A GUIDE TO SETTING-UP AN INSTITUTIONAL REPOSITORY
SEPTEMBER 27, 2002

TABLE OF CONTENTS

INTRODUCTION

An institutional repository (IR) is a digital archive of an academic institutions intellectual output. IRs adhere to an open access model, by centralizing and preserving the knowledge of an academic institution and making it accessible to anyone with internet access. In contrast with Eprints archives, institutional repositories are not discipline-specific, and aim to archive the entire range of a university's intellectual output. IRs also form part of a larger global system of repositories, which are indexed in a standardized way, and searchable using one interface, providing the foundation for a new model of scholarly publishing.

This guide, written for members of the **CARL Institutional Repository Pilot Project,** discusses the major steps that must be taken to set up an institutional repository and examines some of the issues involved in each step**.** The project is an initiative to implement institutional repositories at several Canadian research libraries—ensuring that Canadian institutions remain at the leading edge of innovation in scholarly publishing. The project will facilitate discussions of lessons learned and best practices for implementing IRs and will pave the way for other Canadian universities by examining the feasibility of IRs in the Canadian context.

The assumption is that participating institutions will be implementing their IR using Eprints Version 2 software. The aim is that the IRs will be OAI-compliant, thus being interoperable with each other and other OAI-compliant archives. The steps presented here are based on documentation produced by the Open Archives Initiative (www.openarchives.org), Eprints (http://software.eprints.org); and institutional repositories designed by several other universities:

- CalTech Computer Science Technical Reports (caltechCSTR), California Institute of Technology: A repository of technical reports - on behalf of the Caltech Computer Science Department. http://caltechcstr.library.caltech.edu/

- DSpace, Massachusetts Institute of Technology: A stable and sustainable long-term digital storage repository that provides an opportunity to explore issues surrounding access control, rights management, versioning, retrieval, community feedback, and flexible distribution capabilities. http://web.mit.edu/dspace/live/home.html

- Eprint Repository, University of Melbourne: A self-archiving institutional repository of research papers for the University of Melbourne. http://www.lib.unimelb.edu.au/eprints/home.htm

- Electronic Pre and Post Print Archive, Australian National University: An eprint containing pre and post prints in a number of subject categories such as: Arts; Biological Sciences; Engineering; Humanities; Medicine; Environmental Sciences; Social Sciences; Physical Sciences and Mathematics. http://eprints.anu.edu.au

- eScholarship Repository, University of California: The development of an infrastructure for digitally-based scholarly communication. http://escholarship.cdlib.org

- Glasgow ePrints Service, Glasgow University: An experimental Open Archive set-up to provide access to the full text of the research output of scholars, scientists and researchers at Glasgow University. http://eprints.lib.gla.ac.uk/

- Nottingham ePrints, University of Nottingham: An experimental project to investigate the institutional use of eprints services. http://www-db.library.nottingham.ac.uk/ep1/

PRE-IMPLEMENTATION

Setting up an institutional repository is not a trivial task. However, the biggest issue is actually deciding what you need and what policies the archive will have. The following is a list of the issues that should be addresses prior to software implementation, since they will effect the software configuration:

Content
The first step when creating an institutional repository will be to establish some content guidelines. A review of a number of existing repositories shows that institutional policies regarding content vary substantially. In general, e-prints/pre-prints archives tend to accept one or more editions of "working papers", while institutional repositories accept a wider range of content types, such as post-prints (copies of already published journal articles), conference papers, technical reports, etc. DSpace, MITs institutional repository, (http://web.mit.edu/dspace/live/implementation/guidelines.html), does not accept pre-prints, but rather only material that is "ready for publication". On the other hand, the University of California eScholarship Repository (http://repositories.cdlib.org/escholarship/) offers faculty a central location for depositing their pre-publication scholarship. Related to content policies will be the copyright guidelines, which will be discussed later on in this guide. Key decisions about the type of content accepted by the repository should be made before the implementation of software, as they will have implications on metadata and information fields.

Metadata
Institutional repositories must incorporate, index, and search items from diverse collections in diverse formats; deal with standard vocabularies from many different fields of study; and include metadata regarding digital object structure, administration, and content. Unqualified Dublin Core (www.dublincore.org) is the minimum metadata required for OAI interoperability, however, depending on the type of content in your repository, you may want to include other metadata sets (such as journal name or pagination-for post-prints).

Since OAI is based on the exchange of metadata, getting the metadata right is fundamentally important for a repository. The EPrints software is OAI compliant and will produce the necessary Dublin Core metadata for harvesting by service providers. EPrints configures a new archive with a set of metadata fields aimed at an archive of research papers.

Document Type: EPrints original document types are eprints, books, posters, and conference papers. Institutional repositories may want to expand on the default configuration to include additional document types such as journal article, preprint or thesis., depending on the content guidelines of the repository.

Subject Headings: Identifying a useful set subject headings will be one of the major challenges for repository implementers. Broad subject headings may be appropriate for a single institutional repository, however, as access to institutional repositories becomes federated, it becomes more problematic. How can a user profitably browse papers from a variety of repositories that use very different subject terminologies? The key will be to provide a set of descriptors that are both useful to experts in the field and those who are not.

Format: The default formats accepted by the EPrints software are Postscript, PDF, ASCII, and HTML. Archive managers may want to add to or take away from these

formats. Possible additions may be specialized formats for data sets, or other common formats, such as Rich Text Format. There are open-source utility programs available to convert from non-supported to supported formats. Consideration may also be given as to whether any of these default formats should be switched off. For example, HTML is very fluid format which is difficult to validate easily and some may choose not to accept documents in this format.

The EPrints software enables implementers to design their own subject hierarchy and load it into the database fairly easily. However, it is far more complex to alter this once you have started to upload documents. So it is important to get this right before uploading too many papers. To view some examples of additional content types, formats and subject headings you can visit the advanced search pages of some existing archives: http://eprints.anu.edu.au/perl/advsearch http://eprints.unimelb.edu.au/perl/advsearch and http://eprints.lib.gla.ac.uk/perl/advsearch

IMPLEMENTATION

It is suggested that, as a first step, a demonstration version using the default configuration is set-up. Previous experience has shown that the initial configuration can be quite a bit of work if you have complex requirements, and rather easy if the only part of the default configuration you want to change is the colour.

Installation of Eprints Software

GNU EPrints primary goal is to be set up as an open archive for research papers, and the default configuration reflects this. However, it can also be easily used for other things such as images, research data, audio archives - anything that can be stored digitally, but you'll have make changes to the configuration.

EPrints Version 2.1.1 is available free of charge at www.eprints.org. The software is already OAI-compliant and once it is installed, it is automatically ready to generate metadata in a form which can be picked up by OAI harvesters. Installation of the eprints software takes approximately one to two days. The installation of the simple default version of EPrints is straightforward but many sites want much more than this and the recent volume of traffic on the eprints-tech [6] list is testament to this (http://www.ariadne.ac.uk/issue32/eprint-archives/). Detailed installation instructions are also available on the eprints web site at http://software.eprints.org/docs/php/installation.php

Technical Requirements

The hardware and software requirements for Eprints version 2.1 software are as follows:

- Any computer capable of running GNU/Linux or similar operating system. Obviously, the faster, the better, but any Intel Pentium II processor will give good performance.

- A GNU operating system. GNU/Linux (www.gnu.org or www.linux.org) a very advanced and free UNIX-like operating system works just fine, and is in fact the development platform.

- The Apache WWW server (www.apache.org/httpd.html) another professional-quality free software product, often included with GNU/Linux distributions, such as that produced by RedHat (www.redhat.org).

- The Perl programming language (www.perl.com) also included with most GNU/Linux distributions.

- The mod perl (http://perl.apache.org) module for Apache (www.apache.org/httpd.html), 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 (www.mysql.com), a free database system.

- The EPrints software (http://software.eprints.org) itself!

Potential Problems
EPrints Version 2 is considered ready for use. There are still some issues and bugs, but hopefully not too many or too major. A list of known issues is in the file BUGLIST in the distribution (http://software.eprints.org/buglist.php). The "eprints-tech" discussion list (http://software.eprints.org/tech.php/) is good place to find solutions to problems you may encounter during implementation.

POST-IMPLEMENTATION

Interface Design
Eprints software provides a web interface for managing, submitting, searching/browsing, and downloading documents and it will take another few days to customize the web interface for the repository.

Archiving Policies
Who can submit, and how will submissions be monitored? Will authors self-archive or will authors submit articles to IR staff for mediated archiving? These are also questions that you will want to address before launching your repository. Self-archiving is the ultimate goal for most repositories and the EPrints software offers a very effective self-archiving facility. EPrints Version 2 offers a web-based registration process for submitters, however many institutions may wish to limit potential submitters to certain members of the institution or system administrators. In the case of e-prints archives, submission is generally open, and submitters need only register and deposit their work. Referred to as the distributed model, this model allows individual faculty to upload and manage their own scholarly output. On the other hand, centralized repositories, where

the repository staff are responsible for uploading material, can more closely monitor the content of the repository. Each model has its benefits and drawbacks. Self-archiving can act as a significant barrier for submission and a number of repositories have found that self-archiving is not feasible. Because it requires some level of IT literacy, some repositories have found that it is easier to attract users if the library deposits the items on behalf of users. The University of Nottingham concluded that mediated archiving was the only thing that worked for them (mainly because many users did not have the facilities to convert a word-processed file into a PDF document). The University of California's eScholarship repository uses a semi-distributed model that assigns management responsibility to organizational units (research units, departments) that then assist faculty with uploading their papers. On the other hand, University of Glasgow has found that early adopters (in disciplines such as Music, Life Sciences and Computing Science) had little difficulty in uploading their material into the archive, in a range of formats and they have been very positive about the experience. Each institution will need to consider alternative models in light of its particular circumstances.

Quality Control

The EPrints software has a submission buffer, in which all content must sit before it becomes publicly available. The system administrator can accept, edit or reject a submission at this stage. This allows the administrator to approve deposited material before it goes live, ensuring a certain level of quality control over metadata, formatting and in some cases, content of the deposited material. If there is a problem with a paper that has been deposited, it can be returned to the submitters "workspace", and the author is sent an e-mail explaining the problem. Again, repositories will want to outline policies regarding quality control of submissions, based on staff resources and type of repository.

Documentation

Repositories will also need to provide in some detail documentation to assist the users, both in submitting papers and accessing material in the repository. The Eprints web site provides sample instructions for submitters that can be used as the basis for help documentation by any repository (http://demoprints.eprints.org/help/).

Copyright

The approach taken towards copyright will play a pivotal role in the acceptance of an institutional repository service. Copyright is the number one question which members of the university ask about when introduced to an institutional repository and it is important to address concerns that might be raised in the minds of academics and managers. Copyright is fairly simple for pre-prints, which can be self-archived without seeking anyone else's permission because the author holds the copyright. However, for the refereed post-print, the author must investigate the copyright policies of the publishing journal. In some cases, authors retain copyright and unlimited rights after first publication in the journal. While in other cases, authors retain no rights to their work after it is published. In most cases it will be left up to the faculty member to discover the copyright restrictions applied by publishers. The University of Melbourne Eprint Repository assists faculty by maintaining a web page that links users to the copyright policies of specific journals (http://www.lib.unimelb.edu.au/eprints/home.htm#copyright) If there are

restrictions to copyright, authors can try to modify the copyright transfer agreement to allow self-archiving, or, failing that, can append or link the file to the already self-archived preprint.

Authors should also be reassured that they are not giving up their copyright by submitting their work into the repository. Some examples of repository copyright policies are provided below:

- **University of Melbourne Eprint Repository**
  (http://www.lib.unimelb.edu.au/eprints/home.htm)

  "The UMER administrators do not expect that any papers on the repository have not appeared anywhere else. There are no restrictions imposed by UMER on where a piece of work has appeared before. However, it should be noted that if your paper has appeared elsewhere, you may have transferred the copyright of the work from yourself to the journal publisher. If this is the case, the publisher may not allow you to include your work on UMER. Additionally, some publishers will not publish material that has previously been included on an eprint repository such as UMER. You will need to check with the publishers to whom you have submitted papers before including them on UMER. A list of some publishers and their policies on this matter can be found at the following link"

- **University of California eScholarship Repository**
  (http://repositories.cdlib.org/escholarship/policies.html)

  "Authors retain the copyright for all papers posted in the repository. The author agreement specifies a nonexclusive right to use. This means the author is free to reuse the content elsewhere, either in the same form or in revised form. If a working paper is published in a journal either in the exact same form or, more commonly, in revised form, many journals allow the working paper to continue to be disseminated over the web; however, some journals do require that the working paper be removed. It is up to the faculty member to check the terms of their agreement with the journal to see what is allowed. The repository would constitute noncommercial use. If you are interested in including a reprint of a journal article on your repository site, have the faculty member check their agreement with the journal to see if it is allowed. If it would not violate copyright, you're welcome to do so. You are the gatekeeper for your repository site, and it is up to you to decide what is appropriate--as long as it doesn't violate copyright and conforms to the few policies set by the eScholarship Repository."

In any case, authors should be encouraged, where possible, to retain their copyright by either submitting to journals that do not require sign-over or altering the copyright agreement to retain their copyright (or at least e-distribution rights). For more information about copyright, see "Is self-archiving legal?" (http://www.eprints.org/self-faq/#self-archiving-legal) and "What if the publisher forbids self-archiving the preprint? (http://www.eprints.org/self-faq/#publisher-forbids).

Registering the Repository

Once the software has been installed, the server needs to be registered with the Open Archives Initiative (http://www.openarchives.org/data/registerasprovider.html). The OAI maintains a list of OAI-compliant archives for OAI Service Providers to be able to visit. Before registering the archive, the OAI will perform a set of conformance tests on the repository, to ensure integrity of the registry. When this is completed, they will confirm by email and the archive will added to the public list of OAI compliant data providers (http://www.openarchives.org/Register/BrowseSites.pl) In the case that the repository fails to complete the tests, the repository will be removed from the registry and an email will be sent containing an explanation of why the repository did not conform. The OAI periodically retests repositories for their conformance.

Promotion and Advocacy

Setting an archive up is one thing, but getting users to participate in its ongoing development is quite another. One of the most difficult tasks in setting up the archive is getting the content. The participation of users is critical and will go through two important phases: first, the goal will be to just get enough content in place to set up a demonstrator. One strategy can be to recruit "early adopters" from fields that have a history of using e-prints servers such as physics, computer science, and chemistry. Once the demonstrator is in place, the second step will be to get a critical mass of content in order to provide a useful service and a repository more representative of all of the departments in the institution.

Some helpful tips offered by others are:

- Identify 'champions' in academic departments who can encourage colleagues to take part is often the most valuable approach.

- Faculty will be more comfortable with providing content if they do not think that the e-prints movement will undermine the 'tried and tested' norms of scholarly communication. The fundamental message should be 'do not stop submitting papers to peer reviewed journals - but please deposit them in the e-prints archive *as well*'.

- It is important whatever happens that e-print archives are run in such a way that they address the needs and working patterns of researchers. Things should be made as easy as possible for them to contribute.

- Set up a project web site that is linked to from the archive itself. This can act as a focus for developments and news.

- Publicize and promote the repository through university magazines, including the Library user newsletter; the distribution of literature about the value of institutional repositories, such as the SPARC *Create change* leaflet; and presenting at departmental meetings and university committees.

REFERENCES

Crow, Raym.(2002) *The Case for Institutional Repositories: A SPARC Position Paper*. Association of Research Libraries. Available at (http://www.arl.org/sparc/IR/ir.html). Last visited September 25, 2002.

Pinfield, Stephen;Gardner, Mike and John MacColl. (2002) *Setting up an institutional e-print archive*. Ariadne, Issue 31: April 11, 2002. Available at (http://www.ariadne.ac.uk/issue31/eprint-archives/) Last visited September 25, 2002.

Nixon, William J. (2002) *The evolution of an institutional e-prints archive at the University of Glasgow*. Ariadne Issue 32 , July 8, 2002 Available at (http://www.ariadne.ac.uk/issue32/eprint-archives/) Last visited September 25, 2002.