Harmonizing Heterogeneous WARC Contributions: Normalization Lessons from the 2024 End of Term Web Archive

Sawood Alam - Internet Archive
Mark Phillips - University of North Texas Libraries

WADL 2026
October 13, 2026

End of Term Web Archive

Every four years a group of partners come together to collect content from the Federal .gov and public .mil websites.

Using the transition in the Executive Branch as the event for the archiving.

Document the changes to the websites as part of the change in government focus, branding, and initiatives.

Act as a longitudinal dataset of publicly available content.

Allows us to experiment with new tools and techniques.



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
- make the dataset publicly available for download, replay, and analysis

End of Term Crawling Partners

	2008	2012	2016	2020	2024
Archive Team (AT)			■		■
California Digital Library (CDL)	■				
Common Crawl Foundation (CC)					■
Harvard Library Innovation Lab (HIL)					■
Internet Archive (IA)	■	■	■	■	■
Library of Congress (LOC)	■	■	■		
University of North Texas (UNT)	■	■	■	■	■
Webrecorder (WR)					■

Multiple Crawling Institutions: Different Approaches in 2024

Internet Archive:

- Heritrix based
- Large, long-running crawls
- Shallow and deep crawls
- Specialized video archiving pipeline
- Static crawling

University of North Texas:

- Heritrix and Browsertrix Crawler based
- Human Nominated Seeds
- New crawls every two weeks
- Broad and shallow browser-based crawling
- High fidelity and static crawling

Archive Team:

- Wget-warc/Warrior based
- Large-scale
- Static

Webrecorder:

- Browsertrix based
- Targeted interactive web content
- High fidelity crawling

Common Crawl Foundation:

- Nutch crawler-based
- Global crawl biased by search-engine style ranking
- EOT collection is extracted from our global crawl
- Static crawling

Library Innovation Lab:

- bag-nabit crawler
- focused on datasets
- bagit-based storage

Data Movement

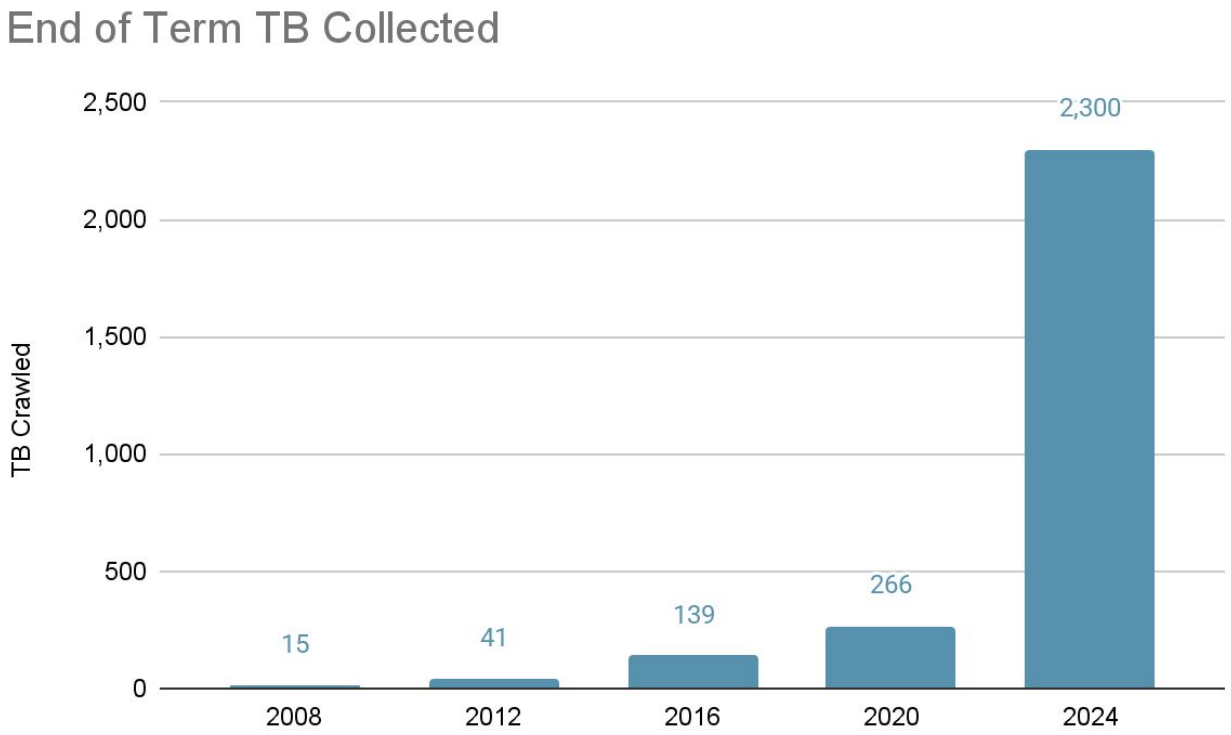
Compiling data into a single dataset.

WARC, WACZ, and BagIt datasets

Hosted at Internet Archive and on Amazon S3

```
AT  (WARC.gz)        -> IA                                -> S3
AT  (WARC.zst)       -> IA -> S3 -> UNT {zst-to-gz} -> S3 [rm WARC.zst]
CC  (Repkg WARC.gz) -> S3                                 -> IA [rebucket]
HIL (BagIt.zip)      -> IA {bagit-to-warc}                -> S3
IA  (WARC.gz)        -> S3
UNT (WARC.gz)        -> IA                                -> S3
UNT (WACZ)           -> IA {wacz-to-warc}                 -> S3
WR  (WACZ)           -> S3 -> IA {wacz-to-warc}      -> S3 [rm WACZ]
```

EOT-2024 - Compared to Previous Years



Results of the work

Crawl	WARC Files	WARC Size (TB)	WAT Size (TB)	WET Size (TB)	CDX Size (TB)	META Size (TB)
EOT-2008	125,704	15	0.4	0.1	0.01	0.07
EOT-2012	78,509	41	0.9	0.2	0.01	0.07
EOT-2016	194,683	139	2	0.3	0.03	0.18
EOT-2020	239,811	266	9	3.0	0.08	0.70
EOT-2024	**1,216,891**	**2,290**	**18**	**7.0**	**0.29**	**1.73**
Total	1,855,598	2,751	30	10.6	0.40	2.73

Traditional vs Computational Access

Traditional:

- Single-page access paradigm via Wayback Machine
- Lexicographical version diffs
- High-level quantitative reports
- Content presentation within its context
- Focusing on “close reading” of content
- Important when studying the layout and presentation of content via the browser

Computational:

- Longitudinal dataset
- Low-barrier access of datasets
- Content and context analysis
- Semantic change analysis
- Quantitative and qualitative analysis
- Collection understanding
- AI ingestion and summarization

Common Crawl datasets feed virtually all major AI models.

“We all use WARC”...

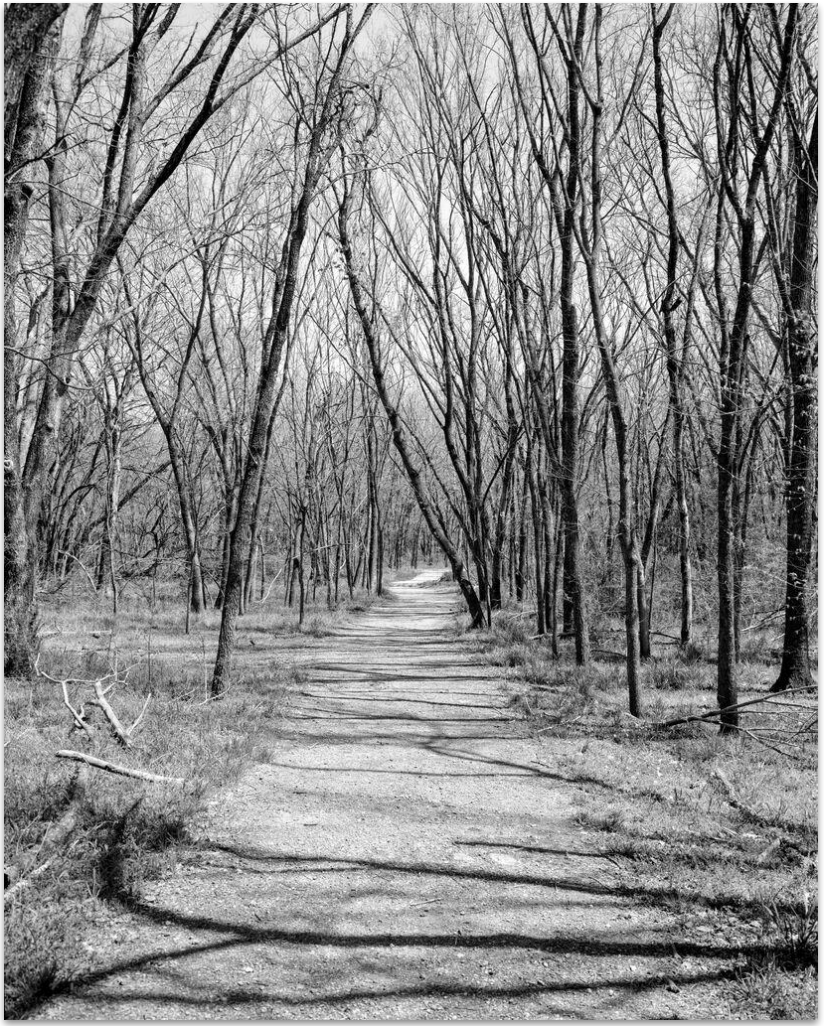
Normalization Pathways

WACZ Unpacking

Zstandard WARC Recompression

BagIt to WARC

Recovering bad WARCs



Goals of the Dataset

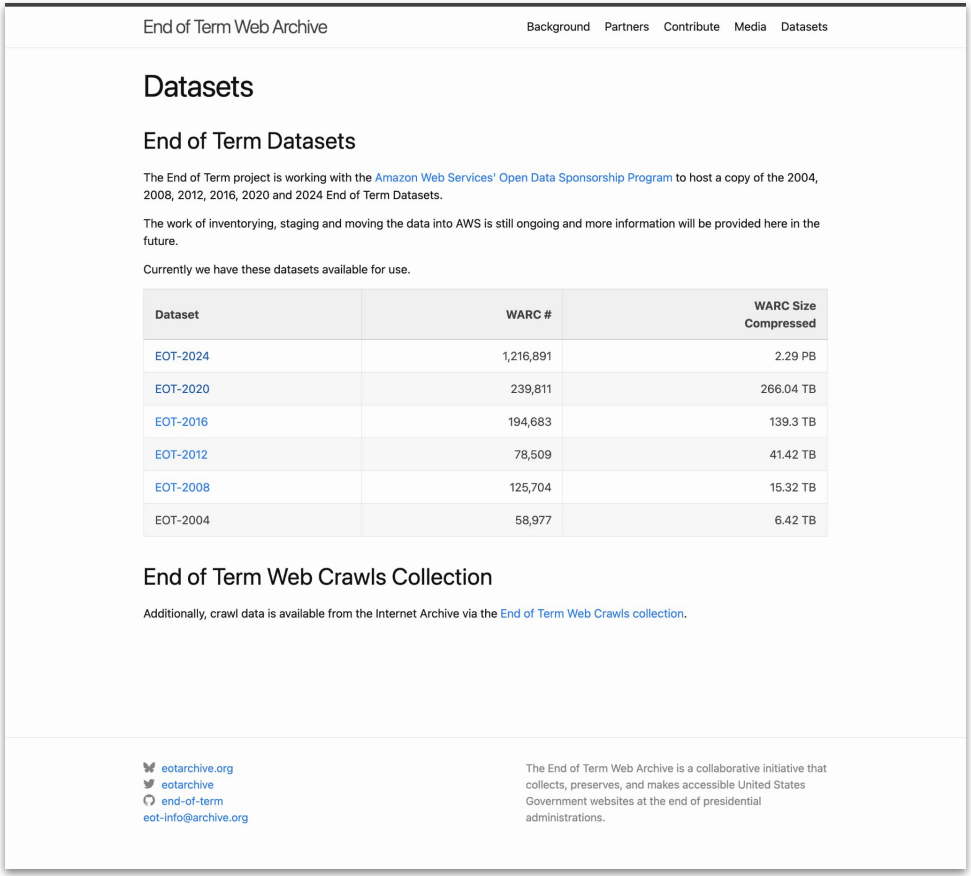
Standard, record-compressed warc.gz

Standardized file format for derivative generation

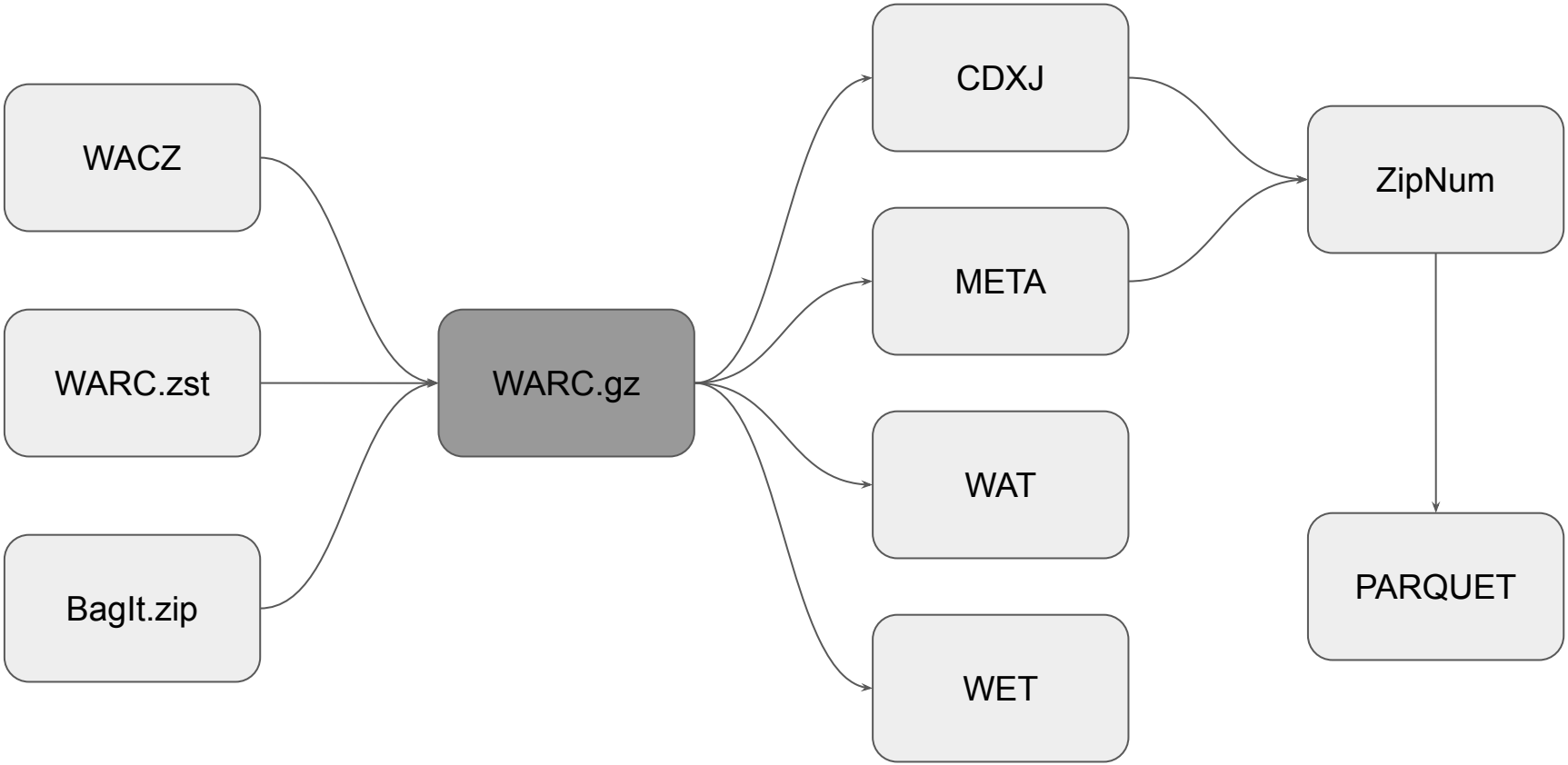
Similar, WARC as primary unit as in previous efforts

Consistency for end users not familiar with the specific formats

Does not introduce new tools or methodologies in the access chain



Normalization and Derivation Transitions



EOT on AWS Layout

```
crawl-data/EOT-2024/segments/
                    |-- AT-{000..014}
                    |    |-- cdx
                    |    |-- meta
                    |    |-- warc
                    |    |-- wat
                    |    `-- wet
                    |-- CC-000
                    |-- HIL-{000..031}
                    |-- IA-{000..071}
                    |-- UNT-{000..004}
                    `-- WR-{000..002}
```

EOT CDXJ Index Layout

```
eot-index/collections/EOT-2024/indexes/
           |-- cluster.summary
           |-- eot24-cdx-00000.cdxj.gz
           |-- eot24-cdx-00001.cdxj.gz
           |-- eot24-cdx-00002.cdxj.gz
           |-- eot24-cdx-00003.cdxj.gz
           |-- ..
           |-- eot24-cdx-00057.cdxj.gz
           |-- eot24-cdx-00058.cdxj.gz
           |-- eot24-cdx-00059.cdxj.gz
           |-- eot24-cdx-00060.cdxj.gz
           `-  eot24-cdx-00061.cdxj.gz
```

EOT Parquet Layout

```
eot-index/table/eot-main/crawl=EOT-2024/
          |-- eot24-cdx-00000.parquet
          |-- eot24-cdx-00001.parquet
          |-- eot24-cdx-00002.parquet
          |-- eot24-cdx-00003.parquet
          |-- ..
          |-- eot24-cdx-00057.parquet
          |-- eot24-cdx-00058.parquet
          |-- eot24-cdx-00059.parquet
          |-- eot24-cdx-00060.parquet
          `-  eot24-cdx-00061.parquet
```

WACZ Unpacking

WACZ packages WARC files into a single zip file.

Didn’t align with “warc-based” derivative pipelines

If we were just providing CDX access it would have worked just fine.

We extracted warc.gz files from WACZ files.

Organized output into segments as normal.

UNT and WebRecorder content.



TABLE OF CONTENTS

Abstract
1. Conformance
2. Status of This Document
3. Terminology
4. Introduction
4.1 Motivation
4.2 Existing Tools
5. WACZ Object
5.1 Directory Layout
5.2 Directories and Files
5.2.1 archive
5.2.2 indexes
5.2.3 pages.jsonl
5.2.4 datapackage.json
5.2.5 datapackage-digest.json
5.3 Other files and directories
5.4 Zip Format
5.4.1 Zip Compression
5.4.2 Zip Format File Extension
6. Processing Model
7. Publishing
7.1 Content-Length
7.2 Partial Requests
7.3 CORS
7.4 Media Type
7.5 Example Response
A. References
A.1 Normative references

Web Archive Collection Zipped (WACZ)

ReSpec

Webrecorder Recommendation 03 June 2021

▼ More details about this document

This version:
https://specs.webrecorder.net/wacz/1.1.1/

Latest published version:
https://specs.webrecorder.net/wacz/latest/

Editors:
Ilya Kreymer (Webrecorder)
Ed Summers (Stanford University)

Additional Documents
Use Cases for Decentralized Web Archives

Previous versions
1.1.0
1.0.0

Repository
Github
Issues
Commits

Copyright © 2021 Webrecorder CC-BY

Abstract

WACZ is a media type that allows web archive collections to be packaged and shared on the web as a discrete file. A WACZ file includes all the data that is needed for the rendering archived content as well as contextual information required for users to interpret it. Rendering software can obtain this data on demand using HTTP Range requests, without requiring the entire file to be fully retrieved, or for it to be otherwise mediated by specialized server side software.

§ 1. Conformance

As well as sections marked as non-normative, all authoring guidelines, diagrams, examples, and notes in this specification are non-normative. Everything else in this specification is normative.

The key words *MAY*, *MUST*, *MUST NOT*, and *SHOULD* in this document are to be interpreted as described in BCP 14 [RFC2119] [RFC8174] when, and only when, they appear in all capitals, as shown here.

§ 2. Status of This Document

This is a stable version of the WACZ standard and is in active use by the Webrecorder project. Please open GitHub issues for questions and suggestions.

What’s in a WACZ?

```
20250324203621503-5376d4c4-540-0/
|-- archive/
|   |-- usgov-archive-www-rferl-org-manual-20250316193306-5376d4c4-540-20250324203624666-2.warc.gz
|   |-- usgov-archive-www-rferl-org-manual-20250316193306-5376d4c4-540-20250324203624711-3.warc.gz
|   |-- usgov-archive-www-rferl-org-manual-20250316193306-5376d4c4-540-20250324203624759-1.warc.gz
|   |-- usgov-archive-www-rferl-org-manual-20250316193306-5376d4c4-540-20250324203624762-0.warc.gz
|   |-- ...
|   |-- usgov-archive-www-rferl-org-manual-20250316193306-5376d4c4-540-text-20250324203632344.warc.gz
|   |-- usgov-archive-www-rferl-org-manual-20250316193306-5376d4c4-540-text-20250324215936732.warc.gz
|   |-- usgov-archive-www-rferl-org-manual-20250316193306-5376d4c4-540-text-20250324230007235.warc.gz
|   |-- usgov-archive-www-rferl-org-manual-20250316193306-5376d4c4-540-text-20250325001737372.warc.gz
|   |-- usgov-archive-www-rferl-org-manual-20250316193306-5376d4c4-540-text-20250325010742953.warc.gz
|   +-- usgov-archive-www-rferl-org-manual-20250316193306-5376d4c4-540-text-20250325022625826.warc.gz
|-- datapackage-digest.json
|-- datapackage.json
|-- indexes/
|   |-- index.cdx.gz
|   +-- index.idx
|-- logs/
|   |-- crawl-20250324203619201.log
|   |-- crawl-20250324215924159.log
|   |-- crawl-20250324225955702.log
|   |-- crawl-20250325001724380.log
|   |-- crawl-20250325010729840.log
|   +-- crawl-20250325022612364.log
+-- pages/
    |-- extraPages.jsonl
    +-- pages.jsonl
```

Zstandard Recompression

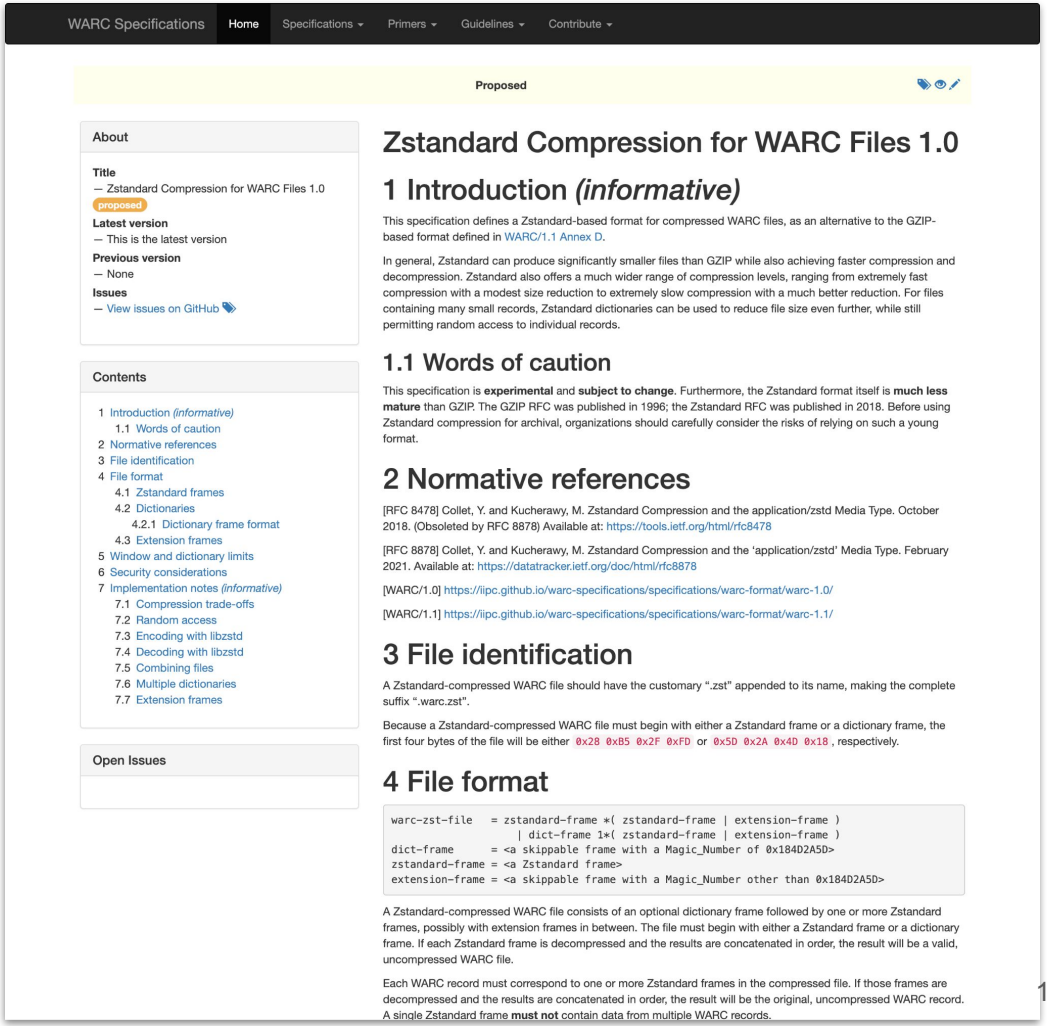
Proposed implementation of Zstandard compression in WARC records.

Differs in how access is expected, requires two lookups to retrieve decompression dictionary then compressed WARC record.

Internet Archive supports the format as well as a couple of other tools, but not broadly supported.

- Uncompress zstandard warc
- Recompress warc with fastwarc

~823 TB of WARC files recompressed



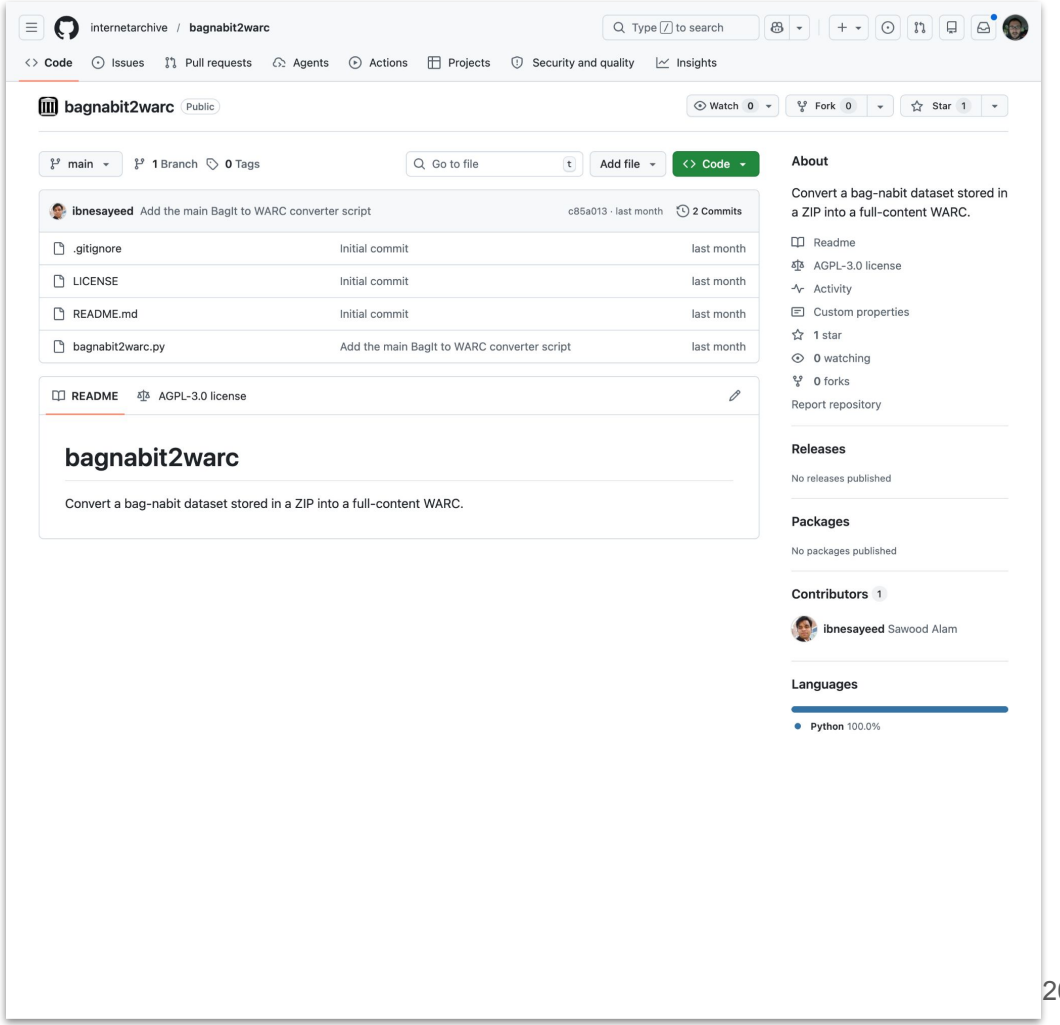
BagIt WARC Generation

bag-nabit tool used by Library Innovation Lab at Harvard.

Stores *response* headers in WARC with *revisit* records with WARC-Profile header pointing to files held in bagit data directory.

Script written to integrate files into WARC payload for standard playback and tool processing.

One WARC per bagit bag to keep things parallel.



HIL Bagit Example

```
3-21-veteran-supportive-community-designation-d63b5_v1/
                |-- bag-info.txt
                |-- bagit.txt
                |-- data/
                |   |-- files/
                |   |   |-- 4.html
                |   |   |-- catalog.json
                |   |   |-- catalog.jsonld
                |   |   |-- cc-by.html
                |   |   |-- dcat-us.html
                |   |   |-- item.html
                |   |   `-- tempegov::3-21-veteran-supportive-community-designation.html
                |   |-- headers.warc
                |   `-- signed-metadata.json
                |-- manifest-sha256.txt
                |-- signatures/
                |   |-- tagmanifest-sha256.txt.p7s
                |   |-- tagmanifest-sha256.txt.p7s.tsr
                |   `-- tagmanifest-sha256.txt.p7s.tsr.crt
                `-- tagmanifest-sha256.txt
```

headers.warc

```
WARC/1.0
WARC-IP-Address: 34.233.149.104
WARC-Type: revisit
WARC-Record-ID: <urn:uuid:5c9807e6-7bad-479f-ad6c-a65cc9bad1f9>
WARC-Target-URI: https://www.arcgis.com/home/item.html?id=ac0056dddc1c4c0ab3152e82e6ae99e3
WARC-Date: 2024-12-10T12:56:30Z
WARC-Payload-Digest: sha1:J2GLFNGTZAVVL4LOEXUNPDRSY6JK5XZ6
WARC-Profile: file-content; filename="files/item.html"
WARC-Block-Digest: sha1:RA4DOPVCTUOI6HFYF35SWSRMOUN26ILB
Content-Type: application/http; msgtype=response
Content-Length: 561

HTTP/1.1 200 OK
Date: Tue, 10 Dec 2024 12:56:30 GMT
Content-Type: text/html
Content-Length: 568
Connection: keep-alive
Content-Encoding: gzip
Last-Modified: Tue, 26 Nov 2024 01:02:54 GMT
ETag: "a9a7bdc53b6d7d280d71cc0681eb4983"
Server: AmazonS3
Cache-Control: private, max-age=0
Expires: -1
Strict-Transport-Security: max-age=63072000
Content-Security-Policy: upgrade-insecure-requests
Referrer-Policy: no-referrer-when-downgrade
X-Cached: HIT
Link: <https://www.arcgis.com/home/item.html?id=ac0056dddc1c4c0ab3152e82e6ae99e3>; rel="canonical"



WARC/1.0
WARC-IP-Address: 34.233.149.104
WARC-Type: request
WARC-Record-ID: <urn:uuid:d994c343-94bd-46fc-b2b9-16429b755091>
WARC-Target-URI: https://www.arcgis.com/home/item.html?id=ac0056dddc1c4c0ab3152e82e6ae99e3
WARC-Date: 2024-12-10T12:56:30Z
WARC-Payload-Digest: sha1:3I42H3S6NNFQ2MSVX7XZKYAYSCX5QBYJ
WARC-Concurrent-To: <urn:uuid:5c9807e6-7bad-479f-ad6c-a65cc9bad1f9>
WARC-Block-Digest: sha1:2TKZEZQ3CUM4LPN7GCHJJ263U64N2WS2
Content-Type: application/http; msgtype=request
Content-Length: 195

GET /home/item.html?id=ac0056dddc1c4c0ab3152e82e6ae99e3 HTTP/1.1
Host: www.arcgis.com
User-Agent: python-requests/2.32.3
Accept-Encoding: gzip, deflate
Accept: */*
Connection: keep-alive
```

Rehydrated WARC

```
WARC/1.0
WARC-Date: 2024-12-10T12:56:30Z
WARC-Target-URI: https://www.arcgis.com/home/item.html?id=ac0056dddc1c4c0ab3152e82e6ae99e3
WARC-Record-ID: <urn:uuid:5c9807e6-7bad-479f-ad6c-a65cc9bad1f9>
WARC-IP-Address: 34.233.149.104
WARC-Type: response
WARC-Payload-Digest: sha1:3NQUOAJJZB32U6TRTGO5HCW37VKBHO7C
WARC-Block-Digest: sha1:R3OZQPAHT474Z4BXD6TD7AXE4G3JZ7UX
Content-Type: application/http; msgtype=response
Content-Length: 2767

HTTP/1.1 200 OK
Date: Tue, 10 Dec 2024 12:56:30 GMT
Content-Type: text/html
Content-Length: 568
Connection: keep-alive
Content-Encoding: gzip
Last-Modified: Tue, 26 Nov 2024 01:02:54 GMT
ETag: "a9a7bdc53b6d7d280d71cc0681eb4983"
Server: AmazonS3
Cache-Control: private, max-age=0
Expires: -1
Strict-Transport-Security: max-age=63072000
Content-Security-Policy: upgrade-insecure-requests
Referrer-Policy: no-referrer-when-downgrade
X-Cached: HIT
Link: <https://www.arcgis.com/home/item.html?id=ac0056dddc1c4c0ab3152e82e6ae99e3>; rel="canonical"

<!DOCTYPE html>
<html>
  <head>

...
```

Recovering WARC Data

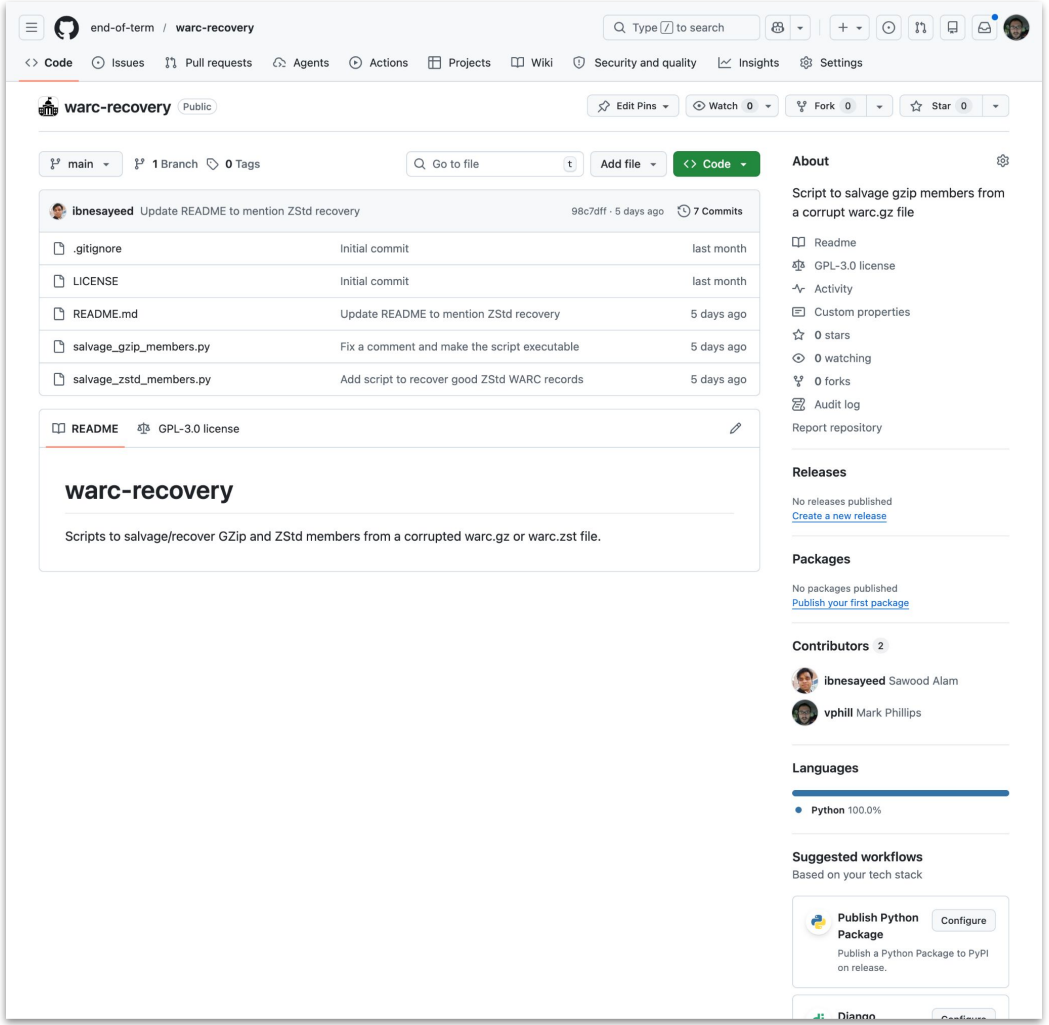
50 WARC files were corrupt to a point that they would not parse with either warcio or fastwarc

3 exhibited gzip compression errors and were unable to be parsed by gzip

47 zstandard WARCs had decompression errors

Two recovery scripts were written to iterate over WARC, test gzip or zstandard compressed blocks, ignore those that were bad and output those that were good.

Ended up with an additional half a TB of content that we would otherwise have ignored.



Tools Used

Small 5-node Local Hadoop Cluster (873 TB) & 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

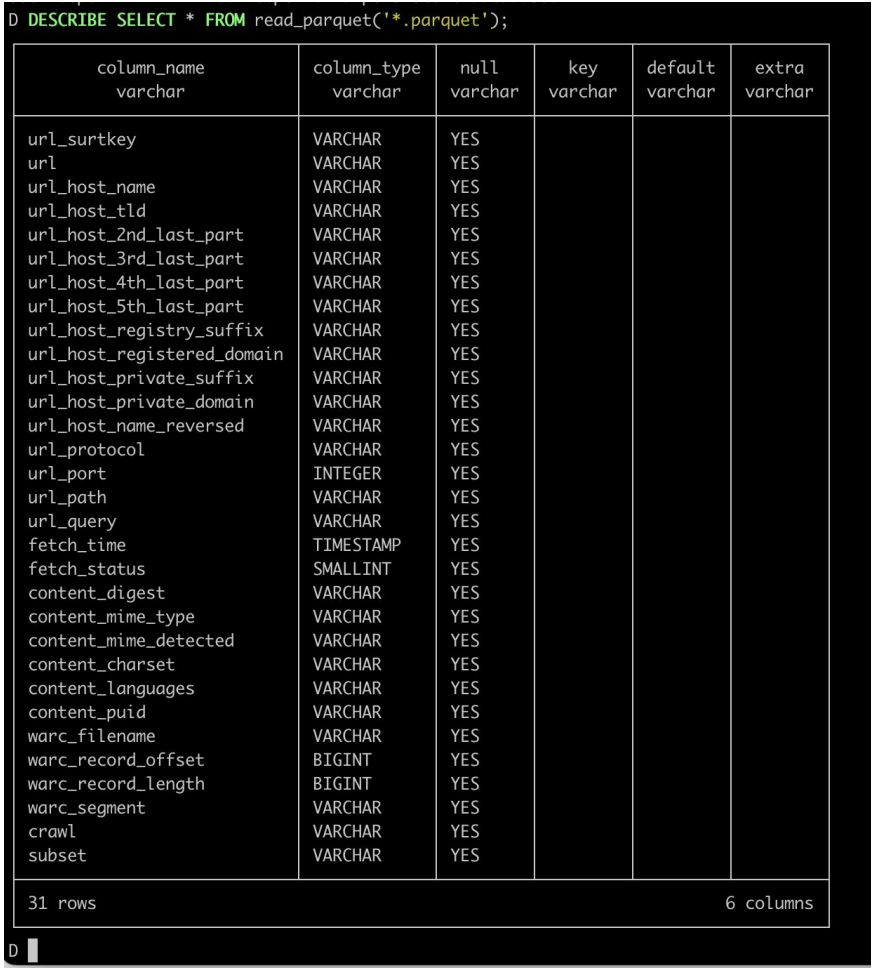
https://github.com/unt-libraries/warc-metadata-sidecar

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
```

New Tools Developed

bagnabit2warc

https://github.com/internetarchive/bagnabit2warc

warc-recovery

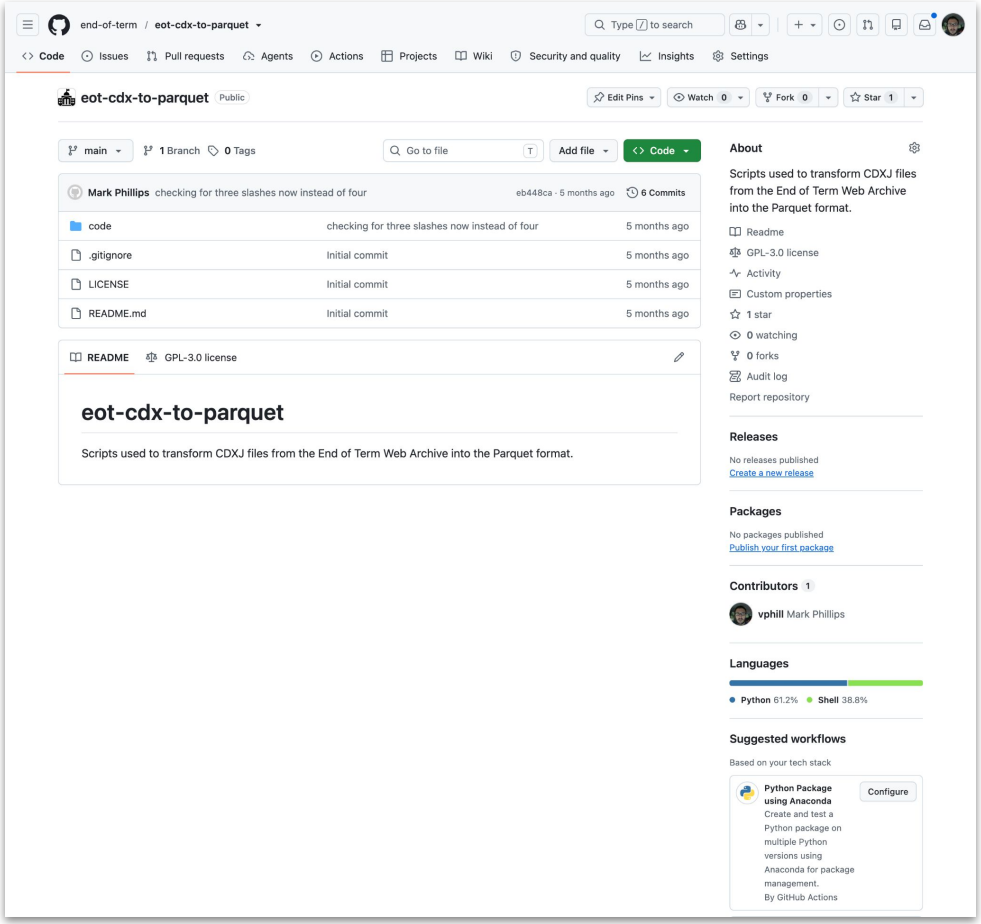
https://github.com/end-of-term/warc-recovery

eot-cdx-to-parquet

https://github.com/end-of-term/eot-cdx-to-parquet

zipnumer

https://github.com/internetarchive/zipnumer



Lessons Learned

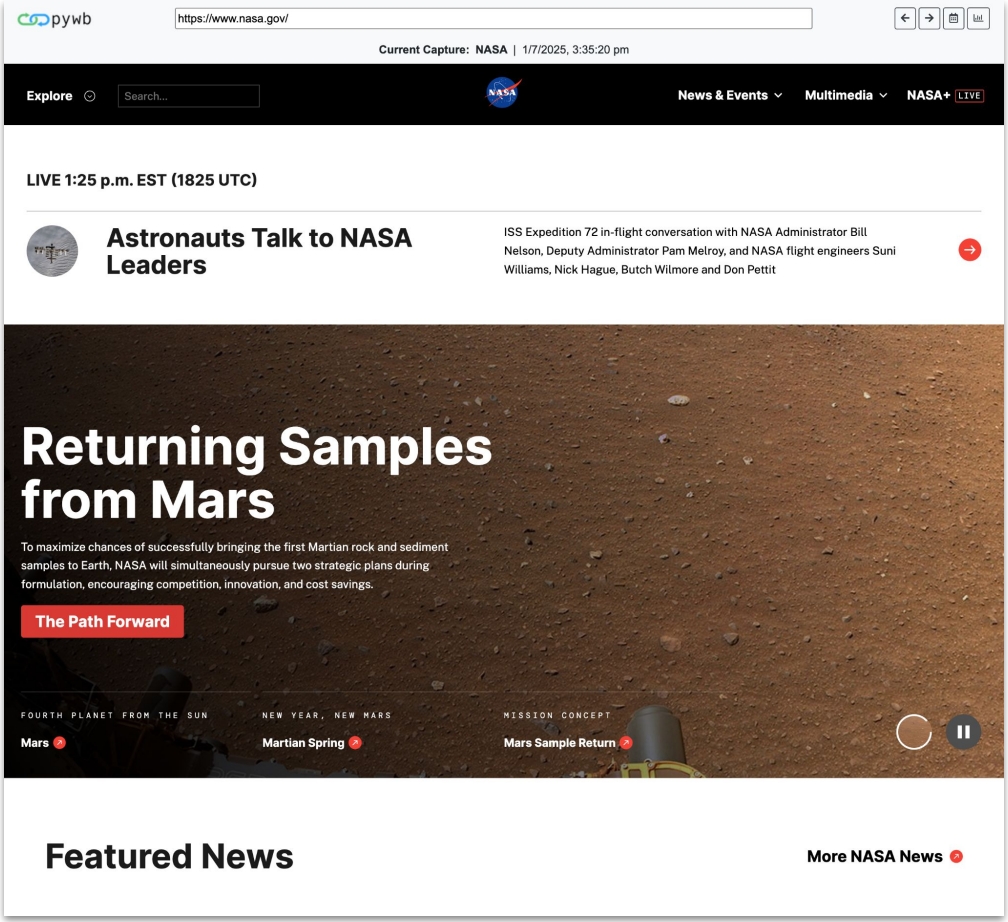
Multiple crawling partners helps distribute the work.

We have different approaches and are interested in different areas of this work.

We need to be more precise about the output formats we want for the final dataset.

Most of the work was machine time, documentation, and watching progress bars.

We've now created these datasets for five EOT instances. We have some ideas for improving the work for EOT-2028.



Next Steps

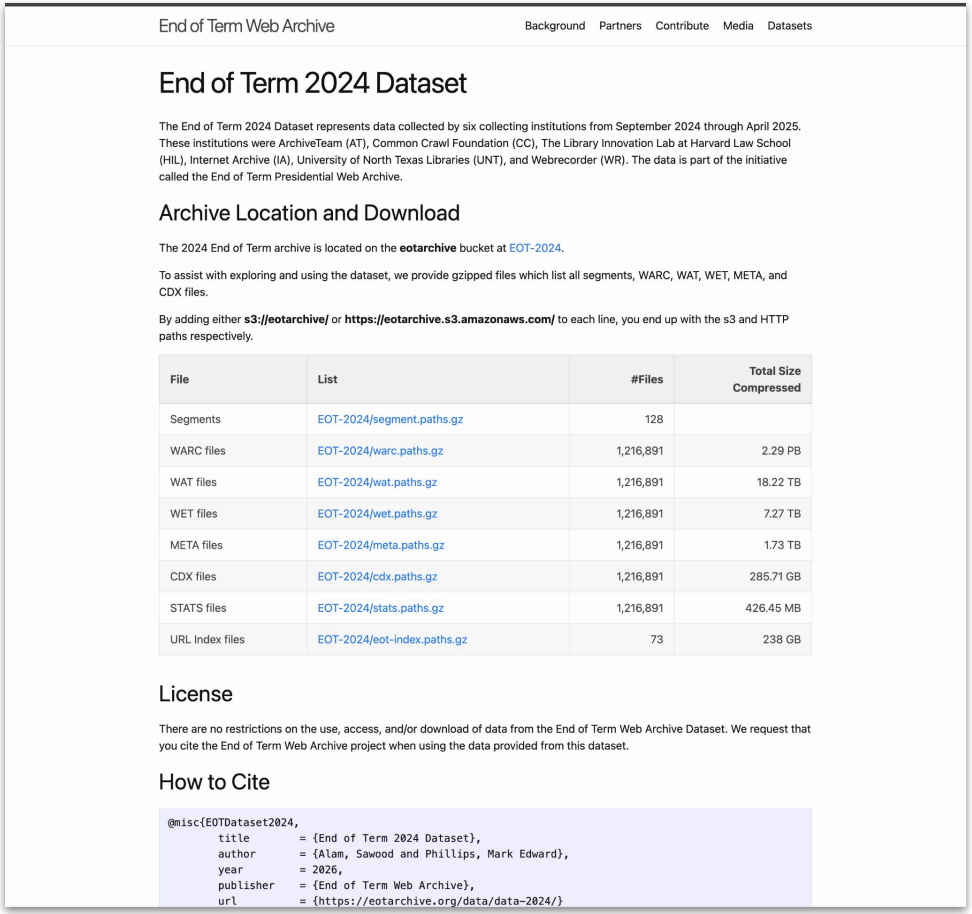
Broad announcement of dataset availability.

https://eotarchive.org/data/data-2024/

Tutorials and documentation.

Build link graph for EOT-2024 dataset.

Work on public, EOT-specific playback access to crawls.



End of Term Web Archive

Background Partners Contribute Media Datasets

End of Term 2024 Dataset

The End of Term 2024 Dataset represents data collected by six collecting institutions from September 2024 through April 2025. These institutions were ArchiveTeam (AT), Common Crawl Foundation (CC), The Library Innovation Lab at Harvard Law School (HIL), Internet Archive (IA), University of North Texas Libraries (UNT), and Webrecorder (WR). The data is part of the initiative called the End of Term Presidential Web Archive.

Archive Location and Download

The 2024 End of Term archive is located on the **eotarchive** bucket at EOT-2024.

To assist with exploring and using the dataset, we provide gzipped files which list all segments, WARC, WAT, WET, META, 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-2024/segment.paths.gz	128	
WARC files	EOT-2024/warc.paths.gz	1,216,891	2.29 PB
WAT files	EOT-2024/wat.paths.gz	1,216,891	18.22 TB
WET files	EOT-2024/wet.paths.gz	1,216,891	7.27 TB
META files	EOT-2024/meta.paths.gz	1,216,891	1.73 TB
CDX files	EOT-2024/cdx.paths.gz	1,216,891	285.71 GB
STATS files	EOT-2024/stats.paths.gz	1,216,891	426.45 MB
URL Index files	EOT-2024/eot-index.paths.gz	73	238 GB

License

There are no restrictions on the use, access, and/or download of data from the End of Term Web Archive Dataset. We request that you cite the End of Term Web Archive project when using the data provided from this dataset.

How to Cite

```
@misc{EOTDataset2024,
        title        = {End of Term 2024 Dataset},
        author       = {Alam, Sawood and Phillips, Mark Edward},
        year         = 2026,
        publisher    = {End of Term Web Archive},
        url          = {https://eotarchive.org/data/data-2024/}
```

Conclusions

End of Term Web Archive:
https://eotarchive.org/

Terms:
2008, 2012, 2016, 2020, 2024

Formats:
WARC, WAT, WET, CDXJ, Parquet

Size:
2.7 PB

Contact:

mark.phillips@unt.edu

sawood@archive.org

