Benjamin Peeples, Peter Sutton, and Benjamin Turkus

Kara Van Malssen

Digital Preservation

29 October 2013

Witness: Syrian Activist Video Web Archiving Policy

"WITNESS uses video to open the eyes of the world to human rights violations. WITNESS empowers people to transform personal stories of abuse into powerful tools for justice, promoting public engagement and policy change."

—The Mission of Witness

I. Introduction

Since its inception in 1992, *Witness* has sought to harness the power of rapidly proliferating video production technologies to present human rights violations to the general public. By aligning personal storytelling with video technologies, *Witness* strives to change the world by raising awareness, documenting, and making accessible stories that might otherwise have been lost. The goal of the *Syrian Activist Video Web Archive*, established in October 2013 by Yvonne Ng, Archivist at *Witness*, and three students of New York University's Moving Image Archiving and Preservation Program, was to broaden and expand the scope of *Witness*'s collections. This project was a major undertaking, and will require ongoing efforts to remain successful.

Syrian activist video is produced in a high-stakes, high-pressure media preservation environment. The situation in Syrian is highly volatile, and these websites can become the victims of shut downs, censorship, threats of violence, and general infrastructural/technological failure. "The web is growing steadily, and at the same time it is continually disappearing. Web sites disappear and the site content tends to change rapidly" (*Library of Congress*, "Web Archiving Policy," 2). This statement, written as part of the Library of Congress's 2008 Web Archiving Policy, holds especially true for this collection. Our goal is not to validate the information we are collecting, but rather to obtain it while still available. By taking advantage of the Internet Library's Archive-It program, a variety of Syrian activist video websites have been crawled, with much information captured. Supplementing the work of *Witness* (already a model and authority in human rights video advocacy), this project will help ensure that the daring, courageous work of Syrian video activists will not be forgotten or lost.

However, as we gather Syrian activist video productions, institutionalizing them at the Internet Archive, transforming from "urgently produced works of cultural activism created in and for the collective space of a social movement" to "archival objects," contained in a space with inherent barriers to access, it is

important that we, as Western human rights advocates, approach these audiovisual documents with both care and self-awareness (Hallas 3). Our ultimate goal is to ensure that this activist history never becomes dead history; we must never stop actively attempting to integrate it into the lives of its audience (Cvetkovich 195).

II. Collecting Policies

Witness is an international non-profit organization dedicated to using video and personal storytelling to inform the public of human rights abuses. Yvonne Ng, archivist at *Witness*, compiled a list of websites related to the Syrian Civil War that were in desperate need of archiving. With Yvonne's assistance, it was determined that this project would gather webpages and videos created by citizen activist journalists operating throughout Syria. By taking advantage of the Internet Archive's Archive-It service, this project expanded upon *Witness's* efforts, preserving and making accessible testimony from this war-torn region.

One of the challenges of this project was targeting and selecting websites that illustrated the range of sources available: from the more reputable to the less reputable, in English and in Arabic, it is impossible to offer a complete portrait of this Syrian activist video movement. Hopefully, the websites chosen for this project provide a sense of the breadth of sources available, and will prompt scholars to dig deeper, continuing to search for overlooked activist works. Eight websites were selected for an initial test crawl: the Shaam News Network Twitter Page, the Syrian Network for Human Rights Website, the Idleb Press Facebook Page, the ANA YouTube channel, the Waw Al-Wasel Alternative Youth Media Website, the Aleppo Media Center Facebook page, the World in the Lens Facebook page, and WATAN (Syrian News Website). Throughout, we attempted to ensure that our efforts to preserve this information not unduly impact the activities of these organizations.

The primary goal of this project was to archive and make accessible webpages and videos from non-mainstream newsgathering organizations. Given the dangerous climate in Syria, it is often these very citizen journalists who are providing major news organizations with the information they present to the world. One of the goals and strengths of this collection is to cut out the middleman, presenting so-called "on-the-ground videos" with minimal mediation. However, this is also proved a challenge, as guaranteeing the authenticity of these videos is extremely difficult. While more research is certainly required, we believe that the sources we have collected accurately reflect the activist movement in Syria. And, from a scholarly perspective, even those videos deemed inauthentic representations of Syrian human rights violations remain worthy of preservation, critique, and commentary. Ultimately, while more verification is in order, we believe that we have provided an unfiltered look at the Syrian activist movement.

The web archiving and crawling plan was designed to reflect these overall collection problems, paying special attention to the highly mutable of the Syrian Civil War, and the time-sensitive nature of preserving information from this ongoing conflict. Rather than allow the crawler to document the entire Web, the parameters

of the test and live crawl were made to be as specific as possible. The test was designed to ensure that the crawler was operating as expected, able to crawl the URLs without any restrictions. During the test crawl, we consulted with a representative of Archive-It, and together we addressed data collection issues. This project is still admittedly in its alpha stage; our efforts will expand as we become more technically proficient, and as the Archive-It program makes technical improvements to its social media archiving abilities. Further consultation with Yvonne and Archive-It representatives will be necessary to create the best possible collection of Syrian activist web videos and information.

III. Acquisition Sources—Current and Future

Due to the increased levels of violence, often the work of these intrepid activists is the only source of information available. In fact, many news sources, including The New York Times and CNN receive most of their information from these websites, further solidifying their historical value.

Current seed URLs:

- Shaam News Network Twitter Page
- Syria Network for Human Rights Website
- Idleb Press Facebook Page
- ANA YouTube Channel
- Waw Al-Wasel Alternative Youth Media Website
- Aleppo Media Center Facebook Page
- The World in a Lens Facebook Page
- WATAN (Syrian News Website)

Not every website we are actively crawling is devoted exclusively to activist video; some seed URLs, such as the Shaam News Network Twitter Page, offers updates for news, gatherings, conversations as well as their video postings. This provides welcome context, information, and stories that go beyond videos. *Witness* listed a number of other websites that will be added to the Syrian web archive after the initial crawls are successful. Websites that contain ancillary or more mediated videos and news stories than the initial run of Syrian activist videos will be added to as well. This will give an image of how more mainstream news outlets document the situation in Syria versus the activist video itself.

IV. Test Crawl and Adjustments

As stated above, this project is in its alpha stage; efforts will expand as needed and along with proficiency in Archive-It and various technical allowances. For instance, right now the crawl is set to update monthly, but during periods of special importance, crawls may be scheduled more frequently, perhaps as often as daily. Keeping in mind that a primary concern is to not unduly impact the activities of these websites, a test crawl with limited parameters and adjustments was run.

After consulting with *Witness*, certain key decisions were made for the test crawl. We determined that Vimeo, not compatible with the Archive-It service, was

off-limits, and that password-protected Facebook pages, a serious web archiving challenge, would be saved for later crawls. For the test crawl, host constraints for Facebook (6,000, or 2,000/per site), Twitter (1,000), and YouTube (2,000) were set, ignoring Robots.txt for these seeds and all others. The parameters of the crawl were also set with definite end points, so it did not go off topic and gather irrelevant information and sites. The test crawl seeds were http://wawalwasel.com/en/, http://www.syrianhr.org/, http://www.youtube.com/user/ANAChannelEng/, https://twitter.com/ShaamNewsEn/, https://www.facebook.com/Idlebpressenglish/, https://www.facebook.com/pages/Aleppo-Media-Centre-English/252724514864976/, https://www.facebook.com/The.World.In.Lens/, and https://www.watan.com/.

Archiving the web resources proved to be a challenge, and the test crawl was of great value. Initially, the crawl was set up with minimal specific parameters to gather reports and see what needed to be adjusted. Later, consultation with an Archive-It representative took place, and adjustments were made to the scope of the crawl. We made extensive use of the "ignore robots exclusion standard" in our test crawl, so as not to disrupt the daily operation of the website. A number of the resources did not utilize the robots exclusion standard, have password-protected content, or other complications that caused the Archive-It crawler to fail. However, social media sites such as Twitter and Facebook presented unique challenges, as well as other resources that either used the robots exclusion standard and had password-protected content. Sites such as YouTube were crawled outside the parameters of the main sites; it was our hope that backing up video and photo files in this way would not drain the bandwidth of the activist sites. Since the bandwidth limits of the activists' websites was unknown, we limited our crawls, sparing their servers from the burden of excessive crawling.

The test crawl was for the most part a success, with 89,427 documents and 13 GB of data collected (with a significant proportion of video). There were difficulties with the password protected Facebook page owned by news organizations which were in turn blocked by robots.txt. YouTube surprisingly posed no difficulty during the crawl, with the majority of the video coming from there.

After the test crawl, a consultation with an Archive-It team member was held with several suggestions to change crawl parameters. The major adjustments after a consultation was to add fbcdn.net and akamaihd.net as host constraints to properly crawl those specific Facebook seeds; they are pages for those sites that actually host the content. The www.youtube.com host constraint was also changed to simply youtube.com with adding ytimg.com to capture more Youtube content.

V. Live Crawl Results

For the live crawl, the above adjustments were completed and a need seed added to crawl the ANA Youtube channel. The parameters for the new seed were listing the

type as RSS/News Feed, named "gdata….." per the Archive-It tutorial. The overall result was a success with 112,502 documents comprising 17 GB data (a total of 1,094 files were video), a substantial increase from the test crawl. Due to time constraints, a patch crawl was run to make sure everything would be backed up smoothly. The Facebook pages caused QA issues again, but most of their content was archived. Those excepted, most of the issues were addressed during this patch crawl. Further consultation to address issues and more analysis of the Facebook QA failures will be conducted.

VI. Conclusion and Future Plans

As stated earlier, crawls are scheduled to take place at the rate of one a month for the foreseeable future. During this time, *Witness* is hoping to add more Syrian activist sites to the main project, and have a working archive dedicated to the ongoing conflict. Additional meetings with Yvonne will be held to review the content of the archive, and obtain more information as to what is desired from *Witness* in this project.

One change will be gathering the metadata from websites. *Witness* uses PBCore, while Archive-It naturally uses the related Dublin Core standard. Consultation with Yvonne about metadata standards did not occur before or after the test to gather their metadata implementation for PBCore. In later crawls, additional fields from *Witness* will be used so that the material is more easily searchable (especially helpful for the thousands of pieces of video material).

Archive-It has proven to be a powerful and easy-to-use tool, and with the exception of password-protected or robot exclusion standard sites, has been successful in backing up data. Further refinements will result in more success and a better web archive for *Witness.*

Works Cited

Cvetkovich, Ann. "Video, AIDS, and Activism." *Art, Activism, Oppositionality: Essays from Afterimage.* Ed. Grant H. Kester. (Durham: Duke University Press, 1998).

Hallas, Roger. *Reframing Bodies: AIDS, Bearing Witness, and the Queer Moving Image*. (Durham: Duke University Press, 2009).

The Library of Congress Collections Policy Statements Supplementary Guidelines, *Web Archiving*. Washington, D.C.: Humanities and Social Sciences Division, 2008