

SHERPA DP: Creating A Persistent Preservation Environment For Institutional Repositories

Introduction

This proposal is submitted by the Arts and Humanities Data Service (King's College London) and the University of Nottingham (lead partner in the SHERPA Project) under Strand Three of the JISC Circular 4/04 Call for Projects in Supporting Institutional Digital Preservation and Asset Management. Its purpose is to create a collaborative, shared preservation environment for the SHERPA project framed around the OAIS Reference Model. The project would bring together the SHERPA institutional repository systems with the preservation repository established by the Arts and Humanities Data Service to create an environment that fully addresses all the requirements of the different phases within the life cycle of digital information.

The collaborative model proposed will take advantage of the skills and expertise developed by the SHERPA development partners which includes the preservation expertise of the Arts and Humanities Data Service (AHDS). By extending this collaboration into a full preservation service the project removes from each individual institutional repository the burden of adding a preservation layer to their repository, and the need for them to seek to employ scarce preservation management skills and expertise. The project will investigate the business case for this model and seek to establish an economic cost model that could be used to ensure its long-term sustainability.

Establishment of the preservation environment will include investigation of the technical challenges, metadata requirements, administrative and workflow processes, and will encompass these within the OAIS reference Model. This will provide a rich set of reports for others to use, and a practical implementation of a preservation environment for SHERPA. The model and working processes that the project will develop and implement is intended to be transferable to other repositories and services; and would be available for other institutional repositories to join in the future.

In summary the Project will:

1. Use the OAIS reference model to develop a persistent preservation environment for the SHERPA consortium, assigning rights and responsibilities and establishing protocols and work flow processes that will ensure the long-term preservation of the repository content.
2. Explore the use of METS as the framework for packaging and transferring metadata held within the institutional repositories, including the preservation metadata created by the preservation service.
3. Establish a coordinated set of protocols and software to be implemented as a working preservation service for a group of institutional repositories.
4. Explore the use of open source software and tools to add functionality to and extend the storage layer of repository software applications.
5. Draw together the experience gained into a Digital Preservation User Guide that will complement the 'The Preservation Management of Digital Material Handbook' created by Maggie Jones and Neil Beagrie, and act as a practical user guide to implementing this type of preservation environment

The proposal will achieve several key objectives outlined in the JISC Preservation Strategy and in the JISC Circular 4/04 Call:

- Implement a preservation environment for several major institutional e-print repositories to ensure long-term preservation of their content

- Demonstrate a collaborative model using the OAIS reference model that brings together local repositories with national services
- Investigate the use of METS as a key element of the preservation environment, both as a metadata framework and as a transfer mechanism for data and metadata
- Investigate the use of open source and grid technologies as tools for the preservation process

Project Partners:

The lead organisation is the Arts and Humanities Data Service (a SHERPA Development Partner and part of King's College London) with the University of Nottingham, (the lead institution for the SHERPA Project) as the named project partner. The SHERPA project is funded by JISC and CURL under the FAIR Programme and aims to investigate issues to do with the future of scholarly communication and publishing. In particular, it is initiating the development of openly accessible institutional digital repositories of research output in a number of research universities. 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. Preservation activities include investigating the requirements for the long-term preservation of e-prints, including metadata requirements and economic models. This latter work is conducted by the Arts and Humanities Data Service.

The Arts and Humanities Data Service is a UK national service funded by the JISC and AHRB to collect, preserve and promote the electronic resources which result from research and teaching in the arts and humanities. By preserving collections the AHDS encourages further research and educational use of its collections. The identification and promotion of shared standards is critical to the AHDS's work. Preserving and exchanging digital information relies upon their widespread adoption and the AHDS endeavours to use open standards and software wherever possible, including in the establishment of its own digital repository. The AHDS seeks to work in fruitful partnerships in order to enhance the production and preservation of high-quality digital resources of whatever type, and to use its skills and expertise in preservation for the benefit of the HE and FE communities.

In addition to the funding already supplied to the current SHERPA Project, the Consortium of University Research Libraries (CURL) Board has agreed to contribute a further sum of £25,000 to this project to fund participation from their members who are either development or associate partners in the SHERPA Project. The AHDS and the University of Nottingham will work with the SHERPA Management Board and the CURL Board to establish criteria for the selection of a further three partners to work with the project. SHERPA project partners will then be invited to submit a bid to the SHERPA Management Board to join this project.

The selection criteria will ensure that both DSpace and Eprints repositories are represented in order to ensure that the project tackles the preservation issues of the two most commonly used repository software applications. The criteria will also require that partners are well advanced in the establishment of their repository infrastructure, and that there is a broad spread of SHERPA partners, including the associate partners. The successful bidders will be required to sign a partner agreement agreeing to contribute as required to the project and to abide by the requirements laid down by the JISC.

Project Governance

We propose that the current SHERPA Management Board be asked to act as the Management Board for this project to ensure the closest coordination between the original SHERPA project and this project. The Project will comply with JISC requirements and will report to the JISC Programme Manager as required.

Project Description

Institutional repositories are a new and high profile area, often feted as the future for disclosing to a wider public the research outputs of Higher Education. In recognition of this JISC has funded the establishment of a number of institutional repositories in the UK as part of the FAIR Programme. Thus far, the initial focus of activity has been on the process of establishing repositories – installing appropriate software and establishing policies and procedures; encouraging deposit of articles and dealing with the associated rights issues; and working to effect the cultural change needed for successful development and population of repositories.

Given the experimental and project-based nature of much of this activity, it is not surprising that less attention has been paid to preservation, and that no repository thus far established would claim to be ‘doing preservation’. Of course the SHERPA project has a specific remit to investigate the requirements for preservation and is producing some valuable outputs, but this falls far short of establishing a coherent and long-term preservation environment for the repositories involved in the project.

The recent JISC-funded Feasibility and Requirements Study for Preservation of E-Prints (James et al, 2003) argued that there is a unique window of opportunity to address the preservation requirements of repositories at the beginning of their adoption rather than leaving it until the lack of preservation management becomes an issue and content is no longer accessible. A key recommendation of the report was the establishment of a repository infrastructure based upon the OAIS reference model. It further recommended that this should be “a practical study that includes implementation at one or more repositories and their partners as appropriate to the chosen organisational model’ (p.56, James et al, 2003)

Furthermore, the Feasibility and Requirements Study identified the diverse range of skills and expertise required to manage and run a preservation environment based upon the OAIS Reference Model. In particular it noted the scarcity of staff and services with practical digital preservation skills and expertise. It therefore suggested that a sensible way forward would be to look to disaggregate the functions and activities identified in the OAIS Model, and to seek collaborative arrangements between repositories and specialist services with each taking responsibility for different functions.

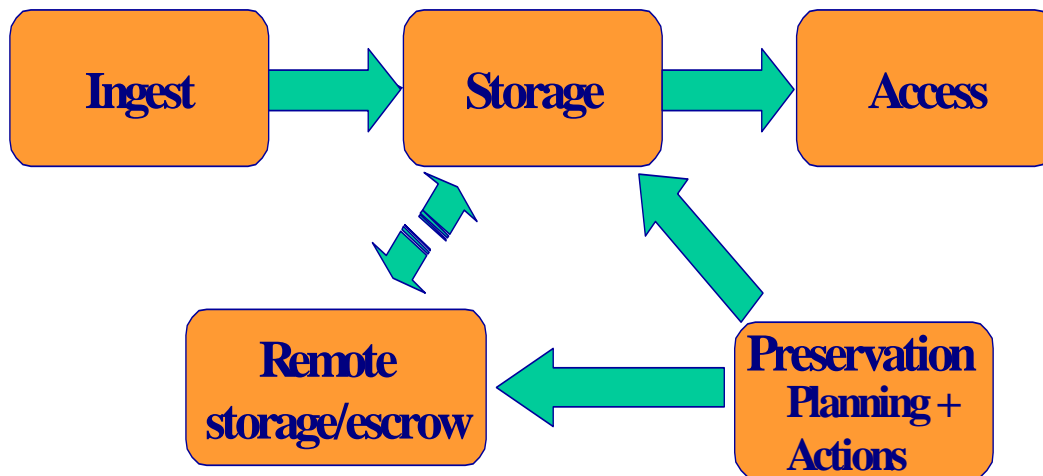
This recommendation fits well with that envisaged in the JISC Continuing Access and Digital Preservation Strategy 2002-2005. Beagrie writes:

“Institutional arrangements may benefit most from third-party or common services being developed to support preservation planning or remote storage.” (p.13)

We therefore propose to create a collaborative, shared preservation environment for the SHERPA project framed around the OAIS Reference Model. The model we are proposing is intended to take advantage of the pre-existing and successful collaboration between the SHERPA repositories and the Arts and Humanities Data Service.

The exact nature of the model to be adopted will be established at the start of the project, but at this early stage it is envisaged that the AHDS will provide a shared preservation store, and undertake preservation planning and preservation functions, whilst the SHERPA repositories will continue their work to raise awareness and promote deposit of content, ingest, storage of content for delivery, and access. To that end we are able to use the simplified model presented in the JISC Preservation Strategy to visualise how the model might look. The top layer would be the continuing responsibility of the institutional repositories, with this project adding the bottom layer, with the addition of preservation actions, through the collaboration with the AHDS.

Institutional Repository Layer (SHERPA Repositories)



Preservation Service Layer (Arts and Humanities Data Service)

Further specification of this model will use the OAIS Reference Model to define a functional model, including assigning rights and responsibilities for the different functions identified in OAIS, and to create the protocols and processes necessary for the implementation of a successful preservation environment. The project is especially interested in exploring the use of open source software and standards to implement the preservation environment, including METS and grid technologies.

Within this model each party will be required to provide an agreed level of functionality in order to ensure successful coordination and interoperability between the parties. Repositories are likely to provide the following functionality:

- Support for publishing metadata to be harvested
- One or more methods for transferring data (e-print content) across the network
- Alerting mechanisms for updated/additional content

The Preservation Service is likely to provide (or provide in collaboration with other preservation services e.g. the DCC) the following functionality:

- Support for harvesting metadata
- Support for harvesting data
- File format conversion tools
- File integrity checking tools
- Preservation metadata extraction tools
- File format obsolescence checking
- Alerting and migration service
- One or more methods for transferring data and metadata back into an institutional repositories

The challenge will be to do this successfully with the different repository software solutions chosen by SHERPA partners and taking into account the individual policies and approaches with regard to content and metadata. To this end the project will be investigating a variety of approaches for interoperating between the institutional repositories and the preservation repository services, and will test and evaluate each during the first phase of the project. A key part of the project will be to add functionality and to extend the storage layer of the Eprints and DSpace repository software applications to enable the necessary preservation actions to take place. Following the testing and evaluation process the chosen solutions will be

implemented. It may well be the case that different solutions are chosen for different institutions and for different repository software.

Project Work Packages: The project will be carried out over a period of two years and comprise of the following activities:

Work Package 1: Project Management; Duration: 24 months;

Months 1-24

This work package will act to manage and coordinate the activities of the partners, to prepare and report as required, and to assess risks and opportunities as the project progresses. This will include drafting and finalising a partners agreement to regulate rights and responsibilities of all partners in a disaggregated model. This work package will also undertake the gathering of time sheets and conduct activity based costings in order to develop a business case for continuation of the process.

Tasks:

1. Develop a detailed work plan with timescales, deliverables, and milestones
2. Develop partner agreement, manage and coordinate activities of the partners and appropriate dissemination activities
3. Monitor progress and identify corrective actions in case of deviation from planned activities, and ensure that the project maintains its schedule
4. Participate in project meetings and project reviews with the SHERPA Management Board and other relevant committees
5. Prepare periodic management reports
6. Develop the business case and cost model

Deliverables:

1. Detailed work plan
2. Progress and risk assessment reports
3. Partner agreement
4. Website and dissemination activities
5. Business case and cost model

Work Package 2: Infrastructure: applying the OAIS Reference Model; Duration 18 months;

Months 1-6 Produce first draft

Months 7-18 Review and refine

This work package will apply the OAIS reference model to the distributed institutional repository / single preservation repository infrastructure proposed by this project. This would identify rights and responsibilities, services and actions and apportion these between the institutional repositories and the preservation repository service. Included would be an investigation of the role (and also act to identify services that might be provided) that the DCC might play. Advice on applying the model would also be sought from the DPC and the DCC and the results disseminated and promoted through both these channels, in addition to the SHERPA project website and the AHDS.

Tasks:

1. Review the OAIS model and its component parts
2. Assign tasks and services to the institutional repositories and the preservation repository
3. Assign rights and responsibilities
4. Investigate role of the DCC and agree services to be supplied through them
5. Liaise with DCC and DPC

Deliverables:

1. Draft report and OAIS SHERPA Preservation Environment infrastructure
2. Final report and OAIS SHERPA Preservation Environment infrastructure

Work Package 3: Digital Repository Handbook; Duration 12 months;

Months 13-18: gather together outputs and document experiences;

Months 18-24: write up as on-line handbook/user guide

This work package is designed to bring together the experiences and outputs from this project, from the first SHERPA project, and from other sources as appropriate, into a practical 'user guide'. The Handbook will recommend standards, best practice, protocol and processes that might be used in the management, preservation and presentation of e-print repositories, and will provide the practical experience of both the SHERPA projects written up as a case study. A detailed specification of the Handbook will be drawn up and circulated for comment.

Tasks:

1. Gather together and review relevant outputs from projects, gather experience of SHERPA projects
2. Draw up detailed specification of the Handbook for comment
3. Edit and compile into an on-line User Guide
4. Circulate for review and comment
5. Publish

Deliverables:

1. Handbook specification
2. Completed Handbook available as a web publication

Work Package 4: Metadata and METS; Duration: 6 months

Months 7-9 METS Framework

Months 10-12 Metadata set

This work package has two components: the first is to review the existing metadata held in the SHERPA repositories and to establish if additional metadata needs to be collected or added. In particular this will address if the administrative metadata is sufficient, and if not, what might be added and how; and will also define and agree a preservation metadata set, based upon the recommendations in the Feasibility and Requirements Study. The second component will be to investigate the use of METS as the framework for combining and packaging metadata, and as a transfer mechanism for metadata and e-print content.

Tasks:

1. Review the use of the METS framework within the SHERPA preservation environment
2. Review metadata contained in the institutional repositories
3. Metadata set agreed as the minimum requirement to ensure long-term access and preservation of the repository content

Deliverables:

1. Report on the use of the METS framework
2. Minimum metadata set agreed for SHERPA institutional repositories

Work Package 5: Repository archiving; Duration 24 months

Months 1-12 Tasks 1-4

Months 12-24 Task 5

This package will investigate and implement automated networked transfers of data and metadata between the Dspace/Eprints self-archiving systems used to set-up institutional repositories and the AHDS preservation repository. The aim is to enable automatic synchronisation of data and metadata resources with a remote preservation repository in order to enable resources to be preserved and maintained within an OAIS framework. Solutions to this problem at several different functional levels will be investigated.

Firstly we will investigate implementing a common grid-enabled storage layer infrastructure. Work is already under way within the Dspace project to enable SRB as a storage medium. We will review this work and look at the Eprints capabilities in this area. The central question to be answered is whether a preservation repository partner can access stored objects in a shared

data grid medium to independently undertake preservation management actions and procedures.

An alternative approach will involve review of public API based mechanisms for transfer of data and synchronisation of Dspace/Eprints systems with the preservation repository. The question will be whether a preservation repository partner can automatically obtain information and resources by harvesting in a secure, manageable and sustainable fashion. One approach might be to look at the existing OAI-PMH API implemented by both Dspace and Eprints and use it for direct transfer of data resources using METS wrappers. A further approach might be to establish harvesting using web crawling mechanisms. The other will be to design and scope add-on modules for transferring data and synchronisation information (for example check sums).

The final approach will be to investigate ancillary services which might be implemented outside the Dspace and Eprint repository environments to enable synchronisation with a preservation repository. At a basic level this could involve procedures for the regular export of data and bulk upload to the remote repository.

All three approaches will involve review of the existing system capabilities, implementing prototype modular solutions and testing them with the institutional repository partners. We anticipate that we will assess solutions with respect to the speed of operation, ease of maintenance, volume of data transfers, security and integration with user management systems.

Tasks:

1. Review of Dspace and Eprint APIs, storage layers and module add-on capabilities.
2. Prototype implementation and testing of SRB as a common storage medium
3. Prototype implementation and testing of API based access mechanisms
4. Prototype implementation and testing of external synchronisation mechanisms
5. Choose, design, implement, test final solution

Deliverables:

1. Review report for task 1
2. Assessment report for tasks 2-4
3. Documentation from task 5

Work Package 6: Preservation Actions; Duration 12 months;

Months 13-15 investigate approach one

Months 16-18 investigate approach two

Months 18-24 Implementation

This package will investigate the processes required to enable changes and updates to the self-archived materials in institutional repositories which will ensure their long-term integrity and preservation. Two different processes will be investigated.

The first approach will be to investigate mechanisms for integrity and security checking. At its most basic level the preservation repository acts as a mirror and back up store. This process will allow the preservation repository to capitalise on this central role by creating alerting or reporting services related to previous archive copies and any changes detected in the structure of the archive or its constituent objects. Checks on the latter would include internal checks on object integrity and format, together with checks on metadata content and the links between objects and metadata records.

The second approach will be to investigate procedures for obsolescence checking and automatic migration or updating. At its most basic level this could be a reporting or alerting service which summarises the state of deposited objects of different classes and any necessary migration actions. It is proposed that there would be a check against a software format

registry together with linking to any available data migration services or knowledge bases. At a higher level we shall investigate the scope for automated updating, transfer and re-versioning of self-archived objects. This process will need to be assessed for its level of sustainability and implications for the security mechanisms of the target institutional repository system.

Tasks:

1. Create repository integrity checking and reporting services
2. Create repository obsolescence checking, reporting and migration services
3. Investigate remote alerting service capabilities
4. Investigate mechanisms for automatic creation of new versions, or migration and re-deposit.

Deliverables:

1. Report on approach one
2. Report on approach two
3. Recommendations for implementation

Work Package 7: Implementation

Implement preservation infrastructure. Implementation will take place during the last 6-9 months of the project and a detailed work package will be assembled following the successful completion of work packages 2, 4, 5 and 6.

Risks

The project is designed to minimise risk. A high level of coordination with the current SHERPA project and the SHERPA Management Board will ensure good oversight of the project. The reports to the Board will include a progress report and an assessment of any risk to the project. In order to minimise risk a key part of the project is the evaluation and testing of a number of approaches, all of which are promising in different ways. Rather than choosing a single solution at the start only to find it doesn't function as well as it might, the project will assess a number of different approaches (as outlined in the work packages) and choose the best solution for the particular circumstances at each institution and for each particular repository software application.

As with most projects, staff loss is always a key risk factor. However, the nature of the consortium minimises this risk in so far as it is possible to do so. The spread of expertise and skills should ensure that the project keeps its momentum until project staff can be replaced.

The project will also address the issue of sustainability. An exit strategy will be developed by the partners based upon a cost model developed during the latter stages of the project. During the final implementation stage staff will be required to keep time sheets to enable activity based costings to be gathered and a cost model will be drawn up.

Value to the Community

Several items of value to the wider community will emerge from this project:

- SHERPA persistent preservation environment established contributing towards the strategic objectives of the JISC
- A series of evaluation reports including using METS, using open source software, assessment and evaluation of a variety of preservation approaches
- A Case Study report
- A Digital Preservation Handbook User Guide that brings together the outputs from the first SHERPA Project, the experience and outputs from this project, potentially outputs from other FAIR projects, and other relevant material into a practical, focused user guide that will assist others to establish preservation environments, or to join the environment established by the project.

Dissemination and Promotion

A key element of the project will be to share the experience, evaluations and results with the wider community. Dissemination will be an on-going activity throughout the project and will take place through a variety of mechanisms. In particular the project will use the following dissemination pathways:

- SHERPA website: the website will be extended to include this project
- AHDS Website: a section will be added to the AHDS for the results and outcomes of this project
- Digital Preservation Coalition: the project will develop an ongoing relationship with the DCC and seek to disseminate its outcomes through the DPC Website. The possibility of a workshop co-hosted with the DPC will also be explored. The DPC will also be used to circulate drafts to the community for comment.
- Digital Curation Centre: the project will develop an ongoing relationship with the DCC and seek to disseminate its outcomes through the DCC website. The DCC will also be used to circulate drafts to the community for comment.
- Email announcements: important milestones and deliverables will be announced on appropriate discussion lists
- Publications: articles will be written for publication in appropriate journals
- Conference presentations: various key conferences will be targeted for dissemination opportunities

Dissemination will take place at an international level in addition to the UK.

Evaluation

Evaluation will be an on-going process throughout the lifetime of the project. The Project partners and the Management Board will evaluate the deliverables and the milestones. A testing process is built in to the work packages to ensure vigorous testing of the project deliverables and software solutions. The implementation process will also produce an evaluation report from the project partners. In addition, the project proposes establishing a 'virtual' evaluation committee with members drawn from the DPC and DCC communities, who will evaluate the various outputs from the project and provide comment and recommendations to the project partners.

IPR will remain with the project partners, and will be specified in the partner agreement. All results and outputs will be freely disseminated and available for use by the HE and FE communities.