

Selection criteria for the preservation of e-prints

Circulation: PUBLIC
Gareth Knight
Arts & Humanities Data Service

Summary

This paper provides guidance on methods of identifying potential risks to e-prints, prioritizing content to be preserved, and methods of implementing preventive measures to ensure content remains accessible in the long-term.

Introduction

Access to information held in a digital form is dependent upon specific combinations of hardware and software. Access to the information, in terms of being able to display and read the content, is therefore highly susceptible to changes in the technological environment. With time, the file format used to store the content of an e-print, and the software used to display it will become obsolete – that is, the file format is no longer supported in new software, while the original software will no longer run on modern computers.

Although the majority of e-print repositories will not yet have encountered problems of software and file format obsolescence, these problems will become increasingly pressing (Lawrence *et al.*, 2000, p. 1). Because such problems tend to become apparent from about ten years after a new format or software product is introduced, even the oldest e-print repositories have had little experience of the difficulties involved in ensuring the long-term survival of digital data in an accessible form.

A cost model for e-prints

The lack of experience makes it difficult to establish the actual costs associated with maintenance and migration issues. Research studies have instead identified cost factors and likely scenarios in which they will occur. The method of obtaining and negotiation for use of e-prints, identification of the correct strategy to preserve and maintain accessibility, creation of discovery and administrative metadata, and the cost of storing e-prints are identified as cost factors relevant to preservation (Granger, Russell & Weinberger, 2000). To an extent, e-print repositories may avoid some of these costs: many require the author to complete a licence agreement form and create discovery metadata on a voluntary basis when submitting their e-print, which reduces the potential costs of pursuing potential authors. Storage costs are also low in comparison to the relatively small size of e-prints (James *et al.*, 2003, p43). In terms of preservation, costs are influenced by the completeness of submitted e-prints & metadata, collections policy of the repository and time period in which any migration process takes place. James *et al.* (2003) identifies four scenarios:

E-print Repository Scenario	Submission (including revisions)	Technical Obsolescence		
		Migrate	Remove	Retain
Preferred format with Complete metadata set (1)	Low	Low	Low	Deferred cost - high
Preferred format with Incomplete metadata set (2)	Medium	Low	Low	Deferred cost - high
Non-preferred format Complete metadata set (3)	Low if accepted as is Medium if migrate on submission	Low if migrated on submission Medium if migrated at this stage	Medium	Deferred cost - high
Non-preferred format Incomplete metadata set (4)	Medium if no migration High if migrated	Low if migrated on submission Medium if migrated at this stage	Medium	Deferred cost - high

Based upon a diagram by James *et al.*, 2003, p47

In most instances the deferred costs of storing an e-print & metadata in its original form will be high if the repository does not act to resolve problems at the submission or technical obsolescence stage. It is useful to take heed of the few actual cases that have occurred. The American Mathematical Society converted its archive of published journal articles from one TeX format to another. A computer program successfully converted 90 per cent of the articles, but 10 per cent had to be converted by hand (Jackson, 2002). For an e-print repository containing half a million articles, that problematic 10 per cent could mushroom into a huge and costly task (James et-al, 2003).

To minimize the risk that file formats will become inaccessible at a later date, or require costly remedial action, e-print repositories need to take an active role in preserving their holdings by validating their metadata and determining the appropriate technical strategy from the moment they are deposited.

Prioritising content to be preserved

One key question associated with the preservation of e-prints is "which documents should be preserved?" This is where selection and retention criteria are important, however there will be many repositories who will question why they need to prioritise e-print - the relatively small file size of e-prints in comparison to increasing hard disk capacity makes the costs of storing e-prints modest. Rather, it is the investment in terms of staff time and use of other resources that can become costly. In practical terms, each e-print is different, consisting of different versioning, layout information, structure, and plug-in components that will make it difficult to adopt a single approach to preservation. It is therefore wise to consider managing preservation by prioritizing papers that will most benefit from being preserved.

Criteria for prioritising e-print content may be derived from three broad categories: its contribution to research in a particular field or specific characteristics of the e-print that distinguish it from a published version (James et-al, 2003; Pinfield & James, 2003):

Contribution to research:

- The paper is cited frequently by other authors;
- The paper contains unique research that contribute to the understanding of a specific field;
- It is part of a wider collection of material deemed worthy of preservation

Distinguishing E-print characteristics

- The e-print allows wider and/or more convenient access in comparison to published journal papers;
- The e-print is a fuller version of the conventionally published paper;

Alternatively, a policy may be developed where only post-prints are preserved rather than pre-prints; papers accepted for publication are prioritised;

Although the prioritisation of e-prints for preservation is a distant consideration in comparison to the immediate problem of attracting depositors, planning for the long-term will help ensure that the migration and maintenance of at-risk content is economical, contributing to the sustainability of the e-print repository.

Measuring the risks to e-print accessibility

File formats define the rules used by application software to convert bits (the fundamental unit of digital data) into meaningful information that can be viewed and manipulated by a user. They should therefore be the focus of an e-print risk assessment that will then form the basis of a repository's preservation strategy.

Status of the file format specification

The specifications for the types of file format used to store e-prints can be classified into four different categories:

- *Proprietary formats* - non-public standards developed by software producers. Popular examples of this category, such as Microsoft Word, are described as de-facto standards.

- *Open formats* – well-documented, public standards that are distributed without restrictions. For example, the ASCII plain text format.
- *Industry standard formats* - public standards maintained and guided by an independent body. The standard is fixed or stable until the next revision is released. Examples include HTML and PDF formats.
- *Open/Industry standard format with proprietary extension* –Publicly documented file formats, merged with proprietary extensions. The extensions may not be public or may be covered by patents that restrict their use and manipulation in non-certified software (e.g., Microsoft's version of XML)

Both migration and emulation — the two best digital preservation strategies currently in use — rely on file format specifications being known and accurate. If it is not, the preservation strategies risk introducing distortion, loss of quality or data, or not being able to render the file usable at all. Open and industry standard formats are regarded as the safest choices for long-term preservation, because details of the format are publicly available, and there are likely to be a range of tools available that can read the format. Proprietary formats and format extensions are tied to the strategic direction and fortunes of individual companies, and tend to change frequently. However, when proprietary formats become widely used de-facto standards they are likely to remain in usage for a long time.

Support within current software

The degree of support within current software is a useful measure of the long-term accessibility of a format. Open, portable, or de-facto formats are likely to remain accessible within contemporary software for the foreseeable future. Contemporary versions of Microsoft Word, WordPad, OpenOffice, and various freely available open source conversion tools are able to display the proprietary Word 95 format – a format that has existed for nine years. In comparison, the Final Writer and WordWorth formats are poorly supported and require access to the original software application to view the content. The National Archives' PRONOM (<http://www.records.pro.gov.uk/pronom/>) or the soon-to-be-launched Digital Curation Centre (http://www.e-science.clrc.ac.uk/web/projects/Data_Curation_Centre) provide public references for identifying and migrating obsolete formats. Alternatively, services such as the Arts & Humanities Data Service may be able to provide advice.

Suitability of format for audience

Repositories should be aware of the suitability of file formats for their targeted audience and the possibility that this will change. The TeX and LaTeX formats, for example, are common within the scientific community but are practically unused by the humanities (James et-al, 2003) reducing the likelihood that tools tailored for these fields will be designed to display them. Repositories that receive e-prints stored in an inappropriate format for their audience should ensure the data is migrated to suitable formats.

The technical considerations contribute to the potential risks involved with storing e-prints. However, they do not necessarily indicate they *will* be rendered obsolete. The repository manager should seek advice from preservation experts to identify e-prints that have a high risk of becoming obsolete and prioritise preservation action to ensure the intellectual content is migrated successfully. For e-prints identified as low-risk, the repository may choose to delay preservation action until a significant concern to accessibility is identified.

Preventative measures to ensure long-term accessibility

The retroactive identification and migration of e-print content can be costly. Therefore, it is cost-effective to identify and resolve potential risks in the short-term. This can be achieved through the introduction of preventive measures to ensure long-term accessibility:

Capture technical and administrative metadata

To retain access to e-prints through multiple generations of hardware and software, it is important to know how the information was originally encoded (Cedars, 2002). A strategy can then be developed to decode it in the future. Unfortunately, very little of this preservation metadata is currently collected by repository software, to the extent that an e-print repository may not even be able to tell exactly what file formats it holds. Software that does collect basic format data (ePrints.org, DSpace) is currently unable to store version number or validate the

data. It may be HTML, but which version and is it suitably well formed? For a repository, it will likely prove more cost effective to collect this information soon after submission, rather than attempting to work it out years later once the formats have passed into history and information and expertise in them is rare.

Process information can also be stored to establish an audit trail of changes performed on an e-print. This allows the identification of when errors have been introduced, particularly when data migration has not been 100% successful. For example, typographical, character set, or formatting errors that may arise when using a particular software revision.

Specify a restricted range of deposit formats

Different communities are likely to self-archive e-print content produced in a diverse range of specialist software. Although this policy may prove effective in attracting depositors (James et al, 2003), it will increase the likelihood that a file format will become inaccessible before it is noticed and preservation action is taken (Granger, Russell & Weinberger, 2000). It will also increase the amount of staff time required to ensure that obscure data has been migrated correctly. It is recommended that repositories establish a restricted list of deposited formats appropriate to the academic community and monitor incompatibilities to ensure content stored in previous format versions can be rendered correctly in modern software.

Store content in an open format

To ensure content is viewable and minimise the risk of format and version incompatibility, repositories should store e-prints in open or well-documented formats that are well supported by free or affordable software available for multiple platforms (Windows, Linux, MacOS) (UKOLN, 2003). This can be achieved by stipulating that e-prints are deposited in specific formats; migrating content to another format after deposit; or performing on-the-fly transformation when the user requests an e-print in a designated format. Rich Text Format (RTF), Portable Document Format (PDF), HTML and ASCII are a few of the formats that meet these criteria.

Enhance descriptive metadata

Metadata is an essential part of the digital repository and good quality metadata offer significant benefits. Even though the practicalities of Dublin Core mean that depositing authors are only required to provide a limited amount of information, research has found that authors often make mistakes when completing metadata entry forms (Boyce, 2000, p. 414). It is therefore preferable if the repository can perform automated or manual reviews of submitted e-print metadata, correcting spelling mistakes, and comparing the supplied metadata with the intellectual content of the e-print to identify differences and extract suitable file format and other identification data (title, subject, author, creator, keywords).

Conclusion

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 implement migration and documentation policies at the earliest point possible when e-print is submitted to the repository. 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.

References

Boyce, P. (2000). For better or worse: preprint servers are here to stay. *College and Research Libraries News*, 61(5), pp. 404-407

CEDARS (2002). Metadata for digital preservation: the Cedars Project outline specification. Retrieved on January 29, 2004 from:

<http://www.leeds.ac.uk/cedars/colman/metadata/metadataspec.html>

Granger, S., Russell, K. & Weinberger, E. (2000). *Cost elements of digital preservation*. Retrieved on May 3, 2003, from <http://www.leeds.ac.uk/cedars/documents/CIW01r.html>

Harnad, S (2001). For Whom the Gate Tolls? Retrieved on January 29, 2004 from:
<http://www.ecs.soton.ac.uk/~harnad/Tp/resolution.htm#1.Preservation>

Jackson, A. (2002, January). From preprints to e-prints: the rise of electronic preprint servers in mathematics. *Notices of the AMS*, 49(1) pp. 23-31. Retrieved on May 3, 2003, from
<http://www.ams.org/notices/200201/fea-preprints.pdf>

James, H. et-al, (2003). Feasibility and Requirements Study on Preservation of E-Prints. Retrieved on January 29, 2004 from:
http://www.jisc.ac.uk/uploaded_documents/e-prints_report_final.pdf

National Library of Australia (1999). Preservation Metadata for Digital Collections. Retrieved on January 29, 2004 from: <http://www.nla.gov.au/preserve/pmeta.html>

National Library of New Zealand (2002). *Metadata Standards Framework – Preservation Metadata*. Retrieved on January 29, 2004 from:
http://www.natlib.govt.nz/files/4initiatives_metaschema.pdf

NEDLIB (2000) Metadata for long-term preservation.
Retrieved on January 29, 2004 from: <http://www.kb.nl/coop/nedlib/results/D4.2/D4.2.htm>

Pinfield, S. & James, H (2003) The Digital Preservation of e-Prints. Retrieved on January 29, 2004 from: <http://www.dlib.org/dlib/september03/pinfield/09pinfield.html#1>

UKOLN (2003). NOF-digitise Technical Advisory Service: FAQ. Retrieved on January 29, 2004 from:
<http://www.ukoln.ac.uk/nof/support/help/faqs/fileformats.htm>

QA Focus (2003). Matrix for selection of Standards. Retrieved on February 20, 2004 from:
<http://www.ukoln.ac.uk/qa-focus/documents/briefings/briefing-31/html/>