

Annex 1: JISC Digital Repositories Review Focus Group Report

Goal of the focus group

To challenge a set of people who have experience-based expertise in repositories to give their opinion on the wider repositories scene as a contribution to the JISC Digital Repositories Review, in order to ensure input from this expert sector of the community.

<i>Time and date</i>	<i>Location</i>	<i>Incentives for participants</i>
3.30-5.30pm Monday 18 October (in association with the DPC/CURL repositories Forum at the British Library on Tuesday 19 October) to be followed by continued discussion over supper.	AHDS in Drury Lane, then a nearby restaurant. A room to accommodate 12 people around a table in a comfortable and fairly informal way was provided; and refreshments arranged for the 3.30pm start, with wine served toward the end of the session.	Opportunity to influence next JISC call for developments in repositories. Opportunity to share views with other experts. An evening meal, plus up to £50 toward expenses if requested.

Participants

There was a good response to the invitation to participate that was issued through the DPC. The first nine volunteers were accepted, on the assumption that at least seven would actually turn up on the day, and that if all nine did attend it will be manageable, though large for a focus group. One volunteer did have to withdraw, leaving the following eight participants:

1. **Richard Boulderstone (RB)** British Library. Director of e-Strategy and Information Systems, active in DPC and ERPANET
2. **Jessie Hey (JH)** University of Southampton School of Electronics and Computer Science. Research Fellow on the TARDis project and very involved with eprints
3. **Martin Moyle (MM)** of UCL Library is on the SHERPA project, and is Engineering Faculty Team Leader of UCL Library
4. **William Nixon (WN)** Deputy Head of IT Services / Project Co-ordinator, Glasgow University Library
5. **Mared Owen (MO)** of National Library of Wales. Systems Manager involved in Web archiving and with experience in digitisation issues
6. **Michael Popham (MP)** is Head of Oxford Digital Library, a core service of Oxford University Libraries Services
7. **Rachel Proudfoot (RP)** of White Rose Eprints. Project Officer and repository administrator (White Rose is part of SHERPA)
8. **Pauline Simpson (PS)** of University of Southampton Oceanography Centre. Chief Librarian, active on ePrints and the TARDis project

The facilitators were Sheila Anderson (SA) of AHDS, Amanda Closier (AC) of UKOLN and Leona Carpenter (LC) a freelance consultant.

Plan for the focus group

1. Introduction

- to facilitators
- to the Repositories Review
- to the aims of the focus group

2. Participants' aims

- who they are and their repositories / organisations / projects
- their own aims/reasons for participating in the focus group

3. Defining repositories – what are the essential features?

4. Relationships– nature and challenges

- relationships among repositories
 - e.g., of eprints, of learning objects/materials, of research datasets
 - e.g., national/institutional/subject repositories
- relationships of repositories to organisations' goals, activities and services
- relationships of repository systems to other systems
 - e.g., VLE, MLE, CMS, RMS, DAMS

5. Issues relating to sustainability of repositories

- e.g., many or one, who manages them, who funds them
- e.g., institutional motivation for investment in them
- e.g., the roles/responsibilities of institutions in relation to datasets
- e.g., buy-in by senior management
- e.g., buy-in by creators of digital assets/objects/data

6. Research and development ideas and priorities for the future

- current and future roles of repositories
- gaps in provision
- challenges

7. Summing up

- key points from discussion
- what's next with the Review
- instructions for meeting up for supper

Record of the focus group

Introduction

Leona Carpenter introduced the facilitators. Sheila Anderson is the Director of the Arts and Humanities Data Service, with additional responsibility for the co-ordination and development of AHDS collections, standards and rights management. Her role in the focus group was to brief the participants about the JISC Repositories Review and the goal of the focus group, to watch out for people who might be wanting to get a word in edgewise, to assist with keeping to schedule and to deal with any interruptions or environmental problems that arise. Amanda Closier is a UKOLN researcher who currently works on the JISC-funded Information Environment Service Registry (IESR), has been involved in the QA Focus service, and has just completed a literature review on repositories. Her role in the focus group was to take notes as someone who understands the subject of discussion – which is very important in such a specialist subject. Leona Carpenter is a consultant with a background in systems analysis and digital library research and development, most recently conducting a user requirements analysis for the Digital Curation Centre. Her role was to plan and lead the focus group and write this report on its outcome.

Participants aims

Leona asked the participants to take a minute or two each to introduce themselves and their repository experience, and their own reasons for participating in the focus group.

Michael Popham

Michael Popham is Head of Oxford Digital Library (ODL) and is involved with the Oxford Eprint repository, part of the SHERPA initiative, including the recently funded JISC Circular 4/04 SHERPA follow-on project on digital archives for preservation of exemplar papers. Oxford also has the AHDS centre for literature, language and linguistics, as well as an e-Science centre.

Richard Boulderstone

Richard Boulderstone is the British Library's Director of e-Strategy and Information Systems. He came to the focus group looking for answers to the questions: "What is a repository?" and "What services does a repository support?" He mentioned that the BL is connected to the SHERPA project. The BL has a digital library programme that is being rolled out in stages and is starting up by building a digital store – which already has some successful functions. Richard is interested in what is happening with other repositories.

Rachel Proudfoot

Rachel Proudfoot of White Rose Eprints is project officer and repository administrator. White Rose is part of SHERPA, and has been live for about three months. Rachel works on the operational rather than the management side of the project, and ends up acting as an advocate to both administrators and academics.

William Nixon

William Nixon is the Deputy Head of IT Services / Project Co-ordinator, Glasgow University Library, involved with the Daedalus project. They have used eprints.org

and DSpace software. Published and peer reviewed papers (subject to IPR issues), theses, preprints and working papers are held, segregated into repositories by type of content. Metadata is harvested from each repository via OAI-PMH as a basis for cross-repository search services. Project management is divided over two posts, one for service development, and one for advocacy.

Pauline Simpson

Pauline Simpson, the University of Southampton Oceanography Centre Chief Librarian is on the TARDIS project. TARDIS is providing a repository for research output. They have found large differences between disciplines in terms of the types of materials people want to deposit and level of interest in repositories. At the University of Southampton, they are co-located with eprints.org repository software development activity, and have fed back into that from their experience. They were glad to have some influence on eprints.org software development. Their repository started as full text only but has evolved to include entries with metadata only (descriptions and abstracts) with full text where possible in accordance with copyright terms. They are looking at RAE-support routes, and are developing a business model for repository sustainability.

Martin Moyle

Martin Moyle of the UCL library is SHERPA Project manager for London LEAP. (London LEAP is a consortium whose members are Royal Holloway, Imperial College, King's College, School of Oriental and African Studies, Birkbeck, and University College London.) They have established an ePrints service, collecting research outputs of University of London colleges. All LEAP repositories share one server, although local differences have to be catered for. Martin said that preservation issues were certainly coming up in relation to their repositories.

Mared Owen

Mared Owen is Systems Manager of the National Library of Wales (NLW). NLW is particularly interested in archives and electronic archives, and has been looking into the Trusted Digital Repository model. They have discovered that there are a lot of theories, models and pilots. NLW is still at the exploratory stage with this. They are concerned about digital preservation, but uncertain as to whether they should attempt to be a trusted digital repository themselves, and whether there should be a certification authority at some point in the future. Mared said NLW would look into the feasibility of this, particularly in terms of cost-benefit. There simply may not be sufficient funds available. NLW is looking at Fedora to deliver digitisation projects. They are also involved in an e-journals project with other legal deposit libraries.

Jessie Hey

Jessie Hey of the University of Southampton School of Electronics and Computer Science is also on the TARDIS project. Her work is on the advocacy side, working with departments on a mixture of technical and softer aspects of repository development. This includes work with the eprints.org software development people. Jessie noted the importance of tools such as the experimental ParaCite tool for improved searching. She will also be involved in a JISC Circular 4/04 project. Jessie has particular concerns about interoperability with other institutions and projects.

1. Defining repositories – what are the essential features?

There was agreement across the group that 'Repository' is too static a word, because it sounds like a place where you dump stuff and then nothing happens to it.

There was a suggestion that **software** is a defining feature. General frustration was expressed that there is no single system out there that does everything. Repositories and the software that supports them are evolving all the time. One participant argued that it is more important to define the protocols and data structures that will ensure interoperability and longer term flexibility, than to worry about which application software is best at the current time.

Some participants were more inclined towards the **services** side of repositories in terms of definition. Others focussed on the current lack of software functionality to support services. There was broad agreement that repository software is in its infancy and the development is still going through an iterative process. Participants also agreed that JISC had demonstrated remarkable forethought in commissioning the FAIR programme. It came in at the right time. ("Infinite wisdom" was mentioned!). Those from FAIR projects said they would do things a lot differently now because of the opportunity that FAIR provided them to learn while doing.

An opposing suggestion was that the defining aspect of repositories lies in **sharing or providing access to information**. There was certainly some support for this view, with one participant stating that with their repository they were focussing on sharing among the different departments and other units within their own institution and also beyond. IPR barriers were said to be stopping some institutions moving forward with filling repositories with **content** other than **metadata** because of publisher issues.

One participant stated that the purposes of repositories are constantly expanding because of growing interest and take up. She questioned the wisdom of the current tendency to accept everything into repositories, and suggested that undertaking digital preservation implies some curatorial control and selection of content. Another said that their institution thought of the repository as trying to bring everything together, including **preservation**. There was general agreement that although a repository was not simply text or literature based, some kind of **selection policy or criteria for types of content to be included** was necessary.

2. Can we articulate Services that makeup a repository?

There was general agreement that successful advocacy for repositories is dependent on being able to offer services that people valued. Services that academics had asked of repositories included search services, support for creating CVs, and support for managing personal web pages. Within institutions, individuals have a positive response to repositories if the services are seen to be of benefit to them. Publishers are quite positive but do want to know about what services are offered.

2.1 Search services / finding aids

Academics are asking repository operators how people are going to find their stuff once they have deposited it. Credibility is lacking when large services are only in the pilot stage and academics question whether it is worth the effort of being involved, at least partly because they are reminded of this by disclaimers on the repository sites. Two participants noted that there was a call at another recent focus group for a harvesting service at a national level so that content could be accessed. That cross searching services are not mature for ePrints is seen as an important current issue.

Some people in the group thought that making repository content visible through search services such as OAISTER and Google make deposit/exposure through OAI more attractive and its benefits more demonstrable. Others would never promote the use of any of these, including (for some) OAISTER. There was some agreement that it would be worthwhile to try to incorporating Google technology or at least Google-like features in search services. An example cited was automatic suggestion of alternative spelling of a search term.

2.2. Usage metrics and statistics as value-added services

Some institutional repositories are looking at statistics (metrics) on which to base value added services and some of these services may prove of wider interest. Further work would be required on statistical output from repositories. Currently there seem to be limited possibilities at the moment to exploit metrics. Some repositories keep track of statistics privately due to various issues, for example the length of time the paper has been in the repository. Metrics would have to be weighted to take account of such issues, for example in providing a "top ten most accessed papers". Opportunities in metrics and service take-up are in a feedback loop, in that repositories need a critical mass of content for statistical outputs to be meaningful, while the availability of statistics is an incentive to institutions to mandate or otherwise support deposit. There was general agreement that it is difficult to put meaningful interpretation on statistics. No realistic or meaningful analysis system is available yet. Sound analysis for statistics needs to be defined. The inevitability of interest in the metrics from management, funders and other means that it is important to develop this area.

2.3. Metadata as a service

The suggestion was made that one potential service is for repositories to re-package, correct, and/or otherwise enhance metadata and re-present such enhanced metadata to the wider community. It was observed that this is already done in the archive community. It is difficult to see how institutions will be able to find the resources to deal with the required metadata creation and copyright clearance. The quality of metadata provided by an institution is representative of that institution and must be of high quality if for this reason alone. There had been an assumption that the time/effort required to create and manage metadata would be swallowed in normal operational budgets, but it is proving too large a task to be accounted for in that way. Ensuring interoperability by provision of high quality metadata is a significant drain on resources. More thought needs to be given to the automatic capture of metadata. It was observed that eprints.org software does not capture preservation metadata.

One challenge is thought to be that there is no consensus within the HE community on how to cite data-sets, there are no widely-used standards for doing so. Group opinion seems to indicate that social and physical sciences are ahead on the documenting of their data.

There was strong feeling that inadequate metadata limits the services that can be offered. Some thought that simple (often commercial) tools might help. These might include such things as clustering and synonym services. The BL was said to be planning to use a commercial solution in one of their websites.

3. Relationships – nature and challenges

From the content providers' point of view, there may be a wide range of different repositories under one roof and this will continue to cause problems with academics unless an interface is designed that is easy to use and will enable them to submit to a range of different sub-repositories. Repository users need a single search interface, but also a single deposit/input interface as a top layer across repositories.

3.1. Relationship of institutional, research data and subject repositories

Interest was expressed in the eBank approach to linking records in different repositories, so that users could move between original research data and the analysis of the data in publications. There was a perception that access to and deposit of research datasets is becoming more attractive to academics and that preserving this material for the future will become more of an issue.

Currently the focus of institutional repositories on peer reviewed, published papers enables the demonstration to academics of the value of using these instead of, or in addition to, subject-based repositories. However it was acknowledged that academics in many fields feel more connected to their subject community than to their institution, especially as many will change institutions from time to time. Thus, subject-based repositories could be seen as a means of taking their papers with them. However, institutional repositories are likely to be sustained but there may be more problems with sustaining subject repositories. It was felt that there would be great interest in services that combine access to these subject repositories and other repositories including institutional repositories.

It was noted that many (most?) formally constituted research data repositories are subject-based, rather than institutional, with AHDS to hand as an example. SCIRUS – a commercial service that captures academic research – was mentioned. There was assent to the suggestion that the community could benefit from JISC profiling projects by combining data and text based repositories, moving on from the position at the moment of data archives and text based archives being developed separately.

Would institutional repositories be willing / able to harvest from other repositories? Infinite potential exists for inter-connectedness of repositories and for overlay services, once repositories are in place and have adequate content. Perhaps the next step is another layer on top of institutional repositories. Then repositories and their content could be clustered in other ways.

3.2. Beyond text-based repositories

Although institutional repositories had originally focussed on ePrints of published papers as the primary target content, it is increasingly difficult to put a narrow limit on the content of institutional repositories. There is interest in the potential for archiving data at an institutional level, and various issues arise in this context due to problems with dynamic material. These problems are similar to those posed by archiving Web sites and pages. Three types of material were identified for inclusion in institutional repositories:

- Static or passive data – not changing over time nor controlled by software, e.g., document files, images.
- Resources that are updated regularly over time.

- Active data that has software embedded in it – e.g., Java embedded in Web pages and resources that have embedded applications.

There are also issues with proprietary formats and research data formats developed as 'home-grown' solutions for which documentation is not always available.

3.3. Relationship with other systems, such as VLEs

Is interoperability between repositories and other systems (e.g. between ePrint repository and Virtual Learning Environments) being considered at an institutional level? Participants say that on the whole – no, it is not happening. Coming from the ground up it appears that a broader view needs to be taken at a strategic level and that is not happening, except among a few early adopters.

4. Issues relating to sustainability of repositories

4.1. Open source software

There was some concern expressed about the long term sustainability of FAIR open source software outputs. Open source solutions were being produced as enhancements to existing tools and systems, but sometimes are not being fed back in to the parent development. DSpace and eprints.org are still immature products in terms of development. The community is looking at ways of managing feed back into the code and sustaining product development. Both of these open source products have been funded on a project basis. One participant argued strongly that a sustainable solution may only come from some commercial input. Organisations are nervous about taking on open source software where support is seen as non-existent, but it was thought there could be commitment to open source if commercial support services were available. Lack of service level agreements is a big disincentive to uptake of open source solutions.

4.2. Collaboration

More exploration should be done regarding the relationships between JISC, the national libraries and other national or regional bodies. What would be better done at local, regional and national levels? For example, the 'Digital Library programme' cannot be done by the BL alone, there is a need for inter-working. The repository task is too huge to be done by BL as a national collection point, but the BL might develop services that could be used/adopted by others. There was some discussion of the BioMed Central model. BioMed Central offers hosting of repositories as separate, charged service. Concerns were expressed about commercial organisations offering such services, because these organisations lack the continuity of purpose, or even existence that is required to ensure continuing availability of content and services. It was suggested that institutions should work with the national archives on some services. The provision of trusted digital repositories with an established reputation is seen as a good thing. Perhaps such 'trusted repositories' might be provided for long-term preservation across particular regions, for example? Trying to avoid duplication is an important consideration. Trusted digital repositories are seen to be the next generation and something to be aimed for. The next JISC Repositories call should crystallise work already done and work with the 4/04 Preservation call to move in this direction. It is important to recognise that many organisations across the UK (for example, FE colleges) cannot afford to create repositories – there needs to be a

solution catering for the requirements of FE, perhaps regional repositories might provide such a solution?

4.3. Overall

Can we say whether the current crop of repository projects will survive as services? The group expressed optimism but no certainty. To general agreement, some participants argued that the external environment, including increased central government interest, is changing constantly and not continuing to develop repositories is not an option. The recent move in Scotland to try to develop an open access declaration was seen as a positive indication of how things are going, and such initiatives in other places could help.

5. *Research and development ideas and priorities for the future*

Participants were asked if they could get money [from JISC] for one more thing – what would it be? The resulting wish list clustered around issues relating to copyright, data quality, and new or improved tools.

5.1. Content, copyright and legal advice

Everyone agreed on the need for content, with many thinking we must persuade academics to self-archive. It was noted that RC UK was meeting to look at open access to research output, specifically with regard to publishers. People wanted more content for their repositories, and saw the current **copyright** situation as a (if not the) major barrier in achieving that. They felt the ideal would be to get changes to the law of copyright, especially so that research output and items of no apparent continuing commercial value could be made available. In the absence of change, and as changes *are* made, concrete legal advice on issues relating to copyright, IPR and so on is required. It was suggested that an alternative to statutory change might be the negotiation with publishers of a license to deposit.

5.2. Data quality – funding for creation, content input

Data quality as barrier to populating repositories – and this could be significantly reduced by **funding clerical effort to get items input, including quality metadata**, moving away from self-archiving by academics. See also 2.3, above for discussion of metadata quality.

5.3 New or improved tools or toolkits

The following were wanted:

- **"smart tools"** for automatic data extraction, automatic classification (STP?)
The national archives[?] concept of 'Straight Through Processing' (STP – a transaction travels through multiple IT systems without a need for manual processing) was attractive to some. This is a concept borrowed from financial applications such as cheque clearance and trading operations and in commercial situations where records management is undertaken for purposes of evidence and to meet disclosure obligations.
- **format conversion tools**
- **increased/improved functionality of repository software**, particular importance was placed on the need for improvements to the user interface for submitting data, including metadata and submission across multiple repositories. In addition more robust exporting functionality is required.

- **tools to facilitate interdisciplinary exploitation** of repositories
- **a repository module to support/manage the RAE.** There was strong agreement that JISC should fund the development of an RAE-support module. It was felt that using repositories to support the RAE would provide leverage to get content into repositories, and would go some way towards raising awareness of repositories nationally. However, it was noted that even building rudimentary RAE information management support into repository systems was hampered by uncertainty about what the RAE categories would be next time around. "It's a marketing problem," said one participant. "What is the service proposition and what are they getting out of it?" He went on to suggest that we should be telling the RAE what sorts of measurements we can provide. Since people were saying that the RAE was likely to ask for things which cannot reliably be measured, we should forestall that problem by arguing for robust measures that can be provided on the basis of repositories. It was agreed that a mechanism to influence RAE categories would be very helpful.