Bayerische StaatsBibliothek
Information for Excellence

BibliotheksVerbund Bayern

Münchener DigitalisierungsZentrum Digitale Bibliothek

Web Archiving and Rosetta

Achievements and Challenges

Tobias Beinert, Bavarian State Library

Rosetta at the Bavarian State Library

Central Library and Repository Library for Bavaria

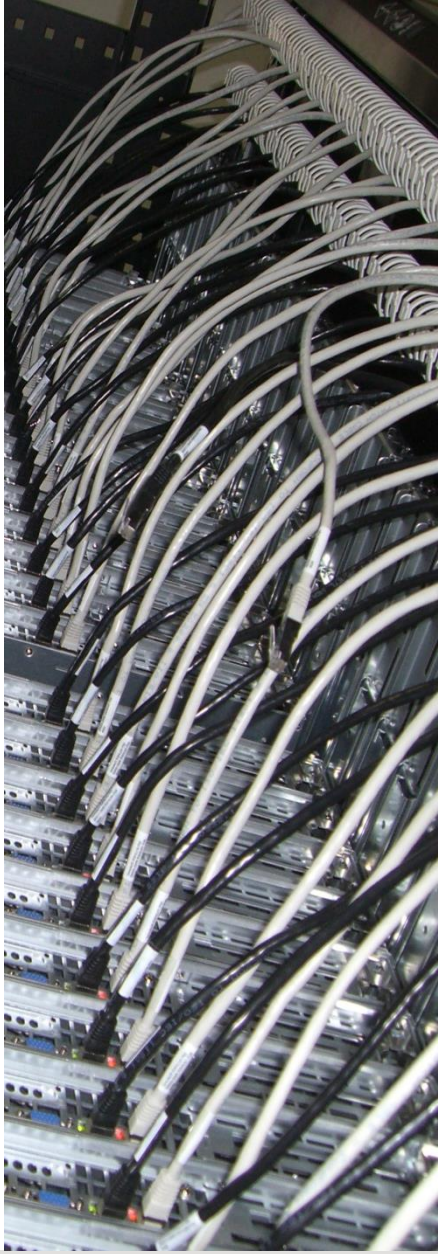
- Last resort for printed and digital works (legal deposit)
- International Research Library
- 10,5 Mio. Printed volumes, 59.000 periodicals (including e-journals),1,9 Mio. digitized works

Routine operation of the existing infrastructure

- Workflows for manual legal deposit and web archiving
- Digitised Works, AV-Media, Images are archived on demand

Rosetta for Bavaria

- License includes Universities



BSB Bayerische StaatsBibliothek
Information in erster Linie

BSB‘s approach for archiving websites

- Selective harvesting for Specialised Information Services and Bavarica

-> Sustainability for ephemeral web resources

- Permissions for harvesting, long-term preservation and access requested
- Captures of a website are done half-yearly
- Approx. 1600 websites archived with several snapshots
- Manual and semi-automated quality control
- Access via BSB‘s catalogue and the gateways of the Specialised Information Services
- Web Curator Tool for Management and Viewer OpenWayback

BSB Bayerische StaatsBibliothek
Information in erster Linie

Integration with Rosetta

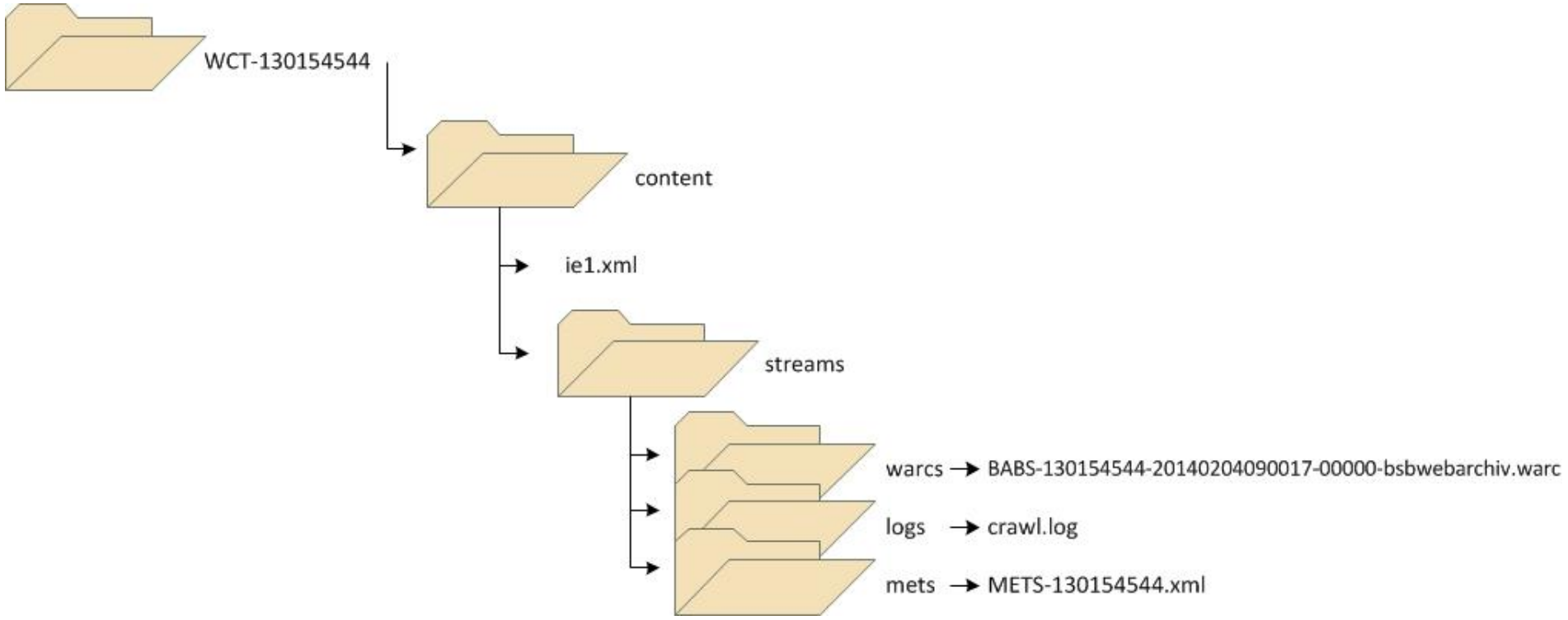
BSB Bayerische StaatsBibliothek
Information in erster Linie

Production of SIPs

- XSL-Mapping from WCT-METS to Rosetta-METS
- Unzip WARCs
- Create Rosetta SIP (WCT METS, crawl.log, WARCs and Rosetta METS)
- Ingest via Submission Job

BSB Bayerische StaatsBibliothek
Information in erster Linie

Rosetta-SIP-Structure



Virus Checking

- ClamAV vs. Sophos: Differences in performance and differences in results
- Running check on storage or via Rosetta
- Actions for infected web archives?

BSB Bayerische StaatsBibliothek
Information in erster Linie

Validation

- Format Identification with DROID
- Technical Metadata Extractor: WARC Metadata Extractor vs. JHOVE WARC Module (Version 1.14)
- JHOVE WARC Module not available in Rosetta

BSB Bayerische StaatsBibliothek
Information in erster Linie

Quoting archived websites

- Process for assigning Uniform Resource Names for every capture
- Development of a Citation Service (java-based)
  - Rosetta SRU for finding IEs
  - Rosetta IE-Update Service for metadata
- Integration in OpenWayback as JavaScript

BSB Bayerische StaatsBibliothek
Information in erster Linie

BSB Bayerische StaatsBibliothek
Information in erster Linie

Externe Links, Formulare und Suchfelder werden innerhalb dieser von der BSB archivierten Website nicht funktionieren.

External links, forms and search boxes may not function within this website archived by BSB.

Diese Webpage zitieren

8 Zeitschnitte
9 Dez 13 - 5 Mai 17

Neuerscheinung am ZMSBw:

Oliver Bange: " Sicherheit und Staat. Die Bündnis- und Militärpolitik der DDR im internationalen Kontext 1969 bis 1990".
» Weiterlesen

58. Internationale Tagung für Militärgeschichte vom 21. bis zum 23. Juni 2017 in Potsdam:

Einsatz ohne
Geschichte

Diese Webpage zitieren:

http://www.zmsbw.de/ in Archiv von: Zentrum für Militärgeschichte und Sozialwissenschaften der Bundeswehr, archiviert von der Bayerischen Staatsbibliothek am 05.05.2017. urn:nbn:de:bvb:12-babs2-0000130212.

Mehr Informationen zum Zitieren von Websites.

Weitere Neuigkeiten aus dem ZMSBw

Geschichte ohne Grenzen? Europäische Dimensionen der Militärgeschichte vom 19. Jahrhundert bis heute.

Zweiter Workshop der Reihe Reihe "Messen - Beurteilen - Entscheiden" fand in Potsdam statt:
Workshop in Kooperation des Zentrums Innere Führung sowie des Zentrums für Militärgeschichte und Sozialwissenschaften der Bundeswehr.

» Weiterlesen

BSB Bayerische StaatsBibliothek
Information in erster Linie

Delivery

- OpenWayback as external viewer on Rosetta Delivery Servers
- Delivery Rule and Plugin for OpenWayback
- Open Wayback indexes and gets data from Rosetta storage
- Rosetta Access Rights are not applied in OpenWayback

BSB Bayerische StaatsBibliothek
Information in erster Linie

Challenges (I)

- Flash, JavaScript, streaming media, database content, dynamic content are still hard to archive
- Focus in Research/Development: Access services vs. curation tools?
- Monopoly for web archiving services?

BSB Bayerische StaatsBibliothek
Information in erster Linie

Challenges (II)

- Improving Access to web archives?
  - Full text indexing of websites
  - Visualization
- Quality control
- Resources?

BSB Bayerische StaatsBibliothek
Information in erster Linie

Perspectives

- Integration of JHOVE WARC Module?
- Integrated viewer for web archives in Rosetta?
- Emulation?
- Legal challenges
  - How to implement Access Rights?
- Preservation Planning

BSB Bayerische StaatsBibliothek
Information in erster Linie

Thank you.

Questions?

beinert@bsb-muenchen.de

Web Archiving and Rosetta

Rosetta Advisory Group Meeting
Sheffield, 12-14 June 2017

Steve Knight – Programme Director Preservation Research & Consultancy, National Library of New Zealand

With thanks to Ben O'Brien (Digital Preservation Web Engineer) and Gillian Lee (Coordinator Web Archives) for putting this presentation together.

Motivation

National Library of New Zealand Act 2003

- E-legal deposit activated August 2006

Mission

To collect, preserve and make accessible New Zealand's documentary Heritage

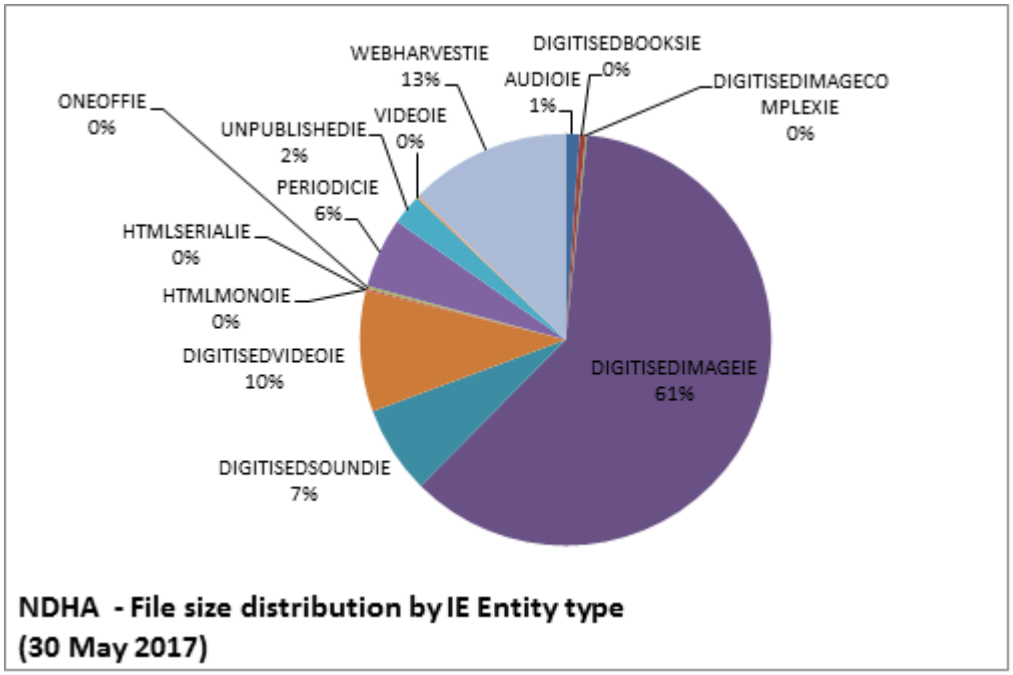
Strategy

- Selective harvesting – Web Curator Tool (2007-)
- Domain harvesting (6 harvests) of .nz domain (2008-)

Some metrics

Number of websites in the collection:

- 7,260 web titles
- 29,943 web instances (includes repeat harvests)
- 15.2 TB storage in Rosetta



NDHA - File size distribution by IE Entity type (30 May 2017)

Number of HTML serials and monographs

- 280 HTML titles
- 3,825 HTML serial issues (195.5 GB)
- 19 HTML monographs(1.7 GB)

In Tray

1 tasks, 13 Notifications

open

Permission Request Templates

open | add new

Harvest Authorisations

183 harvest authorisations

open | add new

Reports

open

Targets

2321 Targets

open

Harvester Configuration

general | bandwidth | profile

Target Instances

334 Scheduled instances, 65 ready for Quality reviews

open | queue | harvested

Users, Roles, Agencies & Rejection Reasons

Users:	open	add new
Roles:	open	add new
Agencies:	open	add new
Rejection Reasons:	open	add new

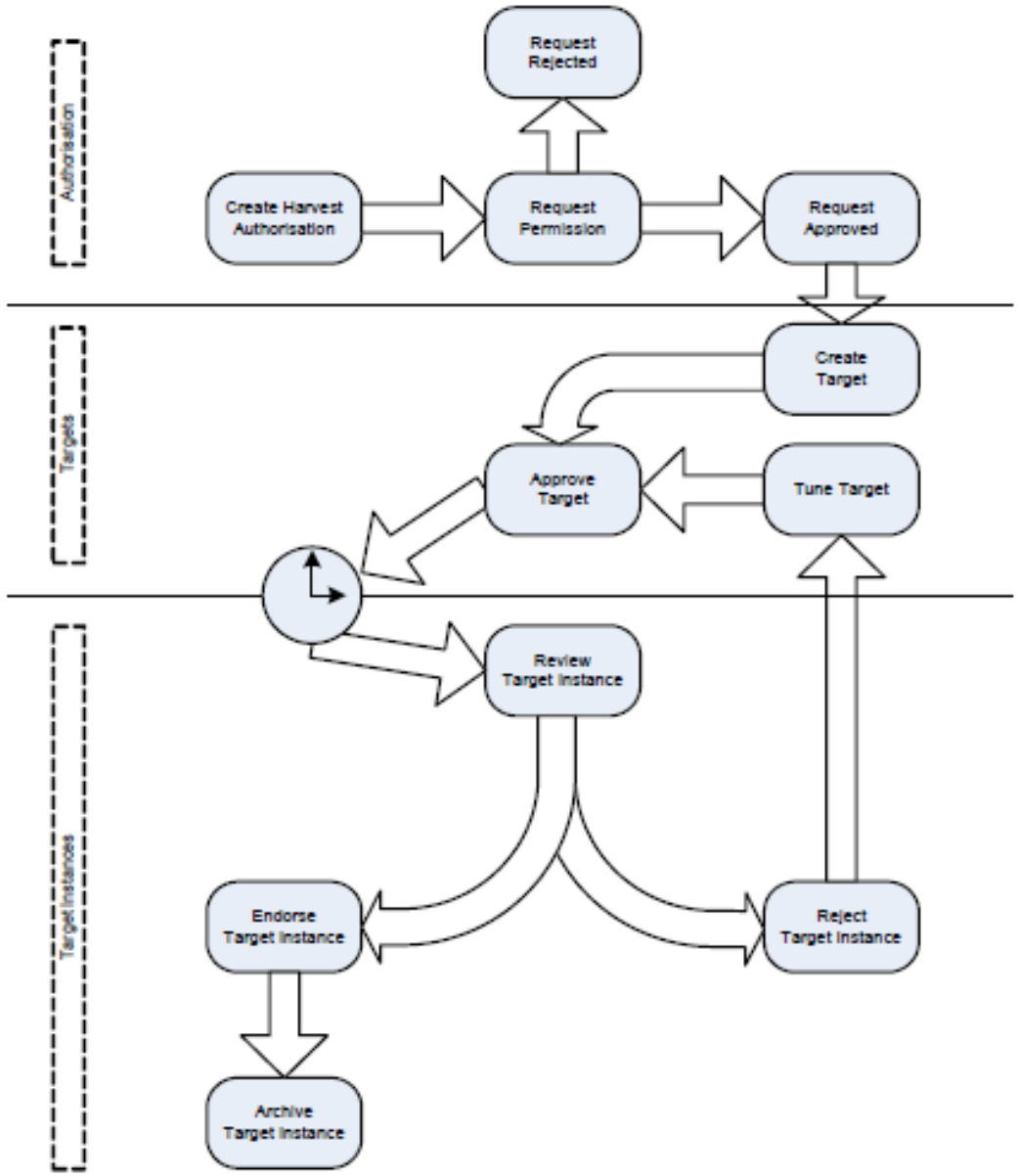
Groups

93 Target Groups

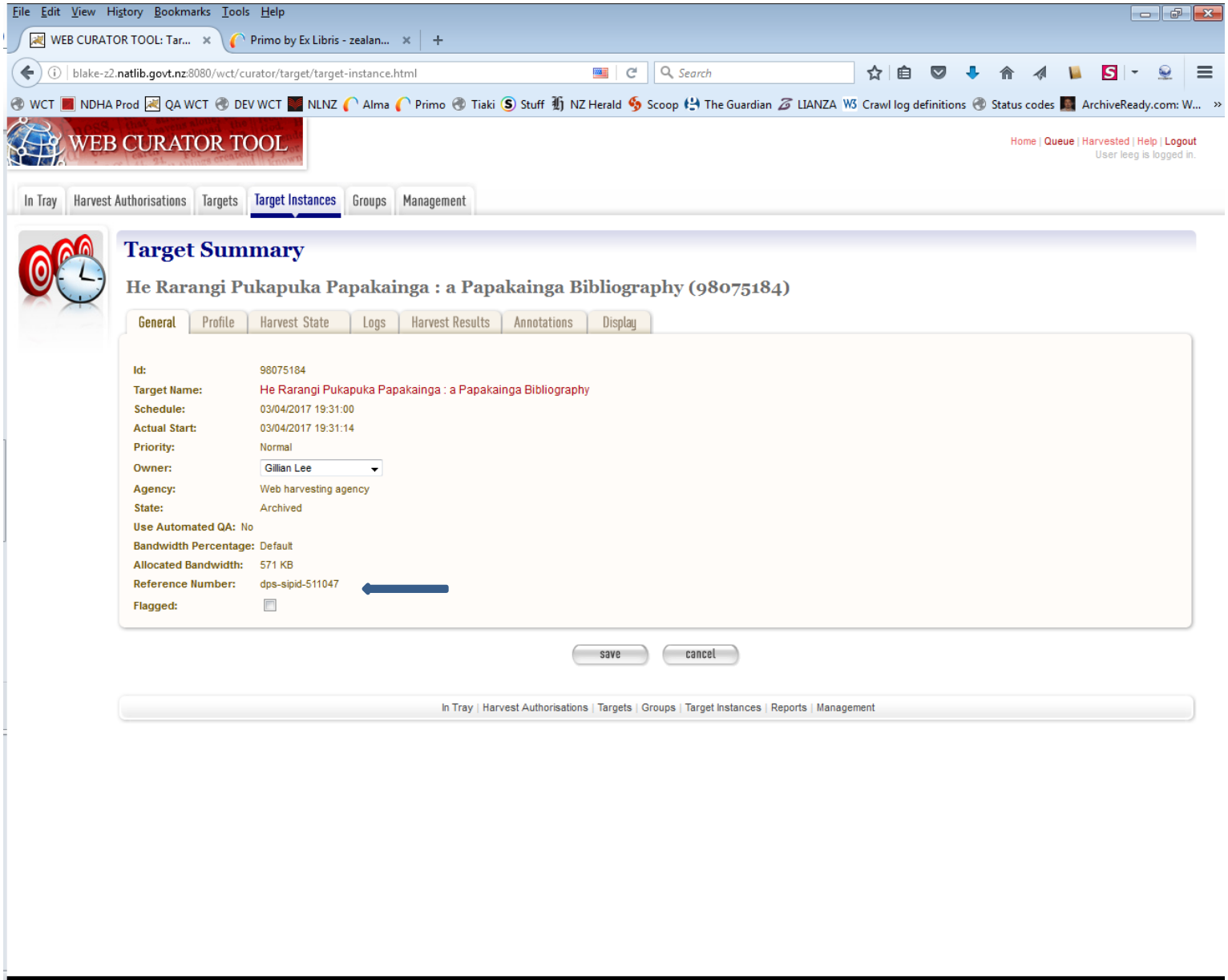
open

General workflow example

The following diagram illustrates a possible flow of authorisations, Targets, and harvests in an institution that requires users to seek permission before initiating any harvests:



DPS-SIP ID in WCT



File Edit View History Bookmarks Tools Help

WEB CURATOR TOOL: Gro... | Primo by Ex Libris - zealan...

natlib-primo.hosted.exlibrisgroup.com/primo_library/libweb/action/search.do?ct=facet&fctN=face

Search

zealand transport agency

All published collections

Advanced Search
Browse Search

RSS
Add page to My Saved Items

Refine My Results

Subject

Author/Creator

Creation Date

Suggested New Searches

by this author/creator:

on this subject:

Results 1 - 10 of **13** for **All published collections**

Sorted by: Relevance

1 2

Refined by: resource type: **Web Sites** x

Website

NZ Transport Agency Waka Kotahi.
NZ Transport Agency.
Wellington, N.Z. : NZTA, 2008
Online access

View It | Details

Open source in a new window — Actions

Open Access

2015 09 29 available at: National Digital Heritage Archive
Open Access

2016 03 29 available at: National Digital Heritage Archive
Open Access

2016 09 29 available at: National Digital Heritage Archive
Open Access

2017 03 29 available at: National Digital Heritage Archive
Open Access

Website

Land Transport NZ Ikiiki Whenua Aotearoa.
New Zealand. Land Transport NZ.
Wellington, N.Z. : Land Transport NZ, 199?
Online access

View It | Details

Website

Feet first walk to school every week = ia wiki hīkoi ki te kura.
NZ Transport Agency.
Wellington, N.Z. : NZ Transport Agency, 2008
Online access

View It | Details

Website

Transit New Zealand Ararau Aotearoa.
Transit New Zealand.
Wellington N.Z. : Transit New Zealand, 199?
Online access

View It | Details

Website

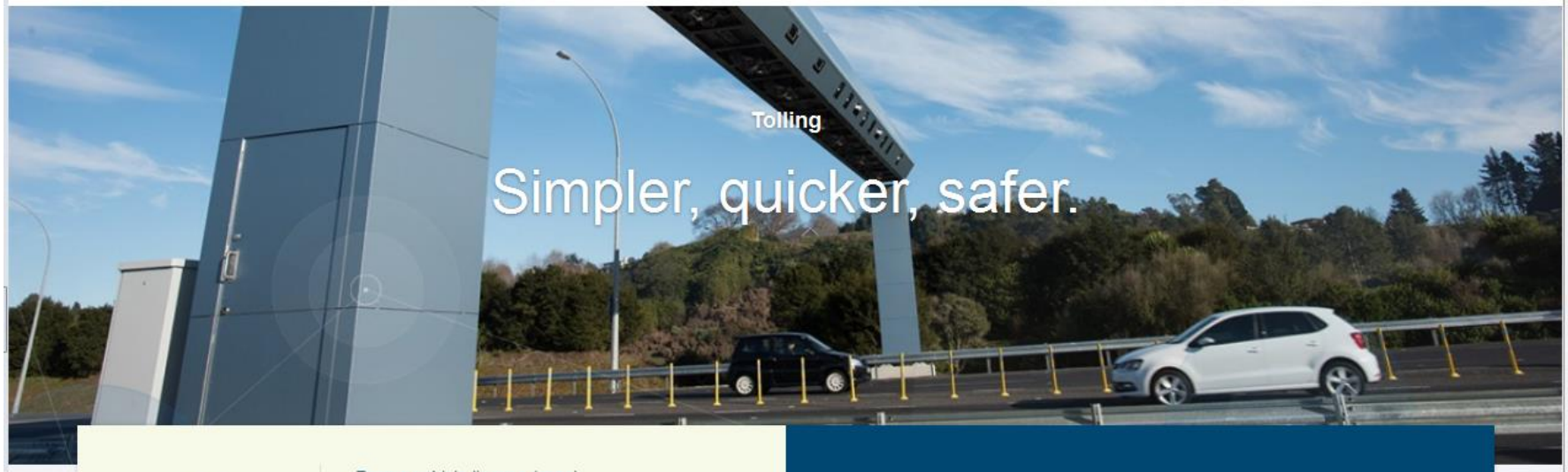
Drive (Web site)
Accident Compensation Corporation (N.Z.),; NZ Transport Agency,
Wellington? : ACC : NZ Transport Agency, 2016]-
Online access

File Edit View History Bookmarks Tools Help
WEB CURATOR TOOL: Gro... | Primo by Ex Libris - zealan... | NLNZ Web Archive Viewer 1.9 -...
ndhadeliver.natlib.govt.nz/ArcAggregator/arcView/frameView/IE28096809/http://www.nzta.govt.nz
Search
Normal copyright restrictions apply.
Search boxes and external links may not function. Some content may not have been captured due to technical limitations.
Harvested by the National Library of New Zealand on: 29 March 2017 at 07:55 PM

NZ TRANSPORT AGENCY
WAKA KOTAHI

What are you looking for?

Menu



Online services

- Renew vehicle licence (rego)
- Book, change or cancel a driving test
- Change your address
- Buy road user charges (RUC)
- Pay a toll
- *More online services...*

Driver licences

Traffic and travel

Vehicles

Tolls, roads and rail

Commercial driving

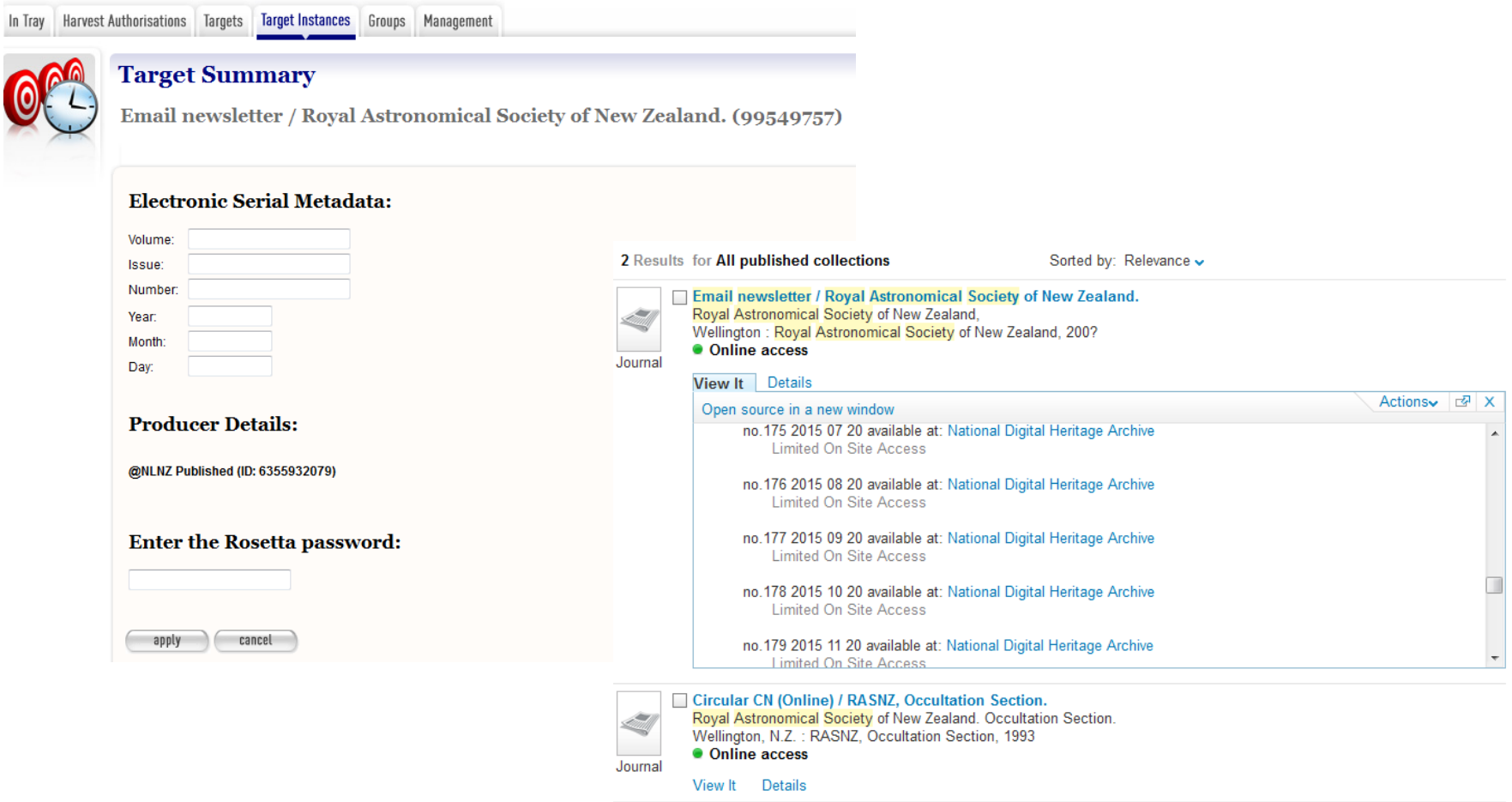
Safety

Walking, cycling and public transport

Planning and investment

Media information

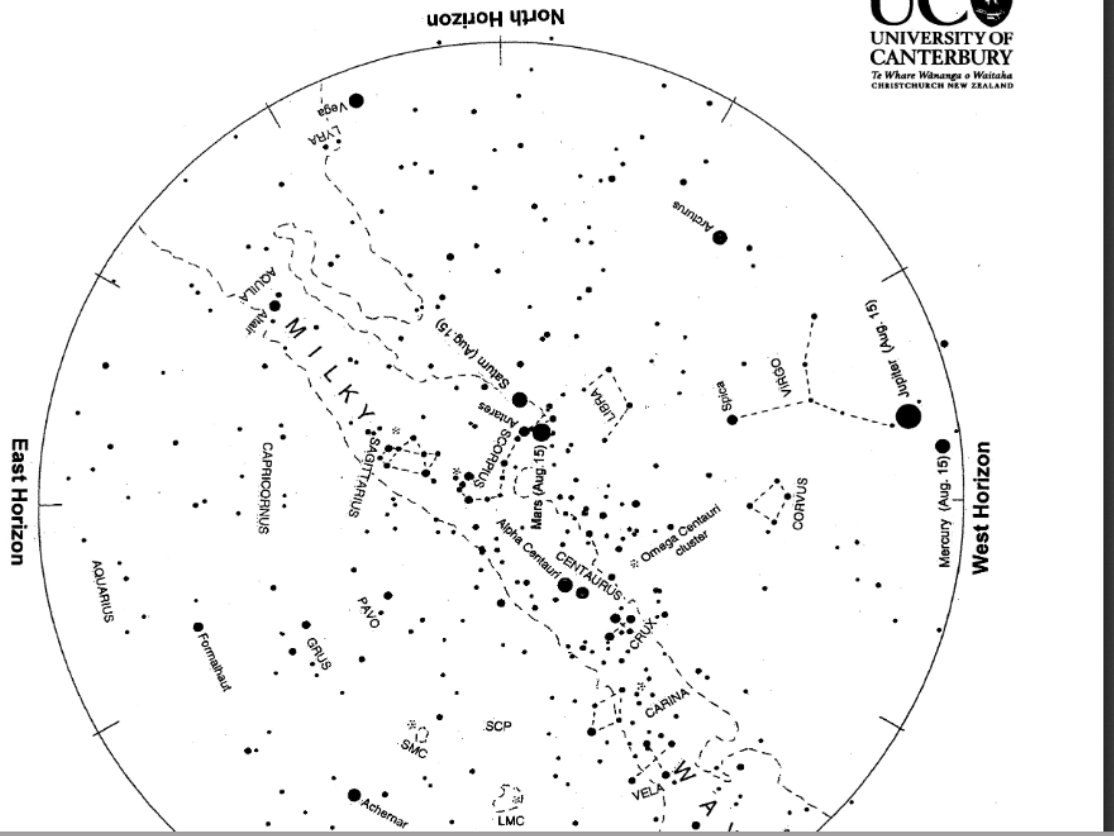
HTML serials



File Edit View History Bookmarks Tools Help
Home Page | DIA Intranet | Primo by Ex Libris - email n... | Email newsletter / Royal Astron...
ndhadeliver.natlib.govt.nz/delivery/DeliveryManagerServlet?dps_pid=IE26803549
WCT NDHA Prod QA WCT DEV WCT NLNZ Alma Primo Tiaki Stuff NZ Herald Scoop The Guardian LIANZA Crawl log definitions Status codes ArchiveReady.com: W...
Te Puna Mātauranga o Aotearoa NATIONAL LIBRARY OF NEW ZEALAND

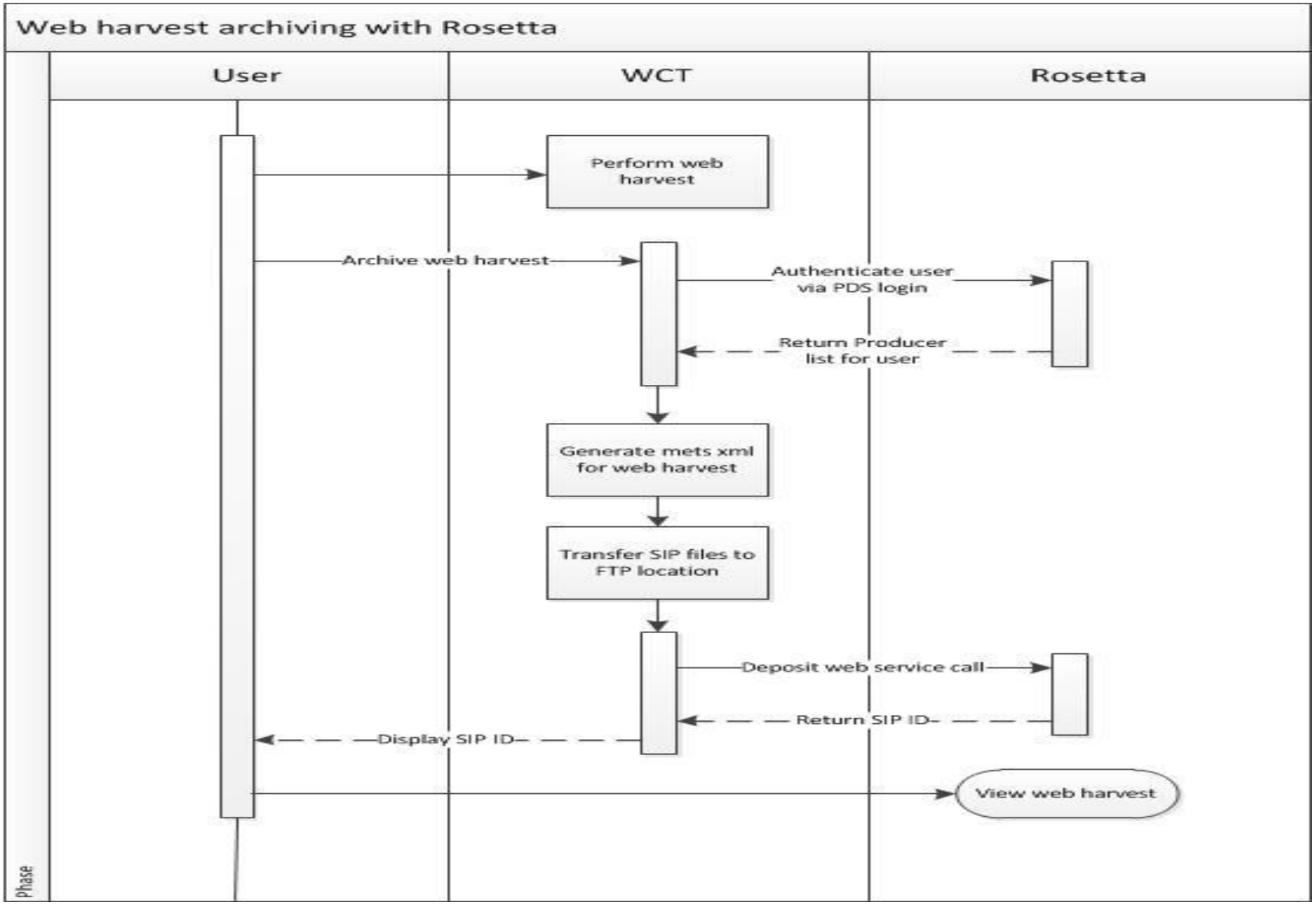
Email newsletter / Royal Astronomical Society of New Zealand.

188
Digital Original
All charts 1608
RASNZ_20_08_2016
Page: 1 of 6 Automatic Zoom



© Ex Libris, a ProQuest Company, 2017
Terms of Use

Integration with Rosetta



Viewers within Rosetta

- **Warc Viewer (Legacy)**
  - One-to-one, Warc file <-> CDX index file
  - Simple to implement
  - Can only handle older warc and cdx formats
  - Struggles to replay a lot of newer websites
  - Live web leakage
- **OpenWayback**
  - More complex implementation
  - Using CDX indexes instead of BDB default
  - Developed and supported by an active community
  - Combined with newer harvesting tools, significantly better replay of web harvests

Challenges

Harvesting

- Websites – more difficult to capture
- WCT – overhaul needed and/or new tools
- Infrastructure

    - systems are slow

    - harvesting at scale problematic

Challenges…

Access

- Improved access is needed:
  - full text indexing of websites
  - Persistent link to content on archived websites (not just at IE level)
  - ongoing development of web archive viewers

Collection Policy

- Traditional scope - comprehensive collecting of NZ and Pacific material
- Sheer volume of web content – prioritisation

Web Archiving and Rosetta

Thank you – any questions?

Steve Knight – Programme Director Preservation Research & Consultancy, National Library of New Zealand

With thanks to Ben O'Brien (Digital Preservation Web Engineer) and Gillian Lee (Coordinator Web Archives) for putting this presentation together.