After the Harvest: Preservation, Access, and Researcher Services for the 2016 End of Term Archive

Jefferson Bailey, Director, Web Archiving, Internet Archive @jefferson_bail | jefferson@archive.org
Abbie Grotke, Web Archiving Team Lead, Library of Congress @agrotke | abgr@loc.gov
Mark Phillips, Associate Dean for Digital Libraries, UNT Libraries @vhill | mark.phillips@unt.edu
it all began a long, long, time ago, in a far away place
original end of term web archive partners

for 2008/2012 - all IIPC & NDIIPP/NDSA partners
# extant gov web archiving efforts

## Capture, Preservation, & Access
- LOC: .gov, election, other
- 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
  - most of those at left plus other feds
- Research Initiatives
  - academic
  - NGO or watchdog
- Citizen Driven
  - grassroots efforts
- End of Term
  - focused but large-scale multi-institutional project
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
eot collaborative distribution of work

- IA: crawling, preservation, access, full-text search, metadata, research tool & support
- LC: crawling, preservation, data transfers
- UNT: nomination tool development, crawling, nomination mgmt, preservation, access
- CDL: web portal, metadata
- GPO: URL nomination, outreach
- All: URL contributions, outreach, project management
- Others: URLs, education, coordination

Some variance of roles between 2008 & 2012 (and for 2016)
major funding brought to you by....
no one
defining the “government web presence”
Community Engagement

**Events**, NYAM, U-Toronto, U-Penn, UC-Riverside,

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

2008: 457 from 26 nominators
2012: 1476 from 31 nominators
2016: 1082 from 118 nominators (so far)
.gov websites proliferate like invasive species
and yes, invasivespecies.gov once existed
some are non-public or unlisted
“web waste” & preservation mentalities

TooManyWebsites.gov
JUNE 13, 2011 AT 2:50 PM ET BY MACON PHILLIPS

Summary: With almost 2,000 separate websites across the Federal Government, Americans often don't know where to turn for information. Macon Phillips, Director of the Office of Digital Strategy, discusses one of the first steps of the Campaign to Cut Waste which will target duplication and waste among federal websites.

http://TooManyWebsites.gov

Hrm.
Wayback Machine doesn't have that page archived.
End of Term Web Archive
US Federal Web Domain at Presidential Transitions

Browse by Timeframe:
- 2008-2009
- 2012-2013
- Browse All
- End of Term 2016

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.

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.

EOT at a Glance
- Crawl dates:
  - September 2008 to May 2008
  - September 2012 to March 2013
- Number of websites captured:
  - 2008: 3,035 websites
  - 2012: 3,247 websites
- Terabytes of data captured:
  - 2008: 16 TB
  - 2012: 21 TB

http://eotarchive.cdlib.org/
affiliated efforts

https://twitter.com/eotarchive

http://www.thinkingprojects.org/rabina_cocciolo_peet_EOT.pdf
affiliated efforts

2016 Election Crowdsourced Collection

White House Web & Social Media Data
In Internet Archive

- **EOT 2008**
  - ~3,000 seeds
  - ~102m URLs (~160m total across partners)
  - 17.95 TB (compressed)
  - multiple crawls & duplication

- **EOT 2012**
  - ~5,500 seeds
  - ~45m URLs (~120m total across partners)
  - 18.60 TB (compressed)
  - more focused crawls & deduped

- Similar data sizes, but 2012 had fewer URLs
- 2012 notable for media richness, uniqueness, density
eot stats 2008 and 2012

http://vphill.com/journal/post/5872/
http://vphill.com/journal/post/5861/
EOT2008 and EOT2012 Crawling Schedule

EOT 2008 and 2012 URLs Harvested By Day

- EOT2008
- EOT2012
TLD Breakdown

EOT2008 - 241 unique TLDs
EOT2012 - 251 unique TLDs
225 common TLDs
Domain Name Breakdown

EOT2008 – 87,889 unique domains
EOT2012 – 186,214 unique domains
30,066 common domains
Subdomain Breakdown

EOT2008 – 140,939 unique subdomains
EOT2012 – 352,679 unique subdomains
50,155 common subdomains
## EOT2008-EOT2012 – TLD biggest change

<table>
<thead>
<tr>
<th>TLD</th>
<th>eot2008</th>
<th>eot2012</th>
<th>Change</th>
<th>% change</th>
</tr>
</thead>
<tbody>
<tr>
<td>com</td>
<td>7,809,711</td>
<td>45,594,482</td>
<td>37,784,771</td>
<td>484%</td>
</tr>
<tr>
<td>gov</td>
<td>137,829,050</td>
<td>109,141,353</td>
<td>-28,687,697</td>
<td>-21%</td>
</tr>
<tr>
<td>mil</td>
<td>3,555,425</td>
<td>16,223,861</td>
<td>12,668,436</td>
<td>356%</td>
</tr>
<tr>
<td>net</td>
<td>653,187</td>
<td>9,269,406</td>
<td>8,616,219</td>
<td>1,319%</td>
</tr>
<tr>
<td>edu</td>
<td>3,552,509</td>
<td>2,442,626</td>
<td>-1,109,883</td>
<td>-31%</td>
</tr>
<tr>
<td>int</td>
<td>135,939</td>
<td>685,168</td>
<td>549,229</td>
<td>404%</td>
</tr>
<tr>
<td>uk</td>
<td>70,262</td>
<td>594,020</td>
<td>523,758</td>
<td>745%</td>
</tr>
<tr>
<td>ly</td>
<td>95</td>
<td>503,457</td>
<td>503,362</td>
<td>529,855%</td>
</tr>
<tr>
<td>org</td>
<td>5,108,645</td>
<td>5,588,750</td>
<td>480,105</td>
<td>9%</td>
</tr>
<tr>
<td>us</td>
<td>840,516</td>
<td>474,156</td>
<td>-366,360</td>
<td>-44%</td>
</tr>
</tbody>
</table>
Largest change by percent

<table>
<thead>
<tr>
<th></th>
<th>ly</th>
<th>me</th>
<th>gl</th>
<th>gd</th>
<th>cat</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>529854.74%</td>
<td>522651.43%</td>
<td>38452.71%</td>
<td>36266.67%</td>
<td>27116.28%</td>
</tr>
</tbody>
</table>
### .gov & .mil biggest change - 2008-2012

<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>
<tr>
<td>usgs.gov</td>
<td>4,896,493</td>
<td>1,690,295</td>
<td>-3,206,198</td>
<td>-65%</td>
</tr>
<tr>
<td>loc.gov</td>
<td>5,059,848</td>
<td>7,587,179</td>
<td>2,527,331</td>
<td>50%</td>
</tr>
<tr>
<td>hhs.gov</td>
<td>2,361,866</td>
<td>366,024</td>
<td>-1,995,842</td>
<td>-85%</td>
</tr>
<tr>
<td>osd.mil</td>
<td>180,046</td>
<td>2,111,791</td>
<td>1,931,745</td>
<td>1,073%</td>
</tr>
<tr>
<td>af.mil</td>
<td>230,920</td>
<td>2,067,812</td>
<td>1,836,892</td>
<td>795%</td>
</tr>
<tr>
<td>ed.gov</td>
<td>2,334,548</td>
<td>510,413</td>
<td>-1,824,135</td>
<td>-78%</td>
</tr>
<tr>
<td>lanl.gov</td>
<td>2,081,275</td>
<td>309,007</td>
<td>-1,772,268</td>
<td>-85%</td>
</tr>
<tr>
<td>usda.gov</td>
<td>2,892,923</td>
<td>1,324,049</td>
<td>-1,568,874</td>
<td>-54%</td>
</tr>
<tr>
<td>congress.gov</td>
<td>1,554,199</td>
<td>40,338</td>
<td>-1,513,861</td>
<td>-97%</td>
</tr>
</tbody>
</table>
## Top 30 .gov & .mil domains present in 2008 missing in 2012

<table>
<thead>
<tr>
<th>Domain</th>
<th>URLs</th>
<th>Domain</th>
<th>URLs</th>
</tr>
</thead>
<tbody>
<tr>
<td>geodata.gov</td>
<td>812,524</td>
<td>egrpra.gov</td>
<td>54,775</td>
</tr>
<tr>
<td>nifl.gov</td>
<td>504,910</td>
<td>4women.gov</td>
<td>45,684</td>
</tr>
<tr>
<td>stat-usa.gov</td>
<td>398,961</td>
<td>4woman.gov</td>
<td>42,192</td>
</tr>
<tr>
<td>tradestatsexpress.gov</td>
<td>243,729</td>
<td>nypa.gov</td>
<td>36,099</td>
</tr>
<tr>
<td>arnet.gov</td>
<td>174,057</td>
<td>nhmfl.gov</td>
<td>27,569</td>
</tr>
<tr>
<td>acqnet.gov</td>
<td>171,493</td>
<td>darpa.gov</td>
<td>21,454</td>
</tr>
<tr>
<td>dccourts.gov</td>
<td>161,289</td>
<td>usafreedomcorps.gov</td>
<td>18,001</td>
</tr>
<tr>
<td>web-services.gov</td>
<td>137,202</td>
<td>peacecore.gov</td>
<td>17,744</td>
</tr>
<tr>
<td>metrokc.gov</td>
<td>132,210</td>
<td>californiadesert.gov</td>
<td>15,172</td>
</tr>
<tr>
<td>sdi.gov</td>
<td>91,887</td>
<td>arpa.gov</td>
<td>15,093</td>
</tr>
<tr>
<td>davie-fl.gov</td>
<td>88,123</td>
<td>okgeosurvey1.gov</td>
<td>14,595</td>
</tr>
<tr>
<td>belmont.gov</td>
<td>87,332</td>
<td>omhrc.gov</td>
<td>14,594</td>
</tr>
<tr>
<td>aftac.gov</td>
<td>84,507</td>
<td>usafreedomcorp.gov</td>
<td>14,298</td>
</tr>
<tr>
<td>careervoyages.gov</td>
<td>57,192</td>
<td>uscva.gov</td>
<td>13,627</td>
</tr>
<tr>
<td>women-21.gov</td>
<td>56,255</td>
<td>odc1.gov</td>
<td>12,920</td>
</tr>
</tbody>
</table>
### Top 30 .gov & .mil domains new in 2012

<table>
<thead>
<tr>
<th>Domain</th>
<th>URLs</th>
<th>Domain</th>
<th>URLs</th>
</tr>
</thead>
<tbody>
<tr>
<td>militaryonesource.mil</td>
<td>859,843</td>
<td>broadbandmap.gov</td>
<td>72,889</td>
</tr>
<tr>
<td>consumerfinance.gov</td>
<td>237,361</td>
<td>saferproducts.gov</td>
<td>65,387</td>
</tr>
<tr>
<td>nrd.gov</td>
<td>194,215</td>
<td>challenge.gov</td>
<td>63,808</td>
</tr>
<tr>
<td>wh.gov</td>
<td>179,233</td>
<td>healthdata.gov</td>
<td>63,105</td>
</tr>
<tr>
<td>pnnl.gov</td>
<td>132,994</td>
<td>marinecadastre.gov</td>
<td>62,882</td>
</tr>
<tr>
<td>eia.gov</td>
<td>112,034</td>
<td>fatherhood.gov</td>
<td>62,132</td>
</tr>
<tr>
<td>transparency.gov</td>
<td>109,039</td>
<td>edpubs.gov</td>
<td>58,356</td>
</tr>
<tr>
<td>nationalguard.mil</td>
<td>108,854</td>
<td>transportationresearch.gov</td>
<td>58,235</td>
</tr>
<tr>
<td>acus.gov</td>
<td>93,810</td>
<td>cbca.gov</td>
<td>56,043</td>
</tr>
<tr>
<td>404.gov</td>
<td>82,409</td>
<td>usbonds.gov</td>
<td>55,102</td>
</tr>
<tr>
<td>savingsbondwizard.gov</td>
<td>76,867</td>
<td>usbond.gov</td>
<td>54,847</td>
</tr>
<tr>
<td>treasuryhunt.gov</td>
<td>76,394</td>
<td>phe.gov</td>
<td>53,626</td>
</tr>
<tr>
<td>fedshirevets.gov</td>
<td>75,529</td>
<td>ussavingsbond.gov</td>
<td>53,563</td>
</tr>
<tr>
<td>onrr.gov</td>
<td>75,484</td>
<td>scienceeducation.gov</td>
<td>53,468</td>
</tr>
<tr>
<td>veterans.gov</td>
<td>75,350</td>
<td>mda.gov</td>
<td>53,010</td>
</tr>
</tbody>
</table>
researcher access to .gov

Researchers: PoliSci, Comms, Legal, Informatics, CS
Project: Mining ~100TB of .gov data
Pros: Data w/ services, subsidized cluster, collaborative structure, some R&D
Cons: Low up-take, tech hurdles, resource constraints
Lessons Learned: Researcher use of “big data” of web archives produce challenges of scale, processing, expertise, and familiarity with context and provenance.
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, whatever)
researcher access to .gov

http://archivesunleashed.com/

http://www.websci16.org/hackathon

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

http://webarchives.ca/
wbm beta access to .gov

https://web-beta.archive.org
"captures": {
  "2001": {
    "text/html": 62,
    "text/plain": 2
  },
  "2002": {
    "text/html": 8622,
    "image/jpeg": 21,
    "image/gif": 152,
    "application/pdf": 29,
    "application/octet-stream": 13,
    "text/plain": 778
  },
  "2003": {
    "application/vnd.ms-powerpoint": 11,
    "image/jpeg": 5048,
    "text/html": 86025,
    "image/gif": 10489,
    "audio/x-wav": 1,
    "image/png": 303,
    "text/css": 46,
    "video/mpeg": 5,
    "text/plain": 7447,
    "application/octet-stream": 26,
    "application/vnd.ms-excel": 41,
    "application/zip": 86,
    "image/tiff": 5,
    "application/pdf": 42827,
    "application/msword": 557
  },
  "2004": {
    "application/vnd.ms-powerpoint": 680,
    "image/jpeg": 219375,
    "text/html": 2715426,
    "image/gif": 332445,
    "video/quicktime": 24,
    "image/png": 10665,
    "application/octet-stream": 518,
    "text/css": 1761
  }
}
wmb beta access to .gov (ppt/pdf)
Military Industrial Powerpoint Complex

United States Military

This collection was a special project done as part of the Internet Archive's 20th Anniversary celebration on October 26, 2016 highlighting IA's web archive. The collection consists of all the Powerpoint files (48,119) from the .mil web domain that were...
eot 2016
eot 2016: more partners!

Federal Government Web Archiving Working Group
timeframe for 2016 project

2016

- **July 2016**: Recruitment of subject experts/nominators to help identify additional websites for prioritized crawling started.

- **Fall/Winter 2016**: Partners will crawl various aspects of government domains at varying frequencies, depending on selection policies/interests. Team will determine strategy for crawling prioritized websites.
  - **September**: UNT crawling began
  - **November**: Bookend (baseline) crawl of government web domains began at IA. Social media archiving by GWU begins.
  - **December**: LOC crawls begin

2017

- **January 2017**:
  - January 13: Cutoff for public nominations, to be crawled before inauguration day
  - Pre- and post-Inauguration crawls done

- **Spring/Summer 2017**: Bookend crawl of all seeds, plus additional crawl of prioritized websites as determined by team.
eot 2016 opportunities

- **Expand Acquisition**
  - distribute crawling
  - deploy new tech
  - build web archiving capacity

- **Nomination and Annotation**
  - community engagement
  - contributed seed lists
  - educational opportunities

- **Researcher Engagement**
  - notable longitudinal breadth
  - good periodicity for data-mining
  - growing community of interest

- **More Partners!**
eot 2016 strategies

• Potential Project Strategies
  • distributed crawling – deduped/replay?
  • coordinated outreach – affiliate communities?
  • more listserv & project interest
  • researcher access – datasets and hosts?

• Access & Preservation
  • updated portal w/ FTS for all 3 eots
  • single replay WB
  • distributed preservation?
eot challenges

- **Same ol’ web 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 = ☹
• 9,000+ social media accounts (scrape of gov SM registry API) 44% FB, 37% TW, 10% YT

• ~190K total domain, subdomains, gov/non-gov sites

• more crowdsourced, curatorial nominations
eot 2016: how you can help

- any and all nominations welcome
- we need particular help with:
  - judicial branch websites
  - government content on non-government domains (.com, .edu, etc.)
  - important content or subdomains on very large websites (such as NASA.gov) that might be related to current Presidential policies
  - Social media
further information and the form to submit:
http://digital2.library.unt.edu/nomination/eth2016
going forward

- **Crawl it All!**
  - Community opportunity for more distributed crawling and acquisition methods

- **Access it All!**
  - Unified portal and search indices
  - New access models, user groups, analytical tools

- **Preserve it All!**
  - Take our WARCs and datasets, please!

- **Join the Fun of it All!**
  - Email: [eot2016@archive.org](mailto:eot2016@archive.org) or [uc3@ucop.edu](mailto:uc3@ucop.edu) (or any of us)

THANKS!