Preserving Public Government Information: The End of Term Web Archive

Abbie Grotke, Library of Congress
Tracy Seneca, California Digital Library

END OF TERM ARCHIVE – FDLC Oct. 17, 2011

Outline

- Background
- Nomination of URLs
- Data transfer
- Preparing for Access
- Demo of public interface
- 2012: what you can do
- Our questions for you / your questions for us!

Themes

- Ad-hoc nature of this project
- Wellspring of web archiving tools
- Testbed for emerging tools

Collaborating Institutions

- Library of Congress
- Internet Archive
- California Digital Library
- University of North Texas
- US Government Printing Office

Why Archive .gov? Why Collaborate?

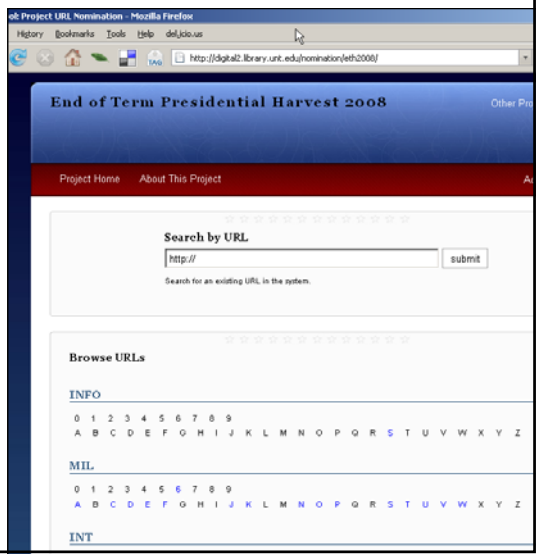
- Fit with partner missions to collect and preserve at-risk (born-digital) government information
- Potential for High Research Use/Interest in Archives
- It Takes a Village
- Experienced Partners

Project Goals

- Work collaboratively to preserve public U.S. Government Web sites at the end of the current presidential administration ending January 19, 2009.
- 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

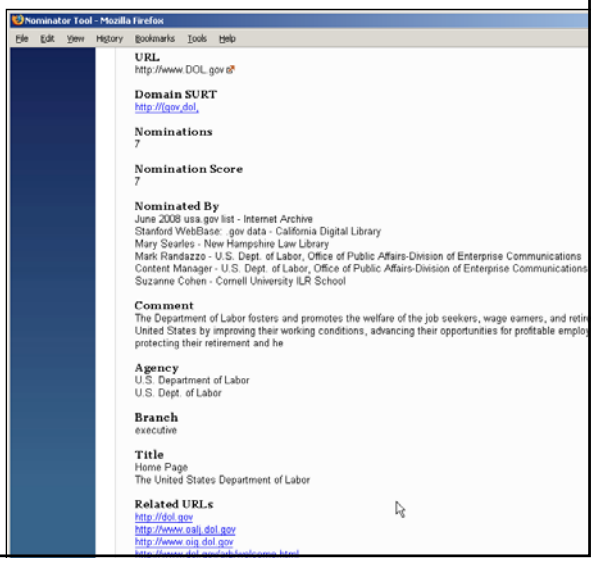


Volunteer Nominators

- 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)

Nomination Tool Lessons Learned

- No coordination of selection or "assignments" so likely gaps in collection
- Start with a blank slate rather than pre-populate the database?
- Admin tools/Reporting features need more work
- Engage more experts to help identify at-risk content

Further Lessons Learned

- CDL: hard to respond to nominations to shape crawler settings & focus – defaulted to getting as much as we could.
- When used for Deepwater Horizon, we found it added an extra task for curators – used "delicious" instead.
- The metadata we did get (branch, description) was **really valuable** after the fact!

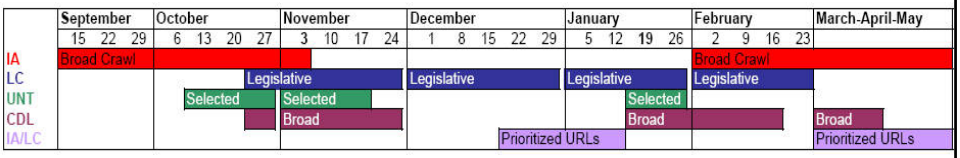
Questions for you:

- What do YOU need from the nomination tool?
- Is it nomination or curation?

Partner Harvesting Roles

- **Internet Archive** – Broad, comprehensive harvests
- **Library of Congress** – In-depth Legislative branch crawls
- **University of North Texas** – Sites/Agencies that meet current UNT interests, e.g. environmental policy, and collections, as well as several "deep web" sites
- **California Digital Library** – Multiple crawls of all seeds in EOT database; sites of interest to their curators
- **Government Printing Office** – Support and analysis of "official documents" found in collection

Crawl Schedule



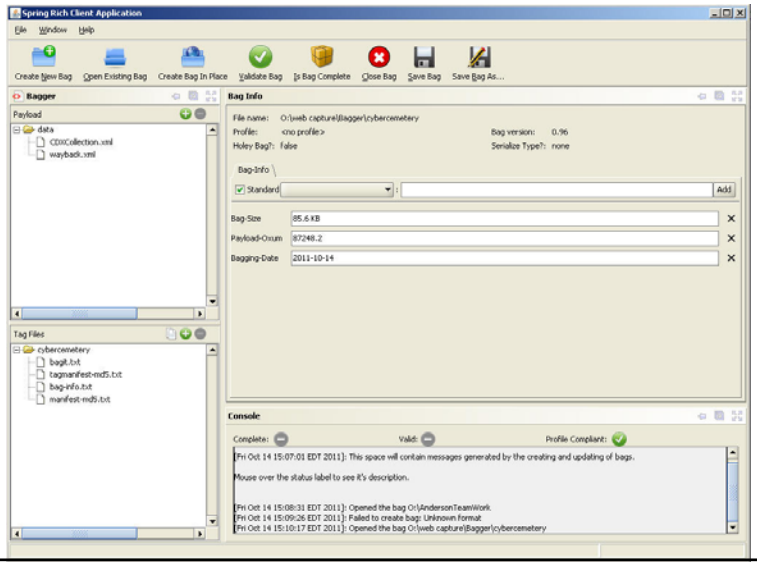
- Two Approaches:
  - Broad, comprehensive crawls
  - Prioritized, selective crawls
- Key dates:
  - Election Day, November 4
  - Inauguration Day, January 20

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:

MODS to DC

```
<title>Visualization of Remote Sensing Data</title>
<creator>Agency name information if you have it</creator>
<identifier>http://crawls-wm.us.archive.org/eot08/*/rsd.gsfc.nasa.gov/</identifer>
<provenance>http://rsd.gsfc.nasa.gov/</provenance>
<date>2008-09-16</date>
<date>2009-08-14</date>
<description>The Vis/RSD website is a showcase for stunning visualizations of satellite data by NASA's Goddard Laboratory for Atmospheres.</description>
<subject>remote sensing</subject>
<subject>data</subject>
<subject>satellite</subject>
<subject>radar</subject>
<subject>telescope</subject>
<subject>earth</subject>
<subject>space</subject>
<subject>land</subject>
<subject>oceans</subject>
<subject>science education</subject>
<subject>land use</subject>
<subject>NASA</subject>
<type>web site</type>
<format>text</format>
<coverage>Executive</coverage>
<relation> http://crawls.archive.org/collections/eot08/</relation>
```

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

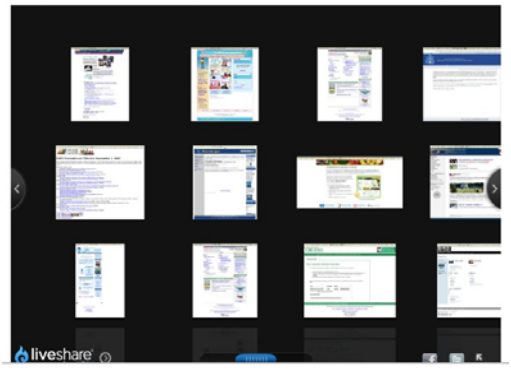
Demo of Beta Interface

Access gateway lessons

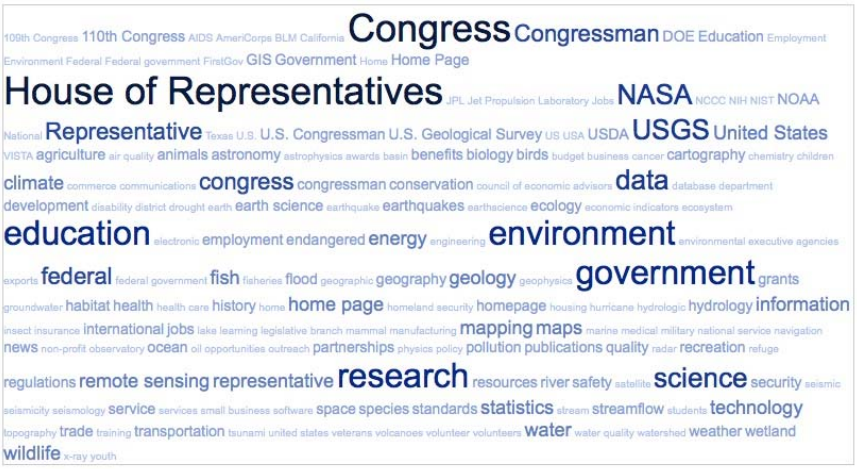
- You could easily use this to integrate:
  - materials from multiple web archives, no matter where.
  - any digital content, whether harvested or scanned.

Forthcoming

- Internet Archive tools for visualizing web archived data ("explore data")



Tag cloud extracted from metadata



More questions to you:

- How might you use this archive?
  - Rediscovering, providing continued access to 'lost' documents?
  - Researching, visualizing trends in government data?

What you can do

- Provide feedback on the Beta site
- 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.)

Timeframe for 2012 project

2012

- **January 2012**: Project team will begin accepting early nominations of priority websites via the Nomination Tool.
- **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 polices/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.

Questions?

eotproject@loc.gov

Follow us on twitter!
@eotarchive

http://eotarchive.cdlib.org/