**Current Events in Africa Web Archive – a Title VI Librarians Pilot Project**

**Current Events in Africa Web Archive**

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

November 2014

The proposal for this three-year pilot project was to archive websites (including blogs, twitter, podcasts, etc.) related to current events in Africa. The original proposal's title was for "African Countries in Conflict", a title which later (based on feedback from the Title VI Librarians group) was changed to "Current Events in Africa Web Archive (CEAWA)". Since we are now about half-way through the pilot phase of the project, this report is intended as a status report to summarize accomplishments, problems encountered, and steps for the remaining one and a half years.

**Accomplishments:**

- The proposal was approved by the Title VI Librarians at the Berkeley spring meeting. Frank-Wilson 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).
- F-W 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).
- Frank-Wilson created a permissions request letter which, with the help of Atoma Batoma was translated into French.
- 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, Marion hired a new graduate student, Paula Mate, a Ph.D. candidate in the School for Informatics. 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 16 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 determined other crawl schedules for the other websites, each one based on how often the websites appear to be updated.
- The sites we decided to archive are 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 has been difficult. At this point, all sites have been successfully captured (see for more detail about each site below). We are currently preserving content of sixteen websites/blogs in addition to a newly added site on the Ebola crisis and have requested permission to archive two more Ebola-related websites.
- In addition to preserving web resources of interest of this pilot project, we built a collection of relevant, current events sources valuable to present and future research.

**Challenges:**

Implementing a digital archive storage website 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 tends to be difficult. Before capturing the online content, we send an email requesting permissions for capture. We try to contact them 3 times. If we do not receive a response, we send a final email including a note that we will be archiving the site unless we hear otherwise.
- During the capture process, some websites prove to be difficult to capture due to their format and/or design. We need to keep in mind that these are international websites built with standards that may not be accommodating to American standards. Moreover, we experience difficulties communicating with webmasters in a timely manner. These problems present delays in capturing content. Furthermore, as you will see below, some content (mostly images and videos) need to be unblocked before the archive-It crawler can grab it. Fortunately, we have the flexibility to capture blocked content (mostly images/media). We always ask for permission before we proceed with the capture of

blocked content. Yet, if we do not hear back from webmasters, we choose to capture with the help of a special an Archive-It tool.

- Archive-It provides a feature called "test crawl." Test crawls allows us to run a "test" (or several) on our seeds (URLs). No data is collected during a test, so a test crawl does not use up any of our crawl budget. A downside to test crawls is that they only generate reports (they provide a number of files: documents and host that will be captured) but, they do not provide a display of what the captured website will look like. Hence, a test crawl does not tell us whether all content will be captured or not. Therefore, we find test crawls to be in some ways limited. They create the illusion of success when in reality, a successful test crawl does not translate into a successful real crawl - as you will see below. The majority of seeds (URLs) with images and videos needed to be patch crawled a few times.
- Another setback is the time constraint. Archive-It does not allow a preview of captured content immediately after capture. We have to wait at least 24 hours to view the results of each crawl. In essence, there is no preview. You only view the final product. And, in order to view, one has to wait at least 24 hours. This constraint presents troubleshooting and communication delays between all parties involved.

**Websites/blogs being archived:**

- Bridges from Bamako. http://bridgesfrombamako.com/
  This is a blog site, created by Bruce Whitehouse, an anthropologist and former Peace Corps Volunteer. The blog is updated on a monthly basis (on and off). Nevertheless, this is one of the blogs we can rely on in terms of consistency. The blog includes an analysis of the 2012 political crisis in Bamako and it describes changes that have been taking place in Bamako since 2011. In addition, since the Mali crisis began in 2012, Whitehouse reports on some political issues as well. We ran our first test crawl for this blog site in mid-March. Our experience with the test was successful. However, running a real crawl was a different experience. Our first real crawl missed style sheets and images. For this reason, we had to troubleshoot, by scheduling patch crawls and setting up our crawls to ignore robots.txt file. After running a QA and patch crawls we were able to capture all content and make the blog live.

- MaliWatch. http://www.maliwatch.org/
  This is a non-political Non-Governmental Organization (NGO). The website lists non active projects but.
  We decided to capture content of this website quarterly. Our test crawl was successful, and we are happy to report that the real crawl was as well. We believe that this success is due to the content of the site. This site's content is heavy in pdf files and text files. Therefore, we did not have to capture many images and video content.

**Central African Republic**

- http://www.radiondekeluka.org/
  We experienced technical difficulties in capturing content of this site. Specifically CSS. The first real crawl did not grab all website content, specifically images. Therefore, we ran a patch and a QA report. After these problems were resolved, we had successful captures. This is one of the websites that we capture on a daily basis.

- http://www.centrafrique-presse.info/site/
  Centrafrique presse's primary goal is to inform the public on real time issues regarding policy, economy, human rights and social issues related to Central Africa.
  This is an example of a site for which we experienced ongoing capture issues on and off. It has become somewhat difficult to predict the outcome of captures, without close monitoring. Besides, the site does not have regular updates. For this reason, we decided to capture only three times since May, 2014. We still have some images that cannot be captured. Archive-It is working on finding a solution to this problem.

- http://www.centrafriquelibre.info/
  Centrafrique Libre is a Central African independent newspaper created by many correspondents and world journalists with the aim of contributing to the freedom of press in Central Africa. We decided to crawl this website on a weekly basis. Test crawls ran successfully however, real crawls were not as successful. After troubleshooting a few times, we successfully gathered all content and we are currently archiving this newspaper site.

- http://centrafrique-presse.over-blog.com/
  Centrafrique Presse's blog was created with the main focus of informing the public on humanitarian aids and economic development issues in Central African Republic (CAR) and around the world. In addition, Centrafique-press promotes democratic governance and economic good for the people of CAR. This blog is updated on a daily basis (on and off) thus, we decided to crawl it daily. Test crawl was successful however, real crawl proved to be difficult as there was hard to capture content. After running a few patch crawls and ignoring robots.txt file, we are now able to capture all content.

- http://reseaudesjournalistesrca.wordpress.com/
  This is a blog administered by the Network for Journalists for Human Rights in Central African Republic (CAR-RJDH). RJDH is an organization consisting of journalists with a goal of defending and supporting human rights and humanitarian organizations in Central African Republic. This blog is updated on a weekly basis. Test crawl along with real crawl were successful. We continue to capture and store content for the blog.

**DRC (Democratic Republic of Congo):**

- Media Congo:
  http://www.mediacongo.net
  Media Congo is a news website which provides news from Congo and its diaspora. We decided to capture content for this site on a weekly basis. Test crawl was successful though, real crawls were not so successful. Since this website is rich in images and media, it proved to be difficult to capture all content without extensive troubleshoot and communication with Archive-It team. Design and layout of the website were part of the problem. Eventually, we captured all necessary content and we continue to do so.

- Infobascongo:
  http://www.infobascongo.net
  InfobasCongo is a non-profit organization comprised of computer and media professionals aiming to convey social-economic and political information about the DRC. This website is updated monthly, hence we crawl its content once a month. Our test crawls were successful however, the real crawls were not. Most images and video content required troubleshooting and working with the Archive-It team for a few weeks until we were able to successfully gather all content.

- Kleber - L'Observateur:
  http://kleber-lobservateur.blogspot.com
  Kleber-L'observateur is a blog that highlights political, economic and social issues related to the African region while paying special attention to the Democratic Republic of Congo. The blogger focuses on regional political issues related to foreign policy.
  We experienced similar problems as with the other sites. After a real crawl, we realized that some content was missing therefore, we needed to run a few patch crawls. After a patch crawl and QA, the website was successfully crawled. This blog is crawled on a monthly basis.

- Congo Siasa:
  http://congosiasa.blogspot.com
  This blog looks to bring to light the trials and tribulations of Congo and its people. The blogger focuses on the conflicts that occur in this area of the world as well as the tensions that exist between the government and the militia. These conflicts are also looked at in light of the interactions between the Congo and other countries around the world.
  Real crawl was unsuccessful. After troubleshooting a few times, we were able to capture all content. This blog is crawled on a monthly basis.

- Societe Civile.cd:
  http://www.societecivile.cd
  La Société Civile is a portal open to the world. It provides peace and democracy, human rights and elections news of the Democratic Republic of Congo (DRC). In addition, the portal offers regional resources from the DRC provinces and information related to health, education, business and trade, agriculture and livestock, nutrition, and social assistance. This website is

updated monthly. Due to the richness of media content we experienced technical problems with captures. After running a few patch crawls and QAs we succeed.

- Churches, Peacebuilding and Women's Rights in DRC:
  http://www.drcongo.nibrinternational.no
  This blog is part of the ongoing research project Religious Civil Society Networks in the Great Lakes region. The project, funded by the Norwegian Ministry of Foreign Affairs is set to help in peace-building process, Peace and Security and in the implementation of the United Nations Security council Resolution 1325 on Women. The research institutions involved in this project are the Norwegian Institute for Urban and Regional Research, Centre for Intercultural Communication in Stavanger, the Evangelical University in Africa in Bukavu and Universite Officielle de Bukavu. Updates for the blog take place yearly. We have only crawled the site once. Crawling experience mirrors that of the other sites in terms of running a few patch crawls.

**Newly added site:**

- Friends of Liberia
  http://fol.org/
  Friends of Liberia (FoL) is a non-governmental, non-profit organization which supports Liberia through socio-economic, humanitarian and education advocacy programs.
  Crawling this site was similar to the other sites above. The site has a marquee with videos, hence, we had to troubleshoot. After troubleshooting, images and video were captured and preserved. FoL is captured monthly.

**Problem sites:**

- Radio Okapi:
  http://radiookapi.net
  Radio Okapi is part of the United Nations in the DRC. We were able to capture some content of this website. This radio station was established with the main goal of spreading independent national, regional and international news in a rigorous, factual and credible form. Content of the website is updated weekly. After we had captured content for three months unexpectedly, we started experiencing problems with the display of images and video captures. After troubleshooting the problem could still not be resolved. For this reason we have asked the web administrator to unblock some content.

- Le Republicain
  http://lerepublicain-mali.com/
  This website covers Mali's recent topics including social, cultural, technological and environmental issues. In, addition it links its audience to external websites such as the UNICEF. Unfortunately, this newspaper website has been disabled and online access was terminated. Therefore, we stopped capturing its non-existing content. Fortunately, we have two months' worth of content preserved.

- Mouvement National de Libération de l'Azawad, MPLA. http://www.mnlamov.net/
  The National Movement for the Liberation of Azawad (MNLA) blog, was created with the intention of alerting the international organizations for human rights protection to the abuses of human rights by the Malian army on civilians.

We set up this site to be crawled quarterly. However, we experienced technical issues capturing videos from the beginning. The main page of the site is composed of a video marquee which proved to be difficult to capture. There seem to be strange redirects that remove parts of the URL. Archive-It has been informed of the problem since Sept 08. They last reported that engineers are working on finding a fix. For this reason, we decided to omit this website from public view, until the problem is resolved.

**What we have learned so far:**

- Logistics/how to develop an efficient workflow to manage a web archiving project.
- We are beginning to have an idea of the amount of time, trouble-shooting etc. that is 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.

**Budget Implications:**

- With the capture and preservation of the websites/blogs listed above, we have spent 50% of our allocated budget. This means we need to increase the number of websites in order to make the best use of the money we have (unspent data budget/subscription money is not returned).
- We are on track with the student hourly budget. Paula works 10 hours per week, which seems appropriate, given the time issues between crawls mentioned above.

**Next steps:**

- Create a website which will provide contextual information for our project.
- Add more websites to make maximum use of our data budget. We will send a request for additional website suggestions to the ALC and Title VI librarian lists.
- Consider expanding the types/formats of websites to include, e.g. podcasts (video and audio as available), Facebook, and perhaps twitter feeds.
- Find ways to promote the project among faculty, graduate students and researchers.
- About 6 months before the end of the pilot, analyze all our crawls and the budget, as the basis of decisions about expanding the pilot project into a larger (possibly grant funded? Or in collaboration with ASA and/or other partners?) project.

**Glossary**

CSS (Cascadian Style Sheets) - is a style sheet language used for describing the look and formatting of a document written in a markup language (html).

Patch Crawl – crawl designed to capture missing URLs.

QA – Quality report (QA) shows why certain content on a page may not be captured or viewable.

Robots.txt - A robots.txt file is a way for a webmaster to direct a web crawler (aka robot) not to crawl all or specified parts of their website.

Seeds – are starting point of URLs for the crawler. Example: http://fol.org/

Wayback Machine - is a digital archive of the World Wide Web and other information on the Internet created by the Internet Archive. Note: the Wayback Machine crawls at its own intervals and does not make special effort to capture hard to capture content, or lower-level websites. Does not create curatedcollections.

Web Crawler (crawl) - is a program or automated script which browses the World Wide Web in a methodical, automated manner.