Preserving webharvests at the National Library of New Zealand Te Puna Mātauranga o Aotearoa

Peter McKinney
Digital Preservation Policy Analyst
National Library of New Zealand Te Puna Mātauranga o Aotearoa
Agenda

• Relationship with National Library of China
• National Library mandate
• Collection analysis
• Preservation of those collections
• Future plans
Agreement on Cooperation: NLZ and NLNZ

- Drawn up in the spirit of friendship
- Principles of equity and reciprocity
- Purpose is to enhance the ability to contribute to improvement of cultural and economic life of their respective nations
- Cooperation on digital library matters

Share knowledge and information in relation to Digital Preservation
Mandate

“The Minister may, by notice in the Gazette, authorise the National Librarian to make a copy, at any time or times and at his or her discretion, of public documents that are internet documents in accordance with any terms and conditions as to format, public access, or other matters that are specified in the notice.” National Library Act 2003
Two web archiving programmes

Selective harvesting of specific websites or parts of websites (since 1999)


http://topics.breitbart.com/fishing+pole/

http://www.trimarinegroup.com/operations/fleet.php
## Selective Harvests (since 1999)

<p>| | |</p>
<table>
<thead>
<tr>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>No. of Websites</strong></td>
<td>14,943</td>
</tr>
<tr>
<td><strong>No. of ARC files</strong></td>
<td>71,374</td>
</tr>
<tr>
<td><strong>Size on disk (Tb)</strong></td>
<td>5.63</td>
</tr>
<tr>
<td><strong>Total number of files in harvest IE</strong> (excluding those within the ARC)</td>
<td>c.291,000</td>
</tr>
<tr>
<td><strong>No. of Files contained within ARCs</strong></td>
<td>87,112,551</td>
</tr>
</tbody>
</table>
Collections

Whole of Domain Harvests

- **Terabytes**
  - 2008: 4
  - 2010: 6
  - 2013: 10

- **No of files**
  - 2008: 5,000,000
  - 2010: 10,000,000
  - 2013: 15,000,000

- **No. of files (within WARCs)**: Increasing over time.
- **Data size**: Increasing over time.
Digital preservation

Thoughts for today
1. How much information?
2. Whole of domain into preservation system
3. ARC to WARC migration
Digital preservation

Simply:

1. What are the technical characteristics of the content?
   - What format is it in?
   - What are the properties the file exhibits?
   - Is it a valid expression of the specification?

2. What is the institution’s capability in terms of rendering the content?
Digital preservation

We use that information for:

• Repository analysis
• Risk assessment
• Preservation planning
• Preservation action
• Evaluation

Validation stack - current
Validation stack - current

Thoughts
1. Mimetype is insufficient
2. Not enough information about the website files to support preservation management
Extracted information - mimetypes

Total number of mime types across selective harvests = 543
Validation stack – proposed

Container file → Bytestream extraction → File 1, File 2, File 3

Validation Stack
- Fixity
- Virus
- Format ID
- MD extraction

Intellectual Entity → Representation
- CDX
- LOG
- XML
- ARC

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

Bitstream Level
- Format
- Significant Properties
- Mimetype
- Virus outcome
- Structure map
- Events

IE level
- Descriptive MD

Rep level
Validation stack – proposed

Thoughts
1. METS file is ‘large’
2. Take only format and fixity?
3. Take only format summary at ARC level?
Domain Harvests into preservation system

Web Curator Tool -> DEPOSIT -> STAGING (TA Assessor, Arranger, Approver) -> Validation Stack -> Enrichment -> PERMANENT REPOSITORY -> Delivery -> Publishing -> Search Tools (Eg. Find / Tapuhi / Voyager / Beta)

Rosetta
Whole of Domain legal considerations

- Films, Videos, and Publications Classifications Act 1993;
- Criminal Procedure Act 2011;
- Defamation Act 1992;
- Human Rights Act 1993;
- Copyright Act 1994;
- Privacy Act 1993
Whole of Domain legal considerations

Legal
- Objectionable/illegal material
- Copyright
- Is giving access “republishing”

Technical/Intellectual
- One corpus?
- Full text search?
- Themed?
- What is the data model?
ARC to WARC conversion

Why?
• Take advantage of most up to date standard
• Normalise all web content

Critical considerations
- Stakeholder engagement
- Development of tools
- Proofs of conversion
Conclusions

• Scale of harvests challenge processes for description, access and preservation
• Preservation of harvests at NLNZ is at a basic level
• Want to move to fuller preservation management of the website files