The impact of OAI-based search on access to research journal papers

Based on a paper given at the UKSG seminar 'The Open Archives Initiative: application and exploitation', London, 14 May 2003

Intuitively, if a product is useful and has both a priced and a free version its total usage rate would be expected to be higher than if there is only a priced version. Evidence is emerging that this is true for online research journal papers. Authors need accessible online sites in which to deposit their published papers, and users need a means of discovering and evaluating those papers. The Open Archives Initiative (OAI) has now produced free software packages for building OAI-compliant institutional archives and OAI search services, including a citation-ranked search and impact discovery service. New data from this service shows that higher usage of free papers leads directly to a higher number of citations and thus greater research impact. Institutional archives need far more papers to be deposited, and one way of bringing this about is to implement institutional and national policies mandating the self-archiving of all funded research output in open access archives. This paper outlines why such policies are beneficial to researchers, their institutions, funders, and to research itself.

STEVE HITCHCOCK, TIM BRODY, CHRISTOPHER GUTTERIDGE, LES CARR AND STEVAN HARNAD
Intelligence, Agents, Multimedia Group, School of Electronics and Computer Science, University of Southampton

Introduction

Alert web publishers will have noticed a fundamental shift in the way users access information in a networked information environment. Instead of navigating web sites, users start with interfaces that allow them to perform particular tasks such as search and select. The most successful example is, currently, Google.

Electronic journals exist not just in a post-Gutenberg world, but a post-Google world too. The ability to locate a specified item of information precisely and instantly among the mass of information available on the web has profound implications. In the electronic environment the search engine has become the *de facto* interface to information, in place of the fragmented packages that have migrated from the print world.

Journal articles will also be accessed directly by search, but while Google's success has been based on an extension of the established scholarly practice of citation ranking, treating web links as citations, Google rankings do not make use of actual bibliographic citations within a paper. Further, most journal papers remain invisible to Google.

Recognising the importance to research of navigating citation space, the Open Citation Project has created a citation-ranked search and impact discovery service, Citebase (http://citebase.eprints.org/), for 'open access' (Suber 2003[1]) journal articles (i.e. accessible for free on the web). Citebase was designed to take advantage of the growing prevalence of the Open Archives Initiative Protocol for Metadata Harvesting (OAI-PMH) for describing the contents of distributed digital libraries. It extracts and indexes citations from published research papers stored in the larger OAI disciplinary archives – currently arXiv, CogPrints and BioMed Central – and is soon to include PubMed Central.

Citebase is now fully operational and is a featured service of the arXiv (http://arxiv.org) physics archives. It is more than a search engine, however. The data it collects offers new and compelling evidence that will induce a change in

the way researchers access published papers, a change that will be every bit as profound as the one induced by Google on the global web.

Citebase: measuring citation impact against usage

Citebase has been described by Hitchcock *et al.* (2002)[2]. A large-scale evaluation of Citebase concluded that web-based citation indexing of open-access e-print archives is closer to a state of readiness for serious use than had previously been realised (Hitchcock *et al.* 2003[3].) The evaluation proved that Citebase can be used simply and reliably regardless of the background of the user showing, despite the bias of the current service towards physics, this powerful new functionality can be extended to all the other disciplines as well.

According to the evaluation, Citebase compares favourably with other bibliographic services, such as ISI's Web of Science, even though its content size and range are still much smaller. Citebase can also provide earlier predictors and measures of impact, at the preprint phase of research.

Citation indexing has had some unexpected consequences: ISI's Science Citation Index has become a career development tool (Guédon 2001[4]). Authors publish for impact, which is classically measured by citations. In choosing a publication authors typically seek, however inexactly, to maximise the impact of their work.

Citebase can provide some of the established scientometric measures of research impact:

- citation counts for the article
- citation counts for the researcher
- co-citation (and eventually co-text) analyses

as well as some new measures of impact:

- citation counts for the preprint phase
- usage measures ('hits', webmetrics) for preprints and postprints
- time-course analyses, early predictors, etc.
- usage/citation correlators and predictors

For the first time, pre- and post-publication citations for individual papers can be measured against usage, i.e. web downloads. According to Kurtz *et al.* (2003)[5]: "Perhaps the most important new information to become available for bibliometric studies is the per article readership information."

Records in Citebase plot usage and citations against time for each arXiv paper indexed, as shown in Figure 1 for a highly-cited example paper. The citations are from all other papers deposited in arXiv. (The usage data ('Hits') are based only on downloads from the arXiv UK mirror server since August 1999, possibly underestimating usage by a factor of 18 across the worldwide network of arXiv mirror sites.) These charts suggest the following cycle of user actions: the preprint or postprint appears; it is downloaded (and sometimes read); eventually citations may follow (for more important papers); this generates more downloads, etc.

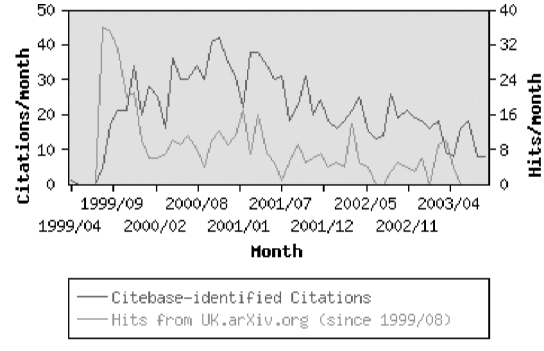


Figure 1. Charts of arXiv usage and citation for Witten, E. (1998) String Theory and Noncommutative Geometry, 'Adv. Theor. Math. Phys.' 2: 253. Generated by Citebase on 10 September 2003. For latest chart see http://citebase.eprints.org/cgi-bin/citations?id=oai%3AarXiv%2Eorg%3Ahep%2Dth%2F9908142

Correlation Generator

With the advent of new online tools authors will not only have greater scope to measure impact, they will quickly recognise the critical factor in enhancing impact, which is to make their published papers openly accessible. Lawrence (2001)[6] showed "an average of 336% more citations to (free) online articles compared to offline articles published in the same venue" for papers in computer science, i.e. free online access improves impact by a factor of over three.

Kurtz *et al.* (2003)[5] reached a similar conclusion for astrophysics, that access increases impact. They measured the impact of the Astrophysics Data System (ADS), a comprehensive collection of journal papers in a fee-based collection that, because it is available to almost all researchers in this field, effectively replicates open access. In this case impact was measured in a novel way: "We

find that in 2002 [the impact of the ADS] amounted to the equivalent of 736 FTE researchers, or $250 million, or the astronomical research done in France."

These startling findings can now be supplemented using the remarkable Correlation Generator based on Citebase data (http://citebase.eprints.org/analysis/correlation.php). This real-time Java tool, which plots the latest data based on user-set criteria, shows that usage impact is correlated with citation impact, i.e. the more often a paper is downloaded the more likely it is to be cited. This correlation is highest for high-citation papers and authors. Results obtained with the correlation generator are shown in Table 1, where the correlation coefficient (r) can be interpreted as the probability that a downloaded paper will be cited. It can be seen that r is higher for high-energy physics (hep), the largest sub-archives, compared to the whole arXiv, and larger for papers in the higher impact quartiles.

The dramatic conclusion from the studies so far is that as open access increases usage compared with fee-based usage and offline usage, this feeds directly into increased impact for authors. If he had measured usage as well, Lawrence would no doubt have found an increase in both usage and citations for free online articles compared with offline articles.

All	**r=.27, n=219328**
Q1 (lo)	**r=.26, n=54832**
Q2	**r=.18, n=54832**
Q3	**r=.28, n=54832**
Q4 (hi)	**r=.34, n=54832**
hep	**r=.33, n=74020**
Q1 (lo)	**r=.23, n=18505**
Q2	**r=.23, n=18505**
Q3	**r=.30, n=18505**
Q4 (hi)	**r=.50, n=18505**

Table 1. Correlation coefficient (r) between downloads and citations for all arXiv physics archive and high-energy physics (hep) sub-archives, also broken down into quartiles Q1 (low impact papers) – Q4 (high impact), n=number of papers

Work is ongoing to substantiate and quantify these results across other disciplines, by comparing citation rates for open-access papers with paired control papers: fee-access papers published in the same journal issue and volume, but not yet made openly accessible through self-archiving by their authors.

Growth of Eprints.org and institutional archives

As results confirming the striking correlation between access and impact become more widely known, a change in the way authors make their papers available can be anticipated. As most journals are not open access, authors will have two options.

Wherever a suitable open-access journal already exists for the subject matter of their article (about 500 such open-access journals exist so far, http://www.doaj.org/), authors can choose to publish in one of these. But even according to the most optimistic estimates, less than 5% of the total number of refereed-journal articles published annually today (at least 2.5 million, in 24,000 journals) as yet have an open-access journal in which to publish them (Harnad 2003[7]).

Most authors will continue to publish in established fee-access journals but they can in addition self-archive their papers in their own institution's open-access e-print archives. An analysis of publisher–author agreements shows that almost 55% (54.6%) of journal titles from the publishers surveyed already "explicitly left proprietary rights with the author" (Gadd *et al.* 2003[8]). In other words, authors of papers in these journals can officially self-archive these papers. For the remaining papers not covered by such agreements, many of the journals will agree to self-archiving if asked.

There are several free software packages that institutions can use to create archives for their research output. The most widely used archive software is Eprints.org (http://software.eprints.org/), now running over 100 archives worldwide, both institutional archives and disciplinary archives. Eprints.org software generates archives that are compliant with the OAI-PMH and, in conjunction with the OAI, Eprints.org has been a primary motivator for new institutional archives of research journal papers.

While the number of archives and self-archived papers is growing, the absolute number of papers accessible in these archives is still small – relative to the 2.5 million papers estimated to be published

annually in peer-reviewed journals. In new institutional archives in particular, after an initial burst of activity, the number of deposits tends to tail off (Figure 2). These curves need to become convex upward if archive growth is to become fast enough to attain the degree of open access that is already within researchers' reach.

This data is from the presentation *The Research Impact Cycle*, which contains further key data on the growth of open access through the self-archiving of institutional (peer-reviewed) research (http://www.ecs.soton.ac.uk/~harnad/Temp/self-archiving.ppt).

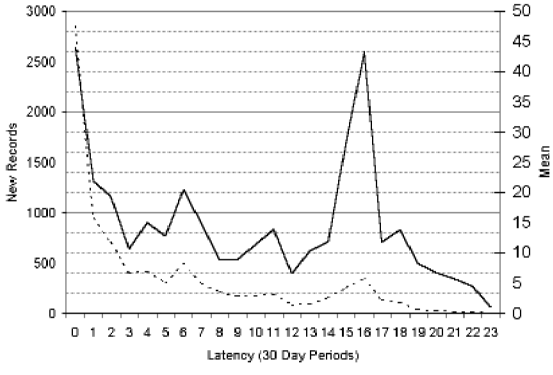


Figure 2. Latency of additions of records to new Eprints.org archives (broken line: new records in latency period; solid line: mean new records per archive)

Institutional archives may be the foundation for an expansion of open access to research papers, but this data shows that creating archives alone is not enough. This needs to be coupled with systematic institutional and national policies focusing on the causal connection between access, usage and impact, to ensure immediate, rapid and substantial growth in self-archived content across all research sectors in institutions.

Institutional and national policies for open access

If current data for the growth of institutional archives and their contents is not encouraging, this is misleading. The concept of open access and its effects on research impact is still new. The research community has not yet absorbed the implications of the findings on the access/usage/impact correlation. It is not only authors who benefit from maximising impact, but their institutions too, and the agencies that fund them.

Within institutions, departments are probably the best placed to implement self-archiving, through local policies, practices, and peer influences. Archive management might be best done either by the department or the institutional library. A sample policy has been formulated for the School of Electronics and Computer Science (ECS) at Southampton University and might serve as a suitable model for other institutions as well:

> 'All research output is to be self-archived in the departmental E-print archive. This archive forms the official record of the Department's research publications; all publication lists required for administration or promotion will be generated from this source.'
>
> *From ECS Research Self-Archiving Policy http://www.ecs.soton.ac.uk/~lac/archpol.html)*

Such policies, with institutional backing, should form the core of all institutional open access and research archiving policies.

In such a scenario the funding agencies are the remaining missing link, because they complete the virtuous circle of funding-research-evaluation-funding. Decisions on what research and researchers to fund, or fund again, are informed and guided by the track record of both the research and the researchers. Track records are in turn based largely on measures of research impact – both past impact and potential impact. So if impact is in turn dependent on access and usage, it stands to reason that whatever improves the impact of research and researchers, and also makes it more measurable, is also beneficial to research assessors and funders. It allows them to decide where to make their funding investment, and helps in evaluating the return on the investment. It also levels the playing field for researchers and their institutions: maximising the visibility and accessibility of a piece of research will not guarantee that it will be more widely used and cited: that also depends on the quality of the research. But open access does guarantee that potential impact will no longer be lost because would-be users could not access it.

In the UK the primary target of research evaluation is the Research Assessment (RA) exercise by the Research Councils. Harnad *et al.*

(2003)[9] proposed mandating online UK Research Assessment CVs linked to university e-print archives. They cite a number of benefits, to authors, institutions and the Research Councils. Just one such benefit – the promise of greater flexibility, speed and precision, all for less effort, if research were assessable by online impact-measuring services like Citebase built on comprehensive open-access archives – would be sufficient for the Research Councils to justify such a mandate. It is not unreasonable to suggest that this idea is likely to receive serious consideration in the ongoing reform of the RA. Such a move in the UK 'will set an example for the rest of the world that will almost certainly be emulated in terms of research assessment and research access'. National policies on open access are being considered in Australia, Germany, and the Netherlands, among others.

Through suitable policies, only access to scholarly papers needs to be transformed. Neither peer review practices nor the practice of publishing in the established journals need to be modified.

Conclusion

Institutional archives are being created, but need to be filled more quickly, by authors, with research journal papers. Attracting authors and their papers requires evidence of services that will improve the visibility and impact of their works. Emerging citation-ranked search and impact discovery services such as Citebase and its offspring usage-impact correlation generator are beginning to do this.

Nor can the role of the OAI be under-estimated. The OAI-PMH has become the shared technical infrastructure for institutional archives and is the enabler for cross-archive discovery services like Citebase. Free software for building institutional archives based on the OAI-PMH is now widely used.

The OAI is gathering momentum within digital libraries but more needs to be done by others across the research and academic community to realise the opportunity of providing open access to all research journal papers:

Universities: Adopt a university-wide policy of self-archiving all university research output, e.g. Southampton (ECS) Research Self-Archiving Policy

Departments: Create departmental OAI-compliant e-print archives

University Libraries: Provide digital library support for research self-archiving and archive-maintenance

Promotion committees: Request a standardized online CV from all candidates, with refereed publications all linked to their full-texts in the departmental archives

Research funders: Assess research impact online (from the online CVs)

Publishers: Ensure they have policies to facilitate authors seeking open access for their published papers, principally by allowing authors to self-archive in institutional archives. Continue to emphasise quality and peer review standards for journals.

Open access to published research papers will prevail because by increasing access and impact it demonstrably serves the interests of authors, users, their institutions and funders, and it will co-exist with high quality peer-reviewed journals whether those journals are open access or not. All participants need to plan accordingly.

References

1. Suber, P., (2003) Removing the Barriers to Research: An Introduction to Open Access for Librarians, *College & Research Libraries News*, Vol.64, no. 113, February, pp 92-94, 2003. http://www.earlham.edu/~peters/writing/acrl.htm
2. Hitchcock, S., Bergmark, D., Brody, T., Gutteridge, C., Carr, L., Hall, W., Lagoze, C. and Harnad, S., Open Citation Linking: the Way Forward, *D-Lib Magazine*, Vol. 8, no. 10, October 2002. http://www.dlib.org/dlib/october02/hitchcock/10hitchcock.html
3. Hitchcock, S., Woukeu, A., Brody, T., Carr, L., Hall, W. and Harnad, S., *Evaluating Citebase, an open access Web-based citation-ranked search and impact discovery service,* Technical Report ECSTR-IAM03-005, School of Electronics and Computer Science, Southampton University, 2003. http://opcit.eprints.org/evaluation/Citebase-evaluation/evaluation-report-tr.html
4. Guédon, J-C., Oldenburg's Long Shadow: Librarians, Research Scientists, Publishers, and the Control of Scientific Publishing, In: *ARL Proceedings, 138th Membership Meeting, Creating the Digital*

Future, May 2001, Toronto.
http://www.arl.org/arl/proceedings/138/guedon.html

5. Kurtz, M.J., Eichhorn, G., Accomazzi, A., Grant, C., Demleitner, M., Murray, S. M., Martimbeau, N. and Elwell, B., 'The NASA Astrophysics Data System: Sociology, Bibliometrics and Impact', 2003 http://cfa-www.harvard.edu/~kurtz/jasist-submitted.pdf
6. Lawrence, S., Free online availability substantially increases a paper's impact, *Nature*, Vol. 411, no. 6837, p 521, 2001. http://www.neci.nec.com/~lawrence/papers/online-nature01/
7. Harnad, S.,On the Need to Take Both Roads to Open Access. *American Scientist EPRINT Forum*, 6 September 2003 http://www.ecs.soton.ac.uk/~harnad/Hypermail/Amsci/2995.html
8. Gadd, E., Oppenheim, C. and Probets, S., RoMEO Studies 4: An analysis of Journal Publishers' Copyright Agreements, *Learned Publishing*, Vol.16, no. 4, October 2003. http://www.lboro.ac.uk/departments/ls/disresearch/romeo/RoMEO%20Studies%204.pdf

See also the list of publishers surveyed http://www.lboro.ac.uk/departments/ls/disresearch/romeo/Romeo%20Publisher%20Policies.htm

9. Harnad, S., Carr, L., Brody, T. and Oppenheim, C., Mandated online RAE CVs linked to university eprint archives: Enhancing UK research impact and assessment, *Ariadne*, issue 35, April 30 2003 http://www.ariadne.ac.uk/issue35/harnad/

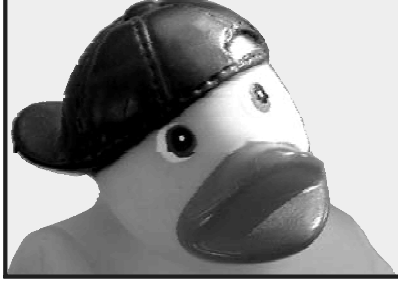
About the authors and paper

The authors worked on the Open Citation Project, funded by the Joint NSF – JISC International Digital Libraries Research Programme, from 1999-2002, and are now involved in extending that work to quantitative studies on changing patterns of access to published papers.

Steve Hitchcock
IAM Group, Department of Electronics and Computer Science, University of Southampton, S017 1BJ
E-mail: sh94r@ecs.soton.ac.uk

DUCKCREATIONS
Design for print and electronic media

Brochures • Branding • Conference and seminar collateral • Advertising Promotional material • Mailshots Flyers • Web • Multimedia



Ducklington, Oxfordshire
Tel: 01993 771428
Tel/Fax: 01993 771058
e-mail: mwalsh@duckcreations.co.uk
www.duckcreations.co.uk