# Collaborating to Preserve Federal Government Websites

THE LIBRARY OF CONGRESS IS WORKING WITH OTHER LIBRARIES AND INFORMATION ORGANIZATIONS TO CAPTURE AND ARCHIVE CONTENT FROM GOVERNMENT WEBSITES BEFORE NEW PRESIDENTS TAKE OFFICE.

**BY ABIGAIL GROTKE**

The Library of Congress began archiving the web in 2000 with a pilot program, collecting content related to the 2000 U.S. election, including sites associated with the presidential, congressional, and gubernatorial elections. Since then, more than a petabyte of web content has been preserved in 90-plus collections covering a variety of topics and different types of websites. Many of these collections are available for researching through the library's website; others are not yet accessible, as they are in various states of processing. All content archived by the library is embargoed for one year.

Through its web archiving, the library builds collections for members of Congress, researchers, and the public and preserves born digital content for long-term research use. Unlike some of our partners and colleagues at other national libraries around the world, we have no easily defined U.S. domain, so we take a selective approach to web archiving. While many of our partners and colleagues at other national libraries around the world can collect their country's domain comprehensively, the U.S. domain is too extensive to permit us to collect everything. We focus on events and thematic web archives covering a variety of topics selected by recommending officers (the subject specialists who select content for the library's collections) according to collection development policies and other guiding documents.

Federal websites have figured prominently in the library's web archives since the program's inception, but it was not until 2015 that a systematic approach to harvesting federal sites was adopted. In addition to comprehensively collecting legislative websites since 2003, the library's web archiving program now seeks to broadly archive websites from all branches of government. We comprehensively harvest all judicial branch websites quarterly; we collect only selectively from the executive branch due to the large number and size of its websites and the commitments by other agencies (GPO, NARA, etc.) to archive them. As a result, the library focuses its archiving efforts on cabinet-level agencies and the affiliated programs that complement the library's judicial and legislative collections, as well as a few smaller agencies. We do not archive national labs or the majority of .mil sites.

## The End of Term Archive

Since the library began web archiving in 2000, we have had a strong history of collaborating with other organizations on everything from developing tools to discussing and formulating policies and approaches to collaborative collections building. In 2003, we became a founding member of the International Internet Preservation Consortium (IIPC)

**ABIGAIL GROTKE** is lead information technology specialist, Digital Collection Management and Services Division, at the Library of Congress. She can be reached at abgr@loc.gov.

and have worked closely with members of the IIPC throughout the years. We are also founding members of the Federal Government Web Archiving Working Group, which is a collection of federal agencies working together to ensure long-term access to historical U.S. government resources through web archiving.

In addition to what the Library of Congress harvests for its own archives, since 2008 we have collaborated to document the changes to U.S. federal websites during presidential transitions through a project known as the End of Term Archive (EOT). EOT seeks to document federal websites prior to a change in administration.

For many years, IIPC members in the United States have been preserving portions of the federal web, depending on their own collection policies. Other than early efforts by the National Archives to preserve .gov content in 2004, no one institution in the United States has attempted to preserve the entire .gov domain comprehensively due to the scale of the effort. Simply identifying all government content on the web is a huge challenge. There is no one list of all .gov domains available for easy reference, and there is much web content produced by the federal government that is published outside the .gov domain, including social media content posted on third-party sites and websites produced by federal agencies on .edu, .mil, and .com.

In 2008 and 2012, the Library of Congress partnered with the University of North Texas Libraries, the Internet Archive, the Government Publishing Office, and the California Digital Library to take on the task of identifying and archiving .gov websites. In 2016, the Stanford University Libraries and the George Washington University Libraries joined the End of Term project. There has been no dedicated funding for this initiative—this is a collaborative effort in which each institution relies on its own available funding, with the work coinciding with archiving efforts that our organizations naturally perform already. Combining efforts on an enterprise of this scale has proved beneficial to all involved, as each partner has expertise in different areas to contribute to the whole.

**A major aspect of the End of Term project has been to open up the nomination and selection of websites for preservation to the general public.**

Partner roles have varied depending on the skills, interests, and availability of staff. There is no one lead institution or person, although the Library of Congress, University of North Texas Libraries, and the Internet Archive typically take leadership and coordination roles. The partners meet and begin planning about six to eight months prior to the start of web crawling, with monthly conference calls continuing throughout the crawling activity and regular communications conducted through a project listserv.

The work is broken out into tasks such as distributed crawling, selection and gathering of seed lists to crawl, nomination tool development, outreach, volunteer recruitment, project management, and access. One early discussion usually involves identifying the parts of the government web each partner might help preserve or what other contributions each can provide (if not crawling). Preservation copies are also stored at any partner institution able to take in a copy of the data, since a goal of the project is to ensure that multiple copies are preserved by various partners, even if access to the entire archive is only provided by one organization.

As an example, the Library of Congress's EOT contribution has varied over the years. As project manager and team lead for the Web Archiving Team, I serve as the primary contact on the project for the library. Other library colleagues contribute time and resources, including subject matter experts on the collections side and technical staff on my team who manage our crawling and transfer activities.

In 2008, Library of Congress staff participated by helping with project management, volunteer recruitment and coordination, and transfer of the complete archive between partners taking in a copy of the data. In terms of contributing crawling resources, the library performed a more in-depth crawl of congressional websites we were already capturing monthly and contributed that data to the collaborative archive.

In 2012, we again performed an in-depth capture of congressional websites and helped with project management and with volunteer recruitment and coordination. We also helped promote the archive, speaking at events with other partners about the effort to raise awareness and recruit volunteers. Transfer was handled a bit differently in 2012, so we didn't get directly involved in managing all of the data.

For the 2016 archive, we again served as one of the project coordinators. We worked with volunteers and promoted the archive, helping respond to increased interest by community members and the press regarding the preservation of federal government websites and data. In 2016 we also expanded our crawling efforts, conducting an in-depth crawl of all of the federal government content that we were already preserving (not just congressional websites, but everything .gov that had been selected for our ongoing archives). This contributed an additional 35 terabytes to the effort.

Between late 2016 and early 2017, End of Term partners archived more than 155 terabytes of government websites and data for the EOT Archive. A related effort at the Internet Archive preserved an additional 100TB of federal FTP files.

User access to the 2008 and 2012 EOT Archive is provided by the Internet Archive and the California Digital Library

through the main End of Term portal. Access to the 2016 archive at that site is still to come, as preservation was the primary focus of the project until recently. That said, every web page the Internet Archive has archived for the recent project is accessible through the Wayback Machine. The Internet Archive has also posted some preliminary statistics for its crawls, which can be found on the End of Term (EOT 2016) summary statistics page. That information and additional data are served by a public EOT 2016 statistics API, and the Internet Archive is making data sets available on its site for broader research use.

With respect to the End of Term data that the Library of Congress is preserving currently, the library has a preservation copy of the complete 2008 EOT archive and portions of the EOT 2012 archive. The library is planning to take in a preservation copy of the 2016 web crawls in fiscal year 2018.

## EOT Outreach and Volunteer Efforts

One of the goals of the End of Term project has been to raise awareness of the importance of preserving federal government websites, and one way we do that is by involving the public in the effort. A major aspect of the End of Term project has been to open up the nomination and selection of websites for preservation to the general public. The project relies on gaining access to available bulk lists, and while more sources were available in 2016 than in prior years, there are still challenges in identifying all federal government web content. To help solicit nominations from the public, the University of North Texas Libraries created a tool in 2008 that is used by End of Term as well as other collaborative web archiving projects. Interested citizens can use the nomination tool to nominate URLs, provide basic metadata, and view nominations that have been received already.

In 2008, we primarily targeted government document subject experts for assistance and received about 500 nominations. In 2012 we enlisted similar types of people to help and expanded our target population to include students at Pratt University, who identified social media content for the project. In all, about 1,500 nominations were submitted.

In 2016, with increased press about the project as well as more public interest in the importance of preserving websites, the nomination tool processed almost 11,400 nominations. Through the EOT's collaboration with DataRefuge, the Environmental Data and Governance Initiative (EDGI), and other efforts, a total of 100,000 webpages or government datasets were nominated by citizens and preservationists for archiving.

This increase in public interest was wonderful for the project and exciting to see, but it also challenged our EOT model of the tasks required to conduct the project. Partners scrambled to manage press requests and answer questions from the public. We developed FAQs and other information resources to help spread the word about the project and set up a new listserv for outreach purposes. Meanwhile, partners from Stanford, the Internet Archive, the University of North Texas, and the Library of Congress fielded questions on a regular basis during the project period. All of this came on top of our other duties to ensure the preservation process proceeded as planned.

## How to Get Involved

The End of Term project has offered the Library of Congress a unique opportunity to collaborate to build collections of federal web data that greatly exceed what we might be able to build ourselves. Collaborative web archiving efforts such as the End of Term Archive offer opportunities to share library resources and expertise with partner institutions, raise awareness of our activities and web preservation in general, and enhance our own collection efforts. While the practice of web archiving has matured and become more of a regular activity for many libraries, archives, and other cultural heritage institutions, the amount of data produced keeps expanding, presenting corresponding challenges for preserving and accessing the archives.

Even as the End of Term partners work to make the entire 2016 archive available for research use, we are beginning to reflect on lessons learned and thinking about ways to improve the project in 2020, including ways we can engage other organizations and enthusiastic individuals to help in this effort. We encourage those interested to contact the EOT project team with any inquiries. You may also want to follow related data rescue efforts and the work of the Libraries+ Network, which has formed as a result of continued interest in the topic of preserving federal data.

And while the specific 2016 End of Term collection has closed, the Internet Archive has continued efforts to preserve the government web. Working with the University of North Texas, the Internet Archive has launched a Government Web & Data Archive nomination form so the public can continue to nominate government websites and data for archiving. **SLA**