Hosting the End of Term Web Archive in the Cloud

Mark Phillips
Library of Congress Designing Storage Architectures
March 24, 2025

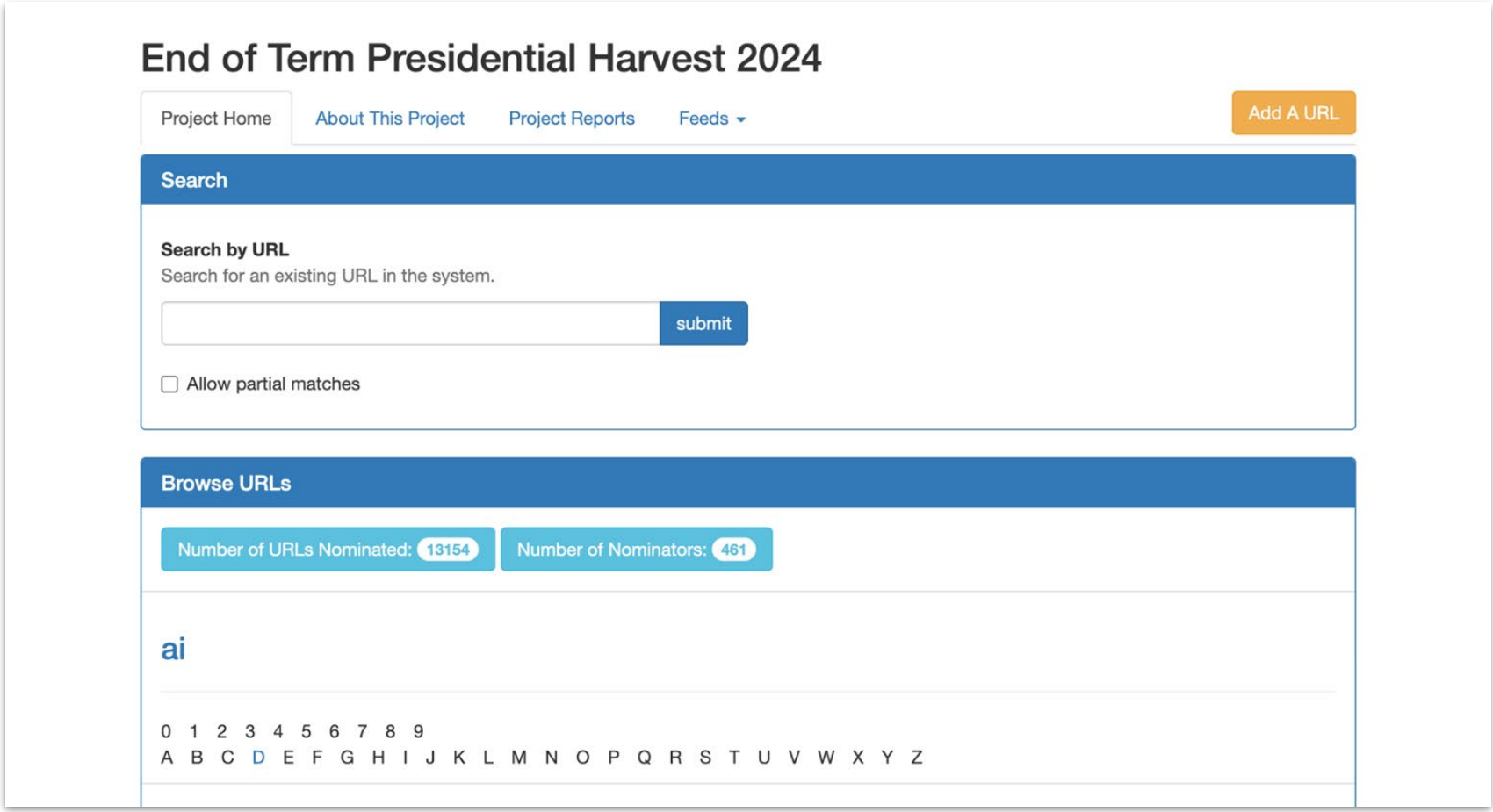
Background

End of Term Web Archive

- Collaborative web archiving activity in the United States since 2008
- Goal to document the transition in the Executive Branch of the Federal web before and after each election cycle
- Serves as a longitudinal snapshot of Federal .gov and public .mil web every four years
- Partners volunteer time, crawling, and storage resources for the project
- Public access provided by the Internet Archives' Wayback Machine
- https://eotarchive.org

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
Common Crawl					Crawl
WebRecorder					Crawl



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

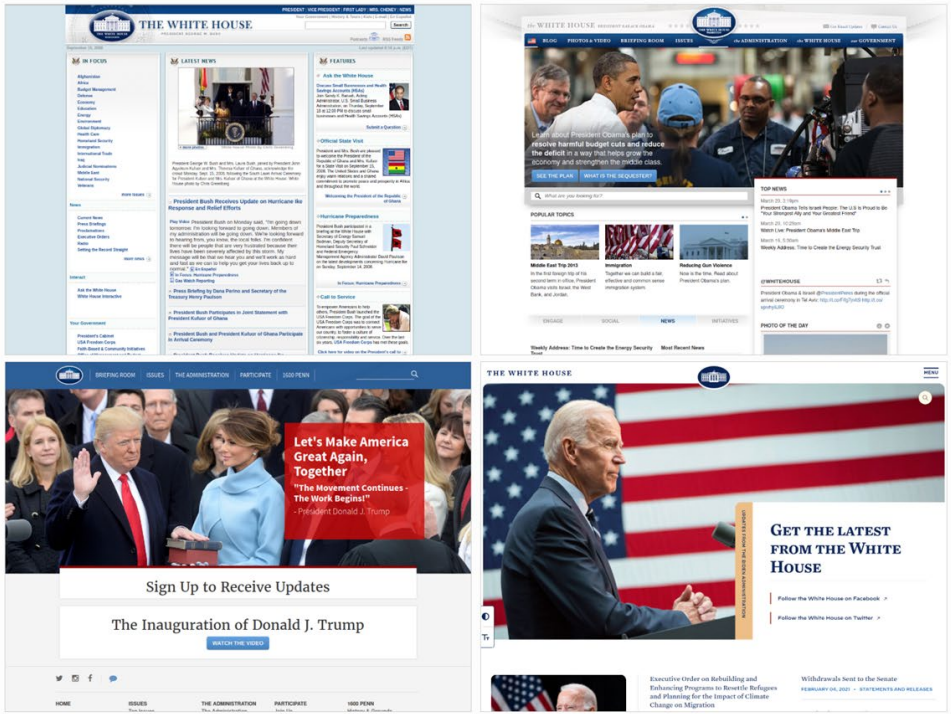
https://eotarchive.org

End of Term Web Archive

Background Partners Contribute Media Datasets

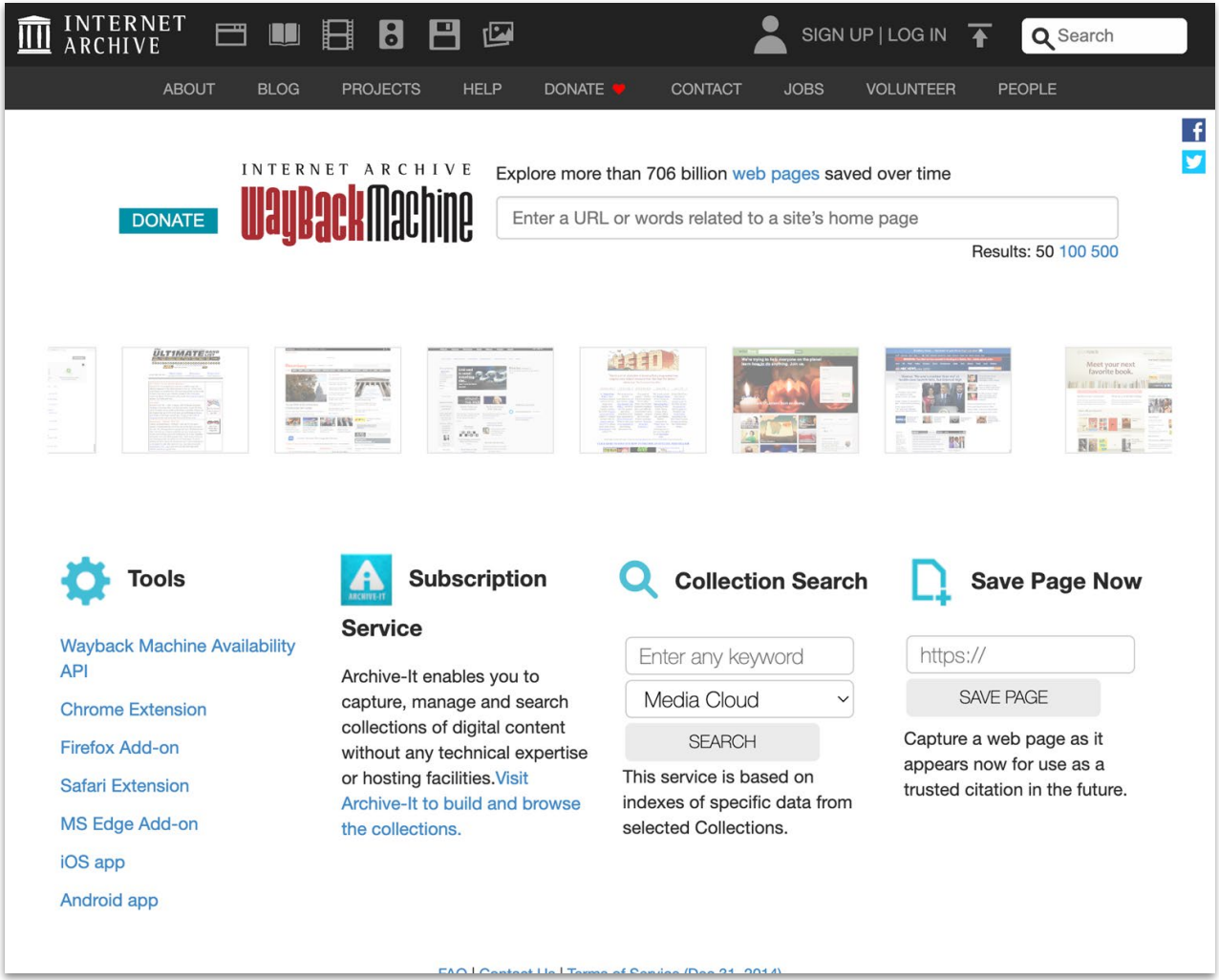
Purpose

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**.



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

https://web.archive.org



24 Results

Clear all filters

Filters

Year Published

2016 - 2016

Media Type

- collection 24
- web 107,356
- texts 72
- movies 16
- data 13
- audio 7

More...

Year

- 2016 2

Subject

- end of term 7
- federal government 6
- 2016 5
- congress 5
- president 5
- US 2

More...

End of Term Web Crawls

125,225 items
1.3 petabytes

End Of Term 2024 Web Crawls

77,510 items
707.5 terabytes

End Of Term 2024 Interim Election to Inauguration Crawls

27,039 items
250.7 terabytes

End Of Term 2020 Web Crawls

29,132 items
316.7 terabytes

End of Term 2016 Web Crawls

15,172 items
228.5 terabytes

End Of Term 2020 Pre Election to Inauguration Crawls

9,649 items
104 terabytes

End of Term 2012 Web Crawls

2,383 items
22 terabytes

End Of Term 2016 Pre-Inauguration Crawls

4,693 items
52.4 terabytes

End Of Term 2024 Pre-Election Crawls

13,635 items
123.9 terabytes

End Of Term 2020 Post Inauguration Crawls

19,103 items
208.4 terabytes

End Of Term 2016 UNT Crawls

1,275 items
23.6 terabytes

End of Term 2008 California Digital Library Crawl

415 items
6.8 terabytes

End of Term 2016 Post-

End of Term 2008 UNT

End Of Term 2016

End Of Term Crawls

End of Term 2020 UNT

End Of Term 2024 Pre-

EOT to AWS

Goals of the Project

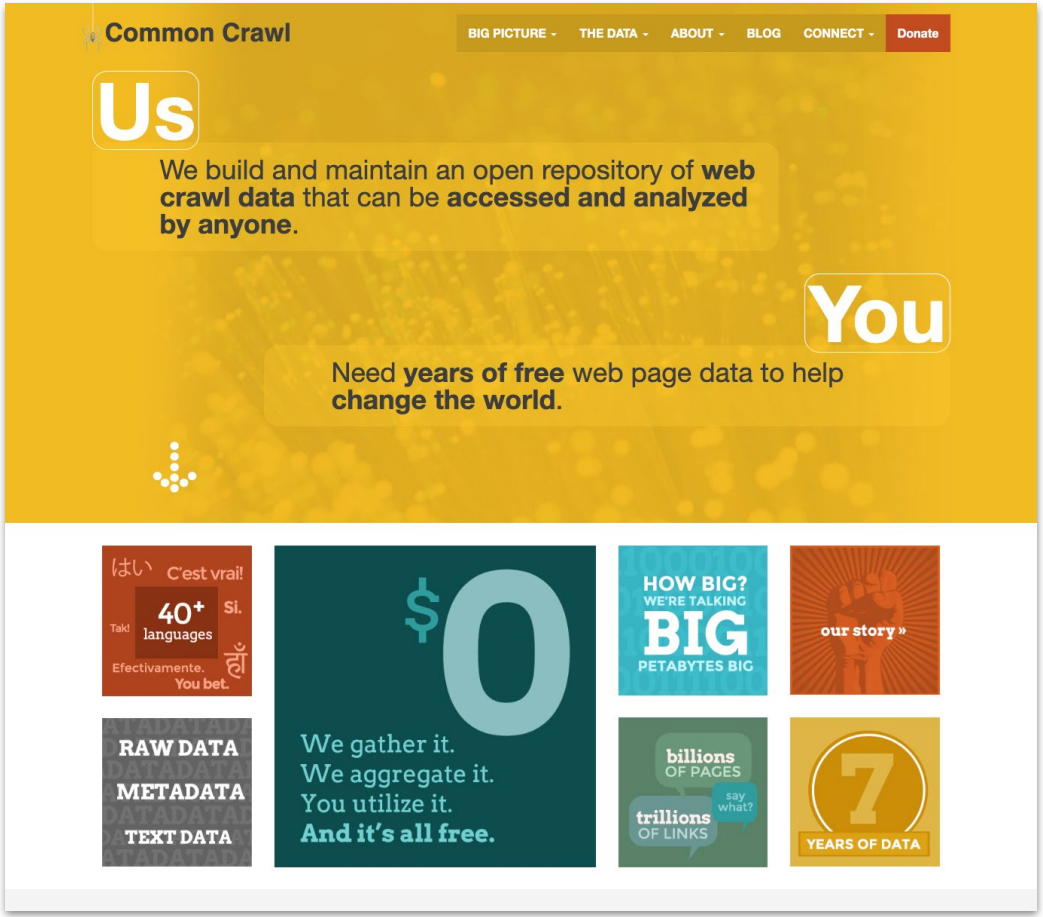
- Provide greater access to the End of Term datasets
  - 2008, 2012, 2016, 2020, and 2024
- Focused on computational consumption of the collection
- Currently challenging because of size, access, storage issues
- Encourage reuse and research with the EOT data
- Position dataset so that we can learn more about our process
- Provide a canonical dataset for each crawl for reference numbers like size, URLs, counts

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



Common Crawl Data

WARC files - content of crawls

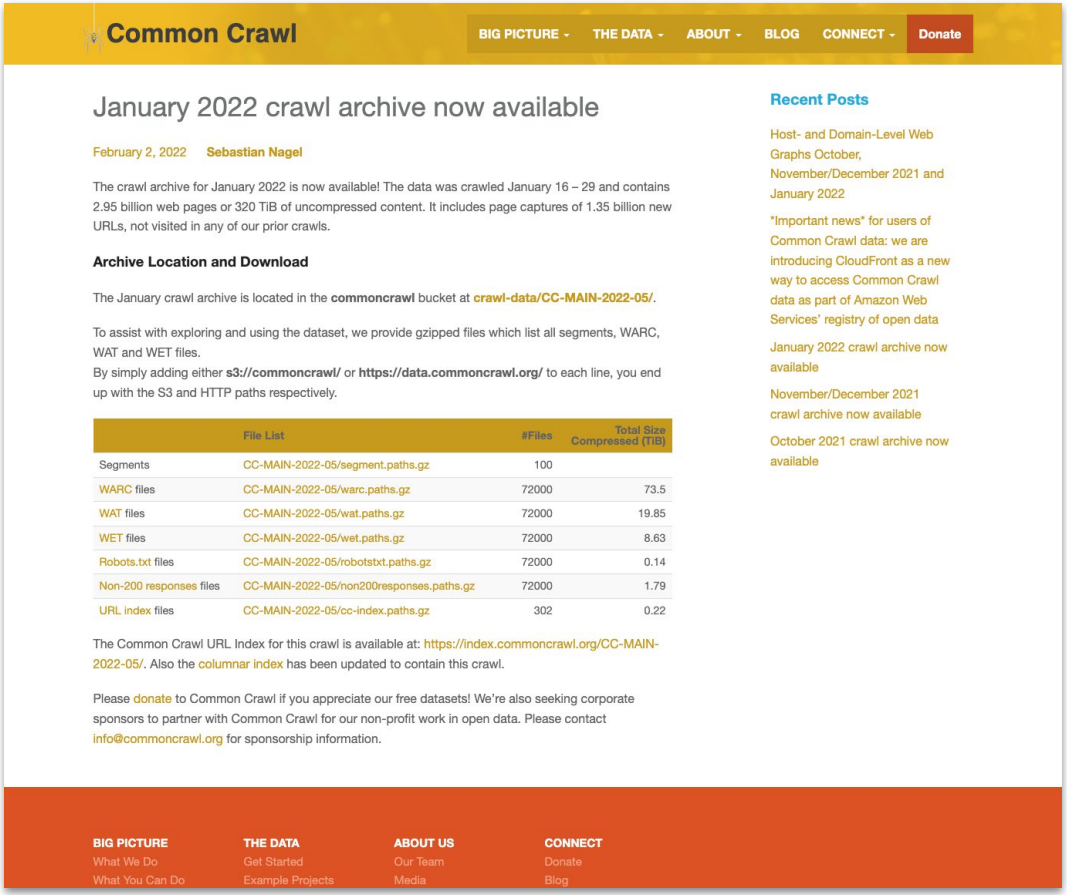
WAT - Extracted metadata from WARC files

WET - Extracted text from WARC files

(WAT and WET limited to HTML and TXT)

CDX Index - ZipNum format

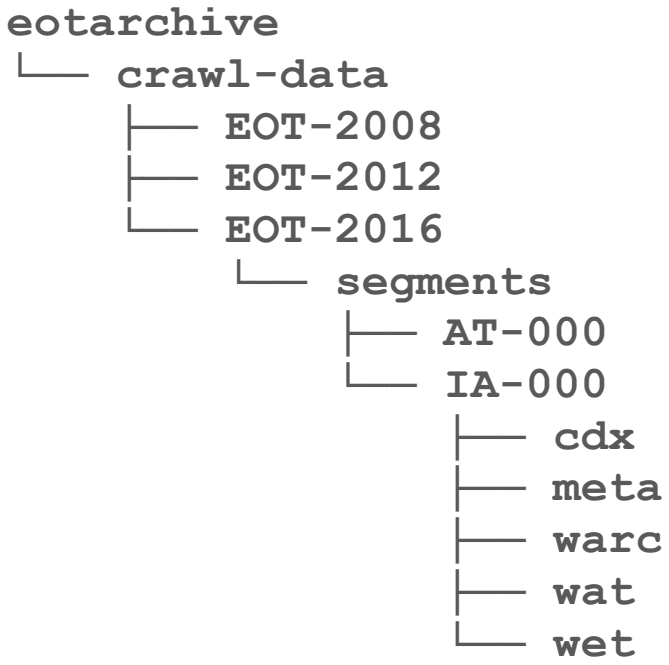
Parquet Index - based on CDX Index



Why Common Crawl Data Structure

- Existing tools can be used for generation of derivatives
  - Well documented tools built around working with large datasets
- Leverage existing Common Crawl community who are heavy users of those datasets
- Reuse documentation about formats, code, processes that exist for Common Crawl
- Great starting point until we have a strong reason to deviate

EOT Structure in Amazon S3



What Are We Moving/Copying?

- Moving EOT crawls dataset to Amazon S3
  - 2008 - UNT Libraries lead
  - 2012 - UNT Libraries lead
  - 2016 - IA Lead
  - 2020 - IA Lead
- Steps include:
  - Identifying scope of collection in repositories
  - Staging and verifying data from local repository
  - Organizing into S3 bucket layout
  - Generating WAT/WET/CDX/META/ZIPNUM/Parquet
  - Uploading to S3
  - Documenting Dataset

Logistics

- Fall 2021 we contacted AWS Open Data Program
- Approval from AWS was pretty quick
- IA is institutional home to the AWS bucket with shared credentials to others working on data
- Currently AWS is providing storage for ~750TB of data
- Weekly meetings between Mark P. at UNT and Sawood A. at IA for planning

Data Movement

- Identifying boundaries of EOT archives can be challenging
  - Requires a bit of digging back into EOT collective memory
  - A little bit of diving into old CDX indexes was required
- Divide and conquer
  - UNT and IA would take different EOT sets
  - UNT - 2008, 2012
  - IA - 2016, 2020
- Two approaches
  - UNT - Upload WARC and derivatives
  - IA - Upload WARC to S3, UNT downloads and generates derivatives

Content in the Cloud…

Crawl	WARC Files	WARC Size	WAT Size	WET Size	CDX Size	META Size
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
EOT-2024	?	700TB+?	?	?	?	?
Total	638,707	461TB	12TB	4TB	130GB	1TB

Tools Used

Small 5-node Local Hadoop Cluster (250TB) & mrjob

WAT/WET

https://github.com/commoncrawl/ia-web-commons

https://github.com/commoncrawl/ia-hadoop-tools

CDXJ

https://github.com/webrecorder/cdxj-indexer

WARC Metadata Sidecar

https://github.com/unt-libraries/warc-metadata-sidecar

Zipnum

https://github.com/commoncrawl/webarchive-indexing

Parquet

https://github.com/commoncrawl/cc-index-table

```
D DESCRIBE SELECT * FROM read_parquet('*.parquet');
```

column_name varchar	column_type varchar	null varchar	key varchar	default varchar	extra varchar
url_surtkey	VARCHAR	YES			
url	VARCHAR	YES			
url_host_name	VARCHAR	YES			
url_host_tld	VARCHAR	YES			
url_host_2nd_last_part	VARCHAR	YES			
url_host_3rd_last_part	VARCHAR	YES			
url_host_4th_last_part	VARCHAR	YES			
url_host_5th_last_part	VARCHAR	YES			
url_host_registry_suffix	VARCHAR	YES			
url_host_registered_domain	VARCHAR	YES			
url_host_private_suffix	VARCHAR	YES			
url_host_private_domain	VARCHAR	YES			
url_host_name_reversed	VARCHAR	YES			
url_protocol	VARCHAR	YES			
url_port	INTEGER	YES			
url_path	VARCHAR	YES			
url_query	VARCHAR	YES			
fetch_time	TIMESTAMP	YES			
fetch_status	SMALLINT	YES			
content_digest	VARCHAR	YES			
content_mime_type	VARCHAR	YES			
content_mime_detected	VARCHAR	YES			
content_charset	VARCHAR	YES			
content_languages	VARCHAR	YES			
content_puid	VARCHAR	YES			
warc_filename	VARCHAR	YES			
warc_record_offset	BIGINT	YES			
warc_record_length	BIGINT	YES			
warc_segment	VARCHAR	YES			
crawl	VARCHAR	YES			
subset	VARCHAR	YES			

31 rows 6 columns

```
D
```

https://eotarchive.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; and February 3, 2017.

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 2020 Web Archive

For the End of Term 2020, The Library of Congress, University of North Texas Libraries, Internet Archive, Stanford University Libraries, and the U.S. Government Publishing Office (GPO) joined efforts again, this time with new partners Environmental Data & Governance Initiative (EDGI) and the National Archives and Records Administrationi (NARA), to preserve public United States Government websites at the conclusion of the presidential administration ending January 20, 2021. This web harvest – like its predecessors in 2008, 2012, and 2016 – was intended to document the federal government's presence on the World Wide Web during the transition of presidential administrations and to enhance the existing collections of the partner institutions. This broad comprehensive crawl of the .gov domain includes as many federal .gov sites as we could find, plus federal content in other domains (such as .mil, .com, and social media content) and FTP'd datasets.

Nominations made by individual URL for inclusion in the End of Term Presidential Harvest 2020 are availabe to view in the Nomination Tool. URLs submitted for consideration in bulk form via files were added to a separate bulk Nomination Tool instance. The files containing the bulk list URLs have also been added to a GitHub repository. The entirety of the archived content is currently being held by the Internet Archive.

Browse the End of Term Web Archive

- 2008-2009
- 2012-2013
- 2016-2017
- Browse All
- Search Full Text

eotarchive
end-of-term

The End of Term Web Archive is a collaborative initiative that collects, preserves, and makes accessible United States

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 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		

End of Term 2008 Dataset

End of Term 2008 Dataset

The End of Term 2008 Dataset represents data collected by four collecting institutions. These institutions were the California Digital Library (CDL), the Internet Archive (IA), the Library of Congress (LOC) 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 2008 End of Term archive is located on the **eotarchive** bucket at EOT-2008.

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

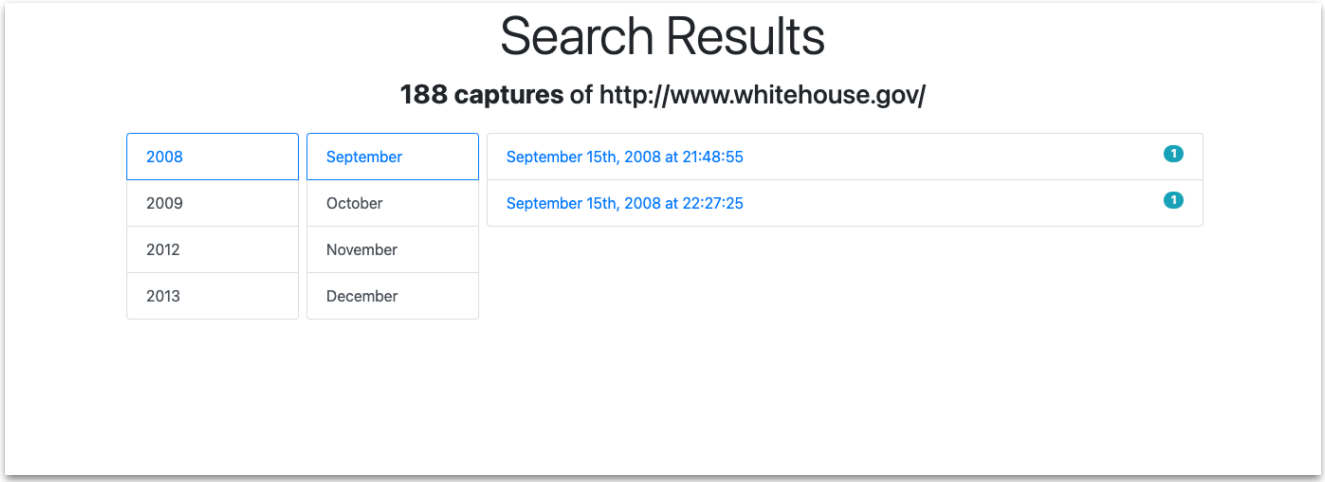
By adding either **s3://eotarchive/** or **https://eotarchive.s3.amazonaws.com/** to each line, you end up with the s3 and HTTP paths respectively.

File	List	#Files	Total Size Compressed (TiB)
Segments	EOT-2008/segment.paths.gz	14	
WARC files	EOT-2008/warc.paths.gz	125704	16.85
WAT files	EOT-2008/wat.paths.gz	125704	0.48
WET files	EOT-2008/wet.paths.gz	125704	0.12
CDX files	EOT-2008/cdx.paths.gz	125704	0.01
URL Index files	EOT-2008/eot-index.paths.gz	50	0.007

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.

We were able to implement a proof of concept for serving directly from AWS with pywb with about 15 lines of configuration.



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

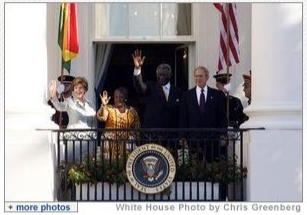
Interact

- 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

+ Press Briefing by Dana Perino and Secretary of the Treasury Henry Paulson

+ 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

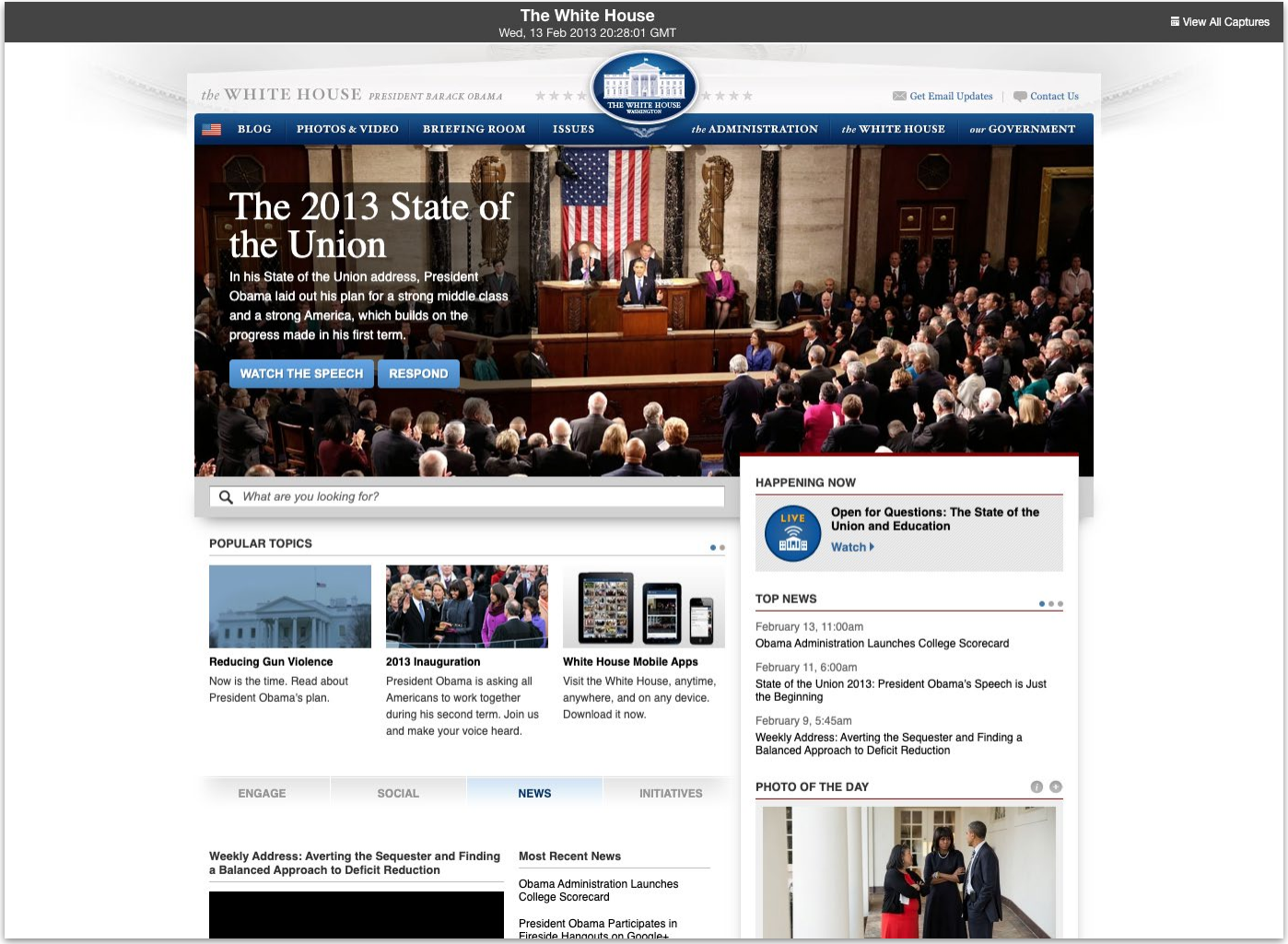
Call to Service

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


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.
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
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+

mark.phillips@unt.edu

@vphill.bsky.social

https://eotarchive.org