Collecting and Archiving the Federal Web

Mark Phillips
Assistant Dean for Digital Libraries
UNT Libraries
May 19, 2015
In most areas of life the Web is the primary platform for the creation and dissemination of information
This rings true for the governmental Web as well.
Municipal, County, State, and Federal Governments reach their constituents via the Web
The United State Federal Government has used the Internet as a publishing platform for over 20 years.
Making it the single largest platform for the dissemination of federal information in existence
The ability to quickly and in a cost effective manner publish content to the Web has made it possible for all branches of the Federal government to reach those in need of governmental information.
But with great abilities to create also come opportunities for loss.
Content that was so easy to make available can just as easily be removed.
With the ability for planned or unplanned denial of access is as easy as a few keystrokes.
It is important for libraries, archives, and cultural heritage institutions to select, collect, organize, provide access to and preserve this important Web content.
End of Term
National Archives and Records Administration (NARA)
Conducted an “End of Term” Web harvest at the end of the first Bush term.
Available at http://webharvest.gov/
2004 Presidential Term Web Harvest

The National Archives and Records Administration (NARA) conducted a harvest (i.e., capture) of Federal Agency public web sites as they existed prior to January 20, 2005. This harvest was intended to document Federal agencies' presence on the World Wide Web at the time that the Presidential Administration term ended in early 2005. Learn more.

The harvest produced a public reference copy of the web sites for the purpose of continual availability to the public, and also produced a record copy to be retained in the holdings of NARA.

Federal Agency Sites Harvested

Web sites included in the harvest were identified from information provided by the U.S. General Services Administration's (GSA) "GOV" Internet Domain Registry (http://www.dotgov.gov) and from a list of unrestricted "MIL" sites provided by the Defense Information Systems Agency (DOD/DISA).

The accuracy of the harvest was affected by these factors:

- Only those unrestricted 1st and 2nd level Federal and military domains registered with the "GOV" registry or the Defense Information Systems Agency (DISA) included on the lists compiled and sent to NARA were included in the harvest,
- Whether addresses from GSA and DISA lists resolved successfully, and
- The capabilities of crawler tools used (see Heritrix at http://crawler.archive.org) and the server environment being crawled. See a report on limitations of capabilities [PDF].

NARA has made every reasonable effort to ensure that web sites' code and programming were captured accurately. NARA is not responsible for the web sites' compliance with Federal laws, regulations, and requirements. NARA is responsible for providing public access to these copied web sites but is not responsible for maintaining code such as links, accessibility features, search or site maps, or other functionality that may have been true of the sites before they were copied.

Mention of commercial products, services, or resources within this notice does not constitute an endorsement by the National Archives and Records Administration or the United States Government.

Learn more: The report requires the capture and transfer of Federal public web sites at the end of a term of a Presidential Administration. To find current Federal agency web sites, please use http://www.usa.gov. For technical issues, contact info@archive.org. For questions about these Federal records, contact cer@nara.gov. For questions about this website, contact brandon.hirsch@nara.gov.

Frequently Asked Questions: For frequently asked questions visit this link FAQ
In 2008 NARA announced it wouldn't be doing the same at the end of the second Bush term
A call to action!
US members of the International Internet Preservation Consortium (IIPC) formed an ad-hoc group to shoulder the task of doing the .gov crawl
California Digital Library
Internet Archive
Government Printing Office
Library of Congress
University of North Texas
Worked to establish a list URLs for the “Federal Domain”
### End of Term Presidential Harvest 2008

**Open Access Symposium**

---

**Number of URLs Nominated:** 4622  
**Number of Nominators:** 29

---

**Search by URL**

http:// submit

Search for an existing URL in the system.

- [ ] Allow partial matches

---

**Browse URLs**

#### COM

<table>
<thead>
<tr>
<th>0</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
<th>7</th>
<th>8</th>
<th>9</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>B</td>
<td>C</td>
<td>D</td>
<td>E</td>
<td>F</td>
<td>G</td>
<td>H</td>
<td>I</td>
<td>J</td>
</tr>
</tbody>
</table>

#### EDU

<table>
<thead>
<tr>
<th>0</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
<th>7</th>
<th>8</th>
<th>9</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>B</td>
<td>C</td>
<td>D</td>
<td>E</td>
<td>F</td>
<td>G</td>
<td>H</td>
<td>I</td>
<td>J</td>
</tr>
</tbody>
</table>

#### GOV

<table>
<thead>
<tr>
<th>0</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
<th>7</th>
<th>8</th>
<th>9</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>B</td>
<td>C</td>
<td>D</td>
<td>E</td>
<td>F</td>
<td>G</td>
<td>H</td>
<td>I</td>
<td>J</td>
</tr>
</tbody>
</table>
Primarily .gov and .mil
Initiated a series of three crawls
Pre Election
Post Election
Post Inauguration
With two goals in mind.
Capture the expected change of the Federal Web caused by the transition in the executive branch
Serve as a catalyst event to create a large-scale archive of .gov and .mil content
In the end 16TB of Web content was harvested
160,211,356 URIs
Copies of all data available to each of the partners
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.

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

loading tweets...
In 2012 the same group initiated the project again
This time Harvard was added to the roster.
A slightly modified process was used to collect candidate URLs for crawling
We also worked with Library Science students from the Pratt Institute to nominate social media sites for the government
1,476 URLs from 31 nominators were collected
Because President Obama was re-elected, additional crawls were not conducted after the initial pre-election crawl
Once compiled the EOT 2012 Crawl data measured 31TB
Almost double that of EOT 2008
We are beginning the discussions for EOT 2016
What has been done with the data to date?
IMLS Funded: Classification of the End-of-Term Archive: Extending Collection Development to Web Archives (eotcd), 2010-2012
Attempt to overlay the SuDoc Classification System on the EOT 2008 Web archive
End of Term 2008 Presidential Web Archive:
PDF Content Analysis
An analysis of the 4.5 million unique PDF documents in the EOT 2008 archive
<table>
<thead>
<tr>
<th>Creation Date</th>
<th>April 11, 2001, 8:06 a.m.</th>
</tr>
</thead>
<tbody>
<tr>
<td>Modification Date</td>
<td>Mon Apr 23 15:26:03 2001</td>
</tr>
<tr>
<td>Producer</td>
<td>Acrobat Distiller 4.05 for Windows</td>
</tr>
<tr>
<td>Num Pages</td>
<td>1</td>
</tr>
<tr>
<td>Entropy</td>
<td>4.68</td>
</tr>
<tr>
<td>Histogram Dispersion</td>
<td>8.16</td>
</tr>
<tr>
<td>File Size</td>
<td>98.6 KB</td>
</tr>
<tr>
<td>Tagged?</td>
<td>False</td>
</tr>
<tr>
<td>Optimized?</td>
<td>False</td>
</tr>
<tr>
<td>Encrypted?</td>
<td>True</td>
</tr>
<tr>
<td>Page Size</td>
<td>2594 x 2376 pts</td>
</tr>
<tr>
<td>Width</td>
<td>36.0 in.</td>
</tr>
<tr>
<td>Height</td>
<td>33.0 in.</td>
</tr>
<tr>
<td>Area</td>
<td>1188 sq in.</td>
</tr>
<tr>
<td>Orientation</td>
<td>Landscape</td>
</tr>
<tr>
<td>Word Count</td>
<td>443</td>
</tr>
<tr>
<td>Words per Page</td>
<td>443.0</td>
</tr>
<tr>
<td>Character Count</td>
<td>2,283</td>
</tr>
<tr>
<td>Average Word Length</td>
<td>5.15</td>
</tr>
</tbody>
</table>
The PDF dataset contains over 60 million pages of content.
1.5 million pdfs (33%) were 1 page in length
Longest PDF is 17,584 pages
Average PDF Length = 13.8 pages
We investigated the number of domains that a PDF was served from.
PDFs hosted on 1 – 25 unique domains
Average 1.1 domain per PDF
PDFs hosted on 1 – 4 Top Level Domains
PDFs hosted on 1 – 1763 unique URLs
Average = 1.2 URLs per PDF
What domains had the most PDFs?
gpo.gov with 1,082,735 PDFs
Followed by:

usda.gov
house.gov
army.mil
bea.gov
census.gov
With PDFs over 1 page?
gpo.gov = 594,430
usda.gov
house.gov
uscis.gov
uscourts.gov
army.mil
tres.gov
noaa.gov
What about 20+ pages
gpo.gov
gao.gov
epa.gov
usda.gov
army.mil
noaa.gov
The PDF work is one small example of ways we have reused this Web archive data internally in the UNT Libraries
We have an interest in mining these Web archives to extract the discrete “publications” and move them into our traditional digital library infrastructure.
We did this in a trial collection
The Defense Base Closure and Realignment Commission
Known as BRAC
Defense Base Closure and Realignment Commission

About the Defense Base Closure and Realignment Commission

The 2005 Defense Base Closure and Realignment Commission was a 9-member independent commission appointed by President Bush with the advice and consent of the Senate. Its purpose was to provide "an objective, non-partisan, and independent review and analysis of the list of military installation recommendations issued by the Department of Defense." Recommendations were evaluated against a list of specific criteria which emphasized military value, but also considered costs and savings, economic impact on communities near military installations, community infrastructure, and environmental impact. The Commission undertook public hearings, site visits, and other information-gathering activities before presenting its findings in a report to the President on September 8, 2005.

Informally known as "BRAC" (Base Realignment and Closure) the process has been through five rounds. The 1988 Commission was authorized by Public Law 100-526. Subsequent Commissions in 1991, 1993, 1995, and 2005 were authorized by Public Law 101-510, as amended by Public Law 107-48.
2 Letters from concerned citizens in response to the recommendation regarding the Defense Office of Hearings and Appeals in Fort Meade

**Date:** July 30, 2005  
**Creator:** unknown  
**Description:** Community Correspondence - 2 Letters from concerned citizens in response to the recommendation regarding the Defense Office of Hearings and Appeals in Fort Meade  
**Contributing Partner:** UNT Libraries Government Documents Department

2 Questions pertaining to DON-0133 and the non-BRAC Scenario (Portsmouth Naval Shipyard)

**Date:** August 12, 2005  
**Creator:** United States. Dept. of Defense.
Statistics for Defense Base Closure and Realignment Commission

877,346 Total Uses / 14,183 Total Items (1,231,138 files) / 14,169 Visible / 14 Hidden

Usage by Month/Year

<table>
<thead>
<tr>
<th>Year</th>
<th>January</th>
<th>February</th>
<th>March</th>
<th>April</th>
<th>May</th>
<th>June</th>
<th>July</th>
<th>August</th>
<th>September</th>
<th>October</th>
<th>November</th>
<th>December</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>2015</td>
<td>31,013</td>
<td>30,741</td>
<td>23,945</td>
<td>26,091</td>
<td>15,363</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>127,153</td>
</tr>
<tr>
<td>2014</td>
<td>22,033</td>
<td>21,568</td>
<td>18,663</td>
<td>21,114</td>
<td>25,673</td>
<td>31,223</td>
<td>103,399</td>
<td>38,818</td>
<td>33,215</td>
<td>29,540</td>
<td>26,703</td>
<td>25,546</td>
<td>397,495</td>
</tr>
</tbody>
</table>
In addition to enabling students, researchers and the general public the ability to go back in time to view the Federal government through the lens of their websites.
In 2014 the UNT Libraries acquired a 110 TB Web archive of all .gov content from the Internet Archive's Global Wayback Machine
We worked with IA to extract this content, which they made available to researchers as a demonstration of the opportunities for research that Web archives offer.
This content represents the .gov domain from 1996 to 2013
We have plans to provide local access to this data as well as researching ways of extracting meaningful subsets of information of integration into the UNT Digital Library and our local government information collections.
In closing...
As the Web continues to play an important role in people's everyday lives,
And as governments at all levels seek to engage with constituents via this Web
It is increasingly important that cultural heritage institutions select, collect, describe, provide access to and preserve the Web publications of our governments.
So that citizens, students, scholars and researchers in the future can view and use these resources as they seek to understand how laws, policies, practices and processes...
Have changed over time.
Questions?

Mark Phillips

mark.phillips@unt.edu