End of Term 2016 Presidential Web Archive

by Mark E. Phillips (Associate Dean for Digital Libraries, the University of North Texas) <Mark.Phillips@unt.edu>

and Kristy K. Phillips (University of North Texas) <kristy.phillips@unt.edu>

Introduction

During every Presidential election in the United States since 2008, a group of librarians, archivists, and technologists representing institutions across the nation can be found hard at work, preserving the federal web domain and documenting the changes that occur online during the transition.

Anecdotally, evidence exists that the data available on the federal web changes after each election cycle, either as a new president takes office, or when an incumbent president changes messages during the transition into a new term of office. Until 2004, nothing had been done to document this change. Originally, the National Archives and Records Administration (NARA) conducted the first large-scale capture of the federal web at the end of George W. Bush’s first term in office in 2004 (https://www.webharvest.gov/). This is noteworthy because, while institutions like the Library of Congress, the Government Publishing Office, and NARA itself have web archiving as part of their imperative, none of their mandates are so broad as to cover the capture and preservation of the entirety of the federal web. On April 15, 2008, NARA released the document “National Archives and Records Administration Web Harvesting Background Information,” which detailed the reasons why the organization decided not to continue this large-scale archival practice during the following election in 2008. As such, a group of interested organizations gathered together to continue the project.

The End of Term (EOT) projects began with the Internet Archive, the Library of Congress, the University of North Texas, the California Digital Library, and the U.S. Government Publishing Office working together to fill the void left by NARA and archive the entirety of the federal web during the transition period in the wake of the 2008 presidential election. Since that first capture, new partners have joined the team, including Harvard University in 2012, and George Washington University and Stanford University in 2016.

continued on page 28
Every year, the process is updated and expanded. Every election brings its own challenges, but the unanticipated outcome of the presidential election of 2016 brought an especially eventful harvest, with people all over the country suddenly interested in what was captured during this particularly divisive transition. The EOT projects have several areas of organization, including seed collection, harvesting, and public outreach, that were affected by the changes brought by the most recent presidential election.

What to Harvest

The first step involved in a successful harvest is deciding what, exactly, needs to be captured. The End of Term project team has experimented with different ways of establishing the scope of the project each time it is completed, and several of them were used during the 2016 EOT project. Web harvesters require a set of starting URLs, or “seeds” that dictate where to begin the crawling process. To start, the harvester downloads the page designated by a seed URL, extracts all of the URLs on that page, then checks whether the extracted URLs have been crawled, and if they have not, it adds them to the list of URLs to crawl. This process is repeated until the list of new URLs has been exhausted, or until the crawler has been stopped by some other means. This can be done by the operator, or based on some threshold like total gigabytes downloaded, number of URLs in the crawl, or length of time crawling. The federal web has a number of high-level websites that are entry points for users into the wide range of content that is available on the federal web. Sites like USA.gov provide an entry point in the format of a search and discovery portal. Unfortunately not all URLs in the federal web are identified in these systems, so the EOT project group first had to work to identify the overall scope of what content we would harvest.

To identify the seed URLs that the EOT project would crawl, and therefore identify the scope of the crawling effort, the team used two primary methods of collecting seeds. These methods were bulk seed lists and URL nominations.

Bulk Seed Lists

It may be somewhat surprising, but there is not a definitive list of all of the domains and subdomains that are managed by the federal government. The U.S. General Services Administration (GSA) has created the U.S. Digital Registry which is an official list of a large number of these domains, but it is by no means exhaustive. Different groups within the government handle the registration of .gov and .mil domains, both of which are in the scope of the EOT project. Outside of the domain names, subdomains are often created and managed within the agency that created them, meaning they don’t make the standard lists of federal websites.

During the 2016 EOT project, the team used seven or eight different bulk seed lists, some from previous web crawls, and others collected from related projects. Once the lists were compiled, they were added to an instance of the URL Nomination Tool that the project team uses to manage them. Ultimately, a total of 43,674 seed URLs from ten different lists were added during the course of the project (http://digital2.library.unt.edu/nomination/eth2016_bulk).
hard drives. The data, stored in WARC files, included fixity hashes to verify file integrity. Altogether, the collecting partners gathered more than 200 TB of data. The Internet Archive loaded the aggregate collection of the 2016 EOT into an instance of the Wayback machine, and access records were added to the projects website (http://eotarchive.cdlib.org/).

End of Term Web Archive Website

Lessons Learned in the 2016 EOT Project

Planning for the project began in January of 2016. The team held monthly calls open to all interested parties. The project was a bit different in this election cycle, as the team knew that there would be a transition in the executive branch of government, given that the previous president had reached his term limit. This allowed for a more concrete plan.

The project began as anticipated in mid-September, and the team was moving forward with content capture. Then, in November, the election happened, and Donald Trump was announced as the 45th President of the United States. The result was unexpected for many people, and some were concerned about the possibility of this new administration removing content from the web after the President took office, especially since the administration’s positions on subjects like climate change were quite different from those of the previous administration.

Some people in academia, particularly the sciences, publicly expressed this concern, and the media published a number of stories discussing the possibility of important content being lost or removed during the transition. A number of initiatives formed in response to this concern, like the Guerrilla Archiving Event: Saving Environmental Data from Trump, which was held during December 2016 in Toronto, and several Data Refuge projects that were conducted during the winter of 2016 and the spring of 2017.

This brought a lot of new attention to the EOT project. The project was suddenly exposed to a much broader audience, and it was a blessing in many ways, as it brought with it publicity and interest in the project itself and in the institutions that were working so hard to collect and preserve this content for future generations. The possibility of losing content from federal websites came to the forefront of many more people’s minds than it had in years past.

This did present some challenges, however. While many people were suddenly thinking of preserving content from the federal web in the first week of November, the EOT project team had been planning the harvest since January, and had done the work for the two elections prior. The community’s sudden desire to participate was unexpected, and the team struggled to find a way to harness all of this public energy in a productive way. Companies were interested in providing storage and computer infrastructure for the project. Individuals wanted to crawl content on their own and then contribute it to the project. People that didn’t know how they could help wanted to talk to the team about ways that they could contribute. The team was almost overwhelmed by eager assistants with nothing specific they could do.

continued on page 30
Finally, the team suggested that the most helpful activity for volunteers was to nominate the URLs of the items that they believed most at risk via the URL Nomination Tool. This influx of nominations helped identify a wide range of content from websites to individual PDFs and datasets. It was a great help, and it allowed people to contribute in a way that they found meaningful. It also exposed a problem with the project: the team needed a better web presence to communicate with the public. Currently, the team has a Twitter account that was active during the project, but that is clearly not enough, as it is difficult to use as the only primary news and information outlet. In addition, the EOT project’s interface, which is hosted by the California Digital Library, wasn’t designed to have a section that listed new content, so updating the public via this resource simply wasn’t possible. Now, one of the major goals for the 2020 EOT project is to have a better news and information platform for communicating with those who are interested, including information about the project and how people can help.

Conclusion
The End of Term projects in 2008, 2012, and 2016 were volunteer efforts by a number of institutions across the U.S. The time, effort, and infrastructure are all donated by the participating organizations. The individuals from these institutions are the ones that moved the project forward and made it successful. The 2016 election cycle offered new challenges and opportunities in relation to project management, channeling user interest, fielding media requests, and gathering and sharing the harvested content. While there were challenges, they were insignificant in comparison to the overall benefit of the project, as well as the accomplishments of the project and its project team.

References
Data Refuge – https://www.datarefuge.org/
End of Term Archive Website – http://eotarchive.cdlib.org/
End of Term Publications Collection – https://digital.library.unt.edu/explore/collections/EOT/
Social Feed Manager – https://gwu-libraries.github.io/sfm-ui/
URL Nomination Tool EOT 2016 Bulk – http://digital2.library.unt.edu/nomination/eth2016_bulk/