



End of Term Web Archive

US Federal Government Websites 2008-2009

The United States End of Term Web Archive

Abbie Grotke, Library of Congress
Kathleen Murray, University of North Texas
Libraries

CNI Project Briefing – April 3, 2012

Building the 2008 Archive

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

Collaborating Institutions

3

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

Why Archive .gov? Why Collaborate?

4

- 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

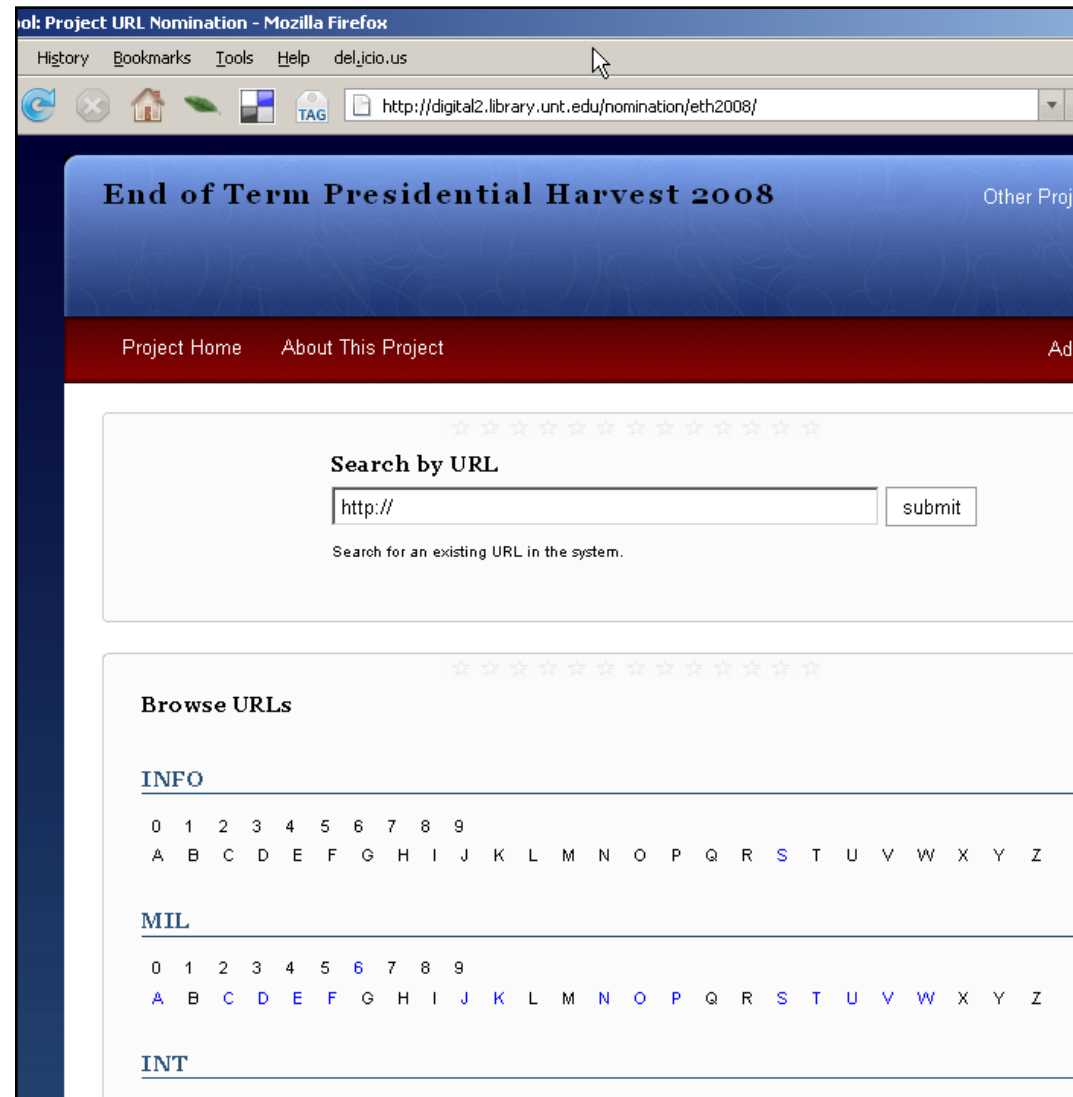
5

- 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

6

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



Volunteer Nominators

7

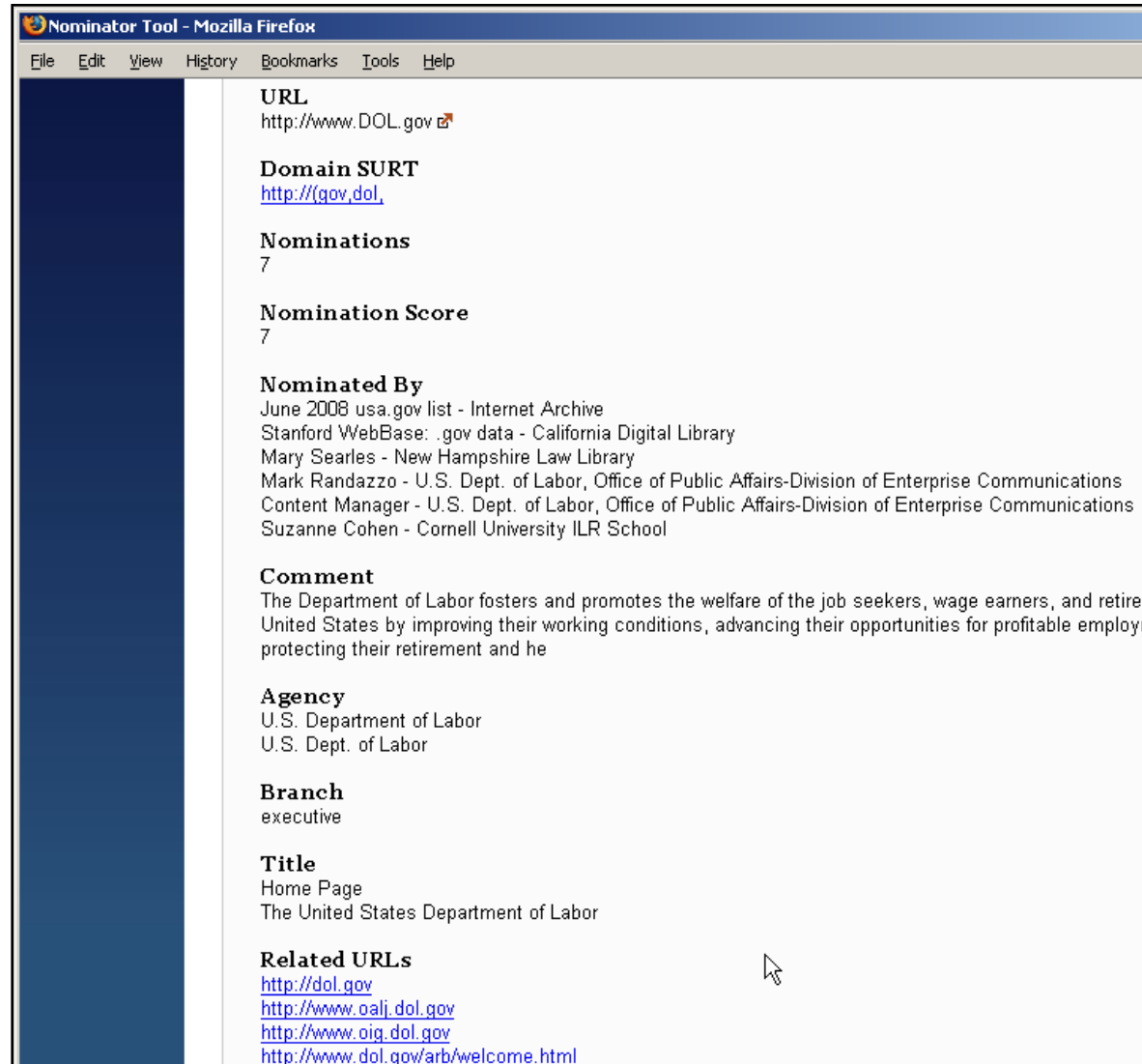
- Call for volunteers targeted:
 - Government information specialists
 - Librarians
 - Political and social science researchers
 - Academics
 - Web archivists

- 31 individuals signed up to help

Nominator To-Dos

8

- 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.



The screenshot shows the Nominator Tool interface in Mozilla Firefox. The browser title is "Nominator Tool - Mozilla Firefox". The menu bar includes "File", "Edit", "View", "History", "Bookmarks", "Tools", and "Help". The main content area displays the following information:

- 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 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 Cohen - Cornell University ILR School
- Comment**: The Department of Labor fosters and promotes the welfare of the job seekers, wage earners, and retire United States by improving their working conditions, advancing their opportunities for profitable employ 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.oalj.dol.gov>
 - <http://www.oig.dol.gov>
 - <http://www.dol.gov/arb/welcome.html>

In Scope vs. Out of Scope

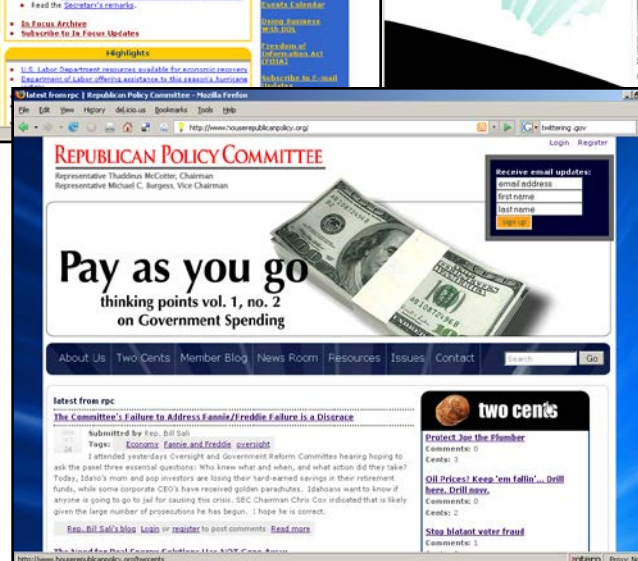
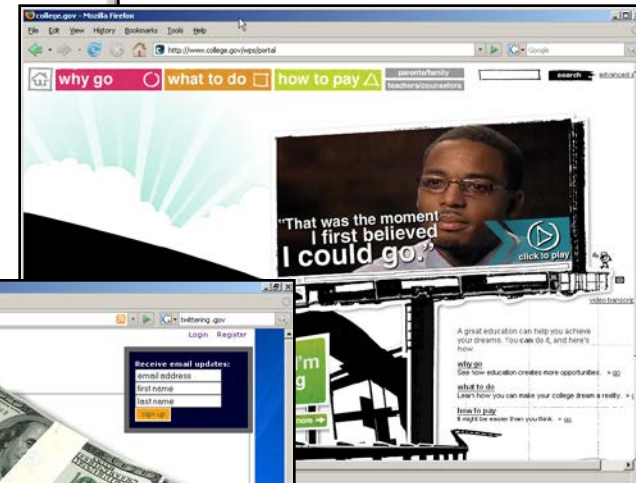
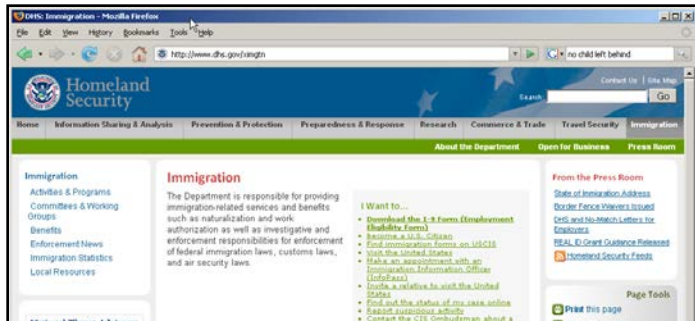
9

- 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

10

□ ~500 URLs nominated by volunteers



Selected Researcher/Curator Interests

11

- 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

12

	September	October	November	December	January	February	March-April-May
	15 22 29	6 13 20 27	3 10 17 24	1 8 15 22 29	5 12 19 26	2 9 16 23	
IA	Broad Crawl					Broad Crawl	
LC		Legislative		Legislative	Legislative	Legislative	
UNT		Selected	Selected		Selected		
CDL			Broad		Broad		Broad
IA/LC				Prioritized URLs			Prioritized URLs

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

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

Results

13

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

```
http://after-school.gov
http://afterschool.gov
http://www.afterschool.gov
http://gears.tucson.ars.ag.gov
http://www.tucson.ars.ag.gov
http://www.uswcl.ars.ag.gov
http://agingstats.gov
http://www.agingstats.gov
http://agoa.gov
http://www.agoa.gov
http://ahcpr.gov
http://www.ahcpr.gov
http://meps.ahrq.gov
http://www.ahrq.gov
http://aids.gov
http://airnow.gov
http://alc.gov
http://amberalert.gov
http://amc.gov
http://amembassy-fiji.gov
http://americaslibrary.gov
http://www.americaslibrary.gov
http://americasoutdoors.gov
http://americasstory.gov
http://americastory.gov
http://americore.gov
http://americorp.gov
http://americorps.gov
http://americorpse.gov
http://www.etd.ameslab.gov
http://www.external.ameslab.gov
http://www.msg.ameslab.gov
http://sc94.ameslab.gov
http://www.scl.ameslab.gov
```

Data Transfer

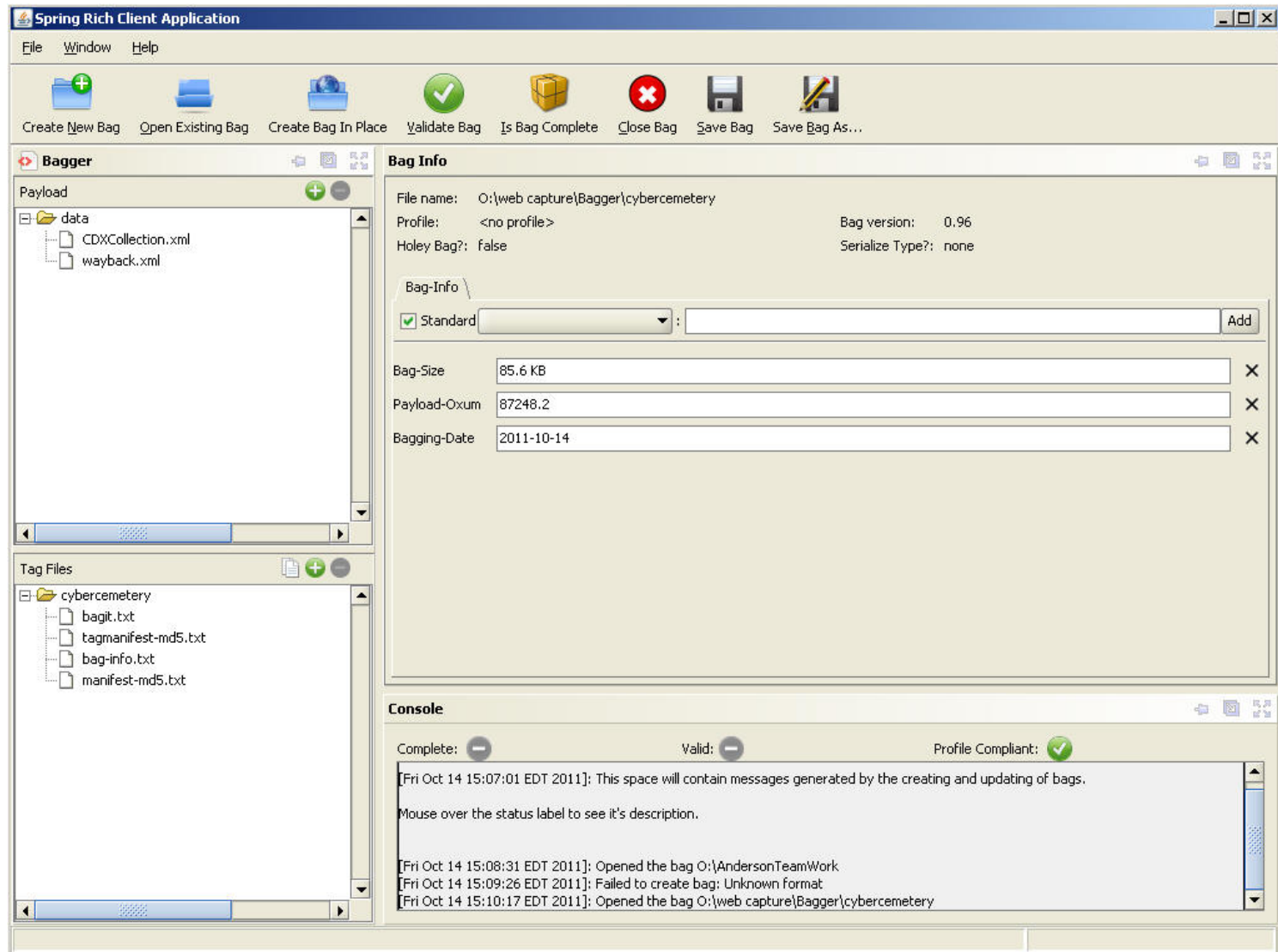
14

- 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

15



Preparing for Access

16

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:

17

- 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:

18

graphic

[Bookbag \(0\)](#)

Browse by: All

Results: 196 Items

Sorted by:

 [RSS](#) | [Modify Search](#) | [New Search](#)

Browse by Facet | [Title](#) | [Author](#)

Page: 1 2 3 4 5 ... [Next](#)

Subject

- [Collection of the Dublin Heritage Museum](#) (196)
- [Schools -- California -- Dublin](#) (39)
- [Pioneer Families of Dublin \(Calif.\)](#) (35)
- [School children -- California -- Dublin](#) (30)
- [Portrait photographs - - California -- Dublin](#) (25)

[more](#)

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](#)

Similar Items: [Find](#)



[Requires cookie*](#)

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](#)

Similar Items: [Find](#)



[Requires cookie*](#)

General Information

Center for African Studies
ias.berkeley.edu/africa

Policies

Submit a Paper

Administrator Login

RSS

Breslauer Symposium on Natural Resource

There are 12 publications in this collection, published between 2004 and 2006


Ryan, Sadie: A spatial location-allocation GIS framework for natural resources in a savanna nature reserve, 2006

Beyene, Zewdineh; Wadley, Ian L.G.: Common goods and the environment: Transboundary natural resources, principled cooperation, and the Millennium Initiative, 2004

Chaudhry, Debajyoti: Privatizing reform at home: Natural resource

University of California

calisphere a world of primary sources and more



view all historical images

Collections for Educators

Themed Collections
Quickly find compelling primary sources that support California history and social studies.

California Cultures
Discover the faces and history of California's diverse populations.

The Letters of John Muir
NEW! Learn about the career of scientist, writer, and activist John Muir through his prolific correspondence.

Browse A-Z
1-9 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

1-9 1906 earthquake 42nd Regimental Central Postal Directory



A Abenathy, Ralph Advertising Aerospace industry Agricultural equipment Agricultural laborers Assembly line Automobile industry

B Bear Flag Republic Bunker, John Black Panthers Business Program Bridges Bridges, Harry

C Central Pacific Railroad Chavez, Cesar Chicano Movement Committee

Selected UC websites

Home | About Us | Privacy Statement | Site Map | State Library

Connect with us  



Home Browse Institutions Browse Collections Browse Map About OAC Help What is OAC?

Welcome to the Online Archive of California

Search OAC go

Browse the Collections

By Title from A to Z
0-9 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

By Institution

- African American Museum and Library at Oakland, Oakland Public Library
- Agua Caliente Cultural Museum
- Alameda County Library, Dublin Library
- Albany Library
- American Jewish University
- Anaheim Public Library
- Arcadia Public Library
- Architecture Collections, Huntington Library
- Archives and Special Collections Department, California State University, Dominguez Hills
- Art Center College of Design Archives
- Arts Library Special Collections, UC Santa Barbara
- Astri National Center Institute for the Study of the American West

Browse Map



Caveats

20

- 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

Beta Interface

21



[Contact Us](#) [Help](#)

End of Term Web Archive

US Federal Government Websites 2008-2009

Home

Search Full Text

Site List

Explore Data

[Overview](#)

[Project Background](#)

[Project Partners](#)

[End of Term 2012](#)

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.

Committee on Natural Resources, Republican Site

Home
Press
Hearings
Reports
About the Committee
Hot Issues
Subcommittees
Links of Interest
109th Web Site

1329 Longworth House Office Building
Washington, DC 20515
Phone: (202) 225-2261
Fax: (202) 225-5929

186 Ford House Office Building
Washington, DC 20515
Phone: (202) 226-2311
Fax: (202) 225-4773

302 Ford House Office Building

Committee News

Friday, September 12, 2008
Removal of OCS Revenue Sharing With Coastal States Confirms Pelosi "Energy" Bill Is A Sham

Thursday, September 11, 2008
Dear Colleague: The Democrat Energy Bill is a Sham. They've gone from 85 percent of the OCS Off limits to More than 85 percent of the Oil Off limits

Thursday, September 11, 2008
Rep. Don Young Statement On Inspector General Report

Read All Press Releases:

GAS Today

Regular 3.84 9

Committee Calendar

Click Calendar Dates for details

Show today's events

September 2008						
Sun	Mon	Tue	Wed	Thu	Fri	Sat
31	1	2	3	4	5	6
7	8	9	10	11	12	13
14	15	16	17	18	19	20
21	22	23	24	25	26	27
28	29	30	1	2	3	4
5	6	7	8	9	10	11

Ranking Member: Don Young (R-AK)
Biography
Welcome Message
Congressional Website

Republican Committee Membership:
Jim Saxton (NJ-3)
Ron Gallaghy (CA-24)
John Duncan Jr., (TN-2)
Wayne Gilchrest, (MD-1)

Committee on Natural Resources, Republican Site – Sep. 15, 2008

Use this archive to:

[Browse](#) over 3000 U.S. Govt. websites
[Search](#) the full text of over 160 million files

The 2008 End of Term web archive contains over 16 terabytes of data collected from U.S. Government agency websites between September 2008 and November 2009.

[eotarchive on Twitter](#)

Join us Monday, Oct 17th at the Federal Depository Library Conference for a project overview & demo! <http://t.co/T0w9ZP6V>

Jul 2011 – Cooperative Archiving: Event Harvesting in Perspective @NDIIPP #pub <http://t.co/myzfecVj>

Jul 2011 – The End of Term Was Only the Beginning @NDIIPP #pub <http://t.co/2gyDHA4>

May 2011 – Curation of the End of Term Archive. #UNT #pub <http://t.co/5OJ3nnW4>

Dec 2010 – Findings of the Web Archive Survey of Federal Depository Libraries #UNT

Full text search

22

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 HTML PDF MS Word

Results 1-10 of about 548,834 (2.341 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 CFDA# Search News... Contacts Budget Annual Reports and Plans Jobs at ED Inspector General **No** FEAR Act Data Site Policies...
<http://www.ed.gov/nclb/landing.jhtml> - 35k - text/html
[All versions \(1\)](#) - [More from ed.gov](#)

2. [Friday, January 12, 2007 - Texas Times: Rein vigorating No Child Left Behind](#)

Friday, January 12, 2007 - Texas Times: Rein vigorating **No Child Left Behind**...
<http://cornyn.senate.gov/calendar/2007/01/12/rein vigorating-no-child-left-behind/index.html> - 1k - text/html
[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**...
<http://www.whitehouse.gov/news/releases/2006/10/20061005-6.html> - 43k - text/html
[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 21 st Century full of hope and... met. In America, **no child** should be **left behind**. Every **child** should be educated to his or her full... establishing annual assessments in grades 3-8, within two years of enacting this plan. "**No Child Left Behind**... of disadvantaged students will be recognized and rewarded with "**No Child Left Behind**" bonuses... in closing the achievement gap will be honored with awards from a "**No Child Left Behind**" school... eligible to receive a one-time bonus. Awards "**No Child Left Behind**" School Bonus... This award... system shared each of look forward to working with Congress to ensure that all children

Site List

23

[Contact Us](#) | [Help](#)



End of Term Web Archive

US Federal Government Websites 2008-2009

[Home](#)

[Search Full Text](#)

[Site List](#)

[Explore Data](#)

Results: 9 Items

Sorted by:

Site List

Page: 1

Site Lookup:

Look up a site by keywords in the title, description or URL.

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)
- digitalpreservation (1)
- memory (1)
- rs6 (1)
- usa (1)



Title: [Library of Congress Home](#)

Archival URL: <http://eot.us.archive.org/eot08/20080915221603/loc.gov/index.html>

IA Site value:

Live URL: <http://www.loc.gov/index.html>

Coverage: September 15, 2008 - August 25, 2009

Description: The Library of Congress. 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.

Matches: ...Legislative [Library of Congress](#) Home <http://...>
5 hits ...index.html 2008-09-15 2009-08-25 The [Library of Congress](#)....
...The [Library of Congress](#) is the nation's oldest federal...



Title: [Library of Congress Online Catalogs \(Upgraded May 19, 2008\)](#)

Archival URL: <http://eot.us.archive.org/eot08/20080916003635/catalog.loc.gov/>

IA Site value:

Live URL: <http://catalog.loc.gov/>

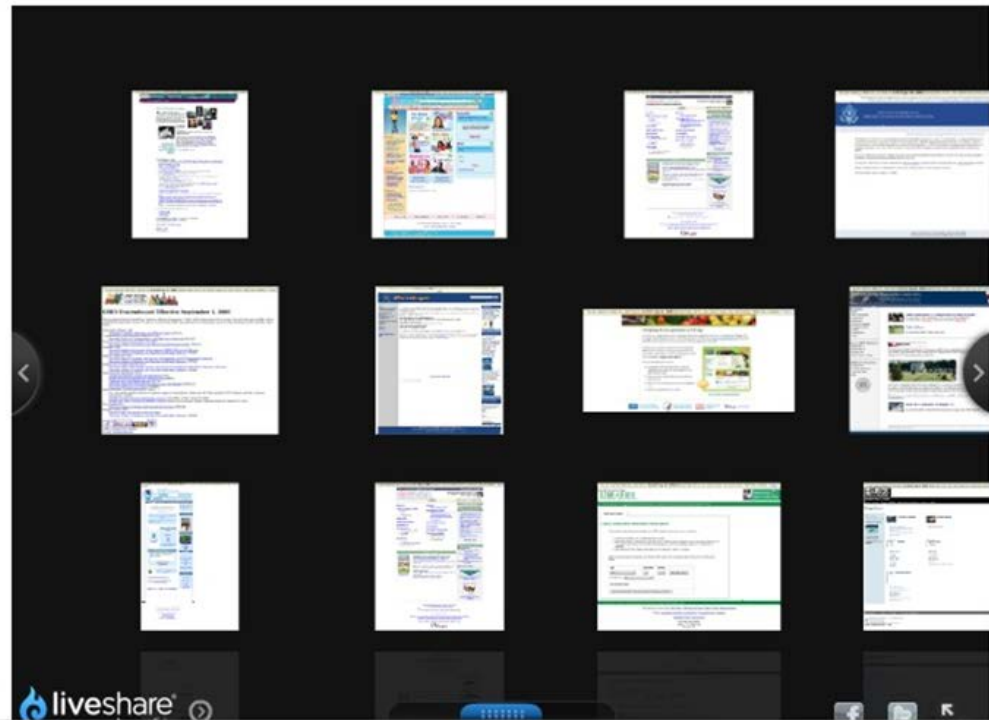
Coverage: September 16, 2008 - May 8, 2009

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

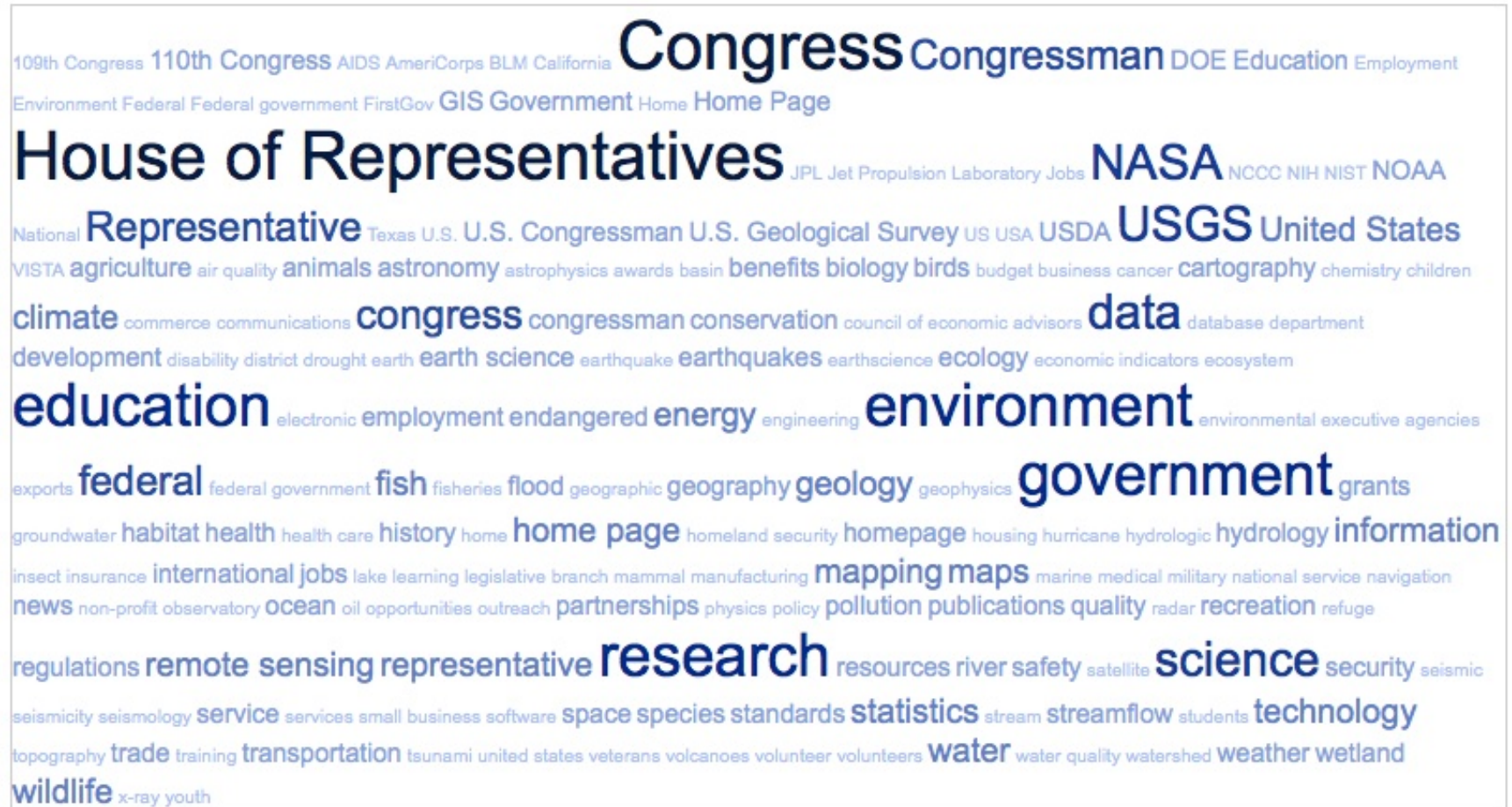
Browse by Image

24

- 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

Background

27

- 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

28

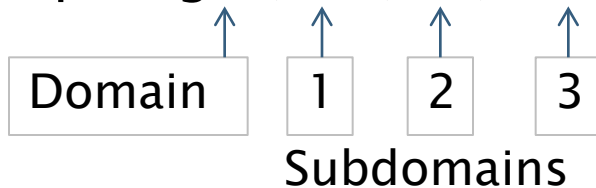
Domains	# URLs	Subdomains
gov	137,847,822	14,339
mil	3,555,425	1,677

16,016

SURTS: Reordering URLs by domain structure

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

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



Unique 1st Level Subdomains = 1,647
After validation = 1,151 Subdomains

Human Classification

29

- 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

30

- 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
 - ▣ Forced to classify at high level

Link Analysis

31

- 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

32

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

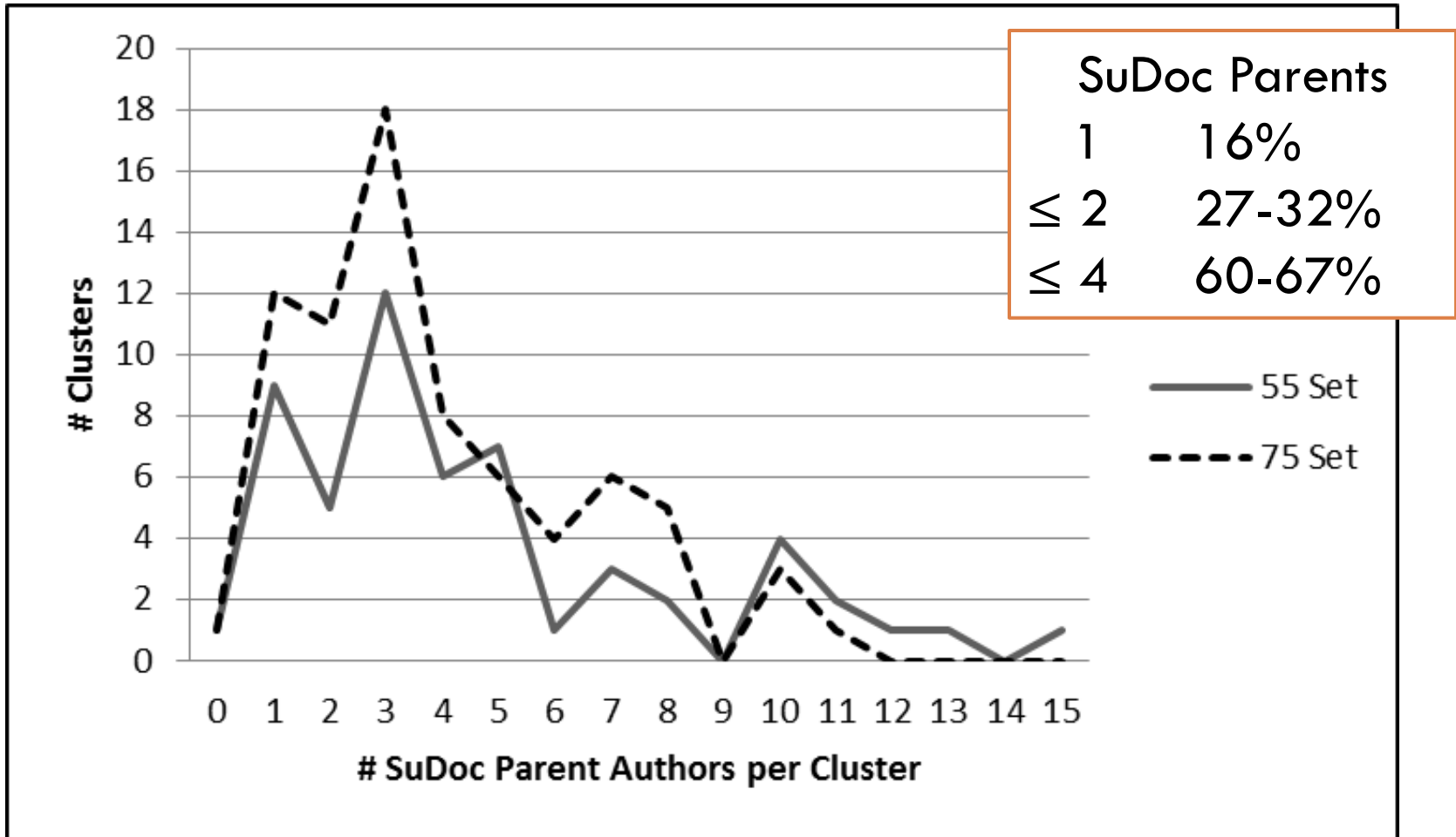
Cluster Analysis		
55-Set		75-Set
39	<i>Identical</i>	39
16	13 / 2 2 / 3 1 / 4	36

Clusters 55-24 & 75-31

Identical Subdomains

- fdic.gov
- fdicconnect.gov
- fdicig.gov
- fdicoig.gov
- fdicseguro.gov
- myfdicinsurance.gov
- egrpra.gov

Findings: SuDoc Classification



Topical Evaluation of Clusters

34

- 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

Cluster Analysis		
55-Set		75-Set
39	<i>Identical</i>	39
16	<i>Unique</i>	36

Tag Analysis

35

- 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
2 Subdomains
<ul style="list-style-type: none">• federalregister.gov• fedreg.gov

Cluster 55-19	SME 40	SME 32	SME 42
RC 3	<ul style="list-style-type: none">• federal regulations• administrative law	<ul style="list-style-type: none">• federal regulations	<ul style="list-style-type: none">• federal regulations

Findings: Topical Evaluation

36

- 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%)

Clusters	RC 1	RC 2	RC 3
130	21%	18%	61%
75-Set	21%	17%	61%
55-Set	20%	20%	60%

- Cluster Analysis successfully identified strongly related subject content in the subdomains of 61% of clusters

Impact of Increasing # of Clusters

37

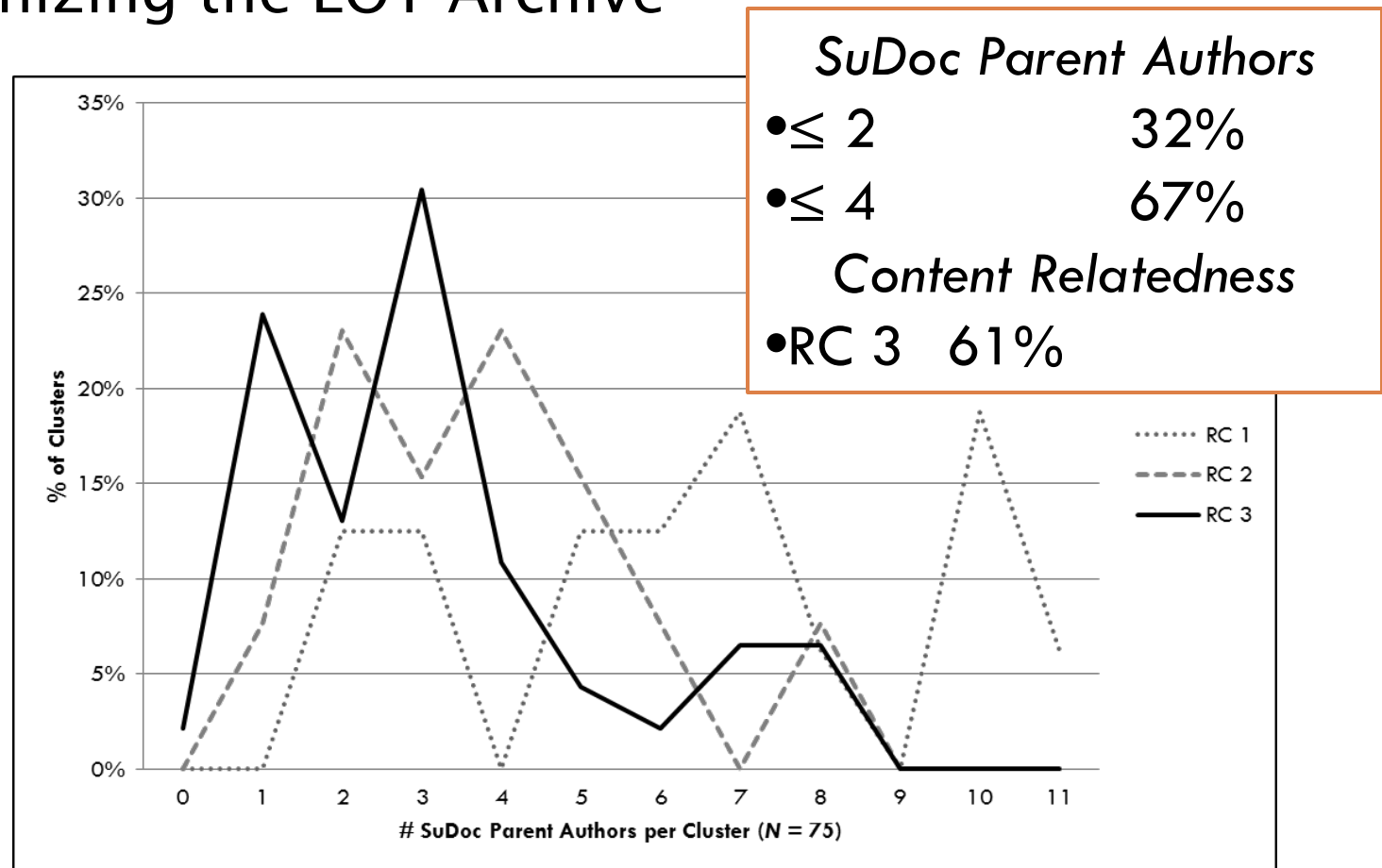
Clusters	#	RC 1	RC 2	RC 3
Identical	39	18%	10%	72%
All	130	21%	18%	61%
Unique in 75-Set	36	22%	14%	64%
Unique in 55-Set	16	25%	31%	44%

- 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%)

Conclusion

38

- Cluster analysis was reasonably effective at organizing the EOT Archive



EOTCD Project Accomplishments

39

- 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

40

- 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?

Building the 2012 Archive

End of Term Presidential Harvest 2012 [Other Projects](#) | [Help](#)

[Project Home](#) [About This Project](#) [Add A URL](#)

Number of URLs Nominated: 0
Number of Nominators: 0

☆☆☆☆☆☆☆☆☆☆

Search by URL

Search for an existing URL in the system.

☆☆☆☆☆☆☆☆☆☆

Browse URLs

Timeframe for 2012 project

42

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

43

- 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.)
 - **Nomination Tool:**
<http://digital2.library.unt.edu/nomination/eth2012/>

Questions?

The screenshot shows the 'End of Term Web Archive' interface for US Federal Government Websites from 2008-2009. It features a search bar, navigation tabs (Home, Search Full Text, Site List, Explore by Image), and a results page for 'National Geospatial Program'. The results list includes site lookups, government branches (Executive, Legislative, Judicial, Quasi-Federal), and URL segments (house, nasa, noaa, uscourts, senate, nsh, gfc, usda, whitehouse). Three specific site entries are shown with their titles, archival URLs, IA site values, live URLs, coverage dates, and descriptions.

End of Term Web Archive
US Federal Government Websites 2008-2009

Home Search Full Text Site List Explore by Image

Results: 3304 items
Sorted by: relevance Go

Page: 1 2 3 4 5 ... Next

Site Lookup:
Look up a site by keywords in the title, description or URL.

Government branch

- Executive (1562)
- Legislative (893)
- Judicial (162)
- Quasi-Federal (3)

URL segment
The following text appears in a section of the site's URL. This can help you narrow results to particular agencies.

- house (307)
- nasa (307)
- noaa (142)
- uscourts (142)
- senate (136)
- nsh (113)
- gfc (98)
- usda (74)
- whitehouse (63)

Site List

Title	Archival URL	IA Site value	Live URL	Coverage	Description
National Geospatial Program	http://eot.us.archive.org/eot08/20080916080621/usgs.gov/ngp/		http://www.usgs.gov/ngp/	September 16, 2008 - August 25, 2009	Home page of the National Geospatial Program
Welcome to the Safety and Mission Assurance (SMA) Homepage	http://eot.us.archive.org/eot08/20080916000624/iaroch-gfrc.nasa.gov/		http://iaroch.gfrc.nasa.gov/	September 16, 2008 - May 11, 2009	The Safety Mission Assurance (SMA) is committed to ensuring the highest probability of mission success.
HEASARC: Suzaku Guest Observer Facility	http://eot.us.archive.org/eot08/20080920011443/heasarc-gfrc.nasa.gov/doc/astroe/astroeogof.html		http://heasarc.gfrc.nasa.gov/doc/astroe/astroeogof.html	September 20, 2008 - August 12, 2009	The responsibility of the U.S. Suzaku/Astro-E2 Guest Observer Facility is to enable U.S. astronomers to make the

eotproject@loc.gov

abgr@loc.gov

krm0028@unt.edu

Follow us on twitter! @eotarhive

<http://eotarhive.cdlib.org/>