Preserv.org.uk
Repository Preservation and Interoperability

JISC

Project Document Cover Sheet

Project Information			
Project Acronym	Preserv 2		
Project Title	(PReservation Eprint SERVices): towards distributed preservation services for repositories		
Start Date	July 2007	**End Date**	March 2009
Lead Institution	University of Southampton		
Project Director	Les Carr		
Project Manager & contact details	Steve Hitchcock sh94r@ecs.soton.ac.uk, Tel. 023 8059 7698		
Partner Institutions	University of Oxford, The National Archives, The British Library		
Project Web URL	http://preserv.eprints.org/, http://preserv.org.uk		
Programme Name (and number)	Capital programme 04/06		
Programme Manager	Neil Grindley		

Document Name			
Document Title	*Final Report*		
Reporting Period	To end March 2009		
Author(s) & project role	Steve Hitchcock, project manager		
Date	23 July 2009	**Filename**	preserv2-finalreport.doc
URL	http://preserv.eprints.org/JISC-formal/preserv2-finalreport.doc		
Access	☑ Project and JISC internal	☒ General dissemination	

Document History		
Version	**Date**	**Comments**
0.1	5 June 2009	Internal project and JISC version, first draft
0.2	9 June 2009	Minor changes, copied to project list, all partners, second draft
1.0	23 July 2009	Final. Permission to release from JISC

Towards repository preservation services

Final report from the JISC Preserv 2 project

Steve Hitchcock, David Tarrant and Les Carr

Contact for the report: Steve Hitchcock (sh94r@ecs.soton.ac.uk)
July 23, 2009

Table of Contents

Acknowledgements

PRESERV 2 was funded by JISC within its Capital Programme (http://www.jisc.ac.uk/whatwedo/programmes/capital) in response to the September 06 call (Circular 04/06), Repositories and Preservation strand.

PRESERV was based in the School of Electronics and Computer Science, University of Southampton, and worked with funded partners at Oxford University Library Services (Systems and Electronic Resources Service) and The National Archives. The British Library acted as an advisory partner.

We are grateful to Sun Microsystems for the provision of storage systems at Southampton and Oxford Universities, which has led to ongoing collaboration on open source storage software.

Find out more about Preserv 2 and its predecessor project at the project Web site http://preserv.eprints.org/

Executive Summary

Preserv 2 investigated the preservation of data found in digital institutional repositories. During the project the number of repositories has grown, as has the volume of material stored, and there is diversification in the types of materials, not just research papers but supporting data, teaching and learning content, arts and audio-visual content. In other words, the risk profile of repositories has changed completely.

Collecting, storing, copying and managing data, together with technical file format support, are the principal requirements for repository preservation. The objectives of Preserv 2 were to identify, build and test, as appropriate, tools and services to enable these processes to be performed cost-effectively by repositories.

Repositories are changing. Repositories both live on the Web and are used on the Web. The infrastructure available to manage content on the Web is expanding. Hardware, software, storage and data tools are not just changing but are being repositioned across a network of services in the ‘cloud’. The underlying assumption is of explosive data growth, not yet tested by many repositories, but with diversification we can assume this will happen.

In particular, storage has been a focus. Between local, mediated, open and cloud storage, offering choices of scale, bandwidth and cost, repositories have an increasing range of storage options. Rather than adopt a single storage approach, with growing data volumes and data types it is likely repositories will choose a combination of services, or 'hybrid' storage. Preserv 2 developed the first repository storage controller, which will be a feature of EPrints version 3.2 software (due 2009). Plugin applications that use the controller have been written for Amazon S3 and Sun cloud services among others, as well as for local disk storage.

It is simple to copy a digital file. It is less simple to copy a file complete with its context. In a breakthrough application Preserv 2 used OAI-ORE to show how data can be moved between two repository softwares with quite distinct data models, from an EPrints repository to a Fedora repository. "Repositories could become interfaces to remote storage where content is represented and accessed via an OAI-ORE resource map ... we can envisage the prospect of many repository softwares (EPrints, Fedora, etc.) running over one set of resources."

Preservation begins with policy. For institutional repositories one area of policy-making that is currently achieving prominence and growth is open access mandates. We examined a registry of such policies for evidence of policies leading towards preservation based on the thesis that preservation policy is more likely to follow from high-level policy initiatives, and these initiatives are more likely to be found for repositories with OA mandates. This is not strongly borne out in practice yet. Just 28% of all mandate policies posted by early March made any reference to preservation. We found that research funder policies are 2.4 times more likely to seek provision for preservation in return for the requirement to deposit papers than are mandate policies from institutions.

The largest area of work in Preserv 2 was on file format management and an ‘active’ preservation approach. This involves identifying file formats, assessing the risks posed by those formats and taking action to obviate the risks where that could be justified. This led to a

new approach combining file format management with automated storage management, which we called ‘smart storage’.

These processes were implemented with reference to a technical registry, PRONOM from The National Archives (TNA). Another tool, DROID (digital record object identification service), also produced by TNA and which can be downloaded and run locally, uses a signature file available via the PRONOM registry to classify files and provide specific details relating to individual files.

Preserv 2 showed we can invoke a current registry to classify the digital objects and present a hierarchy of risk scores for a repository. Critically, however, the registry does not yet have any data for the risks. It has recently been developed and updated with the facility to hold these risk analyses (PRONOM v7), and the detailed criteria for format risk analysis have been elaborated, but the format analyses have not been done yet. It was intended such analyses would be done for selected formats found commonly in institutional repositories.

Classification was performed using the Preserv2 EPrints preservation toolkit. This ‘wraps’ DROID in an EPrints repository environment, also providing a caching database (as an EPrints dataset) and a results page (presented as an EPrints Administration Screen). This toolkit will be another feature available for EPrints v3.2 software.

In tests the Preserv2 EPrints toolkit version was shown to classify more objects than the earlier Web-based classification (PRONOM-ROAR), reducing the number of unknown files. This is achieved primarily by deploying a more recent version of the DROID signature file (v13 against v12 used in PRONOM-ROAR), and by linking DROID more closely with the repository software can ensure successful completion without limiting file size.

The tests also produced findings critical of the tools. PRONOM was found to lack a complete set of MIME-types for its format data, and this is to be rectified. It is vital that tools intended to improve the reliability of digital data management are themselves shown to be reliable.

The result of file format identification, or an indication of a format change, can suggest a file is at risk of becoming inaccessible or corrupted. Preserv 2 developed a repository interface to present formats by risk category, and to enable information on high-risk formats and files at risk to be displayed. Providing risk scores through the live PRONOM service was shown to be feasible. Using thresholds, which can be adjusted by the repository administrator, the file formats are grouped into traffic light-style categories that clearly show the user the risks related to all of their objects, including those where no risk has been found.

To extend format identification services a ‘smart storage’ approach has been demonstrated. Smart storage combines an underlying passive storage approach with the intelligence provided through services. Additional tools - a calendar-based scheduler, a harvester to retrieve the latest stored content from a repository, messaging services to record the history of events and the results, all designed to work with, or 'wrap', the core format identification tool DROID - combine to manage and process the data and results.

Background

Digital preservation is about managing your data over time against identified risks. While there are a generic series of risks that apply to all digital data, the prioritisation of these risks will vary according to the needs and objectives of the organisation responsible for the data. This assessment will also depend on the scale and type of data, and its perceived value. Cost is invariably a constraint.

The Preserv project, most recently as Preserv 2, has investigated the preservation of data found in digital institutional repositories. Such repositories were in their infancy when Preserv began in 2004, and were intended primarily to provide access to research papers. Since then the number of repositories has grown, as has the volume of material stored (generate current charts from Registry of Open Access Repositories, http://roar.eprints.org/). More importantly, there is diversification in the types of materials stored, not just research papers but supporting data, teaching and learning content, arts and audio-visual content. In other words, the risk profile of repositories has changed completely.

Repositories are changing. The process of change for institutional repositories has begun but is not nearly finished. Repositories live and are used on the Web, and the infrastructure available to manage content on the Web is expanding. Hardware, software, storage and data tools are not just changing but being repositioned across a network of tools and services in the 'cloud'. The idea of the cloud is to let someone else manage the infrastructure, while you continue to manage the data but get the advantage of cost-effective scaling of technology and support. The underlying assumption is of explosive data growth, not yet tested by many repositories, but with diversification we can assume this will happen. Another assumption is that the infrastructure will be managed by a specialist organisation thereby reducing that element of risk, but every agent in the chain of data brings a different risk and raises issues of trust. Once again, the risk profile will change.

The core activities involved in managing any digital data are collection, store and copy. Broadly, if these are managed effectively the data should prevail, somewhere. Greater access and use are likely to be contributory factors in improving the chances of data survival. Working against this are barriers to use and restrictions on copying. Some data owners will object to data materialising somewhere not authorised by them. The risk profile adapts for such cases.

There is another factor affecting the continuing usability of data, and it reduces to a technical issue concerning file formats, but it essentially concerns whether a user can open a digital file on the computer they are using at a given moment. If that computer does not have an application, say a word processor or a video player, which can read the file then the user cannot access it. If most or all computers at that time do not have such an application then almost no user can access the data. In this situation we have the data, the digital bits, but no machine to interpret it and render it meaningfully to the user. For many files and formats this problem can be pre-empted by various means, such as format migration, but this is specialised and requires tools or services that we don't yet have. It's another factor to add to the risk profile.

In the course of monitoring the development of repositories Preserv 2 has had to adapt its strategy for supporting their preservation. For example, data growth and therefore storage has become a more prominent issue. The central philosophy of the project has been to identify what repositories are already doing as part of their everyday processes that contributes towards preservation, and then develop services to support and automate in those more specialised areas such as technical file format management.

Simply by bringing an institutional view and commitment to collecting, storing and managing data from across an institution improves the prospects of data survival compared with unmanaged Web servers dotted across that same institution. In that respect, a degree of preservation comes built into the institutional repository. Beyond that prioritising the risk profile is not straightforward. Repositories are changing, and many individual repositories do not yet have a roadmap to direct that change. Most have yet to construct sufficient formal policy to guide decision-making through rapid development and change, and they often work with undefined or short-term cost allocations.

It is against this background that Preserv 2 defined and adapted its primary objectives and plan of work.

Aims and Objectives

If collecting, storing, copying and managing data, together with technical file format support, are the principal requirements for repository preservation, the objectives of Preserv 2 were to identify, and build and test as appropriate, tools and services to enable these processes to be performed cost-effectively by repositories.

Objectives summarised from the plan:

- Store: report on the feasibility of providing a bitstream storage service
- Copy: investigate the interoperability of data, that is, the ability to copy data between different repository platforms running different software
- Manage: survey repositories for policy frameworks on which preservation risks can be assessed and planned and appropriate actions taken.
- Technical: to build and test services to support the technical preservation requirements of repositories covering file format characterisation, plan generation and preservation actions such as migration or emulation.

While the process of collection is clearly a responsibility of the repository, collection policy (for example, restrictions on file formats that can be deposited), interfaces by which preservation-related information might be provided to depositors and repository managers via repository software, and documenting resulting processes and actions through metadata, were also areas of interest in Preserv 2.

Methodology

There is no single approach to repository preservation, as can be judged from the range of preservation activities outlined and the project objectives. The aim was to identify best practices, and provide or enhance, and test, tools and services to support preservation activity either by the repository or by a service provider.

Our starting point was the Preserv 1 project, which had produced the following main outputs (Hitchcock, *et al*., 2007a):

- Modelled preservation services for repositories
- Preservation metadata: mapped a subset of PREMIS to the services model
- Format identification: developed PRONOM-ROAR: a Web service presenting format profiles of 200+ EPrints and DSpace repositories
- Surveyed repository preservation policy and activity
- Added preservation features to EPrints v3.0 repository software: a history module, METS plugin, rights declaration

In particular, the format identification approach had led to a promising Web services-based model combining bitstream storage and ‘active’ preservation, prompting the proposal for Preserv 2 with the prospect of “implementing and testing some of the services identified in the extended model.”

The project partnership meant we had access to some critical elements of the anticipated repository preservation framework: to EPrints and Fedora repository softwares through the involvement of Southampton University and Oxford University, respectively, in their development; to file format management tools, PRONOM-DROID, from the National Archives; and to preservation infrastructure being developed by the PLANETS European project involving both TNA and the British Library among our partners.

In addition, an opportunity arose to work with the then-new JISC Repositories Support Project (RSP), extending our links with repository managers and providing a route by which we could assist repositories with preservation, while at the same time learning about the challenges facing repository managers and understanding the context and constraints for our proposals.

Implementation

The world never turns out as you expect it to. So what actually happened?

Store

Sun, Honeycomb and H2 open source software

Investigation of the feasibility of providing a bitstream storage service was to be based on the Fedora harvesting and interoperability framework. What was required was a large-scale storage system. Sun Microsystems had acquired storage technology (STK5800) that it codenamed Honeycomb, which it was repositioning for its digital library market. The key features of this system were scale, fault tolerance due to a distributed node structure, and an ‘open storage’ approach based, in principle, on platform-independent open source software. A typical STK5800 server included:

- 16 nodes - each with 4 500Gb hard disks
- 1 service node - handles upgrades to software/firmwares (not needed to maintain access to the system)
- 2 switches - perform load balancing: one active, one failover

Two partners in Preserv 2, at Southampton and Oxford, took delivery of Honeycomb servers with the idea of ingesting content from local repositories to test the server. The original technology that became Honeycomb was not specifically designed for digital library applications, so initial implementation was not straightforward, an early clue to a possible market disconnect. Developers from Preserv 2 joined with Sun in California and at Sun's international PASIG conference to build a metadata framework for STK5800 capable of handling repository content.

Such was the scale of the storage available with a Honeycomb server that it could in principle store the content from all repositories worldwide, and there was a second clue to a market disconnect. Sun appeared to target the server at individual institutions rather than service providers, when there were cheaper alternatives, including from Sun itself, with a more appropriate level of storage for single institutional needs. In a Sun Webinar, Preserv 2 had begun to speculate about a possible network of local storage servers linked to large storage services (http://www.sun.com/solutions/documents/video/edu_webinar1.xml, March 2008).

In September 2008 Sun announced it was to end production of Honeycomb, although support would continue for existing systems. Initiated by Preserv 2 developers, work began to save and reuse the Honeycomb code, now known as H2 (http://wiki.preserv.org.uk/index.php/Honeycomb_H2), by exploiting its open storage feature. That work of that group continues today, led by Sun in its work on Object Oriented Storage space.

Cloud storage, more options and a storage controller

The evolution of online computing recognises that computing resources need not be local, and that also applies to storage. Amazon and Google have legitimised the concept of storage on machinery you cannot physically touch because it is somewhere on the world computing network, sometimes called the 'cloud'. More recently DSpace and Fedora repository softwares joined to form DuraSpace, an organisation offering to act as an intermediary for repositories seeking 'cloud' storage services.

Preserv 2 does not provide cloud storage, but noticed that between local, mediated, open and cloud storage, offering choices of scale, bandwidth and cost, repositories have an increasing range of storage options. Rather than adopt a single storage approach, with growing data volumes and data types it is likely repositories will choose a combination of services, or 'hybrid' storage. If there are storage options, we have to manage copy and transfer of content from the repository to the chosen locations, so Preserv 2 developed the first repository storage controller (http://wiki.eprints.org/w/StorageController) for EPrints software.

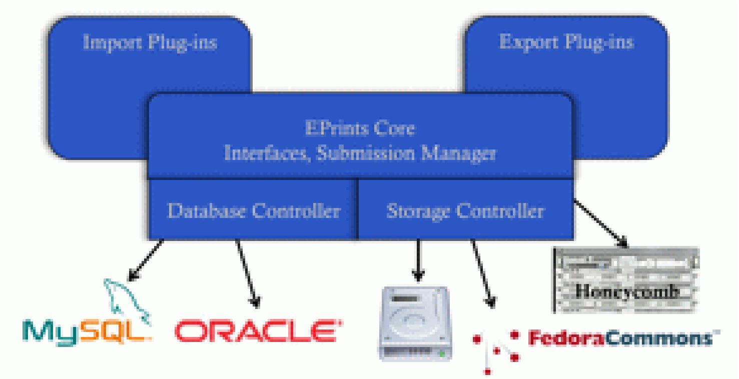


Figure 1. Architecture proposed for EPrints (v3.2) storage controller

The storage controller has been implemented and will be available in EPrints version 3.2 during 2009. Since development of applications that work with EPrints is based on a modular plugin architecture, a series of storage plugins has been written for Amazon S3 and Cloudfront ‘cloud’ services, for Sun Honeycomb, as well as local disk storage. More plugins can be created for any storage service with an application interface, and will work with the storage controller in v3.2.

Copy

ORE: collect and reuse

It is simple to copy a digital file. It is less simple to copy a file complete with its context, which in the case of repositories includes at minimum the metadata describing the item. The Open Archives Initiative (OAI) produced a protocol for sharing metadata describing the contents of repositories. Now OAI has developed a complementary tool for Object Reuse and Exchange (OAI-ORE). This can be used to bind low-level objects into collections. A collection might be an object and its metadata, a collection of objects, even a whole repository. OAI-ORE has many potential applications, but since copy and collection are an intrinsic part of repository data management and preservation practice, it is obvious that OAI-ORE, which has these motivations at its heart, provides a new tool to help in this area. The implications for storage and preservation services are that: 1, collections can be copied to the services; 2, it potentially releases content from being captive to a particular computing platform such as a repository software.

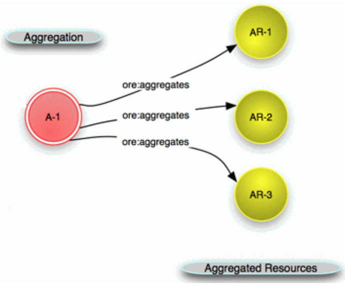


Figure 2. ORE aggregates digital resources

Preserv 2 made an impact using OAI-ORE with an award-winning demonstrator at the JISC CRIG Repository Challenge at Open Repositories 2008. The demo showed how data was moved between two repository softwares with quite distinct data models, from an EPrints repository to a Fedora repository and then back again, using the OAI-ORE (see video http://blip.tv/file/866653, April 2008). "Repositories could become interfaces to remote storage where content is represented and accessed via an OAI-ORE resource map ... we can envisage the prospect of many repository softwares (EPrints, Fedora, etc.) running over one set of resources." (Tarrant, *et al*. 2009). Another detailed analysis of this approach can be found in Rumsey and O’Steen (2008).

Manage

Policy: mandates for preservation?

Preservation begins with policy. At least, it should for an institutional repository. Repository managers have to be given a formal basis on which to make decisions affecting the repository today and in the future. For a start, what type of content does the repository accept, and what is its commitment to look after that content? What resources it has to do this will determine what can be achieved.

A previous Preserv survey found little or no evidence of institutional or repository preservation policy (Hitchcock, *et al*. 2007b). For institutional repositories one area of policy-making that has achieved some prominence is open access mandates. So for a follow-up survey we examined a registry of such policies (Registry of Open Access Repository Material Archiving Policies, ROARMAP) for evidence of policies leading towards preservation based on the thesis that preservation policy is more likely to follow from high-level policy initiatives, and these initiatives are more likely to be found for repositories with OA mandates. This was not strongly borne out in practice. Just 28% of all policies posted on ROARMAP make any reference to preservation. ROARMAP includes mandate policies by research funders as well as institutions. We found that research funder policies are 2.4 times more likely to seek provision for preservation in return for the requirement to deposit papers than are mandate policies from institutions (Hitchcock 2009).

Repositories Support Project: preservation tutorials and workshops

JISC RSP provides guidance, advice and information to repository managers in the UK, and has always prioritised preservation among the range of topics it covers. Steve Hitchcock from Preserv 2 worked with RSP during its first phase to March 2009 and covered preservation in briefing papers, presentations and practical workshops at a number of RSP events:

- Workshop: Preservation and storage management for Institutional Repositories (June 2008) http://www.rsp.ac.uk/events/index.php?page=ThorntonManor-2008-06-18/preservation.php
- Briefing paper: Preservation and Storage Formats for Repositories (April 2008) http://www.rsp.ac.uk/pubs/briefingpapers-docs/technical-preservformats.pdf
- Workshop: Applying Preservation Metadata to Repositories (Jan, Feb 2008) http://www.rsp.ac.uk/events/ProBriefMaterials/BL%20-%20Workshop%202%20-%20preservation%20workshop%20outline.pdf
- Presentation: Policies for Institutional Repositories, including Preservation planning: connecting with policy (June 2007) http://www.rsp.ac.uk/events/SummerSchool2007/policies_2007-06-28.ppt

Presenting at these events, especially the workshops, provided immediate feedback and insights from repository managers, important target stakeholders for Preserv 2. The main lesson: digital preservation can never be simple enough for non-specialists.

For example, preservation metadata, based on the PREMIS standard, was presented twice as a short (less than 2 hours) workshop. At the first event we took a small subset of 20 entries from the PREMIS data dictionary, selected for relevance to repositories, and invited participants in groups to identify whether their repositories collected information to support these entries. After review, for the second event we reduced the number of PREMIS items to 10 to allow more time for discussion on the respective items.

The most engaged workshop was on storage management, in which groups were asked to consider the origination and management of different types of digital data: photographs, music, digitised library content, and an institutional Web site. One group looked at the institutional repository as well. It seems likely that a factor here was the chance to compare a wider range of experiences not just limited to repositories. It is often the case that we are better at solving the problems of others rather than our own, yet we can helpfully reflect those experiences back on our own work.

Outputs and Results

Technical

File formats, 'active' preservation

The largest area of work in Preserv 2 was on file format management and an 'active' preservation approach. Founded on what the National Archives proposed as 'seamless flow' this essentially involves identifying file formats, assessing the risks posed by those formats and taking action obviate the risks where that could be justified. In this work the project went further than implementation, but produced and tested the results.

This led to a new approach combining file format management with automated storage management, which we called 'smart storage'. This anticipates explosive growth in digital data volumes, even for repositories. Smart storage combines an underlying passive storage approach with the intelligence provided through services.

To understand the importance of digital file formats and the threat of format obsolescence to storage and preservation, see this briefing paper produced by Preserv 2 for the RSP (http://www.rsp.ac.uk/pubs/briefingpapers-docs/technical-preservformats.pdf).

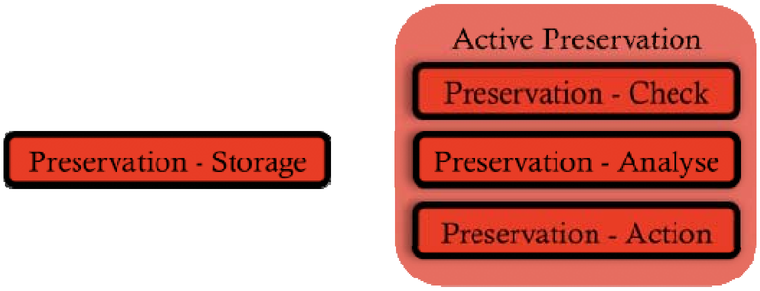


Figure 3. Three stages of active preservation of file formats

Active preservation involves:

1. checking the digital bits, to ensure file consistency and authenticity
2. analysing the files
   - identifying, characterising and validating the format of a digital object

   - invoking a trusted analysis of the risks posed by the identified format to make a recommendation of appropriate action, if any

3. acting on the recommendation

These processes can be implemented with reference to technical registry services, such as PRONOM from TNA, which was used in Preserv 2. Another tool, DROID (digital record object identification service), also produced by TNA and which can be downloaded and run locally, uses a signature file available via the PRONOM registry to classify files and provide specific details relating to individual files. Each file is classified using a PRONOM unique identifier, enabling extra information to be obtained from the registry about the format.

Originally in Preserv 1 we had envisaged applying DROID on a per-repository basis. Instead, applying PRONOM-DROID in conjunction with a Web-based registry of repositories (ROAR) had produced preliminary format profiles for over 200 repositories (Brody, *et al*. 2007). Now we wanted to understand the implications of these profiles for preservation of this content, and demonstrate that presenting this information to repositories and service providers could result in effective preservation decisions with respect to format management.

There were two main limitations to the Web-based approach: high bandwidth requirements for file downloads, especially for large files (above 2 MB), which resulted in many incomplete profiles. With Honeycomb as a prospective storage aggregator we would download content from repositories and apply DROID locally.

Also, we wanted to verify our earlier results, and critically test PRONOM-DROID, since we were unaware of any test data for these support tools. The significance of this is these tools from TNA are in the public domain and widely used, and these results inform the reliability and optimum use of the tools going forward. It is particularly vital that tools intended to improve the reliability of digital data management are themselves shown to be reliable.

Preserv 2 got as far as stage 2 in active preservation, that is, we can classify the objects in a repository, and we have shown we can invoke a current registry to classify the digital objects and present a hierarchy of risk scores for a repository. Critically, the registry doesn't yet have any data for the risks. It has recently been developed and updated with the facility to hold these risk analyses (PRONOM v7), and the detailed criteria for format risk analysis have been elaborated, but the format analyses have not been done yet. It was intended such analyses would be done for selected formats found commonly in institutional repositories.

Testing format classification

Sample data sets, based on the earlier Preserv profiles found in ROAR, were downloaded from 12 selected repositories (100 files from each), together with a 1000-file collection representing the profile of a repository ‘typical’ of all those profiled in ROAR (Figure 4), to our local Honeycomb store.

Classification was performed using the Preserv2 EPrints preservation toolkit. This ‘wraps’ DROID in an EPrints repository environment, also providing a caching database (as an EPrints dataset) and a results page (presented as an EPrints Administration Screen). For more description on how this toolkit works see this Preserv wiki page (http://wiki.preserv.org.uk/index.php/EPrintsPreservation). This toolkit will be another feature of EPrints v3.2 software (due 2009).

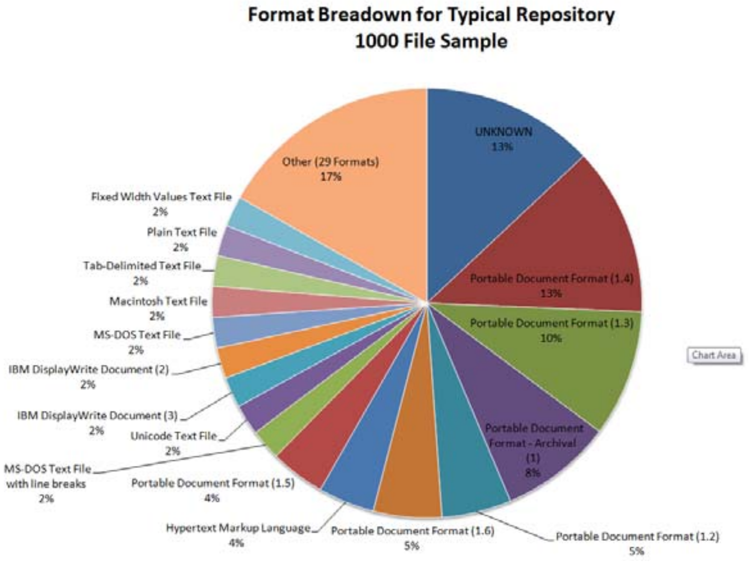
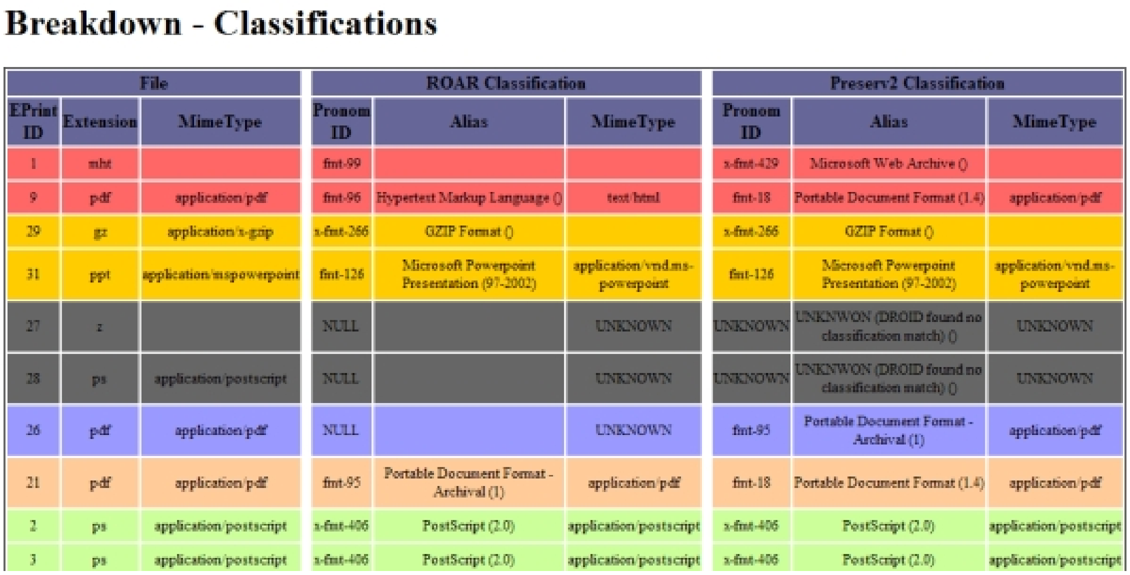


Figure 4. Format breakdown of a 1000-file sample for a 'typical' repository

We were able to compare classifications with the earlier profiles, and as a benchmark against a profile based on inspection of filename extensions (.pdf, .doc, etc.) (Figure 5). File extensions should not be relied upon to get an accurate indication of file format. Formal tools additionally analyse the contents of the file to determine its type and version, and can also verify that the contents conform to the format specification, if there is one.



Breakdown - Classifications

File			ROAR Classification			Preserv2 Classification		
EPrint ID	Extension	MimeType	Pronom ID	Alias	MimeType	Pronom ID	Alias	MimeType
1	mht		fmt-99			x-fmt-429	Microsoft Web Archive ()	
9	pdf	application/pdf	fmt-96	Hypertext Markup Language ()	text/html	fmt-18	Portable Document Format (1.4)	application/pdf
29	gz	application/x-gzip	x-fmt-266	GZIP Format ()		x-fmt-266	GZIP Format ()	
31	ppt	application/mspowerpoint	fmt-126	Microsoft Powerpoint Presentation (97-2002)	application/vnd.ms-powerpoint	fmt-126	Microsoft Powerpoint Presentation (97-2002)	application/vnd.ms-powerpoint
27	z		NULL		UNKNOWN	UNKNOWN	UNKNWON (DROID found no classification match) ()	UNKNOWN
28	ps	application/postscript	NULL		UNKNOWN	UNKNOWN	UNKNWON (DROID found no classification match) ()	UNKNOWN
26	pdf	application/pdf	NULL		UNKNOWN	fmt-95	Portable Document Format - Archival (1)	application/pdf
21	pdf	application/pdf	fmt-95	Portable Document Format - Archival (1)	application/pdf	fmt-18	Portable Document Format (1.4)	application/pdf
2	ps	application/postscript	x-fmt-406	PostScript (2.0)	application/postscript	x-fmt-406	PostScript (2.0)	application/postscript
3	ps	application/postscript	x-fmt-406	PostScript (2.0)	application/postscript	x-fmt-406	PostScript (2.0)	application/postscript

Figure 5. Comparing three classifications for a 100-file repository based on file extensions, PRONOM-ROAR and the Preserv 2 toolkit. Green rows are complete matches, while other colours indicate different types of mismatch. Below the line of this partial list of results all rows are green.

Summary of the main results of format classification testing:

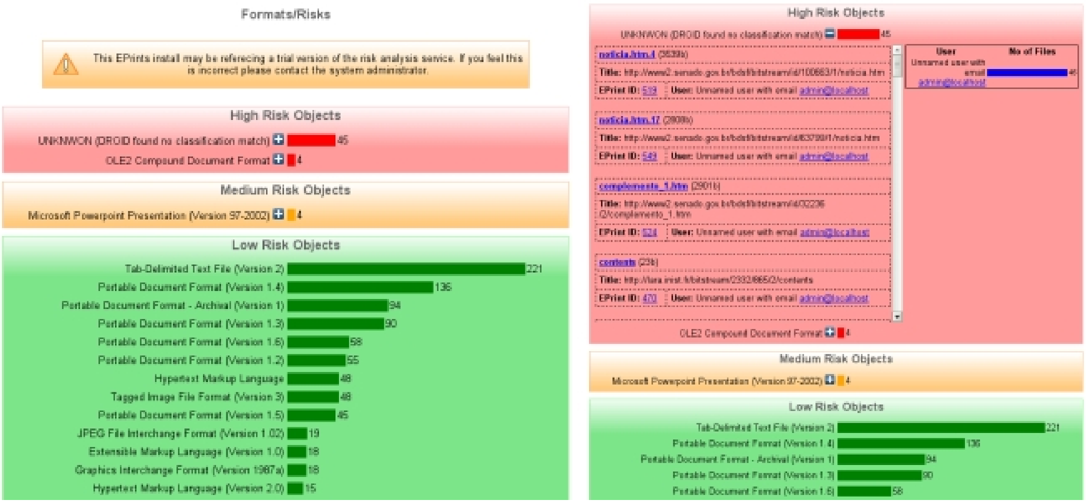
- Using DROID, 146 files could not be classified from a full dataset of 2144 files, a 93.1% classification rate. Including wrongly classified files, this rate comes down to 92.75%.
- Simple examination of file extension will classify an estimated 99.8% of the files correctly (if the extension is correct, not tested for all files but tested on the fringe cases). However, this does not provide any information about the file version.
- PRONOM lacks a complete set of MIME-types for its format data. This matters for files where the classification may have changed between inspections. Such changes could indicate files at risk and prompt investigation. Comparing PRONOM-ROAR classification with that by Preserv 2 suggests that over a quarter of the files in the 1000-item 'typical' repository dataset have changed classification. When MIME-types are applied, however, it was found that 256 files match by MIME-type and only 40 files changed classification.
- The Preserv2 EPrints toolkit version of DROID is able to classify more objects than the PRONOM-ROAR classification, cutting down on the number of unknown files. This is achieved primarily by virtue of deploying a more recent version of the DROID signature file (v13 against v12 used in PRONOM-ROAR), and by linking DROID more closely with the repository software can ensure successful completion without limiting file size.

Format risk analysis: a preliminary interface implementation

The result of a file format identification, or an indication of a format change, can suggest a file is at risk of becoming inaccessible or corrupted. Preserv 2 developed a repository interface to present formats by risk category, and to enable information on high-risk formats and files at risk to be displayed.

The facility for adding risk assessment data is provided in the latest version of PRONOM (v7), which became available to the project during March 2009, the very end of the project. However, the functionality of this version was emulated in advance with the PronomStubCode developed by the project at Southampton University. The risk interfaces (Figure 6) were first developed using this code, after which it was a simple procedure to re-direct the repository classification data (results from the Preserv2 EPrints toolkit, including DROID) to the live version of PRONOM and test this. Two minor changes were needed to make the risk analysis module in EPrints work with the new service. Preserv2 has shown that providing risk scores through the live PRONOM service is feasible. Using thresholds, which can be adjusted by the repository administrator, the file formats are grouped into traffic light-style categories that clearly show the user the risks related to all of their objects, including those where no risk has been found.

The risk scores presented are entirely hypothetical, however. It was expected that risk assessments of selected file formats found commonly in repositories (see Figure 4) would be included in the recent release (v7) of PRONOM, but format risk assessments are very detailed (with dozens of risk factors weighted across several categories), requiring skill and expert scrutiny, and this was not done within the project timescale. It is hoped that format risk scores will be provided through PRONOM or other publicly available services in the near future.



Formats/Risks

⚠ This EPrints install may be referencing a trial version of the risk analysis service. If you feel this is incorrect please contact the system administrator.

High Risk Objects

UNKNWON (DROID found no classification match) ⊞ 45
OLE2 Compound Document Format ⊞ 4

Medium Risk Objects

Microsoft Powerpoint Presentation (Version 97-2002) ⊞ 4

Low Risk Objects

Tab-Delimited Text File (Version 2) 221
Portable Document Format (Version 1.4) 136
Portable Document Format - Archival (Version 1) 94
Portable Document Format (Version 1.3) 90
Portable Document Format (Version 1.6) 58
Portable Document Format (Version 1.2) 55
Hypertext Markup Language 48
Tagged Image File Format (Version 3) 48
Portable Document Format (Version 1.5) 45
JPEG File Interchange Format (Version 1.02) 19
Extensible Markup Language (Version 1.0) 18
Graphics Interchange Format (Version 1987a) 18
Hypertext Markup Language (Version 2.0) 15

High Risk Objects

UNKNWON (DROID found no classification match) ⊟ 45

noticia.htm.4 (3539b)
Title: http://www2.senado.gov.br/bdsf/bitstream/id/100663/1/noticia.htm
EPrint ID: 519 **User:** Unnamed user with email admin@localhost

noticia.htm.17 (2808b)
Title: http://www2.senado.gov.br/bdsf/bitstream/id/63799/1/noticia.htm
EPrint ID: 543 **User:** Unnamed user with email admin@localhost

complemento_1.htm (29011b)
Title: http://www2.senado.gov.br/bdsf/bitstream/id/32236/2/complemento_1.htm
EPrint ID: 524 **User:** Unnamed user with email admin@localhost

contents (23b)
Title: http://lara.inist.fr/bitstream/2332/865/2/contents
EPrint ID: 470 **User:** Unnamed user with email admin@localhost

User	No of Files
Unnamed user with email admin@localhost	45

OLE2 Compound Document Format ⊞ 4

Medium Risk Objects

Microsoft Powerpoint Presentation (Version 97-2002) ⊞ 4

Low Risk Objects

Tab-Delimited Text File (Version 2) 221
Portable Document Format (Version 1.4) 136
Portable Document Format - Archival (Version 1) 94
Portable Document Format (Version 1.3) 90
Portable Document Format (Version 1.6) 58

Figure 6. Format risk interfaces provided in EPrints: left, hierarchy of (hypothetical) format risk scores; right, how high risk formats and files might be presented

Out of this adversity may come opportunity, however, for a new approach to format risk assessment. We are moving towards a semantic Web of data, and work has begun on applying this approach to file formats. Risk assessment involves connecting new and known data on specific file formats. By combining available data from different sources the semantic framework is already proving its worth by revealing more connections than are evident in the native tools. Initial results have been provided in a personal blog (http://davetaz-blog.blogspot.com/2009/05/file-format-risk-analysis-how-hard-can.html) and will be developed further as appropriate.

Clearly the agency assessing the risk will be an important factor in choosing the service. If using PRONOM (from v7) the risk score is calculated by The National Archives (UK). Other services may in future provide format risk scores.

'Smart' storage

To supplement one part of the 'active' preservation services, format identification, a 'smart' storage approach has been developed (Figure 7). Additional tools - a calendar-based scheduler, a harvester to retrieve the latest stored content from a repository, messaging services to record the history of events and the results, all designed to work with, or 'wrap', the format identification tool DROID - combine to manage and process the data and results.

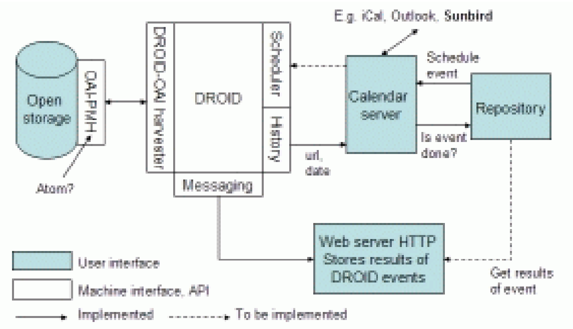


Figure 7. Smart storage schematic

The illustrated approach uses the iCal calendar standard in an implementation called Darwin Calendar server. This controls the timing of planned preservation events, such as file format management, and records the results. The scheduler can also be used to manage preservation actions such as scanning for viruses or malware, and can handle recurring events. Most actions that result from a preservation policy choice can be pre-assigned to an item when it undergoes a change of state (creation, modification, deletion, extension). A full explanation of this smart storage approach is provided by Hitchcock, *et al*. (2008).

Outcomes

If repositories are changing, projects such as Preserv 2 have been among the first to detect the change. We cannot seek to preserve digital content if we do not anticipate and understand the changing framework in which the target digital content is produced, stored and accessed.

In building a storage controller into EPrints v3.2 the project was pre-empting similar work for Fedora, and later the forming of the DuraSpace organisation by Fedora with DSpace to act as an intermediary service for repository storage in the cloud.

Preserv did not fully achieve its goals of a set of tools and services sufficient to provide repositories with a working preservation strategy and motivate emerging preservation service providers prepared to target repositories. Repositories are still confused about their immediate objectives and role, making it difficult to engage with half-fulfilled and still complex approaches to preservation that seem little integrated with their current activities.

Meanwhile the preservation community, which is accumulating an expert and practical knowledge of the detailed and technical requirements of maintaining digital content, preaches fright of the future and urgency while itself seeming pedestrian in developing its core tools and lacking rigour in testing the reliability of key components in the emerging preservation infrastructure (e.g. on JHOVE see http://blog.dshr.org/2009/01/postels-law.html). The digital world will not slow or stand still in the cause of preservation.

Never was it so evident that digital content creation, and its storage, management and preservation are so deeply entwined.

Conclusions

Seeking to identify, build and test tools and services to enable digital preservation to be performed cost-effectively by repositories, Preserv 2 produced some important developments and breakthroughs, but leaves a gap in a key area.

Large-scale data storage options are increasing, and Preserv 2 developed the first repository storage controller, which will be a feature of EPrints v3.2 software. Plugin applications that use the controller have been written for cloud services as well as for local storage.

In a breakthrough application Preserv 2 used OAI-ORE to show how data can be moved between two repository softwares with quite distinct data models, from an EPrints repository to a Fedora repository.

There remain few examples of repository policy driving preservation, and this will continue to hinder uptake of preservation practices. We examined a registry of open access mandate policies based on the thesis that preservation policy is more likely to follow from high-level policy initiatives. This was not strongly borne out in practice. Just over a quarter of all mandate policies posted by early March made any reference to preservation.

File format management is another area of risk for repositories, and although Preserv 2 has produced and tested critical tools, it has not been able to quantify the risk posed by current repository format profiles. To extend format identification services a ‘smart storage’ approach has been demonstrated. Smart storage combines an underlying passive storage approach with the intelligence provided through services.

File format classification was performed using the Preserv2 EPrints preservation toolkit, which will be another feature of EPrints v3.2 software. In tests this toolkit was shown to classify more objects than the earlier Web-based classification (PRONOM-ROAR), reducing the number of unknown files. The tests also identified shortcoming in the current tools. PRONOM was found to lack a complete set of MIME-types for its format data. It is vital that tools intended to improve the reliability of digital data management are themselves shown to be reliable.

It was shown we can invoke a current registry to classify the digital objects and present a hierarchy of risk scores for a repository, but the registry does not yet have any data for the risks. The registry has recently been developed and updated with the facility to hold these risk analyses, and the detailed criteria for format risk analysis have been elaborated, but the format analyses have not been done yet. It is intended such analyses will be done in ongoing work among the project partners for selected formats found in institutional repositories.

Implications

Preservation of digital repositories is an unfinished business. If that is an obvious statement given the nature of preservation – the task never ends – then perhaps a more revealing statement is that it has hardly begun.

Repositories are increasingly positioning for preservation and highlighting this in their promotion, but are not fully prepared or engaged. There is little in the way of policy, even less preservation policy, planning or a costings framework to underpin longer-term decisions and commitments.

Nor are they fully supported by tools and services in the more specialised and technical areas. The work begun in projects such as Preserv 2 needs to go further to complete this process, and sustainable services need to emerge.

A meeting held by the Digital Preservation Coalition on Tackling the Preservation Challenge (December 2008, http://www.dpconline.org/graphics/events/081212RepMngrsWkshp.html) left repository managers asking what can be done to help them preserve their repositories. Who is responsible for repository preservation: repositories and their institutions, or preservation service providers? In many cases the answer is: both. There is no services-only solution.

The market for preservation services for repositories is undeveloped, and is unlikely to be developed until repositories establish clearer requirements, policy frameworks and budgets. It has not so far attracted established preservation organisations, as anticipated in the final report from Preserv 1, and the demise in 2008 of the Arts and Humanities Data Service, an established UK preservation service provider, was ostensibly predicated on institutions taking over the preservation activities of a central service.

Yet some institutions are paying for services such as repository development, customisation and hosting, and currently those repository services providers may be the most likely to extend an offer of support for preservation because of their practical experience and detailed understanding of repository data management.

The technical preservation toolset, even for today's framework, is incomplete and probably too complex. Preserv 2 has contributed to the development of 'active' preservation, an approach to managing technical risks posed by proliferation and possible obsolescence of file formats. A number of preservation-focussed organisations and some large projects are leading this development. Preserv 2 intended to go further than others have to date, to test the practice by quantifying the risk to formats most likely to be found in repositories.

Active preservation seems well suited in principle to the task of format management in today's digital content environment. We cannot even begin to assess its value for tomorrow unless we can test it in practice, so we have to have the data needed to do this. Preservation organisations preach urgency to others, but need to practice this themselves.

In another context it was said: "As novelty spread, old institutions seemed exhausted while new ones seemed untrustworthy" (Shirky, http://www.shirky.com/weblog/2009/03/newspapers-and-thinking-the-unthinkable/). This refers to the last great period of change in our media, to a world immediately before and after the printing press, and it concerns journalism. If we apply this today to digital preservation, where trust is paramount, the danger is all too apparent. Preserv was a joining of the great and the new in this field. Clearly, we need both.

Preserv 2 has provided insights into the imminent transformations in the infrastructure of repositories and storage. We can see this happening more generally, in the use of terms such as Web 2.0, Semantic Web, 'cloud', etc. These represent processes of change. For digital libraries and institutional repositories OAI-ORE, a complement to OAI-PMH which had been fundamental to repository interoperability, promises a radical opportunity in collection building and itself fuelling the transformation of repositories, as Preserv has demonstrated.

The concept of preservation suggests stability, yet the need for preservation is greater when change is all around. But how can these be reconciled; how can we preserve without inhibiting change and progress? Repositories are growing in number, storing more content and diversifying in terms of repository and content types. This simply represents a natural development path for repositories as a still relatively new approach to digital, networked content management and storage. As the Web morphs into Web 2.0 and subsequent identities, so repositories are bound to be transformed to take advantage of new Web services and infrastructure. What is it we are trying to preserve: the repositories, the content, or both? To answer that needs a fundamental reappraisal by each repository, and perhaps each owner or institution, of what it is for and how it can achieve those objectives. From that can be derived policy, plans and the risk profile to facilitate preservation.

Digital preservation is not just about retaining and displaying something as it was, but planning and maintaining an active working platform and access to content through change and into new information environments.

Recommendations

- Promote a joint approach to repository preservation by repositories and preservation organisations and specialists. There is no services-only solution.
- Show repositories how engaging in broad strategic planning and policy development can lead directly to preservation planning and policy.
- Assist repositories to develop specifications and perform requirements analyses.
- Adapt digital preservation training and support to enable the joint approach, recognising that it can never be simple enough.
- Reduce complexity of preservation tools to integrate with repository processes.
- Motivate preservation services by organisations that understand repository data management.

References

Brody, Tim, *et al*. (2007) PRONOM-ROAR: Adding Format Profiles to a Repository Registry to Inform Preservation Services, *International Journal of Digital Curation*, Vol. 2, No. 2, November 2007 http://www.ijdc.net/ijdc/article/view/53/

Hitchcock, Steve (2009) The effect of open access mandates on repository preservation policy, Preserv 2 project, 10 March 2009 http://preserv.eprints.org/papers/mandates/mandates_report.html

Hitchcock, Steve, *et al*. (2008) Towards smart storage for repository preservation services, *iPRES 2008: The Fifth International Conference on Preservation of Digital Objects*, London, 29-30 September 2008 http://eprints.ecs.soton.ac.uk/16785/

Hitchcock, Steve, *et al*. (2007a) Laying the Foundations for Repository Preservation Services, Final Report from the PRESERV project, March 7, 2007 http://www.jisc.ac.uk/media/documents/programmes/preservation/preserv-final-report1.0.pdf

Hitchcock, Steve, *et al*. (2007b) Survey of Repository Preservation Policy and Activity, Preserv project, 21 February 2007 http://preserv.eprints.org/papers/survey/survey-results.html

Rumsey, Sally and O'Steen**,** Ben (2008) OAI-ORE, PRESERV2 and Digital Preservation, *Ariadne*, issue 57, 30-October http://www.ariadne.ac.uk/issue57/rumsey-osteen/

Tarrant, David, *et al*. (2009) Using OAI-ORE to Transform Digital Repositories into Interoperable Storage and Services Applications, *Code4Lib Journal*, Issue 6, 30 March 2009 http://journal.code4lib.org/articles/1062