Describing the Library of Congress Web Archives: A Recent History

Rick Fitzgerald
Librarian
Acquisitions & Bibliographic Access Division

Grace Thomas
Digital Collections Specialist
Web Archiving Team
History of the Web Archiving Program

• 2000
• Selective event and thematic crawling
• Permissions based on country of publication + category of site content
• **Digiboard** (curatorial tool), in-house development - manages thematic collections, sites to crawl, and permissions
• Public access at [www.loc.gov/websites/collections](http://www.loc.gov/websites/collections)
• 17,400+ MODS descriptive records currently available
• 1.9+ Petabytes (PB) of data in the archive
• Some jargon:
  • “crawl” = imposing software on a website to copy and archive its web objects
  • “seed” = the URL at which a crawl starts; can be a homepage, section of site, single document
LC Web Archives Descriptive Records
Descriptive records for web archives at LC 2000-2002

- Earliest collections (Election 2000, September 11th): browse lists of URLs
- Election 2002: item-level description in home-grown XML schema
LC began using MODS to describe at the individual archived websites in 2004, concurrent with the development of the MODS schema itself (we continue to use MODS today).

MARC records for the ILS only at the collection level, point to collection pages at [www.loc.gov/websites/collections](http://www.loc.gov/websites/collections).
2009 Rebrand to “Library of Congress Web Archives (LCWA)"
Early Descriptive Model

- Data represented in 3 separate system layers:
  - **Digiboard**: internal; curatorial tool (initially called “Leaderboard”)
  - **Minerva/LCWA**: public; discovery layer for web archives (represented in MODS)
  - **Wayback Machine**: public; access mechanism to browse snapshots of web archives

- One-to-one-to-one-to-one-to-one relationship between:
  seed URL, collection, descriptive record, and accesspoint

Issues with the Early Model

1. Duplication of effort within and across systems
2. Fragmented results for users
3. The Web is living
4. Changing collecting patterns
# 1. Duplication of Effort & 2. Fragmented Results

<table>
<thead>
<tr>
<th>A</th>
<th>B</th>
<th>C</th>
<th>D</th>
<th>E</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td></td>
<td>title</td>
<td>url_x</td>
<td>genre</td>
</tr>
<tr>
<td>2</td>
<td>379</td>
<td>Democratic Party (CT)</td>
<td><a href="http://www.dems.info/index.php">www.dems.info/index.php</a></td>
<td>Political Parties</td>
</tr>
<tr>
<td>3</td>
<td>2029</td>
<td>Connecticut Democrats</td>
<td>dems.info/</td>
<td></td>
</tr>
<tr>
<td>4</td>
<td>2512</td>
<td>Connecticut Democrats</td>
<td>dems.info/</td>
<td></td>
</tr>
<tr>
<td>5</td>
<td>3577</td>
<td>Connecticut Democrats</td>
<td>dems.info/</td>
<td></td>
</tr>
<tr>
<td>7</td>
<td>129</td>
<td>Florida Department of State Division of Elections</td>
<td>election.dos.state.fl.us/</td>
<td>Government</td>
</tr>
<tr>
<td>8</td>
<td>799</td>
<td>Florida Department of State - Division of Elections</td>
<td>election.dos.state.fl.us/</td>
<td></td>
</tr>
<tr>
<td>9</td>
<td>3106</td>
<td>Florida Department of State - Division of Elections</td>
<td>election.dos.state.fl.us/</td>
<td></td>
</tr>
<tr>
<td>10</td>
<td>3862</td>
<td>Florida Department of State - Division of Elections</td>
<td>election.dos.state.fl.us/</td>
<td>Election 2006 Web Archive</td>
</tr>
<tr>
<td>12</td>
<td>2970</td>
<td>Election Projection: 2008 Elections - Polls, Projections, Results</td>
<td>electionprojection.com/</td>
<td></td>
</tr>
<tr>
<td>14</td>
<td>438</td>
<td>Wisconsin, State Elections Board</td>
<td>elections.state.wi.us/</td>
<td>Government</td>
</tr>
<tr>
<td>15</td>
<td>1188</td>
<td>Wisconsin State Elections Board</td>
<td>elections.state.wi.us/</td>
<td></td>
</tr>
<tr>
<td>16</td>
<td>1985</td>
<td>Wisconsin State Elections Board</td>
<td>elections.state.wi.us/</td>
<td></td>
</tr>
<tr>
<td>17</td>
<td>2931</td>
<td>Wisconsin State Elections Board</td>
<td>elections.state.wi.us/</td>
<td></td>
</tr>
<tr>
<td>18</td>
<td>2119</td>
<td>Constitutionalist Party</td>
<td>home.earthlink.net/~jmarkets/cop.html</td>
<td>Political Parties</td>
</tr>
<tr>
<td>19</td>
<td>3637</td>
<td>Constitutionalist Party</td>
<td>home.earthlink.net/~jmarkets/cop.html</td>
<td></td>
</tr>
<tr>
<td>20</td>
<td>3834</td>
<td>Constitutionalist Party</td>
<td>home.earthlink.net/~jmarkets/cop.html</td>
<td></td>
</tr>
<tr>
<td>21</td>
<td>387</td>
<td>PUERTO RICO - Partido Independentista Puertorriqueño</td>
<td><a href="http://www.independencia.net/">www.independencia.net/</a></td>
<td></td>
</tr>
<tr>
<td>22</td>
<td>388</td>
<td>PUERTO RICO - Partido Independentista Puertorriqueño</td>
<td><a href="http://www.independencia.net/">www.independencia.net/</a></td>
<td></td>
</tr>
<tr>
<td>23</td>
<td>2946</td>
<td>PUERTO RICO - Partido Independentista Puertorriqueño</td>
<td>Independencia.net/</td>
<td></td>
</tr>
<tr>
<td>24</td>
<td>2947</td>
<td>PUERTO RICO - Partido Independentista Puertorriqueño</td>
<td>Independencia.net/</td>
<td></td>
</tr>
</tbody>
</table>
3. The Web is Living

Collecting multiple interrelated URLs over time:
• seed URLs change over time, AND/OR
• multiple seeds can describe one entity, AND/OR
• related social media, AND/OR
• all of the above and more...

<p>| | | |</p>
<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>178</td>
<td>Vermont</td>
<td><a href="http://www.state.vt.us">www.state.vt.us</a></td>
</tr>
<tr>
<td>798</td>
<td>Vermont.gov Home Page</td>
<td>vermont.gov/</td>
</tr>
<tr>
<td>160</td>
<td>Vermont.gov Home Page</td>
<td>vermont.gov/</td>
</tr>
</tbody>
</table>
4. Changing Collecting Patterns

Ability to accommodate a site:

1. Within a time-bound, event-based collection (norm at the time, ex. “Papal Transition 2005”)
2. Added to an ongoing, broader thematic collection (newly emerging practice at the time, ex. “Public Policy Topics” ~2009)
3. Ongoing collection of a website which doesn’t fit into an established collection (ex. “Single Sites” ~2009)

...simultaneously.
2011-12: Discovery layer infrastructure upgrade

Opportunity to fold records from standalone interface into loc.gov, the Library’s cross-format, cross-collection discovery layer
Descriptive records for web archives at LC: 2012-present

- Idea for a new model to group related URLs, and have that group attached to multiple collections simultaneously
- Reducing duplicative effort in the curation tool, including better management of permissions
- Creating a single record in the LC discovery layer which points to all captures for a URL group in the Wayback
- Handling the messy web and equally messy collecting patterns
Entity Model
New Representation to Match the Entity Model

**Digiboard** - full overhaul and import of data from original MODS

**Wayback** - all folded into one /all/ accesspoint

**MODS** - tapping into hierarchical features & utilizing flexibility of the schema to represent the entity model

  `<relatedItem type=“constituent”/>` for each seed URL

**loc.gov** - fitting our data into the digital resources model across loc.gov “item-resource” model

  *Bib record is the item (entity) - individual digital object is the resource (our seeds)*
Digiboard example: Legal Planet blog, seed URL changing over time
Digiboard Example: The White House whitehouse.gov
loc.gov example: The White House

whitehouse.gov

www.loc.gov/item/lcwaN0002688
All captures of whitehouse.gov on one accesspoint

www.webarchive.loc.gov/all/*/www.whitehouse.gov
We’ve solved representation in all systems, now what?

Migrate the data into its new representation in Digiboard!

During the migration process, library selectors continued nominating new sites for archiving, so not only did we have the original seeds to handle, that list continued growing in the interim.

For a variety of reasons, we weren’t able to release or update collections for several years.

Collections were released one by one after manual cataloguing, which became unwieldy with the scale of the backlog.

Especially since entities began overlapping among event-based and ongoing collections...
ARCHIVED WEBSITE

White House

View Captures
Captured from December 27, 1996 to May 3, 2018

About this Item
Title
White House

Summary
None

Created / Published
United States:

Genre
web site

Form
electronic

Repository
Library of Congress, Washington, D.C., 20540 USA

Source URL
https://www.whitehouse.gov/

Part of
Crisis in East 2006 Web Archive (241)
African and Middle Eastern (274)
Iraq War 2003 Web Archive (342)
African and Middle Eastern Division (1,892)
Public Policy Topics Web Archive (2,360)
Researcher and Reference Services Division (10,380)

Format
Archived Web Site

Dates
1999 to 2016

www.loc.gov/item/lcwaN0002688
Minimal-level MODS Records

• The element set was decided in part to align with what could easily be exported from existing fields in Digiboard, with minimal intervention
• Minimal data element set is consistent with similar profiles in MARC/RDA - records lack LCSH, NAF headings

  • Entity ID
  • Entity title
  • Country
  • Abstract, if available
  • Display status
  • Seed URL(s)
  • Seed ID(s)

  • ISO Language Code(s) (three characters)
  • Collection(s)
  • Department(s)
  • Surt(s)
  • Subject(s) (available from the limited Digiboard subjects list, not LCSH)
Details of the Backlog Project

- 4,000+ web archives records out of embargo and available to go online
- ~ 40 collections
- Manual cleaning of data in Digiboard - entity titles, review of permissions elements, merging/splitting of entities as appropriate
- Minimal record approach (does NOT include LCSH, NAF, Alt. Titles) - with data pulled automatically from Digiboard using a MySQL query & Python scripts
- Spot checking, but otherwise limited QA - QA to be done by divisions when adding collection-level metadata, if they desire enhancements
- First release in a “rolling release and updates” model
“Sidecar” mini-MODS records for hand-curated data

- LCSH
- NAF
- Alternate Titles
- Abstracts

```xml
<mods version="3.4" xmlns="http://www.loc.gov/mods/v3" xmlns:xlink="http://www.w3.org/1999/xlink"
  <identifier>lcn0006268</identifier>
  <name authority="naf" type="corporate">
    <valueURI>http://id.loc.gov/authorities/names/no94025215.html</valueURI>
    <namePart>Academia Brasileira de Literatura de Cordel</namePart>
  </name>
  <subject authority="lccn" valueURI="http://id.loc.gov/authorities/subjects/sh85022567.html">
    <topic>Chapbooks, Brazilian</topic>
  </subject>
  <subject authority="lccn" valueURI="http://id.loc.gov/authorities/subjects/sh85103704.html">
    <topic>Poetry</topic>
  </subject>
  <subject authority="lccn" valueURI="http://id.loc.gov/authorities/subjects/sh75128013.html">
    <topic>Brazil</topic>
  </subject>
  <subject authority="lccn" valueURI="http://id.loc.gov/authorities/subjects/sh2010107543.html">
    <topic>Popular literature</topic>
    <geographic>Brazil</geographic>
  </subject>
  <recordInfo>
    <recordContentSource authority="marcorg" dlc=""></recordContentSource>
    <recordCreationDate encoding="iso8601">20190419</recordCreationDate>
    <recordIdentifier source="dlc">lcn0006268</recordIdentifier>
    <languageOfCataloging>
      <languageTerm authority="iso639-2b" type="code">cng</languageTerm>
    </languageOfCataloging>
  </recordInfo>
</mods>
```
Future Rolling Releases & Updates

• Once web archives content is out of the 1-year embargo, we need to automatically create a MODS XML file for the web archive and make it available online
• This will increase visibility to web archives content and allow for better searchability online
• This will increase engagement with selecting librarians & site owners because they’ll be able to see their nominated content sooner
• The records available online will better represent the massive amount of content within the web archives
• We also hope this visibility will encourage users and researchers to dive deeper into the data, possibly using computational analysis, once they know the collection exits
Thank you!

Rick Fitzgerald
rfit@loc.gov

Grace Thomas
grth@loc.gov

Web Archiving Team
webcapture@loc.gov

www.loc.gov/programs/web-archiving/about-this-program/

www.loc.gov/websites/collections