Bootstrapping Web Archive Collections from Social Media

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

Michele C. Weigle
Old Dominion University
Norfolk, Virginia, USA
mweigle@cs.odu.edu

Michael L. Nelson
Old Dominion University
Norfolk, Virginia, USA
mln@cs.odu.edu

ABSTRACT

Human-generated collections of archived web pages are expensive to create, but provide a critical source of information for researchers studying historical events. Hand-selected collections of web pages about events shared by users on social media offer the opportunity for bootstrapping archived collections. We investigated if collections generated automatically and semi-automatically from social media sources such as Storify, Reddit, Twitter, and Wikipedia are similar to Archive-It human-generated collections. This 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? We identified social media sources that may provide similar collections to Archive-It human-generated collections in two main steps. First, we explored the state of the art in collection comparison and defined a suite of seven measures (Collection Characterizing Suite - CCS) to describe the individual collections. Second, we calculated the distances between the CCS vectors of Archive-It collections and the CCS vectors of collections generated automatically and semi-automatically from social media sources, to identify social media collections most similar to Archive-It collections. The CCS distance comparison was done for three topics: "Ebola Virus," "Hurricane Harvey," and "2016 Pulse Nightclub Shooting." Our results showed that social media sources such as Reddit, Storify, Twitter, and Wikipedia produce collections that are similar to Archive-It collections. Consequently, curators may consider extracting URIs from these sources in order to begin or augment collections about various news topics.

CCS CONCEPTS

• **Information systems → Digital libraries and archives**;

KEYWORDS

Social Media, Collection evaluation, Web Archiving, News

ACM Reference Format:
Alexander C. Nwala, Michele C. Weigle, and Michael L. Nelson. 2018. Bootstrapping Web Archive Collections from Social Media. In *HT '18: 29th ACM Conference on Hypertext and Social Media, July 9–12, 2018, Baltimore, MD, USA.* ACM, New York, NY, USA, 9 pages. https://doi.org/10.1145/3209542.3209560

1 INTRODUCTION AND BACKGROUND

Following the 2014 Ebola outbreak in West Africa [6], an archivist at the National Library of Medicine (NLM) collected seeds [26] on Archive-It (a service of the Internet Archive) for the *Ebola virus* outbreak. The seed list is an initial collection of URIs (Uniform Resource Identifiers) representing exemplar web pages for the topic and are subsequently crawled in order to discover more URIs. Human-generated seeds of archived web pages, such as the NLM Archive-It *Ebola virus* collection are time consuming to create. These collections are usually of a high quality because humans do a good job of filtering irrelevant documents. However, important events can unfold at a rapid pace, consequently, we cannot rely exclusively on experts to generate seeds. 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 for users to contribute links to seed Archive-It collections, e.g. the *2016 Pulse Nightclub Shooting* [18], the *2016 U.S. Presidential Election* [16], and the *Dakota Access Pipeline* [17] collections.

It is common practice for users on social media sites such as Storify, Reddit, Twitter, and Wikipedia to share hand-selected stories for events. For example, Table 2 juxtaposes seeds from an Archive-It collection and URIs extracted from Reddit and Wikipedia for the *Ebola virus* topic. We claim these kinds of collections created by social media users offer the opportunity for bootstrapping archived collections. In other words, the URIs extracted from such collections may augment curator-selected seeds for various news events.

To assess the validity of our claim, we investigated if Archive-It seeds are similar to collections created from social media sources for the following topics: "Ebola Virus," "Hurricane Harvey," and "2016 Pulse Nightclub Shooting." Comparing collections is not an easy task especially when the collections are about the same topic. For example, given two collections, e.g., the NLM Archive-It *Ebola virus* collection and a collection of local news stories about the Ebola outbreak from Guinea [24], how can one tell which is the "better" collection? This is a difficult question because both collections cater to different needs and answer different questions. Therefore, to address the problem of comparing collections, we defined a set of seven metrics - Collection Characterizing Suite (CCS) - that objectively characterize individual collections. Subsequently, multiple collections can be compared by computing the distances between their respective CCS vectors. Here is a complete list of the CCS metrics:

(1) Distribution of topics
(2) Distribution of sources (hostnames)

HT '18, July 9–12, 2018, Baltimore, MD, USA

ACM ISBN 978-1-4503-5427-1/18/07…$15.00
https://doi.org/10.1145/3209542.3209560

(3) Content diversity: Doc-Term matrix & List of Entity sets
(4) Temporal distribution: Publication and Content
(5) Source diversity: URI, Domain, Hostname, and Social media
(6) Collection exposure: Archival rate and Tweet index rate
(7) Target audience

Our contributions are as follows. First, we provide a suite of metrics for characterizing collections (CCS). Second, we demonstrate how to compare multiple collections. Third, we provide novel methods for instantiating the metrics in the CCS. Fourth, we used the CCS to compare collections from social media sources and Archive-It collections, showing that these collections are similar. As a result, we propose the extraction of URIs from social media sources to bootstrap archived collections.

Table 1: CCS Metrics derived from the transformation of Library Science Collection Evaluation Metrics

Library Science Metrics	CCS Metrics
Usage statistics, e.g., circulation and interlibrary loan statistics	Exposure (or popularity): 1. Archival rate 2. Tweet index rate Target audience (reading level)
Variety of library collection	Content diversity: 1. Document-Term matrix 2. Entity set matrix Source diversity (policies): URI, Hostname and Domain
Bibliographical set comparison	Distribution of sources (hosts)

2 RELATED WORK

There are many efforts addressing generating collections for specific topics through the use of a focused crawler [7]. Many of these methods require a system that decides whether or not a web page is relevant to the collection topic. Bergmark [3] used a classifier to determine if web pages belonged to various topics in science and mathematics. Farag et al. [10] used a similarity measure that relies on an event model representation of documents in order to determine if web pages were relevant to an event-based collection. Gossen et al. [13] introduced iCrawl, a focused crawler that crawls social media content in order to generate thematically and temporally coherent collections. Similar to focused crawling research, we considered adding the precision metric to the CCS as a means to quantify relevance, but excluded it because we cannot objectively evaluate precision for some collections since relevance can be subjective and there is often no gold standard data available. Also, it may be impossible to automate precision evaluation for arbitrary collections, because evaluating precision requires some notion of relevance. In the absence of these concerns, the user of the CCS may include the precision metric. Many focused crawling efforts did not address how the seeds used to initialize focused crawlers were generated, which is an important part of this work. We do not consider using focused crawlers to crawl seeds to discover more relevant URIs, but focus on how seeds can be generated by exploiting social media collections.

Table 2: 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...	http://blogs.plos.org/dnascience/2014/11/06/eman-reports-ebola-ground-zero/
Human rights and Ebola: the issue of quarantine...	http://blogs.plos.org/globalhealth/2014/11/ebola_and_human_rights/
2014-2016 Ebola Outbreak in West Africa...	http://www.cdc.gov/vhf/ebola/outbreaks/2014-west-africa/index.html
Reddit	
Management of Accidental Exposure to Ebola Virus...	http://jid.oxfordjournals.org/content/204/suppl_3/S785.long
Analysis of patient data from laboratories during...	http://journals.plos.org/plosntds/article?id=10.1371/journal.pntd.0005804
Monkey Meat and the Ebola Outbreak in Liberia...	https://youtu.be/XasTcDsDfMg
Wikipedia	
Proposal for a revised taxonomy of the...	https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3074192
Ebola outbreak in Western Africa 2014...	https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4313106
WHO - Ebola outbreak 2014-2015	http://www.who.int/csr/disease/ebola/en/

Other efforts related to building collections address the difficulty of seed selection. Schneider et al. [34] proposed the continuous selection of seeds for thematic collections about evolving events. Zheng et al. [39] proposed different seed selection algorithms and showed that different seeds may result in collections that are considered "good" or "bad." It is important to note that the discovery of seeds is not the focus of this work, instead we propose to extract seeds from social media collections such as Storify, Reddit, Twitter, and Wikipedia, to augment existing seed selection methods.

An important part of utilizing social media sources to bootstrap archived collections is assessing if the collections generated from social media sources are similar to expert-generated seeds, specifically Archive-It collections. To compare collections, we first proposed the CCS, a suite of seven metrics for characterizing individual collections. Collections are subsequently compared by a distance calculation between their respective CCS vectors. We considered research from Library and Web Sciences about collection evaluation in order to identify widely used metrics to include in the CCS.

In 1974, Bonn [4] 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 [11, 14] 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. More recently (2004), Lesniaski provided a simplification [22] of White's brief tests [37] (comparing a short list of items to a library's collection) in order to make the test more adaptable by smaller college libraries. Additionally, he expressed the idea 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.

The questions proposed by the library sciences such as "How does one evaluate collection strength?" and "What is a good collection?" are applicable to the web domain. The solution offered by libraries to these questions (quantifying the strength of a collection) also inform the web domain through transformations. For example, the need for variety (or diversity) in library collections expressed by Bonn in 1974 is applicable to the web domain. Similarly, we included the *content diversity* metric in the CCS to capture the diversity expressed in web collections. Bonn also expressed the importance of evaluating library collections in order to see if they fulfill the needs of their community of users. Similarly, we included the *target audience* metric in order to estimate the audience a collection targets. Table 1 shows the CCS metrics derived from transforming library collection strength evaluation metrics.

Many solutions offered by libraries for quantifying collection strength can be summarized into two broad categories: collection-centered and use-centered [25]. 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 [15]. At web scale, a gold standard is often absent, but the collection-centered bibliographical set comparison practice informs our CCS *distribution of sources*, which reports the sources (hosts) that were sampled to build a collection. We believe the use-centered metric is a useful metric for approximating the exposure of a collection, which might approximate the popularity of the collection. Consequently, our CCS includes two metrics inspired by the use-centered metric for evaluating collection strength - *archival rates* and *tweet index rates*.

Risse et al. [32] surveyed social scientists, historical scientists, and legal experts in order to extract the requirements they find desirable for building collections. Some of the needs include topical dimension and time dimension, and the need to crawl social media sites. Topical dimension refers to the need to chronicle the evolution of an event over time. Consequently, our CCS includes a metric, *distribution of topics*, which gives insight about the various topics discussed in the collection. The time dimension is related to the topical dimension, but addresses the need to capture documents as events unfold. Some real world events have well-defined times e.g., a sports event and elections. Archivists often need the crawl duration to encompass the real world event time frame. Inspired by the time dimension metric, we added the *publication temporal* and *content temporal distribution* metrics to the suite. Social media is increasingly where the first reports of many events such as protests and popular uprising unfold, consequently, the CCS includes a *social media rate* as part of the broader *source diversity* metric for quantifying the amount of social media sources found in the collection.

3 BOOTSTRAPPING ARCHIVED COLLECTIONS FROM SOCIAL MEDIA

Archived collections begin with a list of URIs, or seeds, that share a common set of topics. The seeds are subsequently crawled to discover more URIs. We believe archived collections can be started or augmented by adding URIs extracted from social media collections from Storify, Reddit, Twitter, and Wikipedia.

Storify is a social media curation service that enables users to create *stories* which consist of hand-selected web resources such as URIs of news articles, images, videos, etc. Unfortunately, Storify is scheduled to go out of service in May 2018 [36], but we are exploring other possible alternatives [19]. We can create a seed list by extracting the URIs from storify *stories* that are relevant to a collection topic. Twitter Moments is a service by Twitter that lets users create topical collections of tweets that may embed URIs and multimedia content. In addition to extracting URIs from the tweets in Twitter Moments collections, we can also generate collections automatically by searching Twitter for tweets related to a topic and extracting URIs from the tweets returned by the Twitter SERP (Search Engine Result Page). 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 can be added to a seed list. 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 2 consists of a sample of URIs extracted from the references of the Wikipedia document [38] about the *Ebola virus* event. A seed list for an archived collection can be generated with URIs extracted from the references of Wikipedia documents [21].

4 COLLECTION CHARACTERIZING SUITE

The CCS provides a means of characterizing individual collections and comparing multiple collections. The various metrics that make up the CCS can be instantiated in different ways - it is a template. Consequently, the main criteria considered for instantiating the various metric was generality.

4.1 Distribution of topics

A "topic" is informally defined as a group of words which frequently occur together. It provides a means to summarize collections and gives us some notion of what the collection is about. It is impractical to manually inspect all the web pages, especially for large collections, in order to discern aboutness, therefore, we need this measure to summarize collections. The *distribution of topics* is 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. A probabilistic language model assigns probabilities to a sequence of words that make up a topic. One goal of a language model is the assignment of high probabilities to frequent topics (or sentences) in a collection. Similarly, we adopted a variant of the n-gram language model. Since collections are organized around specific topics, web pages in the collection include these topics frequently in their vocabulary. For example, we would expect a collection about *sports* events to possess sports vocabulary, e.g., *football*, *basketball*, etc. Inspired by this characteristic of collections, we developed a method to derive the topical distribution of a collection by finding the n-grams in the collections with the highest frequency of occurrence in the collection. The method is described by Algorithm 1 and sample outputs are given in Table 6. Algorithm 1 leads to the possibility of splitting compound word n-grams. For example, given an *Ebola virus* collection, if we choose $n = 2$ to generate bigram topic distributions, it could result in a ranked list that includes "centers disease"

Algorithm 1 : Generate a distribution of n-grams (topics)

```
Input: A collection $C$ of web pages ($|C| = N$), integers $n > 0$, & $m > 0$.
Output: A ranked list of $m$ $n-$grams (topics); the $n-$grams with the highest frequencies at the top of the list.
function GenTopicDist($C$, $n$, $m$)
    0. Represent each document $d_i \in C$ as a $n-$gram document
    1. Create a vocabulary vector $V \in \mathbb{Z}^{1\times p}$, each entry $v_i$ in $V$ represents a unique $n-$gram from $C$ (with $p$ unique $n-$grams).
    2. Create a binary document term matrix $\boldsymbol{M} \in \mathbb{Z}^{N\times p}$. Each row in $\boldsymbol{M}$ represents a document $d_i \in C$, and each column has 1 if $v_i \in d_i$, and 0 otherwise.
    3. Create a ranked list $L$. Populate $L$ ($|L| \leq m$) with $n-$grams ($v_i$) with the highest frequencies of occurrence in $\boldsymbol{M}$ ($\max_{v_i \in V} \sum_{j=1}^{N} m_{j,i}$).
    Populate $L$ with $v_i$ in decreasing order of their frequencies.
    return $L$
end function
```

and "disease control". It is clear that both terms are part of the compound word (trigram) "centers disease control" (stopwords are removed). To solve this problem, we replace multiple lower-order (e.g., bigram) n-grams with their superset higher-order (e.g., trigram) n-grams.

4.2 Distribution of sources

Given a collection of web pages, the *distribution of sources* is a statistical summary of the various sources sampled in order to build the collection. For example, the NLM Archive-It *Ebola virus* collection consists of 18 (12.5%) web pages from *blogs.plos.org*, 14 (9.7%) from *cdc.gov*, and 11 (7.6%) from *twitter.com*. We may conclude that these are the three most influential sources in the collection.

The distribution of sources is instantiated with a simple enumeration of the frequencies of the various hosts that make up a collection. In order to make the description more compact, we chose to report the top 10 hosts that make up a collection, and what proportion of the collection the top 10 hosts account for. For example, the top 10 hosts in the NLM Archive-It *Ebola Virus* collection make up 50% of the collection.

4.3 Content diversity

Given a collection of web pages, the *content diversity* is defined as the degree of self-similarity of the content of the web pages in the collection. For example, if we sample a collection about a *shooting* event one hour after the event, we should expect a high degree of similarity in the web pages. Most of them are expected to report the location of the shooting, the casualty count, possible identity of the perpetrators, etc. However, one year after the event, we may see more diverse content, perhaps discussing the shooting in context to other shootings. The diversity of the content of such events increases with time.

The content diversity is a single metric which summarizes the degree of self-similarity of a collection. A diversity score of 0 means no diversity - duplicate web documents, and a diversity score of 1 means maximum diversity - mutually orthogonal vocabulary of documents.

The input to calculate a content diversity for an arbitrary collection is a *similarity matrix* $\boldsymbol{D}$. The *similarity matrix* consists of the pairwise similarity of the web documents in the collection. We propose two ways of calculating the similarity between a pair of web pages corresponding with the two different ways of representing a collection. First, a collection may be represented as a *Document-Term matrix*: each row represents a document (web page), each column represents the TF or TFIDF value of a unigram in the collection vocabulary. In this representation, the similarity between a pair of documents is the cosine similarity measure. Second, a collection may be represented as a *List of Entity sets*: each document is represented as a set of entities of proper nouns for (people, location, organization, time, date, money, percent, and misc). The entities were extracted using the Stanford Named Entity Recognition System [12]. In this representation, we defined a new similarity measure - weighted Jaccard-Overlap similarity (Eqn. 1) measure to calculate the similarity between a pair of web documents, with a Jaccard weight ($\alpha \in [0, 1]$) of 0.4.

The weighted Jaccard-Overlap similarity $sim(A, B)$ between a pair of documents sets A and B is given by Eqn. 1, where β is the coefficient of similarity, defining the threshold two documents must reach to be considered similar. This threshold was empirically derived from a gold-standard dataset and set to 0.27.

$$sim(A, B) = \begin{cases} 1 & \text{, if } \alpha.J(A, B) + (1 - \alpha).O(A, B) \geq \beta \\ 0 & \text{, otherwise} \end{cases} \tag{1}$$

$J(A, B)$ is the Jaccard index of both documents, $J(A, B) = \frac{|A \cap B|}{|A \cup B|}$, and $O(A, B)$ is the Overlap coefficient of both documents, $O(A, B) = \frac{|A \cap B|}{min(|A|, |B|)}$.

Let a *similarity matrix* of *n* web pages in a collection be represented by $\boldsymbol{D} \in \mathbb{R}^{n\times n}$, and an *all-ones matrix* $\boldsymbol{O} \in \mathbb{R}^{n\times n}$. Given a square matrix, $\boldsymbol{N} \in \mathbb{R}^{n\times n}$, with zeros on the main diagonal and ones everywhere else, for example, if $\boldsymbol{N} \in \mathbb{R}^{3\times 3}$,

$$\boldsymbol{N} = \begin{bmatrix} 0 & 1 & 1 \\ 1 & 0 & 1 \\ 1 & 1 & 0 \end{bmatrix}, \text{the content diversity score } d_c = 1 - \frac{||\boldsymbol{ND}||_F}{||\boldsymbol{NO}||_F}$$

where $||\boldsymbol{A}||_F$ is the Frobenius norm: $||\boldsymbol{A}||_F = \sqrt{\sum_{i=1}^{m} \sum_{j=1}^{n} |a_{i,j}|^2}$ Web documents consist of topics (groups of words that frequently occur together). This means multiple words that belong to the same topic tend to co-occur. We may not always consider our collection diverse by the mere presence of different words, especially if these words belong to the same topic. Instead, we may consider our collection diverse if it consists of different topics. Consequently, if we consider unigrams, we would reward diversity to different terms which occur together, even though they may belong to the same topic, i.e., no new information. The *Document-Term matrix* representation rewards diversity at the term level, while the *List of Entity sets* representation rewards diversity at the topic level.

4.4 Temporal distribution

The *publication temporal distribution* is an aggregation of publication dates that are used to timestamp web pages. The *content temporal distribution* is the collection of time references associated with events being discussed on web pages. The time information may be absolute, (e.g., "On Friday, Nov 17, 2017...") or relative (e.g., "Next month is..."). We normalize relative time information (e.g., if the reference date is "2017-11-17" we represent "next month" as "2017-12-17"). Temporal distributions enable the calculation of the collection age. The ages of web pages may be calculated with respect to the creation date of the collection to indicate how long web pages existed prior to being collected. A short duration between the publication date of web pages and the creation date of the collection may indicate that the curator intended to collect web pages following a recent event. Alternatively, the ages of documents may be calculated with respect to the current date to determine absolute ages of web pages.

The publication dates of web pages may provide useful information about the kinds of events discussed in the document. For example, stories concerning airport security before the September 11, 2001 terrorist attacks are not expected to discuss the TSA (Transportation Security Administration), because the TSA was founded on November 19, 2001. The publication date alone may not be sufficient to give us a full picture of the kinds of events discussed in a document, since documents often discuss events and include the dates of these events in their content. This may be relative, e.g., "last year" or absolute "on Jan 3rd, 2017." Therefore, we also have to pay attention to these dates.

We extract the publication dates of the documents in a collection to form the publication date distribution through the use of CarbonDate [33] which estimates the creation date of web pages based on information polled from multiple sources such as the document timestamps, web archives, Twitter, backlinks, etc. We extract the content dates with the aid of SUTime [8].

NLM (occurrence rate)	Reddit (occurrence rate)
"ebola outbreak west africa" (0.34)	"infected ebola virus disease" (0.25)
"guinea liberia sierra leone" (0.31)	"west africa" (0.21)
"cases ebola virus disease" (0.30)	"public health workers" (0.15)
"public health workers" (0.27)	"sierra leone" (0.15)
"centers disease control prevention" (0.15)	"united states" (0.14)

(a) Distribution of top five topics for NLM Archive-It and Reddit *Ebola virus* collections showing a similar topic distribution.

CCS Metric	NLMś Ebola Characterization	Reddit Ebola Characterization
Dist. of sources	Top 10 hosts fraction of collection: 50%	Top 10 hosts fraction of collection: 46%
Content diversity (Doc-Term matrix / Entity set)	(0.80 / 0.65)	(0.89 / 0.85)
Publication temporal dist. (Median age, where age: Creation date - Pub. date)	36 days	1,450 days (3.9 years)
Content temporal dist. (Median age)	1,144 days (3.1 years)	2,104 (5.8 years)
Source diversity (URI/ Hostname / Social media)	(1.0 / 0.34 / 0.07)	(0.98 / 0.53 / 0.12)
Collection exposure (Archival rate/ Tweet index rate)	(1.00 / 0.72)	(0.78 / 0.40)
Target audience (readability, Q1 / Median / Q3)	(0 / 0.57 / 1)	(0.14 / 0.57 / 0.85)

(b) CCS characterizations of NLM and Reddit *Ebola virus* collections

Table 3: Characterization of two collections Archive-It (144 URIs) and Reddit (150 URIs) *Ebola virus* collections. Each characterization describes the individual collection, juxtaposing multiple characterizations enables collection comparison.

4.5 Source diversity

Similar to content diversity, the *source diversity* metric tells us whether a collection samples a single source, a handful of sources, or many sources. The URI source diversity metric [28], $d_{URI} \in [0, 1]$ tells us the rate of unique URIs; $d_{URI} = 0$ means the collection only has one distinct URI (duplicate web pages). On the other hand, if $d_{URI} = 1$, it means the collection is made up of unique URIs. We also explore source diversity at the domain (d_{domain}) and hostname ($d_{hostname}$) policies.

We deduplicated URIs in collections by trimming all parameters from the URIs as suggested by Brunelle et al. [5] before calculating source diversity. Given a policy set $P = \{URI, Domain, Hostname\}$ for a collection C, and the count of unique URIs in the collection U, the source diversity of a given policy d_p is given by Eqn. 2.

$$d_{p \in P} = \frac{U}{|C|}; d_p \in [\frac{1}{|C|}, 1] \tag{2}$$

The normalized source diversity of a given policy d'_p is given by Eqn. 3.

$$d'_{p \in P} = \frac{d_p - \frac{1}{|C|}}{1 - \frac{1}{|C|}} = \frac{U - 1}{|C| - 1}; d'_p \in [0, 1] \tag{3}$$

The social media diversity metric or social media rate quantifies the proportion of web pages in a collection that are from social media sites. We created a predefined list of social media domains: *twitter.com*, *facebook.com*, *youtube.com*, *instagram.com*, and *tumblr.com*. Given k URIs from social media domains in a collection C, the social media rate is $\frac{k}{|C|}$. For example, a collection composed of 3 URIs from Twitter, 2 from Facebook, and 5 from CNN, has a social media rate of $\frac{3+2}{10} = 0.5$.

4.6 Collection exposure

If a web page is "popular" (used widely), this means there is some need the document fulfills to a wide audience. We approximate popularity with the *collection exposure* metrics - *archival rate* and *tweet index rate*. In our previous work [31], we showed that collections of local news from local news organization, such as the *Caloosa Belle newspaper* (LaBelle, Florida USA), are less exposed, thus less popular than collections of news sources from mainstream news organizations, such as *CNN* and *The Washington Post*.

The archival rate of a collection C is the fraction of C that is archived. For example, if we found 10 archived stories from C (where $|C| = 50$), the archival rate of C is $\frac{10}{50} = 0.2$. Note that when comparing the archival rates of two collections, it is important to consider how old both collections are. For example, a collection A might have a much larger archival rate than a collection B only because A has much older documents than B, and as a result had the greater opportunity to be archived.

Popular (widely used) URIs are more likely to be archived than less popular URIs [1]. This means we could use the archival state of a URI to infer its popularity. This method will not be valid if every URI is archived (e.g. Archive-It seeds). If this were the case (all URIs archived), the magnitude of archived copies of a URI may indicate its popularity. The archive state of a web page can be measured using Memgator [2].

Similar to the archival rate, the tweet index rate of a collection C is the fraction of C found embedded in tweets. For example, if we found 40 URIs from C (where $|C| = 50$) embedded in tweets, the tweet index rate of C is $\frac{40}{50} = 0.8$. Also similar to archival rate, when comparing the tweet index rates of two collections, it is important to consider how old both collections are. For example, a collection A might have a much larger tweet index rate than a collection B only because A includes web pages that are much older than B, and as a result, had a greater opportunity to be tweeted. The tweet index state of a web page is set by searching Twitter for a tweet that embeds the page URI [27].

Similar to the archival rate, popular URIs are more likely to be shared on social media sites (e.g., Twitter) than less popular URIs. Consequently, the tweet index state (in tweet or not) of a web page may indicate the popularity or exposure of the web page. We may also be able to infer the popularity of a URI in a tweet by taking into account how often it is shared on Twitter. The tweet index rate is often a useful alternative to the archival rate when the collections to be compared have the same archival rate. For example, Archive-It seeds have a 100% archival rate. Likewise the archival rate provides an alternative when comparing collections with the same tweet index rates, for example, collections generated from Twitter have 100% tweet index rates.

4.7 Target audience

The *target audience* estimates the target users of the collection. This is not easy to achieve. Our premise is that the readability level of the documents in the collection is a reflection of the target audience. For example, if the reading level of a collection is at the 10th grade level, we conclude that the target audience starts from high school young adults and above. However, if the reading level is at the graduate level (16th grade) level, we may conclude the target audience might be professionals in a subject area.

The target audience of a collection provides important contextual information that may give insight about the composition of the collection, and may reflect the intent of the collection builder, such information is not often readily available.

We instantiate the target audience metric with readability measures. Readability measures estimate the reading level of documents through procedures that include counting syllables, words, and sentences. We employed widely used readability measures that output grade levels. These are the *Flesch-Kincaid Grade level* [20], *Coleman Liau index* [9], and the *Automated Readability index* [35]. For a single document, the readability score is the average score from the three readability measures (normalized between 0 and 1). The higher the readability score, the higher the grade level.

5 COLLECTION CHARACTERIZATION AND COMPARISON

In order to characterize a single collection with the CCS, we simply extract values for the metrics that make up the suite. These values collectively form a characterization for the collection. For example, Table 3 describes two collections. The first, the NLM Archive-It *Ebola Virus* collection, is an archived collection built manually by an archivist at the NLM in October 2014. The second, the Reddit *Ebola Virus* collection, we built by issuing the query "ebola virus" to Reddit from *2017-07-25* to *2017-08-23* and extracting links from the Reddit SERPs and their respective comments. Let us consider both collections to see how the CCS describes both collections.

The top five topics from the NLM Archive-It collection show that the collection addresses issues arising from the Ebola virus outbreak in West Africa (Table 3a, topic 1) and that the main countries affected were Guinea, Liberia, and Sierra Leone (Table 3a, topic 2). Also two major players involved with the outbreak were public health workers and the Centers for Disease Control and Prevention (Table 3a, topic 4 & 5). The Reddit collection also mirrors this sentiment. Both collections are similarly characterized by the fraction of the collections the top 10 hosts make (Table 3b, Dist. of sources). Similarly, both collections target a similar audience (Table 3b, Target audience) since they have the same median normalized grade level of 0.57 (11th grade).

Table 4: Evaluation Dataset comprised of 129 collections from three Topics: "Ebola Virus," "Hurricane Harvey," and "2016 Pulse Nightclub shooting." WSDL represents the collections generated by the authors.

ID	Topic (URI Count)	Source (Author)	Creation Date	Extraction note
1	Ebola V... (144)	(NLM)	2014-10	Archive-It seeds
2	Ebola V... (669)	(WSDL)	2017-11-29	100 sub-collections (IDs 0-99) of URIs from 100 Storify *stories*
3	Ebola V... (669)	(WSDL)	2017-11-29	A Collection created by combining all links in Collection 2.
4	Ebola V... (155)	W (WSDL)	2017-07-25	URIs from references of *Ebola Virus* Wikipedia page
5	Ebola V... (153)	(WSDL)	2017-07-25 - 2017-08-23	URIs from Reddit (& comments) search for query: "Ebola Virus"
6	Ebola V... (152)	(WSDL)	2017-08-02 - 2017-11-28	URIs in tweets from Twitter search for query: "Ebola Virus"
7	Ebola V... (105)	G (WSDL)	2017-11-29	URIs from first 10 pages of Google, for query: "Ebola Virus"
8	Hurricane H...(44)	(IA)	2017-08	Archive-It seeds
9	Hurricane H...(151)	W (WSDL)	2017-09-02	URIs from references of *Hurricane Harvey* Wikipedia page
10	Hurricane H...(14)	(WSDL)	2017-12-08	2 sub-collections (IDs 0-1) of URIs from tweets in Twitter Moments
11	Hurricane H...(14)	(WSDL)	2017-12-08	A collection created by combining all URIs in Collection 10
12	Hurricane H...(94)	G (WSDL)	2017-09-02- 2017-11-29	URIs from first page of Google, for query: "Hurricane Harvey"
13	2016 Pulse...(151)	(IA)	2016-06	Archive-It seeds
14	2016 Pulse...(50)	(WSDL)	2017-12-08	5 sub-collections (IDs 0-4) of URIs from tweets from Twitter Moments
15	2016 Pulse...(50)	(WSDL)	2017-12-08	A collections created by combining URIs in Collection 14
16	Random (500)	UCI ML (Lichman, M)	2014-03-10 - 2014-08-10	10 sub-collections (IDs 0-9) of URIs or random news stories
Total	2,765 URIs			

Table 3b shows that the Reddit collection produced a higher content diversity for both collection representations (*Document-Term matrix* and *List of Entity sets*). The NLM Archive-It collection produced much newer web documents with a median publication age of 36 days, compared to the Reddit collection of 3.9 years. This suggests that the NLM Archive-It collection was created a few months after the Ebola event unfolded. Additionally, the Reddit collection sampled from more hosts (hostname source diversity - 0.53) and had more social media URIs (social media rate - 0.12) compared to the NLM Archive-It colleciton (hostname source diversity - 0.34, social media rate - 0.07). The NLM Archive-It collection indicated a higher exposure than the Reddit collection, with a higher archival rate of 1.0, compared to the 0.78 archival rate of the Reddit collection. The high archival rate of the Archive-It collection is no surprise because it is a collection of seeds; the seeds are meant to be crawled and archived. The NLM Archive-It collection also showed a higher tweet index rate (0.72) than the Reddit collection (0.40).

Table 5: List of collections most similar to three Archive-It collections and three random collections for the evaluation dataset topics.

Gold standard collections	Three most similar		
Ebola Virus (ebo.) ARCHIVE-IT $^{ebo.1}_{0}$	$^{ebo.5}_{0.17}$	S $^{ebo.2.49}_{0.23}$	S $^{ebo.2.57}_{0.25}$
Hurricane Harvey (hur.) ARCHIVE-IT $^{hur.8}_{0}$	G $^{hur.12}_{0.27}$	⚡ $^{hur.11}_{0.32}$	⚡ $^{pul.15}_{0.34}$
2016 Pulse night.. (pul.) ARCHIVE-IT $^{pul.13}_{0}$	⚡ $^{pul.15}_{0.24}$	⚡ $^{pul.14.2}_{0.24}$	⚡ $^{pul.14.4}_{0.31}$
Random news stories 0 (ran.) UCI ML $^{ran.16.0}_{0}$	UCI ML $^{ran.16.8}_{0.16}$	UCI ML $^{ran.16.5}_{0.19}$	UCI ML $^{ran.16.4}_{0.22}$
Random news stories 1 (ran.) UCI ML $^{ran.16.1}_{0}$	UCI ML $^{ran.16.6}_{0.19}$	UCI ML $^{ran.16.3}_{0.22}$	UCI ML $^{ran.16.8}_{0.22}$
Random news stories 2 (ran.) UCI ML $^{ran.16.2}_{0}$	UCI ML $^{ran.16.3}_{0.22}$	UCI ML $^{ran.16.4}_{0.22}$	UCI ML $^{ran.16.5}_{0.22}$

6 EVALUATION

To assess if we could bootstrap archived collections from social media, we measured the distances between archived collections from Archive-It (ARCHIVE-IT) and collections generated from social media sources: Storify (S), Reddit (), Twitter Moments (⚡), Twitter SERP (), and Wikipedia (W). The rationale for this is if collections created by extracting URIs from social media collections are similar (low distance) to expert-created collections on Archive-It, then we may start or augment archived collections with seeds extracted from social media sources.

We generated a dataset (Table 4) of 129 collections (2,765 URIs) from three topics: "Ebola Virus," "Hurricane Harvey," and "2016 Pulse Nightclub Shooting," and 10 collections (500 URIs) for random (multiple topics) news stories from the UCI news aggregator dataset [23]. Random collections (UCI ML) were included to assess if the CCS resulted in clusters of collections of common topics even in the presence of noise. We do not expect collections of random news stories to be more similar to archived collections than social media collections. Additionally, we included baseline collections generated by extracting URIs from Google (G). We believe most users primarily use Google to discover candidate URIs for their collections, so we included Google collections in order to quantify how these compare with social media and archived collections. Our previous work [30] showed that such collections change with time since search engines are biased to produce the latest documents.

Table 6: Dist. of top five Topics for Archive-It Collections.

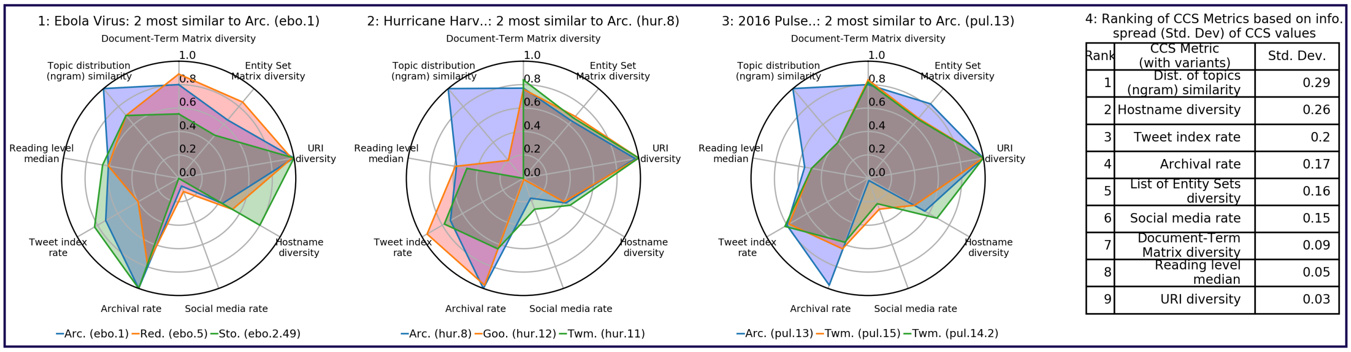
2016 Pulse Nightclub Shooting	Hurricane Harvey
"pulse nightclub orlando florida"	"hurricane harvey photo"
"new york"	"27 2017 houston"
"en la comunidad"	"27 2017 photo"
"mass shooting"	"tropical storm harvey photo"
"omar mateen"	"corpus christi"

The evaluation dataset collections were represented as a vector of CCS values, and a distance was calculated between Archive-It collections (Table 4, IDs 1, 8, and 13) and every other collection irrespective of the topics. The Euclidean distance metric was used (as opposed to cosine) to compute distance because the magnitudes of the respective CCS values in the collection vectors are significant. We normalized (0-1) the Euclidean distances since all possible maximum and minimum CCS values are known. Additionally, the CCS metrics were assessed to identify the metrics which provided the most information in distinguishing the collections. This was done by calculating the spread of values (standard deviation) of the individual CCS metrics for the collections.

We generated a CCS matrix for the evaluation dataset collections. The rows of the CCS matrix represented the collections and the columns represented the CCS metric values. The first and second columns represented the content diversity values calculated with the Document-Term matrix and List of Entity sets collection representations, respectively. The third column represented the URI source diversity, fourth - domain source diversity, fifth - hostname diversity, sixth - social media rate, seventh - collection exposure archival rate, eight - collection exposure tweet index rate, and ninth, the Jaccard similarity score of a given collection's top 10 n-gram distribution of topics to the Archive-It collection. The last column of the CCS matrix represented the normalized median reading level of the collection. Section 4 outlines how to extract the CCS metrics of all the entries, except the Jaccard similarity of the n-gram distribution of topics for two collections. The idea for this method is to find how similar two collections are in terms of their respective n-gram distribution of topics. In other words, if the collections are about a similar set of topics. We focused on finding similar collections based on the content of the collection and not the sources they sample from or the time the collection was built. Consequently, we excluded the distribution of sources and temporal distributions from the CCS vector.

7 RESULTS AND DISCUSSION

Each pictogram in Table 5 represents a collection expressed by an image of the collection source (section 6). The pictogram superscript represents the collection topic abbreviation followed by the collection ID (Table 4). The sub-collection ID follows the collection ID for Storify and Twitter Moments sub-collections. The subscript represents the normalized Euclidean distance of the collection to the specified Archive-It collection. For example, for the *Ebola Virus* topic, the Reddit (*ebo.5*) collection has the closest distance (0.17) to the Archive-It (*ebo.1*) collection.



4: Ranking of CCS Metrics based on info. spread (Std. Dev) of CCS values

Rank	CCS Metric (with variants)	Std. Dev.
1	Dist. of topics (ngram) similarity	0.29
2	Hostname diversity	0.26
3	Tweet index rate	0.2
4	Archival rate	0.17
5	List of Entity Sets diversity	0.16
6	Social media rate	0.15
7	Document-Term Matrix diversity	0.09
8	Reading level median	0.05
9	URI diversity	0.03

Figure 1: Distribution of CCS Metrics for pair of collections most similar to Archive-It collections (Chts. 1-3) and ranking of CCS Metrics based their respective informational values (Cht. 4).

Table 5 shows that the CCS resulted in the clustering of collections of similar topics with a distance ranging from 0.17 to 0.34 across all topics. The Reddit collection ($^{ebo.5}_{0.17}$) was most similar to the Archive-It *Ebola Virus* collections ($^{ebo.1}_{0}$). Since we had more Storify collections in our dataset, the Storify collections have a higher opportunity of outperfoming (lowest distance) other collections. In fact, the Storify *Ebola Virus* collection ($^{ebo.3}_{0.27}$) is 4.3 times the size of the Reddit collection, yet, the Reddit collection was most similar to the Archive-It collection. This suggests that the larger the collection may not always mean the better the collection. This result is potentially consequential: it suggests that we may consider Reddit as a collection source in the absence of Storify. The Google *Hurricane Harvey* collection ($^{hur.12}_{0.27}$) was most similar to the Archive-It *Hurricane Harvey* collection confirming our expectation that collections generated from Google may be similar to social media collections since users may use Google to discover URIs. The Twitter Moments *2016 Pulse nightclub shooting* collection ($^{pul.15}_{0.34}$) was third most similar even though it has no topics in common with the Archive-It *Hurricane Harvey* collection (n-gram topic similarity of 0), indicating a strong similarity across other dimensions. This shows the need for taking topic similarity into consideration before collection comparison. Similarly, the Twitter Moments collections were most similar to the Archive-It *2016 Pulse nightclub shooting* collections.

Random collections were most similar to other random collections due a common set of properties random collections show: all the random collection produced high diversity values for *Document-Term matrix* (0.93 - 0.95) and *List of Entity sets* (0.88 - 1.0) representations. Also, they included no social media sources (social media rate - 0.0) and sampled from a diverse set of hosts (hostname diversity between 0.92 - 0.77).

Across the various topics, the distribution of topics (ngram similarity) CCS metric provided the most information to distinguish the collections, producing the highest variance or spread ($\sigma = 0.29$) across the collections (Fig. 1, Chrt 4). The radar plots (Fig. 1, Chrt 1-3) illustrates this variance. This suggests the importance of collection summaries in distinguishing collections. This was followed by the hostname diversity CCS metric ($\sigma = 0.26$), suggesting multiple ways collections sample hosts. The target audience (readability) and URI diversity provided the least information to distinguish the collections: this may be explained by the idea that the documents in the collection target a common audience and have little or no duplicate links ($d_{URI} = 1$).

8 FUTURE WORK AND CONCLUSIONS

We believe the CCS of seven metrics can be expanded. Any new metric has to provide valuable information to a wide range of users since there are many measures one can easily extract from a collection.

To begin or augment expert-generated archived collections of web pages, we proposed extracting URIs from collections created by users on social media sites: Storify, Reddit, Twitter, and Wikipedia. This required us to assess the degree of similarity of the collections generated from social media sources and archived collection. To achieve such comparison, we developed a suite (CCS) of seven metrics that characterized individual collections. Multiple collections can be compared by computing the similarity or distances with respect to a given collection. The CCS metrics included widely used metrics motivated by the state of the art in collection evaluation such as distribution of topics, content diversity, and publication temporal distribution. We also provided and motivated additional metrics such as distribution of sources, content temporal distribution, source diversity, collection exposure, and target audience. The metrics provide valuable information such as a summary of the collection that indicates whether the collection is on topic and the degree of self similarity of the collection. We consider our collection characterizing suite as a template, and as such, may be realized in different ways, and provided novel options for instantiating the metrics. The CCS distance evaluation results showed that Archive-It collection and social media collections are similar with a distance ranging from 0.17 to 0.34, suggesting that we can start or augment the seed generation process of important events by extracting URIs from social media collections. Our evaluation dataset as well as source code that implements instantiations of the CCS are publicly available [29].

ACKNOWLEDGEMENTS

This work was made possible by IMLS LG-71-15-0077-15, and help from Christie Moffat at the National Library of Medicine.

REFERENCES

[1] Scott G Ainsworth, Ahmed Alsum, Hany SalahEldeen, Michele C Weigle, and Michael L Nelson. 2011. How much of the web is archived?. In *Proceedings of JCDL 2011*. 133–136.

[2] Sawood Alam and Michael L Nelson. 2016. MemGator-A portable concurrent memento aggregator: Cross-platform CLI and server binaries in Go. In *Proceedings of JCDL 2016*. 243–244.

[3] Donna Bergmark. 2002. Collection synthesis. In *Proceedings of JCDL 2002*. 253–262.

[4] George S Bonn. 1974. Evaluation of the Collection. *Library Trends* 22, 3 (1974), 265–304.

[5] Justin F Brunelle, Michele C Weigle, and Michael L Nelson. 2015. Archiving Deferred Representations Using a Two-Tiered Crawling Approach. *Proceedings of iPRES 2015* (2015).

[6] Centers for Disease Control and Prevention. 2016. 2014 Ebola Outbreak in West Africa - Case Counts. https://www.cdc.gov/vhf/ebola/outbreaks/2014-west-africa/case-counts.html.

[7] Soumen Chakrabarti, Martin Van den Berg, and Byron Dom. 1999. Focused crawling: a new approach to topic-specific Web resource discovery. *Computer Networks* 31, 11 (1999), 1623–1640.

[8] Angel X Chang and Christopher D Manning. 2012. Sutime: A library for recognizing and normalizing time expressions. In *LREC*. 3735–3740.

[9] Meri Coleman and Ta Lin Liau. 1975. A computer readability formula designed for machine scoring. *Journal of Applied Psychology* 60, 2 (1975), 283–284.

[10] Mohamed MG Farag, Sunshin Lee, and Edward A Fox. 2018. Focused crawler for events. *International Journal on Digital Libraries (IJDL)* 19, 1 (2018), 3–19. DOI: http://dx.doi.org/10.1007/s00799-016-0207-1

[11] 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.

[12] Jenny Rose Finkel, Trond Grenager, and Christopher Manning. 2005. Incorporating Non-local Information into Information Extraction Systems by Gibbs Sampling. In *Proceedings of Association for Computational Linguistics (ACL 2005)*. 363–370.

[13] Gerhard Gossen, Elena Demidova, and Thomas Risse. 2015. iCrawl: Improving the freshness of web collections by integrating social web and focused web crawling. In *Proceedings of JCDL 2015*. 75–84.

[14] Nancy E. Gwinn and Paul H. Mosher. 1983. Coordinating Collection Development: The RLG Conspectus. *College & Research Libraries* 44, 2 (1983), 128–140.

[15] Terese Heidenwolf. 1994. Evaluating an interdisciplinary research collection. *Collection Management* 18, 3-4 (1994), 33–48.

[16] Internet Archive. 2016. Help build an archive documenting responses to the 2016 U.S. presidential election at. https://twitter.com/internetarchive/status/797263535994613761.

[17] Internet Archive. 2016. What web pages should we save concerning DAPL? Tell us here:. https://twitter.com/internetarchive/status/806228431474028544.

[18] Internet Archive Global Events. 2016. 2016 Pulse Nightclub Shooting Web Archive. https://archive-it.org/collections/7570.

[19] Shawn M Jones. 2017. Where Can We Post Stories Summarizing Web Archive Collections? http://ws-dl.blogspot.com/2017/08/2017-08-11-where-can-we-post-stories.html.

[20] J Peter Kincaid, Robert P Fishburne Jr, Richard L Rogers, and Brad S Chissom. 1975. *Derivation of new readability formulas (automated readability index, fog count and flesch reading ease formula) for navy enlisted personnel.* Technical Report Research Branch Report 8-75. Naval Technical Training Command Millington TN Research Branch.

[21] 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)*.

[22] David Lesniaski. 2004. Evaluating collections: a discussion and extension of Brief Tests of Collection Strength. *College & Undergraduate Libraries* 11, 1 (2004), 11–24.

[23] Moshe Lichman. 2013. UCI machine learning repository. http://archive.ics.uci.edu/ml.

[24] Local Memory Project. 2017. Ebola Virus Collection. http://www.localmemory.org/vis/collections/local-memory-project/queries/guinea-conakry-ebola-virus-10-2017-11-16.

[25] Barbara Lockett. 1989. *Guide to the evaluation of library collections.* American Library Association.

[26] National Library of Medicine. 2014. Global Health Events. https://archive-it.org/collections/4887.

[27] Alexander C Nwala. 2017. Finding URLs on Twitter - A simple recommendation. http://ws-dl.blogspot.com/2017/01/2017-01-23-finding-urls-on-twitter.html.

[28] Alexander C Nwala. 2018. An exploration of URL diversity measures. http://ws-dl.blogspot.com/2018/05/2018-05-04-exploration-of-url-diversity.html.

[29] Alexander C Nwala. 2018. Bootstrapping Web Archive Collections from Social Media - Git Repo. https://github.com/anwala/collection-characterizing-suite.

[30] Alexander C Nwala, Michele C Weigle, and Michael L Nelson. 2018. Scraping SERPs for Archival Seeds: It Matters When You Start. In *Proceedings of JCDL 2018*.

[31] 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 *Proceedings of JCDL 2017*. 219–228.

[32] Thomas Risse, Elena Demidova, and Gerhard Gossen. 2014. What do you want to collect from the web. In *Proceedings of Building Web Observatories Workshop (BWOW 2014)*.

[33] Hany M SalahEldeen and Michael L Nelson. 2013. Carbon dating the web: estimating the age of web resources. In *Proceedings of WWW 2013*. 1075–1082.

[34] 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.

[35] Edgar A Smith and RJ Senter. 1967. *Automated readability index.* Technical Report AMRL-TR-66-220. AMRL-TR. Aerospace Medical Research Laboratories (US).

[36] Storify. 2017. Storify End-of-Life. https://archive.is/DOPFa.

[37] Howard D White. 1995. *Brief tests of collection strength: A methodology for all types of libraries.* Number 88.

[38] Wikipedia. 2018. Ebola virus. https://en.wikipedia.org/wiki/Ebola_virus.

[39] Shuyi Zheng, Pavel Dmitriev, and C Lee Giles. 2009. Graph based crawler seed selection. In *Proceedings of WWW 2009*. 1089–1090.