Current Events in Africa Web Archive (CEAWA) – a Title VI Librarians Pilot Project

https://archive-it.org/collections/4426

Final Report

Editors: Marion Frank-Wilson, Karen Fung, Tim Johnson, Lauris Olson, Jason Schultz Mohamed el Seoud

November 2016

Current Events in Africa Web Archive (CEAWA) is a three-year pilot project (2013-2016) to archive websites (including blogs, twitter, podcasts, etc.) related to current events in Africa, funded by the Title VI Librarians of the ALC. This report is submitted to the Title VI librarians, with a recommendation to consider any future, larger projects within the entire framework of Title VI librarians, ALC and CAMP, to allow us to take full advantage of financial resources and the combined expertise of all members.

Development of the Project

- The proposal was approved by the Title VI Librarians at the Berkeley spring meeting. MFW had offered to administer the project through the Indiana University Libraries, and she worked with CRL to transfer the allocated budget to IU ($2000 for the Archive-It subscription; $3000 for a graduate student assistant).
- MFW hired a graduate assistant (Sarah Keil) in the summer 2013. With the help of IU's Collection Development Office, Marion and Sarah participated in training sessions and Archive-It tutorials.
- Sarah created an initial workflow (spreadsheets, timelines), based on IU's experience with other Archive-It projects. She also did some research into legal implications in case website owners do not respond to requests for permission. At the fall 2013 Title VI Librarians meeting, there was agreement to go ahead and archive websites if after 3 permissions requests no reply has been received. Sarah documented those initial steps so that future student assistants would be able to step into the workflow easily.
- We set up the CEAWA collection in Archive-It (fall 2013). In the process, we learned that Archive-It required a minimum subscription level of $3000 but offered to contribute $1000 for the duration of our pilot ($1000 per year).
- MFW created a permissions request letter which, with the help of Atoma Batoma, was translated into French.
- A number of websites (initially 16; a 17th site was added later) were identified by the editorial team, based on individual knowledge of the sites, consultation with colleagues, and soliciting input on the ALC's listserv.
- Sarah handled the initial permissions request process and set up test crawls but, in the meantime, graduated and accepted a position as a librarian at another institution.
- Early in the spring 2014, MFW hired a new graduate student, Paula Mate, a Ph.D. candidate in the school for Informatics (focus on Africa/Mozambique). Paula had worked as a systems

librarian before and has experience and expertise in working with implementing electronic projects.

- Paula, after going through Archive-It training, began reviewing test crawls and then initiating actual crawls of the initial 16 (later 17) sites identified in the proposal.
- Initially, our goal was to crawl each of the websites once (which included trouble-shooting and re-crawling parts of certain sites that were not captured during the first crawl).
- Once we had completed the first round of crawls (summer 2014), we reviewed all websites and adjusted crawl frequencies. For example, one website is that of a daily newspaper (*Le Republicain*), and over the course of 2 weeks, we scheduled daily crawls to determine the impact on the budget and, ultimately, learn if web-archiving could be a way to archive/preserve newspapers (we found that it is not; daily crawls take up too much of the data budget and online versions of newspapers are not always reliably regular). We determined other crawl schedules for the other websites, each one based on how often the websites appear to be updated. Determining the frequency of crawls was important, since we wanted to capture timely events, and we also wanted to avoid capturing the same content over and over again (which would have resulted in using up our budget unnecessarily).
- Once we had set up the sites to be crawled, we began crawling them one by one. In doing so, we assured that content was fully captured. As this stage, we did quality checks on existing (life) content, on the archived site as well to ensure that all content was captured and displayed properly.
- The sites we had selected to archive were a combination of websites and blogs comprised of images, audio and video recordings with complex objects and texts. Due to a combination of multiple file formats, capture of certain websites was difficult but has been accomplished. By the end of the pilot, the original 16, plus an additional site documenting the Ebola crisis in Liberia (added later) are preserved.

Challenges:

Implementing a web archiving project is a complex undertaking. Challenges range from hard to capture to blocked content and delays in acquiring permission to capture from website creators. Specific challenges encountered:

- Communication with web/blog creators proved to be difficult and we often did not receive replies to requests for permission to crawl.
- Some websites were difficult to capture because of their format and/or design. Moreover, we experienced difficulties communicating with webmasters/owners about these problems which resulted in delays in capturing content. Some content (mostly images and videos) needed to be unblocked before the Archive-It crawler could grab it. We had the ability to capture blocked content – in those instances, we asked for permission before we proceeded with the capture of such content. In cases where we did not hear back from webmasters/owners, we nevertheless chose to capture, using a special Archive-It tool.
- The most challenging aspect of this project has proved to be the weakness of “test crawls”, a feature provided by Archive-It. “Test crawls” allow for a test (or several) on seeds (i.e., URLs). No data is collected during a test, so a test crawl does not use up any of the crawl budget.

However, test crawls have proved to be 'useless' as they don't tell the whole story. We had many successful test crawls and, subsequently, unsuccessful real crawls of the same seed. The majority of seeds with images and video files needed to be patch crawled several times in order to gather all desired content (trouble shooting/monitoring was needed after each crawl to determine how successful a crawl was, i.e., how much of a website was captured and what was missing).

Test crawls do not display a quality assurance (QA) report nor the live page view, which limited our ability to see what the test crawl will actually look like on the live site. The QA report shows why certain content on a page may not be captured or viewable. Note that the QA reports are not automatically generated for each crawl. We had to manually start the QA process for each seed, after which the report is generated within 24 hours. After the report was generated, we were able to select the seed URL we wanted to QA.

Some websites' content is not 'crawlable" and so far Archive-It has not provided a solution to capture difficult content.

- Time constraints – Archive-It does not allow a preview of captured content immediately after capture. We had to wait at least 24 hours to view the results of each crawl. In essence, there is no preview – there is only the view of the final product. This presents troubleshooting and communication delays.

Accomplishments:

- In addition to preserving web resources of interest to this pilot project, we built a collection of relevant, current events sources valuable to present and future research.
- We have learned the logistics involved in a web archiving project, and how to develop an efficient workflow to manage a web archiving project.
- We now have an idea of the amount of time, trouble-shooting, etc. involved in managing such a project in the long-term.
- We have learned how to archive different kinds of websites, and how to deal with blocked content.
- We have gained experience with the selection of websites, on a small scale, by funneling website suggestions to the editorial board.
- We have gained insights into the cost of such projects, which may be helpful in exploring sustainable funding models.

Questions and Considerations:

- Knowledge in Africa is being produced online – related to current, cultural political etc. events and topics. Knowledge disseminated this way is ephemeral – in fact, during our pilot, one the websites we archived disappeared. It is important to preserve this knowledge before it disappears. For area/African studies librarians, this ought to be a part of our collection development activities.

- There is currently a lack of best practices, coordination, both at the national and international levels.
- Some thoughts to consider:
  Archiving, whether it is of print or web resources, is selective. "Archives are the product of a process which converts a certain number of documents into items judged to be worthy of preserving and keeping in a public place, where they can be consulted according to well-established procedures and regulations…The archive is, therefore, fundamentally a matter of discrimination and of selection, which, in the end, results in the granting of a privileged status to certain written documents, and the refusal of that same status to others, thereby judged 'unarchivable'." (Achille Mbembe, "The Power of the Archive and its Limits," in Hamilton, Carolyn, et al. (eds), *Refiguring the Archive*. Cape Town: David Philip, 2002, p. 19.
  With regard to webarchiving, usually the librarian decides what is to be archived, and what is preserved for the future. For example, for CEAWA, we (Title VI librarians) decided on the topic for the archive, how often a website is crawled, what version will be archived, how it will be described and displayed, etc.
- Steps in the selection process:
  What is selected for inclusion in the archive? Who selects, and based on what criteria? Who develops these criteria?
- Once websites have been selected, they are interpreted (by assigning metadata). The metadata determine how the archive/archived websites can be discovered; and by whom.
- Who are the researchers we (i.e., librarians) have in mind for the archives we create, and how does that influence the archives and the data we make available?
- Does our selection, metadata, and web design process determine what kind of research can be done with the archive?
- Ethical considerations – Postcolonial (web) Archiving:
    - Knowledge in many world areas, including Africa, is produced on the web (some speak of a democratization of knowledge production see Dan Hazen, "Lost in the Cloud").
    - The need to archive this knowledge versus arguments of cultural imperialism.
    - Digital heritage is a form of national heritage
    - Who decides what gets archived, who has the power to select websites, and to determine a country's online history?
    - With regard to print and migrated archives, this debate has existed for quite some time in the library and archives world, but in connection with web archiving, it has become more intense, part. concerning the role of the librarian. Whereas before, librarians and archivists have collected materials/archives from other world areas, web archiving changed the collecting role to that of creator.
    - Access: who can access the web archives we create (considering bandwidth issues, user/researcher training – assuming users are able to access the archive, are they aware of their selective nature, omissions, etc.?)
- Need for:
    - Collaboration, at the national, and certainly at the international level
    - Ethical guidelines (perhaps ASA guidelines suffice?)
    - International agreements and laws.

Going forward – what does this mean for the Title VI Librarians, and the ALC?

Web archiving is more than a technology – it is a form of collection development and stewardship. In light of some of the ethical consideration outlined above, it is a form of collection development that would benefit from a collaborative approach, both nationally and internationally. A body such as the ALC is in a good position to take on this responsibility. Based on what we have learned from our pilot project, several possibilities come to mind:

- Increasingly, academic libraries now regard web archiving as an element of collection development. It is as yet unclear what this means in reality. For example, how will such archiving activities be incorporated into their collection development, how will access be provided; will the archive live on the Archive-It server, or in some way have a record in the library's online catalog? What kinds of metadata will be provided, etc. Very few (if any) best practices or policies exist. And yet, this kind of web archiving is taking place (example: a South Asian archive was taken off the web for political reasons; a U.S. librarian stepped in at short notice to rescue the websites and may collaborate with the Indian library at a later point, when the political climate is different; for now, the websites have been preserved).
  Potential role for ALC: serve as clearing house for such initiatives. ALC members could agree to disclose their web archiving initiatives for African websites which could be part of a dynamic website on the ALC website.
- Another possibility is to build on what we already have, i.e., CEAWA. CEAWA could move from pilot to a regular, collaborative ALC project to archive websites in reaction to current events in Africa. This would require ongoing funding and an active editorial board to oversee the selection of websites (ideally in collaboration with colleagues on the continent) as well as a host institution (similar to IU during the pilot). It might be worth exploring whether CRL could play a role in such a project.
- Or create a portal for African websites, in collaboration with a body such as the ASA and possibly CRL. This would require a grant. I believe the effort, if presented within the framework of developing best practices and international collaboration, is innovative and might have a chance with funding agencies. It would have to be based on partnerships with African colleagues/librarians/archivists. Again, the ALC would be very well suited for such a project and could leverage the contacts we already have, and also work through ASA. This could be a very worthwhile project, and at the same time consuming of time, resources, and librarians. It would serve the purposes of preservation, and also of providing access.