

# **SHERPA-DP OAI Report**

## **An OAI compliant model for Disaggregated services**

### **Document Details**

Document type: Project Report  
Author: Gareth Knight  
Draft / Version: Version 1.1  
Date of edition completion: 20/08/2005

---

### **Change history**

1.0 29/06/2005. First draft  
1.1 15/09/2005. First version

## Terminology

Archival Information Package (AIP):	A derivative of the Submission Information Package that has been manipulated by repository staff to make it suitable for preservation. An AIP should, according to the OAIS Reference Model, provide “ <i>all the qualities needed for permanent, or indefinite, Long Term Preservation of a designated Information Object</i> ” (4-33).
Catalogue	A specialised database intended for the search and locate of information.
Depositor	The individual who has submitted the e-print into the E-print Archive. In OAIS terminology, they are described as the Producer
Disaggregated service	The practical implementation of the disaggregated model
Dissemination Information Package (DIP)	An Information Package, derived from one or more AIPs that are intended for use by the Researcher (Consumer). The DIP will be suitable for use by desktop users, but it may not possess all of the qualities of the archival version.
E-print Archive	A generic title to describe project partners within SHERPA DP that hold and distribute E-prints
Information Package	A conceptual container in the OAIS reference model that contains a digital object and associated information.
METS	An XML schema designed as a framework within which all metadata associated with a digital object can be stored.
Preservation Metadata	<i>Preservation metadata refers to information “a repository uses to support the digital preservation process”</i> (PREMIS, 2005). Specifically it should maintain the viability, authenticity, and identify of the resource.
Preservation Service	A generic term to describe the institution that is actively preserving e-print data – the Arts & Humanities Data Service.
Preservation staff	Staff employed by the Preservation Service.
Provenance	In the context of an OAIS, provenance describes events that occur during a digital object’s lifecycle. OAIS Provenance is equivalent to Recordkeeping metadata in the archival community.
Repository staff	Staff employed by the E-print Archive.
Submission Information Package (SIP):	An Information Package that is submitted by a Depositor (Producer) to the OAIS repository for preservation and dissemination.

## Audience

This document is written for use by AHDS Executive staff and Partners within the SHERPA-DP project.

# Contents

Terminology.....	2
Audience .....	2
Contents .....	3
Summary .....	4
1. Introduction.....	4
2. Application of the OAIS model .....	5
2.1. Overview .....	5
2.2. OAIS within the e-print environment .....	5
2.3. Criteria for OAIS Compliance .....	6
2.4. Suitability of the OAIS as a practical model.....	7
3. An Information Model for e-prints.....	9
3.1. Information Package .....	9
3.2. Submission Information Packages .....	10
3.4. Dissemination Information Package .....	13
4. An OAIS-compliant Infrastructure for disaggregated Services .....	14
4.1. A High-Level overview of the Functional Model .....	14
4.2. Identification of functions within the disaggregated model .....	15
5. A simplified workflow for disaggregated services .....	20
References .....	23

## Summary

Institutional repositories are a new area for development that is likely to have significant importance for the preservation of digital research. Although digital preservation is considered to be a significant issue, e-print archives do not currently possess the financial or technical capabilities to ensure digital research is preserved in the long-term. This report outlines a high-level model for describing a disaggregated service that outsources essential preservation service to a third-party. Specifically, it indicates how the OAIS reference model may be refined to fulfill the requirements of a disaggregated preservation service.

## 1. Introduction

*“A trusted digital repository is one whose mission is to provide reliable, long-term access to managed digital resources to its designated community, now and in the future.”*  
<http://www.rlg.org/longterm/repositories.pdf>

E-prints and institutional repositories are a new and high profile area, both for the JISC and for institutions. The initial focus of effort has been on the development of an institutional infrastructure through funding institutional repositories, such as TARDIS and the SHERPA project. However longer-term requirements will inevitably have some influence upon the emerging institutional repositories as they progress beyond the proof of concept and development stage.

Within the institutional repository community there is a growing awareness that digital preservation should be considered a relevant problem and that distinct skills are required to ensure the longevity of digital resources. Few repositories can claim to be actively preserving digital content and the funding model attached to institutional repositories does not cater for this activity.

To ensure the long-term management of a resource, a repository must provide a guarantee that the digital object will remain accessible. Digital preservation, as suggested in the JISC Continuing Access and Digital Preservation Strategy 2002-5 (Beagrie, 2002, p. A13), may be considered a ‘value-added’ service to be provided in a number of ways: an institution may allocate staff to perform preservation processes in-house, could work collaboratively with a group of other repositories, or may contract a third-party agency to provide preservation services for its collection (James et-al, 2004; RLG/OCLC, 2002; Beagrie, 2002). The SHERPA DP project shall examine the latter approach, where the Arts & Humanities Data Service is contracted by multiple institutional repositories to hold, preserve and migrate e-prints.

The primary output of the SHERPA DP project – a sustainable model for the provision of a preservation service for existing institutional repositories – offers several advantages:

- The introduction of preservation practices to current institutional repositories without significant change to existing organizational or technical infrastructure;
- Reduce possible duplication of actions between multiple repositories;
- Implementation of standardized preservation practices across multiple repositories;
- Automation of preservation services, a practice that would be unfeasible for the preservation of a small number of digital resources.
- The use of high capacity, off-site storage facilities

The provision of a long-term management may be considered an important method of ensuring the E-print Archive complies with the RLG/OCLC specification for a Trusted Repository. This report will apply the OAIS reference model to the distributed preservation service proposed by the SHERPA DP project. It will identify rights and responsibilities, services and actions and apportion these between the institutional repositories and the preservation repository service. It should be read in conjunction with the ‘Requirements of a disaggregated service’ document.

Although the report outlines the implementation of a disaggregated model for e-print archives, the organizational model may be applied to institutional repositories that preserve other forms of data.

## 2. Application of the OAIS model

### 2.1. Overview

The Open Archival 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 (2002, p.1-1). It may be applied to all digital repositories that deal with the long-term preservation of information and provides a framework for comparing the architecture and operation of existing and future archives.

The model was developed with three broad objectives:

- To establish a shared vocabulary and common terms and concepts for discussion of archive issues among stakeholders, who may originate from different research communities.
- To serve as the foundation for the development of standards and work practices that supports the archive environment.
- To articulate a core level of functionality that an archive should provide and establish a high-level architecture for describing the relevant system components.

The reference model provides a standard method to describe the functionality of a repository and has been used as the basis for further study, such as the requirements for trust (RLG & OCLC, 2002). As a conceptual model, it is assumed repository implementers will use the model as a guide “*while developing a specific implementation to provide identified services and content*” (OAIS Ch. 1.4).

### 2.2. OAIS within the e-print environment

The OAIS is intended to provide specific functions within the larger information environment. Specifically, the reference model identifies three stakeholders:

1. Producers that supply the information to be preserved by the archive.
2. Management, who establish the objectives of the archive, handles funding processes and performing other functions relevant to the strategic direction of the archive.
3. Consumers, who represent the target audience of the OAIS-compliant repositories:

An OAIS compliant repository must cooperate with the three stakeholders to identify their needs and ensure the archive fulfils their requirements.

To understand how the entities contribute to the maintenance of a repository it is useful to identify the members of these groups and refine the terminology to the academic arena in which they operate – the generic terms *producer* and *consumer* may be substituted with *depositor* (authors or other authorised individuals) and *researcher* respectively. The management aspect of the OAIS compliant service itself require further refinement, to separate them into two distinct entities:

#### E-Print Archive

An E-print archive may, in simple terms, be described as a location to archive digital research, in the form of preprints (pre-refereed, pre-publication drafts of scholarly articles) or postprints (refereed, published articles). The Institutional Archives Registry (2005) indicates there are currently 459 OAI compliant repositories that archives some form of research data. These utilise different software solutions (GNU EPrints and DSpace are the most common) and are implemented within different organisational models.

E-print archives participating within the SHERPA DP (and SHERPA as a whole) may be categorised as institutional/departmental archives that store papers produced by local authors or multi-institutional archives that provide a single access to search multiple locations (Key Perspectives & EPIC, 2004). Resource discovery metadata is harvested through the OAI-PMH (OAI Protocol for Metadata Harvesting), which is made available through search services, such as OAIster, for searching or browsing by the user.

### Preservation Service

Preservation Service is the generic term used to describe the third-party institution that is responsible for the active preservation of research data. The Preservation Service should possess relevant knowledge and the capacity to convert the disaggregated model into a practical implementation. It is primarily responsible for the construction and maintenance of the archival version of the e-print and the creation of supplemental metadata. The Arts & Humanities Data Service will perform the functions required for a preservation service within the SHERPA DP project.

Figure 1, based upon figure 2-1 in the OAIS, offers a high-level overview of the primary stakeholders in the OAIS environment.

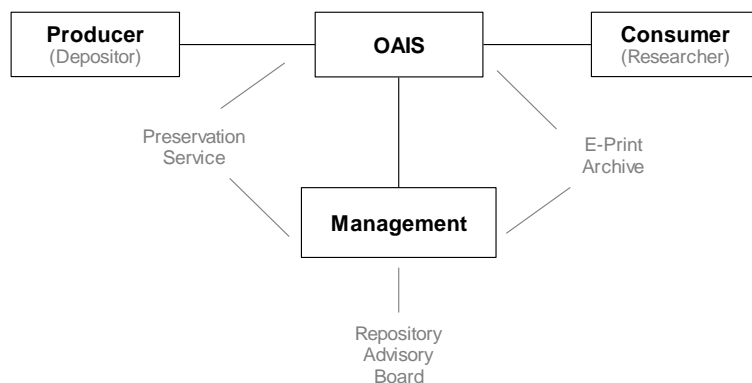


Figure 1: The primary stakeholders within the SHERPA DP environment

In practice, the archive will interact with Depositors, who will submit an e-print and Researchers, who will be from the OAIS Designated Community. The archive will also interact with management of the E-print archive, Preservation Service and a Repository Advisory Board, which will be responsible for guiding the strategic direction of the OAIS and making recommendations on the development of the service infrastructure.

### 2.3. Criteria for OAIS Compliance

The definition of OAIS compliance is intentionally vague within the reference model and subsequent reports by the RLG study on Trusted Repositories (RLG & OCLC, 2002) have sought to clarify ambiguous terminology. Conformance may be applied through implementation of the OAIS Functional and Information model to an in-development archive (i.e. the SHERPA DP disaggregated service) or performed retroactively for an existing service, such as the recent assessment of OAIS and METS compliance within UKDA and The National Archive (Beedham et-al, 2005).

For OAIS compliance, an archive must be able to map its organizational infrastructure to the OAIS functional model and demonstrate it performs six mandatory responsibilities related to the collection, preservation and distribution of digital resources:

1. Negotiate for and accept appropriate information from information producers.
2. Obtain sufficient control of the information in order to meet long-term preservation objectives.
3. Determine the scope of the archive's user community.
4. Ensure that the preserved information is independently understandable to the user community, in the sense that the information can be understood by users without the assistance of the information producer.
5. Follow documented policies and procedures to ensure the information is preserved against all reasonable contingencies, and to enable dissemination of authenticated copies of the preserved information in its original form, or in a form traceable to the original.
6. Make the preserved information available to the user community

In reality, many repositories may claim to comply with some or all of these requirements. E-Print Archives have, at the minimum, considered the user community they will serve, the type of information to be stored within their repository and associated policies and procedures for managing

the information. However, they do not currently implement substantive services to manage preservation tasks – an area that will be resolved by the Preservation Service.

## 2.4. Suitability of the OAIS as a practical model

The OAIS reference model offers a common set of concepts, relationships and processes. It may be compared to an ontology that enables institutions to communicate on common issues related to repository management and preservation without necessarily operating within the same subject field. As a reference model, it does not imply a specific design or formal method of implementation (Gladney, 2004). Instead, it is left as an exercise to the reader to develop their own implementation by analysing existing business processes and matching them to OAIS functions.

OAIS-compliant repositories typically produce a system architecture that remains broadly compatible with the OAIS model, but differ in its implementation. These typically take one of three approaches:

1. Map the six entities of an OAIS-compliant repository (Ingest, Archival Store, Administration, Data Management and Access) onto an existing structure and retain the domain-specific terminology and interactions.
2. Implement an OAIS-compliant structure and ensure that domain-specific terminology can be mapped to the OAIS equivalent.
3. Implement an OAIS-compliant structure and terminology, while ensuring that domain-specific terminology is referenced.

The chosen approach is likely to differ according to the current stage of development at the time that an OAIS mapping is produced. Repositories that possess an existing infrastructure are more likely to map the six OAIS entities to fit their existing model. In contrast, new repositories may use the OAIS model as a starting point. In the absence of a practical implementation, the SHERPA-DP project will implement the third option for the current stage of development and refine the model.

The flexibility of design is an advantage and disadvantage. The OAIS model outlines a common structure to organise the disparate functions of a digital repository. However, certain aspects of its organisation may be confusing, vague, or require clarification. There are likely to be significant differences between OAIS implementations that operate within different research areas. When applying the OAIS model to a disaggregated infrastructure, it was found some degree of refinement was necessary in several areas:

### 1. Vocabulary

The CCSDS provide a common definition of terminology used within the OAIS reference model. However, when used within a specific subject field that has its own vocabulary these terms are often vague or misleading. For example, the use of provenance in an OAIS context differs slightly when describing record keeping metadata.

The disaggregated model represents a specific interpretation of the OAIS and has been refined to take into account the localised vocabulary of e-print archives and the additional complexity associated with the formal interaction of the E-print Archive and Preservation Service. Although the model remains broadly compliant with the OAIS, the model is not universal and may require some reinterpretation if applied within a different organisational structure.

### 2. Clarification of the Information Package

The information to be preserved will differ from the definition provided in the OAIS Reference model through its application to actual data formats and metadata. To apply the model in a practical environment, some modification to the construction of an information package is necessary. In the proposed workflow (see section 5.1.), the dissemination copy of the e-print and associated metadata is derived from the original submission. The AIP is constructed later in the workflow, after the information package has been transferred to the Preservation Service. There are three reasons for this change:

1. The time between the deposit of research data and it being made available is kept to a minimum.

2. The Preservation Service is not required to perform immediate action, which could prove problematic for complex or little understood file formats.
3. The Preservation Service is able to collect information that will only be created after the construction of the dissemination version of an e-print.

Second, the Dissemination Information Package (DIP) differs slightly from that defined by the OAIS Reference model. Although the purpose remains the same, it is not considered practical for the E-print Archive to generate an on-the-fly version of the e-print for dissemination to the Researcher.

Subsequent work will focus upon implementing a model that offers a practical implementation and maintains a degree of compatibility with OAIS terminology.

### 3. An Information Model for e-prints

The primary goal of an OAIS-compliant repository is to preserve information intended for a designated community for an indefinite period of time. To perform this function, an adequate definition of 'information' must be obtained. The OAIS model represents information as a physical or digital 'data object' that can be interpreted. In the context of an E-print Archive, 'information' may be applied to intellectual content – the words, images and layout that compose an e-print. This is often accompanied by additional information necessary to interpret the content and collectively referenced as an Information Package.

A preservation policy for e-prints will be based upon a three-tier understanding of digital preservation:

1. Preservation of the bit stream (the basic sequence of binary data) that represent the information stored in a digital resource.
2. Preservation of the information content (words, images, etc.) stored as bits and defined by a logical data model, embodied in a file or media format.
3. Preservation of the experience (speed, layout, characters, etc.) of interacting with the information content.

There is no definitive approach to the problem of maintaining digital content across multiple generations of technology. Nevertheless, it is important that organizations with a responsibility to preserve research data should declare to their stakeholders the extent to which they are able to perform this function. For the SHERPA DP project, the Preservation Service will preserve digital information, as defined by the first and second criteria.

#### 3.1. Information Package

To adequately preserve the intellectual content of the e-print, the Information Package – the bundle of information that accompanies the e-print – must change over time.

The OAIS Reference Model indicates the Information Package will undergo three distinct stages of development: Submission Information Package (SIP); Archival Information Package (AIP); and Dissemination Information Package (DIP). These move from the producers through the OAIS and on to users.

The workflow required to implement the disaggregated model requires a different approach. Figure 2, adapted from figure 2.4 in the OAIS reference model, demonstrates the information flow that occurs during operation of the disaggregated service.

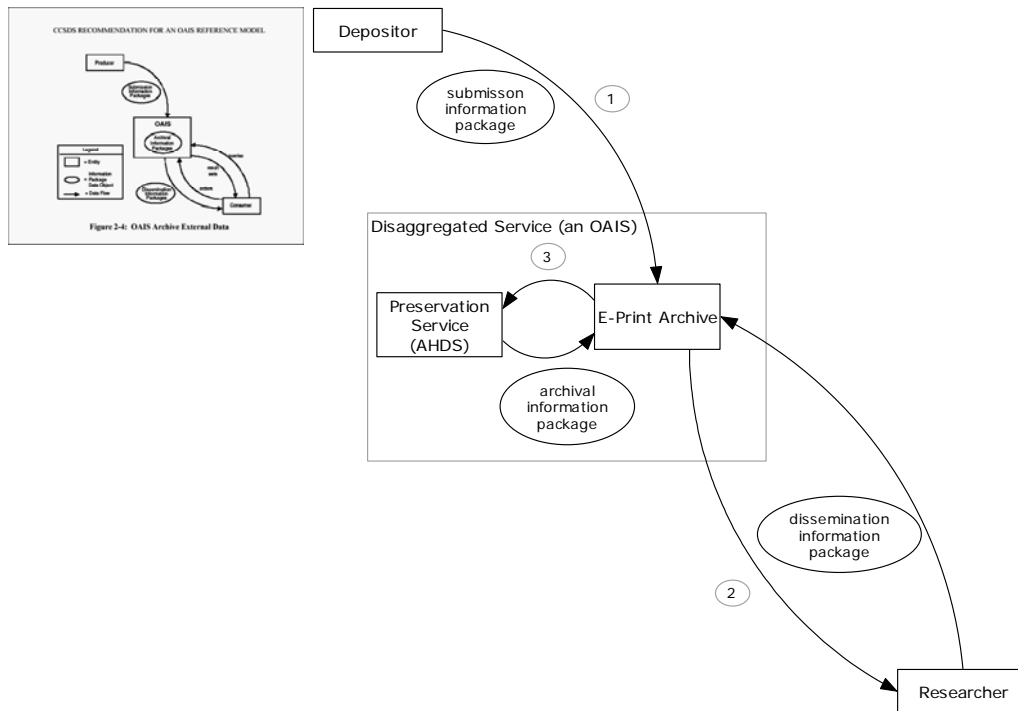


Figure 2: The information flow during operation of the disaggregated service

In the disaggregated model, Depositors submit an e-print and associated metadata (SIP) to the E-print Archive (1). The E-print Archive then prepares the content of the SIP so that it meets their internal requirements (e.g. completing missing metadata fields, migrating content to PDF) and makes it available to the Researcher (2). The Preservation Service will subsequently transfer the submitted e-print and associated metadata to the preservation store, and will generate an archival version (AIP) (3). The Preservation Service will subsequently perform migratory action upon the e-print and deliver the contents to the E-print Archive if the dissemination format is rendered obsolete.

## 3.2. Submission Information Packages

Submission Information Package (SIPs) is the OAIS term to describe data submitted by a Producer to an OAIS repository for preservation and dissemination. For an e-print archive, the SIP is likely to constitute the research output created by an author/s that will be the focus of preservation<sup>1</sup> and relevant discovery metadata requested by the archive during the submission process.

### 3.2.1. File Formats

It is considered good practice to limit the number of file formats accepted by a digital repository and to store data in formats that are open or well documented. E-print archives must balance the practical implications of encouraging authors to deposit their research with the goal of storing a limited number of file formats. The costs and risks associated with digital preservation tend to grow when a digital collection includes a larger number of diverse file formats (James et-al, 2004; Granger, Russell & Weinberger, 2000) and increases the risk of loss of access to the intellectual content over time.

SHERPA DP project partners follow the recommendations of James et-al (2004) by limiting the deposit options to a select number of file formats. These are typically based upon open or documented standards that can be produced and viewed without significant effort (James et-al, 2004). For example, Microsoft Word, plain text, HTML, Rich Text Format (RTF) and Postscript. To a limited extent they also accept other formats on an ad-hoc basis (e.g. images, CorelDraw objects), although they do not provide a guarantee that the content will be made accessible in a suitable format at a later date (ERA Format Support Policy, 2004).

<sup>1</sup> An E-print may be stored in a single or multiple files. The White Rose consortium report they receive research data contained within several different file formats.

In the disaggregated service, the Preservation Service should accommodate existing policies regarding the file formats accepted by the E-print archive and establish procedures to identify the preservation action that should be taken. For example, migration of Microsoft Word documents to XML for archival purposes.

### 3.2.2. Metadata

Existing metadata collected by institutional repositories focus upon resource discovery. The majority of SHERPA repositories, similar to other e-print archives take a pragmatic approach to metadata schemas through the definition of bespoke application profiles to suit the needs of their repository. These profiles retain a certain underlying level of consistency through support for the minimum data recommended for Dublin Core within the oai\_dc record format, but may vary in terms of the minutiae as to how they interpret some unqualified DC elements.

It is unlikely that depositor-created metadata will meet the requirements of the E-print Archive and further work is necessary to enhance it. For example, repository staff may be required to complete missing elements or reformat existing text to a pre-defined layout (e.g. convert 'Richard Jones' to 'Jones, Richard'). The OAIS classifies such enhancements to an existing SIP as resulting in an 'Updated SIP'

### 3.3. Archival Information Package

An Archival Information Package (AIP) is a derivative of the Submission Information Package that has been manipulated by repository staff to make it suitable for preservation. An AIP should, according to the OAIS Reference Model, provide "all the qualities needed for permanent, or indefinite, Long Term Preservation of a designated Information Object" (4-33). Specifically, preservation should maintain three qualities:

1. *Viability* - the archived digital object's bit stream must be intact and readable from the digital media upon which it is stored.
2. *Renderability* – the bitstream should be translated into a form that can be viewed by human users, or processed by computers.
3. *Understandable* – sufficient information should be provided to ensure the rendered content may be interpreted and understood by its intended users.

The construction and maintenance of the AIP is considered a priority for the Preservation Service and will form the focus for analysis. Figure 3 maps the proposed e-print structure onto the abstract OAIS Archival Information Package.

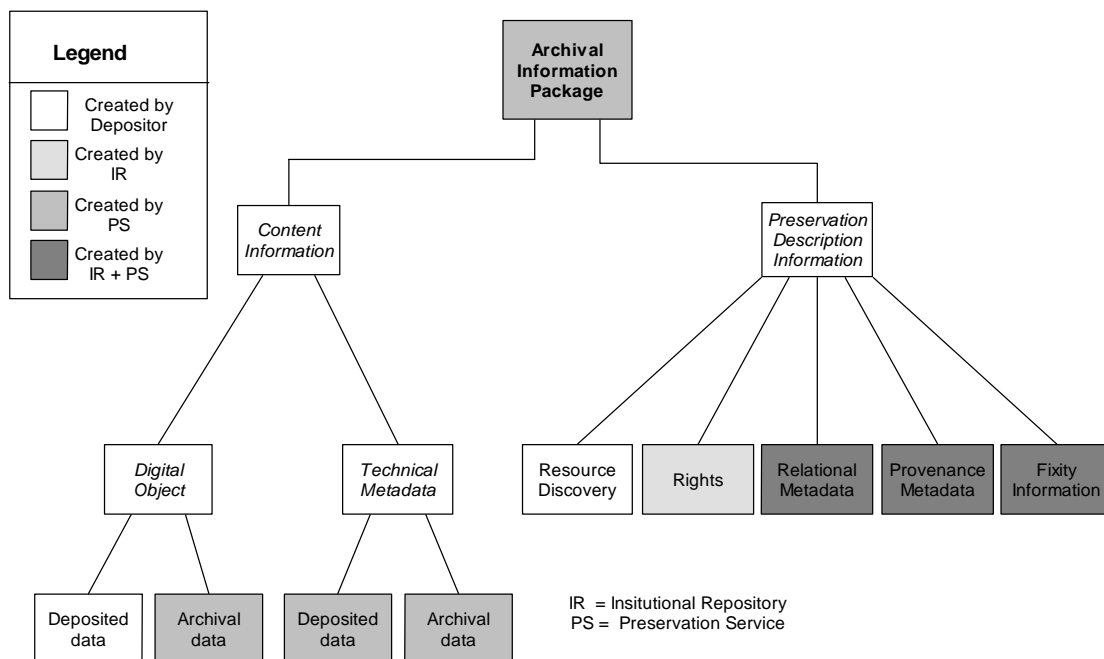


Figure 3: An abstract view of the Archival Information Package

In practical terms, an AIP within the disaggregated model is likely to consist of components constructed by the Depositor, E-print Archive, Preservation Service, or any combination of these roles. It will consist of the following elements:

1. *Digital Object* – The e-print to be preserved. The Preservation Service will store a ‘deposited’ original, as submitted by the depositor and an archival version created and maintained by the Preservation Service.
2. *Technical* – Information to describe the technical infrastructure of the ‘deposited’ and ‘archival’ versions of the digital object. For example, file format and version, etc.
3. *Resource Discovery* - Metadata required to locate and retrieve the e-print. Discovery metadata is created by the Depositor and/or E-print Archive on Ingest.
4. *Rights* - Metadata necessary to describe the rights associated with the e-print. Rights information should distinguish between the Author and publisher version of an e-print.
5. *Relational* – Metadata necessary to identify different versions of a research paper. This may be different revisions of the same document completed at different time periods (pre-prints, post-prints) or multiple versions of the same document (an Microsoft Word ‘original’, an archival RTF version, and a PDF dissemination version).
6. *Provenance* - A description of the content history, including its origins, changes to the object or its content over time, and its chain of custody (if known). Provenance information would include the supplementation of depositor-created metadata or creation of a dissemination version based upon the deposited data by the E-print Archive and the migration of Deposited data to an archival format by the Preservation Service.
7. *Fixity Information* – Metadata necessary to authenticate the digital object and ensure it is unchanged. Authenticity is a key issue raised in the RLG/OCLC study on Trusted Repositories.

METS (Metadata Encoding and Transmission Standard) is an XML framework for the storage of disparate information. To store and transmit metadata and data between the E-print Archive and Preservation Service, METS is to be used to encode the digital object, stored as base64, and associated metadata into a single file.

### 3.3.1. E-prints that cannot be migrated

The E-print Archive is likely to receive research data that possess different copyright requirements. A policy must be developed to identify and handle e-prints that for differing reasons cannot be migrated to a different file format. Specifically, two types may be identified:

1. *Author version* – An e-print that contains intellectual content, layout and other information created by an author and/or other non-commercial entity. In some circumstances, it has been noted the author imposes restrictions that forbid amendments or migration (Tracey Stanley, e-mail correspondence).
2. *Publisher version* – A postprint that has been submitted to a publisher for editing and publication. It is common for publishers to impose limitations on use, including migration of the whole document, or enforce their copyright for certain elements, such as layout.

It has been noted that authors perceive a publisher PDF of the research data to be the definitive version and are, in many circumstances, unaware or uncomfortable with the deposit of author-created versions. The Preservation Service should not oppose the deposit of publisher PDFs. However, the actions to be performed should be clarified. In these circumstances, it is envisaged the Preservation Service will store the submitted version within the dark archive; create preservation metadata; and inform the E-print archive that restrictions will limit the possible preservation action (e.g. the publishers template may be copyrighted material and cannot be migrated).

In order to identify e-prints that E-print archives should clarify copyright statements in their metadata to enable the Preservation Service to programmatically distinguish between e-prints that may or may not be migrated.

### 3.4. Dissemination Information Package

The OAIS Dissemination Information Package (DIP) is an abstract information object that is intended for use by the Researcher (Consumer). The DIP will be suitable for use by desktop users, but it may not possess all of the qualities of the archival version.

The workflow proposed for the SHERPA DP project differs from that defined by the OAIS Reference model and, as a result, the DIP will also differ. In the disaggregated model, the dissemination version of the e-print and associated metadata is derived from the original submission.

In the initial stage of deposit, the e-print archive is responsible for the creation of an e-print suitable for dissemination. The majority of research holdings held by institutional repositories are stored and disseminated in the Portable Document Format (PDF) format <sup>2</sup>.

However, in the event that PDF is rendered obsolete or replaced by a similar, incompatible file format, the Preservation Service will be responsible for exporting the intellectual content held within the SIP/AIP to a replacement format and transferring it to the E-print archive.

---

<sup>2</sup> An exception is the White Rose Consortium that hold 2 records which point to a HTML page.

## 4. An OAIS-compliant Infrastructure for disaggregated Services

To preserve the digital object, an OAIS-compliant repository is expected to meet mandatory requirements necessary to negotiate for appropriate information from Depositors, obtain sufficient control in order to preserve the digital object, ensure the preserved information will be understandable without external assistance, and make it available to the user community (CCSDS, 2002). To perform these roles, the OAIS separates responsibility into six broad entities that perform specific functions:

1. **Ingest:** The services necessary to accept data from a Producer and create an archival version.
2. **Archival Storage:** The services required to store, maintain and retrieve the deposited data.
3. **Data Management:** Functions required to populate a search database, to allow the user community to locate a resource, and administer the archive in its entirety.
4. **Preservation Planning:** Responsible for the development and review of the OAIS Preservation Strategy.
5. **Access:** The facilities available that allow the Designated Community to locate, request and receive data that reside in the Archival store.
6. **Administration:** Responsible for managing the day-to-day operation of an OAIS and coordinating the activities of the previous five OAIS services.

### 4.1. A High-Level overview of the Functional Model

The OAIS model outlines a minimum specification for compliance and the E-print Archive and Preservation Service in its entirety can therefore be regarded as an OAIS. The disaggregated model proposed for the SHERPA DP project may be compared to the 'shared resources' category of archive association. The Preservation Service, to be provided by the AHDS is best viewed as a combination of certain aspects of the OAIS ingest and archival store functions. The main aspects of the OAIS Ingest and Access function shall be provided by the relevant e-print archive.

Figure 4 illustrates the sharing of a common storage and preservation function between two archives, OAIS 1 and OAIS 2. Each archive will serve independent communities and the AHDS, as the preservation service, will not intervene in the delivery.

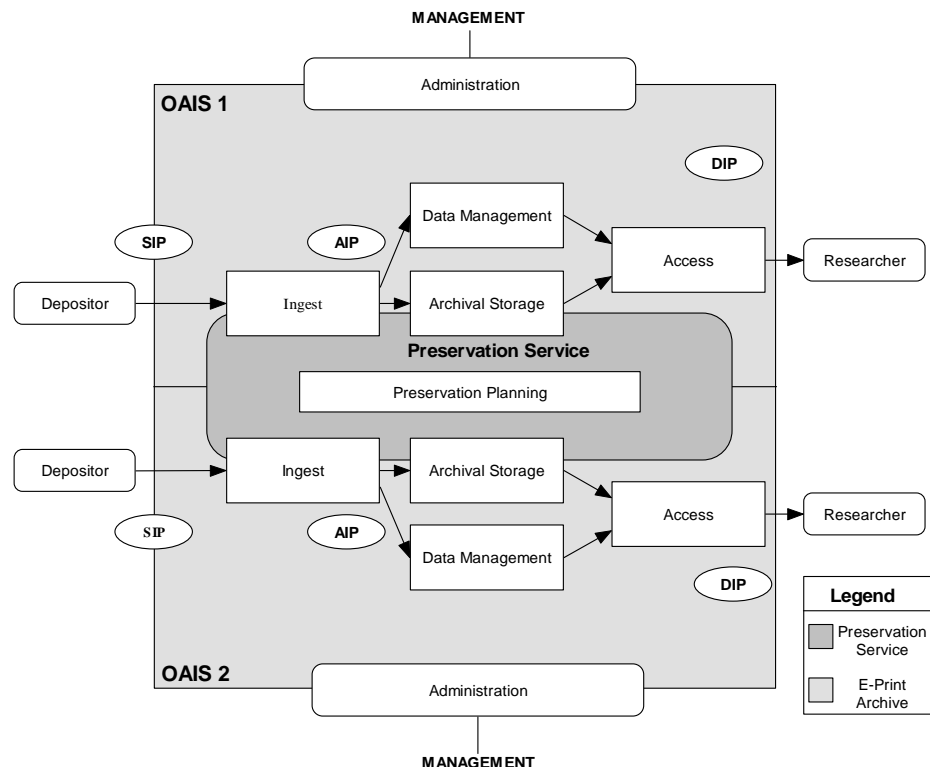


Figure 4: Illustration of two archives that share a common storage function

At a high-level, the progress of an Information Package within the disaggregated OAIS model will consist of six broad stages:

1. The Depositor (OAIS Producer) submits a Submission Information Package (SIP), consisting of an e-print and associated metadata, to the archive.
2. Repository staff refines the resource discovery metadata that accompanies the e-print, as defined by internal archive specifications.
3. On a pre-determined schedule, the updated SIP is transmitted to the Preservation Service, which generate an Archival Information Package (AIP) intended for preservation.
4. The AIP is stored within the Archival Store at the AHDS and an appropriate backup strategy is implemented.
5. The E-print Archive generates a dissemination version intended for use by Researchers (OAIS Designated Community) and make it available via their search catalogue.
6. A user is able to request and download a copy of the Dissemination Information Package.

Further details of the requirements of each stage are provided below.

## 4.2. Identification of functions within the disaggregated model

The disaggregated model proposed for the SHERPA DP project may be considered an OAIS-compliant archive that will implement required services. In this section, we will examine how the E-print Archive and Preservation Service provide the services necessary to be considered a trusted archiving service (RLG-OCLC, 2002). The functions identified at each stage of the disaggregated model are useful to establish the technical infrastructure that must be implemented and the roles and responsibilities that may be delegated to each party. Figure 5 provides a detailed view of how these functions relate to one another within an OAIS-compliant infrastructure.

### 4.2.1. Ingest

The deposit process has been a focus for investigation during the FAIR programme (Focus on Access to Institutional Resources) and, as a result, Ingest is the most clearly defined entity. Ingest, as defined by the CCSDS Reference model for an OAIS “*provides the services and functions to accept Submission Information Packages (SIPs) from Producers*” (CCSDS, 2002, page 4-1). Based upon work performed by the CCSDS (2002), CLRC/RLG study on Trusted Repository (2002) and Lavoie (2004), six ingest functions may be identified:

1. **Accept information transferred to the OAIS-compliant repository**  
The E-print Archive must provide some method for the Depositor to submit their e-print into the E-print Archive. The current test-bed repositories currently implement a mix of DSpace and EPrints that offer mechanisms to ingest a SIP, through direct or mediated deposit.
2. **Confirm that the received information is complete and uncorrupted**  
To ensure the transmission was successful, validation should be performed on the deposit to ensure it has been submitted in the correct file format (as defined by Deposit Guidelines), has not been infected by a virus, and that the file has not been corrupted en-route (Validate transmission).
3. **Confirm that the E-print Archive may preserve the information**  
The E-print Archive must establish permission to preserve the e-print in the long-term (Ingest Policies). Repositories have traditionally been reticent to introduce barriers to the submission process (James et al, 2003). SHERPA repositories have, to varying degrees, implemented a deposit licence that establishes their right to hold and preserve the e-print in the long-term (RLG/OCLC report, 2002), as well as protect the institution from legal action and ensure the process does not infringe the authors’ copyright or their right of integrity (modifications that would be prejudicial to the artist’s reputation) (Knight, 2004). However, these do not currently recognise that e-print repositories may have escrow arrangements for preservation and secure remote storage. It is likely that some refinement of the licence will be necessary to allow the E-print Archive to transfer responsibility for preservation to the AHDS.
4. **Extract and/or create discovery metadata to support search and retrieval**

The depositor is unlikely to create resource discovery metadata suitable for search purposes. The E-print Archive is likely to have to further enhance the metadata to ensure it complies with the repository's policy on Metadata

#### **Transfer of submitted data and metadata to the Preservation Service**

Some method of transferring the updated SIP should be identified. Please consult the *Formal Requirement for a Disaggregated Service* document (Knight, 2005) for possible ways to allocate responsibility.

#### **Transform submitted information into a form suitable for storage and management**

Transformation of the Updated SIP to an Archival Information Package should be considered the most important function within the SHERPA-DP model. The transformation method will vary, according to the results of the Risk Assessment and the Preservation Strategy implemented by the Preservation Service. For example, a risk assessment may indicate the Microsoft Word 95 format is at risk. A review of the Preservation Strategy may recommend the format is migrated to an alternative format (e.g. RTF) and an automated tool will be launched to perform this task.

#### **4.2.2. Archival Storage**

Archival Storage in the OAIS model is responsible for the long-term storage and maintenance of the digital resources entrusted to the repository. This requires validation action to ensure the digital resource remains accessible in the long-term and that it is stored on appropriate storage media. To fulfill its requirements, the Preservation Service should implement procedures to perform the following:

##### **Backup Procedures**

All institutions should implement some form of backup strategy, and the institutions participating in the SHERPA-DP project are no exception. Suitable hardware and procedures should be in place to perform regular backups at the E-print Archive and Preservation Service:

1. Store e-prints on a RAID server and provide multiple redundancy in the event of a hard disk failure;
2. Locate tape backups offsite
3. Perform error checks on a periodic basis to identify media failure.

##### **Migration Strategy**

The long-term accessibility of digital resources is included within the remit of the Preservation Service and so, procedures to validate that the e-print remains accessible are necessary. It is expected that obsolescence checks will be performed on the copy held by the Preservation Service, as opposed to that held in the E-print Archive to reduce complexity.

Similar to the Transformation function in the Ingest entity, the Preservation Service should validate data and migrate it when necessary. The process may be summarized in four stages:

1. Assess risk to file format, based upon a pre-defined criteria provided by internal software tools (such as those developed for the DAAT project)
2. If the file format is considered at-risk, review and perform the preservation actions outlined in the Preservation Strategy.
3. Perform error-checking procedures, to evaluate the outcome of the preservation process. Notify a member of staff at the Preservation Service if the process fails.
4. Update preservation metadata to record the performed actions, as a method of authenticating the process at a later date.

#### **4.2.3. Data Management**

The Data Management function is responsible for administrative functions associated with the internal maintenance of the E-print Archive (and outside the remit for this report). This includes maintenance of repository catalogue system necessary to locate and retrieve e-prints.

The Preservation Service will contribute some metadata to the repository catalogue, when a dissemination file format is rendered obsolete and a new version of the DIP must be generated.

#### 4.2.4. Preservation Planning

Preservation Planning identifies the methods necessary to monitor e-prints held within the archive and develop a strategy to ensure the storage file formats for e-prints are not rendered obsolete. For the purpose of maintaining accessibility, preservation planning should involve the following:

##### 1. Monitor Preservation Community

The Preservation Service must monitor the preservation community and take note of international guidance on the preservation of digital objects. This will provide practical examples on the minimum amount of metadata required to preserve the e-print, as well as recommendations for formats that may be suitable for preservation (e.g. PDF Archive).

##### 2. Develop and Review Preservation Strategies

The Preservation Service, in cooperation with the E-print Archive, must develop procedures to ensure the e-print is accessible in the long-term. A clear preservation strategy should be written that outlines the:

- Method(s) of identifying at-risk content;
- Information required to support the digital object.
- Significant properties of the file format;
- Methods available to preserve the content (i.e. migration or emulation);

In practical terms, each e-print is different, consisting of different versioning, layout information, structure, and plug-in components that may make it difficult to adopt a single approach to preservation.

##### 3. Implement Preservation Strategy

To ensure the e-print remains accessible, the Preservation Service implements procedures outlined in the Preservation Strategy. This is likely to involve one or more of the following stages:

- Migration of the e-print to a file format considered appropriate for preservation;
- Creation of preservation metadata necessary to document actions performed on the digital object. For authentication purposes it is important that the details of any conversion are recorded when a derivative is produced to ensure it can be traced to the original submission.
- The creation of fixity information to allow the Preservation Service and E-print archive to verify that undocumented changes has not occurred during transmission.
- The creation of an agreement between the E-print Archive and the Preservation Service that outlines the terms of service (a description of the service to be provided, the method of implementation, and an exit strategy).
- The creation and implementation of procedures for the creation and/or enhancement of preservation metadata, to be followed by the E-print Archive and Preservation Service.
- The creation and implementation of procedures for the preservation of e-prints that have been transferred to the Preservation Storage Area, to be followed by the Preservation Service.

Further discussion of this subject is provided in the 'Formal Requirements for a disaggregated Service' document.

#### 4.2.5. Access

The Access function manages the services necessary to deliver (or restrict) Dissemination Information Packages (DIPs) to the Consumer (most likely a researcher). The OAIS Reference Model identifies several services necessary to process user queries and forward them to the catalogue held within Data Management, coordinate the retrieval of requested content from Archival Storage and delivery and perform any transformations necessary to render the AIP usable by the Consumer

Although the delivery method should be considered the responsibility of the E-print Archive, the Preservation Service may wish to consider two issues:

1. Automated software tools that will allow the Preservation Service to convert an AIP into a DIP.
2. The extent of information that should be made available to the user when they download the e-print. For example, should they be made aware that content has been migrated?

#### 4.2.6. Administration

Administration is an inclusive term to describe the various services and functions necessary for the overall operation of the E-print Archive and preservation service. The organisation of these administrative functions will be unique to each institution and many aspects of operation should be considered outside the remit of the SHERPA-DP project (e.g. annual reports made by the E-print Archive and Preservation Service to funding bodies, negotiation for university resources, hardware maintenance, etc.) or may be expanded later when a review of Repository procedures is performed. In the OAIS Reference Model, Administration may be applied to three broad areas:

- 1) Policies and procedures;
- 2) Reports on various aspects of the repository; and
- 3) Hardware and software maintenance.

For the purpose of simplicity, we will focus upon the first and second areas within this report, with particular emphasis upon the information necessary to preserve the Content Information.

#### E-print Archive

In order to preserve the digital object, archive staff must implement the following:

- The creation and refinement of a Deposit Licence that enables the E-print Archive to hold and preserve the E-print for a specified time-period, and allocates the work to the Preservation Service. Knight (2004) offers a sample licence for preservation.
- The implementation of procedures to manage the E-Print, including:
  - Creation of persistent identifiers for new accessions.
  - Creation and/or enhancement of resource discovery metadata.
  - Creation of administrative metadata necessary to manage the E-Print within the repository.
- A reporting structure to inform the AHDS of any future technical and organisational changes to the service that may affect the Preservation Service.

It is expected that many of these policies will have been developed in the early stages of the repository's establishment and, therefore it may be necessary to refine them to the requirements established by the Preservation Service.

#### Preservation Service

The agreement between the E-print Archive and the Preservation Service should outline the information required by each party, and some method of monitoring service performance will be required. As a minimum, the AHDS should report the following information:

- The preservation actions performed by the AHDS, with an outline of the time taken and software used.
- The performance of the Preservation Storage area and backup facilities. This should provide an indication of any downtime of the service, the number of Information Packages ingested during an allotted time period (e.g. 6-12 months), the average size of the AIP, and the remaining capacity.

The results of these reports may be shared with the relevant E-print Archive and any management board within the AHDS.

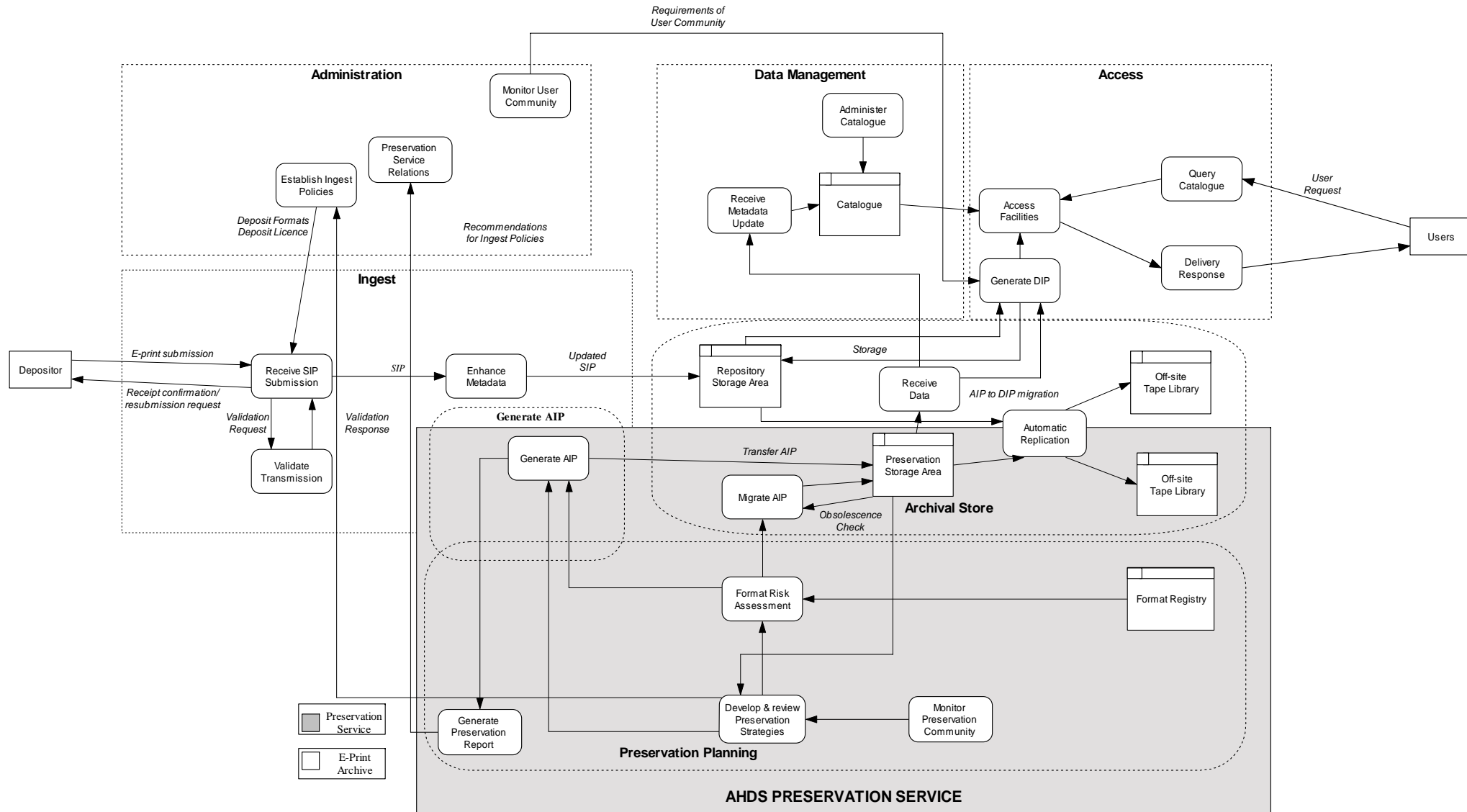


Figure 5: A detailed view of a disaggregated OAIS-compliant model

## 5. A simplified workflow for disaggregated services

The generic approach outlined in this document involves a series of steps undertaken by the Depositor, E-print Archive and the Preservation Service to assemble, describe, transfer and preserve the research data. Figure 6 indicates how the abstract model may be converted into a workflow for staff, indicating the tasks they are expected to perform at each stage. The simplified workflow should be read in conjunction with the previous section on an OAIS-compliant infrastructure.

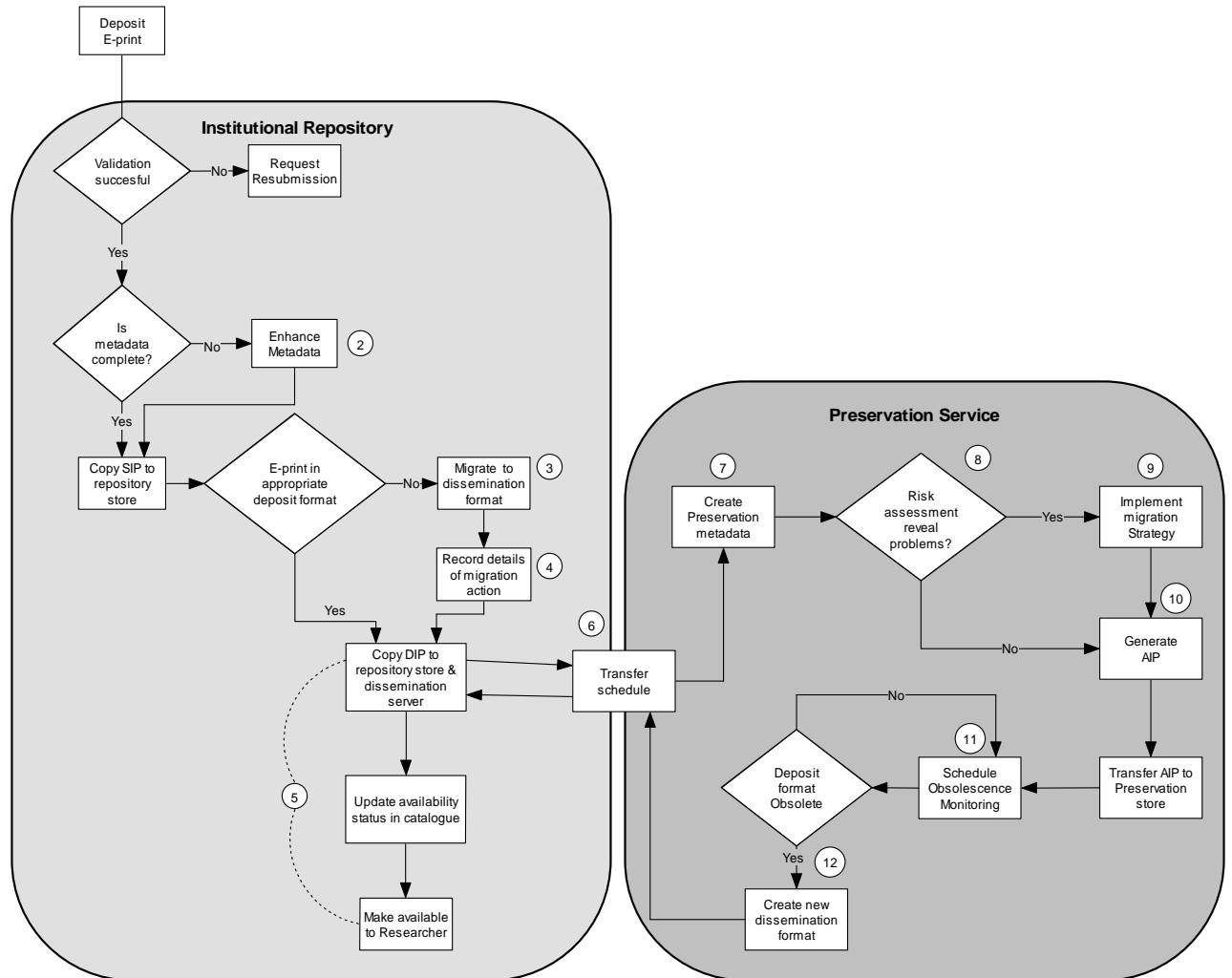


Figure 6. Simplified workflow for the Institutional repositories and Preservation Service

E-print archives are likely to implement a similar workflow during the initial stages of establishing a repository. These steps will typically be performed in a semi-manual, ad-hoc manner. The disaggregated model has been built around an assumption that while these tasks are necessary, they must be formalised to a greater extent, so that ingest can be automated.

In the sequence listed above, the key steps to perform are:

### 1. Validate the e-print

Perform validation checks to confirm the e-print submitted by the depositor is complete, stored in a file format that can be read by repository staff, and have not been corrupted during deposit.

### 2. Enhance metadata

Depositor-created metadata is unlikely to comply with the repository's metadata requirements. Repository staff are likely to be required to complete missing elements or reformat existing text.

### 3. Migrate data to dissemination format

Repository staff should clarify procedures regarding the migration of deposited data to a suitable dissemination format. Procedures should indicate the file format and version to be created and software used to create it. Scenarios should be considered, such as the possibility that data submitted by a depositor will be held within multiple files, and the appropriate action (e.g. merge intellectual content into a single file).

### 4. Create provenance metadata

Details of any migration should be recorded within the relevant metadata elements. Staff should record the agent (a person, group, or software process) responsible for the conversion, the start and completion date of conversion, details of the actions performed, and indication of the software used to perform the conversion<sup>3</sup>.

### 5. Disseminate the e-print

Staff should perform whatever actions are necessary to make the e-print available to the researcher. For example, changing the availability status within the catalogue record.

### 6. Transfer Schedule

A schedule for transfer of submitted e-print and associated metadata must be established between the two parties. Transfer is likely to be initiated by the Preservation Service and may be performed when the metadata record for the e-print is made available for harvest, or based upon a pre-defined schedule (e.g. every four weeks)<sup>4</sup>.

### 7. Create preservation metadata and map into the METS framework

On ingest into the Preservation repository, an automated software tool should extract appropriate technical metadata and map harvested metadata to the METS framework<sup>5</sup>

### 8. Perform Risk Analysis

The file format of data held within the Updated SIP should be examined and a risk analysis generated.

### 9. Implement migration strategy

The Preservation Service must develop procedures and implement automated tools to migrate submitted data into a suitable preservation format (most likely PDF-A or RTF). The Preservation Service should accommodate existing policies regarding the file formats accepted by the E-print archive and establish possible limits on the preservation action that may be taken. It is likely the E-print archives will encourage the deposit of the publishers PDF and the Preservation Service should not oppose this. However, they should be made aware of any restrictions that may restrict the available preservation actions (e.g. the publishers template may be copyrighted material and cannot be migrated).

Any migration performed upon the e-print should be recorded, indicating the agent responsible, the date of completion, and a description of the performed action<sup>6</sup>.

### 10. Generate Archival Information Package (AIP)

Given that the primary purpose of discovery metadata is to locate an e-print, the Preservation Service must create additional information necessary to preserve and authenticate the

---

<sup>3</sup> Institutional repositories do not currently store details of any conversion performed during ingest. The Preservation Service will work with repository staff to implement extensions for provenance metadata.

<sup>4</sup> The Preservation Service should not be expected to provide immediate data backup on ingest. Existing institutional backup policies should continue to duplicate data held in the short-term.

<sup>5</sup> Further details on the options available to connect and transfer data/metadata may be found in work package 5.2 & 5.3.

<sup>6</sup> Details of the minimum metadata set for e-print are included in a forthcoming project output.

content of the e-print<sup>7</sup>. The disparate metadata created at the e-print archive and subsequent transfer to the preservation service should be imported into the METS framework.

### 11. Obsolescence Monitoring

The DIP should be monitored periodically for obsolescence. The practical implementation is unclear, however, it is expected that a software tool will examine metadata created by the E-print archive on the submission to dissemination format conversion and compare it to a list of obsolete formats, as identified by the Preservation Service or an external service provider<sup>8</sup>. If the dissemination format is found to be obsolete, the Preservation Service will inform the E-print archive and arrangements will be made to transfer a replacement dissemination format to the repository.

### 12. Create new Dissemination format

In conjunction with repository staff<sup>9</sup>, the Preservation Service will export the intellectual content and, if possible, the significant properties of an e-print and store them in a suitable dissemination format. A suitable dissemination format should be identified through careful monitoring of the Preservation Community and Archive User Community (see Figure 5).

The workflow offers a simplified outline of the actions to be taken, from initial deposit to dissemination to the user. In many circumstances, actions performed by the E-Print archive and Preservation Service will be performed concurrently and will repeat many times during the lifetime of the project. Further experimentation is necessary before a workflow can be finalized.

---

<sup>7</sup> The RLG & OCLC study on Trusted Repositories indicates When disseminating e-prints that have undergone some form of preservation action the repository must engender trust (RLG & OCLC, 2002) by enabling the identification of digital objects stored within Archival and Dissemination Information Package (AIP & DIP) as a derivative of a submitted digital object and ensuring that a digital surrogate has been faithfully created from the original through pre-established preservation practices.

<sup>8</sup> The AHDS may choose to partner with a third-party that possesses software able to identify obsolescence. Possible partners include The National Archive and its PRONOM tool.

<sup>9</sup> The e-print archive should make recommendations on possible dissemination formats, derived from their examination of the requirements of the user community. See the Administration entity on figure 5 (a detailed view of a disaggregated OAI-compliant model) for further information.

## References

- Beedham, H, Missen, J, Palmer, M & Ruusalepp, R. (2005). Assessment of UKDA and TNA compliance with OAIS and METS standards. Retrieved on August 15, 2005, from [http://www.jisc.ac.uk/index.cfm?name=project\\_oais](http://www.jisc.ac.uk/index.cfm?name=project_oais)
- CCSDS (2002). *Reference Model for an Open Archival Information System (OAIS)*. Retrieved on June 1, 2005, from [http://ssdoo.gsfc.nasa.gov/nost/isoas/ref\\_model.html](http://ssdoo.gsfc.nasa.gov/nost/isoas/ref_model.html)
- Edinburgh Research Archive  
<http://www.era.lib.ed.ac.uk/deposit-guide.jsp>
- Glasgow e-prints Service  
<http://eprints.gla.ac.uk/deposit.html>
- Granger, S., Russell, K. & Weinberger, E. (2000). *Cost elements of digital preservation*. Retrieved on June 1, 2005, from <http://www.leeds.ac.uk/cedars/documents/CIW01r.html>
- Institutional Archives Registry (2005). *Browse Statistics*. Retrieved on June 30, 2005 from: <http://archives.eprints.org/eprints.php?action=browse>
- James, H. et al, (2004). *Feasibility and Requirements Study on Preservation of E-Prints*. Retrieved on June 1, 2005 from: [http://www.jisc.ac.uk/uploaded\\_documents/e-prints\\_report\\_final.pdf](http://www.jisc.ac.uk/uploaded_documents/e-prints_report_final.pdf)
- Knight, G. (2004). *Report on a Deposit Licence for E-prints*. Retrieved on June 30, 2005 from: <http://www.sherpa.ac.uk/advice/licences.html>
- Lavoie, B. (2004) *The Open Archival Information System Reference Model: Introductory Guide*. Retrieved on June 28, 2005, from: <http://www.oclc.org/research/publications/archive/2000/lavoie/>
- Nottingham E-prints Service  
<http://eprints.nottingham.ac.uk/deposit.html>
- PDF-Archive <http://www.aiim.org/standards.asp?ID=25013>
- Research Libraries Group (RLG) & Online Computer Library Center (OCLC) (2002). *Trusted Digital Repositories: Attributes and Responsibilities*. Retrieved on June 1, 2005, from <http://www.rlg.org/longterm/repositories.pdf>
- Royal Holloway Research Online  
<http://www.rhul.ac.uk/information-services/library/eprints/deposit-guide.html#convert>
- SOAS Library E-Prints Repository  
<http://www.soas.ac.uk/eprints/guide/index.html>
- ULCC & AHDS (2004). *DAAT: Digital Asset Assessment Tool*. Retrieved on June 1, 2005, from <http://ahds.ac.uk/about/projects/daat/>
- Van de Sompel, H. Nelson, M. Lagoze, & Warner, S. (2004) *Resource Harvesting within the OAI-PMH Framework*. Retrieved on June 28, 2005, from: <http://www.dlib.org/dlib/december04/vandesompel/12vandesompel.html>
- White Rose University Consortium  
<http://www.leeds.ac.uk/library/sherpa/deposit.html#File>