LINKING UK REPOSITORIES:

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

SCOPING STUDY REPORT

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

CONTENTS

EXECUTIVE SUMMARY

The JISC commissioned the project partners to undertake a scoping study whose aim is to identify sustainable technical and organisational models to support user-oriented services across digital repositories. Open access repositories of interest to UK further and higher education communities were cited as having particular relevance. The study is intended to inform strategies to support access and use of repositories, with a view to the establishment of a national repository services infrastructure or framework.

User requirements and lessons from existing studies

Users and their requirements are summarised as follows:

- **Repository managers**: their needs may include; help to make a business case for a repository within their institution; advice on IPR and copyright issues, on building and maintaining a repository, and on technical issues concerned with running a repository (digitisation, file formats, metadata structure, preservation, data exposure [e.g. OAI standards], name authority systems); access and authentication systems; repository usage services; and help on managing advocacy and providing local repository services to the end user community.
- **End users**: As **searchers** they require resource discovery tools and value-added content. As **content providers**, they need require somewhere to deposit their work; peer review services; and they share with repository managers the need for usage and impact services, technical advice or assistance, and advice on rights and IPR.
- **Content aggregators** primarily require accurate and adequate metadata.
- **Meta-users** (people who use Open Access repository content for analytical work): these people share with the repository managers and end users the requirement for repository usage statistics; they also need tools for research assessment and monitoring.
- **Entrepreneurs** (people who build services upon repositories, such as publishers, re-sellers, and technology transfer specialists): these people need good resource discovery tools, bridging services (navigational/locating tools for identifying repositories, their characteristics and what they hold) and technology transfer expertise

On the basis of the user requirements analysis we have constructed an overall scheme for repository services. This has services located at three main levels. At the ***ingest level*** are the services that cater to the technical and process-based needs of repository managers and depositing authors. At the ***content aggregator level*** are the metadata-production and enhancement services with their associated technologies. Above the aggregator level, at the ***output-level***, are the services that work on repository content, providing for specialised preservation needs, research assessment and monitoring, resource discovery, publishing, overlay journals, meta-analysis and (bridging services).

A number of lessons and insights have been identified from previous or ongoing studies and from expert opinion that would have significance in any scheme that links UK repositories. The main ones are:

At the ingest level: technical capabilities vary widely across institutions as a result of which there is huge variation in the quality of metadata provided by repositories, in the preservation activities being undertaken at repository level, and in the systems in place to capture content. The amount of content in repositories also varies hugely: advocacy work to the author community is critically important in raising the levels of deposition of research postprints. IPR and copyright remain major stumbling blocks in this respect. Some of these obstacles can have a strongly discouraging effect on repository managers seeking progress. They also mean that the volume of Open Access material available for services to use remains low.

At the aggregator level: metadata quality – or even metadata provision itself – remains the major issue. The technical model proposed in this study describes the optimal approach in this respect (see below).

At the output level: specialised resource discovery tools are important and provide a route into repositories for users with specific needs, though users may enter repositories in various other ways, too. They may search a specific repository because they are looking for something they know will be located there. They may use subject-based portals if they are searching a specific discipline or topic, or portals that aggregate repository content by object-type of interest (such as moving images or theses). In many cases they will arrive at repository content via Google or other web search engines. Repository managers and authors also value the exposure that Google and other web search engines bring to their content and these and it is desirable that these be factored into a national scheme, too.

Candidate services and an organisational model for a UK national linked-repository landscape

The elements of a national linked-repository landscape and the candidate services that would be needed are identified as:

Ingest level:

- digitisation services
- services that provide advice on IPR and rights
- services that provide advice and advocacy materials on Open Access
- services that provide help on technical issues
- repository construction services
- repository hosting services

Data level:

- institutional repositories
- national-level 'catch-all' interim repositories for authors with no institutional repository
- subject-specific repositories gathering primary content
- media-specific repositories gathering primary content

Aggregator level:

- Metadata creation and enhancement services

Output level:

- access and authentication services
- usage statistics services
- preservation services
- research assessment and monitoring services
- resource discovery services
- publishing services
- overlay journal services
- meta-analysis services
- bridging and mapping services
- technology transfer/business advice services

With the exception of the very last point in the list, all these activities are carried out to various degrees by existing services or projects, or are currently being scoped. Many are operating in bounded, discrete areas or as demonstrator projects, however. Scaling up such pilot or project-level activities to the level required for a workable national scheme will require careful planning and a strong leadership role from the JISC.

The following have been identified as top priorities:

- interim 'catch-all' repository (or repositories) for authors whose institution does not yet have a repository
- national resource-discovery service
- meta-analysis services, specifically citation analysis and bibliometric analysis services that can inform future national research assessment exercises
- repository usage and statistics services

Second-level priority should be given to:

- preservation services working across areas not already benefiting from specialised services
- a national name authority service
- a national file format/conversion service

A technical model

An aggregation model is proposed to support the development of end-user services. This model builds on previous recommendations of harvesting using the Open Archives Initiative Protocol for Metadata Harvesting (OAI-PMH) as a preferred technical approach. However, the breadth of potential content within open access and other repositories required both a closer examination of OAI-PMH's capabilities and the identification of additional standards/technologies that might be used to achieve the same ends.

Aggregations bring metadata and potentially content together as a basis for end-user services and negate the services having to deal with many individual repositories. Aggregations offer greater control over the data so this basis is a stable one, whilst leaving full control over the content to the originating repository. Regular aggregation permits efficient and up-to-date access for end-user services to build on. Most valuably, aggregations allow re-factoring of the metadata/content to make it better suited for supporting end-user services than working across individual repositories.

Metadata and content

At the heart of all end-user services across repositories is good quality metadata about the digital content repositories hold. Automated generation of metadata in all its forms is an area requiring additional activity, but also lateral thinking to identify where metadata can be generated. Generation should lead where possible to the creation of a rich base metadata record that can be used by a repository internally whilst acting as the basis for externally-facing metadata formats for exposure to aggregators.

The exposure of content for aggregation is less well understood than exposure of metadata. There is a need to model exactly what we wish to do when exposing content so that the technology best suits these requirements. A modelling approach will also be of value in determining the use and granularity for assigning identifiers to digital content and relevant sub-components so that aggregators/end-user services can clearly identify what they are working with.

Repository interfaces

OAI-PMH is used extensively to facilitate access across repositories through OAI service providers. The use of sets and other protocol containers can enhance its value and allow service provider aggregations to better focus what they offer end-user services. RSS and ATOM newsfeeds are mini-aggregations in their own right from individual repositories, whilst RSS/ATOM readers act as aggregators for newsfeeds from many repositories. Web crawlers aggregate available web page information and present this usually through web search engines. These two alternative approaches offer different paths to enabling an aggregation upon which end-user services can be built.

Aggregation and end-user services
Once compiled, aggregations can act as the basis for metadata generation, and offer a more viable option for the processes involved through economies of scale. Aggregations themselves will not normally act as end-user services directly, but rather provide a range of interfaces to enable end-user services to be built on top. This may involve re-exposure via OAI-PMH, RSS or via a web crawler for further aggregation. Once exposed through an end-user service, though, there should always be access back to the originating repository so that additional functionality can be offered and full content sourced.

Architectural approaches
The three main components of the aggregation model are repositories, aggregators and end-user services. Repositories are likely to be independent of aggregators: end-user services, though, are often closely associated with aggregators, though there is an increasing shift to separating these two (e.g., through Web 2.0 approaches). A shift toward viewing the three components as services can facilitate a move towards a service-oriented architecture that can provide maximum flexibility in how the components are implemented. Two specific instances of how the components can be linked are the aDORe and CORDRA initiatives. Both promote the concept of exposing as rich a metadata set as possible to facilitate aggregation and the development of end-user services across repositories. aDORe has practically demonstrated many of the CORDRA concepts and both can benefit future development.

Looking ahead
Communication between the components of the aggregation model is vital to the development of effective end-users services. This can underpin the development of more personalised services that end-users themselves require to suit their varied roles within education and research.

Business models for repository services

Repository services might adopt a range of appropriate business models. Here we focus on five:

- institutionally-supported: appropriate for digitisation, repository provision, preservation at some levels and overlay journal production
- publicly-funded (e.g. from top-sliced money allocated by the JISC): appropriate for all advisory services for interim ‘catch-all’ repositories, metadata creation and enhancement, resource discovery, technology transfer and bridging services
- community-supported: appropriate for subject- and media-specific repository provision, usage, assessment, and meta-analysis services and publishing services (particularly where mediated by learned societies)
- subscription-supported: appropriate for access and authentication, preservation and resource discovery services
- fully-commercial models (including advertising-supported, merchant and utility models): appropriate for digitisation, repository provision and hosting, technical

advisory services, metadata creation and enhancement, technology transfer, and all output-level services (access/authentication, usage statistics, preservation, monitoring and meta-analysis services, resource discovery, bridging services, overlay journal production and publishing services

The highest costs are likely to be incurred by preservation and access/authentication services. Resource discovery services and metadata services will have medium to high costs. Repository provision and hosting, digitisation, usage statistics, bridging services and publishing services could operate at a medium-cost level. Advisory services, monitoring and meta-analysis services, technology transfer, subject-specific and interim repositories, and overlay journal services would be expected to be able to operate at relatively low cost.

Recommendations

A number of recommendations are made to the JISC on how the vision might be achieved in practice. Strong overall management activity will be required with a focus on communication and coordination. The various elements of the system will not fall into place in an effective way just by themselves, chiefly because the need for top-sliced funding will be ongoing for some parts of the system. The natural candidate for coordinating developments is the JISC, which has the vision of the desired outcome and the wherewithal to influence and drive developments in the public sector that underpin the whole system. Opportunities for developments that can and should be left to private enterprise can be communicated as part of that overall plan.

Recommendations therefore call for the JISC to take a strong management role, to construct where necessary the appropriate communication channels with the research community, with repositories and with potential service providers, and to coordinate their efforts.

It is also recommended that further investigation is carried out into how information will be used, since the answers to this affect the way services can develop and how repositories should expose their content. Other issues that will require further study are the automatic generation of metadata and the role of identifiers. The adoption of RSS and ATOM as standards for aggregation in addition to OAI-PMH is recommended, since they have distinct benefits for services targeted at specific user communities.

The full list of recommendations made to the JISC is as follows:

1. The research community should be engaged at the highest level to encourage the establishment of repositories in all HE and FE institutions and the development of policies that will ensure the collection of content.
2. Channels of communication with repository managers should be opened, and the establishment of a community encouraged. This may be done through existing structures: the UKCORR is the most appropriate, and the two main

open source repository softwares (EPrints and DSpace) have their own user communities that could also be used for this purpose. The aim is to have clear and effective communication structures in place between JISC and all operating repositories that will facilitate two-way discussion and enable development.

3. Similarly, an interface or contact point between the JISC and actual or potential service providers should be established. This will enable end-user oriented services to be developed in a coordinated and directed way.
4. Developments of repositories, aggregators, end-user services, and intermediary services should move towards a service-oriented architecture and establish separate layers for the aggregation model to maximise the flexibility available for building end-user services to meet user requirements.
5. Development of end-user services includes an element of investigation of how information to be surfaced through these services will be used. This will assist in helping inform the development of the service and feed back to the underlying repositories being exposed through the service.
6. Additional means to generate metadata using automatic means are required. It is recommended that investigations into relevant techniques and tools be taken forward with some urgency.
7. Further attention to identifiers, specifically location-independent identifiers, and necessary resolution systems is recommended to provide greater understanding of their benefits and use.
8. It is recommended that the use of RSS and ATOM be investigated as additional standards to OAI-PMH for use in aggregating metadata and content. They offer the potential of targeted exposure of repository resources that may be beneficial in the development of end-user services targeted at specific communities. It is also recommended that the exposure of repository contents within web search engines be examined in closer detail to assess the paths of exposure that exist and the implications for repositories of exposure via this route.
9. It is recommended that future work to develop aggregators and/or end-user services include an element of communication and involvement with repositories from the start. This will ensure development does not take place in isolation and increase the interoperability between the three major components of the aggregation model. Where intermediary shared infrastructure is involved those developing this should also be included in relevant communications.
10. It is inevitable that for an optimally-structured set of repository services to be developed on UK repositories, there will be a continuing need for top-sliced funding for some parts of the system. The JISC will need to plan for this for the medium-to-long term.

Alma Swan, Key Perspectives Ltd
Chris Awre, University of Hull
5 June 2006

1. INTRODUCTION

1.1 The Open Access landscape in the UK

Whilst approximately one third of UK universities have an Open Access (OA) repository, most of these have very little content in them, and published research articles are conspicuously low in number amongst the working papers, theses and preprints[1]. The vision that underpins this study is for institutional repositories to form the backbone of Open Access provision in the UK, collecting and exposing the research outputs of every research-active institution so that the entire UK research corpus is visible, usable, and manipulable, to the benefit of future research effort. This is a timely, appropriate and achievable vision, but its conversion to reality remains some way off.

Researchers are not aware of Open Access at all, or are aware but not informed of its benefits, or are aware of its benefits but for one reason or another do not provide it for their work. Institutions are waking up to the possibilities and opportunities that Open Access brings for them, but rather slowly: there will need to be both changes in habit on the part of researchers, and in the academic policies and procedures of institutions, if the use of repositories is to become embedded in the culture of research-based organisations. Research funders are starting to see the merits of Open Access and, when funding research with public money, are starting to pay at least lip service to the moral case for the concept. With the single exception of the Wellcome Trust (a private research funder), however, funders have not yet seen the need to develop policies that will bring Open Access about for the research they fund. This is particularly marked with respect to the UK Government, which has declined to implement the recommendations in the House of Commons Science & Technology Select Committee's report on its investigations into the science publishing scene in the UK (HCSTC, 2004). Furthermore, one step down at research council level the opportunity for an RCUK-wide policy, originally proposed in a draft RCUK policy document first published early in the summer of 2005 (RCUK, 2005), now looks to be lost in favour of piecemeal actions by individual research councils.

Nonetheless, there are promising developments on Open Access in the UK in other quarters. Individual universities are pressing ahead with repositories and developing policies designed to fill them and mandates on self-archiving from a number of universities are imminent. JISC-funded projects on repository software development, interoperability, repository establishment, data repositories, legal aspects of OA, preservation, e-learning and machine services continue to push forward our understanding and knowledge of how Open Access can develop and establish good

[1] http://archives.eprints.org/

practice in active developmental programmes. Individuals and groups within the scholarly communications arena advocate and inform and work to develop the tools to advance the uptake of OA and measure its benefits. And scholars themselves are showing signs of responding at last.

This is the moment, then, to look at how to join up all these initiatives and developments into a coherent whole. The JISC's intention is to develop strategies to support access to and use of repositories, with a view to the establishment of a national repository services infrastructure or framework. This study aims to inform the development of those strategies.

1.2 Overview of the study

The JISC commissioned the project partners to undertake a scoping study whose aim is to identify sustainable technical and organisational models to support user-oriented services across digital repositories. Open access repositories of interest to UK further and higher education were cited as having particular relevance. The study is intended to inform strategies to support access and use of repositories, with a view to the establishment of a national repository services infrastructure or framework.

The JISC Digital Repositories Programme has supported, or is currently supporting, a broad swathe of repository-related studies. The Programme encompasses project clusters that cover data, e-learning, preservation, legal and policy issues, machine services and integrating infrastructure. This present study sits in the last of these clusters but has attempted to take an embracing view of the whole Programme as it has sought to model how technical and organisational issues can be brought together into an overarching schema for linking UK repositories of varying types.

The project has four main elements:

- a user requirements study: this was specified to be an overview approach using and distilling existing information rather than a detailed primary study
- a review of the repository and service level players already in operation with respect to organisational requirements for building viable and sustainable repository services on a national scale
- technical architecture and infrastructure: within the JISC Information Environment and related developing technical standards a range of approaches for creating service models are possible. The project examines the limitations of current technical standards with respect to the demands of interoperability of a range of repositories and repository types, and the possible technical solutions in the light of user needs and preferences
- business modelling for future national services: a set of models was required from the project, though not at a fine level of detail, with attention to scalability, viability and sustainability

The Open Access concept is predicated upon free-to-use information and, in establishing the OAI standards, seeks to provide for distributed information sources to be linked interoperably so that maximal amounts of content can be searched whilst the locations of individual items need not be recalled by the user. It has a huge and unquestionably positive importance for the scholarly community, in research, teaching and learning primarily, but also providing the wherewithal for understanding how these processes may work more optimally and, thus, as a management information tool for the academic sector.

Interoperability is the name of the game but it is not simple to implement. All the constituent elements continue to change as technology advances. Nevertheless, the goal is to link repositories in such a way that services can be built upon them that provide value and utility to the user. At present, repository provision is patchy. Some institutions have one but many do not. Some institutions have several repositories, based perhaps at departmental level, or institution-wide but serving several separate purposes - individual repositories for theses, eprints, research data and so on. The vision requires these to be networked effectively so that distinctions become blurred and the emphasis for differentiation can reside at the level of provision of services.

Whatever the model, there are a set of requirements that underpin the concept of linking UK repositories and building upon them a set of services that are to the benefit of the HE/FE community. The outcome must:

- Collect and present information that people want to use
- Make it possible for users to find what they want to use
- Present information in the form that people require
- Be workable within legal, intellectual property and copyright bounds
- Enable resources to be shared, simply
- Provide and use common standards
- Facilitate the deposition of information
- Encourage and motivate authors to participate by depositing their work
- Have overt and identifiable benefits to authors and users
- Be viable and sustainable

Since the specification for this study required a focus on Open Access repositories there is an implicit assumption that, for the most part, content will be free to use. We recognise that there will be some exceptions to this due to the complexity of some collections and where this is the case the model acknowledges it. We also recognise that repository services developed on top of the Open Access content may themselves display an array of business models, some being free-to-use and others paid-for. The goal here is to develop a linking model that is workable, connects Open Access repositories together effectively, and permits service providers to develop their offerings over the whole corpus of UK Open Access material. Attractive services that result will lead to support by those submitting, and will provide a new cadre of advocates for Open Access since the new service providers will have motivation to encourage its adoption.

We approached this exercise by reviewing existing reports and study outcomes and by seeking advice and input from individuals who have experience in the relevant areas. This was done by means of focus groups, by individual interviews in person or by telephone, by email questionnaire and by lots of email discussions. Some people were involved in more than one of these activities and we are grateful to them especially for their forbearance. The list of individuals canvassed during the course of the work is as follows and our thanks go to each of them:

Stephen Abrams, Harvard University
Sheila Anderson, AHDS
Theo Andrew, University of Edinburgh
Stephen Andrews, British Library
Ann Apps, MIMAS
Anne Atkins, Western Colleges Consortium
Simon Bains, National Library of Scotland
Phil Barker, CETIS Metadata and Digital Repositories SIG
Jonathan Bell, University of Wales, Aberystwyth
Kerry Blinco, DEST
Eddie Boyle, EDINA
Peter Brantley and colleagues, California Digital Library
Tim Brody, University of Southampton
Peter Burnhill, Edinburgh University
Paula Callan, Queensland University of Technology, Brisbane
Debbie Campbell, ARROW, Canberra
Lorna Campbell, CETIS
Les Carr, University of Southampton
Priscilla Caplan, University of Florida
Eric Childress, OCLC
Mark Childs, University of Warwick
Sayeed Choudhury, Johns Hopkins University
Mike Clarke, Higher Education Academy
James Clay, Western Colleges Consortium
Tim Cole, University of Illinois
Sarah Currier, CD-LOR project
Andy Dawson, CDLR
Lorcan Dempsey, OCLC
Gordon Dunsire, Strathclyde University
Ed Fox, Virginia Tech
Morag Greig, Glasgow University
Andrew Grout, Edinburgh University
Kat Hagedorn, OAIster, University of Michigan
Cathrine Harboe-Ree, ARROW, Monash University
Stevan Harnad, University of Quebec, Montreal
Rachel Heery, UKOLN
Jessie Hey, Southampton University
Sarah Higgins, Edinburgh University
Amanda Hill, MIMAS
Tore Hoel, Norwegian Ministry of Education
Bill Hubbard, SHERPA, Nottingham University
Philip Hunter, IRIScotland, Edinburgh University
John Houghton, Victoria University, Melbourne
Arne Jakobsson, NORA, Oslo
Keith Jeffrey, CCLRC
Dean Jones, National Centre for Text Mining
Robert Kiley, Wellcome Trust
Gareth Knight, AHDS
Larry Lannom, CNRI
Norbert Lossau, Bielefeld University
John MacColl, Edinburgh University
Sally MacDonald, Petrie Museum, London
Roddy MacLeod, Heriott-Watt University
Ross MacIntyre, MIMAS
Mark McFarland, University of Texas
Ken Miller, UK Data Archive
Eric Lease Morgan, Notre Dame University
Martin Moyle, University College London
William Nixon, Glasgow University
Jerry Persons, Stanford University
Andy Powell, EduServ Foundation
James Pringle, Thomson Scientific, Philadelphia
Vanessa Proudman, Tilburg University
Peter Raftos, Australian National University
Christine Rees, EDINA
Dan Rehak, Learning Systems Architecture Laboratory
James Reid, Edinburgh University
Robin Rice, Edinburgh University
Griff Richards, Simon Fraser University
John Robertson, Strathclyde University
Peter Robinson, CLIC project, University of Oxford
Steve Rogers, JORUM
Rosemary Russell, UKOLN
Arthur Sale, University of Tasmania, Hobart
Bas Savenije, Utrecht University
Sandy Shaw, EDINA
Frances Shipsey, VERSIONS, London School of Economics
Pauline Simpson, NERC
MacKenzie Smith, MIT
Thornton Staples, University of Virginia
Tim Stickland, EDINA
Amber Thomas, JISC, WM-Share
Andrew Treloar, ARROW, Monash University
Graham Turnbull, SCRAN
Herbert Van de Sompel, Los Alamos National Laboratory
Leo Waaijers, SURF
Caroline Williams, Resource Discovery Network
Andrew Wilson, AHDS
Arnott Wilson, Edinburgh University
Melanie Wright, UKDA
Jeff Young, OCLC
Rowin Young, CETIS Assessment SIG

1.3 National networked repository systems in other countries

There are repository networks established in Norway, the Netherlands and Australia. Each has developed along different lines according to national requirements and attributes but they share, along with JISC, the vision of interoperability, thereby providing a broad-scope national database of Open Access content that can be added to, searched, mined, re-used, exploited for specific interest groups and built upon over time. A very brief description of each follows here.

Norway's network consists of only four research universities at the moment but may extend to the remaining research universities and be linked to the FE sector's own network over time. Each university has a repository exposing content to OAI harvesters, and NORA (Norwegian Open Research Archives)[2] provides a search interface for users. There is currently no mandate in the country, but Norway's universities already each *requir*e researchers to deposit the details of published articles in their own Current Research Information System (CRIS). The aim is now to link NORA and the CRISes so that researchers deposit only once and metadata is migrated between systems. It does not appear that any other services have yet been built upon the network. The system is government funded.

In the Netherlands, the SURF organisation set up DAREnet[3] to link the institutional repositories of all Dutch universities. SURF is also government funded. DARE harvests OAI-compliant content from the repositories and provides the search interface. It has added the Cream of Science service[4], which showcases the work of around 200 top Dutch scientists. Various Dutch universities have built subject-specific services on the system. The preservation of text-based, video, audio, moving images and like files is handled by the Royal Library of the Netherlands. The preservation of research datasets is the domain of the Royal Netherlands Academy of Arts & Sciences.

In Australia, the top research universities have repositories that are linked to form ARROW (Australian Research Repositories Online to the World)[5]. ARROW was developed and is run on government money. The ARROW Discovery Service, developed and operated by the National Library of Australia, provides the search interface.

[2] http://www.ub.uio.no/nora/
[3] http://www.darenet.nl/en/page/language.view/home
[4] www.creamofscience.org
[5] http://arrow.edu.au/

2. USER REQUIREMENTS

2.1 Users of national repository services

Any model for linking repositories needs to be designed with the ultimate purposes of users in mind, and should not fall prey to the temptation to design around attractive technical solutions or for elegance's sake. It is therefore important to take into account from the start the requirements of the potential users of the system. Whilst this element of the study was specified as a minor component of the overall work, it nonetheless determined the outcomes. A model that does not properly provide for the requirements of the potential users is doomed from the start. We have drawn here on previous studies that have examined the requirements of various types of user which are: end users (researchers across all disciplines, teachers, and learners); research administrators, employers and funders, intermediaries working on behalf of any of these groups, and institutions or other bodies with repositories.

The notion of *national* repository services – from the reader side – is a construct that does not sit perfectly with that of Open Access with its emphasis on interoperability across borders. Most end users (researchers, teachers, learners) know little and probably care even less about national aggregator services, preferring to enter the UK repository corpus via large-scale web-wide aggregators such as Google or OAIster. This may not be so much the case for other users, those with different roles, different needs and different motivations. For the purpose of this study, users have been categorised under the following headings:

Repository managers: those who manage institutional repositories, subject-based repositories, object-type repositories or special collection repositories (such as museum collections)

End users as searchers / readers: This category encompasses researchers and scholars across all disciplines, teachers, learners and the interested public

End users as content providers: Content providers are in most cases the same people as the searchers/readers but present new needs in their role as providers of repository content

Content aggregators: These are people who manipulate, select, harvest and modify content from repositories and offer it to their respective users in the appropriate form

Meta-users: Meta-users are people who are entering the corpus not with a simple resource discovery remit, but to carry out analytical activities. Examples would be the

Research Councils and other funders, research assessment investigators including employing institutions, and people (primarily, but not exclusively, in the Open Access community) studying research metrics

Entrepreneurs: this term can encompass people who wish to turn their own projects into services and those who see a way to produce value-added offerings for their own constituency. One example could be the national libraries, which may see opportunity in offering entry points to certain types or subsets of repository content. Others are scholarly publishers or specialist publishers producing bespoke services to certain industries or publics.

Each of these constituencies carries out a set of activities that in turn have certain needs and requirements. These are shown in annotated form in Table 1, after which the requirements of the user groups are discussed briefly and related to the types of repository service that might satisfy that requirement.

User	Requirement	Candidate services
Repository managers	Repository business case	Advocacy advisory services
	IPR advice	Legal advisory services providing guidelines and help on copyright, IPR and associated issues
	Repository creation	Repository construction and/or maintenance services
	Repository hosting	Repository hosting services
	Technical issues:	Technical advice/provision:
	Digital content	Digitisation services
	Metadata:	
	Structure	Metadata creation advisory services
	Controlled terms systems	Authorisation services
	File formats	File management / migration services
	Preservation	Specialist preservation services
	Data exposure (e.g. OAI)	Technical advisory services
	Name authority systems	Name authority services
	Access and authentication	Access and authentication services
	End user services and advocacy:	
	Deposition of content	End user needs analysis
	Use of content	Advocacy advisory services
End users as searchers	Cross-repository search	Resource discovery services
	Subject-specific search	Resource discovery services
	Object-type-specific search	Resource discovery services
	Tailoring to individual needs	Personalisation services
	Purposing	Purpose-specific delivery services
	Payment systems	Revenue-collection services
	Access and authentication	Access and authentication services
	Value-added content	Publishing and overlay journal services

End users as content providers	Peer review Somewhere to deposit Guidance on the best place to deposit Once-only deposition Advice on file formats and associated technical issues Advice on rights issues Usage data Impact data ‘Ownership’ of own content A vision of why	Peer review services Institutional repository / national repository Repository ‘mapping’ services (called bridging services on the diagram) Technical advisory services (e.g. preservation) Rights/IPR advisory services (e.g. SHERPA/RoMEO) Usage statistics services Citation analysis services Advocacy services
Content aggregators	Enhanced metadata	Metadata enhancement services Cataloguing services Text- and data-mining services
Meta-users (employers, funders, research managers, governments, economists, etc)	Usage statistics Research assessment and monitoring Meta-analysis	Usage and feedback services Citation analysis services Data-mining and text-mining
Entrepreneurs (e.g. re-sellers, technology transfer specialists)	Technology transfer mediators Publishers	Specialised resource discovery services Technology transfer services Mapping and bridging services

Table 1: User groups and their repository service requirements

2.2 Repository managers

2.2.1 Business case

The business case for a repository may be part of the remit of some organisations but in the case of universities, colleges and other research-based institutions the case needs to be made from scratch to senior management, usually by library staff (sometimes in concert with researcher 'champions') who instigate the concept within the organisation. Because a repository represents a clutch of intangible assets the case needs a special kind of argument to support it. The ongoing espida project[6] at Glasgow University is developing a model that can help to make such a business case and this is a good example of the type of service that could operate in this arena. In addition, the provision of advocacy resources and background information supports repository managers in formulating and presenting a convincing business case for their institution.

2.2.2 IPR and copyright

Advice on IPR and copyright issues is always sought after by repository managers. With respect to IPR, research-based institutions' technology-transfer offices are usually extremely interested in the idea of a repository making the institution's output open to all, and may seek to influence the development in ways that may not be altogether positive. With respect to copyright, it is still far from easy for repository managers to resolve the problems involved in many instances and authoritative, practicable advice and solutions are extremely valuable.

2.2.3 Repository building and maintenance

Creating the repository itself on-site using an institution's own resources may not be possible for one reason or another. Outsourcing this task to third-party suppliers is one answer. The outsourced tasks may include simply building the repository, or building *and* hosting it on behalf of the institution.

2.2.4 Technical issues

The technical issues that face repository managers must not be understated. They range from obtaining content in digital form at all, through creation of suitable metadata, dealing with multiple file formats from different research communities within the institution, preserving the content in a usable form over time (when file formats change and other technical standards move on) and exposing the metadata in appropriate forms to achieve the proper visibility for the repository's content. Some services already exist in these areas. For example, the American Physical Society offers an XML-conversion service for other publishers, notably small societies. The UK's AHDS[7] advises researchers funded by the Arts and Humanities Research Council on file creation and formats and provides a storage and preservation service to this community.

[6] www.gla.ac.uk/espida/index.shtml

[7] www.ahds.ac.uk

Name authority systems are sought after by repository managers wishing to ensure that all authors are correctly identified, their names spelled correctly, and the various forms in which those names may appear (for example, J. Smith may also be the same person as J.A. Smith or John Smith, depending on how he styles himself from publication to publication) are connected under one single author identity.

File formats are already a problem and this may grow. The National Archives provides the PRONOM file registry service – a service that is in the vanguard worldwide in this arena – which is currently being linked with EPrints. Repository managers need considerably more support in managing file formats in the long term. Shared expertise on accessibility is required.

End users are wary that deposit may take up a lot of time, though we know in reality it does not (Swan & Brown, 2005; Carr & Harnad, 2005). They need advice and help on this. This task falls largely upon repository managers, who can assist end users through advocacy activities (see below). In other ways, getting content into the repository can be maximised: ingest procedures can be streamlined, permitting batch ingest via data feeds and bulk import and export of data, though not all individuals in this management role feel they understand what it possible.

Subject classification is currently rudimentary for the average institutional repository in the UK, if existent at all, yet it is an important prerequisite for harvesting even if it is not used significantly for searching (as behavioural studies have shown). Australia is well ahead in this and a national service in the UK that resolves the problems of classification and taxonomies would be useful.

Finally, various projects found that repositories need to be able to provide better services to individual authors and to research groups or departments. For example, being able to extract information in certain styles or for specific purposes is important and searching by name needs a better system than is currently available.

2.2.5 Access and authentication

In addition to these purely technical issues, there are accompanying problems of authenticating would-be users if some of the content is not to be truly open access, of how to implement access of authenticated users and, in the cases where sensitive or restricted-use data are involved, putting in place the proper controls on how the data are used and what is done with them after use. These sorts of restrictions must apply, for instance, to substantial amounts of data in the social sciences.

2.2.6 Usage

Repository managers and – increasingly – authors want to know how much usage is being made of their content and where this usage is occurring. This sort of feedback is important in encouraging authors to deposit and is used by repository managers to secure commitment and buy-in from senior management in their institutions.

2.2.7 Ownership

Although repository managers are eager to have their content viewed and used, they are also eager to establish some sort of ownership on it. They are in favour of national services that bring searchers to the content but not if the identity of the provenance of the content is lost in the process. People we spoke to said this, and the CLIC (Community Led Image Collections) study on image collections reported the same thing[8].

2.2.8 Advocacy

Finally, having a repository up and running is only part of the story: getting content into it is the rest. Advocacy to the researcher/content creator communities is essential and repository managers seek help to do this in the most effective ways for the communities their institution encompasses.

2.3 End users as readers/searchers

2.3.1 Entry points

Different user groups within the end user community have different entry points to the primary data, some via broad-scope discovery services and some via subject- or object type-specific portals or discovery services. It has been shown by various studies (Swan & Brown, 2005; Sparks, 2005; Day, 2003) that there are differences between scholarly disciplines in the way information is accessed, used and deposited in repositories, and in the nature of the information itself. One point of note is that Google and its like will be the preferred route in for most primary users. This has been demonstrated to be the case in a JISC-funded study on time-based media collections (Asensio, 2003). Also, an examination of the log files for the Southampton ECS repository recently showed that only 11% of searchers entered via the 'front door' of the repository itself, the rest coming in via Google and other aggregator services (Carr, 2006). Indeed, librarians frequently aver that with respect to the majority of their end users (students and some researchers) 'if Google can't find it, it isn't there'.

For some users, though, subject-specific or object type-specific (examples might be digital images or theses) specialised discovery services will be the preferred route. The importance to end users of entering the repository-held literature by means of subject-specific services has been highlighted by a number of authors (Peters, 2002; Stephen & Harrison, 2002; MacLeod, 2005; Heery & Anderson, 2005). JISC has funded a number of projects in this area and these inform this part of the model developed in this study.

2.3.2 Search

Some projects have already reported in detail on user requirements for specific communities. For example, the Geo-Data Browser project recommended that, with users having different skills and expertise levels when it comes to searching for

[8] http://clic.oucs.ox.ac.uk/

information, a portal catering to their needs should support flexibility, providing for searching by item type, keyword, field and stemming for relatively unsophisticated users, and for the use of proximity operators, Boolean operators, nesting and so forth for experienced users. This same study also recommends the inclusion of a controlled vocabulary/thesaurus for users' convenience (Medyckyj-Scott et al, 2001).

Users who will only search using simple strategies require a simple, uncluttered interface and maximised ease of retrieval (Pearce & Martin, 2003). There is some evidence that retrieved metadata can be confusing and indeed overwhelming if too much detail is presented: one of our advisors reported that users had specified that the ideal is for a search service to 'capture more and show less', that is, return an accurate and appropriate set of results by searching rich metadata but *show* the user only a simple metadata set for each item.

Alerts are important to users but they do not wish to be overloaded by them: they prefer them to reside in the portal they use rather than via email, and they prefer to have control over their frequency and volume. RSS may be useful here.

Users do not just search: they also browse, and in some disciplines they do this much more than in others. Arts and humanities scholars and social sciences scholars browse more than natural scientists and discovery services need to take this into account when designing user interfaces.

Users need to extract information from repositories in various forms and styles to suit particular purposes: they may wish to do a simple sort by date, or to include various permutations using this and document type, item type and so forth depending on the intended use, be that a CV, a job application, a funder requirement, a project output, a website entry and so on, and to be able to add or delete. These issues are relatively simple software developments yet they make the difference between adequate and extremely useful for the end user.

2.3.3 Content

Users say they want everything that is relevant to their need delivered by one simple search, with no extraneous or irrelevant material! Since it will take some time for technology to attain this level of performance, pragmatism must play a role. Nonetheless, bigger is better as far as the size and scope of the corpus searched: *'Meta-searching … ranked high among the identified gaps in current services, indicating shared interest in a "one stop-shopping" approach to providing digital resources and services and also highlighting the need for common and consistent high-quality metadata formats to support cross-resource seaching.*' (Halbert et al, 2005)

Cross-searching of multiple resources as a concept appears new to most end users (Pearce & Martin, 2005) (though, contrarily, they almost all use Google [Swan & Brown, 2005]). Users cannot always know what types of material may be relevant to their needs, so there will be a need for resource discovery services for end users that are capable of searching across repositories to return *all item types* in response to a search on a topic. So a search on, say, ‘Van Gogh’ might return not only journal articles but theses, books, learning objects and images. The metadata should clarify very simply what each item is in terms of type. In the arts, humanities and to some extent the social sciences, books and monographs form a body of research literature of greater importance than journal articles: these may form the majority of items returned from a search, and should do so, even if some of these items may not be fully accessible, that is, they are not themselves Open Access, or are available only for a fee. Lurking behind this statement is the issue of non-Open Access material in mainly-Open Access repositories.

2.3.4 Purposing and delivery modes

End users have a variety of purposes for information that they gather for their work from simple current awareness through data sharing, data and text mining, to repurposing for teaching. Information can be delivered through and populate Virtual Research Environments, Virtual Learning Environments and Course Management Systems as well as a user’s own computer hard disk and over time it is expected that repositories and search services will become constituents of such systems, with all components interoperably linked (Awre et al, 2005, Fraser, 2005).

2.3.5 Personalisation

There is debate as to the merits of personalisation services. Evidence from studies of eCommerce services suggests that personalisation is largely unsuccessful in retaining customers or encouraging higher spend (Jupitermedia, 2003). Privacy concerns seem to feature strongly (Nielsen, 2003) and these will apply to academic situations as well as commercial ones. Lynch has argued that in order to allay fears on this score personal information about users should be retained locally, at institutional level, rather than at a centralised system (Lynch, 2001). As well as privacy concerns, the PORTAL project identified concerns over personalisation producing “dangerously narrow views on the information landscape’, presenting to the user as relevant only things they have themselves previously classified as relevant but by this means ‘removing the possibility of serendipitous leaps off into related resources” (Dolphin et al, 2002).

Conversely, Smith, Schmoller and Ferguson (2004) have presented evidence from a number of areas that shows that personalisation can increase the use of a resource ‘if done right’. ATHENS, the authentication, authorisation and accounting system in the UK (administered by Eduserve) works well in general and is widely liked. In specific applications it has proved extremely useful. The NHS uses ATHENS usernames as transportable identifiers for its highly-mobile staff so that ‘favourite journals’ and search histories move with the individual. The Open University has piloted a personalisation system for the OU library portal: it was found that users were reticent about creating their own profiles but if the library did the customisation on their behalf users then modify their profile and, moreover, use the resources to which the OU has subscribed more heavily than before.

This study by Smith *et al* (2004) also showed, however, that end user enthusiasm for personalisation in conceptual terms was not high, especially when a third party is involved in creating the profiles. The privacy issue applies in academia too but, they argue, “academic services will need to take a somewhat different attitude towards their users than commercial organisations. It may be necessary to collect and reuse information about users but there should be clear and open acknowledgment of that so that users understand what is going on. An endless small print agreement to which the user is required to respond ‘OK’ on screen before proceeding is not good enough.”

The same authors also conclude that although the JISC Information Environment is a desirable long-term goal, and would offer the ability for personalisation services to be developed, implementation of these will not be simple until ‘real world interoperability’ issues are resolved. They point out that an LDAP directory service – meant to be a simple shared service – has been much more difficult to implement than envisaged.

The pragmatic view is that personalisation has its place and can enhance the user experience, but at the same time it is expensive to implement and administer and in general does not *need* to be in place, at least at a sophisticated level, if only Open Access material is involved. There are circumstances, however, where it is desirable – where access to material *must* be managed for permissions reasons or for handling information flows efficiently.

Personalisation goes hand in hand with access and authentication services (see below) under which it is subsumed in Chart 1.

2.3.6 Payment

Naturally, payment doesn’t figure highly in a set of end user requirements! And our focus here is on Open Access content. Nonetheless, there will be instances where arrangements will need to be put into place to deal with payment for access to material. Such instances would be when metadata are Open Access but the full object is not, as is the case with material where royalty fees might be applicable, such as digital images. From the user’s viewpoint, where payment procedures are necessary they must be simple, streamlined and trusted.

2.3.7 Access and authentication

Where payment is required, or where repositories must authenticate users before allowing access to their content, services must provide the means to implement these systems. Revenue collection, accounting, and reporting functions must be provided by services either specific to individual repositories or acting as cross-repository services. Similarly, where authentication is necessary services that identify and permit the relevant level of access to each user are needed. Examples, as have been mentioned earlier, are where repositories contain sensitive data that are accessible only under strict conditions of user authenticity, and with strict requirements on the user about the way the data are utilised and then handled when the task is complete. The UKDA already operates a sophisticated service in this regard and may be seen as a model for future instances that may arise.

2.3.8 Value-added content

As end users find it increasingly difficult to manage information in work-related settings they value processed information more highly. Publishers who can add this sort of value to content residing in repositories will enjoy new success. There is plenty of scope for publishers where Open Access research repositories are concerned, since content is available for aggregating in new ways, re-publishing and as an intelligence source: coupled with the content of the growing number of data repositories the new corpus represents a wealth of resources for innovative publishers to mine, add value and present new, useful, important information products to various user communities – both their existing customer base and new market segments. In our model presented in this report we place these envisaged new services, along with the peer review services that many publishers will wish to continue to offer (discussed in the next section), under the heading 'publishing services'.

2.4 End Users as information providers

2.4.1 Somewhere to deposit

As providers of content for Open Access repositories, end users need encouragement on several scores. First, they need to have somewhere to put their content. This may sound trite, but at the time of writing, the UK has 69 OA repositories, of which 54 are institutional or departmental, the rest being subject-based, theses repositories, e-journals or other non-institutional entities[9]. Since the UK has over 200 universities and associated colleges – not counting FE institutions – this means that a large proportion of researchers and teachers who have content they may wish to share do not have a home for that material. The situation *is* changing as educational institutions become aware of the virtues of an institutional repository, but these will only become ubiquitous over the next quinquennium or so. Provision of repository space for the 'homeless' content thus needs to be planned.

At the time of writing a scoping study is being carried out by the University of Edinburgh and the SHERPA Project at Nottingham University on the prospects for a national-level repository that would accept articles from researchers whose institution does not yet have a repository of its own. The intention is that once an institution builds its own repository, content related to that institution's researchers will be migrated from the national repository into the local one. If this plan becomes a reality a valuable service will be provided. The repository being envisaged is primarily for research eprints and associated objects: there is scope, too, for other overarching national-level repositories housing other types of Open Access content.

2.4.2 Worry-free depositing

We know that researchers are dissuaded from depositing their work by a variety of things, amongst them concerns about copyright and other associated issues (Swan & Brown, 2005). They need guidance that they can trust and a clear set of parameters

[9] http://archives.eprints.org/

within which they can operate. Individual institutions have succeeded in dealing with this and have put into place measures that reassure and advise authors on rights and associated issues, but this provision is patchy.

2.4.3 Once-only putting

End users want to deposit only once and they want that process to be as simple as possible. Services that cater to the need for once-only deposit will be important. Metadata may eventually be duplicated across multiple resources and services, but one of the secrets to getting a high level of deposit compliance is to require authors to do it just once. Guidelines for creating metadata are welcomed by authors, though there are reports that where these change too frequently the end user is confused and discouraged. Simple metadata creation can be left to the author (and this is all that the author should be required to create): if rich metadata are required then skilled mediators or machine-created data are necessary. The RDN has successfully implemented a programme of cataloguing using a team of trained professionals to do this, for example.

2.4.4 Technical issues

File formats are something of a bugbear for many depositors. In some institutions, so much difficulty has been met when requiring even just pdf format for research articles that repository managers have had to resort either to tuition sessions (Queensland University of Technology) or to allowing depositors to submit in any format and a mediator reformats the document (some US universities amongst others).

The deposit process should be embedded in workflow and to make this as simple as possible it should take place at the authors' own institutions, into their own institutions' repositories, in order to maximise efficient and happy compliance by authors. Deposit must also make sense with respect to each subject discipline; in other words, if specific requirements are made regarding metadata, these must fit the concepts of a discipline as tightly as possible.

Services that advise on how to maximise the simplicity of the deposit process, educate on metadata creation, help with format issues and, where appropriate, provide a preservation service are important.

2.4.5 Rights

A considerable proportion of authors do not know the copyright position for their work. Studies for JISC by Rightscom and by Key Perspectives Ltd found that up to a quarter of authors were in this category (Sparks, 2005; Swan & Brown, 2005). Publisher copyright agreements are frequently confusing (or even opaque) and can vary from journal to journal even within the same publishing house. The upshot is that anxieties about infringing copyright agreements weigh heavily on authors and present one of the greatest barriers to spontaneous self-archiving. Services providing clear, unequivocal information on copyright with respect to repositories have an important role to play here.

2.4.6 Ownership

The sense of ownership of their material by authors can be very pronounced. Authors regard their output, naturally, as their intellectual capital and are anxious that their work should remain associated with them if it enters the Open Access corpus (Asensio, 2003). This extends to other levels. Repositories themselves may also wish to 'brand' their content; in some cases this is for marketing reasons alone (and these are enough), but in others there may be regulatory reasons for having this in place. Some of the social science data held by the UKDA, for example, is deposited under terms that require information on the identity of the user and for what purposes they are using the data. Services that work across repositories will need to take this requirement into account.

2.4.7 Usage

Authors who have been provided with usage statistics from their repository have expressed the view that these are helpful, informative, and encourage them to deposit further articles because of the increased visibility they perceive their work is getting from the fact that it is available on Open Access terms. Some repositories routinely provide usage statistics for their authors (and users): one excellent example of what can be done on a very simple basis is the University of Tasmania's repository[10]. The JISC is funding one study that will produce a pilot statistics service operating across interoperable repositories worldwide[11] and other services that can provide usage data for Open Access repositories will be needed.

2.4.8 Impact

Just as usage data inform authors and encourage submission, so do data on the impact of their work. Citation analysis services can provide this very effectively. One JISC-funded example is already in operation – Citebase, which currently runs on a few large, subject-based repositories but which has huge potential once a sizeable Open Access corpus is available in the UK and beyond[12].

2.4.9 Peer review

Peer review for the scholarly literature has traditionally been carried out by scholars within a process managed by publishers. This is likely to continue for the foreseeable future, largely in this well-established form. New forms of peer review are, however, being discussed and are in some cases in effect, including experiments with post-publication commentary, pre-publication open peer review and variations on those themes. Some learned society publishers are involved in this, and some are already using repositories as a submission tool, encouraging authors to alert the publisher when an article has been deposited as a preprint so that the publisher can take on and manage the peer review process and formally publish the article subsequently. This constitutes an early-indicator of how publishers may themselves provide services across repositories in the future.

[10] http://eprints.comp.utas.edu.au

[11] http://www.jisc.ac.uk/index.cfm?name=project_irs

[12] www.citebase.org

2.4.10 A vision of why

Finally, users need to understand the reasons and advantages for Open Access. The major problem at the moment remains the lack of content in most repositories. Deposition rates vary by discipline (Sparks, 2005; Swan & Brown, 2005) and by repository, and this is very much to do with author ignorance as to how to go about it or why they should be doing it. It is estimated that only 15% of research articles are deposited. Authors prove willing and able to deposit when required to or when they clearly see the benefits of doing so. Convincing them of the vision, especially where there is to be a coherent, national vision, will be a big step in securing the critical mass of content that is needed to make the vision a worthwhile reality. Services that can provide this kind of advocacy support to the research community are essential. At the same time, repository managers have expressed the view that authors would be further encouraged to deposit were they provided with local repository-level services that provide advice on copyright, on versioning, on what a repository can house and on the whole activity of publishing their work.

2.5 Requirements of other user groups

2.5.1 Content aggregators

The model described in this report is termed an 'aggregation model' and the role of aggregators is described in some detail in the Technical Model and Architecture section. Aggregator requirements are centred on metadata quality and so services that enhance metadata can be of crucial importance. In this context, services that reside just above the data layer, – data-mining and text-mining services, and cataloguing services that enhance metadata – are all important.

2.5.2 Meta-users

Meta-users fall into several groups.

Employers and funders are one such. These people have two motives with respect to research output: to maximise the visibility of the results of the work they have funded and thus to maximise its influence (pounds well spent); and to be able to track the output, analyse and manipulate it and inform their own monitoring and planning activities. Both of these require maximum output to be available in OA repositories, yet this has not so far been achieved, so one of the immediate requirements of this type of user is some mechanism(s) to increase to a maximum the amount of material in Open Access repositories. The Wellcome Trust has already tackled this issue with its recent policy on Open Access, which includes a mandate on its grantholders to place copies of their published articles in the subject-based (life sciences and medicine) repository PubMed Central (PMC)[13]. So seriously does this organisation mean business on this score that it is currently in the procurement phase for contractors to build its own

[13] http://www.pubmedcentral.nih.gov/

satellite of PMC, PMC Europe. It is understood that further PMC satellites will be built in other locations around the world, too.

Institutions that cannot fill their Open Access repositories require assistance to do that. One of the forms that assistance can come in is mandatory requirement from the national funders – the Research Councils – though it now seems unlikely these will act in concert on this issue. Institutions therefore need to act themselves, and advice and guidance would no doubt be welcomed as a service.

If the critical mass of OA content is achieved, employers and funders can then use that content to satisfy another of their requirements, which is to monitor and assess research outputs and progress. For employers, this is mostly a local issue though comparative studies will require access to the networked national Open Access corpus. For funders, except for any who may follow Wellcome along the specified subject-based repository route, there will be the need for discovery services that work across the whole national corpus and that can provide the functionality that permits real comparative analytical work to be carried out on it. The need will thus be for services that can provide such functionality.

Research managers in institutions, governments and agencies that monitor research also require services that enable them to manipulate the Open Access corpus to extract useful data. Usage statistics services and citation analysis services fit this bill.

2.5.3 Entrepreneurs

Finally, there is a disparate set of people who are hard to describe simply but the best term we can think of is ‘entrepreneurs’. These are people who are high on the innovation scale, translating basic research outputs into applications that result in products with huge added-value. For example, they may be companies specialising in technology transfer, or specialist publishers. These are end users with very specific needs and who require, amongst other things, rather specialised resource discovery services to satisfy them.

2.6 The services that will be needed

Various types of service may play a role working within a UK linked-repository network. Existing services are already well-placed to take a position in this scheme and opportunities will open up for new entrants also to develop operations. The roles that needs to be carried out and some of the players – existing or new – that might take responsibility for them are:

2.6.1 Ingest-level services

Digitisation services: Services providing simple digitisation or XML conversion for repositories or content creators who do not have the resources to do this.
Providers: Universities, publishers, other commercial players

Rights and IPR advisory services: JISC Legal currently provides advice and guidance on legal matters to do with digital rights and associated issues. This is probably an area where a centralised service will provide the best answer to user needs but there may be scope for additional, specialised services sharing the stage. *Providers: JISC Legal, universities (probably as part of public sector collaborative organisations), commercial players*

Technical advisory services: Institutions are going to require expert advice on the creation and management of their digital assets. Services do already exist to fill this need in certain areas. For example, the AHDS currently provides workshops for new AHRC grantholders, and ongoing advice and ***preservation*** services to the arts and humanities community; UKDA provides substantial support for the social sciences community; PRONOM[14] provides expert support on ***file formats***, and CERN has a file conversion service that could act as a model for such services in the UK[15]. As e-science and e-research grow, this sort of professional help will be critically important to institutions trying to run repositories. ***Name authority systems*** are also required. There is much work going on in this area at the moment. In the Netherlands a national author identifier service is due to deliver at the time of writing, in May 2006, an authenticating service for Netherlands author names, built by OCLC/PICA. *Providers: universities, either alone or as part of collaborative organisations, AHDS, RDN, UKDA, PRONOM, DCC, commercial players*

Open Access advisory services: Institutions are still unclear or uninformed in many cases about Open Access, the concept, the advantages, the ways to provide it, and how it works. In addition, repository managers need to present a business case to their institution for establishing a repository. RCUK had an opportunity to provide a centralised, authoritative advisory service on OA had it managed to construct a cohesive OA policy but this now appears unlikely to emerge. The ideal is for a trusted national body to provide such a service but in the absence of this, individual organisations will do what they can. Some of the library organisations, individual universities and commercial players offering repository build and host services will continue to offer piecemeal solutions that provide more or less satisfactory results. *Providers: JISC, universities, other public sector organisations, commercial players as part of repository-related businesses*

Repository construction services: Many institutions will build their own repositories as has been the case already, but others are now turning to third parties to do this for them because they either do not have the resources or do not wish to employ them in this way. For the foreseeable future, until repositories have become ubiquitous and demand is fully satisfied, there will be demand for this type of service. *Providers: Eprints, BioMed Central, ProQuest, universities, other commercial players*

[14] http://www.nationalarchives.gov.uk/pronom/

[15] http://cdsconv.cern.ch/

Hosting services: Some institutions will not wish to host their own repository or repositories and would prefer to outsource this activity. This is already happening and both public sector and commercial organisations have begun to provide the service.
Providers: universities (on behalf of other institutions), national libraries, commercial players

2.6.2 Pre-aggregator-level services

Metadata enhancement services: Important now, this kind of service will increase in importance as the years go on. Automated metadata enhancement processes will be developed, mining the full-text or full datasets of deposited items, and these will hugely enrich the metadata and make searching semantically a possibility. There are various entities already working on such initiatives, but it can be expected that many new ones will emerge as the technology is developed.
Providers: The RDN, the National Centre for Text Mining (NaCTeM), universities, commercial players

2.6.3 Output-level services

This is the level that offers the greatest scope and it is at this level that most new services will arise. It is not opting-out of our responsibility here to say that needs – and therefore opportunities to satisfy those needs – will emerge that cannot be included in this overview … because we simply cannot star-gaze that well! Nevertheless, there are a number of things we *can* see happening, or for which we can see the potential, and this is the list of those:

Resource discovery services (including subject portals*):* Resource discovery services may return results that include material from across the whole corpus of UK Open Access material or they may focus on providing subject-specific or object-type specific selections. There are already a good number of players in this field and there is room for many more, serving specialised communities. The RDN is collaborating with SHERPA to develop a 'national' search engine (that will actually search across repositories worldwide): the *raisons d'etre* for this is that current OAI search engines are not particularly attractive to use (and are not well-used) and that a national 'view', or at least a national 'feel' to the front end of a search engine, would encourage UK researchers to use the service
Providers: The RDN, the OAI search engines, universities and commercial players such as Thomson Scientific, CSA, learned societies, other publishers, Scirus, Google, Yahoo! and the like

Name authority /authentication services: where required for administrative reasons, services that can provide identification and authentication procedures will be necessary.
Providers: Athens, Shibboleth, CSA, commercial players

Preservation services: Institutions cannot always be expected to manage the preservation challenges of some of the content that is produced, nor may they wish to accept responsibility for long-term storage anyway. Third party specialists can provide trusted solutions to these issues. There are examples of services already doing this. For example, the AHDS provides workshops for researchers who have been awarded grants by the Arts & Humanities Research Council, advises them on file formats and how to structure and deposit their data and provides a long-term storage and preservation service to the arts and humanities research community. Trusted repositories are also a suitable solution and, amongst others, obvious candidates for this role are the national libraries, especially since they have much experience in spinning off commercial services using material in their collections or their particular in-house information management and information science skills.

Providers: The AHDS, UKDA, the national libraries, commercial players, the Digital Curation Centre (DCC)

Publishing services: Peer review will always be needed for scholarly research output, whether or not it continues in quite the form it has assumed until now. Services will be needed to manage this process and it is likely that many existing publishers will continue to be major players. Already, some publishers are using repositories as submission vehicles for authors: this will no doubt spread as more publishers see the advantages in managing the peer review process. New publishers will enter the scene, too. Some universities are already gearing up for this role and it is an obvious role for the learned societies, even those who may currently only publish on a small scale. In addition to peer review, publishers add value by various means, and much of this value continues to be appropriate to greater or lesser extents for an OA corpus (formatting, front-end content, bundling, etc)
Providers: existing and new publishers, universities, learned societies

Overlay journals (a subset of publishing): Some publishers have already developed overlay journals using repository content. Examples are XXXX. Institutions may also develop overlay journals on their repositories. Lund University, for example, hosts the Lund Virtual Medical Journal[16] that simultaneously provides a convenient-to-use collection of medicine-related articles by Lund authors, showcases Lund's work in this area, and has demonstrably encouraged authors to self-archive their articles (Hultman Ozek, 2005). There is much interest in this concept and it is likely to grow.
Providers: existing and new publishers, universities, learned societies

Bridging services: Services that provide information about repositories and their content to other services that wish to develop their businesses using the UK networked content will be increasingly important. Already some exist (ROAR, OpenDOAR, IESR). There is scope for more providing overview services, pointing services, current awareness (about repositories) and mapping services related to repository content.

[16] www.lvmj.med.fak.lu.se/

Providers: ROAR, OpenDOAR, Information Environment Service Registry (IESR), commercial players

Citation analysis/research assessment services (meta-analysis services): These activities have enormous scope for growth. We are only at the beginning in terms of what can be developed in this area. As demand increases from institutions and from research funders (public or private) for ways of monitoring the outcomes from their investments more and more sophisticated ways of providing the answers will emerge. Bibliometrics is already substantial field of research even in the toll-access age, but it will burgeon in the coming years as the Open Access corpus grows. Citation analysis and other analytical methodologies are still in their infancy and whilst some of the new measures that will be developed can already be envisaged, the techniques of text-mining and data-mining will enrich these hugely. This is definitely a case of 'watch this space'.
Providers: Citebase, Thomson Scientific, other commercial players, universities

Usage statistics and feedback services: One JISC-funded project is underway at the moment, led by Eprints, developing the software to produce usage statistics from interoperable repositories[17]. This project is operating on a global scale, working across interoperable repositories wherever they are. It will be possible to produce a subset of statistics for UK repositories. Repository managers, institutions and authors themselves are eager for this sort of feedback information which informs their operations and enables them to advocate and educate within their own community and services that can provide it will be needed and popular.
Providers: Universities, commercial players

Technology transfer advisory services: Though some projects have successfully made the transition to services, projects frequently struggle in this respect, even though concepts are promising and there is the opportunity to establish viable and sustainable services. Setting up services on a business basis is a specialised skill that cannot be expected to reside within the staff complement of a project in most cases. A professional advisory service is needed at this point to support fledgling services and help them get onto their feet.
Providers: JISC, commercial players

2.7 Gap analysis

The following are the major areas, currently un-provisioned, where repository services may play a role, or areas where there is a need for existing service provision to be boosted:

- ***A place for the homeless.*** As already discussed in the User Requirements section there are repositories in only a minority of research-based institutions in the UK at

[17] http://irs.eprints.org/

present. For the authors in those institutions who wish to deposit their work in an Open Access repository, space needs to be provided. A JISC-funded scoping study is currently underway for such a ‘catch-all’ repository: it is anticipated that this will confirm the necessity for such a repository for the short-to-medium term. Ultimately, the aim is for all institutions to host their own local repository, at which point the data residing in the catch-all one will have been migrated to where it belongs – the institutional repositories of the authors. The British Library is also planning a repository for authors who are not affiliated to an academic institution at all

- ***Digitised content.*** In all fields there are huge amounts of content not yet digitised. Most text-based research resources are now created-digital, but image libraries, for example, have much catching up to do (Pringle, 2005).
- ***Metadata creation.*** For museum collections the situations is extremely patchy. Some collections have structured metadata for all items; some collections have metadata for all or most items but these are inconsistent or not of a standard that can be used for harvesting; and many collections have as yet no metadata or cataloguing at all. They are not all in officially recognised museums, either. University departments, individual researchers and private individuals hold a considerable proportion of the nation’s artefacts; establishing cataloguing and metadata creation procedures for these is a challengng prospect
- ***Non-Open Access research literature***. Some disciplines embrace practitioner communities that carry out and make available their work in different ways to the academic sector. Engineering is an example, where the practitioner base uses technical reports and trade publications for dissemination, and these may be paid-for products. To be truly encompassing, a subject-based service for engineering or other disciplines that operate in a similar way must provide entry to these paid-for bodies of literature as part of its service offering
- ***Bridging services sitting between the repositories and the service providers*** New entrants wishing to develop services on repository content will often need help finding out what data are available, what software repositories run on, and what content types are housed and where. OpenDOAR, ROAR and the IESR are already providing some of this information but as content grows there will be greater demand for services that lead into the corpus and point the way around it
- ***Technology transfer expertise to help convert projects, pilot studies and ideas into successful services.***

2.8 Linked repositories and services overall model

The previous sections have presented an overview of the users of repositories and what they need in the way of services and have briefly described the nature of the services themselves. The issue of where the constituents of a linked network of UK repositories and their services sit in relation to one another can now be addressed. The overall model is shown in the diagram that follows here. We have depicted the repositories and their services as a series of layers.

The data layer contains the repositories themselves which fall into the following categories:

- subject-specific repositories
- institutional repositories, a term which includes cross-institutional or consortium repositories
- national-level repositories that are a ‘refuge’ for content that has no institutional home
- open access journal article collections
- individuals’ repositories (i.e. individuals making available their own repositories/websites via OAI-PMH)
- repositories maintained by learned societies. Most learned societies have so far fought shy of repositories, though if the interests of societies and their members are permitted to dominate over the interests of the publishing arms, societies see that this is a natural development for the future and offers them enormous scope in furthering the interests of their members and fulfilling their missions. There are some pointers already. Two physics societies – the American Physical Society and the Institute of Physics in the UK – have effectively established repositories by building mirror sites for arXiv in their respective countries as a service to physics. The publishing arms of these two societies are already using arXiv in an innovative way. The International Union for Crystallography is also active in this area, helping to roll out repositories in several UK universities, harvesting content from them all in order to collate and validate the data, and building its own society repository for datasets for crystallographers who need somewhere to deposit. Learned society repositories would be expected to provide value-added services themselves in addition to simple exposure of repository content.

Repository hosting services provided by third parties also reside at this level.

Working across the data layer are the ***ingest level services*** that are required by the repositories. These are shown below the data layer in the diagram.

Between the data layer and the ***aggregator layer*** are the metadata enhancement services, into which feed data-mining, text-mining and cataloguing services. The aggregator layer is where the *technical linking* of the repositories takes place and this is described in full in the technical modelling section of this report.

The technology transfer services sit on their own in the ***post-aggregator layer***.

Above this, and containing the services that work across the aggregated data from the linked repositories is the ***output services layer****.* This is where the services that provide preservation, publishing, resource discovery and the other functions described in the section above are found.

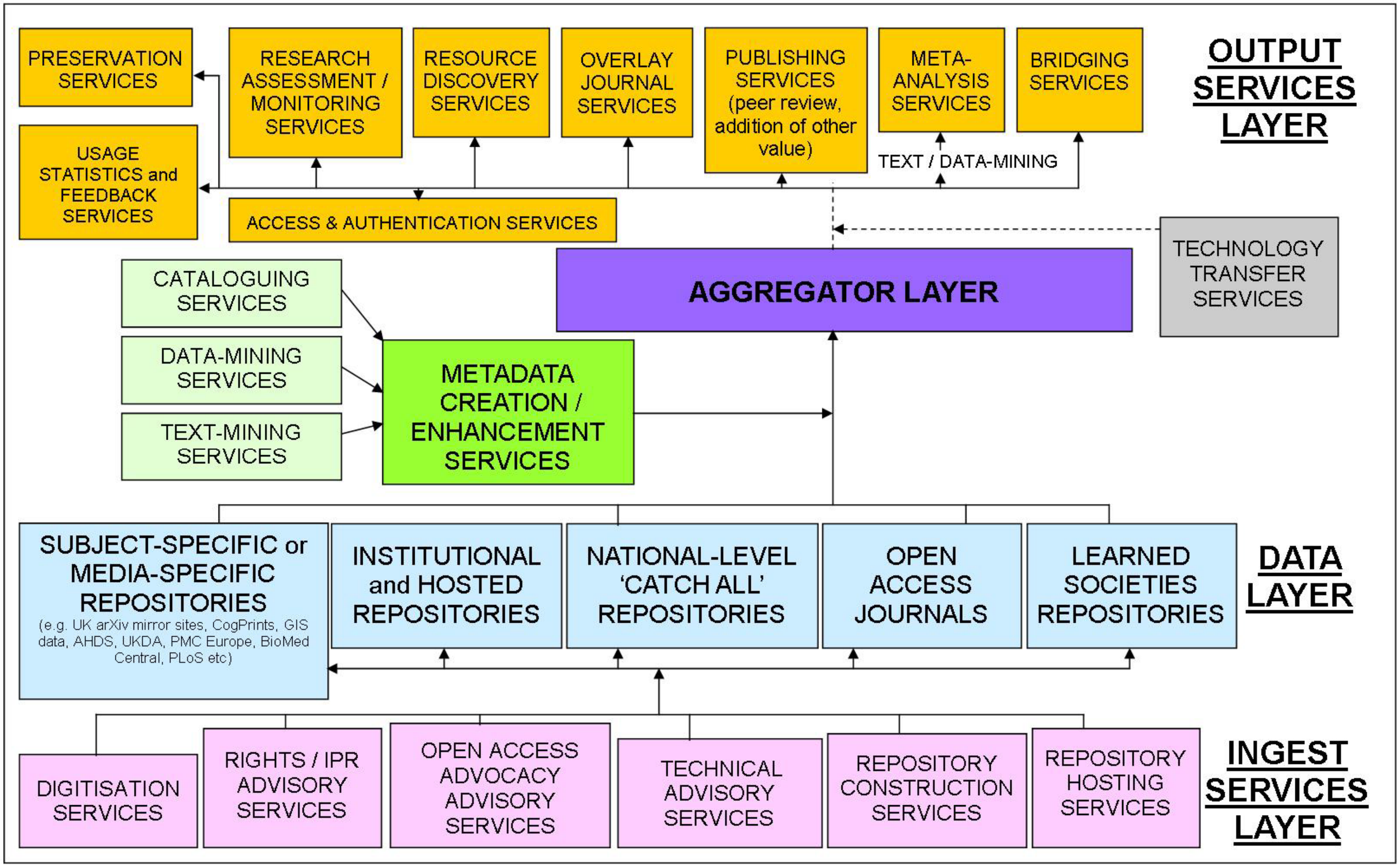

PRESERVATION SERVICES
USAGE STATISTICS and FEEDBACK SERVICES
RESEARCH ASSESSMENT / MONITORING SERVICES
RESOURCE DISCOVERY SERVICES
OVERLAY JOURNAL SERVICES
PUBLISHING SERVICES (peer review, addition of other value)
META-ANALYSIS SERVICES
BRIDGING SERVICES
OUTPUT SERVICES LAYER
TEXT / DATA-MINING
ACCESS & AUTHENTICATION SERVICES
TECHNOLOGY TRANSFER SERVICES
AGGREGATOR LAYER
CATALOGUING SERVICES
DATA-MINING SERVICES
TEXT-MINING SERVICES
METADATA CREATION / ENHANCEMENT SERVICES
SUBJECT-SPECIFIC or MEDIA-SPECIFIC REPOSITORIES
(e.g. UK arXiv mirror sites, CogPrints, GIS data, AHDS, UKDA, PMC Europe, BioMed Central, PLoS etc)
INSTITUTIONAL and HOSTED REPOSITORIES
NATIONAL-LEVEL 'CATCH ALL' REPOSITORIES
OPEN ACCESS JOURNALS
LEARNED SOCIETIES REPOSITORIES
DATA LAYER
DIGITISATION SERVICES
RIGHTS / IPR ADVISORY SERVICES
OPEN ACCESS ADVOCACY ADVISORY SERVICES
TECHNICAL ADVISORY SERVICES
REPOSITORY CONSTRUCTION SERVICES
REPOSITORY HOSTING SERVICES
INGEST SERVICES LAYER

Chart A: Overall model for repositories and the services built across them

To draw together what we have presented so far, the original table appears again overleaf. In this version (Table 2) the table is coloured in accordance with the colour scheme in the repositories/services diagram in Chart A so that the identified candidate services in the table can be clearly identified with respect to their position in the overall scheme.

User	Requirement	Candidate services
Repository managers	Repository business case	Advocacy advisory services
	IPR advice	Legal advisory services providing guidelines and help on copyright, IPR and associated issues
	Repository creation	Repository construction and/or maintenance services
	Repository hosting	Repository hosting services
	Technical issues:	Technical advice/provision:
	Digital content	Digitisation services
	Metadata:	
	Structure	Metadata creation advisory services
	Controlled terms systems	Authorisation services
	File formats	File management / migration services
	Preservation	Specialist preservation services
	Data exposure (e.g. OAI)	Technical advisory services
	Name authority systems	Name authority services
	Access and authentication	Access and authentication services
	End user services and advocacy:	
	Deposition of content	End user needs analysis
	Use of content	Advocacy advisory services
End users as searchers	Cross-repository search	Resource discovery services
	Subject-specific search	Resource discovery services
	Object-type- specific search	Resource discovery services
	Tailoring to individual needs	Personalisation services
	Purposing	Purpose-specific delivery services
	Payment systems	Revenue-collection services
	Access and authentication	Access and authentication services
	Value-added content	Publishing and overlay journal services

End users as content providers	Peer review Somewhere to deposit Guidance on the best place to deposit Once-only deposition Advice on file formats and associated technical issues Advice on rights issues Usage data Impact data ‘Ownership’ of own content A vision of why	Peer review services Institutional repository / national repository Repository ‘mapping’ services (called bridging services on the diagram) Technical advisory services (e.g. preservation) Rights/IPR advisory services (e.g. SHERPA/RoMEO) Usage statistics services Citation analysis services Advocacy services
Content aggregators	Enhanced metadata	Metadata enhancement services Cataloguing services Text- and data-mining services
Meta-users (employers, funders, research managers, governments, economists, etc)	Usage statistics Research assessment and monitoring Meta-analysis	Usage and feedback services Citation analysis services Data-mining and text-mining
Entrepreneurs (e.g. service developers, re-sellers, innovators, publishers)	Technology transfer mediators Publishers	Specialised resource discovery services Mapping and bridging services Technology transfer services

Table 2: Candidate repository services

3. ROLES AND RESPONSIBILITIES

3.1 Open Access repositories in the UK: the current context

In this part of the report the roles and responsibilities involved in establishing repository services are examined. The organisational viability and sustainability of repository services have not yet had the benefit of many clear models. Institutions with repositories may have established them for a variety of reasons. These can vary from one institution to another. In some cases the overriding reason has been to showcase the research activity and output of the institution. In others, there may primarily be the desire to preserve the digital output of the institution, while in yet others the teaching remit provides the case. In some instances, multiple repositories exist within an institution, each fulfilling separate roles, and these may not be linked in any way even within the hosting institution. Other than exposing content in a way that means it can be harvested by service providers, however, in general there has been little service development at the *institutional* level.

Sitting alongside this scattered collection of institutional repositories, which are almost all recently-established, are subject-specific or topic-specific archives, the national data centres, other large archives of material on specific themes and the national and depositing libraries. These are in general more mature, larger repositories and since many share a remit to collect and preserve within their sphere of operation they tend to have organised and standardised content and an emphasis on documentation and procedure.

Where repositories are Open Access there is naturally a desire to maximise usage and to demonstrate the potential of the repository in other ways. Services provided by repositories and third parties can help to increase usage and maximise the benefit of repositories. There are already examples of such services in operation. JISC's desire is to work in partnership with service providers, developing services on whichever bases seem most appropriate and promising.

The service requirements of users have been outlined in Section 1. In this section we look at the way existing or potential services might play a role in an emerging repository services infrastructure.

3.2 Routes into repositories

The research community in general enters the existing Open Access corpus primarily via Web search engines. OAI search engines are used routinely by only a small proportions of researchers – 3% in the case of OAIster, for example (Swan & Brown, 2005). This is not surprising for three reasons. First, the amount of content in OA repositories is still very limited and we presume that once these are better populated more researchers will use them. Second, most researchers are not aware such search services exist. Third, the functionality – of OAIster, for example – is very limited, though it may be expected that the development of such services will proceed as the amount of content makes it appropriate to do so. Indeed, there is some evidence that users are waking up to the existence of repository content and usage is beginning to climb. As far as the majority of current users of institutional repositories are concerned, however, the main route in is clearly via web search engine referrals, as we reported in Section 2.

It is unlikely that this will change significantly in the foreseeable future, especially as the Web search engines continue to develop means of returning ever more relevant results to the searcher. That is not to say that end users eschew completely the additional functionality they can get through other types of discovery tool. We know that 98% of researchers use the 'traditional' abstracting and indexing database services on a regular basis (Swan & Brown, 2005). These offer researchers the means to manipulate and analyse results, providing the tools to do this to a fairly sophisticated level if the user wishes, and that sort if functionality has value. It is probably wise to say, however, that for the time being end users will continue to arrive at repository content for the large part via Web search engines and that in some – perhaps many – cases they are finding repository content 'by accident' this way.

Where preferences have been tested (the Netherlands, for example), though, users have expressed a liking for subject-based entry points and have called for the provision of these. As a result of user demand for discipline-based services in the Netherlands, several have been established by universities. Examples are Connecting Africa (from the African Studies Department at Leiden University), Economists Online (in which Tilburg University is a partner) and Groningen University's developing environmental science service. A note of warning has been set by the Australian experience, however, where subject gateways, while popular with users, have struggled to find a sustainable model and the long-running subject gateway programme in that country has now expired.

In the UK the Resource Discovery Network (RDN) provides for this preference to an extent, as do the subject-based repositories such as arXiv and CogPrints.

Users may also wish to limit their search by object type (e.g. theses, moving images) if they have specific needs in this respect. If those needs are clear-cut then an entry point that offers only objects of particular types is appropriate. Searchers in pursuit of objects in museum collections, of images and of sound clips, for example, are clearly best served by discovery services that work across content of just the single type of interest. This saves the user having to define such requirements in their search protocol and simplifies the search process considerably. We see the provision of resource discovery services that cater to these needs, and to the needs of those who wish to use subject-based entry points to the data level, as desirable. Both have their place, and services that offer both or either have a role in a national networked repository system.

Neither, though, will satisfy the needs of individuals who want to find ‘everything there is on the topic’ as some users say they do; in those instances only a broad-brush discovery service will suffice. There is scope and potential for all three types of service because user needs are so varied, from the very specific to the largely-undefined.

Finally, it is important to remember that the behaviour of ***researchers as users*** can be disconnected and separated from their behaviour as ***depositors***. When referring specifically to their ***user*** behaviour we report that there is a distinct preference for searching and retrieving information from services that cater to their own subject area or scope of interest with respect to object types. As regards the researcher as ***depositor*** there are good practical and pragmatic reasons why deposit should be first and foremost into the institutional repository of a researcher. The arguments for this have been rehearsed before and concern the best ways to ensure that repositories fill with content in the most effective way (Swan et al, 2004).

3.3 Organisational issues in providing repository services

The two main organisational issues that we wish to highlight are discussed below.

3.3.1 Innovation

This is still early in the Open Access era and relevant and sustainable services develop from innovative work at the sharp end. This is a new environment and all the constituents are finding their places. Joining up the dots is the end game, and it is time to work towards it, but the conditions necessary for creative development to continue must also remain in place. Those conditions, at the innovation level, are usually described in the management literature as ‘talent, technology and tolerance’. They would be expected to prevail anyway in commercial enterprises developing repository services. In public sector institutions where innovative activities take place that impact on or may produce repository services this means that funders of projects must operate with a light touch: experimentation must be encouraged; lightweight services should be fostered that can dip their toes in the water to see how users respond. That is not to say that planning and careful user requirement analysis have no place; rather that they have an important role but not always at the level of innovation.

3.3.2 Projects to services

Certain developments in the repository services arena that began as projects have already spun off into services. JORUM is one example. The main factors that influence the performance of such services are their real utility value and the level of continuing innovation associated with them, coupled with the sustainability of the business model upon which they operate. JORUM has found a business model that can work, but it is one of the few in the repository services area that has achieved this so far. The most successful projects that have morphed into services in the past have been the large-scale digital library projects that developed into services such as MIMAS and BIDS (now ingenta) and the access/authentication (ATHENS) and collective purchasing (CHEST) services run by Eduserve, itself a project-turned-service and structured as a non-profit charitable organisation. The vision and dedication of individuals within services like these are undoubtedly crucial components of their continuing success, but these cannot be sustained without the reassurance of revenue or funding into the middle-term, a clear business case and accepted goals. Where projects have faced difficulties two factors – a lack of clarity of vision as to what the project is trying to achieve, and in what time-frame – have been major contributors.

The middle-to-long term view is often an uncomfortable one as far as funders of such projects are concerned but until the Open Access corpus reaches a greater size – a critical mass that offers real opportunities for output-level service providers to develop their operations – then the flow of cash into the system is inevitably going to spring from founts of public money for the most part. These twin necessities – content and funding – will continue to underlie repository developments, the one dependent upon the other, for some time. Once mass is achieved the oft-called 'tipping point' regarding existing business models and the cashflow patterns therein will be reached; cash will begin to flow differently, and sustainability will be a much easier goal to achieve for repository services. We discuss this further in the business modelling section of this report.

The other issue that has been identified as important in this regard is the visibility of the project and thus its ability to *'attract sufficient attention and commitment from potential customers to enable it to be shifted into permanent service mode*' (Brophy, 2006). Brophy cites middleware projects as an example here. In the repository services schema, metadata creation and enhancement services sit in the same sort of largely-invisible position, yet are crucial to the shape and functioning of the model.

Other organisational issues begin at the data layer – at the repositories themselves. We discuss metadata in considerable depth elsewhere in this report and the matter does not need to be visited here in detail, though it is certainly not insignificant: representatives from every repository service we have referred to during the course of this work raised the issue of metadata form and quality as one of the major problem areas of their work.

Other things, though, are also important at the data level. The provision of expertise, particularly technical expertise, at repository level is patchy. Some institutions are well-provided in this sense and can manage sophisticated repositories with ease. Others can operate perfectly satisfactorily at a certain level of requirement, but would struggle, say, to provide for the long-term future of objects created in a variety of formats. Yet others have the will but are severely constrained by resource limitations. As in other countries where a national approach is the focus, in the UK the best-provisioned institutions in terms of IT expertise are the large research universities. Smaller research-based universities and the 'new' universities may need more assistance. The FE sector, whilst in the vanguard in many respects with respect to teaching and learning materials and the sharing thereof, would struggle to fulfil, on its own resources, complex requirements from a national repository network with technically-ambitious plans. There will, therefore, be the need to provide assistance to institutions that cannot manage alone. Already there are services that offer assistance of some kinds: for example, the AHDS provides advice and guidance for those digitising their resources on issues such as file formats, metadata creation and modelling and structuring data. We envisage services that can provide this kind of advice and assistance as an important component of the national system.

These and other issues appear in the lists under ‘Lessons and insights’ below.

3.4 Lessons and insights from existing projects and services

The projects, studies, pilot services and fully-fledged services that have arisen in the area of repository services – many of them funded by the JISC – have a lot to reveal in the way of lessons and insights that are relevant to the establishment of national services on a nationwide network of repositories. We present the pertinent ones below, referencing the source if it is specifically documented. In many instances, though, these insights have been very general, or shared between several sources, and many have been transmitted to us in discussion and thus cannot be referenced as such. Some of the information listed here is detailed, for the sake of completeness and because the pieces of information may be useful to readers. We have categorised the items under headings that relate to the candidate services identified in Section 1:

3.4.1 Ingest-level services

3.4.1.1 Technical and coverage issues

Many projects have struggled because content at the original data source is not standardised and the coverage is very patchy. As a corpus, institutional repositories present an uncoordinated, non-standardised collection of data sources. Where metadata are exposed in an OAI-compliant way harvesting can take place, but the outcome is far from satisfactory in many cases – fields are missing, data have been entered incorrectly, there are typographical errors and so forth. As discussed before, metadata (formats, consistency, even existence) has been a major issue that has had to be addressed by several projects, and remains not altogether resolved, partly because repository managers are, like almost everyone else in this scenario, still only a short way along the learning curve in respect of what the ideals should be and how to attain them. It must be emphasised that although repositories in the UK have come a long way in a very short time – just a few years – there remains a long way to go. Existing and past projects have contributed hugely to the body of knowledge and understanding of what is possible but there is much left to do in this regard as good practice and technology continue on their leapfrogging path.

There is a wide variation in the levels of technical ability and provision of this across institutions and even across specialised services such as the HEA subject centres, some of which *“have sophisticated systems designed to support user needs but others have barely started…”* (Franklin, 2005). Static repositories have been demonstrated to be a low-barrier solution in exposing metadata to OAI search services in cases where the provision of technological expertise is low (Dunsire, 2005).

At the author/depositor level there are also problems to be resolved. File formats have proved to be a difficulty for authors in practice. Many authors are as yet unable to create a PDF document from a Word one, others use an array of ‘exotic’ (Waaijers) formats for objects such as video and audio files, and yet others, unsurprisingly, cannot tackle the deposition challenges of complex objects like relational databases or mixed media objects that comprise more than one element. This is a problem for both authors and institutions and is far from rare (Hey, 2004).

Added to this, as we have also discussed, the content level in repositories varies hugely from one to another. The overall coverage of the UK research literature is poor, and that is before considering the actual levels of learning-object deposition in the context of the potential levels that could be provided, of the patchy nature of metadata provided so far for the nation’s museum collections, and of the technical and cultural difficulties faced when cataloguing special collections such as still and moving images. All these things add up to a repository-content scene that is far from satisfactory, though much progress has been made in identifying these issues, studying their nature and working out ways in which best to deal with them.

There may be some advantage, then, in considering linking repositories to CRISs. In the UK, CRISs (Current Research Information Systems) are not common, though some institutions have implemented one – the CCLRC, for example[18]. In other countries CRISs are well-developed on a national basis and it is likely that this will become the case in the UK over time. Linking a CRIS to an institutional research repository can have benefits in several ways, most pertinent to this study being that authors need only deposit article metadata (and articles themselves) once – in either the CRIS or the repository – and the other can be populated by harvesting from the site of deposit.

3.4.1.2 Advocacy issues

In general, outside the disciplines of computer science and some areas of physics where self-archiving is the norm, authors are nervous of technological applications such as self-archiving and require substantial support.

In situations where authors have been required to deposit the same objects more than once resistance has been encountered. In the Netherlands, where authors were required by their institutions to deposit details of their published output in their institution’s CRIS and were subsequently asked *also* to deposit them in the DAREnet system this was found to be an unworkable approach. It was adjusted so that deposit happens only once, to the author’s own local institutional repository, and harvested for

[18] http://www.itd.clrc.ac.uk/Activity/CRIS

DAREnet and for the institution's CRIS. Such solutions will be critically important in getting author compliance.

The Netherlands' Cream of Science initiative highlighted other, positive, issues the most important of which was that the exercise resulted in improved relations between participating university libraries and faculties, and between libraries and researchers (Feijen & de Kuil, 2005).

There is evidence from widespread sources that much more advocacy is needed within user communities on every level of activity surrounding repositories from filling to use. While some studies have concluded that national user support services are not warranted in their field, evidence from other areas argues in favour of overarching, organised advocacy and user support.

Advocacy services have already been established from some projects, such as the OAISIS service providing advice and guidance to Scottish institutions wishing to set up repositories[19]

3.4.1.3 IPR and rights

Copyright and IPR issues have presented some projects with serious challenges. Authors are in general ignorant and, as a consequence, nervous and wary about legal aspects and requirements of depositing their work. This constitutes one of the biggest barriers to gaining a critical mass of content in Open Access repositories.

Rights issues have shown up as particularly pronounced with ***learning objects***, though this is not to minimise the issues surrounding other types of object too. Though dealing with the rights needs to remain at the data-hosting institution level, these institutions may need external advice when material produced locally contains other material whose copyright resides elsewhere. Such advice is often currently being given piecemeal, but some recommendations have already been made in a study by Charlesworth on the way to proceed on a national scale (Charlesworth, 2005) and a service specifically providing the sort of information and guidance he recommends would substantially aid institutions to manage their repositories more effectively. The Theses Alive! project (Andrew & MacColl, 2002) produced solutions to the rights problems that were a 'significant barrier' to progress in self-archiving of e-theses and published them via the JISC Legal service. That project included a recommendation that institutions change their thesis regulations to include provision for electronic submission, an instance which flags up the uncertain or obstructive nature of traditional arrangements that have not yet moved into the digital age and their impact on any potential national services. The JISC-funded L2L project revealed how difficult it can be to obtain copyright clearance, particularly from public organisations (Brosnan, 2005). One outcome of that project was a publication introducing IPR issues for people producing e-learning materials, which was published by JISC Legal (Casey, 2004). The ongoing JISC-funded TrustDR project is looking at the cultural, legal and technical

[19] http://hairst.cdlr.strath.ac.uk/oaisis/

aspects of setting up DRM systems in learning object repositories, so further recommendations and guidelines can be expected from that in the future.

Finally, a working group from HEFCE, Universities UK and SCOP has published a good practice guidance paper on IPR in e-learning programmes aimed at senior managers (HEFCE, 2003). In all, the e-learning community is actually better provisioned at the moment with advisory material on rights and intellectual property than other communities.

Digital image collections are also severely affected by rights issues, summed up in the report on The Digital Image project (Pringle, 2005): *"IPR in the digital image world is a confused and confusing picture, with far-reaching consequences for getting it wrong."*

The JISC-funded Digital Rights Management study recommended that good practice guidelines and common licences would improve the widespread adoption of DRM by repositories (Duncan et al, 2004).

3.4.2 Pre-aggregator-level services

Cataloguing and indexing, where appropriate, are complex activities that will increasingly be possible by machine though current solutions frequently remain human-mediated. The RDN has developed a solution to cataloguing for the JORUM service by using a team of trained cataloguers. This works well for this application though this is unlikely to scale well if required in other circumstances for a large body of content.

3.4.3 Output-level services

3.4.3.1 Discovery services

Google and other Google-like services (existing and future) need to be factored into a UK national schema. Existing services recognise this and are working to achieve it. The RDN is exploring this possibility, for example, and the Australian national service ADS (ARROW Discovery Service) is talking to Google about its coverage of the ARROW database. As a complication, some discovery services – OAIster is an example – will not accept duplicate entries (i.e. problems arise if a document or its metadata exists in more than one location). National services should broker arrangements between such services and the repositories to save the individual repositories doing so themselves. This brokerage function resides in the 'technical advisory services' location in the overall schema.

ePrints UK looked at various aspects of using the full-text in informing metadata creation, with some success, but had difficulty in some cases in accessing the full-text of documents to run searches across. This will be the case, too, where the original object has been deposited in a trusted repository (e.g. PubMed Central) that falls outside the UK network.

Subject descriptors were the focus of the HILT projects. HILT I showed that there was consensus in the community that a service that mapped between schemes was preferable to the adoption of a single scheme (Nicholson et al, 2001). The HILT II project has followed this up by developing a set of pilot terminologies as a service for the JISC Information Environment (Nicholson et al, 2005). These have the potential to improve interoperability, not only in the UK but globally, since the mapping approach enables functioning across multiple languages. Subject descriptors, ontologies, classifications, thesauri and related systems have increasing importance in semantic web applications.

The RDN and Higher Education Academy have developed a record interchange format that seems to work well (RLLOMAP). RLLOMAP is likely to be re-merged with its source, UK LOM Core, in time.

Interdisciplinary research in both arts/humanities and the natural sciences places new demands upon discovery services and if these are operating across distinct and separate subject areas interdisciplinary material remains invisible. This is an important issue because interdisciplinary and multidisciplinary research is on the increase and is likely to form a major part of research activity in the future.

3.4.3.2 Personalisation and authentication

Authorisation services will be required for access to and licensing of certain types of data, such as that which requires users to commit to agreements about usage and disposal of the data (e.g. some of UKDA's data holdings), data that can *only* be used by certain known parties, and data that the holder does not have copyright for (such as much of the geo-spatial datasets held in the UK where Ordnance Survey data are incorporated).

Personalisation in the form of email alerts has been found to work successfully (i.e. has gained user approval), even if users do generally express a preference for web page-based alerts. The ARROW Discovery Service has developed a daily email alerting system to individuals registered with the service, and sees a spike of usage each day at around 10am, just after the alerts go out[20].

CSA (aka Cambridge Scientific Abstracts), a commercial abstracting and indexing service, has expressed some interest in using its ***Community of Scholars*** database[21] as a name authority system for repositories (MacLeod, personal communication). This represents one possible solution to the problem of name authorisation, though only a partial one since CSA's database will not have complete coverage of the UK author base. This may be yet another area where learned societies have a role, providing name authority services from their own member databases or digital libraries.

[20] Debbie Campbell, ARROW; personal communication

[21] http://www.csa.com/e_products/COScholars.php

3.4.3.3 Publishers

Publishers are already beginning to work with repositories to provide services. The American Physical Society is offering XML-generating services (Kelly, personal communication). The European Physical Society and the Institute of Physics Publishing are encouraging authors to deposit their articles in arXiv and notify the publisher when this has been done so that the publisher can harvest them from the repository for peer review. Yet other publishers have begun using repository content to develop overlay journals, selecting out articles that fit a profile and bundling them for a specific readership.

3.4.3.4 Other issues: Research data

Several significant moves have been made on research data recently. These will undoubtedly represent only the vanguard in a thrust to enforce the making of research data accessible to the community for examination, manipulation, mining etc. In other words, data deposition will increase and repositories will need to plan for this:

- The OECD Committee on Science & Technology has recently developed a Declaration on Access to Research Data from Public Funding, which is now being taken forward by an expert group[22]. This will finalise a draft text in October 2006 which will be taken to the OECD Council towards the end of 2006
- The journal *Nature* and the International Committee of Medical Journal Editors have also issued guidelines to their authors about making supporting data freely accessible in public repositories when articles are published in the journals concerned
- In addition, NIH, NASA, the US Global Change Research programme, the Wellcome Trust, and some of the UK Research Councils have all announced requirements about making data accessible as conditions of grants

3.4.4 Activities undertaken by existing services

In the table below we have summarised where activities needed for a properly linked repository scheme in the UK already take place and may play a role in supporting the repository services outlined in this report.

Candidate services	Existing projects or services that undertake these activities, at least to some extent
Digitisation	HEDS (Univ. Herts Digitisation Services)
Rights/IPR advice	JISC Legal
Open Access advocacy advice	SHERPA; EPrints
Technical advice	SHERPA points to AHDS and other appropriate resources; EPrints
Repository construction	EPrints Services; commercial players
Hosting services	EPrints Services; commercial players
Institutional repositories	Institutions; EPrints Services; commercial players
National-level 'catch-all' repositories	PROSPERO in scoping phase

[22] http://www.oecd.org/document/0,2340,en_2649_34487_25998799_1_1_1_1,00.html

Subject-specific repositories	Institutions; communities
Media/object-specific repositories	Institutions; communities
Metadata-creation and enhancement	RDN; institutions
Technology transfer	
Access and authentication	ATHENS, Shibboleth
Usage statistics	Interoperable Repository Statistics project
Preservation	PRESERV project; AHDS; DCC; UKDA; others
Research monitoring	IRRA project
Resource discovery	RDN, RDN/SHERPA UK search service project; Thomson Scientific
Overlay journals	Institutions; communities; learned societies; commercial players
Publishing services	Institutions; communities; learned societies; commercial players
Meta-analysis	Citebase
Bridging services	ROAR; OpenDOAR

Table 3: Candidate services and some of the existing service providers

3.4.5 Priority services

There are a number of services areas identified as priorities that are not under the remit of existing services, are at project stage, or would benefit from increased support. Because they have been discussed already to an adequate extent, they can be presented as annotated points here. We have made four of them top priorities:

- An interim repository location for those whose institution cannot yet provide one
- Resource discovery/national search service
- Meta-analysis services: specifically citation analysis and bibliometric analysis services that can anticipate, and provide useful metrics for, future research assessment exercises
- Repository usage and statistics services: these can also inform future research assessment exercises but they can do much more, too, in informing and encouraging repository managers, authors and research administrators. The ePrints UK statistics service has been useful for giving the national picture at a certain level. The JISC has one project underway on this (IRS), due to report in a year's time, but nothing yet at service level

Other priority services/activities are:

- Preservation services: this is a complex area and it is not within the remit of this study to go into too much detail. Nonetheless, it is an issue that was raised many times during the course of our research and is clearly an issue for attention for many repository managers and other players. Preservation will become inceasingly important and also challenging, technically, and solutions to the problems should not be expected to be found at institutional repository level. This is something that requires high-level expertise. The JISC has a good deal of project work going on in this area and the national libraries are also very active in this regard. We flag it up here to re-emphasise its importance in the national repository landscape

- Name authority service: a national name authority service developed along the lines of that due to deliver in the Netherlands
- File format registry and conversion service: a service for all UK repositories developed using the examples of the PRONOM service from the National Archives and the CERN file conversion service (see section 2.6.1)

3.4.6 Challenges

An overarching project that seeks to establish repository services would face a number of challenges. There will be inevitable operational issues to deal with at supplier level but the main challenges are those of coordinating and managing such an ambitious programme. In other words, we see the major issues as ones of management process and responsibility. We categorise these as follows:

3.4.6.1 Repository-level issues

The provision of content will be something the JISC will need to consider carefully. Some institutions are successfully populating their repositories but many are failing to gather articles from authors.

- The JISC will need to take a strong leadership role in helping to put advocacy and mandatory policies high on the agenda of institutional management
- Some content will not be full-text because of copyright restrictions and the JISC will need to form a policy and plan to deal with this: the Netherlands' Cream of Science initiative learned lessons here that may be helpful
- Many institutions will need to increase staffing levels and/or skill sets if they are successfully to implement a repository: the JISC will need to have clear guidelines on what is expected from an institutional repository and what implications this will have for institutions

3.4.6.2 Organisational and procedural issues

- Communication will be a major issue at all levels of the scheme.
    - Good communication channels between repositories and services are critical, are currently limited, and will need to be more formally developed
    - Communication with authors and depositors needs to be in their own language and should clearly set out the goals, the rationale and the routes to implementation
    - Formal communication plans should be of any service development activities that the JISC initiates or commissions
- It is not uncommon to find a tension between local interests at repository level and the demands of national services and reducing or eliminating these the tensions can absorb energy, time and cash in substantial amounts

3.4.6.3 Coordination of suppliers

- The state of readiness of chosen service suppliers to begin, continue or resume any work that JISC would require or encourage them to do will vary. Analysing and coordinating supplier response capabilities and planning those into the whole operation will be something that needs to be carefully managed

- Similarly, coordinating contributing projects that are currently at different stages of completion need to be assessed and a proper management plan put in place to dovetail efforts
- If scaling-up of activities is required of existing projects and services a management plan must take account of this

3.4.6.4 Process management

- Specification stage: the JISC must ensure that any specifications to services are tightly drawn up to avoid cost over-runs or omission of elements of the outcome by suppliers. A robust tendering process will be critical if multiple suppliers are to be involved in one overarching scheme. Minimising the number of suppliers is good practice and enables streamlining of overall process management. Detailed contractual arrangements and close managerial control of this stage will be necessary. It would be appropriate to employ a contracts manager/company for the duration of the project.
- Managing developments: the scheme would face all the normal challenges of business development – getting people 'on board'; creating stakeholders who internalise the goals; identifying early adopters who can provide the vision and example; ensuring that roll-out proceeds in a managed and proactive manner, including developing proper incentives so that suppliers deliver in the integrated way that will be required. In our view, the overall operation would require a specific individual or company 'contracted to care' about delivering a successful outcome and the means of getting there; in other words, a marketing plan in the proper sense of the term must be drawn up and put into operation, with all the elements of such a plan – the outcomes, the channels through which they will be coordinated and delivered, the people who are to be involved and the places where the outcomes will reside – included and planned in detail. This management role should coordinate with that of the contracts management outlined above.

4. TECHNICAL MODEL

4.1 Introduction

The JISC-funded Focus on Access to Institutional Resources (FAIR) Programme[23] ran from 2002-5, investigating both the management of and access to institutional assets. It addressed many issues for the first time in UK, and offered up almost as many questions as it answered. These were valuable questions, however, and many are currently being addressed through the JISC Digital Repositories Programme[24]. Together, these two programmes are addressing the needs and requirements of the HE and FE community for the use of repositories. If there is an emphasis on experience so far it lies on the management half of the equation, with a developing body of knowledge of the factors involved in storing digital content and associated metadata. Notwithstanding experience so far, the access half of the equation, facilitating interaction with metadata and content, is less developed. It has been the purpose of this study to explore the issues behind the provision of services that facilitate interaction with repositories and what they store in order to redress this balance and fully share and capitalise on the digital content available.

This section focuses on the technical architecture and infrastructure required to underpin this sharing. It focuses on the factors that will underpin the potential services already proposed, many of which are based in the repository landscape and repositories themselves. The availability and structure of the content and metadata within a repository will affect what can be delivered through a service: how this is exposed for use by services is also important. Looking beyond the individual repository, the architecture within which multiple repositories can be brought together, and relationship between these different repositories, will affect how the services work across to them.

These factors were reflected in the invitation to tender for this scoping study[25]. This invitation also described three key issues that help frame discussion of relevant architecture and infrastructure.

- Firstly, it recognised that the environment in which the scoping activity was to take place is already a heterogeneous one, and that services looking across repositories will need to deal with this background. The situation is not particularly one of shutting the stable door after the horse has bolted: repositories have valid, often important local, reasons for the way they have

[23] JISC FAIR Programme, http://www.jisc.ac.uk/programme_fair.html
[24] JISC Digital Repositories Programme, http://www.jisc.ac.uk/programme_digital_repositories.html
[25] JISC Linking UK Repositories Scoping Study Invitation to Tender, http://www.jisc.ac.uk/index.cfm?name=funding_repositoryservices

been structured which are not necessarily open to easy change. This existing situation needs to be balanced, though, against the benefits of re-visiting these reasons with a focus on making the repository's materials available to wider services. Providing access to your own repository is one thing, providing access to your own alongside others is another.

- Secondly, whilst clearly focusing on user-oriented services the ITT recognised that this scope also needed to include those machine services that would help to underpin these services. The Open Archives Initiative Protocol for Metadata Harvesting (OAI-PMH)[26] itself is a machine interface upon which user-oriented service providers can be established. It is the flexibility that this and other open standards and protocols offer that suggest potential for a range of different services.
- Finally, the ITT emphasised an interest in providing services across open access repositories, as opposed to those repositories requiring various levels of authentication and authorisation. Open access removes a serious barrier to enabling interaction with shared metadata and content. It has to be noted, though, that there are different levels of open access, and restrictions are occasionally put in place alongside open access routes that need to be dealt with.

These issues have been explored through a combination of desk research and interviews with a broad spectrum of individuals involved in repository activity. The outcome from this is a proposed model to underpin the development of services that facilitate interaction across repositories, plus a series of recommendations for the community to consider and take forward. The evolving landscape requires pragmatic choices. Previous work has provided many of the building blocks required: how to put these together is the outstanding challenge.

4.1.1 The user as reader, author, and manager

In addressing the issues involved in developing services across repositories it is important to place the end-user first and consider how they need to interact with and make use of digital content. This has been covered in an earlier section of the report, but is worth re-visiting briefly here to reinforce the needs of end-users when considering technical development.

As highlighted in Table 1, end-users might take on a variety of roles when interacting with repositories. They might be considered readers or searchers, looking to see what others have produced and discovering information for learning, teaching or research. In different circumstances they might be authors or content providers, producing or creating the content the repository will hold for others to access. Some will also be repository managers, dealing with practical issues of interaction with repositories including IPR, advocacy, business cases, and policies behind technical implementation. These may overlap with the content provider role where amendment of

[26] Open Archives Initiative Protocol for Metadata Harvesting (OAI-PMH), http://www.openarchives.org/OAI/openarchivesprotocol.html

metadata and/or content requires management input. These varied roles, which an individual may take on, are complemented by user groups that also have an interest: aggregators, collecting what repositories can provide and potentially enhancing this to form the basis for end-user services; meta-users, who can use repositories to provide valuable management information to guide wider policy; and entrepreneurs, who can add value and enhance repository content to assist with dissemination. The implementation of all these roles will influence how the issues outlined in this technical model and the appendices are implemented.

4.1.2 Services

The term ‘services’ can have many meanings. It can refer to detail at the machine level, where ‘services’ are the description of the individual functions that software components offer. At the other end of the spectrum it can refer to complete packages of functionality that the end-user interacts with. ‘Services’ can apply to many of the points in between as well, some apparent to the end-user and some hidden behind the scenes.

The services being investigated in this study are at the level of complete packages, focused at end-users. This section of the report will refer to them as end-user services. Where services are referred to elsewhere they will be referred to according to the relevant context, for example presentation services for those components that focus on delivering presentation functionality. Where ‘services’ are mentioned generically the implication is that the full spectrum of levels and granularity is referred to overall.

4.1.3 Background

In 2004, the JISC-sponsored study on ‘Delivery, Management and Access Model for E-prints and Open Access Journals within Further and Higher Education’ (Swan et al, 2004) considered three possible models to underpin the exposure of e-print and OA journal content to end-user services.

1. Centralised – both metadata and content are submitted directly to a central agency
2. Distributed – all metadata and content remain in their source locations, and metadata is cross-searched ‘on the fly’
3. Harvesting – a hybrid model: metadata is harvested into a central searchable database but also remains distributed among the original data providers, while the content remains distributed

The centralised model was considered to offer the greatest level of control over the metadata and content, allowing it to be re-factored to facilitate a range of functions including preservation and a range of end-user services. Information latency was reliant on the mechanisms put in place to gather the content into the central agency. However, it was also assessed that this level of control may hinder as much as help the development of end-user services in the long-term because of the heavy costs involved in implementing and running such a model across a large number of distributed repositories.

The distributed model had the advantage of always providing up-to-date metadata as it focused on direct, immediate access to source locations of metadata, including repositories: this avoided replication of metadata records. The main difficulty with cross-searching repositories 'on the fly' was the delivery of the metadata 'as is' without ay opportunity to re-factor this for presentation: this approach is thus entirely dependent on what the repositories can provide. There were also notable concerns about the scalability of this model when accessing large numbers of repositories simultaneously.

The harvesting model offered a compromise between these. The model captures the advantages of centralisation, allowing re-factoring of the metadata to better support end-user services, without exerting the high level of control that a fully centralised model would require. It was accepted that harvesting did not necessarily provide the most up-to-date results when searching as repositories may have been updated since the last harvest. However, regular harvesting could alleviate this to a manageable degree.

The report recommended the harvesting model for wide adoption in the delivery and access of e-print and OA journal materials. It further recommended the use of the OAI-PMH as the underlying standard to support this harvesting model.

Although not considered at the time a fourth model has received a high degree of interest in some repository circles, although not within the open access arena. Peer-to-peer networking has been widely taken-up for the social exchange of music files and other materials. The LionShare project at Pennsylvania State University[27] is examining the use of P2P networking for educational use, and the SPIRE project[28] in the UK has picked up on this. P2P provides valuable control over what metadata and content is distributed throughout the nodes of the network, though the speed at which this takes place can vary. To what extent re-factoring is possible for presentation through end-user services is possible will depend on the policies of the network, though it raises the possibility of versioning between nodes.

P2P offers an intriguing possible solution for controlled open access. However, its immaturity at this time and relative complexity does not permit it to be recommended for wide use in this model.

4.2 An aggregation model

Two years further on, the arguments laid out by this study have not changed and there is every reason to accept that the harvesting model remains the best option. This applies even when taking into account the wider range of materials that might be made

[27] LionShare project, http://lionshare.its.psu.edu/main/
[28] SPIRE project, http://spire.conted.ox.ac.uk/cgi-bin/trac.cgi

available through open access. The evidence amalgamated from the work of this present study thus also recommends the use of the harvesting model to support end-user services across repositories.

Developments, though, since the time of the original report, have provided additional technical evidence for how this harvesting model can be put in place, and what it can offer. The ways in which different standards and technologies can be used have also moved on and, whereas the use of OAI-PMH is still very much warranted as the primary means of carrying out the harvesting, alternative solutions also suggest themselves as possibilities. The term harvesting has become almost synonymous with OAI-PMH because of its methodology. Consideration of alternatives to provide alongside OAI-PMH suggests the need for an alternative, broader term to encompass OAI-PMH and other approaches. The ability to gather metadata, or content itself, from a range of different repositories in order to provide a re-factored or 'shaped' view onto this requires aggregation, regardless of the technology used to enable this. The technical model described here is thus termed an ***aggregation model***.

4.2.1 Component parts of the aggregation model

The findings from desk research and interviews underpinning the development of this aggregation model are presented in four sections as appendices to this report as follows:

- Repository and end-user services overview
- Metadata and content
- Interfaces
- Technical architecture

This section of the report focuses on a description of the model, derived from the evidence and information presented in these four sections. The model places these factors in the context of putting different components in place, the relationship between the different components, and the technical factors affecting sustainability of the proposed model.

4.2.1.1 Metadata and content

Leo Waaijers, from SURF, in considering the value of the OAI, commented, "The data layer of the OAI model is indispensable for the services layer"[29]. This essential relationship can also apply outside of the implementation of OAI, although it was formulation and implementation of the OAI model that gave rise to a deeper appreciation of this link. This view suggests that the starting place for an aggregation model is with the information to be aggregated. This information — or content and its associated metadata — resides in repositories (accepting the variety of these that the use of this term can imply).

[29] Leo Waaijers, Personal email communication, February 2006

Metadata
Content

Figure 1: A repository containing content and metadata

Metadata generation

Content may have associated with it either a single metadata record or multiple metadata records in different formats, depending on requirements or potential use. These multiple records can be created in isolation or generated from a base metadata record that can be used as required to generate different metadata profiles through metadata crosswalks: the richer this base metadata record is the more scope there is to generate different profiles to meet different needs. For example, the University of Virginia uses its own base metadata schema, and derives other metadata records from this[30].

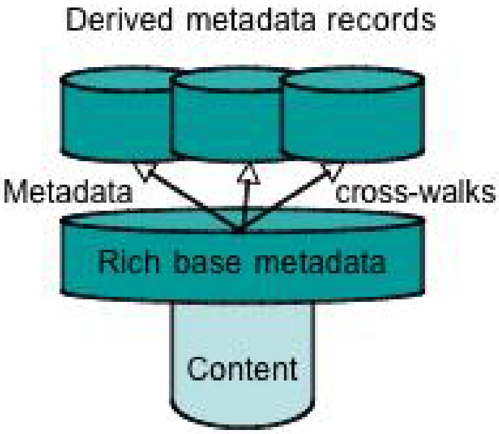


Figure 2: Content and multiple associated metadata records

These metadata can be created either through manual processes or as part of an automatic or semi-automatic submission or editing workflow. Submission can be carried out by the content creator, a designated other or by specific intermediary staff, quite possibly library cataloguers. Different approaches may be required to generate metadata for different purposes, both manual and automatic. Administrative metadata may be sourced locally from other institutional systems, whilst descriptive metadata may be derived from third party information extraction or text-mining tools. The issues around metadata generation are by no means new, but they do require ongoing attention to underpin subsequent developments.

[30] University of Virginia Library Digital Initiatives: Metadata, http://www.lib.virginia.edu/digital/metadata/

It is recommended that tools are promoted and, where necessary, developed to enable repositories to store as rich a base metadata set as possible. This will facilitate the re-use and re-purposing of this metadata through derived metadata records. Additional effort on crosswalks and associated tools to facilitate these is also merited.

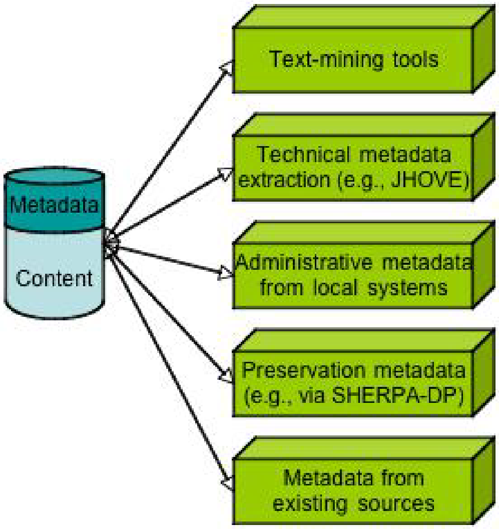

Metadata
Content
Text-mining tools
Technical metadata extraction (e.g., JHOVE)
Administrative metadata from local systems
Preservation metadata (e.g., via SHERPA-DP)
Metadata from existing sources

Figure 3: Non-manual routes for generating metadata

Metadata generation tools are examples of end-user services for the user as content provider. Tools can be hosted locally to the repository or may involve the passing of content to a third party for processing and extraction/generation of the appropriate metadata. Examples of the latter are the approach being explored within the SHERPA-DP project to provide preservation metadata for e-print records and the use of the JHOVE service for extraction of technical metadata. Such services are most likely to be applied at individual repositories, but could potentially work across repositories where access to a larger body of metadata/content would provide the service with more to work with. This is of particular benefit in making use of text mining tools to extract descriptive metadata, which work best with large bodies of metadata/content: descriptive metadata generation has proved particularly problematic to address.

Additional means to generate metadata using automatic means are required in order to provide the rich metadata basis upon which end-user services can operate. It is recommended that investigations into relevant techniques and tools be taken forward with some urgency.

Metadata is often associated with simple content objects (e.g., individual images or document files). If simple objects are combined, for example as component parts of a thesis (documents, images, datasets, etc.), a compound object or, where the components are varied, a complex object is created[31] and metadata for each simple object will become part of the metadata for the compound object. The content to make up the compound object may exist in the same repository, or different components may be stored in separate individual repositories and be held together virtually by the metadata record that may itself be held separately. Multiple metadata records may themselves be combined to form a compound metadata object. For both simple and compound objects the ability to persistently and uniquely identify each component is vital to ensure that the object's integrity and all parts are correctly accessed and made available for end-user services. To facilitate the identification of objects across repositories an identifier resolution service is required that can be used to locate objects on the basis of their identifier. Ideally, identifiers should be location-independent; that is they should not have links with the domain the repository currently sits in. This offers a degree of future-proofing in case content and/or repositories move location, providing confidence that content referenced in metadata records will be available through resolution of the identifier over the long-term.

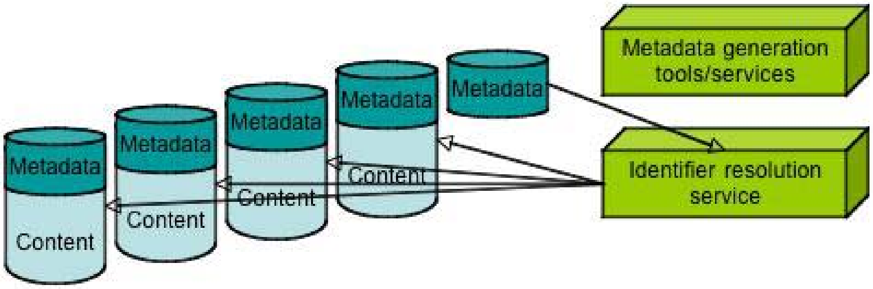


Figure 4: The use of identifiers to enable the relationship between metadata and content

Further attention to identifiers, specifically location-independent identifiers, and necessary resolution systems is recommended to provide greater understanding of their benefits and use. Specifically, the potential use of identifiers by end-user services to provide value-added services should be examined.

Metadata exposure

Once generated, the rich base metadata records and associated derived records can then be used for internal repository management and external access. The former will predominantly be used for internal purposes. There is no reason why this internal record need be the same one that is exposed externally for subsequent use, for which one or more of the derived metadata records can be used. This separation between internal and external roles allows repositories to differentiate between internal data management and external access management, providing flexibility in how repositories present their contents to end-user services.

[31] Complex object definition, http://www.cs.cornell.edu/wya/DigLib/MS1999/glossary.html

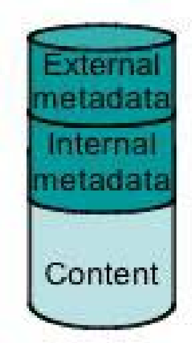


Figure 5: Internal and external metadata layers within a repository

The benefit of doing this reinforces the benefits of generating as rich a metadata record as possible. But it also allows repositories to plan the metadata required for internal purposes and the metadata required for external purposes, and reduces the risk of one need conflicting with the other. This approach should not prevent the base metadata record being exposed in its own right where this is valuable, but does not require that the internal metadata record be the only one that is exposed.

The means through which external metadata is presented may vary. A repository can (i) expose individual metadata records of different formats (e.g., a Dublin Core record, a MODS record, an administrative or technical metadata format) or (ii) it may expose a combination of these, packaged together and exposed for use by other systems. The METS packaging standard[32], which enables the creation of compound metadata objects, can be used for this purpose. Repositories may not know how other systems, such as end-user services, will wish to use the metadata they expose. Offering a combination of options, and the accompanying richness that goes with this, offers flexibility: METS can enable this. The CORDRA model (see section 4.4.2) is based around exposing as rich a metadata set as possible. This model seeks to maintain the richness of the metadata set beyond the repository as far up the information chain as possible whilst also offering flexibility in how this richness can be utilised. Alternatively, there may be a specific need to provide more than one metadata record. As an example, in the Repository Bridge project[33] the University of Wales, Aberystwyth is exposing a METS package for use by the National Library of Wales containing two distinct metadata records related to its theses. A MODS record is included for management and preservation purposes at the National Library, whilst a Qualified Dublin Core record is included for subsequent capture from the National Library by the EThOS project[34] as part of a proposed electronic theses end-user service.

[32] Metadata Encoding & Transmission Standard (METS), http://www.loc.gov/standards/mets/
[33] Repository Bridge project, http://www.inf.aber.ac.uk/bridge/
[34] EThOS project, http://www.ethos.ac.uk

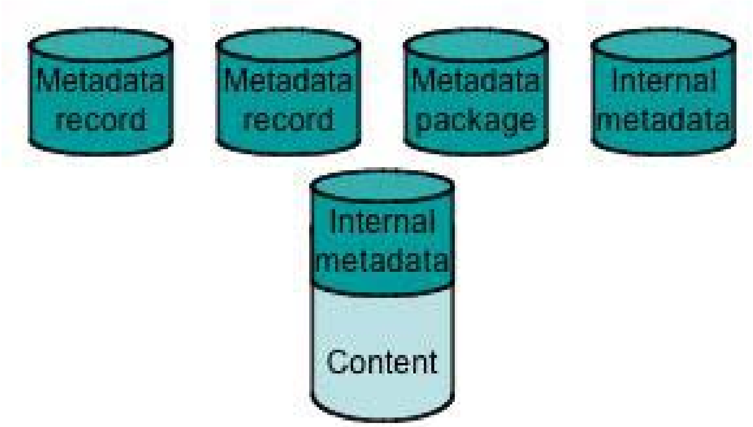


Figure 6: Options for external metadata

It is recommended that additional use cases for the exposure of different metadata records be developed and tested to assess the best and most viable options for repositories to put into practice.

Content exposure

In considering how a repository exposes data for use by other systems, most effort has centred on exposing metadata. As demonstrated by the recommendation of the harvesting model for a UK national eprints/OA journal delivery service [Swan, 2004] and the experience with the OAI-PMH since its inception in 1999, the mechanisms of exposing metadata through harvesting are now well understood. Investigations by Van de Sompel and colleagues at the Los Alamos National Laboratory (LANL), and a series of projects at Virginia Tech, have opened up the possibility of exposing content as well as metadata for subsequent harvesting using OAI-PMH, though the ability to package content together for moving between repositories and other systems is not new in itself. Such exposure of compound objects combines content with its associated metadata and offers a further option for repositories to consider when deciding on how they should make their content and metadata available to others.

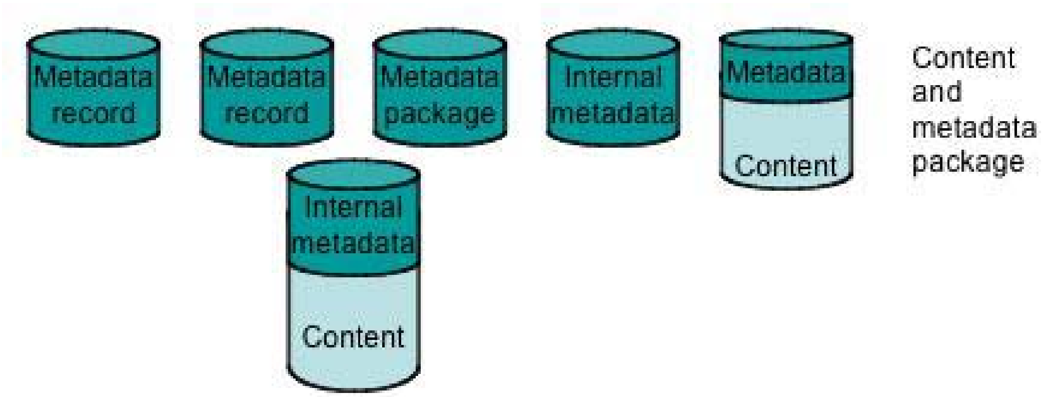


Figure 7: Options for external metadata, including a metadata/content combination

A number of standards exist to facilitate this packaging. MPEG-21 DIDL[35] and IMS CP[36] can contain content either 'by-value', i.e., including it within the package itself, or 'by-reference', i.e., containing a link to the content, which resides elsewhere. IMS CP is a ZIP format that was primarily designed to facilitate the transfer of content between repositories and e-learning systems: there is no reason, however, why its use need be limited to learning materials. The content is associated with a manifest file that determines what the content is and how the different parts relate to each other so they can be used within relevant systems. MPEG-21 DIDL, used in the LANL work, is an XML format that is designed to facilitate both the transfer and use of multimedia materials, especially video, although it can also be used for many other materials as well. It is part of the overall MPEG-21 ISO standard and uses Base64 encoding to include the original content file within the XML. With increasing network bandwidth it has become more feasible to move content around and aggregate it: these standards offer mechanisms to enable this.

Notwithstanding these possibilities, much remains to be understood about the ways in which content should be packaged in this way, how, and why. Both MPEG-21 DIDL and IMS CP have an abstract document model that can be used to guide this process, but consideration of how to apply these models to different content types, and in particular to compound objects, remains uncertain.

It is recommended that the use of a variety of content types, covering both simple and compound objects, with MPEG-21 DIDL and IMS CP be modelled to gather information on the capabilities of these standards to provide value-added end-user services based on exposing content and metadata together (both by-value and by-reference). Alongside this, it is recommended that content exposure also be examined from the perspective of end-user services and how they will best be able to make use of what the packages present: this will assist in validating use cases for exposing content and inform viable and sustainable implementations of such exposure.

Packaging metadata

As indicated, both standards can also package content 'by-reference', i.e., by referring to its location elsewhere, rather than include the actual content within the package. In this they are able to achieve the same as METS described above. As mechanisms for exposing compound metadata objects all three packaging standards are viable. The METS standard lacks an abstract document model, though, and does not offer the same structural capabilities of MPEG-21 DIDL or IMS CP. The structural capabilities that the abstract models provide offer the potential of creating rich packages that contain information about the content that can be used by end-user services (for example the way VLE systems make use of IMS CP packages). METS simply presents

[35] MPEG-21 standard, http://www.chiariglione.org/mpeg/standards/mpeg-21/mpeg-21.htm
[36] IMS CP Content Packaging specification, http://www.imsglobal.org/content/packaging/index.html

what it contains. The benefits of using the structured this will depend on the use cases and requirements. The less structured METS may meet requirements without the need for using a more structured option. It will be necessary to assess specific requirements before deciding on the appropriate standard. Within the aggregation model proposed here, all three packaging standards are valid options.

4.2.1.2 Interfaces

Having considered the options for how repositories can generate and expose their metadata, and/or content, the next step is to consider the interfaces that need to be made available to enable the metadata and/or content to be aggregated. This includes a consideration of the available standards and technologies to support these interfaces.

OAI-PMH

The Open Archives Initiative Protocol for Metadata Harvesting version 2.0 was released in 2002 [Lagoze, 2003]. This version is now established as the standard used by the vast majority of open access repositories exposing their metadata for harvesting. The OAI model contains the concept of data providers and service providers, with the latter harvesting from the former. The Protocol mandates the minimum use of Dublin Core metadata for harvesting, although the protocol document itself and the associated implementation guidelines also mention the possibilities of using alternative metadata formats as well: all metadata to be harvested must be XML, though. Additional metadata formats do not, though, appear to be have widely used so far, except where there is a specific requirement. This may be because OAI service providers have focused their attentions on Dublin Core, making it logical for repositories to do the same: a chicken and egg situation. There are a number of other features of the protocol that do not appear to have been widely implemented within OAI compliant repositories. These are described below. As with many standards and there is additional complexity in implementing additional features, but there are also additional benefits in enhancing access to the metadata and content downstream.

OAI-PMH is a pull mechanism. The repository exposes metadata in accordance with the requirements of the protocol and this can then be harvested by service providers. OAI-PMH offers six verbs that can be used to discover what is available for harvesting and for carrying out the harvesting itself:

- Identify – used to request information from the repository on whether and how it is configured for harvesting using OAI-PMH
- ListMetadataFormats – used to request information about the available metadata formats available for harvesting
- ListSets – used to request information about the specific sets of records available
- ListRecords – used to harvest metadata records
- ListIdentifiers – used to harvest just the headers of records rather than the metadata itself
- GetRecord – used to retrieve single records using the record’s identifier

Through appropriate configuration the OAI-PMH harvester is thus able to discover what is available for harvesting and subsequently harvest the required records for its own needs. This creates a copy of the metadata at the OAI service provider, the harvester that acts as aggregator in this instance. The process can be repeated as necessary in order to update the copy at the aggregator so that any end-user services built upon the aggregation are able to access the most up-to-date content. The harvesting process can capture metadata from across a range of repositories and aggregate the results of this into a single collection.

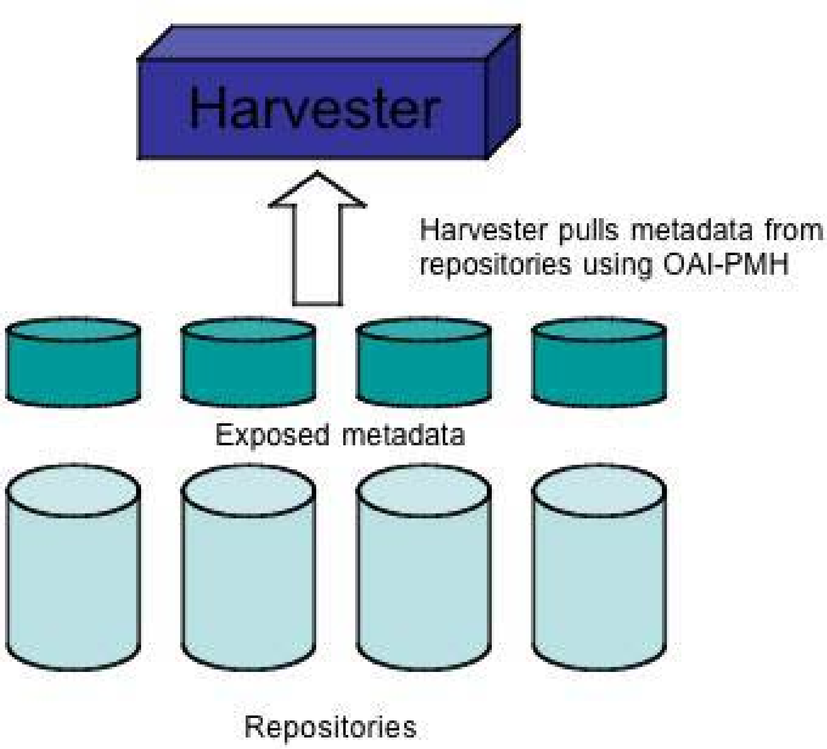


Figure 8: Harvesting using OAI-PMH

OAI containers

When an Identify request is made to a repository the repository as a whole can provide additional pieces of information to assist the harvester. These include the following:

- rightsManifest – information about the rights statements that are attached to metadata records within the repository
- eprints – a means of providing collection description information about e-print repositories
- friends – a means by which a repository can alert a harvester to other repositories that could be harvested
- branding – a means by which a repository can convey branding information related to the metadata being harvested

- gateway – a means for listing associated gateways through which records can be made available for harvesting. An example of this is a Static Repository Gateway (see below).

Containers at the metadata record level can also contain information about the metadata format being used by the repository. The provenance container at record level can contain information about the history of when a record was harvested if implemented.

OAI selective harvesting

The Protocol offers three mechanisms by which selective harvesting can take place, rather than a capture of all the metadata records being exposed.

- By datestamp – When the harvester goes back to a repository it has already harvested it needs to know the changes that have taken place rather than re-harvest everything again (which can take time depending on the size of the repository). This process is common in established harvesting situations, allowing aggregators to maintain a reasonably up-to-date copy of the relevant information.
- By set – Repositories can allocate metadata records to sets, which may provide additional information about the context of the metadata. Sets can provide organisational granularity or subject classification information. This additional information can be used by end-user services. The level of application of sets in the UK varies considerably at this time. An additional container at the set level can contain a description of the set for further information.
- By metadata dissemination type – The metadata format to be harvested has to be specified when carrying out a ListRecords request. This mechanism determines which metadata format the harvester will aggregate and can affect the level of end-user service built on top of the harvested metadata.

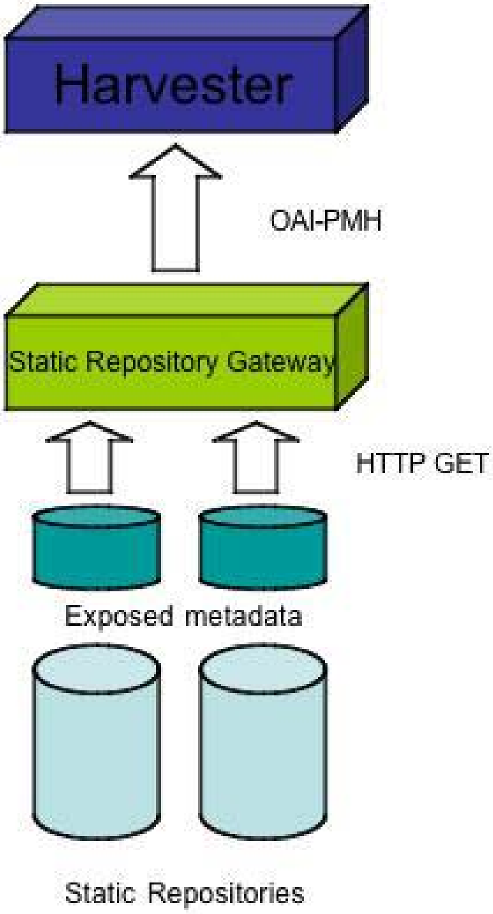
It is recommended that exploration of OAI containers, sets and the use of selective harvesting of different metadata formats is undertaken along with an investigation of tools to facilitate their use. This focussing of the harvesting process can offer much added value in the development of end-user services built on top of harvested metadata.

OAI Static Repositories

Repositories do not have to adhere to the full OAI-PMH protocol if they do not have the wherewithal to implement this. The companion OAI Static Repositories specification[37] offers an alternative for small metadata collections (1-5000 records) where it is not possible to configure a repository for full use of OAI-PMH. Static repositories expose their metadata to a Static Repository Gateway by initiating a link between the two. The

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

Gateway in turn exposes the metadata for harvesting using OAI-PMH. This process, in effect, removes the repository's role in configuring OAI-PMH and passes this to an intermediate, and most likely third party, layer.



Figures 9: Harvesting from Static Repositories

It is noted that Static Repositories are being explored through the STARGATE project[38] in the UK for exposure of publisher metadata into the Information Environment. It is recommended that the results of this work are assessed in tandem with further testing of the specification to see how this alternative means of using OAI-PMH might be more widely applicable to the HE and FE community.

Harvesting metadata and content
The ability to package content alongside metadata, either by-value or by-reference, was described earlier. In the work at LANL, the MPEG-21 DIDL packages were also subsequently harvested and aggregated using OAI-PMH. This is an example of how

[38] STARGATE project, http://cdlr.strath.ac.uk/stargate/

OAI-PMH can be used to harvest far more than simply Dublin Core metadata and is a valuable case exemplar.

It is recommended that harvesting of metadata formats other than simple Dublin Core be tested to demonstrate possibilities for the exposure and harvesting of both metadata and content.

RSS/ATOM

RSS[39] and ATOM[40] are becoming widely known and used as syndication formats to provide news alerts and updates from the repository to the user. A repository offers defined feeds of metadata that individuals subscribe to and access through a browser or desktop tool. Although often commonly perceived and presented as information being pushed to the end-user, RSS and ATOM represent an alternative means through which repositories can present their metadata for exposure: the RSS reader then aggregates the metadata by pulling it from the repositories. The harvesting process assumes metadata will be harvested unless it is specifically withheld. But once it has decided what can be harvested the repository plays a passive role and lets the harvester do the rest of the work. RSS and ATOM require the repository, or content owner, to take a more active role and clearly lay down what can and cannot be exposed through the respective feeds made available. RSS/ATOM readers select what feeds they wish to receive and aggregate what they are given. Exposing metadata through RSS and ATOM can be considered a more controlled way of exposing metadata for aggregation elsewhere.

[39] RSS, http://en.wikipedia.org/wiki/RSS_file_format

[40] ATOM, http://www.atomenabled.org/

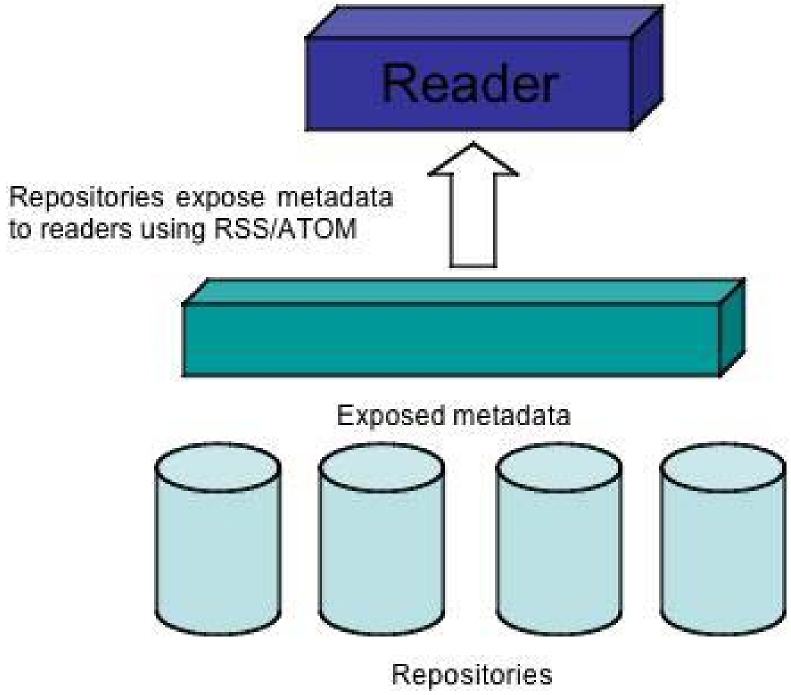


Figure 10: Exposing metadata using RSS/ATOM

In the use of RSS and ATOM it is the reader that is acting as the aggregator. The feeds themselves can be considered mini-aggregations of metadata records, which are added to according to set criteria, often time. The reader can build up a collection of feeds and process these for presentation through end-user services. These may include a web or application interface onto the feeds directly for presentation, but it is also conceivable that the aggregated metadata could be made available to additional end-user services as required for further analysis and use, though there is a lack of widely used tools for this purpose.

RSS is not, unfortunately, a single standard and there are a number of versions of RSS available for use. For repository owners it is a matter of choice and requirements whether metadata adhering to one or more than one version are provided. The same principle applies to a decision about whether to offer RSS feeds or include an ATOM feed as well. In addition to these syndication standards there has been some development of outliner formats, for example OPML. Outline Processor Markup Language (OPML) provides the ability to aggregate metadata about RSS feeds themselves (and many other types of information) for passing between RSS readers. Whilst offering a potentially useful means of aggregating metadata there are some concerns about the structure of OPML records that prevent them from consideration within this model at this time[41]. Outliner formats, nevertheless, offer a potentially future useful means of passing metadata aggregations between repositories and aggregators/end-user services.

[41] OPML, http://en.wikipedia.org/wiki/OPML

RSS/ATOM and content

RSS feeds are intended to be brief, and to connect end-users with greater detail through links back to the main ‘repository’, notwithstanding the fact that in many cases this is a reference to a web page. However, the level of detail within RSS feeds can be extensive. RSS is a common syndication format used by blogs, and it is feasible to include the whole blog post within the RSS feed rather than simply a headline or summary. The boundary between metadata and content blurs at this point,, notwithstanding limits on length and format. In essence, though, RSS provides the ability to feed information at a level determined by the repository source.

The ATOM standard provides an alternative to RSS for the syndication of feeds. Unlike RSS, however, it is not confined to metadata but can contain content using the same Base64 encoding as MPEG-21 DIDL. As such, it can carry out both syndication and packaging functions. It does not offer the same level of structure that MPEG-21 DIDL offers, although experience suggests it is easier to implement (as syndication technologies are intended to be). it does, additionally, offer an alternative means of transferring content for aggregation downstream from the repository, and the potential of using this within end-user services.

It is recommended that the use of RSS and ATOM be investigated as additional standards to OAI-PMH for use in aggregating metadata and content. They offer the potential of targeted exposure of repository resources that may be beneficial in the development of end-user services targeted at specific communities.

Web crawlers

In considering the interfaces that need to be considered by a repository when exposing its metadata and content on open access, it is impossible to ignore the role of web crawlers. Exposure to these offers a route for a repository’s metadata and possibly content to be aggregated with many other resources and easily exposed to end-users through web search engines such as Google and Yahoo!. A number of examples of how this can take place already exist. The University of Glasgow opened up its repository to Google through the use of the inurl syntax and has registered its repository with Google Scholar [Nixon, 2005], whilst OAIster has exposed its aggregated content to Yahoo![42].

[42] OAISter exposure to search engines, http://oaister.umdl.umich.edu/o/oaister/sru.html

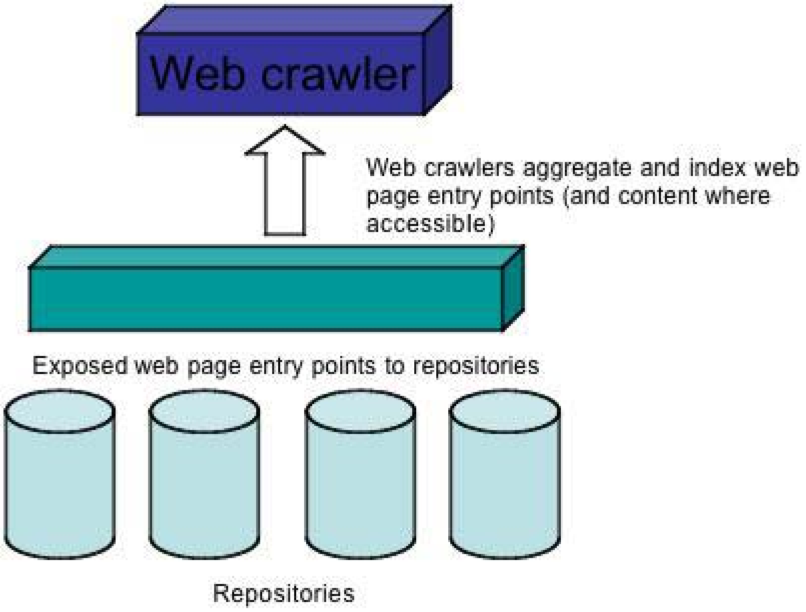

Web crawler
Web crawlers aggregate and index web page entry points (and content where accessible)
Exposed web page entry points to repositories
Repositories

Figure 11: Aggregation by web crawlers

The decision is not always deliberate, though. The University of Edinburgh found that Google was crawling its repository without any agreement or decision on Edinburgh's part. Whilst useful, this does offer the dilemma of how to control what is exposed and not: items deleted from the repository could still be found in Google because of its caching methodology. With OAI-PMH and RSS/ATOM, repositories have some control over how metadata and content are exposed: with web crawlers this is not always apparent.

Exposure of repository metadata and content through web crawlers offers a valuable means of bringing end-users to a repository through commonly used web search engines such as Google and Yahoo!. However, such crawling may not always lead to the level or type of aggregation and exposure that the repository is seeking. It is recommended that the exposure of repository contents within web search engines be examined in closer detail to assess the paths of exposure that exist and the implications for repositories of exposure via this route.

4.2.2 The aggregator view onto repositories

The exposure of metadata and/or content to aggregators has been discussed from the perspective of the repositories so far. This aggregation process can be a one-way path. However, if aggregation is a one-way path there is a risk to the repositories that the aggregation has not taken place correctly and that the repository's resources will be misrepresented when viewed through the aggregation, as can be the case when a web crawler aggregates a repository's content. Likewise, the aggregator may wish to inform the repository that the aggregated metadata/content was not presented correctly and that errors have occurred. Either technical or non-technical ways in which the aggregator and repositories can communicate need to be set in place in order to ensure there is no information loss as a result of the aggregation taking place.

For OAI-PMH harvesters this is a critical function. The protocol itself allows for errors to be reported automatically, and these can be interpreted and acted upon accordingly. The OAI-PMH validator at Cornell University also allows repositories to validate themselves as OAI compliant[43], running a series of standard requests and checks. Depending on the role and purpose of the aggregator, however, these may not provide all the information needed, and in particular will not pick up on errors or misunderstandings in the metadata itself, just its structure.

For RSS bi-directional communication can be established using additional software such as Microsoft's Simple Sharing Extensions[44]. This allows systems (potentially repositories and aggregators) to use RSS as an asynchronous communication protocol between them.

4.2.2.1 Enriching the aggregation

The benefit of being able to automatically generate metadata at the creation stage was discussed earlier. Being able to generate metadata from different sources helps lead to the richer metadata set that can provide greater flexibility in how the repository exposes its metadata and content. Once an aggregator has amassed metadata and/or content from a number of repository sources it is also feasible for similar metadata generation to take place based on this aggregation. This may involve the generation of additional metadata, or the enhancement of existing metadata to improve the quality. In both cases, the aggregator can pass the enhanced metadata back to the host repository to increase the quality of its collection and further enrich what it currently holds. These approaches are geared toward enhancing metadata and it is unlikely that content itself would be altered (certainly not without prior agreement with the content owner). Nonetheless, the content may be used to inform the metadata enhancement process.

Both these approaches were taken by the ePrints UK project[45]. Web services seeking to add subject classification metadata and enhance author names through use of a name authority file were developed, as well as a Web service to automatically parse bibliographic citations within article references into structured forms, using the OpenURL standard. The subject and citation services sought to capture the content as referenced in the harvested metadata record to act as source material. All the services were demonstrated in beta.

[43] OAI Data Provider Validation and Registration, http://www.openarchives.org/Register/ValidateSite
[44] Microsoft's Simple Sharing Extensions, http://msdn.microsoft.com/xml/rss/sse/
[45] ePrints UK project, http://www.rdn.ac.uk/projects/eprints-uk/

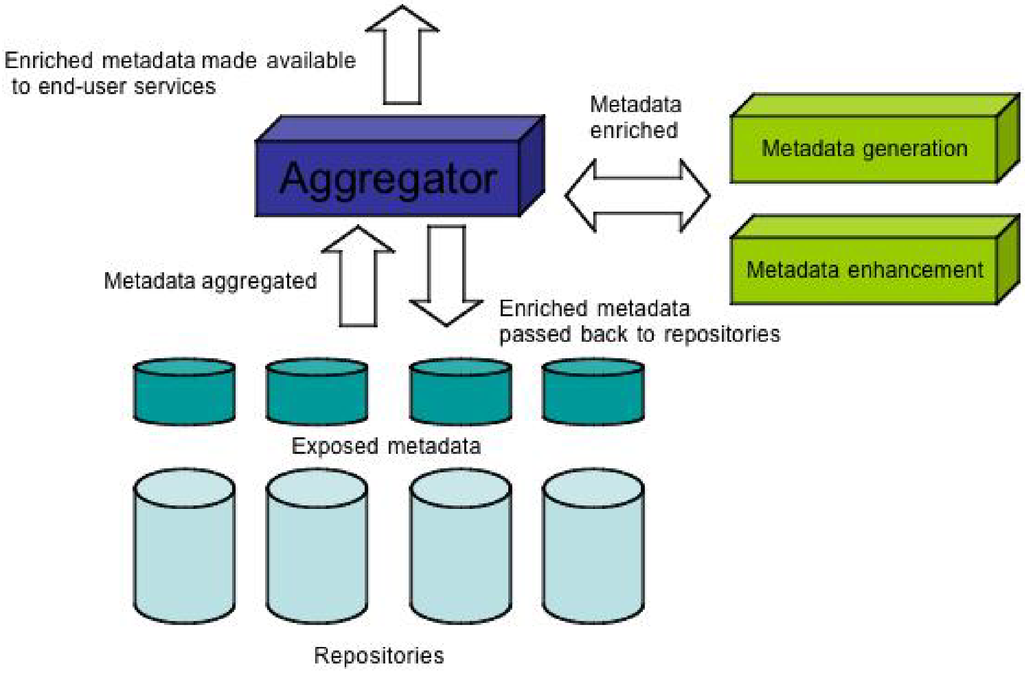


Figure 11: Enriching metadata using aggregations as the source material

Notwithstanding the benefits such enrichment can bring, the way that content was linked from the metadata records led to a number of difficulties in fully developing the subject and citation services. Bearing in mind the ability to package content together with metadata, it would be valuable to re-visit enriching services by investigating whether the use of compound objects would enable a more stable implementation.

The name authority service required the use of a name authority file to check author names against. This is currently lacking in the UK, though it has been suggested that the administrative staff records kept by individual institutions offer a possible alternative. This approach is used at the University of Southampton, which links its repository to the local staff ID database to ensure consistent naming of submitting authors.

It is recommended that the use of compound objects be investigated as alternative sources of information to underpin metadata enrichment services, including the packaging of appropriate institutional and other available information that may be of value to the process. It is also recommended that authoritative institutional lists of author names be investigated as a distributed name authority service.

4.2.3 End-user services

The goal of aggregation is to act as an intermediary between repositories and end-user services. Having considered the role of repositories and how they can make metadata and content available to the aggregator, this section examines the relationship between the aggregator and the end-user services from both perspectives.

4.2.3.1 Aggregator view

Aggregation provides a body of metadata - and possibly content - that can be used through end-user services. The main point of access for the end-user is likely to be via a web page, though it is recognised that desktop and mobile tools and applications could be alternative access points. What interfaces can the aggregations offer to connect them to these access points?

Web crawlers will provide their aggregations through the web search engines that are their main interface. In turn these web search engines might be aggregated further by web meta search engines, providing a combination and comparison view of the different services available. The predominant access point is a web interface onto the aggregation. In the past year the advent of Web 2.0 has allowed web crawler and other web-based aggregations to expose themselves in more flexible ways as well. Google, Amazon, Flickr and eBay have, amongst others, made available APIs that allow others to build services on top of the exposed content (for example a service built on top of Google Maps).

RSS and ATOM feeds and aggregations can be read through a number of alternative routes. Readers can be separate desktop applications, they can be embedded services within web browsers, or they can be embedded within web pages for display as part of a wider end-user service on the web, such as an institutional portal or library catalogue. In many cases these access points are geared toward individual RSS/ATOM feeds, and the level of aggregation is solely within the individual feed. Aggregator tools, however, can bring a number of feeds together and aggregate at a broader level, providing a view across different repository sources. The delivery of RSS/ATOM feeds into ‘readers’ highlights the use of browsing as the main form of user interaction. For many aggregations some form of search capability will also be of value.

An OAI-PMH aggregation offers a wide range of possibilities for provision through end-user services. The aggregation in itself is not a delivery format, as RSS/ATOM aggregations are, but it requires that end-user services interact with it via other interfaces. A number of these are listed in Table 2.

Access point	Notes
Web interface	Direct access for search and browse enabled through web access onto indexed aggregation. This may involve direct web access or embedding of such access in distributed services elsewhere on the web.
SRW/U	Structured search of an aggregation using distributed search protocols
RSS/ATOM	The OAI-PMH aggregation can itself by the origin of RSS or ATOM feeds for delivery through the variety of readers available for these standards
OAI-PMH	The OAI-PMH aggregation can itself be harvested for additional aggregation elsewhere
OpenURL	The aggregation can be used as an OpenURL target to facilitate location of individual items
SOAP	A Web services interface that allows the aggregation to be embedded as part of a wider Web services environment
Semantic web interfaces	Interfaces that present semantic information about content that can be used to build services upon. Often based on RDF.

Table 2: Options for exposing OAI-PMH aggregations to end-user services

The flexibility of the OAI-PMH aggregation is very apparent, with the ability to feed into many different end-user service scenarios. Most have been tested and have been found to work well in different circumstances, though most are not widely used. The use of OpenURL is largely untapped as an interface, particularly the use of the NISO Z39.88-2004 standard[46], and an assessment of how this might be best used with both repositories and aggregators will be of value. RSS/ATOM aggregations too offer a wide range of options for inclusion within end-user services, though there is scope for tools that allow these aggregations to be re-used beyond simple presentation through a reader.

Both OAI-PMH and RSS/ATOM can target the user where they are. Web crawler aggregations, predominantly rely on the user going to the web search engine and searching the aggregation at that point. This latter model offers a relative simplicity of access, but fails to take best advantage of the aggregation and the flexibility this offers. The advent of APIs, though, promises to allow web crawler aggregations to be presented and made available for use in flexible ways.

It is recommended that the use of OpenURL 1.0 over OAI-PMH harvesters and repositories in general, the use of tools that allow re-use of RSS/ATOM metadata feeds beyond presentation, and the use of appropriate web search engine APIs be tested further to establish their value in facilitating interaction with aggregations through end-user services.

[46] OpenURL 1.0 z39.88-2004 standard, http://www.niso.org/standards/standard_detail.cfm?std_id=783

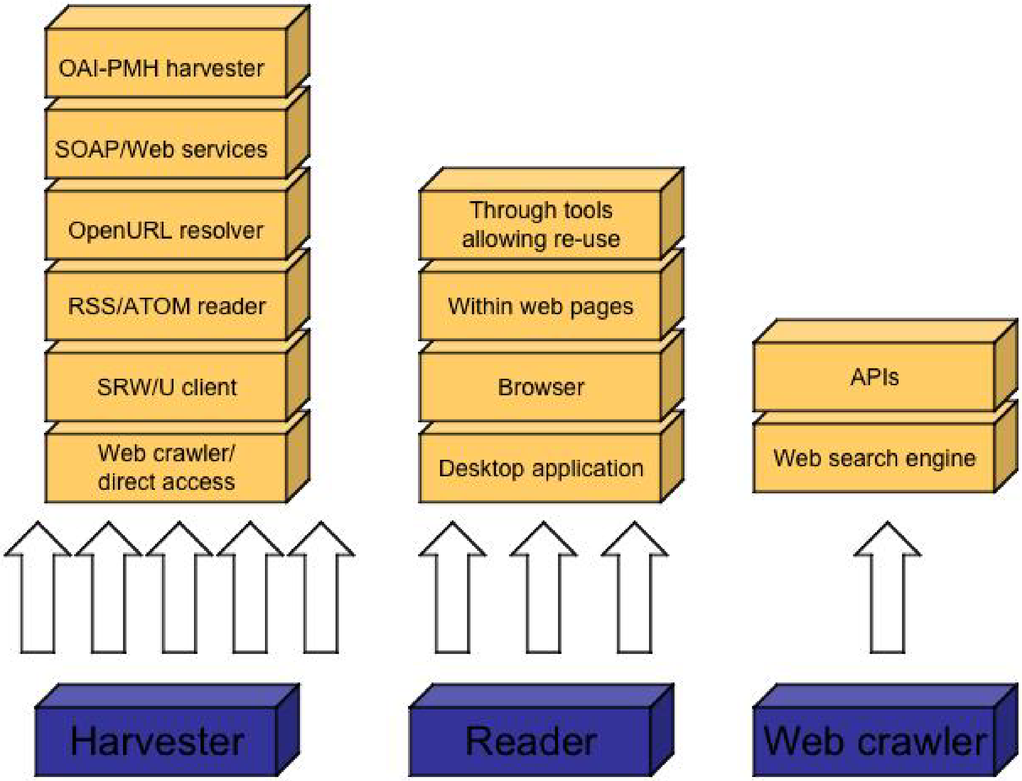


Figure 13: Options for linking aggregations to end-user services

4.2.3.2 End-user service view

From the perspective of an end-user service aggregations offer convenient collections of metadata and content at which to target access. Aggregations bring together resources from across a wide variety of repository sources and allow an end-user service to use these as a primary point of access rather than interact with repositories individually. Where the aggregation is predominantly of metadata the end-user is likely to want to locate and access the full content after discovering the metadata. There is also value in allowing access to the rich base metadata record held by the repository in order to enable extra functionality based on this additional information. This implies a need for the end-user service to establish a connection directly with the repository, albeit one that is brokered through the aggregation. The ability to link through avoids dead-ends for the end-user and the frustration this can bring.

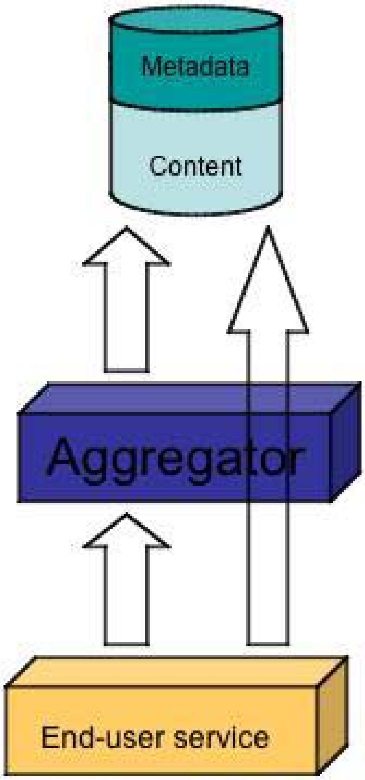


Figure 14: Accessing the repository through the aggregator

4.2.3.3 Intermediary shared infrastructure

There is a range of tasks it may valuable or necessary for the end-user service to carry out to ensure that access to the aggregation(s) and/or repositories of choice will provide the end-user with what they need. A selection of these is listed in Table 3 alongside separate third party services that can help to provide these functionalities. In addition the metadata generation and enhancement services discussed in earlier sections are also applicable here. The tasks can be carried out by the end-user service, though there is value in separating them out to prevent duplication of effort by end-user services. By no means would all of these necessarily be required for every interaction between the end-user service and the aggregation(s), but they may be required in certain circumstances to ensure the interaction takes place correctly and appropriately.

Task	Associated services	Notes
Content authoring and management	Authoring tools or appropriate export functionality from systems producing data	Authoring and editing predominantly takes place at the repository level or across repositories.

	Annotation services	Adding to the metadata may take place at repository, aggregator or end-user level, although all annotations need to be associated with the original object wherever this may be
Content management and preservation	File management services	Functionality that allows the contents of an aggregation or repository to be organised
	File migration services	Services that allow content to be migrated from one format to another for management and/or preservation purposes
Authentication and authorisation	Authentication and authorisation services. Examples include ATHENS and Shibboleth	In the open access landscape this should not be an overriding issue most of the time. However, there will be instances where appropriate AuthN/AuthZ functionality is required, particularly at the deposit and ingest stage and where there are restrictions on the full content when searching
Assess access rights	DRM and legal services	Although there are unlikely to be many rights issues for metadata in an open access environment there may be restrictions on the full content that third party services can help assign and/or manage.
Identify repositories and aggregations to access	Service and collection registries	Registries act as catalogues of service and collection information about aggregations that end-user services can make use of determine the appropriateness of each for the task at hand.
	Repository registry	A specific instance of the above category, but also a specifically important category in the context of this study. Allows aggregators to identify sources to aggregate as well as acting as a source of information to aid the location of objects
	Identifier resolvers	The registries will contain identifiers for each aggregation – these identifiers will need to be resolved to allow access

Building a search	Terminology services	Allow end-user service to map search terms across controlled vocabularies where used or link between common terms, possibly using a Topic Map
	Authority services	Allow end-user service to ensure the search terms match with authoritative terms used within aggregated metadata
Receiving results/output	Format conversion/dissemination services	Enables the results of a search/locate task to be output according to the desired format

Table 3. Intermediary shared infrastructure services supporting interactions between end-user services and aggregations/repositories

It is recommended that further investigation of these intermediary shared infrastructure services take place. Priorities are registries, identifier resolvers, metadata generation and appropriate authoring tools to support ease of interaction with and across repositories

These services can be used by the end-user service as required. They may be part of the end-user service, aggregator or repository, they may not exist at all, but they do add value and provide options towards enabling accurate and useful access across repositories.

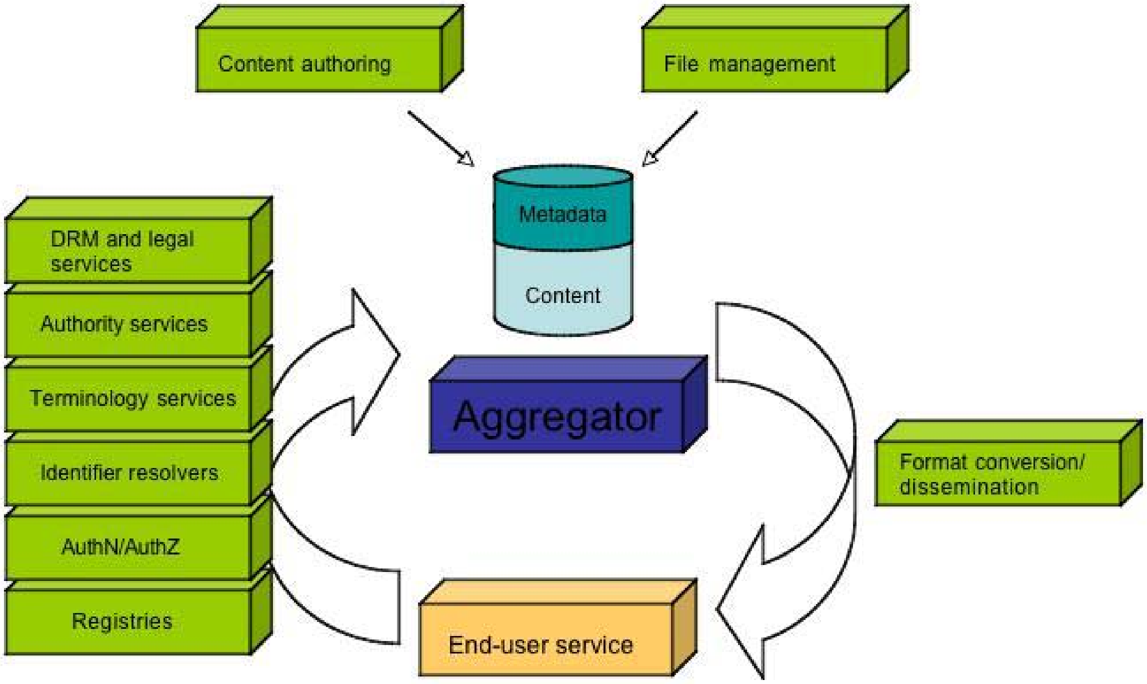


Figure 15: Intermediate services that can be used by end-user services

It is recognised that only a small proportion of the possible intermediary services are currently available for widespread use. This has the potential to hamper full, effective end-user services from being developed. Notwithstanding this, it is noted that one of the possible reasons for this has been the relative isolation of intermediary and end-user service developments thus far.

It is recommended that links be established in future developments between all four components involved: end-user services, aggregators, repositories and intermediary services. This will ensure a better understanding of relative needs, enable the practical demonstration of the benefits such interaction can have, and allow the investigation of workflow throughout the systems concerned.

4.2.2 Model layers

The components and layers of the proposed aggregation model have been discussed so far without reference to where they might sit organisationally. The three layers, repositories, aggregators and end-user services can be encompassed within a single unit, but can equally be separated and operated individually. In an aggregation model the aggregator will normally be separate from the repositories it is aggregating from, simply because it is working across a range of these. The particular circumstances of the components involved and the aims in building the three layers will have an impact on this.

In considering OAI-PMH as the aggregator, where the layers sit will depend on the scale of the proposed end-user service. The University of Glasgow hosts all three layers internally, with the harvester and end-user service both provided through the PKP harvester software being used. On a broader scale OAIster is completely separate to the repositories it harvests globally. Its default configuration combines the harvester with the end-user service onto the aggregated metadata. Recent developments such as exposing metadata to Yahoo! and exposing an SRU target for searching have sought to separate the two top layers.

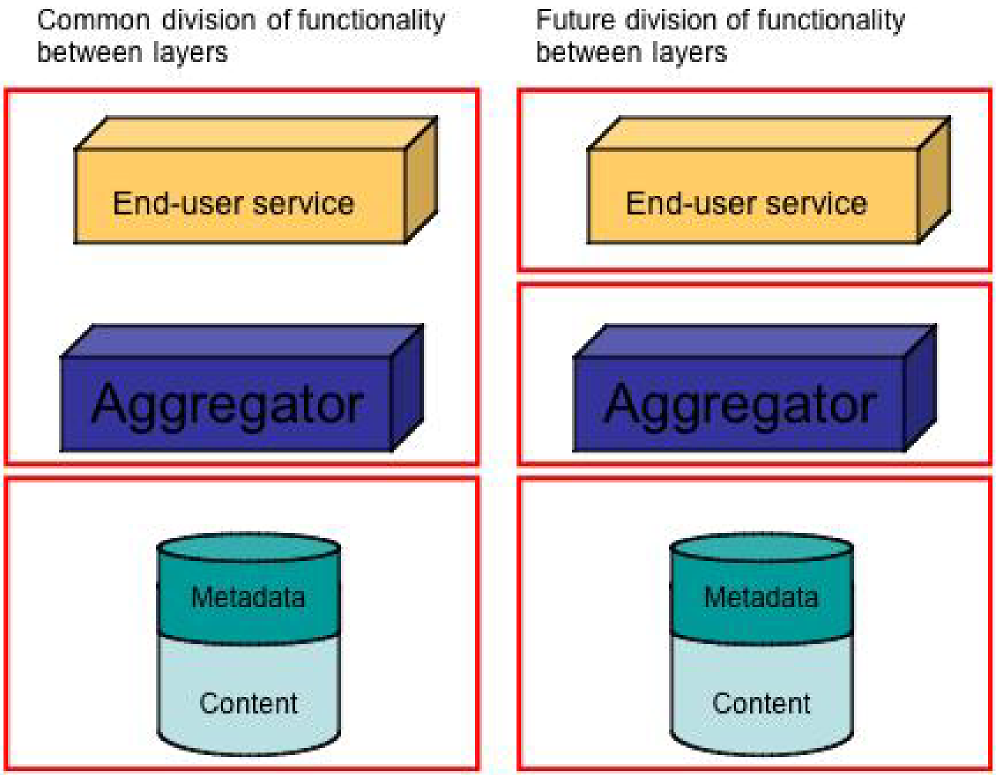

Common division of functionality between layers
Future division of functionality between layers
End-user service
Aggregator
Metadata
Content
End-user service
Aggregator
Metadata
Content

Figure 16: Separation of layers in the aggregation model

The creation of aggregations by RSS readers and web crawlers follows the same approach. RSS feeds are generated by the repository, though are aggregated by a separate RSS aggregator. This aggregator will often provide the end-user service itself, but it may act as a broker and provide the aggregated RSS feeds to another application for use there, e.g., the separation between an RSS aggregator and the institutional portal that presents the results. Web crawlers also aggregate separately from the repositories they are accessing and predominantly provide the end-user service themselves. This represents a combination of the fusion and presentation layers within the JISC Information Environment.

As indicated in section XX, though, a number of aggregators on the web, including search engines like Google, have recognised that there is additional value in separating the role of aggregator and the role of providing an end-user service. Whilst offering what might be considered a basic service directly on top of the aggregator, and this has been important in establishing trust in the aggregation, Google, Amazon, eBay and others have exposed their aggregations through APIs for others to build their own end-user services.

This Web 2.0 approach has at its heart recognition that allowing re-use of aggregated metadata and content brings added value. The OAIster developments mirror this, and both developments mirror the aims of the JISC Information Environment in establishing individual components that can be linked together in flexible ways through the use of standards (open standards where feasible). This approach is also service-oriented, as recommended through the coordinating e-Framework initiative[47], and offers the opportunity to more flexibly combine the different components of the aggregation model toward meeting user needs.

It is recommended that where feasible developments of repositories, aggregators, end-user services, and intermediary services should move toward a service-oriented architecture and establish separate layers for the aggregation model to maximise the flexibility available for building end-user services to meet user requirements.

4.3 Theory into practice

Putting theoretical models into practice often requires necessary pragmatic choices. In the discussions for this study a number of issues arose that may impact on the practical implementation of the aggregation model and which require attention.

1. Aggregation is a federated approach to gathering information together so it can be exposed through end-user services. Where this federation is distributed between institutions network sustainability and stability is necessary to ensure the aggregation works effectively.

2. Persistence of the repository sources available for aggregation is another factor. In considering the use of OAI-PMH-compliant repositories, Graham Turnbull at SCRAN recognised their value whilst also insisting that for an aggregator and service the repository should provide decent metadata, it should allow the service provider to configure a good request, and there should be a persistent location to click through to for access to the full content. Repositories need to take these factors on board to ensure service providers can effectively use them. SCRAN also discovered that users are unforgiving when part of a service fails: attention needs to be given to ensuring that all parts of the aggregation model are available to prevent user dissatisfaction.

3. The software that underpins all parts of the model can have a major impact on how the different parts connect. Selecting software for a repository will have many factors associated with it, but ultimately the repository has to use what is provided once the choice is made. Although it is possible to opt for a flexible software architecture such as Fedora[48], where choices can more easily be slotted in and out through extensions, many institutions will need the software to come out of a box, ready to go. To ensure that repositories are able to best interact with aggregators ongoing communication with software developers is thus necessary. When implementing software there is a balance to be met between a ‘software can do anything’ approach versus a ‘yes, but costs constrain functionality’ reality. In the former case software can do anything so

[47] e-Framework for Education and Research, http://www.e-framework.org/

[48] Fedora Digital Repository system, http://www.fedora.info/

long as it adheres to open standards to enable interaction with it using those standards (e.g., OAI-PMH, RSS, etc.). In the latter case there is a realistic view that complete implementation of open standards is not currently feasible in some software and that this will have an impact on end-user services down the line. The more that can be done to move to the former, the better for practical and long-term implementations. There will always be a legacy issue, however, and there needs to be ways in which older repositories can be included within aggregations: this may involve appropriate software enhancements to convert a legacy system or the use of a third party interface layer such as the OKI DR OSID[49] to overlay the system and provide the interoperability required.

4. The granularity of information has been stated as important. Certainly, granularity of access requires granularity of content, and granularity of identifiers in particular. The level of granularity to be implemented needs to be multilateral where possible. If aggregators cannot make use of high granularity where it is offered there is less incentive to offer it: the more that repositories offer richer metadata the more incentive there is to aggregate it. There is a real need to for communication between repositories and aggregators to agree what will be exchanged in order to facilitate rich services. Identifiers offer an example of how future end-user services can be developed on top of a highly heterogeneous repository environment by pointing to objects wherever they might be. The ability to abstract out identifiers to a common schema, for example infoURI[50], that sits above existing records may help provide commonality across repositories.

4.4 Specific architectural instances

Many of the ideas and suggestions that emerged from the interviews for this study have been encapsulated within exemplar architectures that have been developed to specifically investigate how best to provide services across repositories. The aDORe and CORDRA initiatives are described and compared here.

4.4.1 aDORe

aDORe [Van de Sompel, 2005] is a standards-based, modular repository architecture that has been developed at the Los Alamos National Laboratory in the US. It has evolved out of practice to manage the broad range of repositories within LANL and represents the creation of an interoperable federation across heterogeneous repositories. The architecture has been built in the context of use within LANL, though the components could be provided on a more distributed basis for federations between institutions as well. Version 1.0 of the various software components has been made

[49] Open Knowledge Initiative Digital Repository OSID, http://www.okiproject.org/
[50] infoURI, http://info-uri.info/registry/docs/misc/faq.html

available to the community for use[51]. aDORe is not, however, presented as a repository solution, rather a set of components that can be used to test and showcase the principles applied in its design.

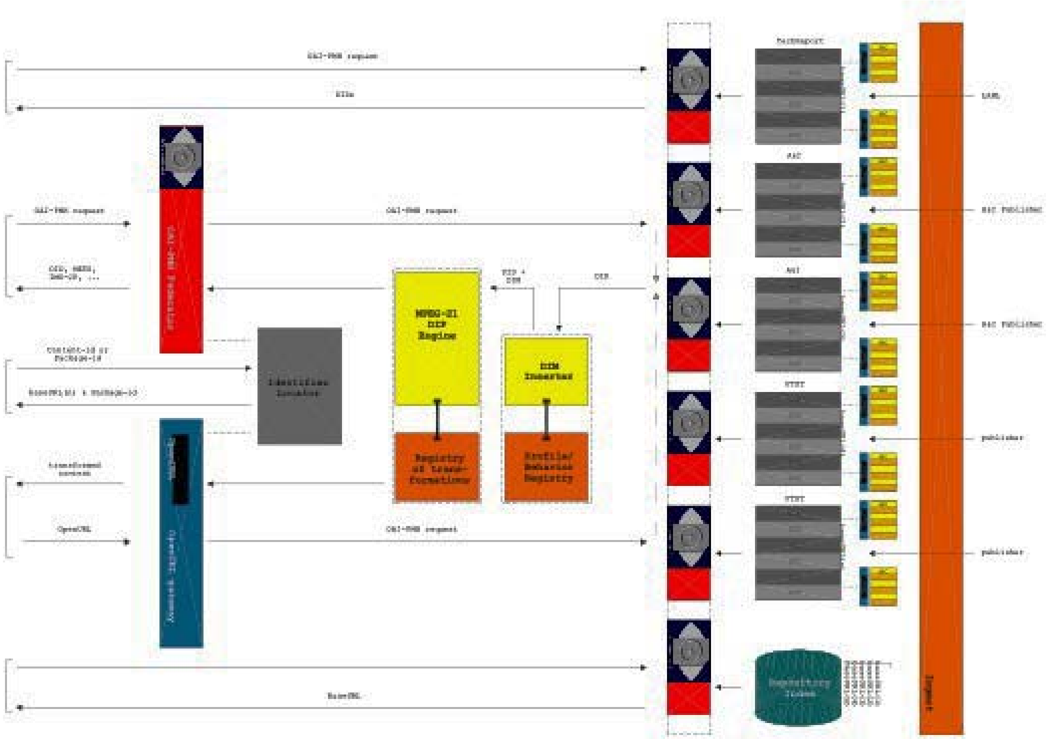


Figure 17: aDORe repository federation architecture

There are three key technical factors at the core of the aDORe architecture:

- All repositories in the federation are exposed via OAI-PMH interfaces. This allows them to be queried and/or aggregated as required.
- The main entry points to the aDORe architecture are an OAI-PMH federator and an OpenURL resolver. These standards-based entry points, and use of OAI-PMH for communication between components, highlight the ability to federate repositories using open standard interfaces, and the ability to present both user and machine interfaces using the same protocols.
- All metadata and content are stored within MPEG-21 DIDL packages, adopting a compound model approach. Such packages may contain simple metadata or compound metadata and content objects, but the use of the standard allows for both to be accommodated equally and alongside each other. All sub-components of the package can be uniquely and persistently identified using an appropriate identifier.

[51] aDORe archive source distribution, http://african.lanl.gov/aDORe/projects/adoreArchive/download/src/index.html, aDORe DIDLTools, http://african.lanl.gov/aDORe/projects/DIDLTools/

Other components of the architecture include an identifier locator/resolver, which keeps a record of all the identifiers within the DIDL packages and can resolve identifier queries to it wherever the object may sit in the federated repositories, and a repository registry to keep track of information about the repositories being federated. These and other management tools help monitor the workings of the federation and would be essential if aDORe was to be implemented on a cross-institutional basis. In order to facilitate use of MPEG-21 DIDL the LANL team have developed tools to allow the creation of such packages, their storage, and their harvesting by an appropriate OAI service provider.

As well as supporting a range of standards-based queries the aDORe architecture has a service overlay dynamic disseminator module that can control the format of how any particular object is disseminated. The object needs to be stored in RDF to enable transformations between different formats, but having this capacity can provide real added value to the end-user. Overall, the ability to organise repository federations in this way allows for flexible implementation that can adapt to changing circumstances over time.

The aDORe architecture has emerged out of practical testing and experience and it works, at least within a single institution. It will be of value to use the aDORe architecture in a cross-institutional scenario using the software components released by LANL. Such testing would provide a testbed for experience in using MPEG-21 DIDL plus experience in the adoption of open standards to support federation of repositories.

4.4.2 CORDRA

CORDRA (Content Object Repository Discovery Registration/Resolution Architecture)[52] is an ongoing initiative and partnership between the Advanced Distributed Learning Initiative (ADL)[53], the Corporation for National Research Initiatives (CNRI)[54] and Learning Systems Architecture Lab (LSAL)[55]. The aim of the work is, as defined on the CORDRA website, to enable:

> “An open, standards-based model for how to design and implement software systems for the purposes of discovery, sharing and reuse of learning content through the establishment of interoperable federations of learning content repositories.”

CORDRA recognises the heterogeneous nature of the current repository landscape and seeks to provide guidance for solutions that can work above and across a broad range of repositories. Note that although the CORDRA team and others have developed implementations [Jerez 2006, Manepalli 2006], CORDRA itself is a model and many details will be addressed by the specific implementation, not by CORDRA.

[52] CORDRA, http://cordra.net/
[53] Advanced Distributed Learning Initiative, http://adlnet.org/
[54] Corporation for National Research Initiatives, http://www.cnri.reston.va.us/
[55] Learning Systems Architecture Lab, http://lsal.org/

Nevertheless, CORDRA has developed registry code as part of its investigations that it is planning to release as open source. Although initially targeted at helping to address federation across learning object repositories the CORDRA model can be used across repository sectors to guide system development.

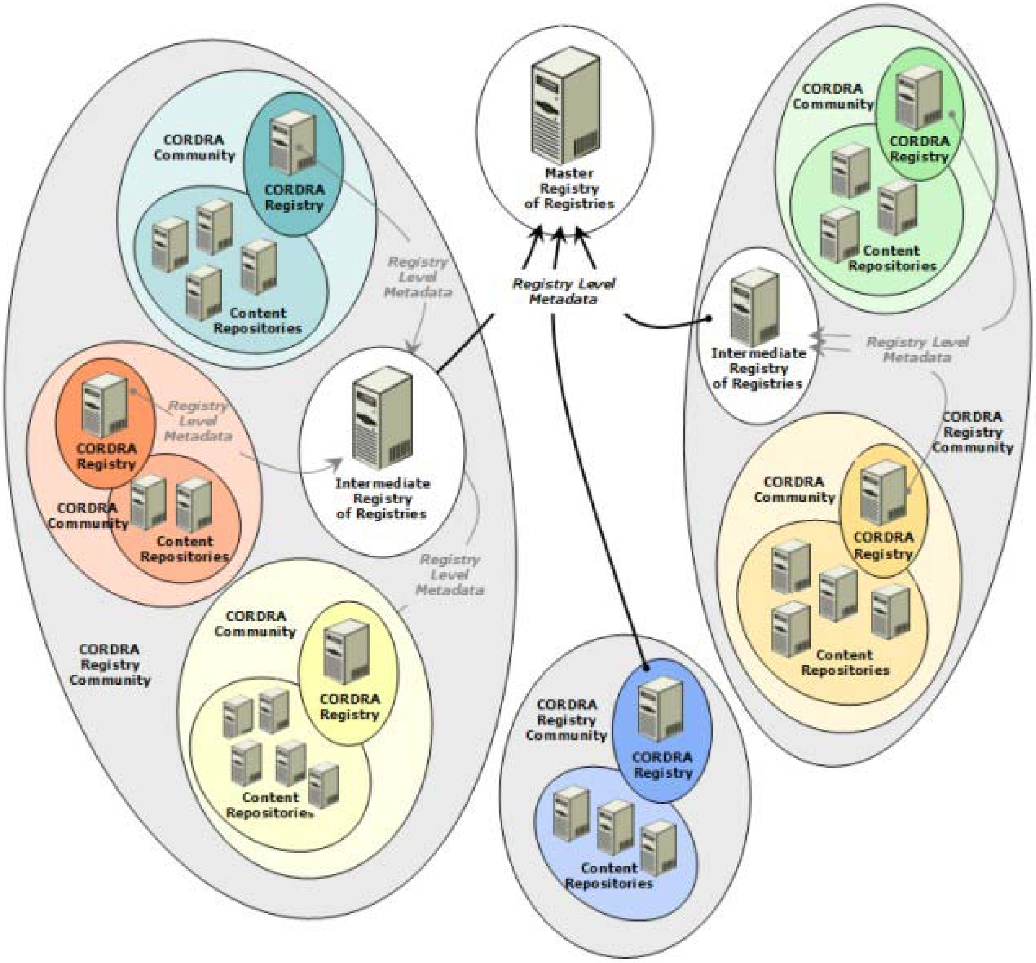


Figure 18: CORDRA community repository federation

The interoperation within CORDRA centres on the role of key registries:

- A master catalogue that holds all metadata and is the primary point of entry to the CORDRA architecture by end-user services.
- A repository registry that keeps track of all the repositories within the federation.
- A system registry that holds information about the CORDRA model and how it is being implemented

The master catalogue aggregates metadata as exposed by the repositories within a federation. This aggregation can be implemented by harvesting using OAI-PMH, as in the FedCOR project [Manepalli, 2006], or could be implemented via push mechanisms. CORDRA overall, though is agnostic in regard of the standards employed. The aggregation at the registry level is the core point of access to the CORDRA model by end-user services. By moving metadata as far up the chain towards end-user services as possible a CORDRA aggregation seeks to minimise the loss of information that can result from bringing together metadata from disparate repositories: metadata is aggregated to maximise access and avoid dumbing down. This pooling of metadata also provides the basis upon which end-user services can be built.

The principles involved in CORDRA are clearly of value in addressing the development of end-user services across open access repositories. The largely theoretical nature of CORDRA means it is difficult to assess what specific factors may be encountered in putting this into practice. In reality, and pending further development of the model and test implementations, CORDRA will be best employed as a checklist of factors that need addressing when establishing an open access repository federation.

4.4.3 Summary

As agreed by those involved the aDORe and CORDRA activities have largely reached a common goal through following different routes. Both seek to provide as much information as possible for use by end-user services through aggregation. Whilst aDORe has nailed its colours to the mast as regards the technologies and standards employed, CORDRA offers a more open-ended model that could be used with a range of technologies and standards. Both architectures espouse the use of a repository registry, a valuable intermediary service that can help to keep track of the repositories within a federation: in a wide open access environment this would be essential to allow the best implementation of either aDORe or CORDRA.

Both architectures also encompass a sense of coordination across a federation to best enable the functionality they are designed to achieve. Open access, through its promotion of rapid exposure and dissemination using predominantly OAI-PMH, tends to favour a light or even at times non-existent coordinating touch. Experience has demonstrated that this can cause problems and successful service providers are those either where there is some sense of communication and coordination between aggregators and repositories or where the service provider has taken it upon themselves to re-factor what has been aggregated. In providing for greater coordination aDORe and CORDRA increase the effort involved in providing open access. But they also provide for the potential of value added and targeted services that could enhance open access.

aDORe components have been made available for testing and it is recommended that this is investigated further. Such testing would also make it possible to contribute to the CORDRA initiative through assessing the aDORe implementation against the CORDRA model. Further work needs to be carried out before CORDRA can be considered a serious guide to setting up federations, but testing of aDORe can help feed into this whilst providing a real, implementable solution to examine the standards, technologies and issues further.

4.5 Aggregation and the user as reader, author and manager

Having described the proposed model for facilitating end-user services across repositories, it is important to ensure that this model meets the needs of the respective user-groups identified in earlier sections. These needs are addressed in Table 4 (with reference to Table 1).

User group	Benefits of aggregation model
Repository Managers	Aggregations relieve repository managers from the maintenance of direct services (albeit that these may still be required depending on circumstances) and dealing with end-user services themselves
	End-users can be directed straight to the required resource from an aggregator rather than through the repository front-end
	Aggregations offer an alternative route for enhancing metadata held within a repository through feeds back from aggregator of enhancements carried out at aggregator level
	The aggregator may address authentication and authorisation issues where required as trusted intermediary
	Aggregation by a third party service can facilitate preservation through appropriate metadata provision and/or content storage
	Repository maintains control over content whilst releasing metadata (though note possibility of exposing content for aggregation as well)
End users as readers and searchers	Aggregations offer breadth of access across many repositories, relieving end-users from accessing each one individually
	Aggregations can offer control and personalisation of content access (e.g., using RSS) allowing the end-user to determine which sources they have access to
	Aggregations offer the capability of developing specific D2D services for specific end-user groups
End users as content providers	Aggregations provide exposure for content providers to make their work available widely
	Such exposure can be focussed around related materials such as aggregations offering subject entry points
	Aggregators can provide preservation and metadata enhancement capabilities to support the long-term storage of and access to the content
Content aggregators	Aggregators can offer added-value services of their own to enhance aggregated metadata and supply this back to the repositories concerned

	Aggregators can also use the amalgamated collections as the basis for analysis, such as text and data mining
	Aggregators can provide a brokering role to facilitate access by end-user services (possibly including a marketing element)
Meta-users	Aggregations can offer a single point of information for statistics about access and downloads of data
	At an individual repository level aggregations can offer a benchmark comparative purposes to support repository management
Entrepreneurs	Aggregations provide a single point of access to multiple sources of research and other materials to aid discovery
	Aggregations also provide suitable collections of materials for possible commercial exploitation through the building of value-added services on top

Table 4. Benefits of the aggregation model for end-users

4.5.1 Discovery-2-Delivery (D2D) paradigm

Where do repositories and end-user services sit in the D2D chain? Many end-user services, including large OAI aggregators such as OAIster and web crawlers like Google, are primarily about providing discovery. Providing accurate and appropriate discovery for different user groups is difficult. Services such as OAIster provide a generic view onto the harvested materials, whilst Google and Yahoo! do the same across the web and linked content they crawl: OAIster aggregated contents are also available through Yahoo! as an alternative. If starting out on a discovery path without having a clear starting point this generic level of discovery is of value, providing rapid feedback and results that can help guide further interaction. The level of usage of web search engines in particular highlight the perceived value of this approach. Even when the end-user knows what he/she is looking for web search engines offer a welcome and usually fast discovery service that is favoured over more considered and structured discovery services. This has been highlighted by recent experience at the University of Southampton where only 11% of accesses came through their local repository-based structured search services, the rest coming through alternative routes (including 64% from web search engines)[56].

The popularity of web search engines suggests that there is less need to focus attention on developing additional discovery services than on other areas. This applies particularly to generic discovery services of global aggregations, which will have difficulty competing with more popular and established services. There is scope for discovery services that are targeted at particular user communities, for example those offering particular subject access or focussing on a key content type such as ETDs. Not

[56] 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

all open access repositories are yet accessible through web search engines, and these will need some type of discovery service onto them. There is also the need to consider the benefits of searching open access repositories in the context of discovery for more traditional information sources such as catalogues and bibliographic databases. But in both cases there needs to be reflection on what relationship there will be between potential discovery services and web search engines to ensure a clear case for development.

Aggregations and associated end-user services can potentially play a greater role in other parts of the D2D chain. Once content has been discovered they need to be located. The use of unique, persistent identifiers can allow clear location of the content, and this is particularly relevant where compound objects are concerned and sub-components of these objects are located in a distributed fashion. Locating objects through the use of OpenURL and the use of COinS[57] can add value to aggregations and lead the end-user from the discovered results to where the content is. There are Firefox extensions allowing the use of OpenURL within Google Scholar already, and COinS have been incorporated into secondary discovery sources such as zetoc.

Once located, objects can be requested and delivered through downloading the relevant object(s), and in an open access environment it is hoped that this direct and immediate route would be the case most of the time. Where restrictions are in place, however, steps are required to ensure the end-user doesn’t hit a dead-end and can still enact the request step. The recent addition of “request copy” buttons to EPrints and DSpace can help to bring about the repository equivalent of an inter-library loan, for example. This less immediate request step may be particularly necessary where the desired content is not easy to download or is not available for download, e.g., a dataset that is only available on request or images being requested in a particular format.

Options for delivery also need to be considered. Delivery of a PDF can be a straightforward affair where browsers are configured to deal with this file format. But other formats may require alternative delivery mechanisms and services. Delivery in alternative formats can be influenced by the intended tools that will be using the materials beyond the D2D chain, for example outputting metadata for import into bibliographic management software or using the information within analysis software. The use of compound objects that encompass content, such as those built using MPEG-21 DIDL or ATOM, provides alternative delivery formats that end-user services need to know how to manage.

[57] OpenURL COinS: a convention to embed bibliographic metadata in HTML, http://ocoins.info/

The development of repositories, aggregations, and end-user services across these in the UK HE/FE community, along with relevant intermediary services, encompasses all the constituent parts of the Information Environment. It is recommended that the development of end-user services take advantage of this end-to-end scenario to test out and give deeper consideration to all parts of the D2D chain, and especially the later stages of this: locate, request and deliver. In particular, it is recommended that the potential role of the OpenURL 1.0 standard be examined to support these extended D2D activities.

4.5.2 To use or not to use?

In a presentation at an Institutional Repositories event at the University of Southampton in January 2005 Richard Boulderstone from the British Library described the information chain and highlighted where the British Library had a role to play (coloured red in Figure 19).

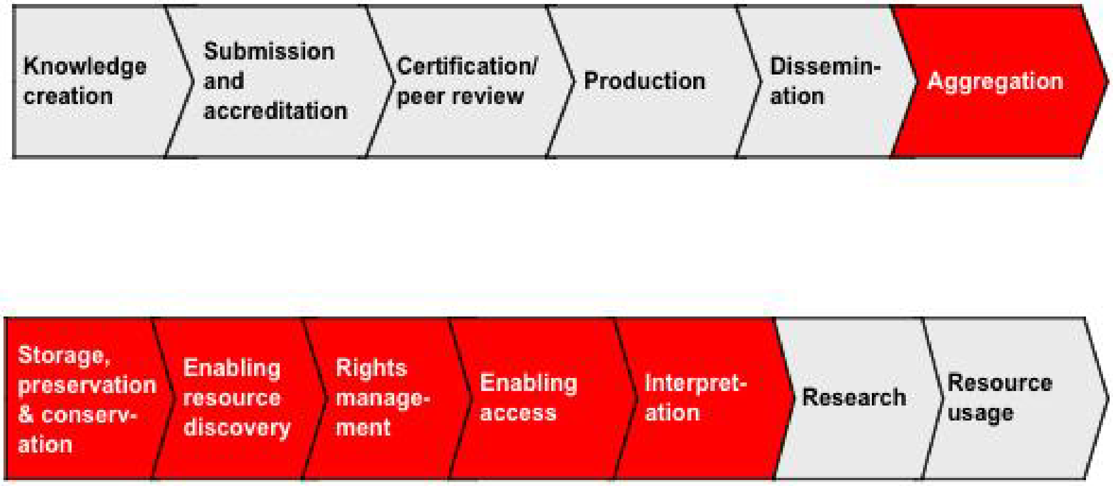


Figure 19: The Information Chain (Richard Boulderstone, British Library)

It is notable that the BL's role starts with aggregation, but stops before use of the discovered information takes place. For a large organisation this is perhaps understandable, as it would be almost impossible to gauge the many ways information found at the BL is subsequently used. It is not an uncommon view. The CORDRA initiative in the US has adopted a similar position. The aim of CORDRA is to provide through its system of linked registries as rich a metadata set as it can for end-user services to serve as they see fit. But no assumptions of use are made, as, again, it is considered that gauging potential use is probably not possible in the many areas that the CORDRA model can be applied. Similarly, the Information Environment Architecture has a presentation layer that focuses on just that, presentation, and has not so far extended into investigating use.

At these abstract levels it is reasonable that consideration of use is a difficult area to address. It is easier in real and smaller scale situations to investigate use and make appropriate adaptations. And yet in the broader scope of information services the reason for making information available is so it can be *used* in some way. Establishing a greater understanding of how information and resources are used can inform both the types of end-user services it is of benefit to develop and also the structure of the underlying repositories themselves so they are better placed to serve the uses required. One way of dealing with potential use is to increase flexibility. CORDRA promotes a high level of flexibility in how it makes metadata available: the aDORe architecture does likewise. Both seek to adapt on the fly to user demands for information. This flexibility is vital, as it offers the most opportunities to react quickly to moving user requirements. But flexibility in tandem with establishing a greater understanding of how information will be used offers a potentially even more powerful tool. There can be a conflict between these, insofar that the available effort may need to focus on one or the other, but they can be valuable together. The information chain depicts information workflow. To fully fit into user workflows an appreciation of what happens at the end of the workflow will help that workflow to proceed more smoothly.

There is a marked contrast between current treatment of open access research content and learning content in this regard. In the learning and teaching environment re-use of materials once discovered is almost assumed: the purpose of looking for and discovering learning materials is in order to use what is found. This sense of use influences discovery systems within learning and teaching: the JORUM national learning object repository has been established with the capability of testing re-use as an added value service on top of discovery, for example. Although the detail of use and re-use may not always be known, an appreciation that use will take place has influenced the development of end-user services and add value to them.

It is recommended that development of end-user services include an element of investigation of how information to be surfaced through these services will be used, to help inform the development of the service and feed back to the underlying repositories being exposed through the service.

4.6 Sustainability

In taking forward the development of end-user services across repositories there are a number of factors that can be borne in mind to maximise the chances of these services being technically sustainable in the long-term. The practicalities of implementing components of the repository/aggregator/end-user chain have been dealt already. This section looks ahead beyond these to see what influences might affect end-user services in the long run.

It is fair to say that in many ways 'long-term' and 'technical' don't go together, as technology change happens rapidly and there is every reason to suspect that new and innovative technologies will emerge over the coming years. Technical sustainability is also linked to cost-effectiveness and the influence of non-technical factors: the recent DLF Aquifer Study on Institutional User Services [Halbert, 2005] found that simple lack of time was a major factor in the lack of new end-user service development.

Nevertheless, this model has attempted to consider approaches that can be applied using different underlying technologies as required. It has adopted a number of the aspects of the Information Environment Technical Architecture, which has been in existence for five years and has stood the test of time well. It is also aligned with the emerging e-framework, following a service-oriented approach and using open standards where relevant. Notwithstanding this, in the context of repositories it is important to remember that web search engines are playing, and are likely to continue to play, a major role in supporting discovery and will continue to act as end-user services onto web crawler aggregations from repositories.

The Information Environment Technical Architecture has stood the test of time because the building blocks with which it was originally conceived are still in place. Many of the open standards promoted, including OAI-PMH, have matured since the Architecture first appeared and the building blocks are still very valid for use in developing end-user services. This study has identified no gaps in the standards required to enable this, but rather has identified gaps where the standards are not being interpreted or implemented correctly or as extensively as they might be. More focused and practical implementation of the building blocks will allow high value end-user services to emerge. Continued adherence to open standards interfaces will support sustainability if underlying repository systems themselves change over time.

Having said that it is important not to rest on our laurels. Standards *have* changed over time to meet changing needs, and there is a need to maintain a watch on where standards face limits to their capabilities that require attention. This is particularly the case in the field of metadata standards, where existing standards still appear to rely too heavily on bibliographic and physical item origins: important though these are there is a pressing need to identify ways of encompassing these alongside metadata standards for born-digital objects, simple and compound, that will allow digital content to be fully utilised. There is also the issue of cross-domain interoperability, where multiple metadata standards co-exist. How these interact will be vital to how cross-domain digital content can be effectively used. A key way forward in addressing these metadata demands is to model the content and metadata we are trying to describe more rigorously so that we can generate metadata standards that meet the needs of the content being created and the end-users wishing to access this.

The OAI-PMH itself has identified limitations. It is bound to HTTP as its transfer protocol, which potentially limits its use in the future. It can only work with XML files, which gives much flexibility, but may not future-proof it in a possible world of Semantic Web and RDF. RSS and ATOM are likely to continue to evolve and it will be valuable for the academic community to feed into this development to ensure these standards can best serve its needs.

In the layered and service-oriented architecture proposed within the aggregation model, with components and functions separated out in distinct layers there is a need to offer guarantees that access to and between the different components, and the network itself, will be reliable. Service level agreements may be required, and certainly a level of non-technical communication that currently doesn't always exist. This is particularly the case where the aggregator and the repository are distinct from each other.

Lastly, sustainability will be affected by the end-users. Will access to repositories meet their expectations? How will repository end-user services fit into their information workflow? Only by meeting these end-user expectations and needs as one service within their wider information landscape will end-user services across repositories truly earn the right to sustainability.

4.7 Looking to the future

In proposing any technical model there is always a need to consider how it may fair over time. As one interviewee put it the terms technical and long-term don't sit together very well. Nevertheless it is at the implementation level that this most applies. The aggregation model, whilst addressing the possibilities that a number of different standards and technologies can provide, is, like to the CORDRA model, not intended to be tied to any specific technology.

Two perspectives can be taken in considering the future development of end-user services across repositories using an aggregation model: the level of take-up amongst other communities and initiatives to assess the breadth of interest; and the potential for the model to meet broader high level views of the technology landscape.

1. Take-up
A number of initiatives making use of aggregations have already been mentioned in this report and the associated appendices. The use of OAI-PMH and its model of data and service providers have driven many of these initiatives in the open access arena. This model has proved successful, and will continue to be with the caveats and recommendations made in this report.

It is not solely in the use of OAI-PMH, however, that aggregation is regarded as a valuable methodology, but in the use cases that have been identified for aggregation. The Research Information Network is investigating the use of data webs, digital information and storage following a lightweight harvesting of metadata about datasets into a central registry (an aggregator), to facilitate access to and awareness of these datasets[58]. They are especially interested in the use of lightweight Semantic Web and Web 2.0 approaches to enable this. The JISC-TIME project establishing an e-books metadata and interoperability testbed developed an architecture that incorporated the

[58] Data webs: new visions for research data on the Web – a Research Information Network workshop, http://www.rin.ac.uk/?q=data-webs

creation of a central aggregation of e-book metadata to support the generation of standard and easily available e-book catalogue records for use in library catalogues[59]. The OpenCourseWare project in the US is developing a model for openly sharing learning materials[60]. The project promotes a model of aggregation for the materials to, currently, make these available through specific websites: they are investigating the aggregation of metadata from these websites to ease discovery across different OCW websites beyond the current web searching that is available.

2. Complementing the technology landscape

All of the examples mentioned have their own detailed technical perspectives on how to implement the aggregations they require for their purposes. One of the most valuable aspects of the aggregating activities indicated in this report is the added value they provide in moving work from individual institutions and organisations up to the network level. They remove the need to manage services at the individual repository level that can be better provided as part of a collaborative aggregation. The aggregations themselves can then provide services that no individual repository would be capable of by themselves. Those Research Councils with such facilities recognised this when setting up their data centres: AHDS, ESDS, etc. The technical model proposed in this report recommends that this successful approach be extended where possible.

In doing so it is necessary to be flexible to ensure that the aggregations generated do not become millstones but are able to adapt to circumstances. Moving towards a service-oriented approach, as recommended by the e-Framework initiative, allows each of the components involved in supporting aggregations – the repositories, the aggregators and the end-user services – to be flexibly interchanged as required. This long-term goal is nevertheless worth pursuing to ensure repository content is utilised to the extent it can be.

SOA centres around communication between different components through machine-to-machine interfaces. One of the strongest responses that came out of the interviews for this study was the need for greater communication between different components, though at the human rather than machine level. Improving and standardising how we humanly describe the interactions we would like to establish between different components will help to define the potential machine interfaces that will allow an SOA environment to communicate for us. The e-Framework initiative to establish common ways for how we communicate, through reference models and related activities, is a valuable step along this road. Improved communication between repositories and aggregators/end-user services will facilitate this in the repository and open access arenas.

Many end-user services today have a personal element to them: they seek to address personal needs. All of us maintain, in more or less organised fashion, a personal collection of digital materials on our computers. We all see the network and the

[59] E-Books Metadata and Interoperability Testbed, http://www.jisc.ac.uk/index.cfm?name=ebooks_metadata

[60] OpenCourseWare OpenContent, http://opencontent.org/

information available through it from our personal viewpoint. The personal aspect to future services will be high, and the provision of services across repositories will be no exception. A future challenge will be personal aggregating and how individuals can exploit these aggregations. How can individuals best interact with the different aggregations available to them for personal information management?

Some have advocated the use of Semantic Web technologies to facilitate this, and the use of RDF to describe the information available. Much remains to be understood about the potential of the Semantic Web, though initiatives such as the investigation of data webs will hopefully open up development paths. RDF may also provide the freedom of structure that metadata generation may require: will individuals feel better able to provide useful metadata about resources they are creating if they can provide this in their own way for later mark-up using RDF rather than a set metadata form? Social tagging suggests it is a route that can be popular.

The use of RDF is not simple, however, and an ongoing source of debate will be the balance between lightweight and more complex solutions for achieving interoperability. Lightweight solutions, using OAI-PMH and RSS for example, can draw people in to using interoperable systems, whilst complex solutions such as the DR OSID require greater initial investment for increased potential gain. The former will encourage take-up and needs to be used as a lead in to more detailed and value-added interactions.

5. BUSINESS ISSUES

Because the JISC's funding is top-sliced from the total budget available to the higher education sector, there is naturally some debate about the extent to which the JISC should fund services and for how long. We perceive a clear difference between investing in projects that might develop into services, or services for which there is a real need but which cannot operate on any business footing other than by being sponsored by public money, and funding the development of services in general. The former are candidates for continued funding from the JISC, whereas there is considerable scope for other services that are distributed over the network to be developed on a self-sustaining basis.

At this stage we are looking still at a UK repository landscape that is not fully populated with repositories. Undoubtedly this will change, and probably quite rapidly, judging by the number of new institutional repositories that have sprung up in the UK and around the world in the last year. There are now around 650 institutional repositories globally, whereas a year ago there were approximately 300 – in other words, at current growth rates each day sees a new repository somewhere in the world. One of the spurs to growth in the UK has been the Research Assessment Exercise: repositories provide institutions with the means to undertake this exercise in a considerably more efficient way than previously. Now that it has been announced by the Chancellor of the Exchequer that in future research assessment for the UK will be based on metrics, we see the usefulness of repositories – and the Open Access corpus that they can provide – being proved even more convincingly. New metrics can be developed that will be far more meaningful for research assessment than any we currently have, but these can only be produced if the research literature (and data) is Open Access. Research assessment will be one of the drivers for Open Access and repositories in the next period.

The outcome in the short to medium term will be that every research-based institution will house its own repository and that FE institutions, with their primary remit of teaching, will also see the strategic advantages of having such an entity as a tool to enhance the teaching and learning functions. Given that scenario, the development of repository services is a natural outcome.

5.1 Developing services for the repository network: costs

As the repository network matures, services will develop that both provide leverage for the investment that has been made and offer institutions and end users a growing range of options suitable for their particular needs. These services may have their roots in publicly-funded projects or be newly developed offerings from the commercial sector. The issue at stake here is what sort of business models these services might most successfully adopt since the viability and sustainability of the repository services scoped in this study will depend upon the proper business and management models being in place. Some of the services already in existence have successfully moved to a fully commercial model, while others are operating on a long-term JISC-sponsored basis. A few are in transition and no doubt the JISC hopes that more will follow along that path. There *is* scope for this in the cases where a clear commercial business case can be seen but in others the long-term model may remain a community or publicly-funded one.

The annual costs of repository services will vary hugely from service to service. It is difficult to arrive at reliable cost figures at this stage though a few examples may serve to give an in-principle idea of the running costs for a service. As some of these are culled from services in development and at project status these need to be viewed with some caution.

Early figures for the cost of establishing an institutional repository were gathered for a previous study (Swan et al, 2004). In that, we reported the actual establishing and annual running costs of four institutional repositories.

Institution	Set up costs	Running Costs
MIT (DSpace)	$1.8m grant: DSpace software developed on-site	Staff $225,000
	3 FTE staff	Operating Costs $25,000
	$400,000 system equipment	Systems equipment $35,000
	Total = $2.4-2.5m	**Annual running costs $285,000**
National University Of Ireland, Maynooth	Software free (EPrints) Grant to hire Computer Science student for set up and customisation 6 months	1 FTE staff member for upkeep and maintenance
	Grant for €5,000 for server	
	Total €20,000	**Total €30,000**
Queens Qspace CARL	Software free (DSpace)	
	Server space at Institution	Library staff: $25,000
	Programmer for 12 months: $50,000	ITS Staff: $25,000
	Staff costs for advocacy work with faculty	
	Hardware: $2,065	
	Total Can$52,065	**Total Can$50,000**
SHERPA: Nottingham	Software: Free (EPrints)	Maintenance absorbed within HEI costs: 5 FTE days per annum
	Standard Server: £1,500	Coordination and collection of material £30,000
	Installation 2-5 FTE days £600	3 year update of hardware and software: 2-5 FTE days and £3,900
	Initial customisation 15 FTE days £1,800	
	Total £3,900	**Total £33,900**

Table 7: the costs involved in setting up and running institutional repositories: actual examples from four repositories in Europe and North America

(From Swan et al, 2004)

The table below shows the costs of depositing articles when done by an intermediary. The example is from the University of Nottingham repository. All costs are in GBP:

Initial set-up costs £		**Technical support / maintenance £**		**Annual operating costs £**		***Article input costs £***	
Software	0	HEI standard Web service maintenance: three year upgrade		Staff salary	30000	Hours per week	*17.7*
Server	1500	Hardware	3000			Articles per hour	*4*
Installation	600	Labour	600				
Customisation	1800						
	3900		***3600***				***4.46***

Table 8: Input costs for the University of Nottingham eprints repository

(From Swan et al, 2004)

It should be noted that not all repositories have mediated deposition: some operate on an author-deposition basis. Carr reported that, based on analysis of Southampton University repository logs, an average researcher would spend 40 minutes per year depositing articles in an institutional repository (Carr & Harnad, 2005).

The House of Commons Select Committee on Science & Technology's own study (HCSTC, 2004) on the cost of repositories concluded that it would cost an institution GBP 3,900 to set up a repository and annual operating costs of GBP 31,300. The study estimated the cost of establishing appropriate repositories nationally to be just over half a million pounds and the annual running costs to be just in excess of GBP 4 million. Operating costs in this study included mediated deposit but did not include technical support costs, which were assumed would be absorbed by institutions, and did not account for any specific preservation costs.

Mornati[61] reported the costs charged by her organisation (CILEA) for setting up and running repositories for Italian educational institutions: CILEA charges € 7,200 for the set-up plus the first year's running costs and € 2,400 per annum subsequently. Kemp[62] reported that a range of figures collected from institutions running repositories showed that set up costs ranged from USD 6,887 to over USD 1 million. Rankin (2005) calculated that a repository might take up to 3 FTEs during year 1 to set up and run the operation, and perhaps 1 FTE to operate it thereafter.

The Cream of Science initiative in the Netherlands had a budget of €100,000 initially, but this was doubled as the project developed and costs were more clearly defined. The final average cost per article for this initiative has now been identified at €50, which includes all the work involved in deposition including digitisation (Feijen & van der Kuil, 2005). It should be noted that this does not include standardisation of metadata or detailed cataloguing.

The latter activity can be very expensive. The average US university probably expects to spend around USD 50-75 on the complete cataloguing of a book, and much the same on a serial[63]. Formal, long-term studies on cataloguing costs for serials titles have been carried out by Dilys Morris and colleagues at Iowa State University and have shown that the basic cost of creating a record is around USD 15 including overhead but that this can rise to several times that amount if substantial authority work is involved (Morris *et al*, 2000). For individual article metadata we might expect the cost to be a little lower, though tight specifications for metadata quality would keep costs on the high side, especially if controlled vocabulary requirements are included. This requires additional expertise over that needed for simple descriptive cataloguing.

[61] http://www.library.yale.edu/~llicense/ListArchives/0509/msg00156.html
[62] http://www.library.yale.edu/~llicense/ListArchives/0511/msg00030.html
[63] Eric Childress, OCLC; personal communication

Further figures have been provided by John Houghton at Victoria University in Melbourne, Australia. John's words in his preamble reflect the state of affairs with respect to collecting data on costs in the scholarly communication arena:

> *"Notwithstanding our costing everything we could think of, attempting to quantify the potential benefits of OA, and comparing those potential benefits with the costs of a national system of IRs in higher education... the bit we are probably weakest on is the cost of institutional repositories. Our extensive literature review and local consultation demonstrated that every case is different... and they cost anything from very little to lots... depending on the level of functionality, etc. In the report we developed a cost model for all scholarly communications activities, along the King & Tenopir lines, based on an extensive literature review, and then refined it for Australia based on local consultations. From the lit[erature] we got a range of annual IR costs anywhere from AUD 4,000 to AUD 80,000... allowing for a 5 year depreciation of hardware & software got us a mean of about AUD 42,500 pa. However, local discussion suggested full costing (salary, oncosts and overheads) for all the related policy and integration activities of up to AUD 240,000 pa. All this gave us a range of costs for a national system of IRs in Australian higher education of anything from AUD 2m to AUD 10m a year... depending on functionality, the level of institutional buy-in, integration with research management and reporting, etc. etc."*

Some costs are also available for metadata creation and for the establishment of relatively simple resource discovery services. The cost of creating the metadata for one (fairly complex) object at the RDN is reported as being GBP 12.50[64]. Care needs to be taken when considering metadata creation costs for an organisation like the RDN, though, since the process of creating metadata for an item may not be a one-off thing: updating may be required as sites change, incurring additional ongoing maintenance costs.

The two year budget for setting up a very capable and well-designed resource discovery service for engineering, mathematics and computer science at Heriott-Watt University[65] has been GBP 66,000. The ARROW Discovery Service, which runs across the ARROW repositories in Australia, has cost around GBP 39,500 in project management time over 18 months; over the same period, software development work has cost GBP 24,000[66].

PerX, a cross repository discovery service for engineering[67], in pilot phase, has a two-year budget of GBP 102,000.

The costs for other types of repository service can only be estimated. Most service types listed in this report would be built upon existing projects that are proof-of-concept exercises or pilots and so costs for these thus far are related to development work and not to mature service running costs. We can only use informed estimates – based upon the expected labour and fixed asset costs – in these cases, and the most useful way to deal with this issue is to classify putative repository service costs as a range of cost levels (see next section).

[64] personal communication

[65] http://www.techxtra.ac.uk/: personal communication from Roddy MacLeod

[66] http://www.arrow.edu.au/: personal communication from Debbie Campbell

[67] http://www.engineering.ac.uk/

5.2 Business models for repository services

There has been some work on business models for entities operating in the e-information/e-commerce arena and it is useful to review briefly the most prominent examples.

Rappa (2001) described a typology of nine models, the most pertinent of which here are:

- the **merchant model**, where services are sold on the traditional retail model. In repository service terms, this means selling a service to repositories or to users that has a cash value and where a straightforward exchange of cash takes place. There is an example of this working already in the Netherlands using the DAREnet network. A small chemistry publishing company is locating doctoral theses protocols from Dutch university repositories (most of these are metadata-only deposits) describing the synthesis of compounds. The universities provide the hard-copy theses that the company wants, the company digitises them and provides the universities with the digital files in exchange for the use of the thesis, thus acting as a digitization service. The company then sells the content of the theses to the chemical industry. In this particular example, money changes hands between the company and its customers in the chemical industry, though the universities get some digitisation carried out for no charge, so there is a quid pro quo. It is possible to envisage similar arrangements where universities may make cash sales. Indeed, MIT is already selling direct access to its digital theses.[68]
- the **subscription model**, which exploits opportunities to sell access to a range of ***software*** and ***content*** services. In repository service terms, subscription models are likely to be targeted at national or regional consortia, individual institutions or departments, in which case the services will be free at the point of use (governed by suitable authentication processes). An example of this type of service is Thomson Scientific's *Web Citation Index*, a citation index analogous to Thomson's Science Citation Index but which indexes institutional repositories rather than journals. Insitutions pay an annual subscription for access to this service.
- There is a possibility that some service providers may attempt to deploy a **utility model**, whereby users may access services on a pay-as-you-go basis, often involving micropayments. It should be noted, however, that this approach has not been especially successful in the past in the higher education marketplace.
- the **infomediary model**, a process whereby data are collected, manipulated in a way that adds value, and sold. Many examples of this type of operation already exist in the search and navigation market and it is likely that existing commercial operators will be able to adapt their operational and business processes not only to accommodate but take commercial advantage of a UK-wide repository network. Selling author lists, data that aid research assessment, or citation data, are all examples of this model.

[68] http://libraries.mit.edu/docs/pricing.html

- the **advertising model.** The extent to which an advertising model can be viable will depend on how popular the UK-wide repository network becomes. The greater the number of people who use it, the more attractive it becomes to service providers that primarily rely upon advertising revenue. Since the network will be part of the World Wide Web, existing commercial services which are based on the advertising model – Google for instance – will reach into the repository network. At least one UK subject hub – EEVL – has succeeded in attracting corporate advertising to its site, though this has not reduced the reliance on public funding via JISC. Nonetheless, it shows that there is promise in this route. Commercial companies have gone this way before: The Nature Publishing Group (NPG), for example, has established the Signaling Gateway, a website for the cellular signaling community, that has sponsorship and advertising to buffer the funding from NPG and the AfCS (Alliance for Cellular Signaling, a consortium of eight US signaling laboratories). More recently, ScienceCommons set up the NeuroCommons[69] site (with sponsorship from Teranode, a laboratory automation company) an experimental service that is using semantic web technologies to provide a knowledge web for the neuroscience community
- the **community model**, in which members of the community of interest invest their own resources, contributions and sometimes cash. An example of this in action is that of open-source repository softwares such as EPrints and DSpace. These both have their own communities contributing ideas, code – and sometimes cash – to ensure the continuing development of the software in the whole community's interest.

Timmers (1998) had already produced a different typology with eleven business models included, upon which Rappa's work builds. There is no need to visit these in detail: the useful point is that Rappa and Timmers agree, despite using slightly different terminology, on the basic types of business model that can be employed in the e-information arena. For our purpose here, we have developed a simplified list of business models that suffice to describe those that might be employed by repository services.

The DAREnet programme established by the SURF organisation in the Netherlands is very clear about the overall business model of the network. It is that the data layer entities represent public content and the infrastructure required is the province of the institutions (in the case of DAREnet, entirely Dutch universities). We concur with this and believe this should work also in the UK for two main reasons. First, there are clear organisational advantages to universities and research institutes in having an institutional repository: the overall business case is relatively simple to make (though may be more complex to carry through 'on the ground' in some institutions) and a repository (or repositories) as the means to collect and take care of the digital assets of HE and FE institutions is a concept that will fairly rapidly become embedded in institutions' ways of operating – a part of everyday life for research and teaching institutions. Second, the costs of establishing a repository can be highly variable, as we have reported above, depending on what the institution requires or expects from its

[69] http://sciencecommons.org/data/neurocommons

repository. The level of elaboration is thus a decision to be made at the institutional level and the appropriate extent of capital and recurrent expenditure is something that is particular to each institution. This is not something that can be standardised and then perhaps paid for from top-sliced funds, as the JANET network has been, for example. So, the data layer provision must be left as an institutional responsibility.

At the services levels there is a different story. Although we expect that some services will need to remain in the publicly-funded domain, paid for by top-sliced money, there are others that can be expected to find other business models under which to operate. Some will be able to assume a purely commercial model, with revenue coming from the market in the form of cash payments, subscriptions, advertising or a combination of these. Others may be able to adopt a community model that is sustained by community involvement, contribution and collaboration. In the case of some services, they may start life in one guise and move to another as their offerings mature: an example of this might be projects that are initially developed into community-model services but which subsequently find a sustainable niche operating on a commercial basis in the marketplace.

Below, we present a scheme showing what we believe are the most appropriate business models for the services discussed in this report, and some additional characteristics, such as the scalability of each, the associated risks and the possibilities for shifting between models over time. The scheme follows the same themes as before in this document; that is, services are categorised as ingest-layer services, pre-aggregator services and output-layer services.

We base this scheme on a business model typology that we believe fits the needs of the JISC in assessing what might be done in the repositories arena. The business models considered are:

- **Institutions own and run the service** to further their own goals and strategies. Services that fall into this category are those that have a perceived advantage to institutions and can be embedded in institutional workflow
- **JISC (and perhaps additional partners) supports the service into the foreseeable future**. Services that fall into this category are those that do not have the basis for revenue-generation and are not appropriate for the institutionally-embedded or community-sustained models
- The service runs on a **community-model basis.** Services in this category are sustained by the communities in which they operate. Cash is not of major importance here: community-based, collaborative effort to sustain the service is the operational basis
- The service runs on a **subscription-model basis**. Services in this category are those that can sell a product or service on an ongoing, subscription basis to paying customers in the marketplace
- The service runs on another **commercial basis**. Services in this category have a clear revenue-generating business model that can operate viably in the marketplace

and are properly in the domain of commercial operators. Commercial models in out typology encompass **merchant, utility** and **advertising** models.

The columns in the scheme below are as follows:

1. ***Cost level***: annotated as high, medium or low. Services with low running costs are those that typically require low staffing levels and low levels of investment in fixed assets: general guideline: up to £100K per annum. Those that might cost up to £250K per annum are termed 'medium' in the table. Those with higher running costs are termed 'high'. Note that these are estimates only in most cases, as described in Section 5.1).
2. ***Appropriate business model(s)***: these are indicated for each service type. Some services have more than one model checked because there are multiple ways of making a business work in those cases. In some of them there is a check mark in the JISC-funded column as well as in other columns. This is largely because we see potential for such services developing from JISC-funded projects; in some cases it is because JISC-funded services might run alongside other services operating with a different business model, specifically commercial or subscription models. This is so in the case of resource discovery, which covers such a broad scope: some commercial (or subscription) services could satisfy the needs of certain market segments but there will remain other segments that will not sustain paid-for services and will require free-to-use resource discovery tools. In the commercial model column we have indicated which types of commercial model might apply
3. ***Scalability***: We have used a simple scale of 1 to 5 for this column. A score of 1 indicates that a service scales up easily if required, simply by incremental adding of the resources required. Middling scores indicate that there would be some careful strategic business planning needed to scale up from a simple-level service to one satisfying more complex needs in the user base. A score of 5 indicates that scaling up a service using the current ways of doing things would prove challenging.
4. ***Associated risks***: Although scalability impacts on this factor, business risks arise from other sources too, such as change in the operating environment, in technologies and in the customer base and its requirements. Most of the ingest-level services have low risks, as do those at output-level that would be selling proven technologies that look to have good utility in the marketplace. The services with medium risk are those that do face scalability challenges but that also face the challenge of continually matching their offerings to a changing user needs base. In the repositories arena it is not easy to see far ahead: technologies and their applications are moving extremely fast and can be expected to continue to do so. Moreover, the amount of digital information will only increase, perhaps hugely, and some of these repository services will need to cope with volume changes as well as new challenges in other ways.

Service	Cost level	Appropriate business model					Scalability 1 = easy 5 = difficult	Associated risks	Comments
		Institutional	JISC-funded	Community	Subscription	Commercial			
INGEST SERVICES LAYER									
Digitisation	M	✓				Merchant	1	Low	Institutions do their own digitisation,or pay a third party operating on a commercial basis
Rights/IPR advice	L		✓				1: but probably not required to scale substantially	Low	Core service for HE/FE sectors
Open Access advocacy advice	L		✓				1: but probably not required to scale substantially	Low	Core service for HE/FE sectors
Technical advice	L		✓			Merchant	1: but probably not required to scale substantially	Low	Core service for HE/FE sectors. Some commercial operators may offer some as part of commercial repository-building service
Repository construction	M	✓				Merchant	1	Low	Institutions do their own construction,or pay a third party operating on a commercial basis
Hosting services	M					Merchant	1	Low	Institutions pay commercial operator

DATA LAYER PROVISION:									
Institutional repositories	L/M/H	✓					N/A	Low	Costs can vary hugely depending on institution's aim and objectives for repository
National-level 'catch-all' repositories	L		✓				1	Medium	Core service for HE/FE sectors
Subject-specific repositories	L	✓		✓			2	Medium	May be set up by institutions or communities
Media/object-specific repositories	L/M	✓		✓			2	Medium	May be set up by institutions or communities
PRE-AGGREGATOR LAYER SERVICES									
Metadata-creation and enhancement	M/H		✓			Merchant Advertising	By machine: 2 By humans: 5	Medium	Existing and future JISC-funded projects may require long-term support Commercial companies will also operate in this niche
POST-AGGREGATOR LAYER									
Technology transfer	L		✓			Merchant	1: but probably not required to scale substantially	Low	Core service for HE/FE sectors: advice on how to translate projects into viable services

OUTPUT SERVICES LAYER									
Access and authentication	H				✓	Informediary	3	Medium	
Usage statistics	M			✓		Merchant	2	Low	
Preservation	H	✓			✓	Merchant	5	High	Challenges will increase. Various models will operate for different user environments
Research monitoring	L			✓		Merchant	2	Low	
Resource discovery	M/H		✓		✓	Utility Advertising	4	Medium	Challenges will increase. Various models will operate for different user environments
Overlay journals	L	✓				Merchant Advertising	1	Low	Institutions can operate here (e.g. Lund Virtual Medical Journal. Otherwise, lots of scope for commercial operators
Publishing	M			✓		Merchant Advertising	1	Low-medium	Publishing services (e.g. peer review) may be provided on a commercial (publishers) or community (learned societies) basis. Value-added products may be produced on both bases too
Meta-analysis	L			✓		Merchant	2	Low	Development costs can be high but ongoing service costs should be low
Bridging services	M		✓			Subscription Advertising	3	Medium	Core services for HE/FE sectors (e.g. ROAR, UKCORR, OpenDOAR) Commercial companies may innovate in this niche

Table 9: Business models for repository services working across UK repositories

RECOMMENDATIONS

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

The following recommendations are made to the JISC:

11. The research community should be engaged at the highest level to encourage the establishment of repositories in all HE and FE institutions and the development of policies that will ensure the collection of content.
12. Channels of communication with repository managers should be opened, and the establishment of a community encouraged. This may be done through existing structures: the UKCORR is the most appropriate, and the two main open source repository softwares (EPrints and DSpace) have their own user communities that could also be used for this purpose. The aim is to have clear and effective communication structures in place between JISC and all operating repositories that will facilitate two-way discussion and enable development.
13. Similarly, an interface or contact point between the JISC and actual or potential service providers should be established. This will enable end-user oriented services to be developed in a coordinated and directed way.
14. Developments of repositories, aggregators, end-user services, and intermediary services should move towards a service-oriented architecture and establish separate layers for the aggregation model to maximise the flexibility available for building end-user services to meet user requirements.
15. Development of end-user services includes an element of investigation of how information to be surfaced through these services will be used. This will assist in helping inform the development of the service and feed back to the underlying repositories being exposed through the service.
16. Additional means to generate metadata using automatic means are required. It is recommended that investigations into relevant techniques and tools be taken forward with some urgency.
17. Further attention to identifiers, specifically location-independent identifiers, and necessary resolution systems is recommended to provide greater understanding of their benefits and use.
18. It is recommended that the use of RSS and ATOM be investigated as additional standards to OAI-PMH for use in aggregating metadata and content. They offer the potential of targeted exposure of repository resources that may be beneficial in the development of end-user services targeted at specific communities. It is

also recommended that the exposure of repository contents within web search engines be examined in closer detail to assess the paths of exposure that exist and the implications for repositories of exposure via this route.

19. It is recommended that future work to develop aggregators and/or end-user services include an element of communication and involvement with repositories from the start. This will ensure development does not take place in isolation and increase the interoperability between the three major components of the aggregation model. Where intermediary shared infrastructure is involved those developing this should also be included in relevant communications.

20. It is inevitable that for an optimally-structured set of repository services to be developed on UK repositories, there will be a continuing need for top-sliced funding for some parts of the system. The JISC will need to plan for this for the medium-to-long term.

REFERENCES and BIBLIOGRAPHY

Andrew T & McColl J (2002) Theses Alive! Final report.
www.thesesalive.ac.uk/archive/ThesesAliveFinalReport.pdf

Asensio M (2003) JISC User requirement study for a moving pictures and sound portal.
http://www.jisc.ac.uk/uploaded_documents/MPSportaluserreqs.doc

Awre C, Dovey MJ, Hunter J, Kilbride W & Dolphin I (2004) Developing Portal Services and Evaluating How Users Want to Use Them: The CREE Project. *Ariadne,* **Issue 41**.
http://www.ariadne.ac.uk/issue41/awre-cree/

Awre C, Waller S, Allen J, Dovey M, Hunter J & Dolphin I (2005) Putting the library into the institution: using JSR 168 and WSRP to enable search within portal frameworks. *Ariadne,* **Issue 45**.
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**).
http://www.dlib.org/dlib/november03/bekaert/11bekaert.html

Blanchi C (2006) DVIA Registry Architecture. Presentation given at the 2006 Defense Technology Information Center Conference.
http://www.dtic.mil/dtic/annualconf/Wednesday/DVIACombined.ppt

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

Brophy P (2006) Projects into services: the UK experience. *Ariadne*, **Issue 46**, p 1-6.
http://www.ariadne.ac.uk/issue46/brophy/

Brosnan K (2005) Final report: Learning to Learn Project.
http://www.daice.stir.ac.uk/l2l/reports/l2l_finalreport.pdf

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

Carr, L (2205) Use of navigational tools in a repository.
http://www.ecs.soton.ac.uk/~harnad/Hypermail/Amsci/5170.html

Carr L. and Harnad S. (2005) Keystroke Economy: A Study of the Time and Effort Involved in Self-Archiving.
http://eprints.ecs.soton.ac.uk/10688/

Casey J (2004) Intellectual Property Rights (IPR) In Networked E-Learning - a beginners guide for content developers. JISC Legal service.
http://www.jisclegal.ac.uk/publications/johncasey_1.htm.

Charlesworth A (2005) Rights in digital environments.
www.jisc.ac.uk/uploaded_documents/JISC%20Rights%20in%20Digital%20Environment%20Report.pdf

Day M (2003) Prospects for institutional e-print repositories in the United Kingdom. ePrints UK supporting study, no. 1.
www.rdn.ac.uk/projects/eprints-uk/docs/studies/impact

Dolphin I, Miller P & Sherratt R (2002) Portals, PORTALs everywhere. *Ariadne*, **Issue 33.**
http://www.ariadne.ac.uk/issue33/portals/

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

Duncan C, Barker E, Douglas P, Morrey M & Waelde C (2004) Digital rights management: final report.
http://www.intrallect.com/drm-study/

Dunsire G (2005) Harvesting Institutional Repositories in Scotland (HaIRST): final project report.
http://hairst.cdlr.strath.ac.uk/documents/HaIRST-FAIR-FP.pdf

Feijen M & van der Kuil A (2005) A recipe for Cream of Science: special content recruitment for Dutch institutional repositories. *Ariadne*, **Issue 45**.
http://www.ariadne.ac.uk/issue45/vanderkuil/

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

Franklin T (2005) Collaboration and interoperability for sharing resource catalogues between the Resource Discovery Network and the Higher Education Academy and is Subject Centres. Report to the HEA, the RDN and JISC.
www.heacademy.ac.uk/InteroperabilityReportFinal.pdf

Fraser M (2005) Towards a research repository for Oxford University.

http://eprints.ouls.ox.ac.uk/archive/00001071/01/fraser_oxford_research_repository_report.pdf

Garrett JJ (2005) Ajax: a new approach to web applications. http://www.adaptivepath.com/publications/essays/archives/000385.php

Godby CJ, Young JA and Childress E (2004) A repository of metadata crosswalks. *D-Lib Magazine*, 10 (12). http://www.dlib.org/dlib/december04/godby/12godby.html

Halbert M, Butler J, Farley L, Furlough M, Sandore B, Walsh J, & Milewicz L (2006) DLF-Aquifer Services Institutional Survey report: Aquifer Services Working Group report on the Institutional User-Services Survey results. Digital Library Federation, Washington DC.
http://www.diglib.org/aquifer/SWGisrfinal.pdf

HCSTC (House of Commons Science and Technology Committee) (2004), Scientific Publications: Free for all*?* Tenth Report of Session 2003-04, The Stationery Office, London.
http://www.publications.parliament.uk/pa/cm/cmsctech.htm

Heery R, and Anderson S. (2005). Digital repositories review.
http://www.jisc.ac.uk/uploaded_documents/digital-repositories-review-2005.pdf

Heery R and Patel M (2000) Application profiles: mixing and matching metadata schemas. *Ariadne*, **Issue 25**.
http://www.ariadne.ac.uk/issue25/app-profiles/

HEFCE (2003) Intellectual property rights in e-learning programmes: report of the Working Group.
www.hefce.ac.uk/pubs/hefce/2003/03_08.htm

Hey J (2004) Targeting academic research with Southampton's institutional repository. *Ariadne*, **Issue 40**, pp 1-14
http://www.ariadne.ac.uk/issue40/hey/

Hultman Ozek Y (2005) *Lund Virtual Medical Journal* makes self-archiving attractive and easy for authors. *D-Lib Magazine*, **Issue 11 (10)**.

Hunter P (2005) Institutional e-print repositories: business and IPR issues. ePrints UK supporting study no. 3.
www.rdn.ac.uk/projects/eprints-uk/docs/studies/business-ipr

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

JupiterMedia (2003) Jupiter research reports that web site "personalization" does not always provide positive results
http://www.jupitermedia.com/corporate/releases/03.10.14-newjupresearch.html

Lagoze C and Van de Sompel H (2003) The making of the Open Archives Initiative Protocol for Metadata Harvesting. *Library Hi Tech*, 21 (2): 118-128.
doi:10.1108/07378830310479776)

Lynch C (2001) Personalization and recommender systems in the larger context: new directions and research questions. *Proceedings of the Second DELOS Network of Excellence Workshop on Personalisation and Recommender Systems in Digital Libraries*
http://www.ercim.org/publication/ws-proceedings/DelNoe02/CliffordLynchAbstract.pdf

Lynch CA (2003) Institutional repositories: essential infrastructure for scholarship in the digital age. *ARL Bimonthly Report* **226**, February 2003.
http://www.arl.org/newsltr/226/ir.html

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

MacLeod R (2005) List of engineering repository resources Version 1.0 (10/11/05).
http://www.icbl.hw.ac.uk/perx/sourceslisting.htm

MacLeod R & Moffatt M (2005) Engineering digital repositories landscape analysis and implications for PerV (Version 1.0; 10/11/05).
http://www.icbl.hw.ac.uk/perx/analysis.htm

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

McCown F, Liu X, Nelson ML and Zubair M. (2006) 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

Medyckyj-Scott D, Chappell C, Pradhan A & O'Hanlon C (2001) A geo-spatial data resource discovery tool for UK Further and Higher Education – Project overview and recommendations
http://edina.ac.uk/projects/geobrowser/Overview_v1-1.doc

Morris D, Hobert CB, Osmus L & Wool G (2000). Cataloguing staff costs revisited. Library Resources and technical Services **44 (2)**, 70-83.

Nielsen J (2003) Intranet portals: A tool metaphor for corporate information.

www.useit.com/alertbox/20030331.html

Nicholson D, Neill S, Currier S, Will L, Gilchrist A, Russell R & Day M (2001) HILT: High-level thesaurus project: final report.
http://hilt.cdlr.strath.ac.uk/Reports/Documents/HILTfinalreport.doc

Nicholson D, Shiri A & McCulloch (2005) HILT: High-Level Thesaurus Project Phase II: A Terminologies Server for the JISC Information Environment: Final Report To JISC
http://hilt.cdlr.strath.ac.uk/hilt2web/finalreport.htm

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

O'Reilly T (2005) What is Web 2.0: Design Patterns and Business Models for the Next Generation of Software.
http://www.oreillynet.com/pub/a/oreilly/tim/news/2005/09/30/what-is-web-20.html

Pearce L & Martin R (2003) Stakeholder requirements for external content in institutional portals: version 1.
http://www.fair-portal.hull.ac.uk/downloads/stakereq.pdf

Peters TA (2002). Digital repositories: Individual, discipline-based, institutional, consortial, or national? *The Journal of Academic Librarianship*, 28(6), pp. 414-417.

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

Pringle M (2005) The Digital Picture: final report.
http://thedigitalpicture.ac.uk/documents/pdf/digital_picture_final_report.pdf

Rankin, J. (2005). Institutional Repositories for the Research Sector:Feasibility Study. National Library of New Zealand
http://wiki.tertiary.govt.nz/static/wikifarm/InstitutionalRepositories.uploads/Main/IR_report.pdf

Rappa M (2000): Business models on the Web: managing the digital enterprise. North Carolina State University, USA, 2000.
digitalenterprise.org/models/models.html

RCUK Position Statement on Access to Research Outputs (2005)
http://www.rcuk.ac.uk/access/statement.pdf

Robertson RJ (2005) Stargate Project plan Version 1.1
http://cdlr.strath.ac.uk/stargate/SGProject_Plan1_1_NB.pdf

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

Smith N, Schmoller S & Ferguson N (2004) Personalisation in presentation services.
http://www.therightplace.plus.com/jp/jp-study-15.pdf

Sparks S (Rightscom Ltd) (2005) JISC Disciplinary Differences Report.
www.jisc.ac.uk/Uploaded_Documents/Disciplinary%20Differences%20and%20Needs.doc

Stephen Tand Harrison T (2002). Building systems responsive to intellectual tradition and scholarly culture. *Journal of Electronic Publishing,* 8(1).
http://www.press.umich.edu/jep/08-01/stephen.html

Stevenson J (2005) JORUM Preservation Watch Report
http://www.jorum.ac.uk/docs/pdf/Digital_Preservation_Report.pdf

Swan A, Needham P, Probets P, Muir A, O'Brien A, Oppenheim C, Hardy R & Rowland F (2004). Delivery, management and access model for E-prints and open access journals within further and higher education (Report of a JISC study), pp 1-121.
http://www.jisc.ac.uk/uploaded_documents/ACF1E88.pdf

Swan, Alma and Brown, Sheridan (2005) Open Access self-archiving: pp1-104. An author study. Published by the JISC.
http://www.keyperspectives.co.uk/openaccessarchive/reports/Open%20Access%20II%20(author%20survey%20on%20self%20archiving)%202005.pdf or
http://www.jisc.ac.uk/uploaded_documents/Open%20Access%20Self%20Archiving-an%20author%20study.pdf

Timmers P (1998) Business models for electronic markets. In: Gadient, Yves, Schmid, Beat F, Selz, Dorian, EM-Electronic Commerce in Europe. EM- Electronic Markets, **8 (2)** 07/98.
www.electronicmarkets.org/modules/pub/view.php/electronicmarkets-183

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

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

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

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

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

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

Waters DJ (2001) The metadata harvesting initiative of the Mellon Foundation. ARL Bimonthly Report **217**, August 2001. http://www.arl.org/newsltr/217/waters.html

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

Further bibliography for the technical issues covered in this report

Bekaert, J. and Van de Sompel, H. A standards-based solution for the accurate transfer of digital assets. D-Lib Magazine, June 2005, 11 (6). Available at http://www.dlib.org/dlib/june05/bekaert/06bekaert.html

Bekaert, J., Balakireva, L, Hochstenbach, P. and Van de Sompel, H. Using MPEG-21 DIP and NISO OpenURL for the dynamic dissemination of complex digital objects in the Los Alamos National Laboratory Digital Library. D-Lib Magazine, February 2004, 10 (2). Available at http://www.dlib.org/dlib/february04/bekaert/02bekaert.html

Campbell, L.M., Blinco, K. and Mason, J. Repository management and implementation. White paper for alt-i-lab 2004. Available at http://www.jisc.ac.uk/uploaded_documents/Altilab04-repositories.pdf

Caplan, P. Preservation rumination: digital preservation and the unfamiliar future. OCLC Distinguished Seminar Series presentation, February 2005. Available at http://www.oclc.org/research/dss/ppt/dss_caplan.ppt

Coleman, A. and Roback, J. Open access federation for library and information science: dLIST and DL-Harvest. D-Lib Magazine, December 2005, 11 (12). Available at http://www.dlib.org/dlib/december05/coleman/12coleman.html

Dempsey, L. Libraries and the long-tail: some thoughts about libraries in a network age. D-Lib Magazine, April 2006, 12 (4). Available at http://www.dlib.org/dlib/april06/dempsey/04dempsey.html

Dempsey, L., Childress, E, Godby, C.J., Hickey, T.B., Houghton, A., Vizine-Goetz, D. and Young, J. Metadata switch: thinking about some metadata management and knowledge organization issues in the changing research and learning landscape. In: Escholarship: a LITA Guide, ed. Shapiro, D., 2005, American Library Association. Preprint available at: http://www.oclc.org/research/publications/archive/2004/dempsey-mslitaguide.pdf

Foulonneau, M., Habing, T.G. and Cole, T.W. Automated capture of thumbnails and thumbshots for use by metadata aggregation services. D-Lib Magazine, January 2006, 12 (1). Available at http://dlib.org/dlib/january06/foulonneau/01foulonneau.html

Gilby, J. Distributed services registry workshop. Ariadne, October 2005, Issue 45. Available at http://www.ariadne.ac.uk/issue45/dsr-rpt/

Hagedorn, K. Service provider (SP) issues. http://kathagedorn.com/SP_issues.html

Halm, M.J., Hatala, M., Morr, D. and Valentine, A. LionShare: a hybrid secure network for academic collaboration. Presentation at Internet 2 Meeting, Fall 2005.

Hochstenbach, P., Jerez, H. and Van de Sompel, H. The OAI-PMH static repository and static repository gateway. In: Proceedings of the 2003 Joint Conference on Digital Libraries (JCDL '03), May 1-2, 2003, Houston, Texas. Pre-print available at http://public.lanl.gov/herbertv/papers/jcdl2003-submitted-draft.pdf

Hunter, P. and Day, M. Institutional repositories, aggregator services and collection development. ePrints UK supporting study, no. 2, 2005. Available at http://www.rdn.ac.uk/projects/eprints-uk/docs/studies/coll-development/coll-development.pdf

Hunter, P., Heery, R., Powell, A., Martin, R., Napier, M. and Day, M. ePrints UK Final Report, 2005. Available at http://www.rdn.ac.uk/projects/eprints-uk/docs/final-report/eprints-uk-final-20050316.pdf

IEEE LTSC CMI Working Group. The RAMLET project – developing a reference model for resource aggregation for learning, education, and training. RAMLET project description, 2005. Available at http://ieeeltsc.org/wg11CMI/ramlet/Pub/RAMLET_project_description.pdf

Kott, K. DLF Aquifer. Presentation at the Digital Libraries Forum Fall 2005 Meeting. Available at http://www.diglib.org/forums/fall2005/presentations/aquifer1105.htm
Kraan, W. and Mason, J. Issues in federating repositories: a report on the First International CORDRA Workshop. D-Lib Magazine, March 2005, 11 (3). Available at http://www.dlib.org/dlib/march05/kraan/03kraan.html

Lavoie, B., Dempsey, L. and Connaway, L.S. Making data work harder. Library Journal, 2006, 15 January. Available at http://www.libraryjournal.com/article/CA6298444.html

Liu, X, Maly, K., Nelson, M.L. and Zubair, M. Lessons learned with Arc, an OAI-PMH service provider. Library Trends, 2005, 53 (4): 590-603

Lossau, N. Search engine technology and digital libraries: libraries need to discover the academic internet. D-Lib Magazine, June 2004, 10 (6). Available at http://www.dlib.org/dlib/june04/lossau/06lossau.html

Lynch, C.A. Where do we go from here? The next decade for digital libraries. D-Lib Magazine, July/August 2005, 11 (7/8). Available at http://www.dlib.org/dlib/july05/lynch/07lynch.html

Mazzocchi, S. On the quality of metadata… Stefano's Linotype Blog, 16th January 2006. Available at http://www.betaversion.org/~stefano/linotype/news/95/

Moffat, M. 'Marketing' with metadata – how metadata can increase exposure and visibility of online content. PerX project report, 2006. Available at http://www.icbl.hw.ac.uk/perx/advocacy/exposingmetadata.htm

Payette, S., Blanchi, C., Lagoze, C. and Overly, E.A. Interoperability for digital objects and repositories: the Cornell/CNRI experiments. D-Lib Magazine, May 1999 5 (5). Available at http://www.dlib.org/dlib/may99/payette/05payette.html

Pearce, J. with Gatenby, J. New frameworks for resource discovery and delivery. National Library of Australia Staff Papers, 2005. Available at http://www.nla.gov.au/nla/staffpaper/2005/pearce1.html

Powell, A. JISC Information Environment Technical Standards. Available at http://www.ukoln.ac.uk/distributed-systems/jisc-ie/arch/standards/

Powell, A. Notes about possible technical criteria for evaluating institutional repository (IR) software. UKOLN report, December 2005.

Powell, A. The JISC Resource Discovery Landscape: a personal reflection on the JISC Information Environment and related activities, May 2005. Available at http://www.ukoln.ac.uk/distributed-systems/jisc-ie/arch/resource-discovery-review/

Research Information Network. Resource discovery – a workshop organised by the RIN, 20 October 2005. Papers available at http://www.rin.ac.uk/?q=resource-discovery-workshop-20-october

Roberts, G., Aalderink, W., Cook, J., Feijen, M., Harvey, J., Lee, S. and Wade, V.P. Reflective learning, future thinking: digital repositories, e-portfolios, informal learning and ubiquitous computing. Report from the ALT/SURF/ILTA Spring Conference Research Seminar, 1 April 2005, Trinity College, Dublin.

Scherlis, W.L. Repository interoperability workshop: towards a repository reference model. D-Lib Magazine, October 1996 2 (10). Available at http://www.dlib.org/dlib/october96/workshop/10scherlis.html

Shreeves, S.L., Habing, T.G., Hagedorn, K. and Young, J.A. Current developments and future trends for the OAI Protocol for Metadata Harvesting. Library Trends, 2005, 53 (4): 576-598

Summann, F. and Lossau, N. Search engine technology and digital libraries: moving from theory to practice. D-Lib Magazine, September 2004, 10 (9). Available at http://www.dlib.org/dlib/september04/lossau/09lossau.html

Van de Sompel, H., Payette, S. Erickson, J. Lagoze, C. and Warner, S. Rethinking scholarly communication: building the system that scholars deserve. D-Lib Magazine, September 2004, 10 (9). Available at http://www.dlib.org/dlib/september04/vandesompel/09vandesompel.html

Van de Sompel, H., Young, J.A. and Hickey, T.B. Using the OAI-PMH … differently. D-Lib Magazine, July/August 2003 9 (7/8). Available at http://www.dlib.org/dlib/july03/young/07young.html

Waaijers, L. From libraries to 'libratories'. First Monday, 2005, 10 (12). Available at http://firstmonday.org/issues/issue10_12/waaijers/index.html