Key Perspectives

Consultants to the scholarly information industry

Open access self-archiving:
An introduction

May 2005

Alma Swan

Key Perspectives Limited
48 Old Coach Road, Playing Place, TRURO, Cornwall, TR3 6ET, UK
(Registered Office)
Tel. +44 (0)1392 879702
www.keyperspectives.co.uk www.keyperspectives.com

Note to readers

The pages that follow constitute the Introduction, Executive Summary and References from a document written in May 2005 reporting the findings of a large-scale survey of scholarly researcher behaviour with respect to open access, specifically the 'green' route to OA via self-archiving. The Introduction serves as a stand-alone starter document for those wishing to acquaint themselves with self-archiving without too much pain. The full study report, for those who are interested, can be found at any of the following URLs:

www.keyperspectives.co.uk/OpenAccessArchive/2005_Open_Access_Report.pdf

http://www.jisc.ac.uk/uploaded_documents/Open%20Access%20Self%20Archiving-an%20author%20study.pdf

http://cogprints.org/4385/

Open Access Briefing Paper
A two-page Briefing Paper on Open Access, which covers the principles and issues of open access in a very concise form, is published by the Joint Information Systems Committee (JISC) and is available at:

http://www.jisc.ac.uk/uploaded_documents/JISC-BP-OpenAccess-v1-final.pdf

EXECUTIVE SUMMARY

This, our second author study on open access, was carried out to determine the current state of play with respect to author self-archiving behaviour. The survey was carried out during the last quarter of 2004. There were 1296 respondents.

The survey also briefly explored author experiences and opinions on publishing in open access journals to follow up our previous study on this topic for JISC and the Open Society Institute. Many of the findings reported here match those of that previous study. For example, the main reasons for authors publishing their work in open access journals are the principle of free access for all and their perceptions that these journals reach larger audiences, publish more rapidly and are more prestigious that the toll-access (subscription-based) journals that they have traditionally published in. The principal reasons why authors have not published in open access journals are that they are unfamiliar with any in their field and that they cannot identify a suitable one in which to publish their work. These reasons, and their rank order, exactly match the findings from our survey that was specifically on open access publishing last year.

The purpose of this present study, however, was to move the focus onto self-archiving, the alternative means of providing open access to scholarly journal articles. Almost half (49%) of the respondent population have self-archived at least one article during the last three years in at least one of the three possible ways — by placing a copy of an article in an institutional (or departmental) repository, in a subject-based repository, or on a personal or institutional website. More people (27%) have so far opted for the last method — putting a copy on a website — than have used institutional (20%) or subject-based (12%) repositories, though the main growth in self-archiving activity over the last year has been in these latter two more structured, systematic methods for providing open access. Use of institutional repositories for this purpose has doubled and usage has increased by almost 60% for subject-based repositories.

Postprints (peer-reviewed articles) are deposited more frequently than preprints (articles prior to peer review) except in the longstanding self-archiving communities of physics and computer science. There are some differences between subject disciplines with respect to the level of self-archiving activity and the location of deposit (website, institutional or subject-based repositories). Self-archiving activity is greatest amongst the most prolific authors, that is, those who publish the largest number of papers.

There is still a substantial proportion of authors unaware of the possibility of providing open access to their work by self-archiving. Of the authors who have not yet self-archived any articles, 71% remain unaware of the option. With 49% of the author population having self-archived in some way, this means that 36% of the total author population (71% of the remaining 51%), has not yet been appraised of this way of providing open access.

Authors have frequently expressed reluctance to self-archive because of the perceived time required and possible technical difficulties in carrying out this

activity. The findings here show that 20% of authors found some degree of difficulty with the first act of depositing an article in a repository, but that this dropped to 9% for subsequent depositions. Similarly, 23% of authors took more than an hour to deposit their first article in a repository, but only 13% took this long subsequently, with most taking a few minutes. Another author worry regarding self-archiving is the danger of infringing agreed copyright agreements with publishers. Only 10% of authors currently know of the SHERPA/RoMEO list of publisher permissions policies with respect to self-archiving, where clear guidance as to what a publisher permits is provided. Where permission is understood by the author to be required, it seems it is being sought (this accounts for around 17% of self-archiving cases); where it is not known if permission is required, authors are not seeking it and are self-archiving without it.

Communicating their results to peers remains the primary reason for scholars publishing their work; in other words, they publish to have an impact on their field. Nonetheless, more than half still do not know what the citation rate is for their most recent articles. Almost all (98%) of authors use some form of bibliographic service to locate articles of interest in closed archives such as publisher websites, but only a much smaller proportion of people (up to 30%) are yet using the specialised OAI search engines to navigate the open access repositories. Nevertheless, at the time of this survey, 72% of authors were using Google to search the web for scholarly articles: the subsequent arrival of GoogleScholar, which indexes the content of open access repositories as well as general websites, and thus retrieves formally-archived open access material, can be expected have a bearing on the level to which open access archives are searched in future and consequently on the eventual impact of articles deposited therein.

The vast majority of authors (81%) would willingly comply with a mandate from their employer or research funder to deposit copies of their articles in an institutional or subject-based repository. A further 13% would comply reluctantly; 5% would not comply with such a mandate.

Alma Swan
Key Perspectives Ltd
Truro, UK
27 May 2005

1. INTRODUCTION

Twelve months ago we at Key Perspectives Ltd completed and reported on a study of authors who had published their work in open access journals, compared and contrasted with authors who had not done this[1,2]. The work was commissioned and funded by the Joint Information Systems Committee (JISC) in the UK and the Open Society Institute. Having thus learned about authors' experience of **open access publishing**, we embarked upon this current study of the alternative means to providing open access — by authors archiving copies of their articles in open access archives or repositories. This process is usually referred to as **'self-archiving'**.

The practice of self-archiving has its roots in the field of computer sciences, where researchers were depositing results in ftp archives some decades ago and, later, on websites. A preprint culture — that is, the distribution of drafts of research articles before they have been peer reviewed to colleagues around the world, to establish ownership of the piece of research, to move the subject along, and to invite critical commentary before final revision and submission of the articles to learned journals — had been in place for many years in print form in the computer science community, and as the digital age arrived the practice simply migrated from paper to electronic form. Today, there are more articles – preprint and postprint (peer-reviewed papers) - freely available through self-archiving in computer science than in any other subject. The computer science 'online library', Citeseer[3], currently has almost 723,000 articles that have been harvested from distributed sites around the world (websites, ftp archives) where authors have deposited their work. Not only does this indicate the size of the corpus of computer science research available on open access, but it clearly demonstrates the success of this mechanism (harvesting from distributed sites) for creating a subject-based open access archive.

There is another mechanism for creating a subject-based archive and that is for authors to deposit their work directly into a centralised repository. In 1991, the first centralised archive, for the high-energy physics community, was established at the Los Alamos National Laboratory. It is called arXiv[4] and today this houses some 300,000 documents, with around 42,000 being added each year. Its main areas of coverage are high energy physics, condensed matter physics and astrophysics: substantial numbers of articles in computer science and mathematics research reside there too, along with, latterly, quantitative biology. It was also, from the outset, the norm for postprints — the peer-reviewed version of each article — to be deposited in arXiv, too. In most cases these are in the form of the author's final version rather than the publisher's formatted file, though some publishers do permit the use of their own copyrighted version for this purpose. The effect, then, was for research articles in the disciplines covered by arXiv to be available to anyone who wished to read them, even if their own institution could not afford to purchase the journals in which they were published. [On a point of terminology, the collective term for an electronic version of an article in draft (preprint) or final, peer-reviewed (postprint) form self-archived by the author is an 'e-print'].

That this practice could be spread to the rest of the scholarly community, freeing up the whole research literature from what he termed 'toll-access', that is, accessible only to those whose library could purchase the journals, was first mooted by Stevan Harnad in 1995[5,6]. Harnad has argued this case ever since, refining the model and rebutting[6] the arguments against the notion, which come not only from publishers, understandably nervous at what they see as a threat to their businesses, but also from the scholarly community itself — from researchers and librarians, both of whom are stakeholders in the developments in scholarly communications[7]. Their concerns have been debated extensively in public fora over the last decade (and continue to be), including the online American Scientist Open Access Forum set up and moderated by Harnad since 1998, the longest-running of all the open access discussion lists[8].

It is useful to lay out here the elements of this debate and the concerns that exercise the various parties. It should be noted that the focus of this present study is self-archiving, not open access publishing (in open access journals), which was extensively covered and discussed in our foregoing study[1,2]. The discussion here, therefore, concentrates on the issues around self-archiving that form the foci of resistance to the practice and which need to be overcome by proponents of open access if the whole research literature is to be 'made free'.

The first discussion point is the definition of what self-archiving is and what it is not. It is *not* an alternative to publishing in learned journals, but an adjunct, a complementary activity where an author publishes his or her article in whatever journal s/he chooses and then simply self-archives a copy. In practice, this means depositing the file, which is usually the author's final version of the article after peer review has been completed, in an open access archive or repository. There are two main types of such archives, which we will come to shortly. The articles are tagged in these archives as peer-reviewed postprints or as preprint drafts, so it is possible clearly to distinguish the two.

This brings us to the second point. Some researchers express a concern about the 'quality' of self-archived articles. Some disciplines use preprints much more extensively than others, but these pre-peer review articles are clearly tagged as such. It is true that some institutional archives may contain lots of other types of material as well (see Section 5.4.6 of this report) but the critical point here is that with respect to the research literature, what is deposited as a postprint is a *copy* of a fully peer-reviewed article whose destiny was to be published in the traditional way in a conventional, quality-controlled journal. It has therefore been peer-reviewed in the usual way. Postprints are not some kind of self-published, second-rate alternative to conventional journal articles: they *are* those articles.

Authors have often cited the issue of copyright as a major stumbling block to self-archiving. They are anxious that, having signed over copyright to the publisher of the journal in which their article appears they will be contravening the agreement if they self-archive the article. To be sure, if they self-archive the publisher's own file (the PDF file supplied by the publisher to the author and

containing the final formatting and layout assigned by the publisher) without permission, then this would in almost all cases be in contravention of copyright, if that resides with the publisher. The publisher has not copyrighted the author's final version, however, and in the vast majority of cases (over 90% is the latest estimate[9,10]) the publisher expressly permits an author to self-archive their own final draft — the version that was finally submitted to the publisher after peer-review revisions and recommendations have been incorporated.

The other main issue that is raised by authors[1,2] and, sometimes, by librarians, is how self-archiving might disrupt the present scholarly publishing model. Naturally, it is the perceived vulnerability of the journals published by learned societies, rather more than those of commercial publishers, that concerns authors. In this respect, it is worth examining what has happened to learned societies that have already had experience in this arena, those publishing in the areas covered by arXiv, alongside which they have had to live since 1991. It has already been said here that arXiv receives around 42,000 deposits per year. The ISI (Institute for Scientific Information) Science Citation Index covers around 420 physics journals, and to give a measure of the total volume of physics research, in 2003 these journals published a total of 116,721 articles: arXiv thus contains a substantial proportion (approximately one third) of the total physics research output and in the specialist areas mentioned earlier — condensed matter, astrophysics and high energy physics — the coverage of arXiv is pretty well complete.

In a separate exercise to this present study, we asked the American Physical Society (APS) and the Institute of Physics Publishing Ltd (IOPP) what their experiences have been over the 14 years that arXiv has been in existence. We asked how many subscriptions have been lost as a result of arXiv. Both societies said they could not identify any losses of subscriptions for this reason. Subscription movements for the journals they publish in the areas covered by arXiv are no different from those of their journals in other areas of physics over the period. Moreover, both societies say that they do not view arXiv as a threat to their business (rather the opposite, in fact) and this is underlined by the fact that the APS helped establish an arXiv mirror site at the Brookhaven National Laboratory – hardly the action of a society with its back to the wall because of that repository. Now it is true that there are only a couple of experiments of this sort carried out so far (physics and computer science), where publishers have to co-exist with a successful open access archive, and so there is always the possibility that there is something of a 'special case' about this example. Quite what might make it such a special case has never been adequately argued, but it is a finite possibility. Nevertheless, the evidence there is to hand points to the likelihood that the peaceful — and perhaps mutually beneficial — co-existence of traditional journals and open access archives is entirely possible; in biological terms, mutualism, rather than parasitism or symbiosis, might best describe the relationship.

The final issue that is raised frequently is the cost to institutions that self-archiving might impose. This is much more in the area of responsibility of librarians and institutional administrators than of authors. Will setting up and

running an open access archive in a research-based university, for example, cost a lot of money? How will it be paid for, whose budget will it fall under, can it be afforded, will it need an open cheque for the future? We collected together some actual figures from various archive managers for a study we undertook recently to develop a model for a national e-prints service for the United Kingdom. The figures varied wildly, as we meant them to for illustrative purposes, for we selected as our examples some of the largest and most ambitious, and some of the smallest and most modest, institutional archives in existence. For the whole range of costs, the reader is directed to the report of that study[11,12]. It is probably most helpful here to say that an average-sized research-based university can set up a functional archive for, say, ten thousand US dollars. Annual running costs vary according to the institution's existing levels of provision of IT services, what level of interventional support administrators are going to give the archive, and how much advocacy activity is to be included, but could amount to half or one FTE if ambitions do not run too high. For all the benefits such an archive brings to an institution (see below), this represents excellent value for money.

So much for the worries and concerns about self-archiving. Let's turn now to the arguments for it and the benefits that it can bring to the scholarly community, for there must be substantial benefits to be realised if the effort is to pay off. The benefits fall into two camps, those for the institution and those for the researchers (and some are shared, of course).

For the researcher, the most obvious benefit of making their work open access is the enhanced citations, and therefore impact, that result[13,14,15,16]. We know from the work reported here and elsewhere[17,18] that authors publish primarily to communicate their research findings to their peers, so that they can be built upon in future research efforts. Depositing an article at the time of acceptance for publication also means that the inevitable delay at the publisher before the article finally appears in the journal is immaterial — the article is already available to anyone who wants to read it and use it for their work. The research cycle is thus shortened. And of course, the article is available to *all* interested parties, not just to readers in institutions that can afford the journal in which it is published.

There are other benefits, too. An institutional repository is a secure storage location for working documents or for research data; it becomes the mediator for a one-input, many-outputs scenario, where a researcher can retrieve whichever elements of his or her own research record are needed for a task-in-hand (perhaps writing a paper, a lecture, preparing teaching materials, preparing a CV). It can also provide the home for research data that cannot be published in traditional journal format but which supports research findings and which the author would like to make available to peers and colleagues, data such as very large datasets, video files, graphical files of various formats, audio files and mixed media output.

For the institution, the benefits are just as substantial. Research-based institutions share with the researcher the wish to enhance the visibility and impact of the research generated within that institution. Institutions also have administrative burdens that require access to, and organisation of, information

about their employees' research records, research grant applications and fulfilment. They also need to carry out research performance evaluation (the Research Assessment Exercise in the UK being one such example,) and an institutional open access archive provides a permanent record of all the research output of that institution (provided that it has ensured all the researchers deposit copies of their articles, of course). An archive can also serve as a marketing tool for the institution, a shop window for potential students, staff and assessors on what is being generated by that institution. In a similar vein an institution can measure itself against other institutions that it sees as 'competitors' when all the outputs are openly visible in institutional archives. And, finally, a repository provides a place for *all* the digital output of that institution, so not just research articles but digital records of academic and cultural life in that institution can be stored there.

This gallop through the world of self-archiving brings us to the final discussion point here, which is the forms that self-archiving repositories might take. In this study we have distinguished the two main types, which are institutional and subject-based archives. Subject-based archives, such as arXiv discussed above, provide a location for the deposition of articles around a disciplinary theme. As well as arXiv (which houses articles in physics, computer science and mathematics), there are other well-known examples, such as Cogprints[19] (cognitive sciences), also a centralised repository. RePEc[20] (economics) is similar but actually works by harvesting articles from distributed archives. Whilst there is the obvious attraction to the appropriate community of such subject-centred services, we have argued that the optimal system for encouraging and achieving self-archiving across the whole scholarly community is via a distributed system; in other words, a global network of institutional archives, all OAI-compliant and thus completely interoperable*, so that a user can locate and be directed to an original article wherever it resides and without having to know anything about its location[11,12]. Subject-based centralised archives have their devotees and can be extremely popular within their communities. They are few and far between, however, and apart from arXiv most have been filling extremely slowly; Cogprints, for example, despite its 8-year existence, still houses only around 2000 articles. Subject-based services can be very useful to researchers, but are probably most effectively created by service providers (search-and-retrieval services) that harvest relevant subject-focused information from *all* repositories and sort and organise it to form a subject-centred offering to the research community.

The reason for arguing for a distributed system is that it is institutions (employers) that can most effectively bring about an effective self-archiving practice across the board. To be sure, research funders can influence the researchers they fund. The Wellcome Foundation is just implementing a self-archiving mandate for its grantholders to self-archive their articles and is setting up a new European PubMed Central archive for this purpose[21]. But external research funds only benefit a fraction of the research carried out in universities, so research funders can only influence a fraction of researchers. The institutions themselves, however, can influence the whole body of scholars, in whatever disciplines they work, funded or not, and if all institutions provide an archive

that is interoperable with every other archive then they are effectively contributing to a global database of freely accessible research — true open access.

*OAI-compliant means that the article metadata (the title, authors, keywords etc) are created in the format laid down by the Open Archives Initiative Protocol for Metadata Harvesting (OAI-PMH). Search engines can then harvest the metadata from all archives making their metadata visible in this form, and present it to users in an appropriate way.

Acknowledgment
The Institute for Scientific Information in Philadelphia donated a mailing list of 25,000 names for this study. The Joint Information Systems Committee (JISC) funded the writing of this report and its publication. Both are gratefully acknowledged.

References

1. Swan, Alma and Brown, Sheridan (2004) Report of the JISC/OSI open access journal authors survey. pp 1-76.
http://www.jisc.ac.uk/uploaded_documents/JISCOAreport1.pdf

2. Swan, Alma and Brown, Sheridan (2004) Authors and open access publishing. *Learned Publishing*, **17 (3)**, 219-224.
http://lysander.ingentaselect.com/vl=15729124/cl=20/nw=1/rpsv/cgi-bin/linker?ini=alpsp&reqidx=/cw/alpsp/09531513/v17n3/s7/p219

3. www.citeseer.ist.psu.edu

4. www.arxiv.org

5. Harnad, Stevan (1995) A Subversive Proposal. In: Ann Okerson & James O'Donnell (Eds.) Scholarly Journals at the Crossroads; A Subversive Proposal for Electronic Publishing. Washington, DC., Association of Research Libraries, June 1995. http://www.arl.org/scomm/subversive/toc.html
(originally posted June 27 1994:
http://www.arl.org/scomm/subversive/sub01.html)

6. Harnad, S (1999) Free at last: the future of peer-reviewed journals. *D-Lib Magazine*, **5**, 12. http://www.dlib.org/dlib/december99/12harnad.html

7. Self archiving FAQ. http://www.eprints.org/self-faq/

8. American Scientist Open Access Forum.
http://www.cogsci.soton.ac.uk/~harnad/Hypermail/Amsci/index.html

9. SHERPA: Publisher copyright policies and self-archiving.
http://www.sherpa.ac.uk/romeo.php

10. Eprints.org: Journal self-archiving policies.
http://romeo.eprints.org/stats.php

11. Swan A, Needham P, Probets P, Muir A, O'Brien A, Oppenheim C, Hardy R and Rowland F (2004). Delivery, management and access model for E-prints and open access journals within further and higher education (Report of a JISC study). pp 1-121.
http://www.jisc.ac.uk/uploaded_documents/ACF1E88.pdf

12. Alma Swan, Paul Needham, Steve Probets, Adrienne Muir, Anne O'Brien, Charles Oppenheim, Rachel Hardy, Fytton Rowland and Sheridan Brown (2005). Developing a model for e-prints and open access journal content for UK higher and further education. *Learned Publishing*, **18 (1)**, 25-40.
www.keyperspectives.co.uk/OpenAccessArchive/Eprints_LP_paper.pdf

13. Lawrence, S (2001) Online or invisible? *Nature* **411**, 6837, p521

http://www.neci.nec.com/~lawrence/papers/online-nature01/ or
www.nature.com/nature/debates/e-access/Articles/lawrence.html

14. Kurtz, M (2004) Restrictive access policies cut readership of electronic research journal articles by a factor of two.
http://opcit.eprints.org/feb19oa/kurtz.pdf

15. Harnad, S and Brody, T (2004) Comparing the impact of open access (OA) vs. non-OA articles in the same journals. *D-Lib Magazine*, 10 (6), (www.dlib.org/dlib/june04/harnad/06harnad.html).

16. Antelman, K (2005) Do open-access articles have a greater research impact? *College & Research Libraries*, **65 (1)**, 372-282.

17. Swan, A and Brown, S (2002) Authors and Electronic Publishing: The ALPSP research study on authors' and readers' views of electronic research communication. pp 1-76. ALPSP, Worthing.

18. Swan, A and Brown, S (2003) Authors and electronic publishing: what authors want from the new technology. *Learned Publishing*, **16 (1)**, 28-33.
http://lysander.ingentaselect.com/vl=15729124/cl=20/nw=1/fm=docpdf/rpsv/cw/alpsp/09531513/v16n1/s6/p28

19. www.cogprints.soton.ac.uk

20. www.repec.org

21. Wellcome Trust and National Library of Medicine in talks for worldwide open access archive (press release).
http://www.wellcome.ac.uk/doc_WTX022826.html

22. Rowlands, Ian, Nicholas, Dave and Huntingdon, Paul (2004). Scholarly communication in the digital environment: What do authors want? Findings of an international survey of author opinion: project report. Centre for Information Behaviour and the Evaluation of Research, City University, London, UK.

23. Publisher copyright policies and self-archiving.
http://www.sherpa.ac.uk/romeo.php

24. UK House of Commons Science and Technology Select Committee: Tenth Report. Scientific publications: Free for all?
http://www.publications.parliament.uk/pa/cm200304/cmselect/cmsctech/399/39902.htm

25. US Government House Appropriations Bill HR 5006 recommendations:
http://thomas.loc.gov/cgi-bin/cpquery/?&db_id=cp108&r_n=hr636.108&sel=TOC_338641&

26. Carr, L and Harnad, S (2005) Keystroke Economy: A Study of the Time and Effort Involved in Self-Archiving. http://eprints.ecs.soton.ac.uk/10688/

27. Pinfield, S (2005). A mandate to self archive? The role of open access institutional repositories. *Serials*, **18 (1)**, 30-34.

28. Pinfield, S (2004) Self-archiving publications. In: Gorman, G E and Rowland, F (eds). *International yearbook of Library and Information Management 2004-2005: Scholarly publishing in an electronic era*. London: Facet. Pp118-145. Available at http://eprints.nottingham.ac.uk/archive/00000142/

29. Registry of Institutional Self Archiving Policies. http://www.eprints.org/signup/fulllist.php

30. Perneger, T V (2004) Relation between online 'hit counts' and subsequent citations: prospective study of research papers in the BMJ. *BMJ* **329**, 546-7.

31. Brody, T and Harnad S (2005) Early web usage statistics as predictors of later citation impact. In press (Journal of the American Society for Information Science & Technology) http://eprints.ecs.soton.ac.uk/10712/