SHERPA and Institutional Repositories

Bill Hubbard

Published in *Serials*, 16, 3, November 2003, pp 243-247
Based on a paper given at the UKSG seminar
'The Open Archives Initiative: application and exploitation', London, 14 May 2003

Abstract
The SHERPA project (Securing a Hybrid Environment for Research Preservation and Access) has been set up to encourage change in the scholarly communication process by creating open-access institutional "e-print" repositories for the dissemination of research findings. The outcomes of the project - advice on building and maintenance of repositories, guidelines on IPR and copyright issues, advocacy materials to publicise an institution's repository - will be available to the whole HE community. This article looks at the terminology involved with such repositories and at the issues that they raise in their construction and use. It reviews the advantages of having an institutional basis for a repository and identifies the key issues that have arisen so far in project work.

The SHERPA Project

SHERPA (Securing a Hybrid Environment for Research Preservation and Access) is being funded by JISC, as part of the FAIR Programme, with additional funding from the Consortium of University Research Libraries (CURL). The partners in SHERPA are the Universities of Nottingham, Edinburgh, Glasgow, Oxford, the 'White Rose Partnership' of the Universities of Leeds, York and Sheffield, along with the British Library and the Arts and Humanities Data Service (AHDS). Seven more Associate Partners, drawn from CURL institutions, will join at the end of this year. All of the institutional partners will be building their own e-print repositories. The project will investigate the practical issues surrounding the creation and use of institutional "e-print" repositories and disseminate the findings to the HE community. The outcomes of the project will include advice on the creation and maintenance of repositories, together with standards-based preservation advice, guidelines on IPR and copyright issues, advocacy materials to encourage repository use within institutions, and open access to the e-prints themselves.

Terminology

There are many on-line collections of material that have developed, including collections of research material sometimes referred to as "archives". In formalising approaches to building such collections for on-line access, it was realised that the term "archive" is not strictly appropriate, given its specific and professional use in the library world. Using the term "archive" implies formal schemas of metadata and preservation strategies which many of the current ephemeral and opportunistic collections of material do not have. Therefore, in common with developing practice, the collections within SHERPA have been termed "repositories". It is intended to develop preservation strategies and formal metadata descriptions, but it remains to be seen whether these repositories will be required to be formal archives.

SHERPA is using the term "e-prints" to refer to an electronic version of a research paper or other similar output. As such, e-prints can be ‘pre-prints’ (pre-refereed papers), ‘post-

prints' (post-refereed papers), or other similar material, such as conference papers, book chapters, reports, etc. The definition of research output varies between disciplines: the definition of the term e-print therefore varies in a similar way. It is important to stress that the material that is being used as e-prints is material that is currently given away for free to publishers and published in scholarly journals - not books or commissioned articles for which the researcher expects a payment.

Institutionally based e-print repositories

SHERPA is focussed on creating significant collections of e-prints in institutionally-based repositories - web servers capable of being freely accessed and searched by researchers world-wide. Such a pilot repository is available at the University of Nottingham, for which a web link is given at the end of this article, as well as some other institutions. Similar subject-based repositories already exist; for example, arXiv for physics (see end-notes), or RePec for economics (see end-notes). These subject-based repositories are accessible through the web, some holding hundreds of thousands of items and many being widely used.

Making research material available through such repositories has a number of benefits for the researcher. Typically, when academics publish their research, they wish their material to be disseminated as widely as possible, and as quickly as possible, to achieve the highest impact. The traditional model of circulating information through journals does not do this. Commercial publishers currently follow a business model based on *restricting* access to information, through subscriptions. This restriction applies to both print-based and e-journals. Print-based journals also have restrictions on space (how many papers can be included per issue) and on time (how many times issues are published per year).

As well as these impact barriers, there are also access barriers. When academics are searching for information, they want easy and convenient access to all of the material that they need. Again, the current model is restricted by limited library budgets and journal costs.

Disseminating and accessing information through on-line open-access repositories drastically reduces these impact and access barriers. There is evidence to suggest that material available in this way is accessed more and referenced more. In addition, such repository systems can offer added-value services - such as recording hit-counts on papers, producing personalised publication lists, or citation analyses.

Additional benefits accrue through using a software protocol called the Open Archives Initiative Protocol for MetaData Harvesting (OAI-PMH). This is a method of ensuring interoperability between repositories. It allows a third party "service provider" to gather content information - to "harvest" uniform metadata - from a large number of repositories. This means that once a repository has been registered as OAI-PMH compliant, harvesting services will automatically search and harvest the associated metadata. Some service providers may bring the harvested metadata together into a single searchable database. A search for information through such a service provider will then encompass all such OAI-PMH compliant repositories and provide results leading through to the material held at individual repositories.

Establishing an e-print repository therefore brings a number of advantages to the individual academic. Impact barriers on their published work are reduced and access barriers on their own research activities are similarly lessened.

Repository use

The use of pre-prints varies between subject disciplines and must be reflected in the repository. Some disciplines have a well-established culture of circulating pre-prints for comment or to assert priority. In these disciplines, academics may wish to mount a pre-print on the repository before going on to the peer-review process for publication. Other disciplines only publicise peer-reviewed material and academics would therefore have their paper refereed through the normal peer review process to a form acceptable for traditional publication before mounting it (a post-print). In either case, the status - pre-print/post-print - is made clear on the repository to maintain clear quality control.

The authors can choose to mount the material themselves, or the institutional library or other agency may offer this as a service. In either case, the process is quick and simple. The paper itself then can be sent to publishers for traditional publication in the normal way. At first sight, use of a repository might be seen to make a traditional journal redundant. However, this has proved not to be the case. The arXiv repository, for example, has been extensively used for both pre-prints and post-prints for the past twelve years and yet physics journals are still published and sold.

Once the e-print is mounted, then the metadata will be harvested by OAI-compliant service providers and accessible through a variety of routes. This could be through the institutional repository itself, or through cross-searching service providers such as OAI-ster, or Arc. Some service providers facilitate automatic searching through the metadata to allow normal search engines, such as Google, to find results (links to these services are given below).

Actually installing an OAI-PMH repository is fairly cheap and straightforward. A standard server, costing under £1,500 will support the service. There is a range of free software available that can be used to run the service: SHERPA is using software from eprints.org (link below) and there are several other sources. The factor that has restricted the wider use of repositories so far is not technical, but cultural. It requires a change in individual working practices and support during the process of change. SHERPA is based on the idea that institutions can provide both the technical facilities and support the process of change.

Subject vs. Institutionally based

Subject-based repositories exist, but cover only a small minority of subjects. In addressing institutionally-based e-print repositories (IBERs), SHERPA is looking at ways of expanding the benefits of repository use for the wider academic community. An institutional basis gives a number of advantages in establishing and maintaining repository use. Institutions have the resources to subsidise the start-up of a repository, and the infrastructure and personnel to develop and maintain the repository. Institutions can kick-start an e-print collection through academic support in mounting material, advocating IBER use, and giving advice on IPR and copyright issues.

It is envisaged that there will be benefits for the institution as a whole. IBERs can play a strong role in identifying and managing institutional information assets. Institutions have a direct interest in efficiently disseminating and publicising the research material of their staff. A research-led university may be producing ten published papers per working day, but the productivity of many institutions in terms of both volume and quantity is often not

recognised. IBERs will provide a continually updated record of research material and could help to develop that recognition and generate an institutional sense of identity in intellectual output.

When it comes to an academic searching for e-prints, if material is mounted using the OAI-PMH, then the actual location of the e-print is identifiable but irrelevant. Importantly, there is no practical conflict between subject-based repositories and institutionally-based repositories: material can exist on both if desired. In the same way, there is no conflict with the idea of Open Access journals. Once the concept of research material being made freely and openly available is accepted, then multiple copies are inevitable and beneficial. Where material is simply a duplicate of traditionally published material the authoritative version can still be seen as the journal version. If, in the future, material comes to exist only as an e-print, then obviously version control would become important and this is again somewhere that institutions could take the lead. If the definitive version is stored on an IBER with institutional guarantees of availability and long-term preservation, this gives the institution a clear role in research provision and authority and gives academics the assurance that their work will be available in the future.

Key Issues

There are four key issues that need to be considered by anyone considering building a repository: collections policy; preservation policy; IPR and copyright, and the process and effects of cultural change.

A repository has to have a clear collections policy that defines what is acceptable in the repository: decisions include the type of material that is acceptable - pre-print, post-print, conference paper, etc. - and the format of that material - as a .pdf, a proprietary word processor format, HTML, etc. Decisions also have to be made as to who can mount what type of material. For example, can post-graduate students mount papers? The user of the repository has to be clear in knowing what material is being held, what status it has, and from whom it has come.

Clearly, with a significant body of e-prints, these repositories will be of immense value for the foreseeable future. Traditionally, preservation of published material has been the responsibility of libraries. There is a compelling argument that says that preservation can be set to one side for now, so as not to impede the collection of material to populate a repository. However, if institutional repositories adopt this approach, then there is an implicit commitment to an unknown amount of work at some point in the future. This might sound uncertain and vague - and that is the point - it is. So many IT-based projects have incurred unforeseen costs and discovered unintended commitments that there is now, quite rightly, a general reluctance to fund institutional scale projects without a clear and well-thought analysis of future implications. This is true for more than financial issues: intellectual property rights in materials mean that a number of permissions and licences have to be quite clear for an institution to be able to maintain a repository in the long term. Certainly it would be a mistake to hold back the foundation and population of repositories until a watertight policy and a foolproof preservation strategy is in place. For one thing, the important issues and trickier questions will only emerge through the population and use of IBERs. To try to answer some of the concerns that have been raised in the field, a significant part of SHERPA's work is in looking at this issue.

The SHERPA repositories are being populated on a pilot basis, knowing that there might have to be an amount of "retro-fitting" of preservation standards, metadata and agreements: this is, after all, a research and development project. It is intended that, in doing the work of population and investigation into preservation issues, guidelines will be developed to enable other repositories to start with preservation standards in place.

IPR issues also need to be addressed. Most universities are now adopting a formal IPR policy to define ownership of copyright in academic materials. To store and allow access to material through an IBER implies clear and agreed rights for that material. Traditionally, institutions have waived any claim of institutional copyright in favour of individual authors, who have then transferred those rights to publishers. An increasing number of publishers are modifying their copyright licenses to specifically encourage mounting a copy of a paper on a web-server. There has to be agreement between all of the stakeholders in the licensing and retention of copyright and IPR, and, again, this is an area of SHERPA work.

Perhaps the single biggest issue is that of cultural change. HE has undergone so many upheavals in the last twenty years that further change is often resisted just because it is change. The traditional journal system has been around for so long and is so strongly entrenched within HE life that even with its restrictions, it is sometimes difficult to see that there can be a different approach. Journals have come to represent a quality mark through their individual brands, so that an academic's individual standing can sometimes be judged against association with brand names rather than against cumulative papers and their results in the peer-review process. Encouraging the take-up of even a supplementary and beneficial method for dissemination is not easy. SHERPA is also looking at the process of advocacy, whereby academics, departments, or research groups can be introduced to, and supported in the use of IBERs.

Hybrid approaches

As mentioned above, IBERs are not seen as an attempt to bypass journals or to make them redundant. The idea of journal publication and use is embedded within research, offering more than just a dissemination route. The peer review process is one example of this. Such peer review is necessary for quality control of e-prints and, currently, review bodies are organised through the journal submission process. Work so far is indicating that hybrid approaches are not only possible, but necessary. One of the lessons from the introduction of IT into very many areas of life is that new IT practices do not replace the old order. They may certainly change it, often radically, but more often act as an alternative or supplement to older methodologies. Mounting e-prints into IBERs should be seen as a supplementary form of research communication. Many far-sighted publishers have already reached this conclusion and see the process not as a threat, but as an opportunity. The project is talking to publishers and web-based service providers to build hybrid models for scholarly communication for the future.

Summary

The partners of SHERPA are building a series of inter-operable IBERs using the OAI-PMH to look at the practical issues involved. Investigations into IPR, copyright, collections and preservation policies are being made to see what can be shared as common guidance, and what sections need localisation for local needs. The work that the partners

are doing in starting the network of UK HE repositories will help other institutions in turn. SHERPA will be working to see that the results of project work are available and disseminated among the whole HE community. SHERPA itself is part of the JISC FAIR programme, which is looking into the provision, management and sharing of institutional resources. The project is also part of a larger world-wide interest into the use of Open Access materials and the development of Open Archives.

The building and use of IBERs offer real benefits for the researcher and for the institution at the fundamental level of research dissemination and scholarly communication. The initial installation of a repository is fairly straightforward and technical demands should not be seen as an issue. The issues which are essential to address relate to a well-founded collections policy; a clear position on preservation; agreement and understanding on IPR and copyright issues between researchers and institutions; and an appreciation from all stakeholders of the cultural change that IBER use will bring.

Links

SHERPA project: - http://www.sherpa.ac.uk
FAIR Programme - http://www.jisc.ac.uk/index.cfm?name=programme_fair
The University of Nottingham's pilot repository - http://eprints.nottingham.ac.uk/
The physics based arXiv - http:/www.arxiv.org
Economics papers at rePec - http://www.repec.org
Open Archives Initiative - http://www.openarchives.org/
ePrints.org - http://www.eprints.org
Service provider OAI-ster - http://www.oaister.org
Service provider Arc - http://arc.cs.odu.edu
Search engine Google - http://www.google.com

Bill Hubbard is the Project Manager of SHERPA, a project funded by JISC and CURL that supports the installation and population of institutional repositories in a number of Higher Education institutions. As part of project work, the team is looking at issues of preservation, IPR, metadata, culture change and the integration of repositories into larger strategic information plans.

Bill can be reached at:

Bill Hubbard
SHERPA Project Manager
IS Divisional Office,
Hallward Library,
University of Nottingham,
University Park,
Nottingham
NG7 2RD

tel: (0115) 846 7657
fax: (0115) 951 4558
email: bill.hubbard@nottingham.ac.uk