LINKING UK REPOSITORIES:

Technical and organisational models to support user-oriented services across institutional and other digital repositories

SCOPING STUDY REPORT: APPENDIX

Alma Swan (Key Perspectives Ltd)
Chris Awre (University of Hull)

Project partners:

Key Perspectives Ltd
University of Hull
SHERPA, University of Nottingham
School of Electronics & Computer Sciences, University of Southampton

APPENDIX:
TECHNICAL ARCHITECTURE AND INFRASTRUCTURE

Contents

A1 The repository landscape

Repository development has taken place separately in different sectors in recent years, for example the often parallel development of institutional and learning object repositories, and much can be learnt from activity across these sectors. Some of the first open access repository developments focused on the needs of particular, often subject-based and dispersed, communities (e.g., the arXiv[1]), and these needs often informed the boundaries of what the repository would store. Many other repositories have been developed to cater for organisational entities, for example universities (many of the SHERPA e-print repositories[2]). Many repositories have focused on a single item type or format, for example the White Rose repository is specifically for e-prints[3], whilst the BioMed Image Archive repository is image-based[4]: others have decided their repository will hold whatever the institution needs it to take (e.g., the Universities of Cambridge[5] and Edinburgh[6]). Repositories may hold simple items, comprising single files (e.g., a PDF document or a JPEG image), or may hold compound objects that comprise multiple linked items (e.g., a thesis comprising documents, images and datasets). Local and specific needs have underpinned these decisions and led to a rich and varied repository landscape within which end-user services must work.

A1.1 Landscape heterogeneity

This rich and varied development has led to a highly heterogeneous landscape as well, with many different repositories being set up for different purposes and containing different content. There has always be heterogeneity for these reasons and it must be assumed this will continue: interviewees were agreed that this is a major issue for end-user services to manage. Heterogeneity may result in fragmented use of the repositories within the landscape, and can have a large impact on the integrity of federated searching.

Many different ways of overcoming this heterogeneity have been suggested to present a more homogeneous view to end-users of the various repositories available.

- ➢ You can harmonise different repositories at the presentation level using a portal.

[1] arXiv, http://arxiv.org/
[2] Securing a Hybrid Environment for Research Preservation & Access (SHERPA), http://www.sherpa.ac.uk/
[3] White Rose Consortium ePrints Repository, http://eprints.whiterose.ac.uk/
[4] BioMed Image Archive, http://www.brisbio.ac.uk/index.html
[5] DSpace@Cambridge, http://www.dspace.cam.ac.uk/
[6] Edinburgh Research Archive (ERA), http://www.era.lib.ed.ac.uk/index.jsp

- You can harmonise at the fusion level, for example Google aggregates metadata and content from across a very heterogeneous environment before presenting this through its website.
- You can harmonise at the metadata level using shared metadata standards or mapping between records using different metadata formats.
- A mixture of the three – it is unlikely that no single way will suffice entirely.

The use of open standards and protocols is an essential tool in helping to overcome this situation, as are agreements at different levels between repositories and services wishing to make use of what they contain. Bringing two repositories together automatically leads to information loss through the mechanisms that are used to address the differences between them: harmonisation at the bottom, metadata, level will potentially address the problem before it occurs.

Although recognising that heterogeneity can be a problem, it is important to understand how it affects different communities before there is any attempt to iron out all the cracks. Heterogeneity can also be seen as the greatest hope for repository evolution. There is ultimately a need to find ways in which different repositories can work together whilst taking their differences into account.

A1.1.2 Open access

When establishing a repository one of the considerations will be how open or closed the content will be for access. Preservation-oriented repositories are sometimes created as dark archives, where short-term access is a minor function. For many, though, the purpose of the repository is to serve the content to end-users in some way. How controlled should this access be? In other words, to what extent will the metadata and content within them be made available on open or restricted access? Looked at broadly, institutional repositories have been developed to facilitate open access to research outputs, most commonly in the form of e-prints of journal articles: indeed, access without restriction is a vital part of their purpose. The term ‘institutional repository’ has become almost synonymous with this open access approach. Some institutional repositories, though, hold far more than simply e-prints and decisions in theses cases need to be made about what will and won’t be made available. There will be occasions when certain items have specific access requirements and restrictions, for example theses that have confidential materials within them.

The extent to which materials are available on open access in other types of repository often equates to whether identification of the end-user is required. Open access, by definition, makes content available openly for anyone to access as they wish. For some materials, though, it is often important to be aware of who is accessing them: a number of repositories holding learning objects operate in this way. Protecting intellectual property rights in content is often the reason behind a desire to impose these access restrictions, though there is also the benefit of using the knowledge of who your end-users are to inform development of targeted services for them. IPR can be also be stated in open access materials, albeit that not knowing the end-users means the repository has less control over how much rights are adhered to when reaching a wider audience.

The following sections highlight developments in repositories where there is a greater or lesser degree of open access built into the implementation. End-user services will need to understand the repositories they are seeking to interact with in order to make it clear to users what is and what is not available.

A2 Repository types and overview

The following sections outline some of the developments taking place in the realm of repositories that may affect the development of end-user services and/or are considering the issues related to providing end-user services.

A2.1 Institutional repositories

Clifford Lynch from CNI proposed a definition for institutional repositories in 2003 [Lynch, 2003].

"In my view, 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."

As indicated above, institutional repositories have at times become synonymous with the concept of a repository to support open access, having a location to place research outputs so they could be easily organised and accessed. However, they appear to be evolving gradually toward the set of services envisaged. These may be focused around different types of possibly overlapping collections, open access being one. This report follows developments that are predominantly open access in nature, though there are differences in strategy being adopted by different institutions, though, as to the purpose of their 'institutional' repository which may affect the nature of how open access they are in the future.

A2.1.1 The JISC FAIR Programme

Amongst the 14 projects within the FAIR Programme, three specifically focused on the development of open access institutional repositories for the dissemination of e-prints: TARDis at the University of Southampton[7], DAEDALUS at the University of Glasgow[8], and SHERPA at the University of Nottingham. The latter project had a number of partners leading to the implementation of repositories at a further 17 institutions plus the British Library (for non-affiliated researchers). Additionally, the Universities of St Andrews and Edinburgh developed repositories through activities in the HaIRST[9] and Theses Alive![10] projects, respectively. These repository developments highlighted the different approaches to an institutional repository that can be taken. All, by and large, focussed on research outputs and e-prints. However, TARDis took a relatively broad view of what this meant, and has sought to include many different types of research output: SHERPA by contrast has focussed very much on pre-prints and post-prints of journal articles. DAEDALUS did both, but split them up between two different repositories, for both management and service reasons.

A2.1.2 UK development

Development in parallel and since FAIR has shown considerable growth in the number of UK open access repositories. The Registry of Open Access Repositories[11] at the University of Southampton listed 70 UK repositories as of May 2006: these include repositories at institutional, cross-institutional, and departmental level, plus a few specific use cases such as repositories operated by journal publisher and repositories offering access to datasets. Many of the departmental repositories can be considered subject-based repositories as well by virtue of their origin. Notwithstanding the last part of the quote by Lynch, almost all institutional repositories have been created using specific software packages that provide a degree of the functionality and services that were envisaged. In the UK this has usually meant adoption of either the EPrints[12] system from the University of Southampton or the DSpace[13] system from MIT.

A2.1.3 International development

This growth in the development of repositories is not limited to the UK. A conference jointly held by JISC, SURF and CNI in May 2005 produced some valuable data on repository implementation across 13 different countries around the world [van Westrienen, 2005]. Although still low in some, repository implementation in many is on the increase. The authors concluded that this trend was leading institutional repositories to take their place as part of institutional infrastructure. They also concluded that the existence of such a repositories infrastructure would support a layer of end-user services. This dependence on the presence of repositories to support the development of end-user services is notable.

[7] Targeting Academic Research for Deposit and Disclosure (TARDis), http://tardis.eprints.org/
[8] Data-providers for Academic E-content and the Disclosure of Assets for Learning, Understanding and Scholarship (DAEDALUS), http://www.lib.gla.ac.uk/daedalus/index.html
[9] Harvesting Institutional Resources in Scotland Testbed (HaIRST), http://hairst.cdlr.strath.ac.uk/
[10] Theses Alive!, http://www.thesesalive.ac.uk/
[11] Registry of Open Access Repositories (ROAR), http://archives.eprints.org/
[12] EPrints, http://www.eprints.org/
[13] DSpace, http://www.dspace.org/

A2.2 Datasets

Research in many disciplines generates large quantities of data. At an international level there has been wide scale agreement through the OECD that access to data from publicly funded research should be openly available[14]. Many of the UK Research Councils store data from research they fund through national data centres. There are a number of these centralised services available for voluntary or required deposit of data outputs from research activities, for example the Arts & Humanities Data Service (AHDS)[15], the Economic and Social Data Service (ESDS)[16], and the British Atmospheric Data Centre (BADC)[17]. Datasets are also held by other organisations and locally within institutions. At the institutional level datasets have tended to have their own storage facilities set up. These may not be coordinated on an institutional basis, but as required by the departments generating the data. As such, linkage between experimental data and publications that arise from these is rare in the same repository space. Indeed, there is little evidence to support holding both data and publications together, with most work looking to hold these separately. Nonetheless, there is value in being able to connect between data and publications in order to get additional information, see how expressed views have been developed or see how data outputs have been described: for example the links between PubMed and the Entrez Nucleotide, Genome and Protein databases are established end-user services built around centralised repositories at the National Library of Medicine[18].

A2.21 eBank UK

Such a linkage is currently being investigated through a number of projects within the Digital Repositories Programme, where the ability to connect distributed repositories is also being particularly explored. These were inspired by the earlier eBank UK project[19] led by UKOLN and the University of Southampton [Lyon, 2003]. eBank UK investigated the deposit of crystallographic datasets in a modified version of EPrints and the generation of associated metadata (using a Dublin Core derived schema). This metadata can be harvested and aggregated with metadata from crystallographic publications. A unified end-user search service was provided over these complementary sources and links established between them where applicable. The second stage of the project is considering the further metadata description of datasets in order to provide end-user services with richer information to present and work with. An underpinning feature of the work is the desire to streamline the presence of datasets within research and learning & teaching workflows.

[14] Organisation for Economic Co-operation and Development (OECD) Declaration on Access to Research Data From Public Funding, January 30, 2004 http://www.oecd.org/document/0,2340,en_2649_34487_25998799_1_1_1_1,00.html

[15] AHDS, http://www.ahds.ac.uk/

[16] ESDS, http://www.esds.ac.uk/

[17] BADC, http://badc.nerc.ac.uk/home/

[18] PubMed, http://www.pubmed.gov/

[19] eBank UK project, http://www.ukoln.ac.uk/projects/ebank-uk/

Experience from eBank UK is being played into the proposed wider EU DRIVER project, which will be looking to establish a testbed for a future knowledge infrastructure for the European Research Area. It is intended that this be complementary with GEANT2, the existing infrastructure for data, and the project will also look at how these data and knowledge infrastructures can interoperate.

A2.2.2 Digital Repository Programme projects

Additional investigations on bringing datasets into the repository landscape, and thereby opening them up for wider dissemination, are as follows:

- CLADDIER[20] is investigating the workflow involved in allowing scientists to move seamlessly from discovery and location to acquisition of data to deposition back into the data repository of any new data generated. Identification of the different publication and data objects is key to this usage.
- GRADE[21] is scoping a repository for the storage of geospatial datasets to facilitate their exchange and linkage to other geospatial resources
- R4L[22] is developing prototype services and tools to address the issues of working with, disseminating and reporting on experimental data. This will involve an investigation of the metadata required to allow effective aggregation for later use.
- SPECTRa[23] is developing tools to facilitate the deposit of experimental data to prevent this being lost, with a view to it being made available on open access to support further research and also teaching.
- StORe[24] is investigating the different needs of researchers for primary research data and publication repositories, and the possible middleware or shared infrastructure that may be required to facilitate linkage between the two.

A2.2.3 The Grid and repositories

The development of the Grid and cyberinfrastructure in general around the world opens up considerable resources for potential exploitation in repository development and associated end-user services. The Grid offers the opportunity to move storage up to the network level and take advantage of the additional capacity available. Two potential cases where this may be of value are as follows:

[20] Citation, Location, And Deposition in Discipline and Institutional Repositories (CLADDIER), http://claddier.badc.ac.uk/

[21] Scoping a Geospatial Repository for Academic Deposit and Extraction (GRADE), http://edina.ac.uk/projects/grade

[22] Repository for the Laboratory (R4L), http://r4l.eprints.org/

[23] Submission, Preservation and Exposure of Chemistry Teaching and Research Data (SPECTRa), http://www.lib.cam.ac.uk/spectra/

[24] Source-to-Output Repositories (StORe), http://jiscstore.jot.com/WikiHome

- Where digital content is being delivered through an end-user service in its own right (in addition or separately to its metadata) there are associated issues of file size and bandwidth required for quick access. Network caching services may be of value in alleviating the potential bottlenecks that may build up if content is frequently being passed around.
- The developers of the ARC OAI-PMH harvester noticed that as they harvested more metadata the system became unable to easily cope with the load. Their solution has been to distribute the load of where the work takes place in the harvesting process, particularly the aggregation itself, across the Grid. Data providers publish their content. Separate Grid nodes harvest this as required, and then a single federated search module searches across all the harvester nodes.

The EU DILIGENT (DIgital Library Infrastructure on Grid ENabled Technology)[25] is more generically using Grid technologies to enable the development of digital libraries for virtual research groups on demand.

A2.3 Multimedia materials

The development of the Internet and the Web in particular has encouraged the widespread use of images, audio and video materials for both social and work-related purposes. Repositories of these are quite often available on open access, although equally many are highly restricted in their availability depending on copyright and the desire of the content owners to commercially exploit the collections. Collections of images and videos are available through national licensing to the HE and FE academic community, and are available on open access at the point of use through end-user services at the national data centres, for example through the Education Image Gallery or EMOL at EDINA[26]. But it is clear that many such resources exist within the community that are not being shared currently, even if the content creators would like to.

A2.3.1 Community Led Image Collections (CLIC) study

Many image collections remain hidden and unavailable for serving through such end-user services. These are predominantly those collections that have been developed in the academic community for local use. The JISC-funded CLIC study[27] set out to examine these collections and how they might be best enabled for sharing more widely. A survey discovered that all collection owners were open to the idea of sharing what they held, but with provisos: a directory listing of what is available in the collections, or a discovery service based around a catalogue or harvested metadata were welcomed (>80%), but the idea of contributing to a national service was not so welcome (~30%). Collection owners wanted their collections to be discovered, but did not want to lose the branding that came from delivering these through their own website it seemed.

[25] DILIGENT project, http://www.diligentproject.org/
[26] EDINA Sound and Picture Studio, http://www.edina.ac.uk/multimedia/
[27] CLIC study, http://clic.oucs.ox.ac.uk/

The study also found that the infrastructure for enabling end-user services was limited. Many collections are simply Microsoft Access files and metadata is not widely created. This has led to recommendations that a lightweight approach to exposing images will be necessary to ease implementation.

A2.3.2 Video and sound

A user requirements study for a moving pictures and sound portal in 2003[28], like CLIC, also supported the implementation of a solution to facilitate the discovery of these materials. But there was also as much interest in a place where community created materials could be stored and discovered. The interest in discovery did not, though, necessarily match the interest in providing the materials for any central discovery service. Participants wished to see what others were doing without necessarily revealing their own hand! The lightweight approach advocated by CLIC would appear to meet the needs of respondents in this case as well.

A2.4 Learning objects

Many of the materials that come under the previous heading, and indeed any of the materials that can be exposed on open access, can be thought of as potential learning objects. However, whereas the development of repositories for research outputs have been mostly structured toward open access sharing, learning object repositories do not necessarily have this as a major priority. Learning object and research repository developments have taken place in parallel, and the different direction taken may be related to the difference in general perceptions of learning objects and research objects, as described in Table 1.

Learning objects	Research objects
Creators like to exert control	Creators want to let go
End-users invited to use them	End-users invited to read them
Disaggregation and re-purposing supported	Disaggregation and re-purposing rare
Repositories not seen as that useful yet	Repositories seen as valuable tools
Repository software is bespoke	Repository software is open source

Table 1: Comparative features of learning and research objects

A different appreciation of what these objects are for leads to different perceptions of what you will allow others to do. Research objects are designed to disseminate a piece of research, and there is an understanding, maybe implicit, that in making them available on open access they will not be changed in any way. Learning objects, in contrast, are intended for use, not just for reading. There is encouragement to take the learning object apart and re-purpose it,

[28] JISC user requirements study for a moving pictures and sound portal, http://www.jisc.ac.uk/index.cfm?name=project_study_picsounds

possibly in conjunction with something else. This freedom to change something can, though, sit alongside a reluctance to let go at times: feedback from the JISC X4L Programme[29] suggested creators often consider that their construction of the learning object should be the way it is used generally without it being broken up. Many also wish to commercially exploit the objects that have been developed. Learning object repositories, therefore, are not always open access, but have some level of access control, based either on registration, cost or membership of a defined community. This appears to be particularly the case in the UK, for example the requirements to use the JORUM repository, although in the US there are a number of examples of open access learning object repositories, for example iLumina[30] and SMETE[31] (though note both label themselves as digital libraries). Learning objects are also found on the Web or available outside repositories: discovery services such as MERLOT[32] are available for creators to submit metadata about their resource to, exposing this whilst maintaining control of the object itself locally.

A2.4.1 Compound objects

Although many learning objects are simple objects, made up from a single image or document, many are also compound objects, comprising a number of individual components. These compound objects can be organised using the IMS Content Packaging specification, which contains a manifest file describing how the components relate to each other. The ability to disaggregate and re-purpose a compound object supports the re-use of such learning objects. These compound objects require a system that is able to mange them and present them in the correct way. Current VLE systems do not always do this effectively, and this may be affecting uptake on learning objects generally.

A3 User-oriented services

A3.1 Overview of existing activity

There have been a number of initiatives to provide access across repositories in the open access arena. Key developments are reviewed in brief here. Due to the close relationship between open access and the OAI many of the services listed are based on the OAI harvesting model and the use of OAI-PMH. However, there are a number of recent developments using non-OAI

[29] JISC Exchange for Learning Programme, http://www.jisc.ac.uk/programme_x4l.html
[30] iLumina digital library of sharable undergraduate teaching materials for chemistry, biology, physics, mathematics, and computer science, http://www.ilumina-dlib.org/
[31] SMETE Digital Library of teaching and learning materials, http://www.smete.org/smete/
[32] Multimedia Educational Resources for Learning and Online Teaching (MERLOT), http://www.merlot.org/

technologies to increase exposure of open access materials that also require consideration.

Arc – Arc was the first end-user federated search service based on OAI-PMH. It was established in 2000 at Old Dominion University, shortly after the OAI-PMH itself was released. It was designed as a proof-of-concept and, whilst being maintained as an end-user service in its own right that will harvest data providers it is aware of, it is probably better known for the harvester and search software it built for the service, which has been used subsequently by a number of other OAI service providers: these include the ePrints UK development described below. This software supports simple search, advanced search, interactive search, an annotation service, and browse/navigation over the search results. Arc service: http://arc.cs.odu.edu, Arc software: http://sourceforge.net/projects/oaiarc/

ePrints UK – A project funded within the JISC FAIR Programme, ePrints UK sought to provide access to e-prints across UK institutional repositories, although additional resources of interest to the UK academic community have also been added, such as BioMed Central. ePrints UK uses the Arc software. The service continues to run beyond the end of its project, though is not actively maintained. The project itself examined the possibilities of capturing the full-text of e-prints for use in enhancing the subject, citation and author metadata for the e-print using web services. Guidelines on how to catalogue e-prints using Dublin Core and how to link to full-text from within the DC record have sought to provide some level of metadata consistency across the repositories being harvested. ePrints UK service: http://eprints-uk.rdn.ac.uk, ePrints UK project: http://www.rdn.ac.uk/projects/eprints-uk/ and http://www.jisc.ac.uk/fairsynthesis_eprintsuk.html

OAIster – The OAIster project was one of seven projects funded by the Mellon Foundation in 2001 to investigate the development of OAI service providers [Waters, 2001]. Three of these were publicly available, of which OAIster at the University of Michigan is the best known (the others being based at Emory University for the American South project and at University of Illinois for the Digital Gateway to Cultural Heritage Materials). OAIster will harvest all repositories where known, but only keeps those records that point to actual objects, i.e. bibliographic records only are not presented. The inconsistency of metadata that occurs across repositories has led to OAIster putting in place a number of normalisation procedures to facilitate access, e.g., mapping records to a controlled list of material types, which currently require manual effort to maintain. OAIster is seeking to promote interfaces to itself other than its web interface, exposing its harvested metadata to Yahoo! and providing an SRU search interface onto the aggregation. OAIster service: http://oaister.umdl.umich.edu/

DAREnet – The DARE initiative led by SURF in the Netherlands has implemented institutional repositories at all Dutch Universities. DAREnet is the OAI service provider that provides a search service across these. There is a clear emphasis on open access as only metadata for objects where the full content is freely available is included within DAREnet. A subset of DAREnet is the Cream of Science service, which focuses on all the publications produced by the top 207 academics in the Netherlands (of which ~60% can be accessed freely on open access). A third service being developed, the Promise of Science, offers access to doctoral theses. National services are being complemented by smaller, subject-oriented service providers including DARC (Distributed Africana Repositories Community)[33]. Holland is currently the only country in the world so far with this level of countrywide open access end-user services. DAREnet service: http://www.darenet.nl, Cream of Science service: http://www.creamofscience.org/

ARROW – The ARROW (Australian Research Repositories Online to the World) initiative in Australia is based at Monash University in Melbourne, though partners are situated across Australia and include both data providers and service providers. In particular, the National Library of Australia is responsible for the ARROW Discovery Service, which provides OAI service provider functionality across a number of repositories. In order to ensure consistency within this service the NLA has laid down policies on how records should be provided for harvesting. The NLA itself offers a range of OAI collections for harvesting, a facility take up by Google amongst others. Images are harvested separately from within ARROW for use by PictureAustralia and MusicAustralia. ARROW project: http://arrow.edu.au, ARROW Discovery Service: http://search.arrow.edu.au/apps/ArrowUI/

University of Glasgow – Providing end-user services across repositories does not need to be at a national or cross-institutional level. The University of Glasgow operates two distinct repositories for published papers and pre-prints/grey literature/theses, respectively. A local OAI service provider, based on the PKP harvester software, is in development to provide a single point of access across these two repositories for ease of use. The contents of the Glasgow repositories have also been exposed for crawling by Google as an alternative access point. DAEDALUS project: http://www.lib.gla.ac.uk/daedalus/ and http://www.jisc.ac.uk/fairsynthesis_daedalus.html

Citebase – This experimental OAI service provider developed by Tim Brody at the University of Southampton provides a generic search across a number of different repositories. It is notable for the ability to view article citations and carry out citation analysis based on these. Citebase service: http://www.citebase.org/

BASE/Omega/Scirus – The BASE (Bielefeld Academic Search Engine) service at the University of Bielefeld in Germany combines OAI data providers alongside

[33] DARC, http://www.surf.nl/en/projecten/index2.php?oid=106

other sources of information and provides a FAST-based search interface across these. This is an example of OAI-PMH being used in tandem with web search engine technology. Links to Google Scholar from search results are available, furthering this connection and allowing the end-user to pursue their investigations. The Omega metadatabase at the University of Utrecht also aggregates OAI harvested metadata from a variety of both internal and external sources, primarily e-journals, for searching across these. The metadata contained within it is subsequently re-harvested for serving through DAREnet. The Scirus web search engine from Elsevier, using the FAST search engine technology also used by BASE, offers similar cross-resource search capability through web crawling, including access to open access repositories where these are known. BASE: http://www.base-search.net/, Omega: http://omega.library.uu.nl/, Scirus: http://www.scirus.com/

Others – A number of additional OAI service providers are listed on the OAI website at http://www.openarchives.org/service/listproviders.html. Notable within this list are:

- DL-Harvest – a service provider developed as part of the Digital Library for Information Science and Technology (DLIST). It has a subject focus around the Information Sciences that is enabled by targetting only appropriate repositories for harvesting. METALIS is another service provider in this field with a more European repository selection.
- MeIND – this offers a general search service across a range of German repositories, though contents are broken down by content type and subject for browsing through a tree structure as well.
- NCSTRL – a service provider that is focussed on access to technical reports alongside e-prints.

Web search engine developments – As already mentioned there have been recent developments to allow generic web search engines to access open access materials. However, there are many moves afoot to enable this more widely as the value of web search engines as primary access points to open access literature becomes more apparent: an analysis of access to the institutional repository at the University of Southampton in March 2006 revealed that 64% of accesses were via a web search engine rather than through more direct or specific search interfaces[34]. A recent study conducted at the Los Alamos National Laboratory found that of a corpus of 3.3 million records harvested from OAI-PMH repositories, 65% were indexed in Yahoo!, 44% in Google, and 7% in MSN Search: 21% were not indexed in any of these three [McCown, 2006]. Two developments have sought to address and support this exposure:

[34] Email from Les Carr to JISC-REPOSITORIES mailing list, 9th March, http://www.jiscmail.ac.uk/cgi-bin/webadmin?A2=ind0603&L=jisc-repositories&T=0&O=A&X=77C61153BB025CA0DC&Y=c.awre%40hull.ac.uk&P=3300

- DP9 – the team behind the Arc service provider have also developed this tool that site between a repository and a web search engine and enables the search engine to index the repository's metadata. This is particularly valuable where there are no fixed URLs for repository records (though many repository software systems do provide this feature, e.g., DSpace and EPrints). DP9: http://arc.cs.odu.edu:8080/dp9/index.jsp
- Google sitemaps – this facility to register sites for indexing now allows the submission of OAI-PMH 2.0 compliant repositories. This is the route used by the National Library of Australia mentioned above. Google sitemaps: http://www.google.com/webmasters/sitemaps/login

The Mellon-funded mod.OAI development is seeking to reverse the equation by allowing web content on Apache web servers to be exposed using OAI-PMH. This would allow a request to web servers for all files added or changed since a specified date, for example, or a request for all files of a certain MIME-type. mod_oai: http://www.modoai.org/

A review of new open access search developments by Peter Suber during November 2005[35] revealed a number of additional developments in this field, covering both OAI and web search engine approaches:

- OJAX – an AJAX-powered metasearch engine for OAI-compliant repositories from University College Dublin School of Information and Library Studies. OJAX: http://ojax.sourceforge.net/
- Chmoogle – an open access web search engine for chemistry. This is a free service, though commercially produced and available for licensing within institutions for local use. As well as textual searches it allows chemical structures to be drawn and searched. Chmoogle: http://www.chmoogle.com/
- Science.gov – a portal and web search engine for open access outputs from US Government agencies. Science.gov: http://www.science.gov/
- Toolbars for browsers – the NCBI has an NCBI Search Toolbar for Firefox and IE that allows searches across all its databases. OAses is a toolbar for IE that searches across open access services (OAIster) and repositories (DOAJ)[36] as well as web search engines. NCBI: http://www.nlm.nih.gov/pubs/techbull/nd05/nd05_toolbar.html, OAses: http://psyplexus.com/oases/
- Creative Commons search engines – the ability to search according to Creative Commons licence allows end-users to immediately find materials they know they can use and the terms under which this is possible. The Creative Commons website[37] has provided this for some time, but both Google and Yahoo! now also provide this capability.

[35] SPARC Open Access Newsletter, December 2005, http://www.earlham.edu/~peters/fos/newsletter/12-02-05.htm#topstories

[36] Directory of Open Access Journals, http://www.doaj.org/

[37] Creative Commons, http://creativecommons.org/

It is beneficial for repositories to have some awareness and agreement with web search engines where possible. Currently the relationship is often undefined because no permission was sought on either side. This can cause difficulties with withdrawn items where the search engine still provides a cached copy.

RSS – There are two ways in which RSS can be used to facilitate interaction across repositories: individual repositories can provide RSS feeds for a separate aggregator to collect for serving to the end-users; or an OAI aggregation can provide RSS feeds of its own based on its aggregated metadata for serving in a similar way. Such feeds will require filtering when content is being deposited regularly, as otherwise feeds will be become difficult to interpret and use. However, RSS does allow end-users to decide on the level and breadth of the aggregation and information they need. Examples include:

- The University of Glasgow offers an RSS feed of new deposits in its published papers repository[38].
- The IMesh project at UKOLN developed a module to allow records exposed using OAI-PMH can be subsequently exposed using RSS [Duke, 2003].

Although not using RSS, the ARROW Discovery Service provides email alerts of new content based on selected subject terms. Both technical approaches are providing selective dissemination of information (SDI), an established approach and service that has proven its worth over the years as a value-added supplement to searching.

A3.2 Planning user-oriented services

Griff Richards from Simon Fraser University in Canada reacted strongly when asked about the need for end-user services across repositories for this study[39]:

> "If services don't build around a repository, then that repository is doomed. If the architecture of a repository prohibits spontaneous development of services by third parties, then that repository is doomed. A healthy suite of services shows that content is alive. If it doesn't interoperate it simply becomes a knowledge cache rather than a usable collection."

The services described in the previous section have all sought to prevent repositories becoming simply knowledge caches. They have had to address issues that underpin the planning development of many such services, not just those onto open access materials, and in this section the range of issues that require consideration in planning services are addressed. In translating these

[38] University of Glasgow Library RSS feeds, http://www.lib.gla.ac.uk/rss/
[39] Griff Richards, personal email communication, March 2006

issues into reality it can be beneficial, as suggested by Larry Lannom from CNRI, to consider superman models and not rely on physical paradigms[40].

A3.2.1 Delivery options for end-user service

Aggregators and repositories need to make decisions about how they will make themselves available to end-users and the interfaces they will need to support this. Delivery through a local web interface brings branding, but also requires the end-users to locate the appropriate website: it asks the end-user to come to the service. Developing web interfaces also adds to the myriad of web interfaces that end-users already have to deal with – adding an e-prints service to the information landscape, for example, may simply add to end-user confusion and reduce usage. Being part of a wider service and fitting within the user workflow increases visibility – if a repository or repositories sit within a wider digital library, for instance, it has its profile raised and there is greater incentive to develop services onto it or them. Once visibility is achieved it is still important to ensure that search or other options made available mean something to the end-user, so it is clear what the service is offering.

Developing machine interfaces as alternatives provides additional routes into the content and, although branding may be less obvious, can lead to a far greater chance of content being discovered and used: such interfaces take the service to the end-user where they happen to be in their web environment. This occurs for those repositories that expose their content to web search engines. Provision of an SRW/U interface, such as that provided by OAIster[41], allows inclusion of open access repository metadata through metasearch services and other routes for structured searching: an SRU target can be used by the d+ tool[42], for example, allowing presentation of searching within an institutional portal or VLE. Inclusion within environments such as an institutional portal may benefit from development of a portlet that allows such embedding cross-platform using WSRP[43]. In making decisions about which delivery routes to adopt there will be quicker wins than others depending on the amount of development effort and support involved. There are no hard and fast rules to which course of action to take, but considering where the intended end-user audience is and working towards exposing services and content there will ensure maximum usage.

A3.2.2 Workflow

The workflow of interaction with repositories in general requires attention when planning end-user services. In the open access arena it is particularly relevant to identify and fit end-user services across repositories into the researcher workflow. Where these repositories are physically located is largely irrelevant, so long as access to them and/or across them can be slotted into the researcher's

[40] Larry Lannom, personal communication, March 2006

[41] OAIster SRU configuration, http://oaister.umdl.umich.edu/o/oaister/sru.html

[42] d+, http://www.jisc.ac.uk/index.cfm?name=dplus

[43] Web Services for Remote Portlets (WSRP), http://www.oasis-open.org/committees/tc_home.php?wg_abbrev=wsrp

wider environment. Establishing such workflows requires QA of the steps involved to ensure that the different constituent parts of the workflow actually do flow (this is particularly relevant where the workflow is underpinned by an SOA approach), but this can provide long-term benefits in easing end-user interaction with the information they need. A particular issue to address when considering workflows is where the workflow stops. User workflows include use of information, though many information provider models stop at the point of presentation to the end-user and do not address use. Establishing possible usage scenarios for information delivered through end-user services can help with structuring how that information is presented in the first place.

A3.2.3 Hiding vs. surfacing repositories

A further step in providing end-user services is to decide whether to make the repositories obvious to the end-user or hide them so the end-user doesn't realise what they are interacting with. Each approach can be of benefit and ease the end-user's path to the information: technically there are benefits in structuring the relationship so that the links between repositories and end-user services are agnostic, possibly via an intermediary repository services layer. In other cases there is benefit in the end-user service acting as simply a thin-layer presentation agent for repositories and passing the end-user from the service to the repository itself at some point in the information chain. This can lead to the provision of greater functionality, richer information and/or access to the full content where the end-user service cannot make this available directly itself.

It is important not to be too affected in this decision by what available software can provide. Although this is a consideration, and particularly so where institutions are needing to make decisions to proceed, it is useful to consider what services can best be managed at the institutional level and which are best provided at the network level (i.e., across institutions). Commercial services onto many information sources work at this network level and these are widely used because of the value this breadth brings. A consideration of which services will work best at the network level will help bring value where it can best be provided. From the end-user's perspective the location of the repositories and end-user services is irrelevant. They are likely to wish to unify the information available to them in one place wherever it originates from.

A3.2.4 To aggregate or distribute

In considering the range of service providers available, as reviewed above, the level of aggregation, and number of aggregators, is increasing. This is the case in both the open access and particularly the commercial sector. Many of these are focused on collecting together everything they wish to offer for searching rather than follow a distributed access model. Conversely, library portal products are proving popular additions to the portfolio of services a library offers. Scalability of what can be searched is one issue to consider: distributed searching does appear to have limitations in this area that aggregations can overcome. A key factor in deciding which approach is most suitable is the flexibility possible to meet user requirements.

A3.2.5 Authentication requirements

Although open access repositories will predominantly contain materials that are freely available to all, there will always remain the need to have in place some level of authentication and authorisation for those materials that may have some restriction on their access, even if this is for a limited period of time. This is particularly the case for theses, which can contain sensitive materials for commercial, IPR or even political reasons. Some level of filtering will thus be necessary, enabled by the repository software itself (for example the Fedora XACML policies[44]) or placing limitations in what can be harvested (for example using OAI sets). The distinction between what is and what is not made available may be the separation of metadata and content: the ARROW Discovery Service makes metadata available for free, though it is accepted that access to the full content be have some level of restriction placed upon it.

A3.3 Functionality within end-user services

There are many areas of functionality that can be provided within end-user services. A range of those considered of value by interviewees for this study is listed in Table 2 below and each subsequently explored in greater detail.

Search	Browse	Harvest/Aggregation
Commenting/Annotation	Edit/Update	Output
Push/Alerting	Preservation	Deposit/Ingest
Linking	Text mining	Conversion
Obtain	Request	Rendering

Table 2: Areas of functionality for possible inclusion in end-user services

A3.3.1 General

Evaluations at SCRAN revealed that the functionality desired by end-users includes search, browse, aggregation, commenting, editing, and output. The workflow progression in this list is of interest, revealing end-user consideration of their own workflow and how they would like information resources to fit into this. In the always available world of the World Wide Web, end-users not surprisingly also expect 24/7 access. In seeking to provide this it is important to focus on deliverable solutions without overreaching what can be achieved, and focus on designing them from a user's perspective.

[44] Fedora authorization with XACML policy enforcement, http://www.fedora.info/download/2.1b/userdocs/server/security/AuthorizationXACML.htm

A3.3.2 Deposit/Ingest

Analysis of gaps in repository functionality at the JISC CETIS Conference in November 2005[45] concluded that there was a lack of standards for how content is deposited within repositories. Deposit is predominantly an area of functionality that is focussed on individual repositories, though the ability to deposit across repositories may address the issue of making content available through different routes, e.g., through institutional and subject repositories. The JISC Digital Repositories Programme Support Team is addressing this issue[46] and it was also discussed at a Mellon meeting in April 2006 entitled "Augmenting interoperability across scholarly repositories"[47].

At the heart of the problem are the different deposit technologies and procedures used by different repository software packages. Deposit is very much embedded within the software, though a standard API would allow it to be abstracted from the repository as a separate service. SRW Update[48] has attempted to provide this, and DSpace has had some success with the use of WebDAV[49], but neither has so far offered the equivalent of a standard deposit API. Combining deposit with other functionality, such as alerting on ingest, would add value to the process.

A3.3.3 Harvest/Aggregation

The range of existing services that facilitate aggregation of repository metadata/content has already been described. The aggregated information can be subsequently exposed for searching and/or browsing and/or further harvesting elsewhere. End-users will most normally access the aggregated information through some other means, and would not initiate the aggregation in the first place. In the case of OAI-PMH harvesting takes place behind the scenes and the ensuing aggregation exposed to end-users and other services. Search engines crawl the web and the end-user interacts with their aggregated cache through search and browse most commonly. RSS aggregation is under the control of the end-user more through selection of the appropriate RSS feeds. Even here, though, the aggregation process sits behind the point of access. Aggregation is not, thus, necessarily an area of functionality for end-user services themselves, albeit that the end-user service and aggregator may be closely linked.

A3.3.4 Search and browse

Search in itself is the most important functionality that end-user services can offer. It allows the end-user to interact with the repositories on their terms, and

[45] JISC CETIS Conference Repositories session, Heriot-Watt University, November 2005, http://www.e-framework.org/events/conference/programme/repositories
[46] JISC Digital Repositories Programme Support Team deposit API activity, http://www.ukoln.ac.uk/repositories/digirep/index/Deposit_API
[47] Augmenting interoperability across scholarly repositories, http://msc.mellon.org/Meetings/Interop/
[48] SRW Update, http://srw.cheshire3.org/docs/update/
[49] WebDAV, http://www.webdav.org/

matches wider expectations about how information is accessed via the Web. What type of search is offered can vary, though, and there are a number of options. Subject searching is most valuable where subject-based communities are being targeted, though it is acknowledged that providing the appropriate metadata to enable such subject searching is difficult. Nevertheless, it is a goal that would benefit from further attention to help meet a clear user requirement. Author searching is also important, though can be considered a specific form of subject searching.

Other options for searching include: searching by content type, as with the EThOS project[50] for ETDs and the OAIster content type access (though this requires a high level of normalisation); searching by data format, which will be of particular importance when accessing multimedia objects; citation searching, like that provided by the Citebase service; and licence searching, as provided by the Creative Commons search engines listed earlier. The ability to search within existing search results, possibly by applying a different search type, and thus a different filter, adds value to the search process.

Searching can be of benefit to end-users who do know what they are looking for, by using known terms, and also potentially for end-users who do not – web search engines are frequently used as starting points in searching through entry of terms to 'see what comes up'. Browsing offers an alternative route into available content, guiding end-users without requiring them to know what they are looking for beforehand. Initial searching can be backed up using browse. Implementation of browse can be complex, as it requires a structure to the metadata/content within the repository and/or aggregator that can be browsed through. Nevertheless, this effort can provide ways into the content that would not have been available otherwise. Browsing is currently underrepresented functionality when considered alongside search, though having both search and browse can help cater for different user groups and needs. End-users don't have to be aware of which one they are using, with options provided to suit the workflow at that point.

Searching and browsing allow the end-user to explore what repositories make available. There has been a lot of interest in recent years in visualisation as an alternative area of explorative functionality. Grokker[51] is an example of how Google searches can be displayed visually, and the ability to search by chemical structure using Chmoogle is of particular interest to the field of chemistry. This targeted approach may work best for visualisation, as take-up of visual alternatives has been limited so far in general. An exception to this is browsing images, which has proved a popular means of providing access to these due to their visual nature.

[50] EThOS project, http://www.ethos.ac.uk/
[51] Grokker, http://www.grokker.com/

A3.3.5 Rendering, output and conversion

Providing search and browse functionality is only one half of a discovery equation. Repositories have to be able to return an appropriate derivative that can be rendered by the end-user service in a useful fashion. Repositories and aggregators need to consider how they will respond to search and browse requests, whilst end-user services need to consider how they will deal with repository outputs. Does the service only display metadata and links to full content for downloading? Or does it enable content to be displayed or converted within the service for possible analysis, comment, and subsequent use? Outputting bibliographic data in different citation styles may be one service made available.

The aDORe architecture developed at the Los Alamos National Laboratory in the US provides for a Pathways InterDisseminator[52] that presents information on what disseminations are available from a repository. End-users select from the available options and the module enables conversion on-the-fly for delivery. A particular issue arises with compound objects, where the end-user service may need to ingest and render different sub-components in the appropriate order for the overall object (a task for which the IMS CP manifest file is designed). In such circumstances aggregating the different components together first may help with this process: alternatively a relationship service could keep track of the links between distributed sub-components to facilitate subsequent actions upon them. In both cases the ability and desirability of doing this will be affected by the reason and purpose for having the sub-components distributed originally.

A3.3.6 Push/Alerting

The ability to push information out to end-users from a repository or aggregator allows the source of the information to exert a level of control and structure to what is being exposed. End-users still make the decision about whether or not to accept the pushed information, and exert their own sense of control through this decision, but on doing so they get what the repository has packaged for them. Pushing information is thus of benefit to repositories and aggregators in defining what they expose. It is also beneficial to end-users in receiving alerts from known information sources that are of value. Receiving alerts is very much a matter of preference for how information is delivered, though has proved very popular at DSpace sites as a second priority behind search and browse. RSS is commonly used to provide alerts. Technically this is not a push technology, as the end-user initiates the alert: it has, nevertheless been commonly perceived at the end-user level as enabling the pushing out of information. The CORDRA model[53] talks of repositories pushing their information to the appropriate registry

[52] Van de Sompel, H. Lessons in Cross-Repository Interoperability learned from the aDORe effort. Presentation at the 4th CERN Workshop on Innovations in Scholarly Communication (OAI4), 20th-22nd October 2005, http://oai4.web.cern.ch/OAI4/

[53] CORDRA, http://cordra.net/

for exposure, though whether push or pull, via harvest or similar, works best in this environment remains to be established.

Push requires effort by the repository to implement, though there can be added value from putting this effort in place. RSS feeds can be embedded in departmental webpages, for example, raising the profile of both the information and the repository within an institution: using RSS for this purpose harks back to the days when SDI services were a staple library offering. Such feeds can be general or, more appropriately, subject-based to meet the needs of the target audience where there is relevant metadata available.

A3.3.7 Request

If it is not possible to render or deliver the desired content once it has been discovered it may be useful to offer end-users the opportunity to request a copy of the content by requesting it from elsewhere. In an open access environment this may not be necessary as the full content would hopefully be available freely. However, there will be occasions where this is not the case. Stevan Harnad has promoted the concept of a request to the author for a copy of an e-print where restrictions are in place and this has recently been implemented in both EPrints[54] and DSpace[55]. Alternatively a request to a repository manager may be required where an end-user is situated at another institution. These requests are akin to traditional inter-library loan requests, albeit more oriented to a repository environment.

A3.3.8 Obtain/Linking

Linking is an area of functionality that facilitates the location of full content and will mainly take place after searching and browsing and from a set of results. It may mean linking from discovered metadata to the full content in the repository: both the ePrints UK service and ARROW Discovery Service have taken this approach. It may also mean linking from metadata to full content available elsewhere as well, for example linking from metadata to full content available through a library subscription or on an alternative website. In an open access environment it will be valuable to know what is available wherever it is.

These links can be made directly, or they can be made in a context-sensitive fashion using OpenURL[56], which can also facilitate further discovery through other resources. Investigations into the creation of OpenURLs for repository content at the University of Southampton have shown that this is not always easy, as the granularity of information required is not always available. However, the University of Glasgow has created OpenURL links from its library catalogue into a local repository, and CERN use OpenURLs to link out from a CDSware repository to other sources. Aggregators have also made use of OpenURL:

[54] EPrints "Request eprint" button, http://www.eprints.org/news/features/request_button.php
[55] DSpace Request Copy Add-on documentation, http://wiki.dspace.org/RequestCopy
[56] NISO OpenURL 1.0 Z39.88-2004 standard, http://www.niso.org/standards/standard_detail.cfm?std_id=783

OAIster, whilst not being fully compliant, provides a technique to allow searches of OAIster initiated from an OPenURL; the METALIS service provider generates OpenURLs based on its aggregated data for linking to other services. The OpenURL 1.0 Z39.88-2004 standard offers greater flexibility than the previous 0.1 version, which was limited to bibliographic materials, and offers a number of possibilities for linking between different content types that are as yet uninvestigated.

A3.3.9 Commenting/Annotation/Tagging and Edit/Update

In these days of Amazon book reviews and ratings etc. there is increased demand for similar functionality elsewhere. Offering similar functionality on top of repositories would allow communities to exchange information and inform each other of views and opinions. It may not be feasible to provide such functionality at the individual repository level, and annotation is more likely a service that can be provided at the network level or at the level of an aggregation. The RESULTs project in the JISC 5/99 Learning & Teaching Programme promoted this type of functionality based on an end-user generated aggregation of links to teaching resources. Tagging of resources with self-selected keywords and developing a folksonomy approach is another means through which end-user generated content can be appended to existing resources.

Appending information to resources is one level of content creation. Editing or updating is another, but requires both greater levels of authentication for security and an audit trail to track the different versions generated by subsequent edits. This level of functionality is more appropriate at the individual repository level where these necessary permissions can be put in place.

A3.3.10 Text mining

Text mining allows deep exploitation of information resources by applying analysis techniques that allow concepts and interpretations to be derived from available data. The greater the granularity of the information available and the greater the body of information available the greater the level of exploitation possible: hence text mining will work better at an aggregated level. The National Text Mining Centre at the University of Manchester[57] has used abstracts in the past as they have been easier to get hold of. In an open access environment there is much potential for exploiting the availability of metadata and content for analysis. It may also be possible to apply text mining across repositories directly, though this would have a greater noise factor as no filtering of the content would have been possible.

It is important in text mining to establish the copyright ownership of the extracted information, as this can be regarded as new information. This may still apply where the original documents were available on open access, though this would need to be tested in practice.

[57] National Centre for Text Mining (NaCTeM), http://www.nactem.ac.uk/

A3.3.11 Preservation

Preservation is not a functionality itself per se. It is a set of functionalities that have as their end aim the preservation of the materials concerned. The individual functionalities around preservation are related to the tasks required to achieve meet the preservation requirements. In the Open Archival Information System (OAIS) reference model[58] there is no search and retrieval, but repositories can accept requests for DIPs (Dissemination Information Packages). This is related to moving these around between repositories as part of the preservation process. This exchange of DIPs could take part via push or pull techniques. Preservation functionality is again more suited at the individual repository level, though there is scope for a third party coordinating the move of DIPs between repositories

A3.3.12 Other

A range of other functionalities can be considered within a repository environment beyond those already described.

- Where end-users are managers as opposed to researchers, students, etc. the ability to gather statistical information on repository or aggregator usage is valuable. The number of citations received by items or the number of downloads of full content items are of particular reference to individual repositories, whilst aggregators may wish to know the number of click-throughs to underlying repositories or the technical routes end-users are using to access the aggregated information. Both EPrints and DSpace have statistics functionality add-ons.
- In the e-science world queries can be placed against remote datasets. These are often the result of collaboration through virtual organisations within or across institutions. Queries are submitted and the job logged at the appropriate Grid resource. Results are returned at some future point in time and can be output in different environments and ways as required for open access or other dissemination and analysis.

A3.4 Levels of end-user service provision

Many different levels of end-user service provision are possible. Services can range from the individual/personal through to global. These different levels do not necessarily require different services, but may equate to the provision of different views across repositories to meet user needs. It is as yet unclear which level or levels are best suited to providing end-user services across open access repositories, though there is increasing experience of how certain levels are working.

[58] OAIS, http://nssdc.gsfc.nasa.gov/nost/isoas/

Developing end-user services onto repositories is affected by the granularity at which the repositories have been established. Repositories are most often established at institutional and subject levels. Initial views at this level are therefore relatively simple to develop and can meet a range of different needs, serving different user groups and purposes such as the RAE. From these repository starting points views can be established onto repository content in a more granular fashion, for example providing views of an individual or department's contributions to the repository. Work at MIT has sought to enable academics to deposit content into the local repository and then extract the metadata to provide a view onto this through their personal or departmental webpage[59]. Personal views onto the repository through interface design, such as that carried out at the University of Rochester [Foster, 2005], can also provide this type of service, and RSS can provide similar focussed views.

Broader views can be established across more than repository through aggregation. ePrints UK has sought to provide a national view onto e-prints within the UK, for example, whilst IRIScotland[60] is seeking to do the same for Scotland – a subset within the overall UK picture. DAREnet has done the same for the Netherlands. The three of these have taken a general line to aggregation, providing a view across all repositories available without any filtering. A subset of DAREnet has also been configured to provide a separate view across Dutch repositories: the Cream of Science end-user service focuses in on content produced by the leading scientists in Holland. National initiatives offer useful organisational structures for collecting and aggregating repository contents.

This national, or regional, view can have its benefits in discovering what is being done where and in promoting the research being carried out in any one country – factors behind both DAREnet and IRIScotland. In the world of open access, though, geographic boundaries can have their limitations. The global research community often looks to where work is taking place regardless of boundaries, and wider aggregations and end-user services are required to make this available. OAIster and ARC take a global view and this may explain their wide awareness and use: it is notable that the other OAI service provider projects funded by the Mellon Foundation are not so widely known, possibly due to the more focussed views across particular repositories. This is not to denigrate their usefulness and value to the communities they served: it is the global view that is of benefit, not necessarily the global audience. Google also takes a global view and this has proved extremely popular. Staring with a global view and then filtering down as required offers the most flexible level of end-user service, though the level of filtering will be dependent on the granularity of metadata.

Flexibility is key to providing end-user services where they are needed. The ability to dynamically bind together institutional repositories offers the chance for

[59] DSpace's Lightweight Network Interface is being used to support this, http://wiki.dspace.org/LightweightNetworkInterface
[60] IRIScotland project, http://www.iriscotland.lib.ed.ac.uk/index.html

end-users build their own aggregations, though this has implications for the possible technologies behind the aggregation: RSS can be easily combined in this way, though OAI-PMH requires greater configuration. End-user services are likely to succeed where the end-user needs to interact with content – hence, addressing the question of where different user groups need access to open access materials will provide information on which end-user services need development. Setting service criteria when establishing end-user services can help with later evaluation of their worth and benefit.

In many cases the interaction required will be at a subject level and, notwithstanding the difficulties inherent in providing this level of end-user service, the demand for subject-based end-user services is high. From a teaching perspective, lecturers think along their subject-lines, and therefore will favour subject-based views of available content. It may be necessary to mask institutional and geographically bounded repositories and aggregations behind subject interfaces to encourage take-up, though this has large implications for the level and detail of metadata required to enable this. This approach was first considered by ePrints UK, which sought to enhance the subject metadata of e-prints so subject-based views could be served through the hubs of the RDN. Although tested in beta the technology was not mature enough at the time to take this further. However, the approach was validated and still applies as a valuable route to follow if solutions to subject classification can be found.

Where there is some control and/or agreement between repositories, and where a common subject classification scheme is available such as in the NEREUS consortium[61], there is scope for providing views based on the subject, in this case the field of economics. OAI sets offer a possible route to classifying repository contents by subject, though this would require widespread adoption and application of both sets and the subject terms agreed. The free text approach followed by Google in crawling everything can be hit and miss when searching, but has proved a valuable starting point. Google Base offers the chance to add structured description to records uploaded to this ‘repository’. This adopts the tagging or folksonomy approach, which is worth exploring further to see if subject-based views cannot take advantage of this.

Norbert Lossau at Bielefeld University in Germany has developed a classification of service levels around repositories. Although designed for e-science they highlight the levels at which academics working in an open access environment may wish to interface with repositories. They could potentially also be applied to aggregations.

- Information gaining/discovery/search/access/navigation
- Information management (personal/organisational, open/closed)
- Information handling (annotation, manipulation)

[61] NEREUS consortium, http://www.nereus4economics.info/index.html

- Scholarly communication (a continuous process through the creation of a piece of scholarly work)
- Publication (the final endpoint)

Services at all different levels can also be made more granular through the application of context-sensitive linking to repository contents. Where the end-user service knows who you are it can apply filters, using the OpenURL ContextObject as a possible technical implementation, that determine whether you can access content, or possibly which version or level of content is available [Blanchi, 2006]. This process is not unlike providing different responses to similar questions at library enquiry desks when approached by a professor and a 1st year undergraduate.

A4 Metadata

The services described in the previous section all use metadata to a greater or lesser degree. Even Google derives metadata from the full text it harvests to assist in the services it offers. Consideration of services cannot take place without an examination of the influence of metadata.

A4.1 Metadata standards

The range of metadata standards available really does epitomise the adage that the wonderful thing about standards is there are so many to choose from. Each has its specific purpose and origin, and these differentiate them. In deciding which one(s) to adopt it is important to bear in mind the balance in requirements between the use of the metadata for organising the content and use of the metadata to allow access to it and the content it is describing. There is also the differentiation between descriptive metadata, technical metadata, administrative metadata, rights metadata, and preservation metadata, each of which has different standards to cater for it. Table 3 summarises some of the many, mainly descriptive, metadata standards available at this time to give a flavour of this range. Notwithstanding this wide range the standards that exist are not enough to allow full and proper description of all content (for example there are limited standards currently to describe datasets, notwithstanding the efforts behind ISO19115 for geospatial data). But there is much experience from different sectors now available to draw on.

Metadata standard	Primary purpose	Notes
Dublin Core[62] (can be simple or qualified)	Supports interoperable search and discovery, and also simple description of multiple content types	Lowest common denominator approach allows interoperability across resources albeit with potential loss of information if mapped from richer metadata.
MARC[63]	Description of bibliographic resources	Designed for physical item description and can struggle to deal with describing digital resources. Also primarily intended for record interchange rather than access.
MODS[64]	Description of bibliographic resources	An adaptation and extension of MARC to allow for a combination of formal MARC tags and more flexible description based on XML.
VRA Core[65]	Description of works of visual culture and images associated with them.	The standard is intended as a starting point to assist in describing visual collections, but is not a complete solution.
MIX[66]	Technical information for still images	Currently undergoing standardisation to assist with the management of digital image collections.
EAD[67]	Description of archival finding aids	Designed to allow the description of inventories, indexes and related tools created by archives and museums
LOM[68]	Description and contextualisation of learning objects	A detailed standard that encompasses descriptive and administrative information about the objects. A UK version, UK LOM Core[69], has been derived from this.
TEI[70]	Representation of electronic texts	Used to assist with the use of electronic texts in research and teaching. Looks to provide a full representation of the document rather than just metadata.
CDWA[71]	Description of core records for works of art and material culture	Extensive set of categories to help describe works of art. Also has CDWA-Lite for interoperability purposes.

[62] Dublin Core, http://www.dublincore.org/

[63] MARC, http://www.loc.gov/marc/

[64] Metadata Object Description Schema (MODS), http://www.loc.gov/standards/mods/

[65] VRA (Visual Resources Association) Core version 3.0, http://www.vraweb.org/vracore3.htm

[66] NISO Metadata for Images in XML (MIX), http://www.loc.gov/standards/mix/

[67] Encoded Archival Description (EAD), http://www.loc.gov/ead/

[68] IEEE Learning Object Metadata (LOM), http://ltsc.ieee.org/wg12/

[69] UK LOM Core, http://www.cetis.ac.uk/profiles/uklomcore

[70] Text Encoding Initiative, http://www.tei-c.org/

[71] Categories for the Description of Works of Art (CDWA), http://www.getty.edu/research/conducting_research/standards/cdwa/index.html

ISO 19115[72]	Description of geospatial datasets	An HE/FE profile has been developed by EDINA for UK academic use[73]. The widely used FGDC standard[74] is being mapped to ISO 19115.
ETD-MS[75]	Description of electronic theses and dissertations	This is an international standard produced by the NDLTD[76]. A UK metadata set for ETDs[77] has been developed through the Electronic Theses project.

Table 3: A sample of the metadata standards currently available

In selecting a metadata standard it is important that it should be fit for purpose. This can mean that the ability to describe the content concerned meets requirements, of which there may be many: access, preservation, management, etc. The standard should also be transparent and preferably widely adopted to allow for easier interoperability, sharing and support. It is notable that many metadata standards can lean towards assisting the management of the content over the requirements for end-user access and/or associated service functionality. Whilst the immediate benefit and purpose of selecting a metadata standard to use may well be for content management, it is valuable to consider the wider purposes of describing the content so as to assist access and preservation further down the line. This is particularly the case where the content may be of value to more than one community, who may wish to access it using different aspects of the metadata. Different purposes may require different standards (for example appending preservation metadata to descriptive metadata or combining two descriptive records for different purposes). As such, systems need to be able to accommodate more than one standard and be prepared to manage these.

A4.2 Metadata richness

In deciding which metadata standard to use there is a need to consider what metadata is to be collected and why it is being collected. In the context of providing metadata for use by an end-user service it is hard to anticipate exactly how the metadata will be used. Providing as rich a metadata set as possible increases the range of options and flexibility available. This rich metadata may derive from a single metadata record, or be exposed through a number of related metadata records, each created for a different purpose: this latter approach is

[72] ISO 19115:2003 metadata for geographic information, http://www.iso.ch/iso/en/CatalogueDetailPage.CatalogueDetail?CSNUMBER=26020&ICS1=35

[73] HE/FE profile, http://go-geo.data-archive.ac.uk/ProfileIndex.htm

[74] Federal Geographic Data Committee, http://www.fgdc.gov/

[75] Electronic Thesis and Dissertation Metadata Standard (ETD-MS), http://www.ndltd.org/standards/index.en.html

[76] Networked Digital Library of Theses and Dissertations (NDLTD), http://www.ndltd.org/

[77] UK Metadata Core Set for Electronic Theses and Dissertations, http://www2.rgu.ac.uk/library/guidelines/metadata.html

being investigated through the CORDRA project[78], which aims to expose as much metadata as possible to maximise the ability and options for end-user services to make use of it.

Such an approach can appear to contradict the desire to create metadata for interoperability between repositories by adhering to the simpler Dublin Core metadata standard. This standard is designed to facilitate search and discovery, and by virtue of presenting a relatively limited standard set of fields it allows access across repositories from many different sectors and institutions, hiding the metadata variety within these. But Dublin Core has limits on how it can be used for description and there will almost always be some loss of information if the Dublin Core metadata record has been derived from something richer. If a repository can hold a richer metadata base record that can be mapped to Dublin Core for search and discovery purposes, allowing a clear entry point for the end-user, the original metadata can still be used subsequently for additional end-user services, such as browse or locate, as part of the Delivery-2-Discovery (D2D) chain. Making use of this richer metadata record requires that there be a link back to the individual repository after initial discovery, so as to provide more detailed information and functionality at the repository level.

A4.3 Metadata granularity

As well as storing rich metadata wherever possible it is also valuable to store metadata and content in as granular a way as possible. In exposing content and metadata to end-user services, the service will make use of what it is given. If the granularity is low, then the service can only deal with the individual objects as they are. If the granularity is high end-user services can potentially make use of all or just parts of what it is provided with, possibly disaggregating an object and re-combining it in ways that add value for the end-user. Content modelling, defining how different content types and their metadata will be dealt with, can help determine the level of granularity that best suits the repository and content at hand.

A4.4 Application profiles

At both simple and rich metadata levels communities will have their own requirements for how the metadata should be used and applied. The use of application profiles to support specific applications, functions, communities or environments to underpin the relationship between rich metadata records and Dublin Core mappings (and relationships between other metadata standards) will support end-user services that are focused around these specific needs [Heery, 2000]. Application profiles can arise from local adaptations or as a result of formal processes. In both cases there is value in knowing about these profiles to help inform the choice of metadata standard to apply. A registry holding details

[78] Content Object Repository Discovery and Registration/Resolution Architecture (CORDRA), http://cordra.net/

of profiles, such as the Information Environment Metadata Schema Registry (IEMSR)[79], can help support community take-up and standardisation of appropriate profiles, as well as providing information about the metadata standards they are derived from.

A4.5 Metadata mapping

Many richer metadata standards have had mappings to Dublin Core formulated (e.g., MARC, MODS[80]). This has been driven by the desire for interoperability. Simple Dublin Core is the most widely used metadata standard for use with OAI-PMH, although it is only mandated as the minimum requirement. As mentioned already such mapping can lead to loss of information, which access to the richer record can potentially alleviate further down the D2D chain. Mappings also help facilitate other aspects of metadata management and presentation to end-user services.

- The metadata used inside the repository does not have to be the same as that exposed by the repository to the outside world. The Fedora repository system makes use of FOXML (Fedora Object XML) internally[81], but maps this to other metadata standards for external presentation and import/export, including Dublin Core and packaged metadata using METS (for more details on packaging see section A4.12.1). This flexibility allows a repository to truly use the metadata standard that is the best fit for purpose, and adapt this as required.
- Storing a rich metadata record as the base record allows a write once, read many (WORM) paradigm to be applied, with derivatives of the base record being generated as required and used for specific purposes, including the provisioning of that metadata to other systems.

Metadata mappings are also known as metadata crosswalks and, as for application profiles, having a common reference location where details of these can be recorded would be a valuable asset to the community. OCLC have set up a registry for such a purpose [Godby, 2004]. It should be noted, though, that crosswalks, whilst facilitating the transition and exchange of metadata, cannot fix bad metadata itself.

In this light, and notwithstanding the flexibility that mappings/crosswalks offer, it is pragmatic to encourage communities to attempt to standardise around a limited set of metadata standards and application profiles to expose, if not one selected option. This would lead to greater consistency of metadata through

[79] Information Environment Metadata Schema Registry (IEMSR), http://www.ukoln.ac.uk/projects/iemsr/

[80] See the homepages of these standards for further information

[81] Fedora XML (FOXML), http://www.fedora.info/download/2.0/userdocs/digitalobjects/introFOXML.html

common application and, potentially, the ability to enable end-user services based on more detailed and consistent information.

A4.6 Metadata syntax

Metadata offers syntax for the information being recorded. Metadata standards are not very good, however, at defining relationships between different metadata components within the overall metadata record. Packaging standards (see section A4.12.1) can provide additional syntax, bringing metadata and sometimes content itself together in an organised fashion. The XML packaging standard MPEG-21 DIDL offers a way to formalise the syntax of the metadata included within it: this standard is one that can also encompass content and provide a structure for this alongside the associated metadata [Bekaert, 2003]. Alternatively, RDF and the Web Ontology Language (OWL)[82] can provide a detailed syntax as well as offering the possibility of semantic interpretation of the metadata and content, what they are and how they can, or should, be used. In both cases, it is the presence of an abstract model that allows such syntax to be put in place. Abstract models, including content models mentioned above, help determine what metadata needs to be collected in the first place. They also enable the development of end-user services: by basing services on the underlying model they are able to then manage variations in actual metadata and content more efficiently. This approach is being put into practice in the SIMILE RDF project at MIT, which is also using RDF to combine different metadata schemes using equivalences[83]. As indicated in section 4.13 these crosswalks allow efficient transition between metadata standards. The abstraction that RDF offers is complex, but offers much flexible potential: development of this potential is encouraged.

A4.6.1 Topic Maps

Topic Maps[84] offer an alternative to RDF as a way of structuring metadata to make it more easily usable. At one level they can act as the infrastructure for web portals (as they are by the National Library of Australia), collating multiple distributed contributions into a coherent web front end. At another level they can enable the drawing together of multiple metadata records and create links between them through a common spine that lays out associations and relationships between pieces of content. Topic Maps do require high metadata quality to work effectively. They require the use of unique identifiers so records can be linked correctly (a feature that is valuable in establishing relationships outside of using Topic maps as well): they also require ontologies to work from, and then add structure around these to hang subsequent resources off. There is much that is not yet understood about how Topic Maps could be used: work to investigate them and ontologies further, as a contrast to the RDF approach, is

[82] Web Ontology Language (OWL), http://www.w3.org/TR/owl-features/
[83] SIMILE project, http://simile.mit.edu/
[84] Topic Maps, http://www.topicmaps.org/

necessary. This can build on experience elsewhere, notably within the Norwegian Ministry of Education.

A4.7 Automatic generation of metadata

Creating the rich metadata records that can act as the basis for end-user services and structuring standards like Topic Maps is far from simple. It is unreasonable to expect content depositors to create all the required metadata, and repository managers are unlikely to have the resources to, especially if repositories are successful. Splitting the task up between different parties, for example between depositors and repository managers following an accepted workflow (as happens at both the Arts & Humanities and Economic and Social Data Services (AHDS and ESDS)), is one way of enhancing manual metadata generation processes, though experience suggests this is not necessarily always the most efficient path.

Automatic metadata generation seeks to address these manual shortfalls. The techniques to allow this vary dependent on the type of metadata in question. Technical metadata for digital objects can be derived using appropriate tools such as the JHOVE service from Harvard University[85] or the DROID service from the National Archives in the UK[86]. This technical metadata can, in part, be used for preservation purposes: another tool from the New Zealand National Library also aims to extract preservation metadata from file headers[87].

Tools to derive descriptive metadata from content are available (e.g., Kea[88]), though the process is currently not so straightforward as for other types of metadata. Natural language processing (NLP) tools are either very domain-specific or require a set of documents to learn from in order to be aware of the terms that should be extracted. Such processing has become well established in the biomedicine field, but is lacking elsewhere. Training NLP tools is not currently an easy task and may be beyond most institutions to implement. NLP tools that use statistics to extract relevant keywords rather than learning are less precise, but may offer a more practical way forward in facilitating automatic descriptive metadata generation.

NLP is one part of the range of services that comprise text mining. The National Centre for Text Mining (NaCTeM) offers a range of tools to facilitate this functionality[89], though again text mining is not a straightforward process and would not currently sit within repository workflows easily. Notwithstanding this,

[85] JSTOR/Harvard Object Validation Environment (JHOVE), http://hul.harvard.edu/jhove/
[86] Digital Record Object Identification (DROID), http://www.nationalarchives.gov.uk/aboutapps/pronom/droid.htm
[87] National Library of New Zealand Metadata Extraction Tool, http://www.natlib.govt.nz/en/whatsnew/4initiatives.html#extraction
[88] Kea automatic keyword extraction tool, http://www.nzdl.org/Kea/
[89] National Centre for Text Mining information extraction tools list, http://www.nactem.ac.uk/software.php?software=infoextraction

there may be feasibility in applying text mining techniques across a range of repositories rather than individual sources, and helping to generate metadata on a large scale. Text mining has issues of its own that would have to be taken into account, for example the rights associated with the metadata generated by the process (who owns this?) and the need to gather the full content of repositories in order to carry out the text mining (although much text mining currently takes place using abstracts rather than full-text).

The automatic generation of metadata is attracting a lot of interest in library circles. In the report ‘Rethinking How We Provide Bibliographic Services for the University of California’ there are a number of statements showing a clear desire to move to automatic generation of metadata where possible [University of California Libraries Bibliographic Services Task Force, 2005]. This includes gathering in metadata from where it exists elsewhere. Whilst this has been common practice in the UK for some years (sourcing bibliographic records from CURL or RLG, for example), it is an approach that is worth considering in the scope of repositories where the content being stored also exists in alternative locations. Is metadata being generated for them as part of any other process? For repositories holding materials that are local and unique this is unlikely, but for e-prints, for example, where there is an equivalent published journal article a metadata record may have been generated by a secondary information provider that could be re-purposed.

A4.8 Metadata quality

Ultimately, the higher the richness and quality of metadata that is entered into repositories, the easier it will be for end-user services to make effective use of this, and the more likely the end-user is to be able to interact with the content being exposed. It would be useful to think that mandating how metadata needs to be entered would solve this issue, and indeed this is the approach being taken in the application of the Integrated Public Sector Vocabulary by the e-Government Unit of the Cabinet Office. However, in the academic world this is unlikely to succeed. There are three levels at which the issue of quality can be addressed:

- At the repository: Guidelines for the implementation of metadata standards are valuable to any community in assisting with the use of the metadata standards selected. Most standards come with some generic guidelines though these can be relatively abstract in their application. Specific guidelines for particular purposes can be of greater value, for example the documents produced by the ePrints UK project on ‘Using simple Dublin Core to describe eprints’ [Powell, 2003] and ‘Encoding full-text links in the eprint jump-off page’ [Tourte, 2004]. Such guidelines are especially valuable for Dublin Core, as this standard is relatively relaxed about what content goes in the different fields – there are no widely accepted and used controlled vocabularies, for instance, leading to high variation.

Widespread awareness and take-up is a follow-on issue from producing guidelines. However, such guidelines do offer the opportunity to harmonise to some degree the metadata being produced for specific purposes and needs to be encouraged: adherence to agreed application profiles as described in section A4.4 is a step along this path. Guidance may be provided at the national level (as promoted by the National Library of Australia as part of the ARROW project for harvesting purposes [Campbell, 2005]) or at the community level (for example the OLAC community have created their own Dublin Core profile for exchange of metadata[90]), dependent on which levels of service are most relevant for the need at hand.

- At an aggregator or end-user service: If an aggregator or end-user service is able to gather the metadata local to itself it has the opportunity to address the quality issue prior to presenting the metadata to the end-user. The OAIster service makes extensive use of normalisation tables (for example mapping free text item types to a controlled vocabulary of these) to introduce consistency across the harvested metadata. This consistency allows end-user services to be built around the metadata, in the OAIster case the provision of an item type search. Consistency is an important feature in providing services: Google carries out normalisation itself by scrapping crawled metadata, which is too inconsistent, and creating its own according to its own guidelines. It is inevitable that there will be variety in what repositories make available, but this variety needs to be managed to enable access, not restrict it.
- At an intermediary service: Possible routes to enable automatic generation of metadata are described in section 4.7. These could be implemented at the repository level or at the end-user level, depending on where it is easier to apply the systems available. Text mining could be applied across repositories, but equally an end-user service that has harvested metadata may have access to links to the full content for similar purposes. This latter approach was tested within the ePrints UK project using web services to assign subject and authoritative name metadata. Normalisation services may also be considered intermediary and applied at the repository or end-user service level.

Raising the quality of the metadata can have issues of its own. Enhancing metadata through adding additional detail runs the risk of lowering the level of interoperability that is possible by providing too much detail. A high quality metadata record provides choices about the path to interoperability, though. Even if some dumbing down is required to facilitate interoperability the availability of a rich metadata provides added value further down the information chain.

[90] Open Language Archives Community metadata, http://www.language-archives.org/OLAC/metadata.html

A4.9 Subject classification

Subject-based end-user services are regarded as extremely valuable, as it is through their subject that many end-users will wish to interact with content. This raises the need for subject classification. Subject terms can be added manually as part of a cataloguing process, for example as carried out at the University of Glasgow, where they use Library of Congress subject terms for consistency across their repositories and library catalogue. An alternative to formal subject classification is the use of informal tags, assigned by the content depositor, as has been the practice within the ERA repository at the University of Edinburgh and was proposed by the RESULTs project in 2002. Tagging content is also a now widely accepted practice across a range of 'Web 2.0' services on the Web (e.g., Flickr). Having both formal and informal would be valuable and opens up possibilities for presenting the metadata in different ways. The combination of formal and informal could be achieved using Topic Maps, as is being tested in Norway currently.

The desire for subject-based access may mean that the lowest common denominator approach of Dublin Core in facilitating search and discovery across repositories may not be sufficient to meet these requirements. Feedback in the interviews conducted for this study suggested that simple Dublin Core is not sufficient for cross-repository retrieval, and that a richer metadata format with greater potential for subject classification may be better suited. Qualified DC can be used, but since qualifiers are not necessarily standard (unless laid down through an application profile) this approach can have the drawback of reducing interoperability rather than raising it.

A4.9.1 OAI sets

One methodology through which broad subject classification can be applied is the application of OAI sets to metadata that is available for harvesting using the OAI-PMH[91]. In February 2006 the SHERPA project carried out an investigation of the sets currently available from repositories in the UK that are registered with the Directory of Open Access Repositories (OpenDOAR[92]) which revealed that these were used, but not to the extent they might be: some offered very little subject breakdown, whilst others offered far too much detail. Any widespread adoption of OAI sets would require common agreement to be most useful to end-user services, but would allow these services to provide valuable subject-based access.

A4.10 Metadata changes and additions

In considering ways in which metadata can be generated it is important to remember that metadata is not always produced just once and that it may change over time. Content may be relevant to events or time, and it may be

[91] OAI Sets, http://www.openarchives.org/OAI/2.0/guidelines-repository.htm#Sets
[92] Directory of Open Access Repositories (OpenDOAR), http://www.opendoar.org/

appropriate to append event-based metadata to a record as required. This is one form of annotation, or layering of metadata, which adds value to the original record and increases its long-term relevance. Other annotations such as feedback, user comments, ratings, and additional metadata specific to different communities being served can also be added. End-user services offering this type of functionality need to ensure that such annotations will be clearly associated with the original metadata record, an issue that requires clear linkage between them using appropriate identifiers. These changes to metadata raise issues of versioning and when one record becomes a new record in its own right rather than just another version. Metadata change management is currently underdeveloped in general, but will need to be developed as digital content and associated metadata becomes the norm.

A4.11 Identifiers

Identifiers are vital components of any metadata record, and indeed of the associated content itself. These identifiers should have the following key characteristics to ensure their long-term value and use:

- They should be persistent. It should be noted that persistency does not necessarily mean forever, but simply for the lifetime of the object being identified. An individual parcel being shipped round the world only requires its shipping identifier until it is delivered, after which it is no longer relevant. Many other identifiers will need to be persistent over many years, though, for preservation purposes.
- They should be unique. Using non-unique identifiers runs the risk of leading to duplication and confusion for the end-user. It should be noted that some common identifier schemes do re-use identifiers, for example ISBNs after a set period of time after cease of publication, and that it is possible to cope in these circumstances. Establishing uniqueness avoids these possible problems further down the line, however.

Having a unique, persistent identifier potentially allows just the identifier to be passed around when processing metadata and content, as the identifier will always know where the relevant components are available. This benefit can be compromised by objects having multiple identifiers (often generated each time an object is ingested into a repository), though the world of physical objects has had this problem for some years and dealt with it accordingly: each identifier usually has a specific purpose (e.g., a journal has an ISSN for the title and a SICI for the issue). In principle, each identifier should identify only one thing.

A4.11.1 Identifier granularity

As stated for metadata granularity in section 4.3, end-user services are able to do far more with metadata and content stored at a high level of granularity rather than at a low level. A number of benefits exist for this. It is valuable that identifiers are clearly associated with the components and sub-components involved, as each can then be easily referred to. High granularity can also support disaggregation of compound objects and recombination as required – you can put components together if each can be identified, but it is far harder to break them up if a low level of granularity is applied. Granular identifiers permit clear links between associated metadata and content records, allowing end-users to clearly see, for instance, which metadata records have associated full-text or not. A high level of granularity would see repositories assigned their own identifier in the same way as the content and metadata within it so that they too can be regarded as individual objects and specifically referred to where required.

A4.11.2 Author identifiers

In identifying components of metadata records and content the ability to uniquely identify authors is an area that is not well understood, though would be a valuable additional element in ensuring correct identification of all aspects of a record. A project in the Netherlands is looking to assign a unique Digital Author Identifier to all scientists in the country. In larger countries this may not be feasible, and maintenance of any authoritative list of author names will not be a light task. Nevertheless, name authority files are regarded as a necessary step in achieving full unique author identifiers. Such a file may be centralised or could be distributed, making use of local institutional records to identify authors. The repository at the University of Southampton has links to the local staff ID database to ensure correct identification at least at this local level: similar use of local staff records have been considered elsewhere as well. On a commercial level Scopus has introduced the Scopus Author Identifier[93] to aid searching by author and CSA has started tagging some authors to facilitate access to all works by a particular author. Further investigation of this issue is required to assess possibilities.

A4.11.3 Identifier resolution

A key question in assigning identifiers is whether they should indicate where they are from or be location-independent. URLs indicate where they are from through their domain name, but cannot be guaranteed to be either unique or persistent (as content at the URL may change and the URL re-used). Ideally, identifiers should be location-independent, and resolve themselves through transparent resolution services (i.e., determined by the identifier namespace). Whatever solution is implemented for resolution it will be of benefit to end-user services if this is global, and remove the need for them to implement their own resolution functionality locally.

A4.11.4 Identifier agreement

As was pointed out by more than one interviewee there is no current common solution on identifiers because there are a number of solutions. Many are waiting on other decisions (often in the same way as Waiting for Godot!), and indeed the only guarantee of true persistence is the policy and organisational support behind the implementation of any scheme. Modelling of identifiers, in the same way as

[93] Scopus Author Identifier, http://info.scopus.com/etc/authoridentifier/

the use of underlying content and data models, will inform what type and level of identification is required and this can inform the decision-making process. Notwithstanding this the global resolution recommended in the previous section suggests it would be beneficial for widespread agreement on one identifier scheme, though this would require high-level agreement and assessment of the relative merits of different identifier standards. There is also a need to agree on identifier syntax and registration procedures for smooth running of the system.

If agreement can be reached there would remain the issue of many identifier standards already being in place and used. It is likely that conversion would almost never work. Abstracting out the identifier to apply a new identifier where required for harmonisation may be a solution, though this is untested. In any work on identifiers exemplars of how to use them and how they can be used will be of wide benefit.

A4.12 Compound objects

Compound objects have been mentioned previously in the context of learning object repositories, where the ability to combine individual objects can lead to a more valuable learning material. Other types of compound objects might include e-research objects involving datasets and multimedia objects or associated publications. Compound objects are discussed here in their own right, and also in the context of packaging standards, which have arisen through the desire to package metadata and/or content (with multiple instances of each forming a compound object) for transfer and use between systems.

Compound objects are potentially very rich. There can also be complex to construct, and it is in the creation of compound objects that automatic generation of metadata finds a valuable use case. Like the automatic generation of metadata there is a need to investigate compound objects further, to assess their role, their construction and their use. Understanding the content and metadata we wish to include within a compound object through developing appropriate data and content models for these will be a valuable step towards making effective use of such objects. Such modelling will provide information on how the different components should relate to each other and how they can be used, as well as provide correct and clear syntax both for the use of metadata standards within objects and the use of the objects through different interface standards. Modelling will also determine the correct granularity of identifiers to use allowing different components of the compound object to used and accessed independently. One clear distinction is the need for separate identifiers for the content and the package itself so that these can be managed accordingly without confusion.

Content modelling is especially necessary for born digital content, as currently used physical paradigms for structuring content may no longer apply. For example, electronic theses can contain multiple content types linked in complex ways. Compound objects may also include both born digital content and analogue references. A new method for dealing with these is required in order to effectively present them to end-users. This is true both for presentation through digital library end-user services and personal repository collections.

Compound objects in their own right do not enhance the quality of the individual components, metadata or content, contained within them. Bad metadata/content will lead to an unuseful compound object. Compound objects also focus primarily on syntax: semantic interpretation of the compound object is an area requiring further research and development and where RDF/OWL may have a major role to play.

A4.12.1 Packaging

The purpose of packaging up a compound object is varied. It is primarily a means of transferring materials from one place to another: this could be for transit between two repositories, or for use via an end-user service. Compound objects facilitate discovery through establishing relationships between individual components, and facilitate delivery by being able to deliver components together. Packaging also facilitates preservation through the ability to move content between repositories. The recent Library of Congress Archive Ingest and Handling Test (AIHT) recommended a need to standardise on the format of the packaging being used for transfer in this way to ensure accurate transfer[94]. Standardisation for a specific use case like this is clearly of benefit. Standardising packages otherwise needs to be equally clear about what is being standardised and why. The transfer of packages relies on repositories being capable of dealing with the packages being submitted to them. This capability could be a set of services at the repository or services provided as a separate layer between the package and the repository, and used generically across and by a range of repositories. For example, DSpace has a separate plugin module to allow the processing and ingest of any packaging standard. This uses RDF as a means to facilitate this interoperability (the packaging itself could be carried out using RDF).

Apart from the content and associated metadata there are a number of other pieces of information it may be useful to include within a package, and which will make up additional components of the overall compound object. Best practice of updating and maintaining additional metadata like this remains unclear.

- Collection description: a way to describe what is held within the package that can be used to determine its usefulness. The manifest file used by IMS Content Packaging (CP) is an example of this.
- Annotations/ratings: metadata within a compound object may not be created just once, as mentioned earlier, but may be appended over time,

[94] Archive Ingest and Handling Test, http://www.digitalpreservation.gov/technical/aiht.html. See also articles in the December 2005 issue of D-Lib Magazine at http://www.dlib.org/dlib/december05/12contents.html

either through event-based metadata, annotations or feedback such as ratings.

The additional complexity of generating packages and the compound objects within them requires a need to understand why the packages are needed and assess the cost effectiveness of generating them. There can be benefit from them, and such benefits need greater examination, but it is unlikely that they are worth generating simply for the sake of it. This is particularly the case for use of packages by end-user services. How will these services deal with such packages and what use can they effectively make of them?

There are a number of standards available to enable the packaging of content, as listed in Table 4. As with metadata standards these have emerged from different sectors and their aims are similar. However, each has its individual characteristics, which will play a role in deciding which one is best in particular circumstances. There is a need to do more work on assessing which packaging standards suit different types of content to help with such decisions.

Packaging standard	Notes
MPEG-21 DIDL[95]	ISO standard. XML-based, it originated in the commercial entertainment industry and primarily designed for multimedia. Highly structured, flexible and generic, though also complex. It is designed for the transfer of assets. Has an abstract model and is good at identification of individual components. Includes content through Base64 encoding.
METS[96]	Community-based open standard, maintained by the Library of Congress. Designed for the packaging of XML metadata formats, not content, and pointers (identifiers) to content where required. Not as structured or as flexible as MPEG-21 DIDL, though possibly easier to implement. It has no abstract model.
IMS Content Packaging[97]	IMS standard. Based on a ZIP file of content and metadata, though there is an XML binding available. Designed for the offline transfer of primarily learning objects and other IMS specifications (for which it has specific features), though can be used more generically. Contains an XML manifest file that provides information and structure about its components. Has an abstract model.
ATOM[98]	An XML-based syndication format currently undergoing standardisation through the IETF. It can include both metadata and content, through Base64 encoding, and can act as a packaging format as well. No abstract model.
XFDU[99]	An adaptation, and possible forking, of the METS standard by the team behind the development of the OAIS reference model. It is unclear at what stage of development this is currently.

Table 4: A selection of packaging standards

[95] MPEG-21 standard, http://www.chiariglione.org/mpeg/standards/mpeg-21/mpeg-21.htm
[96] Metadata Encoding & Transmission Standard (METS), http://www.loc.gov/standards/mets/
[97] IMS Content Packaging specification, http://www.imsglobal.org/content/packaging/index.html
[98] ATOM, http://www.atomenabled.org/
[99] XML Formatted Data Unit (XFDU), http://sindbad.gsfc.nasa.gov/xfdu/

MPEG-21 DIDL has been brought to the attention of the academic and digital library communities by the work of Herbert Van de Sompel and his colleagues at the Los Alamos National Laboratory (LANL) in the US[100]. It has since been used on a limited scale by a number of projects and institutions within the DAREnet project in the Netherlands, where there is some experience of harvesting such compound objects.

There have been concerns about possible patent limitations to the use of MPEG-21 DIDL. Jeroen Bekaert from LANL attended the MPEG meeting in January 2006 where this was discussed. Queries raised at this meeting about patent issues concluded that if there are patent issues then these relate only to the XML format used to express digital rights statements, not the packaging format itself. The uncertainty about the patent is due to a generic ongoing patent issue about the use of XML formats for digital rights expression, which is as yet unresolved. Hence, there is no patent barrier to use of the MPEG-21 DIDL packaging standard. The patent issue of using rights statements within packages generally requires monitoring to assess possible impacts in the future.

METS has become widely used within the library community as a way of bringing together different metadata records about the same object for transfer and possible delivery as part of resource discovery. There is an outstanding need to agree on common usage, not least so that the interchange of materials can occur successfully. There is slight concern that not all library-oriented metadata standards will align with the XML nature of METS, though most commonly used standards do now have associated XML schemas. The lack of an abstract model means that document modelling is difficult and structuring the package beyond its ability simply to hold different metadata records together may not be easy. The PORTICO e-journal archiving project[101] has dealt with this by adapting METS for its own purposes, including some aspects from MPEG-21 DIDL within this.

IMS CP is a robust and successful packaging standard that has been in use for the transfer of resources, primarily learning objects, for some years. The latest version, 1.2, has made adaptations to the manifest file so that this can point to content outside of the package itself as well as referring to resources within the package. The California Digital Library uses METS as an equivalent to the manifest file in much the same way, pointing to distributed resources. This has initiated a ongoing debate within the CDL about the value and need for packaging when you can link through to content as required, so long as it is clearly identified.

[100] Los Alamos National Laboratory Research Library, http://library.lanl.gov/lww/

[101] Portico Electronic Archiving Service, http://www.portico.org/

A4.13 Interoperability

As with metadata a repository's internal compound object structure does not have to be the same as the structure of the packages made available externally. Thus, Fedora uses FOXML internally, but is able to disseminate METS packages externally. The ability to map from internal structures to external ones is valuable. Van de Sompel and colleagues are developing mappings to MPEG-21 DIDL for three of the main open source repository systems: DSpace, Fedora, and EPrints. It is clear that availability of these and mappings to other packaging standards will offer flexibility for repositories in the way they operate. However, there are accompanying needs to coordinate these crosswalk activities and investigate more closely the possible loss of information that such crosswalks can result in. These mappings apply to XML-based standards: IMS CP relies on the ability of a repository to ZIP relevant files together for exposure.

The structure that packages can bring can make interoperability and discovery easier. The NDLTD currently does not offer packages for exposure. Indexing by the Scirus search engine was complicated by this lack of structure and METS is now being considered for future use.

The existence of multiple packaging standards inevitably means that for full interoperability between repositories and end-user services the ability to move between and manage multiple standards is valuable. The RAMLET initiative is examining this possibility, and is currently investigating the use of OWL to facilitate this cross-standard mapping[102]. As mentioned above, such mappings may lead to loss of information and this needs to be closely monitored. The lack of semantic markup within packages indicating the underlying data model, where this exists, may prevent true interoperability.

A4.14 Working with full content alongside metadata

The use of many of the packaging standards involves content alongside metadata being transferred and used between repositories and end-user services. Packaging takes the content to the user directly, though it can equally be pointed at from within a package or from a simple metadata record. In either case, and the end-user will not care which approach is followed, the intention is to allow the end-user to get hold of 'downloadable bits' and gain access to content, a major goal as part of the Discovery-2-Delivery chain. Resource discovery is reduced without access to full content. In considering open access repositories this is particularly the case as access to the full content is one of their main purposes. Other benefits from having access to full content include the ability to provide full-text indexing to support the discovery process, a methodology that Google employs. The metadata enriching services

[102] Resource Aggregation Model for Learning, Education and Teaching (RAMLET), http://ieeeltsc.org/wg11CMI/ramlet/

investigated by the ePrints UK project also relied on access to the full content for analysis in order to enhance the associated metadata record.

Dealing with full content brings issues of its own, though.

Where full content is being pointed to within a metadata record there is a need for clear, persistent, unique identifiers to ensure that the link from the metadata record will succeed and the end-user will not be disappointed. There is a need to standardise on where links from metadata records should go: do they point to the object itself or to a jump-off page that can provide additional context but add an extra step for the end-user? ePrints UK addressed this in seeking to use the content identifiers within harvested metadata records to capture the full content for processing using Web services. This was complicated by the widespread use of jump-off page links rather than links to the content itself. Where a direct link may not be applicable an OpenURL link can provide context-sensitive and persistent links for the end-user and prevent dead-ends in information retrieval. Clear identification will also be the required when identifying relevant components within a compound object package.

Notwithstanding the greater ability to store large amounts of content these days and move this around on the network, space and bandwidth issues have to be borne in mind when considering access to full content, particularly where this includes large items such as 3D, datasets and video/audio. This is particularly the case where content is being moved around, perhaps within packages, to facilitate access to it. In the same way that caching on the web speeds up access to web pages, it may be of value to have object caching services (e.g., such as those provided by Akamai[103] and, originally for e-journal content, LOCKSS[104]). Caching also provides a back-up where access to the original content repository is not available. Third party storage facilities at the network level away from institutions are already of value (e.g., through the Data Services available in the UK, AHDS, ESDS, UK Data Archive etc.). As content increases in size and complexity the need to provide additional network-level storage is likely to increase.

Replication between repositories and services does not scale well and also raises issues of data authenticity, provenance and integrity. If content is updated or its purpose changes as part of its lifecycle, how will this be provisioned to replicated versions elsewhere (if it should be)? For such changes bi-directional standards are required: OAI-PMH is uni-directional in its harvesting, though the originating repository could also act as a harvester for later re-harvesting of the altered content. RSS can also be used in conjunction with an add-on like Microsoft's Simple Sharing Extensions[105] to allow bi-directional, asynchronous

[103] Akamai, http://www.akamai.com/

[104] Lots of Copies Keeps Stuff Safe (LOCKSS), http://www.lockss.org/lockss/Home

[105] Microsoft Simple Sharing Extensions for RSS and OPML, http://msdn.microsoft.com/xml/rss/sse/

exchange. If content is stored elsewhere there are privacy and IPR concerns for content owners. What can end-user services and/or other repositories do with content they access/receive? These are not necessarily technical questions, but they will need technical answers and require attention. There are also preservation issues. Which version will be preserved and where? The ability to move content between repositories/aggregators/end-user services offers a number of possibilities for preservation via a third party, however. The SHERPA-DP[106] and PRESERV[107] projects are examining exactly this type of preservation role for e-print content. The Hybrid Archives project[108] also used this model to capture content whilst it is still served to end-user services by the original content host.

Rights issues have been at the heart of the development of the JORUM learning object repository. This holds both content and metadata, and also acts as the primary end-user service onto these. Content is held in order to facilitate re-purposing and re-use. Acting as a repository of links, as MERLOT does, may well be a far simpler model, but it doesn't allow the additional functionality that JORUM can.

It is sometimes desirable to maintain separation between metadata and content. In the context of resource discovery metadata can provide equal entry points for access to both analogue and digital items. Once discovered the link to either of these can be made. Holding full digital content too close to the metadata can introduce a bias towards digital content access over access to more, though possibly equally valuable, analogue resources.

In order for end-user services to make effective use of content they need to be able to access the delivered content formats. The plugins and tools required for this should thus be as ubiquitous as possible, for example the ability to access PDF files using the relevant plugin. What flexibility there is in delivery format will rely on the availability of 'on the fly' format conversion services between the repository and aggregator or end-user service.

A5 Interfaces

The stated focus of this study is on user-oriented services, those facing end-users that allow direct interaction with them. The machine interfaces that underpin these services are also, however, of interest. The interfaces that repositories make available determine how they provide end-user access to their contents, whether this is through a direct web interface or by exposing repository contents through one or more machine interface options. Thus, both are

[106] SHERPA-DP: Creating a persistent preservation environment for institutional repositories, http://ahds.ac.uk/about/projects/sherpa-dp/index.html
[107] PRESERV: Preservation Eprint Services, http://preserv.eprints.org/
[108] Hybrid Archives project, http://ahds.ac.uk/about/projects/hybrid-archives/index.htm

considered here. Much discussion of interfaces centres on the use of relevant standards and protocols, and this section will address the issues related to these as well as considering the broader aspects of interfaces that repositories offer.

The interfaces that repositories offer also determine the level of interoperability across repositories that can be enabled. Simple web interfaces to individual repositories do not, in general, meet this requirement (unless a metasearch tool is capable of replicating multiple HTTP requests behind the scenes), whereas machine interfaces that expose contents for use elsewhere offer the flexibility and, where standards are adhered to, common platform upon which interoperable services across repositories can be established. The balance in priority given to the development of both web and machine interfaces by repositories is thus important in considering the level of interoperability that is possible. It is important to remember that the interfaces and the services/applications making use of them are not necessarily at the repository itself: there is one or more layers between them, and this layering offers flexibility of its own in how repository contents are exposed.

Overall, interviewees for this study agreed that the simpler the interface standards and the more widely adopted they are, the better for interoperability. Standards are particularly welcome where you can do a lot with a little, for example with RSS. There was also wide agreement that the standards required to enable interoperability across repositories are fairly well in place: it is a combination of application support, policies to determine how they are used, and the need to ensure that they are used consistently that are the greater priorities. Putting policy in place can itself help to identify gaps that were not previously apparent, e.g., the recent recognition of a gap in a standard deposit API through experience of implementing repository systems. There is also a gap in the ability to update repositories: the development status of SRW Update[109] is unclear, whilst although experience with WebDAV[110] has suggested its value in this area there are also concerns based on recognising that repositories are not web servers.

There is a need for repositories to assess what level of interoperability they wish to enable, as this will affect which interface standard(s) will be most appropriate. Beyond generic use it is unclear which interface standards are best for different usage scenarios and content types and both repositories and end-user services need to examine this in greater detail. Offering a range of interfaces can help to assess this: indeed all repositories should consider offering more than one interface to allow for flexible interaction with aggregators and end-user services. An approach based around standards that are extensible and flexible will ensure a repository can adapt to future, and unknown, needs. This flexibility is important to allow for innovation, too, and prevent adherence to standards being a millstone rather than an enabler.

[109] SRW Update, http://srw.cheshire3.org/docs/update/
[110] Web-based Distributed Authoring and Versioning (WebDAV), http://www.webdav.org/

A5.1 Web interfaces

Notwithstanding the need for machine interfaces to facilitate interoperability the issue of web interfaces is addressed first here. This entry point to repository content is the primary one that many repositories develop as a priority, as do aggregators. Although there was wide acceptance of this phenomenon amongst interviewees it was unclear why this should be, as generally, there was also acceptance that machine interfaces are more valuable and, hence, more of a priority. There is perhaps a desire to aim at low-hanging fruit in developing a web interface, guided by user demand (although developing a decent web interface can be a major time and resource commitment); such an interface brings visibility on the web and branding for individual repositories and their owners; or there may be a need to cater for casual users who may not have access to more specific search tools that are making use of the machine interfaces. Web interfaces, with their set URLs, are also ideal for crawling by search engines such as Google, and there is much value in exposing repository contents for this purpose. The ‘Google effect’ can be noted elsewhere as well with user expectations leading to web interfaces onto repositories mimicking the Google interface design.

Web interfaces, ultimately, are good for direct access and are used, whereas machine interfaces, so far, are not being used or exploited as well as they might be. Although machine interfaces appear to offer much potential this remains untapped. In the meantime, web interfaces and web crawler aggregations of these are providing the primary points of entry and interfaces onto repositories.

A5.2 Machine interfaces

The major case for developing machine interfaces onto a repository is to increase the flexibility of how the repository contents can be subsequently used. It is possible to generate services on top of a Google aggregation, though repository metadata tends to be mixed in with all other web content, making it difficult to provide focussed services (though note Google Scholar may be of value here). Machine interfaces permit clearer and more targeted exposure and recombination of repository metadata and content, for example to support subject-oriented access, that allows end-user services and different views onto the contents to be built for specific needs. They can thus add much value assuming repositories are willing for their records to be used elsewhere, possibly out of their control. Providing both machine and web interfaces covers both approaches, and repositories should consider both: for institutional repositories this is made easier by the in built OAI-PMH capability in commonly-used repository software systems: repositories should address their needs and requirements for machine interfaces proactively rather than simply adopt what is provided, though. The level of effort required providing both can be a problem, though if machine interfaces are used as the basis for the local web interface as well as serving remote services duplication can be avoided. Machine interfaces are harder to configure as they are less forgiving, and many content owners have concerns about loss of visibility behind the interfaces, though they can bring wider benefit in making metadata/content more widely available, even if behind the scenes. Focussing on such interfaces as the basis for web interfaces and other end-user services allows a single point of focus for development with maximum output.

A5.3 Interface options

The range of possible standards, specifications and protocols that can be used to provide machine interfaces to allow repositories to expose their contents are discussed in this section. The range of interfaces mentioned in discussions for this study is listed in Table 5. Whilst accepting that there remains some uncertainty about which is best for any one purpose, there is also a recognised need to focus around a constrained set of interfaces to make it clear to subsequent services and end-users what is being made available and to ease implementation.

Interface/standard	Notes
OAI-PMH	Allows metadata, and possibly content if appropriately packaged, to be exposed across repositories for harvesting, aggregation and delivery through a service provider
	http://www.openarchives.org/
RSS	Allows for controlled syndication of metadata by repositories to RSS readers and related applications/services
	http://en.wikipedia.org/wiki/RSS_file_format
ATOM	Like RSS, allows for controlled syndication of metadata, but also content through Base64 encoding
	http://www.atomenabled.org/
SRW/U	Allows for focused searching of repositories and/or aggregations of repositories: SRW is a SOAP-ful standard whilst SRU is a REST-ful standard
	http://www.loc.gov/standards/sru/
Z39.50	A forerunner to SRW/U and pre-dating the Web itself, Z39.50 also allows for focused searching
	http://www.loc.gov/z3950/agency/
OpenURL	Allows contextual linking between different resources, which may involve linking out of a repository or into one
	http://www.niso.org/standards/standard_detail.cfm?std_id=783
SQI	A Simple Query Interface developed to facilitate interoperability across learning object repositories
	http://ariadne.cs.kuleuven.ac.be/vqwiki-2.5.5/jsp/Wiki?LorInteroperability

Table 5: Interface standards for use with repositories

A5.3.1 OAI-PMH

The Open Archives Initiative had its origins in 1999 at the Sante Fe Convention [Van de Sompel, 2000], which sought to describe an approach for the easy sharing of metadata to facilitate sharing of e-prints between researchers. The Protocol for Metadata Harvesting that emerged from these discussions has since been used in a wide number of situations beyond the original e-print remit, although use of the protocol by open access repositories remains a key and established part of the open access process [Lagoze, 2003]. The relative simplicity involved in implementing OAI-PMH has, though, possibly led to a similar scenario experienced in the use of Z39.50: implementation to a certain level allows you to use the standard, however full use of the standard and its capabilities relies on additional effort. For example, the use of OAI sets potentially allows more refined harvesting to take place and better targeted services. Sets are not widely implemented, though, as evidenced in recent work by the SHERPA project on current UK repository implementations.

The OAI model relies on the presence of data providers and service providers. Data providers such as repositories expose information for service providers to harvest. This process is uni-directional: the service provider pulls the metadata from the repository. The process cannot work the other way round unless the data provider and service provider roles are reversed and the repository harvests from the service. Bi-directional use of OAI-PMH, and other standards listed here, would be valuable, and would support the gap in updating indicated in section A5. OAI-PMH has also been mainly used for the harvesting of simple Dublin Core records through its predominant use for e-prints: this narrow focus has possibly been of benefit to interoperability though has limited the protocol’s potential usefulness. Its use with other materials, and specifically with additional metadata standards would be helpful. International standardisation and abstraction of the protocol will help here, as well as helping to address remaining dependencies within OAI-PMH such as its use of HTTP as a transfer protocol.

Once the service provider has harvested from across a range of repositories the resultant aggregation can be accessed either directly via a web interface, or exposed in its own right to other end-user services (and the role of OAI-PMH may be best placed behind end-users services). For example, the aggregation could be made available as an SRW/U target (as OAIster recently has made itself an SRU target[111]). The Bielefeld University BASE search engine[112] places a FAST-based search engine over harvested metadata and content, bringing the benefits of formal standards and web search engine technology together. In general, the use of OAI-PMH to gather metadata and/or content together as a platform for end-user services appears to be gaining acceptance over the distributed searching model, though only the development of further end-user services will provide additional evidence. For BASE, and also the Omega

[111] OAIster SRU target details, http://oaister.umdl.umich.edu/o/oaister/sru.html
[112] Bielfeld Academic Search Engine (BASE), http://base.ub.uni-bielefeld.de/index_english.html

metadatabase at the University of Utrecht[113], there is value in bringing metadata and/or content to a place where there are better end-user services available for the desired audience. In addition to searching, the aggregation approach allows for other services to be applied, such as those highlighted in sections 2 and 3 of the main report.

A5.3.2 RSS

Although relatively simple to implement RSS is not a simple standard, but rather a group of standards that have developed over time. RSS 0.9 and 1.0 represent RSS based on RDF, whilst RSS 0.91, 0.92, and 2.0 represent alternative versions that do not adhere to the RDF model. All are relatively widely used, and the majority of RSS readers will cope with all the standards just in case. The meaning of the acronym RSS has changed with versions, and can stand for RDF Site Summary, Rich Site Summary, Really Simple Syndication or Real-time Simple Syndication (the last two predominantly used for RSS 2.0).

The common factor across all these versions is the role of RSS, which is designed to allow information holders, such as repositories, to push snippets of the available information out to those wishing to read it via an RSS reader, a browser, or within other web applications such as portals. It is notable that this wide implementation has made the use of RSS of interest to publishers, possibly over the use of OAI-PMH for distributing metadata, a trend that the academic community will need to monitor.

As well as alerting based on events and news, for which RSS is mainly used, syndication can also take place of blog entries and even updated search requests (e.g., a search of SCRAN can be embedded in RSS and periodically re-run and the end-user alerted to new results). Wider applications of RSS are also emerging, such as the OpenSearch service from Amazon, which allows the return of search results in HTML, RSS or ATOM formats, allowing flexible presentation and re-purposing[114].

A5.3.3 ATOM

The ATOM syndication standard has its origins in perceived deficiencies with RSS 2.0 and is now being formally standardised through the IETF. It is not as widely used as RSS, though is increasing in popularity, in part brought about by its use within a number of Google's services. As indicated in the table above, it is also capable of pushing content as well as metadata through the use of Base64 encoding. This is akin to the use of such encoding within the MPEG-21 DIDL packaging format. As such, ATOM has the potential to be used for both syndication and packaging.

[113] Utrecht University Library Omega metadatabase, http://omega.library.uu.nl/
[114] OpenSearch, http://opensearch.a9.com/

A5.3.4 SRW/U

SRW/U originated in the ZING (Z39.50 International: Next Generation) initiative led by the Library of Congress. SRW (Search/Retrieve Web service) is a fully-fledged Web Service that makes use of SOAP messaging for query and receipt of results. It uses the Common Query Language (CQL)[115] to structure its searches. SRU (Search/Retrieve via URL) also makes use of CQL, though is easier to implement than SRW through its use of single URLs to build up queries and submit them. Results are, as for SRW, returned as SOAP messages. It is notable that the emphasis between the two standards has shifted to focus more on SRU recently. SRU has also emerged as the frontrunner in discussions within the NISO MetaSearch Initiative[116] as the common denominator that can enable effective and consistent cross-searching of multiple targets (e.g., across repositories). Although of potential benefit in this space there are continuing concerns about the scalability of distributed searching using standards such as SRW and SRU that remain to be addressed. There are also not many native SRW/U targets available yet with many operating through a Z39.50 to SRW/U gateway, for example The European Library [van Veen, 2004].

A5.3.5 Z39.50

The Z39.50 standard originated in the late 1980s as a technology to enable interoperability between different computer systems. Although capable of many aspects of interoperability the standard has been most widely used within the library and information fields for cross-searching of databases. Z39.50 provides a common syntax to allow communication between computer systems, though the level to which this has been conformed to has varied widely and, together with the relative complexity in configuring Z39.50 clients and targets, this has affected its successful use on a wide basis. The standard also pre-dates the Web, and is not designed for use in this arena. For many new developments the use of SRW/U is now favoured: it is notable that none of the commonly used open institutional repository software packages have provided a Z39.50 interface whilst SRW/U is available for at least two of them. The predominant use of Z39.50 within the library and information field is also now seen as a limitation in aiming for wider interoperability. Nevertheless, Z39.50 cannot be ignored completely because of this continuing usage.

A5.3.6 OpenURL

OpenURL has its origins in work carried out at the University of Ghent in Belgium by Herbert Van de Sompel and colleagues in the late 1990s, and subsequent work with Ex Libris leading to the development of the SFX service [Van de Sompel, 2001]. It offers the ability to dynamically link between resources depending on the context in which searchers find themselves. Although originally designed for bibliographic materials the NISO Open URL 1.0 Z39.88-2004 standard has been designed to facilitate linking between almost any types of object, which raises possibilities for how repositories can act as either the

[115] Common Query Language (CQL), http://www.loc.gov/standards/sru/cql/index.html

[116] NISO MetaSearch Initiative, http://www.niso.org/committees/MS_initiative.html

source of OpenURL links or targets for them. OpenURL links from the repository have been tested at the University of Cranfield with DSpace. There is also the possibility of linking into the repository from elsewhere. One of the acknowledged problems is the formatting of citations within the OpenURL: the repository needs to hold the data in as granular format as it can to enable this.

A5.3.7 SQI

SQI has emerged from within the wider EU CEN/ISSS Learning Technologies Workshop initiative to enable interoperability across learning object repositories. The technology is being tested in the Netherlands for use within LOREnet, the Dutch national network of learning object repositories[117].

A5.4 Lightweight approach to interoperability

One of the main advantages that most of the standards listed above have is the relative ease with which they can be implemented, at least on a basic level. This was a highly desirable feature amongst interviewees. Whilst Z39.50 may no longer be flavour of the month even though it has, more or less, proved its worth, OAI-PMH and RSS are being used more because of their openness, allowing subsequent re-use and re-exposure as required. They are more lightweight in their approach, but are able to facilitate interoperability more easily as a result. The OCKHAM initiative is focusing on this lightweight approach, demonstrating how different standards and interfaces can be combined to maximum effect [Xiang, 2005]. Thus, an OAI-PMH harvested collection can be delivered using RSS for updates, be a SRW/U target for specific searching, or even act as an OpenURL target for location of items discovered elsewhere. The standards can even be combined in one go: the DLF Aquifer project[118] is experimenting with "asset actions", the embedding of RSS snippets in OAI harvested metadata containing actionable URLs that allow subsequent actions to take place or be enacted by the end-user.

The lightweight approach is valuable as a methodology for exploring which standards are most appropriate to support the desired interfaces. By presenting more than one interface repositories can test which approach works best for which purpose. For wide applicability a common approach to combining standards is required, though so long as the interface APIs are well-defined and open other parties should be able to pick them up and build on them.

Some of the standards have an improved ability to provide added value if working on bigger aggregations of metadata/content. RSS, for example, can provide more targeted and relevant updates, whilst OpenURL links are more likely to find a result if there is a greater body of objects to link to.

[117] LOREnet, http://www.lorenet.nl/en/page/page.view/home.page
[118] Digital Library Foundation Aquifer Initiative, http://www.diglib.org/aquifer/

A5.5 Repository specifications

The lightweight approach enabled by the interface standards described so far offer clear paths to interoperability. The compromise required is that the standards predominantly focus on specific areas of functionality only (notwithstanding their potential overlap [Sanderson, 2005]): OAI-PMH harvests exposed metadata, SRW/U carries out searching, RSS provides alerts. All the standards can be used on a limited or extensive basis, but the primary underlying aim for each is clearly defined. They permit a particular level of interaction with a repository, and because that level can be replicated across many repositories interoperability can be achieved relatively easily.

If there is a desire to interact with repositories more extensively, and carry out multiple functions across these on an interoperable basis, then the interfaces that the repositories need to present need to be more complex and capable of dealing with different requests. A number of repository specifications have been developed to different degrees to try and enable this deeper level of interoperability, as listed in Table 6.

Specification	Notes
IMS Digital Repositories Interoperability (IMS DRI)	Designed to provide recommendations for the interoperation of common repository functions at a high level. Not developed beyond a set of guidelines.
	http://www.imsglobal.org/digitalrepositories/
JSR 170/283	JSR 170 and its successor JSR 283 are Java Community Process specifications. They provide a content repository API to enable an implementation independent way to access content bi-directionally between repositories and applications at a detailed level.
	http://www.jcp.org/en/jsr/detail?id=283
Open Knowledge Initiative Digital Repository Open Service Interface Definition (OKI DR OSID)	OSIDs in general provide contracts between software systems. The DR OSID addresses issues of storage and retrieval of repository assets.
	http://www.okiproject.org/
eduSource Communication Layer (ECL)	An implementation of IMS DRI from Canada that has been used within the LionShare peer-to-peer project
	http://ecl.iat.sfu.ca/

Table 6: Repository interface specifications

Notwithstanding the development of these repository specifications, many interviewees were unfamiliar with them. All but JSR 170/283 have their origins in the learning object repository field and have sought to facilitate detailed interoperability and interaction between end-user services and learning objects, encompassing many of the tasks that have been specific to such objects. For example, LOREnet in Holland is looking to use the DR OSID to achieve interoperability across Dutch learning object repositories. However, in the field of open access, where there is a desire to simply expose repository contents for discovery the depth of interoperability required has been far less, and repository specifications have not been necessary to meet existing needs. The greater depth of interoperability also requires additional effort over implementation of the more lightweight approaches described above, and many open access repository owners cannot provide this.

As with Z39.50, and potentially OAI, as described in section 5.3 there is a risk that detailed specifications may not be implemented to an equivalent level of detail across all the repositories of interest. As such, implementation within a controlled consortium may be more appropriate where communication and agreement are easier to establish. Assistance is needed to help better understand what the specifications can offer. On one level they can be used to help build repository systems as they offer a standard set of services to which a repository should adhere. But reference models such as OAIS for preservation planning[119] are probably a better starting point since they are not related to any specific technology (as JSR 170/283 is, for instance, with Java) and offer a more abstract description of what needs to be built into the repository.

JSR 170/283 is of interest due to its use of a complex object data model underneath the API. Thus, like MPEG-21 DIDL, it offers the chance to provide clear syntax to repositories and their contents that might be of value. Interoperability is, as metadata and packaging standards, limited to syntax, with semantics being a separate issue. Interoperability is also possibly being achieved by hiding problems: the DR OSID provides wrappers around functionality that allow interoperability without enabling this at a deeper level and possibly obscuring needs and issues at that level. This approach works well for legacy systems and repository specifications may offer a valuable route to include such systems within wider federations.

It is notable that in general and even in the field of learning object repositories implementation of the specifications has been low. Wide adoption is required for useful interoperability, and this may be more applicable within local, controlled and defined scenarios rather than globally due to the relative complexity of the specifications and the effort required for implementation and maintenance. The specifications have their role, but it is a role that needs to be tightly defined to justify the effort required for implementation.

[119] Open Archival Information System Reference Model, http://nssdc.gsfc.nasa.gov/nost/isoas/

A5.6 Web services interfaces

In discussion of interfaces available for repositories so far in this report web services have not been specifically mentioned. However, they are important components in providing services across repositories and are already being used. Many of the interface standards described can be considered web services, albeit at different levels. The development of SRW/U was specifically related to making the functionality of Z39.50 available over the Web, and both versions (SRW and SRU) are web services, SRW using SOAP for its messaging, whilst SRU is a REST-based web service. RSS and ATOM are also web service protocols, designed to pass XML messages between producers and consumers of the web service. OAI-PMH can be considered a web service for the same reason, though may not strictly fit in with some definitions. The repository specifications can be used in a web services environment.

The focussed functionality provided by these interfaces, and their predominant uni-directionality, means they are limited in their capacity as web services. For example, RSS is good for alerting, but additional services around this would be required for more complex interactions, for example the use of Simple Sharing Extensions to enable bi-directional communication. It is also notable that high quality metadata is vital for web services to work well. But it may be possible to build on top of the lightweight starting point to facilitate more complex interactions in a flexible service-oriented way. The flexible approach potentially conflicts with the aims of the repository specifications, which could be considered too heavyweight. Further experience of using different interfaces in a web services environment is required to fully assess the pros and cons of both approaches.

Web services can facilitate true push of content from one system to another. RSS, although considered a push mechanism, has to be initiated by the aggregator or end-user service, but web services can potentially allow content to be pushed based on other criteria that are enacted automatically. As with the repository specifications it is valuable if both ends of a link (repository, service/application) are known if complex interactions are to take place effectively. Web service protocols are still relatively immature, preventing wide applicability outside of controlled environments. The California Digital Library makes use of web services, for example, though predominantly on an internal basis for now. Where they do use them they try to avoid dependencies on SOAP to better guarantee sustainability through technology changes.

Additional web services are also emerging to support interaction with repository software. DSpace has developed a set of REST-ful web services labelled the 'lightweight network interface' that allow deposit, search and withdraw functionality[120]. This was designed to enable academics to deposit materials in the repository, and then extract the associated metadata for presentation elsewhere through a departmental web page or similar. The interface is being

[120] DSpace Lightweight Network Interface, http://wiki.dspace.org/LightweightNetworkInterface

used to motivate content owners to add their own value on top of what the repository itself can offer them. The Fedora digital repository system relies on four separate web service API interfaces (API-M, API-A and Lite versions of both) for all interaction with the repository[121]. Web services can also be used to communicate between the repository and other enterprise systems. The University of Edinburgh has collaborated in the IRRA project[122] to develop a web service that facilitates interaction between their DSpace repository and local research information systems: they are also looking to link the repository to the local personnel system in the same way.

In addition to these specific web services there has also been a recent trend to use lightweight approaches to enable greater interaction on the Web and between the Web and the desktop: the Web 2.0 approach.

A5.7 Web 2.0

Web 2.0 [O'Reilly, 2005] as an approach favours the ability to interact dynamically over the web and allow the flexible re-use of content in different scenarios. These principles, although newly encapsulated under the Web 2.0 banner, have existed for some time and have underpinned many developments described here already. They also match the open access community's desire to expose content for wide use. The interface standards listed above can all enable dynamic interaction and re-use of content: OpenURL allows dynamic linking, whilst OAI-PMH and RSS allow aggregation of metadata that can be re-used for different purposes. Aggregation provides a platform from which end-user services can be built. Google and Amazon to name two major services have recognised this and use the massive aggregations of information they hold to both deliver added value services as well as allow others to do likewise through exposing their APIs. These well-defined connection points allow flexible combinations.

Maximising the potential of lightweight interfaces like OAI-PMH and OpenURL will involve combining them to best effect: a major technology being implemented under the Web 2.0 banner is AJAX [Garrett, 2005] is itself a combination of separate but pre-existing protocols. Allowing interaction between the web and the desktop, as occurs with RSS feeds into RSS reader applications, will also facilitate interaction where the end-user will find it most valuable. Embedding COinS[123] in web pages to automatically generate OpenURL links is another example of how existing standards can be used in flexible ways to facilitate interaction for the end-user.

The ability to interact offers the end-user the opportunity to add to the metadata or content they find. Tagging and annotating can help the individual end-user re-

[121] Fedora Access and Management Web Services, http://www.fedora.info/definitions/1/0/api/
[122] Institutional Repositories & Research Assessment (IRRA) project, http://irra.eprints.org/
[123] OpenURL COinS: a convention to embed bibliographic metadata in HTML, http://ocoins.info/

use resources at a later date, but can also, through services such as del.icio.us and Amazon, allow others to benefit. This creation of additional metadata does require the capability of recording and storing it in a clearly defined way (requiring identification of this new sub-component), but also leads to richer repositories that can act as the basis for other, possibly personalised, end-user services.

The Web 2.0 approach offers many opportunities. The wonderful presentation and end-user services that it promises need stiff underpinning, though, to bring them about, particularly in the academic environment. There are skills issues, the need to ensure end-user services meet accessibility requirements, and the need to identify clear ways to move the complexities of academic information and content into the “evolving super-simplified world”. Web 2.0 is also a reflection of what we could do with the building blocks we have had for many years but which haven’t been fully exploited.

A6 Architecture

Having previously considered the specific issues relating to metadata and interfaces and the impact they can have on establishing end-user services it is necessary to also consider how different components within the ‘repository to end-user service’ chain are put together to ensure viability and sustainability. There have been specific instances of overall repository architectures proposed recently, notably the aDORe and CORDRA initiatives, and these are considered alongside generic architectural aspects to consider when planning both repositories and services across these.

A6.1 Architectural options

There are a number of architectural approaches that can be taken to organise the different components within the repository to end-user service chain. Table 7 briefly lists these.

Architectural model	Notes
Centralised	Aggregating all content and metadata into a single central point to which end-users services are directed
Distributed	Content and metadata are stored in a distributed fashion and end-user services interact with them as required on the fly
Harvested	Metadata, and possibly content, are aggregated to a central point on a periodic basis through a pull mechanism to support end-user service access, with end-users passed through to the originating repository for greater detail following initial interaction
Push	Metadata, and possibly content, are aggregated to a central point on a periodic basis through a push mechanism to support end-user service access, with end-users passed through to the originating repository for greater detail following initial interaction
Peer-to-peer	Individual repositories interact with each other and end-user services on an equal basis

Table 7: Architectural options for organising the repository to end-user service chain

In contrast to the peer-to-peer approach, the other four approaches are hierarchical to a greater or lesser degree in their organisation, with the central aggregator or end-user service at the top of hierarchy. In peer-to-peer all components are equal and interact with each other without hierarchical organisation.

Each approach has its pros and cons. Two main issues that need to be addressed are latency and scalability. Where there is a high level of federation amongst the repositories within the overall architecture there is an increased chance of latency of content taking place. This is as much a matter of how the federation is organised as it is the architecture, but the latter can also have an impact. Federation also affects scalability and the performance of systems.

Architectural model	Latency	Scalability
Centralised	Dependent on organisation of hierarchy in place to formalise centralisation. Low in theory, potentially high in practice.	Potential scalability issues related to size of the centralised store: a large store may slow interaction times.
Distributed	Low latency as repositories are only contacted when needed and always have the most up-to-date information.	Experience of distributed searching has demonstrated scalability problems, e.g., SRW is probably best with no more than 10 targets. Best for focussed searching.
Harvesting	Potential latency as repositories can update between harvests, though lowering the period between harvests can reduce this.	Low scalability issues where only metadata is being harvested: this will increase if content is also included.
Push	Latency unlikely as repositories push out information when new, though possible latency where aggregators/end-user services are slow to ingest pushed information	Low scalability issues where only metadata is being pushed: this will increase if content is also included.
Peer-to-peer	Potential latency where nodes are widely federated due to time for information to spread between nodes.	Ongoing debate on scalability issues for peer-to-peer – unknown factor requiring additional investigation.

Table 8: The impact of latency and scalability on architectural options

A key factor to bear in mind is where the work to maintain relevant systems is situated with these options. Centralised and harvesting approaches require a relatively high centralised workload: push, distributed and peer-to-peer approaches distribute the workload to the repositories more (though note the latter may require central coordination to ensure nodes adhere to accepted practices). Where the workload is placed needs to be balanced against having the required level of coordination in place to ensure end-user services can deliver appropriately.

Each approach has its role and place. A centralised approach allows a greater level of action on the content itself, such as preservation or content enrichment, whilst a harvesting approach potentially allows this if content is also being harvested, but is more focussed on metadata and offers a lightweight means of creating an aggregation upon which end-user services can be built. Push offers a reversal to the pull mechanism of harvesting, though one where repositories themselves would exert greater control over the metadata that is being exposed. A distributed approach is primarily geared towards discovery services. Peer-to-peer is potentially very powerful in enabling a range of actions on the content, particularly back-up, as well as supporting discovery and related end-user services. It also potentially requires the most robust implementation of the supporting architecture to underpin this. Experience within the SPIRE project[124] on the implementation of LionShare peer-to-peer system[125] has highlighted some of the issues that can be encountered.

In the context of open access repositories the harvesting approach offers the best range of functionality to support both access to and management of the metadata and content held in these repositories. Push offers a good alternative, but is not supported through current standards: RSS, push in perception but pull in its technical implementation, offers much of the same capability. The peer-to-peer also offers potential as an alternative where a more controlled level of open access distribution is required. Indeed it would be valuable if a repository could be made available for both harvesting/push exposure for wide exposure and also peer-to-peer exposure for more controlled, and possibly more detailed, controlled exposure. The centralised approach, one adopted by web search engines, can also work, though needs the large resources that the commercial web search engines can provide to make it work effectively. The distributed approach is not best suited as a sole approach to support end-user services across open access repositories, but needs to be considered when placing these repositories in the wider information environment alongside other information sources: in this context, distributed searching across these may well be warranted. Nevertheless, initial aggregation of the open access repository metadata can help limit the scalability issues the distributed approach has.

A6.2 Intermediary shared infrastructure services

Repositories, aggregators and end-user services can be considered the three main links in the information chain between repositories and end-users. The interactions between these can be supported by additional services that act as intermediaries between these links, and which can be shared across them as required. Many of these have been mentioned elsewhere in this document, but are described here in the specific context of the value they can add to these interactions. Many such services have been suggested and a list of these is given in Table 9.

[124] Secure Personal Institutional and Inter-Institutional Repository Environment (SPIRE) project, http://spire.conted.ox.ac.uk/cgi-bin/trac.cgi

[125] LionShare P2P project, http://lionshare.its.psu.edu/main/

Intermediary service	Role	Value
Service registry	A listing of available services (interfaces) onto repositories and aggregators	Allows aggregators and end-user services to find out what repositories are available for interaction with and how these interactions can take place
Collection registry	A listing of information about collections held within repositories	Allows aggregators and end-user services to find out what collections are held within repositories and adjust their interaction around these
Format registry	A listing of formats for objects and information related to these	Provides aggregators and end-user services with information about file formats and how to manage these
Licence registry	A listing of licences that might be used in an open access, or other, environment	Allows repositories to link metadata and content to standard licences and allows aggregators and end-user services to know the terms under which metadata and content can, or can't, be used
Metadata schema registry	A listing of metadata schemas, and possibly application profiles, in use	Allows repositories to identify appropriate metadata schemas and application profiles for local use. Potentially allows mapping between schemas to facilitate interoperability between repositories. Can be used by aggregators and end-user services to structure interaction with repositories.
Metadata crosswalk services	Services that facilitate mapping across metadata schemas to assist in structuring and broadening searching	Facilitates interoperability between repositories and can also allow aggregators and end-user services to structure their interaction with repositories through relevant mapping. This service could be a part of a schema registry.
Terminology/Subject authority service	A specific service mapping across terminologies and acting as an authority for subject terms to assist in structuring and broadening	Akin to metadata crosswalk services, terminology services allow the structuring and expansion of interaction, primarily discovery, with repositories by mapping between different subject vocabularies and promoting

	searching	authoritative terms.
Name authority service	A service that holds an authoritative list of author names for mapping against to support content management and access	An equivalent service to terminology services, but focussing on author names. Could be centralised or distributed according to the location of the appropriate name information.
Identifier resolution services	Services that assist in resolving identifiers to assist in the location of objects	Allows identifiers to be resolved to actual locations of objects within repositories.
Relationship information service	A service to help establish the relationship between objects that are distributed	Allows aggregators and end-user services to be informed of relationships between objects and structure subsequent interaction with repositories
Format conversion services	Services converting objects between formats	Allows objects to be converted between formats, either on-the-fly for delivery or as part of a preservation process
Metadata generation services	Services that assist with automatic generation of metadata to enrich repositories and aggregations	Allows repositories or aggregators to enhance the metadata they hold to facilitate presentation through end-user services
Normalisation services	Services that assist in standardising metadata held in repositories to assist with discovery	Allows aggregators to clan and organise the metadata they have aggregated to facilitate presentation through end-user services
Ratings/annotation service	A service to allow additional user-generated metadata to be appended to existing records	Allows end-users to append metadata to objects they are interacting with and store these in association with the objects

Table 9: A suggested list of possible intermediary services to support end-user services across open access repositories

These intermediary shared infrastructure services will predominantly sit between the three main components of the repository to end-user service chain and interact with them through machine-2-machine interfaces. Existing developments suggest that there is value in providing user interfaces onto them as well, for example the user interface to the Information Environment Service

Registry[126]: it remains unclear, though, whether such user interfaces are valid as end-user services in their own right or whether the demand is related to the ability to ‘see’ how such an intermediary service operates. This latter need is valid in its own right and can potentially demonstrate how intermediary services can work between repositories and aggregators/end-user services. Such demonstration can build confidence that the interoperable interface standards used to present the machine interfaces can achieve the desired functionality and benefit. OpenURL resolvers, a specific instance of an identifier resolver, are positioned in the presentation layer within the JISC Information Environment, and are an example of how an intermediary service can perform this role and that of an end-user service. Resolvers offer an element of choice to the user (where would you like to direct the OpenURL) and positioning intermediary services as end-user services will be valid where this aspect of user choice is of value.

When working as intermediaries, these shared infrastructure services can assist in building bridges between different repositories, and between repositories and aggregators/end-user services. This is an especially important role for service and collection registries where establishing such links may not be that straightforward (often simply due to lack of awareness of the components available). Introducing intermediary services into the interaction between repositories and aggregators/end-user services can enable and make the relationship more effective. The stability and potential dependency on such intermediary services, though, needs to be monitored and developed with care, as end-users are not tolerant of service failure if components they can’t see prevent them from working at their expected level of functionality.

The relationship between intermediary services themselves also requires further investigation, in order to make most effective use of them as a whole. This applies to the relationship between like services (for example, between different service registries such as IESR and OpenDOAR) as well as between unlike services (for example, between service registries and licence registries or between terminology and authority services). Such links may be built into the same service, for example IESR encompasses both service and collection registry roles.

Intermediary services need to be built using accepted standards to order to maximise their capability of interoperation between themselves and repositories/aggregators. There has been little testing of the benefits of intermediary services in practice, though, and work on the combination of open access repositories with such services is warranted.

[126] Information Environment Service Registry (IESR) web search interface, http://www.iesr.ac.uk/service/iesrsrch?type=new

A6.3 Authentication and authorisation

As indicated in earlier sections the open access environment in principle does not require authentication and authorisation. There are occasions, however, where restrictions are required and mechanisms are required to control access to some degree. There is also the need to consider the involvement of open access repositories within the wider information environment and the possibility of interaction across both open and closed access information resources.

In placing restrictions on access there is a need to consider the granularity at which such restrictions will apply. Will individuals be granted access dependent on their role or will access be controlled at a group level? This granular approach can also apply to the objects within the repository, which may have different levels of access associated with them. Shibboleth[127] offers the possibility to implement both levels of authorisation, though institutions need to have appropriate identity management infrastructure in place to take best advantage of this and provide initial authentication. As well as access, authentication will also be required for management activities such as deposit and deletion, particularly in institutional repositories. In planning relevant authentication and authorisation systems it is important to underpin these with relevant models to ensure they meet requirements. Security and identity models exist, but it is unclear whether these are detailed enough in their scope to meet requirements.

Notwithstanding the need or lack of it for authentication and authorisation in an open access environment there is a need to investigate broader trust and provenance issues and models to better understand how these factors affect open access dissemination. This is not to say that open access does not work without these – the evidence is, of course, very much that it does – but that in order to develop further and deeper open access services to additional information and data these are issues that will need taking into account to ensure open access itself is fully trusted over time.

A6.4 Service-oriented architecture

There is much current interest generally in planning systems around the use of a service-oriented architecture (SOA) and the combination of individual service components that this promotes. A SOA approach offers a level of flexibility that it is valuable to work towards when considering the provision of services across repositories. Repositories, aggregators, end-user services, and intermediary services can all be considered as individual service components, each offering different service interfaces. Flexible interaction between these offers the promise of exposing repository metadata and content to better meet end-user needs. Considering the architectural options presented in section 6.1 a SOA approach

[127] Shibboleth, http://shibboleth.internet2.edu/

may be adopted or it may not by any of them: evidence from this study suggests a SOA approach is highly favoured over the long-term.

SOA is not an architecture that can just be implemented, though, but is an approach that requires analysis of what you want systems to achieve and then a design phase to establish what service components will be required. Hence, it is an architecture and approach that it is valid to work towards over time. This approach is affected by the need for all components to adhere, and adhere strictly for best effect, to relevant standards, an aspect that requires coordination in a cross-institutional federated environment.

The need for coordination has informed the ongoing development of the e-Framework for Education and Research[128]. This initiative is supported jointly by the JISC in the UK, DEST in Australia, with partners in the USA, Canada and New Zealand. This multilateral approach recognises both the long-term nature of moving towards SOA whilst also seeking to provide a coordinating framework within which this move can take place.

SOA increases the possibilities of interoperability whilst also providing for greater sustainability through the flexible replacement of individual components. It moves toward a scenario where repository content and services are "loosely organised over relatively consistent plumbing to support lightweight and user-shaped collection and posting and organising capabilities"[129]. Notwithstanding the gradual process over which SOA needs to be implemented, there are examples of organisations currently moving in the direction. The California Digital Library has establishing a plug'n'play approach to adding additional functionality as required over time. SOA has also been taken up within the learning & teaching-focussed LeAP initiative in Tasmania[130]. Learning & teaching has traditionally used monolithic systems: LeAP broke this idea up by instituting a range of component systems that allowed for far greater granularity. This granularity subsequently allowed greater flexibility in how systems were put together to meet user needs, in particular allowing different end-user views onto the same content.

Although there are a number of repositories within the LeAP architecture, a granular architecture does not preclude the presence of a single centralised store: the architecture may be an inverted pyramid with services springing off a central repository store, and indeed there are advantages to storing content once and using it many times (a WORM approach). Aggregations are also single stores from which many services may emerge. A granular approach does allow other repositories to be encompassed in the overall architecture, though, including personal content stores.

[128] e-Framework for Education and Research, http://www.e-framework.org/
[129] Jerry Persons, Stanford University, personal communication
[130] See report at http://www.jisc.ac.uk/index.cfm?name=altilab_papers

SOA, which is very Web 2.0 in its nature, is a valuable long-term goal, and LeAP shows it can work in an educational context. As different components are separated out and the number of layers in the architecture increases there is a need to understand both how these layers interact, and how services between the layers are orchestrated to allow smooth and effective interaction. Greater understanding in these areas is required, though will also enable an architecture that can adapt as required to suit end-user needs effectively.

A6.5 Repository services layer

Many interviewees reflected that they did not consider that overall architecture should impact on repositories or how they are configured. Repositories act as stores of content and metadata: interfaces on top of repositories, which ideally should be based on open standards, act as the exposed service that interacts with other components. These interfaces should ideally hide the repository from aggregators/end-user services. Interface standards are predominantly implemented at individual repositories themselves (e.g., OAI-PMH interfaces built into repository software). If treated as a separate layer this repository interfaces or repository services layer could potentially, though, be used in an agnostic way as an intermediary service serving a range of repositories.

Such a repository services layer was proposed as a generic architectural approach by Sayeed Choudhury and his team at Johns Hopkins University as part of their Technology Analysis of Repositories and Services project[131]. This separation between repositories and aggregators/end-user services means that both parties can focus on interacting with the services layer. Having a common services layer, and ideally a small number of appropriate APIs would be best to encourage take-up, allows interoperability across repositories to be established. It also allows a SOA approach where components can be exchanged and/or replaced as required so long as they are able to communicate with the services layer. Repository specifications such as the DR OSID can act in this way, as does an OAI Static Repositories Gateway[132]. These allow repositories to limit the interactions they need to manage to just those with the specifications themselves, and remove the need to manage contact with aggregators and end-user services: the specification acts as a middleman and broker between the two.

This focus of attention allows repositories to concentrate on managing content rather than access to it. They need to be aware of the demands and requirements end-user services have so they can meet these (end-user services can’t build something if the repository doesn’t support it), but can leave the management of access to the intermediary services layer. This separation between repositories and other components also reduces the dependencies

[131] A Technology Analysis of Repositories and Services, https://wiki.library.jhu.edu/display/RepoAnalysis/ProjectRepository

[132] OAI Static Repositories specification, http://www.openarchives.org/OAI/2.0/guidelines-static-repository.htm

each has on the other. At the object level this can extend to the format of the objects themselves, though this is often embedded in any interaction without scope for alteration. Fedora has the capability of allowing multiple formats to be both stored and disseminated on the fly, though this requires work and careful planning when setting up the repository. The aDORe use of the Pathways InterDisseminator is another mechanism that offers potential for removing the object-type dependency and serving formats flexibly as requested by the end-user.

The separation of end-user services from the underlying component can apply as much to aggregators as to repositories. The development of an SRU interface to OAIster suggests a similar aggregator services layer may be of value in allowing aggregators to better manage their task (aggregation) whilst exposing their collections in a controlled way.

A6.6 Repository integration in local infrastructure

Repositories cannot be considered in isolation, but must be placed in the context of wider infrastructure. At a theoretical level there has been some activity in this area already, with reference models such as OAIS and the DLF Services Framework[133] offering guidance on how to facilitate such integration. On a practical level within an institution this involves positioning the repository, or repositories, in the context of, for example, content management systems, virtual learning and research environments, current research information systems, and institutional portals. The role of the repository will influence the level of integration and interaction required. For example, an open access repository could be an archive of research outputs that is used only at the end of the research process: integration required at this level would be low. Where a repository is being used for multiple content types and as an everyday working tool then greater integration is required to allow it to take on this role. Integration may also be focussed at the presentation level for end-user interaction (e.g., presenting a search or deposit screen within a portal) or can be at the data level for the exchange of information between systems.

To encourage repository use at a local level it will be of value to take the repository to the user rather than require the user to find the repository. This can be achieved by embedding access to the repository through other systems (which also prevents the repository being seen as just ‘another system’ to get used to). Surfacing repository services through an institutional portal, VLE or library catalogue, for example, possibly through the use of portlets and/or Web services [Awre, 2005], may be appropriate routes: it would also be useful if the repository could be accessed from within Word and other commonly used editing software to allow easy access to get, edit, save and delete functions.

[133] DLF Services Framework, http://www.diglib.org/architectures/serviceframe/

The overall architecture in place within an institution will have a major bearing on the ease with which repositories can be integrated. If the architectural layers are kept separate through the use of a SOA approach or similar and good well-defined interfaces built between these then additional components such as repositories can be slotted in as required. Notwithstanding this, other systems may have a knock-on effect on how repository functionality is designed and made available. For example, ingest design may be driven by a VLE need to ingest and disassemble learning objects.

Different systems placing different requirements on the repository will add to the complexity to be dealt with. The open standard interfaces described in earlier sections can be used for interaction with administrative systems as with any other, assuming these systems can make use of such interfaces. These may be too lightweight in nature for the type(s) of interaction required and the use of the more detailed and structured repository specifications (see section 5.3) can offer a greater level of functionality. This situation promotes the case for a separate repository services layer, which the repository specification may provide, to act as broker in between and simplify the interactions each component has to deal with.

Where a repository takes on a wide content role it is valuable if it can act as the feed of content into whichever systems and services it underpins, both internal and external. Clear content identification is vital for effective interaction. In principle clear identification can allow content to sit almost anywhere so long as identifier resolution services are in place to resolve identifiers: until this is available a repository can bring dividends by bringing content together and encourage crossover between different parts and roles within an institution: for example facilitating the use of content for both teaching and research roles, a crossover that might not have been identified previously.

References

Awre, C., Waller, S., Allen, J., Dovey, M.J., Hunter, J. and Dolphin, I. Putting the Library into the Institution: Using JSR 168 and WSRP to Enable Search within Portal Frameworks. Ariadne, October 2005, Issue 45. Available at http://www.ariadne.ac.uk/issue45/awre/

Bekaert, J., Hochstenbach, P. and Van de Sompel, H. Using MPEG-21 DIDL to represent complex digital objects in the Los Alamos National Laboratory Digital Library. D-Lib Magazine, November 2003, 9 (11). Available at: http://www.dlib.org/dlib/november03/bekaert/11bekaert.html

Bibliographic Services Task Force. Rethinking how we provide bibliographic services for the University of California: final report. December 2005. Available at http://libraries.universityofcalifornia.edu/sopag/BSTF/Final.pdf

Blanchi, C. DVIA Registry Architecture. Presentation given at the 2006 Defense Technology Information Center Conference.

Campbell, D. ARROW Discovery Service Harvesting Guide. June 2005. Available at http://arrow.edu.au/docs/files/harvesting.pdf

Duke, M. Delivering OAI records as RSS: an IMesh Toolkit module for facilitating resource sharing. Ariadne, October 2003, Issue 37. Available at http://www.ariadne.ac.uk/issue37/duke/

Foster, N.F. and Gibbons, S. Understanding faculty to improve content recruitment for institutional repositories. D-Lib Magazine, January 2005, 11 (1). Available at http://www.dlib.org/dlib/january05/foster/01foster.html

Garrett, J.J. Ajax: a new approach to web applications. 2005. Available at http://www.adaptivepath.com/publications/essays/archives/000385.php

Godby, C.J., Young, J.A. and Childress, E. A repository of metadata crosswalks. D-Lib Magazine, December 2004, 10 (12). Available at: http://www.dlib.org/dlib/december04/godby/12godby.html

Halbert, M. DLF Aquifer Study on Institutional User Services.

Heery, R. and Patel, M. Application profiles: mixing and matching metadata schemas. Ariadne, September 2000, Issue 25. Available at:
http://www.ariadne.ac.uk/issue25/app-profiles/
http://www.oreillynet.com/pub/a/oreilly/tim/news/2005/09/30/what-is-web-20.html

Jerez, H., Manepalli, G., Blanchi, C. and Lannom, L.W. ADL-R: the first instance of a CORDRA registry. D-Lib Magazine, February 2006, 12 (2). Available at http://www.dlib.org/dlib/february06/jerez/02jerez.html

Lagoze, C. and Van de Sompel, H. The making of the Open Archives Initiative Protocol for Metadata Harvesting. Library Hi Tech, 21 (2): 118-128.

Lynch, C.A. Institutional repositories: essential infrastructure for scholarship in the digital age. ARL Bimonthly Report 226, February 2003. Available at http://www.arl.org/newsltr/226/ir.html

Lyon, L. eBank UK: building the links between research data, scholarly communication and learning. Ariadne, July 2003, Issue 36. Available at http://www.ariadne.ac.uk/issue36/lyon/

Manepalli, G., Jerez, H. and Nelson, M.L. FeDCOR: an institutional CORDRA registry. D-Lib Magazine, February 2006, 12 (2). Available at http://www.dlib.org/dlib/february06/manepalli/02manepalli.html

McCown, F., Liu, X., Nelson, M.L. and Zubair, M. Search engine coverage of the OAI-PMH corpus. IEEE Internet Computing, March/April 2006, 10 (2): 66-73. Preprint available at http://library.lanl.gov/cgi-bin/getfile?LA-UR-05-9158.pdf

Nixon, W.J., Drysdale, L. and Gallacher, S. Search services at the University of Glasgow: PKP Harvester and Google. DAEDALUS project report, 2005. Available at https://dspace.gla.ac.uk/handle/1905/425

O'Reilly, T. What is Web 2.0: Design Patterns and Business Models for the Next Generation of Software. 2005. Available at

Powell, A., Day, M. and Cliff, P. Using simple Dublin Core to describe eprints, version 1.2. 2003. Available at http://www.rdn.ac.uk/projects/eprints-uk/docs/simpledc-guidelines/

Sanderson, R., Young, J. and LeVan, R. SRW/U with OAI: expected and unexpected synergies. D-Lib Magazine, February 2005, 11 (2). Available at http://www.dlib.org/dlib/february05/sanderson/02sanderson.html

Tourte, G and Powell, A. Encoding full-text links in the eprint jump-off page, version 1.0. 2004. Available at http://www.rdn.ac.uk/projects/eprints-uk/docs/encoding-fulltext-links/

Van de Sompel, H. and Beit-Arie, O. Open linking in the scholarly information environment using the OpenURL Framework. D-Lib Magazine March 2001, 7 (3). Available at: http://www.dlib.org/dlib/march01/vandesompel/03vandesompel.html

Van de Sompel, H. and Lagoze, C. The Santa Fe Convention of the Open Archives Initiative. D-Lib Magazine, February 2000, 6 (2). Available at http://www.dlib.org/dlib/february00/vandesompel-oai/02vandesompel-oai.html

Van de Sompel, H., Bekaert, J., Liu, X., Balakireva, L. and Schwander, T. aDORe: a modular, standards-based digital object repository. The Computer Journal, 2005, 48 (5): 514-535. Preprint available at http://arxiv.org/abs/cs.DL/0502028

van Veen, T. and Oldroyd, B. Search and retrieval in The European Library: a new approach. D-Lib Magazine, February 2004, 10 (2). Available at http://www.dlib.org/dlib/february04/vanveen/02vanveen.html

van Westrienen, G. and Lynch, C.A. Academic institutional repositories: deployment status in 13 nations as of mid-2005. D-Lib Magazine, September 2005, 11 (9). Available at http://www.dlib.org/dlib/september05/westrienen/09westrienen.html

Waters, D.J. The metadata harvesting initiative of the Mellon Foundation. ARL Bimonthly Report 217, August 2001. Available at
http://www.arl.org/newsltr/217/waters.html

Xiang, X, and Lease Morgan, E. Exploiting "light-weight" protocols and open source tools to implement digital library collections and services. D-Lib Magazine, October 2005, 11 (10). Available at
http://www.dlib.org/dlib/october05/morgan/10morgan.html