Taming the Web: Transparent Management and Appraisal of Web Archives

A State Archives Perspective

Jamie Patrick-Burns
Digital Archivist, State Archives of North Carolina
August 18, 2018
Appraisal in the State Library/State Archives Web Archiving Program

1. First Step: Appraisal
2. Ongoing Work: Quality Control
3. Periodic Evaluation: Audit
4. Ongoing/Emerging Issues

1. First Step: Appraisal
Collecting Mandate

• **General Statute 121**: Conduct a records management program “to preserve and administer, in the North Carolina State Archives, such public records as may be accepted into its custody.”

• **General Statute 125**: “The State Library shall be the official, complete, and permanent depository for all State publications”

• **General Statute 132**: “’Public record’ or ‘public records’ shall mean all documents, papers, letters, maps, books, photographs, films, sound recordings, magnetic or other tapes, electronic data-processing records, artifacts, or other documentary material, regardless of physical form or characteristics, made or received pursuant to law or ordinance in connection with the transaction of public business by any agency of North Carolina government or its subdivisions”
Web Archiving Scope

• Sites that contain official state government information: websites, blogs, social media

• DO NOT collect websites of local governments and public colleges and universities

• DO NOT collect websites of private organizations that assist state agencies
Web Archiving Tool Decisions

**ARCHIVE-IT**
- No limit on number of seeds, and not worried about data budget
- All websites
- All social media

**ArchiveSocial**
- Only social media
- Limited number of accounts
- Collects direct messages
- Ask agencies to rank top 3-4 social media accounts
Workflow

Agency contacts records analyst, who forwards request

Agency contacts member of web archiving group directly

Web archiving team consults

Web archiving group or analyst identifies a website to be crawled

Seed added to tool and tracking spreadsheet
2. Ongoing work: Quality Control
Approach for each tool

**ARCHIVE-IT**
- Download crawl reports
- Review data
- Check pages in Wayback machine
- Identify issues; test constraints/rules

**ArchiveSocial**
- Monitor number of active accounts
- Contact account administrators if not being captured
Archive-It Quality Control: Current Approach

- Top and bottom 5% of new data and new documents
- Review all errors
- View seed and associated reports for problems
- Record actions taken
Current Approach: Pros and Cons

Advantages
• Reviewing seeds most likely to have errors

Disadvantages
• Recurring errors
• Missing errors
• Time consuming
Archive-It Quality Control: New Approach

- Divide seed list and check a section at a time → each seed checked annually
- Start master list of problems to track recurring issues and share solutions
- “High priority” seeds to check more frequently?
New Approach: Pros and Cons

Advantages
• Each seed reviewed annually to catch less obvious problems
• Better communication of problems and solutions
• (Hopefully) time savings

Disadvantages
• To be determined
3. Periodic Evaluation: Audit
Audit Goals

• Reviewing all active and inactive seeds in Archive-It
• Update redirects or inactive websites
• Reappraisal by records analysts and library clearing house
• Updating/improving metadata
Audit

- November – December 2017
- 5 staff evaluated 214 seeds each
- Included active and inactive seeds (all seeds in Archive-It)
- Identified overlap between Archive-It and ArchiveSocial
- Information gathered:

<table>
<thead>
<tr>
<th>Agency</th>
<th>Active or Historical</th>
<th>Redirected</th>
</tr>
</thead>
<tbody>
<tr>
<td>Frequency</td>
<td>Public Partnership</td>
<td>Redirect URL</td>
</tr>
<tr>
<td>In ArchiveSocial</td>
<td>Page still exists</td>
<td>Notes</td>
</tr>
</tbody>
</table>
Audit Results

- Many old websites still crawled
- Duplicate seeds
- Lots of redirects
- Constraint/scoping errors
Audit Next Steps

• Phase 1: organize seeds (February 2018)
• Phase 2: further evaluation of seeds (March 2018)
• Phase 3: easy cleanup (May 2018)
• Phase 4: appraisal/further cleanup (ongoing)
## Phase 1

<table>
<thead>
<tr>
<th>Active Seeds</th>
<th>Inactive Seeds</th>
<th>Active Seeds that should be Inactive</th>
</tr>
</thead>
</table>
| • All seeds marked “active” in Archive-It that still exist  
• Includes seeds that may be captured by another seed  
• Includes seeds that may need further investigation or modification | • Seeds marked “inactive” in Archive-It | • URL no longer exists on the live web  
• Seed exists, but has not been updated in the past 3-5 years  
• Seed redirects to a completely different URL |
Phase 2

- Seeds sorted by agency and reviewed for duplication
- Staff got approximately 109 seeds to review

<table>
<thead>
<tr>
<th>Administration (9)</th>
<th>Correction (1)</th>
<th>Justice (11)</th>
<th>Public Safety (15)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Administrative Office of the Courts (4)</td>
<td>Environmental Quality (51)</td>
<td>Labor (4)</td>
<td>Revenue (2)</td>
</tr>
<tr>
<td>Agency Unknown (19)</td>
<td>General Assembly (2)</td>
<td>Lieutenant Governor (3)</td>
<td>Secretary of State (5)</td>
</tr>
<tr>
<td>Agriculture (22)</td>
<td>Governor (21)</td>
<td>Natural and Cultural Resources (98)</td>
<td>State Auditor (2)</td>
</tr>
<tr>
<td>Boards and Commissions (80)</td>
<td>Health and Human Services (44)</td>
<td>Office of State Budget and Management (2)</td>
<td>State Comptroller (2)</td>
</tr>
<tr>
<td>Commerce (28)</td>
<td>Information Technology (8)</td>
<td>Office of State Human Resources (3)</td>
<td>Treasurer (11)</td>
</tr>
<tr>
<td>Community Colleges (14)</td>
<td>Insurance (6)</td>
<td>Public Instruction (60)</td>
<td>Transportation (18)</td>
</tr>
</tbody>
</table>
Phase 3

- Addressing all Active and Active-to-Inactive seed URLs
- Fixing easy redirects (http → https)
- Identifying next steps for other seeds:
  - Needs additional constraints
  - Inactive (second review)
  - Seed already captured by another
  - Seed needs URL substantially changed
  - No action needed
Phase 4

• Send the final list to analysts and clearinghouse for their appraisal
  • From their perspective, is the list up-to-date and complete?
  • Do any need to be deactivated?
  • Are we missing any important websites?

• Will meet as a group and complete final cleanup as identified in the spreadsheet
Thinking Ahead

• Review ArchiveSocial accounts
• Update metadata and accompanying procedures
• Reviewing and updating scope rules
4. Ongoing/Emerging Issues
Maturation of a Web Archive

- Scalability
  - Quality control approach (MPLP)
- ArchiveSocial
- Communication with stakeholders
  - Records analysts
  - Agencies
Questions?
Contact

Jamie Patrick-Burns
Digital Archivist, State Archives of North Carolina
Jamie.patrickburns@ncdcr.gov
(919) 807-7355
State Archives Twitter: @NCArchives
State Archives Facebook: https://www.facebook.com/State-Archives-of-North-Carolina-119904548024750/