

E-PRINTS AND THE OPEN ARCHIVE INITIATIVE: OPPORTUNITIES FOR LIBRARIES

Pauline Simpson

Southampton Oceanography Centre
Waterfront Campus, European Way
Southampton, SO 14 3ZH, England
ps@soc.soton.ac.uk

ABSTRACT: The culture of the ‘invisible university’ and the enabling electronic environment coupled with the crisis in scholarly communication has been the impetus for two new initiatives: alternative scholarly publishing and e-Print archives. Centralised subject e-Print archives have achieved cautious success and a complementary institutional based model is now being advocated. The presentation will discuss the opportunities for libraries in setting up an institutional e-Print archive within the Open Archive Initiative <http://www.openarchives.org/> and demonstrate how OAI global harvesting protocol will provide a mechanism to bridge the digital divide and offer free access to research literature for both developed and developing countries. A marine science e-Print archive based on IAMSLIC institutions is proposed.

KEYWORDS: e-print archives; e-publishing; Open Archive Initiative; information management

Introduction

At IAMSLIC 2000 in Victoria BC we were fortunate to have as a keynote speaker, Richard Luce, Director, Los Alamos National Laboratory Research Library New Mexico speaking on “Communicating science in the next generation: new implications for the evolving digital library.” His description and future vision for the Los Alamos e-Print Archive which he and Paul Ginsparg started in 1991 initially for the high-energy physics community, fired me with the determination to build an Ocean and Earth Sciences E-Print Archive at the Southampton Oceanography Centre.

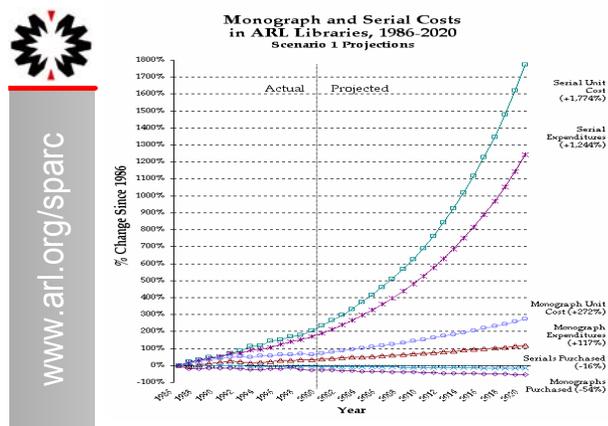
Crisis in scholarly communication

The journal is the primary publication channel for communicating research results and journal publishing is dominated by commercial ventures. Researchers write papers for the journals for no fee (or they pay page charges), they transfer copyright to the journal publishers for no fee; libraries then pay huge journal subscriptions to access those research papers which may have been written by one of their own research staff or they

may not be able to afford the subscription. Universities buy it several times over: in journal subscriptions, in photocopying licences, in study pack charges and in kind by the time contributed by voluntary editors.

Journal prices are spiralling, library budgets are reducing in real terms. Journal price inflation is generally about 9% per year although there are many examples of much higher increases. Since 1986, journal price inflation has been 291% compared to the increase in retail price index (the standard measure of inflation) of 70% (Ayrís 2002). Since 1986, 50% more journal titles are being published, leaving libraries struggling to maintain collections or required to cancel titles. The development of e-journals is not helping the problem because publishers are implementing e-journal pricing models which maintain or increase their income (Pinfield (2001).

Projection of periodical prices to 2020 (SPARC) Blixrud (2002)



Some solutions

1. Alternative publishing models – SPARC (Scholarly Publishing & Academic Resources Coalition), the ARL-initiative is a major player and has made efforts to facilitate competition in scientific communication through the creation of high-quality alternatives to commercial titles New players include:

University startups

- Univ. Arizona: *J. Insect Science*
- Univ. Bielefeld: *Documenta Mathematica*
- Univ. California: *eScholarship*
- Columbia Univ: *Earthscape*
- Cornell Univ/Duke Press: *Project Euclid*

- Univ. Warwick: *Geometry & Topology Publications*

Independent startups

- *Evolutionary Ecol. Rsrch.*
- *Internet Journal of Chemistry*

Hybrids

- *BioOne*

2. e-Print Archives - The success of this initiative can be measured by the collection of the major events in OAI and e-Print archiving activity from the last six months in the FOS Newsletter, August 8, 2002 (Suber 2002). The rationale for e-Print services has been discussed extensively elsewhere (Harnad 2000; Pinfield 2002).

What is an e-Print Archive

First some definitions: **reprints** are published papers, usually in paper copy ; **e-Prints** are electronic copies of any research output (journal articles, book chapters, conference papers etc) They can be: *preprints* – unpublished papers before they are refereed or *postprints* – papers after they have been refereed. An e-Print archive is an online depository of these and is usually Internet-based for free access and dissemination.

OAI-compliant e-Print archives share the same metadata [tags], making their contents interoperable with one another. Their metadata can then be harvested into global “virtual” archives that are seamlessly navigable by any user.

Open Archive Initiative

OAI activities (<http://www.openarchives.org/>) are supported by the Digital Library Federation, Coalition for Networked Information, and the National Science Foundation Grant No. IIS-9817416 (Project Prism). Its mission is to ‘develop and promote interoperability standards that aim to facilitate the efficient dissemination of content.

The Open Archives Initiative has its roots in an effort to enhance access to e-Print archives as a means of increasing the availability of scholarly communication. The facilitating software is the OAI Metadata Harvesting Protocol which creates the framework for interoperability between e-Print servers by enabling metadata from any OAI registered archive to be harvested and aggregated into one searchable database/interface. The metadata format is based on the Dublin Core Metadata Element Set and includes such information as author, date, title, subject and abstract. Archives can be OAI-compliant even if their full-text contents are not open-access. OAI-compliance applies only to their metadata. OAI interoperability is possible for all online content, open-access or not.

OAI categorize e-Print archives into two types: **Data Providers** are archives that expose metadata to harvesters, **Service Providers** harvest metadata and provide value added services with it. Conceptually these are different but in reality archives provide both a

service directly for users and provide metadata for automated harvesting. OAI is not itself a service provider.

Subject vs Institutional e-Print Archives

The best known e-Print archive is arXiv (originally at Los Alamos but now hosted at Cornell), initially set up for high-energy physics but now covering a broad physics base including atmospheric and oceanic physics, math, computing science and nonlinear science. This and other early e-Print services are subject based and hosted by a single domain and rely primarily on researchers to deposit their papers remotely using a self archiving protocol (although some archives do accept papers sent to them by email). Despite the success of the Los Alamos e-Print archive implementation in 1991 there has been cautious uptake by other subject communities and only a few centralised subject-based archives have been successful, e.g., CogPrints at Southampton, Chemistry Preprint Server, etc. As a complementary model, an institutional based archive offering both self and mediated archiving is now being advocated (Harnad 2000) and particularly in the excellent report, *The Case for Institutional Repositories: a SPARC position paper*, (SPARC 2002). Institutions can provide the technical, cultural and organisational framework to support the start-up and maintenance of e-Print archives; it is in their own interests to document and retain the scholarly output of their organisation and to make it available as widely as possible for research profile dissemination.

Libraries' Role

Many scholar-authors have already become active partners with their library, in playing a visible role in making research more accessible. For libraries, the reduction in budgets in real terms has seriously affected their ability to deliver access to the global knowledge base for their researchers. As such, libraries have a vested interest.

Why Libraries?

Information managers are the logical administrators of institutional archives and are now taking a lead role in their implementation. Their professional skills and expertise map to the e-Print support and maintenance profile:

- Positioned in the scholarly communication process
 - Recorders of institutional scientific output
 - Publishers on behalf of the institution
- Collection and dissemination of scholarly resources
- Deliverers of seamless systems, e-Resources etc
- Resource discovery mechanisms in digital environment
- Database expertise
- Records management
- Work with metadata and preservation
- Apply international standards uniformly
- IPR issues

- Central service providers
- Interact at all levels within the institution
- Network culture
- End user of free research corpus

Some early institutional (library) adopters around the world include: Australian National University, Hong Kong University. Universities in Europe: Utrecht, Groningen, Lund, Humboldt in Berlin, Max Planck Institute in Hamburg. In the USA, MIT, Caltech, California Digital Library, Library of Congress etc. In the UK a number of institutions put up demonstrator e-Print archives at Glasgow, Nottingham, Edinburgh, Strathclyde and Southampton.

Growing focus on open access

A useful timeline on e-Print activity over the last six months has been provided by Suber (2002) but no discussion on self-archiving would be complete without mentioning one of the original advocates, Prof. Stevan Harnad of the University of Southampton, who still hosts the discussion list American Scientists Forum 98 and maintains a healthy dialogue concerning developments in e-Print archives. There has also been a growing focus on open access with such services as: PubMed Central (www.pubmedcentral.nih.gov); and BioMed Central (www.biomedcentral.com) and support initiatives like the Public Library of Science (www.publiblibraryofscience.org) and the new International Scholarly Communications Alliance which intends to provide worldwide collaboration with scholars and publishers to establish equitable access to scholarly and research publications. More compelling than words however, e-Print advocacy was strengthened by funding initiatives from the Mellon Foundation who provided US \$1.5 million for seven US projects and the Budapest Open Access Initiative released in February 2002 and supported by the Soros Foundation Open Society Institute's Information Program, which aims to accelerate progress in the international effort to make research articles in all academic fields freely available on the internet. The OSI Information Program has committed funding of US\$1million per year for three years in support of open access projects

In 2002 the UK Higher Education Funding Council (HeFC), Joint Information Systems Committee (JISC) announced a funding call - FAIR (Focus on Access to Institutional Resources) under their DNER Learning and Teaching/Infrastructure Development programme. Successful applicants included Glasgow, the Consortium of University Libraries (CURL), UKOLN (a bid which includes OCLC as a partner), and Southampton University.

University of Southampton e-Print Archive Project

The project entitled **TARDis** (**T**argetting **A**cademic **R**esearch for **D**issemination and **dI**sclosure) has been funded for 30 months beginning August 2002. TARDis will be building a sustainable multidisciplinary institutional archive of e-Prints to leverage the

research created within Southampton University using both self-archiving and mediated deposit measured against discipline culture.

Consideration will be given to including all types of research output in a variety of formats. It is based firmly on the experience of building pilot archives in both Ocean and Earth Sciences and in Electronics and Computer Science.

While developing the archive, TARDIS will be specifically feeding back into the pioneering e-prints software (<http://software.eprints.org/>) developed within the prestigious Intelligence, Agents, Multimedia Group in the University of Southampton. Looking at the adaptive process, we will be gearing it to provide ease of use by archive administrators and end users. Strategies and documentation will address technological, cultural and organizational issues and the development of the e-Print archive concept for use in wider applications.

The technical and management issues relating to electronic authentication will also be addressed in a related JISC funded project led by Information Support Services (ISS) at the University of Southampton and using the TARDIS archive as the test bed.

Requirements for setting up an institutional archive

Hardware and software requirements: (see www.eprints.org for updates)

- Any computer capable of running GNU/Linux or similar operating system.
- A GNU operating system. GNU/Linux (a very advanced and free UNIX-like operating system)
- The Apache WWW server.
- The Perl programming language, (Also a small number of additional modules, detailed in the installation document.
- The mod_perl module for Apache, which significantly increases the performance of Perl scripts. Note that the mod_perl supplied with RedHat 6.2 (i386 architecture) is broken, and should be replaced with this RPM.
- The MySQL Database, a free database system.
- The e-Prints Software itself!

The e-Prints software now V.2, was written at the University of Southampton and is freely available to download from the eprints.org website. Once installed it is automatically ready to generate metadata in a form which can be picked up by OAI

harvesters. Although the software is relatively straightforward to install it does require knowledge of Perl and MySQL so good technical support is required.

Once installed it is necessary to configure the metadata formats and the user interface. These are tasks logically carried out by library staff and require initial decisions on the look and feel of the host web interface; deposit types and document formats, the addition of other metadata and importantly whether a subject listing or thesaurus will be used.

Document formats

The defaults on e-prints software are: PDF, HTML, Postscript and ASCII but others can be added. Specialist document preparation formats often required by publishers, e.g., LaTeX also need to be considered particularly if the deposit is mediated by the library and therefore conversion to supported formats will need to be undertaken. Open source protocols are available to assist in conversion from non supported formats. Unless carefully checked, HTML conversion from Word is unsatisfactory, so decisions on deleting some of the default formats might also need to be made.

At present e-Print archiving addresses the activity of depositing an item rather than addressing preservation strategies. Archive managers will want however to investigate the application of principles stated in the Open Archive Information System reference model for strategies on long-term accessibility, reliability and integrity... (Hirtle 2001).

Subject Index / Thesaurus

The Library of Congress is the default general subject classification but only to three levels. For a focussed marine science e-Print archive there is not sufficient granularity at this level of LoC so consideration might be given to loading the ASFIS thesaurus instead. Experience has shown that too detailed an index will deter would be depositors and if you are part of a wider university archive this may not be sensible. Some e-Print archives have decided not to use a subject classification scheme but to rely on keywords in title and the abstract and additional natural language keywords added by the depositor. These decisions must be made before starting to deposit papers since it is difficult to change once content is loaded.

Metadata

E-prints software was written as a self archiving tool and is OAI compliant and will produce the necessary Dublin Core metadata, however, it does not include some data fields needed for institutional applications, e.g., departments, research groups, etc. Whilst this metadata is not essential for OAI harvesting it is a retrieval requirement to enable the e-Print archive to be used as an institutional research management tool.

All individual implementers of e-Print archives see the need to 'customize' the basic e-prints software. At present there is no easy- to- use editor to enable this and it is often

necessary to climb into code. The Southampton University e-print archive team, in collaboration with the e-prints software designer, has undertaken to look at this requirement as part of the TARDIS Project.

Policy considerations

There are permutations in the institutional e-Print archive model, from total self archiving by the author to full mediated archiving by the archive team. Taking responsibility for the institutional archive will task the team with addressing the administrative and operational load, definitive authentication of depositors, quality control of the metadata and breadth of collection policy. Defining the institutional policy on copyright; and standards for long-term preservation are areas of deep discussion.

Advocacy

Early institutional e-Print archives have experienced problems with the acquisition of content. Discipline culture and the self-archiving protocol have been suggested as barriers. Researchers have also raised issues on copyright, quality control, particularly peer review and undermining the status quo, with an emphasis on their work load. At Southampton Oceanography Centre the response to a call for papers was to receive 50 in the first month which we felt was encouraging. However advocacy discussions with other Schools has shown that some disciplines will take more convincing.. Part of the TARDIS Project will be to document and build on these experiences and it is hoped that mediated archiving will resolve some of these issues. Advocacy methods include most importantly the e-Print archive itself, a web site, briefing papers to management, leaflets, institutional magazine, presentations at departmental meetings and committees, special advocacy events and personal contact.

OAI Registration

Apart from setting up the e-Print archive a key action is to register the archive as an OAI compliant data provider. As a consequence of registering with OAI periodic compliance and robustness testing is carried out. Having done this it is necessary to register with individual service providers such as ARC

Service Providers

Cross archive searching services such as ARC, harvest metadata from e-Print archives registered with it and provides a search engine together with a simple and advanced search interface, providing a resource discovery mechanism similar though not the same as a Z39.50 search interface.

The screenshot shows the homepage of the ARC - A Cross Archive Search Service. At the top left is the 'arc' logo in red and black. To its right, the text reads 'ARC - A Cross Archive Search Service' and 'Old Dominion University Digital Library Research Group'. A navigation menu includes links for Home, Simple Search (highlighted), Advanced Search, Browse, Administration, OAI, and Help. Below the navigation is a 'Simple Search' section with a search box, a 'Group Results By' dropdown set to 'Archive', and a 'Sort Results By' dropdown set to 'Relevance Ranking'. A 'Search' button is located below these fields. To the right of the search area is a 'Related Links' section containing three links: 'Download Source Code', 'Other OAI Service Provider', and 'DP9- An OAI Gateway Service for Web Crawlers'. At the bottom of the search area, a disclaimer states: 'Arc is an experimental research service of Digital Library Research group at Old Dominion University. Arc is used to investigate issues in harvesting OAI compliant repositories and making them accessible through a unified search interface. It is not a production service and may be subject to unscheduled service interruptions and anomalies.'

A Marine Science Cross e-Print Archive Search Service?

One of the main discussion topics within IAMSLIC is how to assist less developed countries to acquire full text access to the marine science literature. The concept of e-Print archives will provide the mechanism:

- IAMSLIC members implement institutional e-Print archives
- Register with OAI
- IAMSLIC set up a cross archive search service (precedent is IAMSLIC Z39.50 Distributed Catalog) with members archives registered as targets
- Harvest members metadata
- Provide search engine and interface

An IAMSLIC Marine Science e-Print Cross Search Archive is born

Bridging the digital divide

e-Print archives can provide free open access to the world's research literature. Implementing e-Print archives is a new challenge for libraries and will provide multiple benefit for :

- Researchers' profile
- Institutions' profile
- Library's profile
- Developing nations

REFERENCES

arXiv. [Online]. Available: <http://www.arXiv.org> [Accessed: 12 Sep 2002].

Budapest Open Access Initiative. [Online]. Available: <http://www.soros.org/openaccess> [Accessed: 12 Sep 2002].

Dublin Core Metadata Elements. [Online]. Available: <http://dublincore.org/documents/dces/> [Accessed: 12 Sep 2002].

eprints software. [Online]. Available: <http://www.eprints.org/> [Accessed: 12 Sep 2002].

Harnad, S. The self archiving initiative *Nature, Webdebates*. [Online]. Available: <http://www.nature.com/nature/debates/e-access/Articles/harnad.html> [Accessed: 12 Sep 2002].

Harnad, S. For whom the bell tolls – how and why to free the refereed research literature online through author/institution self archiving now. [Online]. Available: <http://www.cogsci.soton.ac.uk/~harnad/Tp/resolution.htm>. [Accessed: 12 Sep 2002].

Hirtle, P. 2002. OAI and OAIS: what's in a name? *D.Lib magazine* 7(4). [Online]. Available: <http://www.dlib.org/dlib/april10/04editorial.html> [Accessed: 12 Sep 2002].

Nixon, W. 2002. The evolution of an institutional e-Prints archive at the University of Southampton. *Ariadne* Issue 32. [Online]. Available: <http://www.ariadne.ac.uk/issue32/eprint-archives> [Accessed: 12 Sep 2002].

OAI Metadata Harvesting Protocol. [Online]. Available: <http://www.openarchives.org/OAI/openarchivesprotocol.html> [Accessed: 12 Sep 2002].

Open Archives Initiative. [Online]. Available: <http://www.openarchives.org> [Accessed: 2 Sep 2002].

Pinfield, S., Gardner, M. & McColl, J. 2002. Setting up an institutional e-Print archive. *Ariadne*, 31. [Online]. Available: <http://www.ariadne.ac.uk/issue31/eprint-archives> [Accessed: 12 Sep 2002].

SPARC. 2002. *The Case for Institutional Repositories: a SPARC position paper*. [Online]. Available: <http://www.arl.org/sparc/IR/ir.html> [Accessed: 10 Oct 2002].

Suber, P. 2002. *FOS Newsletter*. [Online]. Available: <http://topica.com/lists/suber-fos/read>. [Accessed: 10 Oct 2002].