THE CONTRIBUTION OF OPEN ARCHIVES INITIATIVES AND INSTITUTIONAL REPOSITORIES TO THE ADVANCEMENT OF KNOWLEDGE

LESLIE CARR[1]; LESLIE CHAN[2;] JEN SWEEZIE[2]

[1]Intelligence, Agents, Multimedia
University of Southampton
lac@ecs.soton.ac.uk
[2]Bioline International and University of Toronto
chan@utsc.utoronto.ca
sweezie@utsc.utoronto.ca

Open access refers to peer-reviewed scientific and scholarly literature that are made available online to readers without access or restrictive barriers. The purpose of this workshop is to provide an in-depth look at one of the most immediate and cost-effective routes of providing open access to the literature, namely institutional self-archiving. Participants will learn about:

1. The principles and benefits of open access, with emphasis on the needs of developing countries (Chan)

2. The technical standards underlying open access: the OAI metadata harvesting protocol and OAI-compliant archiving software (Carr)

3. Technical requirements for installation and customization of Eprints software (Carr)

4. The workflow process of document conversion, metadata input, uploading, verification, and approval (Sweezie)

5. Institutional and policy recommendations (Carr and Chan)

INTRODUCTION

The Budapest Open Access Initiative (http://www.soros.org/openaccess) laid down recommendations for freeing the world's scholarly reviewed literature through both Open Access Publishing (OAP) and Open Access Archiving (OAA). With OAP, the emphasis is on transferring the cost from the readers to the producers , either through author-fee or other alternative revenue streams. With OAA, already published papers (as well as preprints) are archived, simultaneously or subsequently to conventional publication, in institutional archives that are interoperable.

OAA provides enormous and sustainable benefits to researchers and their institutions. The establishment of institutional archives brings greatly increased visibility to the research output of institutions and is already showing a three- to five-fold increase on the research impact of articles archived in this way. This strategy can therefore lead to immediate benefit, and is low-cost, equitable and highly appropriate as a means of levelling the playing field for access to information.

OAA are increasingly accepted by the more advanced countries and a growing number of archives are up and running, making many thousands of publications available to all free of charge (http://archives.eprints.org/eprints.php). Developing countries like India are also beginning to recognize the potential of OAA and the movement is likely to grow rapidly there. In Brazil, a number of OAP initiatives like SciELO and LANTINDEX are well known and are extremely

beneficial to scientific publications in Brazil. However, the benefits and means by mean OAA could advance research impact in Brazil is relatively unknown and so awareness raising is also important. It is important to note that existing journals can co-exist with OAA (as has been done by arXive.org and the major physics journals for over 10 year). And Bioline International has also demonstrated that OAP, OAA and print subscription could all exist in parallel and mutually reinforcing.

The publishing difficulties faced by many authors in the developing world can best be addressed by ensuring that their research is internationally available to all, raising visibility and impact. This way, its importance - to us all, as well as to neighbouring regions facing similar problems - can be globally recognised, partnerships established, the national science base strengthened and the feelings of professional isolation steadily eliminated. Open Access Archiving can begin to achieve this goal and in doing so bring enormous benefits to society through the immediate distribution of research.

Setting up and maintaining Institutional Archives is very low cost as free software are readily available. However, there are issues of personnel and institutional commitments that need to be addressed by each institution.

OPEN ARCHIVE INITIATIVE PROTOCOL FOR METADATA HARVESTING (OAI-PMH)

The Open Archive Initiative (www.openarchives.org) was formed in 2000, with the mission to develop and promote "interoperability standards" that aimed to facilitate the efficient dissemination and discovery of digital content. While the early concern of the Open Archive Initiative (OAI) was the interoperability of well-defined e-prints, it has been growing to extend interoperability to a broad range of digital objects such as datasets, video, databases, theses, technical reports and other grey literature, as well as added applications and services.

The OAI-PMH is designed to harvest metadata and associated resources that are distributed across different OAI-compliant servers, thus connecting all distributed servers into a seamless global digital library. OAI-MHP has built-in support for basic Dublin Core metadata, an internationally recognized standard used in library and digital-resource cataloguing (Dublin Core Metadata Initiative, dublincore.org).

In addition to the facilitation of sharing of metadata, the OAI also promote the model that separates data providers and service providers. As a result, authors and institutions that are only interested in making their publications openly available need not be worried about developing added services, as the latter could be developed by other agencies or institutions. Harvesting services such as the OAIster search engine is a good example of how this model works.

The OAI-MHP is now widely adopted by library, publishing, and scientific communities eager to ensure that their resources on the Web are interoperable with each other. The accessibility, and therefore impact, of materials is greatly reduced if they remain invisible to others because of the lack of interoperable standards

EPRINTS SOFTWARE

EPrints.org archiving software was developed to provide a simple solution for OAA archiving. Developed at the University of Southampton, UK, it is an open source environment which comes pre-configured for the most common requirements of a University institutional archive and can be installed and configured to your institution's particular requirements.

An EPrints archive is partly a database (handled by a genuine Relational Database Management System called *mysql*) and partly a web site (handled by a web server called *apache*). These two parts are yoked together by a collection of scripts (programs written in a programming language called Perl). Some of these scripts have to be run by the archive administrator from the command line, and some are invoked by the Web server in response to one of the users clicking on a link or submitting a form. Whichever way this happens, the effect is to update information in the database or to extract information to show in the Web site.

There are two types of information which the archive needs to store - information about the individual eprints themselves (which is stored in the database) and information about the archive's configuration (stored in separate configuration files in XML format)..

The standard EPrints documentation contains Installation instructions for RedHat Linux, and technically oriented individuals are encouraged to consult the documentation.

http://software.eprints.org/docs/php/installation.php

Skills Needed To Run The Archive

The eprints in an EPrints Archive are the electronically-distributed versions of journal articles and conference papers deposited by researchers to provide maximum access to their own research. The person(s) responsible for an EPrints archive therefore must be able to encourage as many researchers as possible to deposit as many eprints as possible with as much information about them as will make them usefully accessible, all in a timely and sustainable fashion.

At one level, the manager must be able to motivate the researchers for whom they are responsible by helping them to understand the significant benefits for themselves and their institution in terms of visibility and research impact. Without these benefits, the task degenerates into creating new administrative hurdles for an already overworked and naturally sceptical audience.

At another level, the archive manager must have sufficient information management expertise (or editorial experience, or librarianship skills) to be able to understand the issues that may arise with bibliographic metadata in general, as well as the additional information that their institution may wish to collect (is it important to store the identity of the Research Group or Project which produced a paper, how can an author be uniquely identified, are student theses or technical reports to be included, *etc.*).

BENEFITS OF AN EPRINT ARCHIVE

Once the archive is in place and registered with the OAI (Open Archives Initiative), it will automatically be included in a global program of metadata harvesting and other added-value services run by academic and scientific institutions across the globe (including federated searching, bibliographic extraction, citation analysis, subject mapping and visualisation, cataloguing and classification) which will provide many different ways of accessing the material stored in your institution's archive.

THE WORK FLOW PROCESS

Once the technical aspects of an EPrint server are set up, the archive must be customized to reflect the mandate and visions of its owner. Before the first EPrint can be deposited, owners must decide what types of file formats will be used, what subject headings will be made available and what other specific policies will govern the use of the server. Some of these decisions may require small technical adjustments to the free software provided. After establishing the basis for

how the server will be used, users will have to register to deposit articles, and should be made aware of the importance of the quality and accuracy of the metadata that is provided during the submission process. Some of the most important aspects of the submission process include identifying the item being deposited as "published" or not, and whether or not the source of the article is refereed. Users can later return to the same "EPrint" and correct errors in the metadata, add a now-published version of a pre-print that was originally deposited, or add new metadata such as keywords. Depositing an EPrint is only one portion of the work flow process associated with making articles available through an EPrint server. Editors or system administrators must also be designated. After an item is deposited, these editors or administrators are notified of the item and they must review items in the submission buffer. Editors are advised to check each submission carefully for accuracy, as inaccurate metadata will prevent interested users from retrieving relevant articles. Once approved, a successful EPrint will be released into the archive and will be readily available for anyone who wishes to view it.

The EPrints software comes with a number of different editorial features that allow EPrints administrators or editors to monitor the quality of submissions quite easily. Even a file that is accidentally added can be retrieved. Submissions that need more attention can be removed from the archive and returned to the originating user with notes. Subject headings can be added and modified as is necessary to accommodate changes in subject areas.

POLICY RECOMMENDATIONS:
Proposal by Stevan Harnad
Summary of Policy Recommendations:

(1) Universities need to adopt a self-archiving policy—an extension of their existing "publish or perish" policy to "publish with maximal impact." A potential model for such a policy is at <http://www.ecs.soton.ac.uk/~harnad/Temp/archpolnew.html> along with (free) software for creating a standardized online university CV, linking all entries for peer-reviewed articles to their full text self-archived in the university eprint archives <http://paracite.eprints.org/cgi-bin/rae_front.cgi>.

(2) University libraries need to help with the first wave of self-archiving, doing "proxy" self-archiving for those researchers who feel too old, tired, or busy to do the few keystrokes per paper that are involved.
<http://www.ecs.soton.ac.uk/~harnad/Tp/resolution.htm#7.3>.

(3) Research funding agencies such as NSF or NIH (US), HEFCE or EPSRC (UK), NSERC, CFI or FRSQ (Canada), or CNRS or INSERM (France) need to encourage self-archiving as part of the normal research cycle, requiring not only that the research findings be published, as they already require, but that their visibility and usage be maximized by making them openly accessible through self-archiving.
<http://www.ariadne.ac.uk/issue35/harnad/>.

(4) Scientometric performance indicators and analyzers such as <http://citebase.eprints.org/cgi-bin/search>—rather like Google, but based on citation links instead of ordinary links˜need to be created and used to demonstrate, monitor, measure, evaluate and reward the maximization of research impact through open access. Free online accessibility increases citation impact by 336% <http://www.neci.nec.com/~lawrence/papers/online-nature01/>.

(5) Journals need to support self-archiving by modifying their copyright transfer or licensing agreements to encourage self -archiving (as 55% of them already do, with most others agreeing on a per-paper basis if asked: so ask!): <http://www.lboro.ac.uk/departments/ls/disresearch/romeo/Romeo%Publisher%Policies.htm>.