

Institutional repositories, aggregator services and collection development

Philip Hunter and Michael Day
UKOLN, University of Bath, Bath BA2 7AY, United Kingdom
<http://www.ukoln.ac.uk/>

ePrints UK supporting study, no. 2

Version 0.2, 24 January 2005

Abstract: Institutional repositories are managed collections of the intellectual output of university and other research-based institutions. This report introduces collection development issues from two distinct perspectives. Firstly, it highlights issues that may need to be addressed by institutional repositories as OAI data providers. For example, repositories may need to make decisions on the type, quality and format of content, on submission workflows, rights management, access, sustainability and evaluation. Secondly, the report will consider similar issues from the perspective of third party service providers like ePrints UK that harvest selective metadata from institutional repositories. The concluding section will provide some recommendations on best practice for repositories to support such harvesting.

Contents

1. Introduction	2
2. Collection development and institutional repositories	2
3. Repository perspectives	3
3.1 Content	3
3.2 Submission	4
3.3 Intellectual property rights	5
3.4 Access	6
3.5 Sustainability	6
3.6 Evaluation	7
4. Aggregator perspectives	7
4.1 Content	8
4.2 Metadata harvesting	8
4.3 Intellectual property rights	9
4.4 Sustainability	9
4.5 Evaluation	10
5. Conclusions and recommendations	10
References	10

1. Introduction

The ePrints UK (<http://www.rdn.ac.uk/projects/eprints-uk/>) project is funded by the Joint Information Systems Committee (JISC) as part of its Focus on Access to Institutional Resources (FAIR) Programme (Martin, 2003). The aim of the project is to develop a demonstration national service provider repository of e-print records based at the University of Bath, derived by harvesting metadata from institutional and subject-based e-prints archives using the Open Archive Initiative Protocol for Metadata Harvesting (OAI-PMH). In addition, the project aims to provide access to these institutional assets through the eight Resource Discovery Network (RDN) faculty level hubs and the Education Portal based at the University of Leeds. It is also investigating the use of Web Services technologies for the enhancement of metadata and for the automatic linking of citations.

An earlier ePrints UK supporting study introduced the project and assessed the prospects for institutional repositories in the UK (Day, 2003). This study looks in more detail at collection development issues from the perspective of both data providers (repositories) and aggregator services like ePrints UK.

2. Collection development and institutional repositories

Collection development is a term that is used by the library community to refer to "the selection and acquisition of material for an expanding collection and decisions on the material to be included in that collection" (Law, 1991, p. 1). Guidelines from the IFLA (International Federation of Library Associations and Institutions) Acquisition and Collection Development Section say that the primary function of collection development policies is to provide guidance on selecting and deselecting resources from a collection (Biblarz, *et al.*, 2001). Typically, collection development decisions apply at different stages of an institution's workflow, including the selection, acquisition, processing, housing, weeding, retention, preservation, relegation and discarding of library materials. Much of the focus in the library world has been on the development and implementation of written collection development policies and guidelines, although the practical value of these has sometimes been questioned (e.g., Hazen, 1995; Snow, 1996).

Collection development is now viewed as being a part of the wider activity of collection management, which includes "collection policy development, materials budget allocation, selection, collection analysis, collection use and user studies, training and organisation of collection development staff, preservation, and cooperative collection development" (Branin, Groen & Thorin, 2000, 24). In research libraries at least, there has been an increasing emphasis in recent years on the need for cooperative collection development, partly in recognition of the fact that no single institution can aspire to collect the entire record of scholarship, but also because of the development of new delivery technologies. Some libraries, especially in North America have utilised tools like Conspectus for the assessment of collection depth, e.g. for supporting co-operative collection management and resource sharing between libraries (e.g., Bushing, 2001). Ultimately, however, co-operation is dependent on the availability of up-to-date information on acquisitions and holdings (Nichols & Smith, 2001, 87).

The increasing availability and use of information in digital form has led some commentators to conclude that the currently decentralised physical repository-based system will eventually be replaced by a more centrally-organised system, whereby libraries would merely act as the gatekeepers to and aggregators of digital information ultimately managed elsewhere (e.g., Branin, Groen & Thorin, 2000, 28). Collection development in this context may mean co-operating with other research libraries on the licensing of digital content, providing seamless user access to digital content from different providers, or negotiating the right to maintain access to licensed older content even after a subscription has been cancelled. Simultaneously, however, universities and other research organisations are being encouraged to reclaim responsibility for the management and distribution of scholarship, e.g. through the creation of institutional repositories (e.g., Crow, 2002; Lynch, 2003).

Institutional repositories bring universities and other institutions firmly back into the business of managing information, whether this would be research publications or data, learning resources, or administrative records. The setting-up of a repository implies an institutional commitment to the ongoing management of such information. Lynch (2003) calls university-based institutional repositories an 'organisational commitment' to the stewardship of digital materials, "including long-term preservation where appropriate, as well as organization and access or distribution." As such, institutional repositories are collections that have their own set of collection developments requirements.

The architecture of the OAI-PMH means, however, that collection development has to be addressed at two distinct levels, i.e. in terms of what the protocol defines as data providers and service providers (Lagoze, *et al.*, 2002). Data providers or repositories will be concerned with the collection development aspects of their own service, e.g. defining the exact scope of their service, quality levels, etc. (e.g., Wolpert, 2002). Service provider or aggregator services (like ePrints UK) will need to consider a range of different issues, in part based on the information made available by data providers. Collection development issues will include, e.g. the balance of subject coverage, the resource types included, and the quality levels of the content being selected. In this, there are parallels with other broker-type services. For example, the Renardus subject gateway broker service (<http://www.renardus.org/>) requires participating services to have well-defined collection development policies (Renardus, 2002). The following sections of this report will look in more detail at collection development issues from the perspectives of both repositories and aggregator service providers like the pilot developed by the ePrints UK project.

3. Repository perspectives

In the context of the OAI-PMH, 'data providers' are understood as those that administer systems that support the protocol as a means of exposing metadata (Lagoze, *et al.*, 2002). Understood in terms of the protocol, most institution-based repositories using the protocol to expose metadata will be data providers, although they themselves could (in principle) also be service providers harvesting from multiple data providers within a single institution.

There is much more to the setting up of an institutional repository than choosing some repository software, implementing it, and requesting staff to contribute content (e.g., Pinfield, Gardner & MacColl, 2002). Each institution needs to have a common understanding of the purpose of the repository as well as a set of policies that define its intended scope, together with information on issues like deposit, access, and sustainability. These topics can all be seen as aspects of collection management or development. The following sections will elaborate in more detail the types of collection development issues that may need to be considered when setting up an institutional repository.

3.1 Content

The main scope of institutional repositories is defined by the name. Crow (2002, 16) defines a repository as "a digital archive of the intellectual product created by the faculty, research staff, and students of an institution..." Within this general framework, however, various choices need to be made about the exact content of repositories.

A first decision concerns the types of output that the repository should support. Much initial effort has been expended on developing repositories for research papers (or e-prints) and theses. In addition, however, repositories could be used for research data, learning resources, or even the administrative records of an institution. Lynch (2003) suggests that "mature and fully realised" repositories may also contain "documentation of the activities of the institution itself in the form of records of events and performance and of the ongoing intellectual life of the institution." Repository software like DSpace (<http://www.dspace.org/>) is designed to handle a wide range of different resource types. Obviously, the content scope of repositories can evolve as circumstances change, but it is still useful to define clearly the types of output that should be included, not least because different types of output will have different requirements in terms of metadata or access.

Another decision that needs to be made, especially for institutional repositories that include e-prints, concerns peer review. In order to ensure the quality of content, some repositories may deliberately limit their scope to e-prints that have either been published in (or submitted to) peer-reviewed publications. Others may allow non peer-reviewed material (e.g. preprints, working papers, newspaper opinion pieces) to be submitted to the repository, but will distinguish these in some way from the peer-reviewed outputs. Similar decisions could be made about the inclusion of the final draft of an e-print or the version produced by the publisher.

A more practical decision may concern the format of outputs that can be submitted to the institutional repository. A repository may decide that it will only automatically support popular formats like HTML, PDF, PostScript, TeX or TIFF, although it may be able to consider well-documented other formats on request. The range of formats that a repository needs to support will partly depend on the types of output being collected. For example, repositories that collect research data will most likely need to deal with many different format types and will also need to collect the metadata codebooks that define them.

3.2 Submission

Once the collection scope has been carefully defined, thought can then be given to the submission workflow. The practical aspects of the submission interface are largely dependent on the repository software chosen, although an institution could presumably develop its own submission system. Repository software usually has a Web-based submission interface from which it is possible to upload outputs and add essential descriptive metadata [1]. Once content is uploaded, it can be held in a 'buffer' area for checking before being made available in the repository. The exact nature of these checks is up to the institution itself, but it might include simple tests of whether the submitter has the authority to upload an output to the repository [2], whether the output adheres with the scope of the repository, or whether the quality of the associated metadata is adequate.

The manual creation of descriptive metadata can be a time-consuming process and tends not to be a priority for research staff, so some thought has gone into developing processes that may support the submission workflow. This may include the use of library staff to review and enhance the metadata creation process (e.g., Drake, 2004). A survey of data providers by Gadd, Oppenheim & Proberts (2004a) revealed that four out of nineteen repositories undertook accuracy or other quality control checks on the metadata while others provided subject indexing and abstracting support. Other metadata enhancement approaches depend on the automatic capture of metadata or on the development of innovative services that can support the creation workflow. For example, OCLC Research has experimented with the development of various modular services that can be used to support the new environment of distributed repositories. A good example of this is OCLC's name authority Web service, which can be plugged into an instantiation of the DSpace submission interface to produce, where they are available, authoritative forms of names from the Library of Congress Name Authority File (Dempsey, *et al.*, 2005). Similarly, building on earlier experience with projects like Scorpion, OCLC have begun to embed automatic classification tools in cataloguing interfaces like OCLC Connexion (<http://www.oclc.org/connexion/>) and is also experimenting with a subject classification Web service in co-operation with the ePrints UK project (Dempsey, *et al.*, 2005).

On occasion, there may be a need to withdraw outputs from the institutional repository. Some existing repositories like arXiv allow submitters to withdraw a paper, although they require some explanation to be provided and previous versions of the paper will remain available (<http://arxiv.org/help/withdraw>). The withdrawal of a paper from the published scholarly record

1. For an example, see the DSpace submission interface demonstration at:
<http://dspace.org/implement/submit-content.html>

2. Alas, there are already some examples of where people with no official connection with an institution have managed to submit e-prints into an institutional repository.

remains a contentious issue and a change in Elsevier Science policy allowing retracted papers (for whatever reason) to be almost completely erased from the scholarly record drew some critical comment (e.g., Klarreich, 2001; Plutchak, 2002). The archival nature of institutional repositories means that withdrawal policies are likely to be more stringent than for subject-based repositories, although withdrawal of access may be a different issue. There will still be a need for policies and workflows to support the removal or retraction of papers, e.g. in cases of scientific misconduct. In addition, the question of legal responsibility in cases of defamation, infringements of data protection legislation, or other forms of illegal content has not really been considered in any detail with regard to repositories (e.g., Giles, 2003). A general survey of legal issues relating to the preservation of Web resources suggests that institutions may need to take care in this area (Charlesworth, 2003).

3.3 Intellectual property rights

Other legal issues that may need to be dealt with by repositories relate to intellectual property rights. It is perhaps worth noting that this subject has been extensively investigated by a FAIR Programme funded project called RoMEO (Rights METadata for Open archiving). In 2002-2003, the project undertook a series of surveys of the attitudes of academics towards intellectual property rights in e-prints, of publishers' copyright policies, and the opinions of data and service providers; this work all being used to explore the possibility of how information about rights might best be communicated as metadata (Gadd, Oppenheim & Probets, 2003b; 2003c; 2003d; 2003e; 2004a; 2004b). As part of the survey process, Project RoMEO compiled a list of journal publishers' copyright policies on 'self-archiving,' a useful resource that is currently maintained by the SHERPA project (<http://www.sherpa.ac.uk/romeo.php>).

In the changing world of scholarly publishing, the transfer of copyright to journal publishers has become a focus point for a discussion of the respective requirements of authors and publishers. Traditionally, when authors wanted to publish a research paper in a journal they would, on acceptance, assign copyright to the publisher or (more recently) grant publishers some kind of 'exclusive licence' to publish. Initially, many of these agreements would not permit the inclusion of a paper in an institutional repository. Until recently, for example, the license agreement issued by Nature Publishing Group (2003) allowed authors to "re-use the papers in any printed volume of which they are an author; to post a PDF copy on their own (not-for-profit) website; to copy (and for their institutions to copy) their papers for use in coursework teaching; and to re-use figures and tables," but expressly excluded inclusion in "open archival websites, such as those that host collections of articles by an institution's researchers." In response to authors' demands, however, many journal publishers' contracts and licenses are now far more permissive. For example, Nature Publishing Group's current licence actually *encourages* authors to post a copy of a paper in an institutional repository six months after publication of the printed edition, on condition that a hyperlink to the journal Web site is provided (http://npg.nature.com/pdf/05_news.pdf).

The key issue for institutional repositories here is that they attempt to ensure that they do not unintentionally infringe copyright or other intellectual property rights. In order to achieve this, some repositories ask those submitting content to confirm that they have the rights to upload a paper, although the Project RoMEO survey of data providers suggested that this might not be a widespread practice. Gadd, Oppenheim and Probets (2004a) reported that the largest group of respondents were happy to merely take it on trust that the submitter had the right to deposit a paper. In a survey produced for the Open Archives Forum project, Bide (2002, 25-26) has suggested the need for explicit agreements with depositing authors, maybe as an automated part of the submission process. His examples include, "warranties on the part of the author that they are not breaching any third party agreement - or copyright - by posting the eprint." Ascertaining who actually holds rights in a paper is sometimes more difficult than it may first appear. For example, an earlier RoMEO survey found that around a third of academics were not sure who owned the copyrights in research papers (Gadd, Oppenheim & Probets, 2003b). That said, the changing attitudes of publishers mean that copyright assignment to publishers is probably less of a problem than it appeared to be a few years ago.

One response to the copyright problem has been for institutions or research funding bodies to attempt to reassert their own rights. Some universities are beginning to insert corporate ownership of intellectual property rights in university statutes and employment contracts, especially with regard to patents or learning resources. Bide (2002, 23) has described the question of ownership of intellectual property rights of academics as "one of the more contentious issues" facing higher education today. He says that the terms of the UK's *Copyright, Designs and Patents Act, 1988* means that copyright in works made "in the course of employment" would normally pass to the employer. However, in practice, he notes, "most academic institutions do not exercise this right with respect to copyrights in journal articles or in textbooks." Perceptively, Lynch (2003) has warned against universities using institutional repositories as a means of asserting control or ownership of that intellectual work that has traditionally be controlled by academics.

Unlike universities, those organisations that fund research cannot claim copyright over the publications that result from its grants. However, a number of funding bodies have recently expressed their willingness to require grantees to provide open-access to such publications. For example, the Wellcome Trust are proposing that its grantees will be required to deposit electronic versions of research papers in PubMed Central (or its European counterpart) within six months of publication (<http://www.wellcome.ac.uk/assets/wtx022820.pdf>). It is perhaps worth noting that policy changes that 'mandate' the deposit of research papers in institutional repositories could have significant organisational implications for those who manage them.

3.4 Access

Crow's short definition of institutional repositories says that they should be "...accessible to end users both within and outside of the institution, with few if any barriers to access" (Crow, 2002, 16). While this is fully in accord with open-access principles and may be desirable for most research papers, there are a number of reasons why institutional repositories may not make *all* content publicly available. Potentially restricted content might include research papers that have been retracted, reports or theses that contain commercially sensitive information, datasets that are in the process of being refined, learning resources, or administrative records. Some of this material may be legitimately distributed within the institution itself (or parts of it), but the content (and its descriptive metadata) would not routinely be made available outside. Repositories, therefore, may need to define access levels for different types of content, e.g. to place access control mechanisms on restricted content types while ensuring that adequate descriptive metadata is made available for that content that can be shared more widely.

3.5 Sustainability

A more general collection management issue is the long-term sustainability of the repository itself. This has two main aspects; firstly the need for ongoing strategic and financial support from the host institution, secondly the need to ensure continued long-term access to the content of repositories.

Lynch (2003) has said that it is "vital that institutions recognize institutional repositories as a serious and long-lasting commitment to the campus community (and to the scholarly world, and the public at large) that should not be made lightly." He warns that repositories can fail for a number of reasons, e.g. lack of strategic or financial support from institutions, management failure, and technical problems. In addition, institutions are not the stable entities that they sometimes appear to be. For example, new departments or research centres can be opened, old ones can be closed, merged with others, or move to a different institution. Also, while higher education institutions rarely close down completely (although this could theoretically happen), they do increasingly merge with others (e.g., <http://education.guardian.co.uk/universitymergers/>). For this reason it is important that institutional repositories secure high-level political support within institutions. Also that they should develop contingency plans that can be implemented if and when circumstances change. This may mean making arrangements with other institutions or with 'repositories of last resort' like national libraries. In this context, it is perhaps interesting that a growing number of journal publishers have begun to make similar arrangements, e.g. publishers

like Elsevier Science, Kluwer Academic, Blackwell, BioMed Central, Oxford University Press and Taylor & Francis have all recently signed deposit agreements with the National Library of the Netherlands (<http://www.kb.nl/nieuws/>).

Ensuring long-term access to the content of repositories is yet another challenge that will need to be faced by institutional repositories. A JISC-funded feasibility study on the preservation of e-prints highlighted the importance of file formats, metadata and organisational strategies (James, *et al.*, 2003; Pinfield & James, 2003). Existing e-print repositories often only accept a limited number of file formats, usually based on the perceived download preferences of users. These typically include a mixture of proprietary and 'open' formats, e.g. HTML, PDF, PostScript, TeX, MS Word, MS PowerPoint and TIFF. James, *et al.* (2003) recommend that repositories should assess the preservation risks of file formats in their collections and consider format conversion, e.g. to those based on open standards or XML. There is also a need to maintain information about stored file formats, possibly in co-operation with third party format registries like the proposed by Global Digital Format Registry (<http://hul.harvard.edu/gdfr/>). The need for appropriate metadata to support digital preservation processes has been recognised for some time (e.g., Day, 2004). James, *et al.* (2003) recommended that repositories should collaborate on the production of a common set of preservation metadata. Progress on both of these issues will be important but ultimately the long-term preservation of repository content will be dependent on the development of appropriate organisational strategies. This suggests that institutional repositories may eventually need to become trusted digital repositories.

A working group sponsored by the Research Libraries Group and OCLC Online Computer Library Center has defined some of the main attributes of trusted digital repositories. To summarise, these include the need to accept responsibility for the preservation of content, to obtain sufficient control over content in order to be able to preserve it, to demonstrate financial sustainability and organisational viability, to ensure that there are documented policies and procedures that can be monitored and evaluated, and to adhere to standards and best practice (Research Libraries Group, 2002). Most existing repositories would have difficulty fulfilling all of these criteria. Instead, James, *et al.* (2003, 53) suggest that many existing e-print repositories are focused primarily on access rather than preservation and that project-based funding may not be the best way of building long-term sustainability. The best way forward may be for repositories, where necessary, to co-operate with specialist sources of preservation expertise or third party preservation services. It is perhaps worth noting that James, *et al.* (2003, 55-57) have produced a number of useful recommendations in this area.

3.6 Evaluation

Institutional repositories as data providers will also need to provide a full range of services for its user communities. These may include, e.g. managed data storage, metadata creation and enhancement, search and retrieval, the export of metadata about research outputs to research assessment procedures. As such, repositories will need periodic evaluation to ensure that they are fulfilling its basic institutional requirements and user needs.

4. Aggregator perspectives

In the model promulgated by the OAI-PMH, service providers are those that issue protocol requests to data providers and use the metadata as "a basis for building value-added services" (<http://www.openarchives.org/documents/FAQ.html>). Some service providers concentrate on aggregating metadata content from multiple data providers. For example, the experimental ARC service from Old Dominion University harvests metadata from over 80 data providers and stores them in a searchable relational database (<http://arc.cs.odu.edu/>). The University of Michigan's OAIster project harvests metadata describing a wide range of digital content from many different institutions; in January 2005 the service gave access to over 4.5 million records from almost 400 data providers (<http://www.oaister.org/>). Other service providers focus on providing specific types of value-added functionality. For example, the experimental Citebase service developed by the University of Southampton extracts reference data from the full-text of papers and combines this with harvested metadata this to build a citation database (<http://citebase.eprints.org/>). The ePrints

UK project also developed an experimental service provider that was intended to aggregate and federate access to UK-based repositories and test some added value services related to metadata enhancement and citation linking (<http://eprints-uk.rdn.ac.uk/search/>).

The collection development policies needed for service providers will have some superficial similarities with those needed for data providers, however their implementation will be to a large extent dictated by the amount and quality of metadata made available by repositories. Before proceeding, it is perhaps worth noting that institutional repositories can themselves be service providers, e.g. aggregating metadata about content from a number of data providers under their control. In addition, that service providers can also be data providers, to the extent that they can make the content they harvest available through the OAI protocol.

4.1 Content

Services that aggregate content from multiple data providers have to resolve some of the same collection development questions as repositories. In addition to selecting which data providers to harvest metadata from, aggregators also need to make decisions, e.g. about the type, quality-level, subject or geographical origin of the content that they make available through their service. However, there needs to be sufficient metadata available to support these choices. For example, an aggregator wanting to focus on harvesting only metadata about peer-reviewed research outputs would be dependent on these being held in a separate repository from non-reviewed outputs or this information being clearly included in each record's metadata. Similar principles would apply if the aggregator were only interested in harvesting metadata about research datasets, theses or learning resources.

Ascertaining subject coverage is potentially more problematic. Obviously, it may be possible to select content based on its provenance in particular academic departments or schools, or on an analysis of the terms provided in subject fields. An alternative approach would be to harvest the full-text of content (where this is available) and undertake some kind of automatic classification process. For example, the ePrints UK project has considered using automatic classification tools developed by OCLC Research for helping to distribute repository content amongst the eight faculty-level services of the RDN.

Another potential issue related to content is the potential duplication of content. In practice, research outputs are sometimes going to be submitted to multiple repositories, e.g. where they originate in more than one institution. It will be the task of the aggregator to ascertain how much of a problem this might be for users of the service and, where necessary, take steps to de-duplicate. There may be similar problems with different versions of the same content.

Ultimately, the service provider has no direct control over content. Repositories may permit submitters to remove outputs, or may restructure or close down. The collection development strategies of aggregators will have to take this into account.

4.2 Metadata harvesting

The sufficiency and quality of harvested metadata is the *sine qua non* for aggregator services. In this, however, they are again almost totally dependent on the data provider repositories that they harvest from. Unfortunately, studies of metadata usage in OAI contexts suggest that quality varies greatly (e.g., Barton, Currier & Hey, 2003; Halbert, Kaczmarek & Hagedorn, 2003; Ward, 2003), leading to what Halbert (2003) calls 'collisions' between metadata formats and problems with authority control and de-duplication.

Safeguarding the quality of metadata is fundamentally the role of the data provider repositories. As we have noted already, this depends on the provision of well-designed submission workflows, interfaces and quality control, also (perhaps) support from library and information professionals. Unfortunately the last of these, when scaled to the task of growing repositories, may appear expensive. However, in a genuinely distributed repository system, descriptive metadata for each output would only need to be created once, which has some cost advantages over current library cataloguing practice.

Guy, Powell and Day (2004) suggest some practical things that may help improve the creation of good-quality metadata. These include the provision of content guidelines for metadata (e.g., Powell, Day & Cliff, 2003), the improvement of metadata creation tools, and the implementation of appropriate quality control processes

As with repositories themselves, service providers may also have an interest in the automatic generation, capture or enhancement of metadata. Those interested in digital preservation have already experimented with tools that can capture technical metadata about formats (e.g., <http://hul.harvard.edu/jhove/>), but the capture of descriptive metadata directly from the full-text of harvested content is something that needs to be explored further in an OAI context. The use of structured text formats like those based on XML may help with this, as might software based on the Document Object Model (DOM) developed by the World Wide Web Consortium (<http://www.w3.org/DOM/>).

The ePrints UK project explored the use of various proposed third party services (using Web services technology) that could be used to enhance the metadata harvested by aggregators. The first of these was a name authority service developed by OCLC Research that would validate the forms of personal and organisational names in use, thus improving consistency in the use of names. The second, also based on technology developed by OCLC Research, was for a service that would automatically assign subject classification terms, thus helping to ensure consistency in subject metadata across the service provider and enabling harvested content to be accurately distributed to more focused subject-based services. A third service, provided by the University of Southampton, would parse the bibliographic references in the full text of harvested content into structured forms, thus facilitating citation linking and informetric analysis. All of these services have some potential to support the added-value functionality that can be offered by both repositories and aggregators. There needs, however, to be far more work in developing such third party tools and testing their use in realistic harvesting-based contexts.

4.3 Intellectual property rights

While institutional repositories as data providers are concerned with the intellectual property rights vested in both repository content and metadata, service providers are primarily concerned with metadata. This includes both the metadata that they harvest from data providers and any enhancements that they make to it in the process of offering their service (e.g., McClelland, *et al.*, 2002). Gadd, Oppenheim & Proberts (2004a) note that the copyright status of metadata under UK law is uncertain, although individual records could obtain protection as a compilation and collections of metadata as a database. They conclude that, in order to avoid infringing database right, service providers wishing to make use of a data provider's metadata should seek permission to do so. The RoMEO project survey showed that while most data providers agreed that collections of metadata enjoyed database right, they also felt that this might be implicitly waived in OAI contexts. Service providers were equally divided between those who did check the rights status of metadata before harvesting, those who considered it to be implicitly free under OAI, and those who had never thought about it. A majority were happy for others to harvest their enhanced metadata, although some conditions of use were specified.

There is, perhaps, a need for aggregators to define their own working assumptions about intellectual property rights with regard to harvesting metadata (or full-text) from repositories. Service providers should also consider publishing conditions of use where they make their own enhanced metadata available for harvesting by others.

4.4 Sustainability

While aggregator services do not have content that requires long-term management and curation, they do have a need for organisational and financial stability. Service providers mediate between content providers and their users and, as such, their sustainability will depend on whether they provide a service that is required and how well they perform this in a potentially competitive market. As services, however, they can evolve over time, both with regard to the content they harvest and the added-value services they provide.

Many service provider initiatives are funded on a project basis, e.g. services demonstrating the added value that can be provided by metadata harvesting using the OAI protocol. In the future, there is likely to be a 'mixed economy' of service providers, with some supported by public funding while others will be provided on a commercial or quasi-commercial basis.

4.5 Evaluation

As user-focused services, aggregators will need ongoing evaluation to ensure that they continue to fulfil user needs and to help identify new requirements. The results of such evaluation may lead to changes in collection development policies.

5. Conclusions and recommendations

This report has attempted to identify some of the main types of collection development issues that need to be considered when setting up both institutional repositories and aggregator services.

Collection development is largely concerned with *content*. As with any other digital library, decisions need to be made about which types of object should be included and which excluded from repositories. While the primary focus of a repository is defined by its institutional scope, additional decisions need to be made about the exact type of content required, e.g. whether it should just be for peer-reviewed research papers and theses, or for all research outputs and learning resources.

- It is recommended that institutional repositories (as data providers) should consider the production of written collection development policies. These should at the very least define the intended scope of the collection with regard to subject matter, object type and quality (this could be part of a collection-level description). The policy could also contain information on submission workflows (including intellectual property rights clearance), access, and include some consideration of sustainability. Some of this information, for example, is included in the guide for depositors produced by University College London, which includes information on the eligibility of papers, copyright clearance, eligible formats, metadata creation and upload procedures (<http://eprints.ucl.ac.uk/DepositGuide.html>).

Aggregators need to make similar decisions about content to repositories, e.g. about subject coverage, the type of resources to be included, whether it needs to be peer reviewed. However, their choices may be limited by the extent and consistency of metadata made available by repositories. For example, the consistent use of a standard type-list would facilitate the selection of repository content by output type. In the absence of consistent metadata, it is possible that aggregators could use software to support the selection of content, e.g. by using automatic subject classification or output type identification tools. In some contexts, it may be possible for repositories and aggregators to co-operate on the development of common metadata schemas that can support a wider range of functionality than is possible with the simple Dublin Core metadata mandated by the OAI protocol for basic cross-domain interoperability.

- It is recommended that repositories should develop (or adopt) metadata generation tools that facilitate the production of consistent metadata. This may be helped by the use of suitable metadata content guidelines, e.g. those for the use of simple Dublin Core to describe e-prints (Powell, Day & Cliff, 2003). Consideration should also be given to the use of software tools that can automatically capture some types of metadata, also to the integration of third party metadata enhancement services like OCLC's name authority service (Dempsey, *et al.*, 2005). In addition, co-operating repositories and aggregators may need to consider if there are additional types of metadata (e.g. about rights) that are necessary to support value-added services, and make arrangements to share this using the OAI protocol.

References

Barton, J., Currier, S., & Hey, J. (2003). "Building quality assurance into metadata creation: an analysis based on the learning objects and e-prints communities of practice." 2003 Dublin Core

- Conference (DC-2003), Seattle, Wa., USA, 28 September - 2 October. Available at: <http://purl.oclc.org/dc2003/03barton.pdf>
- Biblarz, D., Tarin, M.J., Vickery, J., & Bakker, T. (2001). *Guidelines for a collection development policy using the Conspectus model*. International Federation of Library Associations and Institutions, Acquisition and Collection Development Section, March. Available at: <http://www.ifla.org/VII/s14/nd1/gcdp-e.pdf>
- Bide, M. (2002). *Open archives and intellectual property: incompatible world views?* Open Archives Forum project deliverable D4.2. Available at: http://www.oaforum.org/otherfiles/oaf_d42_cser1_bide.pdf
- Branin, J., Groen, F., & Thorin, S. (2000). "The changing nature of collection management in research libraries." *Library Resources and Technical Services*, 44(1), 23-33.
- Bushing, M.C. (2001). "The evolution of Conspectus practice in libraries: the beginnings and the present applications." CASLIN 2001: Document Description and Access: a New Challenge, Beroun, Czech Republic, 27-31 May. Available at: <http://www.caslin.cz/caslin01/sbornik/conspectus.html>
- Charlesworth, A. (2003). *Legal issues relating to the archiving of Internet resources in the UK, EU, USA and Australia: a study undertaken for the JISC and Wellcome Trust*. Available at: <http://library.wellcome.ac.uk/assets/WTL039230.pdf>
- Crow, R. (2002). *The case for institutional repositories: a SPARC position paper*. Washington, D.C.: Scholarly Publishing & Academic Resources Coalition. Available at: <http://www.arl.org/sparc/IR/ir.html>
- Day, M. (2003). *Prospects for institutional e-print repositories in the United Kingdom*. ePrints UK supporting study, 1. Available at: <http://www.rdn.ac.uk/projects/eprints-uk/docs/studies/impact/>
- Day, M. (2004). "Preservation metadata." In: G. E. Gorman & D. G. Dorner, (eds.), *Metadata applications and information management*. International Yearbook of Library and Information Management, 2003-2004. London: Facet, 253-273. Preprint available at: <http://www.ukoln.ac.uk/metadata/publications/iylim-2003/>
- Dempsey, L., Childress, E. R., Godby, C. J., Hickey, T. B., Vizine-Goetz, D., & Young, J. (2005). "Metadata switch: thinking about some metadata management and knowledge organization issues in the changing research and learning landscape." Forthcoming in: D. Shapiro, (ed.), *LITA guide to e-scholarship* [working title]. Preprint available at: <http://www.oclc.org/research/publications/archive/2004/dempsey-mslitaguide.pdf>
- Drake, M. A. (2004). "Institutional repositories: hidden treasures." *Searcher*, 12(5), May. Available at: <http://www.infotoday.com/searcher/may04/drake.shtml>
- Gadd, E., Oppenheim, C. & Probets, S. (2003a). "RoMEO studies 1: the impact of copyright ownership on academic author self-archiving." *Journal of Documentation*, 59(3), 243-277.
- Gadd, E., Oppenheim, C. & Probets, S. (2003b). "RoMEO Studies 2: how academics want to protect their open-access research papers." *Journal of Information Science*, 29(5), 333-356.
- Gadd, E., Oppenheim, C. & Probets, S. (2003c). "RoMEO Studies 3: how academics expect to use open-access research papers." *Journal of Library and Information Science*, 35(3), 171-187.
- Gadd, E., Oppenheim, C. & Probets, S. (2003d). "RoMEO Studies 4: an analysis of journal publishers' copyright agreements." *Learned Publishing*, 16(4), 293-308.
- Gadd, E., Oppenheim, C. & Probets, S. (2004a). "RoMEO Studies 5: IPR issues facing OAI data and service providers." *Electronic Library*, 22(2), 121-138.
- Gadd, E., Oppenheim, C. & Probets, S. (2004b). "RoMEO Studies 6: rights metadata for open archiving." *Program*, 38(1), 5-14.
- Giles, J. "Critical comments threaten to open libel floodgate for physics archive." *Nature*, 426, 7.

- Guy, M., Powell, A., & Day, M. (2004). "Improving the quality of metadata in eprint archives." *Ariadne*, 38, January. Available at: <http://www.ariadne.ac.uk/issue38/guy/>
- Halbert, M. (2003). "The MetaScholar Initiative: AmericanSouthOrg and MetaArchive.Org." *Library Hi Tech*, 21(2), 182-198.
- Halbert, M., Kaczmarek, J. & Hagedorn, K. (2003). "Findings from the Mellon Metadata Harvesting Initiative." In: T. Koch & I. T. Sølvsberg, (eds.), *Research and Advanced Technology for Digital Libraries: 7th European Conference, ECDL 2003, Trondheim, Norway, August 17-22, 2003*. Lecture Notes in Computer Science, 2769. Heidelberg: Springer-Verlag, 58-69.
- Harnad, S. (2001). "The self-archiving initiative." *Nature*, 410, 1024-1025. Available at: <http://eprints.ecs.soton.ac.uk/archive/00005947/>
- Hazen, D.C. (1995). "Collection development policies in the information age." *College & Research Libraries*, 56(1), 29-31.
- James, H., Ruusalepp, R., Anderson, S., & Pinfield, S. (2003). *Feasibility and requirements study on preservation of e-prints*. London: Joint Information Systems Committee. Available at: http://www.jisc.ac.uk/index.cfm?name=project_eprints_pres
- Klarreich, E. (2001). "Genetics paper erased from journal over political content." *Nature*, 414, 382.
- Lagoze, C., Van de Sompel, H., Nelson, M., & Warner, S., (eds.). (2002). The Open Archives Protocol for Metadata Harvesting, v. 2.0, 14 June. Available at: <http://www.openarchives.org/OAI/2.0/openarchivesprotocol.htm>
- Law, D. (1991). "The organization of collection management in academic libraries." In: Jenkins, C., & Morley, M. (eds.), *Collection management in libraries*. Aldershot: Gower, 1-20.
- Lynch, C. A. (2003). "Institutional repositories: essential infrastructure for scholarship in the digital age." *ARL Bimonthly Report*, 226. Available at: <http://www.arl.org/newsltr/226/ir.html>
- McClelland, M., McArthur, D., Giersch, S., & Geisler, G. (2002). "Challenges for service providers when importing metadata in digital libraries." *D-Lib Magazine*, 8(4), April. Available at: <http://www.dlib.org/dlib/april02/mcclelland/04mcclelland.html>
- Martin, R. (2003). "ePrints UK: creating a national e-print archive." *Ariadne*, 35, April. Available at: <http://www.ariadne.ac.uk/issue35/martin/>
- Nature Publishing Group. (2003). "Nature in 2003." *Nature*, 421, 1.
- Nichols, S. G., & Smith, A. (2001). *The evidence in hand: report of the Task Force on the Artifact in Library Collections*. Washington, D.C.: Council on Library and Information Resources. Available at: <http://www.clir.org/pubs/abstract/pub103abst.html>
- Pinfield, S., & James, H. (2003). "The digital preservation of e-prints," *D-Lib Magazine*, 9(9), September. Available at: <http://www.dlib.org/dlib/september03/pinfield/09pinfield.html>
- Pinfield, S., Gardner, M. & MacColl, J. (2002). "Setting up an institutional e-print archive" *Ariadne*, 31. Available at: <http://www.ariadne.ac.uk/issue31/eprint-archives/>
- Plutchak, T. S. (2002). "Sands shifting beneath our feet." *Journal of the Medical Library Association*, 90(2), 161-163. Available at: <http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=100760>
- Powell, A., Day, M., & Cliff, P., (2003). *Using Dublin Core to describe eprints*, v 1.2. ePrints UK project. Available at: <http://www.rdn.ac.uk/projects/eprints-uk/docs/simpledc-guidelines/>
- Renardus, (2002). *Guidelines for participating services*. Available at: http://www.renardus.org/about_us/guidelines/

Research Libraries Group. (2002). *Trusted digital repositories: attributes and responsibilities: an RLG-OCLC report*. Mountain View, Calif.: Research Libraries Group. Available at: <http://www.rlg.org/longterm/repositories.pdf>

Snow, R. (1996). "Wasted words: the written collection development policy and the academic library." *Journal of Academic Librarianship*, 22(3), 191-194.

Ward, J. (2003). "A quantitative analysis of unqualified Dublin Core Metadata Element Set usage within data providers registered with the Open Archives Initiative." In: C.C. Marshall, G. Henry & L. Delcambre (eds.), *Proceedings of the 2003 Joint Conference on Digital Libraries (JCDL'03)*. Los Alamitos, Calif.: IEEE Computer Society, 315-317. Also available at: http://www.foar.net/research/mp/wardj_quantitative2.pdf

Wolpert, A. J. (2002). "Institutional repositories: key policies creating an infrastructure for faculty-library partnerships." *Institutional repositories: a workshop on creating an infrastructure for faculty-library partnerships*, Washington, D.C., 18 October 2002. Available at: <http://www.arl.org/IR/wolpert/index.htm>

Acknowledgements

UKOLN is funded by the Museums, Libraries and Archives Council (MLA), the Joint Information Systems Committee (JISC) of the UK higher and further education funding councils, as well as by project funding from JISC, the European Union and other organisations. UKOLN also receives support from the University of Bath, where it is based.