

Preservation MD at National Library of New Zealand

Peter McKinney

National Digital Heritage Archive

National Library of New Zealand Te Puna Mātauranga o Aotearoa

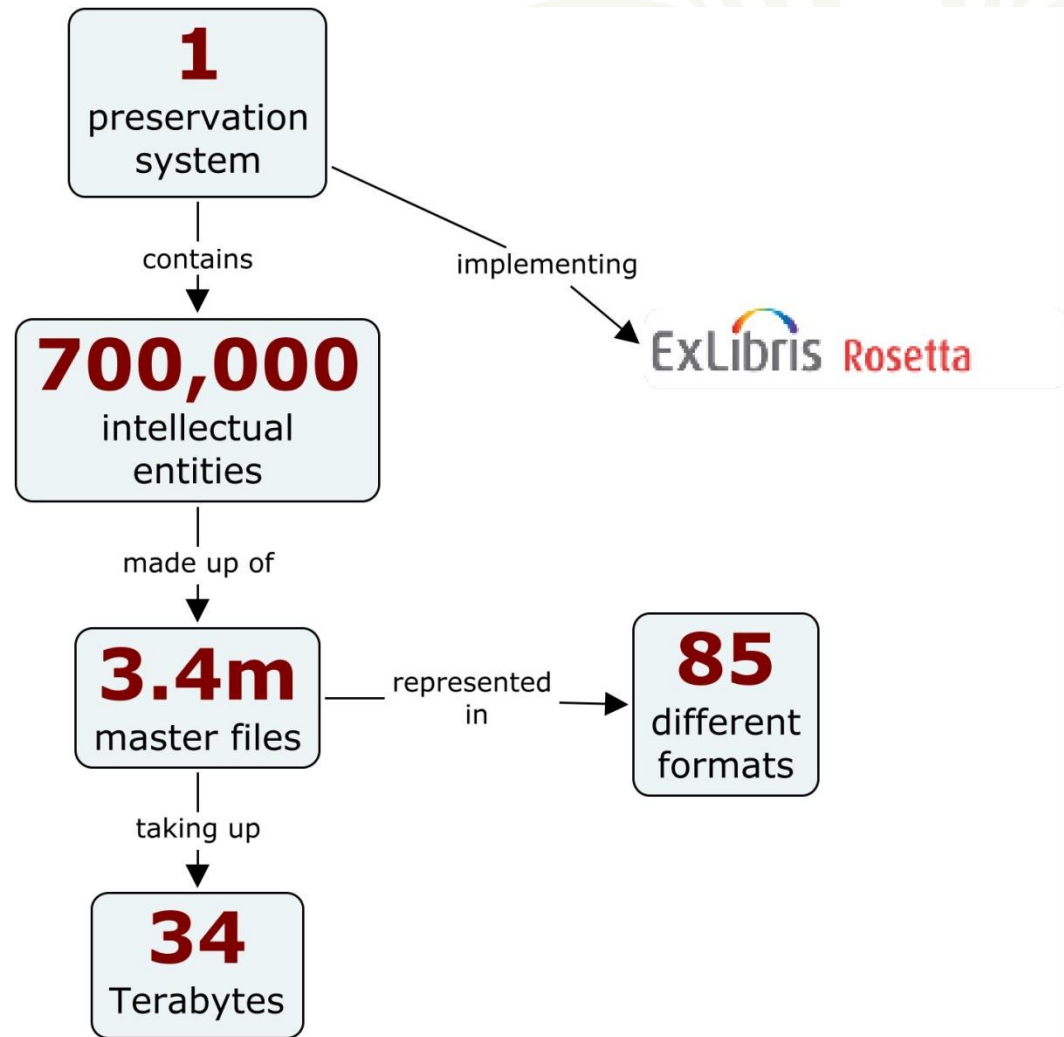
Going to cover...

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

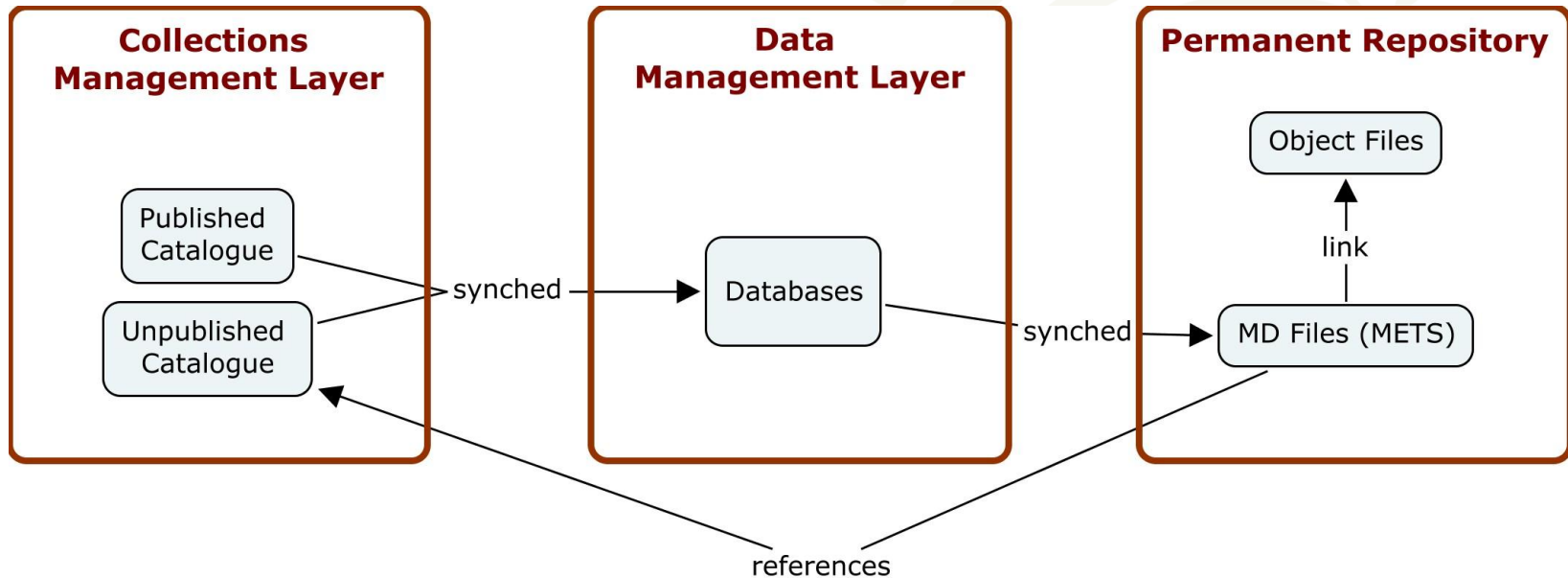


The diagram illustrates the structure of a Permanent Repository. It consists of two main components:

- Layer**: A box on the left containing a partial view of a light blue rounded rectangle.
- Permanent Repository**: A larger box on the right containing:
 - Object Files**: A light blue rounded rectangle at the top.
 - MD Files (METS)**: A light blue rounded rectangle at the bottom.
 - link**: A vertical arrow pointing from MD Files (METS) up to Object Files.

Relationships:

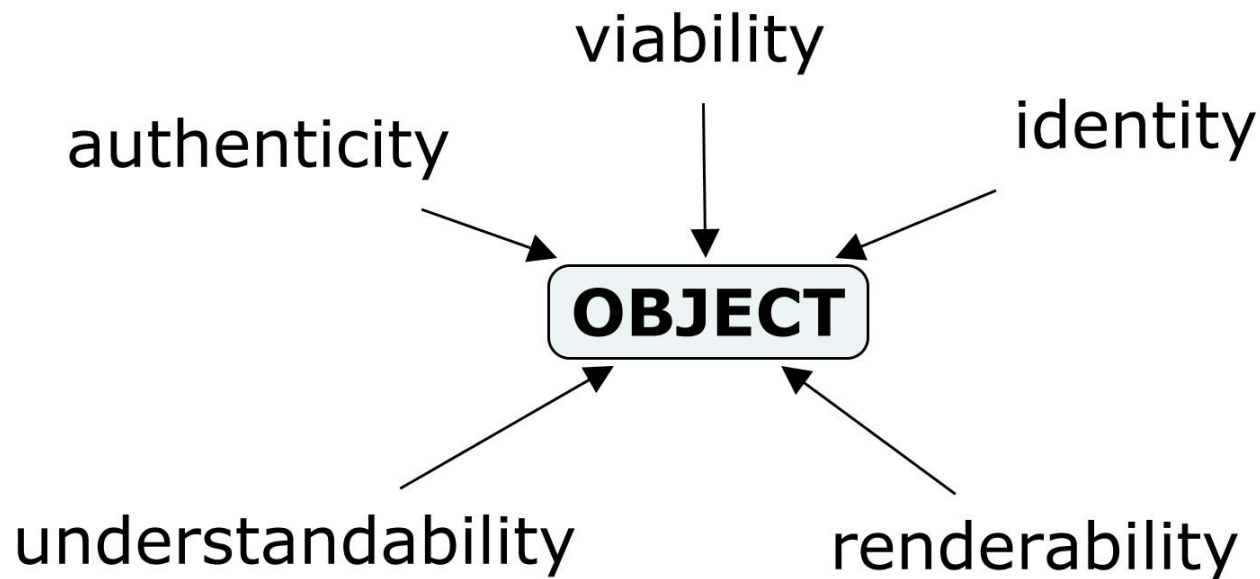
- An arrow labeled **synced** points from the Layer box to the MD Files (METS) box.
- An arrow points from the bottom of the Permanent Repository box to the bottom of the slide.



PREMIS

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).



PREMIS Compliance 2

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

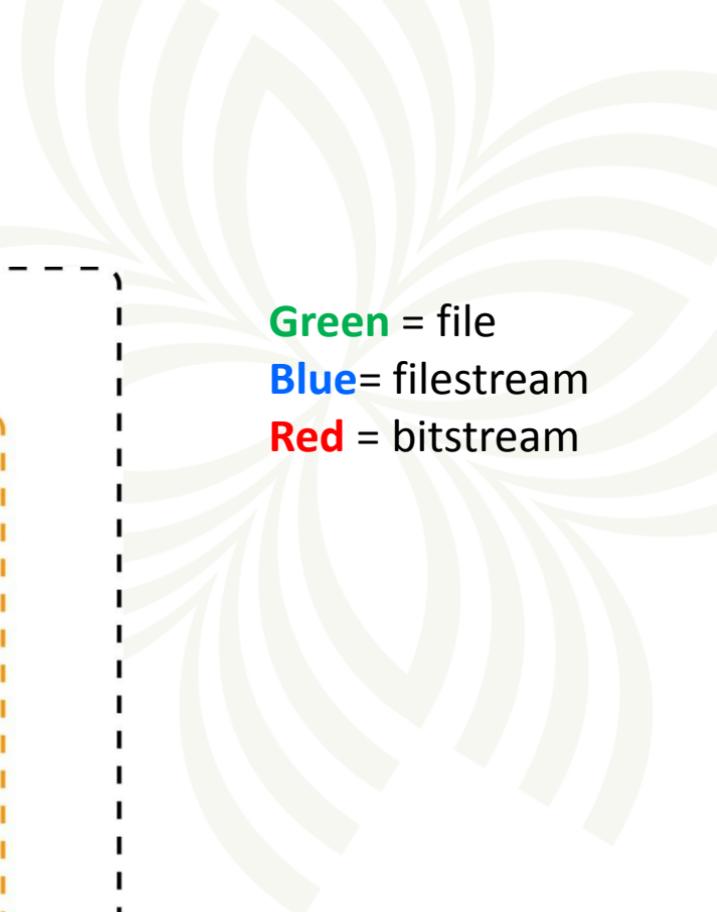
Filestream: a set of bits embedded within a file that **can** be transformed into a standalone file without adding any additional information

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

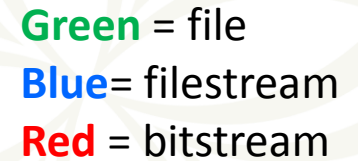
Written to
FILE level

Written to
BITSTREAM
level



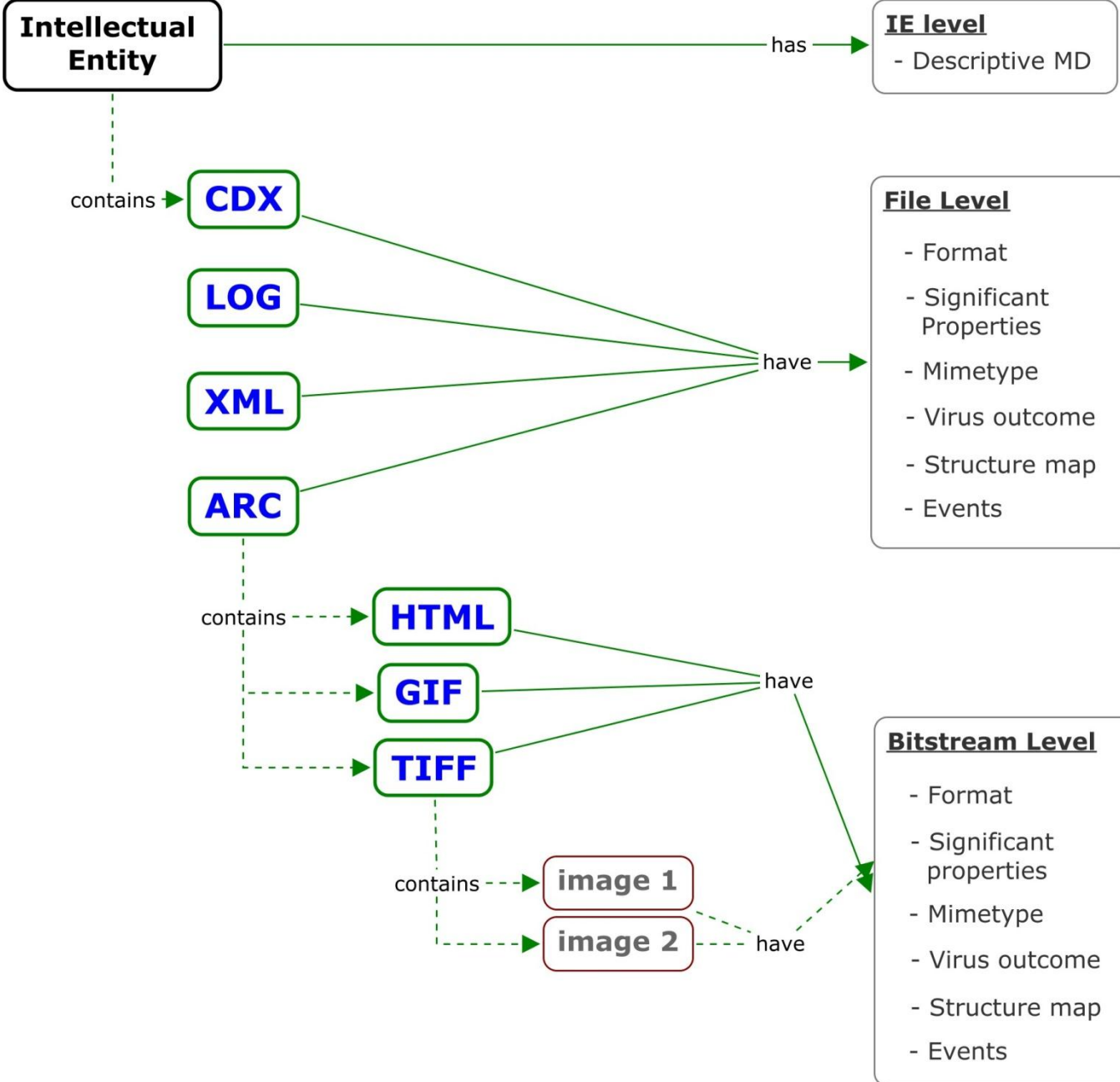


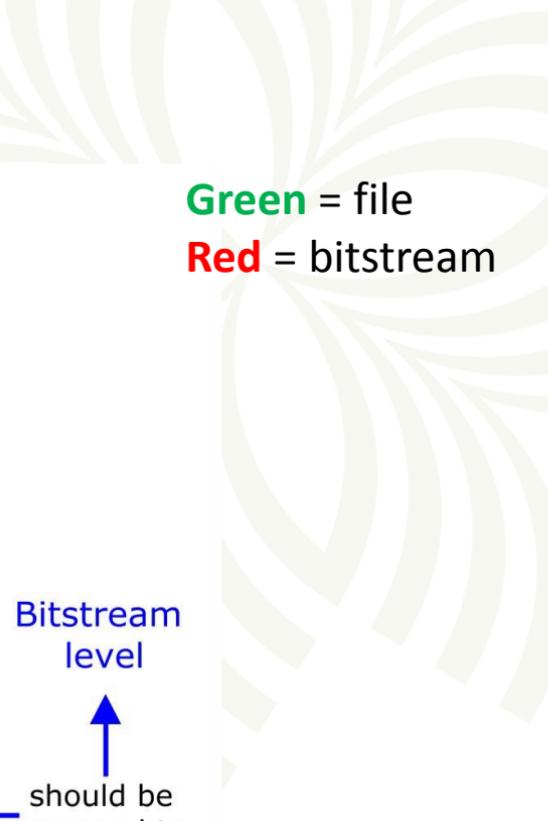
Green = file
Blue = filestream
Red = bitstream





Proposed NLNZ mapping

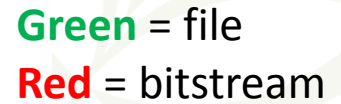




Green = file
Red = bitstream

Bitstream
level

↑
should be
mapped to



Significant properties

1.4 significantProperties

1.4.1 significantPropertiesType

1.4.2 significantPropertiesValue

1.4.3 significantPropertiesExtension

Subtle (?) difference in interpretation

Deviations – significant properties

PREMIS definitions:

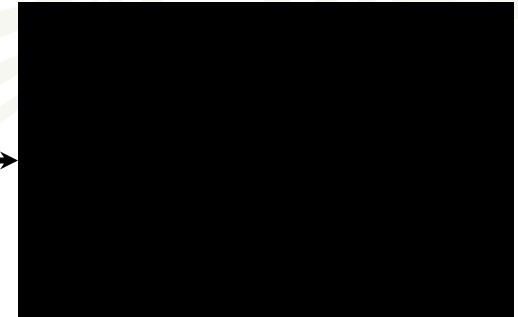
“Characteristics of a particular object subjectively determined to be important to maintain through preservation actions.” P.39

“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.

Derivative creation



www.natlib.govt.nz





Used for:

- 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

Events - PREMIS

- 2.1 eventIdentifier*
- 2.2 eventType
- 2.3 eventDateTime
- 2.4 eventDetail
- 2.5 eventOutcomeInformation*
- 2.6 linkingAgentIdentifier*
- 2.7 linkingObjectIdentifier*

“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.”




Diagram illustrating a transaction being stored in a database:

- A transaction (labeled "5" and "it") is shown on the left.
- An arrow points from the transaction to the text "stored in".
- Another arrow points from "stored in" to the text "Database".

Events in Rosetta

**NDHA**
NATIONAL DIGITAL HERITAGE ARCHIVE

Home | Producers | Submissions | Data Management | Preservation |

**Metadata List**

Intellectual Entity PID	IE6408846	Created on	01/09/2011 10:39:33
Updated on	27/09/2011 09:38:59	Updated by	nga
Version	2		

IE (IE6408846)

Preservation Master Revision 2 (REP6989977)

File (FL6989978)

Preservation Master Revision 1 (REP6408847)

File (FL6408848)

Derivative Copy (REP6408917)

File (FL6408918)

Derivative Copy (REP6408835)

File (FL6408836)

Object SummaryMetadataServices

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**



to

PERMANENT REPOSITORY

un-ect

Fixity events in Rosetta

event Identifier Type	DPS
event Identifier Value	27
event Type	Validation
event Description	Fixity check performed on file
event Date Time	26/09/2011 16:07:05
event Outcome1	SUCCESS
event Outcome Detail1	PROCESS_ID=81176898;PID=FL6989978;SIP_ID=76745;DEPOSIT_ACTIVITY_ID=85062;MF_ID=333919100;TASK_ID=30;ALGORITHM_NAME=MD5;PRODUCER_ID=126755891;



Fixity events in Rosetta

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

Current stats:
1.2m METS files
72Gb

End of Year 2
6.4m METS files
360Gb

End of Year 5
14m METS files
792Gb

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

1.3 *preservationLevel*

1.3.1 *preservationLevelValue*

1.3.2 *preservationLevelRole*

1.3.3 *preservationLevelRationale*

1.3.4 *preservationLevelDateAssigned*

- 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.

1.8 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*

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



Dependency unit

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

1.8.4. dependency

1.8.4.1 dependencyName

1.8.4.2 dependencyIdentifier

1.8.4.2.1 dependencyIdentifierType

1.8.4.2.2 dependencyIdentifierValue

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 processed of decoding or unbundling

1.5.1 compositionLevel



- Haven't found a compelling case to use it yet
(we generally unzip before ingest)
- Complex, in terms of system, to get this information



Format information - discussion

“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

objectCharacteristics	format	formatRegistry	formatRegistryName	PRONOM
objectCharacteristics	format	formatRegistry	formatRegistryKey	fmt/279
objectCharacteristics	format	formatRegistry	formatRegistryRole	specification
objectCharacteristics	format	formatDesignation	formatName	audio/x-flac
objectCharacteristics	format	formatDesignation	formatVersion	-
objectCharacteristics	format	formatNote		-



Format Identification

NLNZ example of
FLAC file

generalFileCharacteristics	fileMIMEType	audio/x-flac
generalFileCharacteristics	formatLibraryID	ExL-Fmt-24417
generalFileCharacteristics	fileExtension	flac
fileFormat	Agent	DROID
fileFormat	formatRegistry	PRONOM
fileFormat	formatRegistryID	fmt/279
fileFormat	formatRegistryRole	-
fileFormat	formatName	ExL-Fmt-24417
fileFormat	formatVersion	-
fileFormat	formatDescription	Free Lossless Audio Codec
fileFormat	formatNote	-
fileFormat	exactFormatIdentification	TRUE
fileFormat	contentType	audio/x-flac
fileFormat	agentVersion	5
fileFormat	agentSignatureVersion	50



Format Identification

Comparison

NLNZ	NLNZ Value	PREMIS Value	PREMIS
fileMIMEType	audio/x-flac		
formatLibraryID	ExL-Fmt-24417		
fileExtension	flac		
Agent	DROID		
formatRegistry	PRONOM	PRONOM	formatRegistryName
formatRegistryID	fmt/279	fmt/279	formatRegistryKey
formatRegistryRole	-	specification	formatRegistryRole
formatName	ExL-Fmt-24417	audio/x-flac	formatName
formatVersion	-	-	formatVersion
formatDescription	Free Lossless Audio Codec		
formatNote	-	-	formatNote
exactFormatIdentification	TRUE		
imeType	audio/x-flac		
agentVersion	5		
agentSignatureVersion	50		



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 State 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