End of Term Web Archives
2008, 2012, 2016 ...

Mark Phillips
UNT Libraries
April 5, 2016
End of Term Presidential Web Archive
2004
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 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
Worked to establish a list URLs for the “Federal Domain”
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
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.
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
## End of Term Presidential Harvest 2012

- **Number of URLs Nominated:** 1476
- **Number of Nominators:** 31

### Search by URL

- **Input:** [http://](#)  
  - **Submit:**

#### Search Options:
- **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>
<tr>
<td>K</td>
<td>L</td>
<td>M</td>
<td>N</td>
<td>O</td>
<td>P</td>
<td>Q</td>
<td>R</td>
<td>S</td>
<td>T</td>
</tr>
<tr>
<td>U</td>
<td>V</td>
<td>W</td>
<td>X</td>
<td>Y</td>
<td>Z</td>
<td></td>
<td></td>
<td></td>
<td></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>
<tr>
<td>K</td>
<td>L</td>
<td>M</td>
<td>N</td>
<td>O</td>
<td>P</td>
<td>Q</td>
<td>R</td>
<td>S</td>
<td>T</td>
</tr>
<tr>
<td>U</td>
<td>V</td>
<td>W</td>
<td>X</td>
<td>Y</td>
<td>Z</td>
<td></td>
<td></td>
<td></td>
<td></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>
<tr>
<td>K</td>
<td>L</td>
<td>M</td>
<td>N</td>
<td>O</td>
<td>P</td>
<td>Q</td>
<td>R</td>
<td>S</td>
<td>T</td>
</tr>
<tr>
<td>U</td>
<td>V</td>
<td>W</td>
<td>X</td>
<td>Y</td>
<td>Z</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
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 and 195,277,904 URIs
Almost double of EOT 2008
All data was shared among partners. LOC, IA, and UNT acquired local copies.
Working with the EOT Archives
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
Not much has been done with the EOT-2012 data
Shared by all three repositories

237,171

117,379,466
Each crawling institution contributed mostly unique content to the project.
There is interest in looking at the 2008 and 2012 crawl data to identify change.
Moving forward to 2016
There will be a presidential change.
The original partners plus new partners are working to plan out crawling this fall and winter.
Additional interest in capturing Social Media for Agencies, Departments, and Projects
Interest in engaging the community to identify content to harvest.
Plan to use the URL Nomination Tool again.
Interest in providing more timely access to all three EOT collections through a single interface operated by CDL/IA

http://eotarchive.cdlib.org/
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.
Copies of the data will be distributed again to interested partners.
How can we reuse this content locally with researchers.
What longitudinal analysis can be carried out with data from three harvests?
The preservation of government information is something that can only be done with collaboration between a variety of institutions and stakeholders.
The EOT partnership is an example of one of those collaborations that has been successful.
Thank you.

Mark Phillips

mark.phillips@unt.edu