Preserving Public Government Information: The 2008 End of Term Crawl Project

Abbie Grotke, Library of Congress
Mark Phillips, University of North Texas Libraries
George Barnum, U.S. Government Printing Office

CNI Fall Task Force Meeting
December 9, 2008
Outline

- Project Goals and History
- Nomination of URLs
- Partner Activities
- Future Work
Project Goals

- Document federal agencies’ presence on the Web during the transition of Presidential administrations.
- To enhance the existing collections of the five partner institutions.
Project History

- **Collaborating Institutions:**
  - Library of Congress
  - Internet Archive
  - California Digital Library
  - University of North Texas
  - US Government Printing Office
Project History

- First Meeting – Canberra, Australia
  - Early April 2008, at the National Library of Australia, International Internet Preservation Consortium (IIPC) – formed the partnership and discussed implications and possible roles for each institution.
  - Agreed from the beginning to share all content with any partner who wished a copy.
Monthly meetings since that time – conference calls and one face-to-face meeting.

- Defined roles
- Released an announcement
- Sought help from specialists to nominate URLs for harvesting
- Shared technology planning
- Developed URL Nomination Tool
Pause for Vocabulary

- **Seed List** – List of URLs fed to the crawler for harvesting.
- **Crawler** – Software which downloads file, parses text to extract URLs, adds URLs to list and repeats.
- **Scope** – Whether a URL should be included or not.
- **Crawl** – Running a crawler on a given seed list.
- **SURT** – Sort-able URL Reversible Transformation.
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 are 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.
AN AMERICAN MOMENT

The story of this campaign is your story. It is about the great things we can do when we come together around a common purpose.

We want to hear your inspiring stories from the campaign and Election Day.

WATCH THE SPEECH

Tell Us Your Story
Change.gov – really a .gov?

- Content copyright © 2008 by Obama–Biden Transition Project, a 501c(4) organization.
- Search of .gov whois says domain was registered by GSA (General Services Administration)
- Other official transition sites:
  - [http://www.opm.gov/transition/TRANS20R–Ch1.htm](http://www.opm.gov/transition/TRANS20R–Ch1.htm)
  - [http://directory.presidentaltransition.gov/](http://directory.presidentaltransition.gov/)
Tool Building

- **URL Nomination Tool**
  - Allows for combining multiple seed lists
  - Allows for collaboration with subject experts
  - Helps create future seed lists
  - Helps to define overall scope of project
URL Nomination Tool

- Allows for collaboration with subject experts
- Ingest seed lists from different sources
- Record known metadata for seed
  - Branch
  - Title
  - Comment
  - Who nominated
- Allow people to help with nomination
  - Search
  - Browse
  - “easy to use”
- Create seed lists for crawls
Tool Concepts

- **URL** – Single instance of metadata in system
  - URL
  - Attribute – (metadata element)
  - Value – (metadata value)
  - Nominator ID
  - Project ID
  - Timestamp

- **Nominator**
  - Nominator Email Address
  - Nominator Name
  - Nominator Institution

- **Project**
  - Project Metadata
List of URLs

1890scholars.program.usda.gov
2001.cancer.gov
2001.nci.nih.gov
acc.nos.noaa.gov
access.usgs.gov
access.wa.gov
accessamerica.gov
accesstospace.gsfc.nasa.gov
acrim.jpl.nasa.gov
acs.oes.ca.gov
acweb.fsl.noaa.gov
adc.gsfc.nasa.gov
List of SURTs

gov.accessamerica
gov.ca.oes.acs
gov.cancer.2001
gov.nasa.gsfc.acessstospace
gov.nasa.gsfc.adc
gov.nasa.jpl.acrim
gov.nih.nci.2001
gov.noaa.fls.acweb
gov.noaa.nos.acc
gov.usda.program.1890scholars
gov.usgs.access
gov.wa.access
Back to the tool…

- **Tool Requirements**
  - Ingest seed lists from different sources
  - Keep track of who nominated seed
  - Record known metadata for seed
  - Allow people to help with nomination
    - Search
    - Browse
    - “easy to use”
  - Create seed lists for crawls
Batch Ingest

- Administrator can import csv files with URLs and associated metadata with batch importer

- An ingest needs to be associated with a Nominator and a Project.

- Arbitrary metadata is recognized and added to the system.
Nomination – In Scope/Out of Scope

- On batch import a URL is given a positive nomination +1

- A user of the Nomination Tool has the ability to nominate a URL as in scope (+1) or out of scope (-1)

- Nominations are calculated to give a possible measure of importance for a project.
EOT 2008 Project Metadata

- Metadata fields defined for EOT 2008 Harvest
  - Branch
  - Title
  - Comment

- Nominators don’t need to register but Name, Email and Institution are required.
End of Term Presidential Harvest 2008

Search by URL

http://

Submit

Search for an existing URL in the system.

Browse URLs

INFO

0 1 2 3 4 5 6 7 8 9
A B C D E F G H I J K L M N O P Q R S T U V W X Y Z

MIL

0 1 2 3 4 5 6 7 8 9
A B C D E F G H I J K L M N O P Q R S T U V W X Y Z

INT
End of Term Presidential Harvest 2008

Browse URLs that begin with W

Browse gov URLs

0 1 2 3 4 5 6 7 8 9
A B C D E F G H I J K L M N O P Q R S T U V W X Y Z

URL
http://wapa.gov
http://www.wapa.gov
http://www.washingtondc.gov
http://waterways.rdi.gov
http://www.wcatwc.gov
http://wdol.gov
http://weather.gov
http://web-services.gov
http://webcontent.gov
http://westnilevirus.gov
http://wethepeople.gov
http://whicoa.gov
http://www.whicoa.gov
http://whitehouse.gov
http://www.whitehouse.gov
http://www.whitehouse.gov/appointments
http://www.whitehouse.gov/ask
http://www.whitehouse.gov/ces
http://www.whitehouse.gov/pres
End of Term Presidential Harvest 2008

URL
http://www.whitehouse.gov

Domain SURT
http://gov.whitehouse

Nominations
2

Nomination Score
2

Nominated By
Stanford WebBase: .gov data - California Digital Library
June 2008 usa.gov list - Internet Archive

Branch
executive

Related URLs
http://whitehouse.gov
http://www.whitehouse.gov/appointments
http://www.whitehouse.gov/ask
http://www.whitehouse.gov/cease
http://www.whitehouse.gov/dpc
http://www.whitehouse.gov/whofc
Scope (Required)

Scope of URL:
- In Scope
- Out of Scope

Optional Metadata

Branch of Government:
- Executive
- Judicial
- Legislative

Department/Agency Name:
White House
The name of the departments, agency, bureau, commission, committee, or office represented by this URL

Title:
Welcome to the White House
The title of the webpage represented by this URL

Comment:
Department of Labor Salutes Our Veterans

WASHINGTON, November 6 — Secretary of Labor Elaine L. Chao hosted the Department of Labor’s annual "Salute to Veterans" program, which highlighted the many DOL Programs that assist Veterans and military personnel. Secretary Chao also took this opportunity to applaud the military veterans who are now employees of the Department of Labor for their dedicated service to our nation.

"Today, the men and women of the U.S. military are the finest, most capable, and technologically-advanced fighting force in the world. Each year over 318,000 military members return to civilian life," said Secretary Chao. "There is no more appropriate way to celebrate Veterans Day than to focus on them, and what we can do to help them make a successful transition to civilian life when they return."

- Visit the Veterans’ Employment and Training Service Web site.
- Read the Secretary’s remarks.
- In Focus Archive
- Subscribe to In Focus Updates

Highlights
- U.S. Labor Department resources available for economic recovery
- Department of Labor offering assistance to this season’s hurricane victims
- DOL Issues Union Transparency Final Rule
- DOL distributes $34 million to 19 states for Trade Adjustment Assistance
- Spring/Summer 2008 Edition of The Labor Advocate

Natural Disaster Recovery Assistance
- Notable Results
- Compliance Assistance
  - FAQs — Employment Law Advisors
- Union Reports
- Budget & Performance Information
- Faith-Based & Community Initiatives
- Newsroom
- Events Calendar
- Doing Business With DOL
- Freedom of Information Act (FOIA)
- Subscribe to E-mail Updates

Secretary’s Corner
URL
http://www.DOL.gov

Domain SURT
http://gov.dol

Nominations
7

Nomination Score
7

Nominated By
June 2008 usa.gov list - Internet Archive
Stanford WebBase: .gov data - California Digital Library
Mary Sears - New Hampshire Law Library
Mark Randazzo - U.S. Dept. of Labor, Office of Public Affairs-Division of Enterprise Communications
Content Manager - U.S. Dept. of Labor, Office of Public Affairs-Division of Enterprise Communications
Suzanne Cohen - Cornell University ILR School

Comment
The Department of Labor fosters and promotes the welfare of the job seekers, wage earners, and retirees of the United States by improving their working conditions, advancing their opportunities for profitable employment, protecting their retirement and he

Agency
U.S. Department of Labor
U.S. Dept. of Labor

Branch
executive

Title
Home Page
The United States Department of Labor

Related URLs
http://dol.gov
http://www.oali.dol.gov
http://www.oig.dol.gov
http://www.dol.gov/arb/welcome.html
http://www.dol.gov/cfbcic
http://www.dol.gov/dolcl/organization.htm
http://www.dol.gov/ebsa
http://www.dol.gov/eseehl/index.htm
Server not found

Firefox can't find the server at after-school.gov.

- Check the address for typing errors such as www.example.com instead of www.example.com.
- If you are unable to load any pages, check your computer's network connection.
- If your computer or network is protected by a firewall or proxy, make sure that Firefox is permitted to access the Web.

Try Again
End of Term Presidential Harvest 2008

URL
http://after-school.gov

Domain SURT
http://gov.after-school

Nominations
2

Nomination Score
0

Nominated By
Nara's list from 2004 - Internet Archive
Kent Norworth - University of Texas at Austin

Comment
No DNS as of 10/14/2008

Title
Afterschool.gov
Project Description:

The Library of Congress, the California Digital Library, the University of North Texas Libraries, the Internet Archive, and the U.S. Government Printing Office have joined together for a collaborative project to preserve public United States Government web sites at the end of the current presidential administration ending January 19, 2009. This harvest is intended to document federal agencies' presence on the World Wide Web during the transition of Presidential administrations and to enhance the existing collections of the five partner institutions.

In this collaboration, the partners will structure and execute a comprehensive harvest of the Federal Government .gov domain. The Internet Archive will crawl broadly across the entire .gov domain. The University of North Texas and the California Digital Library, will supplement and extend the broad comprehensive crawl with focused, in-depth crawls based on prioritized lists of URLs. This dual-edged approach seeks to capture a comprehensive snapshot of the Federal government on the Web at the close of the current administration.

For more information about the project and partners, visit: http://www.loc.gov/today/pr/2008/08-139.html

Project Scope:
The project has two phases: A broad, comprehensive baseline crawl of .gov

Project Name:
End of Term Presidential Harvest 2008

Project Abbreviation:
eth2008

Project Starting Date:
Aug 25, 2008

Project Ending Date:
Feb 02, 2009

Nomination Starting Date:
Aug 25, 2008

Nomination Ending Date:
Dec 01, 2008

Admin Name:
Mark E. Phillips

Admin E-mail:
mark.phillips@unt.edu

Project URL:
http://www.loc.gov/today/pr/2008/08-139.html

Registration Required for Project:
No
#This list of urls was created for the End of Term Presidential Harvest 2008 project. #List generated on 2008-12-05 09:10:16Z

http://www.barackobama.com
http://www.begich.com
http://www.bobbarr2008.com/home
http://www.dtiasociates.com/fbc1
http://figures2008.com
http://www.jamesmadison.com
http://www.jeffsessions.com
http://www.johnmccain.com
http://www.pancanal.com
http://www.pryor2008.com
http://www.tedstevens2008.com
http://www.usps.com/communications/organization/csat.htm
http://votetruth08.com
http://www.cga.edu
http://www.dodea.edu
http://www.nd.edu
http://www.si.edu
http://digital-library.usma.edu
http://www.usma.edu
http://www.usna.edu
http://1903to2003.gov
http://1stgov.gov
http://1stgovt.gov
http://25yeaofenergy.gov
http://4girls.gov
http://4parents.gov
http://www.4woman.gov
http://4women.gov
http://www.4women.gov
http://5aday.gov
http://www.5aday.gov
http://9-11commission.gov
http://911commission.gov
http://aapi.gov
http://www.aapi.gov
http://abmc.gov
http://www.abmc.gov
http://access-board-members.gov
http://access-board.gov
http://www.access-board.gov
http://accessacrossamerica.gov
http://accessamerica.gov
Volunteer Nominators

- Call for volunteers at end of August to lists targeting:
  - Government information specialists
  - Librarians
  - Political and social science researchers
  - Academics
  - Web archivists (IIPC, Archive–IT communities)

- 31 individuals signed up to help
Nominator To-Do's

- Nominations based on their interests/expertise
- 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
What Did They Do?

- Deadline was November 30, 2008
- 24 volunteers nominated at least one site or more (including project team)
- 500 URLs nominated in scope or out of scope
Partner 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
Library of Congress crawl plan

- Legislative: Enhance existing, ongoing monthly crawls (congressional, aoc, loc, gao, gpo and other miscellaneous urls) to include prioritized legislative URLs.
- Identified new congressional websites not on house.gov and senate.gov (speaker.gov) during an intensive pre-crawl review.
- Continue monthly crawl, but crawl deeper, longer (one month rather than one week)
- October – February
- Will also support crawl of all prioritized seeds between election and before inauguration
October EOT Crawl

- Approximately doubled the number of documents harvested during the crawl.

- Finding
  - Twitter
  - Youtube
  - Myspace
  - Flickr
UNT crawl plan

- Crawling selected URLs of interest to UNT Libraries, including:
  - FEMA
  - Energy Information Administration
  - Department of Agriculture
  - Homeland Security
  - Office of Faith-Based and Community Initiatives
  - Department of Education
  - Fuel Economy
  - Environmental ethics and policy materials across agencies

- Around the election; pre-inauguration; post-inauguration; possibly one year later snapshots
CDL crawl plan

- Exhaustive crawl of all seeds in nomination tool
- Before election; after election but before inauguration; shortly after inauguration; and six months after inauguration
- Focused crawling of sites of interest to University of California curators
- Using their Web Archive Service
IA crawl plan

- Performed baseline harvest of 2522 seeds from Sept 15, 2008 until Election Day.
- Plan to support interim captures between Election Day and Inauguration Day to support selective harvests.
- Will begin final comprehensive harvest of all URLs nominated and/or visited on Jan. 21, 2009. Will conclude when full scope of materials have been visited. Will only collect new material, i.e. content that was added or that has changed since the baseline harvest.
Near–Future Work

- Centralizing web data into a single collection at the Internet Archive
- Providing WayBack access to content
- Providing search access to content
- Distributing collection among partners (25–35 TB projected)
- Investigation of browse by Agency/Branch
Other Future Work

- Extracting topical collections from crawl data
- Providing programmatic access for data-mining
- Research in calculating “size” of collection in relation to real world measures
  - Number of pages of text collected
  - Number of 8x10 in equivalent images collected
  - Hours of Audio
  - Hours of Video
  - Number of PDFs
  - Physical library space requirements to hold collection if in physical format.
Questions?
eotproject@loc.gov