



# **Report of the Subject and Institutional Repositories Interactions Study**

**November 2008**

**Catherine Jones, Robert Darby, Linda Gilbert and Simon Lambert**

**Information Services Team, eScience Centre,**

**Science and Technology Facilities Council**

## Contents

1. Executive summary .....	1
2. Introduction .....	3
2.1 Why interact and with what? .....	3
2.2 Project methodology .....	4
2.3 Acknowledgements.....	4
3. Current state .....	5
3.1 Synopsis.....	5
3.2 Repository definitions .....	5
3.3 Rationale for content collection within repositories .....	6
3.3.1 Stakeholders' business cases for collecting publication information .....	6
3.4 Activities and workflows .....	7
3.4.1 Acquisition of content.....	8
3.4.2 Discovery of content .....	9
3.4.3 Repository interactions .....	9
3.4.4 Other uses of repository content.....	9
3.5 Technical development activities.....	9
3.5.1 Deposition .....	9
3.5.2 Location of content .....	10
3.5.3 Linking to data and other related resources.....	10
4. Key findings.....	11
4.1 UK funders.....	11
4.2 UK Institutional Repository managers .....	12
4.2.1 Institutional Repository Managers Survey conclusions .....	13
4.3 Subject repositories .....	14
4.4 Services .....	14
4.5 Interested parties.....	14
4.6 General findings.....	15
4.7 Visions of the future.....	16
5. Some potential future directions for repository interactions .....	17
5.1 DRIVER ONE: Population of repositories .....	17
5.2 DRIVER TWO: Statistics and metrics .....	20
5.3 DRIVER THREE: Preservation.....	20
5.4 DRIVER FOUR: Aggregation of research outputs .....	21
6. Recommendations .....	23
6.1 Standardisation .....	23
6.2 Best practice.....	23
6.3 Community engagement and dialogue.....	24
7. Conclusions .....	25
8. Bibliography and references .....	26

### Change log

Document version	Authors	date
V1 – circulated to JISC	CMJ	7/9/2008
Finalised version (5)	CMJ	20/11/2008

## 1. EXECUTIVE SUMMARY

This report was commissioned by JISC to produce a set of practical recommendations for steps that can be taken to improve the interactions between institutional and subject repositories in the UK. For practical reasons, the report was concerned exclusively with scholarly articles at their various stages of existence, although many of the report's recommendations are capable of a more general application.

We reviewed the history of the development of repositories in the UK, and analysed the current repository landscape, primarily through seeking the views of key stakeholders: repository managers, funders and other parties involved or interested in repository development. Metrical data were gathered from published sources and by means of a survey of institutional repository managers.

### Key findings

- The majority of institutional repositories (IRs) are at an early stage of development and the desired 'critical mass' of content is far from having been achieved;
- despite the declared interest of IR administrators in a co-ordinated approach to the gathering and sharing of information, there is in fact very little interaction between repositories;
- most deposit is initiated and mediated by repository staff, while self-archiving is not yet embedded in author workflows. Technical and administrative solutions for management of research outputs, developments in reporting of article usage statistics, and the requirements of the Research Excellence Framework (REF) are likely to drive cultural change;
- content collection is strongest in established subject/funder repositories;
- there may be scope for greater collaboration with publishers in the development of deposit and distribution procedures;
- repository administrators struggle to identify relevant content/metadata in external sources because identification by author or organisational association is highly problematic;
- content transfer between repositories requires a relationship of trust, which must in turn be based on explicit metadata standards, clear provenance and rights statements, and agreed protocols for transfer and updates to objects and metadata;
- there is considerable interest throughout the community in creating aggregations of content held in repositories and other sources by linking to data and related items. The OAI-ORE web content aggregation specification represents one potentially valuable model of a user-centred content organisation technology;
- there is no coherent approach to content preservation among repositories, and in many cases long-term preservation policy appears underdeveloped. This is a critical issue for the long term;
- there is wide variation between repositories in metadata formats and quality;
- for pragmatic reasons many IRs collect largely metadata-only records. The extraction of metrics to support local and national assessment and administration is an important driver for collection. There is a different imperative to acquire, preserve and make freely available full-text content. There is evidence of a trend towards integration of institutional repositories with research management systems.
- Funding organisations and HEIs share many common purposes and would each benefit from collaboration. That such collaboration is not as yet taking place on any significant scale is attributable less to technical barriers than to the absence of any established structure for the negotiation of co-operative working practices.

## Recommendations

We make a total of seven recommendations, which are intended to be achievable in whole or in part in the immediate future. They are variously addressed to a number of stakeholder groups: JISC, funding organisations, repository managers, software developers and creators of content. This report recommends:

with regard to *standardisation*

1. that continued support be given to implementation of national standards for unambiguous identification of authors, funders and higher education institutions;
2. that the community work towards the adoption of common information interchange standards;
3. that a watching brief be kept on the Trusted Repository certification process and that all repository managers participate in this scheme when fully established;

with regard to *best practice*

4. that records transferred from one repository to another contain clear provenance information;
5. that repositories implement version identification at object and metadata levels;

with regard to *community engagement and dialogue*

6. that a UK repository community forum be established where representatives of subject/funder and institutional repository communities can work to agree and implement standards and protocols for co-ordinated information management;
7. that continued efforts be made to engage with users and ensure that developments address user needs in viable ways.

## 2. INTRODUCTION

This study was undertaken between May and October 2008 and was funded by the Joint Information Systems Committee. The remit was to survey the UK repository landscape examining interactions between subject and institutional repositories and to make concise and practical recommendations on how these could be made to be more productive.

The scope of the project was limited geographically to the UK, but it should be noted that most subject repositories are world-wide and based outside of the UK and a small sample of these was included in the fact finding phase. It was limited by material type to journal articles. While the publishers and commercial secondary sources such as abstracting and indexing services are stakeholders in the journal article environment, interactions with these were out of scope and not directly covered by the project team. There are strong views within the community as to the most appropriate repository for the initial deposit; the project team has not taken a view on this and our recommendations are concerned with effective interactions.

### 2.1 Why interact and with what?

Repository interactions can occur at many different levels, from deposit or information transfer to location tools. The project team has discovered that the one of most interest is communicating at the level of content; usually indicating data transfer from one repository to another. This may be either the metadata with the full text content remaining at the original repository, or both metadata and full text. The table below shows some possible types of interaction. It must be noted that some of these potential interactions are not directly with other repositories, but with third party services which would benefit the repository community as a whole.

Interaction type	Description	Possible partners in interaction
<b>Metadata transfer: set of records (one or more)</b>	Duplicating the content of an identifiable collection of records from one repository to another for the purpose of increasing the completeness of the content	Two repositories; publication databases and a repository; Research management system and a repository.
<b>Metadata and full text transfer: set of items (one or more)</b>	Duplicating the content of an identifiable collection of records from one repository to another for the purpose of increasing the completeness of the content and providing more access points to the full text or a specialised service such as preservation	Two repositories; repository and specialised third party service.
<b>Notification of content</b>	A process where one repository provides an alerting service so that others can collect content or point to content	Repository; repository and research management system
<b>Statistics collection</b>	A process where a repository or third party service collates usage statistics on recognised full text items	Repository and third party service
<b>Information interchange about policies</b>	A process where a third party service collects information on non-content related topics such as policies to be able to build additional services on top of the repository landscape	Repository and third party service
<b>Look-up or resolution services</b>	A process where a URI/link in another service is resolved to content held in the original repository	Third party service and repository
<b>Providing links to related information</b>	A process where the scholarly work can be linked to underlying data or other material to enhance the experience of the searcher	Scholarly works repository and data repository;

This document describes the current state of interactions, summarises the key findings, identifies potential scenarios and concludes with the recommendations.

## 2.2 Project methodology

The stakeholder communities were identified as: authors; end-users; repository managers (both institutional and subject); funders; national projects and initiatives; and experts in the field. Due to the difficulties in identifying authors and end-users who would be interested in contributing to this project in the timescales allotted, it was decided to concentrate on the other stakeholder groups. This does mean that the interests and views of those who need to use the deposit and location tools have not been represented.

We conducted both face-to-face meetings and telephone interviews with stakeholders. In order to ensure we gained an overview of the repository management community, we also undertook an online survey of repository managers using the UKCoRR JISCmail list.

## 2.3 Acknowledgements

Our thanks go to the following people who agreed to be interviewed by the SIRIS team. Their input was invaluable and we appreciated the time they put aside for this task.

.

Theo Andrew	Sophia Jones	Mary Robinson
Chris Awre	Robert Kiley	Sally Rumsey
Juan Bicarregui	Thomas Krichnel & the eLIS	Beverley Sherbon
Michelle Blake	team	Sue Smart
Barbara Búltzmann	Gerry Lawson	Jane Smith
Cormac Connolly	Nadina McShane and the NORA	Elin Stangeland
David Flanders	consortium	Simeon Warner & arXiv team
Jessie Hey	Jayne Morris	Ian Viney
Bill Hubbard	Tony Peatfield	Paul Walk
Neil Jacobs	Jacqui Primrose	Wendy White
Keith Jeffery	Rachel Proudfoot	

We would also like to thank Nicky Ferguson and his colleagues from the study into *Approaches to improve the consistency with which repositories share material* for sharing an early draft of their report and their input into our study.

### 3. CURRENT STATE

#### 3.1 Synopsis

The repository landscape is a complicated one in which there are different types of repositories, serving overlapping communities and providing overlapping services. However there are common themes of collecting and exposing content.

There is no consensus on whether a repository should only hold digital objects with associated metadata rather than metadata-only records. One of the major factors in deciding about this part of the collection policy is the purpose of the repository: whether it is intended to focus on Open Access, and hence digital objects or whether it is intended to focus on a complete collection of information where metadata only records are often required for comprehensive coverage.

There is a mixed economy for repository content: there are a growing number of institutional repositories in the Higher Education Institution (HEI) & research institute communities, some UK-based funders provide repositories for project outputs and there are the international subject-based repositories. It is also becoming more common to have repositories relating to an event, such as a conference, or project, but as these do not focus on journal articles, they have been considered out of scope of this study.

#### 3.2 Repository definitions

The scope of the project discussed subject and institutional repositories; upon investigation the differences between discipline-based and funder-led repositories were significant enough to distinguish them. It is important therefore to note these differences as there are different motivational factors and requirements driving the collection of material and hence their attitudes to interactions. There are additionally repositories using models that do not fit into these three categories, but these are in the minority. It should be noted that there are no clear edges to the categories, it is best to consider it as a continual spectrum.

**Institutional repository:** This is a collection of research outputs with a common link to a particular institution, usually by authorship. These repositories are likely to cover more than one research discipline, to have funders in many if not all the Research Councils and support communities who have different approaches to research dissemination. Whether deposit of content is mandatory is a decision that will be made by each institution. The institutions may have many requirements for the content of the repository, from open access dissemination, through metrics, marketing to strategic planning. It is likely that many of these processes in the past were undertaken through collection of bibliographic information.

**Subject repository:** This is a collection of research outputs with a common link to a particular subject discipline. Subject repositories are likely to cover one broad-based discipline, with contributors from many different institutions supported by a variety of funders; the repositories themselves are likely to be funded from one or more sources within the subject community. Although for some subject repositories the funding may be fragile, if they are of enough importance to the community then funding crises are usually weathered. Deposit of content is voluntary. These repositories are usually concerned with dissemination; for example the emergence of the arXiv repository replacing the practice of circulating paper preprints in the particle physics community.

**Funder repository:** This is a collection of research outputs with a common link to one (or more) funder(s). These are likely to cover the funder's remit, which is usually subject-based but can become indistinct at subject boundaries, and will have authors from many institutions. Deposit of content is usually mandatory and can include project-related material, such as completion project reports. The

funders will have requirements for the content of repository, from metrics through to strategic planning.

Using these definitions, the following categorisation could be made as of September 2008:

Type of repository	Examples
<b>Institutional</b>	Higher Education Institutions Research Council Institutions, such as the NERC Open Research Archive (NORA)
<b>Subject</b>	arXiv for some areas of physics eLIS for information science and technology RePEc for economics
<b>Funder</b>	ESRC's Society Today JISC's Information Environment repository
<b>Difficult to categorise</b>	UKPubMed Central: this is a multiple funder supported subject repository. The Depot: national level repository for those who do not have an institutional one available. Economists Online: defined subject area, contributing content limited to consortium partners.

### 3.3 Rationale for content collection within repositories

There are some easily identifiable primary ends provided by repositories, regardless of the type, but not all of them adopt all of the ends:

- Promotion of open access to full text research outputs;
- Dissemination and promotion of research;
- Long-term preservation of intellectual content;
- Maintenance of a research record for purposes of administrative assessment and evaluation

These aims affect a particular repository's attitude to many policy issues; most importantly whether the repository is full text only or a mixture of metadata and full text records, and need to be considered.

#### 3.3.1 Stakeholders' business cases for collecting publication information

Funders are interested in all the outputs of a project to ensure that the money spent has been effective. Depending on the stance of the funder, the systems to collect the information may just focus on metadata or on metadata and the full text. The Research Councils have an increased focus on outcomes of a project or programme. In this case, outcomes consider the wider and longer lived impacts of a project or programme, trying to identify the more intangible results. The RCUK project on Outputs and Outcomes uses the Treasury Green book definition:

*"Outcomes are the eventual benefits to society that proposals are intended to achieve. Outcomes sometimes cannot be directly measured in which case it will often be appropriate to specify outputs as intermediate steps along the way e.g. value of extra human capital and / or earnings capacity."*

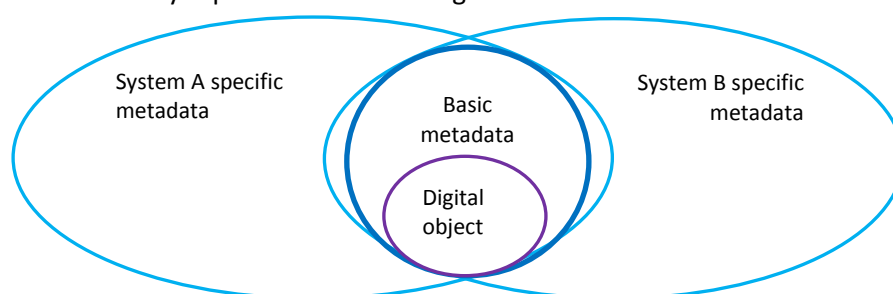
JISC's Information Environment repository is a relatively new one and it is intended to provide a place to collect the outputs of JISC programmes in a concerted and organised fashion to ensure longevity of content and a focus for locating information.

Institutions are interested in both the full text for dissemination, and in some cases preservation, and additionally require a complete record of academic outputs for internal administrative functions. The latter objective may also be a function of the institutional repository, or may be provided by a specialist tool.

Subject repositories collect the full text of publications purely for dissemination.

### 3.4 Activities and workflows

As can be seen from the repository definitions, journal articles form an important part of metrics measurement for most of the stakeholders; however, each stakeholder will require different views of the information. It is only natural that each stakeholder should create a partial view of the information according to their idea of what is useful, while possibly excluding elements of importance to other stakeholders. This is visually represented in the diagram below:



The *digital object* is the intellectual content, regardless of format, and should, ideally, include information within the object so that it can be uniquely identified. This should include title/author/journal, the version details and funder details.

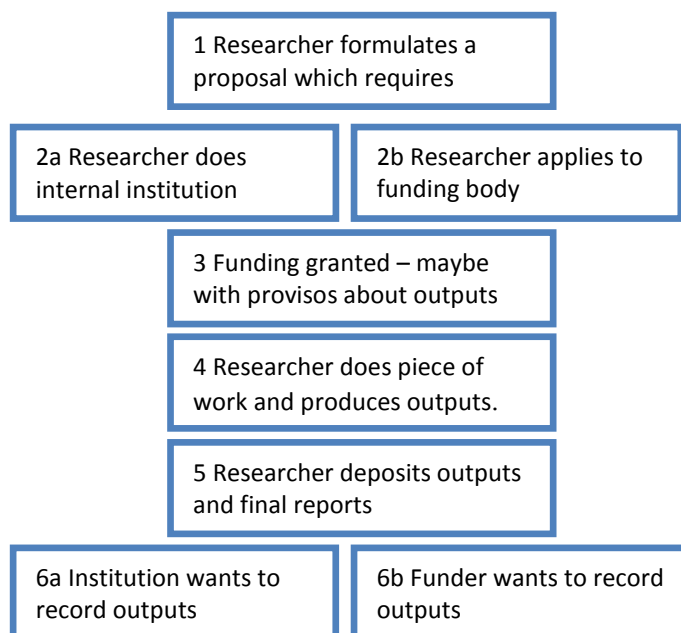
The *basic metadata* is a factual description and would be the same regardless of who input the metadata and which system it was entered in. We suggest that this would include information on: article title, the first author at least, and the journal title. Using text mining tools and with the correct information within the digital object this metadata might be able to be automatically generated in the near future. If the digital object was post-publication then there would be details about the volume/issue/article number or page numbers depending on the nomenclature used by a particular journal.

*Service specific metadata* acknowledges the fact that different systems would have additional information to ensure that all the individual requirements of each system are met. This might include fields such as departmental structures, institution affiliations for each of the authors, subject coding, funder, grant numbers and grant programmes to name but a few. For successful repository interactions then there may be a need for systems to store information not required for internal purposes but that are required for another repository to identify useful content.

It should be noted that not all stakeholders require the digital object for their business processes to be successful.

### 3.4.1 Acquisition of content

The journal articles this study is concerned with form a small part of the research cycle. The different stakeholders interact at different parts of this process, but the output is of importance to all. The diagram below shows the research process:



The author/creator has different calls on their output: dissemination within their community; reporting to their funder and career-/job-related issues within the institution they are affiliated to. The discipline of the community has a big effect on the type of dissemination and this influence on user behaviour needs to be taken into consideration in the repository interaction context.

The author's interest is in fast and effective dissemination, specifically among their peers. The motivation in doing so is to advance knowledge and understanding within the discipline, to develop his/her professional capital and build reputation. Once an output is produced and placed in an appropriate journal of record (stage 4 in the diagram above), his/her interest in it as an object in the research workflow ceases almost entirely; that is to say, stage 6, essential from the repository perspective, is liable to be considered by the author to be outside the core research workflow and essentially bureaucratic.

For the funder the workflow associated with any research outputs and outcomes spans the life of the research from the initial application for funding (at stage 2b, above), and proceeds through to stage 6, the formal record of outputs and outcomes. The interest of the funder is to demonstrate that a value relationship exists between the funding input and the research output. The funder is primarily interested in the intellectual content in so far as it justifies the funding given. Dissemination is certainly a necessary part of this process, but a repository is also a record of data about research outputs and outcomes, and this is the information of value to the funder.

The author's institution has invested in the author and anticipates returns in recognised academic excellence and implications for institutional funding. For this reason the institution also has an interest in collecting information about the research, which can inform both internal assessment and evaluation procedures as well as compliance with national evaluation schemes such as the Research Evaluation Framework (REF).

It is apparent that both funders and institutions are aligned in their requirements for administrative information, and that, while sharing a commitment to the principles of open access and preservation, they also value ownership of content for its 'shop-window' value. In principle they share enough

common purpose to justify a co-ordinated approach to achievement of their objectives. The need to foster this collaborative approach has been highlighted by many of the Institutional managers we have spoken to.

### **3.4.2 *Discovery of content***

The most mentioned tool for locating content in repositories was Google. Intute Repository Search is a useful tool if the end-user knows that the content is likely to be in a repository. There is some discussion about what the user should first see in the results set: should it be a metadata splash page or the full text object? It is likely that this might be different for different people and for different stages of the research fact-finding process. At the initial point the researcher may be uninterested in version/provenance related issues surrounding the digital object but in the intellectual content itself; at a later stage when it is to be formally referred to then these other types of issue become more important.

### **3.4.3 *Repository interactions***

Whilst there are many technical developments which enhance the possibilities for successful interactions between repositories, there is not much evidence that this activity is widespread, at present. There is further information about existing interactions in Section four: Key Findings.

### **3.4.4 *Other uses of repository content***

The dissemination aspects of repositories are satisfied by the discovery of the full text content; there are other direct uses of the content such as providing publication lists for researchers, providing publication lists at a higher level such as departmental, funding programme, institutional, subject based. An area which is becoming more important is the linking of the formal publication to other material, such as the underlying data.

## **3.5 Technical development activities**

### **3.5.1 *Deposition***

The issues around deposition of content and the requirements of different stakeholders have made this area a keen one for technical developments.

The repository standard tool for information transfer has been the OAI-PMH protocol. This enables a repository to expose a defined set of metadata records describing the content which then can be harvested by another repository or third party service. As this protocol requires the set to be defined by the owner of the data, then it is likely to be done in ways which are easiest for that repository to do; such as by material type. It is also not easy for the repository exposing content to identify who has harvested it and reused it. In this environment if another repository or third party service wishes to use this data, they have to use the set provided even if it does not entirely match their requirements. In this case there might be significant post-harvesting processing. An example of this might be a subject repository which exposes a set of journal articles which is harvested by a particular institution which only wishes to add to their repository items relating to their authors. In this case records with the correct affiliation would need to be identified from the complete set. Other examples might include funders harvesting from institutions where identification of the funder would be required. One method of reducing the post-harvesting processing would be for the repository to expose sets matching the requirements of the harvesters; however this has significant maintenance overheads. Although OAI-PMH has some issues in this context, it does have the advantage of being one of the standard interfaces to all repository software.

SWORD is a recent (2007) development which uses the ATOM Publishing Protocol to provide a lightweight protocol for deposit. This enables deposit tools to be built independently of the workflow

of a particular repository and opens up the possibility of one deposit workflow feeding multiple repositories. It does not address issues surrounding the packaging of complex objects.

Open Access Initiative – Object Reuse and Exchange (OAI-ORE) defines standards and protocols for the description and exchange of complex digital objects made up of more than one file. It utilises the concepts of web architecture by enabling complex objects to be described using a resource map and making this resource map identifiable through a URI. It is designed to make reuse achievable and one form of reuse is transferring content from one repository to another. It is in the early stages of uptake and the full impact is not yet quantifiable.

The Dublin Core Scholarly Works Application Profile (SWAP) provides a metadata specification for describing scholarly works at a greater level of richness and consistency than Dublin Core. It includes basic metadata fields and additional system-specific ones such as research funder and grant number. If this were to be adopted more widely then it would be easier to build deposit and location services at a national scale. However as N Ferguson et al discuss in their report 'Feasibility study into approaches to improve the consistency with which repositories share material' the more descriptive power a standard has the more likely it is to have empty fields.

### **3.5.2 Location of content**

The main issue in this area is the discussion around the use of specialised search tools as against Google-type tools. Abstract and Indexing databases provide an effective mechanism for researcher's information seeking behaviours and it is likely that these will adapt to provide access to repository content. If critical mass is achieved in all repositories then there will be a large amount of work in de-duplication search results and accurate information on versioning will be important to be able to understand the version relationships between content from different repositories.

### **3.5.3 Linking to data and other related resources**

Recent data and publication related JISC projects include CLADDIER, StORe and their joint successor project Storelink.

CLADDIER looked at the issues surrounding formal publication and citation of environmental data and developed a mechanism to allow for repository to repository alerting and transfer of citation information. This mechanism was based on the Track back protocol used in the Blog world. This project assumed that the data related to a journal article would be "formally" published and a citation would be part of the bibliography of the paper, and vice versa, and so when a item was deposited with explicit citations the protocol would inform the repository hosting the cited item of the deposit of the item citing and thus closing the loop by providing backwards citation. For this to be taken up within the community there needs to be a trusted environment to allow for the trackback mechanism to work and the relevant data input so that the track back ping can be sent to the right repository.

The StORe (Source to Output Repositories) project had two main phases: a survey of a set of disciplines and a piece of middleware which enables a user to make explicit links between data and publication repositories.

The UK research data service feasibility study (UKRDS) is looking at the UK national strategy for data management and service provision. The interim report suggests that there are three possible models for organisations: 1) no change from the existing state of play; 2) massively centralised service and 3) hybrid or umbrella service which would combine regional and specialist repositories. It suggests concentrating on option three for the next stage. The report also looked at functions which might be provided by UKRDS, however linking data and publications were not explicitly described.

## 4. KEY FINDINGS

This section describes the key findings from our interviews and survey and then identifies common themes.

### 4.1 UK funders

The views of the UK Research Councils, Wellcome Trust and the JISC were sought in this area. Within these Funder communities, there was a strong commitment to subject/funder related repositories as shown by the support of UKPMC, Society Today and the JISC's Information Environment repository. There was a strongly held minority view that this type of repository was more populated, and longer-lived, than institutional repositories. There was consensus from those funders who run subject-related repositories that, at present, content should be deposited directly into their repository and that if institutions were interested in this content they should harvest it from their repository. It should be noted that for full text there might be rights issues involved in transferring this content. There was less consensus around the issue of compliance-monitoring and encouragement to deposit. Some funders already have a process and resources in place to do this, whereas others were still considering the issue.

All funding bodies have Open Access positions, but support the realisation of these in different ways. All funding bodies need in some way to account for themselves and for this the metadata of a digital object is required.

The Research Councils are viewing developments in the repository field with interest. There is growing pressure on Research Councils to show value for money in relating research funding inputs to measurable impact, in terms of both research *outputs*, such as peer-reviewed scholarly works, and *outcomes*, such as patents and related commercial developments. The RCUK Outputs and Outcomes Collection Project (OOC), which reports to the RCUK Operational Management Group in March 2009, is currently investigating the possible development of an integrated system for management of research outputs and outcomes across all Research Councils and is interested in getting information from institutional repositories to support this process.

Research Councils have two roles in the research arena: they provide the funding for projects to explore and expand knowledge in a particular domain, and they provide funding for postgraduate study to individuals. As such the Research Councils are interested in the effectiveness of funding from postgraduate level. Any process to assign identities to researchers needs to start at this point in an individual's career.

Subject classification of the outputs was of importance to many of the funders we spoke to. There was not a consistent approach across organisations to the system adopted or to who actually assigned the terms at present. These subject classifications were used for reporting performance and other administrative functions. There are RCUK level discussions with the Higher Education Statistics Agency about the next version of the Joint Academic Classification of Subjects.

Critical to the effectiveness of any audit system is a mechanism for relating research outputs to funding. RIN's recently-published Guidance on Acknowledgement of Funders in Scholarly Journal Articles (February 2008) proposes a standard format for inclusion in the research object itself of funder and funding code information. If this information were included in research outputs as a matter of course, and made searchable by repositories in a standard metadata format, it would be possible for funders to harvest information about research outputs directly related to their own grants. Funding information is not captured in the simple Dublin Core metadata format, although it could be written into more sophisticated formats and harvested via SWAP.

Even a well-established funder repository such as UKPMC, which is backed by a strong mandate to deposit from the Wellcome Trust, and obtains most of its content by direct transfer from publishers, has a low deposit compliance rate – somewhere in the region of 30% as of June 2008. This is due in part to the systems in place and steps are being put in place to improve these to increase the rate. Institutions, as the employers of the authors who produce the outputs, are well-placed to monitor and enforce compliance through professional assessment procedures. At the same time institutional repository staff can ensure that deposited items record funder and grant code information and that records are automatically shared with the funder repository where appropriate. This is arguably an area where funders and institutions can profitably work together.

Where full text is deposited, there are different positions on which version is preferred. Wellcome Trust would like the author's final manuscript with corrections and the Research Councils are guided by the publisher's terms and conditions. Version identification is complicated by the lack of agreed terminology (see VERSIONS & VIF) but needs to be addressed to assist the location and identification of material.

Although this study is looking at journal articles, there is interest within this community to start to link the publications to the data which generated them. This is in the early stages of exploration. Many of the funders already support data centres for large scale data.

#### **4.2 UK Institutional Repository managers**

A sample of institutional repository managers was interviewed and this process was complemented by an online survey which was circulated to the JISCMail UKCoRR list. One of the biggest concerns for this group of interviewees was transferring policy into content collection and how to approach the different disciplines with their different views on the repository.

An important factor to consider in the interactions between Institutional Repositories and other repositories is the aims and purpose of the Institutional repository. For example, a repository whose main aim is to preserve the institution's intellectual record is more likely to insist on collecting the full text of the item; it would require a higher level of trust before it would accept linkage to full text held elsewhere than would a repository whose main aim is the dissemination of content or providing a showcase for the organisation. The apparent disjunction in some organisations between the publication database recording for metric purposes, run by the Research department and the Open Access repository run by the Library is likely to mean that workflows, information transfer and visibility of the IR will become key issues for some repository managers in the near to medium term. There is also a focus on interoperability within the organisation to leverage other opportunities to collect content.

Each institution has its own policy about whether the content was entirely full-text or not. Generalising from our research, those repositories which are also fulfilling publication reporting roles were most likely to have metadata only records as well as full text items.

Institutional identity was of concern to many of those who we spoke to. Higher Education Institutions are complex bodies with associated research centres and joint appointments and so it can be difficult to establish the research output of a particular organisation in the contents of another repository. There is also the issue of level of detail; it is likely that an institution will want to classify the affiliation to a smaller organisational unit than other stakeholders will want to.

There is interest within this community about linking to data and more uncertainty about who will be responsible for the data which is not destined for a large scale data centre.

#### 4.2.1 Institutional Repository Managers Survey conclusions

There were 24 completed questionnaires and a further 8 partially-completed but saved responses. All of these have been taken into account, giving a response rate of 26% of the entire UKCoRR list membership. Of those respondents 94% worked for an HEI and the remaining 6% for a Research Council establishment. 75% of the repositories had been in existence for less than 3 years, with 9% over five years old. All subject areas were covered by the respondents, although not all repositories covered all areas. The full report of the survey is an appendix to this report.

The table below shows the purposes for the repository as identified by our survey respondents.

Item	Most Important	4	3	2	Least Important	Not considered
To be an open access repository with full-text content	60.0%	28.0%	12.0%			
To be a management tool for institutional research outputs	36.0%	24.0%	28.0%	8.0%		4.0%
To increase visibility and dissemination of the institution's research outputs	76.0%	20.0%	4.0%			
To provide a mechanism for digital object preservation	4.0%	40.0%	36.0%	12.0%	8.0%	

At present most content is obtained by repository staff mediating the deposit process, author self-deposit and other staff mediating deposit. There is some internal and external bulk transfer, but this is in the minority. 18% of the respondents collect content from another repository and 30% have content collected from their repository. In both questions over 40% said that this might happen in the future. Of these respondents, 7% said that they used SWAP to expose data for harvesting whereas 83% said that they used the software defaults.

Further follow-ups were done to see what content was being collected from another repository. One repository had done a one-off collection of institutional material from arXiv; one repository harvested from arXiv monthly and the third plans to collect material from UKPMC and a couple of commercial services.

The most important repositories that IR managers would like to interact with are UKPubMed Central; ESRC Society today; RePEc/Economists Online and arXiv. There were also mentions of commercial databases such as Web of Knowledge.

We asked what technical factors would influence the decision to collect content from another repository; the relative importance was as follows:

- 1) Ability to identify your institution's output easily;
- 2) Able to get full text as well as metadata;
- 3) Minimal processing after information transfer and
- 4) Easy to implement and use interface.

The policy and cultural factors were:

- 1) Trust in the source repository (high quality content, well managed repository etc);
- 2) Clear understanding of the rights & permissions for transferred material;
- 3) Quality of metadata in the source repository;
- 4) Good match between the metadata supplied and your minimum required fields and

## 5) Well defined versioning policy

On the question of versions, most respondents collected the nearest to publication version as allowed by the publisher, a quarter of the respondents were happy to collect multiple versions, but not all of these were publicly visible.

At the end of the survey we asked for free text comments about relationships between repositories and the common themes were copyright issues, single point of deposit, the need for a national approach and some comments about the technology being immature and that setting up the local repository was of higher importance.

### 4.3 Subject repositories

We consulted two well-established subject repositories: arXiv and eLis. Their collection management policies insist on full text content. They both provide an OAI-PMH interface for other organisations to pull content from them. arXiv have developed a SWORD interface.

Any automated deposits from other organisations would be done on a case by case basis and would have to comply with the subject repository's licenses, policies and metadata standards. It is difficult for them to track whether content is being pulled from them, but it is obvious from the research we have done that this is definitely true for arXiv.

Provenance information on the source of the record retrieved from the subject repository is important and additionally for some repositories usage information for full text items transferred is important.

There may be rights issues around allowing the transfer of full text material to another repository as the depositor may not have granted rights to allow this.

They both acknowledge the problems around author affiliation. eLIS are exploring the AuthorClaim system.

### 4.4 Services

Depot is part of a national level infrastructure. It is still in a formative state but is testing protocols such as SWORD to enable the transfer of full text material to other repositories as appropriate. This transfer of material will be done within a context of a trusted relationship with another party. The Depot identifies the affiliation of the depositor and not any of the other authors.

Intute Repository search is in development and has had metadata consistency issues.

The SHERPA team posed the concept of using institutions to aid funders in monitoring compliance.

### 4.5 Interested parties

A small number of interviewees did not fall into the categories listed above, but are interested and involved in the repository community. There are no general conclusions to be drawn from these interviews.

Much was made by some of these interviewees about the possibilities for repositories if the resource oriented view (and REST in particular) was adopted by system developers. Some interviewees suggested the question of whether the present structure of repositories is right for the job they are doing in the long run?

Another interviewee suggested that subject repositories need not exist in future but could be generated by a domain-based service which ran over the top of fully populated institutional repositories. For this to become possible there needs to be some automated method of identifying the subject of the content of the repositories considered. At present there is not enough consensus on the need and utility of subject classification to achieve this. This point of view does not recognise the needs of the funder organisations to produce metrics associated with their funding. Those who run funder-based repositories do not subscribe to this view and in fact feel that a more centralised model as provided by subject repositories is more sustainable in the long-term.

This set of interviewees were also very interested in the use of technology to solve sticking points in the present system, in particular the potential of text mining and automated metadata extraction to enhance the metadata and searching capabilities. There was also some suggestion that having the digital object alone was satisfactory and the metadata would not be of any use in the near future.

#### **4.6 General findings**

- There is a perception that users prefer to deposit in subject repositories and use them if there is a critical mass of material there.
- There are scalability issues around deposit, especially for organisations which validate metadata to an approved standard.
- Metadata standards are important to repository managers; but it is acknowledged that metadata can't both be simple and meet the needs of the organisation collecting it.
- Identifying versions of digital objects is important. If these are updated as part of an automated process then it is important to know what you have.
- Keeping a record of identifiers and their context will help to preserve the context of the object and may help in version disambiguation.
- Transferring content requires levels of trust between the parties and this is perceived to require individually negotiated agreements.
- Importing content from another source needs to adhere to the quality standards and policies of the ingesting repository.
- More clarity about rights to material is required and a better understanding of what the depositor has agreed to allow the repository to do with the material.
- Identifying organisations and individuals within the process is both important and difficult. There are a variety of sources for this data at the moment, a more standardised approach, noting the wider than the UK component, would be beneficial.
- Some repositories support different stakeholder communities with conflicting requirements.
- Repositories need to fit into the workflow of the user, not the user in the workflow of the repository.
- Performance monitoring is important for almost all the stakeholders in this area.
- Most research is not done by a single person at a single institution and research outputs are no different. Each institution and funder will seek to record the research output and the full text of the material may be available in different locations, possibly deposited by different people and so the potential for different versions of the same digital object to be retrieved in a cross-repository search becomes much greater. This adds confusion to the identification of the most appropriate copy for the end-user.

- It can be acknowledged that not all the available content is at present in repositories and as critical mass has not been achieved it is difficult to predict all the types of interaction that will occur in the future. This can be demonstrated by the results of this study which have been mainly focused on content collection rather than discovery; we are sure this is because there needs to be enough content to discover to enable innovative location tools to be implemented and for end-user requirements to be identified.
- Journal articles are a small part of the research workflow and need to be understood in context of the research process. Publication repositories need to be able to link to other types of material such as data and consideration of linking to more ephemeral objects such as blogs and wikis should be made.
- There is a wish for more communication and discussion within the wider repository community.

#### **4.7 Visions of the future**

As part of our fact finding we asked what the ideal future might include, the following list shows how disparate the visions of the community we spoke to are:

- A seamless environment for researchers providing one place/interface to deposit and retrieve information.
- More IRs talking to each other.
- Links to datasets.
- True Open Access in repositories.
- Publishers working with Libraries and academics.
- Standard approaches to systematic communications at a meaningful level.
- Not experimental software.
- Extensible functionality.
- No more journals, this functionality performed by subject overlays of institutional repositories.
- Rapid prototyping to solve issues quickly.
- More involvement in the workflow of the researcher.
- No duplication of input for researchers.
- The current situation is logical.
- More user engagement.
- One interface pushing a deposit to multiple locations.
- Robust versioning mechanisms to identify items.
- Institutions providing compliance monitoring for funders.

## 5. SOME POTENTIAL FUTURE DIRECTIONS FOR REPOSITORY INTERACTIONS

There is no consensus within the communities about what would constitute an ideal situation for repository interactions, as different stakeholders have different expectations and requirements for repositories. This section is intended to explore possible scenarios for the future evolution of repository interactions. These scenarios are conceived in terms of *drivers*, that is, pressures from needs that are likely to be sufficiently common and well recognised that they will influence the direction of developments. The drivers have all arisen from the interviews with stakeholders, with the exception of the preservation driver, which nonetheless is considered of sufficient potential importance to be included in its own right.

Each driver is associated with a number of *enablers*, current or envisaged developments that can provide a basis for meeting the needs expressed by the drivers. These are not necessarily technological, but also include institutional and community developments. The enablers are not intended to be either necessary or sufficient steps for this purpose, but rather to indicate what is already being undertaken or explored and therefore is likely to have some impact in future.

### 5.1 DRIVER ONE: Population of repositories

It is hard to imagine the increased population of repositories, with metadata records if not with full text, as other than a major driver for the future. Indeed it is one of the premises of the SIRIS study. The motivation of course is diverse: from the moral argument about public availability of research outputs, through easing the research process by having more material instantly available online, to the wish for improved metrics. Achieving critical mass is an important aim for the repository community and will enable others to build and innovate on top of this content.

#### Enabler: Common deposit at point of origin

If it is assumed that the best method of populating repositories is for the same digital object to be deposited in multiple repositories, then one way of achieving that is by a simple deposit tool for use by authors. This might be a desktop tool, providing a “folder” into which a researcher could drag and drop a document/collection of files; it would be set up to send the content to a variety of different repositories. The tool would establish the basic metadata (either by automated means or interrogating the depositor) and then automatically process the digital object. This process could take place at any point in the dissemination phase. Alternatively, the word processing software/document saving process could link deposit into a repository to the workflow within the word processing document in a similar fashion to Electronic Record Management systems processes.

#### Benefits

- The deposit process is embedded in the author’s personal research management workflows and self-archiving becomes an integral part of the research process.
- The mechanism has been technically proven: Simple Web service Offering Repository Deposit (SWORD) is the outcome a JISC-funded project to develop a standard mechanism for depositing into repositories and other systems, and has been implemented in test versions of repository software.

#### Costs & Risks

- Any requirement on authors to produce metadata to minimum standards may be resisted as burdensome.
- As mentioned elsewhere in this report, each repository has a common core of metadata fields, but there are others which are personalised to the repository. It would be unfeasible for a common deposit tool to be configured to take account of all local variations, so that the

imported records and object would have to be amended after the transaction to add additional information; this might have scalability issues.

- The author needs to be engaged with the process and own the tool.
- A generalised nationally branded tool might have up-take issues.
- Uncoordinated deposit on multi-author papers might lead to the potential of version confusion during resource location.

#### Main requirements

- Development of a user-friendly deposit tool.
- Community-wide adoption of minimum standards to ensure interoperability.

#### **Enabler: Requirements for external reporting, such as the REF**

External drivers which require organisational collection of material can impact on the deposition rates amongst academics. The potential effect of the REF in HEIs and other performance collection in Research Council institutes should not be underestimated.

#### Benefits

- External driver of importance to authors and their immediate organisational structure.
- Standard information required, leading to convergence in what is collected in a particular sector.

#### Costs & Risks

- Possibility that the organisation chooses a different system to collect this type of information.
- Possibility of being connected to an external process reduces the local branding.

#### Main Requirements

- Development of REF-related reporting in repository software.

#### **Enabler: Greater use of the OAI-PMH protocol to enable content transfer**

OAI-PMH as a protocol to enable the harvesting of metadata is well established; however our study showed that using this to gain content from other repositories is not as well used as we expected. There are two main reasons for this: the issues around establishing the reuse rights to third party material and the problems of identifying the subset of interest to the harvester.

#### Benefits

- Well established protocol and all repositories have implemented it.

#### Costs & Risks

- Issues around the rights and reuse.
- Difficulties of matching what is easy to produce against what is useful to be harvested.
- Difficulties of identification of authors, funders and institutions.

#### Main Requirements

- Adoption of standard naming conventions.
- Common standards to allow for information interchange.
- Development of trusted relationships to enable this to continue.

### **Enabler: Use of the DC SWAP to provide semantic equivalence for data interchange**

To be effective in transferring or sharing metadata between disparate systems, there needs to be an understanding of what the metadata means/is describing; one approach to solving this issue is the development of Dublin Core application profiles and the Scholarly Works Application Profile is designed for use with journal articles.

SWAP aims to be integrated into repository systems and to become the de facto standard for describing scholarly works. This facilitates the exchange of metadata and, where required, associated files between repositories, and supports the development of other cross-repository services.

#### Benefits

- Standard approach to common metadata fields describing journal articles.
- Inclusion of funder information in the basic set.
- Allows the possibility of semantic equivalence.

#### Costs & Risks

- Low take-up within the community and software developers at present.
- Possibility of competition from a different standard (yet to be developed).
- The level of minimum metadata required by SWAP is greater than the resources available to produce it.

#### Main Requirements

- Development of SWAP functionality in the main repository software providers.
- Adoption of SWAP by the repository community.

### **Enabler: Involvement of publishers at the publication stage**

One of the approaches taken by UKPubMedCentral is for the publisher to deposit the object into the repository on behalf of the author. This reduces the need for author intervention but does not involve them in the process. This approach might be applicable in other subject domains with other publishers. There might be sustainability issues for this approach to work for institutional repositories rather than the subject based approach of UKPMC. However as this enabler was not the focus of this report's investigations it is not explained in as much detail as other enablers are.

#### Benefits

- Involvement from a wider range of stakeholders.
- Possibility of greater deposit rates.
- Greater clarity about version deposited.
- Greater consistency of metadata.

#### Costs & Risks

- Possibility of uneven take-up across publishers and subject disciplines.
- Issues surrounding which repositories content would be deposited in.
- Authors being disengaged from the process.

#### Main requirements

- Discussion with publishers about the potential for deposit.
- Identification of a sustainable model for this approach.

## 5.2 DRIVER TWO: Statistics and metrics

Tracking the usage of resources in repositories is an important development area if they are to form the core of the research infrastructure. Stakeholders may be classified as: individual authors; institutions (HEIs); funding bodies.

There are of course different needs for different stakeholders. Authors are interested in the total usage regardless of location. Within the HEI sector there are different approaches to recording the complete research output of the institution. In some cases the IR is used for this purpose and in other specialised publication databases with additional metric producing tools are being used. There is much interest in organisations adopting the latter approach to put processes in place to share information, mostly metadata in these cases. The same is also true of institutions which have adopted Current Research Information Systems (CRIS).

There have been different approaches adopted by the Research Councils to implementing the Open Access principles and collecting information on research funded by them. However the Economic Impact Reporting Framework (EIRF) for which RCs must produce standardised metrics including those related to outputs and outcomes of funding research projects is common to them all. One standard output and/or outcome is the production of journal articles. The RCUK project looking at this area (OOCF), which reports in mid-2009, is considering the possibility of retrieving some of this publication related material from institutional repositories rather than from the Principle Investigator interacting directly with the funding research council. Whilst this is not a subject to IR interaction, it is of interest the wider community. As it is still in the feasibility stage, it is not possible to ascertain all the requirements of this interaction.

### Enabler: COUNTER project: PIRUS for statistics services

The COUNTER organisation has established widely-recognised standards for the statistical reporting of online journals usage, and is now looking at extending a similar initiative to the level of individual articles hosted in publisher and institutional repositories. Its Publisher and Institutional Repository Usage Statistics project (PIRUS), supported by JISC, aims to *'develop COUNTER-compliant usage reports at the individual article level that can be implemented by any entity (publisher, aggregator, IR, etc.) that hosts online journal articles and will enable the usage of research outputs to be recorded, reported and consolidated at a global level in a standard way'* The project runs from August to December 2008. A key challenge is the difference in application of article identifiers from publisher to publisher depending on whether the format warrants a new identifier.

#### Benefits

- Well known standards body developing new standards for repositories.

#### Costs & Risks

- There may be problems implementing the standard.
- There may be slow take-up of the standard.

#### Main requirements

- Completion of the report.

## 5.3 DRIVER THREE: Preservation

One of the possible methods of achieving long-term preservation is to build a relationship with a specialised preservation service, when these become available, and to transfer content to that service for the purpose of preservation activities. This will be a repository-to-repository interaction, and for

this to work on the large scale there needs to be a shared understanding of the semantics of the content and metadata (or in this context the representation information).

Another possible method is for some of the repositories collecting content to be recognised as preservation and trusted repositories, while other repositories then point to the digital object within that repository whilst owning the metadata they require for their internal purposes.

Both of these scenarios require a high degree of explicit trust between organisations for the outcome to be successful. There may be lessons to be learnt from the Virtual Organisation Community in this area.

#### **Enabler: Developments in audit and certification of trusted digital repositories**

Work is progressing towards an ISO standard on audit and certification of trusted digital repositories. The work is being conducted under the auspices of CCSDS, the Consultative Committee for Space Data Systems, though it will have general applicability beyond this particular field. The aim is to define the criteria that a repository must satisfy to have confidence that it can be trusted over long timescales to preserve the resources with which it is entrusted

##### Benefits

- A repository accreditation body with regulation powers recognised and funded by the community.
- Establishment of metadata harvesting and re-use protocols between repositories.
- Processes to identify authors and institutional affiliations.

##### Costs & Risks

- The accreditation process may not be achievable within existing resources.
- There may be patchy take-up of the standard.

##### Main requirements

- Completion of the ISO standard.
- Awareness of this standard within the repository community.

#### **5.4 DRIVER FOUR: Aggregation of research outputs**

Journal articles are a small part of the research workflow and need to be understood in the context of the research process. Publication repositories need to be able to link to other types of material such as data and consideration of linking to more ephemeral objects such as blogs and wikis could be made. Indeed, some views of repositories stress that the potential richness from such linking is one of their chief opportunities, opening up a new prospect of scholarly communication. Research outputs are viewed as aggregations of material, which do not need to be located in the same place but can be associated virtually.

#### **Enabler: Carrying forward of work already done on linking publications and data**

Joining up the different types of material to give a researcher an easier environment to located information will benefit the research process.

##### Benefits

- All parts of the research process are linked so that discovery and reuse are more easily achieved.

##### Costs & Risks

- There may be issues with preparing the data so that it is in a suitable form for deposition and preservation.
- There may resistance to large scale deposition.
- There are issues with assigning meaningful persistent identifiers to some types of data; for example data collected over a long timescale where the year is not one of the common analysis points.
- Adoption of a common infrastructure to link data to publications when each discipline generates different types of data.

#### Main requirements

- Visible persistent identifiers for both the data and the publication.
- Community wide adoption of a linkage mechanism which would work regardless of which digital object was deposited first.

#### **Enabler: A general model for the description and exchange of aggregations of web resources; in particular the OAI-ORE specification.**

ORE provides a viable alternative model to content transfer , where the focus is not on transporting the same information between different locations but on creating aggregated views of information held at different web locations (or URIs) by relating the information units to each other in a Resource Map, itself identified by a URI. The ORE specification is descriptive, in that it can, for example, provide a structural description of a complex object (e.g. a book) and its constituent parts (e.g. chapter files), and functional, in that it can use this description to 'create' the book by aggregating the component files in their proper relationships. There is no need for these files to be grouped together at the same web address. Such an approach is systemic, viewing the web as a single vast database of information capable of sustaining a variety and complexity of interrelationships.

#### Benefits

- Repository administrators can focus on collecting and preserving content and making it accessible to general web search engines, without excessive concern for metadata richness. The services that create relations between items and contexts in which to view them will develop separately.
- ORE offers an economic model of information traffic, being directed towards creation of links between resource locations.
- ORE offers a flexible and sophisticated model for the description of complex relationships between objects, such as version, alternative manifestation and associated data relationships.

#### Costs & Risks

- Potentially less structured and authoritative views of information.

#### Main requirements

- Adoption of this protocol within the community.

The next section which outlines our recommendations builds on the common themes of the drivers and enablers to be progress repository interactions in a successful manner.

## 6. RECOMMENDATIONS

Our recommendations have been formulated with both the original brief to be practical and achievable and with our findings in mind. They are organised into sections with an indication of the type of action and which stakeholders will need to be involved.

### 6.1 Standardisation

These recommendations are concerned with the creation and adoption of standards to aid information exchange and sharing.

#### 1. Clear identification of authors, funders and higher education institutions

Being able to locate with authority and consistency the identity of a person or corporate body attached to a research output is vital for any repository data exchange process and is hugely important for any service running over the top of repositories. Although different stakeholders may need this information at different levels of detail, it is an acknowledged issue. There is already work being done in this area such as the RIN funder affiliation recommendations and the NAMES project. It is important that standardisation work in this area is taken up and that the different stakeholders do not adopt different solutions to the same problem. Additionally there may need to be advocacy work within the community to ensure authors adopt the standards for both the institution and the funder.

*Audience: JISC, funders, repository managers, publishers and content creators*

Next steps: Consider the outcomes of the NAMES project.

Continue to encourage content creators to follow RIN funder affiliation guidelines.

#### 2. Adoption of information interchange standards

We recommend that a common information interchange standard is adopted. At present the Dublin Core application profile for scholarly works (SWAP) is the current solution. It is important that information for common services is built using agreed standards for the semantic meaning and SWAP is a tool which provides for this. It considers the information needs of a variety of stakeholders and is fit for purpose without extensions at present.

*Audience: repository managers and software developers*

Next steps: Further discussions with software developers about the utility of SWAP.

Outline the way forward for information interchange.

#### 3. Trust in other repositories

Trust in the process and content, and preservation processes are key to repository interactions. We recommend that a watching brief on the Trusted repository certification process is kept and that this is encouraged to be taken up by managers of all types of repository once the process has been completed and verified.

*Audience: JISC and repository managers*

Next steps: Consider the outcomes of the Trusted repository project.

### 6.2 Best practice

These recommendations are concerned with achieving consistency through actions.

#### 4. Provenance information within transferred records

For records transferred from one system to another, there should be visible provenance information contained within the bibliographic record. At a minimum this should include the source repository, the transfer date, rights information and source repository identifiers.

*Audience: repository managers and software developers*

Next steps: Explore the technical possibilities to allow automatic generation of this information.

## **5. Clear versioning identification at object and metadata levels**

The Version Identification Framework guidelines for version identification should be promulgated and adopted, so that both the digital object and the metadata have sufficient information for an end user to understand which version of an object they are looking at. This is particularly necessary in an environment where there are likely to be more versions easily accessible.

*Audience: repository managers and software developers*

Next steps: Continue to embed this within the culture.

Explore the technical possibilities to allow automatic generation of this information.

## **6.3 Community engagement and dialogue**

These recommendations are concerned with the wider landscape and engaging with those who interact with repositories.

## **6. Repository community forum**

It would benefit future developments if there were a UK wide group where subject/funder repositories can meet representatives of the Institutional Repository community for further discussion and agreement on standards and protocols.

*Audience: JISC, funders, repository managers, software developers and content creators*

Next steps: JISC to take this forward by:

- a) Identifying possible members;
- b) Setting an agenda and
- c) Organising the first meeting.

## **7. Continued user engagement**

More work in identifying the needs of end-users and authors should be done to ensure that development is not focused on the needs of those who manage the process, either those within the author's institution or those who fund the work or provide national services. We are aware that this is not an easy task to achieve.

*Audience: JISC, funders, repository managers and content creators*

Next steps: Build on user studies funded by the JISC Scholarly Communications Committee announced recently.

Continue to build bridges with professional bodies representing the users.

## 7. CONCLUSIONS

Whilst there are many opportunities and benefits from subject/funder and Institutional Repositories interacting, there are some barriers at present. One of the most obvious is the difference in length of time that individual repositories have been set up and therefore the differences in the stage of development. Many new repositories focus, rightly in our opinion, on getting the internal interactions right and the repository embedded within the organisation before looking outwards.

The table below shows our recommendations, linked to the drivers discussed in section 5 together with a column entitled Immediacy. This column is designed to indicate which recommendations, in our view, are more likely to make significant progress in the near future. Thus three asterisks indicate a shorter time frame than one asterisk.

<b>Recommendation</b>	<b>Drivers</b>	<b>Immediacy</b>
1. Clear identification of authors, funders and higher education institutions	2, 4	**
2. Adoption of information interchange standards	1, 2	**
3. Trust in other repositories	1, 3	*
4. Provenance information within transferred records	2, 3	***
5. Clear versioning identification at object and metadata levels	1, 2, 3	**
6. Repository community forum	All	***
7. Continued user engagement	All	*

Key to drivers: 1 = Population of repositories, 2 = Statistics and metrics, 3= Preservation, 4 = Aggregation of research outputs

We have discovered a general willingness for the different stakeholders to come together to discuss things of importance to everyone. We feel that this should be taken forward, as dialogue should bring better understanding of the constraints of all parties and should enable the repository community to improve things for the most important stakeholders of all: the authors and end-users.

## 8. BIBLIOGRAPHY AND REFERENCES

- Allinson, J., Francois, S. et al (2008) 'SWORD: Simple Web-service Offering Repository Deposit'. Available at: <http://www.ariadne.ac.uk/issue54/allinson-et-al/> [accessed 27/5/2008]
- ATOM Publishing Protocol. Available at: <http://en.wikipedia.org/wiki/ATOM> [accessed 7/11/2008]
- Charlesworth, A., Ferguson, N. (2008) 'Feasibility study into approaches to improve the consistency with which repositories share material'. Draft final report.
- CLADDIER Project. Available at: <http://www.jisc.ac.uk/whatwedo/programmes/digitalrepositories2005/claddier> [accessed 7/11/2008]
- EIRF: Economic Impact Reporting Frameworks. Available at: <http://www.rcuk.ac.uk/aboutrcuk/eirf> [accessed 7/11/2008]
- Feijen, M., Horstmann, W. et al (2007) 'DRIVER: Building the Network for Accessing Digital Repositories across Europe'. Available at: <http://www.ariadne.ac.uk/issue53/feijen-et-al/> [accessed 27/5/2008]
- Heery, R., Powell, A. 'Digital Repositories Roadmap: looking forward'. Available at: [www.jisc.ac.uk/uploaded\\_documents/rep-roadmap-v15.doc](http://www.jisc.ac.uk/uploaded_documents/rep-roadmap-v15.doc) [accessed 3/11/2008]
- Heery, R., Powell, A. (2006) 'Digital Repositories Roadmap: looking forward'. [Available at: [www.jisc.ac.uk/uploaded\\_documents/rep-roadmap-v15.doc](http://www.jisc.ac.uk/uploaded_documents/rep-roadmap-v15.doc) [Accessed 2/11/2008]
- Jones, C. (2007). 'Institutional Repositories: Content and Culture in an Open Access Environment'. Oxford: Chandos Publishing
- Names Project: Pilot national name and factual authority service. Available at: <http://www.jisc.ac.uk/whatwedo/programmes/reppres/sharedservices/names.aspx> [accessed 7/11/2008]
- OAI-ORE: Open Archives Initiative Objects Reuse and Exchange. Available at: <http://www.openarchives.org/ore/> [accessed 7/11/2008]
- OAI-PMH: The Open Archives Initiative Protocol for Metadata Harvesting. Available at: <http://www.openarchives.org/OAI/openarchivesprotocol.html> [accessed 7/11/2008]
- Puplett, D. (2008) 'Version Identification: A growing problem'. Available at: <http://www.ariadne.ac.uk/issue54/puplett> [accessed 27/5/2008]
- Research Evaluation Framework (REF). Available at: [http://www.hero.ac.uk/uk/research/research\\_quality\\_and\\_evaluation/research\\_excellence\\_framework\\_ref\\_.cfm](http://www.hero.ac.uk/uk/research/research_quality_and_evaluation/research_excellence_framework_ref_.cfm) [accessed 7/11/2008]
- RIN Guidance. Available at: <http://www.rin.ac.uk/about> [accessed 7/11/2008]
- Robertson, RJ, Mahey, M. et al (2008) 'An ecological approach to repository and service interactions. Version 1.2. March 2008'. Available at: <http://www.ukoln.ac.uk/repositories/digirep/images/e/e4/Introductoryecology.doc> [accessed 27/5/2008]
- Robertson, RJ. (2007) 'The repository ecology: an approach to understanding repository and service interactions'. Powerpoint presentation given at 5<sup>th</sup> Workshop on Innovation in Scholarly Communication (OA15), Geneva, Switzerland. 18-20<sup>th</sup> April 2007.
- [Science and Innovation Investment Framework 2004-14: annual report on Research Council output and economic impact frameworks. October 2007: http://www.berr.gov.uk/files/file42023.doc](http://www.berr.gov.uk/files/file42023.doc) [accessed 6/11/2008]

Simpson, P. (2005) 'Institutional Repositories and Discipline Based Repositories'. Grade kick off meeting 28<sup>th</sup> September 2005. Available at:  
[www.edina.ac.uk/projects/grade/ppt/GRADE\\_Simpson.ppt](http://www.edina.ac.uk/projects/grade/ppt/GRADE_Simpson.ppt) [accessed 27/5/2008]

StORe: Source-to-Output Repositories. Available at:  
<http://www.jisc.ac.uk/whatwedo/programmes/digitalrepositories2005/store.aspx> [accessed 7/11/2008]

Storelink. Available at:  
<http://www.jisc.ac.uk/whatwedo/programmes/digitalrepositories2007/storelink.aspx> [accessed 7/11/2008]

SWAP. Scholarly Works Application Profile. Available at:  
<http://www.jisc.ac.uk/whatwedo/programmes/reppres/swap.aspx> [accessed 7/11/2008]

SWORD Simple Web Service Offering Repository Deposit (SWORD) . Available at:  
<http://www.jisc.ac.uk/whatwedo/programmes/reppres/tools/sword.aspx> [Accessed 7/11/2008]

Treasury Green Book: Available at: <http://greenbook.treasury.gov.uk/> [accessed 7/11/2008]

Van de Sompel, H. (2006) 'An Interoperable Fabric for Scholarly Value Chains'. Available at:  
<http://www.dlib.org/dlib/october06/vandesompel/10vandesompel.html> [accessed 3/11/2008]

VIF: Version Identification Framework. Available at:  
<http://www.jisc.ac.uk/whatwedo/programmes/reppres/vif.aspx> [accessed 7/11/2008]

UKRDS: UK research data service feasibility study: Interim report. Version v0.1a.030708. Available at  
[www.ukrds.ac.uk/UKRDS%20SC%2010%20July%2008%20Item%205%20\(2\)...](http://www.ukrds.ac.uk/UKRDS%20SC%2010%20July%2008%20Item%205%20(2)...)[accessed 7/11/2008][