From E-Publications Librarian to Web Archivist – a librarian's perspective on 10 years of web archiving at the National Library of New Zealand

Susanna Joe, Web Archivist
Te Puna Mātauranga o Aotearoa
National Library of New Zealand

New Zealand Government

Te Tari Taiwhenua
Internal Affairs

NZ web archiving – legislative context

- Our focus is on selective web archiving but we also do whole of domain harvests
- We collect using both Legal Deposit legislation and through permission-based collecting
  - National Library of New Zealand (Te Puna Mātaraunga o Aotearoa) Act 2003
  - National Library Requirement (Electronic Documents) Notice 2006

Te Tari Taiwhenua
Internal Affairs

Current Staffing and Organisation

- Three Web Archivists are part of the Alexander Turnbull Library's Digital Collections Services team
- One Web Archivist spends half their time selecting born digital NZ music published online
- 1 Digital Preservation Web Engineer in Preservation, Research & Consultancy team (NLNZ digital preservation)
- We also work closely with NLNZ staff in Legal Deposit, collection development, and cataloguing

Te Tari Taiwhenua
Internal Affairs

Alexander Turnbull Library

- Founded on Alexander Turnbull's private collection of books, manuscripts, photographs, paintings and sketches relating to NZ and the Pacific - bequeathed to the nation in 1918
- The 2003 National Library Act confirms the status of the Alexander Turnbull Library as a collection belonging to the Crown; to be maintained in perpetuity
- Its other purpose is to develop research collections with a focus on the history and development of NZ and the Pacific and their peoples
- New Zealand's foremost heritage research library and the country's biggest repository of publications and documents about NZ and its people

Selection approach : documentary history

Highly selective – intellectual content will follow the same priorities selection as for the other published and unpublished collections of the Alexander Turnbull Library

Research value of information content to be determined by considering:

- Quality and depth
- Cultural and social significance
- Research interest (past, current, or emergent)
- Representation of the subject area

Collecting Priorities & Strengths

https://natlib.govt.nz/about-us/strategy-and-policy/collections-policy/newzealand-pacific

- Government
- Politics (elections)
- Māori
- Community Groups
- Music
- Arts & culture
- NZ history
- 2010/11 Canterbury earthquakes & rebuild
- Sports & recreation
- Environment
- Pacific Islands
- Social concerns

Efforts and activities 1999-2003

High priority selection areas:

- Government
- NZ history (national)
- NZ history (Wellington region)
- NZ literature
- NZ education
- Māori

Harvesting pilots/projects:

- 1999 NZ general election
- 2002 NZ general election
- 2002 America's Cup yachting
- 2003 NZ budget

2005 developments

- 2 new E-Publications Librarian/Selectors appointed
  - Selection framework and website appraisal
  - Event harvest on 2005 NZ general election using HTTrack
  - Harvested material stored in interim Object Management System but no access (dark archive)
- NLNZ and BL agree to collaborate in development of the Web Curator Tool (WCT)

2006: Significant milestone #1

E-Legal Deposit = The National Library Requirement (Electronic Documents) Notice 2006

Te Tari Taiwhenua
Internal Affairs

2006 : Significant milestone #2

Web Curator Tool v.1.1 software publicly released September 2006

Te Tari Taiwhenua
Internal Affairs

Significant milestone #3

Launch of the National Digital Heritage Archive (NDHA), NLNZ's digital preservation system in 2008

Te Tari Taiwhenua
Internal Affairs

2007-2009: Web archiving growth

- Escalation in web harvesting:
  - 946 websites archived (2007)
  - 1826 websites archived (2008)
- E-Publications Librarian/Selector (Music) appointed
- British Library releases WCT v1.3
- First NZ Whole of Domain harvest (2008)

2010–2011: Constraints and decline

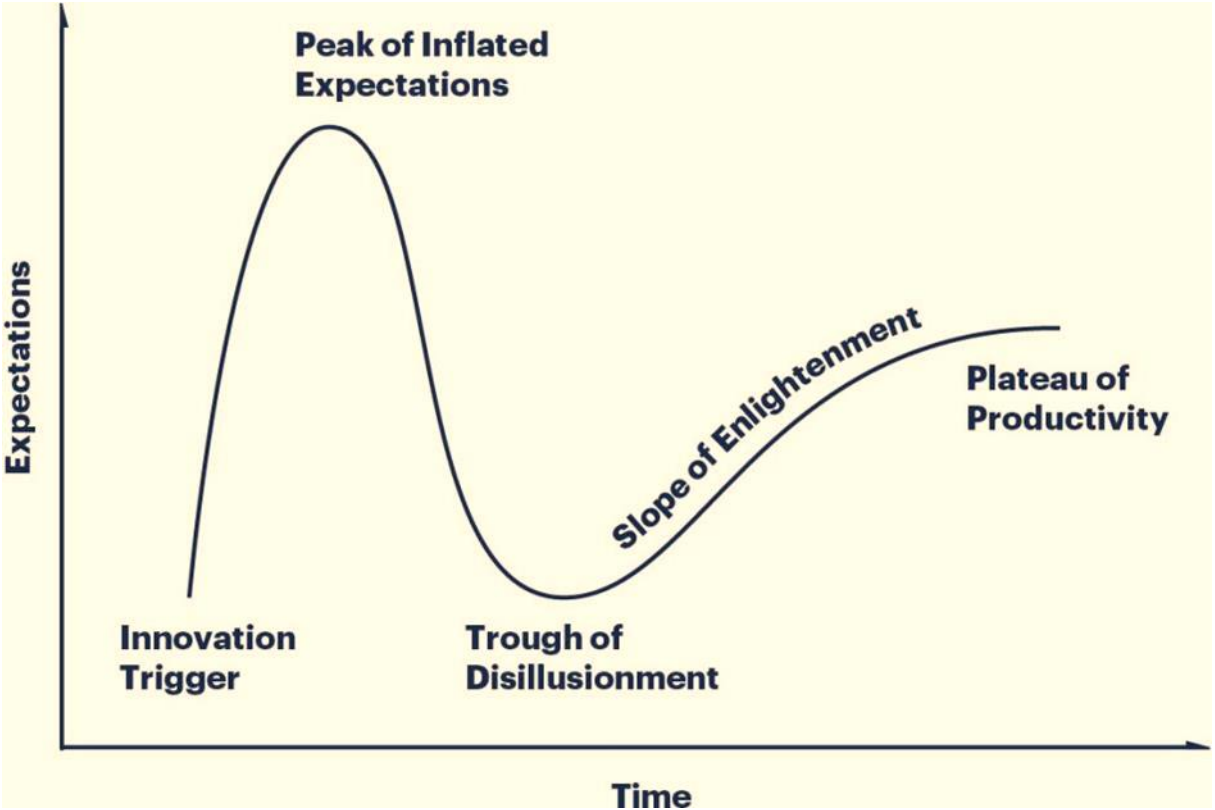
- Economically constrained environment
  - ➢ WCT development slows. BL now pursuing other tools.
  - ➢ Second Whole of Domain harvest (2010)
  - ➢ National Library becomes part of the Department of Internal Affairs (DIA)
- Departure of key NLNZ staff

Te Tari Taiwhenua
Internal Affairs

2012–2014: Stagnation

- Very limited technical expertise for WCT development and support for web archivists
  - ➢ Serious impact on digital collecting capability – WCT ‘end of life’?
- Growing impact of social media but no resources or capability to collect
- Third WoD (2013)
- Difficult working environment - very low staff morale

Trough of Disillusionment?!

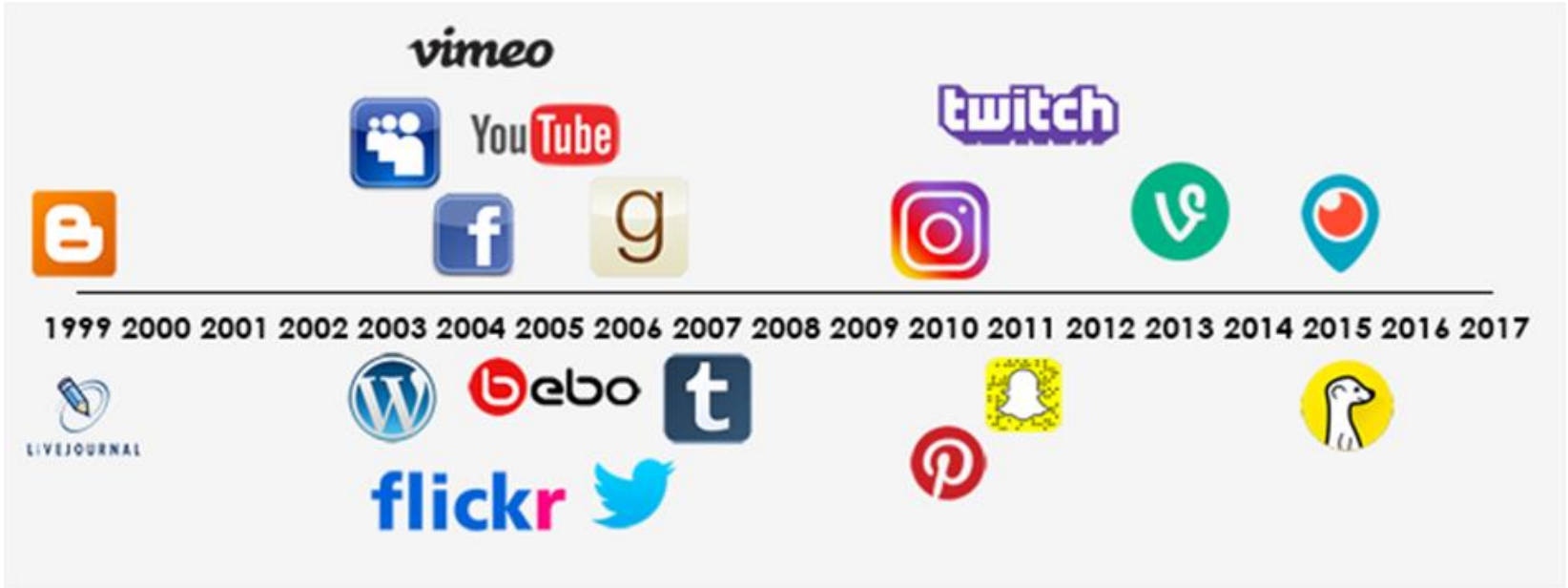


2015/2016: Still catching up…

- Ben seconded to NDHA and starts looking at WCT bug fixes
- Whole of domain harvests recommence
- Research into use of NZ Web Archive: https://natlib.govt.nz/librarians/reports-and-research/use-of-the-nz-web-archive
- Job title change to ‘Web Archivist’
- Recognition of importance of collecting social media..
  - Trials of Archive-it
  - 2016 Kaikoura earthquake - Twitter data collection pilot

Te Tari Taiwhenua
Internal Affairs

Social media development timeline



1999 2000 2001 2002 2003 2004 2005 2006 2007 2008 2009 2010 2011 2012 2013 2014 2015 2016 2017

2017: Wheels are still turning…

- Digital Preservation Web Engineer appointed
- WCT upgraded to use Heritrix3
- WCT v.2.0 development in collaboration with KBNL
- OpenWayBack
- 2017 General Election Twitter data collected

Te Tari Taiwhenua
Internal Affairs

Current state

- H3 in WCT but not yet at full capacity (6 concurrent harvests)
- WCT limitations which H3 cannot solve
  - Exploring other web harvesting tools (Webrecorder)
- Selective web harvesting priorities have not really changed
- Contribute to IIPC collaborative web crawls
- Social media projects:
  - 2017 election Twitter content (re-evaluating workflows)
  - Investigating collecting Facebook
- Staff morale better - renewed optimism

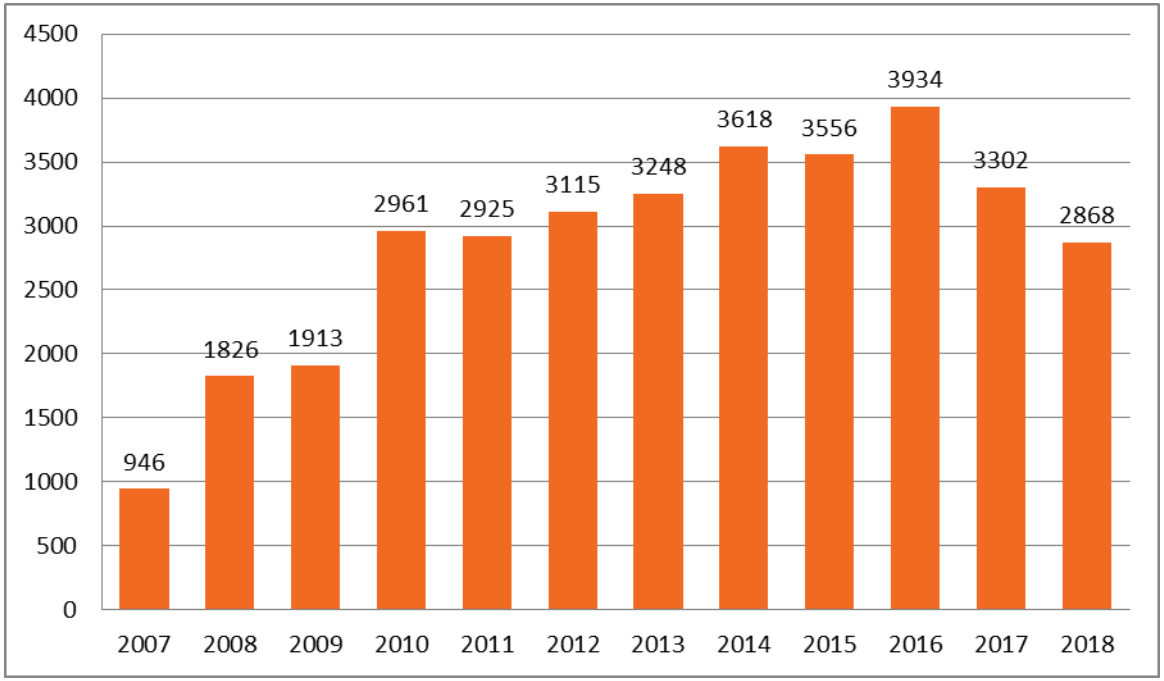
Te Tari Taiwhenua
Internal Affairs

Facts and figures

- 33,500 web instances in the NDHA
  - ➢ Of these around 5,600 are unique titles
- Average of around 56 new website titles harvested per month
- Estimated at least 30% of new websites selected are unable to be harvested and archived due to technical issues

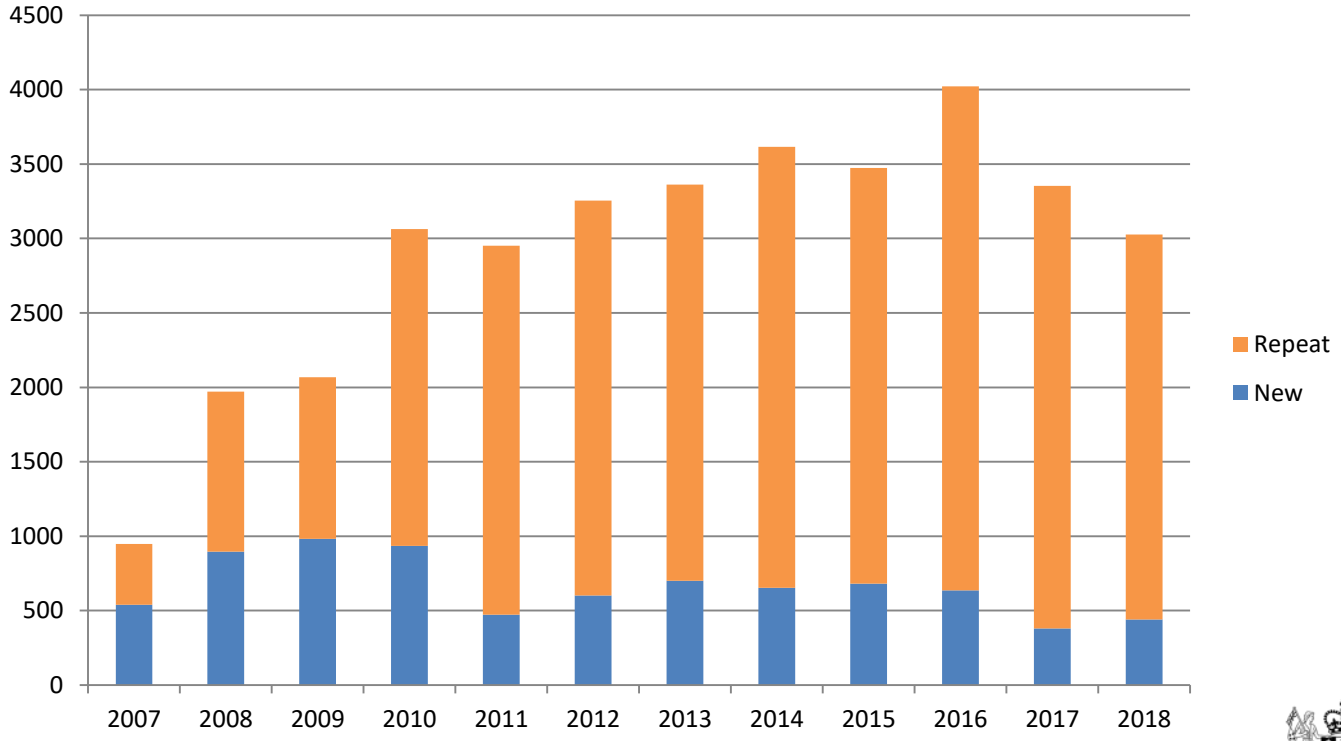
Te Tari Taiwhenua
Internal Affairs

Number of websites archived



Te Tari Taiwhenua
Internal Affairs

Harvested websites : new vs repeat ratio



Te Tari Taiwhenua
Internal Affairs

Permission-based archiving

- Permission-based website archiving very small – around 200 websites recorded in WCT
  - ➢ permission approval rate = 69%
- Digital music:
  - ➢ Percentage of music archived by ATL with permission has been slowly declining in last several years : 50.2% (2014) vs 27.8% (2018)

Te Tari Taiwhenua
Internal Affairs

Current challenges

- Increasing amount of born digital content which we cannot collect due to ongoing constraints
  - Legal issues/risk assessment
- Access –
  - Searching via library catalogue still limited
  - No indexing of archived websites
  - Ongoing viewer issues (H3 harvests are not viewable in NDHA)
  - No public access to Whole of Domain harvests

Looking ahead – challenges & opportunities

- Legal Deposit legislation due for review
- WCT interoperability
- Web archiving – how do we keep up?
- Social media – lots of questions and conversations!
  - ➢ Opportunities for relationship building with content creators and researchers

Te Tari Taiwhenua
Internal Affairs