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’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
Integration with Rosetta
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
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?**
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
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
Neuerscheinung am ZMSBw:

» Weiterlesen

» Weiterlesen

Weitere Neuigkeiten aus dem ZMSBw

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


Einsatz ohne Krieg? Militär, Gesellschaft und Territorialer Wandel der Geschichte vor dem Hintergrund der DDR

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-bvb2-0000130212

Mehr Informationen zum Zitieren von Websites.
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
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?
Challenges (II)

• Improving Access to web archives?
  • Full text indexing of websites
  • Visualization

• Quality control

• Resources?
Perspectives

• Integration of JHOVE WARC Module?

• Integrated viewer for web archives in Rosetta?

• Emulation?

• Legal challenges
  • How to implement Access Rights?

• Preservation Planning
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

Number of HTML serials and monographs
- 280 HTML titles
- 3,825 HTML serial issues (195.5 GB)
- 19 HTML monographs (1.7 GB)
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
HTML serials
# Integration with Rosetta

## Web harvest archiving with Rosetta

<table>
<thead>
<tr>
<th>User</th>
<th>WCT</th>
<th>Rosetta</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Perform web harvest</td>
<td></td>
</tr>
<tr>
<td>Archive web harvest</td>
<td>Authenticate user via PDS login</td>
<td>Return Producer list for user</td>
</tr>
<tr>
<td>Generate mets xml for web harvest</td>
<td>Transfer SIP files to FTP location</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Deposit web service call</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Display SIP ID</td>
<td>Return SIP ID</td>
</tr>
<tr>
<td></td>
<td></td>
<td>View web harvest</td>
</tr>
</tbody>
</table>
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.