

Storage assessment and Synchronisation mechanism Review

Kirti Bodhmaghe
Arts and Humanities Data Service

Summary

This is a combine report on storage assessment for the planned preservation archive and synchronisation mechanism between institutional archive and preservation archive. The storage for preservation archive has been estimated by considering the current size and future expansion of the archives participating in Sherpa-DP project. The synchronisation mechanism review investigates various methods to synchronise the contents between institutional archive and preservation archive.

Storage assessment for preservation archive

The storage and backup of digital data is a key component of preservation and should be the first consideration when planning a preservation repository. The organisation should have a clear understanding of the current and future requirements for storage to allow them to develop a suitable funding model for the project. The role of a preservation archive is to act like a master archive, which is self-sufficient in preserving any file for longer duration, and to provide a backup service. This document outlines the current and future storage requirements to store items held by the 14 institutional repositories participating within the SHERPA DP project.

The storage requirement of a digital archive is based upon two factors – the file size and total number of items held. Based upon these figures the approximate size of the archive may be calculated.

Archive	No. of e-prints	Average file size	Approximate size of archive
Nottingham EPrints + Etheses Archive + Modern Languages and Publication Archive	-	500 Kb	746 Mb
London LEAP Birkbeck University	129(full text archive)	300 Kb	-
London LEAP King's college	41(full text archive)	300 Kb	-
London LEAP LSE	142(full text archive)	300 Kb	370Mb
London LEAP SOAS	25(full text archive)	300 Kb	-
London LEAP Royal Holloway	67(full text archive)	300 Kb	-
London LEAP UCL	860(510 full text+ bibliography records) 1265 Total records	300 Kb	-
White Rose Consortium	614	563 KB	350Mb
Glasgow EPrints	366 full text (1712 total records)	300 Kb	110Mb

Jelit Glasgow EPrints	20	300 Kb	10 Mb
Erpanet Glasgow EPrints	46	300 Kb	30Mb
Edinburgh Research Archive	600 (only full text)	2 MB	-
Estimated storage for preservation archive	1.6Gb		

Table 1: Current size

It is estimated that 1.6Gb of storage space is currently required to store and backup the digital holdings currently held by institutional repositories. The storage requirements of the institutional repositories are likely to increase as a result of higher use within the institutions and various digitisation projects. Table 2 indicates the amount of data that repository staff expects to hold by 2010.

Archive	Estimated No. Of e-prints in archive	Average file size	Estimated size of archive
Nottingham EPrints + Etheses Archive + Modern Languages and Publication Archive	10,000 records		5 GB
London LEAP Birkbeck University	5000 (for 6 London Leap repositories)	-	8Gb (for 6 London Leap repositories)
London LEAP King's college	-	-	-
London LEAP LSE	-	-	-
London LEAP SOAS	-	-	-
London LEAP Royal Holloway	-	-	-
London LEAP UCL	-	-	-
White Rose Consortium	-	1.6Mb	-
Glasgow EPrints	5000 full text records and large collection of bibliographic records	-	-
Jelit Glasgow EPrints	-	-	50 MB
Erpanet Glasgow Eprints	-	-	-
Edinburgh Research Archive	Around 5000 full text	-	10 GB
Estimated storage for preservation archive	Around 26 GB		

Table 2: Predicted growth

The estimates provided by repository staff suggest the AHDS will be expected to provide 26Gb of storage space for their preservation archive within the next five years – a figure unlikely to cause any problem.

Storage Infrastructure at AHDS

A suitable infrastructure must be developed to ensure that data provided by institutional repositories is stored in a suitable storage environment and data management activity to be carried out periodically. The AHDS network will be composed of three components that offer different levels of access and space:

1. *Preservation server* – server that will run preservation service software and will be accessible by users external to the AHDS.
2. *Preservation store* – A SAN (Storage Area Network) unit hosting preservation archive.
3. *SRB framework* – A remote SRB (Storage Resource Broker) server provided by CCLRC will be used to replicate the data.

Figure 1 outlines the conceptual interaction of these servers.

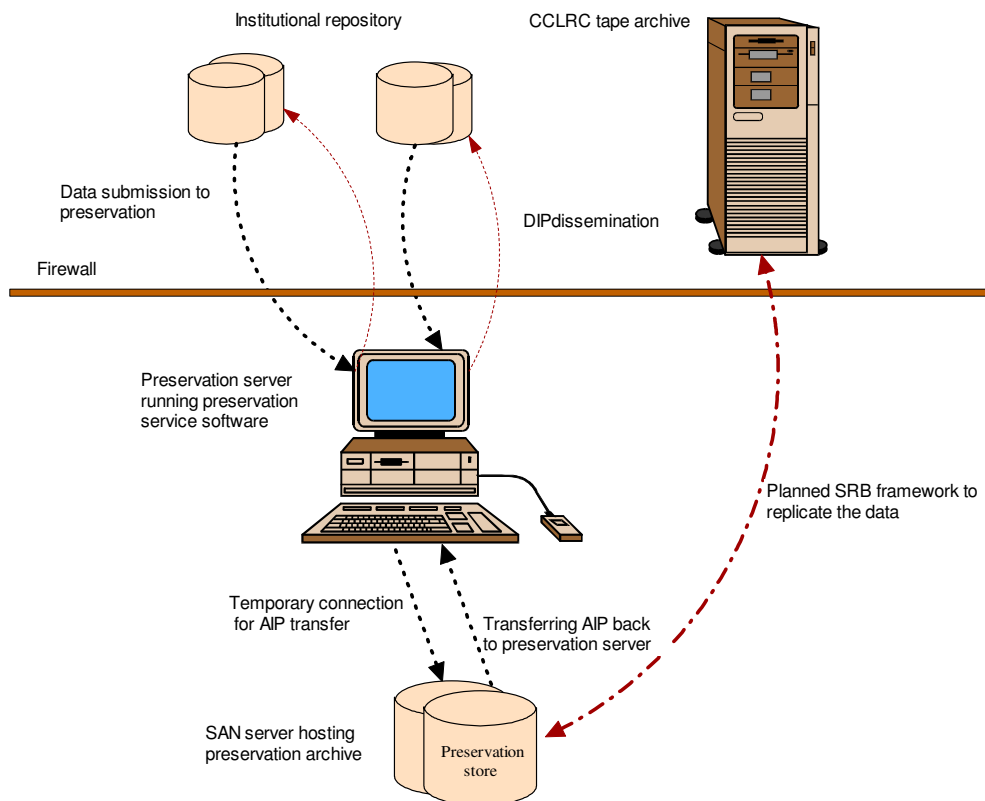


Figure 1: System diagram of Institutional repository-Preservation archive interaction

The preservation process begins when data held by institutional repository is extracted and written to the preservation server. The extracted data is processed to prepare it for storage as an Archival Information Package (AIP)¹. The AIP is transferred to the SAN server. Data replication will be performed on a regular basis to tape, as well as transfer to a secure facility via SRB (under negotiation).

¹ A workflow for preservation is detailed in the Work package 2.2. SHERPA DP-OAIS report.

At a later date, institutional repository may require a modified copy of the item and will be able to request a copy from the preservation service. In this scenario, the request will be queued for action by AHDS staff and a dissemination version will be generated from the AIP stored on the preservation store. Data may be copied to the preservation server for download by institutional archive staff or transferred directly to institutional archive storage area by preservation service.

Synchronisation Mechanism Review

Data synchronisation refers to mechanisms that may be used to transfer the e-print and associated metadata between the institutional repository and preservation service. It should perform three major functions:

1. To identify recent updates in the institutional repository
2. To extract data from the institutional repository and 'pull' data to the preservation server
3. To 'push' a DIP to the institutional repository on request

Several approaches may be identified to achieve these requirements. These are assessed according to existing functionality offered by the EPrints and DSpace software.

1. Identify new submissions to the institutional repository

The preservation service must be able identify new e-prints or updates to existing e-print submitted into the repository. The method to get the updates from the institutional archive to preservation archive can be implemented using OAI-PMH or RSS feed functionality, both of which are already in place in EPrints and DSpace.

1. Using OAI-PMH

The Open Archives Initiative Protocol for Metadata Harvesting enables XML request and response using HTTP protocol. OAI-PMH offers a series of commands that may be used to request information, including ListMetadataFormats, ListRecords, and GetRecords. Additional parameters may be used to specify criteria for the resulting output. The ListRecords method may be used to display metadata between two-time period and type of metadata output it requires (DC). For example, metadata for e-prints submitted after 01/11/2005 may be retrieved by entering the following syntax:

Nottingham EPrints

http://eprints.nottingham.ac.uk/perl/oai2?verb=ListRecords&metadataPrefix=oai_dc&from=2005-11-01

Edinburgh DSpace

http://www.era.lib.ed.ac.uk/dspace-oai/request?verb=ListRecords&metadataPrefix=oai_dc&from=2005-11-01

Metadata in the test DSpace archive at AHDS can be harvested in METS format using OAI-PMH as

<http://ahds.ac.uk/dspace-oai/request?verb=ListRecords&metadataPrefix=mets&from=2005-11-01>

The Preservation service may issue ListRecords requests with parameter 'from' where from=last_update_date and last_update_date is stored into the management database in the preservation service.

2. Using RSS feed functionality

RSS is a format for syndicating news and the content of news-like sites, which updates the sites on regular basis. RSS feed creates an XML document of the latest additions in RDF format.

EPrints archive

EPrints software has a script called latest_tool in cgi directory, which gives latest 20 items in EPrints repository in RDF format. RSS output from the test EPrints archive at AHDS can be viewed using the URL:

http://scout.ahds.ac.uk:86/perl/latest_tool?output=rss. Further changes are necessary (and have been made) to remove the 20 item restriction and to allow it to identify the latest additions to the institutional archive that have not been harvested by the preservation service, dependent upon the value of 'pres-serv' in the database. The following syntax returns the RDF output for e-prints that have a pres-serv flag = 'no'.
http://scout.ahds.ac.uk:86/perl/Search_Chgs?output=rss

DSpace archive

RSS functionality is available for DSpace through a separate patch, available on sourceforge.net. RSS feed is available for collection as well as community level. The patch generates a RDF file for all the items in the archive. However, it needs to be modified to include only the most recent additions in the archive, as indicated in the last_updatation_date. RSS output for AHDS collections can be retrieved by using following syntax:

<http://ahds.ac.uk/dspace/rss/1/2/123456789/3.rdf>

Figure 2 outlines the sequence for identifying new submissions into the institutional repository.

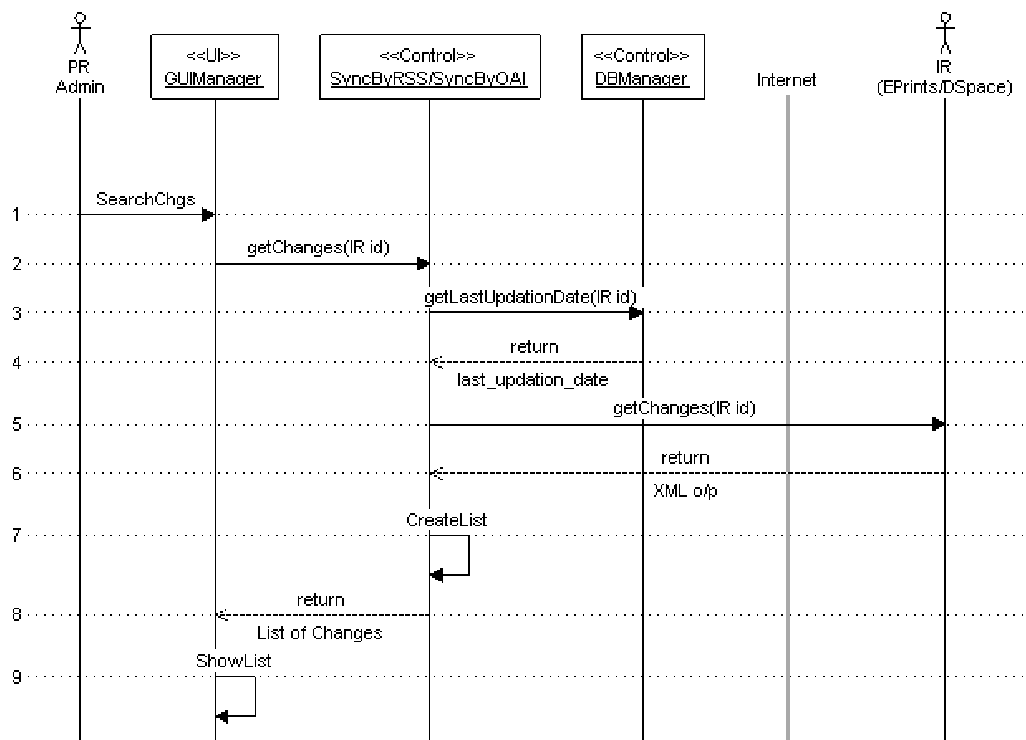


Figure 2: Interaction diagram to identify changes in IR

Recommendation: To identify if new submissions into the IR have been transferred to the preservation archive, it is recommended that an additional Boolean field 'pres_serv' be added to the database. The Boolean flag 'Yes' will identify items that have been transferred to the preservation archive, while those marked 'no' will require subsequent transfer.

2. Pulling the data from institutional archive

Mechanisms that will allow preservation service to pull data from institutional archives have been separately discussed in the EPrints and DSpace review documents. A preferred solution will be designed and implemented in a later work package.

3. Allow IR staff to request a new DIP and ‘push’ it to the institutional archive

The preservation service will allow institutional archive staff to request new versions of existing DIP. Workflow to generate a suitable dissemination package from the archival copy and transmit it to the institutional repository is currently being investigated. Once received, the archive staff will be able to review the new DIP and start the process necessary to make the e-print accessible to the public.

EPrints archive

Archive administrator logs in to UserAreaPage that has a link to check-DIP. If the DIP is Present then a link to start import_xml will be displayed otherwise an appropriate message will be displayed. As a result of import_xml, e-print will be available in submission buffer which archive administrator has to review and move to archive.

DSpace archive

DSpace software includes a batch tool to import the items in simple directory structure where Dublin Core metadata is stored in an XML file. The ItemImport class enables to add, remove and replace the items in a collection. Item can be imported in more than one collection. METSImport is under development at MIT and will be reviewed to see if we could import metadata and content in METS format. Archive staff logs in to MyDSpace page and check for the DIP. If DIP is present then link to import-item will be displayed. Import interface in DSpace can bypasses the submission workflow and item can be accessible to the public.

Sequence for the push scenario can be summarised as follows:

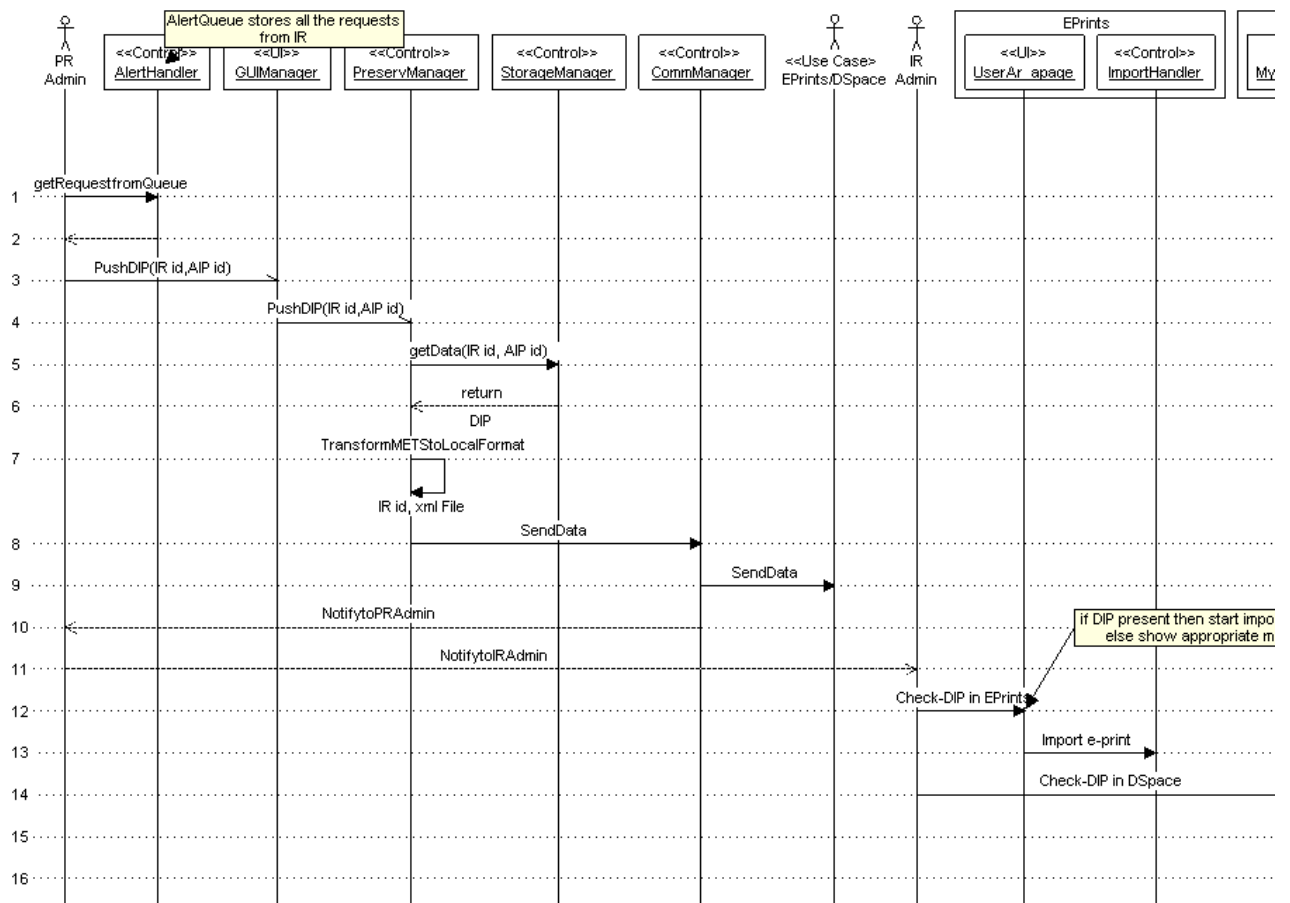


Figure 3: Interaction diagram to disseminate the DIP

Glossary

Terms	Definition
AIP (Archival information package)	Archival information package is a term used in OAIS model to describe the entity, which will be preserved long term. AIP will consist of the original file, preservation version of file and metadata associated with both files.
DIP (Dissemination Information Package)	An Information Package, derived from one or more AIP that are intended for use by the Researcher (Consumer). It will consist of original data file, migrated version if any and metadata file.
RDF (Resource Description Framework)	It is file format for Resource Description Framework, which is used to describe internet-based resources.
SAN (Storage Area Network)	A high speed dedicated network that interconnect different data storage device and has software to configure, monitor and manage the associated hardware.
Preservation archive	The preservation storage that store metadata and content (AIP).
Preservation service	The software, which will be developed to automate the preservation workflow.

References

1. RSS patch is available at

http://sourceforge.net/tracker/index.php?func=detail&aid=1160997&group_id=19984&atid=319984