To appear in: *Interactive Media International* **http://www.epsltd.com/clients/imi.asp**

Citation Analysis in the Open Access World

Tim Brody

Intelligence, Agents, Multimedia Group
University of Southampton

Recent reports by the UK Parliament Committee on Science and Technology and the US House Appropriations Committee have recommended mandating that researchers provide Open Access (OA) to their research articles by self-archiving them free for all on the Web. OA is now firmly on the agenda for funding agencies, universities, libraries and publishers. What is needed now is objective, quantitative evidence of the benefits of OA to research authors, their institutions, their funders and to research itself. Web-based analysis of usage and citation patterns is providing this evidence.

One of the many misconceptions about the OA debate is that it is primarily about economics. Although the *journal pricing/affordability problem* certainly helped draw attention to OA, it has now become a distraction from the deeper problem: *the research access/impact problem:* No institution has the funds to subscribe to every journal that is published (there are about 24,000 according to Ulrich's Directory); most institutions can only afford to purchase access to a small proportion of them for their researchers (see ARL statistics). This would be true even if every journal were sold at-cost, zero-profit. Yet every potential user that an article loses is lost potential impact for its author, its author's institution, its research-funder, and for research itself. This lost impact is the access/impact problem, and the advent of the Web itself has provided the solution (Harnad et al. 2004).

The Web has revolutionised the dissemination of information. Most researchers have not yet recognised or used the web's power to maximise the visibility, accessibility, usage, and hence the impact of their work, but about 10-20% already have. By "self-archiving" a draft free for all on the web, this vanguard of authors has made sure that every would-be user – not only those whose institutions can afford the journal in which it was published -- can access the full text of their article. Lawrence (2001) reported that articles in Computer Science that were freely accessible on the Web received 3-5 times as many citations as their subscription-only counterparts. Similar analyses of physics and mathematics articles in arXiv.org (Harnad & Brody 2004) have confirmed that published articles for which their authors also provide OA by self-archiving them receive 2-3 times more citations than articles (in the same journal and year) for which the author does not self-archive an OA version (Figure 1) (Harnad & Brody 2004). The increased *access* generates the increased *impact*.

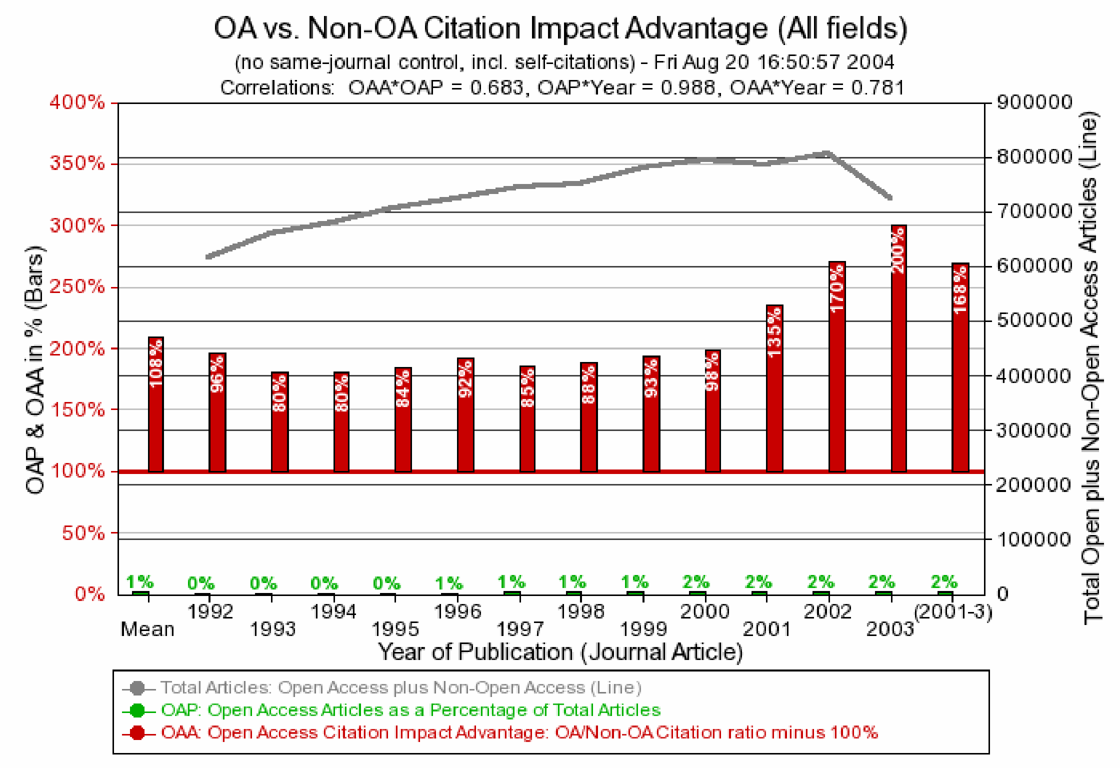

Figure 1 Comparing the citation impact for published articles (in all fields of physics and mathematics) that do and do not have an OA version self-archived in arXiv.org. Articles with OA versions consistently receive more citations than those that do not. This OA advantage is biggest within the year before and the two years after an article is published (an early-access pre-print advantage followed by a new-article post-print advantage), but older OA articles also continue to be cited more in these fields.

These findings are based on comparing OA articles with non-OA articles within the same journal and year. They are hence predicated on the fact that some articles are OA and some are not. Once the message about the OA advantage has got through to all authors, and OA reaches 100%, the relative OA/non-OA advantage itself will of course vanish. But at present, with OA still only 10-20%, the relative advantage is a strong, competitive one. Absolute OA advantages will of course persist even when 100% OA has been reached. Kurtz and co-workers have shown that in astrophysics, a small, field in which there is already effectively 100% OA through institutional licensing, overall usage of articles is doubled over what it was before OA. And of course 100% OA provides a far more rational basis for choosing what to cite and what not to cite than affordability does.

There are other access barriers than just financial ones, however. Paid access requires users to access papers through the publisher's Web site, or through a few aggregating services. Many sites consist of electronic re-creations of the equivalent paper-based journal, offering only inflexible PDF. Some sites have developed linking-services that allow users to click on a reference, but these links are often patchy when the reference is to another journal, and they are not embedded in the PDF full-text. Citation-based navigation and analysis is available on article metadata from the Institute of Scientific

Information's (ISI's) Web of Knowledge but not yet integrated with publishers' full-text; this is only beginning to emerge now with Scopus.

As we approach 100% OA, powerful new services will become possible. Even with the small percentage of OA available today, there are already at least 10 different service providers based on the Open Archives Initiative (OAI) standard, which makes all OAI-compliant Archives (whether they consist of self-archived institutional output or journal/publisher databases) interoperable with one another. Commercial services such as Google and Yahoo! have shown interest in gaining access to the research literature, Google through an OAI-gateway developed by OCLC and Yahoo! in partnership with OAIster (an OAI service provider). We can expect that a range of aggregating services will be developed -- from toll-free, generic search engines such as Google to specialised toll-based services such as Elsevier's Scopus -- which will provide structured searches and improved and augmented metadata. In addition to meta-searches, OA opens the possibility of designing services that analyse patterns in scholarly research, using the built-in citation links of this special literature, without being limited to proprietary databases that cover only a portion of the total literature.

OA is an exciting development for information science ("bibliometrics"). The most comprehensive citation database today is that of Thomson ISI. It covers around 8000 of the world's 24,000 total journals. With new autonomous OA citation tools (e.g., Citebase and Citeseer) information scientists can now build comprehensive citation databases limited only by what has been made OA to date. Citation databases allow the literature to be navigated backwards and forwards in time, following citations to and from any article, guided also by co-citation analysis in order to find related papers (which papers cite the same papers? or are cited by the same papers?). Citation analysis can be used to find emerging fields, to map the time-course and direction of research progress, and to identify synergies between different disciplines. Semantic-web analysis of the full-text content of papers can be used in similar ways to deepen the analysis of the underlying patterns, as well as to aid navigation, search and evaluation. For current users it will at first be just a pleasant surprise to find that the citation links within an article can retrieve the full-texts of the articles it cites (and eventually also those that it is cited by); yet this is just one of the many rich scientometric possibilities that will be provided by OA.

Citebase , a scientometric tool developed at the University of Southampton to explore and demonstrate the potential of an OA corpus, currently harvests self-archived full-texts from 2 *central* OA Eprint Archives, arXiv.org and Cogprints, 2 local *institutional* OA Eprint Archives -- a Southampton University departmental archive (ECS) and Southampton's institutional archive -- plus 1 publisher-based OA archive, Biomedcentral. Citebase parses the reference lists from these papers, linking those references that can be found in its database (i.e. internal references to the archives it harvests from). The linked references create a "citation database", which allows citation impact -- the number of citations to papers or authors -- to be counted. Citebase can be used to display search results rank-ordered on the basis of the citation counts of either the (1) retrieved papers or the (2) retrieved papers' (1st) authors. Users can search and navigate within this entire full-text corpus via citations to/from each article, via co-cited articles, via “hubs and authorities,” and on the basis of graphs of each article’s download and citation history.

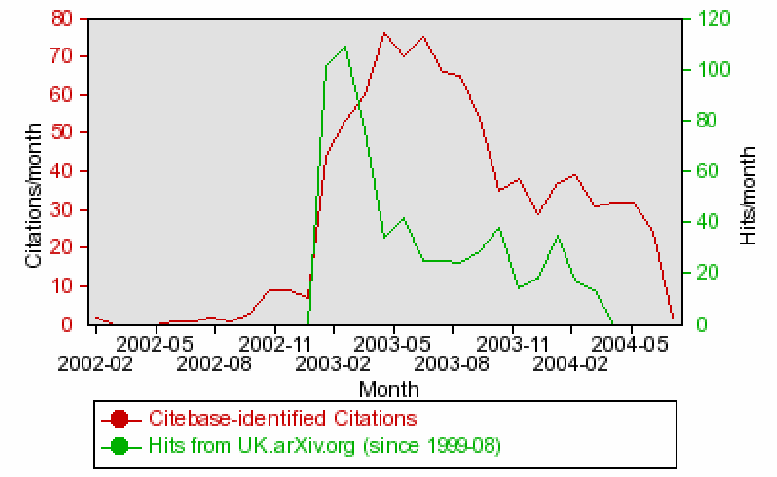

Figure 2 Citation/download history for an **arXiv.org** article (generated from **Citebase**). The article was deposited in 2003-01 (when the download data starts). A rise in downloads leads to a later rise in citations, with both the citation and download rates then decaying over several years. (Citations that appear to originate before the article was deposited are actually from articles that have been revised by their authors to insert citations to articles that were deposited only after they had already deposited their own [citing] article.) Over a longer time interval we might expect to see further, smaller download/citation cycles for the same article, with further citations generating further downloads, and further downloads generating further citations.

In addition to the citation counts for articles in Citebase, data are available for the number of downloads from the UK arXiv.org mirror. This usage indicator provides an additional basis for assessing the research impact of articles. Statistical analysis of the citation and download impact of High Energy Physics papers in arXiv.org reveals a correlation coefficient of **.4** (which is a measure of the degree to which higher download counts are associated with higher download counts, and vice-versa). Six months’ worth of download counts already seem to be as highly correlated with citation counts as two years’ worth of download counts, which suggests that download impact can be used as an early-days predictor of citation impact. Usage data could hence be useful for assessing the impact of very new research, or very junior researchers. Although download data are noisier than citations, the sizeable correlation shows that they are nevertheless quite robust too, and are hence yet another informative new benefit of OA.

The online era has not produced a *substitute* for the traditional research publication system, but a powerful new *supplement* to it, particularly in the area of access provision and impact assessment. It is now important to open the eyes of the research community -- authors, their institutions, and their research funders -- to the vast benefits of providing Open Access to their journal articles by self-archiving their full texts on the Web, in accordance with the self-archiving mandate now being considered by the US, UK, and a number of other countries.

REFERENCES

UK Parliament Select Committee on Science and Technology (2004) “Scientific Publications: Free for All?” http://www.publications.parliament.uk/pa/cm200304/cmselect/cmsctech/399/39902.htm

US House Appropriations Committee (2004) http://www.taxpayeraccess.org/congress.html

Harnad, S., Brody, T., Vallieres, F., Carr, L., Hitchcock, S., Gingras, Y, Oppenheim, C., Stamerjohanns, H., & Hilf, E. (2004) The green and the gold roads to Open Access. *Nature Web Focus*. http://www.nature.com/nature/focus/accessdebate/21.html

Lawrence, S (2003) “Free online availability substantially increases a paper's impact” http://www.nature.com/nature/debates/e-access/Articles/lawrence.html

Harnad, S. & Brody, T. (2004) Comparing the Impact of Open Access (OA) vs. Non-OA Articles in the Same Journals, *D-Lib Magazine* 10 (6) June http://www.dlib.org/dlib/june04/harnad/06harnad.html

Kurtz, M.J. et al. (2003) The NASA Astrophysics Data System: Sociology, Bibliometrics, and Impact. http://cfa-www.harvard.edu/~kurtz/jasis-abstract.html Kurtz, M.J. (2004) Restrictive access policies cut readership of electronic research journals articles by a factor of two. http://opcit.eprints.org/feb19oa/kurtz.pdf

Brody, T (2004) “Citebase Search” http://citebase.eprints.org/

Brody, T (2004) “Correlation Generator” http://citebase.eprints.org/analysis/correlation.php