

7 November 2008

Feasibility study into approaches to improve the consistency with which repositories share material

Final report for JISC

Authors of this report:

Andrew Charlesworth	University of Bristol
Nicky Ferguson	Clax Ltd.
Eric Lease Morgan	Infomotions Inc.
Seb Schmoller	schmoller.net
Neil Smith	Knowledge Integration Ltd.
David Zeitlyn	University of Kent

Contact: Nicky Ferguson, jisc-repocon AT clax.co.uk

Reading this report

The focus of this report is consistency of practice between repositories in the ways they support the sharing (and, by implication, the reuse of) material. We acknowledge that repositories may perform several other functions (administrative, preservation etc.) but this report deals with **sharing and reuse of material**. The introduction defines the boundaries of the report in more detail.

Definitions of key terms as they are used in this report (e.g. *consistency, user, policy expression* etc.) appear in Appendix H, at the end of this report.

URLs appearing in this report were last accessed on 30 October 2008.

This report was commissioned by JISC.

Acknowledgements

We are very grateful to all those who gave up time to help us. Vital to this work were the people who responded to our request for interviews, those who contributed and voted on-line, JISC programme staff and everyone else who responded to our enquiries and made their time, expertise, staff and documentation available to us

1 Executive summary

This report is based on consultation and interviews with those involved in the development of repositories and their infrastructure in the UK and elsewhere. There is a wide variety of repository policy and practice. This is not surprising as repositories differ greatly in resourcing, purpose, coverage, sponsorship, philosophy and community involvement.

In the UK, a large number of Institutional Repositories have been set up very recently. Often, it seems, they lack sufficient clarity of policy and purpose. In interviews with depositors and after conducting a case study of an Institutional Repository, we find different perceptions of the role of the repository, some seeing it mainly as an administrative tool for collecting and collating research at the institution and others believing it is a tool for sharing research and creating open access to the results of that research. If such perceptions are combined with weakly defined policies and/or unclear implementation procedures, then it would be unsurprising to find inconsistencies both within and between repositories. In fact, our respondents tell us that such inconsistency is widespread and are pessimistic that this will change, except where sufficient resources, shared objectives and strong relationships are in place.

In the wider environment in which JISC operates, it would be unwise to attempt to mandate specific technical or organisational approaches. This report instead makes more generic recommendations, although it does include some quite specific suggestions on policy and technical directions.

We strongly recommend that repositories, with the help of their target community, clearly define their purpose and share that definition widely so that there is a common view among that community of what the repository is there for and why it is a good thing. A repository's policies on rights, legal issues and licensing, deposit, preservation and collection scope should similarly be clearly defined, stated and shared. In seeking to allocate resources for present and future activities, institutions and funders should differentiate between materials such as full text scholarly articles (which can be efficiently exposed to crawling and indexing from search engines) and materials such as complex learning objects, images, sound and video files which, on their own and without contextual material or metadata, are currently less 'findable' in this way. In view of resource constraints which were often mentioned by our respondents, priority should be given to:

- populating the repository with sufficient high quality material that your target audience will consider it a 'critical mass';
- creating and exposing robust policies;
- creating and maintaining machine interfaces to metadata, indexes and the full text of items held in repositories, particularly scholarly works;
- creating minimal metadata for all items, with richer metadata for those items which cannot be efficiently crawled and indexed; automation should be used wherever possible to aid and supplement human intervention.

Sufficient consistency for worthwhile collaborative work using standards-based technologies and rich human-created metadata may be achieved only where sufficient staff/human resources exist to create, share and maintain appropriate metadata and policies. In addition, such consistency

may only be found where a federation is able to mandate standards, practices and policies or where participants feel they have a strong common interest in the success of shared objectives.

Taking account of the feedback we received, in order to promote more productive sharing between repositories we also make recommendations to:

- increase the use of automated tools to help, and in some cases replace, human metadata creation;
- create and maintain stronger relationships between repository owners/sponsors and also between them and other metadata holders such as publishers;
- move towards embracing Web standards – as opposed to ‘digital library’ standards – in the production and maintenance of repositories and the sharing of their content.

We make nine detailed recommendations (section 10) which are summarised here:

1. Articulate more precise repository policies:
 - a. Rethink policy presentation and the role of layered policies;
 - b. Consistently use symbols and icons to simplify IPR positions;
 - c. Explore existing cross-repository policies.
2. Clarify goals and purpose and realistically assess future role and likely costs.
3. Identify and define policy principles such as:
 - a. The collection scope, purpose and policies on rights, legal issues and licensing, deposit and preservation;
 - b. The development of a clear and comprehensible institutional copyright policy and embedding repository-friendly IPR practices in institutional operations, including a robust risk assessment, mitigation policies and effective procedures for risk amelioration;
 - c. The clear identification of where the repository fits in the institution’s development strategy.
4. Expose all repository content that it would be desirable to share, including material needing subscription or membership, to search engines and web crawlers as one route to discovery. Where content requires some form of authenticated access, the search engine should link to a metadata record describing the content.
5. Produce human readable guidelines for those who will wish to build innovative services from your repository data.
6. Analyse present and future costs and benefits of metadata, for example:
 - a. Prioritise the population of an absolute minimum of key metadata elements (e.g. title, creator, link to object and rights statement) in the first instance. Where possible, tools which allow these elements to be created by machines or at least auto-populated with ‘best guess’ values for review should be used as part of the ingest process;
 - b. Focus on items of a non-textual nature;

- c. Encourage the further augmentation of metadata through automation and collaboration;
 - d. Focus the future development of SWAP on the active involvement of repository developers and administrators
7. Encourage research and development activities in:
- a. Developing tools and plug-ins to assist with the workflow of repository ingest and/or improve consistency of dissemination; tools which can interoperate seamlessly with commonly used repository products such as EPrints, Fedora and DSpace;
 - b. Exploiting established authority files and existing controlled vocabularies and further pursuing the work begun by RIDIR and Names to populate metadata elements;
 - c. Implementing a 'Cloud' approach to metadata provision and sharing
8. Examine the feasibility of funding practical, evidence-based research into how disciplinary and subject cultures affect consistency issues within and between digital repositories.
9. Encourage repository developers and managers to embrace more widespread Web standards such as:
- a. REST-ful computing techniques;
 - b. Increased use of RDF/XML to mark-up content;
 - c. Exploiting HTTP by taking advantage of content-type headers;
 - d. Use of the ATOM Syndication Format and Publishing Protocol (using the SWORD profile).
10. The full recommendations in this report (section 10), which are summarised above, will require collaboration and partnership activities to achieve. Examples of such necessary activity include:
- a. Working closely with UK institutions to encourage them to clarify and develop not only their repository policies but the wider institutional policies on use, access, ownership and IPR.
 - b. Working with other funders such as Becta, the research councils and Wellcome to develop a common approach to the repository infrastructure towards which the ITT for this work aspires.
 - c. Working internationally on these challenges in standardisation and quasi-standards activities and on collaborative research and development projects.

Contents

1	Executive summary	2
2	Introduction.....	7
2.1	The boundaries of the project.....	7
3	Consultation mechanisms.....	7
3.1	Key interviews	7
3.2	User/depositor interviews and case study	7
3.3	IdeaScale – feedback mechanism	8
3.4	Liaison with subject/institutional repository review	8
4	Background to repositories in the UK	9
5	Types of repository.....	9
5.1	Owner/sponsor	9
5.2	Material type.....	10
5.3	Repository goal/purpose	10
5.4	Repository scope	11
5.5	Technical infrastructure	11
6	Areas of consensus	12
6.1	Define purpose and business case.....	12
6.2	Differentiate materials, treat them differently	12
6.3	Third party creation of metadata.....	13
6.3.1	Building significant content is vital	13
6.3.2	Policies and policy exposure.....	13
6.4	Machine interfaces	13
6.5	User/depositor interviews and case study	14
7	Why consistency?.....	15
8	Feasibility, costs and benefits	17
8.1	Materials and metadata.....	17
8.1.1	Types of metadata	17
8.1.2	Depth and breadth	18
8.1.3	Encoding and distribution.....	21
8.1.4	Content aggregation and packaging	23
8.1.5	Possibilities for the future	24
8.2	Policies.....	27

8.2.1	Differentiating between policies	27
8.2.2	Matching policy form to policy purpose.....	28
8.2.3	Policy elementals: ‘What, who, how, where and when’	29
8.2.4	Supporting policy principles	30
8.2.5	Presenting policy elementals: layering, symbols and education.....	32
9	Examples of repositories.....	35
10	Recommendations.....	39
11	Appendix A – Advisory event attendees.....	42
12	Appendix B – List of key interviewees.....	43
13	Appendix C – Summary of feedback received on IdeaScale	44
14	Appendix D – ‘Long list’ of experts for consultation	45
15	Appendix E – Licensing schemes and conditions	46
15.1	Selection of Creative Commons and AShareNet icons.....	46
15.2	Different licence schemes.....	47
15.3	AShare licensing model (simple example).....	49
16	Appendix F – Repository development projects.....	50
17	Appendix G – e-Prints Soton – qualified Dublin Core fields	52
18	Appendix H – Definitions of terms used in the report	52

2 Introduction

The aim of this study, as expressed by the ITT, is to:

assess the feasibility of approaches to improve the consistency with which repositories share material.

It makes the base assumption that:

To the extent that one of the objectives of national and institutional repositories is to share material, then there are advantages to having some consistency in the ways in which that material is exposed for sharing.

The ITT makes clear that the major component of this study should be consultation. In addition to individual interviews by phone, teleconference and face to face, we attended a number of meetings and conferences where we spoke informally to delegates; we also presented at the UK repository infrastructure strategy meeting on 3 July 2008 at the JISC offices at Brettenham House in London. Attendees are listed in Appendix A. This event was very useful as we were able to put questions to the large number of UK-based experts present and an interesting discussion ensued which enabled us to tailor our questions for key interviewees and add further ideas to our web-consultation mechanism. The majority of the people we were intending to contact during this preliminary phase were present at Brettenham House.

2.1 *The boundaries of the project*

The focus of this report is consistency of practice between repositories in the ways they support the sharing (and, by implication, the reuse of) material. It is worth noting that the ITT points out that the long-term curation of material is also likely to benefit from consistency of practice between repositories, but that long-term curation is specifically excluded from the focus of this work. Similarly, repositories may perform other administrative functions such as internal auditing and collection and collation of materials for submission to government and funders, but these functions are not examined in this work; this work focuses, as requested by the ITT on support of **sharing (and, by implication, reuse of) material**.

Our brief includes the consideration of possible repository contents such as research papers and electronic theses (generally available in a format allowing textual analysis or free text search), images (and other materials generally held as binary files without any integral textual information) and learning materials (which may contain both textual components and binary files). Scientific data is out of scope as JISC felt that this area is too complex to be addressed in this way at the moment.

3 Consultation mechanisms

3.1 *Key interviews*

In consultation with the programme manager, we identified a list of potential key interviewees from the UK plus a further list of international experts. We interviewed 29 key figures in total. The list of interviewees appears in Appendix B.

3.2 *User/depositor interviews and case study*

In addition to the interviews specified in our tender document, we fulfilled our brief to obtain user feedback by conducting a case study of depositors and potential depositors at the University of Kent, UK, together with conversations with some of those involved in the process of encouraging deposit and

the creation of content for the institutional repository there. Details appear in Appendix C. A summary follows in section 6.2.

3.3 IdeaScale – feedback mechanism

At the suggestion of JISC, we have used a wiki-type mechanism called IdeaScale to suggest a number of ideas and allow others to do so and to respond. This has the advantage of allowing people to ‘fly kites’ and respond to suggestions or ideas which are not fully worked out but may spark interesting discussion.

The IdeaScale site was used prior to the advisory meetings and interviews to inform the participants of the areas and issues with which we were working. We also used it to add further ideas as they emerged from interviewees and to allow consultation with a further ‘long list’ of various international experts in the field.

The feedback from IdeaScale is summarised in Appendix C. The full text can be accessed online at:

<http://jiscrepository.ideascale.com/akira/ideafactory.do?discussionID=1719>.

The ‘long list’ of experts appears in Appendix D.

3.4 Liaison with subject/institutional repository review

We are grateful to have received co-operation from Catherine Jones and the team conducting a contemporaneous study of the development of subject and institutional repositories – the report is known as SIRIS, The Report of the Subject and Institutional Repositories Interactions Study. This study only focussed on journal articles and did not take a view on many of the contentious issues surrounding open access repositories. Neil Smith and Nicky Ferguson were able to talk to Catherine after the Brettenham House meeting and we also conducted a joint interview with Paul Walk at UKOLN on 7 August 2008. We were also kindly given sight of their draft report.

In general we agree with their conclusions, particularly their recommendation on clear identification of authors, funders and higher education institutions, which points out that:

Being able to locate with authority and consistency the identity of a person or corporate body attached to a research output is vital.

Jones et al (SIRIS report) 2008

One area where we had a very useful exchange of views was on the Dublin Core application profile for scholarly works (SWAP) and whether it should be generally adopted as a common information interchange format. If there is to be a common interchange format between scholarly works repositories which are well resourced enough to undertake to comply with it, then SWAP may well be the best candidate; but we have serious doubts as to whether it will be feasible or cost effective to expect all repositories holding scholarly works to comply. We particularly urge those working on SWAP to involve the repository software developers in their work and to acknowledge the importance of search engine indexing of full text in the arena of text-based scholarly works.

4 Background to repositories in the UK

There has recently been substantial investment in the development of digital repositories in the UK from JISC, but also from other parties including research funders, publishers and individual HE, FE and research institutions. From BECTA through JISC to Wellcome, from schools through universities to advanced research, money has been spent on researching, supporting, writing guidelines for, and facilitating the creation of, repositories. While it is probably fair to say that content in such repositories has not reached a critical mass, it is reasonable to expect that the effects of the investment will continue to be felt for some years and that both the content and the number of repositories will continue to grow.

The aim of JISC's development work in the area of digital repositories is to bring together people and practices from across various domains (research, learning, information services, institutional policy, management and administration, records management, and so on) to ensure the maximum degree of coordination in the development of digital repositories, in terms of their technical and social (including business) aspects. The work funded by JISC relating to repositories aims to create an interoperable network of repositories for the UK higher education community. Ensuring that these repositories address information management issues within organisations and also access requirements across the UK and beyond is essential to realise the JISC mission.

JISC's ITT for this report

5 Types of repository

The scope for this study, as outlined in section 2.1 above, included the full range of digital repositories (with the exception of research data). Although much of the activity within the sector, and within JISC funded projects, has focused on the development of institutional repositories of scholarly works (or of metadata records about scholarly works), other types of repositories are within the scope of this study and need to be taken into account. Possible dimensions for describing repositories include:

5.1 Owner/sponsor

Repository owners/sponsors identified within this study include:

- institutions
- funders
- publishers
- JISC services (e.g. JORUM)
- other services (e.g. arXiv – run by Cornell with NSF funding).

The owner/sponsor of a repository is likely to set rules about how material is ingested into the repository which may have implications for consistency (e.g. are depositors expected to create metadata or is cataloguing carried out or refined by someone other than the depositor?). The owner/sponsor will also set policies about how material from the repository will be disseminated (e.g. RSS feeds, harvesting, indexing by web crawlers).

5.2 *Material type*

Much of the activity in the sector has been driven by the open access movement and the desire to allow free access to scholarly works (mainly journal articles). The main material types being deposited in repositories today are largely text-based. Text-based materials have the potential for post-deposit processing in a variety of ways: indexing by web crawlers, data mining, automatic extraction of metadata etc., which may reduce the importance of human-created 'external' metadata. In practice, though, different text-based formats vary in the extent to which post-deposit processing can be successfully applied. In general, formats which include explicit markup, such as XML and HTML, are better than plain text. Many 'standard' document formats such as .doc, .docx and .odt can be processed more fully than 'display oriented' formats such as pdf. In general, the algorithms used to extract meaning from a corpus of textual content are improving all the time and this is a very active area of research. However, there are few end user tools on the market and many of the algorithms used (especially those used by general purpose search engines such as Google) are highly prized secrets.

However, not all material deposited in repositories is text-based. JISC has funded work looking at metadata application profiles for a range of other material types including:

- images
- time based media (audio and video recordings)
- geospatial data
- elearning materials (which may include text-based materials).

For all of these material types (with the possible exception of elearning materials) the textual analysis is likely to be of limited value. Whilst it is still possible to surface embedded metadata from within the materials, this is likely to be mainly of a technical nature and limited use for the purpose of, e.g., resource discovery. Although web crawlers are increasingly extracting meaning from time-based media, repositories containing non text-based materials will need to rely more heavily on human-created metadata with consequential impact on the need to adopt some policies with respect to consistency. The material type may impact upon the costs and benefits of these policies; elearning repositories, for example, typically contain relatively large, expensive items, so the costs of creating detailed and consistent metadata are small compared with the costs of creating the resource in the first place.

5.3 *Repository goal/purpose*

Another way of classifying repositories is to look at the purpose or goal that the repository has been established to achieve. Possible goals, which overlap, might include:

- improving resource discovery
- preservation
- assisting with auditing and reporting
- cost saving
- enabling depositors to display a consistently structured and asset-rich publication list.

Obviously, repositories set up with a goal of improving resource discovery are more likely to be interested in the way they interact with other repositories and third party services, and hence consistency, than repositories set up mainly for the purpose of preservation.

The purpose of individual repositories is often not clearly defined or, even if it is defined, may be understood differently by different actors/users. In particular, for many institutional repositories a prime strategic motive may be assisting with auditing, reporting and administrative purposes. This may not be the motive of academics submitting their materials, who may be more concerned with improving resource discovery and access from a potential worldwide audience. Library and faculty staff may be very interested in the preservation capabilities. Staff working on the repository project may be aware of a number of these competing views of its purpose and find themselves 'holding the ring' between these views. It is important that repositories clearly define their purpose and priorities as this will inform decisions on the costs and benefits of actions necessary for consistency.

5.4 Repository scope

Repositories can be distinguished by their scope both in terms of geographic coverage (single institution, regional, national, international) and in terms of their subject coverage (single subject, subject group, inter-disciplinary). In practice the scope is unlikely directly to affect the extent to which consistency is an important factor. This is more likely to be determined by linked criteria such as the owner/sponsor's ingest and dissemination policies.

In the UK, in particular, there has been debate about whether institutional repositories or subject repositories are the way forward. This has been examined in more detail in the SIRIS report mentioned above. The point we are making here is whether the repository's scope makes it classifiable as an institutional repository or a subject repository is less likely to affect its demands for or production of consistency than the goal or purpose of the repository.

5.5 Technical infrastructure

The extent to which repository owners and developers can meet requirements for consistency is mainly governed by resources (e.g. budgets and the availability of suitably skilled staff to input data) and by the technical infrastructure upon which the repository is built. In our research we have encountered the following main types of infrastructure:

- off the shelf (commercial) e.g. IntraLibrary
- off the shelf (open source) e.g. EPrints, Fedora, DSpace
- bespoke or home grown.

Our interviews have shown that the majority of repositories are using off the shelf, open source products as the base for their repository, with EPrints being the market leader for institutional repositories of scholarly works. There is evidence that some open source products are more easily customised than others. Even where it is possible to customise the software, many repository owners are reluctant to do so due to lack of technical resources and potential problems encountered when upgrading to a newer version. For this reason, these products are often treated as 'shrink wrapped' and therefore the potential for consistency is limited to support for what is enabled in the chosen product by default.

6 Areas of consensus

We identify here issues on which there was broad agreement between our interviewees and respondents. Not unanimous perhaps, but a strong suggestion of agreement across people from diverse backgrounds.

6.1 Define purpose and business case

It is vital that managers of repositories and their teams are clear about the purpose and business case of their service. Without clarity about their intended purpose, audience, contributors and the coverage of the repository, it will be impossible to make decisions about standards and consistency. This may sound trite, but many institutional repositories do not make it clear to a web visitor what their purpose is and one suspects that this is because it is not completely clear to the staff.

Certainly serving the audience and fulfilling the sponsor's requirements will determine the approach to consistency.

When building a service, start from use cases of what people might want to use it for and use this to build a business case for participation. The costs of metadata generation and of participation in general (including adhering to federation rules on consistency) need to be proportionate to the benefits the publisher will gain from participation.

Interview Respondent (Australia)

Who are the audiences, the users? Answers to this question inform the levels of interoperability required and the features of the objects in the repository.

Bill Moen

Consistency is theoretically completely possible; whether it is reasonable depends entirely on the context and the use case

Interview Respondent (USA)

These views support our opinion that with JISC's highly heterogeneous community (both of users and repositories) it is highly unlikely that mandating consistency of metadata, materials or policies will be sensible or effective. Making tailored recommendations on *approaches* for different repositories and communities, with a limited amount of detail on specific standards or technologies, is more likely to meet with success.

6.2 Differentiate materials, treat them differently

The popularity of search engines, and the sophistication of their indexing procedures, make it a pragmatic essential to treat research papers, theses and other scholarly texts differently from images, time-based media (audio, video etc.) and other binary and 'closed' files. In general, although there are problems with pdf files and other issues, such as author identification and citation, search engines do a good job at indexing textual material and acting as a first step for users, who may go directly to the item level within a repository, or may end up using an individual or federated repository search after locating the initial destination. Publishers have worked directly with Google to make the full text of articles available for indexing even where the articles themselves are only accessible to human users after passing authentication, payment or subscription challenges. This is an area where repositories could learn from publishers. Repositories for learning objects are an area where significant development and standards work has been done. Learning objects are often hybrids in that some of their content is textual and some binary files. In many cases, the value of a learning object (i.e. the paid-for

development time invested in it) may mean that the relative cost of spending time manually creating metadata is small.

6.3 Third party creation of metadata

It is likely that third party creation of metadata will increase as a phenomenon in the future.

There will be a lot of generation of metadata away from the resource, mostly done remotely using automated tools, also possibly by humans who have a real need at that time – so consistency will only be a small part of the problem.

Paul Walk

Metadata is going to be created by independent parties, some of which are algorithmic parties – it will have a massive impact.

Interview Respondent (USA)

If this is the case, then metadata describing your resource (or a copy or near-copy of your resource) may increasingly be out of your control. The richness this extra activity brings to the environment is likely to outweigh the side effect that you may not like, or agree with, some of the remotely created metadata associated with your resource.

6.3.1 Building significant content is vital

Most respondents recognised that many repositories, particularly institutional repositories which have been relatively recently established, do not have a critical mass of content. Our examination of institutional repositories showed a marked disparity of coverage across subject areas. Within one institution there will often be moderately well populated subject areas and virtually empty ones. Between institutions there are marked variations, with the well populated areas at one institution being far less represented at another and vice versa. It is not within our scope to discover all the reasons behind this. Obviously some institutions are stronger in different research areas, but this on its own does not explain the marked (wild) differences. Clearly having one or more repository enthusiasts within the faculty or department will have a noticeable impact. Another factor is the culture of study within particular subjects and disciplines. One interview respondent tells us that early work done by his graduate students indicates that the culture within some disciplines does not (yet?) embrace the digital object.

6.3.2 Policies and policy exposure

There was general agreement that it would be completely unrealistic to expect institutions and repository funders/owners to even attempt to achieve consistency on policies. However, it should be feasible and reasonable for repositories to consistently expose, both to human and machine users, their policies on rights, legal issues and licensing, deposit, preservation and collection scope. In order to do that, of course, repositories need to define these policies and, where the resources exist to make it possible, employing the DRAMBORA toolkit would be one way of checking that they have done this.

6.4 Machine interfaces

Our respondents were nearly all agreed that creating machine interfaces to metadata, indexes and full text of items held in repositories, particularly scholarly works, was a very high priority. Several pointed out to us that the exposure of full text (even where it occurs behind an authentication wall) is very important, but is merely a subset of the general concern that machine interfaces should be made

available wherever possible so that innovative services can be experimented with and built by third parties.

6.5 *User/depositor interviews and case study*

In the summer of 2008 interviews were undertaken with academics contributing to, and with library staff running, the Kent Academic Repository (KAR). Academic staff were not conscious of issues of consistency, either as depositors or searchers. Their concern was whether full text could or should be deposited (KAR allows bare bibliographic entries without full text), and then accessed.

Library staff were more concerned as an immediate issue to get the repository populated. They felt that consistency issues were being dealt with indirectly by the combination of their work in validating data records and by their use of EPrints software (relying on its developers to resolve metadata sharing issues).

Although the repository had been created by library staff enthusiastic about open publication, it was not widely used until the University mandated the creation of records in anticipation of the Research Excellence Framework (REF) and, in order to encourage uptake, mandated that promotion decisions will use publication evidence from it alone. The repository was actively populated during 2008 and different departments used different means to do this. Some departments ported the contents of existing publication databases to it, some let their individual staff members get on with it themselves, and others employed temporary staff to create records.

In such cases there were different regimes of data checking. One department employed two temporary data entry assistants who worked at different times and so could never check notes. Some data entry assistants were trained by library staff, others by departmental officers who had themselves been trained by the library. Some departmental officers checked all records, others left data checking to library staff. The inclusion of full text was not part of the central brief and concerns about permissions were expressed only by library staff.

Library staff were clear that metadata creation by academics was unreliable. Humans imperfectly remember or imperfectly reproduce publication details such as titles and journal names, even their own (for example, there have been cases where academics accidentally transpose keywords in titles). Although it is more efficient to use a Digital Object Identifier (DOI) to bypass having to track down the official website version of an article, even this is not always straightforward: not all journals publish DOI in paper copy (e.g. *Current Anthropology*), and there can be uncertainty about which of multiple ISSNs applies to a particular reference. Author names, journal titles and article titles are vulnerable to American/UK spelling issues as well as mis-rememberings.

There continues to be a gulf between the concerns of the experts and those of academics, or even normal library staff. One senior academic who has been an enthusiast behind the establishment of KAR, and who is actively concerned with bibliometrics and how they might be used in REF, when asked how repository metadata could be improved, responded 'What is metadata?' If a person like this does not know what metadata is, then how much less can we expect others to know, let alone care? This illustrates how little the debates in information service circles actually percolate out to the wider audience of academics.

7 Why consistency?

JISC has a vision of *a content layer of academic and scholarly resources that are freely available, well structured and searchable* and believes this can be achieved by *working across the UK to provide a federated network of digital repositories*. In order to do this it has articulated *a need to ensure institutional repositories in universities and subject repositories provided by funders such as the Wellcome Trust and the Research Councils work to a shared set of standards and practices*.

This study has been briefed to investigate *the extent to which these standards and policies should and could be implemented* with regard to *sharing (and, by implication, reuse) of the material held in a repository*.

Some interviewees questioned the need for any consistency, or warned against making consistency an aim in itself. However, most agreed that consistency within a repository is desirable, and that some consistency between repositories would be useful, but felt that it is unreasonable to expect it to be an overriding concern for repository managers. The first priority for anyone setting up a repository should be to define in detail, and to examine carefully, the purpose of the repository and its target audience. From this process will emerge designs and prototypes. The internal architecture and standards used will usually flow from these design decisions, although if the repository has a stated role as a member of a community or federation of repositories this process may be shortened, as standards and architecture may already have been defined for federation members. Once the purpose of the repository and the user requirements have been analysed the need for consistency can then be considered. It may be that the repository has a proprietary or unique internal structure; if this is the case it does not rule out presenting a consistent machine interface to data and/or metadata using an open standard such as Dublin Core or a semi-open one such as Google's Sitemaps. Indeed, it may not be necessary to immediately present such an interface, simply to demonstrate that it is possible to create one easily at a later stage – and this creation could be done by a third party, if appropriate.

Although the focus of much discussion on consistency between repositories is on metadata, several of our interviewees pointed out that the more general topic of ensuring consistent exposure to machines and networked software includes exposure to indexing by search engines. While the necessity for this to happen across the JISC community has been recognised by many reports (including several by the current authors) it is disappointing to note that there is still much progress to be made in this regard. It is difficult to know whether this is because previous recommendations have been regarded as naïve (*Google can't solve everything you know*) or because it is regarded in some quarters as the domain of commercial web developers. Nevertheless, many of our interviewees were concerned that the JISC community has not addressed key issues which would increase the visibility of resources held in UK repositories.

One of these issues is that the web does not handle two particular aspects of scholarly works very well – citation and author identity. It is suggested that research work is conducted into the use of Microformats, DOIs and a URI bank or unique naming scheme for people to address these issues. Another issue is the enabling of full text indexing of objects behind authentication walls. Publishers holding final copies of journal articles which are only available to subscribers have nevertheless made the full text of those articles indexable by Google. This allows searchers to 'hit' textual references within the full text item, but when they follow the link they are taken to a login/subscription/payment page. This has probably been enabled for large publishers by negotiation with Google – there is no reason why JISC should not conduct such negotiation on behalf of its community.

There are several examples of specific community approaches where consistency has been largely achieved within a relatively narrow community or federation. A number of Australian federations have

taken this approach such as the ARROW project (e-prints, electronic theses, e-research and electronic publishing), its offshoot the MACAR initiative <http://macar.wikidot.com/> which provides recommendations and advice on metadata requirements for digital repositories, LORN (online training resources from across the Australian vocational education and training sector) and TLF (developing digital curriculum content for all Australian and New Zealand schools); these are all described at <http://fred.usq.edu.au/background.html>. Two other development projects, both supported by the European Commission's eContentplus Programme and both involving the Open University and other UK partners, focus on learning object/elearning repositories: the recently started ICOPER project (<http://www.frepa.org/wp/2008/09/04/kicking-off-icoper/>) which mainly considers HE, and the ASPECT project (<http://aspect.eun.org/>), which focuses on the K12 sector. DINI, the Deutsche Initiative für Netzwerkinformation (German Initiative for Network Information), certifies repositories as meeting quality standards.

Together with the DARE guidelines, the DINI certificate serves as a basis for the DRIVER Interoperability Guidelines for Content Providers. Therefore all DINI certified repositories comply with the DRIVER guidelines. DRIVER – Digital Repository Infrastructure Vision for European Research (<http://www.driver-support.eu/index.html>) – is harvesting and providing a common search to their constituency of member repositories. This parallels Intute Repository Search (IRS), although unfortunately it appears that DRIVER prefers to have relationships with individual repositories rather than using the data which IRS has harvested. It would be preferable to have a national aggregation deal with the interaction with organisations such as DRIVER. This would leave institutional repositories free, but only 'free' as long as they offer enough so that national aggregators can make the aggregation conform to the DRIVER guidelines where that is a priority. One important semiformal aspect to DRIVER's working is mentoring – linking those running newly established repositories with those with more practical experience. We note that all Dutch research repositories (universities and institutes) meet the DRIVER standards and that (as of 02/06/08) 21 German repositories have received the DINI certificate. DRIVER partners in UK HE include SHERPA (Nottingham) and UKOLN (Bath). Whether such a federated approach, where a participant gives up some amount of autonomy in return for the benefits of participating with others, would work in the UK is an open question. Certainly, with SHERPA involved in DRIVER and in the UK's Repositories Support Project – RSP (<http://www.sherpa.ac.uk/projects/rsp.html>) – there is good reason to expect that expertise developed and lessons learnt during DRIVER should be rapidly made available to UK institutions. However, we should not jump to the conclusion that an example of a well-funded relatively tight group of institutions is necessarily applicable on a wider scale:

One problem with the work at Intute Repository Search is that there is no standardisation across UK repositories in how they output their metadata. This means it is difficult even to automatically extract the URLs for full text documents where supplied – which we use to enhance search. And the problems with the metadata mean we can't tell if a document is a preprint or a final copy, a scholarly article or an image or a learning object; whether it's peer reviewed or not – without a lot of intensive manual work. Maybe about 10% of the items we see have links to full text – not all actually available outside their institutions – the rest usually just linking to jump-off pages.

Phil Cross – Intute Repository Search project

Our opinion is that the projects discussed above illustrate that sufficient consistency for worthwhile collaborative work using standards-based technologies and rich human-created metadata may be achieved only where sufficient staff/human resources exist to create, share and maintain appropriate metadata and policies and either where a federation is able to mandate standards, practices and policies or where participants feel they have a strong common interest in the success of shared objectives.

8 Feasibility, costs and benefits

The biggest cost is that asking people to create metadata is a barrier to deposit. How would that impact on scholarly communication and how would you measure it?

Andy Powell

We have so far defined the area and looked at areas of consensus which we found in our consultation. We now move to the more contentious questions of how, when, where and in what circumstances consistency is important, desirable or affordable? This section is divided into two parts, the first dealing with consistency issues in materials and their metadata and the second dealing with consistency in the production and presentation of policies on matters such as rights, ownership, scope and archiving.

8.1 Materials and metadata

Throughout our interviews, the respondents emphasised that there exists a spectrum of metadata types, and the feasibility of consistent metadata across these types varies widely. These metadata types, generally speaking, fall into three categories: 1) metadata of fact, 2) metadata of judgement, and 3) authority lists. Another dimension of metadata is its depth and breadth. To what degree is it expected to be exhaustive or merely accurate? A third aspect of metadata surrounds its encoding and distribution. How is it to be manifested? Fourthly how important is the aggregation and packaging of the content? Finally, there are possibilities for the future. This section discusses feasibility in the context of each of these different aspects of metadata.

8.1.1 Types of metadata

Metadata of fact is analogous to descriptive cataloguing in traditional library work. For the most part this includes determining things such as the size of an item, its date of publication and the publisher. This type of metadata is usually verifiable. In the existing repository environment, these metadata elements correspond to things like the size in bytes of a particular file, the date the file was last saved to a file system, and the location (URL) of an item. Each of these characteristics is a fact – a piece of information that is binary (true or false) in nature. We believe it is highly feasible to expect consistency regarding metadata of fact because of their binary nature and their verifiability through the use of automation.

Factual metadata has the potential to be generated by non-human agents. For example, the size of file in bytes or the length of an audio recording can easily be determined in this way. Other metadata elements such as ‘title’, ‘author’, ‘publisher’ or ‘date of publication’ may be specifically marked up as ‘embedded metadata’ within an item. If so, they are also amenable to machine generation. Even when such elements are not specifically marked up as embedded metadata, textual analysis techniques may be used to extract such data with increasing levels of accuracy. Although ‘shrink wrapped’ tools for repository owners do not exist right now, this technique certainly has potential to ease the workload of metadata creation and is discussed in more detail in section 8.1.5 below.

At the other end of the spectrum is metadata of judgement. In traditional cataloguing practice this is often called analytic cataloguing; it involves the description of an item’s ‘aboutness’ – the assigning of controlled vocabulary terms or classification numbers. This sort of metadata is highly subjective and very expensive to create since it usually requires subject and/or metadata experts (cataloguers). Moreover, it is extremely dependent on the intended audience of the metadata. One person who is describing an item for a learning objects repository may assign one set of terms, and another person describing the same item for a subject repository may describe it another way. Same item. Different experts. Different audience. Different metadata values. Given the expense and nature of this type of content, we believe it

is not feasible to expect consistency with regard to metadata of judgement, except perhaps where it occurs in a tightly controlled, narrow and consistent environment such as a database of drug trials.

In between factual and judgmental metadata lies authority information – names of people, places and titles. In these cases the metadata is often binary (true or false) in nature, but represented ambiguously. For example, the author of a scholarly work may be denoted in any of the following forms:

- * Fredrick Kilgour
- * Kilgour, Fred (1914–2006)
- * F Kilgour
- * Kilgour, F.

In each case the application of the name may be correct, but the implementation of the name is inconsistent. The feasibility of consistency in this regard lies between the feasibility for metadata of fact and metadata of judgement. The feasibility increases with the availability of complete authority lists, time and money, and easy-to-use tools for data entry including computer-generated or aided data entry.

8.1.2 Depth and breadth

In an analogue environment it is necessary to create metadata surrogates to describe information resources, and in order to facilitate the greatest amount of discovery these surrogates are expected to be both deep and broad; in other words, to describe the items in as many ways as possible in both their characteristics of fact as well as judgement.

When considering a digital environment where information resources are largely textual in nature, we learned that depth and breadth are not as necessary as previously thought because full text indexing supplements much of the discovery process.

[In the past,] surrogates played a very important role, but as more full-text is available they are less important. [...] Users want to make the determination of value themselves, based on the full-text.

Jeremy Frumkin

Frumkin gave an example from usability studies against his local metasearch implementation. Students queried the application, got back results, and quickly read the full-text instead of looking at each record's details. Students used the full-text to measure relevance, not the metadata.

Regarding depth, breadth, and full-text indexing:

Depth and breadth of the metadata records are related to the amount of time and money available. But even so, we have to get past hand-crafted records. A metadata bottleneck exists that better tools for metadata creation and automatic metadata generation can help resolve.

Bill Moen

Moen used the work done by Elizabeth D. Liddy (Syracuse University) in automatic metadata generation as one example of resolving the metadata bottleneck. He pointed to other examples such as tools to extract words like 'Normandy' and 'D-Day' from a document and then assign broader subject terms such as World War II.

Eighty percent (80%) or ninety percent (90%) accuracy in metadata element values when automatically generated might be good enough. [...] Certainly, computers can do this [metadata creation] more quickly than humans, and maybe at a good enough standard of completion and accuracy.

Bill Moen

In the UK, influential library figures agree:

The manual addition of metadata is just not viable for the creation of substantial repositories – any strategy which relies on manual creation of metadata in libraries or by repository staff is likely to encounter a very low level of commitment from institutions. We could do A LOT better than we are doing at the moment, using automation, authority file URIs and RDF. However, to date, the library community has not done enough to drive initiatives in these areas, and seems to be very slow to move in this direction.

Owen Stephens

One interview respondent put it more bluntly:

Discussing and taking the time to store bibliographic information is almost a waste of time. [...] Do the Google thing. [...] The effort spent creating bibliographic metadata is largely wasted.

These statements do not mean the creation of deep and broad metadata is not important. In fact it is especially important for non-textual information such as images, audio and video. Moreover, repositories exist for a variety of purposes as well as for a variety of audiences. Nick Weideman and Nigel Ward highlighted that user interfaces based on faceted searching and on interface metaphors such as timelines or maps, which are useful for some material types, require explicit metadata such as temporal or geospatial information which is difficult to extract via full text indexing.

For all these reasons it may be important to spend the time and energy creating ‘hand crafted’ metadata records. In view of the ongoing costs of such activity, however, the benefits should be clearly analysed and set out in a statement of rationale before that decision is taken.

Again, from Bill Moen:

I believe in multiple views of the repository, and repositories have multiple partners.

He elaborated by describing his work on a repository containing learning objects. His local audience will have free access to the underlying materials, but people outside his community will have a different view of the repository and they will also have to pay to access its underlying content. At the same time, he plans to make the metadata describing his content as widely available as possible. He mentioned DLIST as a discipline-specific repository. He compared that to his local repository of learning objects which is a more specific repository:

They each have different audiences and purposes.

This apparent dichotomy between types of metadata and the depth/breadth of metadata can be illustrated through the use of a matrix, such as the one below, where each quadrant represents a different level of feasibility:

	brief records	deep/broad records
factual		feasible w/
metadata	feasible	full-text +
		computers
judgement	traditional	too labour
metadata	library	intensive for
	work	most applications

Given these types of metadata and the range of desired amounts of metadata, we can summarise our conclusions here:

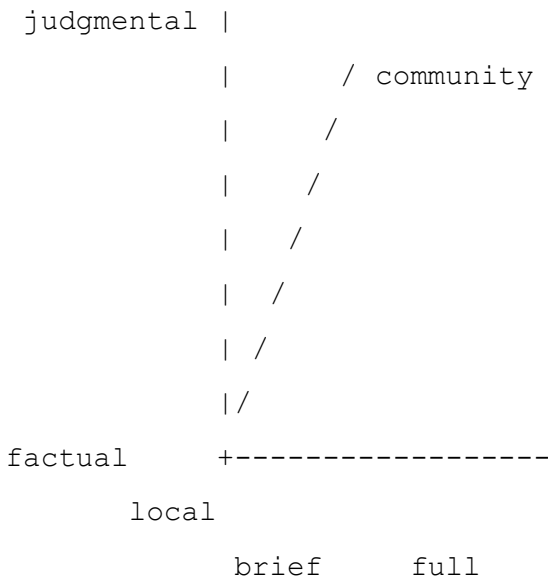
- ✚ factual brief records are feasible with our current methodologies;
- ✚ factual fuller records are feasible, according to our experts, if computers were exploited to a greater degree;
- ✚ brief judgemental records are the domain of humans (and maybe computers) but the content will often be thin and inconsistent;
- ✚ fuller and judgemental records are too labour intensive for all but clearly defined situations where ongoing costs have been accurately estimated and benefits are considered to outweigh them; and computer technology is not yet at a stage to replace human effort in this regard.

When you add yet a third dimension to the matrix – communities – then this casts more doubt on the feasibility of expecting the creation and maintenance of consistent metadata across a wide range of repositories. For example, when creating metadata Martha Yee suggests aiming

at following the standard of the field, but when the needs of the local audience override the standard, then deviate accordingly but only in a reversible way such that a computer algorithm could be designed to put the data back into standard form.

Unfortunately, many (if not most) institutions do not have the necessary expertise or resources to do this adequately. She gave an example. Where the rule states, ‘Enter the title in the language in which it was created’, but your audience does not read the language, then she suggests entering the title in the language of the users and supplementing the record with the ‘foreign language’ title.

So, when all three characteristics of metadata creation are combined, the feasibility of consistent metadata decreases. Consider the following three-dimensional space:



It contains three axes: factual–judgement, local–community, brief–full. If you were to map metadata implementations using these three dimensions, then we assert the closer the mapping is to the origin, the more feasible it is to create consistent metadata. In other words, if the metadata is factual, brief (minimal) and local in nature, then the feasibility of consistency is high. On the other hand, if the metadata is judgemental, full (complete) and designed for a wider variety of communities, then the feasibility of consistent metadata is low.

8.1.3 Encoding and distribution

The types of metadata and the intended uses of metadata are two aspects of consistency. A third aspect is the way metadata is encoded. This encoding is directly related to technical standards, and for the most part those technical standards have revolved around the use of XML bindings of metadata schemes based on Dublin Core (or, in the case of learning objects, IEE LOM) and distributed using OAI-PMH.

Few, if any, of our interviewees had qualms with the philosophy of the Dublin Core Metadata Initiative (DCMI), but there was concern regarding the application of Dublin Core.

Yes, for specific repository applications, there may be deficiencies with the Dublin Core. Addressing specific purposes and content of specific repositories may then require elements from metadata schemes as well as the Dublin Core. Dublin Core, however, can provide a basis for interoperability even with elements from other schemes to meet the needs of specific repositories.

Bill Moen

For example, learning objects are going to have one set of elements and bibliographic objects will have another. At the same time the metadata in all repositories needs to support the following to some degree or another: levels of interoperability, description/features of the objects, user tasks and target audience.

Moreover, it is widely known that the types of values people place into the various parts of the Dublin Core Metadata Element Set vary widely. Probably the best example is with the identifier element. To what degree is the value expected to be a database key or a URL? The creator element is almost as problematic. What is a creator and how is it different from an author, an artist or an editor? Yes, the specification mentions these things but the application is inconsistent.

Which set of Dublin Core metadata elements? the original 11? the DCMI extended set? It's a baseline for sharing across subject boundaries and institutional boundaries only to the extent that those subjects and institutions share agreement as to what the metadata elements mean, and how they should be used. Deciding whether it's a sufficient baseline is more difficult: sufficient for what? not a lot, I suspect.

Lou Burnard

In recent years DCMI has developed an abstract model in an attempt to try to clarify some of these concerns (and also to align DC more closely with RDF and emerging Semantic Web technologies). It has also developed a methodology for developing Dublin Core Application Profiles (DCAPs) which are not predicated on the DC Metadata Element Set at all. It is these guidelines that the JISC projects tasked with developing application profiles for key material types, such as the scholarly works application (SWAP), have followed, using a domain model based on the concepts of 'work', 'expression' and 'manifestation' borrowed from the Functional Requirements for Bibliographic Records (FRBR) model (see <http://www.loc.gov/cds/downloads/FRBR.PDF> for a quick introduction).

The profiles developed are significantly more complex than the 'simple' Dublin Core format (oai_dc) mandated as the minimal level required for interoperability in the Open Archives Initiative Protocol for Metadata Harvesting (OAI-PMH). In the case of SWAP, in particular, two of its original proponents feel this complexity will militate against its widespread adoption:

I don't see it taking off in any real way. [...] application developers won't be able to get their heads round it.

Andy Powell

For SWAP to work we need two things: 1) really good implementation in the repository systems (unlikely) and 2) really good maintenance of metadata at an early stage in the life of the resource (more unlikely), so this may be something of an uphill struggle.

Paul Walk

The almost universal use of XML for encoding metadata, irrespective of the precise details of the metadata scheme or application profile being used, is in itself an enabler of interoperability. The use of XSLT enables content to be transformed from one XML schema to another. Transformation of metadata between different application profiles of Dublin Core is therefore relatively simple so long as mandatory data elements are present in the source data and that vocabularies used are the same or can be easily mapped. It is also possible to use XSLT to transform metadata between completely different base schemas, e.g. from IEEE LOM to DC. In practice, most repository management systems use XSLT extensively as the mechanism for both supporting a range of different metadata formats (e.g. in OAI-PMH) and for transforming XML into human readable HTML in user interfaces.

OAI-PMH is the de facto standard for sharing metadata across repositories of all kinds. While thousands of OAI-PMH data repositories exist and the implementation of such a repository is considered a 'low barrier', the consistent population and maintenance of such a repository is by no means as trivial as once thought and the numbers of such content-rich repositories are too few for commercial players like Google to support a special sitemaps extension relating to OAI-PMH.

Take, for example, the problems surrounding the implementation of OAI-PMH service providers. They are expected to harvest content from data repositories and provide services against the aggregation. Their work is challenged by invalid XML, sparsely populated metadata records and ambiguous metadata values. One interviewee said,

How much time have you spent normalizing dates and reversing inverted names?

From the prize-winning paper given at the 2006 Joint Conference on Digital Libraries:

Over the last three years the NSDL CI team has learned that a seemingly modest architecture based on metadata harvesting is surprisingly difficult to manage in a large-scale implementation. The administrative difficulties result from a combination of provider difficulties with OAI-PMH and Dublin Core, the complexities in consistent handling of multiple metadata feeds over a large number of iterations, and the limitations of metadata quality remediation.
Lagoze, C., et al., 'Metadata aggregation and 'automated digital libraries': A retrospective on the NSDL experience' <<http://arxiv.org/pdf/cs.DL/0601125>>

Often the content of OAI-PMH data repositories exist in the 'hidden' or 'deep' Web, and such things – almost by definition – are not indexed by the 'Big Three' indexers (Google, Yahoo and Microsoft). From the conclusion of a recent article in DLib Magazine:

Google's indexing does not seem to have retrieved more of the hidden web since the publication of the McCown et al. article in 2006. We would venture to conclude that Google has not endeavoured to increase their support and access to OAI materials. Even taking into account the caveats, we would also conclude that aggregations of OAI records are as valuable for user research purposes as they were at least two years ago [...] It is also interesting to note that Google has recently dropped support of OAI for website indexing. Given the resulting numbers from our investigation, it seems that Google needs to do much more to gather hidden resources, not less. (Granted, the OAI for Sitemaps feature may not have been an appropriate approach for Google.)

Hagedorn, K., Santelli, J., 'Google Still Not Indexing Hidden Web URLs', DLib Magazine 14 7/8 (July/August 2008) <<http://www.dlib.org/dlib/july08/hagedorn/07hagedorn.html>>

Given this state of affairs, Dublin Core and OAI-PMH do not yet seem to be making it substantially easier for repository items to be discovered by a mass audience. The time and effort spent on their implementation does not seem to have returned the investment at this stage. For text-based items, focussing on exposure to the 'Big Three' seems to be more cost-effective for discovery purposes.

I don't believe we are capable of building those kind of services [...] I think our approach to date has been to assume we can control the provision of the content and the services that are going to make that content available and I don't think there is much evidence that we do either of those very well, but certainly not the provision of the services, very well. So our thinking needs to be more mainstream rather than less [...] and when we talk about consistency we need to be thinking about consistency in the context of much wider initiatives than just our repositories: that's the only way that we are going to get any significant buy in.

Andy Powell

8.1.4 Content aggregation and packaging

The current drive to establish repositories is based on a world view where the location of content is regarded as a primary concern. Placing an item within an institutional repository reinforces the institution's role in the stewardship of this content. Where content consists of several separate digital objects (e.g. a structured piece of elearning broken up into several modules, an mp3 and realmedia version of an audio recording or a pdf and ms word version of a journal article) these are traditionally 'packaged' together using a domain specific format (e.g. DIDL or SCORM).

As we become used to a world in which related pieces of content are in fact widely distributed across a networked environment, so the focus on packaging things together to be stored locally begins to shift. The Open Archives Initiative draft standard on Object Reuse and Exchange (OAI-ORE) takes a network-

centric view of content aggregation. Based on the RDF syntax and notation, OAI-ORE resource graphs are agnostic about the physical location of the various components (they are simply referenced by URIs) and can thus be used to draw a line around virtual 'packages' that are distributed across a network. As with all things to do with the semantic web, OAI-ORE is deceptively simple to describe but not necessarily easy to implement. In his blog posting 'What the OAI-ORE protocol can do for you' (<http://ptsefton.com/2008/10/14/what-the-oai-ore-protocol-can-do-for-you.htm>), Pete Sefton from USQ describes some promising early work in using OAI-ORE, for example use in conjunction with the SWORD deposit protocol to point to the location of items to be ingested; however, this is still a way off influencing the mainstream. If repository owners find it difficult to get fields like dc.title and dc.creator populated consistently or accurately, the chances of describing even more complex relationship types within the current regime for deposit remain low.

8.1.5 Possibilities for the future

The sections above summarise the present situation and some of the problems encountered with an aspiration towards consistency between repositories, namely: 1) there are different types of metadata used for different purposes, 2) breadth, depth and community strain the feasibility of consistent metadata, and 3) technical encoding standards are difficult to implement. Four threads emerged through the course of our investigations on how to address these issues: 1) the development of better as well as automated metadata creation tools, 2) the establishment of stronger relationships with other information providers, 3) the greater exploitation of Web standards as opposed to digital library standards, and 4) the use of URIs to point to the values of metadata elements.

1. Automated tools

One solution to the lack of resources for human creation of metadata is to implement better and automated tools for metadata creation and maintenance. Such tools might query a person for a name and return the authorised form of the name from an authority list. The same technique can be applied to keywords and authorised controlled vocabulary terms. Another example is a centralised metadata creator. Go to site. Complete form. Get back metadata in one or more flavours. Paste the result somewhere else. This process is akin to the data-entry process employed by the Archives Hub at the University of Liverpool, and apparently by Calames, another archive of EAD finding aids. (See <http://tinyurl.com/5yyg77>, pages 18-19.)

2. Stronger relationships

It is important to remember that metadata is created not only by authors and librarians. It is also created by publishers. One way of applying consistent metadata to information objects would be for the library profession to consider creating stronger relationships with publishers (commercial or open access). UKPMC, which is building such relationships, is described in more detail in section 9. Consider this scenario: Author writes article. Author submits it to a 'gold' or 'green' open access journal. Article gets published and now has unique identifier. Library/repository manager can then search for the article in the publisher's archive and download the metadata, especially if a positive relationship has been previously created. We acknowledge that this approach, if implemented, requires contact with numerous publishers for all disciplines. We also acknowledge there are complicated 'claim back' procedures involved. At the same time we believe the creation of stronger relationships should at least be investigated because publishers are also metadata creators.

3. Embracing Web standards

Several of our respondents pointed to the flexibility and enhanced potential which would be achieved by embracing Web standards to a greater degree and relying less on digital library standards. Examples of this would include:

- 1) the use of Google sitemaps comprised of simply constructed URLs;
- 2) using REST-ful computing techniques such as ATOM to expose content;
- 3) figuring out ways to express metadata through RDF.

The creation of Google sitemaps is relatively easy. Given that many repositories run atop relational databases, write reports against the databases, output lists of simple URLs in the form of sitemaps and expose the sitemaps in robots.txt files. This solution facilitates greater discovery of content and takes advantage of existing infrastructure.

The advent of blogs has given rise to the RSS as a distribution mechanism. This, in turn, has given rise to a plethora of RSS readers and aggregators. Even the OpenSearch query protocol returns RSS as its output. RSS has been expanded and matured into the ATOM Syndication Format and the ATOM Publishing Protocol. Both RSS and ATOM rely on simple HTTP GET requests – REST-ful computing techniques. While OAI-PMH is REST-ful, the returned content is domain-specific. By combining HTTP GET requests with HTTP content-type headers, smart user-agents can be created. Request a URL, and if no content-type is specified, then assume the user-agent is a human and desires HTML output. If a content-type is specified (ATOM, MODS, JSON, RDF/XML, etc.), then return the content and allow the user-agent to process it accordingly. By using REST-ful computing techniques such as the ones increasingly used by the balance of the Web community the possibility of repository content being exposed will also increase.

4. RDF, URIs and the Metadata Cloud

The fourth possible solution was best articulated by two interviewees who live and work on the opposite sides of the automation spectrum. One knows very little about computers (Martha Yee), and the other makes their living as a computer scientist (Herbert Van de Sompel).

The heart of the idea lies in the use of RDF and URIs – not string literals – to denote metadata values. RDF is about creating things called triples. Compare them to tiny sentences. Subject. Predicate. Object. The subject is a thing or information resource (web page, image, pdf document, etc.). The predicate is often equated with 'is a', 'has a' or 'contains'. The object is some sort of value. Here is a rudimentary example:

Subject	Predicate	Object
-----	-----	-----
JISC	has an address	http://www.jisc.ac.uk/
JISC	is a	URI:98701
JISC	has a subject	URI:98756
URI:98701	is an	institution
URI:98756	is the subject	librarianship

The URIs – ‘actionable’ URLs – are the key feature in this model. Since the inconsistency of metadata often lies in how the metadata is expressed (Kilgour, Fred (1914-2006) versus F Kilgour), the use of URIs, like the use of relational database keys, removes the ambiguity. As far as possible, we should associate each metadata value with a URI, allowing people or computers to create these associations. Yes, many URIs will be created but we believe the ‘wisdom of the crowd’ will eventually minimise the number of URLs used to denote similar values. There may be a role for librarians in this regard, to amalgamate or edit the similar values and thus reduce ambiguity.

Van de Sompel elaborated in a number of ways. First he advocated for a tiered approach to the creation of metadata. At the lowest tier information providers would be expected to expose minimal metadata including only a URI/URL, a rights statement and a MIME type. The URI/URL would be used to get the item. The rights statement would be used to describe what people could do with the item. The MIME type would help computers process the item. Considering HTTP servers, these metadata elements are almost trivial to expose since the URL is usually a file name, a rights statement could blanket an entire site, and the MIME type is based on site-wide HTTP configurations.

The next level of the tier of metadata includes some sort of aggregation. All of these items are similar to all of those items. This aggregation includes sets of content all having similar characteristics. To provide this aggregation, Van de Sompel advocates the use of OAI-ORE, which defines a data model for ‘resource maps’ and embraces widely used Web standards such as REST-ful computing techniques and the ATOM syndication format.

The uppermost tier of metadata functions to support the Semantic Web and increasingly includes the metadata values of judgement. This content denotes the ‘aboutness’ of items and represents the sort of metadata creation aspired to by the library profession. Since this tier of metadata is increasingly subjective, Van de Sompel suggests that metadata objects be signed and dated with a URI of the metadata author. For example, suppose music-related metadata elements were assigned to a musical score and signed and dated by Sir Paul McCartney through a URI. One could then assume those metadata elements were more true than if any of the authors of this report were to assign those metadata elements. Since the ‘signature’ includes a date, and since we understand that knowledge changes over time, these signatures can be given different weights considering the time and place they were applied.

This brings us to the Metadata Cloud. Metadata descriptions of objects do not need to be verbose. They can include single RDF statements like ‘This report has an identifier of URI:5673.’ These statements could then be stored in one or more centrally located ‘Metadata Clouds’. Read/write access to these Clouds could be granted to a wide variety of audiences. People or computers could update and add metadata to information objects through the information objects’ URI thus creating relationships between metadata and information objects. Such an approach to metadata creation takes into account the types of metadata (factual to judgemental), supports authority lists (names and subjects), gives credence to subject expertise and, finally, exploits the ‘wisdom of the crowds’ and the ‘network effect’ through the application of ‘many eyeballs’. To some degree, the Web-based name authority services supported by OCLC are a fledgling implementation of this Metadata Cloud idea.

The important work done by the JISC funded Names and RIDIR projects towards the provision of URIs for people and objects respectively would feed into future work in this area.

8.2 Policies

[...] there needs to be a meta-metadata standard. Repositories need to be able to explain their practices to each other in mutually comprehensible terms.

Lou Burnard (via IdeaScale)

8.2.1 Differentiating between policies

When considering the issue of institutional repository policies, it needs to be understood that there will usually be several layers of policy in operation, e.g.:

- institutional policy (or lack thereof) on the allocation of ownership of intellectual property rights within the institution (*legal issues*);
- external policies on the appropriate use of intellectual property rights that have been generated within an institution, but then allocated to third parties, such as publisher intellectual property agreements (*legal issues*);
- institutional policy (or lack thereof) on the role of the repository, such as to facilitate or augment internal institutional management of the RAE process (*collection scope, function*);
- repository operational policy (or lack thereof) on the allocation of responsibility for particular tasks, such as the addition or editing of item metadata (*roles, function, preservation, legal issues*);
- repository operational policy (or lack thereof), such as on the types of material that will be accepted for deposit, the timescales for which materials will be held (*collection scope, function, legal issues, deposit, preservation*);
- repository practice policy (or lack thereof), such as who can deposit items, who can access items and for what purposes (*collection scope, legal issues and licensing, deposit, access, interoperability*).

All of these policies, which may be written or simply observed through custom and practice, will interact to form the institutional repository's policy environment, and to determine its content.

It should be clear that for interoperability purposes it is largely irrelevant whether institutional repositories follow consistent practices in certain areas. To a service seeking to access the institutional repository to harvest metadata, or to an end user seeking to identify and access an item within that repository, it makes no difference whether University A permits its academics to retain all copyright in works they create in the course of their employment, and University B claims all copyright in such works as its own. Equally, neither type of user (service or end-user) will care overmuch about what the internal institutional rationale is for establishing and maintaining an institutional repository in the first place.

Such policies may certainly affect what eventually gets placed in the institutional repository, e.g. is the function of the institutional repository to collect all academic outputs, or a sample (self-selected, or institutionally selected)? To provide a management tool for identification of material suitable for submission for the RAE, or to collect a more eclectic range of outputs demonstrating the vibrancy of the institution's research and teaching culture? It would also be helpful if FE/HE institutions were to finally develop either a consistent sectoral approach to copyright in academic work (whether teaching materials, research data or research outputs), or even to support the establishment of coherent internal policies. In the end, however, to the user these matters are of no concern.

External policies, such as publisher intellectual property agreements, clearly have a direct impact upon deposit, access and use. However, users are rarely concerned with precise details of such policies – they

do not care that Publisher X permits only the deposit of preprint articles; that Publisher Y permits deposit of published articles six months after publication; and that Publisher Z places no restrictions on the deposit and reuse of published material. What users (depositors and accessors) usually want to know is:

- what items can be deposited (e.g. unpublished preprint, final peer-reviewed draft, published version) and are they identified as such? (*value/weight*)
- can specific items/their metadata be accessed and by whom? (*accessibility*)
- what can be done with specific items/their metadata in particular contexts? (*usability*)
- are specific items going to be available in the future? (*viability*)

Certain institutional repository operational policy matters are also of little interest to the broad class of users. They may be interested in a repository's metadata to some degree, e.g. some services may be interested in a repository's use of standards/formats such as DC, OAI-PMH, DIDL, but the vast majority of end-users, if they are interested in metadata at all, are more likely to care about the degree of access to, and the ability to reuse, the metadata contained in the repository.

8.2.2 Matching policy form to policy purpose

Publishing formal policies on IPR, preservation, collection, etc. is of use to service builders, but not to end-users, who will mostly only care about simple, clear statements of a repository's intent.

Jim Downing

This statement needs some unpacking. It is not saying that end-users do not want, or need, information about IPRs, preservation and collection etc. Rather it is saying that seeking to present such information to end-users in the form of a 'formal' (and usually lengthy) policy document is to misunderstand the requirements of both the repository and the end-users. The repository is seeking to ensure that its policy goals can be achieved as effectively as possible *over time*: end-users are seeking a concise and layperson-comprehensible statement of whether the repository and its content can be used for the purpose they *currently* have in mind. A 'formal' policy document whose terms end-users are unwilling to spend time assimilating, or simply don't understand, will do nothing to achieve the repository's longer term goals. Equally, it will increase the possibility that end-users will reject using the repository and its content because they are unclear of the rationale and rules for doing so, or that they will use the repository and its content in inappropriate ways.

Even services that are seeking to utilise institutional repositories and their contents will have differing requirements as to the policy information they require from those repositories. Some services may be risk averse, and thus require a high level of detail and precision; others may be content with a similar level of detail as is provided to end-users. Those parties seeking to create services from repository-based information are likely, initially at least, to be looking for simple human-readable information on the policies, formats and metadata used by repositories.

It is arguable that in circumstances where repositories are only accepting input from institutions, only a 'formal' policy document is required:

Nothing is to be gained by exceeding the level of detail and precision that is required by the funding agency and/or legal counsel. Adding more simply worded policies might be indicated where the 'general public' is contributing or editing objects and/or metadata; if only large institutions are participating, this is probably not necessary.

Karen Calhoun

This is not an uncommon point of view, but it rather assumes that even large institutions have long-term in-house expertise sufficient to interpret such 'formal' policy documents, and/or that they have sufficient long-term in-house experience to know when to seek guidance. This reflects a tendency amongst practitioners in the repository environment to assume a higher level of knowledge and interest amongst individuals and institutions about such matters than is probably the case.

The amount of work/cost involved in putting into place a system which can express a repository's policies via a number of layers of complexity (see below) is likely to be minimal compared to the benefits accrued in terms of repository interoperability, development of third party services and end-user understanding.

The type of information, derived from an institutional repository's policy environment, which it would be most useful for the repository to 'expose' to the outside world to facilitate:

- cross-repository interaction;
- services built on top of networks of institutional repositories; and
- end-user access to, and where possible reuse of, items within the repository

can thus perhaps be described in its simplest form as conforming to a 'what, who, how, where and when' test.

8.2.3 Policy elements: 'What, who, how, where and when'

Terminology is important, as part of communication [...] it is about talking about what we do, not the tools we use. In a sense I'm saying let's not call a spade, a spade, but 'a way of digging your garden'.

O. Stephens (via IdeaScale)

It was noticeable, from the various consultation exercises undertaken during the preparation of this report, that respondents appeared generally disinclined to spend much time discussing the issues of policies and IPRs. There is always the temptation at such times to put this lack of engagement down to the failure of technologists and creative/academic types to pay attention to serious practical legal and administrative questions, preferring as they do to disappear off in different directions chasing their own idealised vision of the future.

However, on closer examination, what appears is often essentially a pragmatic acceptance of the particular environment in which the respondent is embedded. It is thus perhaps less surprising that those operating in, or with experience of, close knit federations of repositories, or in well financed multi-institution projects, may adopt a more sanguine approach to consistency of policies across repositories.

What is important is strong service-level agreements between the federation and the participants. You need to know what is required to participate and remain active. To be sustainable, the federation needs to set expectations on the members.

Dan Rehak

It is neither reasonable nor realistic to expect consistency in how individual institutions publish or otherwise communicate policies. Policies that directly concern the cross-institution service in question should apply to all parties and should be communicated clearly and in a central location.

Karen Calhoun

There are undoubtedly examples of close knit federations of repositories (e.g. the Australian Learning Federation) or well-financed multi-institution projects (e.g. the German DINI project and Dutch

DARE/NARCIS project) that can establish quite detailed consistency between their member repositories' policies. In the rather more chaotic patchwork environment that characterises the heterogeneous, huge (and often fractious) JISC constituency, it is difficult to see the same approach scaling effectively to address the differences between, for example, teaching and research repositories, repositories covering different academic disciplines, and repositories in academic institutions with widely varying incomes. (This does not mean that lessons cannot be drawn from those examples, e.g. the DINI certification process.)

Several key points are often reprised as elemental principles of a consistent repository policy:

- What is (or is not) contained in a repository? (*collection scope*)
- Who can deposit items? (*deposit*)
- Who can access elements (items and metadata) in the repository? (*access*)
- How can they use those elements? (*legal issues, licensing*)
- Where can those elements be used? (*legal issues, environment*)
- When (or for how long) will those elements be there? (*stability, preservation*)

8.2.4 Supporting policy principles

For a consistent cross-repository approach to the policy principles outlined above, individual repositories will need to consider how they approach underlying layers of policy in several areas:

Collection scope

Collection scope will be determined largely by the institutional policy that the repository was designed to facilitate. It can thus be described in part by the role it serves, e.g. 'an institutional or departmental repository', and in part by what the collection contains, e.g. 'all eligible materials submitted by staff for the Research Assessment Exercise', or by what it does not contain, e.g. 'holds all types of materials except: theses and dissertations'. Repositories can thus enhance their visibility and impact by adopting a consistent approach to indicating both their institutional role and the nature of the items in their collection.

Deposit

The question of who can deposit will depend in part on the institutional role of the repository, thus, for example, deposit may be restricted to members of an institution or, more narrowly still, to authors who are a member of the institution. Often the extent to which various classes of depositor are permitted to deposit will be determined by the repository's attitude towards risk, particularly the risk of infringement of intellectual property. In certain cases deposit of particular items with a repository may be compulsory, e.g. as a consequence of acceptance of support from a research funding body.

A consistent policy approach to key deposit issues across repositories would be useful, e.g. defining who can deposit in a given repository; stating the role of the repository in validating and authenticating deposits (if any); disclosing whether or not access to deposits can/will be embargoed for a period of time; identifying where liability for copyright infringement lies; and making clear what happens to items that are subject to infringement claims.

Access

For many repositories, access to items/metadata is intended to be open to the public. Here it is important to ensure that policies consistently and accurately differentiate between 'access' and 'use'. Some repositories have access restrictions, in that only certain groups (e.g. members of an institution, or group of institutions) can access the repository at all, or can access full text items where these are

available. Such access restrictions should be clearly identified for the benefit of users. This would allow informed decisions about deposit, e.g. a depositor might be obliged to deposit in a restricted access institutional repository, but could also choose to then deposit the same item in an open access repository, thus ensuring wider visibility; and informed decisions about access, e.g. a service can decide whether to include a restricted access repository.

Legal issues

For users, a clear understanding of the legal issues relating to their use of materials from a particular repository is often key to their willingness to utilise it, and to their effective and non-infringing use of items and metadata from it. Presenting a clear policy statement of those issues in an institutional repository can be greatly facilitated by clarifying underlying policy on allocation of ownership of intellectual property rights within an institution, and in dealings with third parties, such as publishers. This then permits an institutional repository to make definitive statements about staff contributions, including where the responsibility lies for establishing which IPRs apply to particular deposited items, and the effect those IPRs will have on access to, and use of, those items.

Dealing consistently and effectively with the IPR issues raised by a repository, both at start-up and during operation, will require the clear allocation of responsibility for those determining and addressing those issues within the repository's management team. That responsibility will span both the deposit and access functions of the repository, as changes to the IPR policy on the one side will almost inevitably have repercussions on the other. It is important that processes are in place to ensure that risk management is an ongoing issue, and that responsibility for undertaking such assessment, as well as developing and administering methods of handling any risks identified, is clearly located within the staffing structure of the repository. Effective IPR risk management is key to establishing and maintaining both depositor and user trust in the reliability of a repository. In short, building a flexible copyright/IPR policy framework, based on an initial background and risk assessment, which contains clear and documented processes for deposit and access management, policy and process audit and risk amelioration, and which incorporates the ability to effect coherent change management in the light of shifts in environmental factors, will be essential for long-term sustainability.

In general, an effective underlying institutional/organisational policy on the copyright status of items that may be placed in a particular repository will, in turn, make the construction of consistent policy across repositories and services a simpler task.

Preservation

While preservation (in the sense of ensuring the longevity of items in a repository) has generally attracted little attention from researchers to date, it is likely that this will become an important issue in the future. While no repository is in a position to guarantee that an item will be available indefinitely, it is clear that repository policy on duration of holdings will play a role in where depositors choose to deposit, the type of services that can be viably built upon a repository/network of repositories, and the decision of end-users to use/cite particular resources. As a result, organisations/institutions will need to consider carefully how they intend to:

- ensure the long-term viability of repositories;
- respond to readability and accessibility issues, e.g. by format migration over time; and
- address the issue of withdrawal of particular types of item, e.g. items that breach copyright or other laws.

If there has been significant cross-repository interaction over time, or the construction of services across repositories, the 'knock-on' effect of such preservation issues may result in significant disruptions or

inconsistencies. Developing a consistent preservation policy across repositories will require organisations/institutions to consider a range of issues at both the organisational/institutional and repository levels, including:

- where repository development fits in long-term organisational/institutional financial and developmental strategies;
- contingency plans in the event of withdrawal of repository services, such as measures necessary to permit the transfer of items to another appropriate repository;
- technical issues concerning the effective migration of items, such that there is minimum disruption of cross-repository interaction/service provision; and
- the appropriate form of ‘tombstoning’ for withdrawn resources.

8.2.5 Presenting policy elementals: layering, symbols and education

The promotion of consistency of practice between repositories requires the identification of ‘policy elementals’ – policy principles which, even if consistent practice cannot be reached between repositories, can still provide a common basis for repository owners, service providers and users to understand how those repository practices interrelate. Having suggested a set of policy elementals above, the next question is how to present them as simple human-readable information policies for users and ideally also as machine-readable interfaces.

Layering

The first example draws upon a technique increasingly used in data protection circles for privacy policies, and also by the Creative Commons for its licence scheme. The concept of policy ‘layering’ envisages a set of policy explanations which cover the same policy principles, but which are targeted at particular sub-sets of audience, in terms of technicality of explanation, length of explanation and often both.

In data protection terms, privacy policies are often aimed at lay readers, interested parties and experts. The lay readers are assumed to be simply interested in a very basic explanation of what the policy means for them. The interested parties are assumed to want to know more about the wider implications of the policy and how it affects them in detail. The experts are assumed to want chapter and verse on the precise nature of the policy and how it relates to the Data Protection Act 1998. See, for example, *The Center for Information Policy Leadership, Ten steps to develop a multilayered privacy notice* (http://www.hunton.com/files/tbl_s47details%5Cfileupload265%5C1405%5Cten_steps_whitepaper.pdf).

The layered licences used by the Creative Commons take a slightly different approach. When you create a licence using the Creative Commons licence generator, the program produces three versions of the licence:

- *Commons Deed*. A plain-language summary of the licence, complete with the relevant icons.
- *Legal Code*. The fine print that you need to be sure the licence will stand up in court.
- *Digital Code*. A machine-readable translation of the licence that helps search engines and other applications identify your work by its terms of use.

Creative Commons, License Your Work (<http://creativecommons.org/about/license/>)

A layered policy, if both accurate and well designed, is thus an effective way of communicating an appropriate level of information to a particular audience. A layered policy for repositories might consist of:

- a high-level legal document for repositories, services and institutions;
- a high-level technical document for repositories, services and institutions;
- a mid-level document for managers, administrators and technical support staff combining elements of the first two documents;
- a low-level document for general use explaining, in basic terms, the policy elements.

An example of what the last document might look like can be seen by reviewing the output of the OpenDOAR policy tool:

<http://www.opendoar.org/tools/en/policies.php>

Symbols/icons

With regard to the copyright conditions attaching to use of works held in a repository, the recent BECTA report *Development of Good Practice Guidelines for Repository Owners* (2008) makes specific and detailed recommendations (pages 51-65) about how repositories should deal with the question of intellectual property rights, noting in particular that:

‘Guideline IP5: Repository owners should ensure that the user interface for applying and managing licensing conditions is simple to use. Use of a standard set of icons to indicate key licence terms could be considered.

Rationale: The simpler the user interface to applying and managing licensing conditions is made, the more likely it is that licensors and licensees will understand and internalise the basic copyright requirements.

Commentary: *The CC has been active in driving the use of icons to indicate to licensors and licensees, quickly and effectively, the essential elements of their licences. The use of multi-level licence explanations (the human, machine and lawyer-readable licences) allows licensors and licensees to operate within their legal comfort zone, or to learn more about the CC licensing process if they wish. While it is clear that this has not always been successful (Linksvayer 2007), it seems to be a satisfactory approach for many licensors and licensees. Ideally, repository owners should aim to utilise existing icon sets where possible, to avoid icon proliferation and resulting user confusion.*

Examples of existing practice: The Creative Commons uses icons in combination to indicate the key elements of their licences. AShareNet also uses a set of icons to indicate the key elements of their licences.’

(Italics added)

An important issue in ensuring consistency across repositories is likely to be how they deal with intellectual property rights both at the top level and/or at the item level. While it may be possible for an institution to have a blanket policy level IPR approach:

‘2) Copies of full items generally can be:

- reproduced, displayed or performed, and given to third parties in any format or medium
- for personal research or study, educational, or not-for-profit purposes without prior permission or charge.

provided:

- the authors, title and full bibliographic details are given

- a hyperlink and/or URL are given for the original metadata page
 - the content is not changed in any way
- 3) Full items must not be harvested by robots except transiently for full-text indexing or citation analysis
- 4) Full items must not be sold commercially in any format or medium without formal permission of the copyright holders.'

<http://eprints.nottingham.ac.uk/policies.html>

It is likely that many repositories will have 'mixed' collections. This could be handled in a number of ways, e.g. rather than attempting to have an individual rights statement for each item a repository could state a default copyright position with a menu of alternatives, e.g.

default: If the item is not otherwise marked, you can use and reuse this with appropriate acknowledgement

if marked A: you can use and quote this in full only with appropriate acknowledgement

if marked B: you can read this and point to it but it may not be reproduced without further permission from [publisher etc.]

From the point of view of achieving consistency across repositories, it would be highly desirable if repositories were to seek, firstly, to attempt to keep different rights statements to a minimum, and secondly, where different rights statements are required, to use a harmonised set of symbols/icons to indicate how particular items can be used. Appendix E gives detailed examples.

Education

As was discussed at length in the JISC report *Sharing eLearning Content – a synthesis and commentary* (2007), if there is to be effective cross-repository interaction/provision of services then it is imperative that:

- educational institutions adopt a coherent (and ideally uniform) approach to institutional copyright policies;
- those policies should seek to maximise repository-friendly practices, such as the use of Open Access and Creative Commons licensing, wherever this is appropriate within an institution's business strategy (i.e. Open Access should be the rule rather than the exception); and
- these initiatives should be backed by a coherent and workflow-oriented programme of staff education in the basics of copyright, the mechanics of open access publishing and role of repositories in promoting the dissemination of research and teaching outputs.

9 Examples of repositories

One of the tasks in our brief was to ask our respondents about, and briefly summarise, examples of good practice. In truth, our interviewees almost all struggled to find examples of good practice (apart from the projects they themselves were involved in). The only repository mentioned more than once in this context was arXiv. The following examples are a cross section of repositories we encountered, all good examples in one facet or another but not necessarily examples of good practice throughout.

arXiv

arXiv <http://arxiv.org> is the premier example of a subject repository. It is a collection of preprints from the areas of physics, mathematics, nonlinear science, computer science, quantitative biology and statistics. It was started by Paul Ginsparg at the Los Alamos National Laboratory in 1991. It is now hosted at Cornell University and mirrored all over the world. The repository includes just over half a million items and grows at a rate of about 4,000 items per month.

The acquisition process is straightforward. Registration is open to all. A registered author grants arXiv the non-exclusive right to redistribute their work. They enter title and abstract, select some controlled vocabulary terms in order to classify their submission, then upload their document. The submission is reviewed and usually accepted into the body of work in the repository.

arXiv is very well used by depositors and popular and well-respected amongst end-users. It has been operating for 17 years, has accumulated a critical mass of content and had some sort of first mover advantage. It is also likely that the researchers in these subjects, especially physicists and computer scientists, were open to an innovative and rapid way of sharing research results, came to appreciate the advantages that the technology had to offer and were already very familiar with using technical solutions in their work. Moreover, it directly replaced a vital element of physicists' research culture: a paper system of preprinting which demanded that, for example, 20 years ago the Library at the Rutherford Appleton lab¹ catalogued over 200 technical reports and preprints a week – this is now all done through arXiv. It is common for researchers to use arXiv as the first place to learn about and acquire article literature in their field. arXiv seems to fit very well the needs of these particular communities. There is scope for work to discover whether (and if not, why not) this model might be replicated in other subject cultures.

e-Prints Soton

e-Prints Soton <http://eprints.soton.ac.uk/> is the University of Southampton's Research Repository. It is based on EPrints <http://www.eprints.org/>, an open source repository system that originated at the University and, with Fedora and DSpace, is now one of the three most used open source repository systems. The aim of e-Prints Soton is 'to provide a permanent record of the research output of the University and maximise the visibility, usage and impact of this research through global access'.

In Autumn 2008, e-Prints Soton contained approximately 35,000 records, of which 12% were full text or full item records, and the remainder metadata only records, such as records pointing to the URLs of books or other assets that publisher's rules prevent being hosted in the repository, or to open access versions of an item that the University is not confident can be ingested safely.

Any employee of the University (whether or not an academic) and any student on a higher postgraduate course or above can put content into e-Prints Soton. There is wide variation in the upload rate between different disciplines. Factors affecting the upload rate include: the existence of a strong culture of self-

¹ Thanks to Catherine Jones for this insight.

deposition; the existence of competitor subject repositories (for example arXiv in the case of physics and physical chemistry); the extent to which a discipline (such as medicine) is wedded to the importance of publisher final versions. In the most committed areas, the overwhelming majority of academic staff are depositing item descriptions² and perhaps one third of staff are depositing items.

The University encourages usage of e-Prints Soton in a range of ways including: chasing for items proper from those who have deposited item descriptions; active advocacy of use by staff of the repository; desk-side support for users; email support to users.

Items and item descriptions in e-Prints Soton are exposed to Google, and very well crawled by it, with the possible exception of pdf files in particularly 'locked down' formats. 97% of e-Prints Soton's users arrive from a Google search result. Browsing or searching e-Prints Soton by users is relatively infrequent. A number of services run by others make use of e-Prints Soton. For example, Ethos, when it is fully functioning will harvest theses, and Thompson's Current Contents takes data from e-Prints Soton.

The University encourages consistency in several ways. Firstly, depositors are required to provide a limited set of metadata, which varies by item type (see Appendix G for details). Secondly, every item is subsequently quality assured by e-Prints Soton staff to ensure, when possible, the inclusion of DOIs, ISBN and ISSN, volume and part, pagination, and that event dates are correct for conferences/exhibitions. Thirdly, there is interoperation between e-Prints Soton and one of the main available systems through which staff create their home pages, so that a professional looking and comprehensive citation list can be generated on the staff member's home page³. In particular, this encourages staff to deposit items in e-Prints Soton – it may well be that the time spent on deposit procedures is less than the time it would take to hand-craft a publications list, a powerful incentive. Finally, a detailed internal policy document covering the 'house rules' for depositing items and writing item descriptions is being prepared for use, in particular, in internal training.

Jorum

Jorum <http://www.jorum.ac.uk/> is a UK repository for teaching and learning material. Jorum can be used instead of, or as well as, institutional repositories. It was conceived as a sort of over-arching repository for use by a limited number of institutionally mandated staff, who would upload appropriate materials to Jorum. Individual users (teachers and support staff) could then download specific materials for use in a learning and teaching context and deliver them to students through the institution's VLE or Learning Management System. It seems that this mode of operation may no longer be feasible (i.e. tolerated by users) in the era of widespread sharing of (many sorts of) content, so although originally it required a demanding copyright licensing agreement for all content it hosted, now Jorum is also accepting Creative Commons and other licences. Jorum staff highlight third-party content as a problem in that it may not be obvious to individual academics that third-party content is included in materials that are being adapted or transformed. Institutions also need to assess the level of risk involved in making content available. One role of Jorum has been to raise the issue of medium-term custodianship of learning materials. Jorum staff includes professional cataloguers who complete the full metadata record for resources when they are added.

² e-Prints Soton was used by the University in preparation for the 2008 Research Assessment Exercise, so research active staff were in effect required by the University to use it.

³ Here are three examples of home page citation lists that draw on material in e-Prints Soton:
<http://www.soton.ac.uk/ses/people/staff/StarinkMJ.html>
<http://www.soton.ac.uk/math/people/profiles/pure/bk2.html>
<http://www.soton.ac.uk/philosophy/staff/whiting.html>

UK PubMed Central

UK PubMed Central <http://ukpmc.ac.uk/> is the UK wing of the US-led PMC. UKPMC provides a repository for a consortium of UK health research funders led by the Wellcome Trust who pushed OA publishing of their funded results in 2005. Wellcome promote the model of paying the publisher for a final copy, stored in a central repository (in this case UKPMC) with all rights to use and reuse the full text of the article for non-commercial use. Wellcome pay \$3,000 to the publishers for each article and in return publishers put the final version into the repository, deposit the XML/sgml version along with high-resolution images, ensure that the full text is available for use and reuse for non-commercial purposes and attach a licence which explicitly allows this providing that correct attribution is made. While use of a licence is mandated, the licence itself is not specified. Examples of acceptable licences include the CC-BY-NC (used by Oxford Open, Springer's Open Choice, Biophysical Society etc.) and Elsevier's Sponsored Documents licence: see <http://www.elsevier.com/wps/find/authorsview.authors/supplementalterms>. The rationale for this is pragmatic – some publishers, e.g. Elsevier, object to Creative Commons; as long as the community is allowed to use material and to reuse with attribution, then there's no point in telling publishers what licence to use. Every PMC record will have a licence summary linking back either to a CC licence or the publisher's site to enable the user to see the licence terms. To the authors Wellcome say 'You can put your article anywhere else but you MUST deposit it centrally with us. If you deposit it with us then we'll convert your pdf or Word into a richly marked up XML document, in a much more preservation-friendly format.' If it was submitted by the publisher then end-users automatically get the final version (with the publisher's high quality metadata, authoritative titles of journal author etc.).

They are happy to scale this up to 100% of the articles written from Wellcome grants, reckoning that their total spend on this would be 1 or 2% of their research budget. Robert Kiley of Wellcome quotes an RIN study claiming that this would save the sector £600,000 per year, savings roughly equally split between libraries and publishers. Kiley also believes that this kind of consistency, with inbuilt deduplication and as much full text as possible, is highly desirable if third parties are to build services upon repositories. The metadata is made available using OAI-PMH. To facilitate deposit they have set up the UK Manuscript Submission System (UKMSS) <https://ukmss.mimas.ac.uk/>. UKPMC is currently working with National Centre for Text Mining (NaCTeM) to explore how text mining can be used to extract keywords from the full text in the repository to further enhance the metadata. As yet only 5% of PMC's 1.5 million articles are open access, and although Wellcome have made it compulsory for all authors in receipt of their funding to deposit in UKPMC, by June 2008 they were achieving 30% compliance.

Other funder repositories: The Research Councils are now establishing their own subject specific repositories for research outputs. These differ critically from arXiv in that they only include research funded by the relevant research council (in the case of UKPMC a consortium including two research councils, four medical charities and two government bodies). This gives them an administrative value but a researcher looking for work in their field is usually not concerned about the funding basis of the material, only in the results. However, coverage is not as patchy as one might think, with NORA entries forming the core of the Intute Repository coverage of the field, and UKPMC, since it includes most of the UK funders in life science and is cross linked to PMC, having excellent coverage. Examples include:

STFC ePublications Archive ePubs <http://epubs.stfc.ac.uk>

ePubs has collected the scientific and technical outputs of the Daresbury and Rutherford Appleton Labs of the STFC since May 2004. The Science and Technology Facilities Council (STFC) was formed in 2007, by the merger of the Council for the Central Laboratory of the Research Councils (CCLRC) and the Particle Physics and Astronomy Council (PPARC). The ePublications aims to collect information on the

academic output of the Council, from both authors and facility users. It has been created using an Oracle database, Cocoon and XML, and is OAI-PMH compliant.

NERC Open Research Archive NORA <http://nora.nerc.ac.uk/>

A NERC funded researcher creates records (they are prompted to provide un-controlled keywords and also to assign NORA subject headings which reflect NERC categories) and uploads the pre-publication full text which is then checked by NORA staff before it is made accessible.

The ESRC runs **Society Today** <http://www.esrcsocietytoday.ac.uk/>, which includes research outputs, as well as, for research data itself, the ESDS (the Economic and Social Data Service) <http://esds.ac.uk/>. (In effect the ESDS is a research data repository, although it predates the idea of repositories.) There is a clear difference between the approach of NORA and Society Today. NORA is a standalone repository built using EPrints software. The ESRC has chosen to integrate their repository into the larger ESRC website Society Today, which is a portal site, using the proprietary Autonomy search technology, aiming to present social science research (including the contents of their repository) to a wide audience. ESRC funded researchers like NERC ones are mandated to deposit research outputs. Building on ESDS experience, Society Today repository records use a controlled vocabulary based on HASSET, the social science thesaurus, in conjunction with author supplied keywords. Society Today staff check records before they are published.

Examples of some key repository support and development projects are briefly described in Appendix F.

10 Recommendations

Recommendations 1-5 are aimed at repositories and their institutions/organisations.

JISC should encourage them to:

1. **Articulate more precise repository policies.** Current repository policies appear poorly developed, poorly directed and unlikely to sustain consistent cross-repository policymaking. There are examples of potentially valuable tools/good practice, such as DRAMBORA and OpenDOAR, but the policy environment is typified by a lack of innovation and a failure to take effective advantage of lessons from existing policy initiatives, or from other areas of practice. Several existing developments could be usefully further examined as having potential to enhance existing practice:
 - a. rethink both the rather pedestrian design of repository policy documents, and the possible role of 'layered' policies in enhancing consistent policy practices between repositories: creating simple human-readable information policies for users (services, end-users) is as important as creating machine-readable interfaces;
 - b. consider effective ways to utilise practices such as the use of symbols/icons to provide repository users with a simplified guide to IPR issues. At present, use of such methods is patchy and inconsistent; failure to provide direction to this kind of initiative risks permanently diluting the advantages of rapid recognition of licence terms currently found with CC licences;
 - c. explore the ways several initiatives in Europe (DINI, DRIVER II) have produced cross-repository policies; it is surprising, in context, how rarely such initiatives were mentioned by interviewees and respondents during the empirical phase of this report.
2. **Clarify their goals and purposes at an early stage and then realistically assess their likely running cost budget before committing to consistency strategies involving high levels of human intervention.** They should understand that it is likely that only a substantial body of content will bring attention and visitors to their repository and should adjust the amount of human intervention required per item accordingly. Thus in general it would be preferable for a repository to hold and expose a large volume of high quality content, with a minimum of human-created metadata, than for it to hold a small volume of content, with a large amount of human-created metadata. Exceptions, such as a repository with content of unusual value or critical importance, would need justification on a case by case basis.
3. **Identify the policy principles that have the potential to form the basis for more consistent policy development between repositories.** In order to take advantage of those principles, organisations/institutions should examine closely the impact upon the policy environment that repositories (both institutional and external) are operating within and take active steps (for example: defining a clear institutional role for repositories, clarifying institutional copyright policy etc.) to ensure that their actions do not directly, or indirectly, impede this development. The main policy roles that need to be addressed are:
 - a. to determine the purpose and scope of institutional repositories: the current *ad hoc* approach taken by many institutions to their institutional repositories is unhelpful in building a directed/consistent policy approach;
 - b. to take a firm position on institutional copyright policies within an institution and in dealings with third parties, such as publishers, to provide effective measures to embed repository-friendly IPR practices in institutional operations and ensure that institutions have clear and

documented processes for handling IPRs in the form of deposit and access management, policy and process audit and effective procedures for risk amelioration;

- c. to identify where repository development fits in long-term organisational/institutional financial and developmental strategies, so that an appropriate long-term preservation policy can be adopted.
4. **Expose all repository content that it would be desirable to share, including material needing subscription or membership, to search engines and web crawlers as one route to discovery.** Where content requires some form of authenticated access, the search engine should link to a metadata record describing the content. Rare exceptions to this would clearly be necessary for security reasons, but the default position should be exposure – either of full text or of a suitable metadata record.
5. **Produce human readable guidelines.** In addition to the essential requirements of creating and maintaining machine interfaces to their data, repositories should appreciate that those building innovative services on top of their repository data will require human-readable guidelines explaining the technical bases, configuration and management of the repository as well as the generic and individual rights policies attached to items.
6. **Analyse the present and future costs and benefits of metadata.** When considering allocating resources to metadata creation for repository items, institutions, JISC and other funders should:
 - a. prioritise the population of an absolute minimum of key metadata elements (e.g. title, creator, link to object and rights statement) in the first instance. Where possible, tools which allow these elements to be created by machines or at least auto-populated with ‘best guess’ values for review should be used as part of the ingest process;
 - b. ensure that sufficient resources are allocated so that richer metadata is in place for items of a non-textual nature, such as images, audio and video, which are not easily susceptible to current tools for analysis of content;
 - c. encourage the further augmentation of metadata, either at the time of ingest or at a later stage; for textual scholarly works, where the future budget allows and where specific benefits of human-created metadata have been identified, priority should be given to forming strong inter-institutional relationships between repositories, including common policies and profiles;
 - d. focus the future development of SWAP on the active involvement of developers and administrators of repository software to ensure that the profile is practical in both software and cataloguing terms.
 - e. promote the need, in all circumstances, for specific individual analysis of the costs and the subsequent benefits of human-created metadata before making medium- or long-term commitments.

Recommendations 7 to 10 are aimed at JISC’s future funding directions.

JISC should:

7. **Research and development.** Encourage research and development activities to explore, experiment and foster automated means for metadata creation. In particular:
 - a. use the example of the SWORD project to fund the development of tools and plug-ins which can interoperate seamlessly with commonly used repository products such as EPrints, Fedora and DSpace. Such tools could address issues such as improving the workflow of repository ingest by building on the existing work of document analysis (text based and other), and

improving repository dissemination by facilitating the creation of XSLT mappings to allow metadata to be exposed in more flexible ways;

- b. exploit the existence of established authority files and existing controlled vocabularies to populate metadata elements with globally unique identifiers in URI format incorporating and pursuing recent initial work funded by JISC (e.g. Names and RIDIR) and elsewhere on object identifiers and person identifiers;
 - c. implement a Web Services-based 'Metadata Cloud in the Sky' application – an amorphous collection of RDF triples – allowing for the establishment of relationships between information objects through the use of such URIs.
8. **Disciplinary and subject cultures.** Examine the feasibility of funding practical, evidence-based research into how disciplinary and subject cultures affect consistency issues within and between digital repositories. In particular, a joint feasibility study, with a research funder such as the ESRC, would make it clear whether the Wellcome model (of the 'gold route' to open access and a central authority, negotiating with publishers to achieve consistent material presentation and consistent metadata) would be feasible and acceptable in another discipline community. Examine the consistency benefits of other possible collaborations and metadata sharing initiatives with commercial and open access publishers.
9. **Embrace Web standards.** Encourage repository developers and managers to move away from the present model to embrace more Web standards. This would include, but not be limited to, work to:
- a. employ more REST-ful computing techniques for exposing digital library content;
 - b. increase the use of RDF/XML to mark-up content;
 - c. resist interfaces to repository content that require numerous name/value pairs as a part of an HTTP GET request;
 - d. exploit HTTP and the dissemination of content by taking advantage of content-type headers;
 - e. Use of the ATOM Syndication Format and Publishing Protocol (using the SWORD profile).
10. **Collaboration and partnership activities.** In order to achieve the above recommendations, necessary collaborative activities will include:
- a. working closely with UK institutions to encourage them to clarify and develop not only their repository policies but the wider institutional policies on use, access, ownership and IPR;
 - b. working with other funders such as Becta, the research councils and Wellcome to develop a common approach to developing the repository infrastructure towards which the ITT for this work aspires;
 - c. working internationally on these challenges in standardisation and quasi-standards activities and on collaborative research and development projects.

11 Appendix A – Advisory event attendees

Event	UK repository infrastructure strategy meeting 03/07/2008		
Place	JISC offices at Brettenham House, London		
Attendees	Sheila	Anderson	King's College London
	Theo	Andrew	EDINA
	Ann	Apps	Mimas, The University of Manchester
	Kevin	Ashley	ULCC
	Chris	Awre	University of Hull
	Rachel	Bruce	JISC
	Lorna	Campbell	CETIS
	Leslie	Carr	University of Southampton
	Santy	Chumbe	Institute for Computer Based Learning (ICBL), Heriot Watt University
	Cormac	Connolly	ESRC
	Jim	Downing	University of Cambridge
	Tom	Franklin	Franklin Consulting / JISC
	Neil	Grindley	JISC
	Rachel	Heery	Consultant
	Amanda	Hill	Mimas
	Steve	Hitchcock	University of Southampton
	Philip	Hunter	IRIScotland (EUL)
	Neil	Jacobs	JISC
	Richard	Jones	HP Labs
	Catherine	Jones	Science and Technology Facilities Council
	Stuart	Lewis	Aberystwyth University
	Vic	Lyte	Mimas, The University of Manchester
	Andy	McGregor	JISC
	Peter	Millington	SHERPA, University of Nottingham
	Balviar	Notay	JISC
	Christopher	Pressler	University of Nottingham
	Owen	Stephens	Imperial College London
	Amber	Thomas	JISC

12 Appendix B – List of key interviewees

Ann Apps	Research and Development: Digital Library and JISC Information Environment Services and Standards, Mimas, The University of Manchester
Chris Awre	Integration Architect, e-Services Integration Group, University of Hull
Kerry Blinco	Consultant, Department of Education, Employment and Workplace Relations, Australia
Karen Calhoun, John Chapman and four colleagues	OCLC
Cormac Connolly	ESRC, Society Today editor
Karen Coyle	Library consultant (ex California Digital Library), USA
Phil Cross	Intute Repository Search, ILRT, University of Bristol
Lorcan Dempsey	Vice President and Chief Strategist, OCLC
Jim Downing	Repository Software Architect, Unilever Centre for Molecular Science Informatics, Cambridge University
Jeremy Frumkin	Gray Chair for Innovative Library Services at Oregon State University
Rachel Heery	Consultant (ex UKOLN)
Bill Hubbard	SHERPA Manager, University of Nottingham
Robert Kiley	Wellcome Institute
Carl Lagoze	Senior researcher, Information Science Program, Cornell University
William Moen	Associate Professor, School of Library and Information Sciences, University of North Texas
Andy Powell	Head of Development at the Eduserv Foundation
Dan Rehak	Consultant for information architecture and learning technologies
Owen Stephens	Imperial College, London
Herbert Van de Sompel	Team leader, Digital Library Research and Prototyping Team Research Library of the Los Alamos National Laboratory
Paul Walk	Technical Manager and Team Leader, UKOLN, University of Bath
Nigel Ward	Consultant, Department of Education, Employment and Workplace Relations, Australia
Nick Weideman	CTO, Curriculum Corporation, Australia
Wendy White	Manager of the Southampton University Institutional Repository
Martha Yee	Cataloging Supervisor, UCLA Film and Television Archive

13 Appendix C – Summary of feedback received on IdeaScale

We used a JISC implementation of IdeaScale (an online feedback and voting mechanism) to garner input beyond our interviewed experts. Specifically, we pointed people towards the *consistency* category of the repositories' IdeaScale. Based on the votes and comments found there we can see some trends.

First and foremost, participants thought consistency was not an end in itself. Consistency for its own sake was seen as limiting and impractical:

focusing purely on current service offerings and requirements limits the development of novel services, and is an approach that usually results in focusing on a single solution, which limits the ability to innovate in service delivery.

And the difficulties of achieving consistency on a wide scale were noted:

There are 200 universities in the UK, and perhaps 20,000 worldwide, then there are subject repositories, project repositories, library and archive repositories and commercial repositories [...] To expect any form of consistency of language, of policy, of metadata, of standards, even of legal scope will simply not work.

Consistency was seen as a means to an end, the end being specific service creation:

The value-added service that I can believe in that requires consistency, is the creation of 'virtual subject repositories' by linking across actual institutional repositories.

and:

I would dream of a consistent service, e.g. auto classification to be overlaid at a higher level, e.g. national, to enable the valuable 'virtual subject repositories'.

The idea that Dublin Core is a sufficient baseline for distributing and re-distributing metadata was heavily voted down. We suspect that votes against this idea included those at one end of the spectrum who felt that Dublin Core was not expressive enough and at the other end those who felt that there ought to be a much smaller minimum number of exposed metadata elements.

Third, while there was an acknowledgement that inconsistency exists, users of IdeaScale felt it should be managed rather than embraced:

Inconsistency is a fact of life, and any repository instance or system that wants to avoid bottlenecks is going to have to accept items that have inconsistent metadata [...] That doesn't mean you have to settle for it, though. It's possible to take a progressive approach, where messy metadata comes in, and is then brought into consistency with particular standards.

Some Ideascale respondents felt that there are feasible and worthwhile approaches to improve consistency, e.g.:

Yes, promote good Atom/RSS feeds, and get rid of barriers to deep linking to metadata.

14 Appendix D – ‘Long list’ of experts for consultation

The ‘long list’ of people we did not interview but contacted by email and invited comments by email or via IdeaScale:

Charles W. Bailey Jr.	David Bigwood	Marshall Breeding
Lou Burnard	William Denton	Diana Marcum
Jill Emery	Cindy Hepfer	Thomas Krichel
Natalia Lyandres	Kitty Marschall	Peter Murray
John Mark Ockerbloom	Andrew Pace	Terry Reese
Robert Sanderson	Roy Tennant	David Williamson
Ed Summers	Ian Witten	Peter Binkley
John Blyberg	Jennifer Bowen	Peter Brantley
Eric Celeste	Sayeed Choudhury	Dan Chudnov
Kevin Clarke	Scott Collard	Karen Coombs
Walt Crawford	Jody DeRidder	Marta J Deyrup
Joanna DiPasquale	Thomas Dowling	Brad Eden
Karl Fattig	John Ferreira	Edward Fox
Marilyn Geller	Jimmy Ghaphery	Marcos Andre Goncalves
Carl Grant	Chris Gray	Daniel Greenstein
Martin Halbert	Sebastian Hammer	Per Hansen
Eric Hellman	Geneva Henry	Amanda Hill
Andy Ingham	J.C.McNulty	Paul Jones
Jay Jordan	Mark Jordan	Ranti Junus
Brewster Kahle	Francis Kayiwa	Aaron Krowne
John A. Kunze	Brad LaJeunesse	Amos Lakos
Ralph LeVan	David Lindahl	John Little
Joseph Lucia	Clifford Lynch	Emily Lynema
Gregory Madey	Gary Marchionini	Marua Marx
Robert McDonald	Robert H. McDonald	Mary Claire McKeown
Greg Notess	Nicolas Morin	Chip Nilges
Brandt, D. Scott	Brenda Reeb	Tamar Sadeh
Ross Singer	David Seaman	Debra Shapiro
Tim Spalding	Stephen Sloan	Plato Smith
Ed Summers	Peter Suber	Hussein Suleman
Guenter Waibel	Marc Truitt	John Unsworth
Tyler Walters	David Walker	Jenny Walker
Stanley Wilder	Donald Waters	Paul Watry
Bonnie Wilson	John Price Wilkin	
Jeff Young	Maurice York	

15 Appendix E – Licensing schemes and conditions

15.1 Selection of *Creative Commons* and *AEShareNet* icons

Creative Commons icons



This license lets others distribute, remix, tweak, and build upon your work, even commercially, as long as they credit you for the original creation.



This license lets others remix, tweak, and build upon your work non-commercially, and although their new works must also acknowledge you and be non-commercial, they don't have to license their derivative works on the same terms.



This license is the most restrictive of our six main licenses, allowing redistribution. This license is often called the 'free advertising' license because it allows others to download your works and share them with others as long as they mention you and link back to you, but they can't change them in any way or use them commercially.



This license lets others remix, tweak, and build upon your work non-commercially, as long as they credit you and license their new creations under the identical terms.



This license allows for redistribution, commercial and non-commercial, as long as it is passed along unchanged and in whole, with credit to you.



This license lets others remix, tweak, and build upon your work even for commercial reasons, as long as they credit you and license their new creations under the identical terms.

AEShareNet icons



May be freely used and copied for educational purposes but the owner retains full control of its use for any other purposes.



May be freely copied, adapted and used by anyone. Exact copies must retain the owner's copyright statement and the *AEShareNet-U* mark. Enhancements must not contain the owner's copyright statement and may have a new copyright statement by the Licensee.



Material may be used and enhanced by anyone free of charge but copyright in published enhancements consolidates with the original owner.



The material may be freely copied but only in its original form including the owner's copyright notice.



Licence conditions can be customised by the owner and a Licence Fee and/or Royalties may be charged.

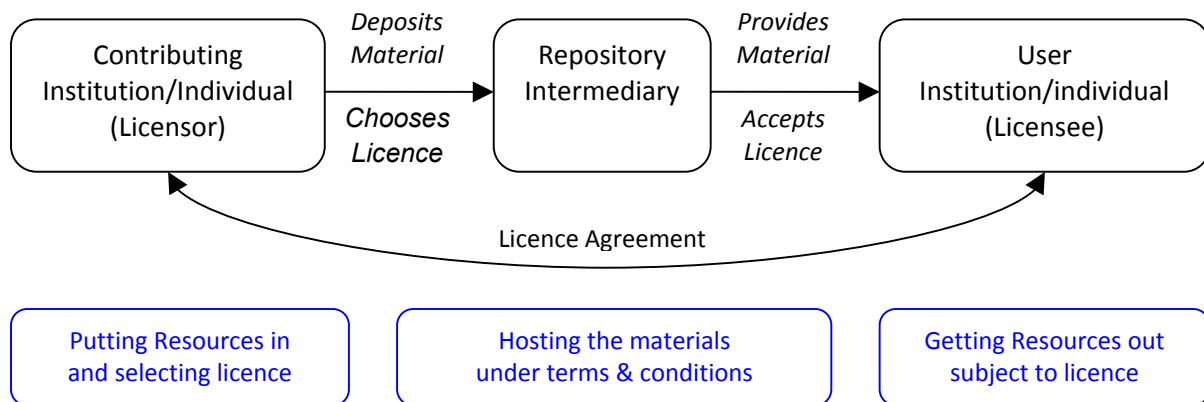


Intended for materials to be used on an as-is basis. An E licence can be taken out by an individual to use the material themselves, or by an organisation for use within the organisation including by an educational organisation for use by students.

15.2 Different licence schemes

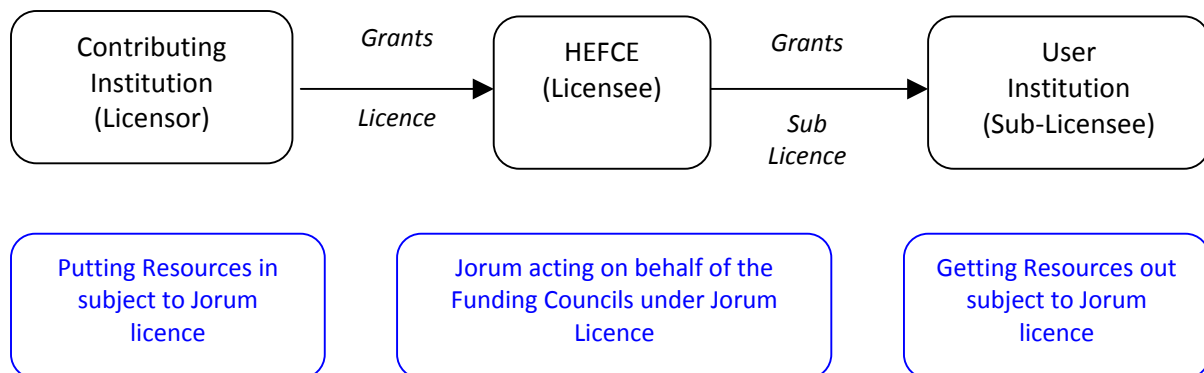
Basic repository licensing model

This model reflects a basic repository arrangement, whereby the contributor licenses the users directly, with the repository acting as a host/intermediary for the material.



Jorum licensing model

In the Jorum model, the HEFCE licenses the work from the contributor and then sub-licenses it to the user. The user does not therefore get a direct licence from the contributor, and HEFCE/Jorum are involved directly in the licensing chain.

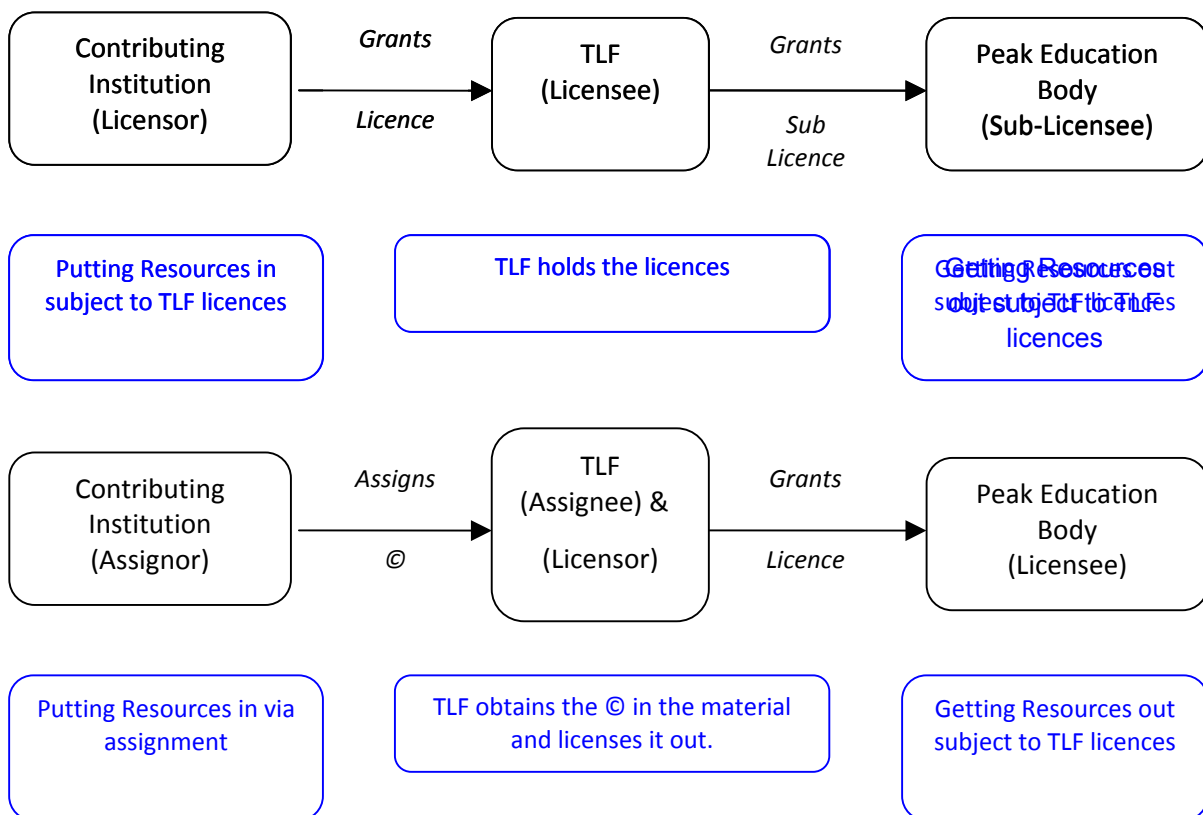


The Learning Federation licensing model

The Learning Federation’s model is similar to that of Jorum in that it includes:

- learning objects where there is a combination of TLF copyright and third-party copyright that has been licensed in;
- learning objects where the copyright vests entirely in a third party. That content is licensed in and made available through TLF’s distribution chain.

However, it also includes material for which contributors have assigned the rights to the TLF. In such cases the TLF is the licensor. The TLF does not deal directly with individuals. At the school level it deals with ‘Peak Education Bodies’, such as state education departments, the Catholic Education Commission, universities and teacher professional associations.



15.3 AShare licensing model (simple example)

AEShareNet-C	AEShareNet-E	AEShareNet-S	AEShareNet-P	AEShareNet-U	AEShareNet-FfE
Commercial Licence	End-User Licence	Share and Return	Preserve Integrity	Unlocked Content	Free For Education
Licence Period:					
Agreed term from 1 month	Agreed term from 1 month	Perpetual	Perpetual	Perpetual	Perpetual
Territory:					
Agreed term can be anywhere	Everywhere, except where specified	Everywhere	Everywhere	Everywhere	Everywhere
Exploitation / Supply Rights:					
Customisable	End user licence - can be used but not exploited	Shared use - encourages reuse	Integrity preserved	Unrestricted use	May be used freely but not exploited
Licensing:					
Fees/royalties may apply	End user licence, fees/royalties may apply	Instant licence, no licence fees	Instant licence, no licence fees	Instant licence, no licence fees	Instant licence, free for educational purposes
Offer mechanism:					
Licensor registers, licensee negotiates/ accepts	Licensor registers product list, licensee selects items and quantities.	Licensor applies mark, licensee uses	Licensor applies mark, licensee uses	Licensor applies mark, licensee uses	Licensor applies mark, licensee uses
Enhancements:					
Enhancements (if permitted) vest in original owner	Enhancements (if permitted) vest in original owner	Enhancements vest in original owner	Enhancements not permitted	Enhanced version vests in licensee	Edited versions permitted for educational purposes only
Example:					
Learning resources	Software, books	Learning resources	Industry standards, curriculum	Professional development materials	Website, policies and general information

16 Appendix F – Repository development projects

Alphabetic list of projects reviewed:

ARROW, <http://www.arrow.edu.au>. Australian project which uses Fedora but is expressly designed to bridge commercial and open source software; see also Meresco below.

ASPECT, <http://aspect.eun.org/>. Focuses on learning object / elearning repositories in the K12 sector

DINI, the Deutsche Initiative für Netzwerkinformation (German Initiative for Network Information). DINI certifies repositories as meeting quality standards (metadata chief among them). Together with the DARE guidelines, the DINI certificate served as a basis for the DRIVER Guidelines for Content Providers. Therefore all DINI certified repositories comply with the DRIVER guidelines. We note that all Dutch research repositories (universities and institutes) as well as 21 German repositories (as of 02/06/08) have received the DINI certificate. See also DRIVER below.

DOAJ, Directory of Open Access Journals, <http://www.doaj.org/>.

DRIVER, Digital Repository Infrastructure Vision for European Research, <http://www.driver-support.eu/index.html>. Having established interoperability guidelines using DINI, they are harvesting and providing a common search to their constituency of member repositories. This parallels Intute Repository Search. One important semiformal aspect to their working is mentoring – linking those running newly established repositories with those with more practical experience.

EThOS, <http://www.ethos.ac.uk/>. Electronic theses once established should be an important and growing part of institutional repositories. The ethos project seeks to establish standards to enable this via the toolkit <http://ethostoolkit.cranfield.ac.uk/tiki-index.php>

FRED, <http://fred.usq.edu.au/background.html>. The Federated Repositories for Education project aims to support deployment of repository federations in Australian education and training communities. It will document generic service-oriented models and produce software toolkits that support development of repository federations.

ICOPER, <http://www.frepa.org/wp/2008/09/04/kicking-off-icoper/>. Focussing on learning object / elearning repositories in HE

IRIScotland, <http://www.iriscotland.lib.ed.ac.uk/>. Implementing repositories for Scotland.

IRS, Intute Repository Search, <http://www.intute.ac.uk/irs/>. This is exploring ways in which text mining can enhance discovery, which might provide a practical, scalable means to overcome consistency issues.

JISC-funded Application Profile projects, e.g. SWAP, http://www.ukoln.ac.uk/repositories/digirep/index/Eprints_Application_Profile, linking DC and other metadata schemes.

JULIET, <http://www.sherpa.ac.uk/juliet/>. Collating the funding agencies' open access policies, mandates and requirements.

LIBER, Ligue des Bibliothèques Européennes de Recherche, <http://www.libereurope.eu/>. Of wider remit than just repositories, being a consortium of research libraries but as such an

important international player with a clear interest in coordination/interoperability of metadata.

LORN - <http://lorn.flexiblelearning.net.au/Home.aspx> - online training resources from across the Australian vocational education and training sector

MACAR initiative <http://macar.wikidot.com/> provides recommendations and advice on metadata requirements for digital repositories,

Meresco, <http://meresco.org/>. Meresco can cross translate between different metadata standards and xml (as can Fedora, DSpace or iNode, <http://www.k-int.com/product.php?id=4>).

Names, <http://names.mimas.ac.uk/>. Providing UK-wide authoritative lists for proper names.

OAI-ORE, <http://www.openarchives.org/ore/>. Leading proponents of web-based exchange protocols, see <http://www.openarchives.org/pmh/>, Protocol for Metadata Harvesting, which seeks interoperability through metadata exchange.

OpenDOAR, <http://www.opendoar.org/> and ROAR (repository directory), <http://roar.eprints.org/>. Both provide directory and summary information about academic repositories.

PILIN Project, <https://www.pilin.net.au/>. This piloted a shared, standards-based, persistent identifier management infrastructure. RIDIR worked with it, and assumed its results in their conclusions (see below).

PREMIS, PREservation Metadata Implementation Strategies, <http://www.oclc.org/research/projects/pmwg/>. Explicitly concerned with long-term preservation in repositories. They have published an xml data dictionary for the metadata believed to be needed to achieve this.

RIDIR, Resourcing IDentifier Interoperability for Repositories, <http://www.hull.ac.uk/ridir/>. Their final report makes recommendations which cover some of the same ground: that JISC should establish a national eframework service to resolve identifiers, probably using OAI-ORE. (Identifiers can be seen as analogues of the metadata we have been considering.)

RoMEO, <http://www.sherpa.ac.uk/romeo.php>. Publishers' open access and copyright policies on journal articles provided by SHERPA.

SHERPA, <http://www.sherpa.ac.uk/>. A project-based repository consortium of UK HE institutions.

SWORD, Simple Web-service Offering Repository Deposit, <http://www.ukoln.ac.uk/repositories/digirep/index/SWORD>. A JISC-funded project (2007-2008), SWORD is a lightweight protocol for deposit, and a profile of the Atom Publishing Protocol. The motivator for SWORD is 'lowering the barriers to deposit', principally deposit into repositories, but potentially deposit into any system which wants to receive content from remote sources.

UK Council of Research Repositories, <http://www.ukcorr.org/> (not linked to by RSP). A professional organisation of librarians and other professionals charged with the running of repositories.

Version Identification Framework, <http://www.lse.ac.uk/library/vif/>. Strictly this concerns issues a step or two beyond our brief, i.e. how to deal with different versions of the same document once they are all in a repository. These issues were also addressed by RIDIR.

17 Appendix G – e-Prints Soton – qualified Dublin Core fields

This appendix relates back to the discussion of the e-Prints Soton repository in section 9.

artefact: School or Centre (internal_group)
article: School or Centre (internal_group); Status (ispublished); Refereed (refereed); Title (title); Journal/Publication Title (publication)
book: School or Centre (internal_group); Status (ispublished); Title (title); Publisher (publisher);
book_section: School or Centre (internal_group); Title of Book (book_title); Status (ispublished); Refereed (refereed); Title (title); Publisher (publisher)
composition: School or Centre (internal_group); Date of Issue (date_issue); Title (title)
conference_item: Presentation Type (pres_type); Event Type (event_type); School or Centre (internal_group); Status (ispublished); Refereed (refereed); Title (title); Event Title (event_title)
exhibition: Subjects (subjects); Event Dates (event_dates); Exhibition Type (exhibition_type); School or Centre (internal_group); Is exhibiting or curating the show? (isexhibiting); Location (event_location); Title (title); Number of works exhibited (exhibition_totalworks)
monograph: School or Centre (internal_group); Status (ispublished); Refereed (refereed); Title (title); Publisher (publisher)
other: School or Centre (internal_group); Title (title)
patent: School or Centre (internal_group); Date of Issue (date_issue); Patent Applicant (patent_applicant); Title (title); Identification Number (id_number)
performance: School or Centre (internal_group); Title (title)
thesis: Department (department); Qualification Name (qualification_name); Thesis Type (thesis_type); School or Centre (internal_group); Date of Issue (date_issue); Status (ispublished); Title (title); Institution (institution)

18 Appendix H – Definitions of terms used in the report

Term	Definition
consistency	Use of recommended or mandated structures, schemes or policies which might enable software designers and service creators to predict with some certainty the type of content to expect.
depositor	An individual who, for whatever reason, uses or intends to use a repository to store, publish or share digital objects.
end-user	An individual seeking to find or use information from a repository.
institutional repository	A repository (see below) which is under the control of an institution and whose primary purpose is to fulfil the institution's objectives. '... a set of services that a university offers to the members of its community for the management and dissemination of digital materials created by the institution and its community members. It is most essentially an organizational commitment to the stewardship of these digital materials, including long-term preservation where appropriate, as well

as organization and access or distribution' (Clifford Lynch, 'Institutional Repositories: Essential Infrastructure for Scholarship in the Digital Age')

policy expression	'Web Services Policy is a machine-readable language for representing the capabilities and requirements of a Web service. These are called 'policies'. Web Services Policy offers mechanisms to represent consistent combinations of capabilities and requirements, to determine the compatibility of policies, to name and reference policies and to associate policies with Web service metadata constructs such as service, endpoint and operation' (Asir S Vedomuthu, et al., 'Web Services Policy 1.5 -Primer', http://www.w3.org/TR/ws-policy-primer/). For the purposes of this work, when we use the specific phrase 'policy expression' we will be referring to the above definition. However, because the tender specifically includes generic reference to the ways that a repository might formulate and make available its policies (e.g. its policies on access to and use of materials, copyright, ownership, archiving etc.) in human readable form as well as machine readable form, we will also discuss the human readable ways in which a repository might express its policies.
precision	Precision and Recall are two widely used measures for evaluating the quality of results in domains such as Information Retrieval and statistical classification. Precision can be seen as a measure of exactness or fidelity, whereas Recall is a measure of completeness. In an Information Retrieval scenario, Precision is defined as the number of relevant documents retrieved by a search divided by the total number of documents retrieved by that search, and Recall is defined as the number of relevant documents retrieved by a search divided by the total number of existing relevant documents (which should have been retrieved). (Wikipedia, http://en.wikipedia.org/wiki/Precision_and_recall)
repository	Any digital store where items (which can be metadata only, content only, or both) are deposited and later accessed. Repositories may be primarily for archiving or publication or coordination. They may be for institutional purposes or run by discipline or subject communities. In education they may be used for research or learning materials. Any combination of all these options is possible.
richness	The tender document for this piece of work talks about: <i>consistency, richness and precision</i> e.g. ' <i>the consistency, richness and precision with which repositories share material</i> '. We define consistency and precision above. Richness in this context is synonymous with the concept of 'deep and broad'. Repositories which are 'rich' are full of scholarly and academic potential; their content exemplifies quality, quantity and variety.
sitemap	A sitemap is a way of portraying the organisation and component pages of a website, identifying the URLs and the data under each section. There are two common usages for the term: 1) refers to a human-readable index of pages on a website, usually an html representation; 2) refers to a machine-readable file, usually in xml, enabling search engines to find data faster and more efficiently. We always use the latter meaning unless we specify 'human-readable sitemap' (see http://www.xml-sitemaps.com/about-sitemaps.html).
user	For the purposes of this piece of work only, we define a user as a depositor and we will did not consult 'end-users' in our user consultation.