Digital Repositories Roadmap: looking forward

Document details

Author:	Rachel Heery, UKOLN, University of Bath Andy Powell, Eduserv Foundation
Date:	2006-04-07
Version:	15
Document Name:	rep-roadmap-v15
Notes:	

Acknowledgement to contributors

The authors would like to thank the following people, who contributed to the roadmap by completing an email questionnaire or commenting on previous versions. The authors take responsibility for interpreting the answers and for any change of emphasis that comes with collating the viewpoints of the various contributors.

- Sheila Anderson, AHDS
- Paul Ayris, UCL
- Phil Barker, CETIS
- Rachel Bruce, JISC
- Lorna Campbell, CETIS
- Fred Friend, UCL
- Mike Hursthouse, University of Southampton
- Bryan Lawrence, CCLRC
- John MacColl, University of Edinburgh
- David Medyckyj-Scott, EDINA
- James Reid, EDINA
- Stephen Rogers, MIMAS
- Andrew Rothery, University of Worcester
- Pauline Simpson, University of Southampton

Acknowledgement to funders

This work was funded by the JISC as part of the Digital Repositories Programme.

UKOLN is funded by the MLA: The Museums, Libraries and Archives Council, the Joint Information Systems Committee (JISC) of the Higher and Further Education Funding Councils, as well as by project funding from the JISC and the European Union. UKOLN also receives support from the University of Bath where it is based.

Eduserv is a not-for-profit IT services group, born from services developed within universities from 1988. Eduserv now delivers innovative technology services predominantly to the public sector and the information industry. Services include access management, software and information licence negotiation, managed web hosting and web applications development. With the contributions generated from these activities the Eduserv Foundation funds initiatives to support the effective application of IT in education.

1 Executive summary

This roadmap presents a vision for 2010 in which a high percentage of newly published UK scholarly output is made available on an open access basis and in which there is a growing recognition of the benefits of making research data, learning resources and other academic content freely available for sharing and re-use. Furthermore, geospatial information will be better integrated with other data through improved licensing agreements. Achieving this vision over a four-year period will not be easy, but it is intentionally set as a challenging aim in order to help focus discussion on what needs to happen to make it a reality.

The authors suggest that while the current technical infrastructure in the UK is in need of some development, it is primarily in the areas of policy (both national and institutional), culture and working practices that changes need to be made. We suggest that the JISC and the wider community need to focus their activities in the following areas:

- Policy – Research councils and other funding bodies need to mandate that all scholarly publications generated by publicly-funded research are made available on an open access basis. The RAE needs to move significantly towards using open access copies of scholarly publications as a primary mechanism to support the assessment exercise. Motivated both by the open access agenda, and by the requirement to manage their digital assets effectively, institutions should build curation of scholarly publications, research data and learning objects into their information strategies. Although the long term preservation of all academic output is an important consideration, the aims and issues in this area need to be clearly articulated separately from (but in relation to) the aims of open access and asset management.
- Cultural – The 'reward structures' and 'professional development' infrastructure within the academic community need to recognise open access as a valuable and important part of the profession. The community needs to find ways to encourage academics to share and re-use publications, research data and learning resources as openly as possible.
- Technical – The technical infrastructure supporting open access needs to be based on a more thorough modelling of the materials being made available, the way such materials are described and identified and the mechanisms for automatically interlinking and manually citing scholarly output, research data and learning objects. There needs to 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. Finally, repositories should be well integrated into institutional and national access management approaches (such as Shibboleth). These activities will provide a solid environment within which a wide variety of software tools (open source and commercial) and added value services can be developed by both the public and private sectors.
- Legal – The licensing of community-developed content needs to protect the intellectual property of institutions, individual academics and third-parties as necessary yet still be supportive of the open access approach. The community needs to find ways to avoid a situation where concerns about IPR are allowed to stifle the creative sharing and re-use of academic content.

2 Introduction

2.1 Purpose of the roadmap

This roadmap is intended to inform the JISC's planning processes and stimulate discussion in the community. It will focus on digital repositories and their role in the information landscape, exploring:

- The starting point — where we are now.
- A destination — where we want to be in 2010.
- A route — what we need to do to get to that destination, including the 'milestones' to be reached. As the document firms up, these milestones may be given target dates and responsibilities.

The document is a first pass at formulating a roadmap. It has been compiled taking into account previous documents and (limited) consultation with various domain experts, who were asked to input their ideas by means of an email questionnaire. The authors have freely used these contributions, but of necessity have interpreted the ideas and, in part, have also added to them. The authors take responsibility for any misinterpretations or changes in emphasis.

There are many unknowns in this area, so the roadmap is aspirational and, to some extent, speculative. This is the first iteration; the intention is to seek further input based on feedback to this draft. It is likely that versions of the roadmap will be produced in future as supporting material for various JISC calls and to inform other activities as necessary.

2.2 Background

For various reasons (political, cultural and financial) the JISC has funded a range of individual digital repository projects, which, whilst they all address technical and organisational barriers to setting up an integrated UK repository system, have not sought to develop that integrated system directly. So, unlike some similar initiatives elsewhere, for example DAREnet[1] in the Netherlands, the JISC repository programmes have not used funding to develop a managed network of institutional repositories, but rather have explored development across a range of areas. This has resulted in programmes, FAIR and the DRP, made up of clusters of projects in various areas (data, learning, legal, preservation, integrated infrastructure) with various common themes (user requirement analysis, metadata standards evaluation, evaluation of software platforms). It has led to a range of innovative developments and to engagement with the international community.

The JISC approach has facilitated innovation across a broad range of areas, however because no central service is under development, there has been no compelling reason to address the full range of issues arising from development of an integrated infrastructure. This is unlike the situation in the Netherlands where the commitment to provide a search service across all repository content has focused attention on integration and highlighted from the start the need for a common approach to various technical issues. With the additional CSR funding now available to the JISC, the intention is to directly support development of infrastructure to maximise investment in digital content. Increased deployment of repositories within the UK will raise organisational, policy and technical issues and a common infrastructure will increase the effectiveness of that activity.

2.3 Scope

This roadmap focuses on UK repositories for research outputs (text, data and other) and learning materials. Administrative records are out of scope. Furthermore, the roadmap is only

[1] DAREnet. <http://www.darenet.nl/en/page/language.view/home>

concerned with objects created, owned and shared by members of the HE/FE community not those made available to HE/FE on a commercial basis.

The roadmap will consider repository services associated with management and dissemination of research and learning outputs of UK institutions offered at institutional, national or subject-based disciplinary level. The roadmap will not include 'repositories' that manage and provide access to information about collections and services, ontologies and terminologies, nor analysis tools (often characterised as 'registry services').

The roadmap looks towards a destination in 2010. It will describe gaps to be addressed between now and then, covering the two main strands of the Information Environment:

- discovery to delivery,
- sharing, curation and management.

2.4 Audience

The principal audiences are:

- the JISC Executive,
- the Repositories, Preservation and Asset Management Advisory Group,
- the relevant JISC Committees.

The roadmap will also be made available from the JISC Web site. It is hoped that it will be useful to HE and FE institutions as they consider their digital repositories and content policies.

3 What is a repository anyway?

It comes as no surprise that there are many understandings of what a 'repository' is, and this roadmap will not try to resolve that debate. However it is worth emphasising that if we are looking ahead over a five year period then current technology and software platforms are certain to evolve. For this reason alone we suggest the emphasis should increasingly be on 'repository services' rather than on the repository as a particular software platform.

As more repositories are implemented there is a realisation of the potential for data to flow between repositories and other systems and for added value services to interplay with repository content.

This perspective was put forward by Cliff Lynch in 2003:

> "*a university-based institutional repository is a set of services that a university offers to the members of its community for the management and dissemination of digital materials created by the institution and its community members. It is most essentially an organizational commitment to the stewardship of these digital materials, including long-term preservation where appropriate, as well as organization and access or distribution. ….. An institutional repository is not simply a fixed set of software and hardware*."[2]

Note that the focus on the services that the repositories provides is very important, and holds true whether the governance of the repository is at a national, agency or institutional level.

4 Role of repositories

4.1 Where we are going – 2010

The authors' overarching vision for 2010 is of a richer scholarly communication environment, based on open access to, and re-use of, scholarly materials. The phrase 'scholarly

[2] Lynch, C., ARL Bimonthly Report 226 <http://www.arl.org/newsltr/226/ir.htm>

communication' is used here in its richest sense to include the life-cycle of information and knowledge from research to learning[3]. While the core meaning of 'open access' is simply that materials are made freely available on the Internet/Web, it is likely that the phrase will also carry with it the notion of exposing supporting metadata about and services on those scholarly materials in order to support the kind of rich infrastructure referred to above. Motivated both by the open access agenda and by the requirement to manage their digital assets effectively, institutions will build managed curation of their scholarly publications, research data and learning objects into their information strategies. The HE and FE community will benefit from a growing number of added value services layered on top of open access materials, such services being offered by both the commercial sector and the education community itself.

Enriched scholarly communication will be supported by repository services operating at a mix of departmental, institutional, regional, national and international levels. Repository services will meet the user requirements of all members of academic institutions, covering teaching and learning materials, scholarly publications, research data, and materials produced by students. As one of this roadmap's contributors says: "*Repositories [will] be demand rather than supply led, and [will] have as their primary aim the fulfilment of researcher, teacher, learner, organisational, and institutional needs*".

It is expected that repositories will continue to focus primarily on serving particular communities, for example subject-based or institutional communities; or be responsible for a particular content type, for example images or learning materials. However, the repositories of the future will be much more interoperable with systems used to support learning and teaching, Virtual/Managed/Personal Learning Environments, assessment systems, ePortfolios, etc., as well as with authoring tools, other repositories, portals and library systems.

In addition to achieving the deposit of a significant proportion of scholarly articles, there will be a expansion in the range of content currently being deposited: more commercially-published research papers, working papers, e-theses, learning objects, primary data, video, film, digitized slides and so on. Increasingly, experimental hardware in research laboratories will be configured to automatically deposit copies of raw experimental data directly into an institutional or departmental repository of some kind. Similarly, desktop tools will be able to 'save' content directly into repositories. Furthermore, there will be widely adopted mechanisms for manually citing and automatically interlinking between this diverse set of resources.

The implication is that by 2010 there will be an extensive network of repositories, both internal and external to institutions, with rich data flows between these repositories and other components in the information landscape. By establishing a network of repositories, the functional components of the information environment will become less distinct. The focus will be on the 'provision of services' rather than on the different 'networked boxes' in which objects reside during the information resource lifecycle.

Repositories will both support and consume other services. They will support aggregation of content (both metadata and full-content) by service providers and will consume services such as content (or metadata) enrichment services. Aggregation services will add value, whether by offering simple search or richer manipulation of data such as interlinking research data and academic papers, visualisation and text and data mining.

There is little consensus on the future role of repositories for preservation, reflecting wider debate in this area. In particular, there are different views as to how far institutions, as opposed to national services, will be responsible for preservation. Some people see long-term digital preservation as an added value service layered onto the network of repositories, provided either by the institution itself or external service providers. On the other hand, some regard the institution as having only a short term responsibility for the curation of research outcomes until these outcomes are formally published or stored in national data centres. From this perspective, the institution's primary responsibility is to give scholars an opportunity to access new material

[3] Lyon, L., Carr, L., Coles, S., Heery, R., Hursthouse, M., Gutteridge, C., Duke, M., Frey, J. and De Roure, D. (2004) eBank UK Linking Research Data, Scholarly Communication and Learning. *In, Semantic Grid Workshop, Global Grid Forum 11, Hawaii, USA, 4-7 July 2004.* <http://eprints.soton.ac.uk/12461/>

before waiting for the publication process, and to access data that would not be otherwise made available. The considerations for data, academic papers and learning material are quite different (see below). This area is further confused by the need for institutions to reconsider their records management and retention policies in the context of the growing body of born-digital information. Whatever the outcome of these discussions, the authors suggest that by 2010 there will be a firmer basis for a national preservation strategy that makes clear who has responsibility for preserving different types of data and who has responsibility for providing open access to resources.

In the area of geographical information systems[4], staff and researchers will be able to discover, locate, access and use geospatial content that is distributed across institutions and other organisations (and that is made available under different licensing regimes) more seamlessly. Access to geospatial data held in repositories will complement data provided through Web services and more traditional content providers. Ideally we will have an integrated and interoperable services layer in UK academia with enabling tools to fully support the academic contribution to the UK Spatial Data Infrastructure.

By 2010, simple metadata will no longer be created 'manually' to the extent that it is now. Techniques such as text and data mining, topic mapping and so on will be used to create metadata and extract information. However, it is still unclear as to who will be responsible for this 'knowledge extraction' and what level of aggregation will be required for it to be effective.

As part of the transition to having a significant proportion of publicly funded research outputs being made available on an open access basis, repositories are likely to become embedded in the publication and peer review process. While it is not yet clear what impact this will have on the business models of traditional journal publishers, it is clear that the academic community will still need peer review to be undertaken in some form. The community will need to find new ways to work with and support publishers as they transition their business models to accommodate the new open access landscape. Furthermore, although open access is usually viewed as a threat to the business models of traditional publishers, it may well be the case that the availability of a significant body of open access material will prove to be the enabler of completely new business models and activities across both the public and private sectors.

Finally, it is worth noting that the institutional business drivers for repositories in 2010 are likely to differ across institutions, just as they do now. As a result, repositories will be established by institutions for a variety of reasons, whether as a showcase for research outputs, to enhance learning outcomes through the sharing and re-use of high quality learning materials, to support RAE submission, to support preservation, as an aid to quality assurance or to provide open access. Furthermore, universities may well adopt different organisational approaches to repositories, incorporating a mix of institutional, departmental, laboratory, learning material repositories and others.

4.2 Where we are now – 2006

4.2.1 Policy/political viewpoint

At a national level, whilst the Department for Education and Skills e-Strategy[5] encourages improved access and availability of e-learning resources and research outputs, the lack of a clear government endorsement of open access has slowed progress towards its adoption. Some specific funding organisations, such as the Wellcome Trust[6], have mandated that outcomes of their funded research must be made available in open access repositories but a

[4] This document gives separate consideration to repositories of geospatial data because of the special nature of the UK licensing situation and the widespread applicability of such data to online services.

[5] Department for Education and Skills e-Strategy "Harnessing technology: transforming learning and children's services', 2005. <http://www.dfes.gov.uk/publications/e-strategy/>

[6] Wellcome Trust position statement in support of open and unrestricted access to published research, 2006. <http://www.wellcome.ac.uk/doc_WTD002766.html>

similar statement from the RCUK has not been forthcoming as yet. With no such mandates from the majority of funding bodies, or from the institutions themselves, motivating the population of repositories will be more difficult.

Not surprisingly, the current publishers of academic journals are suspicious of the open access movement, seeing it as a significant de-stabiliser of their current business models. Given the lack of a central steer and the concerns from the academic publishing industry, policy statements from the key funding bodies have not been finalised, leading to a situation in which there is no clear policy pressure on institutions to get their open access houses in order.

At the same time, there is also some lack of clarity about the relationship between repositories and the network of data centres in the UK in terms of responsibility for preservation and hosting primary data. Similarly, in disciplines where subject-based repositories exist, there is some confusion about whether open access should be achieved through the institutional repository or the subject repository.

4.2.2 Organisational viewpoint

Repository deployment is fragmented, and repositories tend to exist in isolation rather than being embedded into an interoperating network of services. "*We've got bits and pieces but it doesn't operate as a whole and there are big gaps in provision in some areas."* Within institutions, repositories tend not to inter-work with other applications. Nor are they well integrated with other institutional repositories (although there are some examples of innovative workflows, for example between laboratory repository and cross-institutional repository in R4L[7]/eBank[8]). Similarly, institutional repositories do not closely relate to national repositories (data archives).

We are beginning to see the development of effective aggregation services that actively harvest metadata and full-content from repositories (OAIster[9], Google Scholar[10], Scopus[11], etc.) but these are to some extent hampered by the lack of a significant body of content in the available repositories.

The understanding of what a repository is, and the services it offers, is still evolving. Whilst the outline typology presented in the Digital Repositories Review[12] offers a basis for discourse, there is still fuzziness in particular about the nature of an 'institutional repository', what it contains, what services it offers, and how it relates to other repositories within the institution and externally. Arguably, JISC funding is targeted to a great extent on institutional repositories. The pursuit of funding may be exacerbating a process of 'functional creep' in this area. However there appears to be a real variance in what is driving institutions to establish repositories, the selection of the content and the nature of the services they provide.

Often institutions are not clear as to their strategy for establishing repositories. There are real benefits for institutions in effectively managing their digital assets (promoting research outcomes, fulfilling preservation responsibilities, facilitating added value services such as overlay journals, data mining, etc). Such benefits can be assisted by leveraging the open access agenda. Despite this, repositories are not yet fully embedded in institutional strategy and there is perhaps a misplaced confidence that institutions will take on the full range of repository business functions. Interoperability between institutional libraries, repositories, learning management systems and MIS is still rare.

---

[7] R4L: Repository for the Laboratory. <http://r4l.eprints.org/>

[8] eBank UK. <http://www.ukoln.ac.uk/projects/ebank-uk/>

[9] OAIster. <http://oaister.umdl.umich.edu/o/oaister/>

[10] Google Scholar. <http://scholar.google.com/>

[11] Scopus. <http://www.scopus.com/>

[12] Heery, R. and Anderson, S. Digital repositories review. Report to accompany JISC Digital Repositories Programme call, 2005. <http://www.jisc.ac.uk/uploaded_documents/digital-repositories-review-2005.pdf>

4.2.3 Cultural viewpoint

There is a cultural gap between scientists and the library and archives world. This gap is perhaps greater than that between libraries the arts and humanities. The library and archive community does not have experience with 'big computing' and often their ability to tackle significant technical challenges is limited by a lack of resources. There may be a role here for disciplinary informatics specialists or data scientists[13].

Academic institutions are by their nature slow to change and even those with forward looking e-learning and information management strategies are still based on traditional faculty and administrative structures. At the same time, individual researchers and lecturers have no strong motivation to change. The majority of academics do not know what repositories are, nor are they familiar with the issues around new means of dissemination.

The scholarly community needs to encourage a willingness to think differently and to take risks – but this must take place within a framework in which people gain something from their work in terms of their own aims. The 'reward structures' within the academic community also need to recognise open access as a valuable and important part of the profession. The community needs to find ways to encourage academics to share and re-use publications, research data and learning resources as much as possible.

4.3 Milestones - how we get to where we want to be

4.3.1 Policy/political viewpoint

- Ensure that there is a clear open access mandate from the Research Councils and other funding bodies.
- Realise greater national collaboration to develop a common agenda: DTI, HEFCE, JISC, research councils, charities, etc.
- Encourage institutions to define strategy for open access, re-use and preservation of research outcomes.
- Explore national and institutional preservation responsibilities; provide institutions with preservation audit toolkit.

4.3.2 Organisational viewpoint

- Carry out analysis of existing business processes, workflows and dataflows; identify opportunities for innovative inter-working between repositories and between repositories and other applications.
- Clarify the 'business process' and 'business function' aspects of a repository 'reference model'.
- Develop an ecology of repositories (note: this activity is in the Digital Repositories Support team work plan at UKOLN[14]).
- Support institutions in embedding open access and/or repositories into their information strategies.

4.3.3 Cultural viewpoint

- Explore user requirements in greater detail (e.g. the STORE project[15]).

---

[13] Long-lived digital data collections: enabling research and education in the 21st century: report of the National Science Board, National Science Foundation, September 2005. <http://www.nsf.gov/pubs/2005/nsb0540/nsb0540_1.pdf>

[14] JISC Digital Repositories Support, UKOLN. <http://www.ukoln.ac.uk/repositories/digirep/>

- Progress IPR and copyright issues (e.g. TrustDR[16], JISC Legal[17]) in order to develop a licensing framework within which the intellectual property of institutions, individual academics and third-parties is protected while at the same time encouraging the adoption of open access approaches.
- Reach consensus on linking (manually citing) data and academic papers (e.g. start being made in some DRP projects such as CLADDIER[18] and eBank).
- Seek to encourage the involvement of a greater range of academics in the debate about open access.

5 Considerations for different material types

This section discusses some of the issues that are particular to different types of material.

5.1 Academic papers

5.1.1 The vision

As outlined above, the vision for 2010 is that there will be open access to a significant proportion of newly published publicly-funded UK academic papers, primarily through an interoperable network of institutional repositories, subject based repositories, and national services.

Authors will support self-archiving of newly published papers because of:

- Funding body and/or institutional mandate.
- Recognition that reward relies on 'impact'.
- Recognition that impact is significantly enhanced by ensuring open access to scholarly papers.

There are some who believe repositories will play a role in the future publishing infrastructure for open-access non-commercial research materials. In this view, repositories will capture outputs as these are completed by researchers, managing submission to peer review agents or to commercial publishers for 'added value publishing'. There will be an infrastructure to enable acquisition of content with minimum effort by academic authors, and to automate metadata checking and creation. Whether this vision will become reality is not yet clear. However, as indicated above, it is clear that as open access becomes more common, there will be a tipping point beyond which the business models adopted by current academic journal publishers will need to change. The community needs to work with publishers to ensure that such changes do not harm our ability to peer review academic research output.

5.1.2 Where we are now

Many research universities, but not all, have repositories for academic papers. However the growth in usage of these repositories by researchers has been very slow and the number of papers deposited remains small. At a national level, RCUK are still to issue a policy statement on the dissemination of research outputs. There are now statements of support from UUK and the Russell Group, but no real mandate to deposit copies of papers in an institutional repository yet exists.

---

[15] StORe: Source-to-Output Repositories. <http://jiscstore.jot.com/WikiHome>

[16] The TrustDR Project. <http://www.uhi.ac.uk/lis/projects/trustdr/>

[17] JISC Legal Information Service. <http://www.jisclegal.ac.uk/>

[18] CLADDIER Project. <http://claddier.badc.ac.uk/>

The basic technology is in place for ingest and harvesting/access by third parties but seamless technical links between repositories and authoring systems or research information systems do not yet exist.

It is still unclear how far RAE2008 will be facilitated by institutional repositories. Significant copyright issues, technical issues in terms of managing the RAE returns and statistics and cultural and trust issues appear to be hampering the use of open access repositories in the process of submitting returns to the RAE.

The barriers to greater adoption of open access affect each player differently:

- For academic authors: they need to be reassured that the normal system of academic career reward is unaffected. This goes beyond simply providing reassurance that their papers will attract high impact, to the reward value of being published in a top journal (measured differently in different disciplines – not always by impact, but sometimes by rejection rates).
- For publishers: they need reassurance that they still have a viable business model in this new world – one which recognises that research publications are generally destined to be made freely available by their authors. So they need to be primarily in the peer review business, and to be paid viably for that.
- For libraries: they need to be able to rearrange organisational workflows around a new publishing function without detriment to their existing operations (i.e. by having to swap resources to a new area and thereby deprive a still resource-hungry old area).
- For institutions: they need the security of being able to experiment with such new business models without any implied commitment to change their modes of operating unless greater efficiency for their own research management is proved conclusively.

5.1.3 Milestones

- Ensure that open access self-archiving is mandated with all major funding grants.
- Ensure that open access is embedded in the outlook of the Commission and European funding rounds.
- Encourage institutional open access policy commitment in line with above.
- Clearly demonstrate benefits of making papers available on an open access basis in terms of higher profile and more ‘hits’ for author/institutions.
- Citation counts and usage stats need to be unpicked and made to work in an open access environment
- Provide clear guidelines aimed at various stakeholders: institutions, repository managers, depositors. Different guidelines for different material types: scholarly papers, research data, learning objects, etc.
- Ensure that JISC endorses creative commons (and similar) licensing approaches.
- Instigate major national advocacy campaigns.
- Instigate an equivalent of the Dutch Cream of Science project to raise awareness about open access.

5.2 Geospatial data

5.2.1 The vision

The vision for 2010 is for an information environment in which geospatial data will be much better integrated with research publications and data, learning resources and other content through an enhanced technical infrastructure and improved licensing agreements. Licensing of geospatial data collected with public funding will be reformed to make possible re-use by third parties.

5.2.2 Where we are now

The academic community currently faces a number of issues with respect to its use of geospatial data. These are summarised here:

- Licensing issues (especially with respect to licences around derived data). Currently, the high costs of licensing publicly funded geospatial data makes re-use unattainable, even by some commercial organisations. This issue has been highlighted recently by the Guardian 'free our data' campaign[19] [20] which encourages various government agencies to make their data freely available - referring in particular to the Ordnance Survey (mapping data), the UK Hydrographic Office (tidal and naval navigational data), the Highways Agency (traffic data) and the European Centre for Medium Range Weather Forecasting. Note that the JISC GRADE repositories project is investigating the complexities of existing geospatial data licenses and their impact upon derived geospatial data products.
- Poor data management strategies (institutionally and at an individual level) allied to a culture which perpetuates them i.e. data seen as means to an end (publication) rather than an end in itself.
- Heterogeneity in the standards world and confusion over which standards to adopt and lack of interoperability (JISC recommended vs. geospatial information community ISO approved).
- The lack of a 'big stick' to force proper resource custodianship.
- The lack of appropriate 'carrots' and measures geared towards engendering a paradigm shift in resource management.
- The lack of appropriate strategic direction and confusion over who should formulate it.
- Confusion about the role of research council funded data centres vs. institutions.

5.2.3 Milestones

- Ensure that digital geospatial data rights (licensing simplification) and related security issues are addressed.
- Produce clear guidance frameworks in order that the community may more readily appreciate and understand copyright restrictions.
- Generate improved license agreements between commercial data suppliers and academia. This will probably mean data fee increase.
- Ensure that appropriate repository-related training is available to the community.
- Mandate standards and enforce them.

5.3 Learning materials

5.3.1 The vision

The vision for 2010 is for a growing culture of sharing and re-using learning objects, facilitated by a network of repositories at institutional and national levels (e.g. JORUM) and an enhanced technical infrastructure. Furthermore, the licensing of learning materials will be protective of the rights of authors, institutions and third-parties but supportive of an open access approach.

[19] The Guardian, Give us back our crown jewels, 2006. <http://technology.guardian.co.uk/weekly/story/0,,1726229,00.html>

[20] The Guardian, Why a £5m mapping project had to double up on data, 2006. <http://technology.guardian.co.uk/weekly/story/0,,1742097,00.html>

5.3.2 Where we are now

From the perspective of the teaching and learning domain many practitioners have been slow to realise the importance of managing the outputs of their teaching practice. The diversity in how teachers view their practice and the types of resources that they use is huge. There may be some acceptance that "learning objects" belong in "learning object repositories" but learning objects make up a tiny proportion of the outputs and resources generated by this domain. In addition, academic staff traditionally like to control the use of their teaching materials and submitting them to any kind of repository raises all kinds of issues regarding ownership, copyright, quality control, sharing, reciprocity, etc.

Although many teachers are happy to use materials developed by others, there are still barriers to the effective re-use and sharing of online materials. Furthermore, the rapid expansion in the use of elearning in universities presents such a culture change that some teachers are still experiencing difficulty in using elearning resources.

From a policy and political viewpoint, institutions are increasingly becoming aware of the importance of managing their scholarly publications but this awareness has not yet fully translated to the teaching and learning domain.

Finally, from a technical viewpoint, specifications and standards for sharing content are now being implemented by a reasonable range of products, and some practical interoperability is being achieved. Unfortunately, in many cases different approaches are being taken for different types of elearning content. However, there is some awareness that convergence would be desirable (for example the latest version of IMS Question and Test Interoperability specification uses IMS Content Packaging and IEEE LOM metadata).

The key issues related to development of repositories from a teaching and learning perspective have been outlined by Campbell[21].

5.3.3 Milestones

- Institutions should find a way of providing staff with seamless access to the different kinds of online resources that are available.
- Divisions between library and learning and teaching support services need to be overcome.
- Resources and systems need to be put in place that overcome some of the reasons behind the unwillingness on the part of many teachers, especially at HE level, to use materials that they themselves have not developed. For example, teachers should be able to customize resources and embed them in their own materials rather than be forced to take entire courses unmodified.
- Mechanisms that ensure that the time and effort spent discovering suitable resources is minimized need to be developed, for example by integrating resource discovery with other teaching activities, facilitating engagement with other like-minded practitioners and supporting intelligent filtering of resources.
- IPR policies for learning materials developed by academics need to be clarified, recognising the value of the resource to the institution (and so recognising the institution's stake in keeping it available) while not creating barriers to wider sharing.

5.4 Data

5.4.1 The vision

The vision for 2010 is for an information environment in which there is a growing culture of making raw research data available on an open access basis. In many cases this will be done

[21] Campbell, L. M. Repository issues from a teaching and learning perspective. CETIS, 2005. <http://www.jisc.ac.uk/uploaded_documents/repos_issues_cetis_feb05.pdf>

through departmental or institutional repositories, often with direct links to laboratory equipment. The metadata required to access, understand, and manipulate scientific datasets will continue to be largely the preserve of domain-experts. The community's adoption of a common technical infrastructure for repositories will ensure interoperability between all types of repositories, particularly between those holding scholarly publications and those holding research data.

Such an approach will greatly enhance the visibility of the evidence base upon which research is based, encouraging the citation of and re-use of research data and improving the ability of learners to understand the processes on which science is based.

However, the issues raised above regarding preservation of materials in repositories become even more important when one considers research data, since the volume of data forces one to distinguish between repositories, archives, and data centres. There are particular issues as regards curation and preservation of scientific data over time. No single institution is likely to have the appropriate mix of individuals to maintain and migrate for the future all the data and metadata it has produced in the previous 12 months, let alone over the institution's digital lifetime. It is therefore unlikely that departmental or institutional repositories will be the long term home of academic research data for preservation purposes.

5.4.2 Where we are now

Data management within organisations, particularly universities, is not corporately managed. Organisations need to discover what data they hold and document it. Some organisations, e.g. NERC, have designated data centres but others have no organised formal data curation mechanisms, so mixed picture often coloured by funding issues.

Culturally, data repositories have been the property of 'scientists' and here there is some tension between the data and information community. Institutional repositories could fill a gap where there is no data archive, but there needs to be a demonstration of how different repositories treat 'data'.

In policy and political terms, data is now being viewed as an open access candidate. For example, the OECD declaration[22] and working groups are picking up on the repository lead. Technically there are many topics being discussed in this area, including preservation, metadata, persistent identifiers, etc. Many of these issues are currently being addressed in projects. However, there is a need for aggregating the outcomes of these projects for best practice evolution and transition to a production/service environment.

5.4.3 Milestones

- Institutions need to invest in research data repositories.
- Robust acquisition policies and discovery mechanisms need to be developed.
- Agreements need to be reached about unique identifiers and citations for all datasets.
- We need functionality and services that support curation, migration and preservation.
- We need online interaction with research data and links to all related information of all media types.
- We need systems that allow the IPR in research data to be managed properly.
- IT, data and information communities need to work together to support research data repositories and to build on each others skill.
- The community needs to develop mechanisms that foster advocacy, mandates, funding, early adopters and demonstrators with designated responsibilities aggregated at discipline or regional level.

[22] Information Today Inc., 2006. OECD Ministers Support Open Access for Publicly Funded Research Data. <http://www.infotoday.com/newsbreaks/nb040209-2.shtml>

6 Enabling technical infrastructure

6.1 The vision

The vision for 2010 is for 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.

This vision is based on there being more widespread agreement about how the relatively complex objects found in repositories are packaged together and exposed. Such agreement will be instantiated in the software used to deliver repository systems and the other services with which they interact. Although building support for one or more of the current packaging standards into repository software should be relatively straight-forward, software will also need to have some knowledge about the 'complex object models' being used. Without this knowledge, repository software will be able to unbundle a package into its component parts, but it will not understand the relationships between the component parts in order that actions can be performed on them in sensible ways.

In the general case, the issues associated with sharing knowledge about the modelling constructs being used within complex objects are non-trivial. The authors suggest that this is a 'semantic Web' issue that requires significant research work. In specific cases, it may be possible to agree particular 'complex object' models for particular applications (a model for scholarly publications, a model for datasets, a model for lecture objects, etc.). But even if this approach is taken, designers of repository software will need to marry their potentially complex internal data-structures with the externally visible packaging standards according to each of the chosen models, as data flows in and out through the repository APIs (search, harvest, deposit, delete, obtain). It is not clear how easy it will be to do this in an open-ended and flexible way.

Given this complexity, it may be sensible to first develop a very lightweight packaging framework, which can be used consistently across repositories but which is flexible enough to support the more specific packaging requirements of particular domains.

Within the agreed technical infrastructure there will be widespread agreement about how packages and the objects they contain are identified and the mechanisms for creating 'context sensitive' links (e.g. using the OpenURL) between those objects and packages.

Access to institutional and other repositories will be controlled within the same 'single sign-on' access management framework adopted for other internal and external resources.

Finally, the licences under which resources are made available will be more commonly available in a machine-readable form, and therefore more suitable for supporting a DRM-based infrastructure.

Although the conceptual thinking that underpins the technical infrastructure sounds complex, it needs to be instantiated in a relatively simple and intuitive form. This will encourage adoption of the framework by a wide range of developers and service providers, including those creating services outside the academic domain. Content exposed through the infrastructure must be made available in a form that is suitable for use by the 2010 equivalents of Google and Yahoo and in a way that is compatible with the ranking mechanisms that they adopt for 'ordinary' Web sites.

6.2 Where we are now

Two key standards underpin much of the current repository activity. Firstly, the OAI Protocol for Metadata Harvesting is used to support the regular gathering of metadata records from repositories by other service components in the information environment. Secondly, the metadata records exchanged using the protocol are typically based on Dublin Core metadata standard. Unfortunately, it is clear that the current usage of simple Dublin Core metadata and the rather loosely coupled bundles of related objects found in many repositories leads to problems for the consumers of metadata from those systems. As the ePrints UK project found, it is difficult or impossible in many cases to reliably tie identifiers and metadata records to individual 'manifestations' of scholarly material (e.g. the PDF version of a particular scholarly paper), largely because of the widely varying practices across institutions. In general this means that it is often difficult for software robots to move reliably from the harvested metadata record to the full-content of a particular resource. Repositories, particularly those holding scholarly publications, have tended to be developed around relatively simple 'single item' objects. Even where repositories handle multiple versions and/or formats of the same item, there tends to be a single metadata record for the item, linking to the multiple versions/formats.

In the case of learning object repositories it is well understood that much of the content that will be deposited will be in the form of IMS Content Packages (i.e. reasonably tightly-coupled complex bundles of resources). The same is also likely to be true of repositories holding scholarly publications and research data in the fullness of time, where we are likely to see a move towards some form of packaging of 'complex objects'. Consider, for example, a typical 'eprint' (if such a thing exists). Conceptually, an 'eprint' consists of a 'work' and one or more 'manifestations' of that work (a PDF file, a Word document, etc.). Each of these things may also have separate metadata records associated with them. Being able to bundle these separate chunks of content and metadata together in some form, wrapped in a METS or MPEG-21 DID package, will simplify (at least in the long term) the way that these kinds of objects can be deposited, managed and retrieved from repositories.

As indicated above, objects in repositories are currently identified in inconsistent ways across different repositories. For example, in some repositories of scholarly publications the identifier for the 'work' is exposed for use by remote service components. In other cases, the identifier for each 'manifestation' is exposed. This creates a problem not just for services that harvest the metadata and/or full content, but also for those, such as 'appropriate copy' OpenURL link resolvers, that need to create robust linkages into the repository.

The integration of repositories into institutional single-sign-on access control mechanisms is not consistent currently. In some cases, users of institutional repositories must register separately (and are assigned a separate username) for the repository. This is inconsistent with a general trend, both within and across institutions, to move towards single-sign-on.

Finally, the community does not currently adopt a very consistent use of licences for the content in repositories. There is some use of Creative Commons licences, but these are felt to be inappropriate in some contexts (e.g. in the case of Jorum). Furthermore, there is little or no use of automated digital rights management (DRM). As a result, the automatic protection of content in repositories is not yet a reality. While this is not an issue in the context of open access material, it may hinder the take up of repositories in other areas.

6.3 Milestones

- Develop a 'complex object' model (i.e. an agreed way to model arbitrary bundles of objects) in order that the relatively complex objects held in repositories can be dealt with in a more fully automated and interoperable way.
- Develop agreed mechanisms for instantiating 'complex objects' (that conform to the model) in a concrete syntax such as XML.
- Agree mechanisms for identifying 'complex objects' and their component parts. Note: A Dutch national DOI agency is being established to provide DOIs for research papers, learning objects and other content. UK repository projects could benefit from this

experience. Within the UK, the eBank project is piloting use of DOIs for scientific data. The project is registering datasets with the German National Library of Science and Technology (TIB). The authors recommended that collaboration on identifiers for range of resources is taken forward within Digital Repository Programme project cluster.

- Agree mechanisms for creating 'content sensitive' links to 'complex objects' and their component parts (e.g. using the OpenURL standard).
- In the short term, agree mechanisms to use 'qualified' Dublin Core metadata to describe simple (single-item) objects as they are currently held in repositories. For example, agree sets of guidelines for how to use qualified DC to describe scholarly publications, learning objects, research data, etc.
- Agree the machine to machine interfaces (the services) that open access repositories should support in order to ingest and make available content and metadata. Using the language of the eFramework for Education and Research, this means that the community needs to develop a 'repository' reference model, agreeing the protocols and data formats to support at least 'putting', 'getting' and 'deleting' content and metadata in repositories (in single-item and bulk mode)[23].
- Ensure that repository content is well integrated with the large-scale Web search engines (e.g. Google) and that the adopted methods of exposing and interlinking resources does not harm the link-based and other ranking mechanisms adopted by many of them.
- Ensure that repositories are well integrated into institutional and national access management approaches (such as Shibboleth).
- Ensure that content licences are adopted as consistently as possible. Consider setting up a registry of the licences in use across repositories to support and encourage this.
- Work towards DRM solutions that allow software to take decisions based on machine-readable licences.
- Agree mechanisms for manually creating citations (e.g. in scholarly publications) to research data and other resources.
- Develop modular services to be provided by repository software suppliers which can be plugged in to deliver different functions e.g. preservation, RAE outputs, personal profiles (CVs), etc.
- Develop aggregator services that use the features of the technical infrastructure to hide the complexities of the repository landscape and offer a single, seamless view of UK repository content to downstream service components in the information environment.

[23] Powell, A. 2005. A 'service oriented' view of the JISC Information Environment. <http://www.ukoln.ac.uk/distributed-systems/jisc-ie/arch/soa/>

Appendix A

Parameters

The roadmap looks towards a destination that is the UK repositories infrastructure in 2010. It will describe annual milestones between now and then covering:

- The two strands of the Information Environment:
    - discovery to delivery
    - sharing, curation and management
- Technical issues such as metadata and interoperability, federation architecture (including the relation to Shibboleth)
- Legal resources (licences, licence services)
- Policy resources (accepted practice relating to, for example, repository file format policies and retention policies, plus audit and certification – perhaps – to underpin trust in them)
- Barriers and opportunities at the organisational, national, cultural levels
- Principles on which decisions should be made on roles and responsibilities (for example, who collects and quality-assures preservation metadata?)

Scope

The following scope notes define the focus of the document, but in each case it will be important to specify the boundary relations between that which is in scope and that which is out of scope for the roadmap.

- Geographical scope:

    In scope: the UK

    Out of scope: everywhere else

- Object types:

    In scope: research outputs (text, data, other), learning materials

    Out of scope: Administrative records

- Media types:

    In scope: Potentially all (text, image, sound, moving image, simulation, etc)

    Out of scope: None

- Object provenance:

    In scope: objects created, owned and shared by members of the HE/FE community

    Out of scope: objects made available to HE/FE on a commercial basis

Appendix B

Email questionnaire sent to contributors

Subject: JISC repositories roadmap questionnaire

This is a request for you to provide input to a repositories roadmap that we are drafting for the JISC. The roadmap is intended to inform the JISC's planning process for future funding over the next five years or so. The first draft of the roadmap will be presented to the JISC Repository and Preservation Advisory Group in March. Given the tight timescale, the intention is to contact about a dozen people for input at this stage. We are therefore asking for input from one or two people from the perspective of various 'domains' – academic research papers, scientific research data, arts/humanities/social science data, learning materials, GIS, images. If this work is taken forward, we expect that wider consultation will be organised.

We would be grateful for brief replies to the following questions by Friday February 24. Brief replies please!

We are asking you to consider your replies from the perspective of someone with an interest in repositories for [academic research papers][scientific research data][arts/humanities/social science data][learning materials][GIS][images].

1. Where do we want to be in 2010? What are the main business functions that repositories should fulfil in five years time (from the perspective of the domain you represent)?

2. Where are we now as regards fulfilling these functions?
- from an organisational viewpoint?
- from a cultural viewpoint?
- from a policy/political viewpoint?
- from a technical viewpoint?

3. How do we get to where we want to be? What barriers need to be overcome and how?
- from an organisational viewpoint?
- from a cultural viewpoint?
- from a policy/political viewpoint?
- from a technical viewpoint?

Note we have not asked for separate input contrasting institutional as opposed to UK-wide or global services, nor from a subject based repository perspective. Rather, we would ask each person to include consideration of institutional, national, global and subject based requirements as appropriate in their replies.

We are happy for replies to be returned embedded into this mail or as an attachment. We intend to acknowledge all input at the start of the report but will not directly attribute individual replies unless requested to do so. We intend to edit replies into a structured report, adding additional technical and organisational overview.

Thank you for your contribution,

Andy Powell and Rachel Heery