SHERPA DP: Establishing an OAIS-compliant Preservation Environment for Institutional Repositories

Gareth Knight

Arts and Humanities Data Service, 26 - 29 Drury Lane, 3rd Floor
LONDON, WC2B 5RL
gareth.knight@ahds.ac.uk

Abstract. Institutional repositories are considered an important method of capturing research outputs and disseminating them to a wider public. Thus far, the initial focus of activity has been on the process of establishing repositories and handling issues associated with the deposit of academic research. Given the experimental nature of many of these projects, few are able to dedicate staff or funding to the long-term preservation of e-prints. This paper outlines a potential solution, currently in the early stages of investigation, which establishes a shared preservation environment through the outsourcing of essential preservation activities to a specialist service with expertise the field.

Introduction

Institutional repositories are a new and high profile area, often feted as providing a valuable complement to existing scholarly publishing models, and allowing institutions to disclose their research outputs to a wider audience. In recognition of this, the JISC (Joint Information Systems Committee) funded a number of projects, including the SHERPA (Securing a Hybrid Environment for Research Preservation and Access) Consortium, as a method of establishing repositories within higher education institutions. The project has been a success, resulting in the establishment of repositories in 20 partner institutions and the production of a significant body of research on the creation, population and maintenance of e-print collections.

With the establishment of institutional repositories, there is a growing awareness that the academic community should consider the need for digital preservation, to ensure that academic research in the form of e-prints and e-theses, deposited within these repositories remain accessible and offer a guarantee of integrity in the long-term. The recent JISC-funded *Feasibility and Requirements Study for Preservation of E-Prints* [1] 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. However, they note that the funding model currently attached to institutional repositories does not cater for the need for digital preservation, and therefore staff or services with practical skills in this area are absent.

The SHERPA DP project has been funded under the JISC 4/04 call for Projects in Supporting Institutional Digital Preservation and Asset Management, and officially began in March 2005. It aims to test a possible solution through the creation of a collaborative, shared preservation environment for the SHERPA project framed around the OAIS Reference Model. The project is being led by the Arts and Humanities Data Service, with the University of Nottingham, and four project partners – Glasgow ePrints, Edinburgh Research Archive, Nottingham ePrints, London Leap and the White Rose University Consortium (Leeds, Sheffield, York) – who will serve as a testbed for implementation of the service within their existing EPrints/DSpace repositories.

In order to develop a sustainable model for preservation, the AHDS will work with the project partners, to investigate four key areas:

1. Explore the use of METS (Metadata Encoding and Transmission Standard) as a metadata framework and a mechanism to transfer data and metadata between the institutional repository and the AHDS.
2. The development of a metadata scheme to describe the preservation of e-prints, based upon the recommendations of PREMIS and other working groups.
3. An investigation into the APIs provided by EPrints and DSpace software for connecting to these repositories.
4. Identification of specific mechanisms and tools for maintaining the integrity, fixity and security of repository data and performing obsolescence checks and automated migration.

In this paper, I will outline how the OAIS reference model may be applied to the SHERPA DP disaggregated approach, indicating the infrastructure that will need to be developed and possible scenarios for preservation.

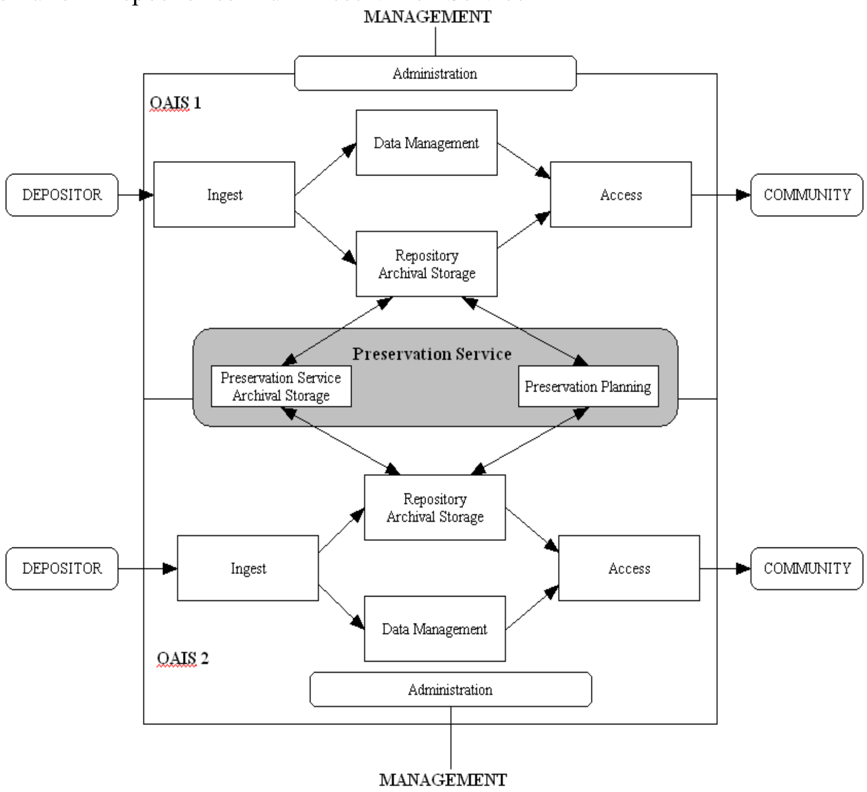
Application of the OAIS model

The Open Access Information System (OAIS) is a high-level reference model that provides a common language and layout for the definition of "*an archive, consisting of an organization of people and systems, that has accepted the responsibility to preserve information and make it available for a Designated Community*" CCSDS [2]. As a reference model, it does not imply a specific design or formal method of implementation. Instead, it outlines six entities (Ingest; Archival Storage; Data Management; Administration; Preservation Planning; and Access) that perform specific activities within the repository. It is left as an exercise to the reader to develop their own model for OAIS compliance by analysing existing processes and matching them to each function.

The model proposed for the SHERPA DP project may be compared to a '*Shared Service'* in the OAIS terminology, or more accurately described as a '*Repository with outsourced Preservation Services*' [1]. However, the disaggregated layout make it distinct from existing OAIS models and the Functional Model must be modified to provision of a second Archival Store. The storage area may be considered a "dark archive" to be used for preservation and back-up of Information Packages held at the Institutional Repository, and no public interface or access will be provided.

The OAIS environment is surrounded by three conceptual groups - the Producer, Management and the Consumer – that perform specific roles in the submission, management and dissemination of an information package. To consider how these roles may apply to the workflow of an institutional repository, it may be beneficial to replace these terms with 'Depositor' (an author or authorized individual who deposits an e-print); 'Repository Management' and 'Preservation Service Management'; and the Community' (most likely to be university students, university staff, and researchers).
Figure 1, adapted from figure 4-1 in the OAIS reference model, maps the OAIS terminology onto the operation of the SHERPA DP model

Figure 1. A functional model that illustrates high-level interactions between two Institutional repositories and a Preservation Service



The primary object within the data flow is an ‘Information Package’ that undergoes various changes between the point of submission and the dissemination. An Information package is essentially a conceptual object within the OAIS model composed of ‘Content Information’ (CI) and associated ‘Preservation Description Information’ (PDI). In practical terms, an Information Package in an institutional repository is likely to consist of an e-print, stored in an appropriate file format (e.g. PDF, RTF, plain text) and a METS record that will contain resource discovery (Descriptive Information) metadata. Further preservation metadata that identify the technical, administrative and provenance of an e-print will be created at varying stages of the ingest and archival process and may be used to audit and authenticate the resource.

To the Depositor and Community, the actions required during the Ingest and Access stages remain the same – the Depositor completes a licence form and deposit a Submission Information Package (SIP), composed of an e-print and associated resource discovery metadata. Repository staff perform basic validation to ensure the metadata meet their internal requirements and an Archival Information Package (AIP) is written to the Archival Storage area. Based upon some action pre-agreed event (scheduled harvest or notification) the Preservation Service transfers the AIP to the Preservation Storage Area and relevant preservation procedures (migration, creation of metadata) will be performed. The modified AIP will be returned to the Archival Storage Area and a dissemination copy, in an appropriate file format, will be generated for use by the User Community.

Implementation of the disaggregated model

In order to implement the model, procedures and software will need to be created or repurposed to manage the connection between the repository Archival Store and the AHDS. Institutional repositories participating within the project currently implement DSpace or EPrints repository software, or a combination of both [3]. Both products are firmly established within the UK educational sector and provide integrated methods of categorising e-prints within the repository and support the harvesting of metadata through OAI-PMH. However, it is not currently in their remit to provide a method for alerting a preservation service when information content requires preservation or a method of returning an Archival Information Package back to the institutional repository. Further investigation is necessary to identify a method of connecting to an institutional repository, either through the implementation of a grid-enabled service [4], the use of complex object formats (METS, MPEG-21 DIDL) [5] or the development of an ancillary service that may be implemented outside the current repository environments.

Scenarios for the harvest and migration of content information

A key decision to be made within the project is the identification of when the Information Package should be transferred to AHDS and who should be responsible for initiating the process. Three scenarios that identify different roles and responsibilities for the Institutional Repository and Preservation Service may be considered.

1. **Harvest and migrate the Content Information on Ingest**
   The Institutional Repository notifies the Preservation Service when a Submission Information Package (SIP) is deposited and a software tool that harvests the Information Package, migrates the e-print to a chosen preservation format and generates appropriate technical and provenance metadata.
   The harvest-on-ingest approach may be beneficial to a Preservation Service that must support several institutional repositories – a scenario that may otherwise require significant resources to monitor for new submissions. Mechanisms must exist that allow the Institutional Repository to notify the Preservation Service when an e-print has been deposited.

2. **Harvest and migrate the Content Information on a regular schedule**
   The Preservation Service harvest recently deposited SIPs (e.g. through OAI-PMH SETS), based upon a schedule (e.g. monthly) agreed between the two parties. An Archival Information Package is generated and resubmitted into the repository storage archive.
   Unlike the alternative approaches, the scheduled harvest approach will not require a method of notifying the other party or monitoring the repository archival store. It may be appropriate for Institutional Repositories/Preservation Services that do not possess the technical infrastructure or staff to implement and support a harvest-on-demand system. However, there may be implications for the repository workflow: there may be a time period of several weeks between the date of deposit and date of harvest by the Preservation Service.

3. **Harvest and Migrate Content Information when it is considered at-risk**
   The Institutional Repository ingests the SIP and generates a dissemination copy from the submitted e-print. The Preservation Service monitors the Content Information and migrate the e-print when the file format is considered to be at-risk of being rendered inaccessible.
   In practical terms, it is not possible to monitor all file formats held by each institutional repository, and so, an automated system, such as the National Archives' PRONOM database [6] may be an option for identifying obsolescence. Use of PRONOM, combined with the projected outputs of the Digital Asset Assessment Tool project, also funded by JISC and currently underway [7], is likely to provide a powerful approach for identifying at-risk file formats held within digital repositories.
   This approach will alter the workflow of the Preservation Service, allowing them to perform a single batch process to convert content in a common format. It may

benefit the preservation of file formats that contain particular functionality that cannot be easily replicated using preservation formats by off-setting migration until a later date.

Conclusion

The practical implementation of an OAIS-compliant persistent preservation environment will offer significant benefits for the institutional repositories funded under the SHERPA project, and should provide useful model that may be adopted by the user community as a whole. The challenge will be to do this successfully with the different repository software solutions and to take into account the individual policies and approaches chosen by SHERPA partners.

References

1. James, H. Ruusalepp, R. Anderson, S. and Pinfield, S: Feasibility and Requirements. Study on Preservation of E-Prints. JISC. 2003. http://www.jisc.ac.uk/uploaded_documents/e-prints_report_final.pdf
2. OAIS: Reference Model for an Open Archival Information System (OAIS). Blue Book. Issue 1. January 2002. http://www.ccsds.org/documents/650x0b1.pdf
3. Nixon, W. DAEDALUS: Initial experiences with EPrints and DSpace at the University of Glasgow. Ariadne Issue 37. October 2003. http://www.ariadne.ac.uk/issue37/nixon/intro.html
4. UCSD Libraries: DSpace/SRB Integration (2004). http://libnet.ucsd.edu/nara/
5. Van de Sompel, H. Nelson, M.L. Lagoze, C. Warner, S: Resource Harvesting within the OAI-PMH Framework. in D-Lib Magazine. December 2004. http://www.dlib.org/dlib/december04/vandesompel/12vandesompel.html
6. The National Archives: PRONOM – The File Format Registry. 2005. http://www.nationalarchives.gov.uk/pronom/
7. Ashley, K: DAAT – Digital Asset Assessment Tool. JISC. 2004 http://www.jisc.ac.uk/index.cfm?name=project_daat&src=alpha