Managing digital objects and their metadata: challenges and responses

Adrienne Kebbell and Douglas Campbell
National Library of New Zealand, Wellington, New Zealand
Tel. +64 4 474-3000
Fax +64 4 474-3140
Mail: adrienne.kebbell@natlib.govt.nz, douglas.campbell@natlib.govt.nz

Abstract: This paper looks at the challenges facing institutions managing ever-increasing numbers of digital objects over the long term. It considers the current environments and challenges for managing preservation digital objects, the metadata used to manage them, and the business processes that support these activities, and documents the National Library of New Zealand's responses to these challenges. These responses include policies, schemas, systems, and procedures in both implementation and development.

Keywords: digital object; standards; metadata; preservation; access; persistent identifier; persistent locator; file format; file role; application profile; crosswalk; OAIS; FRBR; Dublin Core; METS; EAD; DRD; RDF; URI; DOI; Handle; PURL, National Library of New Zealand; Te Puna M_tauranga o Aotearoa.

1 Introduction

Institutions around the world are accumulating more and more digital objects. The task of managing these objects throughout their life cycle, especially for institutions tasked to preserve them in perpetuity, becomes more complex the more deeply it is investigated.

The National Library of New Zealand Te Puna M_tauranga o Aotearoa (NLNZ) has a number of initiatives underway to build its digital library infrastructure. Some of these were piloted in 2002 when NLNZ released its *Discover – Te Kohinga Taonga* (1). Discover presents over 2,000 digitised images, audio and video clips in the context of the nation's school arts curriculum and was built using Dublin Core metadata in RDF/XML format.

NLNZ is now turning its attention to the challenges inherent in dealing with 'complex digital objects'. This includes both digitised physical resources and those 'born digital', such as multi-file e-text publications, digital archival manuscripts or collections of papers. The challenges fall broadly into three areas:

1. Managing, storing, delivering and preserving the digital objects
2. Generating, collecting and managing the metadata used to manage the objects
3. Managing the business processes around both of these.

In this paper we examine the challenges in the context of current global thinking, and look at the strategies NLNZ is pursuing to deal with them. Given that this is a fast-developing area, the solutions presented are interim steps to an overall infrastructure for the National Library's digital library.

2 An NLNZ Framework for Managing Digital Objects

There are many facets to the management of digital objects. NLNZ is establishing a framework for all of the processes needed to select, ingest, describe, manage, disseminate and preserve different kinds of digitised and born digital resources. The Open Archival Information System [OAIS] Reference Model (2) provides a useful framework, but the challenge is to translate this conceptual model into practical processes and supporting systems. At present a number of independent systems support the key OAIS processes. Some of these systems leverage NLNZ's investment in products from Endeavor Information Systems – the "Voyager" integrated library management system and the "ENCompass" digital management system – while others have been developed internally, and still others are identified as a gap.

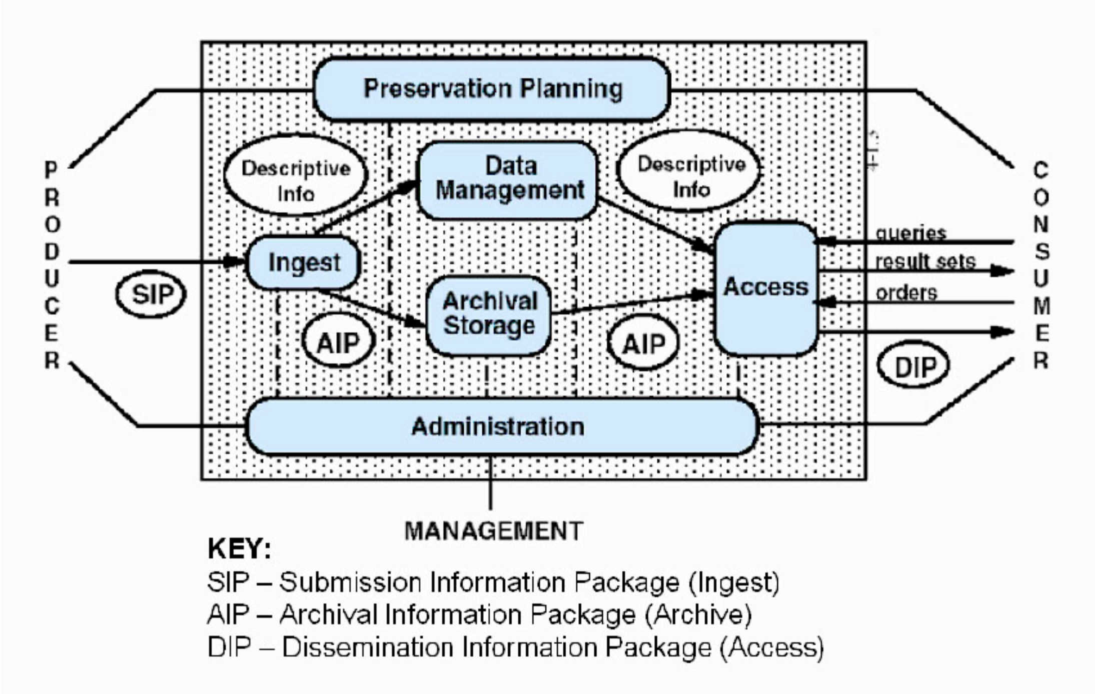


Figure 1 – The OAIS Reference Model for digital preservation (3)

Preservation adds another level of complexity to the set of tasks that organisations face as they attempt to assimilate digital content into business workflows. NLNZ has a legislative mandate 'to collect, preserve and make available recorded knowledge … related to New Zealand' and increasingly this material is digital.

NLNZ must ensure the authenticity and integrity of digital objects in perpetuity. The attributes of repositories with this type of responsibility and the potential for certification in this area have been described by RLG/OCLC in their papers on the 'Trusted Digital Repository' (4).

Other organisations will have a lesser imperative to preserve digital resources for the very long term, but will have some understanding of the technical, legal, organisational and social challenges involved, and collaboration will be essential.

2.1 International protocols, standards and best practices

In developing internal digital protocols NLNZ builds on international standards and experience. In this environment solutions seem always to be tantalisingly just around the corner, though surely never so close as they are now. Some standards and protocols that NLNZ tentatively chose two to three years ago have become much more mainstream. These include the CNRI Handle system, the Metadata Encoding Transmission Standard [METS], and the PDF file format that although proprietary, is now such a common dissemination format that a forward migration path is assured.

2.2 Complexity of multi-file objects
NLNZ's experience in the *Discover* pilot was limited to single image files. The objects NLNZ is now considering often contain multiple component files, for example web sites, CD-ROMs, diskettes of correspondence in word-processing files or accounts spreadsheets. Some of these files are self-contained (e.g. a single spreadsheet), while others are dependent on additional files for their operation (e.g. HTML files requiring GIF image files to build up a web page, or an executable file requiring supporting DLL library files in order to operate).

This inherent complexity can be compounded by actions taken as part of preserving and disseminating digital objects. Ensuring long-term authenticity and accessibility involves migrating and transforming files over time into more manageable formats, and converting them into suitable formats for dissemination.

NLNZ groups objects by their complexity so that processes performed on them are appropriate. These categories are based on the original ‘conceptual object’ as determined and described by the curators, rather than the particular structure of the file(s). We define Simple, Group and Complex objects as follows:

- **Simple digital objects** – A simple digital object consists of a single file that is intended to be viewed as one conceptual object, e.g. a Word document or TIFF image.
- **Digital object group** – A digital object group consists of a set of independent but related files that have been collectively described, e.g. a floppy disk containing 100 letters. Each file is accessible independently (as a Simple object), but its relationship to other objects in the group provides valuable context.
- **Complex digital objects** – A complex digital object consists of a group of dependent files intended to be viewed as a single conceptual object, e.g. a web site or CD-ROM. Often there is only one entry point.

In an attempt to understand the technical and business issues associated with complex objects, NLNZ created a testbed working with a group of objects called the 'Survey Objects'. These were selected as being representative of the spectrum of born-digital materials that NLNZ will acquire (published/unpublished, online/offline, simple/complex; and covering a range of MIME types, e.g. .htm, .pdf, .mdb, .doc, .dot, .exe, .xlm. The challenges that NLNZ has experienced in managing and delivering complex digital objects throughout the digital continuum has informed the practical aspects of our digital framework.

Figure 2 illustrates the growing complexity of the metadata components, even within simple objects. These examples show the use of a single Persistent Identifier [PID] for each 'conceptual object' or independently accessible component of an object group. NLNZ will review this level of identity if it proves impractical for particular kinds of objects not yet tested

.

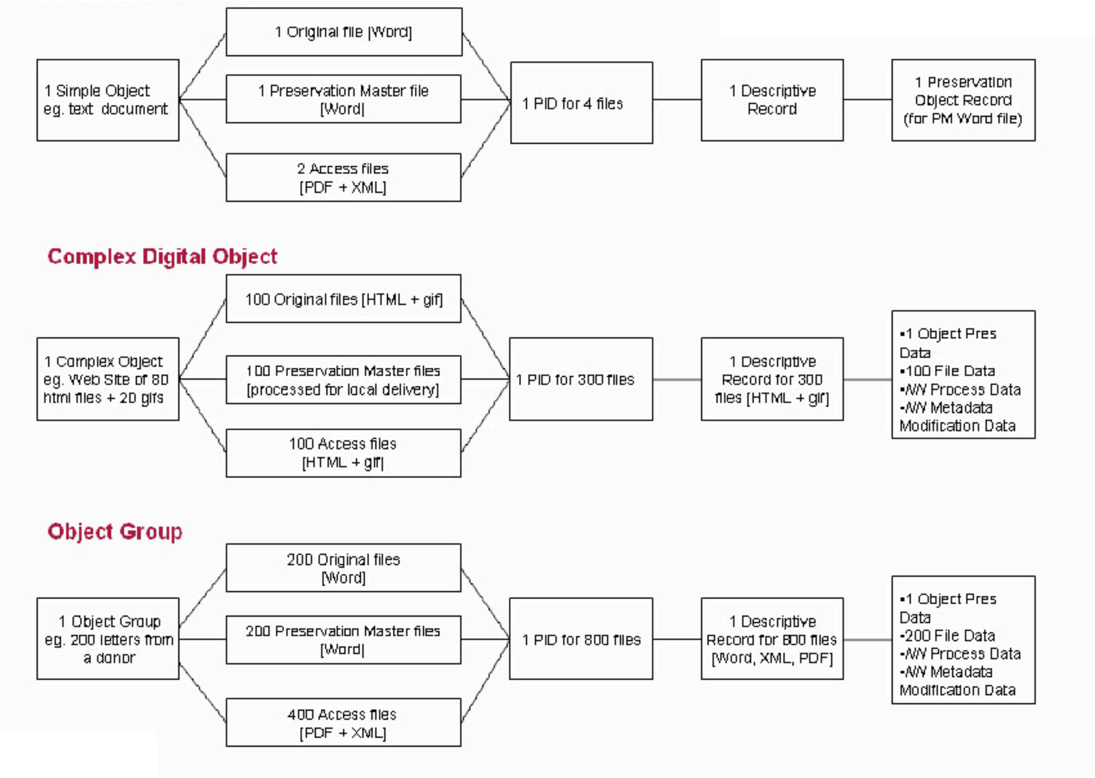

1 Simple Object eg. text document
1 Original file [Word]
1 Preservation Master file [Word]
2 Access files [PDF + XML]
1 PID for 4 files
1 Descriptive Record
1 Preservation Object Record (for PM Word file)
Complex Digital Object
1 Complex Object eg. Web Site of 80 html files + 20 gifs
100 Original files [HTML + gif]
100 Preservation Master files [processed for local delivery]
100 Access files [HTML + gif]
1 PID for 300 files
1 Descriptive Record for 300 files [HTML + gif]
•1 Object Pres Data
•100 File Data
•N/N Process Data
•N/N Metadata Modification Data
Object Group
1 Object Group eg. 200 letters from a donor
200 Original files [Word]
200 Preservation Master files [Word]
400 Access files [PDF + XML]
1 PID for 800 files
1 Descriptive Record for 800 files [Word, XML, PDF]
•1 Object Pres Data
•200 File Data
•N/N Process Data
•N/N Metadata Modification Data

Figure 2 – Simple, Complex and Group objects have multiple components

3 Selection

It is useful to examine the major phases of the digital continuum in more detail.

The selection process for published digital resources is supported by a specialised Voyager database. This is used to record decisions leading to the selection or rejection of a resource for the Digital Archive, and any rights or use negotiations that may have taken place with the publisher. Some of this data is imported into the Integrated Library System when the item is acquired.

Selection of unpublished digital resources continues to be recorded in our archival tool, TAPUHI.

4 Ingest

NLNZ is evaluating a number of tools to support ingest of born digital and digitised, and e-resources of all sorts (e.g. learning objects, e-journals, full-text databases) to the digital archive. These include D-Space from the MIT , the PANDORA Archiving System (PANDAS) from the National Library of Australia, and the Fedora Open-Source Digital Repository Management System from the University of Virginia and Cornell University. We are also exploring the possibility of building our own digital repository to encompass all of the preservation processes from selection to long-term preservation.

5 Identifiers

Persistent identifiers [PIDs] are critical to the successful long-term management of digital objects. Because of their very persistence, PIDs require careful consideration of both form and content from the outset.

NLNZ curators assign identifiers to new archival resources using a purpose built Internal Identifier Database [IID DB]. This is a simple MS Access Client running on top of a MS SQL Server database. The system assigns IIDs (discussed below) that become the locally unique component of a PID, and incorporates this number into the structured filenames for new objects. The IID DB Client is focused on integration into workflows: it provides data to configure uploads of new objects into the Digital Archive, to generate appropriate derivatives, and to support the replacement of digital objects. The Client provides an integrated view for curators of the object together with the associated administrative data.

5.1 Identifier characteristics

In our search for a preferred identifier framework NLNZ investigated the more important characteristics of identifiers (5)(6).

- **Granularity**: The most basic identity question to resolve is: "What do we need to identify?" It appears the best answer is "Whatever we need to identify". Determining what we need to identify is discussed below [see Applying Identifiers].

  Our identifiers will be made globally unique by using URIs, as recommended by the W3C's recent draft *Architecture of the World Wide Web* (7), so they can "stand alone", e.g. "hdl:1727.11/1854" instead of "Alexander Turnbull Library's local reference number EP-1957/643".

- **Intelligence:** The danger of intelligent identifiers lies in being unable to anticipate future changes that may render them inaccurate. NLNZ considered dumb identifiers to be safer in the long-term, though this places more reliance on external intelligence, for example in metadata.

- **Actionable:** The danger of actionable identifiers is the ease with which location and identity can be confused. An entity may exist in multiple locations so using a location as an identifier (e.g. a URL) may mean identifying as multiple resources something that should be considered a single entity. Whether our identifiers should be actionable or not depends on our requirements for the identifiers, but NLNZ makes a clear distinction between the two.

- **Persistence:** We recognise the persistence of our identifiers depends on our requirements for it and our level of commitment to its persistence, as discussed below [see Persistent Identifiers].

- **Extensibility:** We intend to follow as generic a scheme as possible, to follow international standards, and to be application independent.

5.2 Internal identifiers

NLNZ sees the identifier question as having two domains, internal and external. Most external identifiers consist of: a type prefix, an issuing authority and a local identifier, so the internal identifier is seen as the first priority to solve.

NLNZ developed an "Internal Identifier" (IID) scheme, which consists of a simple running number. The characteristics it exhibits are:

- unique – within NLNZ
- dumb
- not actionable – no need to be
- persistent – numbers are not reused
- extensible – simple numbers can be applied to anything.

5.3 Applying Identifiers

The question of which resources get an IID is more complex and includes the granularity question.

To help determine what we were trying to identify and why, NLNZ looked at IFLA's "FRBR" model (8), which is closely related to both the <indecs>' framework (9) and the DOI framework (10). We found FRBR's to be useful for the "big picture" (how large numbers of disparate resources relate to each other) but less developed around components below the manifestation level. However, NLNZ work could map successfully into FRBR, which also provided a workable model for relating digital and non-digital resources to each other [see figure 3].

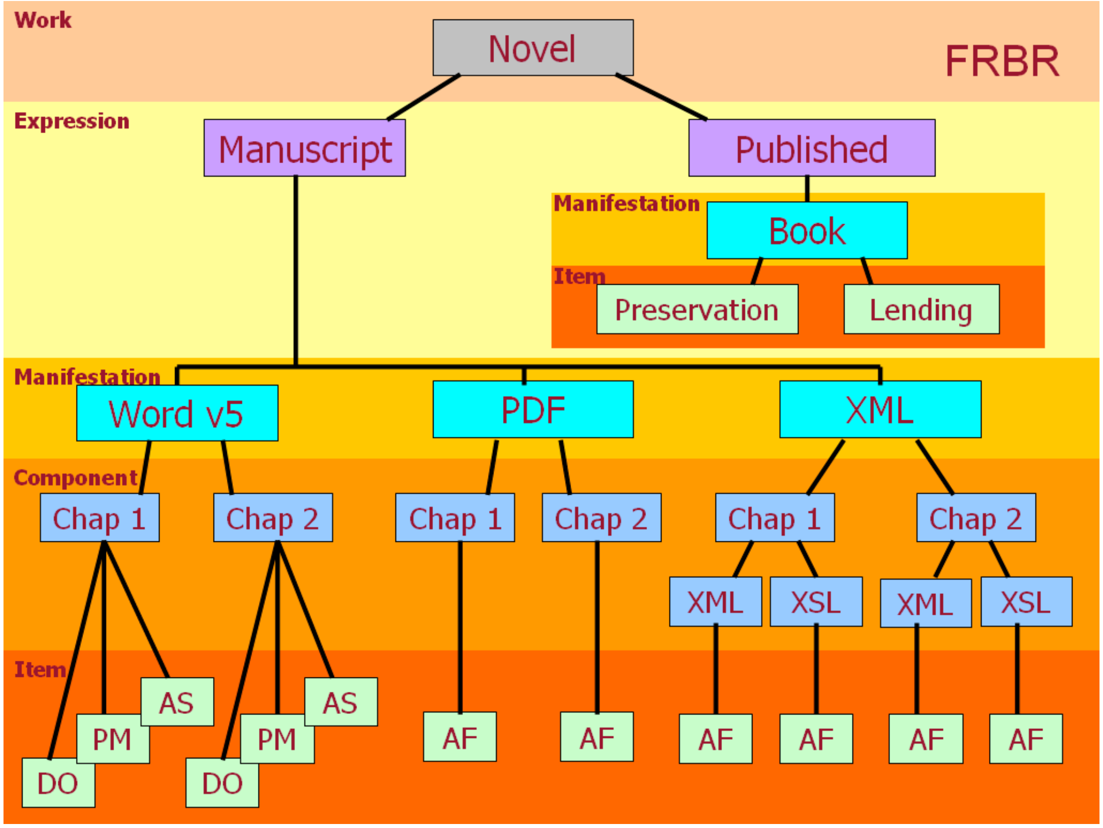


Figure 3 – Digital files occupy the lower levels of FRBR but FRBR provides a relationship with non-digital materials. The item file labels shown use NLNZ Role Codes.

To complement this, NLNZ developed a classification for file roles, as discussed below [see File Naming Conventions].

5.4 Persistent identifiers and locators

Having examined the options for PIDs, NLNZ discarded URNs (not actionable), URLs (persistence is reliant on the level of commitment), and D.I.Y. (do-it-yourself; doesn't follow standards). This left three main contenders: DOIs (expensive and closed/proprietary), Handles (open system), and PURLs (whose future appeared uncertain). We chose to pilot the Handle server, with each Handle consisting of NLNZ's Handle prefix and our IID.

As a result of assigning Handles to the 2,000+ objects in *Discover*, NLNZ discovered the following limitations of Handles:

- Native support in Web browsers is lacking.
- Resolution to multiple locations is supported, but it is not possible to specify which is preferred in an HTTP request.
- There are questions over the scalability of the database.
- A development path is not assured.

NLNZ is now re-evaluating DOIs and PURLs, and is looking at splitting its persistent identifiers into:

- **Persistent identifiers** [PID] – that persist in perpetuity and are assigned at the "conceptual" level of an object
- **Persistent locators** [PL] – file locators that persist, but only for the length of the life of the file

This split acknowledges that "persistent" means different things to the Web community (i.e. no "404 – not found" errors) and the archive community (i.e. in perpetuity). NLNZ would guarantee the PL links to each derivative or component file are persistent while those files are the "best current format", but over the decades, as they become obsolescent, the links would become inoperative. It is the PID links (at the conceptual object level) that are the reliable, permanent access points. At all times the original bitstream would still be available, and transient formats provided so that the current-day audience can view them easily without requiring special resources.

6 Storage and Preservation of Files

NLNZ already has some 750,000 files in its digital store, and it was in anticipation of very large numbers of digital resources that it decided to store them outside of a database, in a structured Unix directory.

Each filename in the digital store must be unique as some batch processes will gather files from separate directories. Unique filenames also make it easier for users downloading files (e.g. there may be a lot of "letter.doc" files). NLNZ decided to include the identifier in the filename, ensuring that all files have an identifier that travels with them. This also alleviates human error and lost files.

The exception to this approach is multi-file or complex objects, where use is dependant on consistent filenames and file locations (e.g. HTML pages pointing to HTML/GIFs or executables pointing to DLLs). These must be stored with the original directory structure and filenames intact.

Issues that need to be addressed at the directory level include:

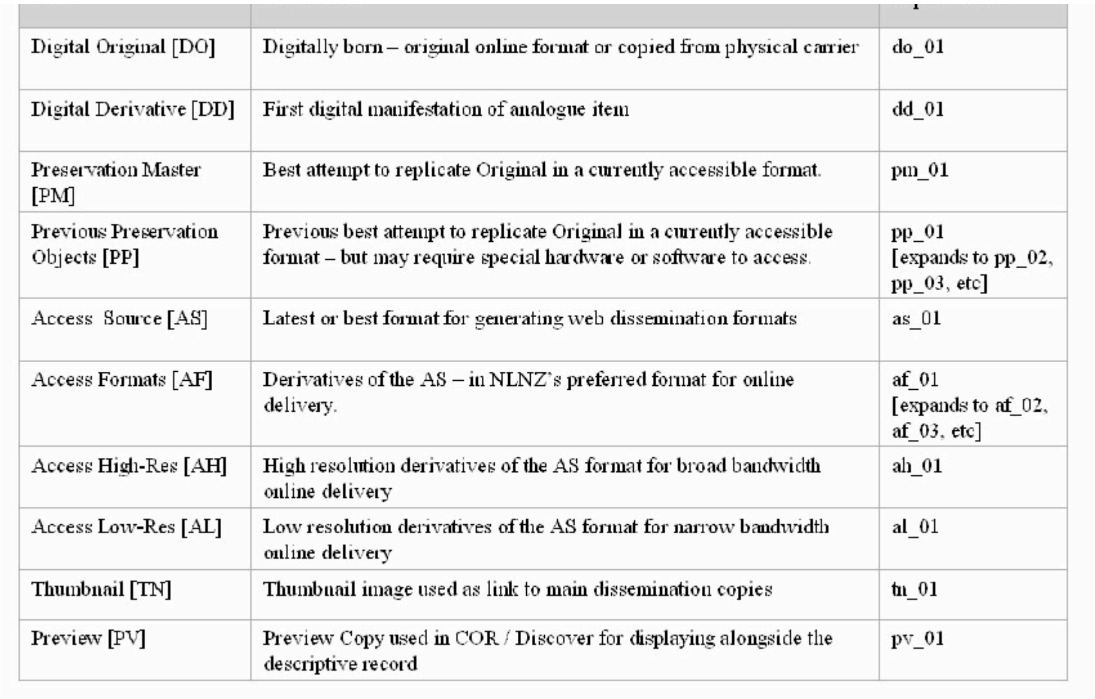
- predictability of the location/path for a given file, e.g. able to predict the path when only the file's identifier is known
- storing multiple derivative formats from an object either together or separately grouped by format
- balancing the maximum number of files in each directory against minimising the empty space in each, and so considering the differing sizes of different file types (e.g. a thumbnail image versus a full length movie)
- setting directory and file permissions
- catering for future growth, including conversion of formats that may change the file size yet need to remain in the same location/path for persistence of access.

Above the directories, NLNZ currently places files into disk partitions that are assigned arbitrarily but will not change over time, and that have headroom for future file conversions.

6.1 File Naming Conventions

In order to make filenames unique within NLNZ, names for simple and group objects follow the format: "IID_Role_Instance.Ext":IID followed by a code representing its role, a number for which instance of that role it is, and the format extension, e.g. "1234_ah_01.jpg". NLNZ considered dropping the format extension as it may have a limited life. In the end, the extension was retained to assist Web browsers to display the file and users to save files locally.

NLNZ developed the "NLNZ Role Definitions" to be able to refer to the classes of the derivatives it generated [see figure 4]. The roles fall into three categories: original, preservation, and access.



Digital Original [DO]	Digitally born – original online format or copied from physical carrier	do_01
Digital Derivative [DD]	First digital manifestation of analogue item	dd_01
Preservation Master [PM]	Best attempt to replicate Original in a currently accessible format.	pm_01
Previous Preservation Objects [PP]	Previous best attempt to replicate Original in a currently accessible format – but may require special hardware or software to access.	pp_01 [expands to pp_02, pp_03, etc]
Access Source [AS]	Latest or best format for generating web dissemination formats	as_01
Access Formats [AF]	Derivatives of the AS – in NLNZ's preferred format for online delivery.	af_01 [expands to af_02, af_03, etc]
Access High-Res [AH]	High resolution derivatives of the AS format for broad bandwidth online delivery	ah_01
Access Low-Res [AL]	Low resolution derivatives of the AS format for narrow bandwidth online delivery	al_01
Thumbnail [TN]	Thumbnail image used as link to main dissemination copies	tn_01
Preview [PV]	Preview Copy used in COR / Discover for displaying alongside the descriptive record	pv_01

Figure 4 – NLNZ Role Definitions for derivatives and surrogates

6.2 Virtualisation

The above solution was not ideal in practice as requirements for a persistent location clashed with the need over time to re-locate or re-arrange storage space. To overcome this we are adding a layer of virtualisation.

The virtualisation layer allows file management – partitioning and location – to be managed dynamically internally, while files are presented externally as if in a persistent location. A proxy-like program is being developed which is optimised to resolve file requests by looking up the file's current location and delivering its content. The approach to physical location and file naming remains as discussed above, but the virtualisation layer presents this externally in a simplified way. This negates the need for a partition number in the location path and moves the role codes into optional parameters, thereby making it more extensible, e.g. "/1234" for the default version, or "/1234?role=TN&size=150" for the 150 pixel thumbnail. These parameters are still being worked on.

This layer offers additional opportunities such as transparent "on-the-fly" conversions and adjusting the MIME type reported, which is useful, for instance, for harvested web pages with file extensions that don't match the reported MIME type.

6.3 File formats and obsolescence

Even lay people are likely to have encountered obsolescence of file formats: some files only a few years old are unreadable already. In addition, media obsolescence can make it difficult to obtain the files in the first place, e.g. 5_ inch floppy disks received require a disk drive now considered "non-standard" to process them and many PCs no longer even have a 3 inch floppy disk drive!

There are three major approaches to tackling obsolescence (11).

- The **Museum** approach is to collect the hardware and software needed to ensure ongoing accessibility to the original file formats. The issues include expertise and offsite accessibility.
- With the **Migration** option, files are converted into current formats. This conversion process becomes ongoing as technologies continue to change. The issues include lossy conversions, deciding what must be preserved (intellectual content,

presentation, the user experience, etc), and balancing the access-driven desire for current formats vs. the preservation requirement to change formats as seldom as possible.
- **Emulation** requires software to be developed that can simulate the original experience using the original file format but with current technologies. The issues include development expertise and lossy emulations.

Of these, NLNZ currently prefers migration because it requires less development resources. However we recognise that migration has its limitations and may not work well with some materials, so emulation is also being considered. Our Preservation Metadata Schema is seen as integral to capturing records of all migration activities.

6.4 Preferred file formats

NLNZ has considered specifying file formats for both archival and dissemination purposes. The practical advantages of limiting the range of file formats that have to be managed over time within the digital archive are obvious. But the tensions between the best format for faithfully preserving the original and that for providing easy ongoing access need to be considered. In some instances it will be possible to specify preferred formats for deposit, but in many instances it will not. Files will be offered to NLNZ in all kinds of formats, often already redundant or impossible to read. These files will need to be transformed or migrated to current formats. Proprietary formats may have to be archived as preservation masters, when other options could alter the content or intended form of the object, but dissemination formats should be non-proprietary to ensure easy and enduring access.

NLNZ's Survey Objects work tested the viability of representing various text documents for web delivery as XML/HTML and PDF. The accurate conversion of many of the Microsoft applications to XML is currently a resource-heavy task, and NLNZ trusts tools for these processes will continue to develop. Other dissemination file formats NLNZ are currently working with include JPEG for images, WAVE and MP3 for sound, MOV and MPEG for video.

7 Metadata Framework

A range of metadata is essential to the successful management of digital resources. NLNZ published the first phase of its *Metadata Standards Framework* (12) in October 2000, focusing on resource discovery. The framework essentially states that the most appropriate international metadata standard will be used in any given circumstances, but it must be able to be mapped to Dublin Core [ISO 15836] at the common resource discovery layer [see figure 5]. The second phase covering a schema and data model for preservation metadata was published in November 2002.

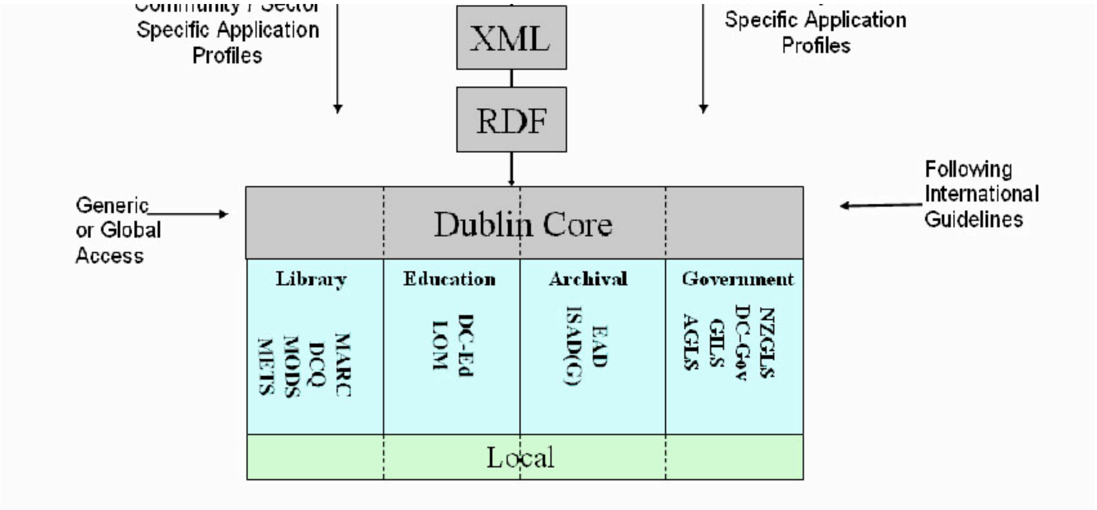


Figure 5 – NLNZ's Metadata Framework with community-specific metadata for vertical access and Dublin Core for horizontal access

There is a plethora of metadata schemas available today, each serving different purposes, so determining the requirements was NLNZ's first step in selecting appropriate schemas. The framework established by NLNZ follows the taxonomy in Kenney and Rieger's *Moving Theory into Practice* (13). This taxonomy describes four key metadata categories for digital objects:

- **Resource discovery** – How do we ensure that the materials we have collected can be found and retrieved by our clients? Dublin Core is an early attempt to provide a 'lingua franca' for resource discovery in an online environment and is still evolving.
- **Structural** – How do we present our objects in context (e.g. as ordered pages of a digitised book) and not just as a bunch of files and how do we navigate within this context (e.g. page 1 to page 2)?
- **Rights management and Access control** – How do we ensure protection of intellectual property rights, authentication of clients and authorisation of clients?
- **Technical and Administrative** – What are the essential attributes of digital objects and the processes and technologies that created them, and which are required for long-term storage, management, preservation and access?

7.1 Building modular metadata

NLNZ's *Metadata Standards Framework* identified Dublin Core as the essential core data for resource discovery. Qualified Dublin Core is used for 'simple digital objects' with the addition of local administrative elements, built in a modular way within RDF/XML. This *Digital Resource Description* (DRD) schema (14) is discussed below.

Descriptive metadata for NLNZ's archival digital collections will be EAD. For complex digital objects, both DC and EAD will be used within the METS framework, with METS providing the structural and behavioral components of the metadata.

This modular approach to building metadata ensures rigid compliance with international standard schema, which is an important factor in achieving interoperability with different constituent communities nationally and internationally.

7.2 Descriptive Metadata

NLNZ sources metadata from a number of systems, but primarily from its MARC-based ILS, and ISAD(G)-based archival system, TAPUHI.

Each of these is optimised for published and unpublished collection management tasks respectively. This means that descriptive metadata is duplicated in the mainstream and digital environments.

Additional metadata is collected as the output of several digital library processes and combined as part of scripted data conversions processes. A number of scripts convert these ISAD(G) and MARC records to XML for loading to ENCompass using either NLNZ's qualified Dublin Core [DC] or Encoded Archival Description [EAD] metadata schemes.

7.3 Metadata Conversion Engine

The expense in creating metadata manually means institutions must leverage their existing metadata when generating metadata in new formats. A lot of new metadata can be generated by "crosswalking" (converting) from existing sources. It can often be mapped successfully but is sometimes fundamentally different, resulting in a "square peg in a round hole". These mis-matches may have to be accepted within budgetary constraints, meaning crosswalking is the only option. The effectiveness of the conversion has to be evaluated against the cost of additional manual work.

NLNZ realised the need to deliver metadata in different formats to various audiences. This metadata also had to be derived from a number of sources (various formats of human-generated descriptive metadata and auto-generated technical metadata) [see figure 6].

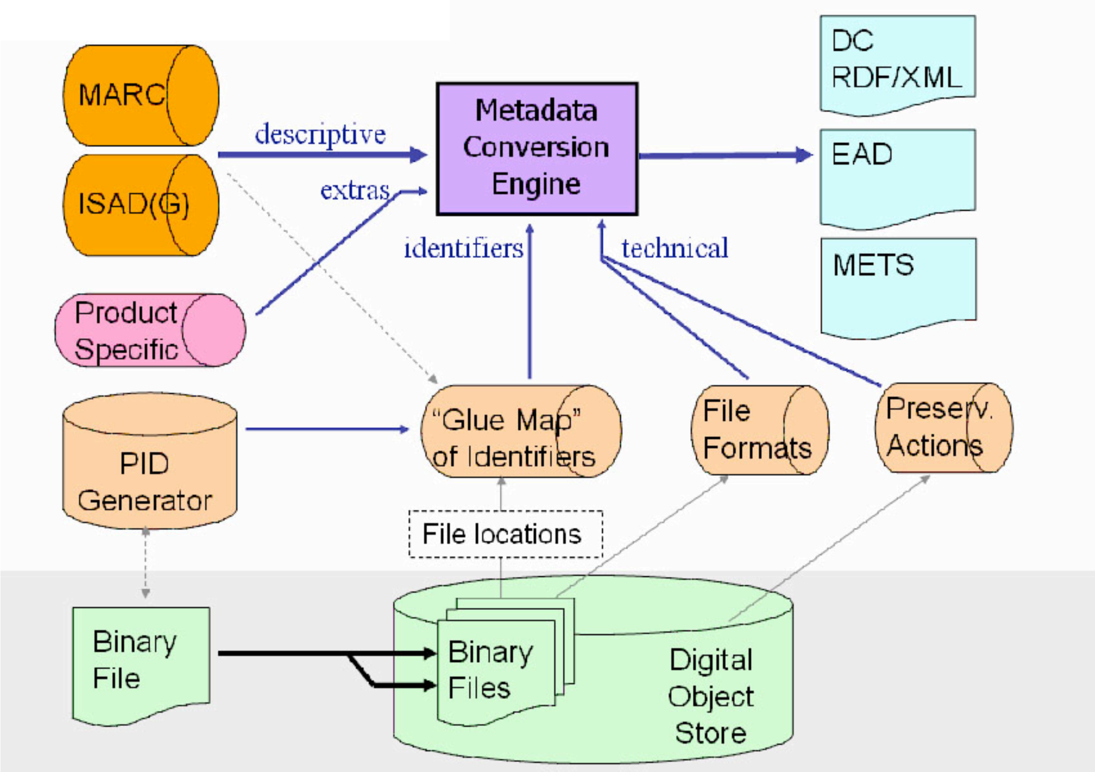


Figure 6 – Multiple sources of metadata feed the conversion

Experience showed that the most efficient way to achieve this was to genericise the conversion processes [see figure 7]. When we attempted to deliver DC metadata for *Discover* using a process developed previously for the

National Library of Australia's *Picture Australia* product–we realised that we had DC data hard-coded to suit *Picture Australia's* interface did not suit *Discover's* interface. Our process now converts the metadata into a base of pure DC, following DCMI's guidelines strictly, and product-specific requirements are then added on as necessary.

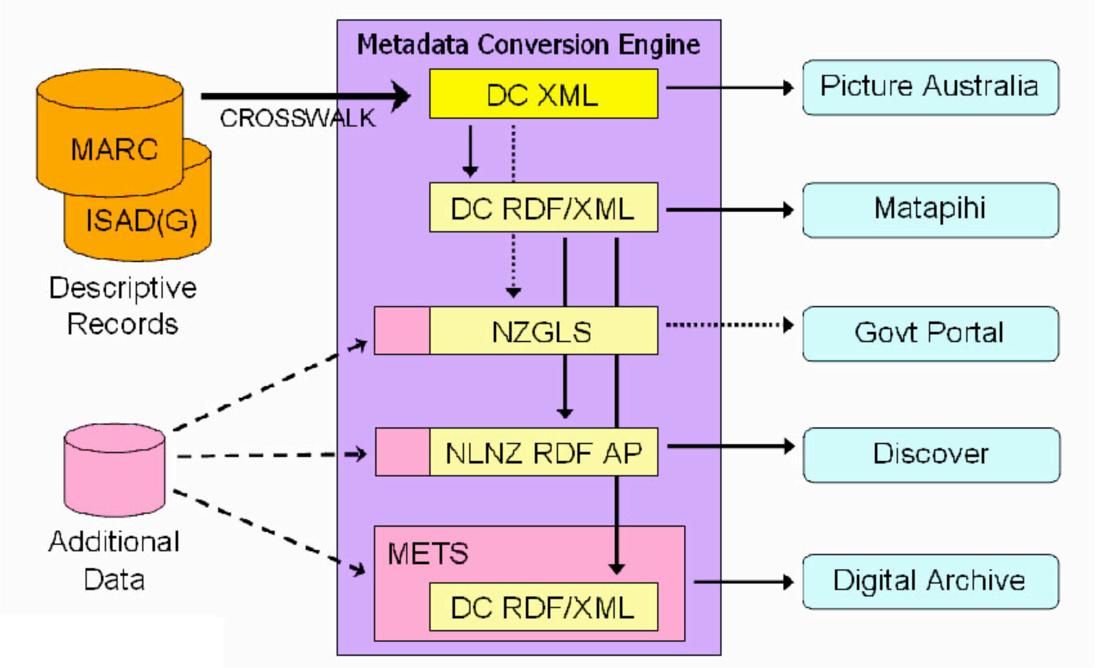


Figure 7 – Modular crosswalking as the most efficient model

Currently NLNZ's Metadata Conversion Engine (MCE) consists of several independent Perl, Java, and XSLT scripts. There is potential for the development of a "Metadata Crosswalk Definition Language" that could capture crosswalk algorithms generically for application within particular MCE implementations or between MCEs.

7.4 DRD Application Profile

NLNZ has developed "Digital Resource Description" (14), an Application Profile based on Dublin Core, for records in *Discover*. One intention for DRD is to provide a "lightweight" alternative to METS for simple digital objects. The rationale is that DC Qualifiers provide sufficient granularity for both finding and using single image files in a Web interface, with two exceptions:

- identifying the multiple derivative files – there is no way of differentiating URLs in multiple Identifier or hasFormat properties
- identifying the type of identifiers in the Identifier property, e.g. local, persistent, or location.

The extensions to DC Qualifiers that NLNZ uses are: Local Identifier, Persistent Identifier, Digital Object Location, XLink 'simple' type attributes (type, href, title, role, arcrole, show, actuate) for more detailed descriptions of the multiple linking elements, and Metadata Rights Ownership (used for tracking the source of metadata not created by NLNZ).

Part of the discipline of creating an

Application Profile is ensuring imported elements comply with the external schema they were sourced from. Each element was researched and appropriate guidelines for use prepared.

The RDF/XML syntax was chosen as it is the Dublin Core Metadata Initiative's preferred syntax and it is part of W3C's vision for the future of the Web.

An opportunity identified during the development of DRD was for a mechanism to make managing Application Profiles easier. NLNZ has begun work on a 'Metadata Schema Description Language' for centralising Application Profile maintenance in XML "super-documents" from which can be derived (using XSLT) any required DTDs, XML Schemas, RDF schemas, HTML documentation, or RDDL directory pages.

7.5 NLNZ Preservation Metadata Schema

NLNZ's preservation metadata schema was developed with practical implementation in mind, and was published for comment in 2002 (12). This metadata supports curators and digital library administrators as they monitor and migrate the different file formats stored in the repository, to ensure they are authentic and accessible over time. Also referred to as digital asset management, this is one part of the work of a 'Trusted Digital Repository' as defined in *Trusted Digital Repositories: Attributes and Responsibilities* (4). NLNZ have not identified a tool for this part of the process.

NLNZ's Preservation Metadata was developed in the light of international research, particularly that of the National Library of Australia, and addresses two functional objectives:

- Providing sufficient knowledge to take appropriate action in order to maintain a digital object's bitstream over the long-term
- Ensuring that the content of an archived object can be rendered and interpreted, in spite of future changes in storage and access technologies.

Preservation metadata will be largely extracted from the digital files in an automated process. However it will also in part be collected manually over a lengthy period, as a digital collection is donated or selected, and objects are appraised, arranged, described, validated and re-formatted. Curators, digital archivists and technical staff will all record components of the preservation metadata, and each will all have different ongoing requirements of the preservation records. Format transformations, digital recovery or conservation processes will require curator approval.

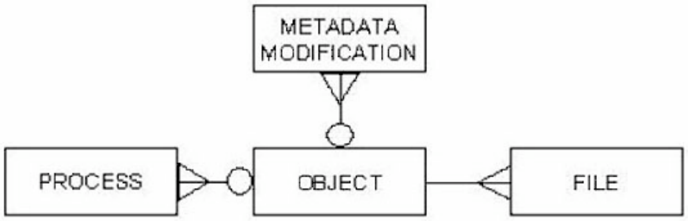


Figure 8 – Preservation Metadata Model

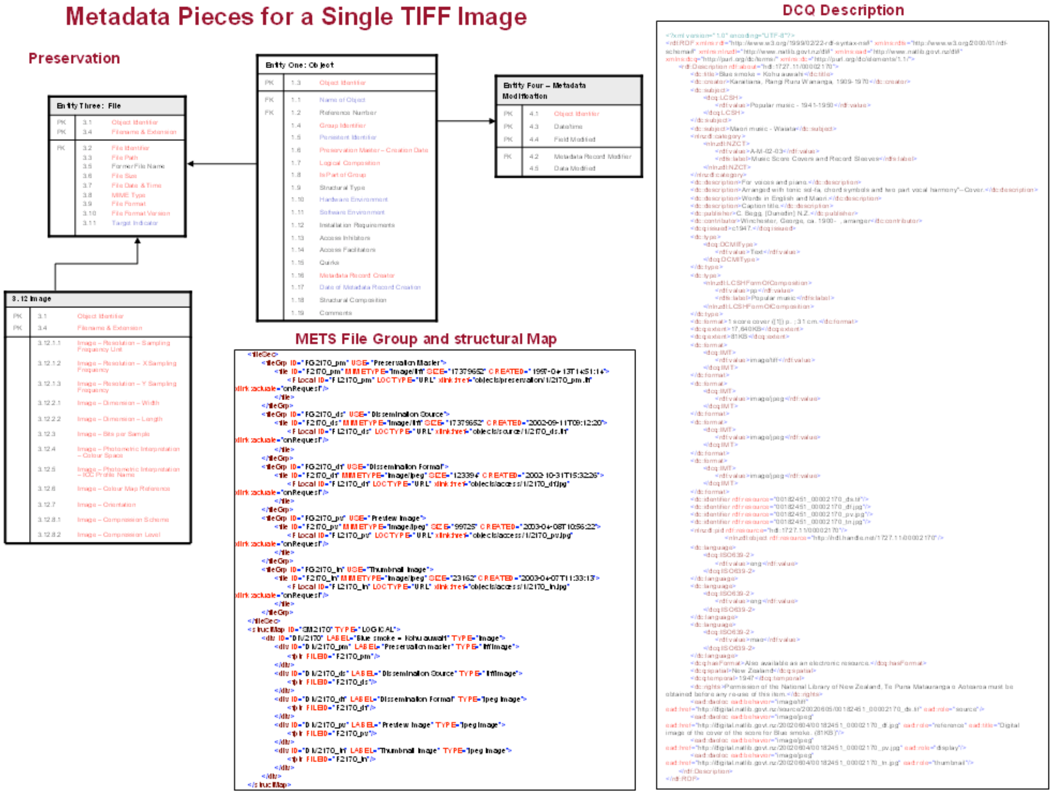
The four entities of the NLNZ schema, as shown in figure 8, represent a minimum set of metadata necessary for preservation:

- **Object** – preservation specific values associated with the whole of the object, e.g. reference number, hardware requirements, software requirements

- **Process** – actions that have happened, and changes made, to an object in NLNZ's care, such as the conversion/migration to a more viable version
- **File** – preservation data associated with the individual files that comprise a digital object, e.g. size, date created, filename
- **Metadata Modification** – all changes made to the metadata record, along with date, time and who made those changes

7.6 Metadata Extraction Tool

NLNZ is concerned about the sheer quantity of metadata needed for the management and preservation for every digital asset [see figure 9]. Automated systems are essential.



Metadata Pieces for a Single TIFF Image
Preservation
METS File Group and structural Map
DCQ Description

Entity One: Object		
PK	1.3	Object Identifier
FK	1.1	Name of Object
FK	1.2	Reference Number
	1.4	Group Identifier
	1.5	Persistent Identifier
	1.6	Preservation Master – Creation Date
	1.7	Logical Composition
	1.8	Is Part of Group
	1.9	Structural Type
	1.10	Hardware Environment
	1.11	Software Environment
	1.12	Installation Requirements
	1.13	Access Inhibitors
	1.14	Access Facilitators
	1.15	Quirks
	1.16	Metadata Record Creator
	1.17	Date of Metadata Record Creation
	1.18	Structural Composition
	1.19	Comments

Entity Three: File		
PK	3.1	Object Identifier
PK	3.4	Filename & Extension
FK	3.2	File Identifier
	3.3	File Path
	3.5	Former File Name
	3.6	File Size
	3.7	File Date & Time
	3.8	MIME Type
	3.9	File Format
	3.10	File Format Version
	3.11	Target Indicator

Entity Four – Metadata Modification		
PK	4.1	Object Identifier
PK	4.3	Date/time
PK	4.4	Field Modified
FK	4.2	Metadata Record Modifier
	4.5	Data Modified

3.12 Image		
PK	3.1	Object Identifier
PK	3.4	Filename & Extension
	3.12.1.1	Image – Resolution – Sampling Frequency Unit
	3.12.1.2	Image – Resolution – X Sampling Frequency
	3.12.1.3	Image – Resolution – Y Sampling Frequency
	3.12.2.1	Image – Dimension – Width
	3.12.2.2	Image – Dimension – Length
	3.12.3	Image – Bits per Sample
	3.12.4	Image – Photometric Interpretation – Colour Space
	3.12.5	Image – Photometric Interpretation – ICC Profile Name
	3.12.6	Image – Colour Map Reference
	3.12.7	Image – Orientation
	3.12.8.1	Image – Compression Scheme
	3.12.8.2	Image – Compression Level

```
  <fileSec>
    <fileGrp ID="FG2170_pm" USE="Preservation Master">
      <file ID="F2170_pm" MIMETYPE="image/tiff" SIZE="17379652" CREATED="1997-04-13T14:51:14">
        <FLocat ID="FL2170_pm" LOCTYPE="URL" xlink:href="objects/preservation/1/2170_pm.tif"
xlink:actuate="onRequest"/>
      </file>
    </fileGrp>
    <fileGrp ID="FG2170_ds" USE="Dissemination Source">
      <file ID="F2170_ds" MIMETYPE="image/tiff" SIZE="17379652" CREATED="2002-09-11T09:12:20">
        <FLocat ID="FL2170_ds" LOCTYPE="URL" xlink:href="objects/source/1/2170_ds.tif"
xlink:actuate="onRequest"/>
      </file>
    </fileGrp>
    <fileGrp ID="FG2170_df" USE="Dissemination Format">
      <file ID="F2170_df" MIMETYPE="image/jpeg" SIZE="123394" CREATED="2002-10-31T15:32:26">
        <FLocat ID="FL2170_df" LOCTYPE="URL" xlink:href="objects/access/1/2170_df.jpg"
xlink:actuate="onRequest"/>
      </file>
    </fileGrp>
    <fileGrp ID="FG2170_pv" USE="Preview Image">
      <file ID="F2170_pv" MIMETYPE="image/jpeg" SIZE="99725" CREATED="2003-04-08T10:56:22">
        <FLocat ID="FL2170_pv" LOCTYPE="URL" xlink:href="objects/access/1/2170_pv.jpg"
xlink:actuate="onRequest"/>
      </file>
    </fileGrp>
    <fileGrp ID="FG2170_tn" USE="Thumbnail Image">
      <file ID="F2170_tn" MIMETYPE="image/jpeg" SIZE="23162" CREATED="2003-04-07T11:33:13">
        <FLocat ID="FL2170_tn" LOCTYPE="URL" xlink:href="objects/access/1/2170_tn.jpg"
xlink:actuate="onRequest"/>
      </file>
    </fileGrp>
  </fileSec>
  <structMap ID="SM2170" TYPE="LOGICAL">
    <div ID="DIV2170" LABEL="Blue smoke = Kohu auwahi" TYPE="Image">
      <div ID="DIV2170_pm" LABEL="Preservation master" TYPE="tiff image">
        <fptr FILEID="F2170_pm"/>
      </div>
      <div ID="DIV2170_ds" LABEL="Dissemination Source" TYPE="tiff image">
        <fptr FILEID="F2170_ds"/>
      </div>
      <div ID="DIV2170_df" LABEL="Dissemination Format" TYPE="jpeg image">
        <fptr FILEID="F2170_df"/>
      </div>
      <div ID="DIV2170_pv" LABEL="Preview Image" TYPE="jpeg image">
        <fptr FILEID="F2170_pv"/>
      </div>
      <div ID="DIV2170_tn" LABEL="Thumbnail Image" TYPE="jpeg image">
        <fptr FILEID="F2170_tn"/>
      </div>
    </div>
  </structMap>
```

```
<?xml version="1.0" encoding="UTF-8"?>
<rdf:RDF xmlns:rdf="http://www.w3.org/1999/02/22-rdf-syntax-ns#" xmlns:rdfs="http://www.w3.org/2000/01/rdf-
schema#" xmlns:nlnzdl="http://www.natlib.govt.nz/dl#" xmlns:cad="http://www.natlib.govt.nz/dl#"
xmlns:dcq="http://purl.org/dc/terms/" xmlns:dc="http://purl.org/dc/elements/1.1/">
    <rdf:Description rdf:about="hdl:1727.11/00002170">
        <dc:title>Blue smoke = Kohu auwahi</dc:title>
        <dc:creator>Karaitiana, Rangi Ruru Wananga, 1909-1970</dc:creator>
        <dc:subject>
            <dcq:LCSH>
                <rdf:value>Popular music - 1941-1950</rdf:value>
            </dcq:LCSH>
        </dc:subject>
        <dc:subject>Maori music - Waiata</dc:subject>
        <nlnzdl:category>
            <nlnzdl:NZCT>
                <rdf:value>A-M-02-03</rdf:value>
                <rdfs:label>Music Score Covers and Record Sleeves</rdfs:label>
            </nlnzdl:NZCT>
        </nlnzdl:category>
        <dc:description>For voices and piano.</dc:description>
        <dc:description>Arranged with tonic sol-fa, chord symbols and two part vocal harmony"--Cover.</dc:description>
        <dc:description>Words in English and Maori.</dc:description>
        <dc:description>Caption title.</dc:description>
        <dc:publisher>C. Begg, [Dunedin] N.Z.</dc:publisher>
        <dc:contributor>Winchester, George, ca. 1900-  , arranger</dc:contributor>
        <dcq:issued>c1947.</dcq:issued>
        <dc:type>
            <dcq:DCMIType>
                <rdf:value>Text</rdf:value>
            </dcq:DCMIType>
        </dc:type>
        <dc:type>
            <nlnzdl:LCSHFormOfComposition>
                <rdf:value>pp</rdf:value>
                <rdfs:label>Popular music</rdfs:label>
            </nlnzdl:LCSHFormOfComposition>
        </dc:type>
        <dc:format>1 score cover ([1]) p. ; 31 cm.</dc:format>
        <dcq:extent>17,640KB</dcq:extent>
        <dcq:extent>81KB</dcq:extent>
        <dc:format>
            <dcq:IMT>
                <rdf:value>image/tiff</rdf:value>
            </dcq:IMT>
        </dc:format>
        <dc:format>
            <dcq:IMT>
                <rdf:value>image/jpeg</rdf:value>
            </dcq:IMT>
        </dc:format>
        <dc:format>
            <dcq:IMT>
                <rdf:value>image/jpeg</rdf:value>
            </dcq:IMT>
        </dc:format>
        <dc:format>
            <dcq:IMT>
                <rdf:value>image/jpeg</rdf:value>
            </dcq:IMT>
        </dc:format>
        <dc:identifier rdf:resource="00182451_00002170_ds.tif"/>
        <dc:identifier rdf:resource="00182451_00002170_df.jpg"/>
        <dc:identifier rdf:resource="00182451_00002170_pv.jpg"/>
        <dc:identifier rdf:resource="00182451_00002170_tn.jpg"/>
        <nlnzdl:pid rdf:resource="hdl:1727.11/00002170"/>
                <nlnzdl:object rdf:resource="http://hdl.handle.net/1727.11/00002170"/>
        <dc:language>
            <dcq:ISO639-2>
                <rdf:value>eng</rdf:value>
            </dcq:ISO639-2>
        </dc:language>
        <dc:language>
            <dcq:ISO639-2>
                <rdf:value>eng</rdf:value>
            </dcq:ISO639-2>
        </dc:language>
        <dc:language>
            <dcq:ISO639-2>
                <rdf:value>mao</rdf:value>
            </dcq:ISO639-2>
        </dc:language>
        <dcq:hasFormat>Also available as an electronic resource.</dcq:hasFormat>
        <dcq:spatial>New Zealand</dcq:spatial>
        <dcq:temporal>1947</dcq:temporal>
        <dc:rights>Permission of the National Library of New Zealand, Te Puna Matauranga o Aotearoa must be
obtained before any re-use of this item.</dc:rights>
        <cad:dacloc cad:behavior="image/tiff"
cad:href="http://digital.natlib.govt.nz/source/20020605/00182451_00002170_ds.tif" cad:role="source"/>
        <cad:dacloc cad:behavior="image/jpeg"
cad:href="http://digital.natlib.govt.nz/20020604/00182451_00002170_df.jpg" cad:role="reference" cad:title="Digital
image of the cover of the score for Blue smoke. (81KB)"/>
        <cad:dacloc cad:behavior="image/jpeg"
cad:href="http://digital.natlib.govt.nz/20020604/00182451_00002170_pv.jpg" cad:role="display"/>
        <cad:dacloc cad:behavior="image/jpeg"
cad:href="http://digital.natlib.govt.nz/20020604/00182451_00002170_tn.jpg" cad:role="thumbnail"/>
    </rdf:Description>
</rdf:RDF>
```

Figure 9 – The metadata pieces for a single TIFF image (the elements in red are derived programmatically)

A major advantage of digital objects is that computer programs can run processes over them with results that could never be achieved by humans. Collecting detailed technical metadata automatically is not uncommon, but NLNZ could find no tool that would open all of the common file formats and extract the metadata embedded inside.

The Library commissioned development of the Java-based *Preservation Metadata Extract Tool.* This can be run with a GUI or in batch, currently handles 5 common file formats with 10 additional formats in development, and outputs the results in any XML format – e.g. processing 10,000 JPEG files per hour. The output is another metadata source for the MCE. The tool was selected as a finalist in the *Pilgrim Trust's 2004 Preservation*

Awards (15).

7.7 Structural Metadata

Having identified all the metadata components needed to support selecting, ingesting, identifying, describing, managing and preserving digital objects, the question was, where to put it? The simple group of XLinks used in NLNZ's DRD schema cannot cater for the complex administrative and technical metadata that multi-layered component and derivative files require.

When the Metadata Encoding & Transmission Standard [METS] came out in an early release it all suddenly seemed possible. NLNZ anticipate the descriptive and administrative metadata would be stored externally from the METS records, though the ability to embed it within METS is attractive from the metadata interchange point of view.

It is METS' 'Structural Map' functionality that NLNZ are primarily attracted to: this provides the first strong mechanism available to track all the digital files across all purposes (administration, preservation, and delivery).

8 Integration into the business

New technologies are often developed within projects that sit outside "business as usual", but eventually the results must be integrated back into the business.

NLNZ has been developing its Digital Library strategies for a number of years and has run a number of pilots. Understanding of the requirements is deepening and the drive now is to move the collective processes from theory to implementation. Some staff who will be with digital objects every day have only been involved in projects on an ad hoc basis – there is a sizable task to up-skill them and determine the best fit of the processes into their work.

8.1 Development of business process workflows

NLNZ has developed a series of business process diagrams that reflect the current workflows. At a very detailed level, these outline tasks and decisions, the business unit and staff member/s responsible, the tools used (which could be particular pieces of hardware and software, but may also be offline tools such as collection policies or printed forms or manuals) and the pieces of metadata that are input/output. Each workflow diagram combines sections that are particular to one type of digital object (the main categories being published online, published offline, unpublished and digitised) with generic workflow processes (e.g. around upload of objects to the digital archive).

These diagrams are a work in progress and are revised constantly. The proposed outcome is a comprehensive end-to-end manual for digital object processes and procedures to support NLNZ's efforts towards trusted digital repository status.

9 Conclusion

NLNZ's understanding of the processes for managing, storing, delivering and preserving digital objects, and of generating and managing the metadata that supports these processes, has progressed a long way in the last five years. The once mammoth-looking list of challenges is being systematically reduced as successive components are resolved. NLNZ has built on its early experiences: pilot projects have brought us face-to-face with the issues and we offer up the responses here to add to the pool of solutions. NLNZ is still refining its digital strategies but feels it has a firm basis on which to build its digital archive.

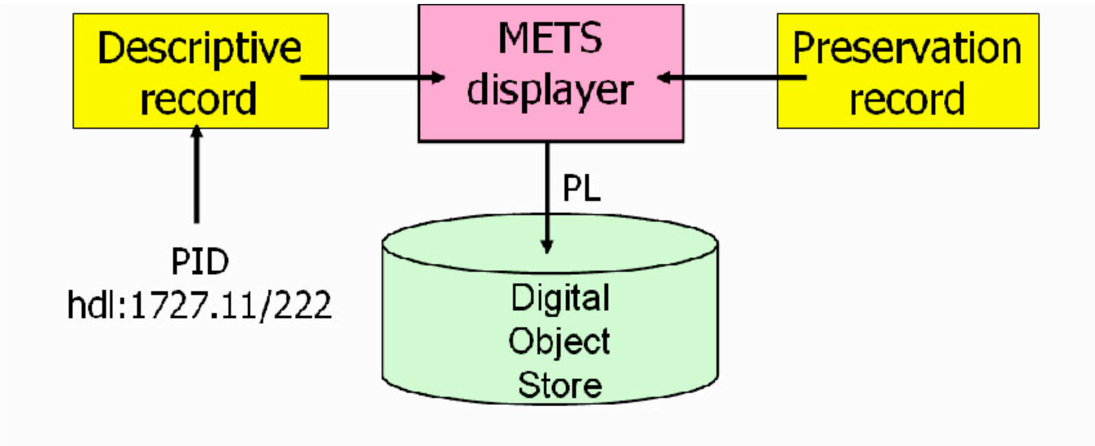


Figure 10 – Simplified view of NLNZ's current solution

Acknowledgement
The authors would like to thank their colleagues in the Library's Digital Initiatives and Digital Strategy Implementation teams for their encouragement, assistance and contributions.

References

1. National Library of New Zealand. Discover – Te Kohinga Taonga. Available at: http://discover.natlib.govt.nz
2. Consultative Committee for Space Data Systems. Reference Model for an Open Archival Information System (OAIS). CCSDS 650.0-B-1, January 2002. Retrieved 1 May 2004 from http://ssdoo.gsfc.nasa.gov/nost/isoas/
3. Lavoie, B. F. The Open Archive Information System Reference Model: Introductory Guide. DPC Technology Watch Series Report 04-01, January 2004. Retrieved 1 May 2004 from http://www.dpconline.org/docs/lavoie_OAIS.pdf
4. RLG/OCLC Working Group on Digital Archive Attributes. Trusted Digital Repositories: Attributes and Responsibilities. 2002.
5. Green, B. & Bide, M. Unique Identifiers: a brief introduction, revised edition. March 1997. Retrieved 1 May 2004 from http://www.bic.org.uk/uniquid.html
6. Paskin, N. On Making and Identifying a "Copy". D-Lib Magazine, Volume 9 Number 1, January 2003. Retrieved 1 May 2004 from http://www.dlib.org/dlib/january03/paskin/01paskin.html
7. W3C Technical Architecture Group. Architecture of the World Wide Web, First Edition, Section 2. 9 December 2003. Retrieved 1 May 2004 from http://www.w3.org/TR/webarch/
8. IFLA Study Group on the Functional Requirements for Bibliographic Records. Functional requirements for bibliographic records : final report. 1998. Retrieved 1 May 2004 from http://www.ifla.org/VII/s13/wgfrbr/wgfrbr.htm
9. Rust, G. & Bide, M. INDECS Metadata Framework: Principles, model and data dictionary, v2.0. June 2000. Retrieved 1 May 2004 from http://www.indecs.org/project.htm#finalDocs
10. International DOI Foundation. The DOI Handbook, Edition 3.3. November 2003. Retrieved 1 March 2004 from http://www.doi.org/hb.html
11. Thibodeau, K. Overview of Technological Approaches to Digital Preservation and Challenges in Coming Years. In The State of Digital Preservation: An International Perspective conference proceedings, 2002. Retrieved 1 March 2004 from http://www.clir.org/pubs/reports/pub107/thibodeau.html

12. National Library of New Zealand. Metadata Standards Framework. October 2000, November 2002 & June 2003. Retrieved 1 May 2004 from http://www.natlib.govt.nz/en/whatsnew/4initiatives.html#meta
13. Kenney, A. & Rieger, O. Moving Theory into Practice: Digital Imaging for Libraries and Archives, Chapter 5. Research Libraries Group, 2000.
14. National Library of New Zealand. Digital Resource Description. Retrieved 1 May 2004 from http://www.natlib.govt.nz/dr/drd.html
15. Digital Preservation Coalition. Pilgrim Trust's Digital Preservation Award 2004. Retrieved 1 May 2004 from http://www.consawards.ukic.org.uk/digital.html