MODULE 14

APPRAISING DIGITAL RECORDS

Geof Huth

SOCIETY OF American Archivists

Appendix B: Case Studies

Case Study 1: Appraising One State Government's Websites

In 2006, in the face of the imminent departure of three-term governor George Pataki, the New York State Archives, in coordination with the State Library, began to capture state government websites, including records from all three branches of government and from those public benefit corporations and public authorities with statewide responsibilities. In 2010, the State Archives expanded the Web crawl to capture the social media presences of these state government entities. In 2011, five years after the crawls first began, the State Archives drafted a Web crawl plan to clarify the schedule at which the archives would capture these websites. The plan was not a de facto appraisal report, but it did include features of one. Its goal was to produce the most complete capture of those websites with the least staff resources.

The original Web crawl plan included the following schedule for crawls:

- Capture of the governor's websites at the end of each calendar year
- Capture of websites of the legislature several weeks after each statewide election
- Capture of all state government entities' websites every four years, just before or after the end of each gubernatorial term of office

The plan also outlined a number of exceptions to these general rules:

- Immediate capture of sites created by constitutional officeholders (governor, attorney general, state comptroller) who leave office before the end of their terms
- Immediate capture of the Web pages of state legislators who leave office before the end of their terms
- Immediate capture of sites created by state entities about to be abolished or merged into other entities
- Capture of any state government sites created or discovered after the initial crawl

In 2013, the archives decided to question its conclusions on this transfer plan by conducting a complicated appraisal that took about

a year of part-time effort to complete. The main questions posed were whether the Web crawl plan brought in too many redundant records and whether it failed to capture all targeted records. An additional reason for this review was that the archives held hundreds of series of Web-based records by 2013, and capturing and describing each full-scale Web crawl required approximately two full-time employees working for the equivalent of eight months with an additional 150 hours of an intern's time.

The archives also used this appraisal to evaluate what records it was actually collecting via these Web crawls. The appraisers first identified a core of ten archival records series common to all state agencies and that were often made available on state entity websites. These included annual reports, operational plans, press releases, publications, and minutes of governing or advisory bodies. The archivists then chose a sample of state government websites, ensuring that the agencies were distinct in terms of mission, size, and complexity. Large agencies, small ones, and even medium-sized ones were in the mix.

The appraisers reviewed the captured websites and the detailed catalog records that described each site to determine the frequency at which agencies were removing older records in a series. The appraisal discovered that most state government entities kept most of the data online long enough that no data were lost between Web crawls. There were exceptions, but those generally came from the lack of a plan to document transitory activities, such as responses to significant disasters. The appraisers recommended two changes to the Web crawl plan and noted the need to evaluate whether to retain all data in the accessioned Web crawls:

- Set up a plan to crawl the websites of any state entities significantly involved in response related to major disasters declared so by the federal or state government.
- Actively monitor the websites of the thirty-three state entities that do not routinely transfer records to the State Archives, and schedule more frequent crawls if the rate at which the entities remove older information increases.
- Weed from the Web crawls websites or subsites that contain few or no records of value.

This reappraisal of sorts thus allowed the archives to justify its current methodology while adding some improvements to that plan. In

the end, the appraisal did not provide solutions to a few issues, and even though those were not meant to be addressed in the appraisal report, this became an important concern of the appraisal team (a larger group of staff that review all appraisal reports). Although the Web crawls preserve hundreds of series, they do not save those series in easy-to-find sets. Instead, each series is broken into overlapping chronologically truncated sets of records, so a user hoping to review an entire series must move from Web crawl to Web crawl to follow that series. In addition, there does not currently exist a quick and easy means of directing users to each Web page for a series, so the users have to search for the series on each Web crawl, and the location of the series within individual Web pages will likely change over time. Finally, the archives continued to accession, directly from state agencies, separate copies of records that were also captured in the crawls. The team decided to continue the practice of duplicate accessioning, because the archives could present the non-Web-based records versions of these series to users in single sequential series that were easier to use.