The End of Term Archive: Archiving the U.S. Government Web

Abigail Grotke, Web Archiving Team Lead, Library of Congress
@agrotke | abgr@loc.gov
It all began a long, long, time ago, in a far away...
Goals of the End of Term Project

- work collaboratively to preserve public U.S. Government websites
- document federal agencies’ presence on the web during the end of Presidential terms
- enhance the existing research collections of the partner institutions
- raise awareness about the need for preservation
- engage with researchers and subject experts
Extant .gov web archiving efforts

Capture, Preservation, & Access

- LOC: legislative branch, some executive
- GPO: agency sites, often ephemeral
- NARA: congressional web harvest every 2 years
- IA: global & curated crawls
- Agency-level: NIH/NLM, DOE, DOL, HHS, CMS, others, using Archive-It or other tools
- UNT & Others: Topical .gov collecting
- Internal agency guidelines

Community Efforts

- Federal Web Archiving Group
- Research Initiatives
  - academic
  - NGO or watchdog
- Citizen Driven
  - grassroots efforts, e.g. Data refuge/rescue
- End of Term
  - focused but large-scale multi-institutional project
Original End of Term Web Archive Partners

for 2008/2012 - all IIPC & NDIIPP/NDSA partners
EOT 2016: more partners!
EOT Collaborative Roles

- Project management and coordination (All, rotates)
- Nomination tool development (UNT)
- URL nomination (All + community/public)
- Crawling (IA, UNT, LC, GW)
- Preservation of full copy (IA, LC, UNT)
- Access: portal, full-text search, metadata, research & support (IA, CDL)
- Outreach, press, twitter account (primarily IA, LC, UNT, Stanford)

The 2016 project brought more partners, capacity for better, distributed crawling, and more community and researcher engagement
Defining the “government web presence”
Community Engagement

Events: NYAM, U-Toronto, U-Penn, UC-Riverside, more

Research: CMU, Georgetown, U-Washington
## Volunteer contributions

<table>
<thead>
<tr>
<th>Year</th>
<th>Nominations</th>
<th>Volunteers</th>
</tr>
</thead>
<tbody>
<tr>
<td>2008</td>
<td>457</td>
<td>26</td>
</tr>
<tr>
<td>2012</td>
<td>1476</td>
<td>31</td>
</tr>
<tr>
<td>2016</td>
<td>15,000+</td>
<td>400+</td>
</tr>
</tbody>
</table>

**End of Term Presidential Harvest 2016**

Plus over 150,000 from DataRescue/EDGI events/tools!
## Size of Archive

<table>
<thead>
<tr>
<th>Year</th>
<th>URLs</th>
<th>Size</th>
<th>Comments</th>
</tr>
</thead>
<tbody>
<tr>
<td>2008</td>
<td>~160 million</td>
<td>17.95 TB</td>
<td>Multiple crawls, deduplicated</td>
</tr>
<tr>
<td>2016</td>
<td>over 429 million</td>
<td>293.91 TB</td>
<td>Includes ~150 TB of FTP crawls</td>
</tr>
</tbody>
</table>
EOT 2016 Content

- Content includes:

  - 9,000+ social media accounts (scrape of gov SM registry API) 44% FB, 37% TW, 10% YT
  - ~190K total domain, subdomains, gov and non-gov sites

- more crowdsourced, curatorial nominations
EOT 2016 Press ++

- **Press**
  - Dozens of articles and interviews

- **Collaborations**
  - Data Refuge
  - EDGI
  - GSA / 18F
  - data.gov
EOT Challenges

• **Typical web archiving challenges**
  • complexity of content
  • volume & proliferation
  • “you get what you get” w/ little cataloging or QA

• **Distribution of work**
  • more partners = more project/partner mgmt
  • contributed seed lists

• **Resource constraints**
  • the “it isn’t anyone’s actual job” problem
  • tech, time limitations & scale of data
  • funding = (there is none)
Using the EOT Archive
End of Term Web Archive
US Federal Web Domain at Presidential Transitions

The End of Term Web Archive captures and saves U.S. Government websites at the end of presidential administrations. Beginning in 2008, the EOT has thus far preserved websites from administration changes in 2008 and 2012, and is currently preparing for the 2016 electoral season.

EOT at a Glance
Crawl dates:
September 2008 to May 2008
September 2012 to March 2013
Number of websites captured:
2008: 3,305 websites
2012: 3,247 websites
Terabytes of data captured:
2008: 16 TB
2012: 21 TB

Archive Scope
The End of Term Web Archive contains federal government websites (.gov, .mil, etc) in the Legislative, Executive, or Judicial branches of the government. Websites that were at risk of changing (i.e., whitehouse.gov) or disappearing altogether during government transitions were captured. Local or state government websites, or any other site not part of the federal government domain were out of scope.
Plans for release of 2016 – 2017

• All web crawl data from IA, LC, UNT has been ingested at IA.

• Derivative datasets for all the data (WATs, WANEs, extracted page text) have been generated.

• Components to integrate new content into portal are being worked on (metadata, search, thumbnails, Wayback indexes). Once finalized, CDL will begin process to update the portal.

• We have a number of researchers interested in the data (IA working with them)
analysis by project team members
Comparing Web Archives: EOT2008 and EOT2012 – Curator Intent

This is another post in a series that I’ve been doing to compare the End of Term Web Archives from 2008 and 2012. If you look back a few posts in this blog you will see some other analysis that I’ve done with the datasets so far. One thing that I am interested in understanding […]

This entry was posted in thinking out loud on June 20, 2016.

Comparing Web Archives: EOT2008 and EOT2012 – What disappeared

This is the fourth post in a series that looks at the End of Term Web Archives captured in 2008.
Comparing PDFs in EOT from 2008 to 2012

http://vphill.com/journal/post/5872/

http://vphill.com/journal/post/5861/
.gov & .mil biggest change

<table>
<thead>
<tr>
<th>Domain</th>
<th>EOT2008</th>
<th>EOT2012</th>
<th>Change</th>
<th>% Change</th>
</tr>
</thead>
<tbody>
<tr>
<td>house.gov</td>
<td>13,694,187</td>
<td>35,894,356</td>
<td>22,200,169</td>
<td>162%</td>
</tr>
<tr>
<td>senate.gov</td>
<td>5,043,974</td>
<td>9,924,917</td>
<td>4,880,943</td>
<td>97%</td>
</tr>
<tr>
<td>gpo.gov</td>
<td>8,705,511</td>
<td>3,888,645</td>
<td>-4,816,866</td>
<td>-55%</td>
</tr>
<tr>
<td>nih.gov</td>
<td>5,276,262</td>
<td>1,267,764</td>
<td>-4,008,498</td>
<td>-76%</td>
</tr>
<tr>
<td>nasa.gov</td>
<td>6,693,542</td>
<td>3,063,382</td>
<td>-3,630,160</td>
<td>-54%</td>
</tr>
<tr>
<td>navy.mil</td>
<td>94,081</td>
<td>3,611,722</td>
<td>3,517,641</td>
<td>3,739%</td>
</tr>
</tbody>
</table>
### Top 15 .gov & .mil domains present in 2008 but missing in 2012

<table>
<thead>
<tr>
<th>Domain</th>
<th>URLs</th>
</tr>
</thead>
<tbody>
<tr>
<td>geodata.gov</td>
<td>812,524</td>
</tr>
<tr>
<td>nifl.gov</td>
<td>504,910</td>
</tr>
<tr>
<td>stat-usa.gov</td>
<td>398,961</td>
</tr>
<tr>
<td>tradestatsexpress.gov</td>
<td>243,729</td>
</tr>
<tr>
<td>arnet.gov</td>
<td>174,057</td>
</tr>
<tr>
<td>acqnet.gov</td>
<td>171,493</td>
</tr>
<tr>
<td>dccourts.gov</td>
<td>161,289</td>
</tr>
<tr>
<td>web-services.gov</td>
<td>137,202</td>
</tr>
<tr>
<td>metrokc.gov</td>
<td>132,210</td>
</tr>
<tr>
<td>sdi.gov</td>
<td>91,887</td>
</tr>
<tr>
<td>davie-fl.gov</td>
<td>88,123</td>
</tr>
<tr>
<td>belmont.gov</td>
<td>87,332</td>
</tr>
<tr>
<td>aftac.gov</td>
<td>84,507</td>
</tr>
<tr>
<td>careervoyages.gov</td>
<td>57,192</td>
</tr>
<tr>
<td>women-21.gov</td>
<td>56,255</td>
</tr>
</tbody>
</table>

### Top 15 .gov & .mil domains new in 2012

<table>
<thead>
<tr>
<th>Domain</th>
<th>URLs</th>
</tr>
</thead>
<tbody>
<tr>
<td>militaryonesource.mil</td>
<td>859,843</td>
</tr>
<tr>
<td>consumerfinance.gov</td>
<td>237,361</td>
</tr>
<tr>
<td>nrd.gov</td>
<td>194,215</td>
</tr>
<tr>
<td>wh.gov</td>
<td>179,233</td>
</tr>
<tr>
<td>pnnl.gov</td>
<td>132,994</td>
</tr>
<tr>
<td>eia.gov</td>
<td>112,034</td>
</tr>
<tr>
<td>transparency.gov</td>
<td>109,039</td>
</tr>
<tr>
<td>nationalguard.mil</td>
<td>108,854</td>
</tr>
<tr>
<td>acus.gov</td>
<td>93,810</td>
</tr>
<tr>
<td>404.gov</td>
<td>82,409</td>
</tr>
<tr>
<td>savingsbondwizard.gov</td>
<td>76,867</td>
</tr>
<tr>
<td>treasuryhunt.gov</td>
<td>76,394</td>
</tr>
<tr>
<td>fedshirevets.gov</td>
<td>75,529</td>
</tr>
<tr>
<td>onrr.gov</td>
<td>75,484</td>
</tr>
<tr>
<td>veterans.gov</td>
<td>75,350</td>
</tr>
</tbody>
</table>
EOT2008 and EOT2012 Crawling Schedule
Extracted Special Web Collections

https://archive.org/details/MilitaryIndustrialPowerpointComplex


eot 2016 “raw” content

https://archive.org/details/EndOfTerm2016WebCrawls
researcher access to .gov

WAT Datasets
(Web Archive Transformation)
Key Metadata from Every Resource

LGA Datasets
(Longitudinal Graph Analysis)
What Links to What over Time

WANE Datasets
(Web Archive Named Entities)
Names of People, Places, Organizations

Web Archive Datasets
(via platform, disk, APIs, etc.)
researcher access to .gov

http://archivesunleashed.com/

http://www.websci16.org/hackathon

https://github.com/vinaygoel/ars-workshop

http://webarchives.ca/
web preservation for content creators
What can website creators do to help?

• Create preservation-friendly websites

• Understand the nature of your organization’s website to determine the extent needed to preserve it:
  • What are the domains, subdomains?
  • What content is being hosted by third-party companies, for instance social media?
  • Is your content changing frequently?

• Archive your own content, using open source or subscription-based tools. Or at least make sure it is being preserved by someone else (check web.archive.org)

• Participate in our next End of Term 2020!
## Creating Preservable Websites

<table>
<thead>
<tr>
<th>✓ Follow web standards and accessibility guidelines</th>
<th>✓ Consider using a Creative Commons license</th>
</tr>
</thead>
<tbody>
<tr>
<td>✓ Be careful with robots.txt exclusions</td>
<td>✓ Use sustainable data formats</td>
</tr>
<tr>
<td>✓ Use a site map, transparent links, and contiguous navigation</td>
<td>✓ Embed metadata, especially the character encoding</td>
</tr>
<tr>
<td>✓ Maintain stable URIs and redirect when necessary</td>
<td>✓ Use archiving-friendly platform providers and content management systems.</td>
</tr>
</tbody>
</table>

For more details see Library of Congress Guide to Creating Preservable Websites
[http://www.loc.gov/webarchiving/preservable.html](http://www.loc.gov/webarchiving/preservable.html)
More Guidance for Website Creators

• International Internet Preservation Consortium Tools and Software
  https://netpreserve.org/web-archiving/tools-and-software/

• Recommended Formats Statement for Websites
  https://www.loc.gov/preservation/resources/rfs/websites.html

• Stanford University’s Archivability Guidelines
  http://library.stanford.edu/projects/web-archiving/archivability

• Columbia University’s Guidelines for Preservable Websites
  https://library.columbia.edu/bts/web_resources_collection/guidelines_for_preservable_websites.html

• Princeton University’s Guidelines for Designing Preservation-Friendly Websites

• Archive Ready, a free website archivability evaluation tool
  http://www.archiveready.com/
EOT Going Forward

• **Processing the 2016 archive:** Full-text search, build index, make derivatives, simple extractions like domains, comparisons against 2012/2008
• **Providing access:** Updating the EOT portal, exploring media search, new access interfaces, sharing derivative data sets with researchers
• **Preservation copies:** Transferring to Library of Congress this year.
• And of course … **End of Term 2020!**

Follow us on Twitter for updates:  [https://twitter.com/eotarchive](https://twitter.com/eotarchive)

Questions?
[abgr@loc.gov](mailto:abgr@loc.gov)
[eot-info@archive.org](mailto:eot-info@archive.org)