Smart Storage and Preservation: How Digital Repositories Can Participate

Sally Rumsey: University of Oxford
Steve Hitchcock: University of Southampton
Neil Jefferies: University of Oxford
Adrian Brown: The National Archives

DLF Forum: November 2008
Crystal ball by Hamachi! CC license. Attribution-Non-Commercial-No Derivative Works 2.0 Generic Available at http://www.flickr.com/photos/mawari/2091456761/
Preserv: Inspiration and Provision of Preservation Services for Digital Repositories

"it is important to build the concept of preservation from the outset*. In the digital era, the 'outset' for most new research and educational materials will be the institutional repository."*

About the project

Fast facts

- **When**: Preserv 2 from July 2007 to December 2008 (Preserv 1 ran from February 2005 to January 2007)
- **Who**: Southampton University (School of Electronics and Computer Science), The National Archives, The British Library and Oxford University (Library Services, Systems and Electronic Resources Service) ... collaborating people.
- **Prior related work by the partners**:
  - Southampton: EPrints repository software, TARDis project, a sustainable multidisciplinary institutional repository
  - TNA: digital preservation at TNA, PRONOM, the file format registry
  - BL: Digital Object Management (DOM) Programme
- **Funded by JISC**, within the programme Supporting Digital Preservation and Asset Management in Institutions
- **Keywords**: digital libraries, institutional repositories, preservation, preservation services, EPrints, PRONOM

Coverage of Preserv

May 1, 2007 Project video. Preserv Us! The story of the Preserv project by the people
Capital programme

The capital programme is additional funding of £90 million over three years to enhance the network infrastructure (SuperJANET 5), to digitise key resources for the academic community, and to support the development of e-learning; e-infrastructure; virtual research environments; users and innovation; and repositories and preservation. Funding for SuperJANET 5 has been provided by all JISC’s funding partners. However, the remaining activities have been funded by the Higher Education Funding Councils for England and Wales (HEFCE and HEFCW).

Funding opportunities 2006/2007

See projects funded under the capital programme.

See the roadmap for calls, summary information for the background to the programme and progress of the capital programme.

July 07: Enterprise Architecture, e-Research: e-Infrastructure and Users and Innovation

Circular 02/07 released 27 July 2007 (Deadline: 2 October 2007).

April 07: e-Learning, Repositories and Preservation and e-Research: e-Infrastructure

Circular 01/07 (Deadline: 21 June 2007) Circular briefing day (9 May 2007, previously known as a Town Meeting) held on for those interested in submitting proposals - FAQs and copies of the presentations are available to download. See supporting information.

September 06: e-Learning, Users & Innovation, Repositories & Preservation and e-Research: e-Infrastructure and Virtual Research Environments
Welcome to Planets

Planets, Preservation and Long-term Access through Networked Services, is a four-year project co-funded by the European Union under the Sixth Framework Programme to address core digital preservation challenges. The primary goal for Planets is to build practical services and tools to help ensure long-term access to our digital cultural and scientific assets. Planets started on 1st June 2006. This website makes available project documentation and deliverables as Planets progresses so that these can be shared with the libraries, archives and digital preservation community.

The strong Planets consortium brings together expertise across Europe from national libraries and archives, leading research universities and technology companies. Coordinated by the British Library, the partners are:

- The British Library
- The National Library of the Netherlands
- Austrian National Library
- The Royal Library of Denmark
- State and University Library, Denmark
- The National Archives of the Netherlands
- The National Archives of England, Wales and the United Kingdom
- Swiss Federal Archives
- University of Cologne
- University of Freiburg
- HATII at the University of Glasgow
- Vienna University of Technology
- Austrian Research Centers GmbH
- IBM Netherlands
- Microsoft Research Limited
- Toscaida Support Services Plc
Overview

Formats matter if digital content is to be accessible now and preservable in the longer term. Institutional Repositories (IRs), which provide access to and store digital objects produced by many creators, will need to manage a range of formats. This briefing paper explains how formats affect preservation, considers which formats repositories should use for deposit and storage, and describes the practical steps repositories can take to produce an initial preservation plan.

How are formats used?

Digital documents are produced, in one form or another, using an application program such as a word processor. These documents are encoded with information to represent characters, layout, and other features. The rules of the encoding are defined by the chosen format of the document. Applications are often closely tied to formats, e.g., Microsoft Word can be used to produce the document (.doc) format, Adobe Acrobat produces the portable document format (.pdf). These may not be the only formats that an application can produce, e.g., Word can also produce Rich Text Format (.rtf), and formats may not be exclusive to one application.

Why are formats important for preservation?

Problems with application-specific formats can arise when users try to open a digital document without access to the application that was used to create it, or without the correct version of an application. This is most likely to happen when opening a document created by someone else. This problem increases over time, that is, it becomes harder to open documents in their original format if the application has changed or no longer exists.

If applications and formats can change over time, it follows that some risk becoming obsolete. This is why formats are a primary focus for preservation actions, and why repositories need to be aware of the formats of the digital objects they store.

Which deposit formats should an IR allow?

There are many different types of digital objects (e.g., texts, images, videos), and many different applications for producing them. There are also different views on which formats are the most “preservable”, and therefore which formats a repository should allow to be deposited. There is one format that an IR should always commit to obtaining: the author’s source format. That is, the version produced by the author directly from the application used at the time of completion.

The most common example of deviating from this approach is a requirement for authors to deposit PDF, which is not an authored format — it is created by converting from another format. By requiring authors to submit the source format for preservation, the repository can then convert to its preferred presentation format, which could be PDF, if that is different from the source format. It is likely this conversion can be automated and, in the process, documented.

Which formats should repositories commit to support in the long-term?

The key phrase that describes the ideal longer-term storage format is open standard, meaning the specification is freely available and implementable. Consequently it is more likely that applications to view and use such formats will be available at any given time, since viewers can be developed by the wider community of users with an interest in the format, and not just the original application developer. Open standard formats include OpenDocument format (ODF), an XML file format for electronic office documents.
Most of the experience of humanity is contained in the past. Medieval and modern history has been studied at Oxford for longer than at almost any other university; a Regius Professor of Modern History was first appointed in 1724, and undergraduate examinations began in 1850.
TNA’s Seamless flow approach to preservation

1. Format identification and characterisation
2. Preservation planning and technology watch
3. Preservation Action
Long Term Storage

Thanks to Dave Tarrant
Southampton University
Open Storage

1. Repository layer

2. Digital object

3. Built-in preservation support

Address object
Smart storage combines an underlying standard storage approach with intelligence provided through services.
Smart storage for preservation in action:
Format identification and characterisation using DROID-PRONOM

Service: DROID
File format & version identification

PRONOM
File registry
Format risk

Communication

Preserv2 Wrapper

Locate new object for classification

Reports

Web Server
(open or restricted)
Outputs/results
(XML docs) DROID

Preserv2 Plug-in

Open storage

OAI-PMH

Repository
Fedora/ePrints
Calendar Service
eg iCal/Outlook
Managing; notification

Darwin Calendar Server
Set schedules
Trigger events
History/provenance

Web Server
(open or restricted)
Outputs/results
(XML docs) from
calendar and DROID

Service: DROID
File format & version identification

PRONOM
File registry
Format risk

Preserv2 Wrapper
Communication

OAI-PMH

Repository
Fedora/ePrints

Open storage

Preserv2 Plug-in

Plug-in aware
of event occurrence

Action
Eg planning
incoming info

Action
Eg migrate; no action

Record of event & location of results posted in calendar

Trigger invokes DROID action

Locate new object for classification

Record of event & location of results posted in calendar

Trigger invokes DROID action

Imports DROID classification into metadata

Notifies
Reports

Adding the scheduler
Continuum of provision of repository preservation services

- Preservation services all in-house
- Some preservation services supplied in-house, others outsourced
- Preservation services all provided by external suppliers
CRIG

Contents
1 JISC Common Repository Interfaces Group (CRIG)
2 Workplan
3 Membership
4 Events (Chronological Order)
5 Channels of Communication

JISC Common Repository Interfaces Group (CRIG)

"The coolest thing to do with your data will be thought of by someone else"

JISC set up the Common Repository Interfaces Group (CRIG) to help identify problem spaces in the repository landscape and suggest innovative solutions. The CRIG consists of a core group of technical, policy and development staff with repository interface expertise. It encourages anyone to join who is dedicated and passionate about surfacing scholarly content on the web. Please see below for joining details.

Workplan

For an interactive version of the diagram below (i.e. the image as a 'table of contents' that links to important pages throughout this wiki) please see: http://www.flickr.com/photos/wocrig/2468077865/
Interoperability in Action

OAI-ORE
EPrints & Fedora

Which is which?
Preserv2 Project Structure

- **Services**
  - TNA API - PRONOM
    - File Format Identification
    - Significant Properties
    - Migration Tools
    (Performance Metrics)
  - Scheduler (Oxford)
    - Services & Invocation API
  - Interoperability
    - OAI-ORE Specification & Mapping

- **Repository Software**
  - EPrints
  - Fedora
  - DSpace

- **Storage Controller**
  - EPrints
  - Fedora

- **Physical Storage**
  - Local Disk
  - Remote Server
  - Cloud Service
  - Honeycomb
  - EPrints
  - Fedora

Application Program Interface (API) + XML
Relation Exclusivity (1 to 1, 1 to Many)
Three fundamental preservation actions on data

Copy

Monitor

Move
This EPrints install is referencing a trial version of the risk analysis service. None of the risk scores are likely to be accurate and thus should not be used as the basis for a program of action.

**High Risk Objects**
- OLE2 Compound Document Format
- **User**: Mr David C Tarrant
- **No of Files**: 1
- **Title**: Towards smart storage for repository preservation services
- **EPrint ID**: 4
- **File**: Hitchcock_a60_doc.doc (278Kb)

**Medium Risk Objects**
- Microsoft Powerpoint Presentation (Version 97-2002)
- **No of Files**: 3

**Low Risk Objects**
- Portable Document Format (Version 1.4)
- **No of Files**: 3
- Portable Document Format (Version 1.3)
- **No of Files**: 2
- ZIP Format
- **No of Files**: 2
PRESERV website

http://preserv.eprints.org/

Sally Rumsey  sally.rumsey@ouls.ox.ac.uk
PRESERV Project Manager: Steve Hitchcock
sh94r@ecs.soton.ac.uk