NATIONAL
DIGITAL HERITAGE
ARCHIVE

iPres London September 29/30 2008

Digital Preservation at the National Library Of New Zealand

Steve Knight,
NDHA Programme Architect

Te Puna Mātauranga o Aotearoa
NATIONAL LIBRARY
OF NEW ZEALAND

The ‘digital native’

NATIONAL DIGITAL HERITAGE ARCHIVE

Why do national libraries care?

“A National Library is a place where a nation nourishes its memory and exerts its imagination –

where it connects with its past and invents its future.”

Pierre Ryckmans. 1996. “Perplexities of an electronically illiterate old man.” Quad-rant, September 1996, No 329.

Te Puna Mātauranga o Aotearoa
NATIONAL LIBRARY OF NEW ZEALAND

The ‘digital native’

NATIONAL DIGITAL HERITAGE ARCHIVE

Isn’t it amazing how much kids know?

- Democratization of information production and access
- ‘New Zealanders connected to information important to all aspects of their lives’.
- Paradigm shift in client expectations of how knowledge and information should be made available to them
- Relevance and viability of national libraries determined by their ability to respond to these changing expectations

Te Puna Mātauranga o Aotearoa
NATIONAL LIBRARY OF NEW ZEALAND

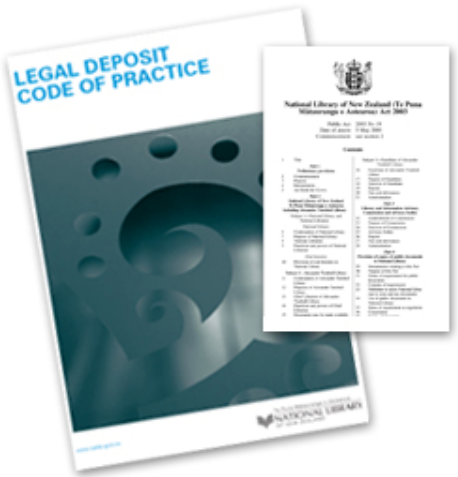
Environment for change

NATIONAL DIGITAL HERITAGE ARCHIVE

Legislation

National Library of New Zealand (Te Puna Mātauranga o Aotearoa) Act 2003

Public Act 2003 No 19
Date of assent 5 May 2003
Commencement see section 2



Strategic Vision

International Community

Te Puna Mātauranga o Aotearoa NATIONAL LIBRARY OF NEW ZEALAND

Ex Libris FROM LIBRARY SYSTEMS TO INFORMATION SERVICES

Sun microsystems

The NDHA challenge

NATIONAL DIGITAL HERITAGE ARCHIVE

If the only goal were to ingest and preserve digital content in complete isolation from the other systems and processes then digital preservation would be a much simpler task.

Organisational readiness

Resources, services and infrastructure supporting digital preservation

Integration with existing systems

INDIGO

WEB CURATOR TOOL

Migration of digital assets

80,000 intellectual entities made up of around 280,000 files

Measuring success

60 key performance measures

Te Puna Mātauranga o Aotearoa
NATIONAL LIBRARY OF NEW ZEALAND

The Technology Response 1

NATIONAL DIGITAL HERITAGE ARCHIVE

Buy or build

Proprietary or open source

Which religious position

- Important to look at the required institutional outcome

- Repository solutions, digital archiving solutions and digital preservation systems are unlikely to be the same thing

- Commercial solution
  - vs. building it yourself
  - vs. project based company
    - User community
    - Enhancements
    - Continuity
    - Open source 80%

Te Puna Mātauranga o Aotearoa
NATIONAL LIBRARY OF NEW ZEALAND

The Technology Response 2

NATIONAL DIGITAL HERITAGE ARCHIVE

It is important from NLNZ perspective that the solution is not NLNZ specific

- Digital Preservation System (DPS)
  - generic software solution for the wider market
  - broad ranging digital preservation solution for a range of community interests
- NDHA is the NLNZ implementation of DPS
  - wider functionality and business change are required for practical digital preservation within any given institution

Te Puna Mātauranga o Aotearoa
NATIONAL LIBRARY OF NEW ZEALAND

Its not just hardware and software

NATIONAL DIGITAL HERITAGE ARCHIVE

Organisational readiness

Legislation and strategies are not sufficient

‘No job will be unchanged’

•Chief Executive/National Librarian

Digital preservation requires interaction with all the organisation’s processes and procedures -

•Business Processes – workflows, procedures and policies

•Capacity & Capability – resources and skills

•Performance Measures – reporting and measuring

•Internal Training – system & staff training

•Producer Management – service, marketing & training

•Business & Technical Support – between departments

•Communication – a constant

Te Puna Mātauranga o Aotearoa
NATIONAL LIBRARY OF NEW ZEALAND

Integration

NATIONAL DIGITAL HERITAGE ARCHIVE

It's not all about the Digital Preservation System

- Deposit Applications development
- Existing Collection Management Systems integration
- Browser based content delivery systems development
- Existing resource discovery and delivery systems integration
- Reporting systems
- Common Services Integration
- Data Migration

Te Puna Mātauranga o Aotearoa
NATIONAL LIBRARY OF NEW ZEALAND

Integration so far

NATIONAL DIGITAL HERITAGE ARCHIVE

Milestone
Y
Y
Y
Y
Y

- Staff deposit application
- HTTrack to ARC converter utility
- Archived website migration tools
- OMS data migration tools
- Content aggregator
- Delivery viewers

Te Puna Mātauranga o Aotearoa
NATIONAL LIBRARY OF NEW ZEALAND

What is INDIGO

NATIONAL DIGITAL HERITAGE ARCHIVE

Indigo

Forms of …

- **Romanic: indicum, indicus**
- **Spanish: indico**
- **Portugese: endego**
- **Dutch: Indigo**
- **NDHA: in dey go**

- Internal Submission Application
  - Submission Information Package (SIP) Creation Tool (Templates, Hotkey support
- Packages up
  - Files (supports complex digital objects)
  - Metadata (Structure map creation – METS)
  - Digital object structure – multiple representations
  - Fixity generation (MD5)
  - Links to descriptive record – CMS integration
  - Links producer records
  - Submits SIP to the NDHA

Te Puna Mātauranga o Aotearoa
NATIONAL LIBRARY OF NEW ZEALAND

Performance Measures

NATIONAL DIGITAL HERITAGE ARCHIVE

How do we measure what we're doing?

From widgets to outcomes

A move to management information with over 60 key performance indicators including:

- Key performance indicators
- Reporting
- Audit
- internal ingest
- + response actions, ie for over/under delivery

Te Puna Mātauranga o Aotearoa
NATIONAL LIBRARY OF NEW ZEALAND

September defect numbers

NATIONAL DIGITAL HERITAGE ARCHIVE

Constant bug finding, verifying, fixing

Defect Tracking (Totals)				
	Current Reporting Period (by defect owner)			
	ExLibris (Israel)	**ExLibris (NZ)**	**Indigo**	**Migration****
Total defects raised in reporting period	**0**	**128**	**3**	**0**
Total defects fixed in reporting period	**64**	**70**	**1**	**1**
Defect Status	**Total Defect Numbers Outstanding** (by defect owner)			
Critical	**0**	**1**	**2**	**0**
High	**7**	**41**	**7**	**0**
Medium	**16**	**50**	**10**	**0**
Low	**13**	**150**	**11**	**0**
Total Defects Outstanding	**36**	**242**	**29**	**0**

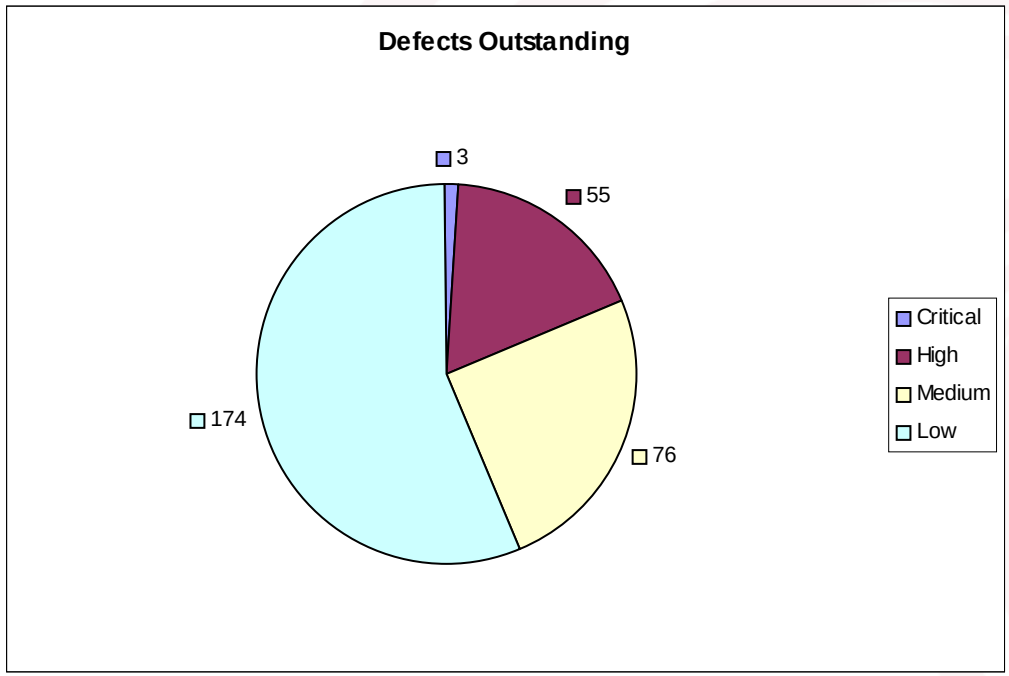
Note: These figures do not reflect issues / defects raised with regard to DROID (application error, memory usage outgoing access, validation errors), JHOVE (outgoing access and validation errors) or enrichment tasks not working due to configuration problems.

Te Puna Mātauranga o Aotearoa
NATIONAL LIBRARY OF NEW ZEALAND

September defect numbers

NATIONAL DIGITAL HERITAGE ARCHIVE

Constant bug finding, verifying, fixing



Pass rate of those run:	90.54%	Passed
	2.95%	Failed

Te Puna Mātauranga o Aotearoa
NATIONAL LIBRARY OF NEW ZEALAND

The OMS

NATIONAL DIGITAL HERITAGE ARCHIVE

Object Management System (OMS) 2005

What to do with all our digital data

Supporting Legal deposit while NDHA is developed

•Published material deposited under Legal deposit

•Digitised material from the Library's digitisation programme

•Websites harvested as part of the Library's web archiving programme

•Material will be migrated as new content

Te Puna Mātauranga o Aotearoa
NATIONAL LIBRARY OF NEW ZEALAND

Migration - principles

NATIONAL DIGITAL HERITAGE ARCHIVE

Object Management System (OMS) 2005

What to do with all our digital data

This will:

- test initial workflows and process configurations
- Impose the same metadata constraints (referential integrity, data validation)
- Impose the same validation checks (fixity verification, virus check, format identification and metadata extraction)
- Impose the same enrichment tasks (CMS identifier association, access derivative generation)

that will be applied in a live operational setting.

- This should give an indication of the amount of effort required to migrate the rest of the National Library's digitized content into the NDHA system.

Te Puna Mātauranga o Aotearoa
NATIONAL LIBRARY OF NEW ZEALAND

Migration - process

NATIONAL DIGITAL HERITAGE ARCHIVE

Ongoing iterative process

Actual run:

3 x 15% + 5%, ie 50% of the OMS extract were uploaded into the production environment in fully load balanced mode (deposit, staging and permanent).

The last 5% run had to be restarted due to Oracle table space problems.

Migration - configuration

Load configuration

Non-functional:
•64-bit mode
•(jmagick and jpeg2000 libraries to be upgraded to 64-bit)
• 8 workers

Functional:
•VS dummy evaluator
•No CMS update and no thumbnail creation
•Thumbnail creation was turned on for the last 5% that are still being processed. Note that it was run with partial VS (fixity check only) and enrichment limited to indexing
•VS evaluator to be activated once automated VS rules have been fully configured.
•Oracle must be monitored for disk usage and for database server status (one of the database servers on the UAT RAC was down for 2 weeks without anybody noticing).

Migration – processing statistics

NATIONAL DIGITAL HERITAGE ARCHIVE

How long does it take to load an object?

And what are we counting?

Average load processing statistics:

- checksum - 2.2 sec per file
- Virus check - 7 sec per file
- File format - 3 sec per file
- Metadata validation - 18 sec per IE
- Object index - 0.1 sec per file , 0.1 sec per representation , 0.1 sec per IE
- Permanent storage - 9 sec per IE

Metrics for migration –

- number of files
- data volume
- number of Ies
- are SIPs TA counted as processed?

Migration – 3rd party tools 1

NATIONAL DIGITAL HERITAGE ARCHIVE

Third party tools –

DROID

JHOVE

MET

Migration testing from OMS to DPS has thrown up a number of problem scenarios related to 3rd party tools:

•DROID pushed significant percentage of files to the TA workbench
•DROID memory usage does not allow to increase number of workers resulting in under usage of CPU usage
•JHOVE has thrown up similar problems.

DROID is being used as a risk analysis tool, not a decision-making tool.

Need to develop VS rules for specific file type errors.

Where relevant, a flag associated with their DROID status is assigned so that we can discover and evaluate later.

Migration – 3rd party tools 2

NATIONAL DIGITAL HERITAGE ARCHIVE

Exemplar Validation rules for DROID

Category	Risk	Action
Objects completely unidentified by DROID	High	- These should be accepted into the permanent repository with the file format 'unknown'. - These should be flagged in order that preservation actions can be undertaken on them as the highest priority by the Business Unit. - Where access copies exist, these should be migrated into the system.
(Two) Files with multiple hits in DROID	Low	- These should be accepted into the permanent repository as the first hit on RTF. - We have manually checked these files and they are RTF.
Objects that have a positive match, but a file extension mismatch	Low	- These should be accepted into the permanent repository with the file type DROID has identified, ignoring the mismatch. The Business Unit will undertake preservation actions.

Migration – 3rd party tools 3

NATIONAL DIGITAL HERITAGE ARCHIVE

The way forward

A more consistent approach to identification and validation:

- one registry that many tools could use
- one tool that uses many registries.

An agreed super-set of risk grading criteria.

This would be at the level of high-level threats to the format.

Local institutions can add in sub-sets of grading information that relate directly to their own situation.

This is a real world problem right now and a solution would generate real value-add.

Migration – virus checking/integration

NATIONAL DIGITAL HERITAGE ARCHIVE

Virus checking (7 sec per file)

Integration with internal systems

Anti Virus checking has a high impact on the performance of the migration loads.

Options:
•Verify that your OMS data is virus free before loading
•Remove the Virus Check from the Validation stack
•After OMS load add the Virus Check to the Validation stack

• Voyager SRU server performed poorly under load from the CMS Update Task (enrichment) during migration.

Turns out it was a configuration issue.

DPS Functionality at Day 1

NATIONAL DIGITAL HERITAGE ARCHIVE

Phase 1 Delivery

From producer management ➔ workflow automation ➔ delivery, audit trails & reporting

- User management
- Producer management
- Deposit 1
- Deposit 2
- Validation stack
- Intellectual Entity (IE) data model
- Submission Information Package (SIP) submission
- SIP processing
- Deposit registration
- Technical analyst
- Workbench
- Consolidated appraisal workbench
- DPS transformers
- Deposit Application Programme Interface (API)
- Audit & provenance
- Process management
- User management API
- Permanent repository
- Delivery
- Meditor
- Reports
- Back office configuration

Te Puna Mātauranga o Aotearoa
NATIONAL LIBRARY OF NEW ZEALAND

NATIONAL DIGITAL HERITAGE ARCHIVE

End result – a fully functioning National Digital Heritage Archive as at 1 November 2008

With

Phase 2 due in 2009

Te Puna Mātauranga o Aotearoa
NATIONAL LIBRARY OF NEW ZEALAND

Steve Knight,
NDHA Programme Architect