Delivery, Management and Access Model for E-prints and Open Access Journals within Further and Higher Education

A joint report by

The Electronic Publishing Innovation Centre (EPIC)*

in partnership with

Key Perspectives Limited

*EPIC is a research group that brings together experts from the Department of Information Science, Loughborough University and Information & Library Services, Cranfield University]

Authors:
Alma Swan (Key Perspectives Ltd, Truro)
Paul Needham (EPIC, Cranfield University)
Steve Probets (EPIC, Loughborough University)
Adrienne Muir (EPIC, Loughborough University)
Ann O'Brien (EPIC, Loughborough University)
Charles Oppenheim (EPIC, Loughborough University)
Rachel Hardy (EPIC, Loughborough University)
Fytton Rowland (EPIC, Loughborough University)

EPIC:
Department of Information Science, Loughborough University, Loughborough, Leicestershire, LE11 3TU, and
Information & Library Services, Cranfield University, Cranfield, Bedfordshire, MK43 0AL

Key Perspectives Ltd
48 Old Coach Road, Playing Place, Truro, Cornwall, TR3 6ET

CONTENTS

1. EXECUTIVE SUMMARY

This study describes a delivery, management and access model for e-prints and open access journal content for UK Further and Higher Education commissioned by the Joint Information Systems Committee (JISC). The target content is (i) e-prints – digital duplicates of academic research articles that are made available online to permit increased access, and (ii) articles published in open access journals. The proposed service would provide immediate and maximal access to scholarly research, supplementing the more limited access provided by subscription-based journals. This would in turn accelerate and enhance the impact of scholarly research, and strengthen and enrich impact measurement and analysis (generating better scientometric performance indicators for research productivity, usage and impact). It would also enable the generation of standardised online CVs for each institution's researchers: these could be used for internal as well as external evaluation purposes, such as the UK national research assessment exercise, as well as to monitor the fulfilment of any research-council funding requirements. A nationally-organised service in the UK for the delivery of e-prints and open access journal content to the scholarly community would therefore be an important development.

Open access scholarly research material takes two forms – e-prints (either preprints, which are articles at the pre-refereed stage, or postprints, which are articles in their final, peer-reviewed form) and open access journal articles. Open access journal publishers make the content of their journals freely available on the Web; many also have an OAI-compliant interface via which other online services, including OAI service providers, harvest their content. Some open access publishers have their own archives and others deposit e-prints of their journal articles in centralised subject-based archives such as PubMed Central. E-prints, on the other hand, are deposited in open access e-print archives by their authors. E-print archives are growing in number around the world. Some are centralised (and usually subject-based), but most are institution-based, covering all of an institution's scholarly disciplines. In some cases, individual departments have also established e-print archives. The software for creating e-print archives is readily available and is free. So far, the development of e-print archives has been mostly *ad hoc*, although national policies requiring the provision of open access to research articles by self-archiving of e-prints are now being considered in several countries, including the UK.

In the UK about thirty individual institutions have so far created e-print archives but there is as yet no organisation of these developments at a national level and the overall number of such institutional archives remains small in comparison with the number of research-active institutions. Moreover, the archives that *are* being created are not being filled with e-prints quickly enough to provide open access to the bulk of UK scholarly literature. There are political and cultural influences responsible for this slow progress, including inertia on the part of authors, most of whom are still not yet voluntarily self-archiving their work. There are ways in which this inertia may be overcome, including mandating the self-archiving of e-prints of published articles by authors in institutional e-print archives. The recent report of the Parliamentary Select Committee on Science and Technology recommended (in Recommendation 44) mandatory depositing of e-prints in institutional repositories. This mandate could be implemented by the institutions themselves or by research-funders.

This study identified three models for open access provision in the UK: (a) the *centralised model*, where e-prints of articles are first deposited directly into a national archive and then made accessible to users and service providers; (b) the *distributed model*, where e-prints are deposited in any one of a distributed network of OAI-compliant institutional, subject-based and open-access journal archives, whose metadata are then harvested and made accessible to users and service providers; and (c) the model we have termed the *'harvesting' model*, a variant of the distributed model in which the harvested metadata are first improved, standardised or enhanced before being made accessible to users and service providers. In considering the relative merits of these models, we addressed not only technical concerns but also how e-print *provision* (by authors) can be achieved, since without this content provision there can be no effective e-print *delivery* service (for users).

For technical and cultural reasons, this study recommends that the centralised model should *not* be adopted for the proposed UK service. This would have been the costliest option and it would have omitted the growing body of content in distributed institutional, subject-based, and open-access journal archives. Moreover, the central archiving approach is the 'wrong way round' with respect to e-print *provision* since for reasons of academic and institutional culture and so long as effective measures are implemented, individual institution-based e-print archives are far more likely to fill (and fill quickly) than centralised archives, because institutions and researchers share a vested interested in the impact of their research output, and because institutions are in a position to mandate and monitor compliance, a position not enjoyed by centralised archives. The study therefore recommends the 'harvesting' model [(c) above], constituting a UK national service founded upon creating an interoperable network of OAI-compliant, distributed, institution-based e-print archives. Such a service, based on harvesting metadata (and, later, full-text) from distributed, institution-based e-print archives and open access journals would be cheaper to implement and would more effectively garner the nation's scholarly research output. The model also permits further enhancement of the metadata to provide improved features and functionality.

The study also makes a further series of recommendations to address other technical and cultural aspects of the problem filling e-print archives:

- The British Library might provide a (central) e-print archive for authors who have no institutional archive in which to self-archive their work.
- JISC should develop a programme to encourage all research-led educational institutions in the UK to establish e-print archives.
- JISC should work to involve non-educational research-based organisations in the provision of e-prints.
- JISC should develop a programme to persuade researchers to self-archive their research results.
- Research funders should be encouraged to mandate self-archiving of research funded by them and perhaps also to provide a backup archive for those researchers who do not yet have their own institutional or departmental archive.
- The relevant stakeholders (data providers, service providers, software developers) should be identified and encouraged to develop a coordinated approach to providing controlled subject metadata.

2. INTRODUCTION

The brief for this study was to forecast a delivery, access and management model for e-prints and open access journal content within Higher Education (HE) and Further Education (FE). This report presents the results of our work. In the first part we provide some background information on the current situation with respect to e-print archives and open access journals. This section provides the context within which any new initiatives by JISC will begin to operate.

The report then moves on to lay out the issues that have a bearing upon the forecasting work. These issues, in our view, centred around three main themes – technical matters, the preservation of digital research information, and the political and cultural influences that will affect the manner and success of implementation of an e-prints service in the UK. Under technical matters we discuss the main models that could be considered for the delivery, management and access of a UK e-prints service, and we argue for the type of model that we term the 'harvesting' model. Arguments for and against each of the three main types of model are presented. Also in this section, technical issues to do with delivery of e-prints are examined in detail.

Preservation of digital information is a complex area with many implications for an e-prints service. It is discussed in section 5, and is followed by a section that covers the cultural and political issues involved in creating and running an e-prints service.

Our detailed recommendations for the 'harvesting' delivery, management and access model follow in Section 7, and this is accompanied by a brief look into the future – at what direction the technology might take and what the outcome would be for the proposed service. Having decided upon the best model to recommend, we present in Section 8 a series of further recommendations for action by JISC and other stakeholders. We argue that if all these can be agreed and implemented, a viable and sustainable service can be achieved in a relatively short period of time. The last sections of the report comprise a cost-benefit analysis for the proposed service and a risk assessment.

Alma Swan (Key Perspectives Ltd, Truro)
Paul Needham (EPIC, Cranfield University)
Steve Probets (EPIC, Loughborough University)
Adrienne Muir (EPIC, Loughborough University)
Ann O'Brien (EPIC, Loughborough University)
Charles Oppenheim (EPIC, Loughborough University)
Rachel Hardy (EPIC, Loughborough University)
Fytton Rowland (EPIC, Loughborough University)
July 2004

3. THE CURRENT LANDSCAPE AND CONTEXT

The early parts of our report set the context for the study by the project team into e-print archives and the development of the models that JISC might adopt for the delivery of e-prints and open access journal content to higher education (HE) and further education (FE) establishments nationwide in England. The work reported here took place simultaneously with the investigation into scientific publication undertaken by the House of Commons Science and Technology Committee (2004).

For the purposes of this report an e-print is defined by JISC as:
"*… a digital duplicate of an academic research paper that is made available on line as a way of improving access to the paper. E-Prints are divided into* pre-prints *(papers that are circulated before they have been formally approved for publication), and* post-prints *(papers that have been approved for publication*)."

This section of the report presents an overview of the current situation with respect to archives that collect and store academic-related digital objects. It includes data on the overall aims of institutional archives, the numbers of these in existence, the types of structure they have adopted, the software available to run them, what kinds of data item are deposited within them, the data formats used, and various issues to do with the inception and subsequent management of archives. The final subsection here outlines the main issues that the project team has addressed in the course of this study. The full results of this exercise are reported in the subsequent sections of this document.

The open access movement is gathering pace around the globe. There are two ways to provide open access to the research literature. One is its publication in open access journals, or in journals that will provide open access to individual articles if the author pays a fee for this provision. The other is for authors to deposit copies of their articles, either as pre-publication drafts (preprints) or as completed, refereed papers (postprints) in an e-print archive. These archives themselves may take one of two forms: they may be institutionally-based in which case they are referred to throughout this report as institutional archives (IAs), or they may be subject-specific archives that have no nominal affiliation to any institution, and in practice the most successful examples of such archives are mirrored at various sites around the world.

3.1 Open access journals

At the time of writing (July 2004), the *Directory of Open Access Journals* (2004) (www.doaj.org) maintained by Lund University Library has 1148 journals in its list which between them contain 53404 articles. Because it is pertinent to discussion about the characteristics of different subject areas with respect to open access discussed later in this report, we include a breakdown of these journals by subject area in the table below. Note that some journals may appear in more than one subject category.

SUBJECT AREA	NUMBER OF JOURNALS	PERCENTAGE OF THE TOTAL
Agriculture and food sciences	63	5.2
Arts and architecture	29	2.4
Biology and life sciences	140	11.6
Business and economics	26	2.2
Chemistry	38	3.1
Earth and environmental sciences	72	6.0
General works (multidisciplinary)	4	0.3
Health sciences	279	23.1
History and archaeology	35	2.9
Languages and literatures	44	3.6
Law and political science	42	3.5
Mathematics and statistics	65	5.4
Philosophy and religion	42	3.5
Physics and astronomy	40	3.3
Social sciences	215	17.8
Technology and engineering	75	6.2

Publications listed in the *Directory of Open Access Journals*

The publishers of these journals vary widely. In some cases, the journal is published by a research group or department in a university, with little or no overheads to speak of, and with an editorial team that works for no cash payment. In these cases, journals do not charge authors for publication but are able to make their content openly accessible because the economic model involved is extremely simple and effectively cost-free, with the journal published in electronic form only from a university server.

At the other extreme, the *Public Library of Science* set out with a mission to launch journals that compete in every way (content quality, production quality and so on) with the top-ranking toll-access journals in biology and medicine. This operation has dedicated offices, publishes in print as well as electronically, has a salaried, professional editorial, production and marketing staff and does charge authors a publication fee (currently of $1500).

Nevertheless, despite the differences in operational detail across the continuum between these extremes, all open access journals share a characteristic in common – they make their article metadata (title, authors, keywords, etc.) available in a format that is OAI- (Open Archives Initiative) compliant so that they can be harvested by OAI service providers like OAIster, of which there will be more later in this document. In other words, e-prints in the form of open access journal content are available to all and the pointers to them are easily harvestable. That is all that needs to be said here for now but this will be discussed in much more detail in the later section of this report on e-print delivery models.

3.2 E-print archives

E-print archives may take any one of several forms (see below), but they all share the characteristic that they are repositories for author-deposited **preprints** (pre-refereed, pre-publication drafts of scholarly articles) or **postprints** (refereed, published articles). Throughout this report we have used the term 'institutional archive' in preference to 'institutional repository'. Partly this is because in many 'official' names the term archive is used (e.g. Institutional Archives Registry, Open Archives Initiative) and partly because it reflects an activity (authors 'self-archive' their work – they cannot 'self-reposit'). We imply here, then, that we could use the two terms interchangeably if desired. However, we are aware that some people use the term repository to denote something bigger than an e-print archive – an institutional collection of material that contains far more than e-prints, such as grey literature, institutional-specific digital collections and so on. This is a third reason why we prefer to use the term archive here, since the remit of this study was to develop a model for the delivery and management of e-print and open access journal content only.

E-print archives may be institutionally-located and administered, in which they are usually called **institutional archives**, or they may be **subject-specific archives** physically located at a suitable site and, commonly, mirrored elsewhere. Although some content of some e-print archives is restricted to certain users, for the most part the content is open to access by all-comers through the expedient of exposing the metadata in an OAI-compliant format, the overall purpose being to provide open access to that body of scholarly research information. Service providers harvest these metadata which are then presented for searching and browsing by users, who can access the original full-text article (or whatever the source digital object is) by means of the simple click of a mouse.

3.2.1 The number of e-print archives in existence

There are currently (June 2004) 206 e-print archives operating around the world: 22 are demonstration sites, 26 are archives for electronic theses, 9 are e-journal archives, 108 are institutional or departmental archives of research articles and 29 are cross-institutional archives for research article content.

(Source: Institutional Archives Registry: http://archives.eprints.org/eprints.php?action=browse)

3.2.2 The form of existing e-print archives

There are four main forms that institutional archives can take.

- *Institutional/departmental archives*

The contents of these archives are created and stored locally in an archive specific to and limited to one institution.

An example of this type of archive is the University of Glasgow's Daedalus service (http://www.lib.gla.ac.uk/daedalus/index.html). This service accepts a wide range of material including published papers, preprints, technical reports, conference papers, grey literature, project reports and theses from researchers at Glasgow University. It will also accept a variety of document types, including HTML, Rich Text Format, PDF, Postscript, XML DocBook and XML TEI. The archive administrators are actively exploring additional document types which would be appropriate for the University community. In June 2004 there were 183 articles deposited in the archive.

- *Supra-institutional (networked sister institutions)*

The contents are created in member institutions, uploaded to a central archive and stored there.

An example of this type of archive is ***The University of California eScholarship program*** (http://escholarship.cdlib.org/wparchives.html) which collects and stores digital objects from ten UC campuses plus other affiliated research institutions. In June 2004 almost 3000 articles had been deposited. The archive accepts any kind of article, even those not authored by UC faculty (such as papers from a conference sponsored by UC faculty). The preferred format for text articles is PDF. Audio clips, video clips and so on can also be deposited to accompany research articles.

- *Centralised: regionally- or nationally-organised, or subject-based*

Contents are created in individual member institutions which upload to one centralised one.

An example of this type of archive is ***DARE***, the Dutch Digital Academic Archives (http://www.surf.nl). This is a collaborative venture between all Dutch universities to make all Dutch research output digital and accessible. Although it is federally organised, it is in reality a national archive: all sites operate on the same platform and formats and are cross-searchable and interoperable, with agreed policies on content and operation. Very recently, a number of Indian research-based organisations (22 universities and 11 research institutes) have announced plans to set up a network of institutional archives throughout India (http://www.ecs.soton.ac.uk/~harnad/Hypermail/Amsci/3749.html) on the same basis as DARE (i.e. OAI-compliant, interoperable archives).

A subset of this type of archive would be centralised, with additional local implementations. This would have a centralised structure, as above, but individual member institutions might also implement local variants as required, the contents of which would remain locally.

Subject-based archives are also of the centralised type. Existing subject-based archives, such as ArXiv and CogPrints, are generally located where the originator works (see section 6.1).

- *Combination*

A combination of any of the three above.

3.2.3 The purpose of institutional/departmental e-print archives

We have already stated that the overall purpose of e-print archives is to provide open access to the body of scholarly research articles contained therein. For subject-specific archives this remains the single purpose. For institutional or departmental archives, there are other, additional aims which may also apply. Proponents of this type of archive usually put forward five distinct aims for these entities. They are:

- **The self-archiving of institutional research output:** this includes preprints, postprints, theses, dissertations, monographs and so forth. The archive provides a place for researchers to make available the findings from their work so that any interested party may access them.

- **Provision of teaching materials online:** Course notes, lecture notes, practical class protocols, supporting material and specimen examination papers can all be presented online for students to access, or as a means of attracting future students to the institution.

- **Digital collection management:** All kinds of digital content can be stored and online within an institutional archive. Librarians have the opportunity with this technology to manage and organise any digital content of worth to the institution.

- **Digital preservation:** in the same way, digital objects of worth to the institution can be preserved in e-print archives, thus safeguarding an institution's intellectual output for the future.

- **Institutional electronic publication**: an e-print archive is also a means for an institution to publish its output electronically – a digital version of the traditional university press. Output might include journals, books, monographs, technical papers, serial works and so on.

3.2.4 The advantages of institutional e-print archives

A number of reasons have been put forward as arguments in favour of institutional archives. They are rehearsed briefly below.

3.2.4.1 Increased access to published research

Even the libraries with the largest budgets in the western world cannot afford to purchase subscriptions to all the journals they would ideally like to have. Spiralling journal prices over the last decade or so, the Big Deals and dwindling library budgets in real terms have made for difficult decisions for libraries, and the overall result has been a cut in the number of journal subscriptions. End users suffer because they are not able to gain access to the research articles they need.

Institutional archives are seen as a means of increasing access to research articles, since they are searchable by anyone within that institution and, provided they are harvested by a service provider, by other researchers worldwide. In this they fulfil the age-old desire of scholars to share data in the interests of furthering scholarly endeavour.

As well as maximising access, depositing an article in an institutional archive means that other researchers are able to see it as soon as possible. In the case of preprints, readers can see preliminary reports of results long before they are finally ready for publication. In the case of postprints, deposition of the article may be after publication in a journal, though it is often possible for an author to deposit the article as soon as it is refereed and approved for publication: in this case, other researchers can see the article before it finally appears in the journal.

3.2.4.2 Increased impact of published research

It seems that if research articles are available to all, then increased citations should follow. This is, indeed, the case: several studies have now shown that if research articles are made freely available online there is an increase in citations to those articles (Lawrence, 2001; Kurtz, 2004). The latest figures show that if articles are freely available electronically citations can increase up to fivefold (Harnad and Brody, 2004). The establishment and population of institutional archives should therefore benefit the impact of research hugely.

3.2.4.3 Provision of enhanced citation analysis (new measures of impact)

Until recently, the only measure of impact in use was the journal impact factor, calculated annually by ISI for all the serials that are covered by that organisation's abstracting and indexing service. More recently, though, new scientometric measures are being developed and these will provide alternative ways of measuring the impact of an article upon subsequent research endeavour in its field. Software that tracks citations to articles in e-print archives, such as Citebase, can be used to give improved information about an author's impact on his or her field.

3.2.4.4 Provision of a tool for the compilation of 'institutional CVs' and institutional impact (a marketing tool for institutions)

This is perhaps one of the most persuasive points for an institution considering setting up an archive. It provides a permanent record of the scholarly output of that organisation forming an 'institutional CV' for research assessment exercises (RAE). It can also be used as a marketing tool by the institution by demonstrating its scholarly worth and its social and financial value and by using the content to promote its scholarly and teaching endeavours.

3.2.4.5 Provision of a tool for the compilation of individual researchers' CVs and personal impact

In the same way as an archive provides a means for institutions to compile an institutional CV, so it provides the means for CVs of individual researchers to be compiled and maintained. The monitoring of download activity also provides researchers with a measure of the impact of their work.

3.2.5 Software types

Software for e-print archives falls into three main categories: open source software, distributed free to anyone who wishes to use it and thus in use on multiple archives around the world; proprietary or commercially-developed software which has a cost attached to it, and locally-developed, bespoke software written for an individual application, usually by a university which has set off on the track of developing a archive in an independent manner. Institutions should select which software to use based on what the requirements of that institution are – what types of object it intends to archive, its IT capabilities and how it sees its authors' needs. The crucial aspect, if true open access to the archive is desired, is that whatever software is chosen it should enable an OAI-compliant archive to be created. The archive or institution then becomes an OAI data provider and the content of its archive can be harvested by OAI service providers, which then make the content of that, and all the other IAs they harvest from, available to anyone. OAI service providers have different regimens for polling data providers, but within a short time of an article being deposited in an IA it should be available for access via all the service providers.

3.2.5.1 Open source software

Several examples of open source software for e-print archives now exist. The best known, and the most used, are Eprints, developed at Southampton University and DSpace, developed by MIT. These are freely available to any institution that wishes to set up an archive. Eprints focuses specifically on digital items of e-print type in Postscript, PDF, ASCII and HTML formats, while DSpace permits the archiving of items of many heterogeneous types, making it suitable for institutions that see a need for presenting and preserving scholarly output in a range of formats. Both Eprints and DSpace offer interoperability via the OAI-PMH (Open Archives Initiative Protocol for Metadata Harvesting). DSpace uses persistent identifiers that, unlike ordinary URLs, do not change when the physical location of the digital item alters.

Other OAI-compliant software systems of note are CDSware, developed by CERN; and Fedora, developed jointly by the University of Virginia and Cornell University, with funding from the Andrew W. Mellon Foundation

The CERN Document Server Software (CDSware) is software developed and run by CERN, which has MARC21 as its underlying bibliographic standard.

Fedora - Flexible Extensible Digital Object and Repository Architecture – is an open-source digital object repository management system that *"demonstrates how distributed digital library architecture can be deployed using web-based technologies, including XML and Web services."* (Fedora, 2004)

3.2.5.2 Proprietary software

One example of a proprietary IA software package is the one from Ebrary, a commercial aggregator that has developed IA software as part of its offering to clients, who normally subscribe to its database products. The software permits institutions to create archives that contain e-prints, theses, technical reports, articles, curricula guidelines and special collections (www.ebrary.com). Another example is BePress, developed by the University of California at Berkeley (http://www.bepress.com).

3.2.5.3 Locally-developed software

There are also numerous individual, locally-developed IA software packages. Some examples are MPG eDoc developed by the Max Planck Gesellschaft; OPUS (Online Publications University of Stuttgart); and MyCoRe, developed by a consortium of universities originally led by the University of Essen. So long as these packages are OAI-compliant (i.e. the contents can be harvested by all OAI service providers) it does not matter which package is implemented on the ground in any one institution, if the primary goal is to make the IA's content accessible by all-comers. Of course, if the primary goal is not open access to all-comers, but access within the institution only, then the condition that the package is OAI-compliant does not apply.

3.2.6 Data objects collected and stored

There is considerable variation between archives in the digital items collected and stored. Some concentrate on e-prints alone, while others extend the content to cover other types of data item deemed desirable for local requirements. For this project, JISC specified that it requires models for the access and delivery of just two types of digital object – e-prints and open access journal articles (highlighted in the list below). This overview covers not only these but other types of object, too, because it seems likely that any model JISC eventually adopts will also be one where the inclusion of these other types of object is deemed desirable. Some examples of the other types of digital item that might be stored in IAs set up by UK educational institutions are:

- **Preprints**
- **Postprints**
- All drafts and working papers plus corrigenda (i.e. a trail from first draft to the postprint: this is sometimes referred to as the 'low threshold' model)
- Ancillary data from research, e.g. video, audio, large datasets. Some archives accommodate this type of data, which cannot be published in a traditional peer-reviewed journal, because there is merit in it being made available to other researchers, and in being preserved digitally in a formal way
- Books and monographs
- Non-published digital objects:
    - Teaching materials
    - Collections (music, images, etc)
    - Research output from specialised subject fields, such as performing arts, where output is usually in the form of performance, video or audio
    - Dissertations and theses
    - Multimedia items
    - Local institutional 'events', e.g. performances, lectures, exhibitions
- **Open access journal articles**

3.2.7 Data formats permitted

The range of data formats permitted varies from one archive to another. The eScholarship archive at the University of California, for example, uploads text-based documents only in PDF format, though it will accept data in a restricted range of other formats for conversion. Another example, eprints@Glasgow accepts (and stores) digital objects in a much wider range of formats (see Section 2.1).

As well as the digital item format, there is also the issue of metadata type, structure and granularity. Metadata in the simplest form might be considered optimal, so long as it is OAI-PMH compliant. In some cases, however, depending on local conditions, institutions may prefer to opt for something more complex, to provide a richer metadata source and to increase its granularity. The greater the granularity, the better, theoretically, the retrieval capability will be. The converse of this is that increasing complexity always brings the danger of obscurity, and of difficulty of use.

From a preservation point of view, it would be better to limit the number of formats archives are willing to accept. However, this policy should not be so restrictive as to discourage depositors from submitting their material. The archive may have to convert some submitted material into a more easily managed format. Another approach is to make it clear to depositors the formats they can expect to be supported and those which the archive cannot guarantee to keep accessible.

The recommendations of James *et al.* (2003) were to:

- Recognise the preservation risks of different file formats
- Use open, standards-based file formats
- Investigate the use of XML formats to describe data and metadata
- Maintain file format information (this could be held centrally – a task of the Digital Curation Centre (http://www.dcc.ac.uk/))
- Plan for migrating rare and obsolete file formats
- Include file format identification functionality in e-print archive software

3.2.8 The global picture

There is considerable variation at the moment around the world with respect to the degree of organisation of e-print archive policies on a national basis. National (or at least supra-regional) policies look like being agreed in Norway (Hauge, 2004) and India (Arunachalam, 2004).

In Norway, where there are only four research-based universities, three universities already have institutional archives running or planned. Work is now being carried out on building a framework of new national research-policy strategies and requirements with the aim of bringing into being a new pattern of scholarly communication.

In India, several publishers (including the Indian Academy of Sciences, which publishes ten journals) have adopted an open access model, with government grants and subscriptions to the print versions covering publishing costs for authors. Moreover, institutional archives are being set up around India, encouraged now that the Ministry of Human Resource Development (MHRD) has set up the 'Indian National Digital Library in Engineering Sciences and Technology (INDEST) Consortium' (see references in Arunachalam, 2004). Sixty-four Government or Government-aided engineering colleges and technical departments in universities have joined the Consortium. The MHRD has advised all the consortium members to set up e-print archives using appropriate OAI-compliant e-print software. MHRD has also recommended that a central server may be deployed to harvest metadata from all such e-print archives.

In Australia, one of the countries furthest forward with respect to e-print archives, the government gave funds of AU$12m last October to make 'Australia's research information….more easily accessible and better managed.' (McGauran, Acting Minister for Education, Science and Training: see McGauran, 2003). The country's major research universities all have institutional archives, and developments in this field have been supported throughout by the Department for Education, Science and Training (DEST). Lobbying from supporters of digital repositories and university librarians led to the acceptance that a national linked-up approach would be best. A DEST working party carried out a scoping study for e-print archives and the AU$12m now supports four projects covering 15 Australian universities, Australian and international libraries, representatives from industry and various international organisations. The Australian Partnership for Sustainable Repositories (APSR) has now been set up and is working through the Australian National University's Centre for Sustainable Digital Collections to develop a national research infrastructure through broad, archive-based architecture. This will ensure access continuity and the sustainability of digital collections, and facilitate national coordination and international linkages.

The Netherlands government has provided financial support for DARE (Digital Academic Repositories), a collective initiative by the Dutch universities, Dutch national libraries and the Netherlands Organisation for Scientific Research, to make all their research results digitally accessible (van Westrienen, 2002).

Elsewhere, consortia of institutions or libraries are developing policies of their own. These may be national in scope, as in the case of the Canadian Association of Research Libraries, CARL (CARL,2003) , whose members aim 'to implement institutional repositories as a coordinated and integrated strategy to aggregate the digital research output of their academic institutions'. Alternatively, they may be supra-national, such as the initiative by the International Scholarly Communications Alliance (ISCA), which has announced a collaborative programme by its members to develop, expand, and leverage initiatives to transform the scholarly communications process, including strategic and advocacy programs including… the establishment of institutional and discipline-based archives that allow public access to content and employ the Open Archives Metadata Harvesting Protocol (Ayris, 2002).

In the UK, the situation is promising but consists of a series of linked pilot projects and a number of already-established institutional e-print archives rather than a coordinated national approach. There are currently 29 e-print archives, of which 17 are institutional/departmental archives, four are cross-institutional archives, four are demonstration sites, three are e-journal publication archives and one is an e-theses archive. At May 2004 there were just short of 25,000 articles in the 20 e-print archives harvested by the RDN/e-Prints UK project (http://eprints-uk.rdn.ac.uk/stats/). One of them had no articles at all and, at the other extreme, over 12,000 articles resided with the open access publisher BioMed Central. By far the best-populated archive is the University of Southampton's ECS EPrints service with 8143 articles. (http://eprints.ecs.soton.ac.uk/perl/oai2). Three other Southampton-based archives also had reasonably high numbers of articles: e-Prints Soton (http://eprints.soton.ac.uk/perl/oai2) had 758, Psycprints, a subject-specific archive (http://psycprints.ecs.soton.ac.uk/perl/oai2) had 720, and the largest and longest-established, CogPrints, another subject specific (cognitive science) archive (http://cogprints.ecs.soton.ac.uk/perl/oai2) had 1987. As yet, there is no involvement in e-print archiving by the British Library, but there are two particularly significant national initiatives operating.

ePrints UK: The ePrints UK project is concerned with developing a series of national, discipline-focused services through which the higher and further education community can access the collective output of e-print papers available from compliant open archive repositories, particularly those provided by UK universities and colleges.

Discipline-focused views of available e-prints will be provided through the use of an automatic subject-classification Web service offered by OCLC. The project will also use 'name authority' and 'citation analysis' Web services (offered by OCLC and the University of Southampton respectively) to enhance the metadata harvested from available archives. So far, the name authority service has had limited success. Licensing issues also need to be sorted out before the service becomes fully operational.

ePrints UK are using the ARC OAI-PMH toolkit to harvest metadata into a Cheshire II database with WebCheshire providing the user interface. (see http://www.rdn.ac.uk/projects/eprints-uk/docs/technical/architecturev1.032003/)

Significantly, ePrints UK already harvest metadata from a number of e-journal repositories, demonstrating that integration of metadata from journal articles and e-prints is a practical and achievable proposition.

SHERPA: The SHERPA project, funded by JISC and CURL, aims to investigate issues to do with the future of scholarly communication and publishing. In particular, it is initiating the development of openly accessible institutional digital repositories of research output in a number of research universities. These e-print archives will contain papers by researchers from the participating institutions. The project will investigate the intellectual property rights, quality control and other key management issues associated with making the research literature freely available to the research community. It will also investigate technical questions, including interoperability between repositories and digital preservation of e-prints.

FAIR: The Focus on Access to Institutional Resources programme, funded by JISC, 'aims to evaluate and explore different mechanisms for the disclosure and sharing of content (and the related challenges) to fulfil the vision of a web of resources built by groups with a long term stake in the future of those resources, but made available to the whole community of learning.' The JISC Information Environment is envisaged as a virtual place where members of colleges and universities can deposit and share useful content (eg, research outputs). The current collection of JISC funded content has the potential to grow to embrace both externally generated content from publishers and aggregators and community-generated resources. To achieve the latter, staff and students will need a 'place' or 'places' in which to lodge suitable content and products and a means for exchanging and adding to it. The FAIR programme has been developed to create the mechanisms and supporting services to allow this process to prosper and these 'places' to be built. The work of this programme has been inspired by the success of the Open Archives Initiative (OAI).

3.2.9 Issues for examination

The project team identified a list of issues that needed to be examined in order that sensible, workable and viable models of eprint/open access journal archives could be developed for JISC's consideration. A brief consideration of each of these is presented below. The issues divide into two groups – technical aspects of the potential models, and cultural/business/management aspects.

3.2.9.1 Technical issues

- *Delivery system structure:* Would a centralised archive collecting content from all HE and FE establishments be the best option, or would it be better to institute a network of independent IAs with a central OAI service provider harvesting from them, collating and presenting the collective content to data seekers?

- *Software:* Which of the available software packages might be the best to run the system on; alternatively, might the best solution be for a new, bespoke package to be developed?

- *Preservation policies:* There are also many issues concerned with preservation of material in archives. One example is what should happen if an author wishes to withdraw an article – should it disappear altogether, or should some metadata 'marker' be left in place? Another is how to handle and track repeated revisions of an article after it is first deposited – should these all remain visible to readers, how should the trail be recorded, what constitutes the final version and how can this be indicated?

- *What costs and resources will be involved in establishing a archive?* From this technical viewpoint, the main costs will arise from the initial outlay on IT equipment and staffing, and from ongoing costs for the same categories. JISC will wish to have estimates of such costs for the models proposed by the project team.

3.2.9.2 Political, cultural and business issues

Pinfield (2003) has suggested that cultural change will be necessary before self-archiving becomes the norm; in addition, there are a number of more concrete issues that will need to be addressed.

- *Institutional attitudes to archives:* Although some institutions in the UK have already established archives, most have not, and even where there are vociferous advocates for such entities within an institution, support from individuals in high places that would guarantee a measure of success for the fledgling archives has not always been forthcoming. We needed to understand better why this situation pertains and what would be needed to persuade HE and FE establishments of the merits of open access to scholarly material.

- *What is the best way to ensure that the archives are populated?* The existing IAs in the UK are, in the main, sparsely populated with digital objects. There have been several reasons put forward for this, concerning author behaviour. One reason may simply be lack of awareness on the part of authors of the opportunity to self-archive in an archive, but this is a relatively straightforward matter to deal with. More complex are other cultural issues: the reluctance of authors to deposit results in existing archives could be because of concerns over their technical ability to prepare and upload their documents, over copyright, over matters to do with quality control, laziness, other priorities and general inertia, and so on. The question of whether authors themselves should be relied upon to deposit their own material, or whether an institution should organise additional resources, such as a dedicated staff member, for this purpose will be addressed, as will the allied issue of mandating the deposition of research output.

- *What sort of agreements will the archive(s) have with authors?* Most IAs have adopted a policy of establishing non-exclusive agreements with authors who deposit documents. In this way, authors are not prevented from using the content in other contexts and this seems to satisfy both parties. There are other issues to do with deposition rights, however, including the thorny one of helping authors around restrictive licensing arrangements imposed by publishers of journals in which their work appears. Whilst a substantial proportion of journal publishers now permit self-archiving (Eprints.org, 2004) not all do, and some very significant publishers retain copyright agreements with authors that expressly forbid self-archiving postprints, a measure that will hinder the population of archives. Nottingham University has developed a standardised agreement for authors to use with publishers, but most authors still sign the default agreement provided by the publisher.

- *Deposition policies:* As well as deciding what software and formats the archive(s) should be prepared to handle, there are other matters to consider if an archive is to function well. Should deposition be made mandatory, for instance, to maximise the rate at which archives are populated, and their effectiveness? With or without mandatory obligations, at what rate might archives be expected to be populated, and what implications will these rates have for the running costs?

- *What costs and resources will be involved in establishing an archive?* From this management viewpoint, the main costs will arise from the staff resources required for planning, promoting and training when the archive is set up, plus the ongoing costs associated with continuing training, advocacy, marketing, development and business planning for the archive. JISC will wish to have estimates of such costs for the models proposed by the project team.

- *Legal issues:*: Educational institutions that use their own archives to carry out their own scholarly publishing activities may be exposed to legal risks, such as defamation.

- *Publishers' attitudes:* Open access publishers should be amenable to the inclusion of their content in any new archives: what needed to be addressed were the details of what might be required from them in terms of formatting or converting data, timing of activities, and so on. Non-open access publishers are a different matter, and their willingness to cooperate, and the degree to which they will do so, had to be assessed. Some of these publishers are already 'green' in that they permit articles published in their serials to be self-archived (Eprints.org, 2004). Others have approached open access in another way, by launching experimental open access journals or adopting a hybrid model, such as for the *Proceedings of the National Academy of Sciences* (Cozzarelli, 2004), where they offer authors the option to pay for their articles to be published in return for those articles being made open access immediately after publication. We needed to ascertain the state of affairs across the scholarly publishing world in general, how things were moving, and what prospects there might be regarding the inclusion of journal content in the new archive(s).

4. TECHNICAL MODELS AND ISSUES

4.1 Service models

The most likely outcome from the continuing development of the two methods of creating open access to scholarly journal articles – open access publishing and self archiving – is the formation of a number of distributed archives of e-prints or other digital objects (resources). In the terminology of the Open Archives Initiative, (an initiative for facilitating interoperability of distributed content), these distributed archives are known as ***data providers***. Because they are distributed, federated access to these archives is needed to provide the services and interfaces required by readers. In general terms, there are three basic models which could support access to metadata (bibliographic records) and the associated scholarly digital resources:

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

These models are discussed in relation to a service incorporating e-prints and open access journal articles. The protocols and standards mentioned in this section are described later in the appendix to this report.

4.1.1 The centralised model

Under this model, authors would deposit their e-prints in a central archive. The service would have a service provision component of its own, which would provide the interface through which readers would search, browse and retrieve articles. The metadata from these articles would also be exposed via OAI-PMH, SRW/SRU and RSS for use by other service providers.

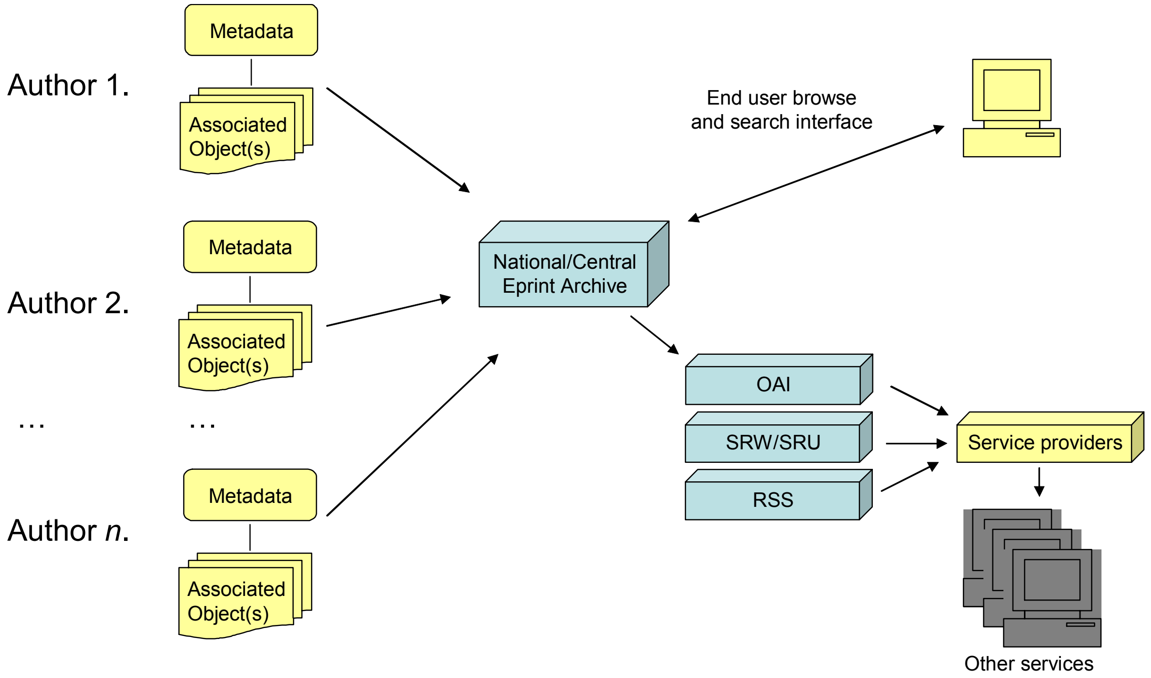

Figure 1. The centralised model

The advantages of this model are that the agency running the service would:

- have overall administration of the whole process, from article deposition through to the user interface
- be able to standardise the protocols used
- be able to select the archive software that provided the most appropriate set of storage and output capabilities
- be able to manage preservation issues
- be able impose requirements for the format in which articles are deposited
- be able to develop facilities that maximised search capabilities (categorisation of the data, subject classification, etc.)
- be able to establish an overall programme of continuing development and improvement

The disadvantages are that:

- With all administrative and maintenance functions centralised, it is an expensive option
- It ignores the existence of, and renders useless, already-established institutional and subject-based archives
- Creating a scheme for nationwide author deposition articles within or across disciplines in one central pandisciplinary archive, or multiple central disciplinary archives, would be extremely difficult if not impossible, for political or cultural reasons (see section 6.1)
- It is not reasonable or practical to expect open access journal publishers to submit articles they publish directly into a central archive

These disadvantages make this option entirely unsuitable for any proposed national service incorporating e-prints and open access journal articles.

4.1.2 The distributed model

In this model proposed services would obtain metadata in real time, as the user asked for it, and point the user at the digital resource which would be located in a distributed archive. The service would cross-search all available archives, using the Z39.50 protocol, or SRW/SRU, and present the results to the user.

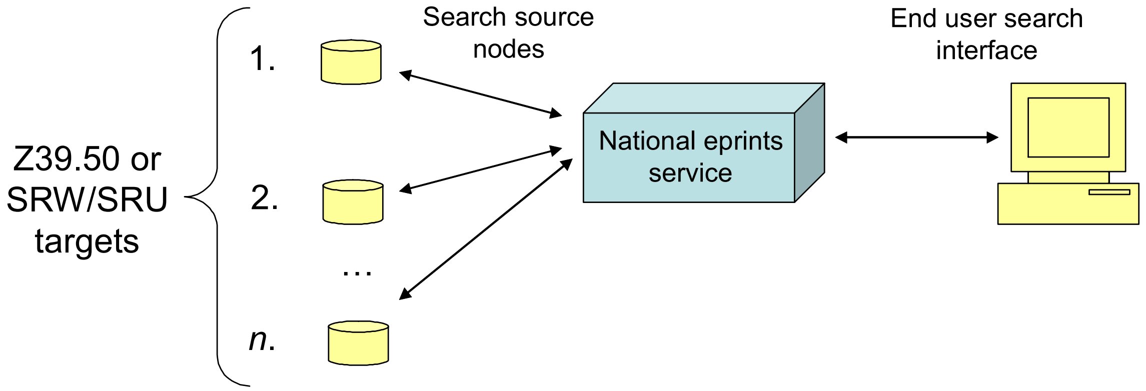

Figure 2. The distributed searching model

The advantages of this model are:

- There is no replication of metadata required
- The metadata retrieved are always current
- It provides a consistent look and feel for searching and retrieving metadata from heterogeneous sources
- It is relatively cheap to implement compared to a centralised solution

The disadvantages are:

- The model does not permit any improvements to be made in the management of e-prints and open access journals
- It does not permit enhancements to the metadata, because these are only grabbed at the time of need (when the user searches)
- As the number of sources to be searched increases, performance decreases – it can only work as fast as the slowest server in the group of archives it is searching
- Query syntax varies across source nodes, and syntax changes over time
- If results are to be returned using relevance ranking, it is difficult to merge results from multiple sets in a meaningful manner
- The institutional and subject-based archives employ software that supports the OAI-PMH. At the time of writing this (July 2004), the vast majority of archives do not support Z39.50 or SRW/SRU.

Again, the disadvantages make this option entirely unsuitable for any proposed services incorporating metadata from e-prints and open access journal articles.

4.1.3 The harvesting model

Under this model, the proposed service (the **service provider**) would harvest and store metadata from available e-print archives and open access journals (the **data providers**), using the Open Archives Initiative Protocol for Metadata Harvesting (OAI-PMH). The service would have a service provision component of its own, which would provide the interface through which readers would search, browse and retrieve articles. The metadata from these articles would also be exposed via OAI-PMH, SRW/SRU and RSS for use by other service providers.

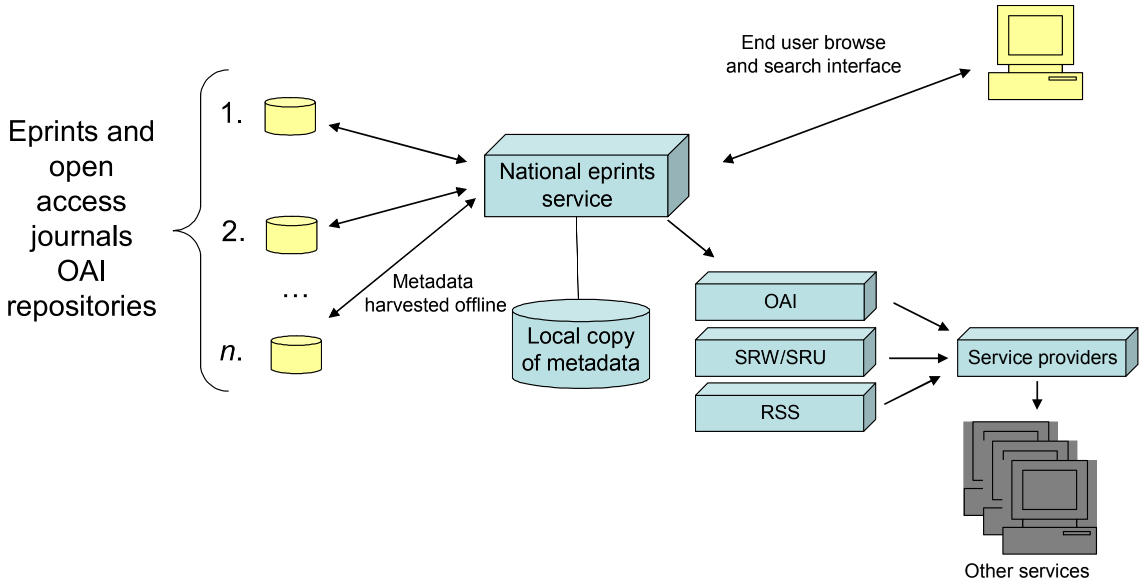

Figure 3. The harvesting model

The advantages of this model are:

- The OAI-PMH is a standard protocol which is easy to implement
- It is flexible – although the use of unqualified DC is mandated to be OAI-compliant, additionally other richer, more complex, metadata schemes may be employed
- The OAI-PMH is designed to allow metadata exchange and the sharing of scholarly knowledge
- The institutional and subject-based archives employ software that supports the OAI-PMH.
- Much of the harvesting can be carried out by automatic scheduled tasks, minimising the need for human intervention

- Once stored in a local database, the metadata can be processed, enhanced and re-exposed both to the original data providers and to other service providers
- It is possible to develop facilities that maximise search capabilities (categorisation of the data, subject classification, etc.)
- It can form the basis for an overall programme of continuing development and improvement
- It is a low-cost option which can work equally well for journal articles, e-prints, journal descriptions and collection level descriptions.

The disadvantages are:

- Unqualified DC, which is mandated as the minimum metadata standard for use by the OAI, is the only metadata scheme in common use as yet. It is a lowest common denominator which lacks semantic richness and limits the possibilities of providing enhancements
- The metadata exposed by the service may not always be the very latest version of that metadata. Changes made to metadata at the IAs, SBAs and OAJs will not be reflected until a subsequent re-harvest

In this model, it is clear that the advantages heavily outweigh the disadvantages. The OAI employs a philosophy whose time has come, and the harvesting model has gained worldwide acceptance. It makes it easy to share information about scholarly resources and to offer enhanced resource discovery tools, and it is being adopted by thousands of institutions and organisations.

In view of this, we recommend that the harvesting model should be adopted to serve as the basis of proposed services. Having laid out this case, in order to set the scene for what follows here, we take it up again in Section 7, where we discuss the proposed model and its implementation in more detail.

Achieving a critical mass of resources in e-print archives is a key task if these archives (and services based on the metadata contained within them) are to become viable sources of information for the UK research community. Much work has already begun to develop and promote institutional archives of resources with associated metadata that can be harvested and used by service providers to provide innovative interfaces to the literature. In the majority of these cases the resources themselves remain distributed through institutional and subject-based archives and the metadata are harvested by service providers, and used as a basis for building access services. These services invariably ‘point’ users to the original data providers where the resource resides, and users download the resource from there. Standards for the harvesting of metadata have been developed by the OAI. These standards are not limited to institutional archives, enabling open access publishers or other information providers to use similar techniques to expose metadata describing their resources.

In a model where services are based on harvested metadata, the quality of metadata harvested is of paramount importance, as it is these metadata that form the limiting factor in determining the level of service that can be provided. In certain instances, resources can also harvested by service providers (e.g. ePrints UK, Citebase): "ePrints UK service retrieves both metadata records and the full text document (in whatever format is available) from the ePrint archive", (see http://www.rdn.ac.uk/projects/eprints-uk/docs/technical/dataflow/). While the full text documents are discarded after analysis, this approach enables the service provider to analyse resources to enhance the quality of the metadata on which they base their service. Citebase, for instance, analyses the reference list within an article to provide a citation analysis service. Although this is an option as part of a national service, and is necessary for certain functions, providing relevant and useful metadata is the most important fundamental element of any service.

It should be noted that e-print archives can be categorised into two main types – institutional archives and subject-based archives – though a third type, personal archives, may also be encountered. It is necessary that any environment for providing access to OA literature must be able to handle data from all of these and from open access publishers.

4.2 Content of e-print archives covered in this study

This report looks at models for managing access to the following categories of resources:

- E-prints
- OA Journal articles

As already noted, an e-print is defined by JISC as:
"… *a digital duplicate of an academic research paper that is made available on line as a way of improving access to the paper. E-Prints are divided into* pre-prints *(papers that are circulated before they have been formally approved for publication), and* post-prints *(papers that have been approved for publication*)."

This is similar to the definition provided by Eprints.org :
"Eprints are the digital texts of peer-reviewed research articles, before and after refereeing. Before refereeing and publication, the draft is called a preprint*. The refereed, published final draft is called a* postprint*."*

This is a slight over-simplification as e-prints may also include working papers, technical reports and so on. This 'grey literature' is also included in the content of many data providers' archives; in the context of this study, however, we have not investigated grey literature apart from where its presence impacts on the nature of information provided to users of a service.

It is assumed that open access to the scholarly literature will be required for all subject areas being researched in HE & FE institutions, and this has implications for the best method of providing rich metadata on which to build services. However, before considering the important options for metadata and subject classification on which to base the services it is necessary to consider the protocols, standards and software that enable the harvesting of metadata from a range of disparate data providers into a centralized service, and the possible methods of making services available.

With respect to publisher content, the 'new' open access publishers, such as BioMed Central and PLoS, already expose OAI-compliant metadata ready for harvesting by service providers. Individual publishers who have adopted the hybrid model (for example, the National Academy of Sciences) expose the OAI-compliant metadata of their open access articles on their own websites. And a very recent announcement by the Directory of Open Access Journals says that article-level searching will shortly be available for the whole of the 1100+ journals in this service and that 'The database records are freely available for reuse in other services and can be harvested using the OAI-PMH….. The article level records will be available for harvesting within two months' (from the time of announcement, June 2004) (DOAJ, 2004).

5. E-PRINT AND OPEN ACCESS JOURNAL CONTENT PRESERVATION ISSUES

5.1 The definition of preservation

Preservation of information includes the technical, financial and managerial issues involved in maximising the useful life of information. It does not necessarily mean keeping it forever or keeping it in its original format. Preservation is an integral part of the management and provision of access to information; it underpins access and cannot be separated from it.

Traditionally, the management of preservation has involved assessing and identifying the preservation needs of and risks to information collections. It is about provision of a suitable and secure physical environment, minimising damage from use of information, repairing damage and perhaps even reformatting content as its physical carrier becomes fragile. All this is equally true for digital information of whatever kind, although the meaning of terms such as "damage" and "use" and actions taken in digital preservation may differ.

Preservation policies should flow from the mission of the organisation or organisations that are responsible for the information to be preserved. Preservation policies are, in essence, a statement of what is to be preserved, for how long, how and by whom. When considering the preservation of e-prints and open access journals, the following questions should be considered.

- *What is the nature and purpose of the collection of digital material?* E-print collections could be organised in various ways. While the prime aim of e-print archives is to improve access, institutional and subject-based archives will differ in focus and this difference should be reflected in preservation policies. The focus of institutional archives is to improve the visibility, accessibility and impact of the output of the institution, whereas subject-based archives are aiming to improve access to research in particular subject areas. Institutional archives may manage material on the basis of individual researchers, research groups, or departments, whereas subject archives may focus on reflecting the development of research areas. Preservation decisions and decisions makers may differ.

- *How will e-print archives and open access journals relate to each other?* There are different possible models for the management of e-prints and the metadata associated with them. The degree of distribution, centralisation and duplication of e-prints and metadata, and existing preservation mechanisms and institutions, will affect the development and content of preservation. Individual archives may develop their own policies, but will have to take into account that individual e-prints may be deposited in more than one archive and consequently take cognisance of the plans of other archives, open access publishers and preservation institutions such as research and legal deposit libraries. Alternatively, preservation policy could be developed in a more centralised way to assist the coordination of preservation actions by these different players.

5.2 Preservation policies

Before the issue of *how* e-prints and open access journals should be preserved can be discussed, it is necessary to consider whether they should be preserved at all. The following section on preservation selection criteria will discuss priorities for preservation and how long different types of material should be preserved. This is followed by a discussion of responsibility for preservation.

5.2.1 Preservation selection criteria

If e-prints and electronic articles are included in archives and open access journals in the first place, then it is reasonable to assume that they should remain accessible for a period of time. Decisions have to be taken about how long these resources should be kept accessible and these decisions are likely to vary according to the following factors:

- The status of the material – preprint, postprint, open access article, accompanying material (e.g. additional data, corrigenda, records of the peer review and development process of a preprint)
- The value of material over time

It is reasonable to assume that formally published research articles should be preserved for posterity as a record of research output. If these articles are supplemented by additional material, then decisions have to be taken on how important these are and whether they also need to be preserved for the long-term. Decisions will have to be taken on how long e-prints need to be kept accessible. If post-prints are mere duplicates of formally published material, then the long-term fate of published articles and post-prints need to be considered together. If long-term access to formal articles is secure, then the long-term preservation of post-prints in archives is perhaps less important.

At present, however, the long-term preservation of articles in electronic form is far from assured. Thus some preservation work for postprints should also be considered as a safety net, should the original electronic-only journals become unavailable or seek to charge an unreasonable amount for the maintenance of back files. . For the short to medium term, preserving postprints should not be too costly, and would avoid the costs involved in making decisions on which postprints to keep and which to delete.

As far as preprints are concerned, the issues to be considered are:

- Whether they are the precursor of formally published material or not
- Whether preprints and records of the peer-review and the development process have any intrinsic long-term value

If preprints are earlier versions of formally published work, then it is perhaps less important to keep them accessible for the long term. However, in some cases both the preprints and the records of the development process may have enduring value, for example in the cases of an author who becomes eminent, a controversial subject or a newly emerging subject. Finally, decisions have to be made about preprints that do not result in formally published articles and where the development process is of little interest. These preprints may not be worth preserving in the long term.

5.2.2 Responsibility for long-term preservation

The following section discusses the various issues surrounding responsibility for long-term preservation, including existing preservation mechanisms and how e-prints and open access journals would fit into these. For the purposes of this section, long term means indefinitely although, as will be discussed below, in the digital environment long term could mean a few years. It also includes discussion of the different roles of players such as e-print archives, publishers, libraries and data archives. This discussion takes into account the different potential models for the organisation of archives and open access journals.

5.2.2.1 Open access journals

Traditionally, scholarly journals have been preserved through various mechanisms. Individual libraries retain and maintain their paper journal collections for as long as they are deemed to have some value to researchers. The UK legal deposit libraries provide last-resort long-term access to journals for researchers and scholars, particularly the UK-published journals that they collect under UK legal deposit law. There is no reason why the output of electronic journals, including open access journals, should not be preserved in the long term as the formal record of the UK's research output on the same basis as printed journals.

However, the UK's legal deposit regime does not cover electronic publishing, although there is enabling legislation in place to allow the scope of legal deposit to be extended, and there is an intention to draw up regulations for the inclusion of electronic publishing. Since 2000, there has been a voluntary system in place and some publishers have been depositing electronic journals on physical carriers or through online means. The British Library has experimented in harvesting Web-based material, including journals, and has started working with electronic publishers.

Since the regulations for the legal deposit of electronic journals do not yet exist, this discussion is speculative. Assuming that electronic journals generally will eventually fall within the scope of legal deposit, then it is likely that open access journals will too, if they are electronic only and are "published" in the UK. The legal deposit libraries may not want to collect both print and electronic versions of parallel published journals if their contents are substantially the same. It is likely they will prefer to take the print version because they know how to preserve print on paper. However, it is also likely that they will want to take electronic-only research journals to ensure they are saved somewhere.

If UK legal deposit is extended to electronic journals, it is likely that open access journal content will be available from both the publisher and legal deposit libraries while the journal is still being published. This is because it is probable that open access journal publishers will be amenable to the provision of much wider access to deposited journals by legal deposit libraries than is the norm for subscription-based journals, because this will not adversely affect their business models. If a particular journal ceases and/or the publisher goes out of business, then legal deposit libraries would still, in theory, be able to provide access to the journal.

The question of what "published in the UK" in the context of legal deposit means in the digital environment is still be to be decided. This is a crucial point and one that is unlikely to be decided until the regulations appear. It is unlikely to mean all material available in the UK no matter where the content is hosted or the publisher is based. UK researchers will want to be able to use to open access journals published in other countries, so there is the possibility that UK researchers could lose access to overseas open access journals when they are no longer available from the publisher. If they are not part of the UK legal deposit collection, then researchers would have to rely on remote access to the legal deposit collections of other countries. This is assuming that these countries have digital legal deposit and that the legal deposit libraries are able to provide access to these journals.

The long-term preservation of open access journals will require discussion and cooperation between legal deposit libraries and open access publishers. If legal deposit is to be extended to electronic journals, then they will have to decide how legal deposit will be implemented and then work together to achieve this. If legal deposit libraries do not take on the long-term preservation of open access journals, the alternative is that the publishers do so. Open access publishers may be more receptive to this than other types of publisher, especially commercial publishers of subscription based journals. These publishers may only ensure that material stays accessible as long as it is in their commercial interest. However, open access publishers can also be commercial enterprises and subject to market forces. If open access publishers go out of business then important research material may be lost.

5.2.2.2 E-prints

There are arguments against the preservation of postprints by e-print archives because postprints are complementary to the published literature, and the published literature will be taken care of by other means. In some cases this will be true. Postprints of articles that have been published in paper form will be preserved for posterity somewhere if there is a functioning legal deposit system in their country of origin. In this case, e-print archives need to develop preservation policies to meet their own needs. However, as mentioned above, e-print archives may need to take some responsibility for the long-term preservation of postprints of electronic-only articles, if legal deposit libraries and electronic-only publishers are unable to. As suggested by the title of the LOCKSS project (Lots of Copies Keep Stuff Safe), at the present early stage of e-journal preservation, properly managed archives ought to take responsibility for the preservation of postprints, with the proviso that this policy is revisited after a defined period of time.

Material accompanying open access articles or post-prints also needs consideration. If it is an integral part of a published article and the article is subject to legal deposit, then perhaps the accompanying material should be deposited alongside the article. On the other hand, if the article is an output of research and the research was funded by a grant, the research funder may require that outputs are deposited in data archives. For example, the Arts and Humanities Research Board requires deposit with the Arts and Humanities Data Service and the Economic and Social Research Council requires deposit with the UK Data Archive at the University of Essex. If accompanying data is deposited in deposit libraries and/or data archives, then e-print archives may not have to concern themselves with long-term preservation of this material.

The case of preprints and records of the process of development of preprints is an interesting one. These would be considered unpublished manuscripts, correspondence and authors' papers in the print environment. As such they would not be eligible for legal deposit and it would be up to the author to decide what to do with them. These papers may be of sufficient interest that libraries or archives would be interested in purchasing them or receiving them as a gift. If this material is deposited in an open access archive and available to anyone who wishes to view them, then this sort of material could be considered as "published", because it is made available to the public. If this were the case, it could potentially be considered eligible for legal deposit in the future. This possibility is very speculative, but it should be considered. If the material is not considered eligible for legal deposit in the future, there is still the question of whether it will have any longer-term value and whether it should be kept by archives or passed over to other archives. Normally it would be the author who would decide if their material is to be given or sold to archives. If the material is deposited in an open access archive, then there is a question of who would decide whether this material is passed over and whether there should be any financial transaction involved.

The discussion above has focused on responsibility for long-term preservation of e-prints and open access articles and has identified the players who might be involved in this. It is clear from the points raised that preservation decisions taken by open access archives will have to take the roles of these players into account and that open access publishers and legal deposit libraries need to work together. However, there is also the question of how open access archives themselves should work together on preservation. The answer to this question depends on how archives are organised. If individual archives work independently, they will still have to take other archives' collections into account. Retention and disposal policies and preservation are closely linked. E-prints may be duplicated in two or more archives. This may have an impact on individual archives' disposal policies. Even if a particular e-print is of no value to one archive, it may be wise to check whether a copy exists elsewhere or whether it would be of value to another archive. An example could be if a researcher has left an institution or an institution no longer carries out research in a particular area. A decision may be taken to remove material from the institutional archive. It might be useful for the researcher's new institution or a subject archive to make that material accessible.

If there is to be federation or centralisation of open access archives, then it would make sense for preservation policies also to have a degree of centralisation. JISC could develop high-level policies or at least guidelines for UK higher and further education archives. JISC could also work with other players, such as publishers, data archives and legal deposit libraries on behalf of these archives. JISC already works closely with The British Library on digital preservation matters.

5.2.3 Preservation in e-print archives

The need for e-print archives to engage in long-term preservation of their collections has been discussed above. The argument that archives should not concern themselves with preserving research output that is also formally published is persuasive because there are already, or soon will be, mechanisms in place to do this. If the aim of archives is to provide quicker and easier access to material, and if most use of material occurs in the first few months of availability via an archive, then arguably there is no need for any longer-term planning. However, the JISC *Feasibility and requirements study on preservation of e-prints* (James *et al.,* 2003, p. 25) found that there was an expectation among authors depositing material with e-print archives that the material would be retained and kept usable for at least 10–15 years if not indefinitely. This report also pointed out that since 10–15 years could be equivalent to at least two generations of hardware and software, with the attendant risk of technological obsolescence of collections, this timeframe could be considered long-term from a preservation point of view. Whether or not e-print archives become involved in longer-term preservation, they still need policies, even if the policy is that they are only concerned with short- or medium-term access to their collections.

5.2.4 Organisational models

The JISC *Feasibility and requirements study on preservation of e-prints study* (James *et al*., 2003) has already investigated the issues surrounding preservation in e-print archives. If and how e-prints are preserved in archives will depend to a great extent on what model or models are developed for the organisation and management of e-print archives. Even if archives take on a preservation role, it does not necessarily mean the individual archives have to carry out the full range of preservation activities themselves. A number of organisation models have been suggested:

- Full e-print archive – located in larger institutions, could take on a full range of preservation activities
- E-print archive with specialist support – call in external specialist expertise for digital preservation
- E-print archive with outsourced preservation services – if collections are to be preserved an external organisation takes full control of preservation activity
- Outsourced e-print archive services – external archive service used by individual researchers, projects or institutions.

Organisations already exist that could potentially take on preservation roles on behalf of e-print archives. As discussed above, there are data archives that could store and preserve material. The recently created Digital Curation Centre (http://www.dcc.ac.uk/) will be a source of expertise and information.

5.2.5 Policies

Whatever role archives decide to take on as far as preservation is concerned, they do need to develop policies, even if the policy is that they are not in the business of maintaining longer term access to e-prints. Since preservation is about maximising the useful life of material, preservation considerations underpin all processes in e-print archives from the form in which material is accepted, the metadata that accompanies it and how it is stored and how material is accessed. Decisions made at earlier stages will have an impact on how well material can be preserved and accessed. There are also rights issues associated with preservation, because keeping digital material accessible for any longer than the very short-term is likely to involve copying activities. Even if archives do not preserve e-prints themselves, the decisions they take will affect the ability of any other organisation to preserve them.

E-print archives need to consider the needs and expectations of their depositors and users. This consideration should be part of any collection management policy. Other issues to be considered in policies are:

- Which encoding formats for e-prints are acceptable? Some digital formats are easier to preserve than others.
- Are metadata required for preservation purposes, what metadata, how will they be obtained and how will they be stored and linked to collections?
- How long different categories of material will be kept for and how this will be decided (who decides, and decisions based on what criteria).
- If material should no longer be publicly accessible, will it be deleted, archived or transferred, and will a record of its existence remain accessible?
- Which technical preservation strategies will be applied and how? In order to deal effectively with technological obsolescence, publishers and archives will have to be aware of changes in technology. It is likely to be more efficient to have some degree of cooperation in areas such as monitoring technology trends and providing registries of file formats and format documentation. The Digital Curation Centre (http://www.dcc.ac.uk/) should have a role in this, but other organisations could also have roles. The UK National Archive is already providing a format registry through its PRONOM initiative (http://www.nationalarchives.gov.uk/pronom/). There may also be a role for the national libraries. Indeed, this is a global issue, so cooperation could be at an international level. There will be a need for coordination to avoid duplication of effort.
- How will integrity and authenticity be ensured? What will be acceptable?
- Rights issues – getting rights from authors to store and make material available to convert and implement technical preservation strategies as appropriate, to delete or transfer material. It is better for archives to make agreements with depositors at the time of submission, rather than have to pursue the relevant rights later on. There is an issue about postprints, in that the publishers then have an interest. They may not allow inclusion of postprints in e-print archives.

James *et al.* (2003) suggest that the OAIS model can be implemented in a disaggregated environment. The different functional entities mentioned above could be contained within individual archives, but they could also be separated out, with some functions carried out centrally and some carried out in a distributed way. They break up the OAIS model into different layers of "services"

- E-print Repository Board – equivalent to administration, formed by managers and users, this would develop policies, and negotiate submissions and deposit agreements
- Infrastructure Services – this would include Ingest, Archival Storage, Data Management and Access – the core functions of current e-print archives, although these could be outsourced to specialist data archives
- Specialist support services – essentially the Preservation Planning function

6. POLITICAL, CULTURAL AND BUSINESS ISSUES

The technical and preservation issues addressed so far are concerned with the *delivery* end of the open access process, but delivery cannot take place without the *provision* of material to deliver in the first place. This provision of e-print and open access journal material is the subject of this section of the report. In Section 3 we have already briefly addressed the reasons why institutions should establish e-print archives, and in Section 8 we present some suggestions for action by JISC to promote this notion strongly to the community. Clearly, the provision of a suitable archive in which any researcher can deposit their e-prints is the foundation of a national service. This does not necessarily need to be at a researcher's own institution, though there are substantial advantages to this, as we discuss below. In the absence of an institutional archive, though, it matters not where a researcher deposits his/her e-prints, so long as they are available on open access and the metadata are harvestable by the proposed service, so self-archiving in alternative locations is a perfectly workable solution.

The House of Commons Science and Technology Committee (2004) was undertaking its enquiry into scientific publication at the same time as our project was taking place, and so it was not possible to discuss the committee's conclusions with our interviewees. Its very long report appeared on 20 July 2004, and we have therefore given consideration at a very late stage only to its conclusions and recommendations, of which there were 82. These have had considerable publicity and discussion in the first week after their publication, and the consensus is that the report, though measured, has taken a position broadly favourable to the principle of open access. A number of recommendations were made to JISC on various points, but for the purposes of this study, the following are the relevant recommendations:

> ”43. Institutions need an incentive to set up repositories. We recommend that the requirement for universities to disseminate their research as widely as possible be written into their charters. In addition, SHERPA should be funded by DfES to allow it to make grants available to all research institutions for the establishment and maintenance of repositories. (Paragraph 115)
>
> 44. Academic authors currently lack sufficient motivation to self-archive in institutional repositories. We recommend that the Research Councils and other Government funders mandate their funded researchers to deposit a copy of all their articles in their institution's repository within one month of publication or a reasonable period to be agreed following publication, as a condition of their research grant. An exception would

need to be made for research findings that are deemed to be commercially sensitive. (Paragraph 117)

45. We recommend that institutional repositories are able to accept charitably- and privately-funded research articles from authors within the institution, providing that the funder has given their consent for the author to self-archive in this way. (Paragraph 118)

46. We recommend that DCMS provide adequate funds for the British Library to establish and maintain a central online repository for all UK research articles that are not housed in other institutional repositories. (Paragraph 118)

47. Institutional repositories should accept for archiving articles based on negative results, even when publication of the article in a journal is unlikely. This accumulated body of material would be a useful resource for the scientific community. It could help to prevent duplication of research and, particularly in the field of clinical research, would be in the public interest. Articles containing negative findings should be stored within a dedicated section of the repository to distinguish them from other articles. (Paragraph 118)

48. In order for institutional repositories to achieve maximum effectiveness, Government must adopt a joined-up approach. DTI, OST, DfES and DCMS should work together to create a strategy for the implementation of institutional repositories, with clearly defined aims and a realistic timetable. (Paragraph 120)." (House of Commons, 2004).

And:

"53. Having taken the step of funding and supporting institutional repositories, the UK Government would need to become an advocate for them at a global level. If all countries archived their research findings in this way, access to scientific publications would increase dramatically. We see this as a great opportunity for the UK to lead the way in broadening access to publicly-funded research findings and making available software tools and resources for accomplishing this work. (Paragraph 131)

54. Peer review is a key element in the publishing process and should be a pillar of institutional repositories. We recommend that SHERPA agree a "kite mark" with publishers that can be used to denote articles that have been published in a peer-reviewed journal. Upon publication, articles in repositories should be allocated the kitemark and marked with the date and journal of publication by the staff member responsible

> for populating the repository. Authors depositing articles in institutional repositories should also be required to declare their funding sources in order to reduce the risk of conflicts of interest occurring. (Paragraph 135)
>
> 55. We recommend that the Government appoints and funds a central body, based on SHERPA, to co-ordinate the implementation of a network of institutional repositories. (Paragraph 136)
>
> 56. A Government-established central body would play a major role in implementing technical standards across institutional repositories to ensure maximum functionality and interoperability. (Paragraph 137)
>
> 57. We recommend that DTI works with UK publishers to establish how the industry might evolve in an environment where other business models flourished alongside the subscriber-pays model. Government also needs to become an intelligent procurer, outsourcing some of the technical work involved in establishing and maintaining institutional repositories to publishers who already have the relevant infrastructure and expertise in place. (Paragraph 140)
>
> 58. We see institutional repositories as operating alongside the publishing industry. In the immediate term they will enable readers to gain free access to journal articles whilst the publishing industry experiments with new publishing models, such as the author-pays model. (Paragraph 143)
>
> 59. For the Government either to endorse or dismiss the new publishing model would be too simplistic. Without any Government action, some authors are already choosing to publish in journals that use author payments to recover costs. Author-pays publishing is a phenomenon that has already arrived: it is for the Government and others to decide how best to respond. (Paragraph 144)
>
> 60. The evidence produced so far suggests that the author-pays model could be viable. We recommend that Government mobilise the different interest groups to support a comprehensive independent study into the costs associated with author-pays publishing. The study could be used to inform Government policy and strategy. (Paragraph 150) " (House of Commons, 2004)

Our research into e-print operations around the world has revealed that as well as the technical issues there are a number of 'softer' issues that need consideration if provision of research material for a new service in the UK is to work successfully. First, there are issues to do with getting e-print archives set up and populated with articles: these are issues that concern institutions and authors. Second, there are other stakeholders and interested parties who can influence e-print delivery – publishers and research funders being the main players. Third, there are issues to do with the heterogeneity of scholarly research – the way that scholars go about carrying out and publishing the results of their work – that impact on an e-prints service in certain ways. This section addresses the main points of interest under each of these topics. The material presented in it was derived from an examination of published articles about the issues under study and from a series of personal interviews with key individuals from the following organisations:

University/college administrators:
King Alfred's College Winchester
University of Leicester
Queensland University of Technology

Librarians/institutional archive administrators:
University of Southampton
University of Oxford
The British Library
California Institute of Technology
Australian National University, Canberra

Research funders:
British Academy
Research Councils UK
Association of Medical Research Charities
Cancer Research UK
The Wellcome Foundation
Arts & Humanities Research Board

Publishers:
Blackwell Publishing
Taylor & Francis Journals
Oxford University Press
BioMed Central

6.1 Central *versus* institutional archives

Central, subject-based e-print archives have been set up largely as a result of the efforts of independent scientists (e.g. ArXiv, set up by Paul Ginsparg and CogPrints, set up by Stevan Harnad), have been hosted at institutions where these scientists worked and are populated with articles as a result of advocacy within the appropriate subject communities. PubMed Central is a slightly different example, set up by NIH under the influence of its former director, Harold Varmus. We anticipate that this sort of activity will continue and that the number of subject-based e-print archives may well increase as researchers see and appreciate their value to their subject community. Certainly the proposed service should harvest from existing subject-based archives and those set up in the future.

Institutional or departmental e-print archives, however, will be far more important to the success of the service proposed here. Subject-based archives have been successful in their particular fields, but still cover only a fraction of the total research output and this is likely to remain the case even if the numbers of such archives increase. Moreover, this approach is the 'wrong way round' with respect to e-print *provision* since for cultural reasons distributed, institution-based archives are much more likely to fill quickly, particularly if institutions adopt mandatory policies across all disciplines, something they are likely to do when the advantages of such archives to institutions become clear (see below and section 6.4).

As well as subject-based archives, however, central archives may take other forms, and there is considerable discussion as to the relative merits of formal, central e-print archives established on a regional or national basis. For technical as well as cultural reasons the project team is recommending that the proposed new service does not adopt this type of centralised archive model itself. The technical reasons are discussed in detail in Sections 4 and 7 of this report, but it is useful here to rehearse the arguments as to why a centralised model – whether subject-based or broad-scope – is also not attractive for academic cultural reasons. Several of these points are discussed more fully in subsequent sections of this report.

- The potential for centralised archives as a basis for a national open access service is suboptimal because:
  - The number of centralised (subject-based) archives is tiny and it is speculative in the extreme to suppose that subject-based archives covering the whole spectrum of scholarly research will be set up within a reasonable time (those that operate at the moment are all in the sciences).
  - The number of articles deposited in them has grown only slowly over the last few years. The growth rate remains linear, which is far too slow to produce a useful body of research literature for a new service in the UK.
  - Getting central archives populated requires advocacy within a subject community, something which can only ever work on the basis of persuasion and appeal, since there is no discipline-based power to mandate content provision. In practice, this is a tried-and-tested approach that has met with only limited success.

  - Research-funders can mandate self-archiving, but they can only mandate it for their own funded research and no funder has the scope of an entire discipline. Research-funders can, however, mandate *institutional* self-archiving by their fundees, and this *does* have the potential to propagate within and across disciplines and institutions.
  - It is difficult to envisage a mechanism by which other types of centralised (i.e. non-subject-based) archives might be filled, either, in the absence of both a mandating authority and research community advocacy.

- Conversely, the potential for institutionally-based archives is much better because:
  - The number of articles in institutional archives *can* grow quickly if the institutions in question simply adopt a formal self-archiving policy that actively and strongly encourages self-archiving
  - The number of institutional archives will grow, quickly, as institutions see the worth of such an undertaking for their own internal purposes as well as to permit open access
  - Institutions have it in their power to introduce mandatory policies on self-archiving across all disciplines

- Finally, many publishers have agreed to permit authors of articles published in their journals to self-archive them locally in institutional archives or departmental or personal websites, but will *not* permit them to place copies in 'third party' archives. A centralised model for the new service would presumably be viewed as belonging in the 'third party' category and would thus suffer from publisher prohibition policies.

So long as archives are OAI-compliant (i.e. interoperable), the physical location of the full-texts of the e-prints themselves is unimportant – both the metadata and the full-text can be harvested from any OAI-compliant archive. What *does* matter for the proposed service is that there *is* content to harvest and, in our view, this is one very good reason why the centralised model is not the most appropriate one to adopt. The proposed service should certainly harvest available content from any centralised archives, but the optimal arrangement – technically and culturally – is for e-prints to reside in a system of distributed archives set up by their authors' own institutions.

6.2 The cost of establishing an e-print archive

How much does it cost to set up an e-print archive? That is almost like asking how long is a piece of string! We have done our best to uncover some figures, though they vary extremely widely. Much of the variation is to do with whether establishing an archive would produce new overheads and what costs might be realistically absorbed into existing operations.

The cost to a sizeable research-led UK university of establishing an e-print archive can be relatively small. The IT infrastructure is already in place and existing IT staff can probably manage to absorb the relatively small amount of work involved. The software is available free (DSpace or Eprints.org) so there is no large outlay required in that direction. If capital investment is required for additional server capacity it will not represent a significant amount of money for an already-large IT service. For a smaller educational establishment, however, new staffing costs may be incurred and capital investment may prove significant. In cases such as this, where small educational establishments are interested and willing to set up e-print archives, JISC may need to offer some financial inducement to get the operation underway. Recurrent costs would be expected to follow the same pattern.

With respect to actual figures, the Wellcome Foundation has made some preliminary costings for setting up an e-print archive for the charity (see later) and estimates a setting-up cost of around £50,000, and an on-cost of around £25,000 per annum to run the service. We can assume that this will represent the cost of a state-of-the-art archive with future capacity planned in. At the other extreme, a consortium of universities in India plans to set up a network of institutional e-print archives for a fraction of this cost. Figures provided to us by Nottingham University estimates that an archive can be set up in an institution that already has a sizeable IT infrastructure in place for approximately £4000 (see below). On-costs in terms of maintenance (technical support, upgrade activities, preservation) can be significant, though, particularly if extensive support and administration activities produce additional FTE requirements.

The following table includes cost examples kindly given by staff from four existing institutional repositories: Nottingham University as part of the SHERPA project (Hubbard, 2003, 2004), DSpace at MIT (Barton and Walker, 2004), National University of Ireland, Maynooth (Redmond Maloco, 2004), and Queens University, Kingston, Ontario's QSpace (member of the Canadian Association of Research Libraries) (Qspace, 2004; Shearer, 2004).

Other projects provided rough ideas for time required for each task but did not have costs available (e.g. Korycinski, 2004). All projects stated that the majority of costs come from staff support required, but that the majority of costs can be absorbed within existing institutional budgets.

Figures for the cost of submitting and storing (i.e. ‘depositing’) and article also follow in a further table. Note that these costs would only be incurred if an individual were specifically employed to carry out this task. In many cases, researchers themselves submit their own articles via a form-based process, thus minimising the cost to the institution of this step.

Institutional Repository Cost Examples

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

The following table shows the costs to individual institutions/communities establishing and operating their own repository/archive of material. The costs are based on experiences of the SHERPA project and include initial set-up costs, maintenance costs and the employment of one staff member to input articles (half the staff time) and maintain the system on behalf of faculty/community members. The table also includes the cost of inputting each article based on the article input of four articles per hour.

INITIAL SET-UP COSTS £		TECHNICAL SUPPORT / MAINTENANCE £		ANNUAL OPERATING COSTS £		ARTICLE INPUT COSTS £	
Software	0	HEI standard Web service maintenance: three year upgrade		Staff salary	30000	Hours per week	17.7
Server	1500	Hardware	3000			Articles per hour	4
Installation	600	Labour	600				
Customisation	1800						
	3900		**3600**				**4.46**

The table below shows the cost of each article deposited as employee salary and article deposit rate vary. It is based on a 35-hour week for an employee.

Cost per Article		Articles Deposited per Hour							
		1	2	3	4	5	6	7	8
Salary of Employee	12000	£7.14	£3.57	£2.38	£1.79	£1.43	£1.19	£1.02	£0.89
	14000	£8.33	£4.17	£2.78	£2.08	£1.67	£1.39	£1.19	£1.04
	16000	£9.52	£4.76	£3.17	£2.38	£1.90	£1.59	£1.36	£1.19
	18000	£10.71	£5.36	£3.57	£2.68	£2.14	£1.79	£1.53	£1.34
	20000	£11.90	£5.95	£3.97	£2.98	£2.38	£1.98	£1.70	£1.49
	22000	£13.10	£6.55	£4.37	£3.27	£2.62	£2.18	£1.87	£1.64
	24000	£14.29	£7.14	£4.76	£3.57	£2.86	£2.38	£2.04	£1.79
	26000	£15.48	£7.74	£5.16	£3.87	£3.10	£2.58	£2.21	£1.93
	28000	£16.67	£8.33	£5.56	£4.17	£3.33	£2.78	£2.38	£2.08
	30000	£17.86	£8.93	£5.95	£4.46	£3.57	£2.98	£2.55	£2.23

In terms of future costs, digital preservation is an area where there is likely to be a significant funding requirement (over the next two to three decades and onwards). Storage capacity will also need to increase: experience of archive administrators to date has shown that this is a difficult area to plan, as growth rates for archives are unpredictable at present, where deposition of articles remains almost exclusively voluntary. There may be substantial storage capacity required as self-archiving develops, particularly when multimedia objects are deposited.

The last area of cost that needs to be considered is that of marketing and advocacy. Institutions that currently operate e-print archives have found that substantial efforts must be made in these directions if authors are to be persuaded to deposit copies of their articles. This issue is discussed from the viewpoint of why this should be so in section 6.4. Here, it is necessary to note that institutions need to devote considerable effort and resources into internal marketing of an e-print archive. Some institutions have reported that to date this has even taken the form of winning support from authors on an individual-by-individual basis. In other cases it has been tackled by series of seminars or presentations. Whatever the details of the approach, it is clear that much effort needs to go into this aspect of setting up an archive and that the cost is certainly not negligible.

6.3 The content of e-print archives

In earlier sections of this document – those on the background/landscape and on technical issues – we have drawn attention to the different types of digital object that may be deposited in an archive. Here, it is useful to raise the issue of how these may vary according to subject area, since this will have a bearing on the nature of archives and how institutions go about establishing and running them.

Most discussion to date about e-print archives centres on articles published in the scholarly literature, given away free by authors. In this circumstance the case for self-archiving is a simple one and the objects archived are going to be, in the main, standard journal articles across the whole spectrum of scholarly endeavour, perhaps accompanied in the archive by additional supporting material such as the large datasets generated in some branches of the sciences, or video or audio clips, perhaps.

For subjects outside the sciences, the type of object that represents research output may vary considerably from this simple model. For example, in the performing arts output is often in the form of a performance: this has always brought problems for assessment but in this context it also presents problems for archives of research results. Video records of performance use large amounts of digital storage space, for example, so that an institution which has a strong representation in the performing arts may well have quite different initial requirements for server space to, say, a university of technology. In this particular case, since the FE sector is home to many well-established and productive performing arts departments, JISC may find it needs to pay special attention to the needs of the FE sector as a provider of archive content.

Another important example for discussion is that of the arts and humanities, which differ from the sciences in the form of much of the research output. In this case, although arts/humanities scholars *do* publish work in traditional journals, there is also a large volume of output in the form of monographs. Whilst these might be archived in the same way as journal articles, there can be differences: first, there may be multiple authors each contributing a chapter to a monograph, possibly from different institutions, which may require separate deposition and submission policies; second, monographs tend to be much larger documents than journal articles, so there is again a space implication here; and third, it is not unusual for authors to be paid royalties by the publisher of an academic monograph and whilst these are usually small, they nonetheless represent payment, so in these cases this is not 'giveaway' literature.

Finally, we return to the needs and interests of the FE sector, which is not in general research-led. Teaching and learning materials are the most likely candidates for inclusion as digital objects in any FE institutional archives and present problems with metadata formats because of their diversity and non-standardisation. It is difficult to see at this stage how learning attributes can be applied in the context of metadata formats, partly because little work has been done in this area to date. Powell and Barker (2004) describe a collaborative venture, funded by JISC, between the RDN and Learning and Teaching Support Network (LTSN) to develop policies for the interoperable exposure of learning materials. This has resulted in the creation of an application profile known as the RDN/LTSN LOM Application Profile (RLLOMAP). Further consideration is required to understand how LOM metadata may be applied to proposed services, and how learning attributes could be added to metadata from HE research-led institutions.

We flag up these issues because they indicate that populating an institutional e-print archive may not be entirely straightforward from a practical point of view and that JISC will probably not be able to formulate a single policy – at the present time – that fits all circumstances.

6.4 Populating e-print archives

In section 3.2.8 we presented some figures that showed the number of articles currently stored in e-print archives in the United Kingdom. They are, by any standards, small and this is not a problem that is confined to the UK: archives around the globe share the same situation. Establishing e-print archives is one thing, but getting them populated with research output is quite another.

Administrators or champions of existing archives have attempted to tackle the problem in various ways. Sustained advocacy is the main one. In some institutions this has taken the form of a formal programme of events to publicise the existence of an archive in the institution and to attempt to persuade authors to deposit their work by logical and persuasive argument. The events that people have mentioned to us as productive are: seminars, workshops, demonstrations, departmental or research-group presentations, and poster campaigns. In many cases it is the institution's library that has taken on the role of advocate, usually because it is also the library that has assumed the role of archive administration, but in our discussions for this study we have learned that the support – tacit or actual – from the pro-vice chancellor (PVC) or provost responsible for research policy is crucial. In institutions that have established an e-print archive but where the PVC is still not persuaded of its merits there is a long uphill struggle to win the hearts and typing fingers of authors. Author inertia is the main enemy of an e-print archive once it is established.

The alternative to the 'author chooses to comply' model is to *mandate* self-archiving. To date, there are a handful of educational institutions that have gone so far as to mandate that its authors deposit copies of all their research articles in the institutional e-print archive (http://www.eprints.org/signup/fulllist.php); the best example of this is Queensland University of Technology (QUT) in Australia. The archive was championed originally by Tom Cochrane, a pro-vice chancellor for research policy, and he recruited assistance from the university library to set up and run the archive. The mandating policy is only recently announced and although it is now officially in place, the university is taking a softly-softly approach to enforcing it in order to avoid alienating faculty members. As discussed earlier in this document, Australia is probably furthest ahead of all countries in terms of national organisation and policy on e-print archiving, and it will be salutary to watch both QUT's progress and also whether this mandatory policy transfers to other Australian research universities.

There are also examples of departmental mandates, one such being the School of Electronics and Computer Science at the university of

Southampton, which has produced a policy that could be used by other departments travelling the same route (http://software.eprints.org/handbook/departments.php). To allay fears about the process of self-archiving and its legality, Eprints.org has produced an FAQ (http://eprints.org/self-faq) and a handbook on the subject (http://software.eprints.org/handbook/).

Other archive champions and administrators have told us that they have stopped short of implementing a mandatory policy for fear of irritating rather than winning over faculty. Nevertheless, there are good arguments why an institution *should* mandate in this way, not least because it enables it to use the archive as an institutional marketing tool, for research assessment exercises, and for monitoring the performance of its own research staff (all these points were discussed briefly in section 3.2.4). A mandatory condition from the institution that researchers lodge full-text copies of all published work in an institutional archive so that the institution can harvest information for the RAE is unlikely to irritate or alienate researchers, who will see it as yet another obligation on their part in the name of educational bureaucracy. And, with respect to author reaction to mandating, the recent study by KPL for JISC showed that the vast majority (about 70%) of authors would *willingly* comply with a mandatory self-archiving requirement from their employer or funder (Swan & Brown, 2004). Anecdotally, we have been told by some sources that the expectation is that RAE drivers will ultimately induce all UK universities to engage in mandating self-archiving by their researchers.

Aside from the institutions themselves, there are other agents that could impose a mandatory self-archiving policy – the research funders. The only funder we have found so far that *requires* the results of research it has supported to be made open access is the Danish Research Centre for Organic Farming (DARCOF) (http://orgprints.org). DARCOF is not a single research centre, but rather organises research on organic farming across a number of research sites and organisations. It has also established its own e-prints archive for researchers to use if their own institution does not provide such a facility (or to use as well, even if it does).

During the course of this study we have spoken to a number of representatives from funding bodies such as the UK research councils and charities that support research in the UK. The present situation is that none of them require their grantees to archive copies of research articles, but all declared that it is a matter for discussion. Some have come further than others in this regard. The Wellcome Foundation has progressed furthest on this and is approaching the matter in two ways: first, it is now actively considering implementing a mandatory policy with respect to self-archiving research that it has supported and second, it has completed a study on the costs involved in establishing and running a Wellcome e-print archive specifically to house e-prints from authors whose institution does not have an archive for them to use. Smaller charities, which are financially not in a position to do this, nevertheless feel somewhat aggrieved at the toll barrier to research they have funded and are now looking at the possibility of mandating self-archiving as a means to obviate this. They feel they are in an iniquitous situation when they cannot have access to the published results of much of the research they sponsor without purchasing research journals. Research Councils UK (RCUK) has set up a discussion group which comprises a representative from each of the eight research councils which is to address the matter of open access and help formulate an RCUK formal policy on this matter, to be announced by the end of 2004.

It is interesting to note in this context that the Arts & Humanities Data Service (AHDS) has been operating a mandatory policy with respect to digital research output since the mid-1990s. It is a requirement of Arts & Humanities Research Board funding that any digital resources produced with funding from this body must be offered for preservation either by the AHDS or by other approved means. Because electronic publication is not yet a norm in this subject area, most digital items produced are databases or datasets that are not formally published material themselves but that underpin it. We were given to understand, though, that this requirement may eventually extend to electronic publications themselves.

Overall, it is our view that the funders may well be the first to pick up this ball and run with it. They have just as much to gain from having grantees' work provided on an open access basis as institutions and are, perhaps, more used to laying down rules as to researchers' obligations to them. Researchers have expressed to us that they would happily comply with a requirement from their funder to archive their published results and would view it in the same vein as the funder's present requirement to, say, produce a report within 3 months of a research programme being completed.

There is one more issue to consider here and that is the role of publishers. One of the reasons authors give for not self-archiving is that their publisher does not permit it. In the past that has been largely true, because most publishers have had strict copyright and ownership policies in place that have prevented authors from using their own work under many circumstances. This is now changing. The latest figures show that around 80% of journals now permit authors to archive a copy of an article on a personal or institutional website, which removes the major barrier in most cases (Harnad & Brody, 2004). Other publishers are actively considering adopting a 'gold' policy on open access instead – that is, permitting authors to pay a publication fee in return for making articles open access. This of course already applies to the Open Access publishers such as PLoS and BioMed Central, but some traditional publishers are also now experimenting in this area. The National Academy of Sciences in New York will make articles in *PNAS* for which the author chooses to pay a fee open access (Cozzarelli, 2004). Oxford University Press has just announced, after some period of trialling on a small scale, that *Nucleic Acids Research* will be an open access (gold) journal from 2005 (http://www3.oup.co.uk/jnls/list/nar/narpressjun04.pdf). Springer (incorporating Kluwer Academic) has also announced that its considerable corpus of journals will adopt the same hybrid model as *PNAS*. (http://www.springeronline.com/sgw/cda/frontpage/0,10735,1-40359-0-0-0,00.html) (accessed 13 July 2004).

In the context of this study, therefore, it is important to note that publishers are now more likely to encourage open access rather than hinder it and JISC will be able to harness this development in its endeavours to create a national s-prints service.

6.5 Institutions that should be involved in a national e-prints project

As well as the universities, which are largely research-led, there are other institutions that should be involved in a national e-prints service from the *provision* side. Further Education establishments, whilst primarily focused on teaching, nevertheless do in some cases produce considerable amounts of research output. A national service should therefore look to include these establishments and assist them in finding a way to expose their research output.

Furthermore, there are the non-university sources of research – government-funded research institutes, independently-funded research institutes, and industry. The latter will or will not participate in a national initiative depending upon individual company attitudes, but JISC should ensure that some effort is made to bring the research institutes on board in this regard. Large volumes of high-quality research results are generated by such organisations and a national service that did not include this would be regarded as wanting.

6.6 'Mopping up': how to serve authors who have nowhere to self-archive

Finally, there is the issue of authors whose institutions do not have an e-print archive. What can they do to make their research results open access? JISC will need to consider this body of authors and make some sort of provision for them because, although it is expected that over the next decade most research-based institutions will set up and operate their own e-print archives, some will not, and authors in these places, or indeed authors working independently of an organisation as such, will need to find a home for their work.

We have already mentioned fledgling plans by the Wellcome Foundation to provide an archive for its own grantees who have nowhere else to place their e-prints. It is possible that other funders, including the research councils, might also consider doing something similar. There may be another possible route here, though. During our research we spoke to a representative from the British Library, who expressed the opinion that that organisation would view very positively a collaboration with JISC in respect of the BL building and maintaining an e-prints archive specifically in the interests of authors who need somewhere to deposit work. This is something JISC may consider developing into a useful initiative. There is one issue to think about here and that is, would publishers consider a British Library archive as a 'third party' site and thus withhold permission for authors to use it? We have begun to ask publishers this question and have not been able yet to get a straightforward answer, mainly, we think, because it is a notion they themselves have not entertained before. Given strong arguments about the reason why such an archive had been established (i.e. just for authors whose institutions have no archives of their own), we suggest that most reasonable publishers would take the view that this is a proxy institutional site and would comply.

Another recent development that may help institutions that do not have an OAI repository is to utilise the newly developed OAI gateway specification. This development is intended to lower the barriers to making metadata available through the OAI. It works on the basic principle that metadata can be encoded in an XML file (conforming to a specific schema) and mounted on a standard web site, e.g. an author's or institution's home page. This file is known as a static repository. The URL of the static repository can be registered with an entity known as a 'static repository gateway'. The gateway reads the metadata file and incorporates it into a fully compliant OAI-PMH service that can subsequently respond to OAI requests. The idea is that metadata can be made available from standard web sites and Incorporated into an OAI environment. This new development is described in guidelines available from the OAI (http://www.openarchives.org/OAI/2.0/guidelines-static-repository.htm - SR_overview).

One of the applications of this development would be to enable institutions or authors who do not have access to an OAI data provider service to make information available as a static repository. Any organisation providing a OAI data provider service for institutions without repositories may wish to consider whether the provision of a static repository gateway service in addition to providing an OAI repository service is something that would benefit the community.

6.7 Legal issues associated with e-prints

The creation and maintenance of e-print archives, whether institutional or subject-based, raise a number of legal issues that have significant implications for those running the archives. The major legal issues are the same as those that face all electronic publishers, namely:

- Breach of confidentiality and official secrets
- Personality and image rights
- Data protection
- Copyright and database right
- Moral rights
- Defamation
- Obscenity and race hate material
- Contempt of Court
- Trade marks and domain name disputes

Further details about these issues can be found in standard texts, such as Armstrong & Bebbington (2003), Gringras (2003), Jones and Benson (2002) and Pedley (2003),but key points are highlighted below.

6.7.1 Breach of confidentiality

There is a general rule that a person who receives information in confidence has a duty to keep that confidence and not disclose the information to others, unless there is a just reason for doing so. Whilst it is unlikely that whoever manages an archive will deliberately breach confidence, it is possible that material offered to the archive does breach confidentiality, and the manager will be a party to a breach of confidence case if it can be shown that the manager acted recklessly in accepting, and then making public, the material in question. Similar rules apply to official secrets. In certain circumstances, it is acceptable to breach such confidentiality – for example, if the information has become public knowledge or if there is a public interest in disclosure – but the manager of an archive would have to take legal advice before going ahead and loading material that he or she believes breaches confidentiality and hopes to rely on such defences.

6.7.2 Personality and image rights

Whilst traditionally those in the public eye have a weaker case than others when complaining about their image appearing in published materials without their consent, that should not be taken as a *carte blanche* to use such images as one sees fit. Certainly those who are not in the public eye will receive a sympathetic hearing from the Courts if they claim their privacy has been breached, notwithstanding the lack of any formal right to privacy in UK law. Certainly, images of patients should never be reproduced in an archive without the patients' express written consent.

6.7.3 Data protection

The Data Protection Act 1998 is designed to ensure that information about identifiable living individuals is not processed (and that includes published on a archive) without their implied or express consent. Furthermore, individuals are given a number of rights to inspect data about themselves, to request amendment of incorrect information, and to sue for damage under certain circumstances. Furthermore, the Act restricts the transfer of personal data to a number of non-EU countries (including the USA) unless permission is obtained from the data subject or certain other conditions apply. Whilst there is no problem in having authors of items within a archive named, as they have given their implicit consent to such publication, issues can arise if the material on the archive relates to other individuals. Jay and Hamilton (1999) provide full information on the Act and its implications.

6.7.4 Copyright and database right

Probably the most problematic area for managers of archives will lie in copyright law. This is because many academics do not understand the law and/or may have signed away copyright in works to publishers prior to submitting the material to an archive. It is therefore essential that those who are depositing materials into the archive fully understand both copyright law and the implications of any contracts they may have signed with other publishers. It is also essential that any material included in the archive is free of plagiarism, as that is copyright infringement and could lead to legal action against the archive.

In addition, there are a number of legal issues associated with copyright ownership of the material in an archive, and the associated metadata. These were explored in the RoMEO Project (RoMEO, 2004) and are touched on elsewhere in this report. Finally, there are legal issues associated with the use of Creative Commons or similar licences that express what may or may not be done by third parties with the material held in an archive. Managers of archives will need to consider both what sorts of licences they should issue and how they intend to police the use of materials from their archive to ensure that the terms of the licence are adhered to and that no unauthorised infringement of copyright occurs.

An archive, in addition to being a series of copyright works, is also a database in its own right under the terms of the Copyright, Designs and Patents Act 1988. The manager of the archive is therefore also responsible for protecting the database rights associated with the archive. These rights are similar to those of copyright, but the manager needs to ensure that he or she is familiar with database law as well (see, for example, Rees & Chalton, 1998).

6.7.5 Moral Rights

The creator of a copyright work has, under many circumstances, the right to be identified as the author of the work, and the right to sue if his or her work is subjected to derogatory treatment. Although not everything in an archive will be subject to Moral Rights, the manager should assume that all of it is. Therefore, the manager must ensure that any materials in the archive do indeed identify the author of the work correctly, and that the material has not been amended in such a way as to impugn the reputation of the author.

6.7.6 Defamation

There is a very real danger that works appearing in an archive defame a third party. Unlike other areas of legal risk, where the manager of the archive is only liable if he or she was reckless in the handling of the materials in the archive, in the case of defamation, the manager is at risk unless he or she can demonstrate that they did not know, or had no good reason to know, that the material was defamatory – a somewhat different test. It is possible for the manager of the archive (or his or her employer) will be successfully sued even if they acted in good faith, but failed to take the necessary steps to ensure that there was nothing defamatory in the text or images loaded. In particular, the manager must always delete the material in question as soon as a complaint about defamation is made, even if subsequently it turns out that the material was innocuous. The law is unforgiving on this matter. Similarly, if a published journal article has had to be withdrawn because of defamation, the archive equivalent must be withdrawn as well.

6.7.7 Obscenity and race hate material

It should be obvious that managers of archives should never upload text or images that might be considered obscene (or otherwise illegal, such as race hate material) without taking legal advice. There are only very restricted circumstances when offering such materials is permissible.

6.7.8 Contempt of Court

Material relevant to on-going Court cases should not be added to the archive except following clear legal advice that it is safe to do so.

6.7.9 Trade Mark and domain names

In general, items that are subject to Registered Trade Marks should always be acknowledged as such, and authors submitting materials should confirm they have done so. Reproduction of logos, images and names is probably acceptable for *bona fide* academic use, but should not be used in the course of business, i.e., for any commercial venture associated with the archive, without the express permission of the Trade Mark owner.

The archive's own URL may find itself the subject of a domain name dispute with another domain name that is confusingly similar. There are now well-established ground rules for deciding which party "wins" such disputes, and the manager should take legal advice should the archive become embroiled in such a dispute.

Furthermore, if any commercial activity occurs at the institutional or subject-based archive (such as charging to view certain parts of the archive), then a number of other legal issues associated with e-commerce arise. These are well reviewed by Tunkel (2000).

It will be clear from this discussion that the maintenance of an archive entails significant legal risks. Most of these can be avoided by a combination of the following actions:

1. Ensure that every author submitting material to the archive provides the archive with a warranty that nothing in the content being offered infringes copyright, is defamatory or breaks any other law. Standard texts on publishing agreements (Owen, 2002) provide an appropriate form of words.
2. Ensure that any complaint about defamatory or copyright infringing material on the archive is dealt with as a matter of urgency, and that the material in question is blocked whilst the inquiry proceeds.
3. Take legal advice in all cases of uncertainty.

6.8 Provision of OA journal content

So far in this section we have addressed only issues concerned with e-prints. The other type of content to be delivered by the proposed service is open access journal content. In many senses this is very simple compared with e-prints: for example, there are no legal issues to be addressed, archives to house this content do not need to be set up, and there are no behavioural or political issues that dissuade the providers of such content from making it available. Open access journal publishers, and those publishers whose journals are embracing the hybrid model where certain articles in each issue are open access, are keen to make their content available wherever possible. In the life and medical sciences (e.g. PLoS, BioMed Central), the content is available on the publishers' websites and is also deposited in PubMed Central. For publishers in all disciplines, their journal content is accessible by OA service providers at all times and it will be a simple procedure for the proposed service to organise routine harvesting of OA journal content from publishers' sites, as discussed in section 4.1.3.

6.9 Integration of content services

The proposed service will be harvesting open access material from e-print archives globally and from open access journals. This means that most of the content of the service will be, at least in early years, work that has been relatively recently carried out. There is, then, the question of the usefulness to researchers of this service and whether this usefulness might be enhanced by the integration of content from other sources that can provide older material. The answer to this question is not simple, for there are discipline-specific differences in how researchers go about using the literature for their work. In the sciences, the main utility will lie in having access to research results immediately upon publication or, in the case of preprints, in advance of formal publication. In this case, then, the proposed service alone, providing as it will access to the latest research findings, will satisfy most of the needs of this group. In the arts and humanities, much of the literature consulted and referred to is much older – sometimes centuries or even millennia back – and researchers in these disciplines do not share the same publishing imperatives as scientists.

There is another perspective on this, too, from the user behaviour angle, and that is that wherever possible, users prefer to access all the information they want through one interface. The proposed service may therefore be enhanced by the integration of selected third-party services. We do not suggest that this should be an early-days step: rather it should be something that evolves as user adoption and usage of the new service is studied and understood. There may be other free content services that provide valuable content for certain subject disciplines, or it may be that the most useful services to integrate are paid-for ones, which will bring with it more complexity for JISC. At the moment we simply flag up this issue as one to watch.

7. THE HARVESTING MODEL AND SERVICES BASED UPON IT

7.1 Introduction

In chapter 4 we established that the harvesting model offered the greatest promise, out of the technical models available to underpin services integrating e-prints and open access journals. We also examined a wide range of related technical issues. Here we discuss services based on this model and their implementation in more detail.

The harvesting of metadata from OAI archives could be organised in a number of ways. Three possible scenarios are:

1. Harvesting from IAs, SBAs and OAJs is carried out at a national, central level. Then subject-based and other service providers harvest or cross-search subsets from the central national service
2. Harvesting within subject disciplines is carried out by subject-based service providers. These then act as data providers to national services.
3. Harvesting by resource types – e-prints/OAJs, e-theses, reports literature – is carried out by agencies dedicated to those types. These agencies act as data providers to national services

While there might appear to be some advantages to the second and third approaches – availability of expertise in the subject or resource type areas and modular management of services being paramount – in practice the first approach is a much more realistic and workable option. Harvesting at national level offers a greater degree of consistency and reduces duplication of effort. The requisite HE/FE Subject Portals, required for option 2, are not in place, nor, with the exception of e-theses which have received much attention from JISC-funded projects, are the hypothetical agencies required for option 3. Furthermore, a prototype for a national service already exists – ePrints UK.

7.2 ePrints UK

A high level overview of the system proposed by ePrints UK is illustrated in Figure 9 (reproduced from http://www.rdn.ac.uk/projects/eprints-uk/docs/technical/architecturev1.032003/).

The key features of this architecture are:

- Metadata are harvested from e-print archives

- The metadata are enhanced through the use of web services
- Data providers can re-harvest their records, which have been enhanced, to improve their local service
- The service has its own native interface and can be embedded into other services

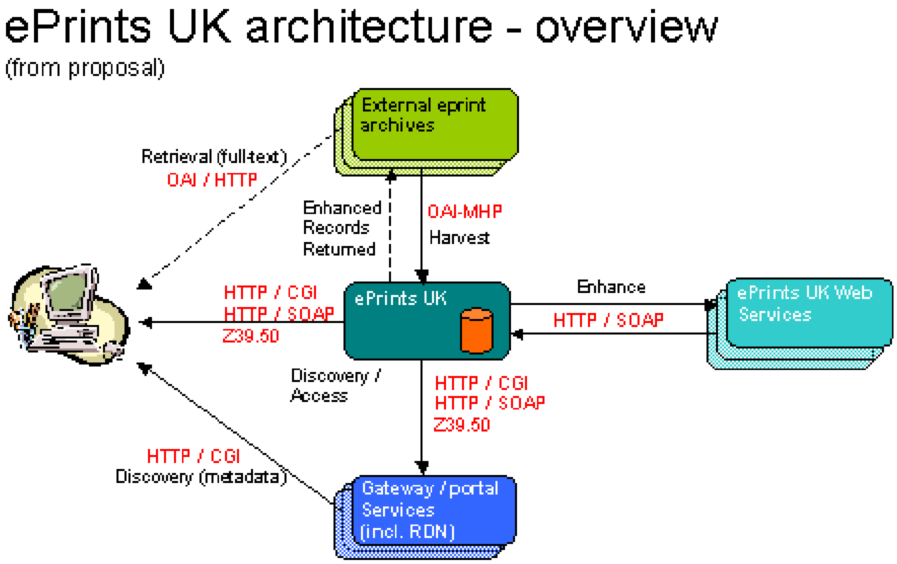

Figure 4. An overview of the ePrints UK architecture.

The following diagram, reproduced from "ePrints UK: Developing a national e-prints archive" (http://www.ariadne.ac.uk/issue35/martin/) illustrates more clearly the web services under trial:

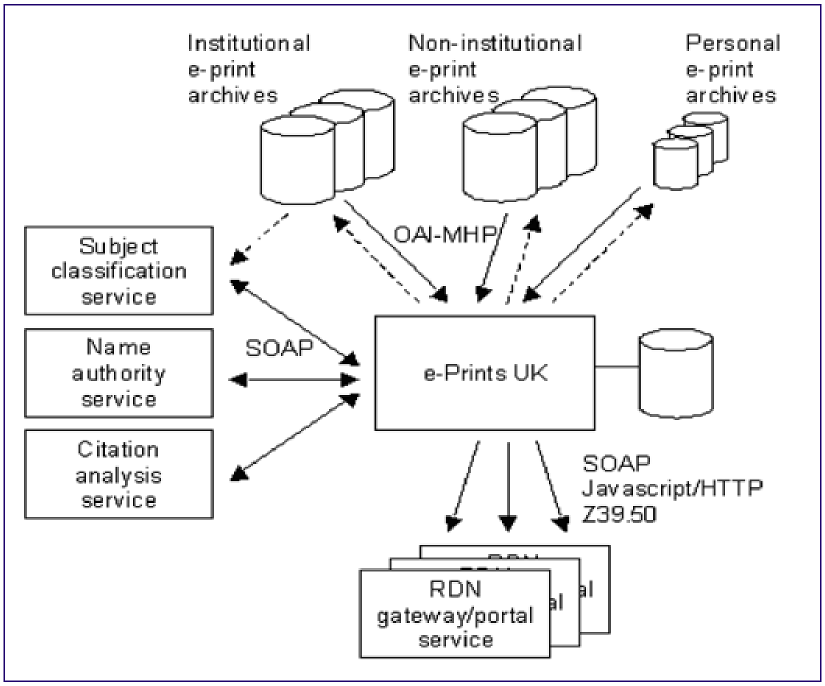
Institutional e-print archives
Non-institutional e-print archives
Personal e-print archives
OAI-MHP
Subject classification service
Name authority service
Citation analysis service
SOAP
e-Prints UK
SOAP
Javascript/HTTP
Z39.50
RDN gateway/portal service

Figure 10. The Project Architecture (Andy Powell, UKOLN)

The subject classification and name authority services are based at the OCLC research centre at Dublin, Ohio. The subject classification service automatically assigns Dewey Decimal Classification (DDC) to the metadata. DDC is used under a research licence. The name authority service checks the author names as they appear in the metadata against authority name files. The citation analysis service, by the Open Citation project team, parses semi-structured citation information in document texts to create OpenURLs.

These web services show some promise, but licensing issues need sorting out before DDC can be used in a fully fledged service, and the name authority service requires further testing. And again, while the citation analysis service has been shown to work, in the shorter term it is just as easy to create an OpenURL from available metadata without reference to the full text.

An ePrints UK Service Demo, which harvests metadata from around 20 institutional archives daily, is available at http://eprints-uk.rdn.ac.uk/. Rather than setting arbitrary limits, all available resource types (e-prints, OAJs, reports, theses, conference papers, etc.) are harvested from the data providers. The database, holding in excess of 26,000 records, demonstrates the power of metadata harvesting and shows that a service based on the work of ePrints UK could be launched immediately.

The search interface is simple and robust, in keeping with the observations of Liu *et al.* (2002):

"Keyword search allows users to search all metadata fields across archives. It is implemented by accumulating and indexing all metadata fields together. Keyword search provides a simple and familiar way to conduct search across all archives, and the input can include Boolean operators (AND, OR, NOT). It is probably the only way to search across extremely variable sources without major work, but it cannot exploit the rich metadata set defined by source archives."

An important point to note is that the ePrints UK service offers a UK-based view of e-prints rather than a view solely consisting of UK resources. While it is appropriate that emphasis should be placed on harvesting, exposing and preserving UK resources, in the digital age, limiting resources to a particular geographical region is arbitrary and restrictive.

7.3 Portal-in-a-browser

In our view, this model (illustrated in figure 11) is simple, elegant and in keeping with the JISC Information Environment architecture. Moreover, it is not that dissimilar to the ePrints UK model, or most models provided by other service providers. We have stripped out the web services offered by ePrints UK, but these can be added in again later when those services have matured. We have also added, tentatively, a central archive that 'mops up' scholarly works written by authors whose institutions do not have their own eprint archives.

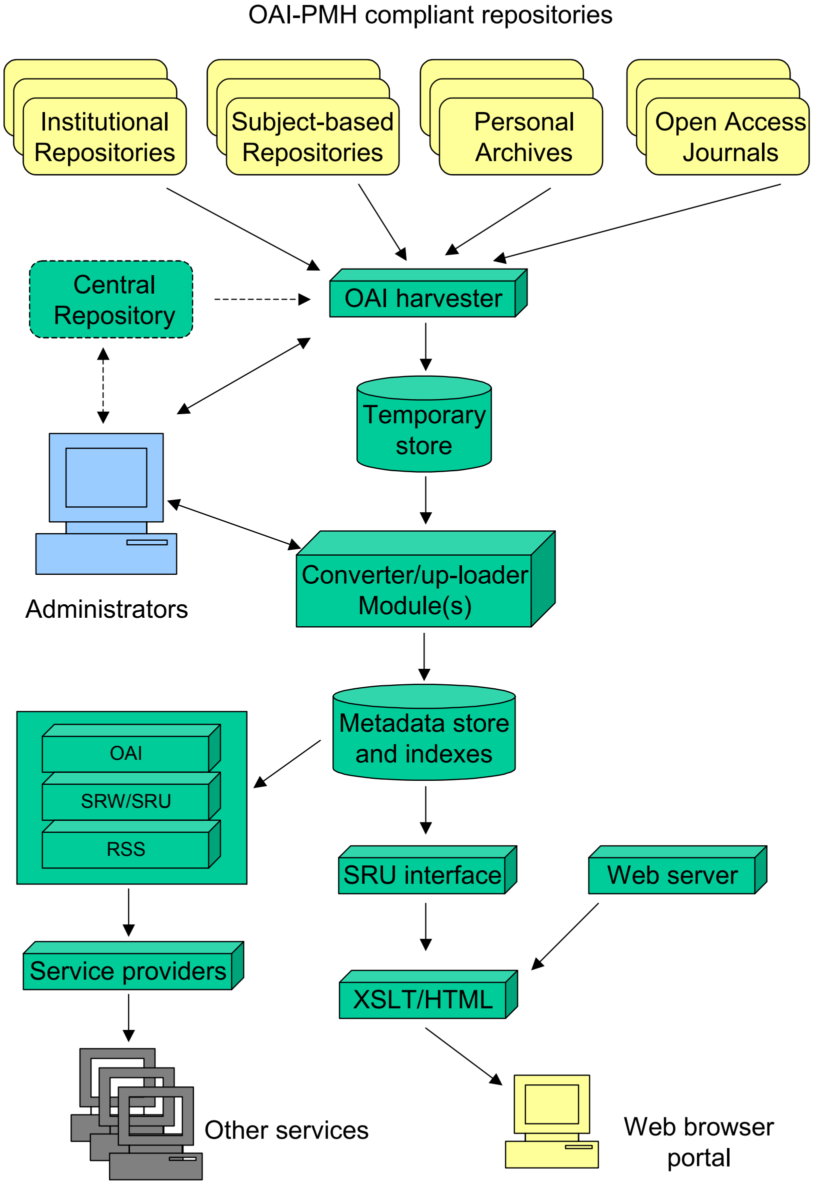

Figure 5. The portal-in-a-browser model

The choice of software tools to support this model has to be driven by the technical expertise and knowledge available at the agency (or agencies) that take on responsibility for any proposed service, but as examples:

- Harvesting could be carried out using ARC, OAICAT or other available OAI tools

- Storage and indexing could employ Cheshire II, MySQL, PostgreSQL, or even flat XML files indexed by Altavista tools

All the protocols used in this model are standard protocols, which are straightforward and inexpensive to implement, making this model worthy of serious consideration.

7.4 The Google model

As Macleod (2004) reports, *"Google has launched a pilot project with 17 leading universities around the world, including Cranfield in the UK, to make academic papers and research more accessible on the internet."*

Work is ongoing to link Google to university archives using DSpace via a search system set up by the OCLC. If the test pilot is successful, then an additional search feature will be added to Google.
If trials prove successful and this idea is expanded to include archives driven by software other than DSpace, then this model may be regarded as an interesting complementary strand to the models discussed above.

7.5 The near future

When considering the near future of services, there are some observations that can be made with (relative) confidence.

The ePrints UK, Portal-in-a-Browser and Google models are not mutually exclusive. ePrints UK have a system in place that could and should be launched, as a simple service, very quickly. Work on developing OAI, SRW/SRU and RSS interfaces to add to the service could continue in the meantime. The Google model may offer an alternative point of entry to scholarly resources that is likely to prove popular with many users.

ePrints UK discard the harvested full-text documents, once the full text has been analysed and the metadata have been updated. However, if a fully integrated service is to be considered, then the full text could not only be harvested and used to support added-value services, but the full text could be used as the basis for preservation. There are obviously rights issues to be considered if this latter opportunity is to be taken, but ongoing developments in the specification of rights over OAI-based resources are currently being considered by the OAI-Rights Working Group. In the short term, the most likely rights statements applied to resources such as e-prints will be Creative Commons licences. The nature of these licences means it is unlikely that rights statements will indicate whether ‘storage of a resource at a service provider for the purposes of preservation’ is explicitly allowed or forbidden. Nevertheless, as long as the conditions and restrictions explicitly stated in the licence are not overlooked, maintaining a copy of the full text purely for the purposes of preservation should be permissible (in fact most creative commons licences allow distribution, but it may be more appropriate for a preservation service to direct downloads to the original data provider rather than fulfilling these requests via the service provider’s copy). However, if there is doubt about the legal position, then agreement should be sought from individual data providers that a centralised service would be permitted to maintain and preserve resources in this way.

Recommendations and guidelines already made by ePrints UK should be pursued and promoted, and the findings of RoMEO, SHERPA and other projects should be incorporated into services as they develop.

With regard to the issue of poor quality of metadata, once metadata have been harvested into a national store, it becomes possible to examine those records in any number of ways. For example, the metadata can be used to carry out automated surveys on:

- Subject – schemes and keyword terms used; to assess the challenges to be addressed in achieving a harmonised scheme (or schemes)
- Format – in what formats are digital objects stored; build a practical ‘real world’ list of formats that should commonly be supported and identify risks to preservation likely to be caused by proprietary or obscure formats
- Type – how have archives categorised their digital objects by type?
- Identifier – how many resources have been assigned persistent identifiers? How many have not? Long term access to these resources is at risk if persistent identifiers have not been assigned

Once such surveys have been carried out, the resulting information can be utilised to inform further developments in improving the quality of metadata exposed, which is key to providing effective browsing and advanced searching facilities.

One issue that needs further attention is the identification of duplicate resources. At the time of writing, duplication of resources is scarcely an issue compared to the need to populate repositories in the first place. However it is an issue that will need to be practically resolved, as repository populations grow.

There are several types of apparent duplicates:

- Different revisions of a resource

- Different formats of a resource
- Mirror copies of a resource in different locations
- Creation of duplicate records or submissions of duplicate resources within a repository
- Records from multiple data providers identifying a single resource in a single location

As discussed in section A.3.4.1, records relating to revisions and formats of a resource should be grouped together, either as loosely bound separate records or as a single structured record. The same approach would also be appropriate for mirror copies.

The creation of duplicate records or submissions of duplicate resources within a repository are problems that should be addressed locally by:

- Repository administrators
- Addition/enhancement of duplicate checking algorithms within the repository software

Giving consideration to records from multiple data providers identifying a single resource:

- Creating algorithms using 'fuzzy' matching to identify duplicates is not a new challenge
- The multiple records should be amalgamated into one record
- In cases where some data providers have enhanced the record they are exposing, this will represent an opportunity to enhance metadata rather than a crisis

As experience with Dspace has shown, it is possible to create and expose functional OpenURLs. While service providers can provide these metadata, in the longer term it would be better if the creation and exposure of OpenURLs were carried out by the data providers.

As the discussion of existing service providers revealed earlier in this document, some features that are deemed desirable and achievable are personalisation, annotation, alerting and linking to related documents in search results.

The Open Access movement and the OAI are both rooted in cooperative and collaborative philosophies. Therefore one certainty is that cooperation and collaboration will be important keys to the development of a future Information Environment.

In particular, archive software developers from different projects need to get 'round the table' and to work together. While administrators of archives can contribute to improvements within their own domains, major steps forward across the board are more likely to be achieved if fundamental improvements are built into archive software packages themselves. Agreement on the labelling of resource types, a set of common high level subject terms, introduction of standard collection descriptions and the adoption of web services that enhance metadata at the time of submission – any or all of these would be helpful.

Future developments to any proposed services will be both evolutionary and revolutionary. While the evolutionary can be predicted, at least to a certain extent, the revolutionary presents an unknown: who knows what paradigm shifts may occur in the next three, five or ten years that will render much of our current thinking invalid?

Despite these reservations, the next section makes a few tentative observations on 'what might be' at sometime in the future.

7.6 The not so near future

This section is written as if ten years hence. The digital archive software development teams – DSpace, Eprints, Fedora and others have been working together co-operatively for several years now. They have integrated name authority, citation analysis and automated subject classification services into their software. Many individual institutions still enter subject terms using their own in-house scheme for local use, but now thankfully all records are marked up with DDC as well.

Nearly all data providers – the IAs, SBAs and OAJs – now assign persistent identifiers, provide collection or journal level descriptions, have automatic versioning control and expose metadata, via the OAI, in a METS wrapper with the descriptive metadata expressed using the DC Library Application Profile. With all these metadata available, a national service can now offer comprehensive management as well as resource discovery functions.

Now, when authors wish to submit an article to journal publishers, as the first step they upload the preprint to their institutional archive. Upon notification from the author, journal publishers pickup the metadata and preprint from the IA – which of course comes complete with a persistent identifier that is used to explicitly to connect the peer-reviewed article with its preprint. Open access journals – as data providers – expose this identifier along with other article metadata in their OAI archives. When service providers harvest and process this metadata, the relationship between the article and preprint can be updated in the preprint record. This, in conjunction with mandatory self-archiving, has also increased the rate of growth of e-print archives. Although the mandatory policy wasn't well received by scholars originally, this change in journal article submission process has given scholarly authors a strong *personal* motive to deposit their works in their institutional archives. The central archive, jointly administered by the British Library and an HE/FE agency is also burgeoning (and generating a healthy income stream from the annual storage charge) for these reasons. And HE/FE institutions are benefiting from simplified, more comprehensive and largely automated RAE returns.

With these rich metadata schemes in place and automated versioning built into archive software (already a feature of Fedora software), bibliographic control of different versions and formats of documents has become almost trivial.

Now that the vast majority of text documents are word processed and published in XML formats, migration and emulation worries are focussed on graphical and other non-text files.

As bandwidth has increased and storage costs have continued to diminish, it has become feasible to routinely harvest metadata *and* the associated digital objects, both documents and ancillary data. This process has been facilitated by the emergence of a protocol for metadata and digital object harvesting related to the OAI-PMH. The availability of the digital objects for replication is providing benefits:

- Preservation based on a modified form of LOCKSS (http://lockss.stanford.edu/)
- Automated subject classification at data provider and service provider level. Crosswalks from DDC to LCSH, MESH and other schemes are now well established as web services

Full text searching plays an important part in enhancing resource discovery.

At last, like repository software, all HE and FE library catalogues incorporate XML-based web services – OAI, SRW/SRU, RSS and others. The library catalogue is now an integrated sub-component of library portals dedicated to cataloguing *local holdings* (access to electronic resources is covered by other sub-components). This, in conjunction with the merging of efforts by teams behind COPAC, SUNCAT, the National GL database, the national theses service and other union catalogues and web services, has led to the establishment of a rich harvested national union catalogue incorporating all types of legacy and electronic resources.

The 'one stop shop' has arrived – but it's not a 'one size fits all' establishment: by employing a 'pick 'n' mix' philosophy - cross searching or harvesting resources according to their own local preferences or needs - everyone can see the content they desire through their own personal or institutional shop window.

8. TARGETED RECOMMENDATIONS FOR ACTION

The harvesting model presented first in Section 4 and then discussed in detail in Section 7 underlies all the recommendations we make in this present section. We present below the series of recommendations for actions that we think JISC and other stakeholders need to make in order to maximise the chances of success of the proposed harvesting model. The tools for doing many of these steps have already been developed (often funded by JISC) and are shown in the table that follows the list below.

1. Give institutions and funders the reasons for adopting an official Open Access provision policy

The main reasons why open access provision policies should be adopted by educational institutions and funders are:

- Open Access dramatically increases research impact
- Institutional archives provide a means for an institution to measure and reward research effort objectively
- Open Access to research articles enables funders to measure and reward research effort objectively

2. Develop a programme to persuade all research-led HE institutions to establish e-print archives

This involves developing both *incentives* and *methods* to encourage UK institutions to provide e-prints.

Incentives would include the provision of the following, to encourage institutions to join a trend that is gathering pace:

- Continually updated data on the numbers of UK e-print archives, their locations and how numbers are growing
- Continually updated data on the numbers of articles stored in these archives and how they are growing
- The latest figures on the increased impact that open access articles enjoy
- The latest information from SHERPA/RoMEO on publisher self-archiving policies, so that institutions can direct researchers accordingly
- Information on how an institutional archive can improve the RAE and make it cheaper and easier (Harnad et al., 2003).

Methods would include:

- Creating a generic demonstration for institutions showing the simple steps to creating an e-print archive
- Showing how simple it is to create and work with a standardised RAE CV from this, and how easy it is to harvest performance indicators from it.

Because there may be institutions willing but unable to create e-print archives for cash reasons, there may be cash implications for JISC here.

3. Develop a programme to persuade researchers to self-archive their work in e-print archives
Again, both incentives and methods can be developed for this purpose.

Incentives would include provision of:

- The latest figures on the increased impact that open access articles enjoy
- The latest information from SHERPA/RoMEO on publisher self-archiving policies, so that authors can easily check whether the journal they are submitting work to permits self-archiving
- A form-based author request to any non-'green' publisher (one that does not explicitly permit self-archiving) asking permission to self-archive a specified article, with wording to the effect 'if refusal is not received within 30 days, then it is assumed that permission has been granted'
- A form-based author request to his/her institution to request that it creates an e-print archive if it doesn't have one

Methods would include:

- Creating a generic demonstration that showed authors the simple steps required to submit their articles to an e-print archive; the demo should also put the general case for Open Access via this route
- Providing an impact correlator that enables authors to predict, from early-days e-print download data, the eventual citation impact from six months later

4. Explore possibilities for cooperation with the British Library on a 'mop-up' archive
The British Library has expressed interest in collaborating with JISC on the provision of an e-print archive to house articles from authors with nowhere else to deposit them. JISC should progress this initiative with the BL.

5. Develop a programme to persuade non-educational research establishments to set up e-print archives
A substantial proportion of UK research output comes from outside educational establishments, from research institutes and government laboratories. JISC should provide the same incentive-and-method information to these bodies as to the universities (as in [2] above).

6. Work with funders to encourage them to mandate self-archiving of their funded research, and perhaps to establish their own e-print archives where appropriate

Funders can influence self-archiving very strongly and a mandate from the main research funders in the UK for it would tip the balance immediately in favour of an effective nationwide service. JISC should work with the main funders (research councils and larger charities) to encourage such a policy, and to facilitate funders to provide e-print archives themselves to provide an archive for use by researchers who do not have one in their own institution. *There may be cash implications for JISC here.*

7. Identify a group of stakeholders to establish the desirability (or not) of a co-ordinated approach to controlled subject metadata, identify appropriate schemes and recommend ways to develop supporting mechanisms

The stakeholders envisaged here are: the data and service providers and the software developers.

The targeted recommendations discussed above are summarised in the table overleaf. In addition to the specific tools listed, we recommend the eprints handbook which comprehensively covers all the steps in setting up and operating an institutional archive (http://software.eprints.org/handbook/).

Action recommended	Tool or methodology where one exists already; notes or comments
Give institutions and funders the reasons for adopting an official open access provision policy	Data to date on increased impact of open access articles from Lawrence, 2001; Kurtz *et* al, 2003; Kurtz, 2004; Harnad & Brody, 2004
Develop a programme to persuade all research-led HE institutions to establish e-print archives. Provision of:	
• Continually updated data on the numbers of UK e-print archives, their locations and how numbers are growing	http://archives.eprints.org/index.php?action=browse and
• Continually updated data on the numbers of articles stored in these archives and how they are growing	http://archives.eprints.org/index.php?action=analysis
• The latest figures on the increased impact that open access articles enjoy	Data to date on increased impact of open access articles from Lawrence, 2001; Kurtz *et* al, 2003; Kurtz, 2004; Harnad & Brody, 2004
• The latest information from SHERPA/RoMEO on publisher self-archiving policies, so that institutions can direct researchers accordingly	http://www.sherpa.ac.uk/romeo.php http://romeo.eprints.org and http://romeo.eprints.org/stats.php
• Information on how an institutional archive can improve the RAE and make it cheaper and easier	Harnad et al., 2003; Bence & Oppenheim, 2004
• A generic demonstration for institutions showing the simple steps to creating an e-print archive	
• Information showing how simple it is to create and	http://paracite.eprints.org/cgi-bin/rae_front.cgi

work with a standardised RAE CV from this, and how easy it is to harvest performance indicators from it.	
Develop a programme to persuade researchers to self-archive their work in e-print archives. Provision of: • The latest figures on the increased impact that open access articles enjoy • The latest information from SHERPA/RoMEO on publisher self-archiving policies, so that authors can easily check whether the journal they are submitting work to permits self-archiving • A form-based author request to any non-'green' publisher (one that does not explicitly permit self-archiving) asking permission to self-archive a specified article, with wording to the effect 'if refusal is not received within 30 days, then it is assumed that permission has been granted' • A form-based author request to his/her institution to request that it creates an e-print archive if it doesn't have one • A generic demonstration that shows authors the simple steps required to	Data to date on increased impact of open access articles from Lawrence, 2001; Kurtz *et* al, 2003; Kurtz, 2004; Harnad & Brody, 2004 http://www.sherpa.ac.uk/romeo.php http://romeo.eprints.org and http://romeo.eprints.org/stats.php

submit their articles to an e-print archive; the demo should also put the general case for Open Access via this route • An impact correlator that enables authors to predict, from early-days e-print download data, the eventual citation impact from six months later	For the method see Harnad & Brody 2004; for the tool see http://citebase.eprints.org/analysis/correlation.php
Explore possibilities for cooperation with the British Library on a 'mop-up' archive	
Develop a programme to persuade non-educational research establishments to set up e-print archives	Government-funded research institutes; privately- or charitably-funded research establishments; industrial research establishments
Work with funders to encourage them to establish e-print archives where appropriate, or to mandate self-archiving in other cases	The UK research councils and larger charities
Identify a group of stakeholders to establish the desirability (or not) of a co-ordinated approach to controlled subject metadata, identify appropriate schemes and recommend ways to develop supporting mechanisms	Software developers, data providers and service providers should be brought together on this

9. COST-BENEFIT ANALYSIS

9.1 Costs

Costs of IAs are discussed in section 6.2. As the four examples discussed there show, there are many causes of variation from institution to institution. The major variable is, of course, the quantity of material that is archived; the Sherpa figures suggest an input cost in the region of £3-£4 per document, but there are also costs associated with the long-term preservation of the material. Nor is it clear to what extent the costs of an IA will be subsumed within the overall IT systems costs of institutions. For example, Loughborough University already maintains the Learn Server, a VLE for teaching-related materials, and an administrative database of references to all publications by Loughborough academic staff, which is kept for RAE purposes. Without pre-empting any future decisions by Loughborough University's senior management, it can clearly be seen that both of these functions could in principle be performed by a possible IA of the future, which could also carry preprints and postprints of the full text of Loughborough staff's publications. In that case, it is not clear what proportion of the costs of the IA might be allocated to the "Open Access" function, as compared with the proportions that might be carried by the Learning and Teaching budget and the Administration budget of the University.

The costs associated with the publication of Open Access journals have also been a matter of controversy. For journals based upon the principle of covering their costs through charges levied to authors (or rather their employing institutions or research funders), publication charge figures ranging from $500 to $3000 are current. The variability is largely a function of the selectivity of each journal – rejected papers are not usually charged, so the higher the rejection rate, the higher the charge to accepted papers. Moreover, these charges are largely arrived at by considering the costs of peer review and editorial procedures, and initial mounting on a server, not the costs of long-term archiving.

There has been extensive discussion on various discussion lists about the ultimate effect on University costs of a switch to Open Access, either through OAJs ("the gold route") or through IAs and/or SBAs ("the green route"). The matter is complicated by issues of internal university accounting; journal subscriptions are paid through university library budgets, while the costs of an IA may be within library, IT services, or administrative budgets. Publication charges payable to OAJs may be paid out of research grants or be debited to the author's academic department, but there are indications that some universities are debiting these charges to the library, thus blurring the distinction between OAJs and toll-access journals. There is also controversy about whether an eventual OA system of scholarly communication will in fact save universities money overall, though probably a majority of those participating in such debates believe that it will. A summary of the debates on the American Scientist discussion forum can be found at http://www.ecs.soton.ac.uk/~harnad/Hypermail/Amsci/3378.html

In addition to the IAs, SBAs and OAJs which are the main concern of this report, costs will also be associated with OAI service providers and with any national system, co-ordinated by JISC, that may emerge from this and other current JISC-supported projects. It is difficult to estimate these costs in advance of decisions about the configuration of any such systems. Nor is it easy to predict what proportion, if any, of any such system will come to be regarded as "infrastructure", like JANET, or whether the scholarly communication system will continue to be regarded as a cost-recovery item. These are policy matters for the funding councils to decide.

It may be that, in spite of the relatively low cost of operating an IA, some institutions may be willing to establish one but cannot find the funds to do so. JISC may wish to create a budget to assist institutions in this situation to establish an IA.

9.2 Benefits

The immediate benefit to an institution of providing an OAI-compliant IA, populated with preprints and postprints of publications written by their academic staff, lies in visibility and impact. If interested parties around the world cannot see, or cannot find, the publications of University X, they cannot take cognisance of them or cite them. There is beginning to be some evidence (Lawrence, 2001; Harnad and Brody, 2004; Kurtz, 2003; Kurtz *et al.*, 2004) that articles that are available on Open Access have greater impact than those that lie behind tariff walls.

Universities regard their websites as marketing tools, publicising the university to various prospective markets: potential students, potential research funders, potential employers of their graduates, and potential business collaborators. Websites, being relatively inexpensive, are generally regarded as good value for money as marketing tools. A research-active university will wish to emphasise its research quality on its website, and one way of doing this is to make all its research reports available in full text through the website. If research papers from members of the university's staff cannot be seen by users of its site, because they are available only on toll-access sites, some of the publicity impact will be lost.

All UK universities need to make RAE returns, and most expend considerable resources in putting their RAE submissions together. If an IA is maintained on a routine basis that contains all of the publications from the academic and research staff of staff of that university, then it is available at no extra cost when RAE time comes around. Currently, universities have to provide to the RAE panels printed copies of all articles that their staff wish to submit for the RAE. With all the publications held on an IA, electronic copies could be selected from the IA and sent directly to the panels at far less cost.

Some of these benefits are easier to quantify than others, but none are very readily quantifiable. The third – RAE submissions – is probably the easiest to quantify (Harnad *et al.*, 2003; Bence and Oppenheim, 2004). Throughout the history of information science, efforts have been made to quantify the benefits of knowing a particular piece of information rather than not knowing it, as a justification for having a library or information service within an organisation (Kingma, 2000). Here we are seeing the situation from the opposite direction – the advantage, to an originator of information, of having that information generally exposed rather than not. The difficulty of quantification remains.

9.3. Costs versus benefits

The costs of IAs are tangible and the benefits largely intangible. A formal cost-benefit analysis is therefore not possible, but the conclusions of this survey seem to suggest that an IA can be provided at a fairly modest additional cost to the host institution, and that considerable benefits to that institution will result from the IA being there. This section can be summarised as follows:

Costs

- Unpredictable for any particular institution because they are dependent on many variables
- Not, in general, high though, versus the likely benefits
- Capital costs of establishing an IA are not high but ongoing running costs are difficult to foresee, being dependent on the degree of take-up by members of the institution
- Some institutions may need financial help from JISC
- Service providers, software developers and the infrastructure will incur costs and it may fall to JISC to cover these
- Long-term assured funding, rather than project funding, will be needed

Benefits

- The impact of British research would be maximised
- Visibility and accessibility would be improved
- The RAE would be cheaper and easier to administer, at both institutional and funding-council level

10. RISK ASSESSMENT

We have performed an early-days risk assessment and this appears in the following table.

RISK	PROBABILITY (1-5)	SEVERITY (1-5)	SCORE (P X S)	ACTION TO PREVENT/MANAGE RISK
Institutions do not set up e-print archives	4	5	20	JISC should develop more effective advocacy programmes using persuasion about the advantages
Institutions do not fill e-print archives	4	5	20	JISC should develop more effective advocacy programmes to persuade institutions and funders to consider mandating self-archiving, including promoting archive-sourced CVs for the RAE exercise
British Library does not establish its own 'catch-all' archive	2	4	8	Positive collaborative action with BL to establish an archive network that can provide 'cover' for the whole of the UK HE and FE community
Funders do not comply by setting up archives	4	1	4	In the short term, this will be more serious than in the longer term. Action by JISC (see section 8) may ameliorate this
Persistence of digital objects - in the sense of the responsible agencies - is a cause for concern	3	5	15	The provision of a national harvesting service for e-prints will help to avoid this. In the future, not only the metadata but the (e.g. full-text) objects themselves will be harvested (see section 7). The British Library may become the electronic legal deposit site for digital articles. LOCKSS will mitigate against catastrophic loss
Persistence of digital objects - in the sense of e-prints without persistent identifiers - is a cause for concern	3	5	15	Cooperation and collaboration between software developers should solve this for the future. At present it is a voluntary add-on task for Eprints and the DSpace handle is not a total solution
Persistence of digital objects - in the sense of digital objects themselves - is a cause for concern	3	5	15	LOCKSS will mitigate against catastrophic loss. JISC should ensure that any new national service follows guidelines that foster good practice in this area
Uncertainty regarding the OA model as sustainable	2	3	6	Some OAJ publishers (BioMed Central, PLoS) are already lodging copies of their archival material with PubMed Central. Others OA journal publishers may be making no provision for copies to persist if

				their business fails. This is a point to watch.
Poor quality of metadata exposed by OAI repositories	5	3	15	This is more serious in the short term than in the longer term. Co-operative action by repository software developers to standardise on their use of metadata fields, and to develop richer, structured metadata schemes may overcome current problems
Uncertainty of long term funding models for repository software developers	4	5	20	As seed-funding of DSpace, Eprints.org and others runs out, these projects need to develop effective exit strategies. In the short term extra funding may have to be found to keep the projects running

REFERENCES

Armstrong, C. J., & Bebbington, L. (2003). *Staying Legal* (2nd edn). London: Facet Publishing.

Arunachalam, S (2004) India's March Towards Open Access. http://www.scidev.net/quickguides/index.cfm?fuseaction=qguideReadItem&type=3&itemid=243&language=1&qguideid=4 (accessed 29 June 2004).

Ayris, P (2002) New International Scholarly Communications Alliance Engages Academics in Broadening Access to Research. http://www.curl.ac.uk/about/isca.html (accessed 20 June 2004).

Barton, M.R. and Walker, J.H. (2003) Building a Business Plan for DSpace: MIT Libraries' Digital Institutional Repository. *Journal of Digital Information*, 4(2). http://jodi.ecs.soton.ac.uk/Articles/v04/i02/Barton/barton-final.pdf (accessed 15 April 2004).

Beckaert, J., Hochstenbach, P. and Van de Sompel, H. (2003) Using MPEG-21 DIDL to Represent Complex Digital Objects in the Los Alamos National Laboratory Digital Library. *D-Lib Magazine*., 9 (11), http://www.dlib.org/dlib/november03/bekaert/11bekaert.html (accessed 7 July 2004)

Bence, V. and Oppenheim, C. (2004) The Role of Academic Journal Publications in the UK Research Assessment Exercise. *Learned Publishing*, 17 (1), 53-68, and references cited therein.

Brody, T. *et al.* (2004) The Effect of Open Access on Citation Impact. http://www.ecs.soton.ac.uk/~harnad/Temp/OA-TAadvantage.pdf (accessed 18 June 2004).

CARL (2003) Position Statement, November 2003. http://www.uottawa.ca/library/carl/projects/ir/selfarchiving.pdf (accessed 18 June 2004).

Cozzarelli, N.R. (2004) An Open Access Option for PNAS. *Proceedings of the National .Academy of Sc*iences, 101 (23), 8509. http://www.pnas.org/cgi/doi/10.1073/pnas.0403554101 (accessed 18 June 2004).

CNRI (2003). Corporation for National Research Initiatives – Handle System (WWW document). http://www.handle.net/. (accessed 5 June 2004).

DOI (2004). The Digital Object Identifier System (WWW document). http://www.doi.org/. (accessed 5th June 2004).

Crow, R. (2002) SPARC Institutional Repository Checklist & Resource Guide. http://www.arl.org/sparc/IR/IR_Guide.html (accessed 20 April 2004).

Directory of Open Access Journals (DOAJ) (2004) Press release, June 3, 2004: Lund University launches Phase 2 of the Directory of Open Access Journals – now with article level search. http://www.doaj.org/articles/040603 (accessed 10 June 2004).

Eprints.org. Journal and Publisher Policies on Author Self-archiving (Eprints/ROMEO version).
http://www.ecs.soton.ac.uk/~harnad/Temp/Romeo/romeo.html (accessed 16 June 2004).

Eprints UK (2004) http://www.rdn.ac.uk/projects/eprints-uk/ (accessed 12 July 2004)

Fedora (2004). The Fedora Project – An Open Source Digital Repository Management System. http://www.fedora.info/ (accessed 10 July 2004).

Gringras, C. (2003). *The Laws of the Internet* (2nd edn). London: Butterworths.

Harnad, S. (2004) Estimates on data and cost per department for institutional Archives? To multiple recipients of: *DSpace Mailing List,* sent: 13 January 2004, time: 17:00.

Harnad, S., Brody, T., Vallieres, F., Carr, L., Hitchcock, S., Gingras, Y., Oppenheim, C., Stamerjohns, H., & Hilf, E.R. (2004) The Green and Gold Roads to Open Access. http://www.nature.com/nature/focus/accessdebate/21.html

Harnad, S. & Brody, T. (2004) Comparing the Impact of Open Access (OA) vs. Non-OA Articles in the Same Journals. *D-Lib Magazine* 10 (6)
http://www.dlib.org/dlib/june04/harnad/06harnad.html (accessed 15 June 2004).

Harnad, S., Carr, L., Brody, T. and Oppenheim, C. (2003) Mandated online RAE CVs Linked to University Eprint Archives: Improving the UK Research Ass4essment Exercise whilst making it cheaper and easier. *Ariadne*, 35
http://www.ariadne.ac.uk/issue35/harnad/ (accessed 15 july 2004).

Hauge, J H (2004) The OA situation in Norway.
http://opcit.eprints.org/feb19oa/hauge-norway.doc (accessed 15 June 2004).

Hitchcock, S., Woukeu, A., Brody, T., Carr, L., Hall, W. and Harnad, S. Evaluating Citebase, an Open Access Web-based Citation-ranked Search and Impact Discovery Service.

http://eprints.ecs.soton.ac.uk/archive/00008204/01/Evaluating_Citebase_TR.pdf (accessed 4 May 2004).

House of Commons Science and Technology Committee (2004) *Tenth Report: Scientific Publications: Free for all?* London: TSO. Available at: http://www.publications.parliament.uk/pa/cm200304/cmselect/cmsctech/399/39902.htm (accessed 26 July 2004)

Hubbard, B (2004) Personal communication to Hardy, R., 27 May 2004.

Hubbard, B. (2003) The SHERPA Project www.nottingham.ac.uk/is/about/news/newsletter/infrom-online-7.6/sherpa.htm (accessed 19 December 2003).

Indian National Digital Library in Engineering Sciences & Technology http://paniit.iitd.ac.in/indest/ (accessed 29 June 2004).

James, H., Ruusalepp, R., Anderson, S. and Pinfield, S. (2003) Feasibility and Requirements Study on Preservation of E-prints: Report commissioned by the Joint Information Systems Committee (JISC). 29 October, 2003. http://www.jisc.ac.uk/uploaded_documents/e-prints_report_final.pdf, (accessed 10 June 04).

Jay, R., and Hamilton, A. (1999). *Data Protection Law and Practice*. London: Sweet & Maxwell.

Jones, H., and Benson, C. (2002). *Publishing Law* (2nd ed.). London: Routledge.

Kingma, B.R. (2000) *The Economics of Information: A guide to economic and cost-benefit analysis for information professionals*. Englewood, CO: Libraries Unlimited

Korycinski, C. (2004) Personal communication to Hardy, R., 1 June 2004.

Kurtz, M.J., Eichhorn, G., Accomazzi, A., Grant, C., Demleitner, M., Muarray, S.S., Martimbeau , N. and Elwell, B. (2003) Worldwide Use and Impact of the NASA Astrophysics Data System Digital Library http://cfa-www.harvard.edu/~kurtz/jasist1-abstract.html The Bibliometric Properties of Article Readership Information http://cfa-www.harvard.edu/~kurtz/jasist2-abstract.html (both accessed 15 July 2004). *Journal of the American Society for Information Science and Technology*, in the press

Kurtz, M.J. (2004) Restrictive access policies cut readership of electronic research journals articles by a factor of two. Paper presented at National Policies on Open Access (OA) Provision for University Research Output: an International meeting, 19 February 2004, New College, University of Southampton, UK. http://opcit.eprints.org/feb19oa/kurtz.pdf (accessed 14 July 2004).

Lawrence, S. (2001) Free online availability substantially increases a paper's impact. Nature (Web Debates) http://www.nature.com/nature/debates/e-access/Articles/lawrence.html (accessed 14 July 2004). Edited version appears in *Nature* , 411, 521 (2001).

Library of Congress, 2001. METS: an overview and tutorial. http://www.loc.gov/standards/mets/METSOverview.html, (accessed 10.6.04).

Liu, X., Maly, K., Zubair, M., Hong, Q., Nelson, M., Knudson, F., and Holtkamp, I. (May 2002) Federated Searching Interface Techniques for Heterogeneous OAI Repositories. *Journal of Digital Information*, 2 (4), Article No. 106, 2002-05-21, http://jodi.ecs.soton.ac.uk/Articles/v02/i04/Liu/ (accessed 10 July 2004).

MacLeod, D. (2004). Google Launches Research Archive Project, *Guardian Unlimited,* http://education.guardian.co.uk/higher/news/story/0,9830,1191090,00.html (accessed 14 July 2004)

Martin, D, and Bide, M. (1997). Descriptive Standards for Serials Metadata and Standards for Terms of Availability Metadata. http://www.ukoln.ac.uk/services/elib/papers/supporting/pdf/serials-metadata-toa.pdf (accessed 7 July 2004).

McGauran, P (2003) $12million for managing university information http://www.dest.gov.au/Ministers/Media/McGauran/2003/10/mcg002221003.asp (accessed 23 May 2004).

National Information Standards Organization (2001) ANSI/NISO Z39.85 - 2001 Dublin Core Metadata Element Set. Bethesda, MD: National Information Standards Organization http://www.niso.org/standards/resources/Z39-85.pdf?CFID=2958242&CFTOKEN=63747668 (accessed 13 July 2004).

Needham, P., Sidwell, K., Bevan, S. and Harrington, J. (2002). The MAGiC Project. Managing Access to Grey Literature Collections. Final Report – October 2002. Sponsored by the BL and RSLP. http://www.bl.uk/concord/docs/magic-final.doc, (accessed 26 July 04).

Odlyzko, A. The rapid evolution of scholarly communication. *Learned Publishing* 15(1), 7-19 (2002).

Owen, L. (2002). *Clark's Publishing Agreements* (6th ed.). London: Butterworths.

Oxford University Press (2004) Nucleic Acids Research: NAR's Open Access Experiment. Press Release: Oxford Journal takes bold step towards free access to research. 26th June 2004. http://www3.oup.co.uk/nar/special/14/default.html (accessed 29 June 2004)

Pedley, P. (2003). *Essential Law for Information Professionals*. London: Facet Publishing.

Pinfield, S (2003) Open Archives and UK Institutions: An overview. *DLib Magazine 9* (3). Available at http://www/dlib.org/dlib/march03/pinfield/03pinfield.html (accessed 25 May 2004).

Pinfield, S. and B. Hubbard (2004) Establishing a National Network of Repositories: Confidential. SHERPA Project: Nottingham University.

Powell, A, and Barker, P. (2004). RDN/LTSN Partnerships: Learning resource discovery based on the LOM and the OAI-PMH. http://www.ariadne.ac.uk/issue39/powell/ (accessed 5 July 2004).

Powell, A, Day, M and Cliff, P. Using Simple Dublin Core to Describe E-prints. UKOLN, University of Bath. Version 1.2 http://www.rdn.ac.uk/projects/eprints-uk/docs/simpledc-guidelines/ (accessed 19 May 2004).

QSpace - Queens Institutional Repository Project Plan *(2004)* http://library.queensu.ca/webir/planning/q_space_planning_document.htm (accessed 13 July 2004).

Redmond Maloco, S. (2004) Personal communication to Hardy, R., 1 June 2004

Rees, C., and Chalton, S. (1998). *Database Law*. London: Jordans.

RoMEO report on transfer agreements between publishers and authors. http://www.lboro.ac.uk/departments/ls/disresearch/romeo/ (accessed 4 May 2004).

Shearer, K. (2004) Personal communication to Hardy, R., 28 May 2004.

Swan, A. and Brown, S.N. (2004) JISC/OSI Journal Authors Survey Report. http://www.jisc.ac.uk/uploaded_documents/JISCOAreport1.pdf. (accessed 10 June 2004).

Tunkel, D. a. Y. S. (2000*) E-commerce: A Guide to the Law of Electronic Business*. London: Butterworths.

van Veen, Theo and Oldroyd, Bill (2004). Search and Retrieval in The European Library (WWW document). http://www.dlib.org/dlib/february04/vanveen/02vanveen.html. (accessed 8 July 2004).

van Westrienen , G (2002) [OAI-prints] DARE: an new age in the provision of academic information. http://lists.openlib.org/pipermail/oai-eprints/2002-November/000009.html (accessed 11 June 2004).

Ware, M. (2004) PALS Pathfinder Research on Web-based Repositories: Final report. http://aims.ecs.soton.ac.uk/pep.nsf/cc4a508424b9c3ff802566dc004e42ff/5c4d447fc4fdeecf80256e46003c0c0e?OpenDocument (accessed 2 February 2004).

XTCat (2002) Experimental Thesis Catalog [WW document]. http://alcme.oclc.org/xtcat/index.html (accessed 2 February 2004)).

Young, Jeff (2003) [OAI-general] System Architecture (Email). http://www.openarchives.org/pipermail/oai-general/2003-February/000252.html (accessed 27 June 2004).

APPENDIX: TECHNICAL ISSUES

A.1 Protocols and standards

A.1.1 OAI

The OAI has developed a protocol for harvesting metadata from compliant archives. The protocol is known as the OAI-PMH (OAI Protocol for Metadata Harvesting) and it has been explicitly designed to enable metadata descriptions of resources to be exposed. Detailed information and instructions on implementing the OAI-PMH are documented on the OAI website (www.openarchives.org). In OAI-PMH terminology, metadata records are shared between *data providers* who expose archives of metadata, and *service providers* who harvest metadata describing resources from the data providers.

In the context of this study, the data providers are the institutions who offer e-print archives as a source of metadata pertaining to e-prints and the OA journal publishers who expose metadata describing journal articles. The service provider is the system proposed in this report.

OAI metadata records are made up of three parts:

- A ***header***, which includes a unique identifier and datestamp
- The ***metadata*** about the resource itself
- An ***about*** section which consists of administrative and rights metadata about the record

To be OAI-compliant, data providers must expose records that conform to the OAI-DC (Dublin Core) XML schema. However, they may also support other richer metadata formats, provided they are appropriately encoded in XML. Software for developing archives that support metadata harvesting has been produced by various players. The most commonly used and best-known software types have already been discussed in brief in section 3.2.5 and include:

Open source software:

CDSware (developed by CERN)
Dspace (developed by MIT)
Eprints (developed by Southampton University)
FEDORA (developed by the University of Virginia and Cornell University)

Proprietary and locally-developed software:

Ebrary (developed by Ebrary.com)
MPG eDoc (developed by the Max Planck Gesellschaft)
OPUS (Online Publications University of Stuttgart)
MyCoRe (developed by a consortium of universities originally led by the University of Essen)

Software such as Dspace and Eprints provide (configurable) user interfaces for creating institutional e-print archives. These interfaces accept submission of e-print resources and automatically create the necessary XML-encoded metadata records. The software supports metadata harvesting by complying with the OAI-PMH requests that service providers use to harvest metadata and build services.

A.1.2 RSS (RDF Site Summary)

RSS is an alternative protocol for providing descriptions of resources. It is the most appropriate protocol for embedding alerts, news or other current items into other web-based services, but has not been widely utilized by the e-print community.

According to the JISC Information Environment Architecture Standards Framework, Version 1.1 (http://www.ukoln.ac.uk/distributed-systems/jisc-ie/arch/standards/):

"*Where news channels are offered within the JISC IE, service components* ***should*** *use RDF Site Summary (RSS) 1.0 [20]. However, if necessary, service components* ***may*** *choose to use either RSS 0.91, RSS 0.92 or RSS 2.0 as alternatives to this format.*"

There are also other commonly-used protocols designed to enable information databases to be queried.

A.1.3 Z39.50

Z39.50 is a long-established protocol in the library community and many library catalogues and other bibliographic databases already have a Z39.50 interface. Though it is usually associated with distributed searching scenarios, there is no reason why it could not be used as a tool for gathering metadata. This is discussed further in section 4.2.3.1.2.

A.1.4 SRW/SRU

SRW and SRU are two closely related protocols, which are effectively 'next generation' implementations of Z39.50. Of the two, SRU is easier to implement. SRW works by employing HTTP POST requests with content wrapped in a SOAP envelope. SRU simply employs HTTP GET requests and returns responses in XML. Both protocols are simpler to implement than Z39.50.

By far the most commonly-used of these protocols within existing open archive developments is OAI-PMH. It would therefore seem to us most appropriate for data providers to use the OAI-PMH to expose their metadata. If service providers wish to further expose their metadata for subsequent use by another service provider or portal then they may opt again for OAI-PMH; in this case, however, Z39.50 and SRW/SRU might also be used for creating a machine-based interface to the metadata – e.g. for allowing access to the services by a third party portal or gateway.

So far, we have discussed only the protocols that support the harvesting and exposing of metadata. There are a range of additional standards, however, which may support any advanced services or service integration that may be required. Such standards include the following.

A.1.5 OpenURL

According to the JISC Information Environment Architecture Standards Framework, Version 1.1 (http://www.ukoln.ac.uk/distributed-systems/jisc-ie/arch/standards/):

"*Where context-sensitive linking is used within the JISC IE, service components* ***must*** *use OpenURLs that conform to the OpenURL version 0.1 specification [21] or the OpenURL version 1.0 specification [22]."*

OpenURL comes into its own in facilitating the search for multiple copies of an e-print resource. The role of OpenURLs in establishing the most appropriate version of an article is considered in detail later in this report.

A.1.6 DOI/HDL

The emergence of e-prints undoubtedly provides greater opportunities for discovery and access; however, because of the widespread use of URLs, the identity of an electronic document often becomes conflated with the location of that document. Of course, while its identity remains, the location of a document can change over time. This does not help the aims of resource discovery and management.

Digital Object Identifiers and Handles are persistent unique identifiers designed to rectify these problems. The DOI enables the identification and exchange of intellectual property in the digital environment, by providing "a framework for managing intellectual content, for linking customers with content suppliers, for facilitating electronic commerce, and enabling automated copyright management for all types of media".(DOI, 2004).

If information about a digital object and its location changes, its DOI will not. The use of DOIs as identifiers makes it much easier to create and maintain automated services.

DOIs are based on the CNRI Handle system (CNRI, 2003), which allows for the unique identification of organisations and resources. So, while a full DOI or handle (HDL) identifies a required resource, the root of that DOI or HDL uniquely identifies the organisation behind that resource.

Many open access journal publishers assign DOIs to articles they publish. Among the software packages available for running IAs, DSpace appears to be unique in assigning a HDL to a resource when it is added to a archive.

In the short term, it seems likely that DOIs/HDLs will be used by data providers to identify and locate resources, but that the use of DOIs/HDLs will not extend beyond this role.

A.1.7 SICI

The SICI (Serial Item and Contribution Identifier) is described in ANSI/NISO Z39.56-199X (1996) as: *"an extensible mechanism for the unique identification of either an issue of a serial title or a contribution (e.g., article) contained within a serial, regardless of the distribution medium (paper, electronic, microform, etc.)."*

In a study commissioned by UKOLN, Martin and Bide (1997) stated: *"The importance of unique identifiers as a key element of metadata in systems interoperability is well recognised; only the ISSN and its derivative identifier, the SICI, have the necessary characteristics to meet the requirement for unambiguous identification of the serials at the title, issue and article level."*

Therefore, it would be helpful, for identification purposes, if open access journal publishers were to routinely provide SICIs in the metadata they expose for journal articles. One mechanism for accomplishing this would be the adoption by publishers of the 'Onix for Serials' metadata scheme, discussed in section 4.2.3.8.

A.2 Existing Service Providers

Existing OAI-based service providers are listed below. They are also discussed again in section 7 where we address service models. We reviewed each of these to assess the features currently offered and which might be appropriate to include in the proposed model(s):

- Arc
- Callima
- CitebaseSearch (Citebase)
- CYCLADES
- MyOAI
- OAIster
- Open Archives Initiative Information in Engineering, Computer Science and Physics (OAIIECSP)
- Perseus (Pers)
- Public Knowledge Project Open Archives Harvester (PKP)
- SAIL-eprints (SAIL)
- Scirus (Scir)
- TORII

A.2.1 Search features

The norm was typically two-levels of searching, basic and advanced. Some systems allowed more wide-ranging access to data such as author affiliation, citation and full text. Some of these are detailed below. Research shows that most people search in a simple manner both when looking for 'known items' i.e., when they know all or some of the information about author or title, and when searching by subject, usually using one or two search terms at a broad subject level. It is therefore essential to understand how much beyond the simple level is required for the proposed model(s) given that a reasonably rich search environment is necessary.

A.2.1.1 Levels of search

Different service providers offer different levels of search, as follows:

- Citebase: can search metadata (standard), citation (search by standard citation), and OAI identifier.
- Perseus: can search metadata or whole content (i.e. search within full text)

A.2.1.2 Fields that can be searched

It is usually possible to search a number of different fields, e.g. author, title, abstract, subject, date. Some services allow searching specifying individual fields, while others only allow all fields to be searched at once.

- PKP also allows searching by author affiliation, sponsor and type of document (e.g. refereed articles, reviewed papers, dissertations), approach / method and coverage.
- OAIIECSP allows searching by report number / journal source
- OAIster permits restricting by resource type (includes image and audio)

A.2.1.3 Searching across multiple archives

Some service providers allow the user to specify which archives to search.

A.2.1.4 Search functionality

A variety of search features is available from the service providers, such as truncation, automatic word stemming, phrase searching, and Boolean searching (some use implicit AND, some require explicit use of operators, some have menus or check boxes that perform the same function):

- MyOAI claims to have one of the widest arrays of search options. Examples include *Soundex / metaphone / phonix* (different algorithms used to search for words that sound the same), *typo* (which searches for documents with terms that contain simple typographical errors), and *thesaurus* (which can expand the search).
- OAIster: This service plans to be able to sort by proximity and institution frequency, and also allow browsing by broad topical categories, and the removal of duplicate records from search results.
- PKP gives the Library of Congress Classification outline to help select search terms and phrases.
- Perseus: This service has an 'alternate names' option (for Greek and Roman materials), a matching field option (distinguishes things from and about particular places), a dynamic clustering option (so can find different forms of a term, and different topics associated with a term), and can deal with Latin and Greek search terms (to a certain degree).
- Scirus: This service includes intelligent query rewrites (rewrites the query entered to produce better results, e.g. adds quotes to common phrases, remove superfluous words), has list of common classification terms formed from top 100 results (and can use these to refine search), and can refine the search by subject area (20 areas in all).

A.2.2 Results options

A.2.2.1 General

In terms of how the results are presented there are a large number of possible permutations. The norm is a choice between 'brief' or 'detailed'. Some services give a list of the different archives from which hits were found, and state the number of hits found in each. Some arrange results by archive.
A wide variety of fields may be included in brief or detailed records and some of these may include external hyperlinks (e.g. URL, Identifier, Institution, Source archive). Examples of fields that may be included are:

- Abstract

  - Author / Creator
  - Contributor
  - Data stamp
  - Discovery date / Year
  - Institution
  - Item identifier
  - Language
  - Note
  - OAI identifier (external link)
  - OAI information
  - Publisher
  - Resource format
  - Resource type
  - Rights statement
  - Source archive
  - Sponsor
  - Subject
  - Title
  - URL

- OAIster returns an annotated list of the collections from which results were found. This includes hotlinks to the collection or the homepage of the hosting institution, and a brief description of the nature and scope of the collection.
- Arc produces results that include an internal link to the author index.

A.2.2.2 Options for display / organisation of results

Many service providers have options on the search screen that affect the way that results will be presented, such as the number of results to display per page, or the way in which they are ordered or ranked. Most allow the results to be viewed in different formats (e.g. 'brief' or 'full').

- Arc: this service can group results by archive / year / subject
- Citebase: can rank by number of citations / date / number of hits for author and can determine whether these are arranged in descending or ascending order
- OAIster: can sort results by 'hit frequency' (that is, it counts the number of instances the terms entered are encountered) or 'weighted hit frequency' (which gives a higher score to occurrences in particular fields). OAIster cannot sort if there are more than 1000 results, however.
- Scirus: Scirus has a separate form for search preferences such as rewriting, or the number of results per page

A.2.2.3 Format in which results are displayed

The results may be displayed in different formats (e.g. HTML, other). It is anticipated that most searchers will want a conventional bibliographical display but there will be a significant number who may want another format such as XML.

- Citebase: displays the format: HTML (default), XML, Refer, BibTeX.
- My.OAI: has a Rich Site Summary, that is, 'lightweight XML format designed for sharing headlines and other Web content'
- OAIster: uses a bold maroon colour to highlight query terms and phrases in search results, facilitating scanning the results

A.2.2.4 Filtering of results

Results can often be refined, for example by author / subject / date. Search refinement or limitation becomes increasingly important when dealing with large volumes of items retrieved.

- Arc permits searching within results using 'search last results' option and includes a horizontal slide to select dates for refining search.
- Callima uses field qualifiers to limit search
- OAIIECSP: this service permits the user to re-access the search strategy from the results page and modify or rerun it.

A.2.3 Other features

A.2.3.1 Annotation

- My.OAI: If the user has a my.OAI account, s/he can create annotations and can specify whether others can view the annotation or not.
- Torii: permits the user to add or view public or private comments

A.2.3.2 Alerting services

A few service providers have alerting services: registered users are emailed with details of new resources that match their enquiry. Examples of service providers that offer this are SAIL and my.OAI.

A.2.3.3 Citation services

- Citebase documents cited references and provides a listing of all articles citing this article (descending order by number of times citing paper has been cited) and a listing of the top 5 articles co-cited with this article.

- Citebase produces a graph of the document's citation/hit history. It also includes a table containing the citations identified, total number of Web hits, and the mean number of hits for an author.
- Citebase has various search and display options: abstract search to display standard, full Citebase record, request list of documents have been co-cited with it, request documents that have cited it.

A.2.3.4 Personalisation

Personalisation is an increasingly important feature of information portals and federated searching and it was important that this should be investigated as part of the proposed model(s).

- My.OAI: the user can access as a guest or as a registered user. There are extra features available to registered users: for example, they can set preferences, save or email records or searches, or add annotations. There is an email alerting service which updates to registered search strategy. Users can store the results in a personalised folder which can be accessed from nearly all pages.
- My.OAI includes various personalisation features, such as user information, search preferences, database selection defaults, data summary preferences, document retrieval preferences, data summary preferences, database selection defaults, and saved search defaults.
- Torii: if the user is registered and signed in, s/he has a personal folder to store documents in for future use. The user can define their profile of interests and the system will then order documents found according to their relevance to that profile. It can also evaluate documents stored and used as part of quality control.
- Cyclades: allows filtering on the basis of individual user profiles/profiles of working communities and is able to send recommendations about articles to others, support collaborative working with shared working spaces for documents, related links, annotations, ratings and recommendations, and so on.

A.2.3.5 Linking to related documents

- My.OAI has advanced features and functionalities; for example, when viewing any given document the user is offered the ability to list similar documents or list recommended documents (identified by search and retrieval patterns of previous users)
- PKP: this service has a research tool kit so that it can link to related research, web sites and databases. It helps the user to interpret studies and pursue a greater understanding of field by retrieving, for example, contextual information such as author biography, related studies, and online forums. It can also email the author of an article.

A.2.3.6 Options for viewing full text

Searchers generally want access to the full text as early and as easily as possible.

- My.OAI permits viewing in 'native' format (e.g. XML).
- OAIIECSP has different full-text access options (e.g. Postscript, PDF, Other)

A.2.3.7 Exporting results

- OAIIECSP allows viewing of the complete metadata record or 'add to a book bag' (collect OAIIECSP records in a separate collection that the user can then print or save)

A.2.3.8 Access statistics

- SAIL gives access statistics in the form of data and bar charts on usage and aggregated and browsing activity

A.2.3.9 Linking to another service

- Arc has a 'service' field with icon links to the DP9 service, which provides the Dublin Core record and links to metadata in alternative formats. For some records, it is possible to link to document references. [DP9 is a service being built by Arc: it is an open source gateway which allows general search engines to index OAI-compliant archives.]

A.3 Metadata

A.3.1 Obtaining the metadata

The primary mechanism for obtaining metadata should be harvesting employing the OAI-PMH. It should be accepted, however, that not all data providers will adopt the OAI, and some may expose their metadata through other protocols, such as Z39.50 or the next generation SRW/SRU protocols.

Data providers will largely consist of institutional archives (IAs), subject based archives (SBAs) and open access journals (OAJs) who expose OAI-compliant metadata. It will necessary to identify the relevant IAs, SBAs and OAJs, using the services OAI registries, OAI friends, and DOAJ (Directory of Open Access Journals).

Unqualified Dublin Core is at the heart of the OAI-PMH, and in fact at the time of writing the vast majority of data providers can *only* offer records in unqualified DC. It should be accepted that in the early days of proposed services unqualified DC would be the only widely available metadata format.

In the longer term, however, it would seem desirable for richer schemes to be adopted for metadata exchange. This is discussed later in section 4.2.4.8.

We recommend that metadata is harvested 'as is' and saved locally in a temporary store prior to further processing. This has a number of benefits when compared to the alternative of processing on the fly:

- It substantially reduces the harvesting time
- The possibilities of encountering network errors are decreased
- It allows for more comprehensive and safer pre-processing of candidate metadata before it enters the database, e.g. setting or verifying metadata semantics (which may vary between data providers), or identifying duplicates or different versions of resources already catalogued in the database

A.3.1.1 Initial harvest vs. ongoing harvesting

The pre-processing requirements for an initial harvest from any given data provider will be much more thorough and labour intensive than for subsequent harvests. Depending on the number of records to be harvested, the process may require a considerable length of time to download the records to a temporary local store. For example, the Experimental Thesis Catalog (XTCat) taken from OCLC's WorldCat, holds over 4 million records. To perform a complete harvest of the whole database could take up to six hours (XTCat, 2002). However, there are not many archives holding anywhere near this number of records, and most initial harvests are likely to take only a few minutes. Of greater importance, when first obtaining metadata from a new provider, is the need to understand the context, structure and semantics of their metadata. EPrints UK (2004) recommends a minimal set of metadata elements for archives exposing e-prints metadata. However, we can expect that not all archives will use all of the recommended elements and may use some or all of the other elements.

When considering subsequent harvests, of course, pre-processing to identify duplicates and so on is still required. However, as the OAI-PMH is designed to allow incremental harvesting after an initial full harvest, subsequent harvests need only pick up new or modified records. Combined with the use of scheduled tasks, this greatly reduces the need for human intervention, and helps to keep down costs.

It is necessary to decide how 'deleted' records be handled. Some archives retain and expose stub-records for deleted resources, others 'quietly' remove them. In the former case, it is easy to identify deleted records and then mark them as such or expunge them from the database. In the latter case, it will be necessary to carry out a full re-harvest periodically, and to run a comparison against previously harvested records in order to ascertain that records have been deleted. Again, the relevant records can be marked or expunged.

A.3.1.2 Harvesting into the local metadata store

Not all potential sources of metadata will necessarily be OAI-PMH data providers. It may be possible to make use of Z39.50 or SRW/SRU interfaces as well, in order to be able to harvest from any appropriate provider. The following diagram illustrates how metadata could be harvested and processed for entry into a metadata store using the OAI-PMH and other protocols.

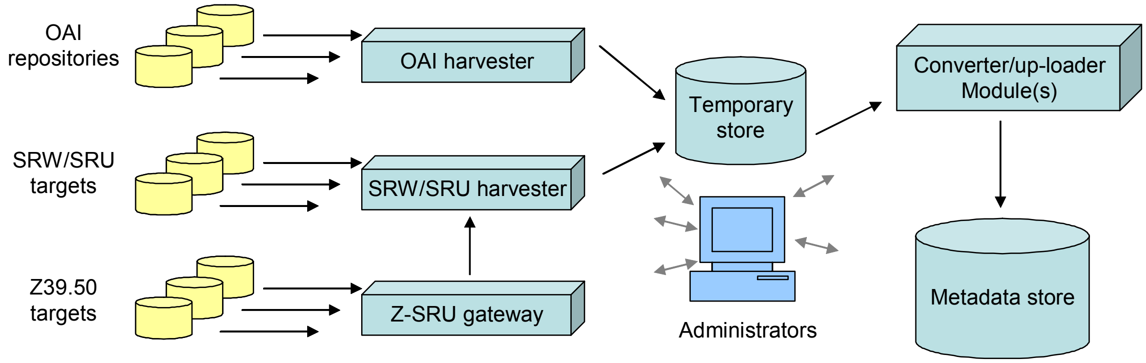

Figure 6. Metadata harvesting into local store

While the use of Z39.50 or SRW/SRU for harvesting purposes appears novel, we suggest that it is a feasible option where an OAI interface is not available. van Veen and Oldroyd (2004) "*anticipate that SRW and SRU will gradually replace Z39.50*". We agree with this statement, however, in the meantime, the use of a Z39.50 – SRU gateway "*allows existing Z39.50 services to be made available as SRU services without any change* [and] *existing investments in Z39.50 targets to be retained.*" Software to implement such a Z-SRU gateway is freely available from the British Library (http://herbie.bl.uk:9080/).

A.3.2 Storing the metadata

A.3.2.1 Storing XML

XML is important and central to the proposed system, providing the key to harvesting and exposing metadata, and it is necessary to consider how the XML metadata can be stored in the system. There are two main approaches to the storage of XML:

- in native file format
- in a relational database management system (RDBMS)

Either approach could underpin the proposed system, but if a hybrid SQL/XML solution is adopted it will be necessary to establish data exchange between the XML and a relational database. This can be difficult, owing to the differences in the structure and type of the data involved. The two main methods of storing XML in a RDBMS are to:

- break up the XML files and store them in a set of tables in a RDBMS
- store XML files in RDBMS as objects (no break-up)

Fortunately, there are a number of ways to deal with the problem of converting XML data into, and out of, relational databases. There are an increasing number of effective, automatic conversion tools which have been developed by database vendors such as IBM, Microsoft, Oracle, and Sybase. The solution chosen for storage will, in practice, be determined by the IT policies and technical skills available within the organisation which takes on operation of the system.

A.3.3 Metadata requirements

While unqualified DC is fundamental to metadata harvesting, we need to consider whether it is adequate as the format to support the services required of a service provider. For a very simple service, the answer has to be 'yes'. Storing the metadata as unqualified DC would remove a lot of the pre-processing requirements and questions about semantics. However, a service based entirely around unqualified DC would offer relatively limited added-value benefits for users. Specifically, the following issues would be difficult to manage

- The relationship between journal articles and e-prints
- The distinction between metadata describing the resource and metadata describing agents of the resource (such as creator)
- Publisher metadata
    - BaseURLs for harvesting
    - Collection level description
    - Journal and journal issue level description

For a service based on harvested metadata to be of any value, then a richer scheme than unqualified DC is required. For maximum interoperability, METS with qualified DC may be leading contenders; though, as discussed below in section 4.2.4.8, there are several other possibilities that should be considered.

Therefore, although many alternative metadata schemes exist, in the *short term* basing a solution on Dublin Core seems to be sensible. Not only does this negate the need for crosswalks between metadata schemes used by different data providers, but as any OAI-compliant archive must provide DC metadata, a base level of metadata can be assumed. In addition, if open access journal publishers can also be persuaded to expose metadata describing their articles in DC then any service provider should be able to integrate OA journals and e-prints. The ePrints UK project have produced guidelines on the usage of DC in the context of eprints (see: http://www.rdn.ac.uk/projects/eprints-uk/docs/simpledc-guidelines/). Based on information derived from those guidelines, the likely differences in semantics behind the different DC elements for open access journal articles and self archived e-print are described in the table below:

Element	Open access journal article	Eprint
DC.Title	The title of the article	The title of the e-print
DC.Creator	An author of the article	An author of the e-print
DC.Subject	The topic of the article	The topic of the e-print
DC.Description	A summary of the content of the article, typically in the form of an abstract	A summary of the content of the e-print, typically in the form of an abstract
DC.Publisher	The publisher of the article, typically the open access publisher	The publisher of the e-print, typically the author's institution
DC.Contributor	A contributor to the article (but not one of the primary authors).	A contributor to the e-print (but not one of the primary authors).
DC.Date	The publication date of the article	The 'last-modified' date of the e-print and/or the date of its accession into the archive
DC.Type	OnlineJournalArticle	The type of e-print (see discussion above or below).
DC.Format	The media-type of the article, e.g. application/pdf	The media-type of the e-print, e.g. text/plain
DC.Identifier	A URI or bibliographic citation for the article	A URI or bibliographic citation for the e-print, typically the URI of the 'jump-off page' for the e-print, as served by the archive
DC.Source	The URI, title or bibliographic citation for a resource from which the article is derived. e.g. Journal identifier	The URI, title or bibliographic citation for a resource from which the e-print is derived, e.g. archive identifier
DC.Language	The language in which the article is written.	The language in which the e-print is written.
DC.Relation	Bibliographic reference to the	The URI of each available

	journal issue	format/version of an e-print
DC.Coverage	The geographic location or temporal period that the article is about.	The geographic location or temporal period that the e-print is about.
DC.Rights	A human-readable statement about the rights held in and over the article, the URI of a Creative Commons [14] licence or the URI of a machine-readable statement.	A human-readable statement about the rights held in and over the e-print, the URI of a Creative Commons licence or the URI of a machine-readable statement.

A.3.4 Metadata relationships

The open access environment does not stand alone from traditional publishing. Open access journals are alternatives to traditional journals (ones in which the publisher raises revenue via alternative mechanisms to subscriptions). Self-archiving, on the other hand, is an alternative method of providing access to research that avoids the toll barriers imposed by traditional subscription-based journals. The co-existence of open access journals, traditional journals and e-print archives means that scholarly research literature may appear in more than one form and the relationships between the three types of publication, mean that, for example, the preprint of an article may be deposited in an institutional archive, the completed paper published in a traditional journal and the postprint also loaded into the archive. If the relationship between these three related resources can be maintained then users may be given appropriate choices and locations as to how to access content.

Relationships between aspects of the publication process also exist at other levels and include:

- preprints, postprints and journal articles – both in OA journals and traditional subscription journals
- journals, their issues and articles
- e-prints, collections and communities
- e-prints, journal articles and ancillary data
- published *versus* non-published research digital objects

A.3.4.1 Preprints, postprints and journal articles (both OA journals and traditional subscription journals)

We assume that a preprint has no formal relationship with a journal, issue or article until that article is published or accepted for publication subject to revision. While, in some cases, there will be an indication that the preprint is intended for publication in a particular journal, it must be remembered that this is only a tentative relationship until the article is actually published and the metadata have been updated to reflect this.

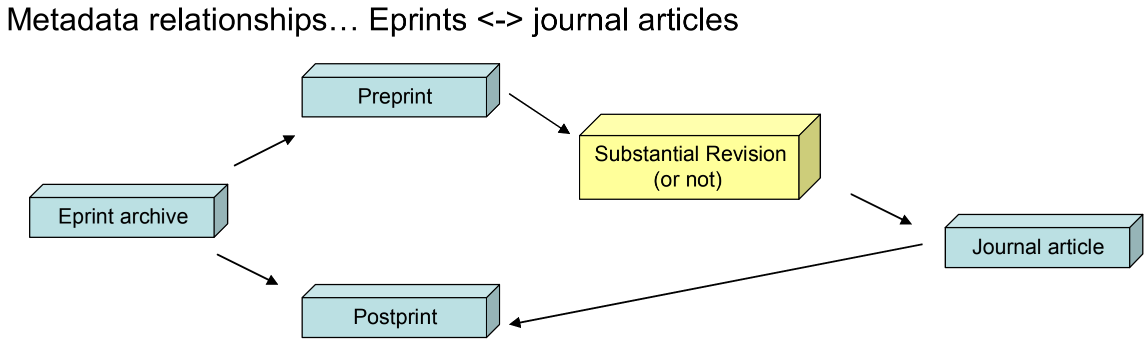
Metadata relationships… Eprints <-> journal articles
Preprint
Substantial Revision (or not)
Eprint archive
Journal article
Postprint

Figure 7. The relationship between e-prints and journal articles

It is possible, though not commonplace, for the preprint and published article to be identical: or it may be that the content is identical in substance, but the article has been edited to correct grammar or reduce the length, etc.; or it may be that substantial revisions have occurred between the preprint and the postprint (indeed, even key attributes such as the title may change). Maintaining a relationship between a preprint and a postprint is therefore a non-trivial task. Given that many institutional archives will operate by getting authors themselves to submit their e-prints, it is unlikely that authors can be relied upon to indicate relationships reliably between e-prints and traditionally published material. This is something that the model will need to take into account.

In addition, it is possible that conditions imposed by certain traditional publishers (e.g. that a preprint is the only e-print that is allowed to be archived), means that postprints may not be self-archived by authors. In cases like this, the Harnad-Oppenheim principle proposes archiving corrigenda alongside the preprint, effectively equating to the postprint in all but name. Under these circumstances it would seem appropriate that the corrigenda and preprint be encapsulated in some way. Again, this involves relating more than one entity in an archive, though in this case it seems likely that the author may in fact be the best person to specify the relationship.

Relating versions of e-prints is another issue that needs addressing. It is possible that there may be multiple versions of a preprint (for example, four authors at four institutions, with slight variations in each uploaded copy). In cases such as this a decision needs to be made as to whether each resource should be treated as a separate entity. If the archives are based on DSpace technology, then each resource is likely to have a different HDL and so presumably should be regarded as a resource in its own right. In these cases the most feasible solution would appear to be to consider each resource as an individual entity, though if a service provider were to harvest metadata describing these four resources it may be possible for the service provider to rationalise the records leaving just one record but with four possible locations.

As well as multiple distributed versions of a resource the possibility of multiple formats of the same resource needs to be considered. EPrints UK raises the following points (http://www.rdn.ac.uk/projects/eprints-uk/docs/technical/issues/):

Questions:

1. Should an e-print archive expose one record for multiple formats of the same publication?
2. If so, how should it disclose the multiple formats and full-text URLs in oai_dc?
3. If not, how are the multiple metadata records linked together in oai_dc?

There are arguments for and against having single or multiple formats.

If one record is to expose multiple formats then one solution is to repeat dc:format and dc:identifier per format (acknowledging that consuming applications can't reliably tie these back together). Alternatively, an e-print archive might wish to disclose the URL for a 'jump-off' page that links to all available formats.

If each format is considered an entity in its own right and has its own metadata record, then one solution could be to use dc:identifier for the format being described and dc:relation to provide the URL of each alternative format (acknowledging that there will be some redundancy in the total information disclosed by the archive because much of the metadata about each format will be repeated).

It may be that either approach would be used depending on the individual situation encountered. However, while the use of unqualified dc is so prevalent, it may be preferable to maintain separate records bound loosely by the DC relation element. Later, as metadata quality improves, these records can be grouped together within a structured METS or MPEG21 DIDL record. This is a matter that requires further consideration.

In summary, there are many relationships between e-prints and associated documents and files. The following list highlights the main reasons for specifying related files within a metadata record:

- To indicate the location of a traditionally published version of an e-print
- To indicate a postprint of a preprint (or vice versa)
- To indicate corrigenda or change for an e-print
- To specify equivalent files (e.g in an alternative format)
- To specify associated files (such as related data files, etc.)
- To specify a jump-off page

There are options as to who should create these relationships. They could be maintained by data providers, deduced by service providers and included in their metadata, or developed by third party agents such as library portals or even 'appropriate copy' (OpenURL) resolvers. However, it seems sensible that as much information as possible about any related files is created and maintained at source – i.e. within data providers' metadata. This could be achieved by using the DC **relation** element to specify related files. In order to distinguish between the various types of related files, qualified DC could be used. Qualifiers allowed for the **relation** element are:

- 'is version of'
- 'has version'
- 'is replaced by'
- 'replaces'
- 'is required by'
- 'requires'
- 'is part of'
- 'is part'
- 'is referenced by'
- 'references'
- 'is format of'
- 'has format'
- 'conforms to'

Although these qualifiers are sufficient to enable some of the above relationships to be expressed they are not descriptive enough in all areas – for instance, to indicate that a preprint has an associated postprint which can be located at a specific URL.

When considering ways to indicate related resources, it should be noted that there is an almost infinite number of combinations available to identify journal issues. Some common ones include:

- Issue *number*, *month, year*
- Issue *number*, *year*
- *month, year*
- Number *number*, *month, year*
- Number *number*, *season, year*
- Number *number*, *year*
- Volume *number*, Issue *number*, *month, year*
- Volume *number*, Issue *number*, *year*
- Volume *number*, *month, year*
- Volume *number*, Number *number*, *month, year*
- Volume *number*, Number *number*, *season, year*
- Volume *number*, Number *number*, *year*
- Volume *number*, *year*
- *year*
- *season, year*

This list is far from complete, and 'Volume' and 'Issue' may also be abbreviated to 'Vol.' and 'Iss.' Extending the dc:relation element to include these tags within the relation element would not appear practical; instead, it may be possible to use intelligent identifiers within the dc:relation tag to specify related journal articles, and one method for dealing with this may be the Serial Item and Contribution Identifier (SICI). The SICI is the most widely accepted identifier for journal articles and other items forming part of a serial publication in any medium. It was developed as ANSI NISO Standard Z39.56, and is based on the ISSN of the serial, the chronology of the article and the title and page number of the article. Alternatively identifiers such as DOIs could be used if known.

Given that in many cases it will be an author's responsibility to maintain (or to inform a data provider's administrator of) these relationships, it is unlikely that this kind of relationship would be effectively managed. Indeed, the relationship between preprints, postprints, and traditionally published articles may well be best managed by using existing 'appropriate copy' services described in a later section.

A.3.4.2 Allowable Formats

The above discussion implies that it is necessary to specify the format of a resource. The ePrints UK project has addressed this problem and states: "*Recommended best practice is to select a term from the IANA registered list of Internet Media Types (MIME type*s)" http://www.isi.edu/in-notes/iana/assignments/media-types/media-types

Likewise, Dspace supports a wide-ranging subset of the IANA list: http://libraries.mit.edu/dspace-mit/mit/policies/format.html

By default, Eprints software supports the following types:

- html
- pdf
- ps
- ascii
- other
- coverimage

A.3.4.3 Published vs. non-published research digital articles

The evolution of the academic publishing process has led to many different 'types' of articles. One area that needs consideration is the importance to readers of knowing the 'type' of article – i.e. whether it is a preprint, a postprint, a published article, a thesis, etc. This is particularly important in the case of federated search services. Although the scope of this document does not include grey literature, grey literature cannot be ignored, since preprints submitted to journals and then subsequently rejected may end up effectively being a permanent piece of 'grey literature' (GL). This begs the question 'what is the difference between preprints and GL?' The DAEDALUS project groups preprints with grey literature (http://www.lib.gla.ac.uk/daedalus/). Moreover, a lot of reports literature is now ending up on Eprint archives despite many definitions limiting e-prints to preprints and postprints. GL is outside the scope of this study, but has to be addressed for these reasons. A useful introduction to the GL landscape in the UK may be found in the MAGiC Final Report (Needham, Sidwell, Bevan and Harrington 2002). The MAGiC (Managing Access to Grey Literature Collections) project, sponsored by the British Library and the RSLP, proposed the creation of a national grey literature service built around the OAI harvesting model, which would incorporate both electronic and hardcopy – legacy – documents. The model, which was proposed to overcome the paucity of GL cataloguing in the UK (and beyond), bears similarity to and may complement services proposed in this study.

One possibility for distinguishing between preprints and postprints is for the 'type' of e-print to be maintained somehow in the metadata (perhaps using the recommendations proposed by eprints.org, discussed later in this report). Grey literature which is never intended for traditional publication could be indicated as such when deposited into a data provider's archive, whilst literature that is intended for publication could be labelled as a preprint. The dc:type element could be used for this purpose. This would enable a service provider to filter out intentional grey literature if required, whilst anything marked as a preprint could be obviously labelled as such for the benefit of users of the service.

The official digital object 'types' that are recognised by the **DCMI** are taken from a different perspective, and include:

- *Collection*
- *Dataset*
- *Event*
- *Image*
- *InteractiveResource*
- *Service*
- *Software*
- *Sound*
- *Text*
- *PhysicalObject*
- *StillImage*
- *MovingImage*

Of these types 'Collection' and 'Text' are applicable to proposed services; however, 'Text' is extremely vague. As ever, our recommended best practice is to select a value from a controlled vocabulary.

Eprints.org states: "*Decide what types of eprint should be stored (for example, refereed journal article, thesis, technical report, unpublished preprint*)" http://software.eprints.org/features.php

The **ePrints UK** project (http://www.rdn.ac.uk/projects/eprints-uk/docs/simpledc-guidelines/) suggests different types again:
Recommended best practice is to take the value of this element from the following list:

- *Book*
- *BookChapter*
- *ConferenceProceedings*
- *ConferencePaper*
- *ConferencePoster*
- *InCollection*
- *TechnicalReport*
- *OnlineJournalArticle*
- *JournalArticle*
- *NewsArticle*
- *Other*
- *Preprint*
- *Thesis*

with a secondary type imposed in the form of:

- *PeerReviewed*
- *NonPeerReviewed*

e.g. <dc:type>JournalArticle</dc:type><dc:type>PeerReviewed</dc:type>

would be used to specify a peer reviewed journal article.

It would be possible to recommend that a type list such as this be adopted when exposing metadata about e-print resources. However, incoming 'types' from institutional archives are uncontrollable and it may be necessary to map from types used by data providers on to those types listed above.

The leading software archive providers (DSpace and eprints.org) provide a mechanism for including dc:type information in exposed metadata. Both these solutions have default options that they use, and unfortunately they are different from each other and from the recommendations above:

Out-of-the-box, DSpace allows:

- Animation
- Article
- Book
- Book chapter
- Dataset
- Learning Object
- Image
- Image,3-D
- Map
- Musical Score
- Plan or blueprint
- Preprint
- Presentation
- Recording,acoustical
- Recording,musical
- Recording,oral
- Software
- Technical Report
- Thesis
- Video
- Working Paper
- Other

Out-of-the-box, Eprints.org software has:

- article
- book_section
- monograph
    - technical_report
    - project_report
    - documentation
    - manual
    - working_paper
    - discussion_paper
    - other
- conference_item
- book
- thesis
- patent
- other

However, individual institutions may customise these options if necessary.

In the case of eprints.org software, the type information is qualified by a "status", which can also be maintained. The status effectively indicates the status of a submission to the archive and whether it is grey literature or not. There appear to be three possible values:

- Unpublished
- In press
- Published

However, these values are not maintained in the metadata. They are used to define OAI “sets”, (but the use is inconsistent and not reliable for our harvesting purposes; see Eprints@bath below). Metadata harvesters, that is, the service providers, can request information from particular sets and therefore can build up a picture of what metadata belong in which set.

For a service provider to harvest e-prints (as pre/postprints of journal articles) and build a service upon the metadata then accurate identification of types is essential.

A brief survey of a sample of archives indicates how sets are used *i) Cogprints*

ListSet	Status = Unpublished
Types	Book Chapter Conference Paper Departmental Technical Report Journal (On-line/Unpaginated) Journal (paginated) Other Preprint Thesis

ListSet	Status = Published
Types	Conference Paper Conference Poster Journal (On-line/Unpaginated) Journal (Paginated) Newspaper/Magazine Article Thesis

ii) Eprints@bath

ListSet	Status = Unpublished
Code=noItemsMatch, Description="No items match. None. None at all. Not even deleted ones."	

ListSet	Status = In Press
Code=noItemsMatch, Description="No items match. None. None at all. Not even deleted ones."	

ListSet	Status = Published
Types	Conference Paper Conference Poster Other

ListSet	Department/Centre = Faculty of Science: Department of Computer Science
Types	Preprint Thesis

Note we have a ‘Preprint’ here, yet nothing ‘unpublished’ or ‘in press’ *iii) Glasgow Eprints Service*

ListSet	Status = ???TRUE???
Types	Journal article (on-line or printed)

ListSet	Status = ???FALSE???
Code=noRecordsMatch, Description="No items match. None. None at all. Not even deleted ones."	

ListSet	Subject = Q Science: QP Physiology
Types	Journal article (on-line or printed)

The use of "Journal article (on-line or printed)" makes it difficult to use the ePrints UK recommended types.

It would seem sensible for a standard approach to be adopted for maintaining 'type' information across e-print resources. Whether this simply uses the dc:type field with a value taken from a controlled vocabulary or combines this with a set-based approach needs further investigation.

If recommendations such as these are adopted, then one point which may need to be addressed is whether there is a difference between an article published in a journal (e.g. in an open access journal) and a postprint of a traditional article available in an archive. Metadata harvested from open access journals will typically describe 'Journal Articles', whereas metadata harvested from archives will be describing 'Postprints' (as well as Preprints). The solution developed to disseminate 'type' information must take into account whether distinctions such as this are important.

A.3.5 Categorisation

The purpose of categorising by subject is to enhance management and retrieval of resources through browsing and subject searching. So classification schemes, keywords, thesauri, and so on are central to the function of the majority of archives, gateways, portals, and services. It must be remembered that subject classification used by archives and journals is in the hands of the institutions and organisations producing the metadata.

Metadata schemes allow for controlled subject systems such as classifications, thesauri and subject headings to be used. When used consistently these allow much more flexible and focused subject retrieval. There are, however, a substantial number of institutions in the UK that have decided not to use such schemes for retrieval of books from their catalogue on the basis that free text (keyword) searching is sufficient. This position needs to be considered.

Part of the reason for the reluctance to use these schemes is to avoid the problems involved in choosing a particular controlled subject scheme. None of them are perfect and only Library of Congress Subject Headings and the Dewey Decimal Classification Scheme are widely used as general schemes. The table below gives a partial summary of some of the issues:

	Issue	Positives	Negatives
No controlled subject scheme	Freetext access on metadata	Facilitates authors assigning metadata. If the metadata includes an abstract, this may be a rich enough source for searching. Simplifies any model	Hit or miss/poor retrieval Ignores a major problem rather than solving it
Adopt a controlled subject scheme	Some subjects have accepted subject systems, thesauri	Authors likely to be more successful at assigning and retrieving	
	Many institutions have their own subject system, thesaurus, classification	Facilitates institutional administration	Lack of knowledge beyond home institution
	Controlled subject systems exist at different levels of specificity		
	Some portals have developed or modified subject systems	If usable, would offer economies of effort. May have automatic assigning	May not be usable for proposed model

		properties.	

Considerable effort is involved in assigning subject metadata if done manually, so clearly automatic or semi-automatic methods are a priority; something which might automatically scan the title of the e-print, identify the subject keywords, map these to a controlled subject scheme and automatically assign the appropriate subject headings or classification is a possible solution. Such possibilities have been developed in the area of library cataloguing schemes and need to be explored further. Qualified DC specifies most of the current general classifications and subject heading systems to be applied.

DSpace allows for:

subject		Uncontrolled index term.
subject	classification	Catch-all for value from local classification system; global classification systems will receive specific qualifier.
subject	ddc	Dewey Decimal Classification Number
subject	lcc	Library of Congress Classification Number
subject	lcsh	Library of Congress Subject Heading
subject	mesh	Medical Subject Headings
subject	other	Local controlled vocabulary.

http://www.dspace.org/technology/metadata.html

By default, **Eprints.org** software uses a subset of the Library of Congress subject as the default general subject classification, but only to three levels.

Tardis carried out a survey of e-print archives (http://tardis.eprints.org/discussion/eprintarchivessubjecttable15103.htm) and found most archives used in-house classification or terminology. Those who use established systems are usually academic institutions that link subject description to current practice with in-house material.

The Self Archiving FAQ (http://www.eprints.org/self-faq/#26.Classification), suggests that where e-prints (in the sense of preprints and postprints) are concerned, there is no need to adopt or worry about classification. While this view may hold water when talking about populating e-print archives at individual institutions, it has little validity when considering how subsets of the resources harvested into proposed services might be embedded in other services. In any case, unqualified DC itself puts limits on how much can be achieved in this area.

A key question is whether it possible to mandate/highly recommend some kind of subject scheme standards, such as 'University X must/should use ddc/lcsh/mesh, etc., to participate in JISC e-prints service'. This is currently unlikely to be acceptable to service providers because of the range of differences in practice which currently proliferate. A more likely approach would be to engender some further discussion amongst a group of stakeholders, to debate the issues and the various alternatives, to consider providing (or developing) a set of resources and tools to be available online and exploring automatic or (more likely) semi-automatic methods to add subject information. Tardis has already identified and anticipated many of these issues (http://tardis.eprints.org/discussion/).

A.3.6 Location of the Resource

It is likely that open access resources will be distributed across a range of data providers – institutional archives that hold resources from an entire institution, subject-based archives that hold resources in specific disciplines, and OA journals. The ability of a service provider (who has harvested records from a number of data providers) to filter them by data provider as part of the search/browse service it provides to a user (or portal) may be a requirement. In order to achieve this, the original location of the metadata record needs to be maintained. DSpace recommend that the dc:source element be reserved for, and used solely by, service providers to specify the original source of the harvested metadata. This recommendation seems appropriate.

Source		Do not use; only for harvested metadata.
Source	Uri	Do not use; only for harvested metadata.

(http://www.dspace.org/technology/metadata.html)

A.3.7 Exposing the metadata

Once a service has harvested metadata from data providers and built value-added services such as the simple search interfaces (e.g. ARC, Callima, etc.) or more complex services such as Citebase's citation indexing system, consideration should be given to the added value that could be provided by enabling portals (such as Metalib) to access the service. Technologies for exposing these metadata include reusing OAI-PMH, developing a Z39.50 interface or implementing SRW/SRU based interfaces.

As stated on the SRW Implementors' website (SRW Implementors, http://www.loc.gov/z3950/agency/zing/srw/implementors.html, WWW document accessed 7 July 2004), "OCLC Research has developed a 100% pure Java implementation of an SRW/U server using the Apache SOAP toolkit. The implementation specifies an interface to underlying databases. Currently the interface has been implemented for MIT's DSpace database (using Jakarta Lucene)". However, it would seem likely that a Z39.50 or SRW/SRU interface would be just as important for facilitating access to service providers' metadata. The principle of enabling service providers' metadata to be made available to external parties is currently in the news with the recent partnership of OAIster with Yahoo: "*The archive—developed through Michigan's University Library OAIster Project—is now available through Yahoo!'s Content Acquisition Program (CAP) and accessible through Yahoo! Search.*" (http://www.umich.edu/news/index.html?Releases/2004/Mar04/r031004)

A.3.7.1 Conventional Portal

Conventional portal technology could be used to provide access to the metadata stores of service providers:

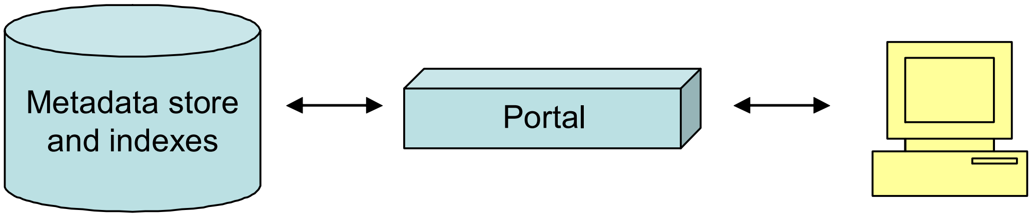

Figure 8. Conventional portal technology

In this scenario, the metadata are stored by the service provider either in a relational database (RDBMS) or a native XML database, and subsequently accessed using a standard portal application. Examples of suitable database applications include:

- Oracle
- IBM DB2
- Microsoft SQL Server

Examples of portal applications that could be considered include:

- Metalib (ExLibris)
- ZPORTAL (Fretwell-Downing Informatics)
- ENCompass (Endeavor Information Systems Inc.)

Any number of combinations is possible, but the potentially high costs of using commercial applications must be considered in addition to functional requirements.

A.3.7.2 Cheshire II Portal

The open source software, Cheshire II (http://cheshire.berkeley.edu/, accessed 13 July 2004) is "a next-generation online catalog and full-text information retrieval system using advanced IR techniques" and is intended to "bridge between the purely bibliographic realm of previous generations of online catalogs and the rapidly expanding realm of full-text and multimedia information resources."

Features of the Cheshire II system include:

- Support for XML as the primary database format
- It is a client/server application where the client communicates with the server using Z39.50
- The search engine supports both Boolean and probabilistic "best match" ranked searching, and permits the combination of the two
- The search engine and graphical user interface support browsing through automatically generated hypertext links, through "nearest neighbour'" searches and relevance feedback.
- User interfaces include a direct manipulation interface, command line and scripting interpreters, and support for WWW access and CGI scripting.

As evidenced by the launch of the ePrints UK Service Demo (http://eprints-uk.rdn.ac.uk/), the Cheshire II system, which underpins the ePrints UK system, may have the features and capabilities required to support proposed services.

A.3.7.3 Portal in a browser

Young (2003) suggested that *"Having harvested and aggregated from a number of OAI data providers into a database, it should be fairly easy to implement an SRW/SRU service (*http://www.loc.gov/z3950/agency/zing/srw/) *to search/retrieve the data. This would give you the power of Z39.50 with a simple HTTP GET or SOAP interface rather than the troublesome traditional Z39.50 interface. The user interface would amount to nothing more than HTML forms with the SRU server as the action target. The responses would come back in XML (SOAP in the case of SRW). It would be possible, though, to include an XSL stylesheet in the response so that it can be rendered as HTML in a browser."*

This novel approach, which has been adopted by the European Library (TEL) project (http://www.europeanlibrary.org), is discussed in an article by van Veen and Oldroyd (2004). It has two variants, which are illustrated below. The first approach uses client-side processing to transform XML responses into HTML, employing an XSL stylesheet and some JavaScript. Drawbacks to this approach are: not all browsers support XSL and JavaScript, and MS Internet Explorer requires a specific security permission to be set to allow XML responses from one server to be transformed by an XSLT script obtained from another server. The second approach uses server-side processing, employing an XSLT middleware component which provides browser independence. By using a combination of approaches, different audiences can be targeted using different architectures it possible to "*serve users who have limited control over their environment or have browsers with limited capabilities (for example, a mobile device).*"

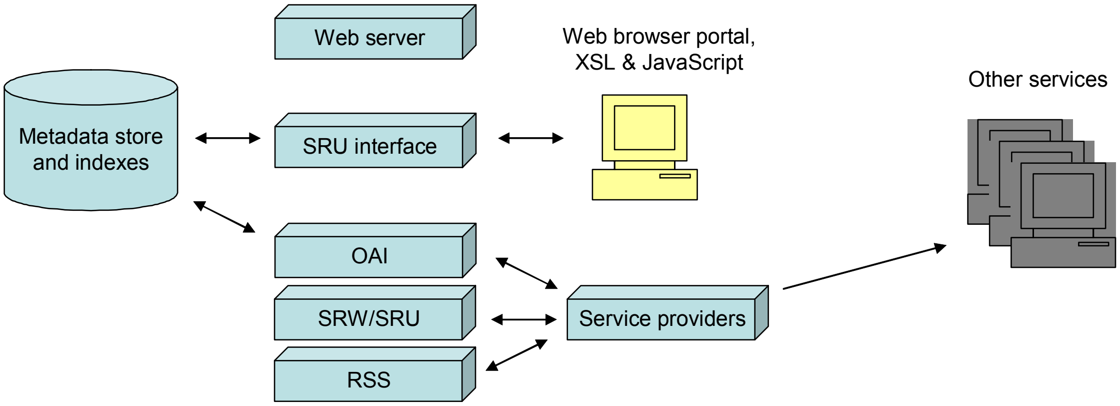

Figure 9. Client-side stylesheet translation

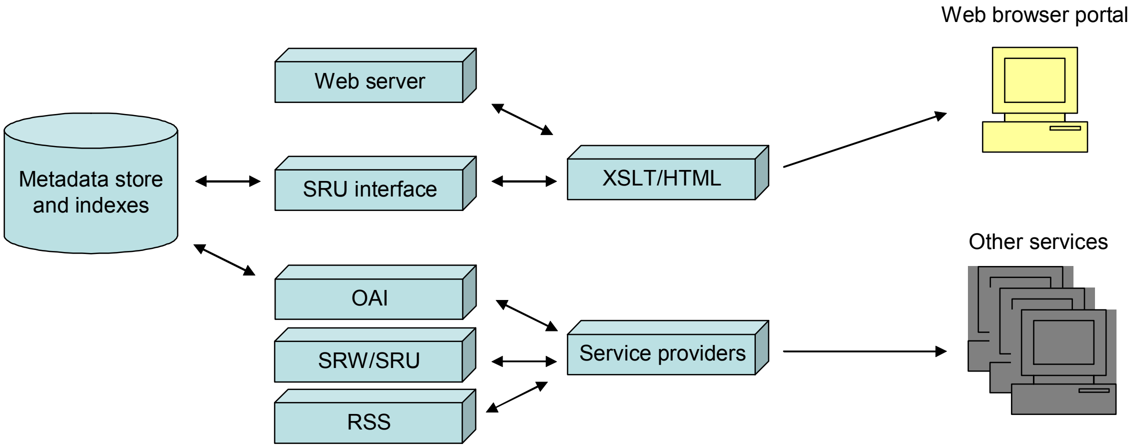

Figure 10. Server-side stylesheet translation

As stated by van Veen and Oldroyd (2004): *"the SRU portal proved to be more promising than the commercial portals—with the specific benefits of* *low cost and low risks—providing TEL with a low barrier of entry for new partners and new collections, and giving TEL full control over the functionality of the portal."*

This method of exposing metadata looks simple and elegant, and is worthy of consideration for use in any proposed national service.

In fact, to take this idea one step further, in addition to SRW/SRU providing the search capability for the native interface of proposed services, it would be possible to use the OAI-PMH to serve up browse pages via the ListSets mechanism. That is to say, the native interface would employ exactly the same protocols as other service providers!

A.3.8 Metadata schemes

The majority of services listed above in section 4.2.3 are based on metadata harvested from data providers. The quality of metadata that is exposed is therefore important in supporting both these services and any services proposed here.

Currently, unqualified DC is the only metadata scheme widely employed by OAI archives, and its prevalence makes it a strong initial candidate for any models proposed in this study.

As the National Information Standards Organization (NISO, 2001) observes, "*The simplicity of Dublin Core can be both a strength and a weakness. Simplicity lowers the cost of creating metadata and promotes interoperability. On the other hand, simplicity does not accommodate the semantic and functional richness supported by complex metadata schemes. In effect, the Dublin Core element set trades richness for wide visibility. The design of Dublin Core mitigates this loss by encouraging the use of richer metadata schemes in combination with Dublin Core. Richer schemes can also be mapped to Dublin Core for export or for cross-system searching. Conversely, simple Dublin Core records can be used as a starting point for the creation of more complex descriptions*".

We need to consider what other metadata schemes could have a place in proposed services in the future. There are many schemes employed for a wide variety of purposes in HE and FE; some of the most relevant ones are considered here.

Qualified DC

Unqualified DC is limited to a set of fifteen elements, qualified DC (http://dublincore.org/documents/dcmes-qualifiers/) allows for enrichment of those elements by permitting the use of two types of qualifier:

- **Element Refinement** – which makes the meaning of an element narrower or more specific.
- **Encoding Scheme** – which identifies schemes that aid in the interpretation of an element value.

The development of a qualified DC schema for use with the OAI-PMH would be a relatively easy and a small but useful step in moving forward towards "semantic and functional richness", which would enhance the functionality of any proposed services.

DC Library Application Profile
The DC-Library Application Profile (http://dublincore.org/documents/library-application-profile/), which is based largely on qualified DC, consists of several namespaces:

- Dublin Core Metadata Element Set, Version 1.1 [http://purl.org/dc/elements/1.1/]
- Dublin Core Qualifiers [http://purl.org/dc/terms/]
- Dublin Core Type Vocabulary [http://dublincore.org/usage/terms/dcmitype/]
- Dublin Core encoding schemes [http://dublincore.org/usage/terms/dc/current-schemes/]
- MODS elements used in DC-Lib application profile [http://www.loc.gov/mods]

This is the application profile which underpins the DSspace system. It allows for the addition of fields such as 'date-captured', 'edition', and 'location', which are useful in the context of proposed services. Again development of a schema for use with the OAI-PMH would offer great potential for increased "semantic and functional richness".

MARC21
In the future, it may be expected that some archives will choose to expose their metadata as MARC21 in an XML wrapper. In the context of proposed services, we feel that MARC21 is overly complex and that the MARC21 records should undergo a crosswalk to another more appropriate schema.

METS
METS (Metadata Encoding and Transmission Standard) appears to be an ideal candidate for future exposure of metadata by archives.

"The METS schema is a standard for encoding descriptive, administrative, and structural metadata regarding objects within a digital library, expressed using the XML schema language of the World Wide Web Consortium. The standard is maintained in the Network Development and MARC Standards

Office of the Library of Congress, and is being developed as an initiative of the Digital Library Federation." (http://www.loc.gov/standards/mets/)

A METS document (www.loc.gov/standards/mets/METSOverview.v2.html) consists of seven major sections:

- METS Header – *"metadata describing the METS document itself"*
- Descriptive Metadata – *"descriptive metadata external to the METS document"* or "*internally embedded descriptive metadata, or both"*
- Administrative Metadata – *"information regarding how the files were created and stored, intellectual property rights, metadata regarding the original source object from which the digital library object derives, and information regarding the provenance of the files comprising the digital library object"*
- File Section – *"The file section lists all files containing content which comprise the electronic versions of the digital object. <file> elements may be grouped within <fileGrp> elements, to provide for subdividing the files by object version."*
- Structural Map – *"outlines a hierarchical structure for the digital library object, and links the elements of that structure to content files and metadata that pertain to each element"*
- Structural Links – *"to record the existence of hyperlinks between nodes in the hierarchy outlined in the Structural Map"*
- Behavior – *"to associate executable behaviors with content in the METS object"*

METS effectively acts as a container for other metadata schemes. The 'Descriptive Metadata' section may contain unqualified DC, MODS or MARC21 metadata, and there is no reason it could not contain qualified DC or DC Library Application Profile metadata. In the context of proposed services, the adoption of METS as a standard for metadata exposure would be a major advance in semantic and functional richness. It offers a structure which can handle versioning and manifestations of digital objects, as well as preservation metadata. It is therefore worthy of further consideration.

MPEG21 DIDL

Another highly structured metadata schema of interest is MPEG21 DIDL. Beckaert, Hochstenbach and Van de Sompel (2003) observe that *"Digital Libraries have reached a point where acceptable architectures must accommodate objects that aggregate datastreams of a wide variety of media-types, and must allow for the association of secondary data – including metadata supporting discovery, digital preservation and rights management – with those datastreams."*

MPEG21 DIDL, which bears similarities to METS, has:

- A modular architecture
- The ability to accommodate any media type and genre
- Applicability to Digital Libraries

As with METS, this appears to a very useful and promising schema, and again deserves further consideration.

Onix for Serials

ONIX for Serials, which is a work in progress, is being designed to describe serial products and subscriptions. At the time of writing (July 2004), it consists of three parts:

- SPS (Serials Products and Subscriptions) is a format for communicating serials product catalogue information and/or details of subscriptions held by specified institution(s).
- SOH (Serials Online Holdings) is a format for communicating electronic serials holdings details from publication access management systems to user libraries.
- SRN (Serials Release Notification) is an article or issue level format for communicating details of the content of a serials release, or indeed of any selected set of journal issues or articles

Of these, in the context of open access journals, the SRN format appears interesting and may provide a format for publishers to expose their article and issue metadata. Of particular interest is the proposed use of the SICI as an unambiguous identifier for journal issues and articles (see section 4.2.1.7). A watching brief should be set to monitor developments.

LOM

In the future the relationship between e-prints, open access journal articles and learning objects may become important. Hence, the Learning Object Metadata (LOM) family, in particular the RDN/LTSN LOM Application Profile is discussed here.

According to Powell and Barker (2004), the RDN/LTSN LOM Application Profile has been designed to support record sharing between RDN and LTSN services using the OAI-PMH. *"However, this application profile will also be used by the Learning and Teaching Portal and may be treated as a candidate application profile for use by X4L projects."*

Importantly, in addition to metadata fields supported by DC it provides information useful in an educational setting – giving the 'learning resource type' and 'context' (higher education, further education, etc.).

While some archives may be set up with a learning context in mind, most are being set up to disseminate institutional research. It is difficult to see how learning attributes would be added, or who would add them, in the context of this study.

Collection Description

Collection Level Descriptions (CLDs) are a useful tool for providing an overview of the content and coverage of collections. CLDs have received little consideration in current archive implementations, though DSpace does have a primitive non-standard type of CLD. Happily, in future releases, DSpace developers are considering introducing the use of standard CLDs which would conform to the RSLP Collection Description Schema (http://www.ukoln.ac.uk/metadata/rslp/schema/) or the DC Collection Description Application Profile (http://www.ukoln.ac.uk/metadata/dcmi/collection-application-profile/2004-02-01/).

The use of CLDs would make it considerably easier to determine the nature and contents of any given collection before harvesting it, and the introduction of subject classification at collection level would facilitate at least some crude level of subject inheritance for the items held within the collection.

Access routes and appropriate copy

Close integration of open access literature with traditionally published literature is an interesting area. Products such as Elsevier's Scirus already index articles from open access journals along with non-OAI papers (http://www.scirus.com/srsapp/). Similarly the DP9 project enables OAI archives to be indexed by search engines.

However, one interesting area is whether current appropriate copy technology can be used to provide users with alternative locations and access routed for the literature. As mentioned earlier, the **dc:relation** element may be used in records describing OA resources to identify the location of related resources, and theoretically it would be possible for the location of a traditionally published version of an academic article to be 'linked' to an e-print record in this way. However it is possible that these relationships will not be maintained in pre-prints as this information may not be known at the time the article is submitted to the archive.

Given this fact, there are two access routes between traditional and open access literature that may be applicable:

- If a user has located an e-print by searching an open access service provider, can a traditionally published version of the article be located?

- If a user has located a traditionally published version of an article, can an open access version be located?

In other words, could appropriate copy technology be used to locate a traditional version of an e-print (and vice versa)? The appropriate copy issue initially seems redundant in the first case as the identified resource would, by default, be open access; however, the resource may be a preprint and it may be useful to give users other locations for the resource (or indeed links to a postprint or a publisher's version if appropriate). A solution to this problem may be possible if sufficient metadata could be returned to an OpenURL-based resolver to enable it to locate published versions. In fact, DSpace already has the capability to construct OpenURLs, which can be passed to a local SFX server, or other OpenURL resolver. While the construction of those OpenURLs needs further work to achieve full compliance with standards, some experiments by the project team, at Cranfield University, have suggested that the OpenURLs do contain sufficient information about the e-print to enable the OpenURL resolver to locate non-e-print versions of the resource. This offers the possibility that portals such as Metalib could provide searching of open access services (through OAI-PMH or Z39.50 interfaces) in the same way that existing databases are searched, and any hits could be related to traditional articles, if need be, by the Open URL resolver.

The approach of ePrints UK to this matter is: *"If possible, repeat this element [the dc:relation] to provide a full bibliographic citation*" and *"If possible, also repeat this element to provide an OpenURL"*. If an OpenURL could be maintained within the dc:relation element, then local resolution of the OpenURL should be fairly simple to implement, (perhaps by implementing the cookie-pusher technique at the data or service providers). However, key questions remains as to whether a 'fully' specified OpenURL is likely to be maintained for certain types of literature such as preprints.

The reverse of this approach is to consider the second case and ask whether it is possible to locate an Open Access resource, given metadata returned to a portal such as Metalib that identifies a traditionally published resource. When the appropriate copy is determined for the hit, it would be nice if locations of open access e-prints with the same title and author were returned as alternatives (with the caveat that they be labelled e-print, preprint, etc.). In order to facilitate this, could OpenURL resolvers such as SFX make a real-time query to a list of known service providers to determine if the resource in question is available through open access methods (presumably the service provider would need Z39.50 or another such protocol enabled for this)? The alternative to this real-time resolution of appropriate copy is to take an approach similar to Scirus, and for portals such as Metalib to harvest metadata either from the service providers or the original data providers and maintain it locally.

A.3.9 Preservation of metadata

There is a need for administrative and preservation metadata as well as metadata for discovery. Administrative metadata is required to manage and administer material. This includes content description and structure, acquisition information and technical dependencies. This data will also help in the preservation of material; it will inform the choice of preservation strategy and facilitate disaster recovery and determination of authenticity. Information about processing and preservation actions is also required.

There has been a great deal of work on the development of preservation metadata in different fields. Library-focused work is relevant to e-prints because they are similar to the sort of digital publications libraries are likely to collect. Initial work on preservation metadata for publications resulted in high level metadata schema. OCLC and the Research Libraries Group have been working on bringing together strands of work and developing a framework based on consensus. The two organisations are now sponsoring work on exploring how preservation metadata can be implemented. James *et al.* (2003) also suggest that record-keeping metadata schemes may also be relevant to e-prints. There seems to be little point of duplicating existing and ongoing work in these areas. It would seem to be more productive to examine this work and decide what would be appropriate to use in an e-print archive setting. Another area of work that would be worth looking at is how to link discovery, administrative and preservation metadata together and with e-prints. James *et al.* suggest use of XML formatting and the Metadata Encoding & Transmission Standard, “an XML document format for encoding metadata necessary for both management of digital library objects within an archive and exchange of such objects between archives (or between archives and their users)” (Library of Congress, 2001).

While archives may depend on authors to supply the bulk of the discovery data for e-prints, the archives will have to create and maintain much of the preservation metadata.