Response to Invitation to Tender: requirements and feasibility study on preservation of e-prints

A proposal to the JISC from the Arts and Humanities Data Service and the University of Nottingham, Project SHERPA

Document Notes

Author: Sheila Anderson, Hamish James, Stephen Pinfield

Date: 22nd November 2002
Version: 1
Name: preservation study bid v1.doc
Notes: This document is prepared for consideration by the Joint Information Systems Committee, through the JISC Preservation Focus

1. Executive Summary

This response to the 'Invitation to Tender: requirements and feasibility study on the preservation of e-prints' is submitted by the Arts and Humanities Data Service and the University of Nottingham as lead site in the SHERPA project. We would aim to commence the study in January and to complete the draft report for submission by 18th April. Ten bound copies of the final report will be presented to the JISC on 6th May. Other deliverables will be attached to the report as appendices.

The study will be carried out under the direction of Sheila Anderson, Director of the AHDS, and Stephen Pinfield, Director of the SHERPA project and Assistant Director of Information Services at the University of Nottingham. Martha Brundin, AHDS Project Manager, will coordinate the study. Hamish James, Collections Manager at the AHDS and Raivo Ruusalepp, preservation consultant at the Estonian Business Archives will be responsible for undertaking the work outlined in the work packages, assisted by the SHERPA preservation officer (to be appointed) and the SHERPA project officer (to be appointed). Hamish James, Sheila Anderson and Stephen Pinfield will be responsible for writing the final report.

1.1 Partner details

The Arts and Humanities Data Service (AHDS) is a national service funded by the Joint Information Systems Committee (JISC) and the Arts and Humanities Research Board (AHRB), to collect, manage, catalogue, preserve and promote the use of digital resources in research, teaching and learning in the arts and humanities. The AHDS provides advice and guidance in the creation of digital resources to quality standards that ensure their suitability for use in research and teaching and their long-term viability. The AHDS identifies and accessions a wide range of digital resources including texts and e-books, and evaluates, validates, adds metadata, and incorporates the collections into its resource discovery, delivery and preservation systems.

In addition to developing significant in-house expertise in preservation issues and processes, the AHDS has been at the forefront of research into digital preservation

undertaking two key research projects: a LIC funded project to establish guidelines for digital archiving, and a follow-up BL funded project to create a workbook for the preservation management of digital materials. The latter resulted in the highly acclaimed 'Preservation Management of Digital Materials Workbook' published by the British Library. Following significant involvement in the CEDARS project the AHDS is intending to establish a central preservation facility based upon the OAIS model. To facilitate this process the AHDS is undertaking a review of its current distributed practice and research into a number of key aspects, including life-cycle management, metadata schema, formats, and naming architecture. The outputs from the review will provide guidance and recommendations on key elements to be included in the AHDS preservation facility. Outputs will be published and will also form the basis of a published 'real-life' case study in establishing an OAIS compliant preservation archive. Results from the review are expected in February 2003.

SHERPA at University of Nottingham
SHERPA is a three year project funded by JISC (Joint Information Systems Committee) and CURL (Consortium of University Research Libraries) and forms part of the JISC FAIR (Focus on Access to Institutional Resources) programme. SHERPA aims to investigate issues to do with the future of scholarly communication and publishing. In particular, it is initiating the development of openly accessible institutional e-print repositories in a number of CURL universities (and beyond). 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 (using the Open Archives Initiative Protocol) and digital preservation of e-prints.

SHERPA is hosted by the University of Nottingham and involves a number of partner institutions in the UK. The partners involved from the beginning are the universities of Edinburgh, Glasgow, Leeds, Oxford, Sheffield, and York, plus the British Library and the Arts and Humanities Data Service. Further partners will come on board as the project progresses.

CURL, the Consortium of University Research Libraries comprises 26 full members, associates and partners, including the British Library. Its vision embraces the aspiration to provide strong leadership and opportunities for innovation for the scholarly library and research communities. CURL has a successful track record in offering services and running development projects on behalf of the HE community. These include COPAC (the CURL OPAC), the HE Archives Hub, and CEDARS. The CEDARS project was an important focus for digital preservation work for a number of years in the UK.

2. Introduction

Management and preservation requirements for digital materials are fundamentally different from analogue materials. Digital materials can be created using a wide range of technologies and formats, whether born digital or digital surrogates of existing analogue materials. They can be described and documented in a variety of ways – or not at all. They are subject to both physical deterioration and technical obsolescence. More than

one copy can be easily and simply created. Access may be provided through more than one point, and may be distributed. All these factors will impinge upon the approach taken to their management and long-term preservation.

These differences present the curators of digital materials with some fundamental challenges. The way in which materials are created, particularly the technologies used, will determine how conducive to long-term preservation the materials are, and will present varied challenges to curators charged with the subsequent management and preservation of the materials. Curators will need adequate metadata and documentation about the resource if they are to successfully manage, preserve and make the materials accessible. Multiple copies may also imply multiple versions – the digital resource curator must somehow ensure the integrity and authenticity of the resource. They must be aware of changing technologies and fragility of media and take these into consideration from an early stage in the ingest process.

Jones and Beagrie define digital preservation as: "…the series of managed activities necessary to ensure continued access to digital materials for as long as necessary. Digital preservation….refers to all the actions required to maintain access to digital materials beyond the limits of media failure or technological change."[1]

All this suggests that digital curation and preservation requires a more pro-active approach beginning at an earlier stage in the material's lifecycle than would traditionally be the case with analogue materials.

Within the digital preservation community, the concept of the life-cycle (or continuum as it is sometimes called) management of digital resources has emerged to describe and document the active management processes that need to take place, and the key decision making and intervention points along the continuum. The life-cycle concept has been incorporated into OAIS Reference Model, recently adopted as an ISO standard for digital preservation. The OAIS model is proving a strong foundation for the development of digital archiving projects and services. In the UK it has been tested and prototyped in digital preservation projects such as Cedars. In America OCLC are developing digital lifecycle management services based on OAIS, and projects such as D-Space are developing tools to manage content based around the model.

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, promoting discovery and access, and developing the OAI schema and tools, 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 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. This feasibility study will address the particular requirements of e-prints.

---

[1] Jones, M and Beagrie, N, 'Preservation Management of Digital Materials: a Handbook', British Library, London, 2002

2. Methodology

Although at a relatively early stage much research has and continues to be carried out into the preservation of a wide range of digital materials. There is therefore a body of literature and other project outputs on which to draw when conducting this feasibility study. This proposal will use these as a starting point for a practical, focused exploration into the feasibility and requirements for the preservation of e-prints. The study will address three fundamental issues:

- The requirements for preserving e-print content
- Organisational models that would best ensure their long-term preservation and continued access
- Implementation scenarios, including life cycle and cost modelling

There are a number of practical issues to address when investigating the requirements for the preservation of e-prints. First among these is the nature of e-prints themselves – their content, format, elements (e-prints may well contain images, tables, scientific formulae, etc.), software dependency and so on. There are also related issues to address in terms of collection development and policy, including both selection and retention criteria and IPR and copyright policy. The study will explore the idea of 'preservation levels' depending upon the selection and ingest processes and conformity with preferred formats, as this will impact upon the cost and viability of long-term preservation. For e-prints, repositories will also wish to evaluate these policies in the light of any future publication of the e-prints within their collections, and the possible transfer of preservation responsibility to the publisher.

Collections policies will also need to identify formats accepted and held, and their impact upon both the possibility and the cost of preserving the content. Repositories undertaking preservation will require a metadata set that not only facilitates discovery and access, but that provides metadata suitable for use within a robust preservation system. OAI metadata provides a metadata set suitable for supporting discovery but will need the addition of preservation metadata if long-term retention is to be supported. The study will draw upon existing work in this area, including work undertaken by UKOLN on behalf of the Cedars project, NEDLIB, the National Library of Australia, RLG and OCLC.

Whilst the nature of digital materials presents us with a number of challenges, it also presents us with opportunities. The ability to create multiple copies, while complicating issues of integrity and authenticity, also alters the nature of what is possible and desirable when looking at institutional and cost models. A number of possible scenarios might be developed that would provide a sustainable preservation infrastructure for e-prints. This study will explore the potential of these models, taking into account the proposed Digital Curation Centre, and will include options for both institutional and national repositories and a combination of the two. The study will explore how and where expertise might be developed and best employed, the capacity of different repositories, timescales, copyright and licensing issues, sustainability and practicalities such as transfer methods and procedures.

In developing these models the study will draw extensively upon the Open Archival Information System (OAIS) Reference Model. The study will concentrate on developing life cycle and cost models, drawing upon existing projects researching cost modelling e.g. Roquade, D-Space, and the work of the Digital Preservation Coalition. It will give serious attention to sustainability and scalability issues, and to possible technical strategies, looking in particular at emulation and migration strategies.

It is proposed to breakdown the study into clearly defined work packages with associated tasks and deliverables. Further details of the work packages may be found in Appendix A. Work packages will be as follows:

1. Gather and collate information on existing e-print services; preservation projects and services, including JISC services; metadata initiatives etc.

2. Investigate properties of e-prints

3. Review collection policies and procedures, selection and retention policies, IPR and copyright issues

4. Metadata review

5. Format review

6. Review of organisational models

7. Development of e-print preservation life cycle and cost models

8. Final Report

3. Timetable

We would propose to commence the project in January 2003 and complete the first draft of the report by 18th April. The final report will be submitted by 6th May 2003.

Actions	Timescale	Partner	Days
1. Information gathering Identify services, projects etc. Collate relevant information	January	AHDS	5 days
2. Investigate properties of e-prints Information from 1. Consult e-print repositories Consult FAIR e-print cluster group	January	Consultant	4 days
3. Review and analysis of policies and procedures Information from 1.	February	AHDS Nottingham	7 days

Consult selected repositories Consult JISC FAIR cluster groups: e-prints, e-theses, IPR/Copyright group			
4. Metadata Review Information from 1. Consult relevant experts Draw up recommendations for e-print preservation metadata	February	Consultant	5 days
5. Format Review Information from 1. and 2. Consult FAIR e-print cluster group Consult other JISC Services	February	Consultant	5 days
6. Review of Organisational Models Consult Preservation Focus and relevant project and services: national and institutional Consult e-print cluster group, CURL, CILIP Use information from 1. and 3. Visits to selected institutional repositories (3) and national services (2)	March	AHDS Nottingham	8 days
7. Development of e-print preservation life-cycle and cost models Reports and analyses from 1-6 OAIS standard Preservation methods (emulation/migration)	March/April	AHDS Nottingham	10 days
8. Final Report	1st Draft: April 18th Final report: 6th May	AHDS Nottingham	12 days 5 days
			61 days

4. Breakdown of Costs

Category	**Breakdown**	**Cost**
Consultant	14 days @ £400 per day	5,600
AHDS/Nottingham staff	47 days @ £325 per day	15,275
Visits x 2 staff 3 to institutional repositories 2 to national services	5 @ £300 per visit	1,500
Cluster Group meeting (1)	10 participants @ £150	1,500
Offices Costs: Phone / email consultation Photocopy/printing 10 bound copies of report		1,000
TOTAL		**£24,875**

Appendix A

Work packages

1. Gather and collate information on existing e-print services; preservation projects and services, including JISC services; metadata initiatives

Description of Work

The project team will conduct an extensive review of existing sources of information, published, and unpublished, starting with relevant services, projects and initiatives such as RoMeo, the OAI Community and JISC Services

Tasks

1. Web search
2. Review web sites of existing services, projects and initiatives
3. Paper and electronic published literature search
4. Gather and collate information, policy papers etc.

Deliverables

Deliverable: annotated bibliography

2. Investigate properties of e-prints

Description of Work

This work package will develop a comprehensive definition of an e-print that will clarify their similarities and differences from other types of digital material, helping to inform the extent to which work packages 3, 4 and 5 can draw on work dealing with other types of digital material, and identifying the key issues unique to e-prints.

Tasks

1. Sample existing e-print repositories in order to describe the typical content and logical structure of an e-print
2. Contact existing e-print repositories and JISC Fair projects for information on the types of software package used to create and access e-prints
3. Investigate any restrictions imposed on the properties of e-prints by the software and hardware systems used to manage e-print repositories

Deliverables

Report providing a breakdown of the properties of an e-print according to content, logical structure, formats and software

3. Review collection policies and procedures, selection and retention policies, IPR and copyright issues

Description of Work

E-prints are likely to be the product of work in progress, and thus selection and retention policies will be complicated by the potentially high turnover and proliferation of versions. E-prints may also be related to subsequent published work raising more questions about retention and rights. These issues will be illuminated through in-depth consultations with selected e-print repositories, a focus group that will be held with the JISC Fair programme E-prints Cluster and general consultation with data services, archives and other digital support services

Tasks

1. Develop consultation interview schedule

2. Contact e-print repositories and conduct in-depth consultations
3. Develop questionnaire for focus group and general consultations
4. Arrange and hold focus group
5. Conduct additional general consultations by correspondence or phone

Deliverables

Deliverable: analysis of existing collection policies; identification of IPR and copyright issues

4. Metadata review

Description of Work

Utilising the deliverables from work packages 1 and 2, the Metadata Review will develop a set of recommendations for the preservation and administrative metadata requirements of e-prints, focusing on the metadata needed to ensure that an e-print can remain technically, legally and ethically accessible.

Tasks

1. Drawing on the deliverable from work package 1 and the properties of e-prints identified in work package 2, a simple outline requirement document for e-print metadata will be developed.
2. Existing metadata schemes for the preservation and administration of digital materials (e.g. those developed by UKOLN, OCLC/RLG, NEDLIB, DDI) will be reviewed in the light of the outline requirements drawn up
3. Recommendations for the use or modification of existing, or development of new, metadata schemes

Deliverables

Report on metadata requirements for the management and long term preservation of e-prints

5. Format review

Description of Work

Building on the deliverables from work packages 1 and 2, the Format Review will access the minimum requirements of a digital format suitable for holding the full range of content found in e-prints. These requirements will then be compared with the formats currently available that are or could be used for e-prints, and the preservation characteristics of these formats (particularly their stability, software independence and openness) in order to recommend preferred and non-preferred formats for e-prints.

Tasks

1. Consult Fair e-print cluster group and other JISC services: AHDS, DA etc.
2. Rank formats in terms of openness, stability, features and available software support

Deliverables

Deliverable: report on preferred and non-preferred formats for likely content of e-prints

6. Review of organisational models

Description of Work

Digital preservation activities are undertaken following a number of organisation models. These models will be examined with consideration to the requirements of an e-print repository highlighted particularly by deliverables for work packages 2 and 3, but also 3 and 4. The OAIS reference model will provide a framework for assessing existing organisational models Extensive consultation will be undertaken, with contact envisaged with Roquade, D-Space, DPC, JISC Preservation Focus**,** JISC Fair program E-print Cluster, CURL and Cilip, among others, and national services within the UK.

Tasks

1. Visit selected institutional repositories and national services, including internet journals e.g. internet archaeology
2. Phone/email consultation

Deliverables

Report on current models, including identification of infrastructure requirements

7. Development of e-print preservation life cycle and cost models

Description of Work

Reports and analyses from work packages 1-6, plus the OAIS reference model, outputs from the CEDARS project and other work, such as the Jones and Beagrie Digital Preservation workbook will be synthesised. A core description of the life-cycle of an e-print will be developed, and the affect of various cost models on this life-cycle examined

Tasks

1. Develop preservation scenarios for e-prints
2. Formulate life-cycle and cost models for e-prints

Deliverables

Report detailing scenarios and curatorial life cycle and cost models, including infrastructure requirements and development needs

8. Final Report

Description of Work

A comprehensive final report will be written, incorporating the deliverables for work packages 1 to 7. The report will cover the organisational, infrastructural and technical aspects of managing e-prints in order to ensure their cost effective long-term viability.

Tasks

1. Identify and address emergent issues when deliverables for work packages 1 to 7 are considered together
2. Develop implementation plan
3. Write conclusions
4. Write introduction and background
5. Write Executive Summary

Deliverables

Final report containing:

a) Executive summary and recommendations
b) Introduction and background
c) Preserving e-print content
    1. Properties of e-prints
    2. Overview of existing practices and procedures, including an overview of collections policies and procedures
    3. Recommendations for e-print collections policies, selection and retention criteria
    4. Recommendations for e-print preservation metadata
    5. Recommendations for preferred formats for e-prints
d) Infrastructure and development needs
    6. Institutional repositories
    7. National services
    8. Infrastructure scenarios, including life cycle management options, cost models and sustainability assessments
e) Conclusions
f) Proposed implementation plan