SHERPA-DP: Distributed Repositories/Distributed Preservation

DAAndrew Wilson
Preservation Services and Projects Manager
Arts and Humanities Data Service
Topics

- Sherpa
- Background and Overview of Sherpa DP
- OAIS (very briefly)
- Disaggregated (distributed) Preservation Service
- Workflow
Sherpa

- 2002-05 JISC project under the Focus on Access to Institutional Resources (FAIR) programme, which aimed to:
  - establish OAIS compliant institutional open access e-print repositories in 20 partner institutions
  - investigate key issues in creating, populating and maintaining e-print collections
  - work to achieve technical, metadata and collection management standards for the effective dissemination of the content
  - investigate digital preservation of e-prints using the Open Archival Information System (OAIS) Reference Model
Sherpa DP Project

- **acronym**: Securing a Hybrid Environment for Research Preservation and Access: Digital Preservation
- **project partners**: AHDS at King’s College London (Lead), Nottingham, Glasgow, Edinburgh, White Rose Consortium, London Leap Consortium
- **duration**: 2 years, March 2005 – February 2007
- **funding**: JISC and CURL
- **JISC programme**: Supporting Digital Preservation and Asset Management in Institutions (4/04)
Sherpa DP Aims

- develop prototype preservation environment for the Sherpa DP partners based on the OAIS reference model, including shared protocols and software tools
- establish a comprehensive workflow and set of procedures to suit the needs of institutional repositories and the preservation service
- provide guidance on the ingest process in order to encourage the deposit of file formats that will minimise long-term operational costs and maximise preservation potential
- develop an exemplar for an outsourced preservation service
- create a digital repositories handbook that will set out best practice standards and processes for resource creation and ingest, preservation planning and management, and provision of access for the holdings of institutional e-print repositories in the UK
Sherpa DP Methodology

- map the six entities of an OAIS-compliant repository (ingest, archival storage, administration, data management, preservation planning and access) onto an existing structure
- model the implementation of a disaggregated preservation service within the OAIS framework
- identify rights and responsibilities, services and actions, and apportion these between the IR and preservation repository service
- develop tools and processes to implement the preservation services and actions
OAIS Functional Model as applied by Sherpa DP
Why disaggregate preservation functions?

- Institutional repositories lack the time to implement preservation
- Scarcity of staff with necessary preservation skills and expertise
- Seeking to remove duplication of services
- Potential cost savings in terms of staff time and equipment
- Preservation is not inherent in most repository software
- DSpace and EPrints software primarily about submission, basic storage and access
Repository Landscape

Leeds, York & Sheffield

Nottingham EPrints

Glasgow ePrints Service

Birkbeck, Imperial, Kings, LSE, Royal Holloway SOAS, UCL

London Leap

Preservation Service

EPrints Software

EPrints Software

EPrints Software

EPrints Software

Electronic Thesis & Dissertations

DSpace Software

Edinburgh Research Archive
## Quantities

<table>
<thead>
<tr>
<th>Archive</th>
<th>Total number of e-prints in archive</th>
<th>Average file size</th>
<th>Approximate Size of archive</th>
<th>Estimated growth around next 5 years</th>
</tr>
</thead>
<tbody>
<tr>
<td>Nottingham EPrints + Etheses + Modern Languages Publication Archive</td>
<td>-</td>
<td>500 KB</td>
<td>746 MB</td>
<td>10,000 records (5 GB)</td>
</tr>
<tr>
<td>London LEAP Birkbeck University</td>
<td>129 (full text archive)</td>
<td>300 KB</td>
<td>-</td>
<td>Expected to grow to 5000 items per year for London LEAP</td>
</tr>
<tr>
<td>London LEAP King’s College</td>
<td>41 (full text archive)</td>
<td>&quot;</td>
<td>-</td>
<td>8 GB total for London LEAP</td>
</tr>
<tr>
<td>London LEAP LSE</td>
<td>142 (full text archive)</td>
<td>&quot;</td>
<td>370MB total size</td>
<td>-</td>
</tr>
<tr>
<td>London LEAP SOAS</td>
<td>25 (full text archive)</td>
<td>&quot;</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>London LEAP Royal Halloway</td>
<td>67 (full text archive)</td>
<td>&quot;</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>London LEAP UCL</td>
<td>860 (510 full text+ bibliography records)</td>
<td>&quot;</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>White Rose Consortium</td>
<td>614</td>
<td>500 KB</td>
<td>300MB</td>
<td>File size is expected to grow to 1.5 MB</td>
</tr>
<tr>
<td>Edinburgh Research Archive</td>
<td>600 (only full text)</td>
<td>2 MB</td>
<td>3.5 GB</td>
<td>Around 5000 full texts. Expected size: 10 GB</td>
</tr>
<tr>
<td>Total Size (estimated) on the preservation server</td>
<td></td>
<td></td>
<td></td>
<td>5-6 GB</td>
</tr>
</tbody>
</table>

**Total Size (estimated) on the preservation server**

5-6 GB Around 25 GB
Establishing responsibility

- Who is responsible for creating the AIP?
  - preservation service, institutional repository, or both?
- What type of metadata is created & needed?
  - descriptive, technical, structural & administrative metadata
- How will AIP be used?
  - identification of at-risk formats, migration
- When will the AIP be created?
  - on ingest, schedule, or when the resource is at-risk
Establishing responsibility: Institutional Repository

- implement appropriate repository software (*all use Eprints except Glasgow which uses DSpace; AHDS uses Fedora*)
- develop selection, retention and ingest policies
- develop a rights framework
- specify a minimum metadata set, and provide details to the Preservation Service
- quality control for descriptive metadata
- support mechanisms for metadata harvest
- support for extension schemes to enable preservation.
- creation of technical metadata (possibly)
- alerting mechanisms for updated/additional content?
Establishing responsibility: Preservation Service (AHDS)

Storage:
- provide a permanent storage facility and disaster recovery capabilities
- manage storage hierarchy

Preservation Planning:
- Evaluate contents of archive and undertake risk assessment
- develop recommendations for preservation standards and policies
- life cycle management. Monitor changes in technology environment, users’ service requests, and knowledge base

Preservation Action:
- develop and implement migration plans
- create and manage multiple copies of objects, including off-site storage
- record appropriate information on any changes to the objects
Data transfer
(IR ↔ AHDS)

- investigate methods to identify new submissions or new content in IR.
- implement transfer mechanisms between institutional repositories and preservation service (DSpace and Eprint APIs, storage layers and module add-on capabilities)
- examine the capabilities of OAI-PMH for complex object formats
Create new, or refine existing, automated tools to perform:

- file format migration
- metadata extraction
- obsolescence checking and migration services
- identification and tracking of versions
- synchronisation of versions across repositories
- integrity checking and reporting
Components of technical solution

- metadata addition to EPrints software
- plug-ins for DSpace and EPrints
- java scripts to manage retrieval of metadata and data from IRs
- use of METS for storing metadata
- Fedora to be implemented in preservation repository
- customise existing web interface for Fedora
Simplified workflow (preliminary)
Further Information

URL:
http://www.ahds.ac.uk/about/projects/
http://www.sherpadp.org.uk

Contact
andrew.c.wilson@ahds.ac.uk
gareth.knight@ahds.ac.uk