Preservation MD at National Library of New Zealand

Peter McKinney
National Digital Heritage Archive
National Library of New Zealand Te Puna Mātauranga o Aotearoa
Going to cover...

Preservation metadata at National Library of New Zealand
Te Puna Mātauranga Aotearoa, with particular reference
to PREMIS.

- (Brief) Background to NLNZ preservation programme
- How we use some PREMIS units
- PREMIS units we do not currently use
- Elements we use that are not in PREMIS
Background to NLNZ

1 preservation system
contains

700,000 intellectual entities
made up of

3.4m master files
represented in

34 Terabytes
taking up

implementing

85 different formats

ExLibris Rosetta
Background to NLNZ

Collections Management Layer
- Published Catalogue
- Unpublished Catalogue

Data Management Layer
- Databases

Permanent Repository
- Object Files
- MD Files (METS)

Synched

References
Preservation metadata is:

“The information a repository uses to support the digital preservation process” – p. 3 PREMIS Data Dictionary
PREMIS Compliance 1

What does it mean to be compliant with PREMIS?

Must follow all requirements and constraints at **unit** level

1. If it shares a name, it must share the definition
   - You can have different names with shared definitions as long as they are mapped

2. Usage requirements must be observed (repeatability, obligation and applicability requirements can be made more stringent, but not more relaxed).
What does it mean to be compliant with PREMIS?

Must follow all requirements and constraints at dictionary level

1. Include the mandatory semantic units for any Data Model Entity (Objects, Events, Agents) supported by the repository.
2. Be able to recover all of the information specified in the mandatory PREMIS semantic units from the repository system and associate it with its corresponding Entity.
differences/deviations/(mis)interpretations

- Filestream and bitstream
- Significant properties (difference in meaning, and not yet using behaviours)
- Events not fired if re-running fixity (virus)
- Fixity mandatory
- File format
Files, filestreams and bitstreams

Does it matter if we map filestreams to bitstream level?
Files, filestreams and bitstreams

**File**: a named and ordered sequence of bytes that is known by an operating system

**Filestream**: a set of bits embedded within a file that can be transformed into a standalone file without adding any additional information

**Bitstream**: a set of bits embedded within a file that cannot be transformed into a standalone file without the addition of file structure (e.g., headers) or other reformatting to comply with some particular file format specification
Files, filestreams and bitstreams

Green = file
Blue = filestream
Red = bitstream

Intellectual Entity

Representation 1

CDX LOG XML

ARC

HTML TIFF Multipage
GIF TIFF1 TIFF2
MP3
Proper PREMIS mapping

- **Intellectual Entity**
  - contains
    - CDX
    - LOG
    - XML
    - ARC
    - contains
      - HTML
      - GIF
      - TIFF
      - contains
        - image 1
        - image 2

- **IE level**
  - Descriptive MD

- **File Level**
  - Format
  - Significant Properties
  - Mimetype
  - Virus outcome
  - Structure map
  - Events

- **Bitstream Level**
  - Format
  - Significant Properties
  - Mimetype
  - Virus outcome
  - Structure map
  - Events
Proposed NLNZ mapping

- **Intellectual Entity**
  - contains
    - CDX
    - LOG
    - XML
    - ARC

- **IE Level**
  - Descriptive MD

- **File Level**
  - Format
  - Significant Properties
  - Mimetype
  - Virus outcome
  - Structure map
  - Events

- **Bitstream Level**
  - Format
  - Significant properties
  - Mimetype
  - Virus outcome
  - Structure map
  - Events

- contains
  - HTML
  - GIF
  - TIFF

- have
  - image 1
  - image 2
Files, filestreams and bitstreams

- Green = file
- Red = bitstream

Diagram showing the relationship between files and bitstreams, with file streams and bit streams being represented in different layers, indicating the mapping between file level properties and bit stream level.
Significant properties

Subtle (?) difference in interpretation
Deviations – significant properties

PREMIS definitions:
“Characteristics of a particular object subjectively determined to be important to maintain through preservation actions.” P.39

“Listing significant properties implies that the repository plans to preserve these properties across time.” p.39

NLNZ definition:
Technical properties that we may or may not want to keep across time.
The problem with properties

Unknown photographer, Apirana Turupa Ngata leading a haka at the 1940 centennial celebrations, Waitangi, 1940, B&W original negative, Alexander Turnbull Library, Reference: MNZ-2746-1/2-F
For example:

- Risk analysis
- Preservation evaluation (how do we know what changes we’ve made?)
- Problem analysis (search)

* *we might want to deliberately track their ‘demise’ across preservation actions **
Events

- Example of preservation events
- Events on re-run
“The event entity aggregates information about an action that involves one or more object entities. Whether or not a preservation repository records an event depends on the importance of the event.”
Events in Rosetta

Events

154

Number

18 Provenance

136 Audit

Covering

Split into

Stored in

Database

METS file (permanent repository)

Areas of system

Staging

Deposit

Web editor

Back office

Preservation

Delivery

Process automation

Set management

Format library

General
Events in Rosetta

Metadata List

Intellectual Entity PID: IE6408846
Updated on: 27/09/2011 09:38:59
Version: 2

Created on: 01/09/2011 10:39:33
Updated by: nga

Object Summary | Metadata | Services

Name | Type
DNX | DNX
Mets Section | Structure Map
Events in Rosetta

eventIdentifierType=DPS
eventIdentifierValue=355
eventType=Creation
eventDescription=Representation was added by preservation plan execution
eventDateTime=29/2011 09:28:29
eventOutcome1=SUCCESS
eventOutcomeDetail1=IE_ID=IE6408846;PLAN_NAME=AMY'S MP3 PLAN 3;
ALTERNATIVE_ID=77726268;PLAN_ID=77726199;REP_ID=REP6989977
linkingAgentIdentifierType1=Software
linkingAgentIdentifierValue1=MP3toWaveMigrationTool
Events in Rosetta

relationshipType=derivation
relationshipSubType=has source
relatedObjectIdentifierType=None
relatedObjectIdentifierValue=FL6408848
relatedObjectSequence=1
Fixity events in Rosetta
### Fixity events in Rosetta

<table>
<thead>
<tr>
<th>event Identifier Type</th>
<th>DPS</th>
</tr>
</thead>
<tbody>
<tr>
<td>event Identifier Value</td>
<td>27</td>
</tr>
<tr>
<td>event Type</td>
<td>Validation</td>
</tr>
<tr>
<td>event Description</td>
<td>Fixity check performed on file</td>
</tr>
<tr>
<td>event Date Time</td>
<td>26/09/2011 16:07:05</td>
</tr>
<tr>
<td>event Outcome1</td>
<td>SUCCESS</td>
</tr>
<tr>
<td>event Outcome Detail1</td>
<td>PROCESS_ID=81176898;PID=FL6989978;SIP_ID=76745;DEPOSIT_ACTIVITY_ID=85062;MF_ID=3339191900;TASK_ID=30;ALGORITHM_NAME=MD5;PRODUCER_ID=126755891;</td>
</tr>
</tbody>
</table>
Fixity events in Rosetta

If we ran fixity twice a year, *and* kept the events

<table>
<thead>
<tr>
<th>Current stats:</th>
<th>End of Year 2</th>
<th>End of Year 5</th>
</tr>
</thead>
<tbody>
<tr>
<td>1.2m METS files</td>
<td>6.4m METS files</td>
<td>14m METS files</td>
</tr>
<tr>
<td>72Gb</td>
<td>360Gb</td>
<td>792Gb</td>
</tr>
</tbody>
</table>
PREMIS units we don’t use (but are available to us)

- Preservation level
- Environment information
- Dependency units
- Composition level
Preservation Level

Information indicating the decision or policy on what can/should be done to the object in terms of degrees of care

- All objects currently at the same level = FULL PRESERVATION
- Have not (cannot?) taken account of the ‘preservability of the object’
- Will probably use as we head towards other institutions’ content coming into the system.
Environment Information

“Hardware/software combinations supporting use of the object” p.80.

1. Collecting this information problematic/painstaking
2. Repeatability – necessary to keep for every file? [as noted by PREMIS]
3. Use our ‘own’ format library

1.8 environment*
1.8.2 environmentPurpose
1.8.3 environmentNote
1.8.4 dependency*
1.8.5 software*
1.8.6 hardware*
1.8.7 environmentExtension*
Dependency unit

Use ‘when one object requires another to support its function, delivery, or coherence of content’ p.14

We really want to use them!
- Are moving towards trying to implement
- New scenarios in Papers Past – xml and dtds
- How do we get this automatically?
Composition Level

An indication of whether the object is subject to one or more processes of decoding or unbundling

- Haven’t found a compelling case to use it yet
  (we generally unzip before ingest)
- Complex, in terms of system, to get this information
"Identification of the format of a file or bitstream" p.53

"a specific, preestablished structure for the organization of a digital file or bitstream." p.204

The most crucial piece of information for preservation?
### Format Identification

**PREMIS example of FLAC file**

<table>
<thead>
<tr>
<th>objectCharacteristics</th>
<th>format</th>
<th>formatRegistry</th>
<th>formatRegistryName</th>
<th>formatRegistryKey</th>
<th>formatRegistryRole</th>
<th>formatDesignation</th>
<th>formatName</th>
<th>formatVersion</th>
<th>formatNote</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>audio/x-flac</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
## Format Identification

**NLNZ example of FLAC file**

<table>
<thead>
<tr>
<th>Type</th>
<th>Value</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>generalFileCharacteristics</strong></td>
<td></td>
</tr>
<tr>
<td>fileMIMEType</td>
<td>audio/x-flac</td>
</tr>
<tr>
<td><strong>generalFileCharacteristics</strong></td>
<td></td>
</tr>
<tr>
<td>formatLibraryID</td>
<td>ExL-Fmt-24417</td>
</tr>
<tr>
<td><strong>generalFileCharacteristics</strong></td>
<td></td>
</tr>
<tr>
<td>fileExtension</td>
<td>flac</td>
</tr>
<tr>
<td><strong>fileFormat</strong></td>
<td></td>
</tr>
<tr>
<td>Agent</td>
<td>DROID</td>
</tr>
<tr>
<td><strong>fileFormat</strong></td>
<td></td>
</tr>
<tr>
<td>formatRegistry</td>
<td>PRONOM</td>
</tr>
<tr>
<td><strong>fileFormat</strong></td>
<td></td>
</tr>
<tr>
<td>formatRegistryID</td>
<td>fmt/279</td>
</tr>
<tr>
<td><strong>fileFormat</strong></td>
<td></td>
</tr>
<tr>
<td>formatRegistryRole</td>
<td></td>
</tr>
<tr>
<td><strong>fileFormat</strong></td>
<td></td>
</tr>
<tr>
<td>formatName</td>
<td>ExL-Fmt-24417</td>
</tr>
<tr>
<td><strong>fileFormat</strong></td>
<td></td>
</tr>
<tr>
<td>formatVersion</td>
<td></td>
</tr>
<tr>
<td><strong>fileFormat</strong></td>
<td></td>
</tr>
<tr>
<td>formatDescription</td>
<td>Free Lossless Audio Codec</td>
</tr>
<tr>
<td><strong>fileFormat</strong></td>
<td></td>
</tr>
<tr>
<td>formatNote</td>
<td></td>
</tr>
<tr>
<td><strong>fileFormat</strong></td>
<td></td>
</tr>
<tr>
<td>exactFormatIdentification</td>
<td>TRUE</td>
</tr>
<tr>
<td><strong>fileFormat</strong></td>
<td></td>
</tr>
<tr>
<td>mimeType</td>
<td>audio/x-flac</td>
</tr>
<tr>
<td><strong>fileFormat</strong></td>
<td></td>
</tr>
<tr>
<td>agentVersion</td>
<td>5</td>
</tr>
<tr>
<td><strong>fileFormat</strong></td>
<td></td>
</tr>
<tr>
<td>agentSignatureVersion</td>
<td>50</td>
</tr>
</tbody>
</table>
## Format Identification

<table>
<thead>
<tr>
<th>NINZ</th>
<th>NINZ Value</th>
<th>PREMIS Value</th>
<th>PREMIS</th>
</tr>
</thead>
<tbody>
<tr>
<td>fileMIMEType</td>
<td>audio/x-flac</td>
<td></td>
<td></td>
</tr>
<tr>
<td>formatLibraryID</td>
<td>ExL-Fmt-24417</td>
<td></td>
<td></td>
</tr>
<tr>
<td>fileExtension</td>
<td>flac</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Agent</td>
<td>DROID</td>
<td></td>
<td></td>
</tr>
<tr>
<td>formatRegistry</td>
<td>PRONOM</td>
<td>PRONOM</td>
<td>formatRegistryName</td>
</tr>
<tr>
<td>formatRegistryID</td>
<td>fmt/279</td>
<td>fmt/279</td>
<td>formatRegistryKey</td>
</tr>
<tr>
<td>formatRegistryRole</td>
<td>-</td>
<td>specification</td>
<td>formatRegistryRole</td>
</tr>
<tr>
<td>formatName</td>
<td>ExL-Fmt-24417</td>
<td>audio/x-flac</td>
<td>formatName</td>
</tr>
<tr>
<td>formatVersion</td>
<td>-</td>
<td>-</td>
<td>formatVersion</td>
</tr>
<tr>
<td>formatDescription</td>
<td>Free Lossless Audio Codec</td>
<td>-</td>
<td>formatVersion</td>
</tr>
<tr>
<td>formatNote</td>
<td>-</td>
<td>-</td>
<td>formatNote</td>
</tr>
<tr>
<td>exactFormatIdentification</td>
<td>TRUE</td>
<td>-</td>
<td></td>
</tr>
<tr>
<td>mimeType</td>
<td>audio/x-flac</td>
<td></td>
<td></td>
</tr>
<tr>
<td>agentVersion</td>
<td>5</td>
<td></td>
<td></td>
</tr>
<tr>
<td>agentSignatureVersion</td>
<td>50</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
One or two lessons

**NLNZ only terms**

**HardwareUsed**

Ideal: Operating system used to capture material (sound preservation and digitisation).

**PhysicalCarrierMedia**

Ideal: The media that content was transferred to us on.
One or two lessons

Issues

- Outdated vocabulary
- Hard to maintain vocabulary
- No clear ownership of vocabulary
- Refining can cause reconsideration (changing older entries?)

- **Magnetic disc** [I assume this is a sound one relating to magnetic recording?]
- **Optical disc** [We've been using this for CDs and CDRs I think. But this could logically also cover DVD+R. DVD-R and CD-ROM, so if we have optical disc, do we need those others?]  
- **Solid Sate Media** [Another sound one? Please provide a definition and some examples?]  
- **Hard drive** [What is the difference between hard drive and portable hard drive?]  
  - 3.5 inch IBM formatted floppy disk  
  - 3.5 inch IBM formatted floppy disk (double density)  
  - 3.5 inch Mac formatted floppy disk  
  - 5.25 inch IBM formatted floppy disk  
- **CD-ROM** [what relationship does this have to optical disc?]  
- **DVD+R** [what relationship does this have to optical disc?]  
- **DVD-R** [what relationship does this have to optical disc?]  
- **Portable hard drive** [What is the difference between hard drive and portable hard drive?]
Some concluding thoughts

Key driver is the **use** the information (significant properties)

Differences can result because of:
1. deliberate reinterpretation
2. ‘accidental’ reinterpretation

Some barriers to implementation
1. How to get the information to fill the elements?
2. System design can make it hard to follow exact compliance on all elements
Thoughts/questions?

peter.mckinney@dia.govt.nz