Moving the End of Term Web Archive to the Cloud to Encourage Research Use and Reuse

MARK E. PHILLIPS, University of North Texas, United States
SAWOOD ALAM, Internet Archive, United States

The End of Term Web (EOT) Archive is a collaborative project with a goal of collecting the United States federal web, loosely defined as .gov and .mil, every four years coinciding with presidential elections and often a transition in the Executive Branch of the government. In 2021 the End of Term team began to process the longitudinal web archive for EOT-2008, EOT-2012, EOT-2016, and EOT-2020 to move into the Amazon S3 storage service as part of the Amazon Open Data Program. This effort adopted tools, structures, and documentation developed by Common Crawl in an effort to maximize potential research access and reuse of existing tools and documentation. This paper presents the process of organizing, staging, processing, and moving these collections into the Amazon cloud.

Additional Key Words and Phrases: web archives, cloud storage, research datasets, web archive datasets

ACM Reference Format:
Mark E. Phillips and Sawood Alam. 2022. Moving the End of Term Web Archive to the Cloud to Encourage Research Use and Reuse. In . ACM, New York, NY, USA, 4 pages. https://doi.org/XXXXXXX.XXXXXXX

1 BACKGROUND

The End of Term (EOT) Web Archive [1] is a collaborative project with a goal of collecting the United States federal web, loosely defined as .gov and .mil, every four years coinciding with presidential elections and often a transition in the Executive Branch of the government. Starting in 2008 [14], this project documented the federal web before the transition of the Bush administration to the Obama administration, then documented the transition from one Obama term to another in 2012, the transition from Obama to Trump in 2016 [13], and the transition from Trump to Biden in 2020. In total, the EOT has collected nearly 500TB of content in its four iterations. The EOT is an ad-hoc collaboration that comes together every four years to plan, publicize, and execute the crawls related to this effort. Long-term access to the content is often a more challenging component of the process. So far, access has been provided by the Internet Archive through different configurations of their Wayback Machine and currently access is provided by the Global Wayback collection. Additionally, some members of the EOT team have curated and hosted secondary access points to the crawled content in their own infrastructure to provide redundancy in access. This provides a minimum level of access to the harvested resources for general users, but over the years the EOT team has found that there are logistical limitations in place when users want to use the EOT archives to answer computationally focused research questions that would require larger portions of the archive. In the Fall of 2021, the EOT began to explore working with the AWS Open Data Sponsorship Program [3] to host a copy of the four EOT web crawls as a longitudinal dataset.

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

Web Archiving and Digital Libraries '22,
,

ACM ISBN 978-x-xxxx-xxxx-x/YY/MM…$15.00
https://doi.org/XXXXXXX.XXXXXXX

2 ORGANIZING THE COLLECTION

Before moving the EOT collections to the cloud, there were a number of decisions to be made about how the data would be organized and made available. The AWS Open Data Sponsorship Program provides high-speed cloud storage for open datasets through the Amazon Simple Storage Service (S3) [1]. The EOT team looked for prior work in this area and decided on the organizational structures in place in the Common Crawl program [9]. Common Crawl broadly crawls the web and provides the data freely to users for research and analysis. In addition to the crawl data, Common Crawl will create derivative formats for each WARC file consisting of a Web Archive Transformation (WAT) [5] which provides content-metadata about the crawled resources such as out links, anchor text, and overall structural information. Another file provided is Web Extracted Text (WET) files that present just the text for formats like HTML and TXT to the user. These WAT and WET files are created for every WARC file in the dataset. An index of all captured content is provided in the CDXJ format [6] and organized using the ZipNum structure. Finally, the CDXJ data is processed and compiled into a columnar data format called Parquet [2] that provides another entry point into the collection that can be used by many common tools and services. By adopting these structures for the EOT collections, the team was able to build on existing workflows and leverage tools used by Common Crawl in their processing pipeline. Another goal was to work with researchers already using Common Crawl data to allow the EOT dataset to fit into their research workflows and tools with little modification. Finally, the ability to reuse and adapt existing documentation about the formats and processes was also a benefit of using existing formats.

3 LAYOUT OF THE CRAWLS

One goal of this project is to provide self contained versions of each crawl. To enable this, each of the four End of Term crawls, 2008, 2012, 2016, 2020 would have their own path structure in the Amazon S3 buckets. The EOT-2008, EOT-2012, EOT-2016, and EOT-2020 crawls can be seen as analogous to the Common Crawl monthly crawl structures in a straightforward way. The next thing the EOT team wanted to maintain was the provenance of which institution was responsible for the crawling of the data. For example, in the EOT-2008

[2] Apache Parquet https://parquet.apache.org/

dataset, the crawls were conducted by the California Digital Library (CDL) [3], the Internet Archive (IA) [4], the Library of Congress (LOC) [5] and the University of North Texas Libraries (UNT) [6]. The Common Crawl organizational structure includes a concept of segments that divide a crawl into subsets. These segments are holdovers from the distributed Nutch crawler that they use for collecting content. Each segment in the Common Crawl dataset contains under 10,000 WARC files and the EOT team felt that provided for a reasonable limit for others looking to download and work with the content locally. Many of the crawl partners generated over 10,000 WARC files in the EOT project and it was necessary to create several segments per crawling partner. An example of this structure can be seen in the example below.

```
crawl-data/EOT-2008/segments/CDL-000/
crawl-data/EOT-2008/segments/CDL-001/
crawl-data/EOT-2008/segments/CDL-002/
crawl-data/EOT-2008/segments/IA-000/
crawl-data/EOT-2008/segments/IA-001/
```

This structure works for organizing files in a normal POSIX filesystem as well as in an object store like S3 by making use of the common concept of a prefix. Inside each segment, another prefix/folder structure for WARC, WAT, WET, and CDX files was created. This results in the final structure of a segment as you can see below.

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

The ZipNum and Parquet indexes are stored in a similar layout but with a path structure separate from the crawl data. The layout is presented below.

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

4 INVENTORY OF DATA

One of the surprisingly challenging parts of this project was to completely identify the crawl data within different institutions' respective repository infrastructure. Many institutions will include their web archives in both a preservation repository to provide long-term stewardship, replication, and structured access to the files, as well as locating them in a secondary access system where they are indexed and served using a replay system such as Open Wayback [7] or pywb [8]. Because of this, it can be challenging to identify all of the components of a large web crawl within a repository. Additionally, because the EOT crawls were completed by different institutions and then aggregated into single collections, it was challenging to identify contributed WARC files from the locally crawled content. Finally, because the EOT projects generally last from September until March of the following year, there are many individual crawls during that period that all need to be accounted for during the process. In the EOT project, the EOT-2008 and EOT-2012 crawls were fully replicated by three institutions: the Internet Archive, Library of Congress, and the UNT Libraries. For the EOT-2016 and EOT-2020 crawls, the only complete copy of the data is held at the Internet Archive, with the crawling partners maintaining a copy of their own crawled data. Because of this distribution of content, the EOT team was able to split the responsibility of inventorying all content between the Internet Archive for the EOT-2016 and EOT-2020 crawls and UNT for the EOT-2008 and EOT-2012. As an example to demonstrate this effort, the EOT-2008 crawl totalled 16TB of data and was distributed across 110 Archival Information Packages in the UNT Libraries' Coda preservation repository [11]. Similar divisions of a complete EOT crawl across dozens to hundreds of archival packages occurred at both IA and UNT for the other crawls.

Once inventories were complete, the next step was to download the archival packages, verify that everything was complete and valid based on package checksums, and then reorganize the WARC content into the structures mentioned above. The EOT-2008 crawls contained both WARC and ARC files, whereas the remaining crawls only held WARC files. The decision was made early in the project to not rename files but to leave them as they were originally contributed for better provenance and lineage. There were a small number of WARC/ARC files (36) from the CDL dataset in EOT-2008 that had duplicate filenames but different content and those were renamed with "-duplicate-name-" inserted into the filename to allow them to be included. The EOT team decided early in the process not to concatenate, modify, or convert the content files into other formats to preserve integrity at file and record levels. This would have included converting the many 100MB WARC/ARC files into larger 1GB files, format conversion of ARC files to the modern WARC format, updating legacy WARC versions to modern versions of the specification, or including additional metadata records inside existing content files. When additional metadata is desired for these files, it will be generated and stored in a separate file being referred to as a "metadata sidecar file" [12]. This decision allows for the provenance of the files to be maintained for archival purposes, but does require additional attention to be paid to the early crawls when building tools because of mixed use of the ARC/WARC format and early versions of the WARC standard that are present in the EOT-2008 crawls.

5 CREATING THE DERIVATIVES

In order to provide datasets that can be used in a wide variety of applications, there was a need to create derivatives of the ARC/WARC files for different use-cases. The Common Crawl organizational structure mentioned above includes standard derivative formats in the web archiving community of WAT and WET files. For the EOT-2008 and EOT-2012 crawls, data was aggregated and processed

[3] California Digital Library: https://cdlib.org/
[4] Internet Archive: https://archive.org
[5] Library of Congress: https://loc.gov
[6] UNT Libraries: https://library.unt.edu/
[7] Open Wayback: https://github.com/iipc/openwayback
[8] Webrecorder pywb: https://github.com/webrecorder/pywb

at the UNT Libraries before uploading into the AWS S3 service. For the EOT-2016 and EOT-2020 crawls, the EOT teams planed to upload the WARC data to AWS S3 and then create the derivative files afterward. Derivatives were generated using the Common Crawl branch of the ia-hadoop-tools [4] package and specifically the WEATGenerator functionality. No modifications were made to this tool before running it on the collections. The index files were generated by the CDX-indexer from the pywb project. We used the CDXJ format with flags to sort each output file and also include the full relative path to the WARC/ARC file based on the Common Crawl organizational structure. The ZipNum format was generated using scripts in the webarchive-indexing repository [7] again from Common Crawl. Finally, the cc-index-table repository [8] was used to generate the Parquet format from the ZipNum index [10]. All of these tools required no modification to the base code and only small configuration changes to make them work in our various processing environments.

6 UPLOADING THE DATASETS

Once datasets had been locally staged, verified, and derivatives created, the next step in the process was to load the data into the AWS S3 infrastructure. The AWS Open Data Sponsorship Program provides access to a storage bucket in this case called "eotarchive" where the original and derived data of various crawls would be uploaded. The EOT team made use of the AWS Command Line Interface [2] to load data into the service. At UNT this was generally accomplished one segment at a time after all derivatives were generated. At the Internet Archive, WARC files were loaded individually as a sequential process that was later run in parallel. These two approaches were necessary due to local organizational structures and infrastructure constraints for staging large datasets. While the project was only nominally interested in understanding the throughput of the different approaches, it can be noted that at UNT the upload speed of data to AWS was limited by local IO from disk, and at IA was generally limited by network bandwidth. The resulting datasets for EOT-2008 and EOT-2012 took about a month each to stage, create derivatives, and upload at UNT. EOT-2016 and EOT-2020, being much larger in size, are still ongoing from IA and are expected to take upwards of six months to complete the initial upload.

7 DOCUMENTATION AND PUBLIC ACCESS

The final step in the process includes the documentation of the datasets for researchers and also providing guides and examples on how to use these datasets to answer research questions. The two datasets that are completed are available at the End of Term website [9], where the others will be added in the future. Documentation of the formats as well as guides and examples for using these data formats will be based on the previous work of the Common Crawl project. It is expected that during the summer of 2022 this documentation will be generated by the EOT team and the final datasets will be registered with the AWS Open Data Sponsorship Program in their Open Data on AWS catalog. The team also has an interest in working with tools from the Archives Unleashed [10] program to document how these datasets can be used with their tools and services. For a complete listing of dataset sizes see Table 1.

Dataset	WARC #	WARC Size Compressed (TiB)
EOT-2008	125,704	16.85
EOT-2012	78,509	45.57
EOT-2016	TBD	159 + 150 (FTP)
EOT-2020	TBD	300

Table 1. Summary of End of Term Datasets on Amazon S3

8 CLOSING

This effort by the collaborative End of Term Web Archive 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 team is excited to make these datasets available more broadly to researchers who are interested in using the End of Term web archives in their research and scholarship.

ACKNOWLEDGMENTS

This effort would not have been possible without the 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 these datasets and providing access to them.

REFERENCES

[1] Amazon. 2022. *Amazon S3.* Retrieved May 1, 2022 from https://aws.amazon.com/s3/
[2] Amazon. 2022. *AWS Command Line Interface.* Retrieved May 1, 2022 from https://aws.amazon.com/cli/
[3] Amazon. 2022. *Open Data Sponsorship Program.* Retrieved May 1, 2022 from https://aws.amazon.com/opendata/open-data-sponsorship-program/
[4] Internet Archive. 2020. *ia-hadoop-tools.* Retrieved May 1, 2022 from https://github.com/commoncrawl/ia-hadoop-tools
[5] Jefferson Bailey. 2016. *WAT Overview and Technical Details.* Retrieved May 1, 2022 from https://webarchive.jira.com/wiki/spaces/ARS/pages/90997503/WAT+Overview+and+Technical+Details
[6] International Internet Preservation Consortium. 2016. *OpenWayback CDXJ File Format 1.0.* Retrieved May 1, 2022 from https://iipc.github.io/warc-specifications/specifications/cdx-format/openwayback-cdxj/
[7] Common Crawl. 2019. *WebArchive URL Indexing.* Retrieved May 1, 2022 from https://github.com/commoncrawl/webarchive-indexing
[8] Common Crawl. 2022. *Common Crawl Index Table.* Retrieved May 1, 2022 from https://github.com/commoncrawl/cc-index-table

[9] End of Term Web Archive Datasets: https://eotarchive.org/data/
[10] Archives Unleashed: https://archivesunleashed.org/

[9] Common Crawl. n.d.. *Common Crawl.* Retrieved May 1, 2022 from https://commoncrawl.org/
[10] Sebastian Nagel. 2019. Accessing WARC files via SQL. In *2019 International Internet Preservation Coalition General Assembly and Web Archiving Conference* (Zegreb, Croatia). https://digital.library.unt.edu/ark:/67531/metadc1608961/
[11] University of North Texas Libraries. 2022. *Coda.* Retrieved May 1, 2022 from https://github.com/unt-libraries/coda
[12] University of North Texas Libraries. 2022. *warc-metadata-sidecar.* Retrieved May 1, 2022 from https://github.com/unt-libraries/warc-metadata-sidecar
[13] 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
[14] 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.