Assistant Editor: Sarah Dorpinghaus, University of Kentucky. Contact Sarah at sarah.dorpinghaus@uky.edu if you would like to guest author a column or have a good idea to share.

Programmatic Challenges of Web Archiving: A Graduate Student Perspective

Adapted from a guest blog series for the SAA Web Archives Section

By Grace Moran, University of Illinois

There is a small window for preserving websites; finding a way to preserve them and ensure accessibility for end-users is one of the greatest challenges facing recordkeeping institutions today. This article details my efforts over this past academic year to develop, standardize, and envision a future for the University of Illinois web archiving program.

To start, a bit about myself. I graduated on May 15 from the iSchool at the University of Illinois with my MS in library and information sciences. I have been working for my supervisor, Dr. Chris Prom (associate dean, Office of Digital Strategies) since January 2019. Last year, he gave me the opportunity to transition from working on digital special collections and workflows to focusing on our young web-archiving program. Taking this opportunity led to an incredible learning experience that I want to share widely.

The University of Illinois has been an Archive-It partner since 2015. Prior to this, the library was running ad hoc event-specific web crawls with other technology and capturing content through a contract with the California Digital Library prior to 2015. The program has been inconsistent in its stewardship, with varying levels of support and staffing. This year, I was charged with evaluating the status of the University of Illinois subscription to Archive-It and the creation of a plan for its continued success. The culmination of the role was a final report outlining my accomplishments for the year, an evaluation of programmatic needs for the web-archiving program, and recommendations for the future of web archiving at the University of Illinois. The challenges facing a burgeoning web-archiving program are numerous and varied, ranging from policy to accessibility. With over 5 TB of data archived since 2015, the care and keeping of these collections is of utmost importance; the time invested by the University of Illinois Library must not be wasted.

My experience this past year made it clear that managing a web archive is no small feat. (Is this an obvious statement? Yes, but that doesn't make it any less true.) Two of the most urgent issues I have identified are metadata and access. The debate around metadata best practices for web archiving is not new. As it stands now, we don't have a standard for documenting archived websites. The OCLC Research Library Partnership Web Archiving Working Group has dedicated a tremendous amount of time to this issue (see Dooley, Farrell, Kim, and Venlet, 2018). Archival and bibliographic metadata standards don't fully account for the unique nature of websites, so some institutions have adopted a hybrid approach. The question remains whether or not a hybrid approach is sufficient. Is it necessary to create a new standard? Would this over complicate things?

What sort of description does Archive-It crawling technology support? The Archive-It platform has three possible levels of metadata:

1. Collection-level
2. Seed-level
3. Document-level

Our collections at the University of Illinois tend to have collection-level and seed-level metadata. At the minimum, the goal is to have collection-level metadata. Depending on the number of seeds (saved websites) in a collection, it is also common to implement seed-level metadata for an additional layer of description. The inconsistency of metadata below the collection level is due to a few factors. First, the labor of editing seed metadata takes enough hours without also creating document metadata—the law of diminishing returns. Second, what is/isn't a document on Archive-It remains unclear, prompting the question of what benefit there is in creating document-level metadata. The most urgent metadata need not only for the University of Illinois but for other collecting institutions is to create a consistent policy for description.

Deeply tied to the issue of metadata is that of access. How do we make sure that content reaches end-users? If a tree falls in the forest . . . no I'm not going to finish that; too cliché. You get the idea. My point is this: taking the time to preserve digital materials is nearly meaningless if no one gets to enjoy the fruits of the labor. Currently, very few access points to University of Illinois web collections exist. One resides on Archon (our local archives online catalog) and leads directly to a lower-level collection on Archive-It: "University of Illinois at Urbana-Champaign Web Archives." Furthermore, the International and Area Studies Library, thanks to graduate assistant Nathan Sonnenschein, now links out to its collections in Archive-It. Otherwise, users must navigate to the public-facing Archive-It website and either search "University of Illinois" or stumble across one of our collections as a result of a search. We are not in any way alone in this; institutions all over are struggling with the idea of making collections searchable.

So, what's the solution? How do we implement an access system that doesn't overburden end-users? In my final report to the library, I recommended both short- and long-term solutions. Short-term, Archive-It allows subscribers to add a search bar to their own local discovery systems—with only a few lines of html—creating another method of discovery for researchers. This remedies the disconnect between local discovery systems and the Archive-It discovery system. This simple, elegant, and quick solution would make a significant difference for researchers. A long-term, more involved solution would be to index all of our pages locally and work to allow full-text search of collections. In the era of Google, our users are used to full-text search that does much of the heavy lifting for them.

Policy and personnel are also central to web-archiving programs. Bragg and Hanna argue that policy affects every layer of a web-archiving program (Bragg and Hanna, 2013, 3). Based on my experience this past year, I believe institutional web-archiving policies must include the following among other things:

- A collection development policy unique to the institution's web-archiving program (a general organization-wide collection development policy does not suffice given the unique nature of the content being collected)
- A clear, centralized workflow outlining how crawls are to be run, troubleshooting documentation, and chain-of-command for web archiving
- A statement on copyright and ethics in web archiving (Niu, 2012)

(Continued on page 20)

(Continued from page 19)

Though it may be extremely obvious to some readers, it is worth saying: policy should be public. As someone who has worked for a public university, I am painfully aware of the importance of policy accessibility for our stakeholders. Furthermore, Belovari observes that this sort of information is also useful to researchers (Belovari, 2017).

What about personnel? Who should run a web-archiving program? How many people should be involved? Of course, this varies from institution to institution; however, my experience has made clear the need for a dedicated point person. This job could be

- A graduate web-archiving position, like my own, for 20 hours a week to coordinate crawls across units, run quality assurance, and populate metadata fields; or
- A civil service or academic professional position with at least a 50 percent appointment to web archiving. If an institution is looking to grow its web-archiving program, it should consider making this a 100 percent appointment for the first couple of years and then slowly transition the point person into additional activities related to digital strategies of the library.

If you are just beginning your program, you may find that a part-time position does not fulfill its needs. Additionally, long-term employees have the advantage of institutional knowledge and memory and therefore understand the administrative history of digital programs within the organization. Time is lost when retraining someone for a position annually or biannually is necessary.

Like any burgeoning field, web archiving presents a unique set of challenges. Investment of both capital and personnel now will save collecting institutions from later regrets when domains have gone extinct and the historical record is deprecated.

References

Belovari, S. 2017. “Historians and Web Archives.” *Archivaria* 83, no. 1: 59–79.

Bragg, M., and K. Hanna. 2013. *The Web Archiving Life Cycle Model*. Archive-It. https://archive-it.org/static/files/archiveit_life_cycle_ model.pdf

Dooley, J. M., and K. Bowers. 2018. Descriptive Metadata for Web Archiving: *Recommendations of the OCLC Research Library Partnership Web Archiving Metadata Working Group*. OCLC Research.

Dooley, J. M., K. S. Farrell, T. Kim, and J. Venlet. 2018. *Descriptive Metadata for Web Archiving: Literature Review of User Needs.* OCLC Research.

Dooley, J. M., K. S. Farrell, T. Kim, and J. Venlet. 2017. “Developing Web Archiving Metadata Best Practices to Meet User Needs.” *Journal of Western Archives* 8, no. 2: 5.

Dooley, J., and M. Samouelian. 2018. *Descriptive Metadata for Web Archiving: Review of Harvesting Tools*. OCLC Research.

Niu, Jinfang. 2012. “An Overview of Web Archiving.” *School of Information Faculty Publications*, 308. http://scholarcommons.usf.edu/si_facpub/308.