Living Movements, Living Archives: Selecting and Archiving Web Content During Times of Social Unrest

Sylvie Rollason-Cass & Scott Reed

Routledge
Taylor & Francis Group

Living Movements, Living Archives: Selecting and Archiving Web Content During Times of Social Unrest

SYLVIE ROLLASON-CASS and SCOTT REED
Archive-It Internet Archive, San Francisco, California, USA

The ease of creating and sharing content on the web has had a profound impact on the scope, pace, and mobility of social movements, as well as on how the documents and evidence of these movements are collected and preserved. This article will focus on the process of creating a web based archive around the #blacklivesmatter movement while exploring the concept of the "living archive" through collaborative collection building around social movements. By examining this and other event-based web collections, best practices and strategies to improve the process of selection and capture of web content in Living Archives are presented.

KEYWORDS cultural responsibility, web archiving, social movements, living archives

INTRODUCTION

On August 15, 2014 web archivists from Archive-It, a subscription web archiving service of the Internet Archive, began archiving web content related to the shooting death of Michael Brown by Officer Darren Wilson in Ferguson, Missouri and the ensuing unrest in the community and around the United States. Currently that collection, titled the #blacklivesmatter Web Archive, contains over nine hundred URLs related to the movement against the mistreatment of African Americans and other minorities at the hands of law enforcement. By employing a "living archive" model, this collection

Address correspondence to Sylvie Rollason-Cass, Archive-It, Internet Archive, 300 Funston Ave., San Francisco, CA 94118, USA. E-mail: sylvie@archive.org

has been able to evolve and grow just as #blacklivesmatter evolved out of specific events and grew to become a large-scale movement.

BUILDING THE #BLACKLIVESMATTER WEB ARCHIVE

The process for building the #blacklivesmatter Web Archive followed Archive-It's Spontaneous Events model, which was developed in 2007 in collaboration with Virginia Tech to archive web content related to the April 16th shooting on their campus. This model provides Archive-It with a framework for creating collections of web content related to a specific event and archiving content that may be at risk during times of crisis, as well as for providing public access to those collections. Since its inception in 2007 the Spontaneous Events model has been used to build over 15 collections on topics such as the Earthquake in Haiti, the 2013 U.S. Government Shutdown, and the attack on the Charlie Hebdo offices in Paris. Depending on the nature of the event, nominations for URLs to be archived are accepted directly from librarians, archivists, or other subject matter experts with a knowledge base relevant to the event in question, from crowdsourcing via social media, or a combination of the two sources. The nature of the event also dictates the frequency and duration of web crawls.

Because the events unfolding in Ferguson, Missouri and across the country were occurring at the community level, crowd-sourcing URLs for the collection was deemed the most appropriate avenue and Archive-It began soliciting URL nominations through social media via a Google form. Shortly after the URL nomination form was shared, Washington University in St. Louis, Missouri and Maryland Institute for Technology in the Humanities (MITH) began to submit URLs directly. Both institutions had also recently begun their own Ferguson related digital archiving projects. Later, the Wisconsin Historical Society would also directly contribute URLs to the collection. The nomination form contained a field for a URL along with fields for title, description, type of web resource (news article, social media, etc.), and subject all using fields from the Dublin Core element set. This metadata would be used to help describe the resources on the publicly available collection page. It could also help facilitate interoperability with other collecting projects and potentially contribute to the collection's longevity, as recommended by digital archivists at the Roy Rosenzweig Center for History and New Media (RRCHNM) at George Mason University (Brennan and Mills 2009). The nominated URLs were then crawled using the Heritrix web crawler, with crawled data stored in the WARC format, and made accessible via the Wayback software at https://archive-it.org/collections/4783.

The initial title of the collection was "Ferguson, Mo—2014"; however, in the following months it appeared that reactions around the country were no longer focused solely on the events in Ferguson. In December of

2014 protests were reignited after a Staten Island, NY grand jury decided not to indict Officer Daniel Pantaleo in the chokehold death of Eric Garner. At this point it became apparent that this was a growing movement focused on race and police tactics. Later that month the collection title was changed to "Police Involved Deaths: Acquittals and Reactions" and another call for crowdsourced URLs was issued via social media. Over the next few months specific events were still spurring protests and reactions but were also carrying the broader message of mistreatment of African Americans by law enforcement and began to solidify under the banner of #blacklivesmatter. Once again, the collection had outgrown its title. Archive-It decided #blacklivesmatter Web Archive best described the scope of the collection and allowed for content from events in major cities where high profile deaths had occurred (Ferguson, New York, Baltimore) but also from areas where community organizing around #blacklivesmatter was significant, for example the Blackout Collective based in the San Francisco Bay Area in California.

WHAT IS A LIVING ARCHIVE AND WHY IS IT IMPORTANT?

The concept of the Living Archive is not a widely researched topic; however, it is crucial to understanding the trajectory of a project like the #blacklivesmatter Web Archive. In "A Living, Breathing Revolution: How Libraries Can Use 'Living Archives' to Support, Engage, and Document Social Movements" Tamara Rhodes (2013) describes a variety of uses of the term and provides examples of projects using variations of it. Many of the ideas she discusses are specifically applicable to, and provide some structure for, the event-based web archiving projects led by Archive-It. A unifying theme of all projects Rhodes discusses is the collection and curation of materials while an event is occurring. Another is a democratic or collaborative approach to collecting that highlights community members as key contributors and stakeholders in the collection. Naira Antoun, in her description of efforts to archive the 2011 Egyptian Revolution, describes the Living Archive as an attempt to "record popular contemporary memory at a moment when that memory itself is in a moment of flux and contestation" (Antoun 2012, 1). This idea aligns well with a third understanding of a living archive mentioned by Rhodes, as the need for fluidity and the "ability for change within the collection" (2). Capturing content about an event as it occurs, using content contributed by members of the communities that are actively participating in or are affected by events, and allowing the collection trajectory to evolve if it becomes necessary have been the three principles on which the #blacklivesmatter Web Archive were constructed.

Earlier examples of collaborative digital archives also lend precedence to the #blacklivesmatter Web Archive. The September 11th Digital Archive

TABLE 1 Occupy Web Archive

Type of Website	Sites Archived in 2011–12	% Available on Live Web in 2014
Movement Sites (e.g., Occupy Seattle)	582	41
Social Media Pages	203	85
News Articles	163	90

and The Hurricane Digital Memory Bank, both collaborative projects built by RRCHNM, are examples that utilize the Omeka content management system to collect and arrange content from visitors with the specific intent of archiving born digital content and firsthand accounts of these major social events from individuals or groups who experienced them. Occupy Wall Street is an often cited example of a social movement that spawned a number of collaborative digital and web archiving projects, including the collaborative OWSArchives, Occupy Archive from RRCHNM, and the Occupy Movement 2011/2012 collection from Archive-It. Even three short years later these collections can provide a tangible example of why collecting this content is important. For example, Archive-It conducted an informal study in 2014 comparing the archived versions of local and national Occupy related websites in their Occupy Movement 2011/2012 collection with the versions that were available on the live web (Reed and LaCalle 2014). The findings, outlined in Table 1, elucidate the ephemerality of web content even after only a few years, as well as the necessity of capturing web content as the movement is taking place.

COLLABORATIONS TO BUILD THE #BLACKLIVESMATTER WEB ARCHIVE

As previously referenced, at least two different projects were initiated to archive digital content related to the events in Ferguson. One was an effort led by Ed Summers (2014) of the Maryland Institute for Technology in the Humanities. In the week after Michael Brown's death, Summers began to collect tweets containing the term "Ferguson" using a Python command line program called twarc, and Twitter's search API that allowed him to collect tweets from up to about one week prior. Running this program, Summers was able to collect 13,480,000 tweets from August 10th through August 27th. From that corpus he was able to extract URLs that were linked from individual tweets. Archive-It received the 400 most frequently referenced URLs and crawled them as a part of the #blacklivesmatter Web Archive collection. The entire list of URLs was crawled as a part of the Internet Archive's Wayback Machine and the tweet IDs are available in a collection on archive.org (Ferguson Tweet IDs).

Another project, titled Documenting Ferguson, was conducted by Washington University in St. Louis, Missouri. Because they are located in the vicinity of Ferguson, they were able to collaborate with other institutions and organizations in the area to collect and preserve content related to events with local significance that carried national impact. This project employs a version of the living archive seen in the event-based archives of RRCHNM, including the use of Omeka and digital content focused on firsthand accounts of the event. A unique quality of this archive is that it also mobilized further collaboration and contribution to the #blacklivesmatter Web Archive by linking to the Archive-It URL submission form and to the Archive-It collection, allowing their users to also submit web resources and see them archived via one portal.

CHALLENGES, LIMITATIONS, AND POTENTIAL FOR COLLABORATIVE EVENT-BASED WEB ARCHIVES

The #blacklivesmatter Web Archive illuminates the potential of archiving a movement on a national scale with multiple invested institutional partners and the assistance of a varied crowdsourced approach to curation. It also presents multiple challenges that are not unique to any one web archiving project.

The web archiving community is an experienced group of practitioners and technologies, and yet, as the web changes rapidly, most capture is "best effort" rather than completely comprehensive. The most common and mature method for capture is the Heritrix web crawler, used for the #blacklivesmatter Web Archive. But as the MITH Twitter archive shows, other methods can be mobilized in particular instances where Heritrix cannot act quickly or comprehensively enough.

Until a more complete and comprehensive archive of dynamic web content can be captured, it may be necessary to document the technical specifications and decisions made during the process of collecting the content. For example, some URLs are intended to be a complete website capture but may be missing key components (e.g., embedded video) because a particular format is difficult for the crawler to capture. Because of the time and technical expertise required for quality assurance, a collection of this size will inevitably be incomplete in some respects. Future projects could dedicate more technical and engineering resources to get a better capture of content, as well as resources to document the quality assurance process and any scoping decisions made during capture.

Further education and outreach around digital and web archiving can be achieved during an event to encourage wider participation in URL nomination. For at risk URLs to be archived they need to be nominated by informed individuals, preferably those locally involved in social movements or events.

Efforts like the tweet archive generated by MITH can provide alternative solutions to URL selection when events are at a particularly large scale and more manual URL nomination efforts are not effective or lose momentum over time.

CONCLUSION

As a living archive, the #blacklivesmatter Web Archive suggests that web archives focused on social movements should actively seek collaboration and flexibility to be accountable for both the limitations of web archiving and the constant change of the movement it seeks to document. By employing crowdsourcing campaigns and with the help of institutional partners, the archive can have a wider and more accurate scope. By employing various technologies to capture content on the web, the archive can achieve a more robust and diverse capture of fast moving data.

Spontaneous event collections through Archive-It and the Internet Archive have continually improved through focused efforts to capture web content while events occur. Living Archives such as these can improve by investing resources into the aforementioned challenges. However, as they stand, they provide a valuable resource for researchers, activists, and historians who study the web and its impact on large social, political, and environmental events.

ACKNOWLEDGMENTS

REFERENCES

Antoun, N. *The Archive of the Revolution: A Living Archive*. Arabic Literature (In English). 30 Sept. 2012. Web. 25 May 2015. <http://arablit.wordpress.com/2012/09/30/the-archive-of-the-revolution-a-living-archiv>.

Archive-It Spontaneous Events, n.d. Web. 25 May 2015. <https://archive-it.org/learn-more/spontaneous-events>.

#blacklivesmatter Web Archive, n.d. Web. 25 May 2015. <https://archive-it.org/collections/4783>.

Brennan, S., and T. Mills Kelly. *Why Collecting History Online is Web 1.5*. Essays on History and New Media, 2009. Web. http://chnm.gmu.edu/essays-on-history-new-media/essays/?essayid=47

Documenting Ferguson, n.d. Web. 25 May 2015. <http://digital.wustl.edu/ferguson/>.

Ferguson Tweet IDs, n.d. Web. 25 May 2015. <https://archive.org/details/fergusoncrawl>.

Hurricane Digital Memory Bank, n.d. Web. 25 May 2015. <http://hurricanearchive.org/>.

Occupy Archive, n.d. Web. <http://occupyarchive.org/>.

Occupy Movement 2011/2012, n.d. Web. 25 May 2015. <https://archive-it.org/collections/2950>.

OWS Archives, n.d. Web. <https://owsarchives.wordpress.com/>.

Reed, S., and M. LaCalle. *Only 41% of Occupy Movement Websites Accessible on Live Web*. 2014. Archive-It Blog. Web. 25 May 2015. <https://archive-it.org/blog/post/only-41-of-occupy-movement-urls-accessible-on-live-web>.

Rhodes, T. A Living, Breathing Revolution: How Libraries Can Use "Living Archives" to Support, Engage, and Document Social Movements. *Paper presented at: IFLA WLIC 2013 — Singapore — "Future Libraries: Infinite Possibilities" in Session 84 - Social Science Libraries*. 2013. Web. 25 May 2015. <http://library.ifla.org/52/1/084-rhodes-en.pdf>.

Summers, E. *A Ferguson Twitter Archive*. Inkdroid.org. 2014. Web. <http://inkdroid.org/journal/2014/08/30/a-ferguson-twitter-archive/>.

September 11th Digital Archive, n.d. Web. 25 May 2015. <http://911digitalarchive.org/>.