Institutional Repositories Supporting Cyberscholarship: Building Content

Joanna Richardson [HREF1], Projects Manager, Digital Repositories Team, Division of Information Services, Griffith University [HREF2], Queensland, 4111, Australia. j.richardson@griffith.edu.au

Abstract

Recent reports point to an increased focus on evolving forms of research and scholarship, known as "cyberscholarship". Building content in institutional repositories is integral to supporting the future of scholarly communications and thereby supporting cyberscholarship. However the literature consistently points to the basic failure to date to embed the institutional repository in the intellectual life of the scholar / researcher. The author examines several initiatives in Australia designed to address some of the well documented obstacles to engagement. This paper concludes that one of the keys to changing the current culture lies in changing the mind set -- not by concentrating solely on the advantages of open access but rather by proactively facilitating the process for participation and then demonstrating the resultant benefit(s).

1. Introduction

In 2004 Malcolm Gillies described the advent of the Web and subsequent electronic and digital innovations as "every bit as significant as Herodotus or Gutenberg" [HREF3]. It is not surprising, therefore, that much attention is currently being focused on the new opportunities and strategies for managing the information created --and used-- by researchers and scholars. Given the widespread availability of digital content, new forms of research and scholarship are evolving. This is often referred to as "cyberscholarship" [HREF4]. New forms of research inherently drive new infrastructure models since the content being captured and subsequently managed is used in ways which are markedly different from traditional approaches.

One of the interesting trends is the importance of "data-driven" science. In other words, computers analyse vast amounts of information that could never be done manually. As a result tools have been developed that seek patterns in databases, indexes and other compiled resources. Arms [HREF4] says:

In the cyber age, collections of digital content and the software to interpret them have become the foundation for discovery; they have entered the realm of infrastructure. When content becomes infrastructure, there is value in investment to support it. The preservation and organization of information for new forms of scholarship enable others to discover unexpected and novel associations without having to replicate the primary data.

Lesk [HREF5] echoes this sentiment, adding that this is leading to the creation of new interfaces for access and retrieval.
In some disciplines, digital content is the norm, thereby accelerating the need to build the infrastructure to support it both now and in the future. Some of the literature discusses the requirement for new superdata centres with substantial computing power to support the analyses inherent in data-driven science. Strong collaboration --both nationally and internationally-- is required to develop the required cyberinfrastructure. Inevitably some collaborators will be from the university sector and yet universities compete with each other for research funding, resulting in some interesting "tensions". At a seminar delivered in February 2008 in Australia, Ralph Schroeder from the Oxford Internet Institute talked about new e-research technology tools driving research and interdisciplinary collaboration. He gave examples of how research can progress more rapidly because supporting tools can move across disciplines, e.g. social sciences and health sciences in Sweden.

As has been noted widely in the literature, it is important to not just replicate traditional methods currently used for handling physical media in the digital arena. For example, in future text / data needs to be in formats that support machine processing, e.g. XML rather than PDF. It has been suggested that a new kind of data journal is needed which offers peer review of data sets, for example. Concurrently with the far-reaching discussion about new approaches, there are sobering reality checks. For example, Peter Murray-Rust is well-known for his research in molecular informatics, which brings tools from computer science to chemistry, biosciences and earth sciences, in order to manage information. Despite the advances in systems created to read the current chemical literature, he concludes:

Our thesis is that the current scientific literature, were it to be presented in semantically accessible form, contains huge amounts of undiscovered science. However the apathy of the academic, scientific and information communities coupled with the indifference or even active hostility and greed of many publishers renders literature-data-driven science still inaccessible.

2. The Role of Institutional Repositories

Crane has discussed the role of institutional repositories in supporting the future of scholarly communications. This includes offering long-term access to digital objects with persistent value, providing reliable open access to research outputs, and exporting the institution's scholarship. Institutional repositories remain an important focal point, which, according to Crane, will never approach their full potential without institutional cyberscholarship. Categorising the repository movement as having failed to achieve any significant impact, he advocates a system of interoperable institutional repositories for scholarly production that are actively used by scholars.

To their credit some institutional repository services are endeavouring to achieve greater impact by customising their metadata to optimise their Google performance. For example, the DRIVER (Digital Repositories Infrastructure Vision for European Research) Project reported in 2007 that "Soton has analysed Google rankings in order to determine whether it can influence the place of its material in future; Minho has done the same for Google Scholar and hopes to analyse more similar services to better target the visibility of its research. CERN has agreed on metadata specifications with Google Scholar whereas HAL plans to insert its Dublin Core metadata into the html
meta-field of each HAL publication page to ensure the better information retrieval by such services." [HREF8]

The actual nature of institutional repositories is currently under debate. Should repositories be all inclusive or specialised? Should the model be a few very large repositories or many repositories with small specialities? Indeed would there be just one model? In the final analysis, though, it may be quite premature to worry about attempting to resolve these questions now. The repository structures of today will not be the structures of tomorrow. They will evolve just as the research technology tools are evolving --although perhaps not at the same prodigious rate. Ultimately one of the drivers will be the ability to create the infrastructure required to enable research collaboration.

3. The Repository Environment in Australia

In a 2007 article Margaret Henty [HREF9] has described ten major challenges for set up and sustainability of institutional repository services in the higher education sector in Australia, including defining collections, roles and responsibilities, service quality and reporting requirements. University research increasingly involves the use, generation, manipulation, sharing and analysis of digital resources. This raises questions of the relationship between institutional repositories and eResearch and provides challenges to repository managers to "broaden their thinking" still further to help meet these needs. As she points out, new paradigms of ICT-enabled research have become mainstream in all disciplines, and the repository has emerged as a key piece of eResearch infrastructure in providing enduring access to revolutionary new collections of research data.

In Australia --as elsewhere in the world-- there is an underlying tension / dichotomy between the tendency of scholars to sign away all their rights when an article --or other content format-- is published, and the pressure to make research publicly available. It has been suggested in the international arena that not only rewards (in some disciplines) persist for writing in traditional formats but also some institutions are more bound up with the preservation of "cultural capital" than with academic performance as such. Juxtaposed to this culture is the 2006 report by the American Council of Learned Societies (ACLS) which recommends that all content be freely available under open access, even if no plan has been put forward for addressing the IP issues surrounding many formats [HREF10]. In 2007, a joint NSF/JISC [HREF4] report observes that projects which use public funds to generate data, etc. have a responsibility to make that information available to other researchers. Since the public funds so much of universities' research, the latter have a particular responsibility in this regard.

In Australia, several important research funding bodies, e.g. Australian Research Council (ARC) and the National Health and Medical Research Council (NHMRC), have clearly outlined in their respective funding rules for 2008 an expectation that "any publications arising from a research project [will be deposited] in an appropriate subject and/or institutional repository wherever such a repository is available to the researcher(s)" [HREF11]. In addition the new Labour government has just launched its Excellence in Research for Australia (ERA) initiative. The minister charged with
the portfolio, Senator Kim Carr, has delivered several speeches which pick up on the issues outlined above:

The Australian Government recognises the importance of encouraging collaboration between research organisations as a means of meeting national research challenges. This collaboration will be central to building an effective national research capability. (6 Feb 2008) ([HREF12]

We want the research conducted in universities and public research agencies to inspire and inform fresh thinking across the community. The more collaboration and interaction there is between researchers and the society around them, the better. It follows that research and research data should be widely disseminated and readily discoverable. The results of publicly funded research should be publicly available. (7 Feb 2008)[HREF13]

Clearly the Australian government wants to ensure that it is part of the worldwide move towards establishing frameworks which can optimise access to research, i.e. be a stakeholder in both cyberscholarship and cyberinfrastructure.

4. Building Content

4.1 Lack of Critical Mass

If the institutional repository is a cornerstone in scholarly / research production, then it needs to have more engagement from scholars and researchers. One of the fundamental challenges is to develop strategies to move creators from thinking about knowledge production as inherently linear to embracing more open-ended knowledge processes. Undoubtedly the opportunity for more research collaboration --especially of an interdisciplinary nature-- could act as a catalyst as researchers working with more "traditional" formats are exposed to new tools and new practices.

Green suggests that the future depends on changing the practices of multiple arrays of individuals [HREF14]. Yet consistently the literature points to the basic failure to date to embed the institutional repository in the intellectual life of the scholar / researcher. If self-archiving, i.e. relying on academics to either deposit their own works themselves or allocate the task to someone else such as a research assistant, serves as the basis for populating the repository, then this concept / workflow has failed to fulfil initial expectations. The author has detailed some of the reasons in another paper [HREF15]. Certainly the report from the DAEDALUS Project [HREF16] that "The majority of academic staff felt that they did not have the time to self-deposit, and were particularly unwilling to do this where they had already provided publication details to a departmental administrator, for example" is as valid today as it was in 2005 [HREF17]. The 2007 report by Davis [HREF18] on the underutilisation at Cornell University of its institutional repository covers a range of factors, including:

. . . redundancy with other modes of disseminating information, the learning curve, confusion with copyright, fear of plagiarism and having one's work scooped, associating one's work with inconsistent quality, and concerns about whether posting a manuscript constitutes "publishing".
In Australia—in an effort to address the issue raised above of double data entry—Woodland and Ng [HREF19] have reported on a project at Curtin University of Technology to design and implement an "integrator system" to share data between an institutional eprint repository and a University publications management system. While not intended to eliminate self-archiving, it builds on the fact that data entered by an academic into Curtin's PUB system can be used to automatically populate their espace@Curtin eprint repository. The major impetus for entering data into the University publications management system is that this data is utilised as part of the government's annual Higher Education Research Data Collection (HERDC) exercise. The latter collects research income and research publications data submitted by universities each year. Data collected is then used, along with data from the Higher Education Student Collection, for determining allocations to universities under various grants and awards schemes. As a model, the Curtin approach starts to address the issue of duplicate data entry.

An important Australian initiative is the Open Access to Knowledge (OAK) Law Project led by Professor Brian Fitzgerald, Queensland University of Technology, and funded by the former Department of Employment, Science and Training. It aims to ensure that "people can legally and efficiently share knowledge in an open access manner across domains and across the world" [HREF20]. As part of that project the "OAKList" was launched in February 2008. Interoperable with the UK's RoMEO/SHERPA database of publisher agreements, OAKList is a database of journal publishers which allows authors of journal articles to see, at a glance, publishers' agreements and their open access policies [HREF21]. As part of the project project staff are gradually introducing a methodology for building database content to heads of institutional repositories. This approach will leverage the efforts of all universities to contact publishers, especially Australian publishers.

### 4.2 Overview of Griffith Research Online

The author, as part of a team to implement an institutional repository at Griffith University, realised in 2004 that descriptive data being captured electronically in the PeopleSoft Research Administration Module for the annual HERDC exercise should logically be used to create the core metadata content for an eprints collection. As a result of a major project to support this integration, Griffith Research Online (GRO) -- the University’s institutional repository for published research material using DSpace -- was launched in late January 2007 [HREF22]. A number of technical innovations have been applied to populate the database, customise the integration and extend statistical reporting. These innovations were detailed by Martin Borchert at the Clever Collections Seminar in November 2007 [HREF23].

To quickly seed the database and provide authors with an incentive to upload full-text files, 10,200+ metadata records from previous HERDC surveys were migrated from the PeopleSoft Research Administration Module (known to academics as "My Research Publications") to GRO. Populating GRO with full-text files was identified as a challenge --as already discussed earlier in this paper-- because academics would not support two separate upload processes for HERDC and GRO. A more integrated approach was achieved by disabling the file upload function of DSpace and replacing it with an online workflow between the My Research Publications database (PeopleSoft) and GRO (DSpace).
For the 2006 HERDC survey and beyond, academics upload metadata and content files via the My Research Publications system. GRO staff manage copyright compliance and, where appropriate, enhance records with abstracts and links to publishers’ websites, prior to the records being authorised for public display in the institutional repository. In last year's 2006 HERDC survey process, academics flagged 17% of submitted publications to be uploaded to GRO; the latter now provides 13,700+ records and 1,200+ full-text documents. This new workflow --known internally as "continuous verification"-- is available throughout the year.

An enhanced statistical reporting system (in addition to AWStats) provides usage reports for brief views and downloads based on time period, connections from external (to Griffith) and internal IP ranges, and country of origin. The system counts only one download of each record and file, and the download of multiple files attached to a single record as a single download, per unique IP address, thus avoiding self-promotion.

4.3 Critical Success Factors

Between March and November of 2007, 32 information sessions were conducted throughout a broad range of organisational elements within the university in order to introduce the service and explain the benefits of participation, i.e. contributing full-text content files. This initiative has continued to form part of a multi-faceted marketing strategy which focuses not only on group presentations to organisational elements but also on personally contacting Griffith authors who have bibliographic publication records within GRO, encouraging them to provide their postprints, where permitted.

Another useful strategy has been to identify publishers who have recently revised their attitude toward open access and now permit authors to upload the publisher’s PDF. Many of these publishers are either university presses or associations. The GRO Team has been both surprised and heartened to discover the increasing number of publishers who not only allow but also actively encourage the uploading of the publisher’s PDF rather than the author’s postprint. The Team has then used that information to contact Griffith authors who publish in those journals and conferences, inviting them to authorise the upload of the publisher’s PDF. This has been particularly advantageous because many Griffith authors do not normally retain their author’s postprint version. In addition, the Team has changed the default request to a journal publisher from "author's postprint" to "publisher's PDF".

The corollary to this approach has been for the Team to "harvest" publications by Griffith authors from sources which allow the publisher PDF version in institutional repositories. The relevant Griffith author is then contacted by a Team staff member, requesting permission to upload the publisher's PDF on their behalf. An enormous boon for busy researchers and academics. The importance of this strategy has been reinforced by Jens Vigens, CERN institutional repository head: "Had we not aggregated material from outside we would have failed dramatically" [HREF8].

In the recently released DRIVER's guide to European repositories [HREF8], the authors have emphasised the importance for institutional repositories to be indexed by external search engines:
The primary purpose of digital repositories, however, is to provide a seamless database of worldwide content, searchable by all. In this context, the best marketing tool available for a repository is to ensure that it is indexed by Google/Google Scholar and other similar web services. We know that over 70% of researchers use these services to look for work-related information and that the majority of referrals to a repository are from external search engines. For driving usage of a repository, therefore, Google and its ilk cannot be bettered.

An important part of making Griffith University’s research available to the widest possible audience, therefore, was to ensure that Griffith Research Online was registered with the leading “harvesters” of data in open access repositories. Because the service is relatively new, it is difficult to determine --other than looking at download statistics-- exactly what use is being made of it. The fact that Google Scholar has been indexing Griffith Research Online since mid-2007 should enhance resource discovery. It has definitely proved a useful internal marketing tool to Griffith researchers. Standard preparation for a presentation about GRO to a given faculty, school, department or research centre always involves locating a record in GRO by a relevant author, which has a full-text content file attached. The publication title is then copied into Google Scholar as an "exact title" to demonstrate the link to the full-text in GRO. This seldom fails to achieve the desired result, i.e. capture the attention --and later participation-- of at least several members of the audience. In addition short flyers are left with attendees, summarising the benefits of open access and outlining the policies of major publishers relevant to the discipline coverage of that particular group.

In summary the critical success factors for Griffith Research Online have been:

- "harvest", i.e. find content published by Griffith authors and then ask authors for relevant files
- allocate one staff member to contact publishers' for permission as well as keep abreast of which publishers now allow publisher PDF version
- upload files on behalf of Griffith authors
- undertake all copyright checking
- provide easy access to a range of statistics relating to each publication by a given author
- demonstrate how searching in Google Scholar returns an entry in GRO
- utilise both marketing and support strategies which are tailored to meet the needs of different "cultures", e.g. disciplines

The important point is that an estimated 70% of content files uploaded to date have been a result of the GRO Team proactively searching for Griffith publications and then working with the respective authors to upload the correct version as opposed to relying solely upon Griffith authors to initiate the process. The end result has seen 1,200 full-text content files uploaded within the first 12 months --a reasonable effort for an institution with no previous exposure to "eprints" and "open access" as part of its corporate culture.

5. Conclusion
Building content in institutional repositories is integral to supporting the future of scholarly communications and thereby supporting cyberscholarship. While many scholars and researchers are actively engaging with exciting new forms of research and scholarship, an even larger number have yet to engage. The reasons have been well documented in the literature. This paper has suggested that one of the keys to changing the current culture lies in changing the mind set -- not by concentrating solely on the advantages of open access but rather by proactively facilitating the process for participation and then demonstrating the resultant benefit(s). In this way the institutional repository will become embedded in the intellectual life of the scholar and researcher.

**Hypertext References**

HREF1  
http://jprglobal.tripod.com/

HREF2  
http://www.griffith.edu.au

HREF3  

HREF4  

HREF5  
http://www.academiccommons.org/commons/essay/michael-lesk

HREF6  
http://www.sis.pitt.edu/~repwkshop/papers/murray.html

HREF7  
http://www.academiccommons.org/commons/review/gregory-crane

HREF8  
http://dare.uva.nl/document/93898

HREF9  
http://www.dlib.org/dlib/may07/henty/05henty.html

HREF10  
http://www.acls.org/

HREF11  

HREF12  

HREF13  

HREF14  
http://www.academiccommons.org/commons/essay/cyberinfrastructure-introduction

HREF15  

HREF16  
http://www.lib.gla.ac.uk/daedalus/index.html

HREF17  
https://dspace.gla.ac.uk/bitstream/1905/510/1/daedalusfinalreport.pdf
Copyright

Joanna Richardson, © 2008. The author assigns to Southern Cross University and other educational and non-profit institutions a non-exclusive licence to use this document for personal use and in courses of instruction provided that the article is used in full and this copyright statement is reproduced. The author also grants a non-exclusive licence to Southern Cross University to publish this document in full on the World Wide Web and on CD-ROM and in printed form with the conference papers and for the document to be published on mirrors on the World Wide Web.