

# Feasibility and Requirements Study on Preservation of E-Prints

Report Commissioned by the Joint  
Information Systems Committee (JISC)

October 29, 2003

Hamish James, Arts and Humanities Data Service  
Raivo Ruusalepp, Estonian Business Archives  
Sheila Anderson, Arts and Humanities Data Service  
Stephen Pinfield, SHERPA, University of Nottingham

## Arts and Humanities Data Service (AHDS)

The AHDS is a UK national service funded by the Joint Information Systems Committee and the Arts and Humanities Research Board. Organised via an Executive at King's College London, and five Centres at various Higher Education institutions, the AHDS aids the discovery, creation and preservation of digital collections in the arts and humanities.



## Estonian Business Archives

Estonian Business Archives is a privately owned archives service. Estonian Business Archives offer consultancy in digital preservation, digital document and archives management.



## SHERPA at the University of Nottingham

SHERPA is a three year project funded by JISC and CURL (Consortium of University Research Libraries). SHERPA aims to investigate issues to do with the future of scholarly communication and publishing. The project is investigating the IPR, quality control and other key management issues associated with making the research literature freely available to the research community. It is also investigating technical questions, including interoperability between repositories and digital preservation of e-prints.

SHERPA is hosted by the University of Nottingham and involves a number of partner institutions in the UK.



# 1 Contents

1	Contents .....	2
2	Executive Summary .....	3
2.1	Background .....	3
2.2	Properties of E-Prints .....	3
2.3	Reasons to Preserve E-Prints .....	4
2.4	E-Print File Formats .....	4
2.5	Infrastructure and Development Needs .....	5
2.6	General Conclusions .....	6
3	Recommendations .....	8
3.1	E-Print File Formats .....	8
3.2	Preservation Metadata .....	9
3.3	E-Print Preservation within the JISC IE .....	9
4	Acknowledgements .....	12
5	Introduction .....	13
5.1	Background to JISC Strategy .....	13
5.2	Preserving E-Prints .....	13
6	Properties of E-Prints .....	15
6.1	Defining the Term 'E-Print' .....	15
6.2	E-Print Repositories .....	16
6.3	Subject Based and Institutional Collection .....	16
6.4	Self-Archiving, the Open Access Movement and Publication .....	17
6.5	Lifecycle of an E-Print .....	18
6.6	Why Preserve E-Prints? .....	20
6.7	The Future of the E-Print .....	21
6.8	E-Prints in the UK Academic Domain .....	22
7	E-Print File Format Review .....	25
7.1	File Format Requirements For an E-Print .....	25
7.2	Current File Formats in E-Print Repositories .....	25
7.3	Risk Assessment of Common E-Print File Formats .....	27
7.4	Reducing Preservation Risks Associated with Current E-Print File Formats .....	31
7.5	Recommendations for E-Print File Formats .....	32
8	E-Print Metadata Review .....	34
8.1	Documentation as Metadata .....	34
8.2	Resource Discovery Metadata for E-Prints .....	35
8.3	Preservation Metadata for E-Prints .....	35
8.3	Recommendations for E-Print Metadata .....	40
9	Cost Models for Preserving E-Prints .....	41
9.1	Introduction .....	41
9.2	Cost Elements .....	41
9.3	Taxonomy of Archives .....	44
9.4	E-Print Lifecycle Cost Elements .....	45
10	Organisational Models .....	48
10.1	Responsibilities and Roles .....	48
10.2	Functional Requirements .....	48
10.3	Non-Functional Requirements .....	50
10.4	Preserving E-Prints in a Disaggregated Environment .....	53
10.5	Organisational Models in the JISC IE .....	55
10.6	Recommendations for Preserving E-Prints in the JISC IE .....	55
11	References .....	58
12	Appendix I: Additional Sources For File Formats .....	63
13	Appendix II: Survey Documents .....	65

## 2 Executive Summary

### 2.1 Background

E-prints and institutional repositories are a new and high profile area, both for the JISC and for institutions in the UK and elsewhere. The initial focus of activity has been on the process of establishing repositories, depositing articles and promoting discovery and access, together with an emphasis on encouraging the cultural change necessary for successful development of e-print repositories. This focus is reflected in the JISC funded Focus on Access to Institutional Resources (FAIR) programme. However, if the e-print content of these repositories is to continue to be made available into the future, the concept of preservation needs to be brought into the equation.

This *Requirements and Feasibility Study on Preservation of E-Prints* seeks to do just this, providing recommendations for further research and the development of services and tools to support the long-term preservation of UK e-print content, in the context of the JISC Information Environment (IE) and the JISC Continuing Access and Digital Preservation Strategy 2002-5 (Beagrie, 2002).

The Continuing Access and Digital Preservation Strategy 2002-5 details JISC's continued commitment to the development of the digital preservation agenda within the UK Higher and Further Education sectors. The strategy revolves around the need to move from research to service delivery as the volume and significance of digital resources continues to grow. An initial target of the strategy has been to complete digital preservation risk and retention criteria assessments during 2002 and 2003. This study is one of the series of assessments initiated under the strategy.

### 2.2 Properties of E-Prints

An e-print is a digital duplicate of an academic research paper that is made available online as a way of improving access to the paper. E-Prints are divided into *preprints*, papers that are circulated before they have been formally approved for publication, and *postprints*, papers that have been approved for publication.

A key characteristic of e-prints is that they are created and managed in a way that emphasises quick and easy distribution to a wide audience. Preprint e-prints are not necessarily subject to formal quality control procedures, such as peer review, the absence of which speeds up the dissemination of research results. All e-prints benefit from being electronic documents made available via the Internet, which provides instant access for readers globally.

E-Prints are not distinguished from other digital material collected by libraries or archives by any unique technical characteristics. The file formats, metadata requirements and software applications used to manage and view e-prints can all be used to manage and disseminate other forms of digital content.

These characteristics, coupled with the idea of authors *self-archiving* their own work in e-print repositories, form the basis of a future envisioned by many in the open access movement where e-print repositories will provide unlimited access to research literature, providing an alternative, less expensive and more effective method of distributing research than traditional journal publication (Harnad, 1994). Equally, the value of e-prints as a quick and easy way of sharing information means that commercial e-print services are also being established.

E-Prints are found in a variety of settings, including large formally managed e-print repositories, smaller more informally managed repositories, but also in scattered collections stored in the Web sites of individual projects or academics. The term *e-print repository* is typically associated with online services that operate using specialist management software. The first e-print repositories were set up in large research centres and organised around specific subjects. More recently, individual institutions have started setting up their own e-print services, thus creating *institutional e-print repositories* that aim to collect and disseminate e-prints written by members of the institution. Institutional e-print repositories

may be seen as a limited type of *institutional repository* (a digital repository broadly conceived of as an archive of the total, digital, research output of an institution).

## 2.3 Reasons to Preserve E-Prints

A recurring sentiment we found during this study was that the requirements for running successful e-print repositories in the long-term are still being formulated, and that e-print repository managers have yet to fully engage with the challenges of digital preservation. UK e-print repository managers are aware of the need to consider preservation, but are unsure of how to proceed, and are looking for guidance.

Not all e-prints need be preserved. Some e-prints are early drafts of papers that are superseded with later versions; some authors are not interested in keeping their work for posterity; the intellectual property rights held in some existing e-prints may not be stated clearly enough to permit their long-term retention. Five criteria can be identified that provide a basis for determining if the long-term preservation of an e-print should be considered:

- The e-print provides wider and/or more convenient access than alternatives such as published journal papers
- The e-print is cited by other scholars
- The e-print contains unique information, not recorded elsewhere
- The e-print forms a significant part of the research record (for example, early drafts of important research)
- The e-print is part of a wider collection deemed worthy of preservation

The criteria above are offered as the basis for discussing the archival appraisal of e-print collections. To make practical use of these criteria, repository managers will need to develop objective measures that take account of local conditions and the interests of relevant stakeholders. Key considerations will include authors' wishes regarding retention period, which may be ascertained at deposit through a formal e-print deposit agreement, and reader's wishes regarding continued availability of e-prints (especially preprints not available elsewhere). These are very likely to vary between research communities, particularly if scientific subjects are compared to the humanities. The role of e-print repositories within wider institutional information management policies will also be important, as the preservation of e-prints, like any other digital material, will require a steady long-term commitment of resources.

There are only approximately 5,000 readily identifiable e-prints currently housed in subject based or institutional repositories within the ".ac.uk" domain. Consequently the preservation of the current UK e-print collection should not be the main concern. Instead, this report makes recommendations focused on improving current e-print repository practices so as to reduce the preservation risks associated with e-prints collected in the future.

## 2.4 E-Print File Formats

The usability requirements of readers, rather than the level of risk they present for long-term preservation, largely determines the choice of file formats for e-prints. Repositories usually accept a mix of proprietary and open standards formats. Although e-print repositories can technically accept files in any format, most repositories limit the list of accepted formats to those most commonly used by their user community. Frequently occurring e-print file formats are PDF, PostScript, TeX, HTML, MS Word, TeX and TIFF.

Proprietary file formats usually present the greater risk to the preservation of digital content over the long-term (Wheatley, 2001; Jones & Beagrie 2000, p. 131). Conversion to open standards based formats (for example, XML based formats) offers a good option for minimising the level of risk. The nature of the content of e-prints and their authenticity conditions do not set high requirements for conversion or preservation processing – adequate content replication in a format that can be used with current software is the ultimate aim of the preservation of e-prints.

While the adoption of open standards based formats for e-prints is probably the best option for minimising the risk of losing access to e-print content in the long-term, from the point of view of widespread use and compatibility, PDF and RTF are appropriate formats for e-prints. Although some issues remain with long-term preservation of these formats, their almost universal acceptance and the availability of specifications somewhat lowers the risk of the future loss of e-print content held in these file formats.

The content within an e-print typically adheres to the limitations of paper printing: they are mainly textual documents, but may also include equations, formulas and static images. In a digital environment, these restrictions no longer exist and it seems unlikely that the content and layout of e-prints will continue to adhere to these artificial limits in the future. E-Prints that contain databases, audio, video or other types of content will require far more attention in terms of description and preservation than is currently practised by e-print repositories.

### 2.4.1 Metadata

The metadata efforts of e-print repositories are at present centred on the Open Archives Initiative (OAI) Protocol for Metadata Harvesting (OAI-PMH) and the creation and maintenance of simple Dublin Core resource discovery metadata. Little or no administrative or preservation metadata is created.

Few e-print repositories have been in existence for a sufficiently long period of time to encounter significant digital preservation problems, or to have had to migrate substantial portions of their collections. Hence, the virtual absence of preservation metadata has not, as yet, proved a problem for e-print repository managers, but the lack of administrative and preservation metadata in e-print repositories is probably the biggest obstacle to the successful long-term preservation of e-prints.

Preservation metadata ... is the information necessary to maintain the *viability*, *renderability*, and *understandability* of digital resources over the long-term. Viability requires that the archived digital object's bit stream is intact and readable from the digital media upon which it is stored. Renderability refers to the translation of the bit stream into a form that can be viewed by human users, or processed by computers. Understandability involves providing enough information such that the rendered content can be interpreted and understood by its intended users.  
(OCLC, 2002)

In short, preservation metadata ensures that the content of an e-print can remain accessible in the long-term. Its value becomes clear when the formats used to record e-prints and the software used to display them becomes obsolete, a point not yet reached for most e-prints. The collection of preservation metadata should not, therefore, be seen as conflicting with the access orientated goals of e-print repositories, but rather as a sensible precaution that will help ensure that e-prints remain accessible.

Most e-print repositories rely on authors to provide resource discovery metadata, and this carries an inherent incentive to keep the metadata schema simple and short. Administrative and preservation metadata would need to be attached to e-prints by repository managers. Extending repository management software to support these types of metadata should encourage the creation of adequate metadata to support the long-term preservation of e-prints.

The report makes some recommendations for preservation metadata elements for e-prints.

## 2.5 Infrastructure and Development Needs

An institutional repository can fail over time for many reasons: policy (for example, the institution chooses to stop funding it), management failure or incompetence, or technical problems. Any of these failures can result in the disruption of access, or worse, total and permanent loss of material stored in the institutional repository.  
Lynch (2003)

Recent work has begun to establish the archival requirements of digital repositories (OAIS 2002, pp. 3-1-3-5; RLG 2002 [Is this, by any chance the same as the OCLC 2002 also referred to on this page (above)?]). E-Print repositories do not meet these requirements, and thus are not currently best placed to provide for the long-term preservation of e-prints. It is not necessarily the case that this should be a matter for concern. E-Prints are primarily promoted as a means of making research literature freely and globally available, not as a means of preserving that literature (Harnad, 2001). Nevertheless, given the poor arrangements for digital preservation in most e-print repositories, there is little guarantee that they will be able to ensure that the contents of e-prints remains accessible in the long-term. [RR: I meant to distinguish the aspect of technical access to the content of digital objects from their general availability that Harnad talks about when he says 'access'.]

In the UK, subject based e-print repositories have not yet made a significant impact, while institutional repository development is still at an early experimental stage. The scale and organisation of e-print repositories as a means of disseminating research literature is still developing. JISC is currently funding a number of projects with the FAIR programme that are investigating the development of e-prints as a means of scholarly communication. The SHERPA project in particular includes within its remit an investigation of digital preservation issues relevant to institutional repositories. Given these current activities, it might be considered premature to devote additional resources to the preservation of e-prints. However, a decision to defer action is just that, and it would be necessary to revisit the preservation issue soon. Given that any eventual decision to start preserving e-prints is likely to become more expensive to implement the longer it is left, there are good reasons to continue work in a number of areas:

The varying settings and responsibilities for managing e-print repositories means that the provision of support and services for the preservation of e-prints must be flexible and capable of responding to different levels of need depending on the situation. Specifically, e-print collections managed by large institutions (whether as part of a disciplinary based collection or an institutionally based collection) are likely to require less external support than collections housed in small institutions. What may emerge in the UK is a mix of institutional repositories, consortia of repositories and some subject base repositories, backed by national services providing varying levels of support and services according to the resources and expertise available to repositories.

Regardless of the manner in which they are provided, preservation services should not add to the real or perceived barriers that discourage authors from depositing their work in e-print repositories.

This report makes recommendations to improve the provision of preservation services for e-prints in the context of the JISC Information Environment.

## 2.6 General Conclusions

It is too early to recommend a single approach to the preservation of e-prints, or to even assess the full scale of the issue. In the UK, subject based e-print repositories have not yet made a significant impact, while institutional repositories' are still at an early experimental stage. Given the small size of the current UK academic e-print collection, attention can be focused on preparing for the preservation of future e-prints rather than securing the preservation of existing e-prints.

Considering the technical characteristics of e-prints, the organisational environments they are managed in, and the focus of current developments, this report has reached three general conclusions:

### Technical Characteristics

E-Prints do not present unique technical challenges for preservation. Generic preservation strategies, such as conversion to standard formats, migration, migration on demand, and emulation, could all be employed to preserve e-prints over long-term. Lessons from other studies funded by JISC will help to inform the choices to be made (Curl Exemplars in Digital Archives [CEDARS], <http://www.leeds.ac.uk/cedars/>; Creative Archiving at Michigan and Leeds Emulating the Old On the New [CAMiLEON], <http://www.si.umich.edu/CAMiLEON/>; "The File Format Representation and Rendering Project, [http://www.jisc.ac.uk/index.cfm?name=project\\_fileformat](http://www.jisc.ac.uk/index.cfm?name=project_fileformat)).

The static, mainly textual, content of e-prints suggests that conversion to standard formats and/or migration will adequately preserve the intellectual content of e-prints. In the future, other approaches such as emulation may become more important depending on the extent to which dynamic features appear in e-prints. The development and implementation of technical strategies for the preservation of e-prints should be treated as part of the wider development of technical strategies for preservation in a networked environment. Starting a separate program to look only at the development of technical strategies for the preservation of e-prints would be of limited value.

### **Organisational Environment**

Short term project funding has supported UK experiments in both subject based and institutional repositories for e-prints. A more stable organisational infrastructure would facilitate long-term preservation. Ultimately, this means that long-term preservation requires long-term funding.

### **Starting Preservation**

In order to manage e-prints in the longer term, e-print repositories need to start addressing preservation issues now. Repositories need to begin collecting the administrative and preservation metadata that will underpin collections management. Leaving this issue too long will only increase the cost of preserving e-prints in the future.

Perhaps unusually for the rapidly evolving scholarly digital world, there is an opportunity to address the preservation of UK e-print collections before the issue becomes urgent. At the present time UK e-print repositories have yet to encounter significant preservation problems, and they hold only a very small proportion of academic research output. However, although the future is uncertain, e-print repositories are likely to become home to more and more significant material that is difficult to obtain elsewhere, or simply not held elsewhere. E-Prints can represent the corporate memory of research communities – hypothesis, experiment, critique and synthesis. It is difficult to see how this material can be viewed as anything but worthy of long-term preservation. Efforts to preserve this material should begin now.

## 3 Recommendations

### 3.1 E-Print File Formats

#### 1. Recognise the Preservation Risks of File Formats

E-Print repositories should be encouraged to assess risks associated with each file format in their collections and consider how this will affect the possibility of the repository providing long-term preservation of, and access to, the intellectual content held in each format. E-Print repositories should reserve the right to convert e-prints deposited in unsuitable formats to others that can be successfully retained for longer periods.

#### 2. Adopt Open, Standards-Based File Formats

Proprietary file formats present the greater risk to the preservation of e-prints over the long-term, but conversion to open standards based formats (such as XML) offers a safe option for minimising the level of risk. Consequently, e-print repositories should seek to adopt open standards-based file formats, and to encourage their authors to deposit e-prints in file formats that are based on open standards, by providing them with information on the advantages of such file formats.

#### 3. Investigate the Use of XML formats to describe data and metadata

E-Print repositories should also be encouraged to research the possibilities that using XML offers for creating bundles of files and their associated metadata. An XML 'wrapper' around each bit stream in the e-print collection could contain all the necessary metadata for preservation and resource discovery, and could also include information about the file format of the bit stream and potentially how to use and convert it. Text archives, for example, are using the Text Encoding Initiative (TEI) headers and increasingly also XML, as the best practice for describing and storing their collections (Standards – Electronic Text Center, n.d.).

#### 4. Plan for Migrating Rare and Obsolete File Formats

Repositories should invest time and effort into describing file formats in their collections and planning for the migration of rare or obsolete file formats. The planned Digital Curation Centre, funded by JISC and the e-Science Core Programme, will have a key role in supporting this work.

#### 5. Maintain File Format Information

E-print repositories should maintain a list of all file formats that are held in their collections. This will serve as the basic information needed to plan for the migration of particular file formats. The planned Digital Curation Centre, funded by JISC and the e-Science Core Programme, will have a key role in supporting this work.

#### 6. Include File Format Identification Functionality in E-Print Repository Software

E-Print repository software should be expanded, or provided with plug-in modules, that will automatically identify file formats that are deposited into a repository and e-print repositories should investigate the use of automatic file format conversion tools to reduce the variety of formats that will require long-term preservation. The OAI-PHM could be used as a basis for sharing technical metadata about file formats needed for preservation with specialist preservation services providing technology watch and file format registry services. The OAI-PHM development should be

informed by research and development into file formats preservation issues. (Leeds, 2003; Public Record Office, n.d.).

## 3.2 Preservation Metadata

### 7. Define E-Print Preservation Metadata Schemas

E-Print repositories should seek to agree a common set of standards for the technical preservation metadata that should accompany each e-print through its life. These standards should be developed in consultation with wider digital preservation communities, and may need to vary to cater for different categories of file format such as binary word processor files and text based mark-up documents.

### 8. E-Print Resource Discovery Metadata Standards

E-print repositories should develop explicit policies on their description principles and produce metadata schemas that are based on internationally accepted description standards. Interviews conducted as part of this study suggested that “the SHERPA project could identify an agreed metadata template for e-prints, which the community could use”.

An outline scheme for e-print preservation metadata is presented in section 8.3.

### 9. Collect Preservation and Administrative Metadata

E-Print repositories should start creating and managing administrative and technical preservation metadata.

E-Print repository software developers should be encouraged to develop tools for automatically creating technical preservation metadata and assisting the repository managers with creating and managing the administrative and preservation metadata. The common e-print file formats should all be automatically recognisable; fixity metadata could be made explicitly part of the collections management functions; support for administrative metadata elements could be linked with the preservation planning functionality.

## 3.3 E-Print Preservation within the JISC IE

### 10. Encourage Preservation Planning in Existing E-Print Repositories

E-Print repositories should be encouraged to incorporate preservation planning functions into their operations. However, *preservation requirements should not add to the real or perceived barriers that discourage authors from depositing their work in e-print repositories*. E-Print repositories that lack the infrastructure to undertake preservation planning and relative activities should be encouraged to develop collaborative arrangements with preservation and data services

### 11. Funding for E-Print Repositories

Existing or planned e-print repositories established through project funding do not necessarily have a secure future. Institutions and national funding bodies should clarify their plans for future contribution to e-print repositories.

### 12. E-Print Repositories should provide Clear Collection and Retention Statements

E-Print repositories should make available to authors and readers clear statements of their

collection and retention policies. The retention period should be discussed with each submitting author, and the repository should make clear the details of their retention commitment.

Specifically, the repository should make clear how long they will hold the e-print and make it available online, and whether they will undertake to migrate the e-print if it becomes inaccessible due to technological obsolescence. As a corollary to this, e-print repositories should clarify arrangements for the transfer or disposal of e-prints in the event of the repository's closure.

### 13. Develop a Model Licence for E-Prints

JISC should commission the development of a model licence for the deposit of e-prints into e-print repositories.

### 14. Advice and Outreach

JISC should provide advice and outreach to repository managers to make them more aware of preservation issues and current best practice that could be applied to their repository.

Specific actions include:

- Summarise key findings of this report in a briefing document for repository managers
- Establish single point of contact for e-print repository managers to coordinate relevant advice from all JISC advisory services
- Run a risk assessment and preservation planning workshop for repository managers

### 15. E-Print User Needs Analysis

JISC should consider research into e-prints that may be held in settings other than formal e-print repositories.

This analysis should:

- Establish an accurate baseline of current e-print usage, and provide well supported projections for future usage
- Determine the wishes of individual research communities regarding minimum retention periods for e-prints
- Establish whether or not e-print readers want long-term access to the e-prints
- Establish whether or not e-print authors want their e-prints to be held in the long-term
- Establish in what situations information professionals believe e-prints should be preserved

### 16. Pilot of a National E-Print Preservation Service

JISC should consider funding a longer-term project to develop a fully costed e-print repository infrastructure that is based on the OAIS Reference Model. It is recommended that this is a practical study that includes implementation at one or more e-print repositories and their partners as appropriate to the chosen organisation model.

The infrastructure pilot study should seek to:

- Identify the actual costs of implementing different preservation options across the life-cycle of an e-print
- Establish standards, best practice, processes and procedures for the management, preservation and presentation of e-prints, and to articulate these in an e-prints Digital Repository Handbook (much of this could be compiled from outputs from FAIR projects)
- Investigate requirements for software automation to perform collections management, data and metadata transfer, and preservation actions
- Expand existing e-print repository software and provide with plug-in modules, to assist in a range of preservation tasks (tools that can automatically identify file formats, tools to convert file formats, and tools to collect preservation metadata would be useful)

- Trial a licence agreement for e-print preservation (building on the RoMEO project)
- Implementation of the repository infrastructure at one or more e-print repositories either at a single institution or in collaboration with one or more JISC-funded services as appropriate
- Trial a preservation service for e-prints provided in informal settings

It is envisaged that the Handbook, together with the infrastructure and associated tools would have wider uses beyond this project and could be employed by other e-print repository managers or their partners to manage and preserve their content.

Storage requirements for a pilot are unlikely to be significant. Based on an estimated size of 0.5 – 1.0 MB per e-print, a pilot storing 5,000 e-prints (approximately the number of e-prints in the UK academic domain) would only require 5 GB of storage per copy. Staffing costs will be far more significant. The pilot will need to provide staffing for:

- Evaluation or development of automation tools
- Systems administration
- Repository system development
- Coordination between partners

## **4 Acknowledgements**

The study team would like to thank everybody who assisted us, and in particular the members of the JISC FAIR (Flexible Access to Institutional Resources) programme e-prints and e-thesis cluster group, especially Christopher Gutteridge, University of Southampton.

# 5 Introduction

## 5.1 Background to JISC Strategy

UK Higher and Further Education institutes have invested considerable effort and resources into intellectual assets that are held in digital form. To secure the long-term future of these digital resources, significant effort must now be urgently put into ensuring that they are preserved and continue to be accessible in the future.

Since 1995, JISC has played a significant role in advancing the digital preservation agenda in the U.K:

- Funding a series of seven digital preservation research studies as part of the eLib programme
- Jointly (with the Arts and Humanities Research Board) funding the Arts and Humanities Data Service (AHDS)
- Funding the Cedars digital preservation project
- Jointly (with the US National Science Foundation) funding the CAMiEON digital preservation project
- Establishing the JISC Digital Preservation Focus in June 2000 as a means of:
  - Developing a long-term retention strategy for digital materials of relevance to HE/FE institutions in the UK
  - Providing a UK focus for the development of practices, policies and strategies for the preservation of digital materials
  - Generating support and collaborative funding from and promoting inter-working with appropriate agencies worldwide
- In partnership with other organisations and sectors, establishing a Digital Preservation Coalition aimed at developing the UK digital preservation agenda in an international context

JISC recognises that the increasing scale and complexity of digital resources now requires a shift in emphasis from relatively modest funding for research into digital preservation towards the establishment and on-going support of shared services and tools. Digital preservation represents a complex set of challenges, which are exceptionally difficult for institutions to address individually. National action in this field is therefore appropriate to the community and UK wide remit and mission of the JISC.

JISC's continuing commitment to developing the UK digital preservation agenda is set out in the JISC *Continuing Access and Digital Preservation Strategy 2002-5* (Beagrie, 2002). JISC foresees responsibility for digital preservation activities spread between national services, individual institutions and, potentially, institutional consortia. The planned Digital Curation Centre (DCC) will act as a conduit for sharing expertise and developing best practice. The DCC will itself not hold digital resources, but will provide a set of central services, standards and tools for digital repositories.

An important initial stage of the JISC Continuing Access and Digital Preservation Strategy 2002-5 is to complete preservation risk and retention criteria assessments for, and to inform and prioritise the development of, future services and calls in digital preservation. An initial target of the strategy has been to complete digital preservation risk and retention criteria assessments during 2002 and 2003. This study is one of the series of assessments initiated under the strategy.

## 5.2 Preserving E-Prints

E-prints and institutional repositories are a new and high profile area, both for the JISC and for institutions in the UK and elsewhere. The initial focus of activity has been on the process of establishing repositories, depositing articles and promoting discovery and access, together with an emphasis on encouraging the cultural change necessary for successful development of e-print repositories. This focus is reflected in the JISC Focus on Access to Institutional Resources (FAIR)

programme. However, if the e-print content of these repositories is to continue to be made available into the future, the concept of preservation needs to be bought into the equation.

The *Requirements and Feasibility Study on Preservation of E-Prints* has sought to do just this, providing recommendations for further research and the development of services and tools to support the long-term preservation of UK e-print content, in the context of the JISC Information Environment (IE) and the JISC Continuing Access and Digital Preservation Strategy 2002-5.

The study was conducted from January to May 2003 by a team from the Arts and Humanities Data Service (AHDS), Estonian Business Archives and the University of Nottingham, as lead site in the SHERPA (Securing a Hybrid Environment for Research Preservation and Access) project. This report presents the findings of the study, arranged into three main areas:

- Properties of E-Prints
- Technical Characteristics of E-Prints
- Cost Models and Organisational Models for Archival E-Print Repositories

The study focused on the requirements for the *long-term* preservation of e-prints, which is defined, for the purposes of this report, as: the period of time during which the hardware, software, and standards used to create and access digital objects, such as e-prints, become obsolete and can no longer be obtained.

Digital objects store meaningful information encoded as a stream of binary digits (bits). In addition to preserving the bit stream (bit preservation), and ensuring that it is not destroyed or corrupted, digital preservation involves ensuring that the bit stream can be correctly decoded and converted into meaningful information again (functional preservation). This report considers both aspects of digital preservation.

Digital preservation involves active intervention across the life-cycle of a digital object. The long-term survival of an e-print will be affected by the priorities and actions of all those who have an interest in it. There are three main stakeholder groups with an interest in e-prints: authors, readers and repository managers. The study team sought to contact, and received feedback from, a wide range of e-print stakeholders, focusing on those managing or planning e-print repositories within the UK Higher Education sector. The recommendations made in this report will primarily affect repository managers, but in some cases are also highly relevant to authors. Even when recommendations are targeted at those who fund, plan and manage repositories, the importance of ensuring appropriate involvement from authors and readers of e-prints should not be forgotten. Ultimately, the value of e-prints is in their value to authors and readers. Repository managers must meet the needs of long-term preservation in a way that does not conflict with the requirements of the authors and readers of e-prints.

Comments on this report may be directed to:

Hamish James, Collections Manager  
Arts and Humanities Data Service  
75 - 79 York Road (8th Floor)  
King's College London  
LONDON SE1 7AW

[hamish.james@ahds.ac.uk](mailto:hamish.james@ahds.ac.uk)

## 6 Properties of E-Prints

### 6.1 Defining the Term 'E-Print'

The question 'what is an e-print' is an important one. In the course of this study we have encountered a range of views, each slightly different from the others, but there is a widely agreed core to the definition of an e-print:

'E-prints' are electronic copies of academic research papers. They may take the form of 'pre-prints' (papers before they have been refereed) or 'post-prints' (after they have been refereed). They may be journal articles, conference papers, book chapters or any other form of research output.

Pinfield, Gardner & MacColl (2002)

An eprint is an 'electronic publication', usually an electronic copy of a research article. Eprints are usually either 'pre-prints' ... or 'post-prints' .... The significant element is 'peer-review'.

Fraser (2003)

Eprints are the digital texts of peer-reviewed research articles, before and after refereeing. ... (as well as any significant drafts in between, and any post publication updates).

Eprints.org (2002)

An e-print is a digital duplicate of an academic research paper that is made available online as a way of improving readers' access to the paper.<sup>1</sup> An e-print may be a copy of a born digital document, or may be a digitised copy of a hardcopy document. In addition to the main text of the paper, an e-print may contain other elements such as equations and static images, including reproduced photographs, maps and graphs. What an e-print does not include is the raw data (such as sensor outputs, survey responses and interview transcripts) on which the research was based.

The concept of an e-print encompasses both draft and final versions of a research paper. A *preprint* is a draft version of a research paper, before it has been approved by a formal quality assessment process, such as peer review. A *postprint* is the final version of a research paper after it has been approved. In this report the, admittedly awkward, phrase 'formal quality assessment process' is used to avoid implying any specific method of assessing the quality of a research paper. Thus, an e-print may or may not have been subject to a formal process of quality assessment and, where this process is linked to publication, may or may not have been published. E-Prints are generally regarded as duplicates of research papers that are, or will be, published elsewhere.

E-Prints are not distinguished by their technical characteristics. None of the file formats, metadata schemas and software used to manage and view e-prints are limited to only working with e-prints. This is amply demonstrated by the case of *Psycoloquy*, an e-journal running using the University of Southampton's EPrints software (<http://psycprints.ecs.soton.ac.uk/>).

For the purposes of this study, a key point is that hardware and software considerations – central to digital preservation – are not significant to the definition of an e-print. Put another way, e-prints do not represent a separate class of digital preservation problem. Instead, they share preservation issues with other types of digital material. The fine detail of how an e-print is defined will not alter the generic problems of technological obsolescence caused by the rapid development of computing hardware and software.

---

<sup>1</sup> The term "e-print" (with the hyphen) was coined in 1992 by Greg Lawler (Suber, 2003b)

## 6.2 E-Print Repositories

Essentially, an e-print repository is a collection of e-prints made available online. An e-print repository is realised as a collection of files – the e-prints – that are managed through a content management system, typically one designed specifically with characteristics of e-prints in mind – that is a mainly textual document incorporating static images, which is displayed and read, rather than accessed and interacted with. The core functionality of these software systems includes submission (including description), discovery (browsing and searching), delivery (display and download), and metadata interoperability, developed through the Open Archives Initiative (OAI) Protocol for Metadata Harvesting (OAI-PMH). An **OAI compliant** e-print repository makes use of the OAI-PMH to enable it to share metadata about its collections with other OAI aware e-print repositories (see Section 7: Metadata below for more).

Most e-print software is simple content management software with an emphasis on the deposit (of data and resource discovery metadata), discovery and delivery of digital documents that are analogous to hardcopy research papers (C. Gutteridge, personal communication, March 21, 2003). E-Print repository software automates some routine tasks and is useful in larger repositories, but an e-print repository can also be run without the aid of specialised software. Some small repositories are managed as static HTML pages with hard coded links to each e-print.

## 6.3 Subject Based and Institutional Collection

Paul Ginsparg started the first official e-print repository, now called arXiv, at the Los Alamos National Laboratory in the US in 1991. The arXiv repository (<http://xxx.arXiv.cornell.edu>) began as a electronic preprint server for physics, but now also contains many postprints. Despite the subsequent proliferation of e-print repositories, there is little doubt that arXiv remains the most successful,<sup>2</sup> and it now contains over 230,000 e-prints (arXiv.org Monthly Submissions, 2003). A number of reasons for the success of arXiv have been suggested focusing on specific aspects of the physics research community that arXiv supports. In particular, the rapid pace of research in physics has been offered as a reason for the enthusiastic adoption of arXiv as an electronic preprint service. The CERN Document Service (CDS) contains another successful e-print service supporting physics, with preprints being collected from 1993 (<http://cds.cern.ch/>).

Citation analysis suggests that arXiv has reduced the period of time from dissemination of a research paper to citation by others to a few months (Jackson, 2002), and it has been argued that “in some subjects, where rapid transmission of knowledge is critical, electronic dissemination of preprints is an absolute necessity, with subsequent traditional publication becoming almost a formality” (Luce, 2001; Langer 2000). Outside of physics and related areas of study, however, efforts to establish e-print repositories in other scientific disciplines, particularly the biomedical subjects, although widespread, have had less impact (Lawal, 2002; Till, 2001).

Services such as RePEC (Research Papers in Economics, <http://repec.org/>) suggest that e-prints have made a noticeable impact in economics. Perhaps this is due to the existing culture of writing working papers that exists in economics, similar to the preprints writing culture of physics. Elsewhere in the social sciences, and generally throughout the arts and humanities, subject based e-print repositories do not appear to have achieved the same impact.

Many early e-print repositories, such as arXiv, collected e-prints according to subject areas. While this has worked well in some disciplines, it has failed to attract scholars in others. A more recent development, partially in response to this, is the creation of institutional repositories.

In cases where the disciplinary practice is ready, institutional repositories can feed disciplinary repositories directly. In cases where the disciplinary culture is more conservative, where scholarly societies or key journals choose to hold back change,

---

<sup>2</sup> For citation as one measure of this, see Brown (2001). See Carr, Hitchcock Hall & Harnad (2000) for a usage analysis of CoRR (Computing Research Repository)

institutional repositories can help individual faculty take the lead in initiating shifts in disciplinary practice.  
Lynch (2003)

**Institutional repositories** are broadly conceived as digital repositories that will hold the total (digital) research output of an institution; this output includes e-prints, e-theses, conference proceedings, datasets, learning and teaching resources, audio and video recordings, and other types of digital material (see Crow, 2003 and The Fedora Project, n.d. for information on popular institutional repository management software). E-Prints, however, remain a core part of the planned content of institutional repositories, and this, along with the similarities in the functionality of software used by subject based e-print repositories and institutional based repositories, can generate a certain amount of confusion. The TARDIS (Targeting Academic Research for Deposit and Disclosure) project at the University of Southampton, for example, aims to develop a “sustainable multidisciplinary institutional archive of e-Prints”, but this institutional repository will consider “all types of research output in a variety of formats” (TARDIS, 2003), including datasets, audio and video clips and other types of material (P. Simpson, personal communication, March 20, 2003). Similar views were expressed by representatives from other UK universities in response to the question “Can an e-print file contain material other than text (e.g., images, audio, datasets)?”

- “I would have thought yes, using our working definition an e-print can contain material other than text.”
- “Yes, and it is highly likely that, in the scientific and medical environments, it will do”
- “Yes, I think it has to be widely defined or we shall find all kinds of things excluded. There may be some exclusions, but it cannot be defined too narrowly”

These responses highlight the lack of firm technical boundaries between e-prints and other types of digital material. To avoid making this confusion worse, the term *institutional e-print repository* has been used throughout this report to describe an e-print repository managed by an institution with the purpose of collecting e-prints written by members of that institution, while the term *institutional repository* is used in the broader sense described above.

If institutional repositories accept a wide range of material they will encounter a wide range of digital preservation issues, many of which will be more challenging than the task of preserving a traditional text based e-print.

## 6.4 Self-Archiving, the Open Access Movement and Publication

Put simply, **self-archiving** is the practice of scholars depositing their own work into e-print repositories. The term is somewhat misleading because the use of the word ‘archiving’ suggests that the work is being deposited into a secure environment suitable for long-term retention. Few, if any, e-print repositories meet the emerging understanding of the standards necessary to be regarded as trusted repositories, suitable for the long-term preservation of digital material (OAIS 2002, p. 3-1; RLG 2002). Self-archiving is therefore better characterised as the backing up of files onto a publicly accessible server.

Indeed, self-archiving is primarily promoted as a means of making research output more widely available. Self-archiving into e-print repositories is presented as a means of eliminating the barriers that restrict access to traditional subscription based scholarly journals<sup>3</sup> by making research output available globally to anyone, providing they have an Internet connection.

Self-archiving is advocated by a variety of individuals and groups who loosely form what is known as the Open Access movement or the Free Online Scholarship movement (Suber, 2003). The objectives of the open access movement are to remove access and impact barriers to research literature (Harnad, 2001) by radically altering the way in which research communities share their work and results.

Self-archiving in e-print repositories effectively decouple the task of formally assessing the quality of a research paper from the task of distributing it to reader. These two tasks are currently tied together

---

<sup>3</sup> Subscription based electronic journals can be included in this category.

through the traditional scholarly journal publication process. Harnad and many others in the open access movement envision a world where OAI compliant e-print repositories provide unlimited access to the 'give away' research literature (Budapest Open Archive Initiative [BOAI], 2002), providing an alternative, less expensive, and better method of distributing research than traditional journal publication (Harnad, 1994). It is, though, important to note that not all e-prints are made freely available, and commercial preprint services already exist (see, for example the Social Science Research Network, <http://www.ssrn.com/>).

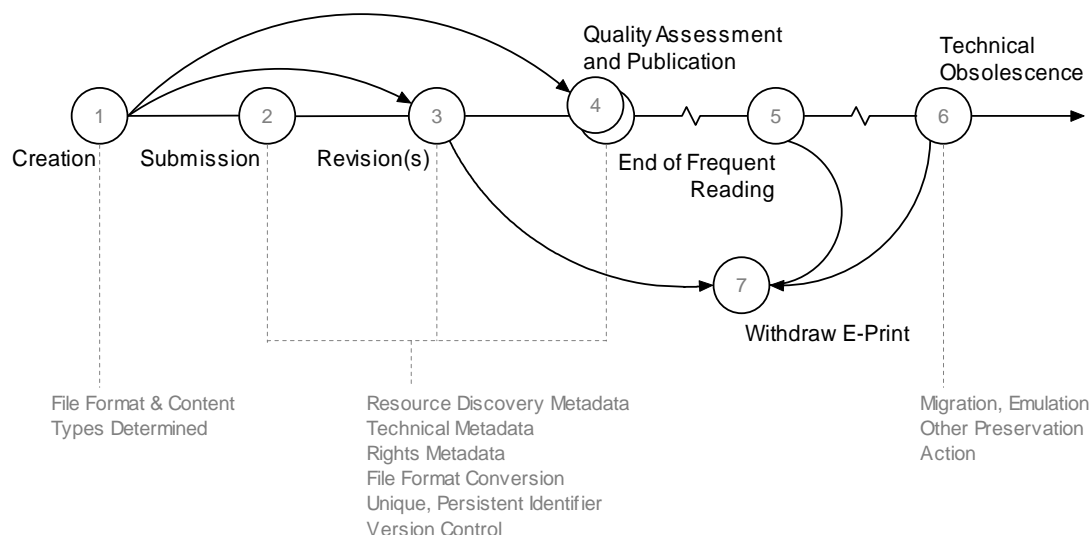
The relationship between publication and e-prints is complex. E-Prints have evolved from electronic preprints, which in turn have their origins as a faster, cheaper replacement for paper preprint services. Many e-print repositories now contain preprints and postprints that have been published in both traditional paper journals or in online electronic journals. The policies of journals, which remain the focus of scholarly dissemination, can complicate the deposit of material into e-print repositories (see Harnad, 2001) and can affect the decisions of potential depositing authors (Gadd, Oppenheim and Proberts, 2003).

At the same time, journals – the traditional home of postprints – have also been changing. The Open Access movement also endorses the development of open access journals (BOAI, 2002). These are essentially electronic journals that do not charge readers for access. Open access journals differ from e-print repositories in that they provide similar styles of dissemination (both are online and available for free) but they still combine distribution with quality assessment and link these to the traditional idea of publication.

Prominent online journal article repositories, like JSTOR ([www.jstor.org/](http://www.jstor.org/)) and PubMed Central ([www.pubmedcentral.nih.gov/](http://www.pubmedcentral.nih.gov/)), complicate the situation further by providing articles from back issues of contributing journals. In the case of JSTOR there is a charge for access, while in the case of PubMed Central, access is free. The archiving of e-publications is the subject of a separate JISC study (Archiving E-Publications, 2003). The preservation of e-prints and the preservation of e-publications overlap when considering the technical strategies and requirements for preservation, but they are less likely to intersect when considering responsibilities and organisational models. Publishers and libraries have long standing roles in the preservation of publications, and this is starting to be carried over into the digital environment (see, for example, Elsevier Science: News Items, n.d.).

## 6.5 Lifecycle of an E-Print

Figure 6.1: Lifecycle of an E-Print



There are seven key events that may occur in the full lifecycle of an e-print (Figure 6.1):

1. Creation
2. Submission
3. Revision(s)

4. Quality Assessment and Publication
5. End of Frequent Reading
6. Technical Obsolescence
7. Withdraw E-Print

At each key event shown in Figure 6.1 a range of actions are, or should, be taken that will affect the future of the e-print. Many of these actions will affect the longer term survival of the e-print and will determine if it is merely a collection of bits, or a readable research paper.

### **Creation**

The *submission* and possible *revision(s)* (of a preprint) are events that are commonly associated with e-prints, but *creation* is seldom explicitly considered. It is highlighted here because decisions made during the creation of an e-print, just as with the creation of any other type of digital object, can have far reaching consequences that affect the long-term preservation risks associated with the object. At creation, the author makes choices about which software package to write the e-print in, which file format to store it in, and what types of content (text, images etc.) it will contain. These decisions will affect the cost and feasibility of preserving an e-print in the long-term.

### **Submission**

The submission of an e-print is the first, and often only, point of substantial contact between the author and the e-print repository. This is a crucial opportunity to provide feedback to the author that may improve the preservation characteristics of later pre- and postprints that may be submitted. It is the repository's best opportunity to collect resource discovery and administrative metadata needed to manage the e-print in the long-term, but more importantly, it is the repository's only real chance to establish a formal agreement with the author to govern the long-term care of the e-print

At submission, the e-print also needs to be established as an item in the repository collection. A unique, persistent identifier should be assigned, and, in anticipation of future revisions, a means of establishing the version should be recorded. This might take the form of an assigned version number or key phrase (e.g. 'first draft') or could be a date stamp.

### **Revision**

More than one version of an e-print may be deposited in an e-print repository. In the end, several preprints and a postprint may be included in the repository. In addition to the requirements of submission, the revision(s) event introduces the possibility that the author may wish to withdraw the earlier version from circulation. Repositories should consider the wishes of their authors and readers when deciding how to handle this action.

### **Quality Assessment and Publication**

Formal quality assessment, which will probably occur elsewhere, leads to the acceptance of a *final* version of the e-print. This final version may be deposited in the e-print repository, in which case all the actions of submission and revision are relevant to this event. Alternatively, the final version may be published elsewhere, and only the metadata in the e-print repository must be updated.

### **End of Frequent Reading**

The life-cycle assumes that the period of frequent reading of the e-print will typically not extend to the time when the e-print begins to become difficult to use due to changes in technology. This arrangement is certainly true for the scientific e-print repositories where the period of frequent reading appears to be measured in months. The period may be much longer in the social sciences or arts and humanities, and for e-prints that represent seminal research, where the number of readers may increase with time.

### **Technological Obsolescence**

Technological obsolescence, is the central problem to overcome when planning for the long-term preservation of any digital object, but not one that has yet caused major problems for e-prints. The

combination of the relatively short period of time since the first e-print repository was set up (12 years) and the relative ease with which text, the main constituent of e-prints, can be preserved across software generations means that, to date, few problems have been encountered. However, when an e-print does become technologically obsolescent, then either some action, such as migration or emulation, must be taken to restore the usability of the e-print, or a decision should be made to withdraw the e-print from circulation.

### **Withdraw E-Print**

Figure 6.1 highlights three points at which a conscious decision may be made, by the author or the repository, to withdraw an e-print from circulation.

- An earlier draft may be replaced by a later draft
- An e-print may be withdrawn when it is no longer read frequently
- An e-print may be withdrawn when it has become inaccessible due to technological obsolescence

When an e-print is withdrawn it may be physically deleted from the repository, but the better practice already followed by many repositories, is to maintain the original e-print, but mark it as superseded and point readers towards the newer version of the same e-print. For an interesting discussion of these issues see Harnad and Goodman (Eprint version removals. 2003).

## **6.6 Why Preserve E-Prints?**

Indeed, the notion that this material is tentative or ephemeral is disappearing. Some believe that such archives will evolve to become the custodians of the primary research literature. Jackson (2002, p.24)

The literature on the open access movement and debate about the future of scholarly communications is considerable (see Bailey, 2003), but preservation is one factor that has been, largely, ignored. A recurring sentiment we found during this study was that the requirements for running successful e-print repositories in the long-term were still being worked out, and e-print repository managers have yet to fully engage with the challenges of digital preservation. UK e-print repository managers are aware of the need to consider preservation, but are unsure of how to proceed, and are looking for guidance.

At the base of this uncertainty is the simple question, should e-prints be preserved at all? E-Prints and the repositories that hold them have evolved as a method for sharing information, not as a way of preserving it. E-Print repositories are seldom presented as a replacement for traditional scholarly publication (either paper based or electronic). They are most often seen as a means of improving access to either work in progress (preprints) or formally approved work (postprints) published elsewhere. This suggests an *a priori* case for not preserving e-prints. On the one hand, preprints do not describe completed research, which should be the target of preservation, while on the other hand, postprints that do describe completed research are published in journals that are already the focus of an established system of preservation, although this system does not yet clearly encompass electronic journals (Jones, 2003, p. 3).

Against this case, several points can be made in favour of preserving e-prints. The first argument arises directly from the role of e-prints as a means of improving the availability of scholarly research. If a postprint research paper is made easier to obtain by depositing it in an e-print repository, then this is likely to remain the case in the future. In this situation, work to preserve the e-print clearly becomes part of the effort needed to ensure that it remains accessible in the future. Of course, like other material, if the readership of the e-print drops too low, there is justification for considering removing it from the repository.

Authors also need to recognise that when they deposit a preprint in a publicly accessible e-print repository they have, in one sense, 'published' it, and made it part of the record of research in their field of study. Even if the author feels the e-print is of only short-term interest, the readers may disagree. Certainly, where others have cited the e-print, there is a reason to retain it, even if it has

been superseded by a later draft of formally published version. A particular case in point is preprints that contain more material than the final published postprint.

From these points, five criteria can be identified that provide a basis for determining if and when an e-print the long-term preservation of an e-print should be considered:

- The e-print provides wider and/or more convenient access than alternatives such as published journal papers
- The e-print is cited by other scholars
- The e-print contains unique information, not recorded elsewhere
- The e-print forms a significant part of the research record (for example, early drafts of important research)
- The e-print is part of a wider collection deemed worthy of preservation

At present there appears to be a reluctance to engage with preservation issues. Exchanges such as that between Harnad and Sargent (EPrints, DSpace or ESpace?, 2003) suggest that while preservation issues are not being discounted, they are being deferred to some future date. Neither authors of e-prints or repository managers have reached firm decisions about the long-term preservation of e-prints.

The criteria above are offered as the basis for discussing the archival appraisal of e-print collections. To make practical use of these criteria, repository managers will need to develop objective measures that take account of local conditions and the interests of relevant stakeholders. Key considerations will include authors' wishes regarding retention period, which may be ascertained at deposit through a formal e-print deposit agreement, and reader's wishes regarding continued availability of e-prints (especially preprints not available elsewhere), which are highly likely to vary between research communities, particularly if scientific subjects are compared to the humanities. The role of e-print repositories within wider institutional information management policies will also be important, as the preservation of e-prints, like any other digital material, will require a steady long-term commitment of resources.

## 6.7 The Future of the E-Print

The authors encountered clear views that defined an e-print as a research paper, but we also encountered a wider viewpoint that defines an e-print as a scholarly research output. This difference of opinion can be roughly characterised as a difference between those involved primarily in the open access debate, and those planning institutional repositories. Those more closely associated with the open access movement tended to place the emphasis on an e-print as a digital equivalent to a traditional hardcopy research paper. Those associated with the development of institutional repositories were more willing to include ancillary materials, such as scientific animations and datasets, as part of the definition of an e-print. E-Print repositories vary in the limits they place on the type of material they will accept; while the main e-print file is expected to be a textual document, with static images embedded or supplied as separate files, supplementary files in a wide range of formats may also be allowed (see for example, The Chemistry Preprint Server, 2002, Information for Authors).

The convention of regarding e-prints as primarily textual documents is bound up with the restrictions imposed by printing. In a digital environment, the restrictions of paper no longer exist, and it seems unlikely that the content and layout of e-prints will continue to adhere to these artificial limits in the future. Writing about arXiv Luce expresses this view well:

Increasingly, merely preserving the article itself cannot capture the value of an electronic article. Rather the value is in the associated contextual links, associated graphics, multi-media and connecting databases that have become intrinsic parts of modern scientific literature. Given this fact, in the very near term, the print versions of journals will not be the true archives. Luce (2002)

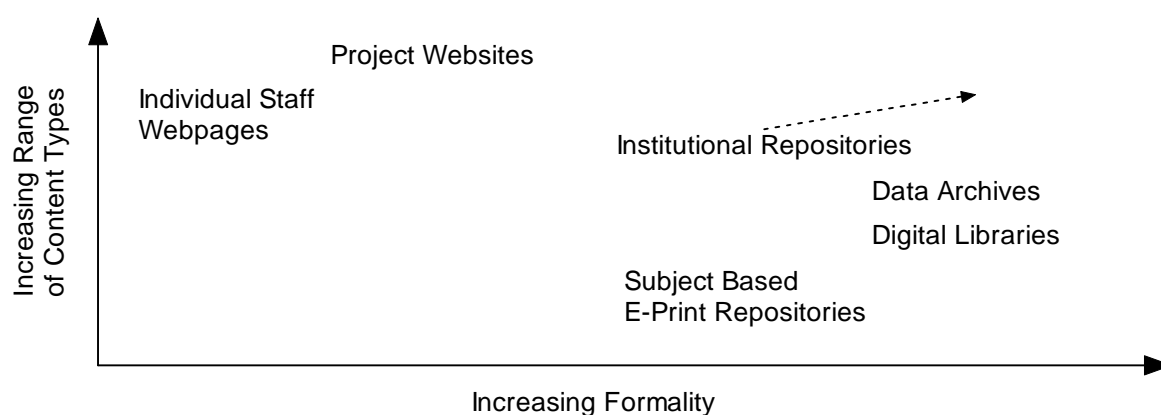
In small ways, divergence from fundamental characteristics of paper printing is already widespread. HTML files, for example, are not constrained by considerations of page size.

Two factors are likely to encourage authors to create e-prints that cannot be fully represented as paper documents. Firstly, e-print repository software can manage any file format. E-print repository managers may choose to limit the acceptable list of file formats for submission, but the design of the software does not preclude the submission of file formats for databases, audio, video or other types of content. Secondly, many of these types of material can already be embedded into file formats that are accepted by e-print repositories. Microsoft Word, for example, can have embedded spreadsheets, audio clips, video clips and many other types of material. Thus there is already a mechanism available to authors for, perhaps unintentionally or surreptitiously, depositing a wide variety of material into e-print repositories. These types of content present far greater preservation challenges than text, and it will be in the long-term interests of the repository to ensure that they are aware of any such material in their holdings, and manage its deposit through an explicit submission process. Repositories could make use of application export functions and file format conversion tools to exclude unwanted content types by converting deposited files into simpler file formats.

The digital environment is a very flexible one, and it encourages the blurring of boundaries that appear solid in a paper-based world. The increasing complexity of e-prints is one likely outcome of this blurring, the integration of e-prints into a range of broader repository environments, as discussed earlier, is another. The development of institutional repositories using e-print repository software is one example of this, but there are many other possibilities. Many e-print repositories already show a tendency to accumulate additional functionality and roles. CERN holds e-prints within the CDS which also holds meeting minutes, administrative documents and a large number of bibliographic records. The RePEc economics service holds a considerable number of bibliographic records and provides contact details for individual researchers and organisations. Others, such as The History and Theory of Psychology Eprint Archive (<http://httpprints.yorku.ca/>) supports a “commentary/response threads” concept for discussion about a particular e-print.

At the present time, formal quality assessment and publication processes are not closely connected to the e-print repository infrastructure. Whether quality assessment and publication do become more closely linked to e-print repositories is an issue beyond the scope of this report, but it is certainly a clear possibility. The E-LIS e-print repository (<http://eprints.rclis.org/faq.html>), for example, already notes, that submitted documents, “can be either approved by the Editorial Board, rejected outright or returned to the author for modifications”.

## 6.8 E-Prints in the UK Academic Domain



**Figure 6.2: E-Print Management Environments**

E-Prints are found in large formally managed e-print repositories, smaller more informally managed repositories, in scattered collections stored in the Web sites of individual projects or academics, and also within the collections of data archives and digital libraries. In short, e-prints are an integral part of the wider academic digital landscape, and the e-print collections held in large e-print repositories, such as Cogprints (<http://cogprints.ecs.soton.ac.uk/>) are only part of the total body of e-prints. Jackson (2002) talks of hundreds of preprint servers in the fields of mathematics alone, and suggests that less

than half of the existing e-prints are held in the most visible repository, CoRR (Computing Research Repository, <http://xxx.lanl.gov/archive/cs/intro.html>).

Figure 6.2 provides an impressionistic view of how these settings vary in terms the formality with which they are operated and the range of content they may contain.

There is a high risk that the e-print collections held in informal settings will not survive in the long-term. Their continued availability is tied to factors such as the time individual academics have to devote to their up-keep, the availability of computing resources within institutions and the vagaries of project funding.

While it is beyond the scope of this study to assess the numbers and significance of e-prints not held in formal repositories, this is a task that should be performed. It would be extremely unwise to assume that only the largest and most visible e-print repositories hold material that is valuable and worthy of long-term preservation. To take one example, the Disability Archive UK repository (<http://www.leeds.ac.uk/disability-studies/archiveuk/index.html>) is highly valued by its user group as a key source of difficult to obtain published and unpublished material, but its future relies on the efforts of a single scholar (C. Barnes, personal communication, April 16, 2003). Outreach and awareness-raising activities may be needed to ensure that informally managed e-print collections are incorporated into appropriate repository infrastructures where they can both be made available to a wider community and be properly preserved. Because many of these collections of e-prints are not managed using e-print repository software, but are instead presented as static Web pages, it may also be possible to treat them as a web archiving issue rather than an e-print issue. Web archiving has been addressed in a separate study funded by JISC (Day, 2003).

In table 6.1 an attempt has been made to quantify the number of visible e-prints in the UK academic domain. The table is based on a search for Web sites within the “.ac.uk” domain that identify themselves as e-print or preprint servers and hold e-prints locally. A total of over 5,000 e-prints, or apparent e-prints (some may not contain the full text) were found in these repositories. Of these, over 3,000 are located in two e-print repositories based at Southampton University, Cogprints and the ECS EPrints Service. Southampton is also home of the widely used EPrints software, and is a centre of e-print activity in the UK.

**Table 6.1: UK Academic E-Print Repositories<sup>a,b</sup>**

E-Print Repository	E-Prints
Armagh Observatory Preprints/Reprints Series <a href="http://www.arm.ac.uk/home.html">http://www.arm.ac.uk/home.html</a>	238
University of Bath : Mathematics Group : Preprints from the Mathematics Group <a href="http://www.maths.bath.ac.uk/MATHEMATICS/preprints.html">http://www.maths.bath.ac.uk/MATHEMATICS/preprints.html</a>	186
Bristol Centre for Applied Nonlinear Mathematics <a href="http://www.enm.bris.ac.uk/anm/publications.html">http://www.enm.bris.ac.uk/anm/publications.html</a>	134
University of Cambridge : Isaac Newton Institute for Mathematical Sciences Preprints series. <a href="http://www.newton.cam.ac.uk/preprints.html">http://www.newton.cam.ac.uk/preprints.html</a>	41
University of Cardiff : School of Mathematics : Hoyle-Wickremasinghe Reprint Series on the Internet <a href="http://www.cf.ac.uk/math/wickramasinghe/contents.html">http://www.cf.ac.uk/math/wickramasinghe/contents.html</a>	11
University of Durham : Geometry and Arithmetic Preprints <a href="http://fourier.dur.ac.uk:8000/pure/preprint.html#viewnote">http://fourier.dur.ac.uk:8000/pure/preprint.html#viewnote</a>	82
University of Edinburgh : Theoretical and Applied Linguistics eprints archive <a href="http://archive.ling.ed.ac.uk/">http://archive.ling.ed.ac.uk/</a>	89
University of Glasgow : Glasgow ePrints Service <a href="http://eprints.lib.gla.ac.uk/">http://eprints.lib.gla.ac.uk/</a>	60
Lancaster University : Department of Mathematics and Statistics Spatial and Computational Statistics Network : Network Preprints <a href="http://www.maths.lancs.ac.uk/dept/stats/essn/preprints.html">http://www.maths.lancs.ac.uk/dept/stats/essn/preprints.html</a>	81
University of Leicester : White Dwarf Group Preprint Server <a href="http://www.star.le.ac.uk/wd/preprint.html">http://www.star.le.ac.uk/wd/preprint.html</a>	16
University of Leicester : X-ray Astronomy Group and the Astronomy Group <a href="http://ledas-www.star.le.ac.uk/Preprint/">http://ledas-www.star.le.ac.uk/Preprint/</a>	172
Department of Mathematical Sciences : Loughborough University : Preprint Archive	194

E-Print Repository	E-Prints
<a href="http://www.lboro.ac.uk/departments/ma/preprints/index.html">http://www.lboro.ac.uk/departments/ma/preprints/index.html</a> University of Manchester Institute of Science and Technology (UMIST) : Department of Physics.	114
<a href="http://www.umist.ac.uk/departments/physics/research/preprint.htm">http://www.umist.ac.uk/departments/physics/research/preprint.htm</a> MCMC Preprints	469 <sup>c</sup>
<a href="http://www.statslab.cam.ac.uk/~mcmc/pages/list.html">http://www.statslab.cam.ac.uk/~mcmc/pages/list.html</a> University of Nottingham : Nottingham eprints.	45
<a href="http://www-db.library.nottingham.ac.uk/ep1/view-ROOT.html">http://www-db.library.nottingham.ac.uk/ep1/view-ROOT.html</a> University of Oxford : Mathematical Institute RAND-APX Thematic Network : Preprint Series.	15
<a href="http://www.maths.ox.ac.uk/rand-apx/preprint.html">http://www.maths.ox.ac.uk/rand-apx/preprint.html</a> University of Southampton : Cogprints : Cognitive Sciences EPrint Archive	2936
<a href="http://cogprints.soton.ac.uk/">http://cogprints.soton.ac.uk/</a> University of Southampton : ECS EPrints Service	895
<a href="http://eprints.ecs.soton.ac.uk/">http://eprints.ecs.soton.ac.uk/</a> St. Andrews Astronomy Group Preprint Server	57 <sup>c</sup>
<a href="http://star-www.st-and.ac.uk/astronomy/preprints.html">http://star-www.st-and.ac.uk/astronomy/preprints.html</a> University of Strathclyde : Strathprints	0
<a href="http://eprints.cdlr.strath.ac.uk/">http://eprints.cdlr.strath.ac.uk/</a> University of Ulster : Formations	56
<a href="http://formations2.ulst.ac.uk/view-ROOT.html">http://formations2.ulst.ac.uk/view-ROOT.html</a> University of Wales, Bangor : School of Informatics : Maths Preprints	66
<a href="http://www.informatics.bangor.ac.uk/public/mathematics/research/preprints/preprint.html">http://www.informatics.bangor.ac.uk/public/mathematics/research/preprints/preprint.html</a>	

Source:

Based on an Internet search using the Google ([www.google.com](http://www.google.com)) search engine for any of the terms "eprint e-print preprint postprint post-print" within the domain "ac.uk" on 30/04/2003, plus additional information.

Notes:

- Table excludes the trial e-print service at the University of Bath
- Sites that do not hold e-prints locally (same parent URL) are not included
- Total includes some e-prints held remotely

While no great claim is made for the accuracy of the figures in table 6.1, it does indicate the current size of the UK academic e-print collection is very small (the number of e-prints written by UK scholars may be considerably higher, as many could be held in overseas e-print repositories).

Allowing for the unknown number of e-prints which are not clearly identified as such, and are held on project Web sites and so forth, we can conclude that preservation of the current UK academic e-print collection is not a significant issue. E-Prints are a relatively new idea and advocates of e-prints are still working towards their acceptance as a viable part of the scholarly communications landscape. A great deal is happening, but with the focus on encouraging the use of e-prints as a means of providing access to research papers, preservation issues have not yet been fully considered, let alone addressed. A typical response to the question "*Once deposited in a repository, should e-prints be stored indefinitely?*" was "we are unsure whether our initial efforts in this area will promote our e-print repository as an archive of all research output for the University and preserve it for future generations."

Consequently, there is a high preservation risk associated with these early institutional repositories. To reduce, or at least clarify, this risk, institutional repositories should move quickly to establish collection and retention policies that define what will be collected and how long it will be kept.

The importance of preserving e-prints will grow as e-prints grow in significance as a means of scholarly communication, but it is not necessarily the case that all, or even most, e-prints will need to be preserved in the long-term. When thinking about the preservation of e-prints it is important not just to consider the e-print itself, but to also consider the repository that holds it and the reasons why authors and readers make use of these e-print repositories. These issues must be put to individual research communities. The answers they give will determine if and when e-prints are preserved, but general advances in digital preservation practice will inform how e-prints will be preserved.

# 7 E-Print File Format Review

## 7.1 File Format Requirements For an E-Print

E-Prints are created and submitted to repositories with ease of use in mind. Being able to use an e-print file means, in majority of cases, being able to display and read the intellectual content of the file, that is, the text. The predominant file formats chosen for e-prints are textual file formats or ones that can be created with word processing software packages. The underlying idea being that the file retrieved by a user from the e-print repository should be viewable without any unusual effort and using only common desktop software. Hence, file formats like Portable Document Format (PDF), Rich Text Format (RTF), ASCII, and others are widely used.

The primary use value of an e-print is often in its first months of being submitted to an e-print repository. An e-print is deposited in order to provide quick and easy access to research findings, but the e-print repository is not normally the place of formal publication. However, interviews carried out as part of this study revealed there is a perception among authors that their e-prints will be retained by repositories for at least 10-15 years, if not forever. The implicit assumption is that an e-print repository will maintain the same level of accessibility to the intellectual content of the e-print throughout the whole life-cycle of the e-print (or the stated retention period of the e-print). In other words, the e-print repository should be able to preserve the file in an *easily usable* condition in the long-term. Even a retention period of 10 to 15 years qualifies as long-term, because within this time period at least two generations of hardware and software (and realistically several more) will pass.

The functionality required of an e-print file format will change depending on the retention period considered. In the short-term file formats need to be easily interoperable, have free or low cost viewing software, and provide support for wide range of platforms. In the medium term, as it becomes somewhat more difficult to access the original file format of the e-print, backward compatibility from newer software and import/export facilities to other software packages become important. In the long-term, file formats need to be suitable to migrate or emulate without loss or damage of the intellectual content of the e-print.

Successful e-print repositories can be expected to hold at least thousands of individual e-prints, so the ability to manipulate e-prints automatically, to extract technical or resource discovery metadata, or to convert to another format for example, is important.

It is likely that authors, repository managers, and readers will have slightly different expectations of e-print file formats. While the repository managers would, or at least should, be interested in easy handling of files for both managing the collection and preserving it over the medium to long-term, the authors and users are much more likely to be concerned only with the immediate usability and easy handling of files for both input to and output from the repository.

## 7.2 Current File Formats in E-Print Repositories

E-Print repositories can and do contain a variety of file formats. The concept of self-archiving, plus large and diverse communities of both depositors and users, makes it difficult for e-print repositories to impose strict restrictions on the use of file formats. To an extent, the range of file formats in e-print repositories is self-regulated by the most popular software packages used by specific academic communities: e.g., sciences e-prints make wide use of TeX and LaTeX whereas humanities practically do not use these formats. The repositories try to be open to the *de facto* standard formats in their target user communities and are increasingly restrictive towards less known or used file formats (cf. Crow, 2002b, p. 37).

E-Print repository software allows a repository manager to limit the range of file formats accepted for submission, and many repositories make use of this feature. Widely used or standard file formats are usually chosen as the acceptable formats with the aim to ensure the widest possible usability of deposited files: users of e-print repositories would be reluctant to download files that require non-

standard or expensive software, or if using the files would require extra skills or training. From the point of view of long-term preservation of e-prints, limiting the number of file formats and using formats based on open standards, is good practice and reduces the risk of loss of access to the content of files over time. The costs and risks associated with digital preservation tend to grow when a digital collection includes a larger number of diverse file formats (Granger, Russell & Weinberger, 2000). However, it cannot be confidently said that the requirements of long-term retention are currently dominant, or in some cases at all involved, in making the choices of acceptable file formats for e-print repositories. Only a few comments can be found in e-print literature referring to the issue of the explicit expectation that the content managed by the e-print repository system will have to survive the system itself and should be possible to migrate as new technologies evolve (Crow, 2002b, p. 35). No file format has so far been specifically chosen as the preservation file format for e-prints.

Typically, e-print repositories are accepting submissions in following file formats:

- Portable Document Format (PDF)
- Rich Text Format (RTF)
- PostScript
- TeX
- LaTeX
- ASCII
- HTML
- XML
- MS Word
- MS PowerPoint

It is common practice to accept widely used formats such as ASCII, PostScript, Rich Text Format, and PDF into e-print repositories. Additionally, the repository content policy or the administrator will determine whether the repository will accept other generic formats (e.g., HTML), proprietary word processing formats (e.g., MS Word), and more discipline-specific text editors (e.g., TeX or LaTeX), images and streaming media. A Survey of e-print repositories conducted as part of this consultancy revealed that e-print repositories may also accept multiple-file digital objects (e.g., HTML file with embedded image files (JPEG), LaTeX file with its Math module for formulas, etc.).

Some repositories accept a wider array of file formats and also some specialised formats, provided they have translation programs available to convert files from submission formats to supported dissemination formats. For example, open source utility programs exist to convert LaTeX to PostScript or PDF. Repositories are interested in simplifying deposit procedures to encourage participation among academic depositors and may want to accommodate a wide range of file formats popular with various academic departments. At the same time, the repository needs to balance the desire to accommodate content contributors with the complications that migrating some of those file formats may present in the future as new standards evolve.

Most institutional e-print repositories exercise a review process of submitted material before making it public through their services. Perhaps most importantly, this review verifies and, if necessary, improves the depositor-supplied metadata, but the review also includes checking the submitted e-print file formats. The submitted files must be readable and suitable for on-line distribution; if they are not, then the administrators of the repository may choose to convert the files into formats that conform to best practices. It is also possible to return the e-print with comment on file formats to the author. Most e-print repositories refuse the publisher produced PDF or other format versions produced for official publication. However, no technical requirement sets these files apart from other files in the same format (e.g., a PDF) that are readily accepted.

While many of the early institutional repository implementations have deferred decisions about long-term digital preservation, some of the more recent versions of software are capturing automatically the file format information from submitted e-prints. One such system maintains a registry of known file formats, and automatically identifies the format of an e-print when possible. For unknown formats, the system queries the submitter requesting additional information. System administrators maintain the registry of known format types and the preservation service level available for each format type. However, where the format of the bit stream is unknown, the repository can make no claims regarding preservation and future use of the file (Bass *et al.*, 2002, p. 5).

Information on individual file formats and issues related to their preservation is being collected and published by digital preservation projects. JISC has commissioned a “Survey and assessment of sources of information on file formats and software documentation” (Leeds, 2003) and is preparing the establishment of a Digital Curation Centre as a pilot development of long-term preservation planning tools, and services for recording and monitoring file formats (JISC, 2003). The National Archives (formerly the Public Record Office & Historical Manuscripts Commission) has been developing a database system (PRONOM) that stores and provides information about file formats and the application software needed to open them. The preservation of e-prints should make use of such information and, where possible, provide feedback and input into these developments.

### 7.3 Risk Assessment of Common E-Print File Formats

File formats continue to evolve, becoming more complex as revised software versions add new features and functionality. It is not uncommon for software enhancements to render files generated by earlier versions unreadable. The threat to aging digital information has surpassed the danger of unstable media or obsolete hardware, the most pressing problems confronting managers of digital repositories are data format and software obsolescence (Lawrence *et al.*, 2000, p. 1).

File (or data) formats define the rules used by application software to convert bits (the fundamental unit of digital data) into meaningful information that can be viewed and manipulated by a user. Most application software developers produce file format documentation for the formats they design and develop. Not all of them make this documentation available and even if they do, it is not always accurate (see Lawrence *et al.*, 2000, pp. 13-15 for examples of attempts to retrieve the Lotus 1-2-3 and TIFF file formats from their developers.).

Based on the availability and stability of the format specification, file formats can be classified as proprietary, open or standard formats. **Proprietary file formats** are not public and are developed and maintained by software producers. Larger software producers may sometimes publish their format specifications (PAS – Publicly Available Specification) or several firms may join together in a consortium to define interface standards so that they can develop mutually compatible products. These are called **open or public file formats**. Some file formats are developed to become international standards (**standard file formats**) which are then public and fixed or stable until the next release of the standard. It is not unusual that software companies produce their own modified, proprietary, versions of standard file formats – these will be based on standards, but will have extensions that are proprietary and generally not public (e.g., Microsoft’s version of XML). Many proprietary formats are, nevertheless, widely used and provide extensive compatibility with application software – these formats are often classified as *de facto* standards (cf. DLM 1997, pp. 50-52).

Successful and cheap long-term preservation of a digital file depends on the openness, level of standardisation and compatibility with other software products of the file format. Without a format specification the vital rendering tools that enable the use of digital files over longer time cannot be developed. Reverse engineering of software or the digital objects themselves can provide some answers, although legal constraints may well prevent this kind of action. Even where reverse engineering is possible, without any file format documentation, the process is likely to be too laborious and expensive (Leeds, 2003, p. 4).

The preservation risks associated with file formats are mostly related to loss of data and cost. Both migration and emulation — the two best digital preservation strategies currently in use — rely on file format specification being known and accurate. If it is not, the preservation strategies risk introducing distortion, loss of quality or data, or not being able to render the file usable at all. The risk management of file formats for preservation has to account for all these considerations. An assessment of more popular e-print file formats follows.

Even the oldest e-print repositories have been in existence for just over ten years and the majority of repositories only a few years. E-Print repositories have, thus far, had almost no experience of content migration for preservation and can be characterised as relatively unaware of dangers and risks inherent in conversion or emulation based digital preservation strategies. Statements like “if you build 5 percent into your budget for conversions, then you are ‘covered’ for all eternity” have been cited in literature, but the awareness and understanding that digital resources “cannot be left ignored for a

century if they need to be used” is on the rise (cf. Jackson, 2002, pp. 30-31). Only a couple of references could be found to actual content migration that has been undertaken by e-print or similar repositories and the results should offer warnings to other repositories: The American Mathematical Society converted its archive of published journal articles from one TeX format to another. A computer program successfully converted 90 per cent of the articles, but 10 per cent had to be converted by hand, requiring substantial effort by highly trained personnel. For an e-print repository containing, say, half a million articles, that problematic 10 per cent would mushroom into a huge and costly task (Jackson 2002, p. 31). Smaller e-print repositories would benefit from specialist preservation services where such complex migration tasks could be solved more easily and efficiently (see Section 10: Organisational Models).



The **Portable Document File Format (PDF)** is designed to replicate a document exactly as it appeared to the creator of the document. PDF is a platform-independent document format developed by Adobe as a follow-up to its PostScript language. Although the PDF specification is open and freely published (Adobe Systems Incorporated, 2003a, File Format Specifications) the format is maintained by the Adobe Systems Inc. who considers it the open *de facto* standard for electronic document distribution world-wide (Adobe Systems Incorporated, 2003b) PDF documents can be protected against editing or revising which makes it a safe format in the sense of protected content, but it is often not recommended as a safe format for long-term preservation. Although PDF has seen widespread take up across the preservation and archival communities, there are problems associated with some of the more complex PDF tags as well as issues surrounding availability and use of fonts. Adobe has recently started developing conversion tools from PDF to XML, which would generally be a safer format for long-term preservation. PDF-Archive is an initiative to specify a subset of PDF tags for archival purposes as an ISO standard.

Because the basic file format standard is public, Adobe does not have a monopoly on PDF tools, and third parties have developed tools that also work with PDF. However, because the standard is complex and changes from time to time, much of the support for the format, as well as the tools that are considered to set the standard for PDF use, comes from Adobe (Ockerbloom, 2001). Nevertheless, wide, almost universal, use and published specification render the Portable Document File Format a medium or low-risk preservation file format for e-prints.

Table 7.1 lists the positive and negative preservation considerations of the PDF format.

**Table 7.1: A Checklist of Preservation Consideration for the Portable Document File Format**

Positive	Negative
File format specification is public	Proprietary format developed by one company only
Wide acceptance of the format	Frequent releases of new format versions (Florida Centre for Library Automation [FCLA], 2003, p. 2)
Format is platform independent and conversion tools are beginning to appear	Limited range of tools for creating the file format
PDF to XML conversion possible	Problems with conversion between earlier versions of the format

John Warnock and Chuck Geschke of Adobe Systems Inc developed the **PostScript** language in 1985 as a written description of a printed page interpreted by a computer chip placed inside a laser printer. It is a simple interpretative programming language with powerful graphics capabilities. The language is still maintained by Adobe Systems Inc., but the documentation is freely available. The primary application of PostScript is to describe the appearance of text, graphical shapes, and sampled images on printed or displayed pages according to the Adobe imaging model. A program in this language can communicate a description of a document from a composition system to a printing system or control the appearance of text and graphics on a display. The description is high-level and device-independent (Adobe Systems Incorporated, 1999). An Encapsulated PostScript (EPS) file is one in which the PostScript code for either a whole page or a single image is saved as an ASCII text file (Adobe Systems Incorporated, 1992).

Preservation concerns with the PostScript format are similar to those of the PDF format that has superseded and incorporated the PostScript language. Archivists have been relatively unconcerned

with this format, but very little evidence has been published regarding conversion and problems with converting PostScript to other formats.

**Rich Text Format (RTF)** was developed by Microsoft as an interchange format designed to retain the full formatting of a document. RTF is a proprietary product developed and maintained by Microsoft, but the file format specification has been made public (Microsoft Corporation, 1999). Its widespread use and compatibility with nearly all word processing software qualifies RTF as a *de facto* standard format that is suitable for medium- to long-term preservation of textual material.

**Table 7.2: A Checklist of Preservation Consideration for the Rich Text File Format**

Positive	Negative
File format specification is public	Proprietary format developed by one company only
Wide acceptance of the format	Microsoft updates the RTF specification each time they release a new level of software
Format is platform independent	Exact version is not easily discerned from file
Easily compatible with many software products	

**TeX** is a computer language designed for use in typesetting and in particular, for typesetting mathematics and other technical material. It was first developed by Donald E. Knuth at Stanford University in 1978 to deal with revisions to his book series “The Art of Computer Programming”. The idea proved popular and Knuth produced a second version (in 1982), which is the basis of TeX that is used today (version 3). The language is device and platform independent which makes the files highly portable, but since it was developed as a typesetting language it is directed more towards printed output rather than on-screen viewing, although the latter is possible.

TeX also provides powerful facilities for compiling structured sets of macros. Most users generate documents that are coded using TeX macro sets, of which LaTeX (<http://www.latex-project.org/>) is by far the most popular. LaTeX is a TeX macro package that provides a document processing system that allows mark-up to describe the structure of a document, so that the user need not think about presentation. The current TeX source code is still maintained by its original developer (but no further versions are promised) and the basic source code is freely available.<sup>4</sup>

**Table 7.3: A Checklist of Preservation Consideration for the TeX File Format**

Positive	Negative
File format specification is public	The original file format is proprietary, different versions developed by multitude of individuals, interest groups and companies
Format is platform independent (the TeX DVI) and easily portable	Requires specific software to create and render files in this format

The **plain text** format could be called the “simplest” file format from the preservation point of view and although some difficulties may arise if the character encoding system: 7 bit ASCII (ISO/IEC, 1991), 8 bit ASCII (ISO, 1998-2001) or Unicode (ISO, 2000a) is unknown to the repository, these can, as a rule, be overcome relatively easily. However, since a plain text file cannot preserve the formatting of text nor layout, it is used less frequently than other text file formats, for example mark-up formats that are still based on plain text.

**Table 7.4: A Checklist of Preservation Consideration for the Plain Text File Formats**

Positive	Negative
File format specification is a public standard	File format cannot retain text formatting
Format is platform independent and easily portable	Limited text processing tools available
File format is compatible with almost any software package	

For a long period, **SGML (Standard Generalised Mark-up Language)** (ISO, 1986) was considered the only “safe” format for long-term storage of complex textual data files. It remains a safe format for

<sup>4</sup> For example, from <ftp://cam.ctan.org/tex-archive/systems/> for various operating systems and hardware platforms

preservation, as long as the structure of the file is described in a Document Type Definition (DTD) and retained alongside the text file itself. The same applies for the World Wide Web Consortium's (2000) **Extensible Mark-up Language (XML)** and other mark-up based formats. The **Hypertext Mark-up Language (HTML)**, (W3C, 1999) although plain text-based and therefore technically not difficult to preserve, is still not considered to be very stable for archival purposes (although a stabilised version has been defined by ISO/IEC 15445 – ISO, 2000), nor very suitable for long documents and, therefore, a low- to medium risk file format for long-term retention of textual data.

Table 7.5 lists the positive and negative preservation considerations of mark-up file formats in general. The risks associated with mark-up file formats increase when the file includes links to objects outside the file that are considered to be part of the same document or data resource. While the marked-up text files usually preserve the structure of the document, they may not always be able to retain the original presentation of the document and/or the more complex functionalities offered by word-processing packages.

**Table 7.5: A Checklist of Preservation Consideration for Mark-up File Formats**

Positive	Negative
File formats specifications are public and standardised	File format standards are developing and new features are added at a rapid rate
File formats are platform independent and easily portable	For most cases requires a DTD to be preserved alongside the text file
File structure can be described with simple tools and can be human-readable	Can only be used to preserve text, any other objects or data types need separate treatment for preservation
File formats use plain text that is easy to preserve for long-term	Files may include links to external objects that cannot be preserved as one whole
File formats are compatible with many software packages	

**Microsoft Word, Microsoft PowerPoint** and other similar office software products are highly proprietary and their file formats are being updated at short intervals. Although some backward compatibility is offered by the new versions of software, it is unclear how long compatibility will be maintained with superseded versions of the software. Microsoft provides information for developers through its MSDN web site (<http://msdn.microsoft.com/library/default.asp>) which contains some documentation aimed primarily at migrating data to Microsoft formats, but it does not offer specifications for the MS Office file formats. Some Microsoft file format specifications (MS Office 97) have found their way into public domain and can be obtained but are considered to be incomplete.<sup>5</sup>

Although widely used, MS Word and other MS Office family file formats are not suitable for long-term preservation. These proprietary formats will, nevertheless, be popular for e-prints because of the widespread use of MS Office software. E-Print repositories that receive and hold e-prints in these formats should consider conversion to other more open formats such as RTF and XML. Microsoft Office software is generally compatible with other file formats and can output files in other formats without significant risks to the content of files. Using the OpenOffice package (<http://www.openoffice.org/>) is another, open file formats based, alternative.

**Table 7.6: A Checklist of Preservation Consideration for Microsoft Office File Formats**

Positive	Negative
File formats are widely used	File format is proprietary and not public
File format is compatible with some software packages	File format is being developed and changes at short intervals
Tools to convert the file formats into XML are being developed	

E-Prints may also be accompanied by static images delivered in file formats such as JPEG and TIFF.

<sup>5</sup> Two sources are Wotsit's Format (<http://www.wotsit.org/default.asp>) and ffe (<http://pipin.tmd.ns.ac.yu/extra/fileformat/>).

**Joint Photographic Experts Group (JPEG)** format is a lossy compression image file format that earlier contained a proprietary (IBM) component which rendered it unsuitable for authentic long-term preservation of image data. The JPEG 2000 initiative was set up in 1998 to make improvements to the JPEG format by using better compression algorithms. The JPEG 2000 format became an international standard (ISO, 2000b). Only the first part of the standard of the intended eight has been published so far and the work on it continues.<sup>6</sup> The new format “desires” that both lossless and lossy compression be available at time of saving a file. The feature of lossless compression, combined with standardised format specification, makes JPEG 2000 a well-suited format for long-term preservation.

The **Tagged Image File Format (TIFF)** is widely used and is a *de facto* standard for storing image data. Aldus and Microsoft developed the format to provide a basis for importing scanned images into desktop publishing packages. Aldus originally owned the file format specification before its merger with Adobe Systems. Consequently, Adobe Systems now holds the copyright for the TIFF specification, but have made it public (Adobe Systems Incorporated, 1992).<sup>7</sup> TIFF is a lossless image format but different versions of it offer a limited level of compression options (LZW/CCITT compression, JPEG compression was introduced in version 6). TIFF files can also contain contextual metadata including information about the author and copyright.

Image file formats are likely to create fewer problems when migrated for preservation and both file formats chosen for e-prints are suitable for medium to long-term preservation.

## 7.4 Reducing Preservation Risks Associated with Current E-Print File Formats

The driving force behind selecting acceptable formats for e-prints is the need to encourage the widest possible dissemination and readership of the e-prints. Despite this, existing e-print repositories appear to be quite cautious in defining which formats they will accept, although some respondents to this study indicated that exceptions might be made for an especially valuable text.

Positive aspects of current e-print repository practices for long-term preservation of e-prints are:

- The list of recommended submission file formats is often limited
- The common file formats are widely used and some are open standards-based
- The repositories reserve the right to review and if necessary, change the deposited file format;
- The repositories reserve the right to prescribe a limited list of acceptable submission file formats
- The future e-print file formats can be guessed from the popularity of software among the depositor and user communities

Negative aspects of current e-print repository practices for long-term preservation of e-prints are:

- There is an incentive to avoid limiting the list of acceptable submission file formats, as this may discourage submissions
- Selection of acceptable submission formats is geared towards satisfying the requirement for easy accessibility in the short-term, rather than retention and accessibility in the long-term
- Not all file formats in use are based on open standards
- Long-term preservation needs rank relatively low in defining the submission policies of repositories

The nature of the content of e-prints and their authenticity conditions do not set high requirements for conversion or preservation processing – adequate content replication in a format that can be used with current software is the ultimate aim of the preservation of e-prints. Current e-print file formats are relatively risk-free for this purpose, but with more dynamic file formats and e-prints likely to occur in the future, stricter preservation-driven policies should be established for defining acceptable file formats for e-prints.

---

<sup>6</sup> For more information see the official JPEF homepage (<http://www.jpeg.org>)

<sup>7</sup> The specifications for versions 4.0 and 5.0 can be downloaded from the “Unofficial TIFF Home Page” (<http://home.earthlink.net/~ritter/tiff/>)

An ideal file format for an e-print would thus have to be:

- Easy to create, using standard software
- Able to render all content that the user wants to communicate (including, e.g., formulas, graphs, drawings, etc.)
- Of relatively small size in relation to available, affordable, storage and network transfer speeds
- Easy to use with the widest choice of standard software
- Open to preservation over long-term (i.e., preserving the functionality of accessing the contents of the file)

The best-suited file format for an e-print should be based on an open standard and be widely used by different scholarly communities. There is no single file format to match all these requirements, but XML holds a promise to develop into such a language for defining flexible formats for documents in the future. Current practices could be improved, without placing additional burdens on depositing authors, if e-print repositories recognised three categories of file formats:

- Submission formats
- Preservation formats
- Dissemination formats

Submission formats would be selected according to the preferences of authors. Dissemination formats would be selected according to what is currently most convenient for readers. To manage and simplify the task of converting from submission format to dissemination format, the repository would define a, small, set of standards-based preservation formats. E-Print repositories could convert their more problematic submissions in proprietary file formats to a smaller set of standards-based preservation formats. This will reduce the risks and costs associated with e-print file formats over the long-term. Copies in a wider range of dissemination formats could be created from either of the other two formats. Given the relatively limited and consistent range of e-print file formats currently in use, most of these conversions could be performed automatically on request by the repository software. A number of e-print repositories and software packages already support automatic file format conversions, especially into PDF (for example, The Berkley Electronic Press, 2002, p. 4). Because the repository has a well-understood set of preservation formats, it does not need to maintain either the submission formats or dissemination formats in the long-term.

## 7.5 Recommendations for E-Print File Formats

### **Recognise the Preservation Risks of File Formats**

E-Print repositories should be encouraged to assess risks associated with each file format in their collections and consider how this will affect the possibility of the repository providing long-term preservation of, and access to, the intellectual content held in each format. E-Print repositories should reserve the right to convert e-prints deposited in unsuitable formats to others that can be successfully retained for longer periods.

### **Adopt Open, Standards-Based File Formats**

Proprietary file formats present the greater risk to the preservation of e-prints over the long-term, but conversion to open standards based formats (such as XML) offers a safe option for minimising the level of risk. Consequently, e-print repositories should seek to adopt open standards-based file formats, and to encourage their authors to deposit e-prints in file formats that are based on open standards, by providing them with information on the advantages of such file formats.

### **Investigate the Use of XML formats to describe data and metadata**

E-Print repositories should also be encouraged to research the possibilities that using XML offers for creating bundles of files and their associated metadata. An XML 'wrapper' around each bit stream in the e-print collection could contain all the necessary metadata for preservation and resource discovery, and could also include information about the file format of the bit stream and potentially how to use and convert it. Text archives, for example, are using the Text Encoding

Initiative (TEI) headers and increasingly also XML, as the best practice for describing and storing their collections (Electronic Text Center, n.d.).

### **Plan for Migrating Rare and Obsolete File Formats**

Repositories should invest time and effort into describing file formats in their collections and planning for the migration of rare or obsolete file formats. The planned Digital Curation Centre, funded by JISC and the e-Science Core Programme, will have a key role in supporting this work.

### **Maintain File Format Information**

E-print repositories should maintain a list of all file formats that are held in their collections. This will serve as the basic information needed to plan for the migration of particular file formats. The planned Digital Curation Centre, funded by JISC and the e-Science Core Programme, will have a key role in supporting this work.

### **Include File Format Identification Functionality in E-Print Repository Software**

E-Print repository software should be expanded, or provided with plug-in modules, that will automatically identify file formats that are deposited into a repository and e-print repositories should investigate the use of automatic file format conversion tools to reduce the variety of formats that will require long-term preservation. The OAI-PHM could be used as a basis for sharing technical metadata about file formats needed for preservation with specialist preservation services providing technology watch and file format registry services. The OAI-PHM development should be informed by research and development into file formats preservation issues. (Leeds, 2003; Public Record Office, n.d.).

## 8 E-Print Metadata Review

The type of documentation currently attached to e-prints reflects the general purpose and profile of e-print repositories — their aim is to provide fast and open access to scholarly output. It is often perceived that the immediate usage value of an e-print is for a short-term, hence the metadata that describes it, is predominantly for resource discovery, rather than for long-term preservation. More often than not, the number of e-prints in (institutional) repositories is manageable without the more advanced tools as used by digital libraries and archives.

### 8.1 Documentation as Metadata

Metadata is an essential aspect of preservation and collections management. It is required to support the following functions:<sup>8</sup>

#### Resource discovery

Users must be able to locate and retrieve resources through a searchable index or catalogue. Information must also be available about any terms and conditions that are attached to those resources, and the means by which they may be accessed.

#### Administration

Information is required to manage and administer digital resources and to support the use of a resource, including details of its content, structure, acquisition information and technical dependencies.

#### Preservation

The preservation and management of digital resources requires the maintenance of a wide variety of technical information, as well as administrative information and their processing/preservation histories.

The resource discovery metadata is often understood as the catalogue and index record component of an archive or repository. The administrative metadata or resource management metadata comprise information that is required for management of a digital resource throughout its life-cycle, including its preservation and processing history.

Preservation metadata is intended to support and facilitate the long-term retention of digital information.

Preservation metadata ... is the information necessary to maintain the *viability*, *renderability*, and *understandability* of digital resources over the long-term. Viability requires that the archived digital object's bit stream is intact and readable from the digital media upon which it is stored. Renderability refers to the translation of the bit stream into a form that can be viewed by human users, or processed by computers. Understandability involves providing enough information such that the rendered content can be interpreted and understood by its intended users.  
OCLC, 2002

The lack of preservation metadata in e-print repositories is probably the biggest obstacle to the successful long-term preservation of e-prints. Preservation metadata captures technical, administrative and other information that will help ensure that the content of an e-print remains accessible in the long-term. Its value becomes clear when the formats used to record e-prints and the software used to display them becomes obsolete, a point not yet reached for most e-prints. The collection of preservation metadata should not, therefore, be seen as conflicting with the access orientated goals of

---

<sup>8</sup> Alternatively, metadata for managing digital collections is often categorised into resource discovery, structural, and administrative (including preservation) metadata.

e-print repositories, but rather as a sensible precaution that will help ensure that e-prints remain accessible.

Preservation metadata serves as the blueprint for the chosen preservation strategy and disaster recovery procedures. It provides a record of the file location and properties required to authenticate and migrate files to new formats and media. It also documents decisions and actions performed in processing, converting and migrating digital resources, forming an audit trail of sorts for each e-print. The authentication information of a digital resource in a repository is often described separately as fixity metadata that documents the means by which a digital resource can be authenticated, and safeguarded from undocumented alteration.

## 8.2 Resource Discovery Metadata for E-Prints

The e-print repositories investigated during the course of this study are all creating resource discovery metadata. The reviewed institutional e-print repositories are creating resource discovery metadata that meets the Open Archives Initiative (OAI, <http://www.openarchives.org>) mandatory requirement for simple Dublin Core metadata. Dublin Core provides a generic set of basic metadata for common elements such as date, author and title. This metadata can then be shared between repositories which support the Open Archives Initiative Protocol for Metadata Harvesting (OAI-PMH). The OAI-PMH defines a mechanism for harvesting XML-formatted metadata from individual repositories (OAI, 2003). The protocol provides a mechanism for harvesting content that is encoded in XML only. Co-operation between the OAI and the Dublin Core Metadata Initiative has led to a common XML schema for simple Dublin Core (15 metadata elements) that has evolved into a *de facto* standard for simple cross-discipline metadata (Dublin Core Metadata Initiative [DCMI], 2003a). This standards based approach to metadata disclosure facilitates consistent results when users are searching across holdings in multiple repositories or browsing metadata records gathered from multiple repositories. Services such as ARC - A Cross Archive Search Service (<http://arc.cs.odu.edu>); Cite-Base Search (<http://citebase.eprints.org/cgi-bin/search>) and OAIster (<http://oaiSTER.umd.umich.edu/o/oaiSTER/>) provide access to distributed content through interoperable metadata records.

The OAI-PMH does not, however, mandate the means of association between that metadata and the related content in the repository.

Commonly used e-print repository software packages support simple Dublin Core by default, but managers of repositories are allowed to configure the resource discovery metadata elements during installation of the software. They can decide both what metadata elements to use for describing an e-print and what metadata elements to present to the user during a search, as well as set up subject hierarchies that provide meaningful browsing options to users. After configuration, the repository has to be registered with the OAI and the metadata will be checked for consistency and compliance with the OAI-PMH.

While the e-print repositories may choose to collect more metadata than the 15 elements required by the OAI-PMH, most of them adhere to these basic elements. Guidance has been issued for interpreting and filling the OAI compliant metadata both by e-print repository software creators, as well as by e-print services (DSpace, 2002b; Powell, Day & Cliff, 2003).

E-print repositories rely on authors to fill in the necessary description of the e-print they are submitting, but there are problems collecting even the short and simple metadata needed to create a Dublin Core metadata record. Obtaining accurate metadata from authors has proved a challenge for e-print repositories: authors have not proved particular good at tasks such as entering references correctly or making consistent use of keywords (Boyce, 2000, p. 414). It has been observed that more automated tools could fix many things, but there is no antidote to a lack of common sense (Luce, 2001).

## 8.3 Preservation Metadata for E-Prints

Unfortunately, in contrast to the support for resource discovery metadata, managers of e-print repositories have practically no preservation metadata support provided by the common repository software packages. The software has built-in mechanisms for uniquely identifying files and tools for file management, and may include fixity metadata (e.g., the eprints.org software creates and keeps and

incremental log of checksums with deposits of every user). The metadata generated at the submission of an e-print into the repository, include file name, file format, and date of submission (sometimes using date stamps). Some repositories have established their own internal file naming conventions that carry meaning (e.g., an unique identifier, depositor's code, etc.). The software usually keeps track of the number of files in repository and the total disk space taken up by files, but more advanced collections management tools have not been implemented in any of the software packages studied.

The bulk of e-print repositories have not been in existence for a sufficiently long period of time to encounter significant digital preservation problems. Hence, the virtual absence of preservation metadata has not, as yet, proved a problem for e-print repository managers, but it can be said with confidence that the issue will become a serious concern in the future.

Several digital preservation metadata schemas and standards have emerged over the last years, most notably from the digital library community (National Library of Australia, 1999; CEDARS, n.d.; California Digital Library, 2001; National Library of New Zealand, 2002). The evolution of the Open Archival Information System (OAIS) Reference Model into an international draft standard has spurred development of digital preservation metadata in this context (Lupovici & Masanès, 2000; Online Computer Libraries Center [OCLC], 2002). National archives and other government bodies have been developing standards for recordkeeping metadata (Public Record Office [PRO], 2002; National Archives of Australia, 1999; Public Record Office Victoria, 2000) and information interchange (Office of the e-Envoy, 2002; METS: Metadata Encoding and Transmission Standard, 2003).

The services offered by e-print repositories are easily comparable with those of digital libraries, making the preservation metadata recommended by libraries highly relevant for e-print repositories. At the same time, e-print repositories have functionality that is closer to electronic records and content management systems, which means the metadata standards used in these systems can also be usable when managing e-prints collections. Metadata interoperability standards like METS will become very important if e-print repositories are to transfer and store their collections for long-term preservation to third parties.

The use of simple Dublin Core metadata elements to describe e-prints has been explained in detail elsewhere (Powell et al, 2003) as well as examples of using other subsets of Dublin Core (the Dublin Core Libraries Working Group Application Profile (LAP), for example) (DSpace, 2002). Table 8.1 below (on next page) presents an extension to the simple Dublin Core metadata schema (the first 15 elements in the table) and offers a minimum set of administrative, and preservation metadata elements. Table 8.1 also indicates whether a metadata element can be treated as serving primarily the purpose of resource discovery, administration or preservation, or several of these.

**Table 8.1: A Suggested Minimum Set of Digital Preservation Metadata for E-Prints**

No	Element	Definition	Comment	Resource Discovery	Adminis- tration	Preser- vation
<b>Resource Discovery Metadata (OAI/DC)</b>						
1	Title	A name given to the resource (compare with DCMI, 2003b)	Typically, a Title will be a name by which the resource is formally known	x		
2	Creator	An entity primarily responsible for making the content of the resource	Examples of a Creator include a person, an organisation, or a service	x		
3	Subject	The topic of the content of the resource	Typically, a Subject will be expressed as keywords, key phrases or classification codes that describe a topic of the resource	x		
4	Description	An account of the content of the resource	Description may include but is not limited to: an abstract, table of contents, reference to a graphical representation of content or a free-text account of the content	x		
5	Publisher	An entity responsible for making the resource available	Examples of a Publisher include a person, an organisation, or a service	x		
6	Contributor	An entity responsible for making contributions to the content of the resource	Examples of a Contributor include a person, an organisation, or a service.	x		
7	Date	A date associated with an event in the life-cycle of the resource	Typically, Date will be associated with the creation or availability of the resource	x	x	x
8	Type	The nature or genre of the content of the resource	Type includes terms describing general categories, functions, genres, or aggregation levels for content	x		
9	Format	The physical or digital manifestation of the resource	Typically, Format may include the media-type or dimensions of the resource. Format may be used to determine the software, hardware or other equipment needed to display or operate the resource	x	x	x
10	Identifier	An unambiguous reference to the resource within a given context	Recommended best practice is to identify the resource by means of a string or number conforming to a formal identification system	x	x	x

No	Element	Definition	Comment	Resource Discovery	Adminis- tration	Preser- vation
11	Source	A reference to a resource from which the present resource is derived	The present resource may be derived from the Source resource in whole or in part	x		
12	Language	A language of the intellectual content of the resource	Recommended best practice is to use RFC 3066, which defines two- and three-letter primary language tags with optional sub-tags	x		x
13	Relation	A reference to a related resource	Recommended best practice is to reference the resource by means of a string or number conforming to a formal identification system	x	x	x
14	Coverage	The extent or scope of the content of the resource	Coverage will typically include spatial location (a place name or geographic co-ordinates), temporal period (a period label, date, or date range) or jurisdiction (such as a named administrative entity)	x		
15	Rights	Information about rights held in and over the resource	Typically, a Rights element will contain a rights management statement for the resource, or reference a service providing such information. Rights information often encompasses Intellectual Property Rights (IPR), Copyright, and various Property Rights	x	x	x
<b>Technical Preservation Metadata</b>						
16	File format version	Version of the file format	An optional extension to element 9 (Format) to identify the particular version of a file format (e.g., MS Word 97) (c.f. 9)	x		x
17	Extent	The size of the resource	The file size of the resource		x	x
18	Application software	Software program capable of displaying the resource, or accessing its intellectual content	The software necessary for rendering or retrieving the content of the resource. May include a version number for the application software (e.g., MS Word 2000)		x	x
19	Operating system	Designation of the software platform upon which Application software programs operate	Identifies the operating environment used by the Application software programs of the resource. May include a version number for the operating system		x	x

No	Element	Definition	Comment	Resource Discovery	Adminis- tration	Preser- vation
<b>Administrative Metadata</b>						
20	Fixity method	Type of authenticity error detection technique used	Enables the checking of the integrity of an archived resource		x	
21	Authentication date	Date of most recent archival use of this fixity method	Date of the fixity method used to authenticate the resource		x	
22	Fixity information	Value of the fixity method used	Value of the fixity method used to authenticate the resource		x	
23	Status	The position or state of the resource	Typically the status of the resource would record a stage in its life-cycle, e.g. public, withdrawn, archived, deleted, etc.		x	
24	Disposal	The retention and disposal instructions for the resource	Typically this element would include a condition or date when the resource can be either deleted or the retention decision reviewed		x	
25	Preservation action date	The date and time at which defined preservation action on a resource takes place	Provides validation of preservation actions carried out on resources		x	x
26	Preservation action type	A preservation action carried out on the resource	Typically, this element records one action from a pre-defined list of preservation action types, e.g., media refreshment, migration, moving off-line, checking for readability, etc.		x	x
27	Preservation action description	The specific details of the preservation action	The description should include information about the original status of the resource, the changes made to it, the reasons for the changes, and authorisation for the changes		x	x
28	Next preservation action	The next preservation review, check or action that the resource needs to undergo	This elements acts as a tool for preservation planning and co-ordination		x	x
29	Next preservation action due	The date that the resource is due for preservation action review, or that the next preservation action is due	This elements acts as a tool for preservation planning and co-ordination		x	x

# Recommendations for E-Print Metadata

## **Define E-Print Preservation Metadata Schemas**

E-Print repositories should seek to agree on a common set of standards for the technical preservation metadata that will accompany each e-print through its life. These standards should be developed in consultation with wider digital preservation communities, and may need to vary to cater for different categories of file format such as binary word processor files and text based mark-up documents.

## **Define E-Print Resource Discovery Metadata Schemas**

E-print repositories should develop explicit policies on their description principles and produce metadata schemas that are based on internationally accepted description standards. Interviews conducted as part of this study suggested that “the SHERPA project could identify an agreed metadata template for e-prints, which the community could use”.

## **Collect Preservation and Administrative Metadata**

E-Print repositories should start creating and managing administrative and technical preservation metadata.

E-Print repository software developers should be encourage to develop tools for automatically creating technical preservation metadata and assisting the repository managers with creating and managing the administrative and preservation metadata. The common e-print file formats should all be automatically recognisable; fixity metadata could be made explicitly part of the collections management functions; support for administrative metadata elements could be linked with the preservation planning functionality.

# 9 Cost Models for Preserving E-Prints

## 9.1 Introduction

It is notoriously difficult to predict costs for preserving any type of digital material. Partly this is because there is not a great deal of practical experience on which to base cost estimates, and partly because much depends upon the nature of the material to be preserved. E-prints are not exempt from these problems. Moreover, the situation is further complicated by the fluidity and difficulty in defining the nature of an e-print. This report has already highlighted the significant likelihood that e-print repositories will move towards incorporating both the text of the research paper, and the underlying digital resource on which the text is based. Whilst it is important to mention this issue, which has potentially significant effects on the costs of preserving e-prints, it is beyond the remit of this study to address the possible cost issues associated with these wider definitions of an e-print. This report seeks only to present cost models for e-prints defined as digital duplicate of preprints or postprints.

Major influences on the likely future costs of preserving e-prints are the collection policies, and decisions contained therein, of repositories, and the decisions taken for rights clearance issues. Other factors affecting costs will be the availability of software tools to (semi-) automate some of the functions of an e-print repository and the existence of specialist support services that can provide technology watch and format registry services. The CEDARS working paper *Cost Elements of Digital Preservation* (Granger, Russell & Weinberger, 2000) highlights the 'shrinking timeframe' between creation and the need for preservation actions for digital materials. The report also asserts that "commitment ... will depend on the archiving model in which preservation occurs and how responsibility is allocated, the technical strategy chosen for preservation and the type of access required" (Granger, Russell & Weinberger, 2000, p.1).

In this section an attempt is made to identify cost elements; to model likely repository collecting scenarios; to highlight cost events and decision points that are likely to occur in the lifecycle management of e-prints; and to give some indication of the impact on costs of decisions taken at these key points. This report further recommends that a longer term study be undertaken, utilising an existing or recently established e-print repository, that would be tasked to develop a fully-costed business model, placing actual figures against the cost points (see Section 10.6: Recommendations for the JISC IE). This study should also include an assessment of the cost of deleting or removing e-prints and their associated metadata. The latter is often regarded as a method for saving costs, but it is likely that in some circumstances removing or deleting an item and the associated records may incur costs that exceed those of retaining the item, particularly given the decreasing cost of storage.

## 9.2 Cost Elements

It is possible to start to assess the costs associated with an e-print repository using the elements identified in the CEDARS working paper project report *Cost Elements of Digital Preservation* (Granger, Russell & Weinberger, 2000):

### Selection

Costs here will depend upon the approach taken by an e-print repository, and in particular, whether the repository chooses to encourage a self-archiving approach with a standard licence that is signed and submitted with the e-print as a part of the submission process, or a more pro-active approach that identifies and requests deposit. Costs will be relatively low in the former case, but much higher if a pro-active approach is taken. It seems likely that most repositories will start with a more pro-active approach, largely because the concept of e-print repositories is relatively new and unknown and the focus

for many institutions will be on encouraging authors to submit content. However, it might be safe to assume that over a period of time when the concept is better established the move will be towards self-archiving.

### **Negotiation**

Negotiation includes the costs of obtaining the rights to both provide access and to perform any actions necessary to ensure long-term preservation. Further consideration of the costs of rights negotiation is provided Under Taxonomy of Archives.

### **Technical strategy for preservation and access**

This includes the submission process, and preparation for preservation and access. It involves identifying the significant properties of the e-print and determining the technical requirements for preservation. It may also include the purchase or creation of software to assist and manage these processes. This report indicates that for e-print repositories dealing only with texts and with a relatively restricted range of formats, these costs are likely to be moderate. However, should repositories choose to impose no restriction on formats, then costs may be significantly higher.

### **Validation**

This includes the time taken to check the object received against any documentation and ensuring that the e-print is intact and complete. For e-print repositories this is unlikely to require significant resource.

### **Metadata**

This is likely to include the production of resource discovery, administrative and preservation metadata. It seems unlikely, given current repository practice, that sufficient metadata for technical and administrative purposes will accompany deposited e-prints. Therefore, this is likely to involve fairly high costs for a repository both to create, check and structure the required metadata elements.

### **Storage**

Following the CEDARS definition, storage costs include the long-term management of e-prints and ensuring their accessibility for as long as required. This will include the process of migration, refreshment of media, and the development and implementation of software tools to assist this process.

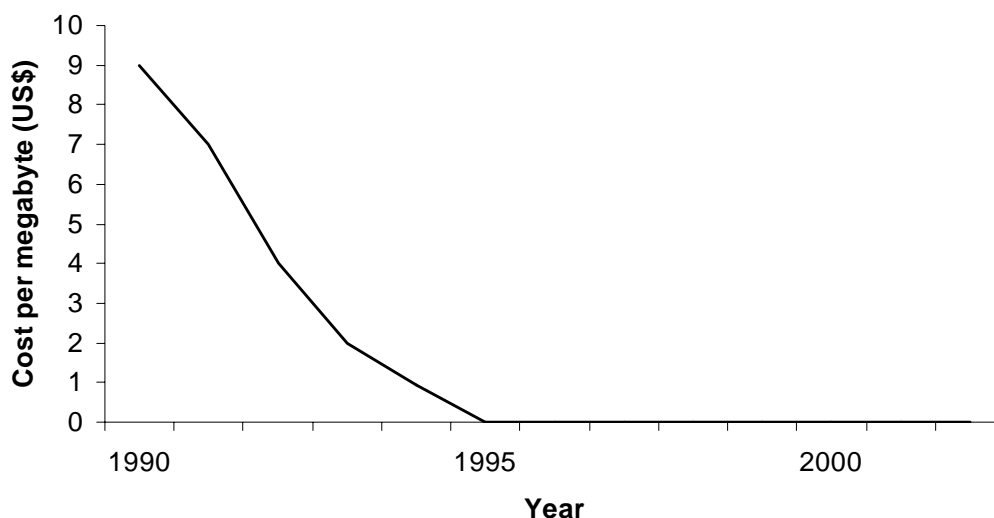
The primary purpose of e-print repositories is access and it is safe to assume that migration policies will be predicated on retaining accessibility. It is also safe to assume that access to e-prints is unlikely to require complicated interfaces, and that migration would therefore be expected to concentrate on maintaining e-prints in easily accessible file formats through a process of periodic upgrades. Appropriate tools should be developed to identify file formats and to automate as far as is possible the process of migration from version to version. Requirement to migrate should be linked to a file format registry and technology watch service (the authors are assuming that the DCC will provide these functions). Wherever possible batch migration of e-prints should be undertaken rather than manual migration of individual e-prints. Migration should only take place when the e-print falls into a retention category defined in the repository collections policy.

Physical storage costs can be calculated and planned for based upon two factors: The expected (likely) submission rate for an e-prints archive and their average size; and the cost of storage hardware and medium. Repositories should estimate the number of submissions expected in any one year, calculate the average size of each e-print, and assess the likely increase in storage requirements over a defined period of time.

Institutions that are establishing e-print repositories might calculate likely submission rates on the basis of RAE returns and current published output from their research-active staff and to use this to estimate the maximum number of articles likely to be submitted to an e-print repository each year. The average size of an e-print is estimated to be between 0.5 and 1.0 megabyte (C. Gutteridge, personal communication, May 2, 2003). By calculating the maximum number of e-print submissions together with the average size it is possible to assess the volume of storage space that will be required over a defined period of time.

It is also possible to estimate the costs of storage equipment. Figure 9.1 illustrates the sharply decreasing costs associated with storing a megabyte of data. The cost of both hardware and associated storage medium (CDs etc) has plummeted over recent years.

**Figure 9.1: Storage Costs per Megabyte (US Dollars)**



Taking the estimate of the space required and the estimated cost of storage and hardware it should be possible for e-print repositories to predict with reasonable certainty the costs of storing content over a defined period of time. These are likely to be static, if not decreasing costs, and relatively insignificant<sup>9</sup> in comparison to the staff costs required to operate an e-print repository.

In summary, the cost elements that are the most significant factors affecting costs in e-print repositories are:

- The cost of negotiating rights
- The potential costs involved in managing proprietary formats should repositories choose to accept whatever is offered to them
- The cost of creating additional metadata, particularly that associated with the technical and administrative needs for long-term management of e-prints

What is also clear is that a pre-requisite to assessing costs with any degree of certainty is for repositories to have developed and implemented clearly defined collection and collection management policies. We strongly recommend that repositories undertake such work when establishing e-print repositories.

<sup>9</sup> For example, the authors estimate that the cost of storing all the content within arXiv.org (some 240,000 prints) is approximately £300. To gain a true figure the cost of necessary servers would need to be added but again, it is estimated that an annual amount of approximately £1500 would cover these costs.

## 9.3 Taxonomy of Archives

The nature of a digital repository will clearly have a significant impact upon digital preservation costs. The Cedars report provides a conceptual taxonomy of archives based upon three categories: content, data types and formats; rights; and control. Content, data types and format is self-explanatory – the more varied and complex these are, the higher the cost to the archive of managing and preserving them. Rights refers to ownership – if an archive owns the rights to the content then costs are significantly lower than if they have to negotiate for permission to preserve and provide access. Control refers to the level of control an archive has over the nature, form and completeness of the content that is submitted. The more control an archive has the less the cost. The authors of the Cedars report summarise these categories into a simple table indicating potential impact upon the costs of running an archive (figure 9.2).

**Figure 9.2: Cedars Cost Model for Archives**

	<b>Increasing, complexity, cost</b> →	
	Simple, low cost archive	Complex, high cost archive
Data Types & Formats	Limited Number	Large Number
Rights	Ownership	Non-Ownership
Control	Full Metadata	

Source: Based on Granger, Russell & Weinberger, 2000, p. 5

Figure 9.3, below, matches the nature of e-print repositories identified in this report against the Cedars model to give an indication of the position of e-print repositories within this matrix and to start to establish

**Figure 9.3: E-Print Repositories Cost Model**

	<b>Increasing, complexity, cost</b> →	
	Simple, low cost archive	Complex, high cost archive
Data Types & Formats	Text only formats	
Rights	Preprint	Postprint
Control	File Only	Metadata

For the first category, e-prints are clearly low-cost archives as they accept mainly relatively simple textual data (although this may change in the not too distant future) in a limited and fairly standard range of file formats. However, the picture becomes more complicated when we look at rights and control. E-print repositories as a rule, do not own the content they hold. However, current practice indicates that repositories regard the act of submission as an indication that they may provide access to the content, and none specifically address the issue of the right to preserve deposited e-prints. In addition, few repositories appear to have policies in place to handle rights issues and the associated negotiations with authors and publishers once an e-print is published.

A recommendation of this report is that rights should be agreed with submitting authors and a model deposit licence developed for use by repositories (see Section 10.3: Non-Functional Requirements and Section 10.6: Recommendations for the JISC IE). Once these arrangements are in place rights negotiation for pre-prints should be a low to medium cost activity. Rights remain a far more complex issue when dealing with post-prints. Negotiation to continue to provide access and to undertake necessary preservation actions could prove difficult and costly. Should these negotiations prove necessary for large numbers of e-prints then it is probable that an e-print repository would fast become financially difficult to maintain. By and large, e-print repositories have ignored this issue to date, but, to the knowledge of the authors, no cases of infringement of copyright have yet come to Court.

The level of control held by e-print repositories is more difficult to ascertain. Theoretically, they should have control and the ability to define in their collections policies what they will and won't accept, their preferred formats and the required metadata. Based upon current or predicted repository behaviour we have modelled likely repository collecting behaviour into four scenarios:

1. Repository accepts only e-prints in preferred formats with complete metadata set.
2. Repository accepts e-prints only in preferred formats but accepts incomplete metadata sets.
3. Repository accepts e-prints in both preferred and non-preferred formats but accepts only complete metadata sets.
4. Repository accepts e-prints in both preferred and non-preferred formats and accepts complete and incomplete metadata sets.

Of these four options the most likely scenarios would appear to be option 2 or option 4. Imposing some form of control on acceptable formats seems more likely due to the limited number of formats in which e-prints are created. However, requiring a full metadata set as specified in this report seems extremely unlikely if authors are to be encouraged to submit e-prints to repositories. In a climate where the concept of e-print repositories is relatively new and unknown the focus for many institutions will be on encouraging authors to submit content, and to minimise the barriers. Thus many will receive content in a variety of formats, some of which are likely to be proprietary and difficult to manage, and many (if not all, should the requirement include preservation metadata) of which will come with limited and incomplete metadata.

## 9.4 E-Print Lifecycle Cost Elements

The life-cycle model for e-prints presented in this report identifies a number of key events. At each stage in the life-cycle, management decisions will be made that affect costs, including whether to accept, retain or delete items, the technical strategies to be employed, and the benefit of any action relative to its cost. Five key events are relevant to costs:

- Submission (including submission of revisions)
- Publication
- Retention assessment
- Technological obsolescence
- Rights negotiation (at submission and on publication)

### Submission & Revision

This is the point where a repository may accept or reject an e-print. Decisions taken here should be guided by the collections policy of the repository. Criteria for inclusion in the repository are likely to be based upon content, format, metadata, and rights. As outlined

in the section on archive taxonomies the latter three are likely to have a significant bearing upon the costs associated with accepting an e-print.

### **Publication**

The normal assumption with a preprint deposited into an e-print repository is that it will, in due course, be formally published elsewhere. When a formally published version of the e-print become available, the repository should decide whether to retain or remove the preprint, depending on the accessibility of the published version, including its long-term accessibility, or rights issues. If retention relies upon negotiating continued rights to access and preserve the e-print with a publisher, then a repository would be advised to undertake a cost-benefit analysis to assess if the benefits of continued retention outweigh the costs.

### **Retention Assessment:**

Collections policies should include decisions on review and retention/removal criteria. Repositories may well wish to remove items from their collection, perhaps based on usage rates, some form of value assessment, or other criteria specific to their repository or institution, for example an institution may withdraw services for a particular discipline and no longer wishes to retain the e-prints relating to that discipline.

### **Technical Obsolescence**

This event requires the repository to make decisions on whether to migrate, remove, or retain the e-print 'as is' (inaccessible), with future migration or emulation to retain accessibility in mind. This latter option goes against the spirit of e-print repositories as a method of disseminating knowledge, and is therefore not recommended.

Table 9.1, below, brings together the three significant cost elements and the five key cost events that have been identified so far and attempts to classify each combination of repository scenario and life-cycle event in terms of low, medium or high costs. Without adequate data it is not possible to put actual figures against these costs and it is a recommendation of this report that a further more detailed study is undertaken with an e-print repository to more accurately assess the costs associated with each stage.

At the point of submission a repository will decide what action to take to produce metadata, to validate files, and will determine the appropriate technical strategy. The concept of up-front and deferred costs is useful in describing where costs might fall in the lifecycle of e-prints. Up-front costs are those costs associated with actions undertaken at the submission stage in the lifecycle process. Deferred costs are those costs that are deferred until a subsequent event in the lifecycle.

Repositories should seek to defer costs wherever there is uncertainty about long-term retention (such as rights issues arising on publication, or changes in institutional policies), and seek to undertake batch processing wherever possible. To that end it is recommended that files are validated and that any essential additional metadata is created at the submission stage, and that costs associated with these tasks are budgeted for as up-front costs. However, repositories may wish to defer the cost of migrating files for preservation purposes until such time as they are certain that all necessary rights have been cleared, and that they wish to retain the e-print beyond the point of technological obsolescence.

**Table 9.1: E-Print Cost Drivers**

E-print Repository Scenario	Submission <sup>a</sup> (including revisions)	Publication		Use Assessment		Technical Obsolescence		
	Accept/reject	Retain	Remove	Retain	Remove <sup>b</sup>	Migrate	Remove	Retain
1 Preferred format with Complete metadata set (1)	Low	Low	Med	Low	B = Low M = Med	Low <sup>c</sup>	Low <sup>d</sup>	Deferred cost - high
2 Preferred format with Incomplete metadata set (2)	Med	Low	Med	Low	B = Low M = Med	Low <sup>c</sup>	Low <sup>d</sup>	Deferred cost - high
3 Non-preferred format Complete metadata set (3)	Low if accepted as is  Med if migrate on submission	Low	Med	Low	M = Med	Low if migrated on submission  Med if migrated at this stage	M = Med	Deferred cost - high
4 Non-preferred format Incomplete metadata set (4)	Med if no migration  High if migration	Low	Med	Low	M = Med	Low if migrated on submission  Med if migrated at this stage	M = Med	Deferred cost - high
Rights negotiation	Low-Med <sup>e</sup>	High <sup>f</sup>						

Notes:

- a) When assessing costs there is an assumption that the DCC will be in existence and providing format registry and other services to assist institutions or preservation/data services undertaking this work
- b) B= batch processed whereby tools exist to identify and remove e-prints in the same format; M= manual removal where costs are significantly higher
- c) Assumes tools are available to assist in the migration process
- d) Assumes tools are available to assist in the removal process
- e) This assumes a standard licence in place and voluntary submission of e-prints by authors
- f) This assumes that negotiation with publishers is required. At this point a repository would be advised to undertake a cost/benefit exercise to ensure that the value of the e-print outweighed the cost of negotiation and retention. If the author has negotiated the right to continue to provide access to the e-print via the repository, then no costs are incurred

# 10 Organisational Models

## 10.1 Responsibilities and Roles

One of the problems that the e-print movement has had to confront is who should be responsible for the e-prints and how. Many supporters of e-prints want to see the traditional publishers removed from the role of caretakers of this scholarly communication genre, substituting either the professional societies or universities. However, this raises the major question of archiving. Luce (2001)

While libraries and publishers undertake established roles in the preservation of printed material, the allocation of responsibilities for the preservation of digital material is still evolving. E-Prints are collected, stored, and delivered within a variety of organisation settings, some of which are better equipped than others to meet the functional and non-functional requirements for the long-term preservation of e-prints.

Within the library and archive community there is a growing awareness that in the digital world responsibility for preserving information will need to be distributed in new ways (Crow, 2002a; Smith, 2003; Luce, 2001; Beebe & Meyers, 1999, Lynch, 2003). The way forward envisioned by many is to disaggregate the tasks undertaken by a digital repository, so that not all repositories need undertake all tasks.

Fundamental to implementing this disaggregated model is the logical separation of the content and service components.... This separation allows for distributed open access content repositories to be maintained independently of value-added services fulfilled discretely by multiple service providers. Crow (2002a)

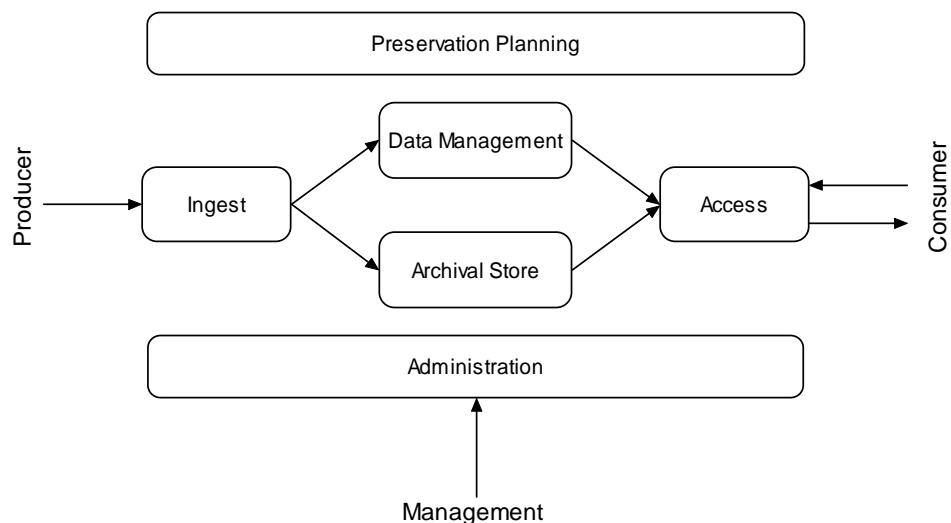
Digital preservation could be seen as one of these 'value-added' services, and could be provided in a number of ways, as suggested in the JISC Continuing Access and Digital Preservation Strategy 2002-5 (Beagrie, 2002, p. A13). An e-print repository could undertake preservation activities itself, could work collaboratively with a group of other repositories, or could rely on an external agency to provide preservation services for its collection.

Before attempting to consider the likelihood of each of these possibilities in more detail, it is useful to have a more detailed understanding of what the functional and non-functional requirements of an *archival* e-print repository – one that is capable of preserving e-prints in the long-term – would be.

## 10.2 Functional Requirements

The functional requirements for the preservation of digital information have been the focus of considerable recent attention. One can turn to the recent, but widely discussed and accepted, standard, the *Reference Model for an Open Archival Information System (OAIS)* (Consultative Committee for Space Data Systems [CCSDS], 2002).

**Figure 10.1: OAIS Functional Entities (Simplified)**



Source: Based on Figure 4-1 in CCSDS, 2002, p. 4-1

The OAIS functional model, shown in Figure 10.1, identifies the main tasks that any type of repository must perform in order to secure the long-term preservation of digital material. The model defines six main functional entities that describe the activity of a digital repository as a flow of digital material, from the arrival of new material in the repository, its storage and management, and through to its delivery to a user (consumer). When thinking about the OAIS model specifically in relation to e-prints, it may be helpful to replace the generic terms *producer* and *consumer* with *author* and *reader* respectively.

### **Ingest**

Ingest includes the physical transfer of files and the legal transfer of rights through the signing of licences or other agreements that establish the OAIS repository's right to maintain the ingested material. During ingest, descriptive information (resource discovery metadata) should be created to describe the material, and the submitted files are checked to ensure that they are consistent with the OAIS repository's data formatting and documentation standards. This may include tasks such as file format conversions or other changes to the technical representation and organisation of the submitted material.

### **Archival Storage**

This functional entity is concerned with the bit storage of the submitted digital material including tasks such as backup, mirroring, security and disaster recovery.

### **Access**

All the services and functions needed for users to find and access the contents of the repository.

### **Data Management**

Data management involves the collection, management and retrieval of both resource discovery, administrative and preservation metadata about the OAIS repository's content.

### **Administration**

The administration functional entity involves the entire range of administrative activities that an archival organisation should undertake. Notable tasks include managing,

monitoring and developing the repository's software systems, negotiating submission agreements with producers (authors), and the establishment of policies and standards for the repository.

### **Preservation Planning**

This functional includes four sub-entities associated with identifying preservation risks and developing plans to address them:

*Monitor Designated Community* – the designated community is an OAIS term that refers to the community of stakeholders who have an interest in the content of the repository. An OAIS repository needs to monitor its designated community's adoption of new technology, and other trends that may affect preservation of the community's digital output.

*Monitor Technology* – The monitor technology function ensures that the OAIS repository is constantly aware of technological changes that may render its current holdings obsolete or difficult to access.

*Develop Preservation Strategies and Standards* – The development of strategies and standards for preservation that are informed by the current and future requirements of the producers and consumers of the OAIS repository.

*Develop Packaging Designs and Migration Plans* – This function accepts standards for file formats, metadata and documentation (generated as part of the administration functional entity) and creates tools or defines techniques that apply these standards to submissions.

The core functionality of most e-print repositories is based on a simple process model of upload – store – download. In comparison to the functional requirements described in the OAIS model, deficiencies are apparent in a number of areas.

E-Print repositories provide rather basic ingest functionality. Most important is the typical lack of any formal agreement made between the submitting author and the repository that will govern the repository's role in maintaining the e-print. Resource discovery metadata is often provided by the e-print author and is not subjected to any quality assurance process. In an environment of interoperable e-print repositories, metadata of inconsistent quality could affect the ease with which e-prints can be located. As noted earlier in this report, administrative and preservation metadata is often not collected at all, and this will affect the ability of e-print repositories to manage their collections.

The lack of preservation metadata will also affect the ability of the e-print repository to conduct preservation planning. On occasion, the belief has been expressed that the problems that this functional entity is meant to address will not occur with e-prints, but there is at least now a growing awareness of the issues (Day, 2001), although very little action in practice.

With regard to archival storage, e-print repositories present no unusual challenges, and a failure to provide adequate archival storage will most likely occur because of a lack of formal financial and organisational support for a repository, an issue taken up below.

## **10.3 Non-Functional Requirements**

The reasons for the deficiencies in the preservation practices of e-print repositories discussed above can, perhaps, be found by looking at the non-functional requirements of an archival e-print repository. Recent work examining the requirements for successful long-term digital archiving, notably the RLG/OCLC report *Trusted Digital Repositories: Attributes and Responsibilities* (RLG, 2002), serves as a useful basis for describing a core set of non-functional requirements that an e-print repository must possess if it is to provide a suitable

archival home for e-prints. The six attributes listed here represent a synthesis of attributes identified in the RLG-OCLC report and the OAIS (OAIS, 2002, pp. 3-1-3-5) model.

**Obtain sufficient control of the deposit material to permit preservation actions such as storage, duplication and migration.**

The ethos of the open access movement is to downplay the role of any intermediaries between the author and reader. Perhaps as a consequence of this, e-print repositories often do not require the depositing author to sign any type of formal agreement with the repository. Interestingly, this is in stark contrast to the educational data archiving sector where such agreements are common, and are considered a vital method of establishing ownership and controlling access, as well as establishing the data archive's obligations and rights. Admittedly, the situation with e-prints is somewhat different. The creators of datasets may often need or wish to impose restrictions on who can make use of the data, whereas the authors' of e-prints want to disseminate their work as widely as possible. But in other ways, non-exclusive licence agreements, such as those found in the educational data archiving sector, can serve as a useful model. These non-exclusive licences establish the rights of the repository to duplicate, transfer and, most crucially, alter the deposited digital material through actions such as migration. Licence agreements also minimise the repository's legal liability by formally establishing that the depositor holds the necessary legal rights to deposit the material. For e-print repositories that hold postprints, this is clearly very important.

Failing to establish a formal deposit agreement with authors may cause a number of problems for the e-print repository, including: legal challenges from publishers; disputes over withdrawal of papers with authors; confusion over permitted reuse of the e-prints; difficulties transferring the e-prints to third parties for preservation. Retrospectively establishing rights will be far more difficult than establishing them when the e-print is deposited.

Several US institutional repositories do have publicly available agreements that can be reviewed:

- Caltech Library System Papers and Publications - Author Permission Agreement (<http://caltechlib.library.caltech.edu/archive/00000006/>)
- DSpace – MIT Libraries - Non-Exclusive Distribution Licence (<http://dspace.org/mit/policies/license.html>)
- University of California eScholarship Repository – CDL-ORU Agreement (<http://repositories.cdlib.org/escholarship/join.html>)

In the UK, the RoMEO (Rights METadata for Open archiving) project, funded as part of the FAIR programme, has investigated rights issues relating to open access, and has provided a comprehensive listing of the rights issues that affect self-archiving (RoMEO, 2002).

**Demonstrate financial sustainability.**

The preservation of digital objects requires active and ongoing management. Multiple copies of files must be maintained to guard against the loss or corruption of data. Media must be regularly refreshed and data monitored for corruption. Periodically, data must be migrated to new file formats or emulators must be developed to ensure that the information encoded in a file can be decoded and used.

To provide these archival services e-print repositories need to operate in a secure and relatively stable funding environment. This is not presently the case.

Within the UK Higher Education sector, e-print repository developments have been funded as projects, not services. The JISC eLib programme, for example, funded four e-

print repositories, including Cogprints.<sup>10</sup> These e-print repositories continue to operate, but in the case of Cogprints at least, they operate without dedicated funding (S. Harnad, personal communication, April 15, 2003). Even ArXiv relied on a series of grants from the US National Science Foundation and support from Los Alamos National Laboratory before its recent move to Cornell University (Jackson, 2002).

Currently, JISC is funding a number of experimental institutional repository projects as part of the FAIR programme.<sup>11</sup> The study team received a strong message from these projects that the short-term nature of their funding prevents them from making any type of long-term commitment to preserving e-prints. The apparent assumption is that if these projects are successful they will garner financial support from their institutions in the future. Indeed, one of the reasons for the change in direction away from centralised subject based repositories and towards distributed institutionally based repositories is that the institutional repository model ties the interests of the repository more closely to the interests of the institution, a potential source of relative financial stability and security.

**Provide services within a viable organisational setting, including an appropriate legal status, mission and staffing level.**

In addition to financial sustainability, archival repositories must also be able to operate as effective organisations.

... says Greg Kuperberg, "What I think sets the arXiv apart [from other preprint servers] is its oversight even more than its software. It has a full-time paid staff, several dedicated volunteer helpers, and an array of moderators and advisors. This escalation of policy and structure is just what you would expect for a system that now gets 30,000 submissions a year."  
Jackson (2002, p. 25)

An archival e-print repository will need to take on many of the characteristics of a professionally run library or archive. Repositories run on an informal or voluntary basis will be vulnerable to many threats such as legal challenges from publishers, simple lack of time or loss of interest from those managing the repository, lack of dedicated technical support and reliance on the goodwill of computing support staff for server space.

**Perform the functions of the repository according to documented policies and procedures that are monitored and can be externally assessed.**

Kuperberg, quoted in Jackson above, notes the need for policy and structure when a repository grows in size. Many repositories appear to operate with a minimum of formal policies and procedures. It is difficult to be certain as many repositories do not provide easily identifiable policy and procedure documents. Those that are available can often be interpreted as help documents for users, so it is difficult to judge if they are adhering to good practice and appropriate standards in their management of the repository collection.

Some of the range of issues that should be addressed in appropriate policy and procedure documents are illustrated by the DSpace at MIT Web site policy section (DSpace, 2002c).

**Perform the functions of the repository according to relevant standards and best practices.**

E-Print repositories are aware of, and making use of, appropriate standards and best practices for access functionality. There is wide spread adoption of the OAI-PMH and the

---

<sup>10</sup> The four preprint repository projects were: CogPrints: The Cognitive Sciences Eprint Archive; Education-line: Electronic Texts in Education and Training; Formations; and WoPEc: Working Papers in Economics. See <http://www.ukoln.ac.uk/services/elib/projects/> for links to the project Web sites.

<sup>11</sup> These projects are grouped together in the programme's *E-Prints and E-Theses* cluster. See [http://www.jisc.ac.uk/index.cfm?name=programme\\_fair](http://www.jisc.ac.uk/index.cfm?name=programme_fair) for links to the individual project Web sites

use of Dublin Core metadata as a basic interoperability standard. In other areas, the situation is less clear.

### **Accept responsibility for the long-term preservation of material deposited in the repository.**

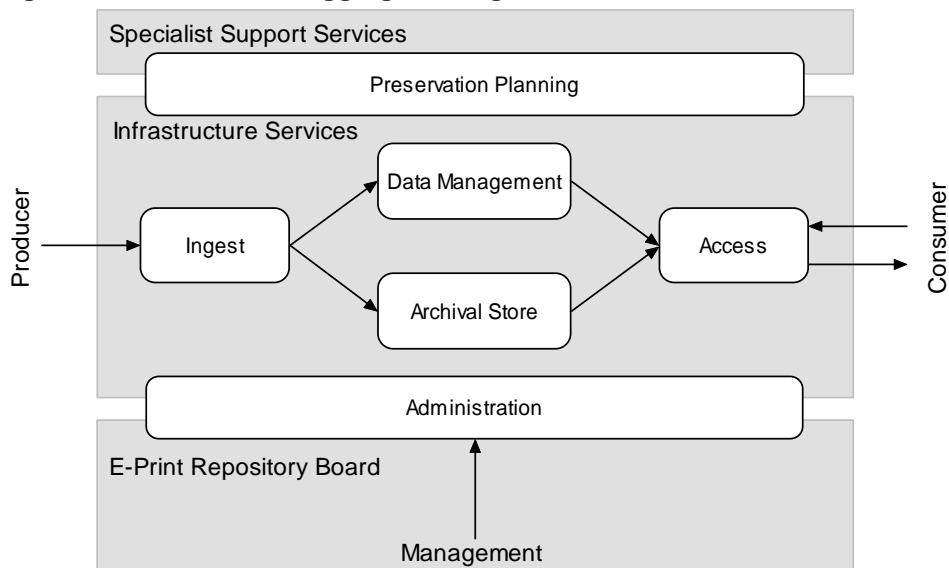
Most e-print repositories do not see it as their role to ensure the long-term survival of the e-prints they make available and therefore they make no commitment to preserve e-prints in their collections. However, even in situations where a repository wishes to take on a preservation role, it is an open question if the repository will have the capability to do so. To be able to accept responsibility for the long-term preservation of e-prints, the repository will need to have met the five non-functional requirements just discussed. Of these, the requirements for financial sustainability and a viable organisational setting are probably the most important, and the present project based funding of e-print developments in the UK Higher Education sector does not assist in achieving these requirements.

Simple actions could still be taken to improve the situation. A valuable first step would be for e-print repositories to ensure that authors fully understand what the repository is, and is not, capable of providing in terms of preservation. Additionally e-print repositories could make provisions to transfer or return e-prints in the event of the repository shutting down. Of the two suggestions, transfer to another repository is preferable to return to the author, although this may only be possible if the repository has an appropriate formal agreement with the depositing authors.

## **10.4 Preserving E-Prints in a Disaggregated Environment**

Institutional libraries and archives have traditionally taken a key role in preserving research, both unpublished and published, but they are unprepared to take on responsibility for the preservation of digital material (Hedstrom & Montgomery, 1998). The functional preservation of digital material draws upon a range of skills that may be spread across a number of specialities and departments, particularly those involved with information technology. Staff with specific knowledge and experience of practical digital preservation may not be available at all within the institution and these skills may need to be brought in from a third party.

**Figure 10.2: E-Prints Disaggregated Organisation Model**



Exactly how roles and responsibilities might be shared between different specialists, departments or organisations can be examined by turning again to the OAIS model. In figure 10.2, the functional requirements of an archival e-print repository are overlaid by a categorisation of the key groups that will provide this functionality.

We envisage the involvement of three *groups* of specialists, who may or may not be located within the same organisation, combining their expertise to form a complete archival e-print repository. In a nutshell, the disaggregated model presented in Figure 10.2 recommends separating out issues to do with the content of an e-print repository, particularly collection and retention policy, and assessment of submissions, from the technical management and delivery of the e-print. A similar separation of tasks is proposed by MIT Libraries' for their implementation of DSpace (Barton & Walker, 2000, p. 3). At one extreme all three groups of specialists may be located in the same organisation, although perhaps spread across a number of sections of that organisation. At the other extreme, the specialists may be distributed across multiple organisations. The need to ensure that work practices are compatible, communications and management are efficient, and services are technically interoperable will place some practical limits on the disaggregation of an archival e-print repository, but there is still considerable scope for a variety of solutions to emerge.

The interests of its authors and readers should guide an e-print repository, therefore the most important component of this disaggregated repository model is the E-Print Repository Board, which should be formed from members of the repositories community (for example, representatives of repository users, institutional representatives or representatives of subject based professional bodies). In terms of the OAIS model, the Repository Board would undertake some of the management and administrative entities tasks, notably defining the repository's collection policy and negotiating submission agreements with authors. Aspects of this idea can be seen in existing e-print repository services such as the eScholarship programme at the University of California (<http://escholarship.cdlib.org/>).

*Infrastructure Services*, namely the ingest, data management, archival storage and access functions that are required to operate the repository, could be divided up in various ways. Most obviously, there are many commercial and non-commercial organisations capable of providing the archival storage function. In an institutional setting, it is likely that computer services will take on responsibility for archival storage, but that ingest, data management and access might be controlled by library services. At the inter-institutional level, the RePEc collaboration involves centralised data management, in the form of the RePEc database, but leaves ingest, archival storage and access distributed across the contributing repositories.

Recognising the current scarcity of digital preservation expertise and services, the model assumes that specific preservation tasks might be outsourced to specialist support services. In practice, infrastructure services have *prime* responsibility for ensuring the security and preservation of the e-prints, and will therefore need to be well informed about preservation issues, and capable of deciding when to make use of specialist support services. There are three main ways in which specialist support services, such as those the planned DCC may offer, could be utilised. Infrastructure services may choose to:

- Subscribe to specialist services that provide information relevant to e-print collection management, such as technology watch and file format registry services
- Outsource specific preservation actions, such as bulk migration of files, to specialist services
- Seek specialist advice and support for their preservation planning activities

Infrastructure and specialist support services may be provided by a single organisation in some situations. The recently launched OCLC Digital Archive (<http://www.oclc.org/digitalpreservation/services/archiving/digital/>) offers this type of unified service, while within the JISC IE existing services such as the AHDS (<http://ahds.ac.uk/>) and the Economic and Social Data Service (ESDS, <http://esds.ac.uk/>) could provide a similar combined service.

## 10.5 Organisational Models in the JISC IE

In the JISC IE, a number of organisational models for the provision of archival e-print repositories could develop. These models are not mutually exclusive, and disaggregated provision of archival repository functions does not necessarily require the establishment of national services. Institutions, or consortiums of institutions could provide their own preservation services, while commercial solutions could also play a role.

### Full E-Print Repository

E-Print repositories operated within larger institutions may be in a position to undertake the full range of OAIS functional requirements. Projects such as DSpace at MIT and [DSpace@Cambridge](http://www.lib.cam.ac.uk/dspace) (<http://www.lib.cam.ac.uk/dspace>) are examples of institutional repositories that are likely to be self-contained.

### E-Print Repository with Specialist Support

Otherwise self-contained e-print repositories may need, or prefer, to call upon external services with specialist expertise in digital preservation. In the JISC IE, the JISC data services themselves further supported by the planned DCC, could provide these services.

### E-Print Repository with Outsourced Preservation Services

E-Print repositories do not currently provide preservation services, and do not see it as a core part of their activity. Therefore, where it is concluded that particular collections of e-prints should be preserved, it may be that an external organisation takes full control of this activity.

### Outsourced E-Print Repository Services

An individual academic, project, interest group or institution could make use of an external e-prints repository service. More than one supplier of e-print repository services may emerge, and commercial services, such as Ingenta's plans to make a version of the Southampton Eprints software available as a service to institutions may be important.<sup>12</sup>

Outsourcing e-print repository services could prove a cost-effective solution for the e-print collections of smaller institutions, projects and individual academic staff. This model has particular potential as a way of providing a secure archival home for e-prints not currently held in a formal e-print repository.

## 10.6 Recommendations for Preserving E-Prints in the JISC IE

### Encourage Preservation Planning in Existing E-Print Repositories

E-Print repositories should be encouraged to incorporate preservation planning functions into their operations. However, *preservation requirements should not add to the real or perceived barriers that discourage authors from depositing their work in e-print repositories*. E-Print repositories that lack the infrastructure to undertake preservation planning and relative activities should be encouraged to develop collaborative arrangements with preservation and data services

---

<sup>12</sup> See the Ingenta Press Release at:

[http://www.ingenta.com/isis/general/Jsp/ingenta?target=/about\\_ingenta/press\\_releases/southampton.jsp&WebLogicSession=PrlyjkSzErP959EoDI5T\]-6087404270695523202/-1052814329/6/7051/7051/7052/7052/7051/-1](http://www.ingenta.com/isis/general/Jsp/ingenta?target=/about_ingenta/press_releases/southampton.jsp&WebLogicSession=PrlyjkSzErP959EoDI5T]-6087404270695523202/-1052814329/6/7051/7051/7052/7052/7051/-1)

### **Funding for E-Print Repositories**

Existing or planned e-print repositories established through project funding do not necessarily have a secure future. Institutions and national funding bodies should clarify their plans for future contribution to e-print repositories.

### **E-Print Repositories should provide Clear Collection and Retention Statements**

E-Print repositories should make available to authors and readers clear statements of their collection and retention policies. The retention period should be discussed with each submitting author, and the repository should make clear the details of their retention commitment.

Specifically, the repository should make clear how long they will hold the e-print and make it available online, and whether they will undertake to migrate the e-print if it becomes inaccessible due to technological obsolescence. As a corollary to this, e-print repositories should clarify arrangements for the transfer or disposal of e-prints in the event of the repository's closure.

### **Develop a Model Licence for E-Prints**

JISC should commission the development of a model licence for the deposit of e-prints into e-print repositories.

### **Advice and Outreach**

JISC should provide advice and outreach to repository managers to make them more aware of preservation issues and current best practice that could be applied to their repository.

Specific actions include:

- Summarise key findings of this report in a briefing document for repository managers
- Establish single point of contact for e-print repository managers to coordinate relevant advice from all JISC advisory services
- Run a risk assessment and preservation planning workshop for repository managers

### **E-Print User Needs Analysis**

JISC should consider research into e-prints that may be held in settings other than formal e-print repositories.

This analysis should:

- Establish an accurate baseline of current e-print usage, and provide well supported projections for future usage
- Determine the wishes of individual research communities regarding minimum retention periods for e-prints
- Establish whether or not e-print readers want long-term access to the e-prints
- Establish whether or not e-print authors want their e-prints to be held in the long-term
- Establish in what situations information professionals believe e-prints should be preserved

### **Pilot of a National E-Print Preservation Service**

JISC should consider funding a longer-term project to develop a fully costed e-print repository infrastructure that is based on the OAIS Reference Model. It is recommended that this is a practical study that includes implementation at one or more e-print repositories and their partners as appropriate to the chosen organisation model.

The infrastructure pilot study should seek to:

- Identify the actual costs of implementing different preservation options across the life-cycle of an e-print
- Establish standards, best practice, processes and procedures for the management, preservation and presentation of e-prints, and to articulate these in an e-prints Digital Repository Handbook (much of this could be compiled from outputs from FAIR projects)
- Investigate requirements for software automation to perform collections management, data and metadata transfer, and preservation actions
- Expand existing e-print repository software and provide with plug-in modules, to assist in a range of preservation tasks (tools that can automatically identify file formats, tools to convert file formats, and tools to collect preservation metadata would be useful)
- Trial a licence agreement for e-print preservation (building on the RoMEO project)
- Implementation of the repository infrastructure at one or more e-print repositories either at a single institution or in collaboration with one or more JISC-funded services as appropriate
- Trial a preservation service for e-prints provided in informal settings

It is envisaged that the Handbook, together with the infrastructure and associated tools would have wider uses beyond this project and could be employed by other e-print repository managers or their partners to manage and preserve their content.

Storage requirements for a pilot are unlikely to be significant. Based on an estimated size of 0.5 – 1.0 MB per e-print, a pilot storing 5,000 e-prints (approximately the number of e-prints in the UK academic domain) would only require 5 GB of storage per copy. Staffing costs will be far more significant. The pilot will need to provide staffing for:

- Evaluation or development of automation tools
- Systems administration
- Repository system development
- Coordination between partners

# 11 References

- Adobe Systems Incorporated (1992a). *Encapsulated postscript file format specification: version 3.0*. Retrieved on May 3, 2003, from <http://www-cdf.fnal.gov/offline/PostScript/5002.PDF>
- Adobe Systems Incorporated (1992b). *TIFF Revision 6.0*. Retrieved on May 3, 2003, from <http://partners.adobe.com/asn/developer/pdfs/tn/TIFF6.pdf>
- Adobe Systems Incorporated (1999). *Postscript language reference: third edition*. Retrieved on May 3, 2003, from <http://partners.adobe.com/asn/developer/pdfs/tn/PLRM.pdf>
- Adobe Systems Incorporated (2003a). *What is Adobe PDF?*. Retrieved on May 3, 2003, from <http://www.adobe.com/products/acrobat/adobepdf.html>
- Adobe Systems Incorporated (2003b). *Acrobat 5.0 SDK documentation*. Retrieved on May 3, 2003, from <http://partners.adobe.com/asn/developer/acrosdk/docs.html#filefmtspecs>
- arXiv.org monthly submission rate statistics* (n.d.). Retrieved on May 2, 2003 from [http://arxiv.org/show\\_monthly\\_submissions](http://arxiv.org/show_monthly_submissions)
- Bailey, C. W. (2003). *Scholarly electronic publishing bibliography*. Retrieved May 1, 2003, from <http://info.lib.uh.edu/sepb/sepb.html>
- Barton, M. R. & Walker, J. H. (2002). *MIT Libraries' DSpace Business Plan Project Final Report to the Andrew W. Mellon Foundation*. Retrieved on October 16, 2003 from <http://libraries.mit.edu/dspace-mit/mit/mellon.pdf>
- Bass, M., Stuve, D., Tansley, R., Branschofsky, M., Breton, P., Carmichael, P., Cattey, B., Chudnov, D., & Ng, J. (2002). *DSpace – functionality*. Retrieved on May 3, from <http://www.dspace.org/technology/functionality.pdf>
- Beagrie, N. (2002). *A continuing access and digital preservation strategy for the Joint Information Systems Committee (JISC) 2002-2005*. Retrieved on April 30, 2003, from [http://www.jisc.ac.uk/uploaded\\_documents/dpstrategy2002b.rtf](http://www.jisc.ac.uk/uploaded_documents/dpstrategy2002b.rtf)
- Beebe, L. & Meyers, B. (1999). The unsettled state of archiving. *The Journal of Electronic Publishing*, 4(4). Retrieved April 29, 2003, from <http://www.press.umich.edu/jep/04-04/beebe.html>
- Boyce, P. (2000). For better or worse: preprint servers are here to stay. *College and Research Libraries News*, 61(5), pp. 404-407
- Brown, C. (2001). The Coming of Age of E-Prints in the Literature of Physics. *Issues in Science and Technology Librarianship*, Summer 2001. Retrieved on April 30, 2003, from <http://www.library.ucsb.edu/istl/01-summer/refereed.html>
- Budapest Open Archive Initiative [BOAI] (2002). *Budapest Open Access Initiative*. Retrieved on May 3, 2003, from <http://www.soros.org/openaccess/read.shtml>
- California Digital Library [CDL] (2001). *Digital object standard: metadata, content and encoding*. Retrieved on May 3, 2003, from <http://www.cdlib.org/about/publications/CDLObjectStd-2001.pdf>
- Consultative Committee for Space Data Systems [CCSDS] (2002). *Reference model for an open archival system*, CCSDS 650.0-B-1 Blue Book Retrieved on May 3, 2003 from <http://www.ccsds.org/documents/650x0b1.pdf>

CEDARS (n.d.). Metadata for digital preservation: the Cedars Project outline specification. Retrieved on May 3, 2003, from <http://www.leeds.ac.uk/cedars/colman/metadata/metadataspec.html>

CEDARS (2002). *Cedars Guide to Digital Preservation Strategies*. Retrieved on May 3, 2003, from <http://www.leeds.ac.uk/cedars/guideto/dpstrategies/dpstrategies.html>

Crow, R. (2002a). *The case for institutional repositories: A SPARC Position Paper*. The Scholarly Publishing and Academic Resources Coalition. Retrieved on April 25, 2003 from <http://www.arl.org/sparc/IR/ir.html>

Crow, R. (2002b). *SPARC institutional repository checklist & resource guide*. The Scholarly Publishing and Academic Resources Coalition. Retrieved on May 3, 2003 from [http://www.arl.org/sparc/IR/IR\\_Guide\\_v1.pdf](http://www.arl.org/sparc/IR/IR_Guide_v1.pdf)

Crow, R. (2003). *A Guide to Institutional Repository Software*. Open Society Institute. Retrieved on October 29, 2003, from <http://www.soros.org/openaccess/software/>

Day, M. (2001). Regular columns: e-print services and long-term access to the record of scholarly and scientific research. *Ariadne*, no. 28. Retrieved on 31 May, 2003, from <http://www.ariadne.ac.uk/issue28/metadata/>

Day, M. (2003). *Collecting and preserving the World Wide Web: A feasibility study undertaken for the JISC and Wellcome Trust*. Retrieved on May 3, 2003, from [http://library.wellcome.ac.uk/projects/archiving\\_feasibility.pdf](http://library.wellcome.ac.uk/projects/archiving_feasibility.pdf)

Dublin Core Metadata Initiative [DCMI] (2003a). *XMLS schemas*. Retrieved on May 3, 2003, from <http://dublincore.org/schemas/xmls/>

Dublin Core Metadata Initiative [DCMI] (2003b). *DCMI metadata terms: a complete historical record*. Retrieved on May 3, 2003 from <http://dublincore.org/usage/terms/history/>

DLM-Forum (1997). *Guidelines on best practices for using electronic information*. Luxembourg: Office for Official Publications of the European Communities. Retrieved on May 3, 2003, from <http://europa.eu.int/ISPO/dlm/documents/gdlines.pdf>

DSpace (2002b). *Metadata*. Retrieved on May 3, 2003, from <http://dspace.org/technology/metadata.html>

DSpace (2002c). *MIT DSpace policies and guidelines*. Retrieved on October 16, 2003, from <http://libraries.mit.edu/dspace-mit/mit/policies/>

Elsevier Science: News Items (n.d.). *National Library of The Netherlands and Elsevier Science make digital preservation history: press release*. Retrieved on May 2, 2003, from <http://www.elsevier.com/homepage/newhpgnews/preview/KB/links/link5.htm>

Eprints.org (2002). *What is an eprint?*. Retrieved on May 2, 2003, from <http://www.eprints.org/self-faq/>

Florida Centre for Library Automation [FCLA] Digital Archive (2003). *Action plan background: PDF version 1.4*. Retrieved on May 3, 2003, from [http://www.fcla.edu/digitalArchive/pdfs/action\\_plan\\_bgrounds/pdf\\_1\\_4.pdf](http://www.fcla.edu/digitalArchive/pdfs/action_plan_bgrounds/pdf_1_4.pdf)

*The Fedora Project*. (n.d.). Retrieved on October 29, 2003, from <http://www.fedora.info/index.shtml>

Fraser, M. (2003). *The RTS and eprints*. Oxford University Computing Services. Retrieved on May 2, 2003, from [http://www.oucs.ox.ac.uk/rts/eprints/eprints\\_intro.xml](http://www.oucs.ox.ac.uk/rts/eprints/eprints_intro.xml)

- Gadd, E., Oppenheim, C. & Proberts, S. (2003). *RoMEO Studies 1: The impact of copyright ownership on academic author self-archiving*. Retrieved on May 3, 2003, from [http://www.lboro.ac.uk/departments/ls/disresearch/romeo/RoMEO\\_studies1.pdf](http://www.lboro.ac.uk/departments/ls/disresearch/romeo/RoMEO_studies1.pdf)
- Granger, S., Russell, K. & Weinberger, E. (2000). *Cost elements of digital preservation*. Retrieved on May 3, 2003, from <http://www.leeds.ac.uk/cedars/documents/CIW01r.html>
- Harnad, S. (1994). *Scholarly journals at the crossroads: A subversive proposal for electronic publishing*. Retrieved on May 3, 2003 from <http://www.arl.org/scomm/subversive/sub01.html>
- Harnad, S. (2001). The self-archiving initiative. *Nature* no. 410, 1024-1025. Macmillan Publishers Ltd. Retrieved on May 3, 2003 from <http://www.ecs.soton.ac.uk/%7Eharnad/Tp/nature4.htm>
- Harnad, S. (2002). *Scholarly journals at the crossroads: a subversive proposal for electronic publishing*. Association of Research Libraries, Washington, DC. Retrieved on May 3, 2003 from <http://www.arl.org/scomm/subversive/sub01.html>
- Harnad, S. & Goodman, D. (2003). Online transactions ["Eprint versions and removals"]. Messages posted to American-Scientist-E-PRINT-Forum
- Harnad, S. & Sargent, D. (2003). Online transactions ["EPrints, DSpace or ESpace?"]. Messages posted to JISC-DEVELOPMENT@JISCMail.AC.UK
- Hedstrom, M. & Montgomery, S. (1998). *Digital preservation needs and requirements in RLG member institutions*. Research Libraries Group. Retrieved on May 3, 2003, from <http://www.rlg.org/preserv/digpres.pdf>.
- ISO (1986). *Standard Generalized Mark-up Language (SGML)*. ISO (1986) 8879:1986
- ISO (1991). *7-bit coded character set for information interchange*. ISO/IEC 646:1991
- ISO (1998-2001). *8-bit single-byte coded graphic character sets (Parts 1-16, 1998-2001)*. ISO 8859
- ISO (2000a). *Universal Multiple-Octet Coded Character Set (UCS)*. ISO (2000a) 10646:2000.
- ISO (2000b). *JPEG 2000 image coding system (with corrections and amendment from 2002)*. ISO ISO/IEC 15444-1:2000
- Jackson, A. (2002). From preprints to e-prints: the rise of electronic preprint servers in mathematics. *Notices of the AMS*, 49(1) pp. 23-31. Retrieved on May 3, 2003, from <http://www.ams.org/notices/200201/fea-preprints.pdf>
- JISC (2003). *Discussion paper: draft ITT for a digital curation centre*, ver. 2.2. Retrieved on May 2, 2003 from [http://www.jisc.ac.uk/uploaded\\_documents/digitalcurationcentrev3.pdf](http://www.jisc.ac.uk/uploaded_documents/digitalcurationcentrev3.pdf)
- Jones, M. *Archiving E-Journals Consultancy Final Report*, Consultation Draft, Report Commissioned by Joint Information Systems Committee (JISC) (2003). Retrieved on October 14, 2003 from [http://www.jisc.ac.uk/uploaded\\_documents/ejournalsdraftFinalReport.pdf](http://www.jisc.ac.uk/uploaded_documents/ejournalsdraftFinalReport.pdf)
- Jones, M. & Beagrie, N. (2001). *Preservation Management of Digital Materials: A Handbook*. British Library
- Langer, J. (2000). Physicists in the new era of electronic publishing. *Physics Today Online*. Retrieved on May 3, 2003 from <http://www.aip.org/pt/vol-53/iss-8/p35.html>

Lawal, I. (2002). Scholarly communication: the use and non-use of e-print archives for the dissemination of scientific information. *Issues in Science and Technological Librarianship*, Fall 2002. Retrieved on May 3, 2003, from <http://www.istl.org/02-fall/article3.html>

Lawrence, G., Kehoe, W., Rieger, O., Walters, A. & Kenney, A. (2000). *Risk management of digital information: a file format investigation*. Retrieved on May 3, 2003, from <http://www.clir.org/pubs/reports/pub93/contents.html>

Luce, R. E. (2001). E-prints Intersect the Digital Library: Inside the Los Alamos arXiv. *Issues in Science and Technical Librarianship*, Winter 2001. Retrieved on May 3, 2003, from <http://www.istl.org/istl/01-winter/article3.html>

Lupovici, C. & Masanès, J. (2000). *NEDLIB Metadata for Long-term Preservation*. NedLib Report Series, no. 2. Retrieved on May 3, 2003, from <http://www.kb.nl/coop/nedlib/results/NEDLIBmetadata.pdf>

Lynch, C. A. (2003). Institutional Repositories: Essential Infrastructure for Scholarship in the Digital Age. *ARL Bimonthly Report*, no. 226. Retrieved on May 3, 2003, from <http://www.arl.org/newsltr/226/ir.html>

*METS: Metadata Encoding and Transmission Standard* (2003). Retrieved on May 1, 2003, from <http://www.loc.gov/standards/mets/mets.xsd>

Microsoft Corporation (1999). *Rich Text Format (RTF) Specification, version 1.6*. Retrieved on May 3, 2003, from <http://msdn.microsoft.com/library/default.asp?url=/library/en-us/dnrftspec/html/rftspec.asp>

National Library of Australia (1999). *Preservation Metadata for Digital Collections*. Retrieved on May 3, 2003, from <http://www.nla.gov.au/preserve/pmeta.html>

National Archives of Australia (1999). *Recordkeeping Metadata Standard for Commonwealth Agencies*. Retrieved on May 3, 2003, from <http://www.naa.gov.au/recordkeeping/control/rkms/contents.html>

National Library of New Zealand (2002). *Metadata Standards Framework – Preservation Metadata*. Retrieved on May 3, 2003, from [http://www.natlib.govt.nz/files/4initiatives\\_metaschema.pdf](http://www.natlib.govt.nz/files/4initiatives_metaschema.pdf)

Open Archives Initiative [OAI] (2003). *The Open Archives Initiative Protocol for Metadata Harvesting*. Retrieved on May 3, 2003, from <http://www.openarchives.org/OAI/openarchivesprotocol.htm>

Ockerbloom, J. (2001). Archiving and Preserving PDF Files. *RLG DigiNews*, 5(1). Retrieved on May 2, 2003, from <http://www.rlg.org/preserv/diginews/diginews5-1.html - feature2>

OCLC [Online Computer Library Center] (2002). *Preservation Metadata and the OAIS Information Model: A Metadata Framework to Support the Preservation of Digital Objects*. OCLC/RLG Working Group on Preservation Metadata. Retrieved on May 3, 2003, from [http://www.oclc.org/research/pmwg/pm\\_framework.pdf](http://www.oclc.org/research/pmwg/pm_framework.pdf)

Office of the e-Envoy (2002). *e-Government Interoperability Framework*, ver. 4.0. Retrieved on October 16, 2003, from <http://www.e-envoy.gov.uk/assetRoot/04/00/22/35/04002235.doc>

Pinfield, S., Gardner, M., & MacColl, J. (2002). Setting up an institutional e-print archive. *Ariadne*, no. 31. Retrieved on April 20, 2003, from <http://www.ariadne.ac.uk/issue31/eprint-archives/>

Powell, A. (2003). *RDN admin metadata*, ver. 1.0. Retrieved on May 3, 2003, from <http://www.rdn.ac.uk/publications/cat-guide/admin/>

Powell .A., Day, M. & Cliff, P. (2003). *Using simple Dublin Core to describe eprints*, ver. 1.2. Retrieved on May 3, 2003, from <http://www.rdn.ac.uk/projects/eprints-uk/docs/simpledc-guidelines/>

Public Record Office (n.d.). *Digital Preservation: PRONOM*. Retrieved on October 16, 2003, from <http://www.pro.gov.uk/about/preservation/digital/pronom/default.htm>

Public Record Office (2002). *Requirements for Electronic Records Management Systems, 2: metadata standard*. Retrieved on May 3, 2003, from <http://www.pro.gov.uk/recordsmanagement/erecords/2002reqs/2002metadafinal.pdf>

Public Record Office Victoria (2000). *VERS Metadata Specification*. Retrieved on May 2, 2003, from <http://www.prov.vic.gov.au/vers/standards/pros9907/99-7-2toc.htm>

Research Libraries Group [RLG] (2002). *Trusted Digital Repositories: Attributes and Responsibilities. An RLG-OCLC Report*. Retrieved on May 3, 2003, from <http://www.rlg.org/longterm/repositories.pdf>

RoMEO (2002). *Project RoMEO*. Retrieved on May 2, 2003, from <http://www.lboro.ac.uk/departments/ls/disresearch/romeo/>

Suber, P. (2003). *Guide to the Free Online Scholarship Movement*. Retrieved on May 1, 2003, from <http://www.earlham.edu/~peters/fos/guide.htm>

*Standards – Electronic Text Center* (n.d.). Retrieved on May 3, 2003, from <http://etext.lib.virginia.edu/standard.html>

TARDIS (2002). *Targeting Academic Research for Deposit and Disclosure*. University of Southampton. Retrieved on May 3, 2003, from <http://tardis.eprints.org/>

Till, J. E. (2001). Predecessors of preprint servers. *Learned Publishing*, 14(1) pp. 7-13. Retrieved on May 3, 2003, from <http://www.catchword.com/09531513/v14n1/contp1.htm>

University of Leeds (2003). *Survey and assessment of sources of information on file formats and software documentation*. Unpublished draft.

W3C (1999). *HTML 4.01 Specification*. Retrieved on May 3, 2003, from <http://www.w3.org/TR/html401/>

W3C (2000). *Extensible Mark-up Language (XML) 1.0 (Second Edition)*. Retrieved on May 6, 2003, from <http://www.w3.org/TR/REC-xml>

Wheatley, P. (2001). *Migration - a CAMiLEON discussion paper*. Retrieved on Oct 15, 2003, from <http://www.ariadne.ac.uk/issue29/camileon/>

# 12 Appendix I: Additional Sources For File Formats

Adobe Systems Incorporated (2002). Recommendations for Creating PDF Files in Excel 97 . Retrieved on May 6, 2003, from <http://www.adobe.com/support/techdocs/13132.htm>

Adobe Systems Incorporated (2002). Recommendations for Creating PDF Files from Word with Acrobat 4.05x. Retrieved on May 6, 2003, from <http://www.adobe.com/support/techdocs/f8be.htm>

Adobe Systems Incorporated (2001). PDF File format specification, version 1.4. Retrieved on October 16, 2003, from [http://partners.adobe.com/asn/acrobat/docs/File\\_Format\\_Specifications/PDFReference.pdf](http://partners.adobe.com/asn/acrobat/docs/File_Format_Specifications/PDFReference.pdf)

Adobe Systems Incorporated (2003). PDF reference, forth edition, Adobe Portable Document Format, version 1.5. Retrieved on May 6, 2003, from [http://partners.adobe.com/asn/acrobat/sdk/public/docs/PDFReference15\\_v6.pdf](http://partners.adobe.com/asn/acrobat/sdk/public/docs/PDFReference15_v6.pdf)

American Mathematical Society TeX Resources Home Page (2003). Retrieved on May 6, 2003, from <http://www.ams.org/tex/>

Bennett, J. (1997). A Framework of Data Types and Formats, and Issues Affecting the Long-term Preservation of Digital Materials. *British Library Research and Innovation Report*, no. 50. Retrieved May 3, 2003, from <http://www.ukoln.ac.uk/services/papers/bl/jisc-npo50/bennet.html>

Coleman J. & Willis, D. (1997). *SGML as a framework for digital preservation and access*. Washington, DC: Council on Library and Information Resources.

Digital Preservation Testbed (2001). Migration: Context and Current Status. Retrieved on May 6, 2003, from <http://www.digitaleduurzaamheid.nl/bibliotheek/docs/Migration.pdf>

Digital Preservation Testbed (2002). XML and Digital Preservation. Retrieved on May 6, 2003, from [http://www.digitaleduurzaamheid.nl/bibliotheek/docs/white-paper\\_xml-en.pdf](http://www.digitaleduurzaamheid.nl/bibliotheek/docs/white-paper_xml-en.pdf)

eScholarship Repository (2000). A Guide to PDF for Scholars. Retrieved on May 6, 2003, from <http://repositories.cdlib.org/howpdfol.pdf>

Gilheany, S. (2000). Permanent Digital Records and the PDF Format. Retrieved on May 6, 2003, from <http://www.archivebuilders.com/whitepapers/22025p.pdf>

Hughes, M. (2002). PDF as an Archive Standard. Retrieved on May 6, 2003, from <http://www.aiim.org/documents/standards/pdfa2.pdf>

Jones, M. & Beagrie, N. (2001). *Preservation Management of Digital Materials: A Handbook*. British Library

McDowell, K. (2001). Export a Word Document to XML. Retrieved on May 6, 2003, from [http://msdn.microsoft.com/library/default.asp?url=/library/en-us/dnword2k/html/odc\\_expwordtoxml.asp](http://msdn.microsoft.com/library/default.asp?url=/library/en-us/dnword2k/html/odc_expwordtoxml.asp)

Mellor, P., Wheatley, P., & Sargent, D. (2002). Migration on Request: A Practical Technique for Preservation. Retrieved on May 6, 2003, from <http://www.si.umich.edu/CAMILEON/reports/migreq.pdf>

Microsoft. (1999). Converting Files Between Different Versions of Office Software: A File Format Matrix. Retrieved on May 6, 2003, from <http://www.microsoft.com/office/previous/deployment/whitepapers/ConvFile.doc>

PDF-A Working Group (n.d.). PDF Archiving Requirements. Retrieved on May 6, 2003, from [http://www.aiim.org/documents/standards/pdf\\_archiving1.pdf](http://www.aiim.org/documents/standards/pdf_archiving1.pdf)

PDF Zone (2003). The online authority for PDF and document management professionals. Retrieved on May 6, 2003, from <http://www.pdfzone.com>

Potter, M. (2002). XML for Digital Preservation. Retrieved on May 6, 2003, from [http://www.digitaleduurzaamheid.nl/bibliotheek/docs/mp\\_erpanet\\_xml.pdf](http://www.digitaleduurzaamheid.nl/bibliotheek/docs/mp_erpanet_xml.pdf)

Ritter, N. (1997). The Unofficial TIFF Home Page. Retrieved on May 6, 2003, from <http://home.earthlink.net/~ritter/tiff/>

Thomas, S. (2002). File Formats for Electronic Text. Retrieved on May 6, 2003, from <http://www.library.adelaide.edu.au/~stthomas/papers/etext-formats.html>

The Joint Photographic Experts Group. (2003). Retrieved on May 6, 2003, from <http://www.jpeg.org>

The TeX Users Group [TUG] (2003). Retrieved on May 6, 2003, from <http://www.tug.org>

## 13 Appendix II: Survey Documents

As part of the study, a short questionnaire (first document below) was distributed to members of CURL (Consortium of University Research Libraries) and contacts at a number of e-print repositories.

A slightly longer discussion document (second document below) was used to initiate discussion with members of the FAIR programme E-Prints and E-Theses cluster group.

# Feasibility and Requirements Study on Preservation of E-Prints

## 13.1.1 Study Overview

The Arts and Humanities Data Service (AHDS) and the University of Nottingham, as lead site in the SHERPA (Securing a Hybrid Environment for Research Preservation and Access) project, have been funded by JISC (the Joint Information Systems Committee, funded by the UK Higher and Further Education Councils) to conduct a requirements and feasibility study for the preservation of e-prints. Further background to the purpose of the project can be found on the JISC Web site at: [http://www.jisc.ac.uk/index.cfm?name=project\\_eprints\\_pres](http://www.jisc.ac.uk/index.cfm?name=project_eprints_pres)

The study will produce a report addressing the following main areas:

- Properties of e-Prints
- Policies and Procedures
- Metadata
- Formats
- Organisational Models
- E-Print Preservation Life-cycle
- Cost Models

We would appreciate your answers and thoughts on the following questions. Please send any replies to [hamish.james@ahds.ac.uk](mailto:hamish.james@ahds.ac.uk)

## 13.1.2 Questions

1. How would you define an e-print?
  - a. Can an e-print file contain material other than text (e.g., images, audio, datasets)?
2. Is your organisation planning/does your organisation operate an e-print repository?
  - a. Does, or will, the repository have a separate collection policy?
  - b. Does, or will, the repository have a retention or removal policy?
  - c. Does, or will, the repository have a preservation policy?
3. Once deposited in a repository, should e-prints be stored indefinitely?
  - a. If yes or no, why?
  - b. If no, who or what determines how long an e-print should be retained?
4. What types of metadata need to accompany an e-print?
  - a. Is metadata supplied by the e-print author sufficient?
  - b. What metadata standard(s) are used for e-prints?
5. Are there particular file formats that are especially suited to/not suited to use for e-prints?
  - a. Does/would your repository accept any file format submitted?

## 13.1.3 Other comments and thoughts

Please include any other comments or thoughts you have.

Thank you!

Hamish James  
[hamish.james@ahds.ac.uk](mailto:hamish.james@ahds.ac.uk)

Raivo Ruusalepp  
[raivo@eba.ee](mailto:raivo@eba.ee)

# Feasibility and Requirements Study on Preservation of e-Prints

## 13.1.4 Study Overview

The Arts and Humanities Data Service (AHDS) and the University of Nottingham, as lead site in the SHERPA project, have been funded by JISC to conduct a requirements and feasibility study for the preservation of e-prints. Further background to the purpose of the project can be found on the JISC website at:

[http://www.jisc.ac.uk/index.cfm?name=project\\_eprints\\_pres](http://www.jisc.ac.uk/index.cfm?name=project_eprints_pres)

The study will produce a report addressing the following main areas:

Properties of e-Prints  
Policies and Procedures  
Metadata  
Formats  
Organisational Models  
e-Print Preservation Life-cycle  
Cost Models

## 13.1.5 Properties of E-Prints

1. What is your definition of an e-print?
  - a. Is it defined technically, or by its relation to publication, or some other dimension?
  - b. Can an e-print file contain material other than text: images, audio, moving images, datasets?
  - c. How does an e-print differ from an e-journal paper or an e-thesis?
2. What is the main purpose of creating an e-print?
  - a. Could e-prints replace traditional publishing methods?

## 13.1.6 E-Print Repositories

### 13.1.6.1 Collection Policies

3. Do you anticipate a slow, moderate or fast growth in the number of e-prints produced (in your repository; in general)?
4. Do you anticipate a slow, moderate or fast growth in the number of e-print repositories?
5. Is there an optimum size for an e-print repository?
6. What are the relative merits of subject based versus institutional repositories?
  - a. Are there cost reasons for preferring one approach to the other?
7. Once deposited in a repository, should e-prints be stored indefinitely?
  - a. If yes, why?
  - b. If no, why?
  - c. If no, what is the likely lifespan of an e-print?
  - d. If no, who or what determines how long an e-print should be retained?
8. How much, and what type of, control should an e-print repository exercise over submissions?

### 13.1.6.2 Retention Policies

9. Does the repository take responsibility for the submitted e-print for a fixed/limited period of time, indefinitely, or is it undetermined?
  - a. How will the cost of maintaining the repository and its content be met?
10. Does the repository have the right/permission to convert/migrate the deposited e-print to new file formats (for the purposes of preserving continuing access to the file)?

11. Do you believe e-print repositories will encounter difficulties with publishers and others over copyright?
12. Does the repository have a preservation policy or guideline or manual?

### **13.1.6.3 Metadata**

13. What types of metadata need to accompany an e-print?
14. Is metadata supplied by the e-print author sufficient?
  - a. Are there any tools or procedures that can improve the quality of metadata created by the author?
  - b. Is any technical and administrative metadata attached to a deposited e-print file (e.g., for the purposes of preservation management or collection management, etc.)?
  - c. What is the cost associated with adding metadata to an e-print?

### **13.1.6.4 Formats**

15. Are there particular file formats that are especially suited to/not suited to use for e-prints?
16. What formats are popular for e-prints, why are they popular?
  - a. What are the most popular formats in your subject area?
17. Have you experienced/do you see any problems with the use particular formats for e-prints?
  - a. Does PostScript present any problems?
  - b. Does PDF present any problems?
  - c. Does Microsoft Word present any problems?
  - d. Does HTML/XHTML present any problems?
  - e. Do image formats (JPEG, GIF etc.) present any problems?

### **13.1.7 System Openness**

18. Does e-print repository software place restrictions on the types of formats that can be accepted?
19. How would you describe/assess the openness of the software system that is used for the e-print repository? Would it be easy to change to a new system and migrate all the archived e-prints to a new system?

### **13.1.8 Other comments and thoughts**

Any other comments or thoughts you have.

Thank you!

Hamish James  
[hamish.james@ahds.ac.uk](mailto:hamish.james@ahds.ac.uk)

Raivo Ruusalepp  
<mailto:raivo@eba.ee>