End of Term Web Archive Dataset: Longitudinal Web Archive of .GOV and .MIL Domains

Mark E. Phillips
University of North Texas
Denton, TX, USA
mark.phillips@unt.edu

Kristy K. Phillips
University of North Texas
Denton, TX, USA
kristy.phillips@unt.edu

Sawood Alam
Internet Archive
San Francisco, CA, USA
sawood@archive.org

ABSTRACT

The End of Term (EOT) Web Archive Dataset presents a longitudinal dataset of the US federal web which includes publicly available `.gov` and `.mil` domains, created during the 2008, 2012, 2016, and 2020 presidential elections in the United States. Based on the End of Term Web Archive, this dataset presents 461TB of WARC data and accompanying derivative files in WAT, WET, and CDX format. A metadata sidecar file is also provided that contains content-based characterizations, including languages, character sets, format identifiers, and mime types. In addition to these derivative formats, CDX indexes in the ZipNum and Parquet formats that provide additional functionality to the dataset are included. The EOT dataset is freely available on the Amazon S3 platform as part of the Amazon Open Data Program.

KEYWORDS

End of Term, Archive, Web Archiving, Dataset, Research Dataset, AWS, S3, WARC, CDX, Parquet

1 INTRODUCTION AND MOTIVATION

Since 2008, the Internet Archive, the Library of Congress, the University of North Texas, and many other organizations have worked on a collaborative project to document the United States federal web. The output of this collaboration is the End of Term (EOT) Web Archive, which consists of a web crawl of all publicly available federal websites on the `.gov` and `.mil` domains collected concurrently with each presidential election. In years in which there is a presidential transition, this serves as a way to document the effect of the transition on public websites. In years in which the current president is re-elected, the web crawl serves as an archive of the changes made over the four years since the previous election.

In 2022, the Internet Archive and the University of North Texas began working to create the End of Term Web Archive Dataset, a more accessible dataset of the content found in the EOT Web Archive. This dataset overcomes the logistical challenges faced by users interested in using the archive for computationally-focused research and allows open access to a large, longitudinal dataset of the federal web.

The full dataset is available with a Creative Commons CC0 1.0 Universal (CC0 1.0)[1] Public Domain Dedication and is downloadable from the End of Term Website in the data section[2]. A record for the dataset is also available in the Registry of Open Data on AWS[3].

[1]Creative Commons CC0 1.0 Universal https://creativecommons.org/publicdomain/zero/1.0/
[2]End of Term: Data https://eotarchive.org/data/
[3]Registry of Open Data on AWS https://registry.opendata.aws/eot-web-archive/

2 BACKGROUND

Every four years, concurrent with United States presidential elections, a group of librarians, web archivists, and technologists work to document the change in the administration of the Executive Branch of the United States government. The End of Term (EOT) Web Archive[4] is both an ad-hoc collaborative project that comes together every four years to plan, publicize, and execute web crawls of the United States federal web, and an archive of the data collected during the project. For this project, the federal web consists of all publicly available websites on the `.gov` and `.mil` domains. Since its inception in 2008 [13], the EOT project has documented the federal web during four presidential elections: the transition from Bush to Obama in 2008, the transition from one Obama term to another in 2012, the transition from Obama to Trump in 2016 [11], and the transition from Trump to Biden in 2020.

In total, the EOT Web Archive consists of over 600TB of content collected during the four EOT web crawls. Providing long-term access to this content is often the most challenging aspect of the project. Since the EOT Web Archive was created, the Internet Archive has provided access to the EOT Web Archive data, first through various custom configurations of their Wayback Machine created specifically for the EOT, and currently through the global version of the Wayback Machine. In the interest of data redundancy, some members of the EOT team have also provided access to portions of the web archives via their own hosting infrastructure. The access provided by these two methods is sufficient for a general user, but for those interested in more complex research, this type of access is limiting. Users with an interest in computational research often need a larger portion of the EOT dataset than was available via previous methods of access. In order to meet this need, the EOT team began to work with the AWS Open Data Sponsorship Program [3] in the fall of 2021. This sponsorship has allowed a team from the Internet Archive and the University of North Texas to host the End of Term Web Archive Dataset, a longitudinal dataset of the four web crawls in the EOT Web Archive, via Amazon S3.

3 RELATED WORK

Over the past decade, interest in computational research using web archive data has increased in conjunction with the processing power of computers. Initiatives such as the Archives Unleashed project have developed tools and services to enable research over large collections of WARC files [8], which is an ISO standard format used by organizations around the world to store and transfer web archives and their associated derivative files [10]. The Archives Unleashed project found that users interested in utilizing web archives for research had difficulty accessing the data in these archives for

[4]End of Term Web Archive https://eotarchive.org/

scholarly work. Researchers interested in using web archives encountered two major issues with access. First, the underlying web archive files, stored in their native WARC format, were not openly available for download. The access points provided by the Internet Archive and other host organizations instead played back the collected websites as they appeared when they were crawled in a web browser, without ever offering access to the underlying files. Secondly, the sheer amount of data stored in a large web archive like the EOT makes dealing with the size and scale of the archives challenging, even for some large research institutions.

Some organizations like the Library of Congress [9] and the UK Web Archive [15] have found a way to offer more access to web archives without providing the primary datasets. These organizations offer access to derivative datasets, which are made of sampled data taken from a web archive [12]. Derivative datasets can include things like a collection of PDF files from a web archive, a Geoindex, or crawled URL indexes. Researchers are able to download these derivatives and process them with different tool kits to get an overview of the data in the greater web archives.

In 2008, Common Crawl[5], a non-profit foundation in the United States, began to make large web archives available for free download and reuse. Since 2013, Common Crawl has offered free access to the web crawls it has completed in the WARC format on Amazon S3 as a part of Amazon Web Services' Open Data Sponsorship Program. Common Crawl's monthly datasets provide access to billions of crawled web pages and present users with a variety of derivative formats. These derivative formats make it easier for the data be reused in many ways, without the need for the full WARC datasets, though those are also available. The Common Crawl data has been used by hundreds of projects and the Common Crawl team maintains a growing list of citations to publications that make use of the underlying datasets [5].

Common Crawl's work with Amazon Web Services' Open Data Sponsorship Program, as well as the access they provide to their web crawl datasets, offers an excellent example of how to share web archive data, and serves as the inspiration for the End of Term Web Archive Dataset.

4 METHODOLOGY

There were several goals for making the End of Term dataset available to researchers.

1) *Overcome challenges to accessing the underlying data* - The EOT Web Archive is held in different repositories, by different organizations. Because of this, it was challenging to provide copies to users. The size of the data presented additional logistical challenges in staging content for others to use.

2) *Offer access to the original WARC data from the EOT crawls* - Files in the archive come from different time periods, use different crawler infrastructures, come from different crawling partners, were stored in different file formats (ARC, which is WARC's predecessor, then WARC), and use varying standard WARC file sizes for storage. The goal is to maintain these characteristics whenever possible and therefore not to rename files, thereby providing users with the data essentially "as-crawled."

3) *Provide derivative files for each of the WARC files in the datasets* - Derivative files allow users access to subsets of the archive for research without requiring them to download the full WARC data. Generally, the WAT (Web Archive Transformation) and WET (Web Extracted Text) derivative files are only fractions of the size of the original WARC files. For situations where a research project only needs access to text data, they might only have to download the WET files, or if they are only using the link metadata, then the WAT files might serve their needs. If a user only wanted to understand the high-level presence of a domain or subdomain in the EOT dataset, the CDX files might be sufficient for that need.

4) *Follow existing practices where available* - Following existing practices allows for ease of use and compatible access structures. This meant replicating the decisions Common Crawl made around data layout, base derivatives, and documentation with a goal of leveraging existing tools, documentation, and hopefully communities.

5) *Augment legacy data with content-based characterization tools* - Resources in the EOT Web Archive have not been consistently characterized. The legacy crawl data needed to be augmented by applying content-based file format identification, character set identification, language identification, and Soft 404 identification. These techniques, while not new, have not been consistently applied over time while crawling.

6) *Enable archival playback and CDX server access to dataset* - Providing archival playback to the dataset allows for crawl-specific access points into the data that can be hosted directly from S3, with no need to stage another copy of data. A CDX server allows for API access that provides enhanced URL querying to the underlying indexes. In order to enable these services, ZipNum indexes for each web archive (2008, 2012, 2016, and 2020) and an index for the entire EOT dataset are necessary.

7) *Experiment with column-oriented formats for archive indexes* - In addition to formats that are common within the web archiving field, like WARC, WAT, WET, and CDX, other formats such as Parquet, which is used in cloud-based or big-data oriented environments, are making inroads into the field of web archiving [16]. Apache Parquet files are a column-oriented file type that is used to store and retrieve data more efficiently. Given that this type of file is used frequently in computational research, it made sense to provide Parquet files to store data found in the CDX and metadata sidecar (META) file formats so that this data can be queried and analyzed using frameworks that recognize Parquet.

4.1 Data Layout

A major goal of this project is to provide self-contained versions of each of the four web crawls. To enable this, each crawl, EOT-2008, EOT-2012, EOT-2016, and EOT-2020, have their own path structure in the Amazon S3 buckets. The EOT dataset team wanted to maintain the provenance of the data, including which institution was responsible for crawling the data. For example, in the EOT-2008 dataset, the crawls were conducted by the California Digital Library (CDL)[6], the Internet Archive (IA)[7], the Library of Congress

[5]Common Crawl https://commoncrawl.org/

[6]California Digital Library: https://cdlib.org/

[7]Internet Archive: https://archive.org

```
crawl-data/EOT-2008/segments/CDL-000/
crawl-data/EOT-2008/segments/CDL-001/
crawl-data/EOT-2008/segments/IA-000/
crawl-data/EOT-2008/segments/IA-001/
crawl-data/EOT-2008/segments/LOC-000/
crawl-data/EOT-2008/segments/UNT-000/
```

Figure 1: Example segment paths for the EOT-2008

```
crawl-data/EOT-2008/segments/CDL-000/cdx/
crawl-data/EOT-2008/segments/CDL-000/meta/
crawl-data/EOT-2008/segments/CDL-000/warc/
crawl-data/EOT-2008/segments/CDL-000/wat/
crawl-data/EOT-2008/segments/CDL-000/wet/
```

Figure 2: Example derivative paths for the EOT-2008

```
eot-index/collections/EOT-2008/indexes/
eot-index/collections/EOT-2012/indexes/
eot-index/collections/EOT-2016/indexes/
eot-index/collections/EOT-2020/indexes/
eot-index/table/eot-main/warc/crawl=EOT-2008/
eot-index/table/eot-main/warc/crawl=EOT-2012/
eot-index/table/eot-main/warc/crawl=EOT-2016/
eot-index/table/eot-main/warc/crawl=EOT-2020/
```

Figure 3: Example index and Parquet paths for the EOT dataset

(LOC)[8], and the University of North Texas Libraries (UNT)[9]. This dataset uses a replicate of the the Common Crawl organizational structure that uses segments to divide a crawl into subsets. Each segment in the EOT Dataset contains 10,000 or fewer WARC files. Because many of the crawl partners generated over 10,000 WARC files in the EOT project, it was necessary to create several segments per crawling partner. An example of this structure can be seen in Figure 1.

This structure works for organizing files in a normal POSIX file system, as well as in an object store like S3 by making use of the common concept of a directory prefix. Inside each segment, another prefix/folder structure for WARC, WAT, WET, CDX, and META files was created. This results in the final structure of a segment as can be seen in Figure 2.

The ZipNum and Parquet indexes are stored in a similar directory prefix layout but with a file path structure separate from the crawl data. The layout is presented in Figure 3.

To provide easily downloadable lists of paths, each EOT year and derivative type has a path file that contains the S3 paths to all of the files of that type. For example, the EOT-2008 WARC path file contains one line per WARC file in the EOT-2008 dataset. These path files can be iterated over to download the dataset from either the command line S3 interface, or via HTTPS.

[8]Library of Congress: https://loc.gov
[9]UNT Libraries: https://library.unt.edu/

4.2 Derivative Data

As mentioned above, one of the major goals of the project was to provide derivative formats for each of the WARC files that could be used in scenarios where the full content payload was not required. With this goal in mind, a WAT file that includes link structures, anchor text, and pertinent metadata from the payload file was generated. In addition, WET files that contain just the text from HTML and TXT formats from the WARC files were generated.

A metadata sidecar file was also created for each WARC file that includes content-based characterizations for each *response* and *resource* record in the dataset. These characterizations include content-based language identification using Python bindings for the Compact Language Detector 2 (CLD2) library [2], character set detection using the Python library Chardet: The Universal Character Encoding Detector [4], file format identification using Format Identification for Digital Objects (fido) [6], and mime type detection using python-magic [7], which is a Python interface to the libmagic file type identification library. Finally, in order to experiment with identification of the Soft 404 phenomenon, this project used the Python tool Soft404 [14] . The Soft 404 phenomenon occurs when a web server responds with an HTTP response code of `200 OK`, but returns a page that indicates that the content is not available instead of returning a `404 Not Found` response code. The output of each of these characterizations were combined into a single metadata record in the metadata sidecar file, which is a WARC file itself. This packaging of derivative files (WAT, WET, META) as WARC files allows for logical alignment to *response* and *resource* records in the original WARC files.

Finally, a CDX file in the CDXJ format was generated for each of the WARC files. The CDXJ format allows for additional metadata from the WARC records to be included in the final index.

5 MAJOR CHARACTERISTICS OF THE DATASET

A total of 461TB of WARC/ARC data were staged for the EOT dataset. This dataset represents four large crawl events occurring in 2008, 2012, 2016, and 2020. An overview of the dataset size is provided in Table 1.

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
Total	638,707	461TB	12TB	4TB	130GB	1TB

Table 1: Summary of End of Term Dataset on Amazon S3

The derivative data is much smaller than the original WARC data. For example, the EOT-2008 WARC data is 15.32TB in size, with the WAT derivatives being only 447GB, which is a 71% decrease in size. The WET and CDX derivatives offer a 95% and 99% size reduction respectively. When a researcher has questions that can be answered by the derivative formats, providing generated files like this can make the process of transferring and storing the data much easier.

6 SUMMARY AND FUTURE WORK

The EOT Web Archive Dataset is a longitudinal dataset of harvested web content from the federal web captured every four years as part of the End of Term Web Archive. This dataset presents the content harvested by a number of different institutions, using different crawling methods, scopes, configurations, and tools. The data includes all content as-crawled, though it may not be a a complete representation of the `.gov` or `.mil` space because things may have been missed at crawl time.

Future work for this dataset includes generation of host-level and domain-level network graphs that will show the relationships between domains within the EOT Web Archive. Future work is expected to continue to leverage existing tools and processes developed by Common Crawl for graph generation for this portion of the work. With the complete dataset available in CDX format, overviews of each of the EOT crawl years using CDX summarization tools [1] can be generated. These can be helpful in communicating the contents of this dataset to others.

As the EOT Web Archive grows with future U.S. presidential elections, the decisions and processes established for these first four web crawls can be replicated to further extend the longitudinal dataset.

This effort by the EOT dataset team to stage a copy of the four EOT crawls in the cloud has been helpful in understanding many of the challenges that organizations will face when thinking about staging content for large-scale computational use. The greatest challenge encountered during this project was accounting for the collections that had been stored in various repositories and infrastructures for over a decade. In many situations, those repository structures have changed in ways that are forgotten to the current EOT team and required investigation and the rebuilding of a knowledge-base of previous operations. The decision to base this work on the Common Crawl organizational structure and subsequently leverage existing tools and documentation was a major benefit to this project. If those tools had not been in place and previous examples were not available, the whole process would have been more challenging and required greater allocations of time and resources. The EOT dataset team is excited to make this dataset available more broadly to researchers who are interested in using the End of Term Web Archive in their research and scholarship.

ACKNOWLEDGMENTS

This effort would not have been possible without storage support from the Amazon Open Data Program, which provided S3 storage for this initiative. Likewise, this project leaned heavily upon the prior work of the Common Crawl team and adopted their organizational structures, tools, and documentation in building this dataset and providing access to it.

REFERENCES

[1] Sawood Alam and Mark Graham. 2022. CDX Summary: Web Archival Collection Insights. In *Linking Theory and Practice of Digital Libraries - 26th International Conference on Theory and Practice of Digital Libraries, TPDL 2022, Padua, Italy, September 20-23, 2022, Proceedings (Lecture Notes in Computer Science, Vol. 13541)*, Gianmaria Silvello, Óscar Corcho, Paolo Manghi, Giorgio Maria Di Nunzio, Koraljka Golub, Nicola Ferro, and Antonella Poggi (Eds.). Springer, 297–305. https://doi.org/10.1007/978-3-031-16802-4_25

[2] Rami Alrfou. 2022. PYCLD2 - Python Bindings to CLD2. https://github.com/aboSamoor/pycld2

[3] Amazon.com, Inc. 2022. *Open Data Sponsorship Program.* Amazon.com, Inc. https://aws.amazon.com/opendata/open-data-sponsorship-program/

[4] Dan Blanchard. 2022. Chardet: The Universal Character Encoding Detector. https://github.com/chardet/chardet

[5] Common Crawl. 2023. *Examples using Common Crawl Data.* Common Crawl Foundation. https://commoncrawl.org/the-data/examples/

[6] Open Preservation Foundation. 2022. Format Identification for Digital Objects (fido). https://github.com/openpreserve/fido

[7] Adam Hupp. 2022. python-magic. https://github.com/ahupp/python-magic

[8] International Internet Preservation Consortium. 2022. *WARC Specifications.* International Internet Preservation Consortium. https://iipc.github.io/warc-specifications/

[9] Library of Congress. 2023. *Web Archive Datasets.* Library of Congress. https://labs.loc.gov/work/experiments/webarchive-datasets/

[10] Ian Milligan, Nathalie Casemajor, Samantha Fritz, Jimmy Lin, Nick Ruest, Matthew Weber, and Nicholas Worby. 2019. Building Community and Tools for Analyzing Web Archives Through Datathons. In *2019 ACM/IEEE Joint Conference on Digital Libraries (JCDL).* Association for Computing Machinery, New York, NY, USA, 265–268. https://doi.org/10.1109/JCDL.2019.00044

[11] Mark E. Phillips and Kristy K. Phillips. 2017. End of Term 2016 Presidential Web Archive. *Against the Grain* 29, 6 (2017), 27–30. https://doi.org/10.7771/2380-176X.7874

[12] Nick Ruest, Samantha Fritz, and Ian Milligan. 2022. Creating order from the mess: web archive derivative datasets and notebooks. *Archives and Records* 43, 3 (2022), 316–331. https://doi.org/10.1080/23257962.2022.2100336

[13] Tracy Seneca, Abbie Grotke, Cathy Nelson Hartman, and Kris Carpenter. 2012. It Takes a Village to Save the Web: The End of Term Web Archive. *Documents to the People* 40, 16 (2012), 16–23.

[14] TeamHG-Memex. 2017. soft404: a classifier for detecting soft 404 pages. https://github.com/TeamHG-Memex/soft404

[15] UK Web Archive. 2023. *UK Web Archive Open Data.* UK Web Archive. https://data.webarchive.org.uk/opendata/

[16] Xinyue Wang and Zhiwu Xie. 2020. The Case For Alternative Web Archival Formats To Expedite The Data-To-Insight Cycle. In *Proceedings of the ACM/IEEE Joint Conference on Digital Libraries in 2020* (Virtual Event, China) *(JCDL '20).* Association for Computing Machinery, New York, NY, USA, 177–186. https://doi.org/10.1145/3383583.3398542