Journal of Western Archives

Volume 8
Issue 2 *Web Archiving Special Issue*

Article 1

2017

Introduction to the Special Issue on Web Archiving

Nicholas Taylor
Stanford University Libraries, ntay@stanford.edu

Introduction to the Special Issue on Web Archiving

Cover Page Footnote

Many folks contributed to make this Special Issue not just possible, but also great. Thanks to J. Gordon Daines III and, previously, Trevor Alvord for the invitation and opportunity to assemble a special issue on web archiving. J. Gordon Daines also deserves praise for his consistent editorial support throughout the production. Thanks to the submitters for their diverse and high-quality contributions to the open access literature on web archiving. Thanks to the reviewers for their deliberate and constructive feedback on multiple manuscript drafts; the improvements this yielded were tangible. Thanks to the Journal of Western Archives Editorial Board for their participation in meta-reviewing. Thanks to Cory Nimer for copyediting and formatting support. Thanks finally to the web archiving community, whose spirit of sharing and collaborative problem-solving is an inspiration for all of this work.

Introduction to the Special Issue on Web Archiving

Nicholas Taylor

There has never been as much web archiving as there is now, in terms of community breadth, grant funding, collaborative initiatives, collected content, and web archive repositories. Respondents to the biennial NDSA Web Archiving Survey have increased from 77 in 2011[1] to 106 in 2016,[2] with 44% of respondents having initiated their programs in the two years preceding the most recent survey.[3] The Society of American Archivists Web Archiving Roundtable now counts 968 members,[4] placing it in the top ten roundtable groups by size. Institutional membership in the International Internet Preservation Consortium has grown from its twelve founding members[5] in 2004 to fifty today.[6] Web and social media archiving

1. National Digital Stewardship Alliance Content Working Group, "Web Archiving Survey Report," National Digital Stewardship Alliance, June 27, 2012, http://www.digitalpreservation.gov/documents/ndsa_web_archiving_survey_report_2012.pdf#page=2, https://web.archive.org/web/20170107233005/www.digitalpreservation.gov//documents/ndsa_web_archiving_survey_report_2012.pdf (accessed February 12, 2017).

2. Nicholas Taylor, "2015 NDSA Web Archiving Survey Report Highlights," SlideShare, August 9, 2016, https://www.slideshare.net/nullhandle/2015-ndsa-web-archiving-survey-report-highlights/3, https://web.archive.org/web/20170215045554/www.slideshare.net/nullhandle/2015-ndsa-web-archiving-survey-report-highlights/3 (accessed February 15, 2017).

3. Jefferson Bailey et al., "Web Archiving in the United States: a 2016 Survey," National Digital Stewardship Alliance, 2017, http://ndsa.org/documents/WebArchivingintheUnitedStates_A2016Survey.pdf (accessed March, 15, 2017).

4. Society of American Archivists Web Archiving Roundtable Leaders, "Society of American Archivists," Society of American Archivists, http://saa.archivists.org/4DCGI/committees/Roles.html?Action=Show_Comm_Roles&CommCode=SAA**TBL-WEBRT&Name=Officers&Status=Active& (accessed February 12, 2017).

5. International Internet Preservation Consortium, "iipc – about – members," last modified May 2004, http://web.archive.org/web/20040603043437/netpreserve.org/about/members.php (accessed February 12, 2017).

6. International Internet Preservation Consortium, "Members | IIPC," http://netpreserve.org/about-us/members, https://web.archive.org/web/20161002195443/netpreserve.org/about-us/members (accessed February 12, 2017).

are the subject of major grants funded by the Andrew W. Mellon Foundation,[7] the Institute of Museum and Library Services,[8] the Knight Foundation,[9] the National Historical Publications and Records Commission,[10] and others. Initiatives such as ArchiveTeam,[11] Documenting the Now,[12] the End-of-Term Web Archive,[13] and other grant-funded and spontaneously-organized collaborations engage in targeted collection of Web content, within a diversity of community, organizational, policy, and technical frameworks. The Internet Archive's Wayback Machine index recently surpassed 510 billion Web objects.[14] While the Internet Archive Wayback Machine was once the only Web archive repository, the number of natively Memento-

7. Andrew W. Mellon Foundation, "Advanced Search | The Andrew W. Mellon Foundation," https://mellon.org/grants/grants-database/advanced-search/?amount-low=&amount-high=&year-start=&year-end=&city=&state=&country=&q=web+archiving&per_page=25, https://web.archive.org/web/20170215050042/mellon.org/grants/grants-database/advanced-search/?amount-low=&amount-high=&year-start=&year-end=&city=&state=&country=&q=web+archiving&per_page=25 (accessed February 15, 2017).

8. Institute of Museum and Library Services, "Awarded Grants Search | Institute of Museum and Library Services," https://www.imls.gov/grants/awarded-grants?field_program[]=57&field_institution=&field_city=&field_state=All&field_fiscal_year_text=&field_recipient_type=All&search_api_views_fulltext=%22web+archiving%22&search_api_log_number=&sort_by=field_program, https://web.archive.org/web/20170215050133/www.imls.gov/grants/awarded-grants?field_program[]=57&field_institution=&field_city=&field_state=All&field_fiscal_year_text=&field_recipient_type=All&search_api_views_fulltext=%22web+archiving%22&search_api_log_number=&sort_by=field_program (accessed February 15, 2017).

9. Knight Foundation, "Grants – Knight Foundation," http://www.knightfoundation.org/grants?text=web%2520archiving&content_sources=grant&page=1 (accessed February 12, 2017).

10. National Historical Publications & Records Commission, "District of Columbia | National Archives," last modified December 9, 2016, https://www.archives.gov/nhprc/projects/states-territories/dc.html, https://web.archive.org/web/20160322232151/www.archives.gov/nhprc/projects/states-territories/dc.html (accessed February 12, 2017).

11. ArchiveTeam, "Archiveteam," last modified June 28, 2015, http://www.archiveteam.org/index.php?title=Main_Page&oldid=23630, https://web.archive.org/web/20161206202237/www.archiveteam.org/index.php?title=Main_Page&oldid=23630 (accessed February 12, 2017).

12. Documenting the Now project team, "Documenting the Now," last modified January 20, 2017, http://www.docnow.io/, https://web.archive.org/web/20170126180841/www.docnow.io/ (accessed February 12, 2017).

13. End of Term Web Archive project team, "2016 : End of Term Web Archive," last modified December 15, 2016, http://eotarchive.cdlib.org/2016.html, https://web.archive.org/web/20170210154624/eotarchive.cdlib.org/2016.html (accessed February 12, 2017).

14. Vinay Goel, "Defining Web pages, Web sites and Web captures | Internet Archive Blogs," Internet Archive Blogs, October 23, 2016, https://blog.archive.org/2016/10/23/defining-web-pages-web-sites-and-web-captures/, https://web.archive.org/web/20170204101120/blog.archive.org/2016/10/23/defining-web-pages-web-sites-and-web-captures/ (accessed February 12, 2017).

compatible archives included in the Memento Aggregator alone is now up to fifteen.[15] The tremendous growth in Web archiving that these trends evince reflects, though no doubt trails, the growth of the Web itself, both in terms of its primacy and size.

There has never been as much of a need for Web archiving as there is now. The many, comparatively benign reasons for link rot have been lately compounded by more conspicuous, deliberate efforts to erase Web information,[16] which have provoked anxiety about the prospect of more ambitious deletions.[17] A growing quantity of materials that have typically been of concern for preservation—ephemera, government information, gray literature, long-tail scholarly communications—only appear online. As social media has democratized communication, social media archiving provides a means to bolster the representation of marginalized groups within the archive[18, 19] and to preserve a unique record of contemporary social activism.[20] We are increasingly successful at connecting researchers with Web archive data to support novel research in the humanities and social sciences.[21] Cultural

15. Memento project team, "Memento Depot," last modified September 22, 2016, http://mementoweb.org/depot/, http://mementoarchive.lanl.gov/twa/memento/20160920154801/mementoweb.org/depot/ (accessed February 12, 2017).

16. Coral Davenport, "With Trump in Charge, Climate Change References Purged From Website," *New York Times*, January 20, 2017, https://www.nytimes.com/2017/01/20/us/politics/trump-white-house-website.html, https://web.archive.org/web/20170209090543/www.nytimes.com/2017/01/20/us/politics/trump-white-house-website.html (accessed February 12, 2017).

17. Brady Dennis, "Scientists are frantically copying U.S. climate data, fearing it might vanish under Trump," *Washington Post*, December 13, 2016, https://www.washingtonpost.com/news/energy-environment/wp/2016/12/13/scientists-are-frantically-copying-u-s-climate-data-fearing-it-might-vanish-under-trump/, https://archive.fo/20161213174451/https://www.washingtonpost.com/news/energy-environment/wp/2016/12/13/scientists-are-frantically-copying-u-s-climate-data-fearing-it-might-vanish-under-trump/ (accessed February 12, 2017).

18. Bergis Jules, "Confronting Our Failure of Care Around the Legacies of Marginalized People in the Archives," On Archivy, November 11, 2016, https://medium.com/on-archivy/confronting-our-failure-of-care-around-the-legacies-of-marginalized-people-in-the-archives-dc4180397280, https://web.archive.org/web/20170215050802/medium.com/on-archivy/confronting-our-failure-of-care-around-the-legacies-of-marginalized-people-in-the-archives-dc4180397280 (accessed February 12, 2017).

19. Kate Theimer, "Gaps in the Past and Gaps in the Future: Archival Silences and Social Media - #acaubc2016 talk," ArchivesNext, February 17, 2016, http://archivesnext.com/?p=4018, https://web.archive.org/web/20161202053904/http://archivesnext.com/?p=4018 (accessed February 12, 2017).

20. Bergis Jules, "Documenting the Now: #Ferguson in the Archives," On Archivy, April 8, 2015, https://medium.com/on-archivy/documenting-the-now-ferguson-in-the-archives-adcdbe1d5788, https://web.archive.org/web/20160208172807/medium.com/on-archivy/documenting-the-now-ferguson-in-the-archives-adcdbe1d5788 (accessed February 13, 2017).

21. Emily Reynolds, "Web Archiving Use Cases," International Internet Preservation Consortium, March 7, 2013, http://netpreserve.org/sites/default/files/resources/UseCases_Final_1.pdf, http://webarchive.loc.gov/all/20160204051235/http://netpreserve.org/sites/default/files/resources/UseCases_Final_1.pdf (accessed February 12, 2017).

heritage institutions continue to embrace digital technologies, and a concern with Web preservation is a logical extension of their historical work.

With the growth of Web archiving has come a broadening of participation and a maturation of practice. Internet Archive's Archive-It service and the California Digital Library Web Archiving Service have done much to lower the barrier to entry, by making such an infrastructure-intensive endeavor practical for an institutional investment of even fractional staff time. Archive-It now supports Web archiving by more than 400 organizations.[22] The expanded Web archiving community of practice has predictably translated to expanded community attention to Web archiving practice. While we have not transcended the foundational preoccupation with the efficacy of our technical methods for Web capture, preservation, and replay, there is more, fruitful contemplation of other dimensions of the work: access, collaboration, ethics, metadata, policy, research use, staffing, and so on. This is reflected by the articles in this Special Issue.

A sign of this advancement is surely that an institution new to Web archiving can make major progress in preserving collections of Web content that matter to them with only part-time staff effort over the course of a single year. In "Case Study: Washington and Lee's First Year Using Archive-It," Alston Cobourn documents the application of Archive-It to the archiving of interactive Web-based scholarship produced by students, and other institutional Web content at Washington and Lee University. This includes an accounting of their decision-making on common aspects of a Web archiving program, such as validation of collecting scope and determination of notification and opt-out policies, as well as a detailed review of Archive-It capture configuration and performance, descriptive metadata creation, and access and discovery strategies. While the 3D visualizations, maps, and timelines in many of the works proved challenging to archive, they were nonetheless pleased with how much they were able to achieve and optimistic about continued programmatic progress going forward. Key takeaways are that Web archiving is iterative, demands experimentation, and benefits from ongoing attention.

These characteristics are as true for institutions long engaged in Web archiving as for those just starting out. In "Using RSS to Improve Web Harvest Results for News Websites," Gina Jones and Michael Neubert share their problem-solving for better archiving of large and frequently-updated news websites by the Library of Congress. The effort was motivated by consideration of use cases articulated by the Web archiving community, the affordances of platforms serving digitized newspaper content (e.g., Chronicling America), and consultation with reference and collection development staff. It is heartening to hear that aspects of the work have become routine, even if they concede that the RSS-driven capture approach is still a work-in-

22. Internet Archive, "Archive-It: A Web Archiving Service," last modified May 30, 2016, https://archive-it.org/blog/files/2016/06/Archive-It_Brochure-digital.pdf, https://web.archive.org/web/20170209195926/archive-it.org/blog/files/2016/06/Archive-It_Brochure-digital.pdf (accessed February 12, 2017).

progress. The detailed explanation of their methodical crawl engineering and contemplation of access implications may be insightful for Web archiving institutions operating at many different levels of scale. Also vital is their observation of the inadequacy of strictly browse-based access to the Library of Congress Web archives—an archived RSS feed is a sub-optimal entry point for this kind of exploration, and users of Web archives would in any case benefit from more sophisticated exploration tools such as full-text search and ngram visualizations.

Other institutions are focusing no less systematically on strategies for effective collaboration around Web archiving. In "Collaboration Made It Happen! The Kansas Archive-It Consortium," Cliff Hight, Ashley Todd-Diaz, Rebecca Schulte, and Michael Church unpack the history, mechanics, rationale, and trade-offs of a consortial approach by Kansas institutions to curate Web archive collections. Selective Web archiving is a natural fit for collaboration, given that materials captured by any one institution via Archive-It can be made world-accessible; co-curated collections are likely to have a more diverse and thorough representation of resources; and individual institutions typically do not dedicate more than fractional staff time to it. In this context, it is great to be able to understand in detail what makes this collaboration successful, as a possible template for others. The benefits they highlight include cost savings, shared best practices, shared collection development, and strengthening professional relationships. A collaboration of this scale benefits from co-determined financial setup, goals, governance, meeting logistics, and overall administration, as well as agreement on access policy, collecting policy, and metadata approach. Challenges foremost include the time commitment, but also varying levels of resources and proficiency with Web archiving.

Collaboration is taking place at broader scales, as well, with potential impact across the field. In "Developing the Web Archiving Metadata Best Practices to Meet User Needs," Jackie Dooley, Karen Stoll Farrell, Tammi Kim, and Jessica Venlet report on the progress of a community-wide effort to shape metadata best practices for Web archives. This is a work-in-progress update in advance of releasing the project's planned deliverables, including a review of Web archiving tools with metadata-related features, a report on user needs and behaviors, and the metadata guidelines themselves. As such, it serves more to highlight issues and questions than to offer recommendations. They observe a couple of high-level challenges with respect to Web archive metadata: the mixed application across the field of archival versus bibliographic approaches to description and the disjuncture between the perceived importance of metadata and levels of Web archiving staffing. Of particular value is an elaboration of the ambiguities of particular descriptive metadata fields (e.g., creator, publisher, extent, and provenance) as applied to archived Web content. Working through these ambiguities, informed by strong notions of what different types of users may want or expect, seems highly valuable.

These articles together demonstrate some of the diverse and exemplary work taking place throughout the Web archiving community—programmatic bootstrapping, capture engineering, nurturing collaboration, and harmonizing

community practices. The relentless evolution of the Web both complicates and inspires Web archiving, but ensuring that our work remains ethical, strategic, sustainable, and useful will take much more than an attention to technology. In this context, I am grateful to be part of an expanding, diversifying, and ever more thoughtful community of practice. I commend the authors, reviewers, and Editorial Board of the *Journal of Western Archives* for their high-quality contributions to this Special Issue and the collective bibliography of Web archiving resources.