Bootstrapping Web Archive Collections of Stories from Micro-collections in Social Media

Alexander C. Nwala
Old Dominion University
Norfolk, Virginia 23529, USA
anwala@cs.odu.edu

ABSTRACT

Archive-It collections of archived web pages provide a critical source of information for studying important historical events ranging from social movements to terrorist events. The ever-changing nature of the web means that web archive collections preserve some of our collective digital heritage, and thus provide the means of studying events no longer present on the live web. There are many methods for collection building. Some adapt manual efforts, such as seed nomination on Google Docs, to begin the collection building process. The seeds are subsequently crawled (e.g., with focused crawlers) to discover more URIs. The different methods for generating collections solve some aspects of the collection building problem, but, irrespective of the method of collection building, most methods begin with seeds - an initial representative list of URIs (Uniform Resource Identifier) for the collection topic. Consequently, the discovery of seeds is a critical aspect of collection building. The traditional method of seed discovery requires manual effort and is an arduous process that often requires some domain knowledge about the collection topic, such as disasters and popular uprisings. This potentially limits the number of collections generated for important newsworthy events. We propose a seed generation method that extracts seeds from user-generated collections (*micro-collections*) in social media such as Twitter, Wikipedia references, and Reddit. The discovered seeds may augment existing collections or bootstrap new collections, thus accelerating the collection building process that largely relies on a few curators to start. Additionally, we propose a Collection Characterization Suite (CCS) to characterize and evaluate the collections generated. An important part of collection generation is the characterization or description of the collections that are generated and not just the generation of collections. The CCS provides a means of characterizing individual collections and serves as a baseline for comparing multiple collections.

CCS CONCEPTS

•Information systems →Digital libraries and archives;

KEYWORDS

Seeds; Collection building; Web Archiving; Micro-collections; Crawling

ACM Reference format:
Alexander C. Nwala. 2018. Bootstrapping Web Archive Collections of Stories from Micro-collections in Social Media. In *Proceedings of Joint Conference on Digital Libraries Doctoral Consortium, Fort Worth, TX, USA, June 2018 (JCDL'18),* 8 pages.
DOI: 10.1145/XXXXXX.XXXXXX

JCDL'18, Fort Worth, TX, USA

DOI: 10.1145/XXXXXX.XXXXXX

1 MOTIVATION AND INTRODUCTION

A few months after the 2014 Ebola virus outbreak that severely affected Guinea, Liberia, and Sierra Leone, an archivist at the National Library of Medicine (NLM) collected seeds on Archive-It [19] for the Ebola outbreak. A seed list is an initial collection of URIs (Uniform Resource Identifiers) of exemplar web pages for a topic and is subsequently crawled to discover more URIs. Archive-It, a collection development service deployed by the Internet Archive in 2006, enables members to build thematic collections of archived web pages, and these collections begin with seeds. Human-generated seed collections of archived web pages, such as the NLM Archive-It *Ebola virus* collection, are time consuming to create and often require domain knowledge about the collection topic, ranging from disasters (e.g., 2011 Deep Water Horizon oil spill) to popular uprisings (e.g., 2010/2011 Arab Spring). This makes it difficult for non-specialists to quickly create collections as events unfolds. To cope with the problem of a shortage of curators amidst an abundance of world events, various organizations such as the Internet Archive (IA) routinely request (Fig. 4) for users to contribute links to seed Archive-It collections e.g., the *2016 Pulse Nightclub Shooting* [12], the *2016 U.S. Presidential Election* [9], and the *Dakota Access Pipeline* [10] collections.

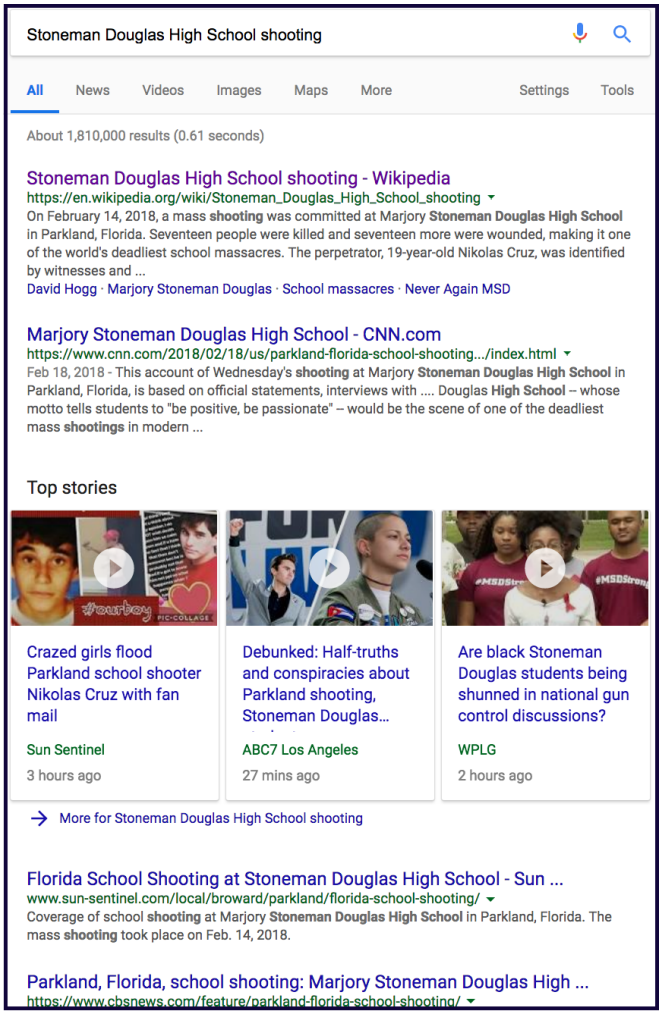
In contrast, it is common practice for users on social media sites such as Storify[1], Reddit, Twitter, and Wikipedia to share stories or commentaries about news events consisting of hand-selected URIs of news stories, tweets, videos, etc. For example, on February 14, 2018, there was a tragic shooting that claimed the lives of 17 people at Stoneman Douglas High School. On the same day as the event, a Wikipedia[2] page (Fig. 3a) was created for the event. Almost 10 months after the event, the references from the *Stoneman Douglas High school shooting* Wikipedia page (Fig 3b) had over 230 URIs pointing to news articles and other relevant web pages about the shooting event. Similarly, one day after the shooting event, a Twitter Moment [33] (Fig. 1a) was created. It consists of URIs of news stories as well as videos, images, and tweets about the event. As of December 8, 2018, there was no Archive-It collection about the event, thus URIs from the Twitter Moment (Fig. 1a) and Wikipedia references (Fig. 3b) for the shooting event may be used as seeds to bootstrap an Archive-It *Stoneman Douglas High school shooting* collection.

Comparably, Fig. 2a shows a story on Storify (created January 2014) about the riots in Kiev, Ukraine. This was before the incident became a crisis in late February 2014 when Russia began the annexation of the Crimean Peninsula. In contrast, the Archive-It collection about the Ukraine conflict (Fig. 2b) started in February 2014, and potentially omits some of the prelusive contents in the Storify story (Fig. Fig. 2a) which could be used to augment the Archive-It collection.

The *Ukrainian conflict* event highlights a common scenario in which users on social media express early interest and build *micro-collections* for events before they gain prominence in the public discourse. It is also part of this research effort to find these high-quality *micro-collections*

[1] Storify is scheduled to go out of service in May 2018.
[2] https://en.wikipedia.org/wiki/Stoneman_Douglas_High_School_shooting

(a) A Twitter Moment [33] about the Stoneman Douglas High School shooting created the day after (February 15, 2018) the tragic incident. Social media collection such as this provides the opportunity for augmenting existing archived collections or bootstrapping archived collections. This is especially useful when no archived collection for the event exist; as of December 8, 2018, there was no Archive-It collection for the *Stoneman Douglas High School* shooting event. This page has been edited to show more detail.



(b) Google SERP (Search Engine Result Page) for query: "Stoneman Douglas High School shooting." Since SERPs are a common means of URI discovery, we propose evaluating *micro-collections* against seeds extracted from SERPs.

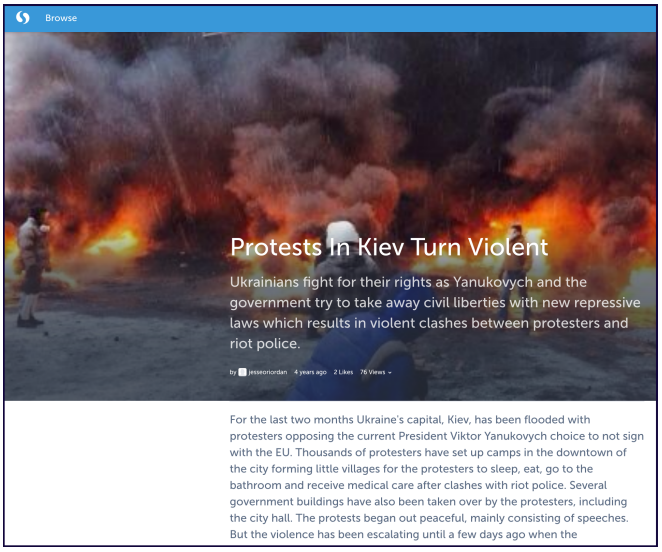
Figure 1: Links extracted from social media collections such as Twitter moment (left) can be used to augment existing archived collections or bootstrap new collections.

in social media to bootstrap archived collections. Kleinberg [16] introduced the concepts of *authorities* (information sources) and *hubs* (provide links to authorities) in the web graph. Similarly, we consider *micro-collections* as valuable *hubs* that could provide high-quality URIs that could be leveraged to generate seeds. Even though search engines are the primary means of discovery on the web, and could be seen as *hubs*, search engines prioritize recency, thus, produce the most recent documents with respect to the time a query is issued

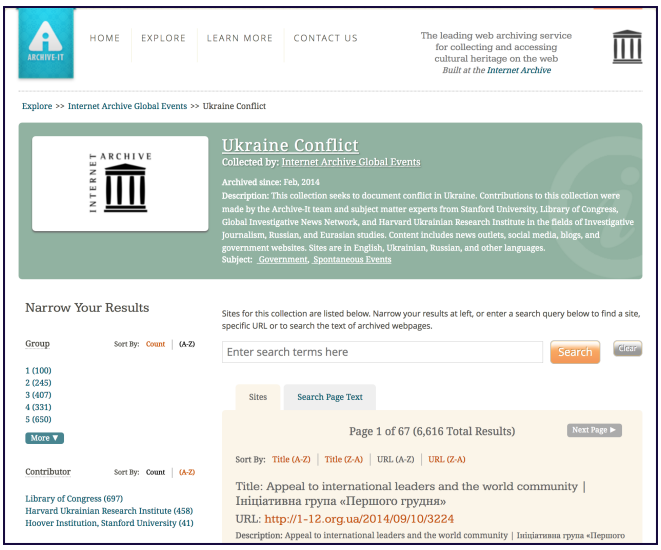
Collections created by social media users offer the opportunity for bootstrapping archived collections. Therefore, we propose a method of exploiting the collective domain expertise of web users by using collections they are already creating to augment or bootstrap archived collections. In other words, the URIs extracted from such collections may augment curator-selected seeds for various news events. For example, Table 1 juxtaposes seeds from an Archive-It collection and URIs extracted from Reddit and Wikipedia[3] for the Ebola virus topic. URIs from Reddit and Wikipedia can also be used to augment existing *Ebola virus* collections or bootstrap new ones. Since important events occur at a rapid pace, we cannot rely exclusively on archivists and curators for generating collections. Generating seeds from user collections on social media provides the opportunity for building a larger number of collections faster for important news events and for assisting archivists and curators in the collection building process.

Seed collections are the beginning of archived collections. Archived collections are a critical source of information for studying past events that may no longer be present on the live web due to link rot and content drift. There are many studies that highlight the link rot and content drift problem that plagues the web. For example, SalahEldeen and Nelson [29] showed that 11% of resources shared on social media are lost

[3]https://en.wikipedia.org/wiki/Ebola_virus_disease



(a) A story [13] from Storify: "Protests In Kiev Turn Violent," published in January 2014. We propose extracting URIs from *micro-collections* such as this to generate seeds.



(b) The *Ukraine Conflict* Archive-It collection [11] created February 2014.

Figure 2: The micro-collection from Storify (left) for the *Ukrainian crisis* event was created in January 2014 and highlights incidents such as riots before the event became a prominent news event. Russia began the annexation of Crimea in late February coinciding with the creation of the Archive-It collection (right). The Archive-It collection potentially omits some of the prelusive contents in the Storify *micro-collection* (a).

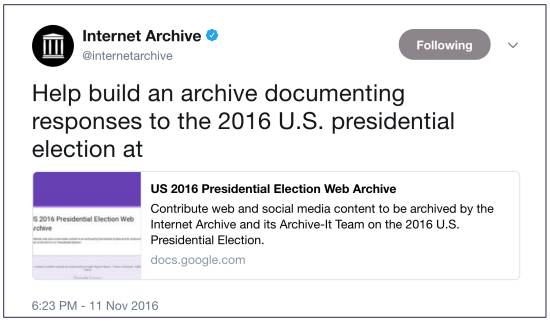
(a) The Wikipedia page about the Stoneman Douglas High School shooting was created the same day as the shooting event (February 14, 2018).

(b) References from Wikipedia Stoneman Douglas High School shooting page. As of December 8, 2018, it had over 230 references. We propose extracting URIs from *micro-collections* such as this to generate seeds.

Figure 3: As of December 8, 2018, there was no Archive-It collections for the Stoneman Douglas High School shooting. URIs from user-generated collections in social media such as Wikipedia page references provide the opportunity for augmenting existing collections or bootstrapping new archived collections.

after one year, and Klein et al. [15] showed that one in five scholarly articles suffers from reference rot. These studies highlight the importance of archived collections as a vital part of addressing the link rot and content drift problem.

The proposed method of collection generation in this work leverages social media *micro-collections* for collection generation to assist curators in the collection building process. However, it is very important to assess the quality of the collections generated from the proposed method. In other words, generating collections is not sufficient, the generated collections must be comparable to expert-generated collections. This requires a means of characterizing individual collections and comparing multiple collections (e.g., curator-generated vs. social media and SERP-generated). Consequently, as part of this work, we propose a Collection Characterizing Suite (CCS) to provide insight about the characteristics of a collection which forms the basis for comparing collections.



(a) A tweet from the Internet Archive requesting seeds for the U.S. Presidential Election collection.

(b) A tweet from the Internet Archive requesting seeds for the Dakota Access Pipeline collection.

Figure 4: The Internet Archive has on multiple occasions requested that users submit seeds to bootstrap collections.

2 RESEARCH QUESTIONS AND GOALS

The primary goal of this research effort is to bootstrap web archive collections by generating seeds extracted from *micro-collections* (user-generated collections) in social media. The primary research questions are as follows:

(1) **Research Question 1:** Are seeds that are generated automatically from *micro-collections* in social media comparable to curator-generated seeds?
(2) **Research Question 2:** If we consider curator hand-selected seeds the gold standard for collections, could this lead to the definition of what makes a collection good?

The seeds generated from *micro-collections* can be used to augment existing curator-generated seeds or bootstrap new archived collections. This means that the automatically generated seeds should be of a comparable quality to curator-generated seeds.

In order to assess if the seed collections generated from social media are similar in quality to seed collections created by curators or archivists, a collection similarity measure is required. In addition to assessing the similarity of collections, it is crucial to provide details about the characteristics of a "good" collection and how to distinguish it from a "bad" collection. If we consider curator hand-selected seeds collections the gold standard for collections, this may provide insight on the definition of what makes a collection good.

The contributions of this research effort are summarized as follows: an automatic method of generating seeds from *micro-collections* in social media and a Collection Characterization Suite (CCS) that provides a means for characterizing collections and a forms the basis for comparing multiple collections.

Table 1: Sample of seed URIs from Archive-It *Ebola virus* collection, URIs extracted from Reddit SERP and comments for query "Ebola virus," and URIs extracted from the references of the Wikipedia *Ebola virus* document.

Title	URI
Archive-It (seed URIs)	
Eman Reports From Ebola Ground Zero...	blogs.plos.org/dnascience/2014/11/06/eman-reports-ebola-ground-zero/
Human rights and Ebola: the issue of quarantine...	blogs.plos.org/globalhealth/2014/11/ebola_and_human_rights/
2014-2016 Ebola Outbreak in West Africa...	www.cdc.gov/vhf/ebola/outbreaks/2014-west-africa/index.html
#EbolaResponse (@ebola_response)...	twitter.com/ebola_response/
WHO — Situation assessments: Ebola virus...	www.who.int/mediacentre/news/ebola/en/
Reddit	
Liberia: Catholic Hospital Boss Tested Positive...	allafrica.com/stories/201407310957.html
Ebola plagues Africa nearly four decades...	america.aljazeera.com/articles/2014/8/1/ebola-explainer.html
Management of Accidental Exposure to Ebola Virus...	jid.oxfordjournals.org/content/204/suppl_3/S785.long
Analysis of patient data from laboratories during...	journals.plos.org/plosntds/article?id=10.1371/journal.pntd.0005804
Monkey Meat and the Ebola Outbreak in Liberia...	youtu.be/XasTcDsDfMg
Wikipedia	
Proposal for a revised taxonomy of the...	www.ncbi.nlm.nih.gov/pmc/articles/PMC3074192
Ebola outbreak in Western Africa 2014...	www.ncbi.nlm.nih.gov/pmc/articles/PMC4313106
Ebola data and statistics	apps.who.int/gho/data/view.ebola-sitrep.ebola-summary-latest
WHO - Ebola outbreak 2014-2015	www.who.int/csr/disease/ebola/en/
Ebola virus entry requires the cholesterol...	www.nature.com/articles/nature10348

3 RELATED WORK

There have been many research efforts addressing collection building. Not much effort has focused on assessing collections except from the Library Sciences domain. Also, it is important to note that existing research efforts address a single aspect of the two components of this proposed work such as: seed generation or collection characterization and evaluation. Combining these two components is one of the novel contributions of this research.

Not many efforts addressing the seed selection process of collection building exist. Schneider et al. [30] proposed the continuous selection of seeds for thematic collections about evolving events, but does not address the source for selecting the seeds. Du et al. [3] proposed a seed selection process based on user-interest ontologies. Priyatam et al. [23] proposed a seed discovery method from URIs in tweets for initializing domain-specific search engine crawlers. Extracting URIs from tweets or extracting tweets discovered through search or crawling a hashtag is also another common strategy used for social media collection building and seed generation. For example, the Integrated Digital Events Archive and Library has collected millions of tweets [8] for various topics such as the 2014 *Ebola virus* outbreak. Most focused crawling is performed on the live Web. Unfortunately, the live web is plagued by link rot, consequently, Klein et al. [14] demonstrated that focused

crawling on the archived Web results in more relevant collections than focused crawling on the live Web, for events that occurred in the distant past. Additionally, similar to this work, Klein et al. proposed extracting seeds from external references contained in the Wikipedia page of an event. However, instead of utilizing the live version of the Wikipedia page, they proposed using the version of the Wikipedia page that corresponds with the datetime after which the edit frequency drastically decreases. The method proposed in this effort utilizes *micro-collections* on social media (e.g., Twitter Moments) which are hand-crafted collections created by users as opposed to URIs returned from the Twitter SERP or crawling a hashtag. Vieira et al. [34] proposed a seed selection method that utilizes search engines and utilizes a Pseudo-Relevance Feedback to discover more queries to issue to the search engine for discovering more seeds. Other efforts [36] address the seed selection process of collection building but do so in the context of generalized crawling, as opposed to focused crawling for thematic collections.

Collection characterization is an important part of collection building. The second component of this research effort is the characterization and assessment of collections. In the Web Science domain, efforts related to characterizing or assessing generated collections are few unlike in the Library Sciences domain. Thus, the efforts addressing the collection characterization from the Library Science domain inform the methods proposed in this research effort.

Due to insufficient funding and limited shelf space, libraries strive to maximize the quality of their holdings to satisfy the needs of their users. In the web domain, storage is cheap but quality still needs to be maximized. The questions proposed by the library sciences such as "How does one evaluate collection strength?" and "What is a good collection?" are applicable and can inform this research. In 1974, Bonn [2] presented different quantitative methods for evaluating various library collections and expressed the need for library collections to be varied in order to fulfill the needs of various academic programs. In the 1980s, the Research Libraries Group (RLG), a consortium of libraries in the U.S, published the RLG six (0-5) collecting levels [4, 6]. The collecting levels were used to quantify the strength of collections. In summary, level 0 means the library collection is out of scope with respect to a subject, and level 5 means the collection is comprehensive. In 1995, White published the *Brief Tests of Collection Strength* [35] in which he outlined a systematic method of comparing a short list of items (brief tests) to a library's collection. Thereafter, he scored the library's holding on the RLG scale. More recently (2004), David Lesniaski provided a simplification [17] of White's brief tests in order to make the test more adaptable by smaller college libraries. He also expressed the notion that there is not a single meaning of a "good" library collection since the meaning is defined by the user or target audience of the collection. Many solutions offered by libraries for quantifying collection strength can be summarized into two broad categories: collection-centered and use-centered [18]. Collection-centered methods include comparing a collection against an expert-provided gold standard bibliographical set. Use-centered methods include assigning the strength score to a collection based on circulation and interlibrary loan statistics and patron surveys [7]. These methods for evaluating collections and details about what characteristics that make a "good" collection are relevant and could inform the second component of this research effort - collection characterization and evaluation.

4 PRELIMINARY WORK

This section presents the progress that has been made towards this research effort.

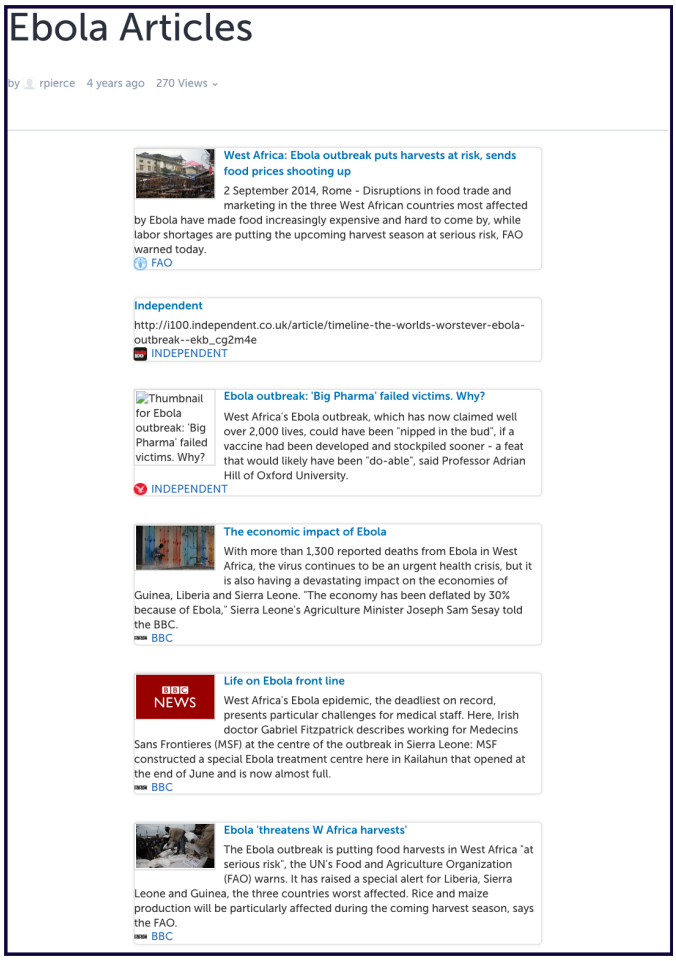


Figure 5: Example of a micro-collection [28] on Storify titled: "Ebola Articles," created in August 2014 consists of 20 URIs of news articles related to the 2014 Ebola Virus outbreak.

4.1 Bootstrapping Web Archive Collections from Social Media

We implemented a prototype system [20] for generating seeds from the Google SERP and the following social media sites: Storify, Twitter Moment, Reddit, and Wikipedia. The following is a summary of how the seed collections were generated from various sources.

The high-quality Google SERP (Fig. 1b) provides an opportunity to generate collections for news stories and events. We investigated the discoverability of news stories on the Google SERP [21] and found out that the probability of finding the URI of a news story diminishes with time. It ranged between 0.34 – 0.44 daily and between 0.01 – 0.11 weekly. Consequently, the process of generating seeds from the Google SERP has to begin days after events in order to capture the first stages of events and should persist in order to capture the evolution of the events, because it becomes more difficult to find the same news stories with the same queries on Google as time progresses.

Storify is a social media curation service that enables users to create *stories* that consist of hand-selected web resources (e.g., Fig. 5) such as URIs of news articles, images, videos, etc. Unfortunately, Storify went out of service in May 2018 [32], but we are exploring other possible alternatives [31]. Storify provides search a functionality on their

Table 2: Summary of research schedule plan

Time Frame	Phase Description	Evaluation	Status
2016 – 2017	Local Memory Project (LMP): Highlighted need for using local news sources to build collections for local events.	Measured precision, archival and tweet coverage, temporal range, and overlaps for local and national story stories. Also, showed that local news sources are less exposed than non-local news sources	Completed (Published: JCDL 2017 [22])
2017 – 2018	Demonstrate feasibility of generating seed collections from social media and define Collection Characterizing Suite (CCS)	Measured distance between Archive-It seeds and user-generated collections on social media, showing they are comparable with distance range between 0.17 to 0.34	Completed (Submitted)
2017 – 2018	Investigate discoverability of URIs of news stories on SERPs	Calculated the probability of finding a story as a function of time and the new story rate on the Google SERP	Completed (Accepted: JCDL 2018)
2018 – 2019	Identify hubs and authorities in social media	Assess similarity (with CCS) between seeds generated from hubs and gold-standard Archive-It seeds and SERP collections	Pending
2018 – 2019	Identify and evaluate sources of social media *micro-collections*	Assess similarity (with CCS) between seeds generated from *micro-collections* and gold-standard Archive-It seeds and SERP collections	Pending
2018 – 2019	Candidacy Proposal	Submit and defend candidacy proposal	In progress
2019	PhD Defense	Graduation	Pending

website, but their content is more discoverable via Google [1]. Consequently, Google search (with the *site:storify.com* directive) was used to search for Storify stories, and URIs from the stories were used to populate seed lists.

Twitter Moment is a service by Twitter that lets users create topical collections of tweets (e.g., Fig. 1a). The tweets may embed URIs and multimedia content. Similar to Storify, Google search was used to search for Twitter Moments for various topics, and the URIs from the Twitter Moments were used to populate seed lists.

Reddit is a service that allows users to post URIs for various topics. Reddit users rate the URIs and post comments that may also include URIs. Reddit provides search, thus, the URIs from the Reddit SERP and their respective comments for relevant topics were added to seed lists.

The Wikipedia encyclopedia is a service that enables multiple contributors to create documents about various topics ranging from politics to science and technology. Wikipedia documents often include URIs of external references that are relevant to the document topic. For example, Table 1 consists of a sample of URIs extracted from the references of the Wikipedia document for the *Ebola virus* event. The URIs from the Wikipedia references were used to populate a seed list.

4.2 Collection Characterization

Characterizing and comparing collections is a challenging task because it requires comparing collections that may cater to different needs. It is also challenging to compare collections since there are many possible measures to use as a baseline for collection comparison: how does one narrow down this list to metrics that reflect if two collections are similar or dissimilar? Inspired by the state of the art in collection characterization in Library and Web Sciences, we defined a suite of seven measures (Collection Characterizing Suite - CCS) [20] to describe the individual collections and compare different collections. With the CCS, we showed that Archive-It seed collection were similar to collections generated from social media and SERPs. The CCS consists of the following measures:

(1) **Distribution of topics:** a ranked list of topics in a collection with the most frequent topics (most important summaries) at the top and the least frequent topics (least important summaries) at the bottom.
(2) **Distribution of sources (hostnames):** a statistical summary of the various sources sampled in order to build the collection.
(3) **Content diversity - Document-Term matrix & List of Entity sets:** the degree of self-similarity of the content of the web pages in the collection.
(4) **Temporal distribution - Publication and Content:** an aggregation of the dates in a collection.
(5) **Source diversity - URI, Domain, Hostname, and Social media:** indicates whether a collection samples a single source, a handful of sources, or many sources.
(6) **Collection exposure - Archival rate and Tweet index rate:** approximates popularity.
(7) **Target audience:** approximates the target audience of a collection with readability scores.

5 PROPOSED WORK

Table 2 is a summary of the work plan for this research effort. We have demonstrated the feasibility of bootstrapping web archive collection from seeds extracted from user-generated collections in social media [20] and introduced the CCS [21]. Section 4 outlines the details of these milestones. The following is a description of the remaining work preceding my candidacy proposal and defense.

5.1 Identifying hubs and authorities in social media

Similar to Kleinberg's method that provides hub and authorities scores in the web graph, we propose a method for dynamically identifying *hubs* and *authorities* for various topics on the web. We consider *micro-collections* in social media as *hubs* that point to *authorities* for various topics. Consequently, a crucial part of generating collections from

(a) Tweet published on March 22, 2018. **(b) Tweet published on March 26, 2018.** **(c) Tweet published on March 26, 2018.**

Figure 6: As of March 2018, three recent tweets from Ron Fonger linking to three new stories ([25–27]) about the Flint water crisis. Twitter accounts such as his could be potentially labeled as a *hub*. In contrast, the last news stories from *CNN* and *FoxNews* about the Flint water crisis were in November 2017.

micro-collections in social media relies on the identification of *hubs* and *authorities*.

One approach of identifying *authorities* involves labeling mainstream news media organizations such as *CNN* or *FoxNews* as authorities for various news topics, but this approach is flawed. The mainstream news media organization is driven by sensationalism, and has a short attention span. Many important news events start well before they become reported by mainstream news media organizations. For example, in April 2014, state officials in Flint, Michigan switched the city's water source from Lake Huron of the Detroit water system to the Flint River. This was reported by local media such as the Flint Journal-MLive, Michigan Radio, and the local TV affiliates in Flint (WEYI, WJRT, WSMH, and WNEM) [24]. Almost two years later in January 2016, Michigan Governor Rick Snyder declared a state of emergency for the city of Flint due to dangerously high levels of lead contamination in the drinking water. The declaration was preceded by a series of events such as complaints by residents about the water's taste and smell [5] and three boil advisories [24]. The preceding events such as the complaint about the water's taste and smell were reported by local media but not the mainstream media which took about one year to report the Flint story. The Flint water problem has not been solved and mainstream news media has moved on, but the story is still being reported by local media. For example, as of March 2018, the Twitter account of Ron Fonger, the reporter at Flint Journal-Mlive who reported the complaints about the water's taste and smell included links (Fig. 6) to recent (March 2018) developments for the Flint story ([25–27]). In contrast, the last report about the Flint event by *CNN* and *FoxNews* as of March 2018 was in November 2017. We propose investigating if such Twitter accounts could be labeled as *hubs* and/or *authorities* and dynamically identify them.

5.2 Identify sources for social media *micro-collections* and evaluate existing sources

We also propose identifying other possible sources for social media *micro-collections*. This is especially crucial given that some of services such as Storify could shutdown. Thus, it is pertinent that we identify various possible sources on social media that possess *micro-collections*. For example, threaded conversations (tweet threads) on Twitter include public exchanges between web users on various topics. The tweets in such exchanges include URIs hand-selected by the tweet author in the context of the conversation. However, it is also important to consider that tweet conversations could include a lot of noise such as off-topic content and spam. Consequently, evaluating the kinds of collection and/or filtering irrelevant content is an important aspect of generating seeds from *micro-collections* on social media.

Collection evaluation is not an easy task since this involves comparing collections that may cater to different needs. However, we could evaluate the *micro-collections* from new sources such as Twitter conversation thread against other known sources such as Archive-It seeds and SERP baselines collections. Collection building often begins with the use of a search engines to discover seeds. For example, this can be done by issuing queries to Google and extracting URIs from the SERP (Fig. 1b). The URIs extracted from Google can serve as seeds in Archive-It. Thus, we consider utilizing SERP collections as baselines in evaluating *micro-collections* in addition to Archive-It seeds.

6 CONCLUSIONS

Web archive collections provide the valuable opportunity of studying past events no longer present on the live web due to link rot and content drift. The traditional way of collection building involves the manual selection of seeds by curators or archivists for various collection topics. Given the rapid pace at which new important events unfold, many events and stories go unnoticed. Additionally, curators may not build collections that require some domain knowledge they lack, spanning from politics (e.g., *Brexit* event) in foreign countries to natural disasters (*2011 Japan Tsunami and resultant nuclear meltdown*). Some organizations such as the Internet Archive cope with the shortage of curators by requesting for the public to recommend URIs to bootstrap collections. Similarly, we propose a method for bootstrapping web archive collections by leveraging the domain knowledge of users on social media.

Specifically, we propose generating seeds to bootstrap archived collections by extracting URIs from *micro-collections* created by users on social media sites: Twitter Moment, Reddit, and Wikipedia. Additionally, we consider collection characterization another important aspect of collection building since it provides us with a means of describing or characterizing the collections that are generated by the collection building system.

We have demonstrated the feasibility of generating collections from user-generated collections in social media through prototype systems and proposed a means for characterizing and evaluating collections [20, 21]. The next stage of this work involves developing a method for dynamically identifying *hubs* and *authorities* on social media for various topics as well as the evaluation of seeds generated from various sources.

ACKNOWLEDGEMENTS

I am very grateful for the guidance and support of my PhD advisor Dr. Michael Nelson and co-advisor Dr. Michele Weigle. I also want to thank Christie Moffat at the National Library of Medicine and the reviewers of this extended abstract and welcome all feedback. This work was made possible by IMLS LG-71-15-0077-15.

REFERENCES

[1] Alexander C. Nwala. 2016. Can I find this story? API: Yes, Google: Maybe, Native Search: No. http://ws-dl.blogspot.com/2016/05/2016-05-31-can-i-find-this-story-api.html.
[2] George S Bonn. 1974. *Evaluation of the Collection.* Graduate School of Library and Information Science. University of Illinois at Urbana-Champaign.
[3] YaJun Du, YuFeng Hai, ChunZhi Xie, and XiaoMing Wang. 2014. An approach for selecting seed URLs of focused crawler based on user-interest ontology. *Applied Soft Computing* 14 (2014), 663–676.
[4] Anthony W Ferguson, Joan Grant, and Joel S Rutstein. 1988. The RLG Conspectus: its uses and benefits. *College & Research Libraries* 49, 3 (1988), 197–206.
[5] Ron Fonger. 2014. State says Flint River water meets all standards but more than twice the hardness of lake water. http://www.mlive.com/news/flint/index.ssf/2014/05/state_says_flint_river_water_m.html.
[6] Nancy E. Gwinn and Paul H. Mosher. 1983. Coordinating Collection Development: The RLG Conspectus. *College & Research Libraries* 44, 2 (1983), 128–140.
[7] Terese Heidenwolf. 1994. Evaluating an interdisciplinary research collection. *Collection Management* 18, 3-4 (1994), 33–48.
[8] Integrated Digital Events Archive and Library (IDEAL). 2015. Tweet Collections. http://www.ctrnet.net/node/942.
[9] Internet Archive. 2016. Help build an archive documenting responses to the 2016 U.S. presidential election at. https://twitter.com/internetarchive/status/797263535994613761.
[10] Internet Archive. 2016. What web pages should we save concerning DAPL? Tell us here:. https://twitter.com/internetarchive/status/806228431474028544.
[11] Internet Archive Global Events. 2014. Ukraine Conflict. https://archive-it.org/collections/4399/.
[12] Internet Archive Global Events. 2016. 2016 Pulse Nightclub Shooting Web Archive. https://archive-it.org/collections/7570.
[13] Jesse O'Riordan. 2014. Protests In Kiev Turn Violent. https://storify.com/jesseoriordan/ukraine-fight-for-their-rights.
[14] Martin Klein, Lyudmila Balakireva, and Herbert Van de Sompel. 2018. Focused Crawl of Web Archives to Build Event Collections. In *Web Science Conference (WebSci 2018).*
[15] Martin Klein, Herbert Van de Sompel, Robert Sanderson, Harihar Shankar, Lyudmila Balakireva, Ke Zhou, and Richard Tobin. 2014. Scholarly context not found: one in five articles suffers from reference rot. *PloS one* 9, 12 (2014), e115253.
[16] Jon M Kleinberg. 1999. Authoritative sources in a hyperlinked environment. *Journal of the ACM (JACM 1999)* 46, 5 (1999), 604–632.
[17] David Lesniaski. 2004. Evaluating collections: a discussion and extension of Brief Tests of Collection Strength. *College & Undergraduate Libraries* 11, 1 (2004), 11–24.
[18] Barbara Lockett. 1989. *Guide to the evaluation of library collections.* American Library Association.
[19] National Library of Medicine. 2014. Global Health Events. https://archive-it.org/collections/4887.
[20] Alexander C Nwala, Michele C Weigle, and Michael L Nelson. 2018. Bootstrapping Web Archive Collections from Social Media. In *Submitted for publication.*
[21] Alexander C Nwala, Michele C Weigle, and Michael L Nelson. 2018. Scraping SERPs for archival seeds: it matters when you start. In *Accepted: Joint Conference on Digital Libraries (JCDL 2018).*
[22] Alexander C Nwala, Michele C Weigle, Adam B Ziegler, Anastasia Aizman, and Michael L Nelson. 2017. Local Memory Project: Providing Tools to Build Collections of Stories for Local Events from Local Sources. In *Joint Conference on Digital Libraries (JCDL 2017).* 1–10.
[23] Pattisapu Nikhil Priyatam, Ajay Dubey, Krish Perumal, Sai Praneeth, Dharmesh Kakadia, and Vasudeva Varma. 2014. Seed selection for domain-specific search. In *International Conference on World Wide Web (WWW 2014).* 923–928.
[24] Denise Robbins. 2016. ANALYSIS: How Michigan And National Reporters Covered The Flint Water Crisis. https://mediamatters.org/research/2016/02/02/analysis-how-michigan-and-national-reporters-co/208290.
[25] Ron Fonger. 2018. Edwards puts blame for Flint water crisis at doorstep of Michigan DEQ. http://www.mlive.com/news/flint/index.ssf/2018/03/edwards_testimony_could_spark.html.
[26] Ron Fonger. 2018. Edwards testimony could spark battle of scientists in Flint water crisis. http://www.mlive.com/news/flint/index.ssf/2018/03/edwards_testimony_could_spark_1.html.
[27] Ron Fonger. 2018. Flint water plant manager says he put state on notice with bad info on lead. http://www.mlive.com/news/flint/index.ssf/2018/03/flint_water_plant_manager_says.html.
[28] rpierce. 2014. Ebola Articles. https://storify.com/rpierce/ebola-articles.
[29] Hany M SalahEldeen and Michael L Nelson. 2012. Losing my revolution: How many resources shared on social media have been lost?. In *International Conference on Theory and Practice of Digital Libraries (TPDL 2012).* 125–137.
[30] Steven M Schneider, Kirsten Foot, Michele Kimpton, and Gina Jones. 2003. Building thematic web collections: challenges and experiences from the September 11 Web Archive and the Election 2002 Web Archive. *Third Workshop on Web Archives* (2003), 77–94.
[31] Shawn M. Jones. 2017. Storify Will Be Gone Soon, So How Do We Preserve The Stories? http://ws-dl.blogspot.com/2017/12/2017-12-14-storify-will-be-gone-soon-so.html.
[32] Storify. 2017. Storify End-of-Life. https://archive.is/DOPFa.
[33] Twitter Moments. 2018. 17 people are dead after school shooting in Florida. https://twitter.com/i/moments/963863619271254016.
[34] Karane Vieira, Luciano Barbosa, Altigran Soares Da Silva, Juliana Freire, and Edleno Moura. 2016. Finding seeds to bootstrap focused crawlers. *International conference on World Wide Web (WWW 2016)* 19, 3 (2016), 449–474.
[35] Howard D White. 1995. *Brief tests of collection strength: A methodology for all types of libraries.* Number 88. Greenwood Publishing Group.
[36] Shuyi Zheng, Pavel Dmitriev, and C Lee Giles. 2009. Graph based crawler seed selection. In *International conference on World Wide Web (WWW 2009).* 1089–1090.