics and sociology. The mean number of journal and author self-citations for OA articles was 2.57, and for TA articles, this was 1.83. Consistent differences between the mean number of journal and author self-citations were also evident at subject level and are consistently greater for OA articles than TA articles; the mean OA/TA journal and author self-citation counts were respectively for ecology 4.78/3.61; economics 1.41/1.17; applied mathematics; 2.14/1.61; and sociology 1.69/1.17. These means were compared using the independent 2 sample t-test; the result showed all four to be from populations with different means ($p<0.001$).

The mean number of authors for all of the articles was 2.34. For all TA articles the mean number of authors was 2.21 and for OA articles this was 2.46. At subject level in every case OA articles had a slightly higher mean number of authors than TA articles. Table 3 gives a breakdown of the mean author counts by subject.

TABLE 3. OA/TA counts by country and subject.

		Subject and article count				
		Ecology	Economics	Applied Math	Sociology	Total
Open Access	N America	383	520	293	190	1386
	Europe	121	95	254	21	491
	UK	65	69	43	24	201
	Rest of World	49	55	88	10	202
Total		618	739	678	245	2280
Toll Access	N America	236	221	159	621	1237
	Europe	147	65	189	65	466
	UK	80	72	23	149	324
	Rest of World	90	44	109	83	326
Total		553	402	480	918	2353

The results of a Chi-square test ($\chi^2(8) = 88.83$, $p < .001$) showed there was a significant association overall between the number of authors and the OA/TA status of an article. However, the association between the number of authors and the OA status of an article showed that there was a tendency towards OA status only when there was more than one author. Hence, there is a strong association between single authorship and articles being TA. Of the 1356 single authored articles 61.36% were TA and the remaining 38.64% were OA. The situation is reversed for articles having more than one author, with these articles having a tendency towards OA status. For those OA articles having two to five authors, the differences between them and the TA articles is, however, less marked and ranges, dependent on the number of authors, from 53%-56% to the advantage of OA articles. This result however, breaks down at subject level where the association is not significant for ecology or sociology, but is for applied mathematics and economics ($p < 0.001$), although even this association is not entirely consistent throughout the range of author counts for the TA and OA articles.

Examining the origin of articles by first author affiliation shows that authors from North America provided the majority of articles, accounting for 57% of all the articles published in the 65 journal titles. Figure 3 gives a breakdown of the OA status of articles by country and subject. North America has the highest rate of OA (60.8%) followed by continental Europe (21.5%), with the UK and the Rest of the World at 8.8% and 8.9% respectively.

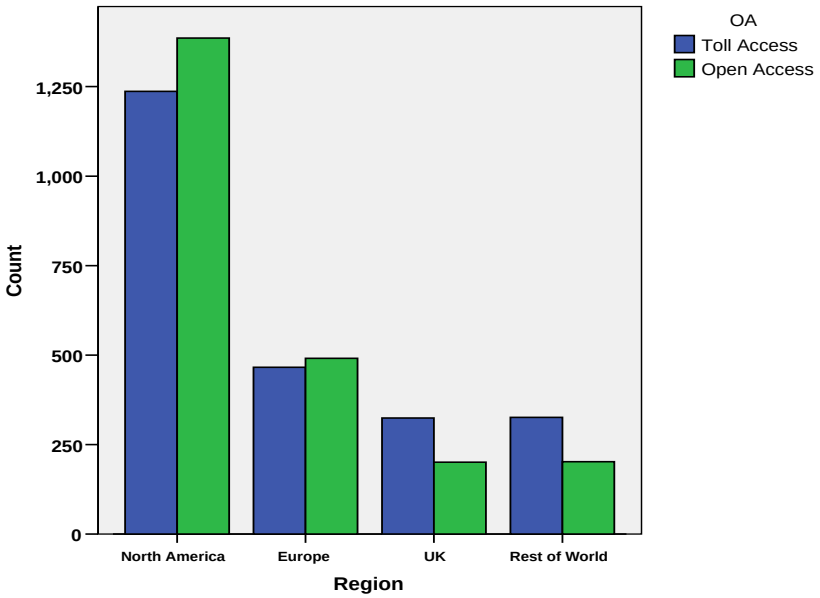

FIG 3. Number of OA/TA articles by region

The results of a Chi-square test ($\chi^2(3) =$, $p<0.002$) showed there was a significant association overall by region, subject and split between OA and TA articles. There is a tendency towards OA in North America and continental Europe, and a tendency toward TA in the UK and the Rest of the World.

The correlation between the number of authors and the total number of citations was tested. For OA articles, this was 0.19 and for TA articles, this was 0.21. Whilst these correlations were significant ($p<0.01$) given the relatively large sample size, the actual results were poorly correlated. Taking just journal and author self–citations and comparing this total to the level of authorship revealed no substantial differences between the two sets of data; the correlations were 0.309 and 0.288 for OA and TA articles respectively ($p<0.01$). Similarly, correlation between journal impact factor and the number of authors was for OA 0.16 and for TA 0.25.

Search engine success

As a by-product of the data gathering, the success of OAIster and OpenDOAR in retrieving OA versions of articles was measured. In comparison to the success of Google and Google Scholar, their success overall were relatively poor. Only in economics and applied mathematics could OAIster and OpenDOAR be considered a

relative success, finding 21% and 22% respectively of the hits between them. A breakdown of hits shows that, of the total 2280 OA items, only 14% overall were found between OAIster and OpenDOAR, with Google and Google Scholar finding the other 86%. A particularly useful feature of Google Scholar was the way that it grouped multiple finds of an article into a single hit and from it, if present, an OA article could readily be found without having to search several pages of records.

A difficulty faced by all web searches is the consistency of web links to, in this case, OA articles. Many articles which look to have viable OA web addresses have broken links, and hence were counted on the day of interrogation as TA even though on another search they may appear as OA. In the case of Google, its results are relatively consistent. In a study of web citations by Vaughan and Shaw (2005, p. 1078), the stability of their initial search results from Google were checked by subsequent repeat searches and found to be fairly constant. The other side of this problem is failing to find OA versions of articles when they are, in fact, available. This appears at first as a positive feature because not finding an OA article would suggest that the OA citation advantage is being understated. However, it can be argued that an OA article that is hard to find and remains unfound will have a lower citation count than those that are easily found, and by default will become coded as TA; if this effect applied in large numbers of cases, it would artificially widen the citation advantage of OA articles. This is an issue for all the research that has been undertaken so far, and it is argued here that searches for OA articles through two general search engines, the metadata of an international repository and a surrogate search of international repositories through Google's indexes have minimised this problem.

Discussion and Conclusion

Out of the 2280 OA articles identified, 86% were found using Google or Google Scholar; at 14%, the finds by OAIster and OpenDOAR were relatively modest, suggesting that the majority of open access articles are not deposited in institutional archives where either OAIster or OpenDOAR can find them. For these subjects at least, in these higher impact journals, the best strategy to find an OA article would be to use Google Scholar followed by Google and then use OAIster or OpenDOAR.

The results found in this work agree with earlier studies that have examined the broad citation advantage of a range of subjects where OA articles are dispersed across the Internet rather than confined to a single subject repository. Notably, the work reported by Antelman (2004), Harnad et al. (2004) and Hajjem et al. (2005) has given similar results. Other work which has found an OA citation advantage has either concentrated on the results from a single journal (Eysenbach, 2006) or particular subjects which have a preprint culture, such as high energy physics, and almost exclusively, in the case of arXiv, their own repository.

The results show a statistically significant difference in the mean number of citations that OA articles received when compared to TA articles. This is apparent for all of the subjects for both the gross citation count and when journal and author self-citations are removed. There are however, variations in both the degree of OA and the citation advantage within the four subjects, with sociology having the smallest number of articles that are OA but having the highest citation advantage. Similarly, Hajjem et al. (2005) reported from their large-scale study that sociology had, at 172%, the largest citation advantage from the ten subjects they examined, although unlike the result here, it also had the highest OA rate as well. In the results here, economics had, at 65%, the highest OA rate and this, it is suggested, is related to the frequency with article metadata is deposited in RePEc and the frequency with which this is found through OAIster and OpenDOAR and hence the ease with which the work can be accessed. Antelman (2004) reported an OA frequency for mathematics of 69% compared to a similar result for applied mathematics of 59% in the results here.

The majority of citations that the articles received were not author or journal self-citations, although in the case of mathematics a substantial number of them were (41%). In all subjects, however, OA articles were self-cited more frequently than TA articles. This however, does not account for the overall gross OA mean citation advantage over TA articles. Indeed the OA citation advantage is even more marked when self-citations are removed from both sets of counts.

Overall, there is a significant association between the number of authors an article has and whether it is OA or not; generally single authored articles are more likely to be

TA. The results, however, become inconclusive when considered at subject level, and for example, there is no significant association between author count and OA/TA status in sociology.

Most articles originated from North America (57%) when first author affiliation was used to identify their origin. Although a little mixed, there was a noticeable bias in favor of authors making there work OA in North America and continental Europe, this was evident at subject level as well. This was not the case for the UK or the Rest of the World where OA rates were generally lower; however, applied mathematics had almost consistently more OA articles than TA articles in all four regions, and the reverse was true for sociology. Despite this relatively poor position for sociology, its OA citation advantage was the highest, suggesting that where scholars can find what few articles are OA, they are cited heavily. In a similar finding to Antelman (2005, p. 377), who found in her sample that mathematics had the highest number of OA articles and that it also had the lowest citation advantage, our results show that applied mathematics had the second highest number of OA articles, but the third lowest citation advantage.

Other measures of association or correlation were generally inconclusive, leaving the issues of causation of any OA citation advantage unclear. Whilst there was an obvious association between single authors and TA status, this was much less decisive when there was more than one author. Likewise, measures of correlation between impact factor and OA status were found to be weak as was the correlation between the number of authors an article had, and the number of citations it received.
It is evident that the level of OA is subject dependent, and that within these subjects, there are different levels of authorship and citation practices thereby making it difficult to explain the cause of any OA citation advantage. The idea that early access to preprint articles is an explanation for OA citation advantage is not proved, since unlike articles posted to arXiv, the subjects we examined are less well served by a recognised subject repository, except possibly RePEc for economics. Likewise, solely ascribing the advantage to a quality bias is difficult to sustain, since with the exception of sociology, well over half of the articles were OA. As Harnad (2007b) suggests, it is likely to be combination of factors.

Although the reasons why there is a citation advantage for OA articles has still not been satisfactorily explained, it is clear that the advantage exists and occurs regularly across a range of subject areas. Further data collection is planned to investigate the possible cause of this advantage. This may allow some conclusions to be drawn on the reasons for any OA advantage.

References

Antelman, K. (2004). Do open-access articles have a greater research impact?. College and Research Libraries, 65, 372-382. Retrieved March 22, 2007, from http://eprints.rclis.org/archive/00002309/

Carr, L. (2006, October 22). Access to self-archive via Google Scholar. Message posted to http://listserver.sigmaxi.org/sc/wa.exe?A2=ind06&L=american-scientist-open-access-forum&D=1&O=D&F=l&S=&P=81868

Craig, I., Plume, A., McVeigh, M., Pringle, J., & Amin M. Do open access articles have greater citation impact? A critical review of the literature.

Retrieved July 21, 2007, from http://www.publishingresearch.net/Citations-SummaryPaper3_000.pdf.pdf

Davis, P., & Fromerth, M. (2006). Does the arXiv lead to higher citations and reduced publisher downloads for mathematics articles? Retrieved March 22, 2007, from http://arxiv.org/abs/cs.DL/0603056

Eysenbach G. (2006). Citation advantage of open access articles. PLoS Biology, 4, 692-698. Retrieved March 25, 2007, from http://biology.plosjournals.org/perlserv/?request=get-document&doi=10.1371/journal.pbio.0040157&ct=1

Garfield, E. (1979). Citation indexing – its theory and application in science, technology, and humanities. New York: John Wiley.

Garfield, E. (2005). The agony and the ecstasy – the history and meaning of the Journal Impact Factor. Retrieved January 8, 2007, from http://garfield.library.upenn.edu/papers/jifchicago2005.pdf

Hajjem, C., Harnad, S., & Gringras, Y. (2005). Ten-year cross disciplinary comparison of the growth of OA and how it increases citation impact. Retrieved January 8, 2007, from http://eprints.ecs.soton.ac.uk/11688/01/hajjem.pdf

Harnad, S. (2006). Opening access by overcoming Zeno's paralysis. In N. Jacobs (Ed.), Open access: Key strategic, technical and economic aspects (pp. 73-85). Oxford: Chandos Publishing. . Retrieved April 3, 2007, from http://eprints.ecs.soton.ac.uk/11688/01/hajjem.pdf

Harnad, S. (2007a). The open access citation advantage: quality advantage or quality bias?. Retrieved April 3, 2007, from http://openaccess.eprints.org/index.php?/archives/191-The-Open-Access-Citation-Advantage-Quality-Advantage-Or-Quality-Bias.html

Harnad, S. (2007b). Citation advantage for OA self-archiving is independent of journal impact factor, article age, and number of co-authors. Retrieved April 5, 2007, from http://openaccess.eprints.org/index.php?/archives/2007/01/17.html

Harnad, S., & Brody, T. (2004). Comparing the impact of OA (OA) vs. non-OA articles in the same journals. D-Lib Magazine, 10. Retrieved March 23, 2007, from http://mirrored.ukoln.ac.uk/lis-journals/dlib/dlib/dlib/june04/harnad/06harnad.html

Harnad, S., Brody, T., Vallieres, F., Carr, L., Hitchcock, S., Gingras, Y., Oppenheim, C., Stamerjohanns, H., & Hilfet, E. (2004). The access/impact problem and the green and gold roads to open access. Serials Review, 30, 310-314. Retrieved April 5, 2007, from http://users.ecs.soton.ac.uk/harnad/Temp/impact.html

Hicks, D. (2004). The four literatures of the social science. In H. K. Moed., W. Glanzel, & U. Schmoch (Eds.), Handbook of quantitative science and technology research (pp. 473–495). Dordrecht: Springer.

Hinton, P. (2004). Statistics explained. (2nd ed.). London: Routledge.

Hirsch, J. (2005). An index to quantify an individual's scientific research output. Proceedings of the National Academy of Sciences, 46, 16569-16572. Retrieved April 5, 2007, from http://www.pnas.org/cgi/reprint/102/46/16569

Jacso, P. (2005a). Google Scholar: the pros and the cons. Online Information Review, 29, 208-214.

Jacso, P. (2005b). As we may search – Comparison of major features of Web of Science, *Scopus* and *Google Scholar* citation-based and citation-enhanced databases. Current Science, 89, 1537-1547. Retrieved April 5, 2007, from http://www.pnas.org/cgi/reprint/102/46/16569

Journal Citation Reports (2007). Retrieved January 8, 2007, from http://portal.isiknowledge.com/portal.cgi?DestApp=JCR&Func=Frame

Kurtz, M., Eichhorn, G., Accomazzi, A., Grant, C., Demleitner, M., Henneken, E., & Murray, S. (2005). The effect of use and access on citations**.** Information Processing and Management, 41, 1395–1402

Lawrence, S. (2001). Online or invisible. Nature, 411, 521. Retrieved April 11, 2007, from http://citeseer.ist.psu.edu/lawrence01online.html

Metcalfe, T. (2006). The citation impact of digital preprint archives for solar physics papers. Solar Physics, 239, 549-553. Retrieved April 12, 2007, from http://www.springerlink.com/content/3485x525622j0801/fulltext.pdf

Meho, L., & Yang K. Impact of data sources on citation counts and rankings of LIS faculty: Web of Science vs. Scopus and Google Scholar. Journal of the American Society for Information Science and Technology. [In press] . Retrieved July 22, 2007, from http://www.slis.indiana.edu/faculty/meho/meho-yang-03.pdf

Moed, H. (2005). Citation analysis in research evaluation. Dordrecht: Springer.

Moed, H. (2006). The effect of 'Open Access' upon citation impact: An analysis of ArXiv's Condensed Matter Section. Retrieved March 20, 2007, from http://arxiv.org/abs/cs.DL/0611060

Myhill, M. (2005). *Google Scholar*. Retrieved February 15, 2007, from http://www.charlestonco.com/review.cfm?id=225

Nederhof, A. (2006). Bibliometric monitoring of research performance in the Social Sciences and the Humanities: A review. Scientometrics, 66, 81-100.

Norris, M., & Oppenheim, C. (2007). Comparing alternatives to the *Web of Science* for coverage of the social sciences' literature. Journal of Informetrics, 1, 161-169.

Notess, G. (2005). Scholarly web searching: Google Scholar and Scirus. Retrieved February 15, 2007, from http://www.infotoday.com/Online/jul05/OnTheNet.shtml

OAIster. (2007). Retrieved January 29, 2007, from http://OAIster.umdl.umich.edu/o/OAIster/about.html

OpenDOAR (2007). Retrieved January 29, 2007, from http://www.opendoar.org/about.html

Opthof, T. (1997). Sense and nonsense about the impact factor. Cardiovascular Research, 33, 1-7.

The RePEc Project (2007). Retrieved July 22, 2007, from http://repec.org/docs/RePEcIntro.html

Sale, A. (2005) Comparison of IR content policies in Australia. Retrieved February 12, 2007, from http://eprints.comp.utas.edu.au:81/archive/00000230/01/Comparison_of_content_policies_in_Australia.pdf

Schwarz, G., & Kennicutt, R. (2004). Demographic and citation trends in astrophysical papers and preprints. Retrieved March 19, 2007, from http://arxiv.org/PS_cache/astro-ph/pdf/0411/0411275v1.pdf

Swan, A., & Brown S. (2005). OA self archiving: an author study. Retrieved March 26, 2007, from http://eprints.ecs.soton.ac.uk/10999/01/jisc2.pdf

Van Leeuwen, T. N., Visser, M. S., Moed, H.F., Nederhof, T.J., & Van Raan A. F. J. (2003). The holy grail of science policy: exploring and combining bibliometrics tools in search of scientific excellence. Scientometrics, 57, 257-280.

Vaughan, L., & Shaw, D. (2005). Web citation data for impact assessment: a comparison of four disciplines. Journal of the American Society for Information Science and Technology, 56, 1075-1087.

Web of Science. (2007). Retrieved March 23, 2007, from http://portal.isiknowledge.com/portal.cgi?DestApp=WOS&Func=Frame

Wren, J. 2005. Information in Practice. BMJ, 330. Retrieved March 23, 2007, from http://bmj.bmjjournals.com/cgi/reprint/330/7500/1128

Appendix A

Journal titles and their 2005 impact factors: Applied Mathematics

Title	Impact Factor
ACM Transactions on Mathematical Software	1.463
Chaos	1.760
Communications on Pure and Applied Mathematics	1.841
Inverses Problems	1.541
Journal of Cryptology	2.280
Journal of Mathematical Imaging and Vision	2.197
Journal of Non-Linear Science	1.556
Journal of Scientific Computing	1.653
Mathematical Models and Methods in Applied Sciences	1.248
Mathematical Programming	1.497
Physica D-Non linear Phenomena	1.863
Siam Journal on Applied Dynamical Systems	2.159
Siam Journal on Numerical Analysis	1.392
Siam Journal on Optimisation	1.238
Siam Journal on Scientific Computing	1.509
Siam Review	7.213

Journal titles and their 2005 impact factors: Ecology

Title	Impact Factor
American Naturalist	4.464
Conservation Biology	4.110
Ecology	4.506
Journal of Applied Ecology	4.594
Journal of Ecology	4.277
Molecular Ecology	4.301
Trends in Ecology & Evolution	14.864

Journal titles and their 2005 impact factors: Economics

Title	Impact Factor
Econometrica	2.626
Economic Journal	1.440
Health Economics	1.919
International Economic Review	1.284
Journal of Accounting and Economics	1.877
Journal of Econometrics	1.579
Journal of Economic Geography	3.222
Journal of Economic Growth	2.577
Journal of Economic Perspectives	2.634
Journal of Environmental Economics and Management	1.529
Journal of Financial Economics	2.385
Journal of Health Economics	2.708
Journal of International Economics	1.667
Journal of Law and Economics	1.609
Journal of Monetary Economics	1.661
Journal of Political Economy	2.245
Journal of Risk and Uncertainty	2.100
Mathematical Finance	1.345
Resource and Energy Economics	1.541
Review of Economic Studies	2.035
Review of Economics and Statistics	1.518
World Development	1.504

Journal titles and their 2005 impact factors: Sociology

Title	Impact Factor
American Journal of Sociology	3.262
American Sociological Review	2.933
British Journal of Sociology	1.49
Economy and Society	1.125
Global Networks – A journal of Translational Affairs	1.340
Journal for the Scientific Study of Religion	1.039
Journal of Marriage and the Family	1.350
Language in Society	0.902
Law and Society Review	1.396
Leisure Sciences	1.045
Politics and Society	1.100
Population and Development Review	1.076
Rural Sociology	1.067
Social Networks	1.382
Social Problems	1.796
Society and Natural Resources	1.339
Sociological Methods and Research	1.032
Sociology of Education	1.222
Sociology – The Journal of the British Sociological Association	1.096