Report on Preservation Standards

Summary

Digital preservation involves keeping information accessible through changes in hardware and software. This document provides a summary of existing standard practices, guidelines, and procedures that will allow the repository to consider long-term implications when storing e-prints.

Framework for a repository structure

The objective of digital preservation is to extend the useful life of information resources in terms of its longevity, integrity and access. When handling a large number of e-prints, stored in many different formats, a detailed preservation framework is important to ensure data is preserved in a timely manner and according to existing preservation standards, guidelines and practices.

The functions of a digital repository are not unique to e-prints. The OAIS (Open Archival Information Systems) has proven useful as a reference model for repository implementers or simply as a checklist for those already in place. It does not specify the correct method of implementing a digital repository, but instead establishes a high-level framework for understanding the structural organisation of a repository and the most effective method of maintaining content. By applying the OAIS model, e-print repositories (as well as libraries and archives) can benefit from the use of common terminology and a common conceptual framework, which simplifies the process of sharing ideas and experiences.
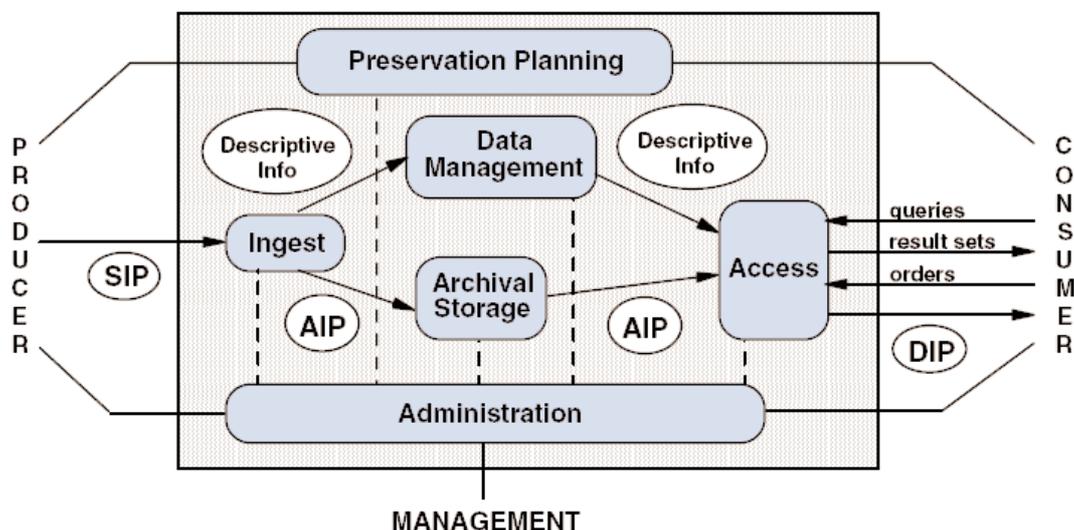

Figure 1: OAIS Functional Entities (CCSDS, 2002, p38)

The OAIS *Functional Model* (figure 1) indicates the entities necessary to ensure the "Information Package" (the e-print) is handled in a consistent and efficient manner. These entities, when overlaid onto the repository management structure, may be used to identify the relationships and responsibilities necessary to accept an e-print at submission (Ingest), store and maintain it as necessary, and make it available to the target audience. To ensure the repository is able to distribute and preserve content in the long-term, the OAIS reference model identifies five functions that should be performed by a repository (CCSDS, 2002):

- Negotiate for and accept e-prints from authors and rights holders.
- Obtain sufficient control of the information provided to support long-term preservation.
- Determine the target audience for e-prints and ensure dissemination copies are "independently understandable" by these users without requiring the assistance of experts.
- Establish and follow documented policies and procedures to ensure the information is preserved and enables the information to be disseminated as authenticated copies of the original or as traceable to the original.
- Make the preserved information available to the Designated Community.

The OAIS model provides some perspectives on the requirements to preserve digital objects, but it does not describe the processes necessary to perform this tasks. As formats, and the viewers needed to interpret and render these formats, become obsolete measures to preserve the content of an e-print and all related aspects such as look and feel, layout, structure and functionality, need to be taken. To this end, several strategies may be followed, such as migration and emulation to ensure continued access to the text itself and any other important characteristics. The repository implementer should consult practical examples provided by NedLib, the Library of Congress (Cundiff, 2002) and Harvard University Library (Abrams, 2002), to identify the most effective method of mapping the OAIS to their own organisation structure, and consider the issues raised by the OAIS model.

Rights to store and preserve e-prints

To protect themselves against claims of copyright infringement, repositories should ensure they have permission to preserve e-prints and maintain a record of such agreements. The legal requirements of preservation are often overlooked – a survey by the RoMEO project indicates that 31.8% (the largest group of respondents) took it on trust that the author had the right to deposit with them and provided permission to migrate the Work at a later date, without explicitly asking in a written statement (Project RoMEO, 2002). Unless the institution itself owns the copyright to deposited works, as is possible with e-theses repositories (e.g. Theses Alive!), consent will be required from the copyright holder to allow the repository to store and migrate the e-print as necessary.

The approach that the repository uses to acquire permission will vary in accordance with the type of content being stored (e-theses, journal articles, etc.) and the method of acquiring deposits. If the repository is actively contacting potential depositors (e.g. those who have had their work published in a journal), they will be required to confirm the copyright holder wishes to store their e-print within a repository and there are no other restrictions that prevent its use or migration. If the copyright holder is encouraged to voluntary deposit data within the repository – a deposit model used by many e-print repositories – they should be required to complete a deposit licence, similar to those used by the TARDis (http://tardis.eprints.org/) e-print archive or the University of California eScholarship Repository (http://repositories.cdlib.org/escholarship/).

The deposit licence can have many functions: it will establish the depositor (author or assignee) as copyright holder; clarify the repository's rights to remove, copy and alter an e-print for purposes of preservation and backup; and reduce their legal liability if a paper is found to be infringing copyright. For the author, this can provide reassurance that the repository is not claiming their rights in the paper, and make them aware of the type of services the repository is providing.

Preservation strategy

It is often unwise to delay preservation decisions until the value of an e-print, in terms of frequency of use, is established. The postponement of preservation action may prove to be more complex, labour intensive and costly if delayed (Research Libraries Group, 2001; James et-al, 2003). To ensure an e-print is maintained, the repository is encouraged to implement a consistent preservation strategy that establishes the requirements of an e-print at various stages in its lifecycle (James et-al, 2003).

The short-term requirements are quite limited, indicating the necessity that the repository provide some form of archival storage facility, make regular backups and store sufficient information to authenticate the digital master (Research Libraries Group, 2001).

The long-term use of e-prints is more problematic. Unlike print resources, software and hardware is required to decode the file format and display the intellectual content. This may be complicated by the variety of formats stored within the repository. It is common practice for e-print repositories to accept widely used formats such as ASCII, HTML, PostScript, Rich Text Format (RTF), and Portable Document Format (PDF) as well as, in some circumstances, formats that are used mainly in particular disciplines, such as TeX or LaTeX. As a result, the potential costs and risks for preservation are likely

to increase when a large number of diverse file formats are stored (Granger, Russell & Weinberger, 2000). It is therefore wise to establish a policy that allows the file formats of e-prints to be reviewed, either when they are submitted or at a set point in time (e.g. annually), and migrate the content as necessary according to specific criteria.

To assess the risks associated with e-print file format in their collection, the repository should consider how it would affect the repository's ability to provide long-term preservation of, and access to, the intellectual content held in each format. Potential criteria for assessment include the status of the file format specification, the extent to which the internal structure of the format is known, the availability of third party conversion tools to migrate the format, and existence of free viewers on different operating systems. Proprietary file formats are considered to present the greater risk to the preservation of e-prints over the long-term (James et-al, 2003), in some cases being poorly documented, requiring specific plug-ins or software that may not be available for all machines, and may (in some circumstances) require commercial software to view.

To minimize the threat of obsolete and proprietary formats, E-Print repositories should seek to adopt open, portable, or de-facto formats (such as XML, RTF and PDF) that are likely to remain accessible within contemporary software for the foreseeable future. The e-print should be designed so that its core content is independent (or as near as possible) from the means of access (CCSDS, 2002). The repository should also encourage authors to deposit e-prints in file formats that are based on open standards by providing them with information on the advantages of such file formats (RLG, 2001). Closed or specialist formats may be accepted provided that software tools are available to convert files from submission formats to supported dissemination formats (James et-al, 2003). For example, open source utility programs exist to convert LaTeX to PostScript or PDF. The National Archives' PRONOM (http://www.records.pro.gov.uk/pronom/) provides public references for identifying and migrating obsolete formats. Alternatively, services such as the Arts & Humanities Data Service (http://www.ahds.ac.uk) or the soon-to-be-launched Digital Curation Centre (http://www.e-science.clrc.ac.uk/web/projects/Data_Curation_Centre) may be able to provide advice on the most effective migration method and warn of potential problems that may occur when migrating data into a different or later revision of a format.

An e-print's long-term preservation can be safeguarded best when it is considered at the earliest point possible in the deposit process. Yet few repositories impose formal preservation procedures. To minimize costs and potential loss of intellectual content, repositories should impose formal preservation procedures that identify file formats at risk and establish the most effective method of migrating the content. Avoiding this issue is likely to increase the time and financial costs that will be incurred later when the file format or software an e-print is reliant on becomes obsolete.

Use of suitable metadata standards

The creation of suitable descriptive and technical metadata is an effective method of ensuring the preservation, as well as the interoperability of e-prints. One great advantage of the OAI (Open Access Initiative) protocol is the possibility to access any number of electronic archives in a uniform way and get records in at least one common metadata schema. All OAI compliant archives must provide unqualified Dublin Core metadata – a 15-element set of semantic terms (title, creator, subject, description, publisher, and type). Qualified Dublin Core offers further granularity of descriptors, to define elements more accurately.

For metadata it is useful to understand that preservation problems are not introduced by different methods of storing and accessing data, but by different definitions of semantic terms and incorrect use of data elements. Though Dublin Core provide a degree of interoperability by defining standard metadata fields, use of Dublin Core elements vary significantly. For example, "creator" may be defined as either the person that created the e-print, or the institution. Elements may also contain information that will vary in its expression. The content of the language tag for English documents may range from "en gb" over "English" to "19th century English with passages in French" (Fischer & Fuhr, 2001). To minimize potential issues when transferring metadata between systems, it is advisable to follow best practice guidelines whenever possible to ensure consistency. This will vary for different elements: the DCMI recommend the RFC 1766 for languages (e.g. 'en' for English, 'fr' for French, or 'en-uk' for English, as used by the United Kingdom), the ISO 8106 format for Date, the URI (Uniform Resource Identifier), DOI (Digital Object Identifier), or ISBN (International Standard Book Number) for the

Resource Identifier, and other formal classification methods (DCMI, 2003). When suitable standards do not exist, a controlled vocabulary (a restricted list of terms) and layout guidance should be adopted and documented to provide uniformity.

If a complex resource requires description and the Dublin Core is insufficient for the task, an extended metadata schema may prove useful. To ensure interoperability with other repositories, the metadata schema may be stored internally, and then mapped to Dublin Core so that other repositories and services (e.g. portals) can access information about the e-prints in the repository using the default OAI-PMH (at the loss of some information that cannot be stored as unqualified Dublin Core). The Theses Alive! project (http://www.thesesalive.ac.uk/), for example, utilizes the more specialised Electronic Theses and Dissertations (ETD) schema which includes elements that are not present in unqualified Dublin Core (display title, department, document type) or have less obvious meanings. Other specialized metadata formats provide subject or material specific elements (e.g. MARC records for books, ISAD (G) records for archival material, or TEI headers for electronic texts, IMS/IEEE LOM for e-learning).

E-Print Repositories typically rely on the OAI-PMH and the Dublin Core schema to define their metadata requirements and make the metadata interoperable. However, it should not be automatically assumed that Dublin Core will meet all of a repository's requirements and will automatically comply with existing standards. It is preferable to identify the most appropriate standards when implementing the repository, rather than performing the costly process of crossmapping and improving metadata at a later stage.

Summary

The legal and technical considerations necessary to preserve content within a repository are not unique to e-prints and can be identified in advance. Detailed planning is required to establish the repositories right to identify and migrate problematic or obsolete file formats. The use of suitable open standards for data and metadata will also limit the potential costs associated with preservation and is likely to improve accessibility to the resource.

References

Abrams, S. (2002) *An Archival Submission Information Package for E-Journals.* Retrieved on March 30, 2004, from: http://www.rlg.org/longterm/forum02/abrams.html

AHDS. (2003). *AHDS Licence form for Depositing Data.* Retrieved on March 29, 2004 from: http://ahds.ac.uk/depositing/licence.htm

Consultative Committee for Space Data Systems (CCSDS) (2002). *Reference Model for an Open Archival Information System (OAIS).* Retrieved on March 29, 2004, from: http://ssdoo.gsfc.nasa.gov/nost/isoas/wwwclassic/documents/pdf/CCSDS-650.0-B-1.pdf

Cundiff, M. (2002). *Towards an Archival Information Package for Audiovisual Materials.* Retrieved on March 30, 2004, from: http://www.rlg.org/longterm/forum02/cundiff.html

Dublin Core Metadata Initiative. (2003). DCMI Metadata Terms. Retrieved on March 29, 2004, from: http://www.dublincore.org/documents/dcmi-terms/

Fischer, G & Fuhr, N. (2001). *Heterogeneity in Open Archives Metadata.* Retrieved on March 29, 2004 from: http://www.is.informatik.uni-duisburg.de/bib/fulltext/ir/Fischer_Fuhr:01.pdf

Granger, S., Russell, K. & Weinberger, E. (2000). *Cost elements of digital preservation.* Retrieved on Mar 3, 2004, from http://www.leeds.ac.uk/cedars/documents/CIW01r.html

James, H, Ruusalepp, R, Anderson, S. & Pinfield S. (2003) *Feasibility and Requirements Study on Preservation of E-Prints.* Retrieved on November 25, 2003, from: http://www.jisc.ac.uk/uploaded_documents/e-prints_report_final.pdf

Morrison, A, Popham, M. & Wikander. K (n.d.). *AHDS Literature, Language & Linguistics guide to Creating and Documenting Electronic Texts.* Retrieved on March 29, 2004 from: http://ota.ahds.ac.uk/documents/creating/

National Archives (n.d.) *PRONOM*. Retrieved on March 31, 2004 from:
http://www.records.pro.gov.uk/pronom/

NedLib (2000). *Applying the OAIS Reference Model to the Deposit System for Electronic Publications* (DSEP). Retrieved on March 29, 2004 from:
http://www.kb.nl/coop/nedlib/results/OAISreviewbyNEDLIB.html

Project RoMEO (2002), *RoMEO Studies Series*. Retrieved on November 25, 2003, from:
http://www.lboro.ac.uk/departments/ls/disresearch/romeo/

Research Libraries Group (2001). Attributes of a Trusted Digital Repository:
Meeting the Needs of Research Resources. Retrieved on March 29, 2004 from:
http://www.rlg.org/longterm/attributes01.pdf

Tardis, (2003). *e-Prints Soton - Deposit Agreement*. Retrieved on March 2004, from:
http://tardis.eprints.org/discussion/e-Prints%20Soton%20-deposit%20agreeement.doc

Theses Alive! (2003). Theses Alive! Introduction. Retrieved on March 2004, from:
http://www.thesesalive.ac.uk/ta_home.shtml

SHERPA Project Document

Report on Preservation Standards
Gareth Knight
Arts & Humanities Data Service
31/03/2004