The United States End of Term Web Archive

Abbie Grotke, Library of Congress
Kathleen Murray, University of North Texas Libraries
Building the 2008 Archive

- Background
- 2008 Project & Nomination of URLs
- Demo of public interface
- Data transfer
- Preparing for Access
Collaborating Institutions

- Library of Congress
- Internet Archive
- California Digital Library
- University of North Texas
- US Government Printing Office
Why Archive .gov? Why Collaborate?

- Fit with partner missions to collect and preserve at-risk (born-digital) government information
- Potential for High Research Use/Interest in Archives
- It Takes a Village
- Experienced Partners
Project Goals


- Document federal agencies’ presence on the Web during the transition of Presidential administrations.

- To enhance the existing research collections of the five partner institutions.
URL Nomination Tool

- Facilitates collaboration
- Ingest seed lists from different sources
- Record known metadata
  - Branch
  - Title
  - Comment
  - Who nominated
- Create seed lists for crawls
Volunteer Nominators

- Call for volunteers targeted:
  - Government information specialists
  - Librarians
  - Political and social science researchers
  - Academics
  - Web archivists

- 31 individuals signed up to help
Nominator To-Dos

- Nominate the most critical URLs for capture as "in scope"
- Add new URLs not already included in the list
- Mark irrelevant or obsolete sites as "out of scope"
- Add minimal URL metadata such as site title, agency, etc.
In Scope vs. Out of Scope

- **In scope**: Federal government Web sites (.gov, .mil, etc.) in the Legislative, Executive, or Judicial branches of government. Of particular interest for prioritization were sites likely to change dramatically or disappear during the transition of government.

- **Out of scope**: Local or state government Web sites, or any other site not part of the above federal government domain.

- **Not captured**: intranets, deep web content.
Prioritized URLs

- ~500 URLs nominated by volunteers
Selected Researcher/Curator Interests

- Homeland Security
- Department of Labor
- Department of Treasury
- Education/“No Child Left Behind”
- Health Care Reform
- Stem Cell Research
- Bush Administration Budget Justifications
- Federal Program Assessments (ExpectMore.gov)
## Crawl Schedule

### Two Approaches:
- Broad, comprehensive crawls
- Prioritized, selective crawls

### Key dates:
- Election Day, November 4
- Inauguration Day, January 20
Results

Between September 2008 and November 2009:

- Over 3,000 “sites” archived (or is it 4,622?*)
- 160 million files/documents
- Over 15 TB of data

*counting is tricky: sites are loosely defined
Data Transfer

- Goal: Distribute 15.9 TB of collected content among partners

- LC’s central transfer server used:
  - “Pulled” and “pushed” data from and to partners via Internet2, May 2009 – Mid 2010
  - Common transfer tools, specifications were key

More info here: http://blogs.loc.gov/digitalpreservation/2011/07/the-end-of-term-was-only-the-beginning/
Transfer Tools: Bagger
Preparing for Access

1st Tuesday of each month, 12:00 pm:

Anything to report on public access?

No, nothing to report on public access.
Internet Archive also had:

- A full copy of the content from all EOT partners
- A QA “Playback” tool (takes screen images of archived materials)
- An export of the Nomination Tool metadata from UNT
- MODS record extractor tool
CDL had:

1. **Author:** Alfred Greene, Photographer  
   **Title:** Murray Public School (1935), photograph  
   **Published:** 1935  
   **Subjects:** School children -- California -- Dublin | Teachers -- California -- Dublin | Schools -- California -- Dublin | Students -- California -- Dublin | Collection of the Dublin Heritage Museum

2. **Author:** Unknown  
   **Title:** Elizabeth Flanagan Nevin (1901-1902), photograph  
   **Published:** 1902  
   **Subjects:** Teachers -- California -- Dublin | Nevin, Elizabeth Flanagan | Collection of the Dublin Heritage Museum
Caveats

- As with any web archive, the crawler is good, but not always perfect!

- Full-text index of 16 TB of data
  - Some behaviors designed to help rank and navigate such a large body of content
The End of Term Web Archive documents the United States Government's World Wide Web presence during the transition between the administrations of President George W. Bush and President Barack Obama.
Full text search

Archival Search Sample XSLT

This simple XSLT demonstrates the transformation of OpenSearch XML results into a fully-functional, human-friendly HTML search page. No JSP needed.

Search for: no child left behind  
Search  HTML  PDF  MS Word

Results 1-10 of about 540,034 (2.041 seconds)

1. No Child Left Behind - ED.gov
   No Child Left Behind - ED.gov Advanced Search Students Parents Teachers Administrators NCLB Overview Stronger Accountability More Local Freedom Proven Methods Choices for Parents A-Z Index About ED Budget & Performance Press Room Publications Teaching Resources FAQs Contact Help Jobs at ED Online Services Recursos en español State Information Web Survey Features A-Z Index Find your way around the NCLB section. GO > NCLB Policy Policy documents on accountability, choice, SES, teacher quality, and more. GO > NCLB State Status See the status of your state’s efforts to achieve NCLB goals. GO... Program NCLB Policy Letters No Child Left Behind Programs By Subject By Title By CPDB# Search News... Contacts Budget Annual Reports and Plans Jobs at ED Inspector General No FEAR Act Data Site Policies...
   http://www.ed.gov/index/landing.html  35k  text/html

2. Friday, January 12, 2007 - Texas Times: Reinforcing No Child Left Behind
   Friday, January 12, 2007 - Texas Times: Reinforcing No Child Left Behind...
   All versions [1]  More from senate.gov

3. President Bush Discusses No Child Left Behind
   President Bush Discusses No Child Left Behind Skip Main Navigation PRESIDENT | VICE PRESIDENT... Audio Photos En Español Fact Sheet: The No Child Left Behind Act: Challenging Students... objective, and the No Child Left Behind Act was all part of making sure that we get it right in the schools... No Child Left Behind says, look, we trust the local folks. I don‘t want Washington, D.C. running... the education system functions for all. And that’s the spirit of No Child Left Behind. By measuring... things that I think is most important about the No Child Left Behind Act is that when you measure. . unfair to the children. And the No Child Left Behind Act demands result for every child, for the good... time to reauthorize the No Child Left Behind Act, my attitude is, instead of softening No Child Left...
   All versions [1]  More from whitehouse.gov

4. No Child Left Behind
   No Child Left Behind Skip Main Navigation PRESIDENT | VICE PRESIDENT | FIRST LADY | MRS. CHENEY... Education So That No Child is Left Behind As America enters the 21st Century full of hope and... met. In America, no child should be left behind. Every child should be educated to his or her full potential, establishing annual assessments in grades 3-8, within two years of enacting this plan. "No Child Left Behind... disadvantaged students will be recognized and rewarded with "No Child Left Behind" bonuses... closing the achievement gap will be honored with awards from a "No Child Left Behind" school... eligible to receive a one-time bonus for their efforts, which will be used to support educational programs in their school. As the years passed, the system will continue to improve and the gap will continue to close. As the years passed, the system will continue to improve and the gap will continue to close.
Site List

End of Term Web Archive

Site List

Site Lookup:
- Library of congress

Government branch:
- Legislative (5)
- Executive (1)

URL segment:
The following text appears in a section of the site's URL. This can help you narrow results to particular agencies:
- loc (7)
- thomas (2)
- catalog (1)
- digitization (1)
- memory (1)
- rss (1)
- use (1)

Title: Library of Congress Home
Archival URL: http://web.archive.org/web/20080916003635/loc.gov/index.html
IA Site value:
Live URL: http://www.loc.gov/index.html
Description: The Library of Congress is the nation's oldest federal cultural institution, and it serves as the research arm of Congress. It is also the largest library in the world, with more than 120 million items. The collections include books, sound recordings, motion pictures, photographs, maps, and manuscripts.

Title: Library of Congress Online Catalogs (Upgraded May 19, 2008)
Archival URL: http://web.archive.org/web/20080916003635/catalog.loc.gov/
IA Site value:
Live URL: http://catalog.loc.gov/
Coverage: September 15, 2008 - May 8, 2009
Description: The Library of Congress collections include over 110 million items in a variety of formats and languages. The catalog information for many of these items has appeared in a number of traditional card catalogs located in the Library. Much of the information in these catalogs is also searchable over the Internet. This page links to the two methods for search the main Library of Congress Online Catalog and to several libraries...
Browse by Image

- Internet Archive tools for visualizing web archived data ("explore data")
Forthcoming: Tag cloud extracted from metadata
Classification of the End-of-Term (EOT) Archive: *Extending Collection Development Practices to Web Archives*

University of North Texas Libraries | IMLS National Leadership Grant

- Project Background
- Archive Classification
  - SMEs: SuDocs Classification Scheme
  - Link Analysis: Web graph
    - Cluster Analysis
  - SMEs: Cluster Tagging
- Conclusion
Problem
- The absence of descriptive metadata or classification schemes thwarts discovery & access
- WARC files (ISO 28500)
  - Specifies formats needed for storage, management, and exchange of data objects (or resources); Not designed for user access
- Wayback access
  - Need to know a resource’s URL

Objective: Classify materials in accord with the Superintendent of Documents (SuDocs) Classification Numbering System

Outcome: Enable librarians to utilize existing selection practices to identify materials in the EOT Archive
## Classification: Size Challenge

<table>
<thead>
<tr>
<th>Domains</th>
<th># URLss</th>
<th>Subdomains</th>
</tr>
</thead>
<tbody>
<tr>
<td>gov</td>
<td>137,847,822</td>
<td>14,339</td>
</tr>
<tr>
<td>mil</td>
<td>3,555,425</td>
<td>1,677</td>
</tr>
</tbody>
</table>

**SURTS: Reordering URLs by domain structure**

**URL:**  http://marriagecalculator.acf.hhs.gov/marriage/

**SURT:** http://(gov,hhs,acf,marriagecalculator,)

**Domain** 1 2 3

---

**Unique 1st Level Subdomains = 1,647**

**After validation = 1,151 Subdomains**
Human Classification

- SuDocs Classification Scheme
- 10 SMEs classified 1,151 URLs (230/SME)
  - 70% agreement ($n = 808$)
    - Unable to classify: 18 – in scope; 36 – out of scope
  - 30% disagreement ($n = 343$)
- 3 arbitrators classified 343 URLs
  - Assigned SuDocs authors to 286 URLs
  - Unable to classify: 42 – in scope; 15 – out of scope
Classification: Findings

- Overall, SuDocs Scheme worked well
- Assigned SuDocs authors to 1,040 subdomains
  - 1,111 authors (1,040 + 71 multiply authored sites)
- Major Classification Challenge
  - Determining primary author among multiple authors
- Weaknesses
  - Lacks sufficient granularity for subordinate agencies
Link Analysis

- Web graph
  - Identified # of outlinks and inlinks for each URL

- Subdomains
  - 1,151 1st level subdomains within .gov & .mil domains
  - Multiple URLs per subdomain

- Explored cluster analysis algorithms
  - Best result: Linlog Coordinates with Agglomerative Hierarchical Clustering
Cluster Analysis

- Set limit on number of clusters to identify
  - First analysis: Set of 55 clusters
  - Second analysis: Set of 75 clusters

<table>
<thead>
<tr>
<th></th>
<th>Cluster Analysis</th>
<th>Clusters 55–24 &amp; 75–31</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>55–Set</td>
<td>75–Set</td>
</tr>
<tr>
<td>39</td>
<td>Identical</td>
<td>39</td>
</tr>
<tr>
<td>16</td>
<td>13 / 2</td>
<td>36</td>
</tr>
<tr>
<td></td>
<td>2 / 3</td>
<td></td>
</tr>
<tr>
<td></td>
<td>1 / 4</td>
<td></td>
</tr>
</tbody>
</table>

Identical Subdomains
- fdic.gov
- fdicconnect.gov
- fdicig.gov
- fdicoig.gov
- fdicseguro.gov
- myfdicinsurance.gov
- egrpra.gov
Findings: SuDoc Classification

SuDoc Parents
- 1  16%
- ≤ 2  27-32%
- ≤ 4  60-67%

# Clusters

# SuDoc Parent Authors per Cluster
Topical Evaluation of Clusters

- Total of 130 clusters tagged (55 + 75)
  - 12 SMEs: Each cluster tagged by 3 SMEs
    - 52 Clusters tagged 3 times
    - 39 Clusters tagged 6 times

<table>
<thead>
<tr>
<th>Cluster Analysis</th>
<th>55–Set</th>
<th>75–Set</th>
</tr>
</thead>
<tbody>
<tr>
<td>39</td>
<td>Identical</td>
<td>39</td>
</tr>
<tr>
<td>16</td>
<td>Unique</td>
<td>36</td>
</tr>
</tbody>
</table>
Tag Analysis

- How topically related are the tags?
- Two researchers independently assigned “relatedness category” (RC)
  - RC 1 = little or no relation
  - RC 2 = somewhat related
  - RC 3 = strongly related

Cluster 55–19

<table>
<thead>
<tr>
<th>Cluster 55–19</th>
<th>SME 40</th>
<th>SME 32</th>
<th>SME 42</th>
</tr>
</thead>
</table>
| RC 3          | • federal regulations  
                • administrative law | • federal regulations | • federal regulations |

- 2 Subdomains
  - federalregister.gov
  - fedreg.gov
Findings: Topical Evaluation

- Relatedness Categories ($N = 130$)
  - RC 1 = little or no relation ($n = 27; 21\%$)
  - RC 2 = somewhat related ($n = 24; 18\%$)
  - RC 3 = strongly related ($n = 79; 61\%$)

<table>
<thead>
<tr>
<th>Clusters</th>
<th>RC 1</th>
<th>RC 2</th>
<th>RC 3</th>
</tr>
</thead>
<tbody>
<tr>
<td>130</td>
<td>21%</td>
<td>18%</td>
<td>61%</td>
</tr>
<tr>
<td>75-Set</td>
<td>21%</td>
<td>17%</td>
<td>61%</td>
</tr>
<tr>
<td>55-Set</td>
<td>20%</td>
<td>20%</td>
<td>60%</td>
</tr>
</tbody>
</table>

- Cluster Analysis successfully identified strongly related subject content in the subdomains of 61\% of clusters
## Impact of Increasing # of Clusters

<table>
<thead>
<tr>
<th>Clusters</th>
<th>#</th>
<th>RC 1</th>
<th>RC 2</th>
<th>RC 3</th>
</tr>
</thead>
<tbody>
<tr>
<td>Identical</td>
<td>39</td>
<td>18%</td>
<td>10%</td>
<td>72%</td>
</tr>
<tr>
<td>All</td>
<td>130</td>
<td>21%</td>
<td>18%</td>
<td>61%</td>
</tr>
<tr>
<td>Unique in 75–Set</td>
<td>36</td>
<td>22%</td>
<td>14%</td>
<td>64%</td>
</tr>
<tr>
<td>Unique in 55–Set</td>
<td>16</td>
<td>25%</td>
<td>31%</td>
<td>44%</td>
</tr>
</tbody>
</table>

- Identical clusters had the highest percentage of topically related subdomains (72%)
- Unique clusters had a substantially higher percentage of topically related subdomains after subdivision (64% v. 44%)
Cluster analysis was reasonably effective at organizing the EOT Archive.

**SuDoc Parent Authors**
- \( \leq 2 \) 32%
- \( \leq 4 \) 67%

**Content Relatedness**
- RC 3 61%

![Graph showing distribution of SuDoc Parent Authors per cluster.](image)
Selection of Materials in Web Archives

PROBLEM:
- Absence of descriptive metadata or classification schemes thwarts discovery & access; URL must be known

RESULT:
- Cluster analysis holds promise for organizing Web archives into topically related groupings
- Involving SMEs in limited-scope classification activities may generate meaningful descriptive metadata for resources in focused Web archives
What’s Next

- Using the Web graph
  - How do we leverage the graph for identifying content?
- Describing the collection
  - How can we engage faculty with our Web archives?
- Identifying change
  - How is the .gov Web changing over time?
- Full-text search
  - What other improvements to Web archive search can be made?
Timeframe for 2012 project

2012

- **Summer 2012**: Recruitment of curators/nominators to help identify additional websites for prioritized crawling.
- **July/August 2012**: Bookend (baseline) crawl of government web domains begins.
- **Summer/Fall 2012**: Partners will crawl various aspects of government domains at varying frequencies, depending on selection policies/interests. Team will determine strategy for crawling prioritized websites.
- **November – February 2012–13**: Crawl of prioritized websites.

2013

- **January 2013**: Depending on the outcome of the election, focused crawls will be conducted as needed during this period.
- **Spring or Summer 2013**: Bookend crawl, plus additional crawl of prioritized websites as determined by team.
What You Can Do to Help

- Help nominate URLs for 2012:
  - Any nominations welcome, any amount of time you can contribute
  - Need particular help with:
    - Judicial Branch websites
    - Important content or subdomains on very large websites (such as NASA.gov) that might be related to current Presidential policies
    - Government content on non-government domains (.com, .edu, etc.)
- Contact eotproject@loc.gov to sign up
Questions?
eotproject@loc.gov
Follow us on twitter! @eotarchive
http://eotarchive.cdlib.org/