

Your Place Or Mine? Showcasing eResearch via an Institutional Repository

Chris Taylor, Executive Manager, Information Access Service, University of Queensland Library

Belinda Weaver, Co-ordinator, ePrints@UQ and eScholarship@UQ, University of Queensland Library

Australian research is increasingly appearing in digital form, but much of it is hard to access. Journal articles may be accessible only to subscribers. No clear pathway exists to the research publications posted by academics on personal Web sites, nor to much of the digital "grey" literature, such as working papers and technical reports. Even where it is systematically reported, it is not made available in a national, searchable dataset. At the institutional level, it usually metadata only that is made available, not the content.

The above realities limit the research impact of much of academia's output. The development of institutional repositories is one way to collect, organise and make searchable, research that might otherwise be virtually "lost". The University of Queensland Library established its repository, ePrints@UQ, in 2002. The centralised repository is a "destination" service, providing not only metadata, but also the full text of preprints, journal articles and conference papers.

ePrints@UQ is part of a wider project called eScholarship@UQ, a testbed program for the Australian Partnership for Sustainable Repositories whose aim is to examine and provide guidance on the critical issues impacting on the sustainability of local digital collections.

Hidden University Research Output

University research output is increasingly appearing in digital form, but that doesn't necessarily guarantee that it is easier to find. According to the UK Wellcome Trust, an independent health research charity, "Access to research is restricted and the means to gain access are determined by a market in which a small number of publishers have a dominant position."¹ We probably all know who they are – Elsevier, Springer/Kluwer, Taylor and Francis, Blackwell, Wiley and Sage. Anecdotal evidence indicates that many academics are largely untroubled by these restrictions, believing that the people they want to reach, their academic and scientific peers, will be able to get hold of their published research somehow. In some cases, the researcher may either distribute a copy of the article to his or her community of interest, or make it available via a personal or academic department website. In other cases, the researcher assumes that his or her colleagues will have access through a subscription to the database or electronic journal by their university's library.

Don't be fooled by initiatives such as Google Scholar², which describes itself as enabling one "... to search specifically for scholarly literature, including peer-reviewed papers, theses, books, preprints, abstracts and technical reports from all broad areas of research." It is a good thing that search engines such as Google are increasingly able to spider such resources. However, the reality of what you actually get is less impressive. Firstly, Google has only been able to scratch the surface in its negotiations with publishers. An example will illustrate the point. A political scientist at UQ performed a spot check in November 2004:

"I checked out a theorist I use a lot, and who I'd checked out in terms of times cited via the websites via the library (Web of Science I think). The latter netted me 13 citations of the article I use most, while Google scholar only netted me three. Long live the Library!"³

No doubt the coverage will increase as publishers realise that it is a cheap way to market their product and make more money. This leads to the second reality. Unless your library has a subscription to the database or electronic journal, you will still be denied access to much of the material

Whatever the preferred access method, research output currently reaches a limited audience. There are no standards for how this kind of casual posting is done. Much of this material has very abbreviated citation information. Nor is posting uniform across all disciplines. Search engines may or may not pick up such resources on local websites. It's very much a hit or miss scenario.

Of course, academics in large, well-off institutions are generally protected from want by the buying power of those institutions. If their library does not subscribe to a journal they want, the article can generally be obtained via inter-library loan. Outside of these wealthy institutions, the picture is very different. According to the World Summit on the Information Society, the breaking point (for research access) has already been reached in "all third-world Universities, in almost all European Universities and in many second-tier US Universities".⁴

Perhaps just as importantly, it ignores the increasing demand for the general community to have access to taxpayer funded research output. The Science and Technology Committee of the United Kingdom House of Commons produced a report in 2004, *Scientific Publications: free for all?* which states:

"It is not for either publishers or academics to decide who should, and who should not, be allowed to read scientific journal articles. We are encouraged by the growing interest in research findings shown by the public. It is in society's interest that public understanding of science should increase. Increased public access to research findings should be encouraged by publishers, academics and Government alike."⁵

Equally lost to sight is much of the digital “grey” literature emanating from academic departments, schools and research centres, such as seminar presentations, working papers and technical reports. These may be posted on websites, but may lack essential metadata that would boost their ranking in a search engine, if indeed a search engine ever finds such material at all. Other forms of research output, such as X-rays, laboratory results, pathology slides, reports of experiments, research datasets, and software are also under-reported at all levels, not least at the institutional level.

Even where research is systematically gathered by an institution, for example, as part of reporting statistical data to the Commonwealth Department of Education, Science and Training (DEST)⁶, this massive output is not made available in a national, searchable dataset. At the institutional level, some make their metadata available via a database, but most do not. Even in the former case, the actual content is rarely available.

The above realities limit and blunt the research impact of much of academia's output. It may also result in wasteful duplication of effort.

Open Access Movement

The development of the open access publishing movement in the UK, the US and now worldwide is a response to just these problems of hard-to-access research. The movement developed in support of the principle that the published output of scientific research should be available, without charge, to everyone.

The open access movement consists of two planks:

- Open access journals. Open access journals look just like the existing commercial journals, except that everyone may view the content. Peer review still takes place. Projects such as BioMed Central⁷ have proved that the model can both work and make money. As at November 2004, the *Directory of Open Access Journals*⁸ had 1,366 titles listed.
- Open access archives or repositories. Repositories allow researchers to self-archive their research output in an organised manner, thereby improving access and safeguarding its long term availability. Repositories may be developed at the institutional level or be discipline-specific. An example of the former is ePrints@UQ⁹, showcasing research output at the University of Queensland.

For a comprehensive overview of the open access movement, see the report produced by the Electronic Publishing Innovation Centre (EPIC) and Key Perspectives Ltd.¹⁰

Open Access Repository Models

There are four basic models for open access repositories:

- **Institutional or sub-institutional** – the content and associated metadata are created and stored on a central server. An example is ePrints@UQ.
- **Subject-based** – the content and associated metadata is contributed by a decentralised community to a central server. An example is Cogprints¹¹, covering cognitive science.
- **Multi-institutional** – the content and associated metadata is created locally by member institutions and uploaded to central archive. An example is the California Digital Library’s eScholarship¹².
- **Harvested or gathered** – content and associated metadata is created and stored on local servers, with the content and/or the metadata harvested or gathered into a central server. An example is OAIster¹³ which harvests more than 3.7 million records from 363 institutions (as at November 2004).

Combinations or hybrids of the above models are also being developed, for example eScholarship@UQ¹⁴ is deploying a combination of institutional and harvester-based models (see below).

While IT should never be the main driver in development of a new service, the choice of infrastructure does make a difference. The vast majority of institutional repositories are underpinned by open-source software packages. Preference for open-source software is probably due to the “frontier” mentality of many of the early developers of open access repositories. As part of the Australian Partnership for Sustainable Repositories project (APSR)¹⁵, of which the University of Queensland Library is a member, a comprehensive summary of open-source software packages that support open access repositories has been compiled. As of November 2004, there were 22 packages¹⁶.

Commercial players are now starting to enter the marketplace, suggesting that institutional repositories are more than a passing phase. However, most of the commercial players have not built their own products from scratch. Instead, they have partnered with open-source developers to provide the basic product, leaving them to concentrate on value-added services, such as hosting and more sophisticated search interfaces. Such products include:

Open-source product	Open-source developer	Commercial product	Commercial developer
Keystone Digital Library Suite ¹⁷	Danish Electronic Research Library ¹⁸	Keystone Digital Library Suite	Index Data
Fedora ¹⁹	University of Virginia and Cornell University	iVital ²⁰	VTLS
D-Space ²¹	MIT	Open Repository ²²	BioMed Central
bepress ²³	University of California, Berkeley	Digital Commons @ ²⁴	ProQuest

It is interesting that the commercial players are coming from widely differing sectors, for example, a software consultancy company, an integrated library

system vendor, a publisher of open access journals and a publisher of commercial databases.

Regardless of the model chosen, institutional repositories generally conform to the standards developed by the Open Archives Initiative. The metadata contained in such repositories is 'open', and can be harvested by other services via the Open Archives Initiative Protocol for Metadata Harvesting (OAI-PMH)²⁵. Harvesters may be sent out by other repositories, possibly in a similar discipline; by the cross-archive search tool such as OAIster; or they may be crawlers from search engines such as Google and Yahoo, which regularly index this type of research material. For example, the Googlebot harvests from UQ's ePrints@UQ repository daily.

The main advantage of OAI-PMH is that it is relatively lightweight, i.e. compliance is relatively easy to achieve. The downside is that it lacks sophistication, particularly in specifying metadata standards. All you need is Dublin Core²⁶ metadata in order to play.

The remainder of this paper examines the experiences at the University of Queensland in developing an institutional repository within the context of the worldwide movement to improve access to research.

ePrints@UQ

The University of Queensland Library established its repository, ePrints@UQ, in 2002. It showcases the research work of academic staff and postgraduate students at UQ, and is part of a wider Group of Eight²⁷ initiative to make Australian scholarship more visible and accessible. To date, three other Go8 universities, apart from UQ, have established repositories – Australian National University, Monash University, and the University of Melbourne. Repositories have also been established at Queensland University of Technology and Curtin University. In New Zealand, while there are currently no true institutional repositories, the University of Waikato has developed a demonstrator site for Internet archives²⁸. The aim of these institutional repositories is to provide a 'destination' service: they provide not only metadata but also the full text of a wide range of resources, including journal articles, conference papers, book chapters, preprints, theses, working papers and technical reports.

Getting such repositories up and running is no easy task. While the cost of setting one up can generally be absorbed into existing infrastructure, they are not cost-free. They need people to go out and sell the idea of such a repository to academics, to develop use and training materials for it, and they need technical staff to establish and maintain the service. Hard decisions must be made as to the purpose of such a service. Will it be full text or metadata only? Will it seek to get 'everything'? If not, what will be excluded? And what decisions need to be made about preservation?

Copyright is an issue. While many academics would be happy to see all their materials in an ePrint repository, the reality is that many will not be able to. Too often we find that academics have signed away their copyright and the publisher is not about to give it back.

Two years on, ePrints@UQ is experiencing limited success. The service has more than 1,160 papers deposited, with many more waiting to be loaded. However, this represents but a fraction of UQ's annual research output. While the service has been valuable in making a range of previously undiscovered material available (download statistics show this material is very heavily used), it does not encompass the full range of UQ's research output.

Part of the problem lies in what has up until now been considered "acceptable" content. In librarian's speak, it equals ePrints@UQ's collection development policy. Content has been largely restricted to so-called "document-like" objects such as books, journals and conference papers. As mentioned above, research output from an institution is more than just words on a page. Such things as the experiments that were run, the surveys that were conducted, the spreadsheets and datasets created and the images captured in a laboratory, are genuine research outputs as well. Hitherto, they have gone unreported, and would have continued to be at the University of Queensland despite the creation of ePrints@UQ.

eScholarship@UQ

Enter eScholarship@UQ, a project which seeks to grapple with some of these issues. eScholarship@UQ is a testbed program for the Australian Partnerships for Sustainable Repositories, a three-year, DEST-funded project, which aims to examine and provide guidance on the critical issues impacting on the sustainability of local digital collections. UQ is partnering in this project with ANU, the University of Sydney, the National Library of Australia and the Australian Partnership for Advanced Computing. Each institution contributes different strands to the project, though all three testbeds focus on sustainability.

For our part, eScholarship@UQ seeks in the first instance to identify all UQ research output and then to test the feasibility of capturing its associated metadata with a view to making it available through a central portal that can be searched or browsed by discipline. Of course, this is easier said than done. We therefore decided to start with what we already had control over within the Library, namely ePrints@UQ and the local version of the Australian Digital Theses Program (ADT)²⁹. With ePrints@UQ being OAI-PMH compliant, the process was relatively straightforward.

The same could not be said for extracting metadata from ADT. While ADT is not OAI-PMH compliant, extracting the metadata was not really the problem, as the data was already being exposed for gathering by the central ADT portal at the University of New South Wales. The real problem rested with the quality of the metadata. While ADT uses Dublin Core as its metadata

encoding schema, it says very little about the standards for metadata itself, apart from mandating selection from a controlled subject vocabulary, the Australian Bureau of Statistics (ABS) Australian Standard Research Classification³⁰. Some effort was therefore required to normalise the data by programmatic means.

Shaken, but not stirred, our first venture into “foreign” territory came with successful negotiations with the University’s Office of Research and Postgraduate Studies to capture the research data annually reported to DEST. The metadata is stored in an Oracle database, with further negotiations required with the IT managers to give us rights to gather the required data. While the data is highly structured, we anticipate that some normalisation will be required.

ePrints@UQ, ADT and the UQ research quantum are just the beginning. The project will also pull in citation and other material from the range of Web-based publication lists, personal Web sites, departmental publication pages and other means by which research data is locally housed and reported. Knowing that we have done the easy stuff and virtually every other repository of research information at UQ will offer challenges, both in term of extracting and normalising the metadata, leaves us facing the future with some trepidation. Our solution is to tackle one source of data at a time.

The end result will be a metadata gateway that will identify what research has been done at UQ. eScholarship@UQ does not aim to be a destination service for content. However where content is available locally, eScholarship@UQ will point to where that content is stored, so that users can retrieve it if they so wish.

In any case, making the previously under-reported citation and other information available will greatly improve the visibility of much of this work, and will do so in an open access, harvestable environment.

Why build eScholarship@UQ when we already have ePrints@UQ?

The answer to the above question is summarised in the title of this paper – “Your place or mine?” Our experience with ePrints@UQ is that while some schools find it easier to allow ePrints@UQ to host and manage content they had formerly housed locally, other schools prefer to continue managing their own content. In the former case, the web page from which the data had originally been linked was altered to point people towards the data’s new home in the ePrints@UQ repository. The benefits for the school are threefold - a more secure environment for the data; greater visibility of the work through harvesting and search engine crawling; and decreased Web storage and management costs.

In the latter case, because ePrints@UQ mandates the housing of content, there is no opportunity for ePrints@UQ to become involved. The result is valuable research remaining virtually hidden.

eScholarship@UQ offers the best of both worlds. Schools now have a choice. They can transfer the management of content and access to ePrints@UQ, or continue managing their own material, but allow eScholarship@UQ to harvest their metadata. In both cases, the schools benefit from the greater visibility that indexing in eScholarship@UQ will confer, and from the centralisation of UQ research indexing generally.

Proof of the flexibility of this arrangement came towards the end of 2004. In 2005, the University of Queensland is piloting a formal assessment of research quality, to be run as part of the normal school and centre reviews. The two schools to be involved in the pilot are the School of Human Movement Studies and the School of Physical Sciences. The assessment will be conducted by a panel of internationally recognised experts. Each eligible member of staff will be required to submit three pieces of work from the last five years.

The Library was approached to provide a gateway for the international panel to remotely access the papers submitted for assessment. Our solution was to offer two gateways, ePrints@UQ and eScholarship@UQ and let the individual school decide which one best suits their needs. Because the School of Human Movement Studies does not have a history of managing its own content, they are very happy for the Library to do so via ePrints@UQ. The papers submitted for assessment is just the beginning, as the School can see the obvious benefits from adding all their papers to ePrints@UQ.

Nor is the choice of “your place or mine” cast in concrete. We expect that over time, schools that were once nervous about giving up management of their research files, will become more comfortable with the advantages of getting ePrints@UQ to do the job.

Thinking Globally and Acting Locally

While we are acting locally with the development of eScholarship@UQ, we are also thinking globally. In developing strategies to showcase institutional research, one cannot simply rely on users to come to your place. The “mountain must also go to Muhammad”. One avenue is to expose your metadata for spidering by search engines, such as Google and Yahoo. For example, the Googlebot harvests from ePrints@UQ repository daily. Another avenue is harvesting by federated repositories, such as OAIster. In both cases, compliance with OAI-PMH streamlines the whole process.

Benefiting From Our Experience

The development of institutional repositories is still very much in its infancy. A sustainable pathway forward is not yet clearly marked. It is therefore vital that the IR community communicates its collective experiences, good and bad, for the benefit of all. That is precisely the modus operandi of the APSR project, particularly in the area of sustainability.

As previously mentioned, eScholarship@UQ has made available the most comprehensive listing and summary of open-source IR packages. We will also be reporting on our experiences testing a number of these packages, or components thereof. We already have extensive experience with GNU Eprints³¹ developed by the University of Southampton, which underpins ePrints@UQ. There has been a lot of publicity and take-up of DSpace. Our APSR partner at ANU migrated from GNU Eprints to DSpace. Others believe that Fedora, with its sophisticated content management based on the Reference Model for an Open Archival Information System (OAIS)³², will become the IR package of choice.

Our preliminary conclusion is that currently, there no one solution that can meet the needs of our project. For example, very few of the 22 packages currently available (and none of the three packages mentioned above) support the harvesting of metadata. Accordingly, project staff are working to find the best components that can be put together to create the best solution. Whatever solution will ultimately come up with, we plan to make the package available to other organisations wanting to set up similar services.

Conclusion

University research is a valuable commodity that is currently under-exposed. Concerted effort is required at the institutional, national and international level, to address the issue. The open access movement and, in particular, institutional repositories, offers a tangible method to showcase university research. As a fledgling industry, the institutional repository movement needs to share its knowledgebase at every opportunity. The Australian Partnership for Sustainable Repositories is providing leadership on pathways for sustainable IR development. The “your place or mine” model being piloted by the University of Queensland Library is but one of the possible pathways.

References

¹ Wellcome Trust, 2004, *Costs and business models in scientific research publishing*: http://www.wellcome.ac.uk/en/images/costs_business_7955.pdf.

² Google Scholar, < <http://scholar.google.com/>>.

³ Private email, 2004.

⁴ World Summit on the Information Society, < <http://www.wsis-si.org/oa-facts.html>>.

⁵ [House of Commons, Science and Technology Committee, Tenth Report, 2004, *Scientific publications: free for all?*](#) <<http://www.publications.parliament.uk/pa/cm200304/cmselect/cmsctech/399/39902.htm>>.

⁶ Commonwealth Department of Education, Science and Training: <<http://www.dest.gov.au/highered/research/herdc.htm>>.

⁷ BioMed Central: <<http://www.biomedcentral.com/>>.

⁸ Directory of Open Access Journals: <<http://www.doaj.org/>>.

⁹ ePrints@UQ: <<http://eprint.uq.edu.au/>>.

-
- ¹⁰ Electronic Publishing Innovation Centre and Key Perspectives Ltd, *Delivery, Management and Access Model for E-prints and Open Access Journals within Further and Higher Education*, 2004:
<http://www.jisc.ac.uk/journals_work.html>.
- ¹¹ Cogprints: <<http://cogprints.ecs.soton.ac.uk/>>.
- ¹² California Digital Library, eScholarship Repository:
<<http://repositories.cdlib.org/escholarship/>>.
- ¹³ OAster: <<http://oaister.umdl.umich.edu/>>.
- ¹⁴ eScholarship@UQ: <<http://www.library.uq.edu.au/escholarship/>>.
- ¹⁵ Australian Partnership for Sustainable Repositories:
<<http://www.apsr.edu.au/>>.
- ¹⁶ eScholarship@UQ, *Software options*:
<<http://www.library.uq.edu.au/escholarship/software.pdf>>.
- ¹⁷ Index Data, *Keystone Digital Library Suite*: <<http://www.indexdata.dk/>>.
- ¹⁸ Danish Electronic Research Library: <http://deff.dk/index_eng.html>.
- ¹⁹ University of Virginia and Cornell University, *Fedora*:
<<http://www.fedora.info/>>.
- ²⁰ VTLS, *iVital*: <<http://www.vtls.com/>>.
- ²¹ MIT, *D-Space*: <<http://www.dspace.org/>>.
- ²² BioMed Central, Open Repository: <
<http://www.biomedcentral.com/info/libraries/product-page?product=opr>>.
- ²³ University of California, Berkeley, *bepress*: <<http://www.bepress.com/>>.
- ²⁴ ProQuest, *Digital Commons @*: <
<http://www.il.proquest.com/umi/digitalcommons/>>.
- ²⁵ Open Archives Initiative, *Open Archives Initiative Protocol for Metadata Harvesting*: <<http://www.openarchives.org/>>.
- ²⁶ Dublin Core Metadata Initiative: <<http://dublincore.org/>>.
- ²⁷ Group of Eight: <<http://www.go8.edu.au/>>.
- ²⁸ University of Waikato, *New Zealand Digital Library Project*:
<<http://www.nzdl.org/>>.
- ²⁹ Australian Digital Theses Program: <<http://adt.caul.edu.au/>>.
- ³⁰ Australian Bureau of Statistics, *Australian Standard Research Classification*:
<<http://www.abs.gov.au/ausstats/abs@.nsf/0/51C2BDB99EBA43E8CA256889001E9D9E?Open&Highlight=0,rfcd>>.
- ³¹ University of Southampton, *GNU Eprints*: <<http://software.eprints.org/>>.
- ³² Open Archival Information System:
<<http://ssdoo.gsfc.nasa.gov/nost/isoas/>>.