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

Abbie Grotke, Library of Congress
Kris Carpenter, Internet Archive

DLF
November 13, 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
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.
Project History

- 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
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-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.
- Deadline: November 30, 2008

- 23 volunteers have nominated at least one site or more (including project team)
- 479 URLs nominated in scope or out of scope so far
- Nominations based on their interests/expertise
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.
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
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.
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-id.gov
- http://www.wcalwc.gov
- http://wdoj.gov
- http://weather.gov
- http://web-services.gov
- http://webcontent.gov
- http://westnilevirus.gov
- http://wethpeople.gov
- http://whcoa.gov
- http://www.whcoa.gov
- http://whitehouse.gov
- http://www.whitehouse.gov
- http://www.whitehouse.gov/appointments
- http://www.whitehouse.gov/ask
- http://www.whitehouse.gov/cea
- http://www.whitehouse.gov/policy
End of Term Presidential Harvest 2008

Project Home  About This Project

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 2006 usa.gov list - Internet Archive

Branch
executive

Related URLs
http://whitehouse.gov
http://www.whitehouse.gov/apointments
http://www.whitehouse.gov/ask
http://www.whitehouse.gov/cea
http://www.whitehouse.gov/ceq
http://www.whitehouse.gov/dpc
http://www.whitehouse.gov/hci
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:
Immigration

The Department is responsible for providing immigration-related services and benefits such as naturalization and work authorization as well as investigative and enforcement responsibilities for enforcement of federal immigration laws, customs laws, and air security laws.

I Want to...
- Download the I-9 Form (Employment Eligibility Form)
- Become a U.S. Citizen
- Find immigration forms on USCIS
- Visit the United States
- Make an appointment with an Immigration Information Officer (InfoPass)
- Invite a relative to visit the United States
- Find out the status of my case online
- Report suspicious activity
- Contact the CIS Ombudsman about a long-standing case issue

Activities & Programs
- Immigration Policy
- Secure Border Initiative
- Worksite Enforcement
- Submit a case problem to the CIS Ombudsman
- More Activities & Programs »

Benefits
- Electronic filing for immigration benefits
- Learn about lawful permanent residency
- Immigration forms and fees
- More Benefits »

Immigration Statistics
- Yearbook of immigration statistics
- Profiles on local permanent residents
- Border Apprehensions, 2005 (PDF, 2 pages - 208 KB)
- More Immigration Statistics »

Local Resources
- USCIS field offices addresses and information
- ICE Office of Detention and Removal Operations Offices
- More Local Resources »
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 300,000 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

More Resources:
- The White House
- USA.gov
- GovBenefits.gov
- DisabilityInfo.gov
- HireVetsFirst.gov
- CareerVoyages
- Wirtz Labor Library

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
- e laws — 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
?

Nomination Score
?

Nominated By
June 2008 usa.gov list - Internet Archive
Stanford WebEase: .gov data - California Digital Library
Mary Searles - 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 Chen - 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.cali.dol.gov
http://www.cig.dol.gov
http://www.dol.gov/arb/welcome.html
http://www.dol.gov/cfaci
http://www.dol.gov/dol/organization.htm
Recent Developments

- Approval of proposal by American Express Company and American Express Travel Related Services Company to become bank holding companies
- Federal Reserve will offer $150 billion in 17-day credit through its Term Auction Facility today
- Federal Reserve Board and Treasury Department announce restructuring of financial support to AIG
- Speech by Governor Warsh on the promise and peril of the new financial architecture
- Federal Reserve Banks announce reduced number of check processing sites and accelerated restructuring schedule

Features

Foreclosure Resources for Consumers
If you are having difficulty making your mortgage payment, one of the most important things you can do is seek assistance. The following resources provide information and links to agencies and organizations that may be able to help you.

Federal Reserve Consumer Help
If you have a problem with a bank or other financial institution, contact the Federal Reserve for help.
Welcome to the Select Committee on Energy Independence and Global Warming website. This unique committee was established by Speaker Nancy Pelosi in early 2007 to add urgency and resources to the commitment of this Congress to address the challenges of America's oil dependence and the threat of global warming. Use this website to learn about how these issues are impacting America and the world, and what you can do to help solve these twin challenges.

The Select Committee will be working every day to make a difference, but it will also take concerned citizens to enact the significant changes we need to effectively address two of the most important issues of our time: global warming and energy security.

Edward Markey
Chairman

Global Warming Impact Zones

Welcome!
United States Courts of Appeals

The 94 U.S. judicial districts are organized into 12 regional circuits, each of which has a United States court of appeals. A court of appeals hears appeals from the district courts located within its circuit, as well as appeals from decisions of federal administrative agencies.

In addition, the Court of Appeals for the Federal Circuit has nationwide jurisdiction to hear appeals in specialized cases, such as those involving patent laws and cases decided by the Court of International Trade and the Court of Federal Claims.
Reports to the President

2007

- Report for the President on the Use and Results of Alternative Dispute Resolution in the Executive Branch of the Federal Government
  April 2007

- Letter from the Attorney General to the President, transmitting the Report for the President on the Use and Results of Alternative Dispute Resolution in the Executive Branch of the Federal Government
  April 27, 2007

  April 27, 2007

2000

- Report to the President
  May 8, 2000
“That was the moment I first believed I could go.”

A great education can help you achieve your dreams. You can do it, and here’s how:

- **why go**
  See how education creates more opportunities.

- **what to do**
  Learn how you can make your college dream a reality.

- **how to pay**
  It might be easier than you think.
Pay as you go

thinking points vol. 1, no. 2
on Government Spending

latest from rpc

The Committee’s Failure to Address Fannie/Freddie Failure is a Disgrace

Submitted by Rep. Bill Sali
Tags: Economy, Fannie and Freddie, oversight

I attended yesterday’s Oversight and Government Reform Committee hearing hoping to ask the panel three essential questions: Who knew what and when, and what action did they take?

Today, Idaho’s mom and pop investors are losing their hard-earned savings in their retirement funds, while some corporate CEO’s have received golden parachutes. Idahoans want to know if anyone is going to go to jail for causing this crisis. SEC Chairman Chris Cox indicated that is likely given the large number of prosecutions he has begun. I hope he is correct.

Rep. Bill Sali’s blog Login or register to post comments Read more

Protect Joe the Plumber
Comments: 0
Cents: 3

Oil Prices? Keep ‘em fallin’... Drill here, Drill now.
Comments: 0
Cents: 2

Stop blatant voter fraud
Comments: 1
Cents: 0
“Today we begin in earnest the work of making sure that the world we leave our children is just a little bit better than the one we inhabit today.”

– President-Elect Barack Obama

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.

Obama & Biden in Grant Park
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://directory.presidentialtransition.gov/
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 Norwothy - University of Texas at Austin

Comment
No DNS as of 10/14/2008

Title
Afterschool.gov

Score (Required)
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 congressional crawls (house.gov, senate.gov, committees, etc.) to include any prioritized legislative URLs
- 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
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