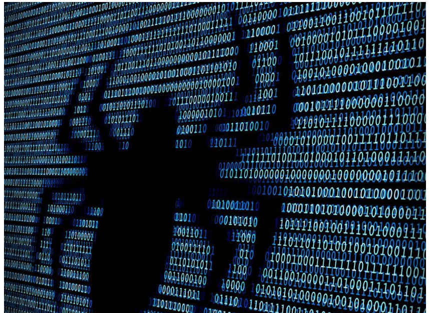
End of Term Web Archive Collaborating to Preserve the US Federal Web Domain

James Jacobs, Stanford University (jrjacobs@stanford.edu)
eot-info@archive.org

December 05, 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

Goals of the end of term project

UNITED STATES CENTRAL COMMAND
United States Central Command – Sep. 16, 2008

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

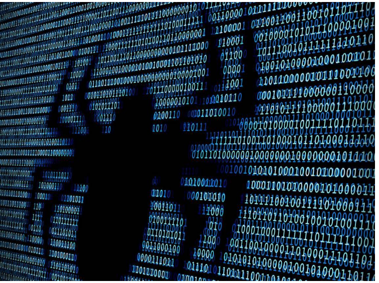
□ document federal agencies' presence on the web at the end of Presidential terms

□ 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

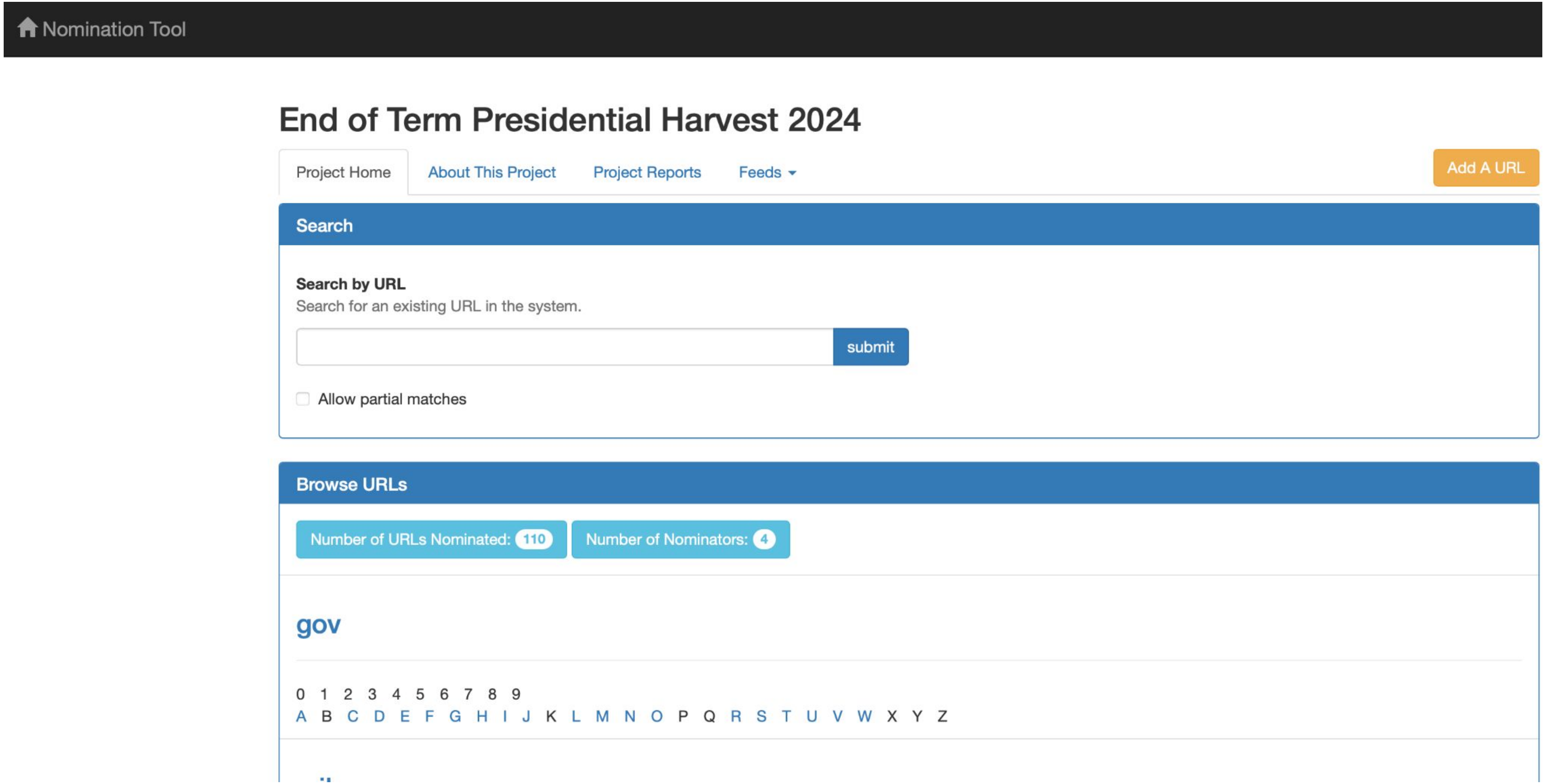


General Timeline of EOT Process

- 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



End of Term Presidential Harvest 2024

Search

Search by URL

Search for an existing URL in the system.

□ Allow partial matches

Browse URLs

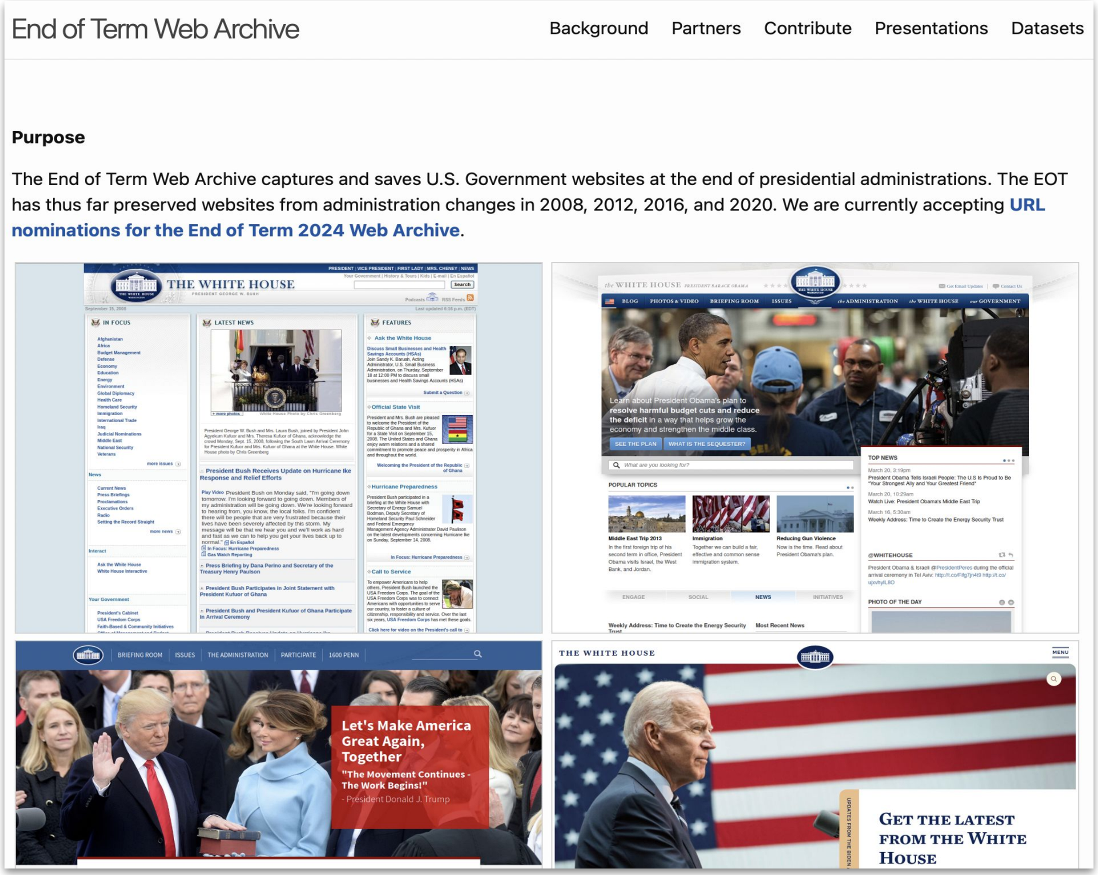
Number of URLs Nominated: 110

Number of Nominators: 4

gov

0 1 2 3 4 5 6 7 8 9
A B C D E F G H I J K L M N O P Q R S T U V W X Y Z

https://digital2.library.unt.edu/nomination/eth2024/



End of Term Web Archive

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. We are currently accepting **URL nominations for the End of Term 2024 Web Archive**.

THE WHITE HOUSE
PRESIDENT GEORGE W. BUSH
IN FOCUS
LATEST NEWS
FEATURES
THE WHITE HOUSE PRESIDENT BARACK OBAMA
TOP NEWS
POPULAR TOPICS
BRIEFING ROOM
ISSUES
THE ADMINISTRATION
PARTICIPATE
1600 PENN
Let's Make America Great Again, Together
"The Movement Continues - The Work Begins!"
- President Donald J. Trump
THE WHITE HOUSE
MENU
Get the latest from the White House

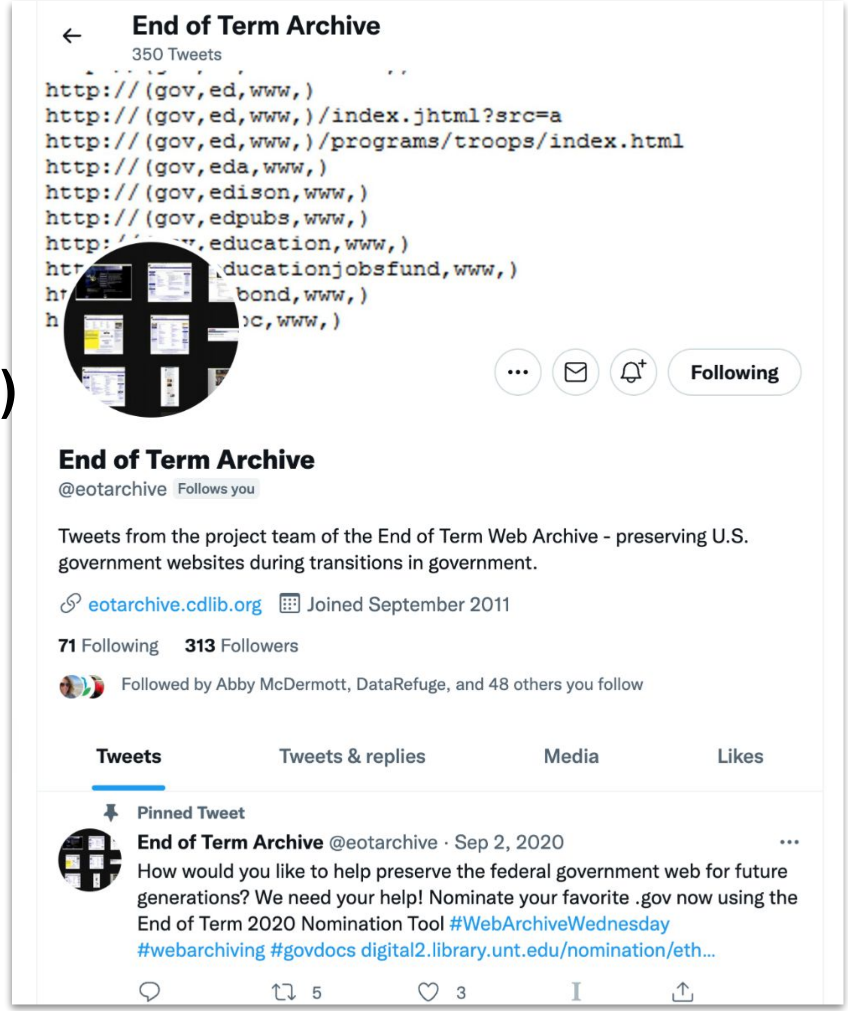
https://eotarchive.org, https://github.com/end-of-term
eot-info@archive.org

Goals of the Project

- Provide greater access to the End of Term datasets
  - 2008, 2012, 2016, 2020
- Focused on computational consumption of the collection
- Currently challenging because of size, access, storage issues
- Encourage reuse and research with the EOT data
- Position dataset so that we can learn more about our process
- Provide a canonical dataset for each crawl for reference numbers like size, URLs, counts

x.com/eotarchive (@eotarchive)

bsky.app/profile/eotarchive.org (@eotarchive.org)



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:/      ,education,www,)
htt          ducationjobsfund,www,)
ht            bond,www,)
h             oc,www,)
```

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

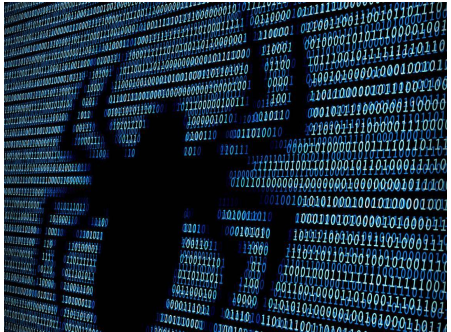
Followed by Abby McDermott, DataRefuge, and 48 others you follow

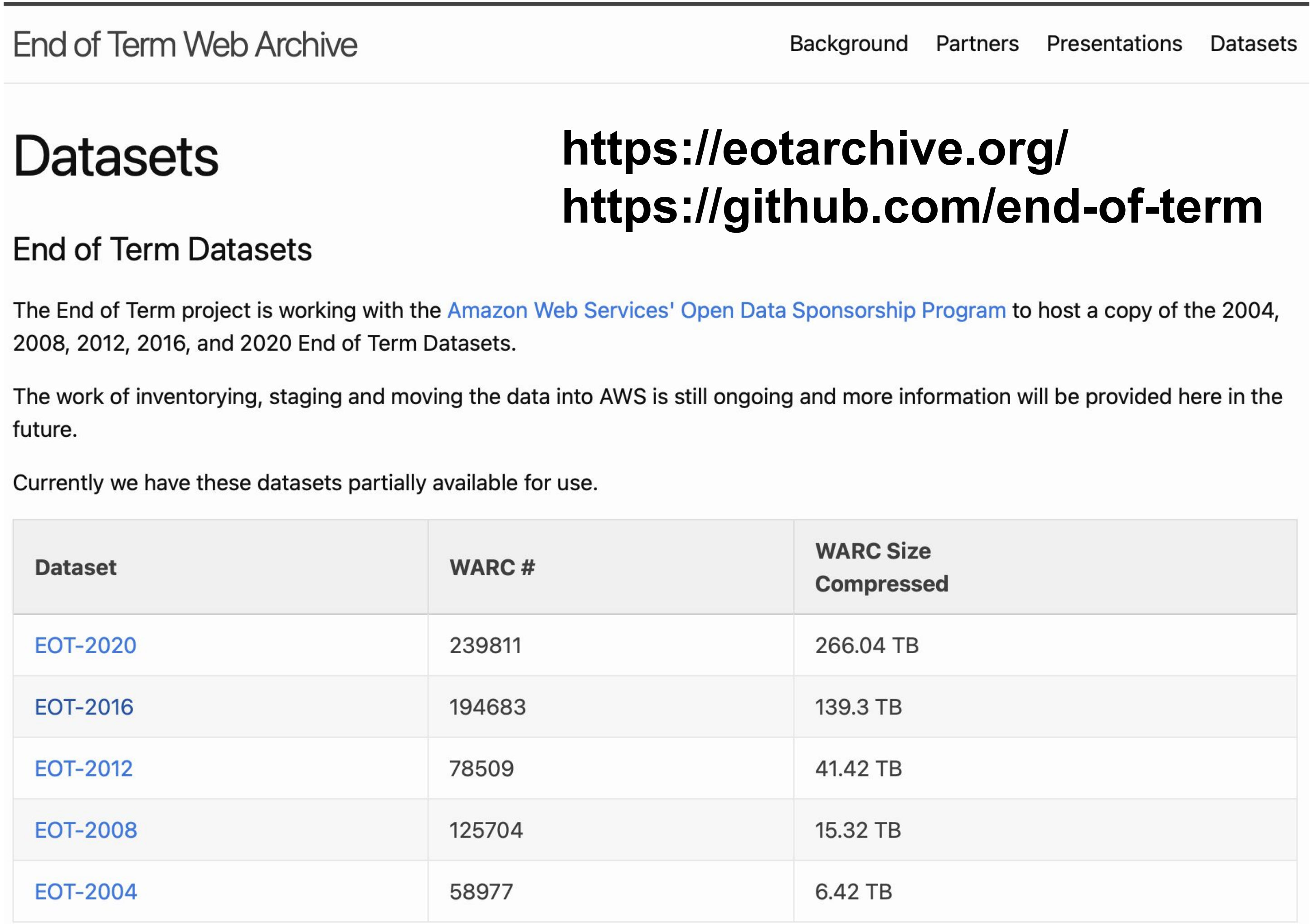
Pinned Tweet

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...

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

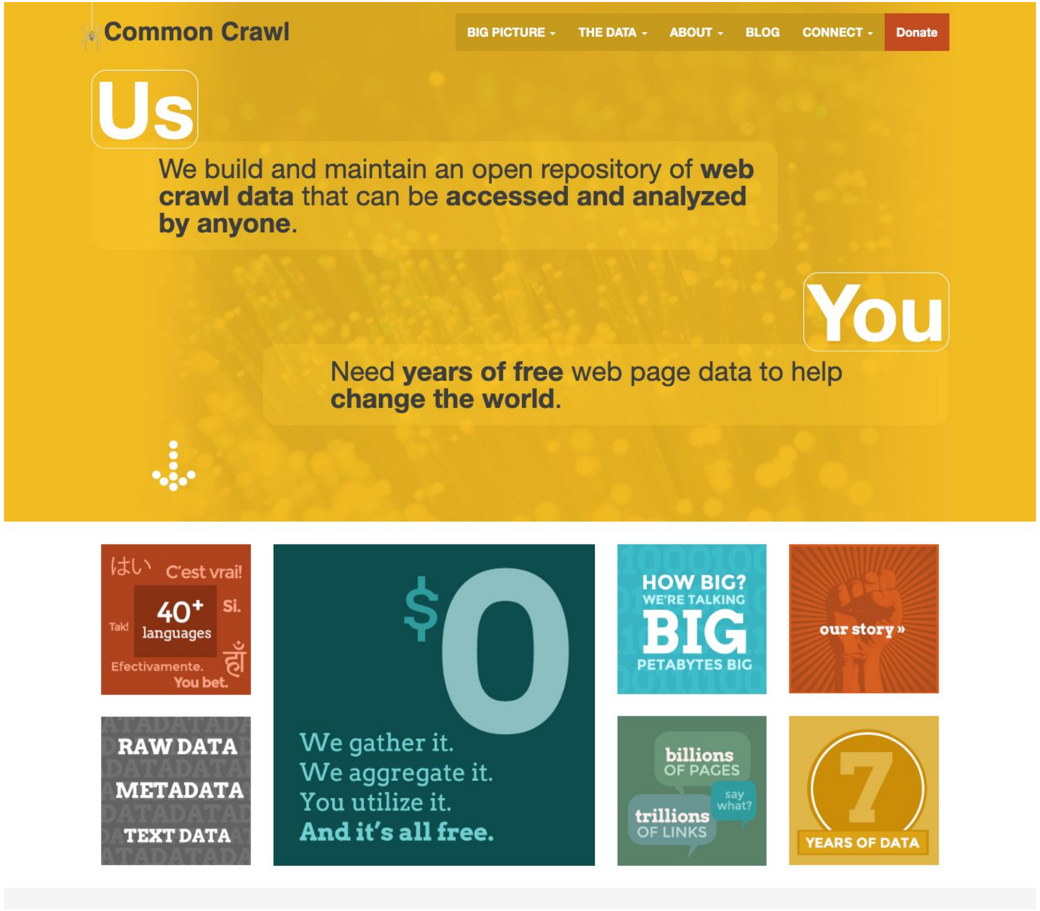
Crawl	WARC Files	WARC Size	WAT Size	WET Size	CDX Size	META Size
EOT-2004	58,977	7TB	108GB	18MB	6GB	36GB
EOT-2008	125,704	15TB	447GB	108GB	9GB	68GB
EOT-2012	78,509	41TB	885GB	217GB	12GB	82GB
EOT-2016	194,683	139TB	2TB	331GB	25GB	178GB
EOT-2020	239,811	266TB	9TB	3TB	84GB	713GB
Total	638,707	468TB	12TB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

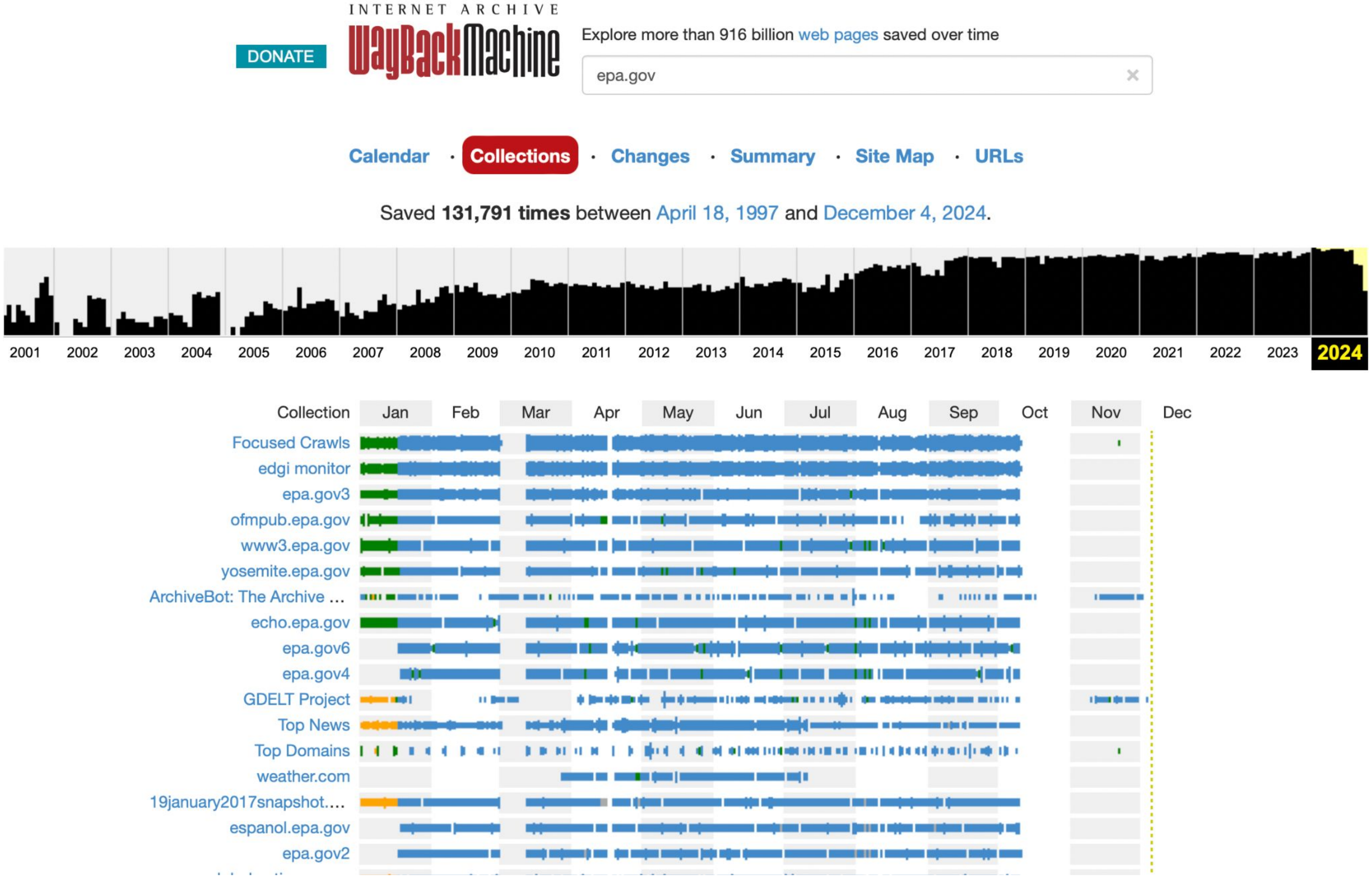
- Monthly large (~300TB) crawls of the web
- Uses Nutch for crawling
- Stores data in WARC files
- Openly shares their data via AWS Open Data Sponsorship Program



We build and maintain an open repository of **web crawl data** that can be **accessed and analyzed by anyone**.

Need **years of free** web page data to help **change the world**.

We gather it.
We aggregate it.
You utilize it.
And it's all free.



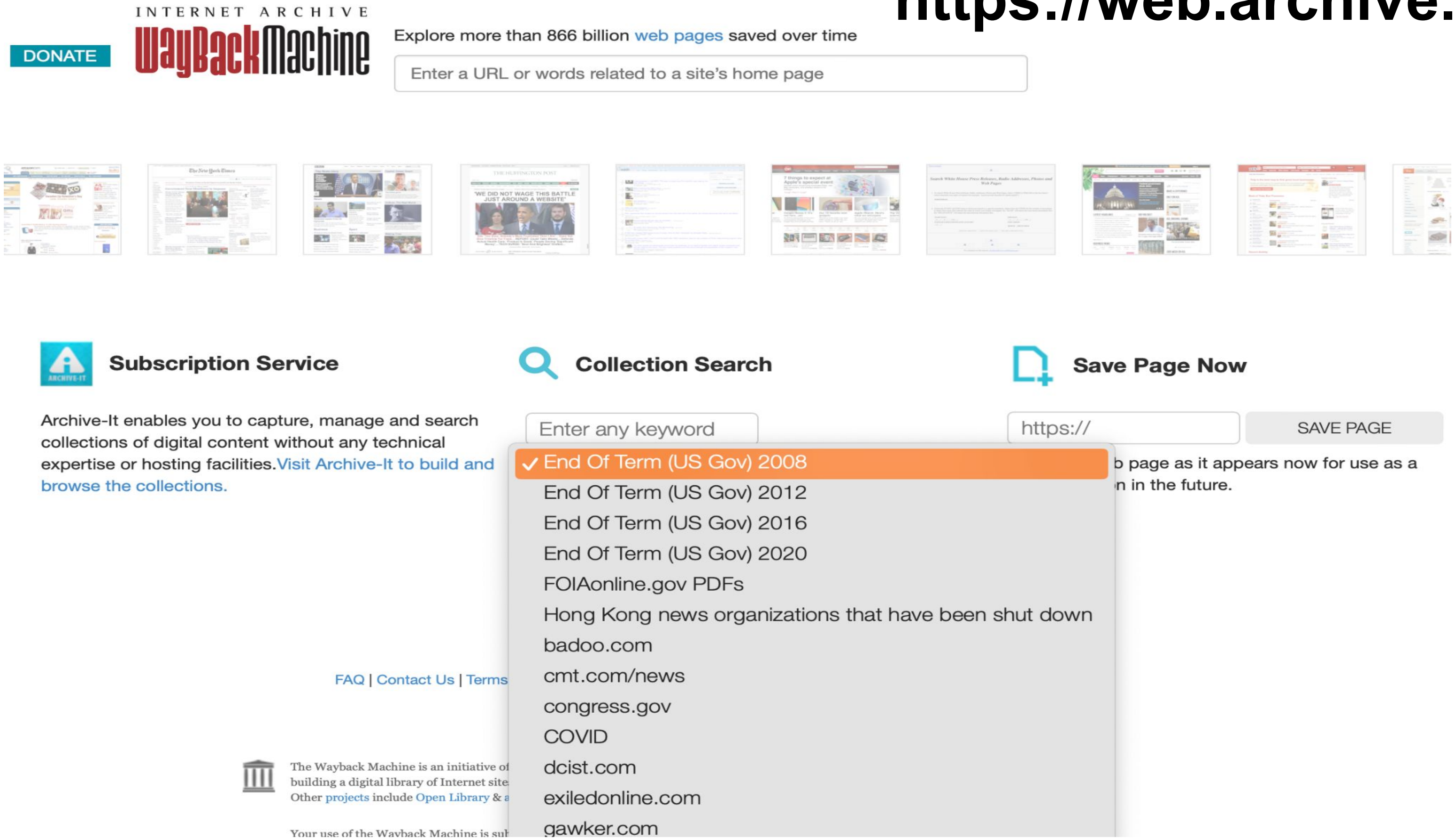
Explore more than 916 billion web pages saved over time

epa.gov

Saved **131,791 times** between April 18, 1997 and December 4, 2024.

https://web.archive.org

https://web.archive.org



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
- gawker.com

Save Page Now

https://

SAVE PAGE

b page as it appears now for use as a
n in the future.

FAQ | Contact Us | Terms

The Wayback Machine is an initiative of
building a digital library of Internet site
Other projects include Open Library & a

Your use of the Wayback Machine is sub

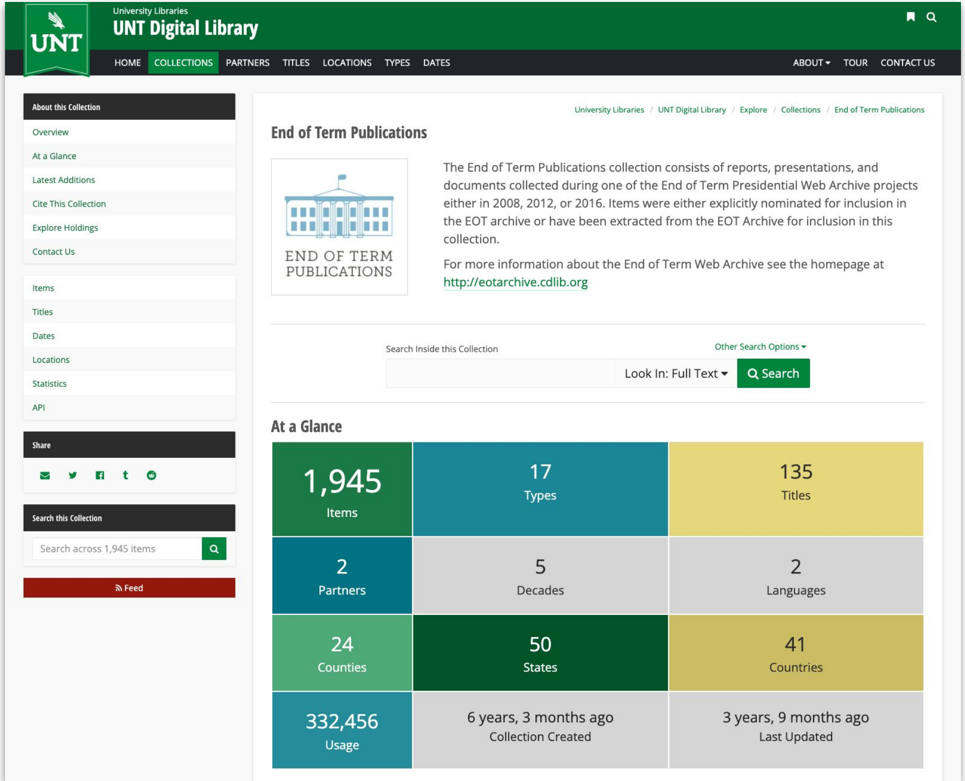
End of Term Publications

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

These totalled 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.



End of Term Publications

The End of Term Publications collection consists of reports, presentations, and documents collected during one of the End of Term Presidential Web Archive projects either in 2008, 2012, or 2016. Items were either explicitly nominated for inclusion in the EOT archive or have been extracted from the EOT Archive for inclusion in this collection.

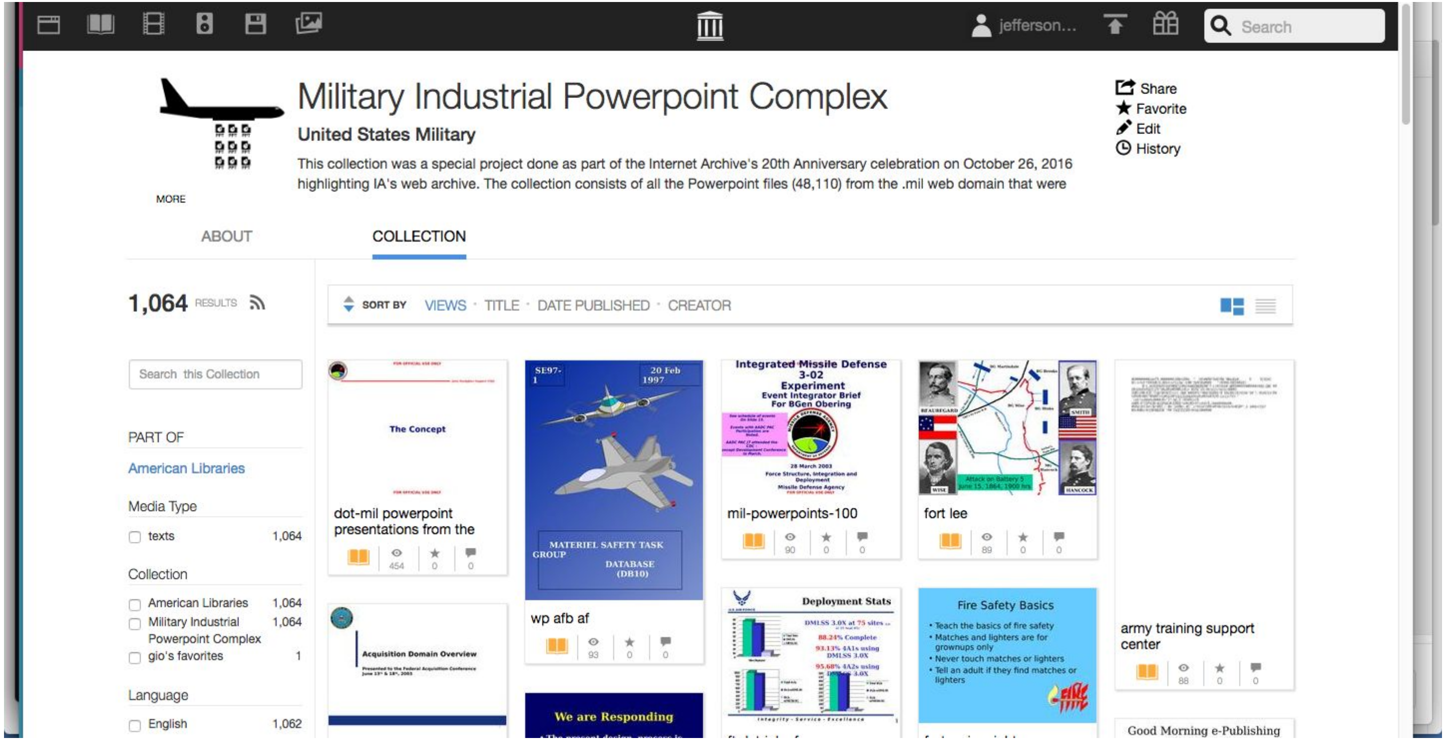
For more information about the End of Term Web Archive see the homepage at http://eotarchive.cdlib.org

At a Glance

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

PART OF

American Libraries

Media Type

□ texts	1,064

Collection

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

Language

□ English	1,062

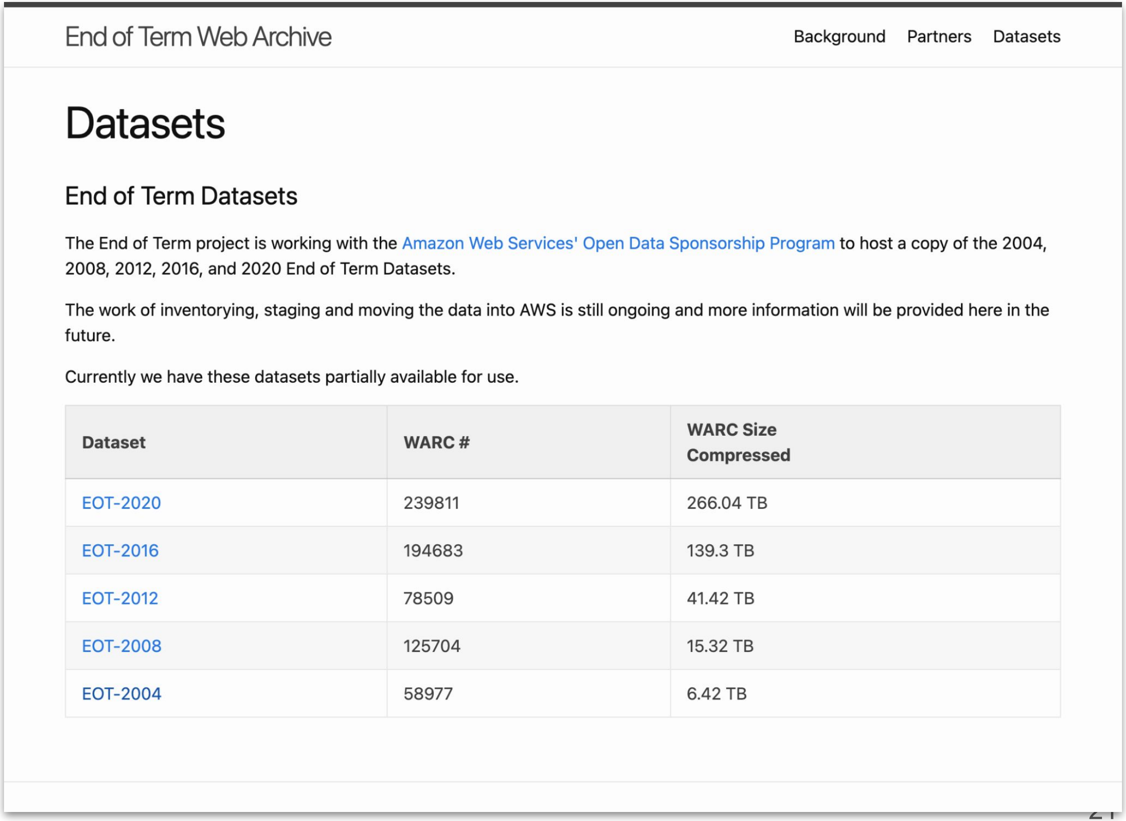
Title	Views	Favorites	Comments
dot-mil powerpoint presentations from the	454	0	0
wp afb af	93	0	0
mil-powerpoints-100	90	0	0
fort lee	89	0	0
army training support center	88	0	0

https://archive.org/details/MilitaryIndustrialPowerpointComplex

EOT Datasets

Where to get the datasets

https://eotarchive.org/data/



Datasets

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

Dataset overview

Download with HTTP or S3

Path files contain full paths to each file in dataset.

Download path files and then iterate over all lines in file to retrieve full dataset

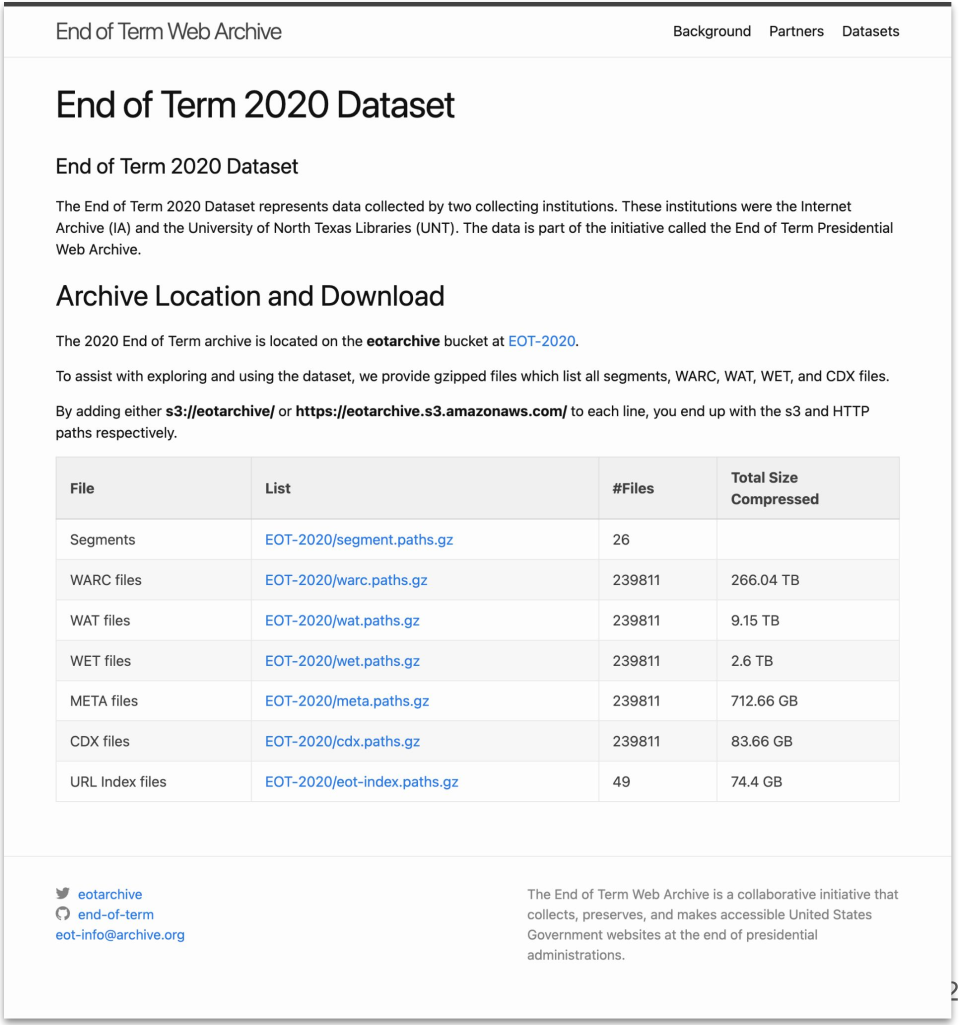
Take the parts you need

If you have questions reach out.

mark.phillips@unt.edu

kristyphillips@my.unt.edu

sawood@archive.org



End of Term 2020 Dataset

End of Term 2020 Dataset

The End of Term 2020 Dataset represents data collected by two collecting institutions. These institutions were the Internet Archive (IA) and the University of North Texas Libraries (UNT). The data is part of the initiative called the End of Term Presidential Web Archive.

Archive Location and Download

The 2020 End of Term archive is located on the **eotarchive** bucket at EOT-2020.

To assist with exploring and using the dataset, we provide gzipped files which list all segments, WARC, WAT, WET, and CDX files.

By adding either **s3://eotarchive/** or **https://eotarchive.s3.amazonaws.com/** to each line, you end up with the s3 and HTTP paths respectively.

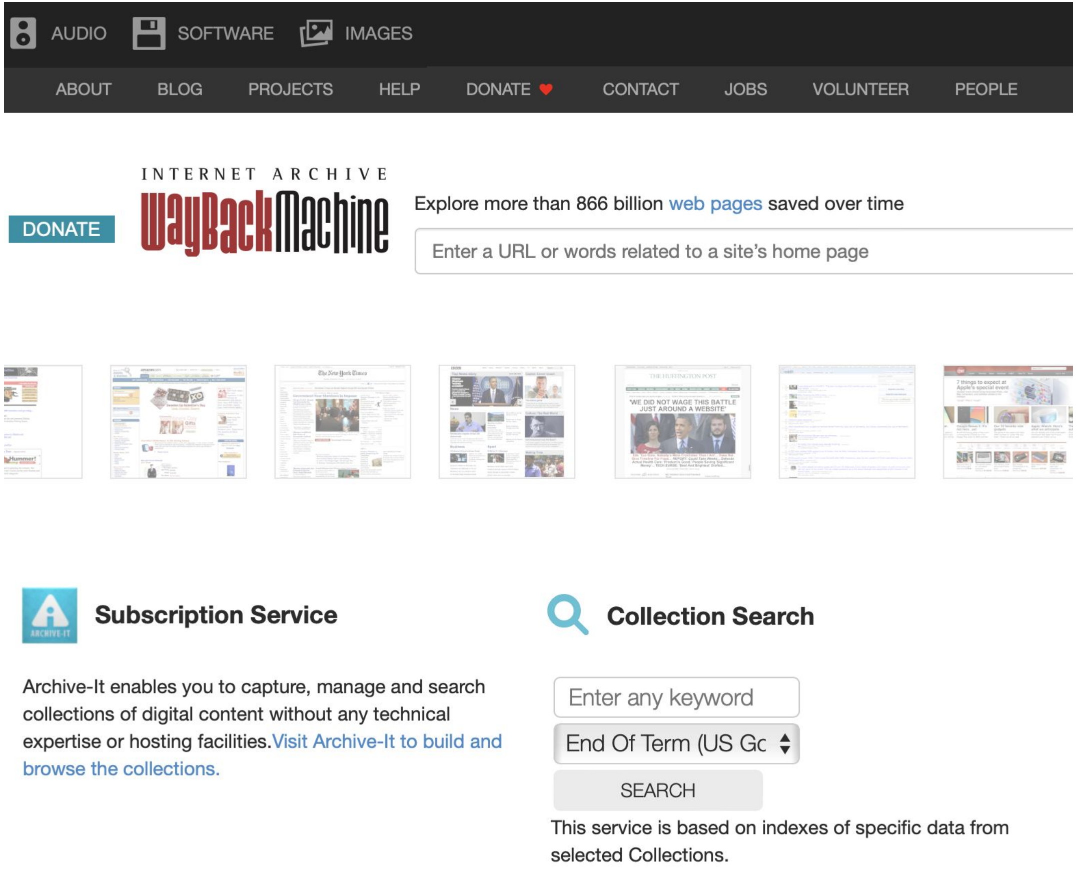
File	List	#Files	Total Size Compressed
Segments	EOT-2020/segment.paths.gz	26	
WARC files	EOT-2020/warc.paths.gz	239811	266.04 TB
WAT files	EOT-2020/wat.paths.gz	239811	9.15 TB
WET files	EOT-2020/wet.paths.gz	239811	2.6 TB
META files	EOT-2020/meta.paths.gz	239811	712.66 GB
CDX files	EOT-2020/cdx.paths.gz	239811	83.66 GB
URL Index files	EOT-2020/eot-index.paths.gz	49	74.4 GB

eotarchive
end-of-term
eot-info@archive.org

The End of Term Web Archive is a collaborative initiative that collects, preserves, and makes accessible United States Government websites at the end of presidential administrations.

Where to search the data

https://web.archive.org/



Explore more than 866 billion web pages saved over time

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

Subscription Service

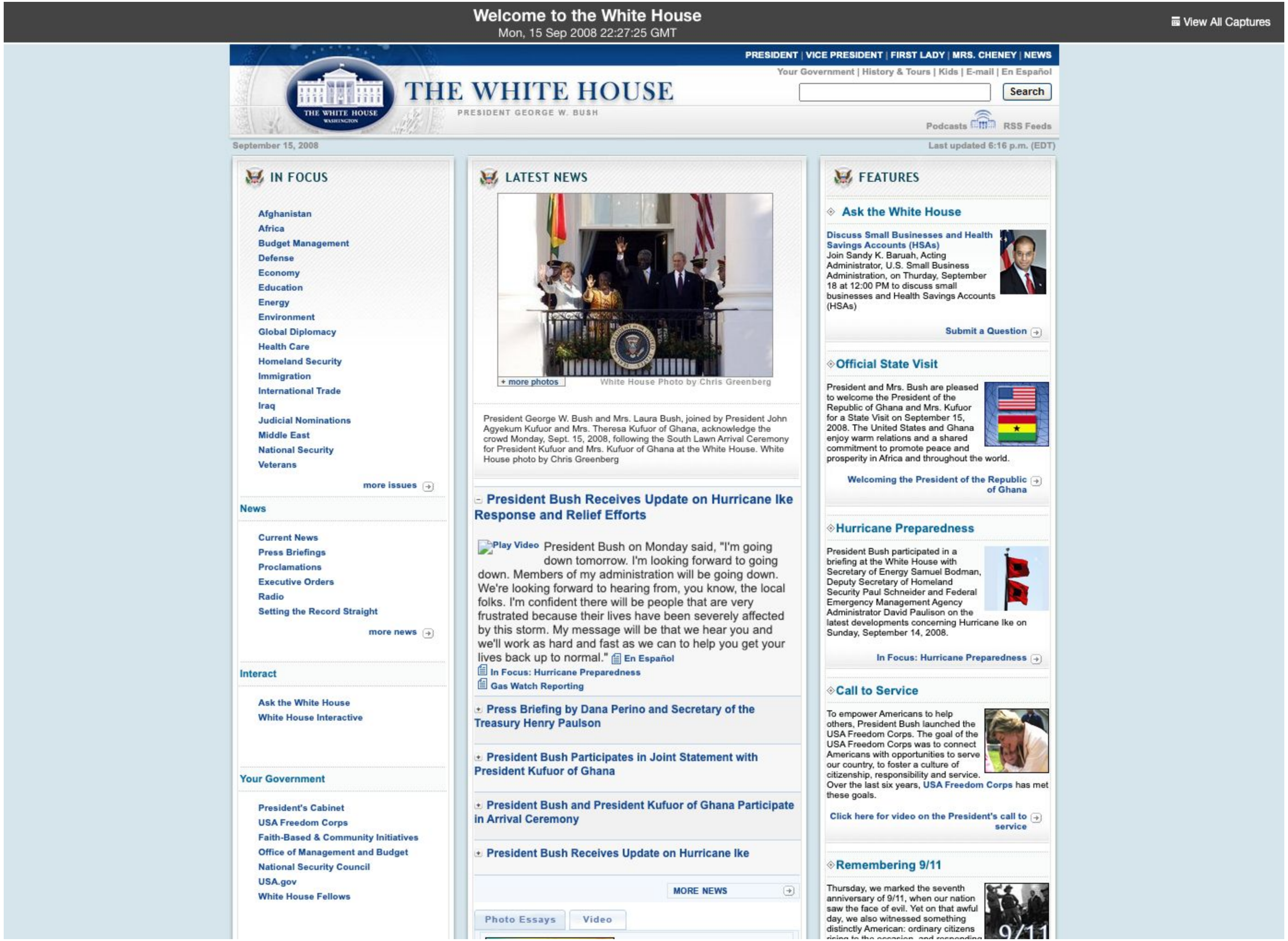
Archive-It enables you to capture, manage and search collections of digital content without any technical expertise or hosting facilities. Visit Archive-It to build and browse the collections.

Collection Search

Enter any keyword

End Of Term (US Gc

This service is based on indexes of specific data from selected Collections.



Welcome to the White House
Mon, 15 Sep 2008 22:27:25 GMT

View All Captures

PRESIDENT | VICE PRESIDENT | FIRST LADY | MRS. CHENEY | NEWS

Your Government | History & Tours | Kids | E-mail | En Español

THE WHITE HOUSE

PRESIDENT GEORGE W. BUSH

Search

Podcasts RSS Feeds

September 15, 2008

Last updated 6:16 p.m. (EDT)

IN FOCUS

- Afghanistan
- Africa
- Budget Management
- Defense
- Economy
- Education
- Energy
- Environment
- Global Diplomacy
- Health Care
- Homeland Security
- Immigration
- International Trade
- Iraq
- Judicial Nominations
- Middle East
- National Security
- Veterans

more issues

News

- Current News
- Press Briefings
- Proclamations
- Executive Orders
- Radio
- Setting the Record Straight

more news

Interact

- Ask the White House
- White House Interactive

Your Government

- President's Cabinet
- USA Freedom Corps
- Faith-Based & Community Initiatives
- Office of Management and Budget
- National Security Council
- USA.gov
- White House Fellows

LATEST NEWS

+ more photos

White House Photo by Chris Greenberg

President George W. Bush and Mrs. Laura Bush, joined by President John Agyekum Kufuor and Mrs. Theresa Kufuor of Ghana, acknowledge the crowd Monday, Sept. 15, 2008, following the South Lawn Arrival Ceremony for President Kufuor and Mrs. Kufuor of Ghana at the White House. White House photo by Chris Greenberg

President Bush Receives Update on Hurricane Ike Response and Relief Efforts

Play Video President Bush on Monday said, "I'm going down tomorrow. I'm looking forward to going down. Members of my administration will be going down. We're looking forward to hearing from, you know, the local folks. I'm confident there will be people that are very frustrated because their lives have been severely affected by this storm. My message will be that we hear you and we'll work as hard and fast as we can to help you get your lives back up to normal." En Español

In Focus: Hurricane Preparedness

Gas Watch Reporting

President Bush Participates in Joint Statement with President Kufuor of Ghana

President Bush and President Kufuor of Ghana Participate in Arrival Ceremony

President Bush Receives Update on Hurricane Ike

MORE NEWS

Photo Essays | Video

FEATURES

Ask the White House

Discuss Small Businesses and Health Savings Accounts (HSAs)
Join Sandy K. Baruah, Acting Administrator, U.S. Small Business Administration, on Thurday, September 18 at 12:00 PM to discuss small businesses and Health Savings Accounts (HSAs)

Submit a Question

Official State Visit

President and Mrs. Bush are pleased to welcome the President of the Republic of Ghana and Mrs. Kufuor for a State Visit on September 15, 2008. The United States and Ghana enjoy warm relations and a shared commitment to promote peace and prosperity in Africa and throughout the world.

Welcoming the President of the Republic of Ghana

Hurricane Preparedness

President Bush participated in a briefing at the White House with Secretary of Energy Samuel Bodman, Deputy Secretary of Homeland Security Paul Schneider and Federal Emergency Management Agency Administrator David Paulison on the latest developments concerning Hurricane Ike on Sunday, September 14, 2008.

In Focus: Hurricane Preparedness

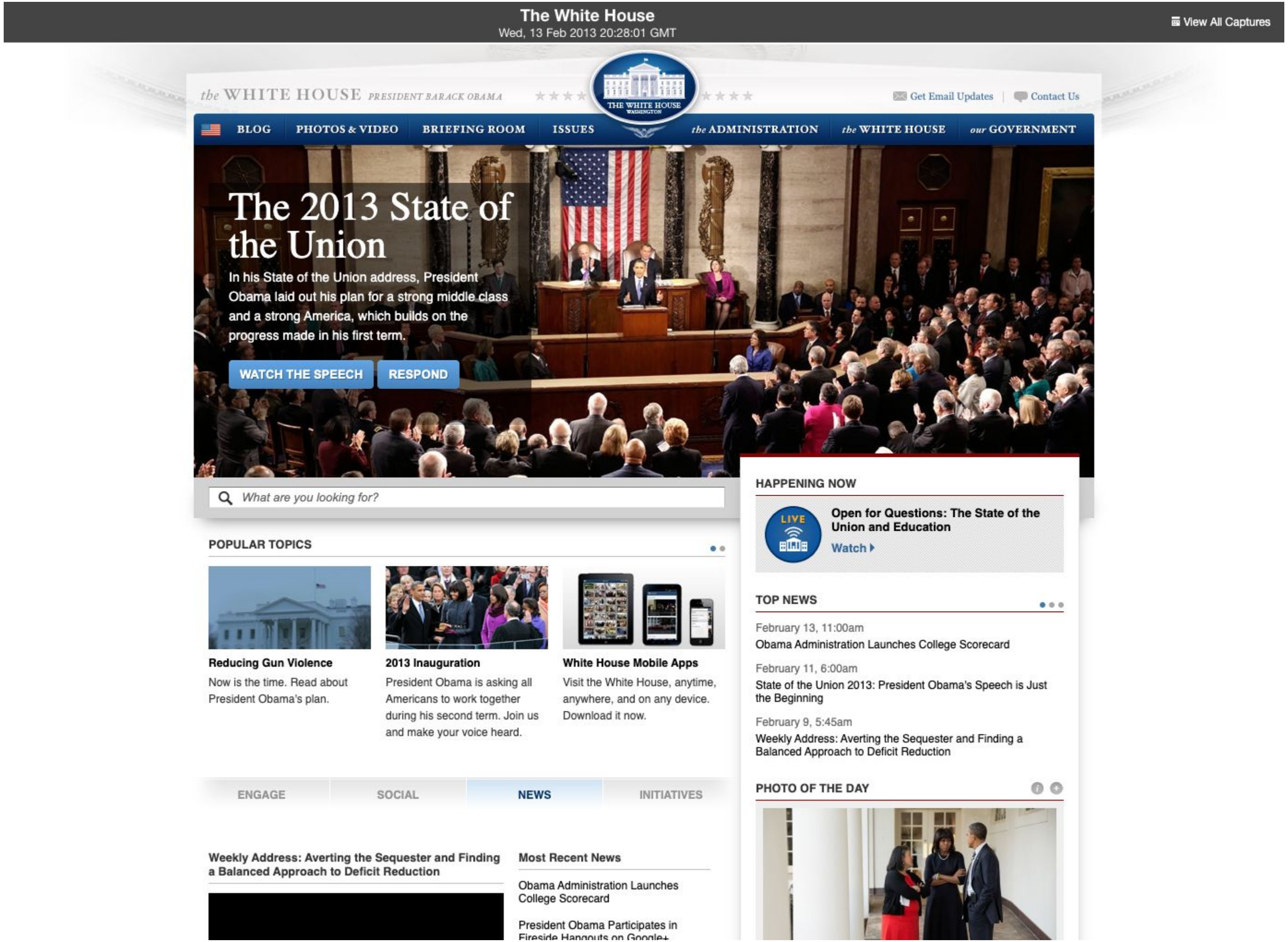
Call to Service

To empower Americans to help others, President Bush launched the USA Freedom Corps. The goal of the USA Freedom Corps was to connect Americans with opportunities to serve our country, to foster a culture of citizenship, responsibility and service. Over the last six years, USA Freedom Corps has met these goals.

Click here for video on the President's call to service

Remembering 9/11

Thursday, we marked the seventh anniversary of 9/11, when our nation saw the face of evil. Yet on that awful day, we also witnessed something distinctly American: ordinary citizens rising to the occasion, and responding



The White House
Wed, 13 Feb 2013 20:28:01 GMT

View All Captures

the WHITE HOUSE PRESIDENT BARACK OBAMA

Get Email Updates | Contact Us

BLOG · PHOTOS & VIDEO · BRIEFING ROOM · ISSUES · the ADMINISTRATION · the WHITE HOUSE · our GOVERNMENT

The 2013 State of the Union

In his State of the Union address, President Obama laid out his plan for a strong middle class and a strong America, which builds on the progress made in his first term.

WATCH THE SPEECH · RESPOND

What are you looking for?

POPULAR TOPICS

Reducing Gun Violence
Now is the time. Read about President Obama's plan.

2013 Inauguration
President Obama is asking all Americans to work together during his second term. Join us and make your voice heard.

White House Mobile Apps
Visit the White House, anytime, anywhere, and on any device. Download it now.

ENGAGE · SOCIAL · NEWS · INITIATIVES

Weekly Address: Averting the Sequester and Finding a Balanced Approach to Deficit Reduction

Most Recent News

Obama Administration Launches College Scorecard

President Obama Participates in Fireside Hangouts on Google+

HAPPENING NOW

LIVE

Open for Questions: The State of the Union and Education
Watch ▸

TOP NEWS

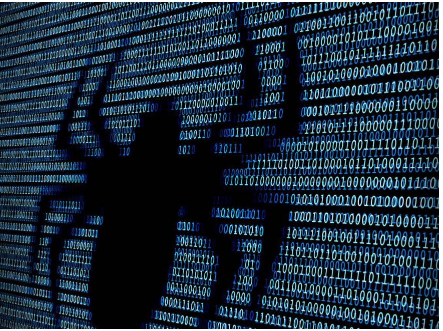
February 13, 11:00am
Obama Administration Launches College Scorecard

February 11, 6:00am
State of the Union 2013: President Obama's Speech is Just the Beginning

February 9, 5:45am
Weekly Address: Averting the Sequester and Finding a Balanced Approach to Deficit Reduction

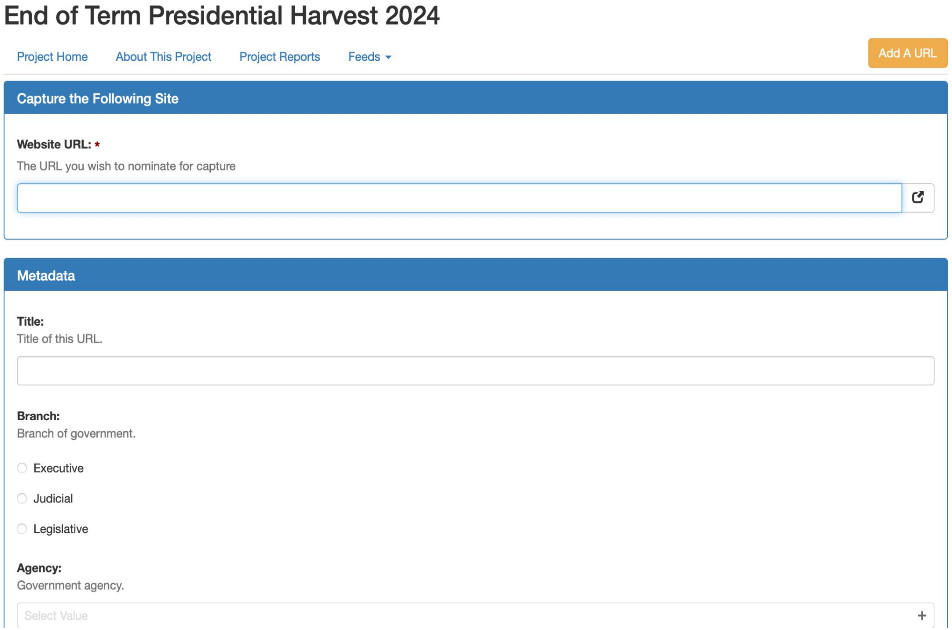
PHOTO OF THE DAY

Next Steps



How you can help with EOT

- □ any and all nominations welcome (single urls or bulk seed lists!)
- □ 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

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.

Select Value +

https://eotarchive.org/contribute/

How to nominate seeds

https://www.energy.gov

https://www.energy.gov/ig/calendar-year-reports

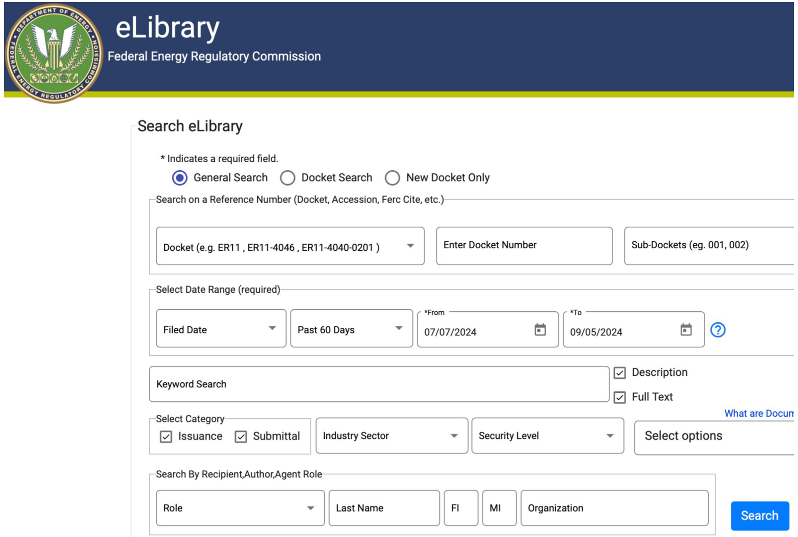
https://www.energy.gov/ig/office-inspector-general

https://www.energy.gov/lm/historical-resources

https://www.energy.gov/management/articles/fehner-and-gosling-atmospheric-nuclear-weapons-testing-1951-1963-battlefield

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

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) | Enter Docket Number | Sub-Dockets (eg. 001, 002)

Select Date Range (required)

Filed Date | Past 60 Days | *From 07/07/2024 | *To 09/05/2024

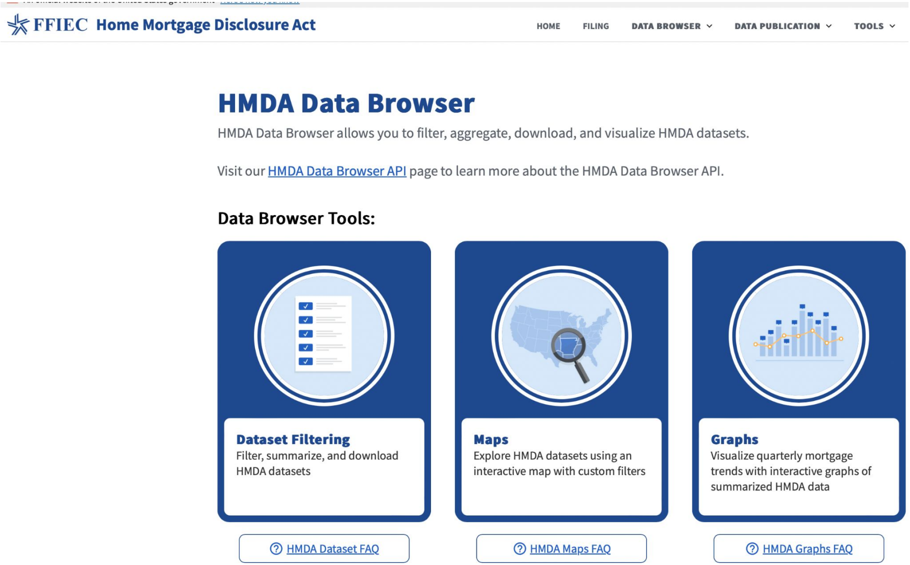
Keyword Search | ☑ Description | ☑ Full Text

What are Docum

Select Category: ☑ Issuance ☑ Submittal | Industry Sector | Security Level | Select options

Search By Recipient,Author,Agent Role

Role | Last Name | FI | MI | Organization



FFIEC Home Mortgage Disclosure Act

HMDA Data Browser

HMDA Data Browser allows you to filter, aggregate, download, and visualize HMDA datasets.

Visit our HMDA Data Browser API page to learn more about the HMDA Data Browser API.

Data Browser Tools:

Dataset Filtering
Filter, summarize, and download HMDA datasets

Maps
Explore HMDA datasets using an interactive map with custom filters

Graphs
Visualize quarterly mortgage trends with interactive graphs of summarized HMDA data

https://forms.gle/sg2UDSn6BjYoPsKv9

https://eotarchive.org/contribute/

Questions / Discussion

AMA (about EOT!)

James R. Jacobs jrjacobs@stanford.edu

eot-info@archive.org