

End of Term Web Archive Collaborating to Preserve the US Federal Web Domain

Abbie Grotke, Library of Congress (abgr@loc.gov)

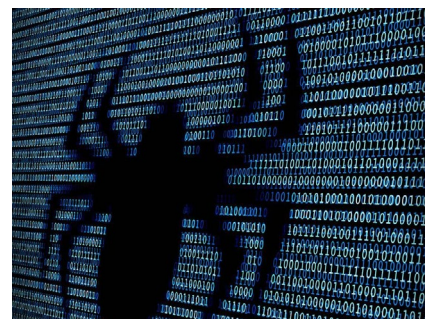
James Jacobs, Stanford University (jrjacobs@stanford.edu)

eot-info@archive.org

October 22, 2024

Agenda

- Background and history of the End of Term Archive (<https://eotarchive.org>)
- Timeline and moving parts
- Various ways to access EOT
- Next steps
- How you can help!
- Q&A



it all began a long, long, time ago, in a far away place



<https://flic.kr/p/4N2jHU>



<https://flic.kr/p/4JNkLE>

National Library of Australia

nla.int-nl39859-a11-v

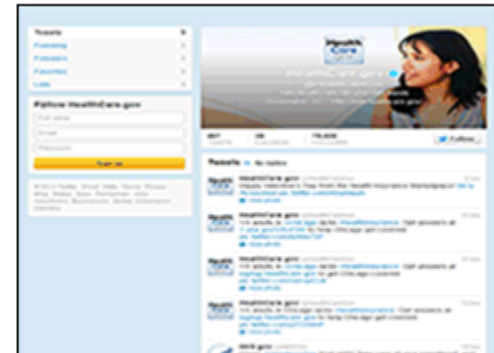
Goals of the end of term project



United States Central Command
Sept 16, 2008



U.S. Department of State Official Blog
Feb 13, 2013



Healthcare.gov Twitter
Feb 15, 2013

- Work collaboratively to preserve public U.S. Government websites
- Document federal agencies' presence on the web at 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 nuts and bolts



General Timeline of EOT Process

- Jan - Mar - Begin meeting to discuss upcoming EOT
- Mar - Apr - Set up Nomination Tool Instance
<https://digital2.library.unt.edu/nomination/>
- Apr-Sept - Begin seeking nominations
- Sept - Begin “bookend crawl” (broad scope/comprehensive crawls)
- Sept - Begin Human Nominated/Prioritized Crawls (updated every two weeks)
- Dec - Generally Initial bookend completes.
- Feb - Begin second bookend crawl
- Mar - May Copy and stage data for access at IA
- Following 3 years - Rest.

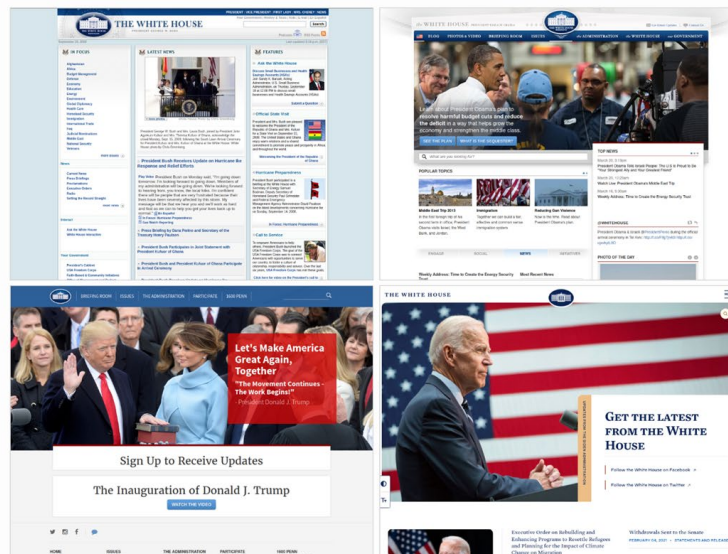
EOT Crawling Partners

	2008	2012	2016	2020	2024
Archive Team (AT)			Crawl		
California Digital Library (CDL)	Crawl				
Internet Archive (IA)	Crawl	Crawl	Crawl	Crawl	Crawl
Library of Congress (LOC)	Crawl	Crawl	Crawl		
University of North Texas (UNT)	Crawl	Crawl	Crawl	Crawl	Crawl

<https://eotarchive.org>
<https://github.com/end-of-term>
eot-info@archive.org

Purpose

The End of Term Web Archive captures and saves U.S. Government websites at the end of presidential administrations. The EOT has thus far preserved websites from administration changes in 2008, 2012, 2016, and 2020.



Whitehouse.gov captures from: September 15, 2008; March 21, 2013; February 3, 2017; and February 5, 2021.

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 government websites, or any other site not part of the federal government domain were out of scope.

U.S. Federal Government Domain End of Term 2024 Web Archive


For the End of Term 2024, The Library of Congress, University of North Texas Libraries, Internet Archive, Stanford University Libraries, U.S. Government Publishing Office (GPO), and the National Archives and Records Administration (NARA) have joined efforts again to preserve public United States Government websites at the conclusion of the presidential administration ending January 20, 2025. This web harvest – like its predecessors in 2008, 2012, 2016, and 2020 – is intended to document the federal government's presence on the World Wide Web during the transition of presidential administrations and to enhance the existing



<https://x.com/eotarchive>

@eotarchive

← **End of Term Archive**
350 Tweets


http://(gov,ed,www,)
http://(gov,ed,www,)/index.jhtml?src=a
http://(gov,ed,www,)/programs/troops/index.html
http://(gov,eda,www,)
http://(gov,edison,www,)
http://(gov,edpubs,www,)
http://(gov,education,www,)
http://(gov,educationjobsfund,www,)
http://(gov,bond,www,)
http://(gov,nc,www,)




⋮   **Following**

End of Term Archive
@eotarchive Follows you

Tweets from the project team of the End of Term Web Archive - preserving U.S. government websites during transitions in government.


eotarchive.cdlib.org  Joined September 2011

71 Following 313 Followers





 Followed by Abby McDermott, DataRefuge, and 48 others you follow

Tweets Tweets & replies Media Likes

📌 Pinned Tweet

 **End of Term Archive** @eotarchive · Sep 2, 2020

How would you like to help preserve the federal government web for future generations? We need your help! Nominate your favorite .gov now using the End of Term 2020 Nomination Tool [#WebArchiveWednesday](#) [#webarchiving](#) [#govdocs](#) digital2.library.unt.edu/nomination/eth...

  5  3 

Access to EOT Archive



Datasets

<https://eotarchive.org/>
<https://github.com/end-of-term>

End of Term Datasets

The End of Term project is working with the [Amazon Web Services' Open Data Sponsorship Program](#) to host a copy of the 2004, 2008, 2012, 2016, and 2020 End of Term Datasets.

The work of inventorying, staging and moving the data into AWS is still ongoing and more information will be provided here in the future.

Currently we have these datasets partially available for use.

Dataset	WARC #	WARC Size Compressed
EOT-2020	239811	266.04 TB
EOT-2016	194683	139.3 TB
EOT-2012	78509	41.42 TB
EOT-2008	125704	15.32 TB
EOT-2004	58977	6.42 TB

Datasets to date

Crawl	WARC Files	WARC Size	WAT Size	WET Size	CDX Size	META Size
EOT-2004	58,977	7TB	108GB	18MB	6GB	36GB
EOT-2008	125,704	15TB	447GB	108GB	9GB	68GB
EOT-2012	78,509	41TB	885GB	217GB	12GB	82GB
EOT-2016	194,683	139TB	2TB	331GB	25GB	178GB
EOT-2020	239,811	266TB	9TB	3TB	84GB	713GB
Total	638,707	468TB	12TB	4TB	136GB	1TB

Common Crawl

<https://commoncrawl.org>

“Common Crawl is a 501(c)(3) non-profit organization dedicated to providing a copy of the internet to internet researchers, companies and individuals at no cost for the purpose of research and analysis.”

- Monthly large (~300TB) crawls of the web
- Uses Nutch for crawling
- Stores data in WARC files
- Openly shares their data via AWS Open Data Sponsorship Program

The image shows the top portion of the Common Crawl website. At the top, there is a navigation bar with the following links: "BIG PICTURE", "THE DATA", "ABOUT", "BLOG", "CONNECT", and "Donate". Below the navigation bar, the word "Us" is displayed in a large, white, rounded font. To the right of "Us", there is a text box that reads: "We build and maintain an open repository of **web crawl data** that can be **accessed and analyzed by anyone**." Below this, the word "You" is displayed in a large, white, rounded font. To the right of "You", there is a text box that reads: "Need **years of free** web page data to help **change the world**." At the bottom center of the main content area, there is a small icon of a downward-pointing arrow with three dots below it.

A graphic with a dark red background. It features the text "はい C'est vrai!" at the top, "40+ languages" in the center, and "Si. Efectivamente. You bet." at the bottom. There are also some small icons and symbols.

A graphic with a dark grey background. It lists three types of data: "RAW DATA", "METADATA", and "TEXT DATA" in white, uppercase letters.

A graphic with a dark teal background. It features a large, light blue "\$0" at the top. Below it, the text reads: "We gather it. We aggregate it. You utilize it. **And it's all free.**"

A graphic with a light blue background. It features the text "HOW BIG? WE'RE TALKING **BIG** PETABYTES BIG" in white and dark blue.

A graphic with a dark green background. It features the text "billions OF PAGES" and "trillions OF LINKS" in white and light green. There is also a small speech bubble that says "say what?"

A graphic with a dark red background. It features a stylized fist icon and the text "our story »" in white.

A graphic with a yellow background. It features a large number "7" inside a circle and the text "YEARS OF DATA" in white.

Explore more than 866 billion [web pages](#) saved over time

Enter a URL or words related to a site's home page



Subscription Service

Archive-It enables you to capture, manage and search collections of digital content without any technical expertise or hosting facilities. [Visit Archive-It to build and browse the collections.](#)



Collection Search

Enter any keyword

- ✓ End Of Term (US Gov) 2008
- End Of Term (US Gov) 2012
- End Of Term (US Gov) 2016
- End Of Term (US Gov) 2020
- FOIAonline.gov PDFs
- Hong Kong news organizations that have been shut down
- badoo.com
- cmt.com/news
- congress.gov
- COVID
- dcist.com
- exiledonline.com
- qawker.com



Save Page Now

https://

SAVE PAGE

Save page as it appears now for use as a snapshot in the future.

[FAQ](#) | [Contact Us](#) | [Terms](#)



The Wayback Machine is an initiative of building a digital library of Internet sites. Other projects include [Open Library](#) & [Internet Archive](#).

Your use of the Wayback Machine is subject to our [Terms of Use](#).

End of Term Publications

During the 2016 End of Term project we identified all of the PDF documents that had been nominated for capture.

These totaled over 1,900.

We extracted these from our crawls and built a digital collection for these in the UNT Digital Library

We worked with volunteers to create metadata records these documents so they could be easily accessed.

The screenshot shows the UNT Digital Library website. The main content area is titled 'End of Term Publications'. It includes a search bar with the text 'Search inside this Collection' and a 'Search' button. Below the search bar is an 'At a Glance' section with a table of statistics:

1,945 Items	17 Types	135 Titles
2 Partners	5 Decades	2 Languages
24 Counties	50 States	41 Countries
332,456 Usage	6 years, 3 months ago Collection Created	3 years, 9 months ago Last Updated

<https://digital.library.unt.edu/explore/collections/EOT/>

Extracted Special Web Collections

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 (48,110) from the .mil web domain that were

1,064 RESULTS

Search this Collection

PART OF
American Libraries

Media Type
 texts 1,064

Collection
 American Libraries 1,064
 Military Industrial 1,064
 Powerpoint Complex
 gio's favorites 1

Language
 English 1,062

COLLECTION

Sort by: VIEWS · TITLE · DATE PUBLISHED · CREATOR

The Concept
dot-mil powerpoint presentations from the
454 0 0

SE97-1
20 Feb 1997
MATERIEL SAFETY TASK DATABASE (DB10)

wp afb
93 0 0

Integrated Missile Defense 3-02 Experiment Event Integrator Brief For BGen Obering
28 March 2009
Force Structure, Integration and Missile Defense Agency
mil-powerpoints-100
90 0 0

fort lee
89 0 0

Deployment Stats
DINLNS 3.0K of 73 items
88.2% Complete
93.1% 44.1a using
DINLNS 3.0K
95.68% 4.2a using
8.6%

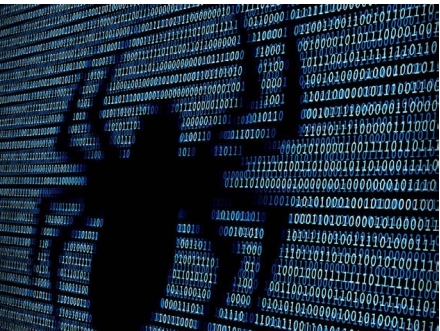
Fire Safety Basics
• Teach the basics of fire safety
• Matches and lighters are for groupups only
• Never touch matches or lighters
• Tell an adult if they find matches or lighters

army training support center
88 0 0

Good Morning e-Publishing

<https://archive.org/details/MilitaryIndustrialPowerpointComplex>

Next Steps



How you can help with EOT

- 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
 - Official social media accounts

End of Term Presidential Harvest 2024

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

Capture the Following Site

Website URL: *
The URL you wish to nominate for capture

Metadata

Title:
Title of this URL.

Branch:
Branch of government.

Executive
 Judicial
 Legislative

Agency:
Government agency.

<https://digital2.library.unt.edu/nomination/eth2024/>
Or <https://eotarchive.org>

We're also targeting databases!



eLibrary

Federal Energy Regulatory Commission

Search eLibrary

* Indicates a required field.

General Search Docket Search New Docket Only

Search on a Reference Number (Docket, Accession, Ferc Cite, etc.)

Docket (e.g. ER11 , ER11-4046 , ER11-4040-0201) Sub-Dockets (eg. 001, 002)

Select Date Range (required)

Filed Date *From *To ?

Keyword Search Description Full Text

Select Category Issuance Submittal Industry Sector [What are Docum](#)

Search By Recipient,Author,Agent Role

Role FI MI Organization



<https://forms.gle/sg2UDSn6BjYoPsKv9>

Questions / Discussion

AMA (about EOT!)

James R. Jacobs jrjacobs@stanford.edu

eot-info@archive.org

