Preservation for Institutional Repositories: practical and invisible

Steve Hitchcock, Tim Brody, Jessie M.N. Hey and Leslie Carr

Preserv Project, IAM Group, School of Electronics and Computer Science,
University of Southampton, SO17 1BJ, UK
sh94r@ecs.soton.ac.uk

With good prospects for growth in IR contents, in the UK due to the proposed RCUK policy on mandating deposit of papers on funded work, and internationally due to the Berlin 3 recommendation, it is timely to investigate preservation solutions for IRs. The paper takes a broad view of preservation issues for IRs - based on practice, experience and visions for the future - from the perspective of Preserv, a JISC-funded project. It considers preservation in the context of IRs. Based on the OAIS preservation model, an architecture is proposed to support distributed preservation services for IRs. Work performed so far involves adapting the IR user deposit interface in a pilot version of EPrints software for building IRs, and determining accurate file format information using PRONOM software. The paper looks ahead briefly at the role of preservation service providers, working for the IR, within this architecture. The strategy is to take practical steps that are, as far as possible, invisible to all but those concerned with the preservation process for IRs.

Introduction

Digital preservation is a paradox (Wiggins 2001), perhaps more so for institutional repositories (IRs). The overarching paradox of digital preservation is cost: "people think of digital content as inexpensive--and inexpensive things are not worth preserving" Wiggins notes. Preservation might not be inexpensive. For IRs the paradox extends beyond cost to content.

Hardly anyone would claim that digital preservation is not desirable. Some would call it 'vital', even 'critical', although cries of a sense of imminent crisis are no longer appropriate. Instead, a view is emerging that "digital preservation is not an isolated process, but one component of a broad aggregation of interconnected services, policies, and stakeholders which together constitute a digital information environment." (Lavoie and Dempsey 2005)

Managers of IRs naturally have a responsibility for the longevity of the materials they are charged with managing for their institutions and researchers. With cost constraints, priorities have to be managed carefully and sensitively, however. Preservation concerns should not be allowed to become a barrier to the deposit of new

content in IRs. Then there is the issue of the purpose and need for digital preservation in IRs. It can be noted that the materials in IRs that are the most likely targets of preservation activity are in fact author drafts or copies of traditional peer-reviewed journal publications (Self-Archiving FAQ), and these are already subject to preservation actions by publishers and national libraries (Steenbakkers 2004), in other cases if only by virtue of legal deposit of a print edition.

The community of authors and producers of this open access (OA) content should be confident of depositing digital materials in IRs: "the proliferation of experience, research, and infrastructure throughout the cultural heritage community has made trustworthy digital repositories conceptually realistic." (RLG-NARA 2005) IRs that comply with the Open Archives Initiative-Protocol for Metadata Harvesting (OAI-PMH) have a particular advantage. IRs do not have to become "trustworthy digital repositories", but will be able to interface to such services through the OAI-PMH. Separation of content provision and distribution of services is fundamental to OAI.

This paper takes a broad view of preservation issues for IRs - based on practice, experience and visions for the future - from the perspective of Preserv, a project funded within the JISC programme for Supporting Digital Preservation and Asset Management in Institutions. It considers preservation in the context of IRs, and looks at support for preservation already built into IR software. Based on the OAIS preservation model, an architecture to enhance support for preservation in IRs is proposed. Work performed so far involves adapting the IR user deposit interface and the identification of file formats at the point of deposit. The paper looks ahead briefly at the role of preservation service providers, working for the IR, within this architecture.

For IRs the work is just beginning in this project and is ongoing in other projects concerned with preservation in IRs (Sherpa-DP, DSpace). With good prospects for growth in IR contents, in the UK due to the proposed RCUK (2005) policy on mandating deposit of papers on funded work, and internationally by the Berlin 3 (2005) recommendation, now is the time to investigate preservation *solutions* for IRs.

The likelihood is that IRs will be served by a range of solutions for their preservation needs, and growth can continue unhindered. These solutions should be there when needed, practical and invisible, and not an impediment to the growing body of content they are intended to serve.

What is an IR?

In the broadest terms IRs are intended to provide managed access to the digital outputs, or resources, produced and self-archived by the members of an institution. Typically IRs are adopted by universities, in particular by research-intensive universities. The extent of this managed environment and what it does depends on the purpose and the local priorities for the IR. Lynch (2003) offered one view:

> *It is most essentially an organizational commitment to the stewardship of these digital materials, including long-term preservation where appropriate, as well as organization and access or distribution.*

The first and most obvious benefit of an institutional commitment is that a managed environment provides a greater degree of assurance of continued access than personal Web sites that have been so popular with many researchers (Swan and Brown 2005). Since preservation is as much a management requirement as a technical requirement, the institutional backing of IRs provides a platform in principle to adopt good practice.

What that practice might be is the subject of this investigation. From experience of two large and growing repositories at Southampton, institutional and departmental, we can project that whatever priorities are adopted, IRs will have two common characteristics:

- **Heterogeneous data formats**. Although the primary target will be refereed research papers, content need not be limited to this type of content: for example, data and correspondence supplementing published results are likely to start appearing with greater use of IRs by authors (e.g. eBank UK project)
- **Low cost per item deposited**. IRs must keep costs low enough not to jeopardize open access.

The implications of these for preservation are the need for automation, in the collection of preservation metadata, and in the efficient labelling, selection and delivery of content to preservation services, and in any preservation actions taken subsequently.

These constraints are likely to prove severe against audited checklists required for the certification of specialised preservation service providers, or "trusted digital repositories" (RLG-NARA 2005). As a result, it cannot be assumed that this will be *best* practice, as might be advocated by preservation experts, but the most appropriate and cost-effective approach depending on the priorities of the IR.

To consider the role of preservation in IRs we have to try to establish what is meant by preservation.

What is Preservation?

"Digital preservation" has been defined as (RLG-OCLC 2002):

> *the managed activities necessary for ensuring both the long-term maintenance of a bytestream and continued accessibility of its contents.*

How this is achieved is less certain.

There is a tendency among those working in the field to treat digital preservation as a primary concern. In the interpretation below, preservation is a third-level activity (Lord and Macdonald 2003, p12):

- Curation: The activity of, managing and promoting the use of data from its point of creation, to ensure it is fit for contemporary purpose, and available for discovery and re-use. For dynamic datasets this may mean continuous enrichment or updating to keep it fit for purpose. Higher levels of curation will also involve maintaining links with annotation and with other published materials.
- Archiving: A curation activity which ensures that data is properly selected, stored, can be accessed and that its logical and physical integrity is maintained over time, including security and authenticity.
- Preservation: An activity within archiving in which specific items of data are maintained over time so that they can still be accessed and understood through changes in technology.

The latter two definitions seem in application to be based on a very traditional view of the archiving-preservation range of activities. Two awkward terms here are the *selection* of *specific items*, as highlighted in the following scenario involving the FBI (Talbot 2005):

> "In 1972, Weinstein (sworn in as the new Archivist of the United States in February) was a young historian suing for the release of old FBI files. FBI director J. Edgar Hoover--who oversaw a vast machine of domestic espionage--saw a Washington Post story about his efforts, wrote a memo to an aide, attached the Post article and penned into the newspaper's margin: "What do we know about Weinstein?" It was a telling note about the mind-set of the FBI director and of the federal bureaucracy of that era. And it was saved--Weinstein later found the clipping in his own FBI file.
>
> "But it's doubtful such a record would be preserved today, because it would likely be "born digital" and follow a convoluted electronic path. A modern-day J. Edgar Hoover might first use a Web browser to read an online version of the Washington Post. He'd follow a link to the Weinstein story. Then he'd send an e-mail containing the link to a subordinate, with a text note: "What do we know about Weinstein?" The subordinate might do a Google search and other electronic searches of Weinstein's life, then write and revise a memo in Microsoft Word 2003, and even create a multimedia PowerPoint presentation about his findings before sending both as attachments back to his boss."

How would these items have been selected for archiving and preservation, and for the specific items would it be possible to reestablish the external links between them in the absence of any implicit connections?

Clearly this example does not just apply to government and similar examples could be substituted from IRs. The Preserv project is performing a stakeholder survey - including authors and users, IR managers and administrators, heads of institutions, research funders - to discover their views on the preservation needs of IRs. Early anecdotal evidence supports the need to consider scenarios such as painted above.

Q: What Would It Mean To Preserve Your Research?
A1: "One would want to understand the context - the whys and motivation for the work - both personal and in terms of the calls for research proposals."
A2: "Capture a copy of everything I have ever published: in context, i.e. including a copy of the thing it which my article was published."

We have to imagine architectures and systems for curation (to use this definition) that take a broad, digitally-oriented view of selection before we can satisfy the needs of archiving and preservation.

Preserv, EPrints and Distributed Preservation Services

The Preserv Project is investigating the delivery of preservation services based on EPrints by proposing to adapt elements of the software. The project has three partners - The British Library, The National Archives, and Oxford University. The roles of each partner will become apparent as the proposed architecture for IR preservation services is revealed. The project formally began in February 2005, and will be ongoing for two years. The work described here focusses on early implementation, principally involving the IR user deposit interface.

Architecture for IR preservation

The distributed architecture envisaged (Figure 1) is based on a typical IR workflow, between author and reader, with two additional components:

- a modified import (author deposit) interface
- an OAI export interface to a preservation service

The blacked-out components in the architecture represent:

- Pilot IRs, in the project represented by repositories at Oxford University and Southampton University (eprints.soton.ac.uk).
- The preservation service component, provided in the project by the BL, could potentially involve a range of activities associated with digital preservation: storage media, media refreshing, reformatting, backups and disaster recovery, environment, audit, security, preservation strategy, migration, technology preservation, emulation, records management, etc.

These activities are managed entirely by partners and, apart from acting as test interfaces and providing feedback to influence the evolving architecture, are not the subject of the investigation by the project. The arrow bounded by a dashed line represents a feedback-influence loop for preservation partners and IR stakeholders via a control mechanism.

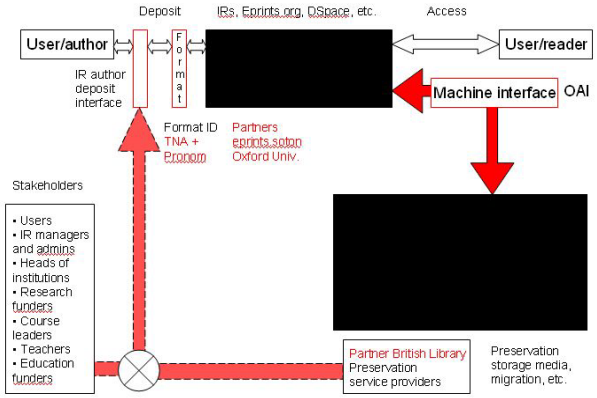

Fig. 1. Preserv architecture for distributed preservation services

Informing EPrints author deposit: determining file formats with PRONOM

A primary determining factor in long-term accessibility - the ability to read and present any digital object - is the format of the object. With the growth and commercialisation of computer usage in the general population, the number of formats has proliferated. Formats are often tied to applications, such as word processing, are not always well documented or standardised, and may have a lifespan as short as the application. While openly published standards are ideal, commercial pressures often mitigate against these.

For preservation purposes, therefore, certain actions may be required that can be inferred from the file format. What is needed is an accurate and reliable description of the format and its versioning. One approach is to automate the determination of file format and version by comparison of content bits with a database, or registry, of known formats.

To support its work in maintaining government records, The (UK) National Archives (TNA) has produced such a tool, called PRONOM (Darlington 2003, Brown 2005). The Preserv Project is working with TNA to integrate and evaluate the use of PRONOM in the deposit procedure for EPrints-based IRs.

PRONOM is a Java tool for identifying the file format of arbitrary files. To determine what format a file is, PRONOM uses the file extension and searches for simple expressions within the file. These file extension maps and simple expressions are stored in a signature database, in essence a 'fingerprint' database for file formats.

Integration of PRONOM into the Preserv Project is two-fold. Initially PRONOM will be used to determine file format versions, hence to inform users and administrators of document type (in particular how old the format is) as part of a preservation service.

While the execution of PRONOM occurs in the background during the deposit process, the PRONOM output is stored as an EPrints metadata field such that it can be exported for use by a preservation service (Figure 3).

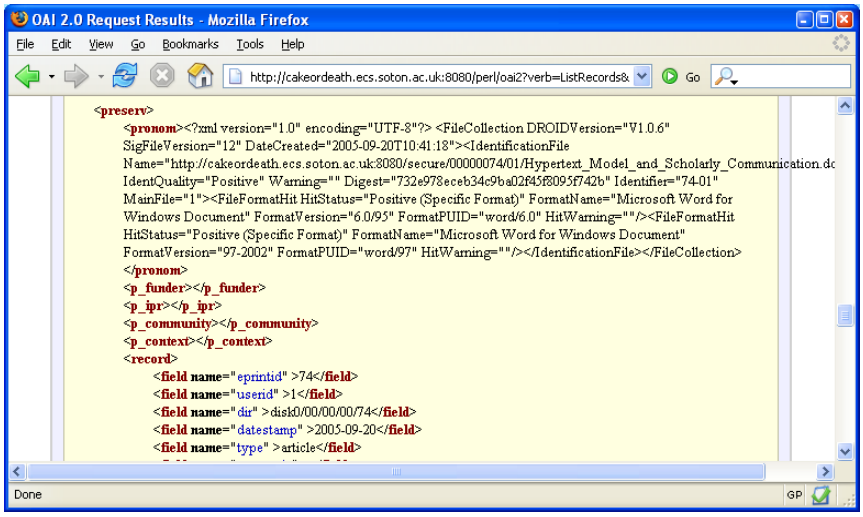

Fig. 2. PRONOM output exported as part of a 'preservation' OAI export from EPrints. This OAI export uses the existing EPrints XML export function that encodes all of the system metadata associated with an eprint as field-value data elements

Next steps

We now need to investigate what other information can be captured at the deposit stage, from authors or directly from the content to be deposited, to inform later preservation actions. What those preservation actions might be needs to be informed by assessing the requirements of all stakeholders in IRs. It cannot be assumed, for reasons described above, that the preservation requirements of IRs will be the same as any other digital production system. Based on the survey of stakeholders, we can begin to consider the types of preservation services and business models that might be viable for IRs.

Due to the specialisation often required of preservation services, and for economy of scale, it is anticipated that service providers will be external to the IR, although this may not be so in all cases. OAI-based harvesting has been identified as the most likely means for disseminating digital objects between IRs and service providers (and back again!), although this is not yet a complete solution (Van de Sompel *et al*. 2004).

Conclusions

> "Access is still not the primary purpose of a preservation system" (Cornell 2003)

Access *is* the purpose of an IR. While preservation has a role in assuring long-term accessibility of the contents of an IR, it is important that the two purposes do not conflict. Availability and access to a resource *now* must not be compromised, either within an IR or by raising a barrier to deposit. The Preserv Project, along with its partners, is investigating how to manage these requirements, by enhancing information capture at the IR deposit stage, principally by determining accurate file format information using PRONOM, and subsequently automating the dissemination of data for preservation to distributed service providers via an adapted OAI harvesting mechanism. Although much remains to be done to assess the needs of all stakeholders, identify service requirements and business models for the preservation of IR contents, and test and evaluate this approach, the underlying strategy must remain constant: practical steps that are, as far as possible, invisible to all but those concerned with the preservation process.

References

Berlin 3 (2005) Outcomes: Agreed Recommendation. Berlin 3 Open Access: Progress in Implementing the Berlin Declaration on Open Access to Knowledge in the Sciences and Humanities, Southampton, February 28 - March 1
http://www.eprints.org/events/berlin3/outcomes.html

Brown Adrian (2005) Automating Preservation: New Developments in the PRONOM Service. RLG DigiNews, Vol. 9, No. 2, April 15
http://www.rlg.org/en/page.php?Page_ID=20571#article1

Cornell (2003) The OAIS Reference Model. In Digital Preservation Management: Implementing Short-Term Strategies for Long-Term Problems, section 4B, Cornell University, September
http://www.library.cornell.edu/iris/tutorial/dpm/

Darlington, Jeffrey (2003) PRONOM - A Practical Online Compendium of File Formats. RLG DigiNews, Vol. 7, No. 5, October 15
http://www.rlg.org/legacy/preserv/diginews/diginews7-5.html#feature2

DSpace Federation
http://dspace.org/

eBank UK project
http://www.ukoln.ac.uk/projects/ebank-uk/

EPrints Free Software

http://www.eprints.org/software/

Lavoie, Brian and Lorcan Dempsey (2004) Thirteen Ways of Looking at ...Digital Preservation. D-Lib Magazine, Vol. 10, No. 7/8, July/August
http://www.dlib.org/dlib/july04/lavoie/07lavoie.html

Lord, Philip and Alison Macdonald (2003) e-Science Curation Report Data curation for e-Science in the UK: an audit to establish requirements for future curation and provisionJISC
http://www.jisc.ac.uk/uploaded_documents/e-ScienceReportFinal.pdf

Lynch, Clifford A. (2003) Institutional repositories: essential infrastructure for scholarship in the digital age. ARL Bimonthly Report, No. 226, February
http://www.arl.org/newsltr/226/ir.htmlhttp://www.archives.gov/press/press-releases/2005/nr05-112.html

Preserv Project (2005)
http://preserv.eprints.org/

RCUK (2005) RCUK Announces Proposed Position on Access to Research Outputs. RCUK press release, 28 June 2005
http://www.rcuk.ac.uk/press/20050628openaccess.asp

RLG-NARA (2005) An Audit Checklist for the Certification of Trusted Digital Repositories (draft for comment). RLG-NARA, August
http://www.rlg.org/en/page.php?Page_ID=20769

Self-Archiving FAQ: Preservation
http://www.eprints.org/openaccess/self-faq/#1.Preservation

RLG-OCLC (2002) Trusted Digital Repositories: Attributes and Responsibilities. RLG-OCLC Report, May
http://www.rlg.org/longterm/repositories.pdf

Sherpa-DP Project
http://ahds.ac.uk/about/projects/sherpa-dp/

Steenbakkers, Johan F. (2004) Treasuring the Digital Records of Science: Archiving E-Journals at the Koninklijke Bibliotheek. RLG DigiNews, Vol. 8, No. 2, April 15
http://www.rlg.org/en/page.php?Page_ID=17068&Printable=1&Article_ID=990

Swan, Alma and Sheridan Brown (2005) Open access self-archiving: An author study. Report to JISC Scholarly Communications Group, May
http://www.jisc.ac.uk/uploaded_documents/Open%20Access%20Self%20Archiving-an%20author%20study.pdf

Supporting Digital Preservation and Asset Management in Institutions, a JISC programme
http://www.jisc.ac.uk/index.cfm?name=programme_404

Talbot, David (2005) The Fading Memory of the State. Technology Review.com, July
http://www.technologyreview.com/articles/05/07/issue/feature_memory.asp

Van de Sompel, Herbert, Michael L. Nelson, Carl Lagoze and Simeon Warner (2004) Resource Harvesting within the OAI-PMH Framework, D-Lib Magazine, Vol. 10, No. 12, December
http://www.dlib.org/dlib/december04/vandesompel/12vandesompel.html

Wiggins, Richard (2001) Digital Preservation: Paradox & Promise, Library Journal NetConnect, April 15, 2001
http://libraryjournal.reviewsnews.com/index.asp?layout=article&articleid=CA106209&publication=libraryjournal