Article for **Research Information**
Proposed Title: **Digital Repositories : building an infrastructure for the UK**
Authors: Julie Allinson, Digital Repositories Support Officer, UKOLN, University of Bath and Roddy MacLeod, Senior Subject Librarian, Heriot-Watt University
Status: Draft

Repositories are everywhere. But are they here to stay, or merely this season's buzzword?

JISC (the Joint Information Systems Committee) are demonstrating their commitment to repositories by funding directly into repository-related research and development since 2002. In the current JISC Information Environment repositories really are everywhere and come in many shapes and sizes[1]. Subject repositories such as Cogprints and the Chemistry Eprint Server in the UK; data archives and data centres such as those offered by AHDS, the NERC data centres and the UK Data Archive; Research Council-approved repositories such as PubMed Central; learning object repositories like Jorum; digital libraries; and more than 50 UK institutional eprint repositories all fall within the broad scope of these digital stores which manage and provide access to resources and metadata.

Of the 200 plus Higher Education Institutions in the UK, around 50 have Institutional and/or Department repositories according to the OpenDOAR directory[2]. And that figure is set to change as JISC embark on the £14m Repositories and Preservation strand of their Capital Programme, building on outcomes of its FAIR, X4L and Digital Repositories Programmes[3].

JISC's Digital Repositories Programme, ongoing, has taken forward this commitment to repository developments and it is that Programme which forms the focus for this article. Comprising 25 projects, plus a number of external studies and a package of cross-Programme Support activities, the overarching ethos of the Programme is to enable Institutions to make better use of repositories across research, teaching, information and administration. By fostering a greater understanding of the current repository landscape, the Programme seeks to guide the future of repositories by informing practice and enabling interoperability and coordination between existing and developing repositories.

A year into the Programme, it would be impossible to capture the full range of work currently happening or the expected deliverables from projects with up to a year to run. One project already completed, though, is Repository Bridge, whose aim was to facilitate the automatic deposit of electronic theses from the University of Wales to the National Library of Wales. From its final report, the project "has succeeded in its primary aim of developing software to enable the automatic import of theses from institutional to archival repositories. Along with the similar work undertaken by the EThOS project, we have demonstrated that combining OAI-PMH and METS is an appropriate approach to achieving this export, especially because of the support for alternative encodings of descriptive metadata."[4]

By successfully demonstrating that the proposed deposit mechanism can be achieved using open source software and open standards, this project is an important milestone for the Programme. Yet, this is not its only deliverable as, on the road to that outcome, the project has documented its understanding of processes (theses deposit), technologies (DSpace and Fedora) and standards (METS and OAI-PMH). They have also fulfilled roles of dissemination, awareness-raising and advice

and have collaborated with another project, EThOS, ensuring interoperable development.

Collaboration, advocacy and awareness-raising are key for projects investigating different aspects of the same topic and help to ensure .duplicate activity or divergent practices do not arise. In addition, other JISC Programmes, international activity and practical implementations in institutions make it essential to utilise opportunities to share and collaborate. One mechanism is through the work of the Support team, whose remit includes dissemination and synthesis, as well as facilitating a set of cross-Programme cluster groups to help projects working on similar themes share knowledge and experience.

For example, the data cluster includes a number of projects examining issues surrounding the deposit and curation of primary research data, such as workflows for the deposit of experimental data (R4L, SPECTRa), linking research data with research publications (R4L, StORe), mechanisms for citing research data (CLADDIER) and the reuse of geospatial data (GRADE). These projects have also developed links with related activities such as the Digital Curation Centre, the e-Bank project and DART, an Australian initiative. Cluster activities include a forthcoming information and collaboration event, to be held at one of the national data centres.

Beyond the cluster groups, many project outputs relate to work of other projects. One example is the GRADE, one of whose outputs is a report entitled *Use Case Compendium of Derived Geospatial Data.* This detailed report presents a series of geospatial use cases as a basis for an examination of copyright issues relating to selected data sets. Other projects looking at digital rights issues are TrustDR and Rights and Rewards, whose focus is on repositories of learning objects and teaching materials. Rights and Rewards are also examining barriers and potential reward mechanisms to motivate deposit into repositories. Both projects have already produced a number of interesting reports and papers with further models, guidelines, briefings and development materials to follow.

To truly help the cause of interoperability between repositories, both in the UK and worldwide, the Digital Repositories Programme has established a Support team whose remit involves supporting projects through training and guidance and, in addition, engaging in dissemination and strategic activities within the wider repositories community. Their work to date has included training for projects on writing scenarios and use cases, establishing an international repositories discussion list, creating a Dublin Core Application Profile to facilitate the exchange of metadata relating to scholarly publications (eprints) and a producing roadmap which envisions a way forward for repositories over the next 5 years.

A key relationship for the Programme is with the JISC (UK) and DEST (Australia) e-Framework (www.eframework.org), a service-oriented framework for education and research. One project, ASK (Accessing and Storing Knowledge), is utilising the e-Framework to document a reference model and design for a repository software system. This kind of service-oriented approach to establishing interoperable services is of significant interest to a swathe of other projects falling within the remit of the ‘Integrating infrastructure’ cluster. Here, EThOS are working with the British Library towards a fully electronic e-theses service; SPIRE are investigating the use of peer-to-peer technology for sharing resources; MIDESS will enable three universities to collaboratively manage their image collections; and IRIScotland are scoping a National repository search service for Scotland. The PerX project offers a subject perspective on cross-searching of repositories. The following project profile highlights the progress PerX has made.

Project Profile : PerX

The PerX project (Pilot Engineering Repository Xsearch) www.icbl.hw.ac.uk/perx/ is based at Heriot-Watt University, with partners at Cranfield University, the Institution of Civil Engineers/Thomas Telford Ltd, University of Arizona and RSC East Midlands. It is one of the JISC Digital Repositories Programme (DRP) projects, and has been funded for two years from June 2005 to explore various issues associated with the provision of subject-based resource discovery services.

What has the PerX Project achieved so far, what are some of the issues it has discovered relating to subject-based access to digital repositories, and what can be learnt from its findings?

One of the first tasks of the project was to produce a listing of the most significant repositories of relevance to a particular subject area (engineering) and to provide examples of repositories via type and coverage. This listing is available at www.icbl.hw.ac.uk/perx/sourceslisting.htm It reveals that there is a wide variety of existing and potential digital repositories of interest to the subject in question, including repositories where actual content has been deposited and also metadata repositories which contain only metadata about content. To give a flavour of the findings, these can be classified according to content type: Research data, e-Theses, Learning Materials, Multimedia, Assessment Materials, Technical Reports; and coverage: Personal/Informal, Journal, Institutional Departmental, Inter-Institutional, National, Geospatial, and Subject Access.

An analysis of the engineering digital repositories landscape www.icbl.hw.ac.uk/perx/analysis.htm as revealed in the 'Listing', plus reference to existing literature on the topic, reveals several interesting things. Firstly, despite the overall number of repositories, there are some significant gap areas of provision in engineering (including research data, subject-based access, technical reports (in the UK and Europe), journals, and assessment materials). Secondly, the means and levels of interoperability of the identified repositories varies from un-interoperable, to non-standard interoperability (i.e. proprietary APIs), to fully functional interoperability based on established standardised means such as Z39.50, SRW and OAI-PMH. Thirdly, the information landscape of engineering is quite complex. It includes resource-types such as technical reports, standards, patents and trade literature, alongside more obvious types such as peer-reviewed scholarly articles. Lastly, differences between disciplines, and also the real information needs and information retrieval habits of subject communities, needs to be carefully considered when developing subject-based resource discovery services.

In the context of digital repositories, what is meant by 'interoperability' and 'metadata? Why are they important for data providers and service providers? Why is a standardised approach to interoperability important? What is the difference between harvesting of metadata and distributed searching of metadata? How can Z39.50, OAI-PMH, Static OAI Repositories and SRU/SRW facilitate content syndication? In order to answer these questions, and also to encourage data providers to expose their data via standardised means, PerX published a document entitled 'Marketing' with Metadata - How Metadata Can Increase Exposure and Visibility of Online Content www.icbl.hw.ac.uk/perx/advocacy/exposingmetadata.htm . This explains, in non-technical language, all of the above and outlines ways by which content providers can share, or embed, their descriptive data (metadata) with

other websites, in standard and reusable ways. This document has received favourable feedback, and continues to attract a considerable number of downloads.

A major landmark in the PerX Project was the creation of a Pilot service www.engineering.ac.uk/ allowing numerous (at the time of writing, 29, though this will increase in the future) digital repositories to be cross-searched from one interface www.engineering.ac.uk/index.html?action=basic . The repositories vary considerably in size, content and type, and range from a large subset of arXiv.org to the much smaller Geotechnical, Rock and Water Resources Library (GROW) Digital Library.

Results Summary	Search status	No of results
Aeronautical Research Council Technical Reports	Finished	5 records - go to results
Australian Research Repositories Online to the World (ARROW)	Finished	3 records - go to results
arXive - ePrint Archive	Finished	1 records - go to results
Caltech Earthquake Engineering Research Laboratory Technical Reports	Finished	0 records
CISTI - Canada Institute for Scientific and Technical Information	Finished	30 records - go to results
Copac - Library Catalogues	Finished	105 records - go to results
CSA Discovery Guides	Finished	8 records - go to results
Digital Library Network for Engineering and Technology	Finished	0 records
Directory of Open Access Journals (DOAJ)	Finished	19 records - go to results
DSpace at MIT	Finished	5 records - go to results
Energy Citations Database	Finished	26 records - go to results
EEVL Ejournal Search Engine (EESE)	Finished	400 records - go to results
EEVL - Best of the Web	Finished	1 records - go to results
EEVL Website Search Engine	Finished	400 records - go to results
ePrints UK	Finished	15 records - go to results
Geotechnical, Rock and Water Resources Library (GROW)	Finished	1 records - go to results
Inderscience Journals	Finished	4 records - go to results
IoP Electronic Journals	Finished	158 records - go to results
JORUM	Finished	0 records
NACA Technical Reports	Finished	4 records - go to results
NASA Technical Reports	Finished	37 records - go to results
Networked Digital Library of Theses and Dissertations (NDLTD)	Finished	83 records - go to results
National Engineering Education Delivery System (NEEDS)	Finished	0 records
OneStep Jobs	Finished	0 records
OneStep Industry News	Finished	0 records
Pearson Education - Books	Finished	0 records
Recent Advances in Manufacturing (RAM)	Finished	0 records
SearchLT Engineering	Finished	0 records

Figure 1: PerX Pilot cross-search results table

Feedback about the Pilot, through both a web-based survey and face-to-face focus groups, has shown that there is considerable agreement about the need for a subject focused service which cross-searches numerous collections in engineering. Various suggestions for improvements to the Pilot interface have been made, but what is most obvious is the need to increase the number of digital repositories of various kinds being cross-searched. This mirrors the findings of the 'Analysis' mentioned above, which concluded that a service which focused only on materials in repositories, and ignored materials found in other sources for which metadata repositories may be available, would be unlikely to be regarded as an essential information retrieval tool.

On the technical side, PerX has found that metadata harvested from OAI-compliant repositories too often contains non-valid or ill-formed XML documents which need to be corrected before further use. Another limitation, which is especially important in the context of subject-based services, is the lack of uptake of OAI 'Sets' by many data providers. A very basic subject-type standard for sets would make the identification, by aggregators/subject-based services, of relevant records from multi-disciplinary repositories much easier. Quality of metadata is an issue which needs further attention.

The future for repositories

Looking forward, the *Repositories Roadmap,* produced within the Digital Repositories Programme, offers a vision for 2010 of …

> a technical infrastructure that supports the deposit, discovery, access and use of objects in repositories by software applications". Such an infrastructure needs to work across both open access and closed repositories and be based on a more thorough modelling of the objects being made available, the way such objects are described and identified and the mechanisms for automatically interlinking and manually citing scholarly output, research data and learning objects. There will be widespread agreement about the machine to machine interfaces (the services) that open access repositories should support in order to ingest and make available content and metadata. This activity will provide a solid environment within which a wide variety of software tools (both open source and commercial) and added value services can be developed.[5]

Another important report from the Programme is *Linking UK Repositories*, produced by Alma Swan and Chris Awre. It concludes thus:

> The creation of a system of Open Access repositories across the UK with user-oriented services built across them will not happen properly unless it is led by an organisation with vision and focus. The essential issues in the process are planning, communication and coordination. The task is complex and will require firm management combined with the ability to project the overall vision to all constituencies that might be involved. The outcome is a most worthwhile goal, and provides a host of opportunities for all the players and stakeholders. Coordinating their activities is the challenge that needs to be tackled.[6]

Delivering on these visions for an interoperating infrastructure of repositories and services is no easy task, but the excellent work that has been and is being done by JISC-funded projects is already having an impact. The JISC Capital Programme is set to build on this, through initiatives such as the UK repositories search service which will offer a single point of access to search repositories from across the UK. Using the Eprints metadata application profile[7], developed under the aegis of the Digital Repositories Support team, will enable the service to offer a much richer set of search features. Funding to support institutions to implement or enhance repositories and for a national repository advisory and support project are also in the pipeline, and many more projects will come on stream over the next three years, offering new tools and mechanisms for support widespread open access to resources.

Note

Information about all of the projects mentioned above, along with links to their web sites can be found from the Digital Repositories Programme wiki: www.ukoln.ac.uk/repositories/digirep

[1] As noted in the *Digital Repositories Review* (www.jisc.ac.uk/uploaded_documents/digital-repositories-review-2005.pdf), ongoing work on a typology and ecology by the Digital Repositories Programme Support team (www.ukoln.ac.uk/repositories/digirep/index/Typology_and_ecology) and the CD-LOR project *Report on Learning Communities and Repositories* (www.ic-learning.dundee.ac.uk/projects/CD-LOR/CDLORdeliverable1_learningcommunitiesreport.doc)

[2] OpenDOAR (www.opendoar.org) lists 56 UK repositories; ROAR (archives.eprints.org) lists 72, 44 of which are classed as Institutional or Departmental

[3] FAIR (Focus on Access to Institutional Resources) Programme (2002-2005); X4L (Exchange for Learning) Programme (2002-2005); Digital Repositories Programme (2005-2007), see www.jisc.ac.uk

[4] Repository Bridge. *Final Report*, v1a, 28/06/2006 www.jisc.ac.uk/uploaded_documents/Repository_Bridge_Final_Report.pdf

[5] Heery, Rachel and Powell, Andy. *Digital Repositories Roadmap: looking forward*, April 2006 www.ukoln.ac.uk/repositories/publications/roadmap-200604/

[6] Swan, Alma and Awre, Chris. *Linking UK Repositories,* 2006 www.jisc.ac.uk/uploaded_documents/Linking_UK_repositories_report.pdf

[7] See www.ukoln.ac.uk/repositories/digirep/index/Eprints_Application_Profile