

The International Journal of Digital Curation

Issue 1, Volume 2 | 2007

Modelling OAIS Compliance for Disaggregated Preservation Services

Gareth Knight, Mark Hedges,
SHERPA DP Project, AHDS Executive.

June 2007

Abstract

The reference model for the Open Archival Information System (OAIS) is well established in the research community as a method of modelling the functions of a digital repository and as a basis in which to frame digital curation and preservation issues. In reference to the 5th anniversary review of the OAIS, it is timely to consider how it may be interpreted by an institutional repository. The paper examines methods of sharing essential functions and requirements of an OAIS between two or more institutions, outlining the practical considerations of outsourcing. It also details the approach taken by the SHERPA DP Project to introduce a disaggregated service model for institutional repositories that wish to implement preservation services.

Introduction

Institutional repositories (IRs) represent a particular form of digital archive that is implemented for use within an institutional setting. The purpose of a repository, within the wider context of the organisation, is to capture and make available the research output of an institution (i.e. a university). This may include materials such as research papers, student theses, presentations, or other digital assets. Given the importance of actively advocating the use of the repository, it is not surprising that less attention has been paid to preservation. Several reasons have been identified for this omission, including the need to embed repositories into the institutional infrastructure before considering additional issues, the limited funding available to institutional repositories, and the absence of staff and services with practical experience of preservation issues (James, Ruusalepp, Anderson, & Pinfield, [2003](#)).

This paper provides an overview of how the OAIS may be applied to multi-institution configurations. Through the development of a disaggregated service model that allocates core components of the OAIS to different institutions, digital repositories may benefit from the provision of services that they could not perform in isolation or which do not fit into their core funding. We describe the work undertaken by the AHDS to provide preservation services to institutional repositories participating in the SHERPA DP Project.

OAIS as a Repository Framework

The OAIS (Open Archival Information System) has proven useful as a high-level model within which to frame the structural organisation of a repository. The conceptual framework serves as a community- and technology-independent model that defines the core components of a repository, including the people and automated systems necessary to manage digital content in the long-term and make it available to the user community. A key advantage of the OAIS reference model is its emphasis on abstract design and subject-independent terminology. It offers significant flexibility to system designers wishing to map the OAIS to their own repository, enabling them to interpret its use in a manner that is relevant to their field of expertise and content type. Any repository that actively accepts a digital resource, stores and manages it in a controlled environment, and makes it available to an end user can claim to be OAIS-compliant. It has also underpinned subsequent work by the RLG and OCLC on the responsibilities of a trusted digital repository and the role of certification (RLG/OCLC, [2002](#)). The PREMIS implementer group's investigation into the requirements of descriptive information appropriate for preservation has also made extensive use of it.

The active discussion of OAIS compliance in the research community has resulted in a diverse range of institutions implementing the model. Methods of OAIS implementation vary significantly, according to organizational structure, funding level, availability of staff, software infrastructure and subject domain. An OAIS-compliant repository may be organized in a number of ways appropriate to circumstances:

1. A repository operated by one department in an institution, e.g. the Archaeology Data Service.
2. A repository operated by two or more departments in an institution, e.g. the British Atmospheric Data Centre.
3. A repository operated by one department in many institutions, e.g. the Arts

& Humanities Data Service

4. A repository operated by two or more departments in many institutions, e.g. UK Data Archive

Each of the above repository structures may provide ‘end-to-end’ services, performing actions necessary to accept submitted data, perform appropriate action necessary to manage the data, and make it available to the designated community. However, the organizational structure necessary to manage the process and workload required is likely to differ. A repository that operates in a single department will be required to perform all actions necessary to ingest, manage, preserve and distribute the digital resource. The second and third organizational model, in contrast, allocates key components to different individuals, departments, or institutions. For example, one department may be responsible for producing archival and dissemination copies of an information package, while a second department is responsible for management of the technical systems. To ensure that the institutions can perform the tasks required, the technical and managerial components must exist to facilitate co-operation.

Existing Institutions and Projects That Have Implemented a Disaggregated Service Structure

The notion of many different institutions co-operating to deliver a shared service is not a new idea. The Arts & Humanities Data Service was established in 1996 on the basis that departments in universities around the country would establish subject centres that would provide expert knowledge on a particular subject area. However, technological changes in the last 5-10 years have enabled different types of interaction, enabling data to be transported and managed in many ways. At a system level, the process has been simplified by effort on the part of repository software developers to provide frameworks that may be mapped to appropriate OAIS terminology¹. Several permutations of the OAIS model may be identified, that establish different methods to manage the Ingest, Data Management, Archival Store, Preservation Planning, Administration and Access functions. The projects described below have similar requirements, but define how services should be allocated on a case-by-case basis and, therefore, do not fit into simple categories.

- **Shared Access functions** – Consistent implementation of discovery facilities, such as federated searching or OAI metadata harvesting by repository implementers may be considered a well-established method that has enabled institutions to co-operate, enabling local and global communities to search resources. By actively sharing metadata, repository content may be noticed and used by researchers in a range of subject areas, beyond the boundaries of the OAIS ‘designated community’.
- **Shared Archival Storage/Administration** – Several institutions are investigating technologies to enable the OAIS ‘archival store’ to be distributed at different locations. The San Diego Super Computer Centre (SDSC), University of Maryland, and the National Archives and Records Administration (NARA) collaborated to build a persistent digital archive, located at the three sites, each running different database management software connected through

¹ For example, the OAIS Archival Information Package (AIP), Content Information and Content Data Object may be mapped to the Fedora Object XML (FOXML) document, Fedora digital object and Fedora datastream, as well as to the DSpace 2.0 METS document, DSpace item and DSpace bitstream (Bekaert & Van de Sompel, 2006).

the Storage Resource Broker (SRB) middleware. Through an abstraction layer, a user (OAIS Consumer) located at one of the three sites would be able to search and locate resources (Smorul et al., 2003). A similar approach is being taken by the British Atmospheric Data Centre (BADC), which is using Atlas Petabyte Storage Services (APS) as its long-term storage system. When considered in combination, these two facilities fulfill the primary requirements of the OAIS. The BADC fulfills the OAIS requirements of ingest, local storage, preservation, and access to the Consumer, while the APS performs the function of archival store. Jointly, the BADC and APS fulfill the requirements of Data Management and Administration (Corney et al, 2004).

- **Shared Data Management/Administration** - The Cornell University Library (CUL) and Göttingen State and University Library (SUB) are collaborating to develop a long-term preservation repository for digital journals. The project is building an interoperable implementation of the OAIS administration function that enables both institutions to administer the repository system, verify content and repair any faults (Rosenkrantz, 2003).
- **Shared Preservation Planning/Archival Storage** – The Florida Center for Library Automation (FCLA) has a relationship with publicly funded colleges and universities in Florida, whereby it is responsible for the preservation of digital assets. In a possible mapping to the OAIS model, the libraries may perform Ingest, Data Management and Access functions, while the FCLA is responsible for the essential requirements of Archival Storage and Preservation Planning (Thibodeau, 2006). The San Diego Supercomputer Centre (SDSC) Persistent Archives Testbed Project (SDSC, 2004) has established a similar relationship, using Storage Resource Broker (SRB) to provide preservation services for several institutions.

The impetus to establish a distributed service between two or more partners is unique to each institution. However, some general principles may be identified. A consistent and clearly defined understanding must be established detailing the nature of the co-operation. This must detail the task to be performed, the technical and organizational environment, and the length of contract. Additionally, technical services must be available to support the co-operation required by the partnership.

Establishing Preservation Services for Institutional Repositories: The SHERPA DP Project

The SHERPA DP Project is investigating a disaggregated model for the provision of preservation services to institutional repositories. The project is lead by the Arts & Humanities Data Service, working with five partners – the University of Edinburgh, University of Glasgow, University of Nottingham, White Rose Consortium and London Leap – that operate as a mix of single and multi-institution repositories running the ePrints and DSpace repository software.

Assessing the Need for a Disaggregated Preservation Service

For a repository, with a commitment to preserving the academic research of its institution, the need to be seen to comply with the mandatory requirements of the OAIS reference model may be considered a convincing argument to establish a partnership with another institution capable of providing particular services, particularly if it is required for certification as a Trusted Digital Repository (RLG,

2006). Repository staff are often occupied with the task of advocating its use in an institution, handling submitted data and performing regular maintenance. In many circumstances, they do not have sufficient time to actively manage the preservation process. The argument is supported by Lavoie and Dempsey (2004), who suggest “*long-term stewardship may be beyond the means of an individual institution*”.

The disaggregated services model represents a particular method of implementing preservation functions within smaller institutional repositories, allowing a mixture of institutional and community-wide services. As a side effect of the process, several additional benefits may be identified: the standardization of preservation practices across multiple repositories; reduction in possible duplication of effort by different repositories; automation of management services that would be unfeasible for the preservation of data held by smaller repositories; as well as some limitation on the funding requirements that an oversight body (e.g. JISC or other funders) must provide for institutions to perform the same actions.

A Disaggregated OAIS Model for Preservation Services

The disaggregated model in use by the SHERPA DP Project may be described as a modified version of the co-operating archive’ model in OAIS terminology, or more accurately a ‘Repository with outsourced Preservation services’ (Bekaert & Van de Sompel, 2006). It is composed of two types of institution – a Content Provider and a Service Provider that maintain a formal relationship, most likely a contract that specifies the type of work they will perform. These may be supported by additional services that are provided for use by other institutions. In combination, the institutions fulfil the requirements of the OAIS that could not otherwise have been performed in isolation. The relationship may be mapped using several methods. Hitchcock, Brody, Hey, and Carr (2007) define the relationship between the two institutions as a single OAIS that shares common functions. Knight (2006) adopts a detailed model, depicting the relationship as two incomplete OAISs that operate a closely linked workflow. Figure 1, adopted from figure 4-1 in the OAIS reference model, demonstrates the latter, mapping the functions of SHERPA DP onto the OAIS ‘shared services’ model.

For SHERPA DP, e-print institutional repositories participating in the project fulfil the functions of Content Provider, while the AHDS Preservation Service operates as a Service Provider. In isolation, institutional repositories fulfil many of the core functions of an OAIS: as content providers, they are able to accept and ingest digital objects (Ingest); manage the objects in a controlled environment (Data Management); and make them available to the user community in a suitable format (Access). However, many institutional repositories are unable or unwilling to perform managed preservation action (Preservation Planning and Archival Storage) necessary to ensure the longevity of digital content over time. To resolve the discrepancy between repository configuration and the OAIS model, the Arts & Humanities Data Service provide a “dark archive”, taking responsibility for aspects of the Ingest, Archival Storage, Data Management and Administration functions. Data will be stored in a “dark archive” and appropriate preservation action will be performed.

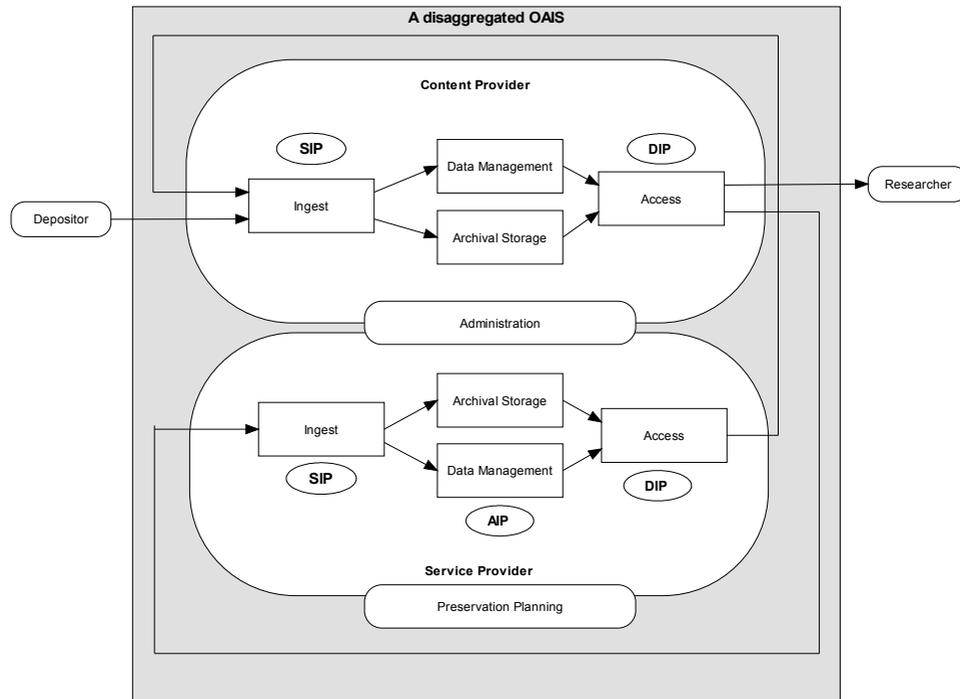


Figure 1 An OAI-compliant model for the provision of outsourced preservation services

Overall System Architecture

The OAI is a conceptual framework that may be used to inform the design of system architecture. However, a discussion of OAI compliance must relate to the practicalities of establishing a distributed repositories model. The approach taken during the project was to develop an architecture that would allow the Service Provider to perform preservation functions in a managed environment. Although some modification to the Content Providers as clients would need to be made to allow successful export of the required data, most development focused on the preservation repository.

The Content Providers that participated in the SHERPA DP Project may be categorized as institutional archives that store and make available electronic research data, such as pre-prints, post-prints, electronic theses and other types of research paper. In most circumstances, they have been in operation for a number of years and have adopted repository software and practices appropriate to the type of data that they store and the research community that they serve. The majority of the repositories operate the EPrints v2.x software ([n.d.](#)) developed by the University of Southampton. An exception is the Edinburgh Research Archive, which is using DSpace ([n.d.](#)). These repositories are tailored to the requirements of the research community – they operate workflow processes that allow users to submit research papers and make them available in a short time frame. However, they do not perform activities necessary to manage and preserve the research data in the long term.

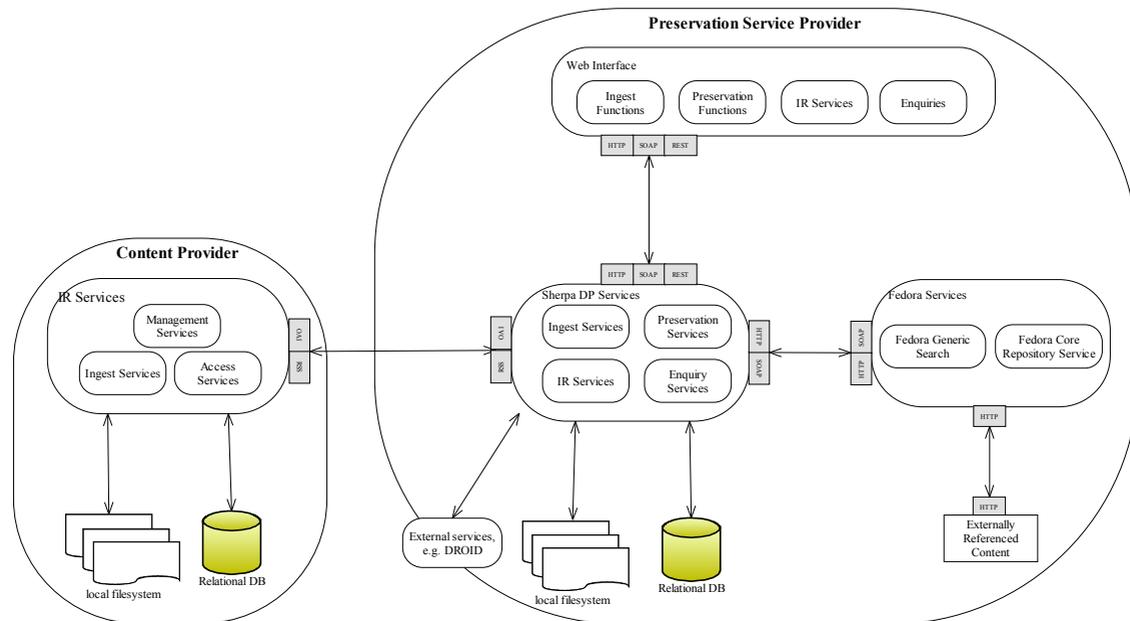


Figure 2 A simple system architecture of the SHERPA DP preservation system

The Preservation Service Provider (Figure 2) that serves as the basis of the project consists of repository management software, supported by a set of atomic web services. The Fedora management system serves as the kernel of the preservation repository, operated in a Xen virtual machine and supported by a CX300 SAN storage system. Fedora was selected for its flexible data model, which allows arbitrary data and metadata to be associated with an object as datastreams. It is also capable of representing a diverse set of relationship information, which is essential for management of a resource, through built-in support for RDF. Although Fedora supplies the core repository functionality, it does not implement specific ingest and management services required by the project. Instead, it incorporates the Fedora Service Framework, which may be used by developers to build and integrate appropriate services intended to perform particular functionality. These services run as atomic modular web applications that are independent, but interact with the core repository service, providing additional functionality that facilitates the integration of the basic repository kernel into a broader application environment. These additional services interact with the Fedora repository via SOAP, and in turn provide services that are made available via SOAP. They fall into two broad categories:

1. Services that perform preservation-related processing, which are used primarily by the preservation repository. These include services to generate preservation (e.g. technical provenance, etc.), to normalise files that are not in a format suitable for preservation, and to migrate files held in an obsolescent format. These services call upon external registries, such as PRONOM and GDFR, to make best use of their expertise and to decouple processing control from domain-specific knowledge about file formats.
2. Web-based services that may be used by institutional repository staff to make enquires, request reports, order replacement Dissemination Information Packages (DIPs), or other services.

To reduce demands on the repository administrators' time, it is possible to invoke web services automatically, either by a timer or as part of a workflow that is activated automatically by events such as the capture of a new object. In addition, user access to the services is facilitated by a lightweight web interface, which communicates with the service layer using the SOAP interface.

Enabling Co-operation Between the Content and Service Provider

The co-operation of two digital repositories that operate different software is problematic, raising numerous issues that must be addressed. Notable issues include:

1. The method of enabling machine-to-machine transfer between two repositories;
2. Maintaining consistent identifiers between the digital repositories;
3. Maintaining authentic records between the digital repositories.

Several methods of enabling machine-to-machine transfer are currently or will shortly be available, including OAI-PMH (Open Archives Initiative Protocol for Metadata Harvesting), OAI-ORE (Open Archives Initiative Object Reuse and Exchange), and RSS, which provide simple methods of expressing methods of expressing information; SOAP and WSDL web services may be developed to enable transfer, or SRB (Storage Resource Broker) may be used as a method for transmitting information. In the absence (at the time of investigation) of pre-existing tools that integrate these technologies with the various repositories and the time required to develop tools, the project team adopted OAI-PMH as a basis for transfer of metadata, and the subsequent http transfer of associated digital objects. The OAI export facilities of the EPrints and DSpace repositories were modified using appropriate software patches to export a full record of metadata. The expanded record contains database fields that are not present in the OAI-DC output, such as the EPrints note field that is commonly used by institutional repositories to store PDF export passwords, provenance information that may be used to record format conversion or other events in the object lifecycle, as well as checksum values that are essential for validating that the digital object is unchanged.

The storage of metadata and data provided by the Content Providers is relatively straightforward. The project adopted an atomic model in which each record provided the institutional repositories equates to a Fedora object. The persistent identifier for each Fedora Object is assigned on ingest into the repository. Each Object is likely to contain several datastreams, including PREMIS Object, PREMIS Event, format-specific and relationship metadata, as well as the metadata and digital object transferred from the institutional repository that are the objects of preservation. Datastreams may be updated or appended at a later date, subsequent to the performance of migration activities. Each stage of processing will result in event metadata being recorded that provides a full audit trail that can be queried to verify the integrity of the object being preserved. The resubmission of metadata and data into the Content Provider was not fully addressed in the project. A prototype system was developed that allowed repository staff to initiate download of all of the digital objects and metadata stored by the repository and metadata. However, further work is necessary to integrate the transfer process with the disparate repository software and versions in use. The project outputs of the DepositAPI Project (UKOLN, [2006](#)) may also prove useful.

Workflow Management

The development of a sustainable and scalable digital repository requires some consideration of the workflow that may be developed to manage activities, as well as the development of appropriate services to provide the functionality. The creation of a concise workflow is particularly important if two organizations wish to collaborate to achieve a common objective. An effort was made to avoid causing unnecessary changes or disruption to pre-defined work practices in the institutional repository – they may continue to accept research data, process it, and make it available in an appropriate timescale. The workflow of the Service Provider may only begin subsequent to the research data being made available. In abstract, the workflow for the project remains broadly compliant with the OAIS. However, the cyclical relationship that is inherent in the SHERPA DP Project requires some changes to that outlined in the OAIS Reference Model. Institutional repositories have developed a workflow that is pragmatic - the Submission Information Package (SIP) is often used as the basis for the Dissemination Information Package (DIP). The Archival Information Package (AIP) in the disaggregated model is created by the Service Provider, subsequent to ingest of the Content Provider's DIP into the preservation repository. There are pragmatic organizational and technical reasons for these changes. As a disaggregated service, the software does not yet exist to connect directly to the repository submission buffer and extract recently submitted data in a secure manner. In addition, a requirement for the preservation service provider to produce an archival version upon which a dissemination surrogate could be based would inevitably lengthen the time that elapses between deposit of the research paper into the repository and its eventual availability.

Many software tools have been developed that allow the definition of automated, or semi-automated workflows. These enable the performance of a series of tasks without manual intervention, improving the likelihood that a preservation system may scale to handle many Content Providers and object types. In the initial Sherpa DP system, these functions are implemented in a control layer above the individual service implementations. However, during subsequent investigation it has been found that Java Business Process Management (jBPM) ([n.d.](#)) – a processing language that allows the creation of workflows constructed of a series of components that have dependency relationships – may prove useful for developing appropriate models and work is underway to define preservation workflows. For example, it is possible to define a workflow that defines a common set of actions for format migration that uses pre-defined software applications to migrate digital objects, validates the success of migration and records appropriate metadata. In the event that the migration process fails, an exception may be defined that halts the process and notifies an administrator. At present, some operations, such as format recognition and metadata generation, use tools that are still under development, and require user input before processing can continue. However, the long-term objective is to reduce the degree of manual intervention to a minimum.

Conclusions

The OAIS is a useful model that offers significant flexibility when modelling a repository environment. However, it is evident that many repositories are unable to comply with the mandatory requirements in their entirety and may require additional support from a third-party service provider. By partnering with a preservation service provider, such as that developed by the AHDS for the SHERPA DP Project, institutional repositories may consider preservation issues in a sustainable manner that, most importantly, does not disrupt their core function of accepting and providing access to research data. On a larger scale, the academic community will benefit from further research into the practicalities of disaggregated repository models, particularly if they are to be considered a practical approach for meeting certification requirements as a trusted digital repository.

References

- Bekaert, J. & Van de Sompel, H. (2006). *Access interfaces for open archival information systems based on the OAI-PMH and the OpenURL Framework for context-sensitive services*. Retrieved on June 9, 2007 from Cornell University Library arXiv e-prints service: <http://arxiv.org/abs/cs.DL/0509090>
- Corney, D., de Vere, M., Folkes, T., Giaretta, D., Kleese van Dam, K., Lawrence, B., et al. (2004). *Applying the OAIS standard to CCLRC's British Atmospheric Data Centre and the Atlas Petabyte Storage Service*. Paper presented at UK e-Science Programme All Hands Meeting (AHM2004), Nottingham, UK. Retrieved on June 9, 2007 from <http://epubs.cclrc.ac.uk/bitstream/487/156.pdf>
- DSpace Federation. (n.d.). MIT Libraries & Hewlett Packard. Retrieved on June 9, 2007 from <http://www.dspace.org/>
- EPrints: Supporting Open Access. (n.d.). Retrieved on June 9, 2007 from <http://www.eprints.org>
- Hitchcock, S., Brody, T., Hey, J.M.N., & Carr, L. (2007). *Digital preservation service provider models for institutional repositories: Towards distributed services*. Retrieved on June 9, 2007 from <http://preserv.eprints.org/papers/models/models-paper.html>
- James, H., Ruusalepp, R., Anderson, S., & Pinfield, S. (2003). *Feasibility and requirements study on preservation of e-prints*. Report commissioned by the Joint Information Systems Committee (JISC). Retrieved on June 9, 2007 from http://www.jisc.ac.uk/uploaded_documents/e-prints_report_final.pdf
- jBPM Development team (n.d.). jBpm.org - java Business Process Mgmt. Retrieved on June 9, 2007 from <http://sourceforge.net/projects/jbpm/>

