Self-archiving publications
Stephen Pinfield

Published in: Gorman, G.E. and Rowland, Fytton (ed.s). *International Yearbook of Library and Information Management 2004-2005: Scholarly publishing in an electronic era*. London: Facet, 2004, pp 118-145.

Abstract

This chapter discusses the historical development, current practice and future prospects of the self-archiving of research papers in open-access repositories (so-called 'e-print archives'). It describes how the development of interoperable e-print repositories in a number of subject communities has shown that self-archiving can benefit academic researchers (and potentially others) by enabling quick and easy access to the research literature and therefore maximising the impact potential of papers. Realising that the possible benefits are high and the technical entry barriers low, many organisations such as universities have recently tried to encourage widespread self-archiving by setting up institutional repositories. However, major barriers to self-archiving remain – most of them cultural and managerial. There are concerns about quality control, intellectual property rights, disturbing the publishing status quo, and workload. Ways in which these issues are currently being addressed are discussed in this chapter. A number of self-archiving initiatives in different countries have been set up to address the concerns and to kick-start e-print repository use. However, issues remain which require further investigation; those discussed in this chapter include discipline differences, definitions of 'publication', versioning problems, digital preservation, costing and funding models, and metadata standards. The ways in which these issues are resolved will be important in determining the future of self-archiving. Possible futures are discussed with particular reference to journal publishing and quality control. If widely adopted, self-archiving might come to assume a central place in the scholarly communication process, but a great deal of restructuring of the process needs to take place before this potential can be realised.

Introduction

The self-archiving of publications has the potential to revolutionise scholarly communication, making it more efficient and effective. But a great deal needs to be done before that potential can be realised. This chapter discusses some of the key issues associated with self-archiving. It analyses the ways in which self-archiving has so far developed, examines the possible benefits and drawbacks of self-archiving, and outlines the potential impact of the practice on scholarly communication.

'Self-archiving' (or 'author self-archiving' as it is sometimes known) is "a broad term often applied to the electronic posting, without publisher mediation, of author-supplied research" (Crow, 2002, 11). The term was first used in the literature in 1999 by leading advocates of the practice, Stevan Harnad and Paul Ginsparg (for example, Harnad, 1999a, 1999b; and Ginsparg, 1999a). It was used by them a year earlier in email discussion lists (for example, Harnad, 1998). It seems they were (knowingly or unknowingly) adapting a term already in use amongst computer scientists meaning a

program that archives files automatically. Ginsparg and Harnad were now applying the term to authors and their research papers.

What Harnad, Ginsparg and other proponents of self-archiving were (and still are) arguing is that authors of research papers should mount their work on the web so that all potential readers have free and unrestricted access to it. Such 'open access' to research literature would, Harnad (2001a) suggests, ensure that it is "freed" from the "unwelcome impediment" caused by "toll-gating access" in the form of conventional subscriptions, site licences and pay-per-view charges. Some practitioners have objected to the 'archiving' part of 'self-archiving' being used in this way to mean simply mounting a file on the web – the word implies to many a high degree of curation and preservation which may not be present in the self-archiving scenarios discussed by Harnad. Nevertheless, the label has now been widely adopted within the information community and beyond.

History of self-archiving: arXiv

Whilst the *term* 'self-archiving' may only have come into use in this field in 1998, the *practice* of self-archiving is much older than this. High-energy physicists have been posting their papers in an open-access repository since 1991. That was the year when arXiv, as it is now known, was set up at the Los Alamos laboratories by Paul Ginsparg and colleagues. Since then arXiv has become the most important vehicle for scholarly communication in High Energy Physics and related areas of Mathematics and Computer Science. It now contains over 300,000 papers, is mirrored on several continents, and is widely used. It is hailed by its managers, now based at Cornell University, as an exemplar of effective open-access web-based research communication.

ArXiv was originally designed to automate a pre-existing paper-based practice – the circulation of 'pre-prints'. Pre-prints are early versions of research papers before they have been refereed or formally published. Prior to the creation of arXiv, it was the practice of physicists to circulate hard-copy pre-prints to colleagues in other research groups worldwide as a preliminary stage of scholarly communication. The circulation of pre-prints achieved three main objectives. Firstly, it was a way of establishing priority. Physicists (like most other researchers) are eager to lay claim to an idea if they first thought of it. Pre-prints were a way of registering that claim without having to wait for formal peer-reviewed publication. This leads to the second objective of pre-print circulation – rapid dissemination. Pre-print circulation, even in a paper-based world, was fast. It meant scientific progress itself (which involves building on the work of others) could also be fast. Thirdly, circulating a pre-print was a way of improving the finished article. Authors of pre-prints would often receive comments from colleagues which could be incorporated in the final versions of papers submitted to peer-reviewed journals. This 'informal peer review', it was said, often led to better published papers.

Hard-copy pre-print circulation had a major limitation. Regular circulation could only ever include a relatively small number of institutions, and so some researchers (at other institutions) would miss out. The Los Alamos archive was designed to address this problem. Instead of circulating paper copies, authors were able to FTP their papers to a central server. Others could then easily download the papers, thus making

the research rapidly and widely available. With the advent of the web, this service became even easier to use for author and reader alike.

However, as with many things designed for a particular purpose, arXiv developed in unexpected ways. It was originally designed as a kind of 'bulletin board' to facilitate pre-print circulation. Pre-prints it was envisaged would be held for a temporary period only. However, it soon became clear that authors wanted it to be a long-term archive for papers. It also became clear that they wanted the repository to include not just pre-prints but also 'post-prints'. Post-prints are final versions of papers which have been revised in response to referees' comments and accepted for publication in journals. Electronic post-prints, of course, had a paper-based precedent. Authors would commonly circulate off-prints of their published papers, normally in response to a request from an interested colleague, albeit after the paper was formally published. The Los Alamos archive was seen as a convenient way of automating this and had the added benefit that the paper could be posted before it itself had appeared in the journal. It is common to see notes on papers in arXiv that the article has been accepted by and is forthcoming in a particular journal. It is clear then that arXiv has become a repository for electronic versions of papers both pre- and post- refereeing.

Electronic pre-prints and post-prints have become known collectively as 'e-prints', a term with a chequered history. Ginsparg (1999b) has described the history of the term. It was originally coined in the early 1990s by a mathematician, Greg Lawler, to describe electronic pre-prints. It was used more generally in the mid 1990s to mean "electronic versions of anything". Ginsparg was, however, influential in redefining the word at that time to mean "an article either in draft or final form SELF-ARCHIVED by the author" (original emphasis). This is the way the word is now generally used, particularly by users of arXiv, the largest e-print repository.

Brown (2001) and Pinfield (2001) have described how arXiv is currently used by physicists. The workflow is illustrated in Figure 1. The left hand column summarises the well-established process leading to publication in a peer-reviewed journal. This begins when an author writes a paper and submits it to a journal editor for consideration. The journal editor then sends the paper to one or more (usually two) referees who are working in the same field. The referees report back to the editor advising on whether or not the paper should be published. Assuming they recommend publication (as in Figure 1), they will normally suggest revisions to the paper. These suggestions are forwarded to the author, who is then expected to make the required changes. Following the submission of the revised version of the paper, it will be prepared by the publisher for publication with copy editing and formatting. Finally, it is published in an issue of the journal. The whole publication process can take twelve months, sometimes more. It extends beyond this, of course. The journal publisher or a secondary publisher will usually create metadata describing the paper which may be incorporated into separately published finding aids.

Self-archiving in a repository such as arXiv commonly takes place at two points in the publication process, as Figure 1 shows. Firstly, before the paper is refereed, the author may post it on the e-print repository as a pre-print. Secondly, when the paper has been revised in response to referees' comments, the final version of the paper can be mounted on the repository as a post-print. The post-print is normally formatted by the author and so may not incorporate changes made by the publisher at the pre-

publication stage. In fact arXiv does not normally accept publisher-produced files for copyright reasons (the publisher will normally own copyright in the layout of the formally-published article). At each of these stages the author creates metadata describing the paper which can be searched by users.

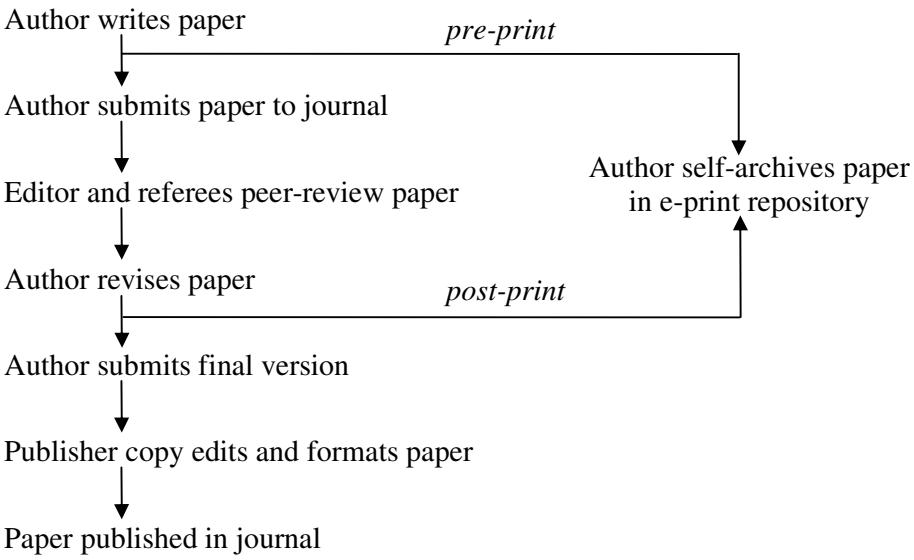

Figure 1: The journal publication process and self-archiving

Benefits of self-archiving

A service such as arXiv creates benefits for the individual researcher and for the research community in general. The benefits stem from the fact that arXiv lowers barriers created by the conventional publication system. These barriers are often divided into two related categories: 'impact barriers' and 'access barriers' (see Harnad, 2001b). These labels perhaps need some further explanation.

'Impact barriers' exist where a work is prevented from reaching all of its potential audience. Such a situation is not normally in the interests of the author. The author of a research paper usually publishes in order to make an 'impact' – to be read and cited by other researchers. Authors would not normally expect to make any income from publication, for this reason research papers are sometimes referred to as 'give-away literature'. It is in the author's interests that a paper should be distributed as widely as possible in order to maximise its potential impact. However, the current system of publication involves *publishers* making an income from the distribution of academic papers and it is therefore in *their* interests that circulation of a paper should be restricted to paying subscribers only. These restrictions limit the potential impact of a paper. They also create 'access barriers' which affect researchers in their capacity as readers of the scholarly literature. Readers want easy access to all publications in their field. However, the restrictions placed on access to the literature by publishers prevent this from happening. No academic institution can afford to subscribe to all peer-

reviewed journals and so its members cannot gain easy access to all publications required for their research.

Where these impact and access barriers are lowered for researchers by services such as arXiv, the benefits for the individual researcher soon become clear. The first and most obvious benefit is that papers are disseminated widely, thus maximising their impact potential. Evidence is beginning to emerge that papers which are openly accessible are more likely to be cited (see for example Lawrence, 2001). A second benefit is that research is disseminated quickly. Formal publication in a peer-reviewed journal can take twelve months or more. Even after a paper is accepted for publication in its final form there can be long delays until space can be found in the journal. Self-archiving, by contrast, is virtually instantaneous. A pre-print or post-print can be disseminated quickly by the author and can therefore have an immediate impact on research. For the reader, access is also quick and easy. The latest research literature is available in an unrestricted way from the desktop. Metadata describing papers can be searched and may even be pushed to registered users via email alerts.

Benefits for individual researchers translate into benefits for the research community as a whole. The speed of dissemination means that scientific progress itself can be accelerated. Researchers have access to the latest results in their field. They can also be confident (and increasingly so as self-archiving on arXiv becomes widespread) that they have access to the full breadth of the research literature available. Better communication enables better science. An example of this might be that unintentional duplication of research can be avoided since scientists are more likely to be aware of each other's activities.

Such benefits would ostensibly carry over into other disciplines. Since the creation of arXiv, other similar services have been set up for separate subject communities. CogPrints for cognitive sciences, and RePEc for economics, are examples of these. Like arXiv both were set up by pioneering enthusiasts: Stevan Harnad in the case of CogPrints and Thomas Krichel in the case of RePEc. Both reflect the slightly different communication cultures of their disciplines but have in common that they both contain pre-refereed versions of papers as well as post-refereed. However, neither CogPrints nor RePEc has yet been as successful as arXiv. They do not contain as many full-text papers, nor have they achieved the same importance to researchers that arXiv clearly has to high-energy physicists. The long-term success of each, of course, remains to be seen.

The development of other e-print repositories has prompted some of the supporters of self-archiving to speculate on the potential benefits of the practice were it to be adopted in a wide range of subject disciplines. The benefits of maximising impact and access potential for individual researchers and the consequent benefits for their research communities would apply more widely. Scholarly communication as a whole would operate more efficiently and effectively. The benefits of this would be felt globally, not least in countries, such as in the developing world, where institutions currently find it difficult to afford access to more than a few peer-reviewed journals. Researchers in these countries would be better able to contribute to the on-going development of scientific knowledge.

Some have suggested the benefits would spread more widely still. There are wider social and economic benefits which might follow from making high-quality research more easily available. Fundamental curiosity-driven research could more easily have an impact in applied areas. For example, there would be greater opportunities for knowledge transfer between the academic sector and the commercial sector. There would also be the potential for the public understanding of science to be enhanced. Science journalists and popular science writers would have better access to original research. Also, high school students could begin, with the guidance of a knowledgeable teacher, to become acquainted with the primary literature. The real weight of such arguments is difficult to assess at the present time, especially as the best example of self-archiving is currently in a discipline area (High Energy Physics) which is rather remote from commercial applicability and usually impenetrable to the uninitiated. Nevertheless, the attraction of making publicly-funded research more easily available to the public is in principle a strong one. Whilst such arguments are probably a long way from the original aims of the creators and users of arXiv, the benefits of self-archiving are potentially far-reaching.

Interoperability

With a number of separate e-print repositories beginning to appear in the late 1990's, it became clear that their usefulness would be enhanced by the development of interoperability between them. The Open Archives Initiative (OAI) was set up in 1999 to provide this. The OAI "develops and promotes interoperability standards that aim to facilitate the efficient dissemination of content." The most important technical outcome of this initiative has been the OAI Protocol for Metadata Harvesting (OAI-PMH).

The OAI Protocol (described in detail in the following chapter) was designed as a way of transporting data. It facilitates an exchange of information between 'Data Providers' and 'Service Providers'. Data Providers expose structured data (such as bibliographic records) on the internet so that it can be harvested by third parties. Service Providers harvest the data (normally in the form of simple Dublin Core) from a number of different Data Providers, organise it and then make it available to users in various ways. They often make it available in a searchable form so that an end-user can carry out a search encompassing a large number of Data Providers by interacting with a single Service Provider system. In the real world, a Service Provider, such as ARC, harvests data (or strictly speaking metadata which describes full-text papers) from a large number of OAI-compliant e-print repositories (and other similar services). The metadata is processed and presented to the end-user in a searchable form via a web interface. The end-user can search for e-prints held on a large number of different servers worldwide, by keywords from the title, abstract, subject terms or author names. If the user finds an item of interest, the record delivered by the Service Provider contains a clickable link to the full text of the paper held by the Data Provider. Tools are now available to ensure that OAI-compliant metadata can also be converted into HTML so that it can then be crawled by robots from mainstream web search engines, such as Google. This means that papers available in OAI-compliant e-print repositories are accessible to users not just via specialised OAI Service Providers but also via standard web search engine services.

The OAI Protocol is technically simple and this has facilitated widespread adoption. Since the release of version 2 of the Protocol in June 2002 (the experimental version 1 first came out in January 2001), it has been stable and can be implemented with confidence by information managers for production services. Many existing services, such as arXiv, were able to retro-fit their systems to become OAI-compliant relatively quickly. New Data Providers can achieve OAI-compliance easily and cheaply, especially as there are now several pieces of free repository software which come out of the box OAI-ready. Crow (2004) lists seven such software packages which are now available. The most well-established of these is the GNU e-prints (eprints.org) software, produced at the University of Southampton. This was originally based on the software used to deliver CogPrints. Pinfield, Gardner and MacColl (2002) have described the setting up of a GNU e-prints repository which took about five person days using an inexpensive server installation. The more recently released repository software from MIT and Hewlett Packard, DSpace, is now also being widely adopted. With the stabilisation of the OAI Protocol and the release of free OAI-compliant software, the technical barriers for setting up an interoperable e-print repository have become very low.

The 'openness' of the Open Archives Initiative is strictly speaking a technical one. The OAI Protocol has introduced a technology for systemic openness allowing services to talk to each other. Nevertheless, the combination of open access and the OAI Protocol is a powerful one which creates a number of benefits. The first benefit is enhanced accessibility. The content becomes more easily locatable and navigable for users. With repositories worldwide sharing interoperability standards there is the potential for a global virtual archive of research papers, entry into which can be gained from a single access point. The simplicity of the OAI Protocol (whilst a key to its success) does, however, create limits to interoperability. The Protocol deals in unqualified Dublin Core and this means the metadata from different Data Providers may be structured in different ways. As a consequence it is very difficult for Service Providers to say create a meaningful browse index, since names, or dates will be structured differently, and subjects will be described using different controlled vocabularies (or none at all). But despite this limitation the search tools created by existing Service Providers (such as ARC) are impressive.

Service Providers can do more than just deliver search services. A second major benefit created by OAI is the potential for other developments, such as analysis of the literature. Key metrics, such as citation analysis of self-archived papers, can already be delivered. Citebase is an interesting example of this. The potential of such tools is enormous. They can create useful post-publication quality indicators which could complement pre-publication quality assessment mechanisms such as peer review.

Institutional Repositories

The centralised subject-based approach to self-archiving has not always been successful. Warr (2003) has described the low take-up of an experimental Chemistry pre-print repository set up by Elsevier Science. Whilst still up and running, this service has met with scepticism from most chemists, although Warr also reports on a limited amount of support and interest. In other subject disciplines, e-print repositories may not exist at all or may be no more than a 'pet project' of one or two enthusiasts. Some commentators have, however, seen latent demand for e-print

repositories in many disciplines judging by the number of researchers who 'informally self-archive' their publications on their personal or departmental websites.

Since the majority of subject communities have not yet set up e-print repositories or adopted the practice of self-archiving on a significant scale, many supporters of self-archiving have come to favour an alternative strategy: institutional repositories. These are open-access archives set up and run by organisations such as universities which contain work by members of the institution. Institutions are ideally suited to support this kind of development for a number of reasons. Firstly, they have the resources to subsidise the start up of repositories and to fund their maintenance. Secondly, institutions have infrastructures (technical and organisational) to support them. Thirdly, institutions are able to provide a policy drive to encourage self-archiving amongst their members. Finally, it is argued, institutions have an interest in doing so. An institutional repository could potentially enhance the profile and prestige of an institution acting as an attractive shop window for research activities. A repository could also become part of a systematic information-asset management initiative to be used in activities such as community outreach, media relations, or accreditation management. There is also the potential benefit to the institution of long term cost savings in periodical subscriptions if the literature becomes widely available on open access.

In some ways then the institutional approach to self-archiving is a pragmatic one. It is seen as a pragmatic way to try to encourage the wider adoption of self-archiving. Despite the fact, as it is sometimes commented, that researchers may identify more with their subject community than with their institution and would therefore be more inclined to self-archive in a subject-based repository, it is institutions which are more likely to foster self-archiving on a large scale. Nevertheless, with OAI functionality in place, the location of the full text of a paper (whether in an institutional or subject repository) is in fact largely irrelevant. If papers are self-archived on institutional servers, it is easy to imagine that subject communities may provide subject-specialist Search Provider views of the data.

It is possible, however, that institutions may use repositories in additional ways, in addition that is to facilitating the self-archiving of scholarly papers. Repositories may also be used to store other digital objects associated with research (or indeed teaching) activity. These could include image, audio, and video files or data sets of various kinds. Lynch (2003) has described some of the potential. One simple development favoured by many researchers is that a published paper could be placed in a repository alongside the raw data produced during the research. The two could be linked in useful ways such that the communication of research results could be enhanced.

As well as providing immediate access to it, institutional repositories might also be used as a vehicle for preserving the scholarly output for the long term. This is a controversial idea amongst supporters of self-archiving. As Pinfield and James (2003) have described, some regard any concern with digital preservation to be a distraction from the central aim of encouraging self-archiving and achieving immediate access to the scholarly output. Harnad has argued this case (his views are outlined in Pinfield and James, 2003). On the other hand, some would say that preservation ought to be at the centre of the institutional repository mission. Crow (2002) described preservation as one of the key features that define an institutional repository. More work needs to

be done in this area but there is clearly a potential to use repositories as a means of achieving the systematic preservation of digital objects of all sorts, including research papers.

Barriers to self-archiving

Despite the apparent benefits of self-archiving and the recent growth in support for the practice, significant barriers to its widespread adoption remain. The first barrier is lack of awareness. This is demonstrated by Swan and Brown (2004) who report on the results of a survey of authors sponsored by JISC and the Open Society Institute published in February 2004. Their respondents are divided into two groups: "OA authors" (that is those who have published in open-access journals) and "non-OA authors" (those who have not). They report that 71% of OA authors and 77% of non-OA authors were not aware of any electronic repositories. This is a significant finding for supporters of self-archiving which indicates that there is a major awareness-raising job to do.

Even where researchers are aware of repositories, there is still considerable inertia when it comes to self-archiving. Many are cautious about practising self-archiving and sceptical about its potential benefits. Their objections normally fall into four main categories: quality control (particularly peer review), intellectual property rights (particularly copyright), concern about disturbing the publishing status quo, and work load. These objections will be discussed in turn below along with possible immediate responses. A more detailed discussion of how the system of scholarly communication might develop when self-archiving is widespread will be reserved for the section on the future.

Quality control is normally uppermost in the minds of researchers. There is a common suspicion that self-archiving undermines peer review. Because e-print repositories distribute content independently of any formal peer-review process they are often seen as a way of self-publishing without quality checks. There is a particular dislike of pre-prints in some disciplines. Of course, pre-prints are not a *necessary* part of an e-print repository. It is perfectly possible to set up a repository and only accept post-prints (or other documents which have been formally published or accepted for publication). Repositories are in themselves neutral with regard to quality control and so they can accommodate any form of quality assessment including peer review. The scenario advocated by most of the supporters of self-archiving – authors should submit their papers to peer-reviewed journals and also self-archive them – certainly takes into account the importance of peer review. Peer review is acknowledged to be important but it is recognised that at present it is carried out outside the e-print repository environment. In this case, repository managers should carry out low-level checks on quality before making a paper live on the system, but they can assume that the real quality checks occur elsewhere. Warr (2003) quotes Paul Ginsparg as describing this process for arXiv.

> "We still think it's important to have a minimal level of screening, to keep the material at least "of refereeable quality", and avoid material that is manifestly irrelevant, offensive, or silly." (Warr, 2003, 367-368)

Under institutional management, e-print repositories could, however, be managed (if necessary) with higher levels of quality control. Self-archiving need not be the anarchic activity it is sometimes assumed to be (although some commentators would regard the supposed anarchy of self-archiving as a good thing). It is possible that schools or departments within the institution could have to give formal authorisation before a paper is made live on the institutional server. Most repository software already has an authorisation procedure built into the workflow. This could be implemented with a light or heavy touch depending on the preferences of stakeholders. There is no suggestion that this would replace peer review but rather that it could provide an additional first-line quality check which could screen-out obviously inappropriate material.

The second area of concern for many researchers is that of IPR and copyright. Most research institutions allow their employees to dispose of the copyright of their own papers as they (the authors) choose. Some journal publishers require authors to sign-over exclusive rights before their papers are published. Publishers policies do, however, vary (see Gadd, Oppenheim and Probets, 2003). Other publishers do not require exclusive rights to be transferred by the author and may even explicitly allow the posting of pre- or post-prints on the web. Authors need assistance at a local level in order to deal with the complexities of copyright. Most institutions have research support offices which could expand to provide this sort of support. They could help to change the existing system where many authors are willing to sign more-or-less anything put in front of them by publishers in order to get their paper published. Authors would then be supported in ways that would allow them to maximise the potential impact of their work without unnecessary restrictions.

Some authors have another concern in the area of IPR. They are concerned that their work is more likely to be plagiarised if they self-archive it on an open-access server. There is, however, no empirical evidence to support this fear (although some publishers claim that they ask authors to sign over copyright in order to enable them to protect authors from plagiarism). It may be true that making material available online makes cut-and-paste plagiarism easier but this applies to all electronic information not just that which is openly accessible. What can be said in favour of open access is that it makes *detection* of plagiarism easier. Many automatic plagiarism detection services can operate better when they can move around documents without barriers. For this reason, some have suggested that open access is actually more likely to discourage serious attempts at plagiarism. A determined plagiarist is most likely to choose a more obscure work to copy so that his or her dishonesty will not be noticed. In any case, it would be rather perverse of authors to prefer their work to remain in relative obscurity (limiting its potential impact) in order simply to guard against a hypothetical risk of plagiarism.

Perhaps the major barrier to widespread self-archiving is that authors just do not see the point. There are two related issues here. The first is the argument that the existing system of scholarly communication works, and that self-archiving will disrupt it without replacing it with anything workable. The second is that, whether the existing system works optimally or not, it is the reality within which researchers are required to work – all the reward-mechanisms within their institutions and subject communities (promotion, peer recognition etc) are based on it, not on self-archiving.

At best then self-archiving is an unnecessary distraction. At worst, it is a dangerous innovation which has the potential to weaken the 'tried-and-tested' system.

The fear that self-archiving will fatally undermine the existing system is a common one but it does not seem to be borne out by the empirical evidence. The arXiv service has not destroyed journals. Journals are still valued by physicists for the quality certification function they perform. Taking this on board, most advocates of self-archiving support the idea that e-print repositories should complement rather than replace the existing system of peer-reviewed publication – at least in the short and medium-term. However, it is reasonable to expect that over time the character of journal publishing will alter if self-archiving becomes widespread. Publishers are likely to become managers of the peer review process (and perhaps providers of copy editing and formatting services) rather than distributors of content. However, this will probably happen gradually over a number of years and in the meantime self-archiving can proceed immediately. There is no immediate need for alternative business models to be in place. These will evolve naturally as practices change (their possible final shape is discussed in the section on the future, below).

Even if self-archiving is not likely to undermine the strengths of the existing system of scholarly communication, the fact remains that there is often little personal incentive for researchers to self-archive. Benefits of self-archiving apparent to physicists may seem rather remote from the concerns of other researchers and only seem to accrue if everybody does it rather than a few enthusiasts going their own way. Inertia rather than opposition is the biggest barrier to self-archiving at the moment. It needs to be taken seriously. Supporters of self-archiving are beginning to recognise that this can only be addressed by sustained advocacy within subject communities, institutions, and other stakeholder organisations (including national and international agencies).

Another way of encouraging self-archiving is to put real support services in place to facilitate it. Making self-archiving as easy as possible for researchers will help to ensure that it is not just another administrative burden (the final, common objection to self-archiving). This has already been discussed in relation to legal advice on copyright but other services are also useful. Important among these is 'self-archiving by proxy'. Rather than expect authors themselves to self-archive (convert their files into acceptable formats, create the appropriate metadata, and deposit their work in the repository), institutional support services could offer to do this for them (if provided with the original file). Anecdotal evidence seems to indicate that such measures are going to be necessary if self-archiving is to enter the mainstream.

Self-archiving initiatives

In order to create some kind of momentum for the self-archiving movement, a number of initiatives have been set in motion in various countries. For example, in the UK, the FAIR programme has sponsored a series of development projects investigating (amongst other things) e-print repositories, e-theses services and associated intellectual property rights. FAIR (Focus on Access to Institutional Resources) is funded by the UK higher education funding councils' Joint Information Systems Committee (JISC). The FAIR programme builds on previous JISC activities, such as the development of the GNU e-prints software. FAIR began in the summer of 2002,

with a completion date of the end of 2005. It has funded a total of 14 different projects in UK universities which between them cost about £3 million (excluding institutional overheads). The programme was "inspired by the vision of the Open Archives Initiative" and aimed "to support the disclosure of institutional assets" (JISC, 2002). Pinfield (2003) has described the main features of the programme and evaluative accounts of its progress will begin to emerge in 2005.

Similar programmes are underway in other countries. One of the first was the Mellon-funded programme in the USA. In the US, there is also the DSpace initiative, the California Digital Library eScholarship Repository, and the Ohio State University Knowledge Bank. In the Netherlands, the DARE programme (Digital Academic Resources) is now up and running, and this has been followed in Germany by the DINI (Deutsche Initiative für Netzwerkinformation) initiative. In Canada, the Canadian Association of Research Libraries is sponsoring the Institutional Repositories Pilot Project, and in Australia, the Australian government is funding the Research Information Infrastructure Framework for Australian Higher Education programme. These different initiatives have slightly different aims and emphases but they are all attempting to address in practical ways some of the barriers to self-archiving outlined above and to kick-start self-archiving in institutions and subject communities.

Many individual institutions also now have local initiatives in the area of self-archiving. These are often run by the library and information service but are beginning to capture the interest of researchers (albeit sometimes slowly). A few advocates of self-archiving have managed to secure institutional policy-level backing. A good example of this is the Queensland University of Technology in Australia where there is now a policy in place which requires authors in the institution to self-archive their work in the institutional repository if the publisher copyright agreement permits it (Queensland University of Technology, 2003).

All of this is happening in a climate of greater interest in and support for open-access. 2003 was the first year when there were regular items on the scientific literature about open-access issues, covering both open-access repositories and journals. Scientific publishing in general and open-access in particular also featured regularly in the financial, education, and mainstream press. These news articles were partly generated by interest in the initiatives already mentioned. They were also a response to the increasing number of policy statements supporting open-access launched during 2003. These included the Bethesda statement (2003, from US research funders), the Berlin Declaration (2003, from German funders), and the Wellcome Trust statement (2003, from a leading funder in the UK).

The number of repositories and e-prints has grown rapidly since 2001 but is still relatively small. Mark Ware (2004) provided a snapshot of the field in January 2004. He identifies about 250 OAI Data Providers, 45 of which are institutional repositories. The median number of records in each repository was 314. For institutional repositories (excluding the CERN pre-print server), the mean number of documents per site was 1250 and the median number 290.

Unresolved issues

A number of significant unresolved issues remain in the field of self-archiving. Some of the most important of these are discussed below. They include discipline differences, definitions of 'publication', versioning issues, digital preservation, costing and funding models, and metadata standards.

The issue of discipline differences in perhaps the most important issue which requires further work. Will all disciplines naturally converge on a single model of communication based on e-print repositories? Ginsparg (1999a) argues that this is the case:

> "Regardless of how different research areas move into the future (perhaps by some parallel and ultimately convergent evolutionary paths), I strongly suspect that on the one- to two-decade time scale, serious research biologists will also have moved to some form of global unified archive system, without the current partitioning and access restrictions familiar from the paper medium, for the simple reason that it is the best way to communicate knowledge, and hence to create new knowledge." (Ginsparg, 1999a)

This optimistic view is, however, questioned by Kling and McKim (2000). They argue that different disciplines have developed different cultures of communication, and that those differences are likely to persist for the foreseeable future in the electronic era. They provide an account of the different practices which persist in different disciplines which lead them to conclude that e-print repositories may not be universally adopted.

The question of whether certain disciplines are more inclined than others to accept self-archiving needs considerably more work. Some have observed that there appears to be a correlation between disciplines that have pre-existing pre-print cultures and those that have developed e-print repositories. It is, however, difficult to know how to read this. Whilst an inclination to communicate informally through circulation of pre-peer reviewed research may explain early adoption of self-archiving, it may not necessarily be an indicator that other disciplines will not adopt e-print repositories in the medium or long term. Other disciplines may adopt self-archiving when it has become more formalised and will perhaps limit their postings to post-prints only. E-print repositories could be (and need to be) set up and managed in such a way that they can accommodate the variety of cultures of communication that exist in different subject communities.

A related question is important here: does self-archiving a paper constitute publication? Stevan Harnad (2001b) argues forcibly that it does not. He defines 'publication' in a very particular way to mean the appearance of a paper in a peer-reviewed journal. Self-archiving a pre-print is therefore not publication. However, a number of journals have policies that they will not publish papers already made available on e-print repositories, regarding this as prior publication. The fact that the situation is ambiguous indicates that self-archiving is part of a trend in which the whole notion of publication is becoming more fluid. Publication may become more a process than a single event and the norms of such a process still need to be worked out.

This leads to another issue: versioning. At present, peer-reviewed journals provide 'the version of record' – the definitive version of the author's work which can be cited and archived. This version has normally been revised by the author following peer review, and copy-edited and formatted by the publisher. In an open-access e-print repository environment, what is the version of record? Post-prints are the final version of the paper produced by the author but have not been altered by the publisher. Does this matter? In the short term, researchers will probably need to continue to cite the article as published by the journal publisher (even if they initially access the paper via an open-access repository). In the longer term, it is possible that papers held in repositories may become the version of record. Copy-editing does not necessarily have to be carried out by publishers. It could be provided (if it considered to be essential) as a stand-alone service to authors before the final version is deposited in an e-print repository. However, even in a situation where such arrangements are in place, versioning remains an issue. In an open-access environment it is probable that many copies of the same paper may be made and then stored in different places (the so-called 'many copy problem'). Suber (2004) has outlined the pros and cons of this phenomenon and has suggested ways in which the problems might be addressed. However, more work on this is certainly needed.

In a system (still hypothetical) where the paper held in a repository becomes the version of record it will, of course, be necessary to ensure that version is preserved. Even before this, the issue of preservation of self-archived material is an important one. Pinfield and James (2003) have put forward arguments to support the case for the preservation of selected e-prints in the current situation. However, issues surrounding the costing, funding and management of preservation still need a great deal of further work.

Preservation is not the only activity associated with self-archiving where there needs to be more work on costing and funding models. In fact, the whole field of self-archiving requires further economic analysis and modelling. Barton and Walker (2003) have published some work on the costs of setting up and running an institutional repository, but these include a number of elements which do not necessarily have to be included in a simple e-print archive. On a larger scale, work needs to be done on the costs of a whole scholarly communication system which has open-access e-print repositories at its centre. Apart from running the repositories, the main essential cost would be administering the peer review process (assuming that the hidden costs of author time, and referee time continue to be covered in other ways). Traditional journals will continue to provide these services in the short term but if their subscription incomes fall as content becomes more easily available on open access, they would need to secure their income in different ways. It is possible that publishers could continue to provide quality-control services but would perhaps need to secure their income at the input stage, charging for the peer-review process, rather than for subscriptions. In the long term, other stakeholders, such as learned societies or consortia of institutions, may provide peer-review services on a cost-recovery or even profit-making basis. The assumption is often made that e-print repositories are likely to result in cost savings for institutions since publishers would not be able to charge such high prices for content but this assumption requires further testing and analysis.

At present, publishers (either the primary journal publishers or secondary publishers) generate metadata to enable article searching. The self-archiving process involves the creation of metadata, either by the author or a proxy. The structuring of that metadata is at present, however, very variable. More work needs to be done on the question of how this metadata could be standardised. Standardisation could occur at Data Provider level, with Data Providers agreeing detailed standards. Alternatively, it could occur at the Service Provider level, with post-harvesting normalisation (or even enhancement) of metadata. Both create potential technical and organisational challenges which need to be addressed.

The future of self-archiving

The future of self-archiving, particularly in relation to peer-reviewed journals, remains to be seen. If self-archiving is widely adopted in the way that its supporters expect there will be both discontinuities and continuities with the existing system of scholarly communication. Free and unrestricted access to the research literature will be a revolutionary discontinuity. However, there will also need to be important continuities, the most significant of which is perhaps peer-review. Scholarly communication needs robust quality control mechanisms. The majority of self-archiving advocates regard peer-review as a given which might be streamlined but should not be undermined. It remains to be seen whether some elements of the publishing system can be radically changed whilst at the same time leaving others intact.

The questions of what a new system with self-archiving at its centre would look like and how the transition might occur are important ones. Some have speculated on how the changes could unfold. Harnad (2001a) cautiously suggests possible scenarios. Once the scholarly literature has been self-archived, he then suggests:

> "One possible outcome is that that will be the end of it. The refereed literature will be free online for those who want it and cannot get it any other way, but those who can afford to get it the old way via paying journals will continue to do so. In this event, the access/impact problem will be solved...
>
> An alternative outcome is that when the refereed literature is accessible online for free, users will prefer the free version (as so many physicists already do). Journal revenues will then shrink and institutional savings grow, until journals eventually have to scale down to providing only the essentials (the quality-control service), with the rest (paper version, online PDF version, other 'added values') sold as options." (Harnad, 2001a, 1025)

Harnad's account is intentionally sketchy. He does not attempt to go into any detail of what the economics of the system might be, for instance. Harnad has in fact always been reluctant to speculate in any detail on the long-term future, fearing that discussion on hypothetical scenarios may be a distraction from the immediate imperative to ensure the scientific literature is self-archived in the short term.

Others have not been so cautious. Crow (2002) describes a model where the different components of peer-reviewed journal publishing are disaggregated and could potentially be carried out by different parties. Simplifying this kind of analysis, it is

clear that journals currently provide two essential features of scholarly publishing: peer review and distribution of content. In a new model, open-access e-print repositories could become vehicles for the distribution of content. The issue of who would then provide peer review is a moot point. At present, peer review is carried out by researchers, overseen by an editor and editorial board under the umbrella of a journal. The same parties (expert researchers overseen by a group of senior academics) could continue to provide peer review with or without a journal title as an umbrella. Learned societies or consortia of institutions could form peer-review groups to provide refereeing of papers outside of the traditional journal environment. Papers in repositories would then be 'quality-stamped' in some way to indicate to users that the work had undergone peer-review.

Such a practice does not necessarily mean the end of journals. It does mean journals would be different. 'Overlay journals' may develop, where papers located in archives are selected and brought together in virtual journal issues. This process of selection could involve peer review. Smith (2000) has described how journals might transform themselves in order to coexist with e-print archives. Journals in this view can "'overlay' what already exists, as opposed to communicating new, original content." (Smith, 2000, 47). They do, however, continue to provide peer review.

Peer review will remain a central feature of scholarly communication for the foreseeable future but other forms of quality assessment may also develop to complement it. At a pre-publication stage, as already discussed, institutional archives may put in place quality screening of various kinds before a paper is made live on the repository. Quality indicators may also be developed at the post-publication stage. If content is available on open access, counts of downloads and citations could be easily calculated at the article level for all the literature. Such metrics could be used to provide a post-publication assessment of the significance of a paper.

A system might develop then with several layers of quality control through which research output would pass, in response to which the content would potentially go through a number of iterations. There would be an initial quality screening before a paper was posted on a repository. This might prompt some changes to the paper. Following this, there would be a stage in which the author might receive comments from colleagues in the research field on the pre-print, and make any necessary changes. The paper would then be submitted for formal peer review. Changes would normally be expected at this stage. Mounting the post-print might generate further scholarly discussion and possible corrections, rebuttals or updates. The final stage of quality control, that of citation analysis and related metrics, would be less likely to produce changes in the article but would be likely to prompt further work. It would certainly help identify the key papers on which the subject community was building its ongoing work.

Guédon (2002) has described a multi-layered system like this. He argues that open-access archives should be developed to incorporate quality certification mechanisms so that they can exist apart from traditional journals. He goes further, suggesting that widespread use of open-access archives may even lead to the demise of the traditional scientific paper.

> "In its place may gradually emerge a more fluid and flexible mode of scientific communication where a given individual could contribute as little or as much as he/she wants, so long as it is significant and accepted by his/her peers." (Guédon, 2002, 12)

Scholarly communication would consist in an ongoing flow of information facilitated by interoperable open-access repositories.

Conclusion

Guédon's vision may be some way off becoming reality even though the technology is already in place to achieve it. What needs to develop now are communication cultures and management frameworks which take advantage of the technical possibilities. Considerable progress has already been made, but a great deal remains to be done. Many institutions and other organisations have begun to implement practical repository initiatives. The next two to three years will tell us a great deal about whether or not a scholarly communication system based on e-print repositories will work. In the meantime, it is worth while to keep an eye on the big picture. The prospect of a scholarly communication system where academic authors can easily achieve the rapid and wide dissemination of their output and where readers can gain free and unrestricted access to the literature is worth pursuing.

Acknowledgements

Thanks to Bill Hubbard for his useful comments on drafts of this paper.

About the Author

Stephen Pinfield is Assistant Director of Information Services at the University of Nottingham. He is Director of the SHERPA institutional repository project. He is also a member of the CURL (Consortium of University Research Libraries) and SCONUL (Society for College, National and University Libraries) groups on scholarly communication.

References

Barton, Mary R. and Walker, Julie Harford (2003). 'Building a business plan for DSpace, MIT libraries' digital institutional repository'. *Journal of Digital Information*, 4, 2. Available at http://jodi.ecs.soton.ac.uk/Articles/v04/i02/Barton/.

Berlin Declaration (2003). *Berlin Declaration on Open Access to Knowledge in the Sciences and Humanities*, Berlin, Max-Planck-Gesellschaft. Available at http://www.zim.mpg.de/openaccess-berlin/berlindeclaration.html.

Bethesda statement (2003). Available at http://www.biomedcentral.com/openaccess/bethesda/.

Brown, Cecelia (2001). 'The e-volution of pre-prints in the scholarly communication of physicists and astronomers'. *Journal of the American Society for Information Science and Technology*, 52, 3, 187-200.

Crow, Raym (2002). *The case for institutional repositories: a SPARC position paper.* Washington, DC: SPARC. Release 1.0. Available at http://www.arl.org/sparc/IR/ir.html.

Crow, Raym (2004). *A Guide to Institutional Repository Software*, New York: Open Society Institute. 2nd ed. Available at http://www.soros.org/openaccess/software/.

Gadd, Elizabeth, Oppenheim, Charles and Probets, Steve (2003). 'RoMEO Studies 1: The impact of copyright ownership on academic author self-archiving'. *Journal of Documentation*, 59, 3, 243-277. E-print available at http://www.lboro.ac.uk/departments/ls/disresearch/romeo/RoMEO%20Studies%201.pdf.

Ginsparg, Paul (1999a). 'Journals online: PubMed Central and beyond'. *HMSBeagle*, 3-16. Available at http://www.biomednet.com/hmsbeagle/61/viewpts/page5. Cited in: Kling and McKim (2000).

Ginsparg, Paul (1999b). 'Re: The significance of the LANL preprint server'. AmSci Forum Email Discussion List, 23 July. Available at http://www.ecs.soton.ac.uk/~harnad/Hypermail/Amsci/0347.html.

Guédon, Jean-Claude (2002). 'Open access archives: from scientific plutocracy to the republic of science'. *Proceedings of the 68th IFLA Council and General Conference, August 18-24, 2002*. Available at http://www.ifla.org/IV/ifla68/papers/guedon.pdf.

Harnad, Stevan (1998). 'Re: Savings from Converting to On-Line-Only: 30%- or 70%+ ?'. AmSci Forum Email Discussion List, 31 August. Available at http://www.ecs.soton.ac.uk/~harnad/Hypermail/Amsci/0052.html.

Harnad, Stevan (1999a). 'Advancing science by self-archiving refereed research'. *Science dEbates* 31 July. Available at http://www.sciencemag.org/cgi/eletters/285/5425/197#EL12.

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

Harnad, Stevan (2001a). 'The self-archiving initiative'. *Nature*, 410, 26 April, 1024-1025 and *Nature: webdebates*. Available at http://www.nature.com/nature/debates/e-access/Articles/harnad.html.

Harnad, Stevan (2001b). 'For whom the gate tolls? How and why to free the refereed research literature online through author/institution self-archiving, now'. Available at http://www.cogsci.soton.ac.uk/~harnad/Tp/resolution.htm.

JISC (2002). *Circular 1/02: Focus on Access to Institutional Resources Programme (FAIR)*, Bristol: Joint Information Systems Committee. Available at http://www.jisc.ac.uk/index.cfm?name=circular_1_02.

Kling, Rob and McKim, Geoffrey (2000). 'Not just a matter of time: field differences and the shaping of electronic media in supporting scientific communication'. *Journal of the American Society for Information Science*, 51, 14, 1306-1320. E-print available at http://arxiv.org/abs/cs.CY/9909008.

Lawrence, Steve (2001). 'Free online availability substantially increases a paper's impact'. *Nature*, 411, 31 May, 521 and *Nature: webdebates*. Available at http://www.nature.com/nature/debates/e-access/Articles/lawrence.html.

Lynch, Clifford A. (2003). 'Institutional repositories: essential infrastructure for scholarship in the digital age'. *ARL Bimonthly Report*, 226. Available at http://www.arl.org/newsltr/226/ir.html

Pinfield, Stephen (2001). 'How do physicists use an e-print archive? Implications for institutional e-print services'. *D-Lib Magazine*, 7, 12. Available at http://www.dlib.org/dlib/december01/pinfield/12pinfield.html.

Pinfield, Stephen (2003). 'Open archives and UK institutions: an overview'. *D-Lib Magazine*, 9, 3. Available at http://www.dlib.org/dlib/march03/pinfield/03pinfield.html.

Pinfield, Stephen, Gardner, Mike and MacColl, John (2002). 'Setting up an institutional e-print archive'. *Ariadne*, 31, March-April. Available at http://www.ariadne.ac.uk/issue31/eprint-archives/.

Pinfield, Stephen and James, Hamish (2003). 'The digital preservation of e-prints'. *D-Lib Magazine*, 9,9. Available at http://www.dlib.org/dlib/september03/pinfield/09pinfield.html.

Queensland University of Technology (2003). *Policy F/1.3 E-print repository for research output at QUT*, Brisbane: Queensland University of Technology. Available at http://www.qut.edu.au/admin/mopp/F/F_01_03.html.

Smith, Arthur P. (2000). 'The journal as an overlay on preprint databases'. *Learned Publishing*, 13, 1, 43-48. Available at http://www.ingentaselect.com/alpsp/09531513/v13n1/contp1-1.htm.

Suber, Peter (2004). 'The many-copy problem and the many-copy solution'. *Open Access Now*, 14, 15 March. Available at http://www.biomedcentral.com/openaccess/archive/?page=features&issue=14.

Swan, Alma, P. and Brown, Sheridan, N. (2004). *JISC/OSI Journal Authors Survey: Report*, Truro: Key Perspectives Ltd. Available at http://www.jisc.ac.uk/uploaded_documents/ACF655.pdf.

Ware, Mark (2004). *Publisher and Library/Learning Solutions (PALS): Pathfinder Research on Web-Based Repositories: Final Report*, Bristol: Mark Ware Consulting Ltd. Available at http://www.palsgroup.org.uk.

Warr, Wendy A. (2003). 'Evaluation of an experimental chemistry pre-print server'. *Journal of Chemical Information and Computer Sciences*, 43, 362-373.

Wellcome Trust (2003). *Scientific Publishing: A Position Statement by the Wellcome Trust in Support of Open Access Publishing*, London: Wellcome Trust. Available at http://www.wellcome.ac.uk/en/1/awtvispolpub.html.

Web sites

ARC http://arc.cs.odu.edu/
arXiv http://uk.arxiv.org/
California Digital Library eScholarship Repository http://escholarship.cdlib.org/
CERN Scientific Information Service http://cds.cern.ch/
Chemistry Preprint Server http://www.chemweb.com/preprint
Citebase http://citebase.eprints.org/cgi-bin/search
CogPrints http://cogprints.ecs.soton.ac.uk/

DARE http://www.darenet.nl/en/toon
DINI http://www.dini.de/dini/arbeitsgruppe/arbeitsgruppe_details.php?ID=9
DSpace http://www.dspace.org/
eprints.org (GNU eprints) http://www.eprints.org/
FAIR http://www.jisc.ac.uk/index.cfm?name=programme_fair
Institutional Repositories Pilot Project (Canada) http://www.carl-abrc.ca/projects/ir/
JISC http://www.jisc.ac.uk
Ohio State University Knowledge Bank https://dspace.lib.ohio-state.edu/index.jsp
Open Archives Initiative http://www.openarchives.org
Open Society Institute http://www.soros.org/
RePEc http://repec.org/