How much do you really want to know? Harvesting to build institutional knowledge resources

Edmund Balnaves

Prosentient Systems

Library as publisher

The journey of a thousand miles begins with one step. Lao Tzu

The cautious journey of libraries into the institution publishing space: the IFLA metaphor

- IFLA 2007
  - Building the architectures: Digital Libraries and OAI/PMH
- IFLA 2017
  - Pervasive presence of IT in sections
    - Digital Humanities Special Interest Group
    - Big Data Special Interest group
    - Library as publisher
- Today: from Viki: Queensland memory – Simon – Published library technician

Digital Content

Institutional content

- Print (candidate for scanning?)
- Born digital (integration candidate)

External content by institutional authors

- Harvesting – content published by staff in the organisation externally

External content relevant to the organisation

- Harvesting via Google, Twitter and other sources
- Media monitors
- Discovery services (eg EBSCO, Proquest)
- Linked / open data

Supporting Institutional knowledge

- Changing role of the library
- Changing role of the library staff
- Information literacy support
- Working with research areas and the embedded librarian
- Managing knowledge resources
- Harvesting information resources into an institutional repository

- Building the bridge to knowledge

Where to begin?

7 billion websites

35 trillion web pages

Ratio of librarians to websites:

1 : 100,000,000

Ratio of librarians to web pages

1 : 3,500,000,000

… and that's just the beginning.

Facebook

Twitter

Where to begin

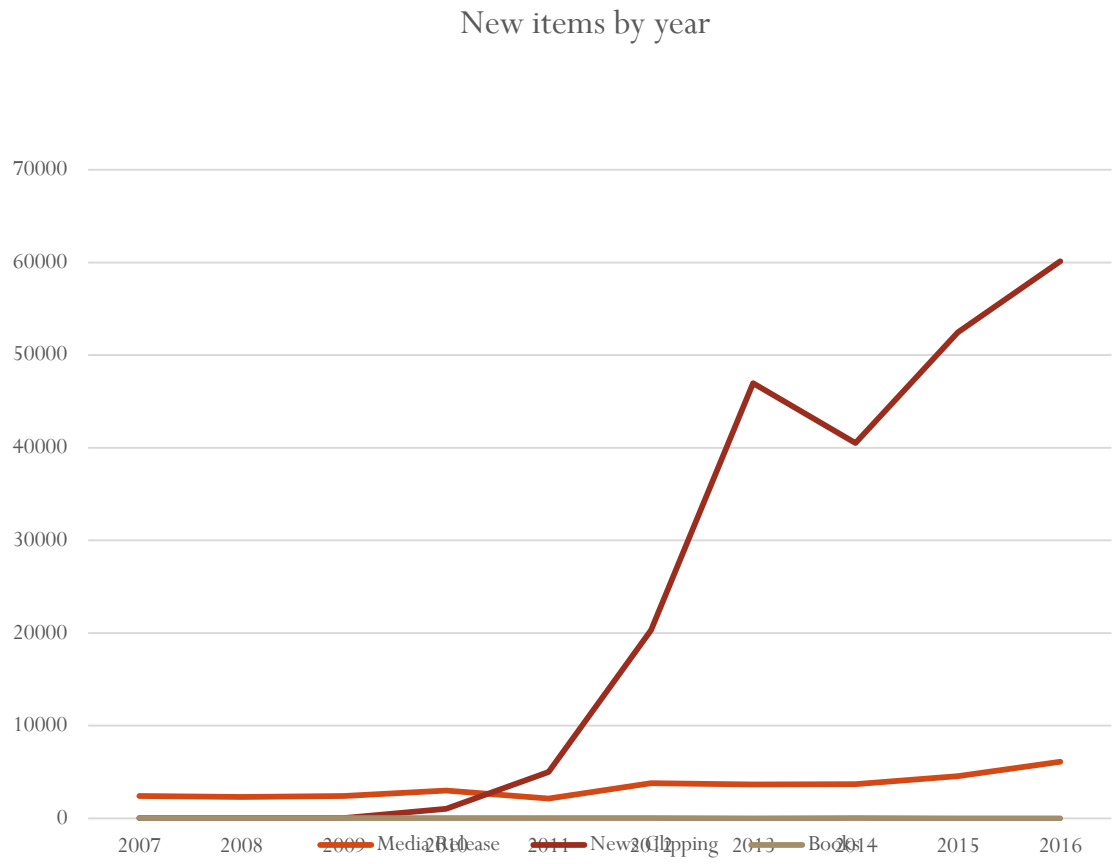
- Information audit
  - What are you interested in?
  - What do you have in the organisation
  - What do you have access to?
  - Advocacy
    - Who can you team up with?
    - Do they know what they are missing?

Your friends and enemies

- The publisher as friend
  - Innovations in digital delivery
  - Collaboration with discovery providers
  - Innovations in open access and open source journals
- The publisher as enemy
  - Licensing
  - Copyright
- Open Access, Creative Commons and pre-print versions
- Open source

In special libraries the change to Digital is definitive and complete

NSW Parliament library – the change



Meeting the challenge: Automated Harvesting

- Limitations of one-at-a-time
- Some examples
  - NSW Parliament
    - 1,025,836 digital objects (possibly the largest DSpace library around)
  - Territory Stories (state library)
    - 245,560 digital objects
  - Austin Health (medical research)
    - 1,714 digital objects

Example 1: Harvesting from pubmed

- ((((((austin health[Affiliation] or austin hospital[Affiliation] or heidelberg repatriation[Affiliation] or royal talbot[Affiliation])))) OR (((austin[Affiliation]) AND ((((((((((((melbourne brain[Affiliation]) OR ludwig[Affiliation]) OR national stroke[Affiliation]) OR epilepsy research[Affiliation]) OR university of melbourne[Affiliation]) OR parent-infant[Affiliation]) OR institute for breathing[Affiliation]) OR brain research[Affiliation]) OR latrobe[Affiliation]) OR la trobe[Affiliation]) OR florey[Affiliation]) OR northern health[Affiliation]) OR liver transplant unit[Affiliation]) OR victorian spinal cord[Affiliation])))) OR ((((heidelberg[Affiliation]) AND ((((liver transplant unit[Affiliation]) OR Florey Institute[Affiliation]) OR victorian spinal cord[Affiliation])))))) OR (((austin pathology[Affiliation]) NOT texas[Affiliation]))) OR olivia newton-john cancer[Affiliation]

Austin: importing into DSpace

- DSpace as a RIS importer that can be used in conjunction with PubMed
- Imported content can be fed into an approval workflow
- Post import curation tasks can do things like mandatory metadata checking and virus checking

Harvesting methodologies

- **RSS**, an interface popular for news syndication, and very relevant to this project - many corporate and government websites implement RSS (webreference.com, 2013);
- **OAI/PMH** - a protocol for bibliographic and record interchange between digital repositories (openarchives.org, 2013);
- **JSON-based information sources**. This has emerged as the data format of choice for many applications, given the success of JavaScript toolkits such as JQuery (a powerful JavaScript Library) and AJAX (Asynchronous JavaScript and XML)(json.org, 2013);
- **schema.org** and similar. Websites that embed specialised tags to enable harvesting, such as proposed by schema.org (a collaboration between Bing, Google, Yahoo! and Yandex);
- **Screen scraping with XPath** with XML and PHP coding to isolate portions of a web page of interest.
- **Google search** (and similar)**.** Search Application Programming interfaces such as Google Custom Search and Bing Search (https://developers.google.com/custom-search/);
- **Email** E-mail APIs (especially IMAP processing tools for PHP).
- **Database** queries
- **SOLR** – direct solr search queries
- **Document property** parsing of metadata embedded in born digital documents, images, videos.

Northern Territory Government

Home | Help

Sign on to:

NORTHERN TERRITORY LIBRARY

Territory Stories

Browse | Advanced Search | Map

Keyword search

Or search by Title search

Babies on the carriage See

About

Territory ANZAC Collection

Discover the personal stories of over 400 soldiers who enlisted in the Northern Territory, traveled from here to enlist or had a strong association with the Territory before or after the First World War.

The Library's **Territory ANZAC collection** pays tribute to those who lost their lives and honours those who served and returned to Australia. Biographical details about enlistees' lives in the Territory before the war plus their Service Records, Red Cross Missing in Action reports, letters home, photos and Unit War Diaries are provided in this collection.

Territory Stories is a repository of historical and culturally significant digital objects. Territory Stories contains photographs, documents, audio and video and allows users to contribute content to enrich the collection.
Territory Stories provides stories that record the history and development of the Northern Territory from the early days, to the present. These stories will entertain, inspire and contribute to our understanding of who we are.

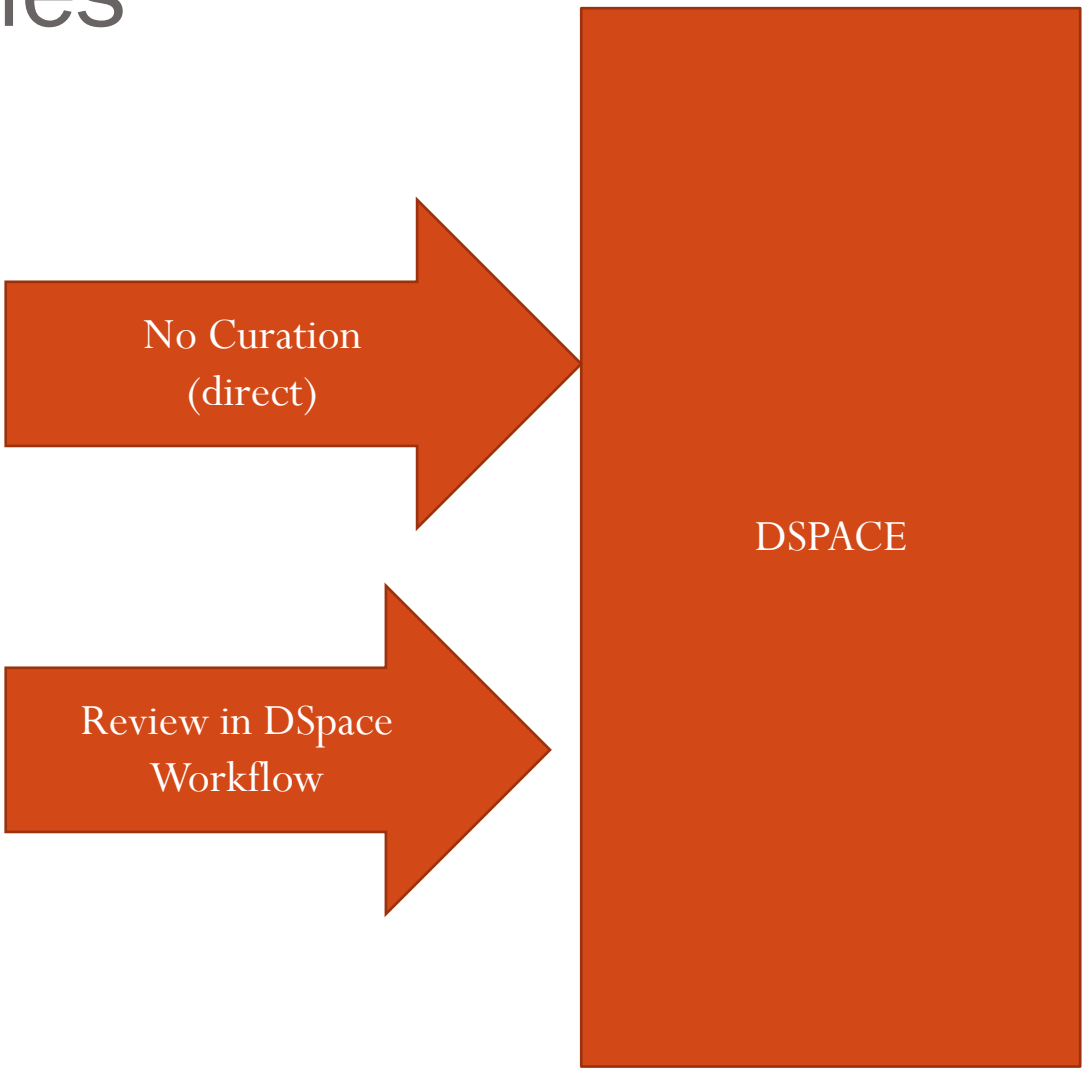
Territory Stories

OAI/PMH harvesting from Government collections

Hansard folder harvesting

PDF folder harvesting from born-digital newspapers collections

Scanned photographic image folder harvesting



NSW PARLIAMENTARY LIBRARY

koha Cart Lists

Search Library catalogue All libraries Go

Advanced search

Home

Catalogue search

Search Media releases

Search Newspaper archive

Search journal articles

Library Home

Search for Media Releases

Filter by date: TO

Keywords and

Keyword Phrase and

Title and

Title Phrase and

Issued By and

Portfolio

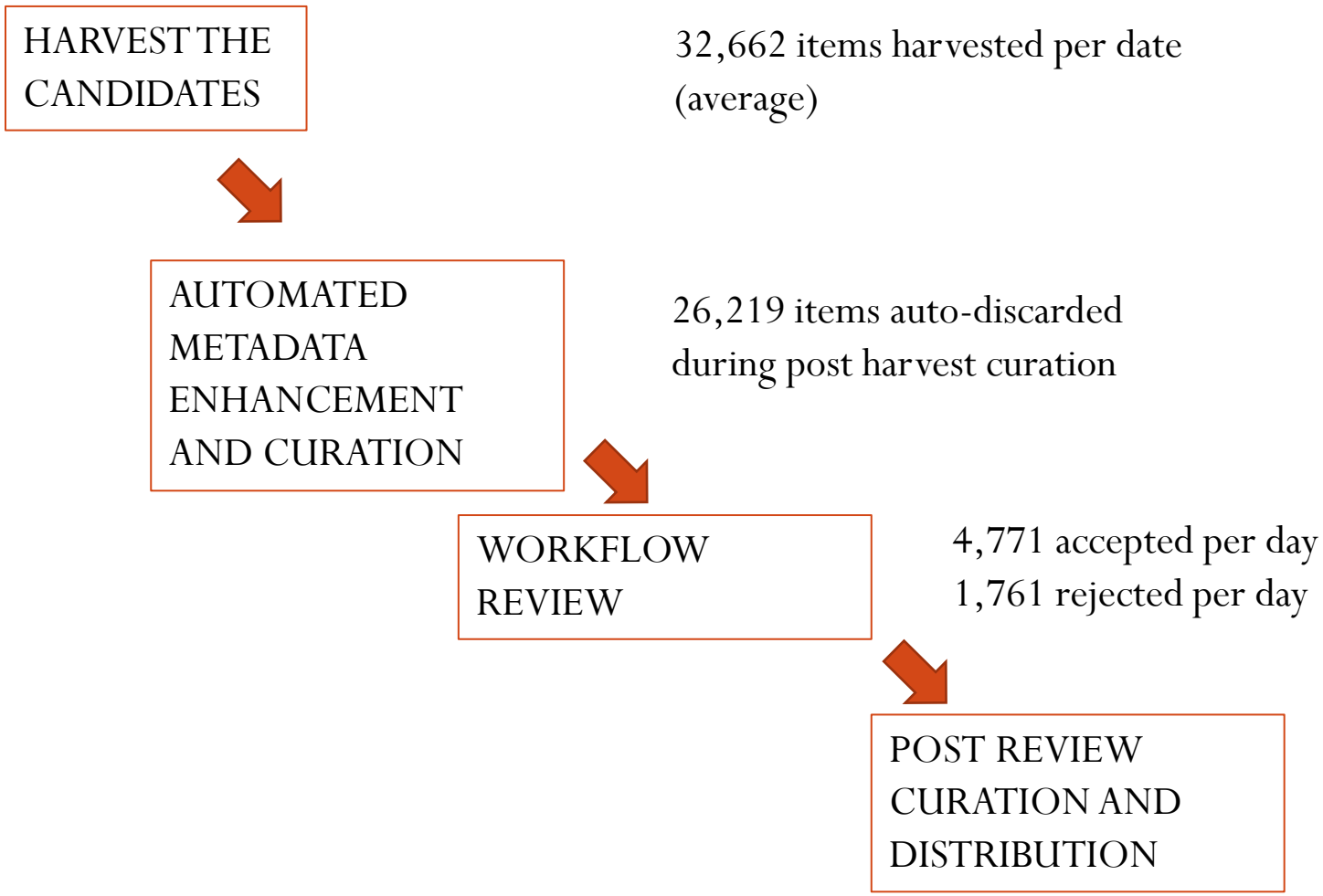
Search Reset Form

About the Media Releases archive

It contains releases issued by Members of the New South Wales Parliament and held in the Parliamentary Library. The database contains records of releases collected since 1991 and PDFs of releases from January 2000. Copies of Press Releases in hard copy issued from late 1976/77 can be obtained with assistance from the Library staff. The Parliamentary Library does not guarantee that every media release issued by Members is held in this database.

The Parliamentary Library is not responsible for the content of the media releases.

Building a workflow



Accept/reject workflow

- 80% of harvested items auto-discarded
- 21% of reviewed items discarded after staff review

Edit news item 845983587

Save and close | Save | Cancel

SAVE & review next | Publish & review next | PUBLISH

DISCARD this item | DISCARD & review next

ID 845983587

Publication Sydney Morning Herald

Publication Date 18/09/2017

Headline Ghost of city's past makes a brief appearance

Section General News

Page 8

Author Tim Barlass

Metadata Matches

Member +

Subject Advertising

Heritage

Planning and Development

isentia

Sydney Morning Herald
Monday 18/9/2017
Page: 8
Section: General News
Region: Sydney Circulation: 93,403
Type: Capital City Daily
Size: 974.00 sq.cms.
Frequency: MTWTFS-

Brief: #NSWPARL
Page 1 of 3

Ghost of city's past makes a brief appearance

Catch a glimpse of the old Peapes sign while you can, writes **Tim Barlass**.

The demolition of a 1960s tower block at the old Menzies Hotel at Wynyard has delivered an historic surprise.

Exposed for the first time in more than half a century is the state-of-the-art (at least it was then) advertising facade of a gentleman's department store that was once a landmark of George Street.

Such works are frequently described as ghost signs, but there is nothing too ghostly about the Peapes gents' and boys' outfitters artwork because it has been covered

"It would have been a real landmark on George Street. Peapes menswear was an institution that was established in 1866 and by the early 20th century it was known for both tailoring and ready-made clothing for men. Before the Sydney Harbour Bridge opened (1932), Wynyard was quite a retailing area so the building itself is a marker of our retailing history down that way."

The Peapes 1922 Christmas Catalogue, describing their new building which opened the following year, states: 'It will be a landmark very easy to find. The position is unique, not only because it commands a full view up Hunter Street but also because it is contiguous to the leading hotels, clubs, financial, commercial and shipping offices and it is the centre of the business area most frequented by

Author Vanessa Berry, whose book *Mirror Sydney* to be published next month contains a section on ghost signs, said they were a window into the past enabling people to imagine the city as it once was.

"People ... really notice this one because it is so bright," she said.

"The '60s was an era in Sydney's history when the prevailing idea was to remove the Victorian city and replace it with a more modern city.

"I wasn't around at the time but from what I have researched it seems that it would have been an unusual idea that this was, perhaps, worth preserving."

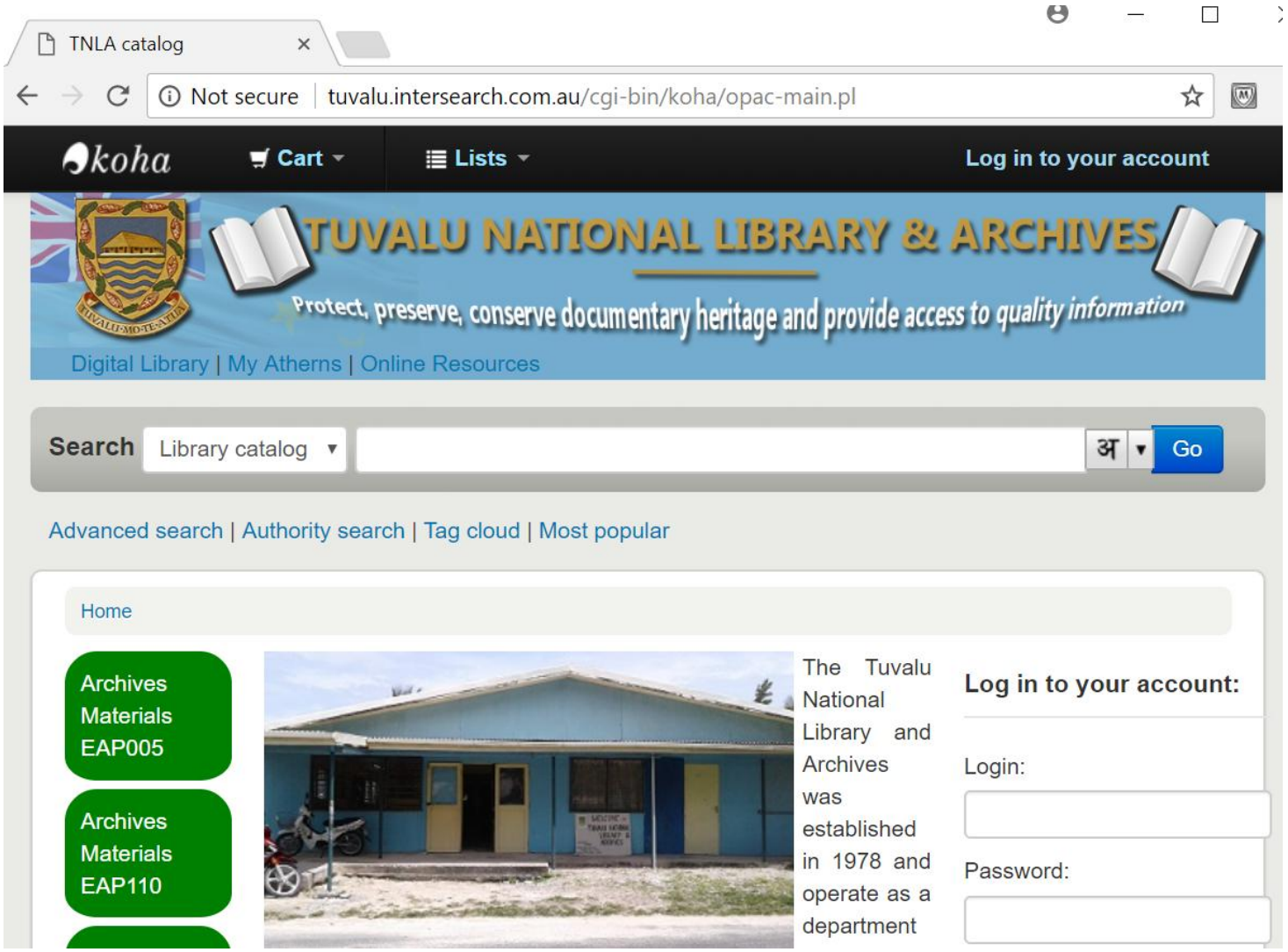
Murray said there were other good examples to be seen around the CBD.

On Sussex Street, where another building has been demolished, is a black-and-white sign for tent-

Sri-Lankan Parliament

- Building a legislative database by harvesting from internal born digital content systems (Hansard and others).
- Content indexing of Hansard content
- Digital scanning of Parliamentary documents

National & Archives Tuvalu



Harvesting existing colonial resources

Existing digital collections from

- British Library
- ANU

Digitisation of the print collection

Library Informatics

- Be across the technology
- Try it out
- Share experiences
- Be adventurous
- Be mindful (long term governance)

What else is out there?

- What about the Dark Web
  - TOR "The Onion Router"
- You may hear about it even when you don't want to know about it! Especially if you find your client data being sold there!
- Is this also an information source that needs to be scanned?

Attributions

- DSpace: http://www.dspace.org/
- Koha: http://koha-community.org/
- OAI-PMH - https://www.openarchives.org/pmh/
- Prosentient: http://www.prosentient.com.au
- TOR https://www.torproject.org + imagery https://pixabay.com/en/deep-web-dark-web-darkness-binary-1106648/

Thank you everyone!

Edmund Balnaves

ebalnaves@prosentient.com.au