ARTICLES

D-Lib Magazine
April 2003

Volume 9 Number 4

ISSN 1082-9873

Preservation Metadata

Pragmatic First Steps at the National Library of New Zealand

Sam Searle
Digital Library Projects Leader
National Library of New Zealand
<sam.searle@natlib.govt.nz>

Dave Thompson
Digital Library Resource Analyst
National Library of New Zealand
<Dave.Thompson@natlib.govt.nz>

Introduction

The National Library of New Zealand Te Puna Mätauranga o Aotearoa (NLNZ) has a legislative mandate "to collect, preserve and make available recorded knowledge, particularly that relating to New Zealand" [1].

In common with other cultural institutions, the Library is undergoing a period of intense change brought about by the quantity of digital resources that must be managed and the knowledge that the rate at which we accumulate this material will dramatically increase year on year. The complexity of digital objects is a concern, as is the rising proportion that are "born digital" rather than as digital copies of analogue items from the Library's collections.

NLNZ is adopting a holistic approach to the long-term management of its digital assets. The Library has established a Digital Library Transition Team to:

- Develop and implement business process workflows
- Specify infrastructure for digital material, e.g., storage, access, data authentication
- Research and develop a range of Digital Library activities, e.g., metadata (resource discovery, preservation, structural) and persistent identifiers
- Pilot web harvesting for the capture and preservation of New Zealand web sites
- Implement production processes for bulk digitisation of textual materials.

The primary objective is that processes for digital objects become "business as usual" for the Library.

This includes activities relating to digital preservation. In contrast to the inertia that may seem an understandable response to what some describe as an unmanageable flood of digital materials, NLNZ is developing pragmatic business-oriented processes for managing this material, for the long term.

One component of the digital preservation puzzle is preservation metadata. NLNZ has developed a Preservation Metadata Schema [2] designed to strike a balance between the principles expressed in the OAIS Information Model [3] and the practicalities of implementing a working set of preservation metadata. This tension has informed a recent OCLC/RLG report [4] and work at the University of North Carolina [5]. A pragmatic response to this environment is required, one that recognises the need to implement a workable solution within existing resources and organisational structures.

This article introduces the NLNZ schema, describes the environment in which it was conceived and identifies areas of further development, which will include:

- Developing data definitions for the elements in the schema
- Designing a repository based on those data definitions
- Investigating and developing tools for automatically extracting metadata to populate the repository.

The NLNZ Preservation Metadata Schema

The NLNZ schema identifies the data that the Library will collect and maintain. This relates to the Preservation Master held in the Digital Archive, but could also cater for an object that is not or is no longer a Preservation Master, e.g., the CD-ROM on which the Library received the original digital object or a previous Preservation Master that has been superseded through hardware or software obsolescence.

The Preservation Master will be a "best effort" creation of a working preservation object. It will be a rendition of some form of "original" as supplied to or acquired by the Library, in a file format that can be preserved, managed and disseminated over time. The Preservation Master is dynamic and will be subject to processes such as migration during a lifecycle of creation, use and eventual replacement. At any time there can be only one Preservation Master for an object and maximum preservation effort will be applied whilst it has that status.

As shown in the diagram below, the NLNZ schema is split into four entities.

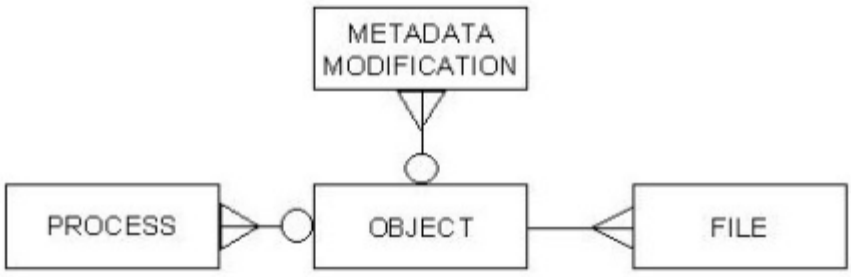

Figure 1. The four entities of the NLNZ schema.

Entity 1 - Object contains 18 elements describing the logical object, which may exist as a file or aggregation of associated files. These elements identify the object and describe those characteristics relevant to preservation management.

Entity 2 - Process contains 13 elements that record the complete history of actions performed on the objects. It includes the objectives of a process, who has given permission for the process, critical equipment used, and the outcomes of the actions taken. An audit trail of date/time stamps and responsible persons and/or agencies is constructed.

Entity 3 - File contains technical information about the characteristics of each of the files that comprise the logical object identified in Entity 1. Nine elements are common to all file types, and further elements are specified for certain categories of file (e.g., image, audio, video, text). Entity 3 will develop further in light of emerging standards such as *NISO Z39.87 Technical Metadata for Still Images* [6].

Entity 4 - Metadata modification contains 5 elements and records information about the history of changes made to the preservation metadata. This acknowledges that the record is itself an important body of data that must be secure and managed over time.

Although an object may have multiple files or processes for which data is recorded, each set of preservation metadata will pertain to a single logical object. This is an arbitrary construct allowing the Library to differentiate between the following types of digital objects:

- **Simple objects**: One file intended to be viewed as a single object (e.g., a word-processed document comprising one essay).
- **Complex objects**: A group of dependent files intended to be viewed as a single object (e.g., a website or an object created as more than one file, such as a database), which may not function without all files being present in the right place.
- **Object groups**: A group of files not dependent on each other in the manner of a complex object (e.g., a group of 100 letters originally acquired on a floppy disk). This object may be broken up into (described as) 100 single objects or 4 discrete objects containing 25 letters each, or it may be kept together as a single logical object ("Joe Blogg's Letters").

Practice will determine the viability of this model, especially in relation to complex objects.

More information about the schema and each of the elements it contains is available at <http://www.natlib.govt.nz/en/whatsnew/4initiatives.html#meta>.

Schema development: international context and local business requirements

The schema developed in light of international endeavours relating to preservation metadata, particularly work undertaken by the National Library of Australia [7], as well as initiatives through the CEDARS programme [8], the OCLC/RLG Preservation Metadata Working Group [9] and the emerging consensus regarding the role of OAIS. These efforts provided a useful framework, but they were not immediately applicable to NLNZ's business requirements and environment.

Much of the work to date has occurred at a largely theoretical level. The standards and guidelines available are not yet backed up by attention to business processes. This more practical focus will inevitably emerge as organisations develop and maintain working preservation metadata repositories, but the current lack of documented experience poses some risks for organisations needing to implement preservation metadata schemas sooner rather than later. As Seamus Ross has argued, "Concentration on the definition of metadata divorced from the processes that need to be undertaken using metadata, will result in the creation of guidelines of limited value because they will not reflect the data environment" [10].

The development of digital preservation activities requires in-depth knowledge of the Library's

business activities, which revolve around processes such as:

- Acquiring digital objects, both published and unpublished, in a variety of file formats
- Storing many digital objects comprising gigabytes or terabytes of data
- Processing large volumes of material, e.g., migrating multiple objects to avoid format obsolescence
- Disseminating digital objects to users in easy, secure, and meaningful ways.

We recognise that we will need to achieve our aims with limited resources, in a budgetary environment where responses to a changing electronic world are being developed within baseline funding. There is little concrete information about the immediate and long-term costs of digital preservation, although it is generally acknowledged that the costs will be greater than for analogue materials. The schema addresses these limitations by stressing that the resources we do have must be used efficiently. In this context, NLNZ's approach is an integrative one in line with Joint Information Systems Committee/National Preservation Office (JISC/NPO) findings: "Rather than attempting to isolate a global preservation cost, we should assume that there are some preservation costs associated with all the elements involved in the lifecycle of a digital resource" [11].

We believe that if digital preservation is to be incorporated into the Library's routine business, it must become as productionised as other processes such as cataloguing. We acknowledge upfront that bulk processing (mass migration and/or emulation) will be required and that hand-crafting, while sometimes necessary, will be avoided wherever possible. The schema suggests that successful implementation within a resource-constrained environment will require at least three things: 1) limiting the scope of preservation metadata; 2) maximising potential for automation; and 3) ensuring change control for metadata.

Limiting the scope of preservation metadata

We focus only on the data that is key to digital preservation. Wherever possible we have stripped out elements that more properly support other activities such as the preservation of analogue formats or the resource discovery and rights management of disseminated digital objects. We envisage that these types of metadata will rarely be required in order to undertake digital preservation activities and that they can be drawn upon as needed from other sources rather than routinely collected as part of our preservation strategy.

Where preservation and other functions do require a common element, this has been identified within the schema, so that further rationalisation can occur during implementation. For example, where elements are collected for preservation and also required for resource discovery (or vice versa) — for example, **1.2 Reference Number** and **1.8 Structural Type** — this has been noted so that duplication can be avoided during repository development.

The schema also removes the need to collect preservation metadata about dissemination formats. This differentiates the NLNZ Schema from some other schemas that require the identification and categorisation of the relationships between different manifestations of the object. It is clear that certain metadata required for preservation — for example, details of the internal structure of a complex object — may also be needed to manage dissemination formats. However, we believe that relationships between preservation objects and the dissemination formats that they generate can be efficiently documented through the use of persistent identifiers and consistent file storage structures. This effectively removes the need to collect dissemination-related metadata against preservation objects.

Maximising potential for automation

The drive towards automatic population of the maximum number of elements is most clearly demonstrated in the NLNZ Schema's focus upon the Preservation Master. As noted above, the Preservation Master will be a "best effort" representation of the material acquired by the Library, distilled into one of a small number of preferred file types. It is the Preservation Master, not the "original", that will be subject to preservation processes to continually transform it from obsolete into current formats.

The concept of the Preservation Master is in line with guidelines provided by Resource: "Narrowing the range of file formats handled streamlines the management process and reduces preservation costs" [12]. Through working with preferred file types we envision that most preservation-related processes will become more standardised and the number of required processes limited. As a consequence, the range of values that need to be collected as preservation metadata will also be reduced to a more manageable set amenable to automation.

Change control for metadata

In developing the schema, an important question arose: why provide an audit trail for the object but not for the metadata? The decisions relating to preservation processes and the steps involved in those processes will not stand up to scrutiny in ten, fifty or a hundred years time if the metadata records in which they are documented are not subject to similar processes of change control as the preservation objects themselves.

It is for this reason that the NLNZ Schema has a built-in audit trail. Entity 4 - Metadata Modification will enable the Library to track changes to the metadata record and the person responsible for them. This ensures that the goal of long-term integrity is not only applied to digital objects but also to the related metadata records.

Although these characteristics of the NLNZ Schema may seem superficially to depart from existing work, in fact they reflect the widespread tension noted above, between high-level conceptual models for preservation metadata and the pragmatism required to actually implement them.

Further work

We are engaged in a variety of other activities to support the work done on the Preservation Metadata Schema and to ensure that management of digital objects continues to be aligned with the business objectives of the Library. Three pieces of work are key to this: 1) an implementation data model; 2) a preservation metadata repository; and 3) a preservation metadata extract script.

Implementation data model

Whilst the Preservation Metadata Schema offers a generalised conceptual model of preservation metadata, it is not an implementation model. NLNZ is currently undertaking data modelling work that will inform the implementation of the schema. This work is due for release in May 2003.

Preservation metadata repository

Following the production of data definitions, NLNZ will develop a metadata repository, with a view to integrating it with our existing systems for other types of metadata. Ultimately we hope to incorporate preservation metadata into our core portal product, Endeavor Information Systems' Encompass. Until that takes place, it is likely that the Library will need to develop an interim solution.

Preservation metadata extract script

NLNZ is also developing a tool that automatically extracts metadata embedded in commonly found file types. This automation is essential given the number of files involved and the complexity of their associated metadata. The script, which is currently moving beyond the proof-of-concept phase, produces an XML report of that metadata identified as important to preservation. This will then be uploaded to the metadata repository. The script's flexible modular architecture will allow the addition of extraction components for new file types and for the fine-tuning of the XML output as required. This tool is due for completion in June 2003.

Conclusion

The desired outcome of the activities described in this article is the integration of digital objects into the NLNZ collections as simply another type of material we collect, preserve and make available. To move digital preservation into a business-as-usual framework requires a change in language and in thinking, away from describing the requirements of digital preservation as 'problematic' and the accumulation of digital material as "an unmanageable flood" [13]. The risk of such rhetoric is that digital preservation continues to be perceived as outside the norms of business processes.

There is no doubt that the Library will face many challenges in ensuring that digital objects remain functional into the future. Our Preservation Metadata Schema will need to evolve as these challenges arise and are resolved. In the immediate future, the schema is particularly likely to be influenced by the following:

- Research and development in the area of emulation (especially of complex objects)
- The evolution of METS, the Library of Congress Metadata Encoding and Transmission Standard [14]
- The practical experience that NLNZ and other organisations will gain from managing a wide range of digital objects.

In the meantime, we are working to implement the Preservation Metadata Schema. This is one of the first steps to ensure that the preservation of our digital objects takes place within a set of agreed processes and policies. In time we hope these policies and processes will become as standardised as those that currently relate to other Library activities such as acquisitions, collection management and bibliographic description.

References

[1] National Library of New Zealand Te Puna Mätauranga o Aotearoa. 2001. *The 21st Century: The Strategic Direction of the National Library of New Zealand Te Puna Mätauranga o Aotearoa. A Revised Framework for Planning.* <http://www.natlib.govt.nz/en/about/1pubframework.html>

[2] National Library of New Zealand Te Puna Mätauranga o Aotearoa. 2000. *Metadata Standards Framework: Preservation Metadata Schema.* Available at <http://www.natlib.govt.nz/en/whatsnew/4initiatives.html#meta>

[3] International Organisation for Standardisation. 2003. ISO 14721:2003: *Space data and information transfer systems - Open archival information system - Reference model.* Also available as: Consultative Committee for Space Data Systems. 2002. CCSDS 650.0-B-1. *Reference model for an Open Archival Information System (OAIS).* Blue Book. Issue 1. January 2002.

<http://www.ccsds.org/documents/650x0b1.pdf>

[4] OCLC/RLG Working Group on Preservation Metadata. 2002. *A Recommendation for Preservation Description Information.* <http://www.oclc.org/research/pmwg/pres_desc_info.pdf>

[5] North Carolina ECHO, 2003. *Exploring Cultural Heritage Online.* <http://www.ncecho.org/>

[6] National Information Standards Organisation. *NISO Z39.87 Data Dictionary - Technical Metadata for Digital Still Images.* <http://www.niso.org/standards/resources/Z39_87_trial_use.pdf>

[7] National Library of Australia. 1999. *Preservation Metadata for Digital Collections -Exposure Draft.* <http://www.nla.gov.au/preserve/pmeta.html>

[8] Cedars: CURL Exemplars in Digital Archives. 2002. *Cedars Guide to: Preservation Metadata.* <http://www.leeds.ac.uk/cedars/guideto/metadata/guidetometadata.pdf>

[9] OCLC/RLG Preservation Metadata Working Group, <http://www.oclc.org/research/pmwg>

[10] Ross, Seamus. 2000. *Changing Trains at Wigan: Digital Preservation and the Future of Scholarship.* London: National Preservation Office. Also available at <http://www.bl.uk/services/preservation/occpaper.pdf>

[11] Mary Feeney, ed. 1999. *The Digital Culture: Maximising the Nation's Investment. (A synthesis of JISC/NPO studies on the preservation of digital materials).* <http://www.ukoln.ac.uk/services/elib/papers/other/jisc-npo-dig>

[12] Jones, M. & Beagrie, N., for Resource: The Council for Museums, Archives and Libraries. 2001. *Preservation Management of Digital Materials: A Handbook.* The British Library: London. Also available at <http://www.dpconline.org/graphics/handbook/index.html>

[13] University of Heidelberg Institute for Chinese Studies. *Digital Archive for Chinese Studies: About DACHS.* <http://www.sino.uni-heidelberg.de/dachs/intro.htm>

[14] Library of Congress. Metadata Encoding and Transmission Standard. Official Website. <http://www.loc.gov/standards/mets/>

(All URLs accessed 20 March 2003.)

(On April 16, 2003 the email address for Sam Searle was corrected.)

DOI: 10.1045/april2003-thompson