

Local Memory Project

Providing tools to build collections of stories for local events from local sources

Alexander C. Nwala, Michele C. Weigle and
Michael L. Nelson
Old Dominion University
Norfolk, Virginia 23529, USA
{anwala,mweigle,mln}@cs.odu.edu

Adam B. Ziegler and Anastasia Aizman
Harvard Library Innovation Lab
Cambridge, Massachusetts 02138, USA
{aziegler,aaizman}@law.harvard.edu

ABSTRACT

The national (non-local) news media has different priorities than the local news media. If one seeks to build a collection of stories about local events, the national news media may be insufficient, with the exception of local news which “bubbles” up to the national news media. If we rely exclusively on national media, or build collections exclusively on their reports, we could be late to the important milestones which precipitate major local events, thus, run the risk of losing important stories due to link rot and content drift. Consequently, it is important to consult local sources affected by local events. Our goal is to provide a suite of tools (beginning with two) under the umbrella of the Local Memory Project (LMP) to help users and small communities discover, collect, build, archive, and share collections of stories for important local events by leveraging local news sources. The first service (Geo) returns a list of local news sources (newspaper, TV and radio stations) in order of proximity to a user-supplied zip code. The second service (Local Stories Collection Generator) discovers, collects and archives a collection of news stories about a story or event represented by a user-supplied query and zip code pair. We evaluated 20 pairs of collections, Local (generated by our system) and non-Local, by measuring archival coverage, tweet index rate, temporal range, precision, and sub-collection overlap. Our experimental results showed Local and non-Local collections with archive rates of 0.63 and 0.83, respectively, and tweet index rates of 0.59 and 0.80, respectively. Local collections produced older stories than non-Local collections, at a higher precision (relevance) of 0.84 compared to a non-Local precision of 0.72. These results indicate that Local collections are less exposed, thus less popular than their non-Local counterpart.

CCS CONCEPTS

•Information systems →Information retrieval;

KEYWORDS

Local news, Web Archiving, Digital collections, News, Journalism, Collections building

ACM Reference format:

Alexander C. Nwala, Michele C. Weigle and Michael L. Nelson and Adam B. Ziegler and Anastasia Aizman. 2017. Local Memory Project. In *Proceedings of Joint Conference on Digital Libraries, Toronto Ontario, Canada, June 2017 (JCDL '17)*, 10 pages.

DOI: 10.XXX/XXXX

Permission to make digital or hard copies of part or all of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for third-party components of this work must be honored. For all other uses, contact the owner/author(s).

JCDL '17, Toronto Ontario, Canada

© 2017 Copyright held by the owner/author(s). 123-4567-24-567/08/06...\$15.00

DOI: 10.XXX/XXXX

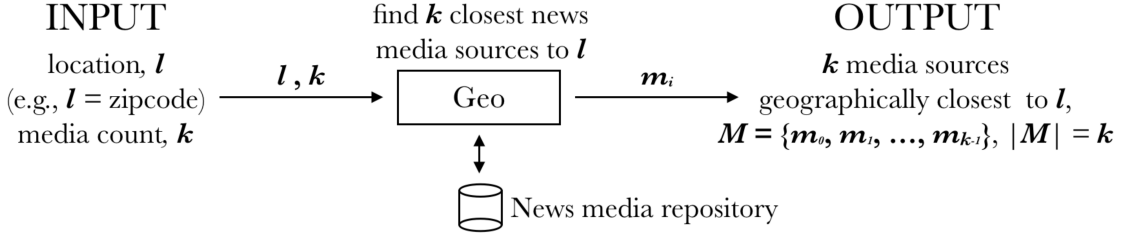
1 INTRODUCTION

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 Michigan Radio, The Flint Journal-MLive, and the local TV affiliates in Flint (WEYI, WJRT, WSMH, and WNEM) [18]. On May 23, 2014, Ron Fonger of Flint Journal-MLive reported [7] about complaints by city residents about the water’s taste and smell. Between August and September 2014, the city issued three boil advisories [18] to residents of Flint after finding fecal coliform bacteria (*E. coli*) in the water. Following multiple other incidents, Michigan Governor Rick Snyder declared a state of emergency for the city of Flint on January 5, 2016, due to dangerously high levels of lead contamination in the drinking water. A chain of events about the Flint water was reported by local media, but most of the non-local media did not report this crucial story until 2016 [18]. In fact, according to Denise Robbins’ report [18], Michigan Radio and The Flint Journal-MLive published over 500 articles before the State of Emergency was declared, and it took the national media one year after the *E. coli* outbreak to report the Flint story.

Important news stories begin in different ways. Some are explosive and have well defined timeframes - these are usually aggressively reported by the national news media. But some other important stories start slow and unsensational. The Flint disaster started small, was tracked by local media, but did not become a national story until it became a complete disaster. Similarly, in April 2016, activists joined Indian tribes in protest at the Standing Rock Indian reservation in North Dakota, but the national media was mostly absent at the beginning. If we rely exclusively on national media for building collections, we could be late to the important milestones which precipitate major events. Consequently, we run the risk of losing important content due to link rot and content drift. The national media deserves criticism [20] for the delay in reporting Flint, but any such criticism should be weighted by the idea that national and local news have different priorities. Local media such as the *Caloosa Belle newspaper* (LaBelle, FL) cover news stories that would not naturally be of interest to another locality, such as the annual Swamp Cabbage Festival. Non-local news organizations such as CNN cover stories of a broader (national/international) scope such as Obamacare and the Syrian refugee migrant crisis. Since the interests of local and non-local media often intersect, it is common for both media to cover the same stories. However, local media focuses on news that have a local impact. Also, it is well known that multinational news organizations routinely cite the reports of smaller local news organizations for many stories. Consequently, local news media is fundamental to journalism. Given the existential crisis of print media in the digital age that has seen the rapid decline of local media [13], measures such as the Local Memory Project which attempt to shed light on local media are pertinent.

We begin by presenting two services. First, our Geo service informs users about local media (newspaper, TV and radio) that serve a location

STAGE 1: NEARBY NEWS MEDIA DISCOVERY



STAGE 2: LOCAL STORIES COLLECTION BUILDING

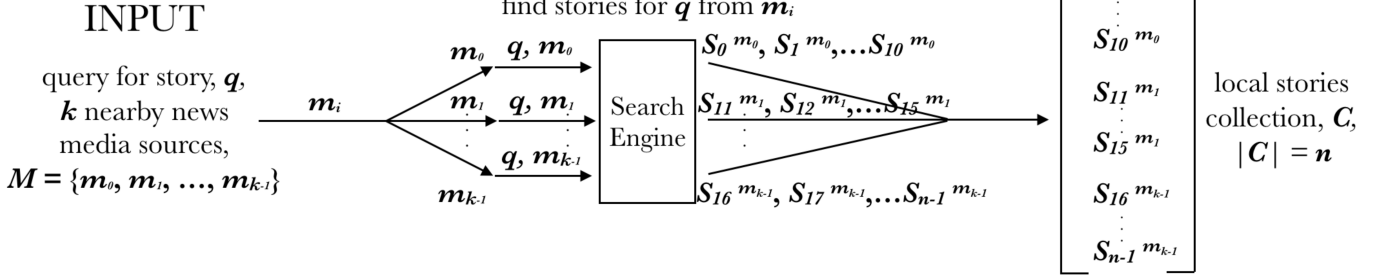


Figure 1: LMP’s Local Stories Collection Generator Architecture

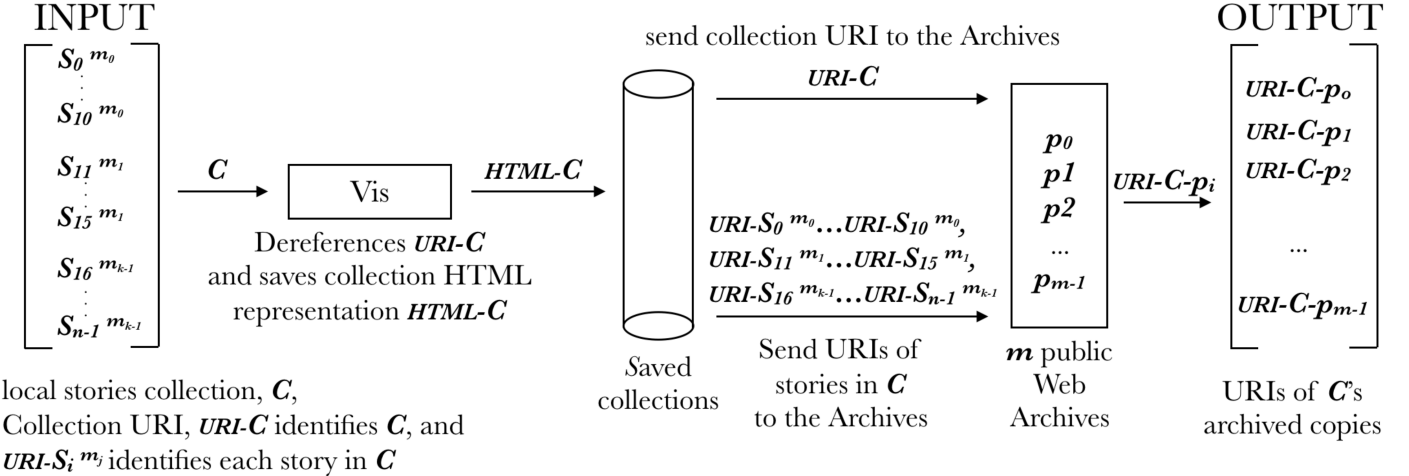


Figure 2: LMP’s Local Stories Collection Generator Archiving Architecture (Ideal). The current implementation does not include Render and sends $URI-C$ and $URI-S_i^{m_i}$ to one Web Archive - Archive.is)

(currently for US media). Second, our Local Stories Collection Generator builds high quality stories from local media sources by leveraging the Google Search Engine Result Page (SERP). Additionally, collections built locally may be archived in public web archives and shared on Twitter. We believe these open source tools could contribute to increase the exposure of collections from local news sources.

2 RELATED WORK

There are many efforts for building collections especially based on some variant of a focused crawler. Bergmark [3] introduced a method for building a collection by downloading webpages in a focused way and classifying them into the various topic areas. Similarly, Gossen et al. [8] proposed a focused crawler called iCrawl, which uses social media

content to discover fresh and relevant content, and Yang et al. [21] introduced a prototype system for building archives by using seed URIs collected from tweet collections. Qin et al. [17] proposed a meta-search enhanced focused crawling to address the problems of collection building by traditional crawlers. Traditional crawlers can be restricted to sub-graphs of the Web due to the limitations of local search. Also, Farag and Fox [6] proposed an Event Focused crawler, which collects relevant documents which fulfill the criteria specified by an event model. Blackburn et al. [4] emphasized the importance of longitudinal data collection for studies of social media and presented a means of collecting this dynamic social data from multiple sources.

The tools we propose under the Local Memory Project differ from the state of the art in four major ways. First, most of the tools proposed

by the state of the art target expert users and researchers based on their complexity of use. Our tools target average users. This is realized by implementing our collection building tools in form of a Google Chrome browser extension [11]. Expert curators can also benefit from our tools and the high quality collections generated through its use. Second, we do not process (i.e., attempt to understand) the user query and we do not crawl to discover relevant stories. Processing the user query is a challenging task and may require classifying the query. Crawling also poses challenges such finding “good” seeds and deciding the parameters for the web graph exploration. Instead, we leverage the high quality results of the Google SERP to discover relevant stories [16]. Third, instead of performing a generic Google search in which we may not have control of the news sources consulted, we localize collecting relevant stories from local media. Fourth, we integrate collection discovery, building, archiving, and sharing capabilities in the same application.

3 LMP LOCAL STORIES COLLECTION GENERATOR ARCHITECTURE

Our Local Stories Collection Generator can be summarized in two stages: discovering local media, and collection building (Fig. 1).

3.1 Stage 1: Nearby News Media discovery

The first stage of the Local Stories Collection Generator discovers local news media organizations that serve a particular location. This function is performed by the tool Geo (Fig. 1, [10]). Given a zip code l and an integer k , this stage returns a set M ($|M| = k$), of newspapers, TV, and radio stations in order of proximity to l . Radio stations are excluded from local stories collection building. Each element m_i of M is a newspaper or TV media organization represented by its website, parent city, social media sites, etc. For example, for $l = “23529”$ (Norfolk, Virginia, USA) and $k = 10$, stage 1 returns Table 1. This stage relies on our public local news media repository which consists of:

- 5,992 newspapers
- 1,061 TV stations, and
- 2,539 radio stations

The local news media repository includes the longitude and latitude coordinates of the parent city of all media sources. This allows mapping from zip codes to nearby news media outlets. The repository was built by scraping data from [19], and currently consists of US local news media. We plan to expand the repository to include other countries [14]. Our local media repository is publicly available [15].

Algorithm 1 : Local Stories Collection Generator - generate a collection of local stories for a query.

Input: Query q and set M of nearby local news media.

Output: Local news collection matrix C .

function GENCOL(q, M)

for each $m_i \in M$ **do**

$q'_i \leftarrow q + “site : ” + m_i.website.domain$ \triangleright (stm. 1)

$S^{m_i} \leftarrow SE(q'_i)$ \triangleright find stories from m_i (stm. 2)

for each $s_j \in S^{m_i}$ **do**

$C_i \leftarrow s_j$ \triangleright Add story s_j to col. C (stm. 3)

end for

end for

return C

end function

3.2 Stage 2: Collection Building

The second stage of the Local Stories Collection Generator finds and collects news stories from the local news sources discovered in stage 1 (M). We use Google search to discover local news stories. Given a story or event represented by a query, q , and a set of local news sources, M , the collection building process (Fig. 1 and Algorithm 1) is outlined as follows.

- (1) For a given media m_i modify the original query as such: concatenate the site directive (“site:”) to the domain of the news media website and the original query. For example, if query, $q = “zika virus”$ and m_i is the Community News of Miami (domain - communitynewspapers.com), due to stm. 1 of Algorithm 1, $q'_i = “zika virus site: communitynewspapers.com”$. The site directive instructs Google to return links only from the specified domain.
- (2) For a given media (m_i) issue the modified query q'_i to Google (Algorithm 1, stm. 2).
- (3) For each story ($s_j^{m_i}$) from a news media m_i , add to collection C . (Algorithm 1, stm. 3)

The Local Stories Collection Generator is implemented by a Google Chrome extension. The source code is publicly available [15]. See examples of two collections generated with the same query: “protesters and police” - The Local collection¹ (Fig. 4), generated by the Local Stories Collection Generator consists of multiple local news sources from Virginia, such as *Virginia Pilot*, *WHRO-TV*, and *WTKR-TV*. The non-Local collection² (Fig. 3), consist of a mix of non-Local sources (e.g., *CNN* and *NBC News*), Local sources (e.g., *ABC7 Chicago* and *Chicago Tribune*), and a Youtube source.

3.3 Collection Archiving

To mitigate the problems of content drift and link rot, as well as preserve collections for future users and researchers, the LMP’s Local Stories Collection Generator implements a variant of the archiving architecture outline by Fig. 2. The ideal archiving architecture is outlined by the following steps.

- (1) Every collection, C , is identified by a collection URI, $URI-C$, and every story in the collection is identified by the story’s URI, $S_i^{m_j}$. The collection is sent to a service (*Vis* - Fig. 2, [12]) which dereferences the $URI-C$ and saves the resultant HTML representation, $HTML-C$. The HTML representation includes thumbnails for each story from each media source in the collection. In addition, $HTML-C$ includes other details such as the author of the collection, the date and time the collection was built, the collection name and the query that generated the collection. It is important to save (via *Vis*) collections as soon as possible after creation to avoid temporal violations [1].
- (2) The collection URI, $URI-C$ is sent to m public Web archives such as the Internet Archive. Additionally, the URIs for each story, $S_i^{m_j}$ is sent to the Archives.
- (3) The outcome of sending the collection URI, $URI-C$ to public Web Archives is a set of URIs representing archived copies of the collection from m public Web Archives.

Our current archiving implementation differs from the described architecture in two ways. First, we do not save the HTML representation of the collection. Instead we save a JSON representation that is parsed

¹See complete: <http://www.localmemory.org/vis/collections/local-memory-project/queries/usa-23529-protesters-and-police-10>

²<http://www.localmemory.org/vis/collections/lmp-test/queries/country-na-zipcode-na-protesters-and-police-0-2017-02-04>

















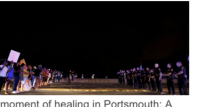



<p>U.S. UNCUT TOM CAHILL</p>  <p>How Protesters Actually Helped the Dallas Police - US Uncut (Jul 8, 2016)</p> <p>Exclude (saved/archived)</p> <p>After snipers killed five police officers last night in Dallas, the right wing has rushed to blame Black Lives Matter protesters for the officers' deaths.</p>	<p>ALTERNET</p>  <p>Millions of people are heading Alternet. Find out how many and support us.</p> <p>Why Are Police Attacking Peaceful Protesters? How OWS Has Exposed ... (Oct 20, 2011)</p> <p>Exclude (saved/archived)</p> <p>As the number of OWS arrests nears 1000, instances of police brutality continue to pile up. Now all of America is seeing the result of police militarization.</p>	<p>CHICAGOTRIBUNE.COM PATRICK M. O</p>  <p>Hundreds protest police brutality, gnarling traffic on downtown streets ... (Jul 12, 2016)</p> <p>Exclude (saved/archived)</p> <p>For the fifth consecutive day, crowds of demonstrators gathered downtown Monday to protest police violence against African-Americans. A sit-in in Millennium ...</p>	<p>CHICAGOTRIBUNE.COM GRACE WONG AND MEGAN CREPEAU</p>  <p>Demonstrators gather to protest police shootings - Chicago Tribune (Aug 6, 2016)</p> <p>Exclude (saved/archived)</p> <p>Protesters gathered at police headquarters and marched through the South Side on Aug. 5, 2016, to protest the shooting of 18-year-old Paul O'Neal, ...</p>	<p>CNN ASHLEY FANTZ AND STEVE VISSEF</p>  <p>Hundreds arrested in protests over shootings by police - CNN.com (Aug 4, 2016)</p> <p>Exclude (saved/archived)</p> <p>(CNN) After a weekend of confrontation, sporadic violence and arrests during protests against the police shootings of black men, the mother of one of the victims ...</p>
<p>CNN</p>  <p>What started out as peaceful protests turned into a heated standoff with police.</p> <p>Exclude (saved/archived)</p>	<p>NBC NEWS</p>  <p>Protesters and Police Face Off Outside Trump Speech in California ... (Apr 30, 2016)</p> <p>Exclude (saved/archived)</p> <p>Police pushed back protesters who rushed the hotel while elsewhere a group overran the barricade and reached the building.</p>	<p>USA TODAY AAMER MADHANI AND AND MIKE JAMES</p>  <p>Protesters taunt police on Chicago streets amid anger over teen killing (Nov 25, 2015)</p> <p>Exclude (saved/archived)</p> <p>CHICAGO — Angry protesters hit the streets and taunted police for a second night Wednesday following the release of a dramatic video showing a white ...</p>	<p>WASHINGTON POST RADLEY BALKO</p>  <p>After Ferguson, how should police respond to protests? - The ... (Aug 14, 2014)</p> <p>Exclude (saved/archived)</p> <p>There are also reports, video, and images of police taunting, arresting, and otherwise intimidating peaceful protests all over the town. While it's true that there ...</p>	<p>YOUTUBE BEARKAT4160</p>  <p>San Diego, California Donald Trump Rally. Crazy Trump haters attack the police and try to jump over ...</p> <p>Exclude (saved/archived)</p>
<p>ABC7 CHICAGO EVELYN HOLMES</p>  <p>Marchers swarm downtown protesting police brutality abc7chicago.com (Jul 10, 2016)</p> <p>Exclude (saved/archived)</p> <p>More than 100 protesters swarmed downtown Chicago for hours on Saturday, starting at the Taste of Chicago, moving to Water Tower Place and then into the ...</p>	<p>ABC7 CHICAGO</p>  <p>RAW VIDEO: Protesters throw bottles at police, horses abc7chicago.com (Jul 9, 2016)</p> <p>Exclude (saved/archived)</p> <p>Protesters threw bottles at police and their horses Saturday night in the South Loop. The group had marched for hours throughout downtown Chicago protesting ...</p>	<p>ABC7 SAN FRANCISCO LYANNE MELEN</p>  <p>Black Lives Matter protesters chain themselves to doors of Oakland ... (Jul 21, 2016)</p> <p>Exclude (saved/archived)</p> <p>Following a more than 12-hour occupation of the Oakland Police Officers Association building, Black Lives Matter demonstrators took their cause to police ...</p>	<p>ALTERNET</p>  <p>Police Gone Overboard: Militarized Cops Arrest 200 Non-Violent ... (Jul 13, 2016)</p> <p>Exclude (saved/archived)</p> <p>Since the police killing of Alton Sterling, thousands of people in Baton Rouge have been peacefully protesting day and night all over the city. There has been no ...</p>	<p>CBSNEWS</p>  <p>Protesters and police clash outside California Trump Rally - CBS News (May 28, 2016)</p> <p>Exclude (saved/archived)</p> <p>San Diego police seen beating demonstrator with batons as presumptive GOP nominee tells crowd in Fresno "There is no drought"</p>

Figure 3: A subset of a non-Local collection for query: "Protesters and Police".

3. Virginian Pilot, Newspaper, (2.98 miles, VA - USA)

<p>VIRGINIAN-PILOT KIMBERLY PIE</p>  <p>Portsmouth council meeting punctuated by hymns, praise, tension ... (Jul 12, 2016)</p> <p>Exclude (saved/archived)</p> <p>Just moments before, elected leaders and members of the public almost universally praised Portsmouth's police chief and the protesters for ...</p>	<p>VIRGINIAN-PILOT MIKE HIXENB</p>  <p>A moment of healing in Portsmouth: A black protester hugging a white ... (Jul 13, 2016)</p> <p>Exclude (saved/archived)</p> <p>The photograph, taken during a weekend protest in Portsmouth, spread ... Police and protesters stand divided after a march that began at I.C. ...</p>	<p>VIRGINIAN-PILOT ROBYN SIDERSK</p>  <p>Gathering outside the Scope arena in Norfolk to protest recent shootings of black ... (Jul 9, 2016)</p> <p>Exclude (saved/archived)</p> <p>On Saturday morning, a small group gathered near the water fountain outside the Scope Arena in Norfolk to protest recent shootings of black ...</p>	<p>VIRGINIAN-PILOT JONATHAN EDWAR</p>  <p>Local Black Lives Matter leaders, police praise one another after ... (Jul 11, 2016)</p> <p>Exclude (saved/archived)</p> <p>A Black Lives Matter leader and a Norfolk police lieutenant negotiated Sunday night in a scrum of protesters who were blocking six lanes of ...</p>	<p>VIRGINIAN-PILOT AMIR VERA</p>  <p>Protesters of police shootings organize in Portsmouth, take to the ... (Jul 11, 2016)</p> <p>Exclude (saved/archived)</p> <p>At least 20 people, along with Chief of police Tonya Chapman, gathered at I.C. Norcom High School Monday. However they weren't meeting to ...</p>
--	---	---	--	---

4. WHRO, TV, (2.98 miles, VA - USA)

<p>WHRO</p>  <p>WHRO - Baton Rouge Protester On Arrest: 'I Didn't Know If I Was ... (Jul 13, 2016)</p> <p>Exclude (saved/archived)</p> <p>Akeem Muhammad stands with a group of protesters in Baton Rouge, La., on ... disrupted on Sunday, and police arrested more than 120 people.</p>	<p>WHRO SHEREEN MARISOL MERAJ</p>  <p>WHRO - A Letter From Young Asian-Americans To Their Families ... (Jul 27, 2016)</p> <p>Exclude (saved/archived)</p> <p>A police officer patrols during a protest in support of the Black Lives Matter ... The protesters said Liang was being treated as a scapegoat at a ...</p>	<p>WHRO LEAH DONNELLA</p>  <p>WHRO - Roundup: Reactions To Bill Clinton's Exchange With Black ... (Apr 8, 2016)</p> <p>Exclude (saved/archived)</p> <p>Bill Clinton's response to Thursday's protesters was peppered with ... and campaigning with mothers who've lost sons to police violence, Bill ...</p>	<p>WHRO MEG ANDERSON</p>  <p>WHRO - Politicians React To Two Deaths Of Black Men By Police (Jul 7, 2016)</p> <p>Exclude (saved/archived)</p> <p>Politicians React To Two Deaths Of Black Men By Police ... Earlier that day, Dayton addressed protesters outside his home, calling the killing a ...</p>	<p>WHRO EYDER PERALTA</p>  <p>WHRO - Black Lives, Blue Lives: Political Conventions Reveal A ... (Aug 1, 2016)</p> <p>Exclude (saved/archived)</p> <p>It felt like all of the protests, the killings of police and of young black men, and the riots had led to this moment of bald-faced confrontation.</p>
---	---	--	---	--

5. WTKR, TV, (2.98 miles, VA - USA)

<p>WTKR.COM MATT KNIGHT AND C</p>  <p>After riots, protesters and police ensure peace in Baltimore WTKR.com (Apr 29, 2015)</p> <p>Exclude (saved/archived)</p> <p>While some protesters defied the curfew and faced off with police, the confrontation was essentially a staring contest -- each side waiting to see ...</p>	<p>WTKR.COM CNN WIRE</p>  <p>Trump protesters smash door, break through barriers WTKR.com (May 25, 2016)</p> <p>Exclude (saved/archived)</p> <p>Albuquerque police say bottles and rocks were thrown at a Donald Trump rally there Tuesday night, shattering a glass door as tensions ...</p>	<p>WTKR.COM MELISSA STEPHENSON AND CNN WIRI</p>  <p>Five officers killed during protests against police in Dallas; suspect ... (Jul 7, 2016)</p> <p>Exclude (saved/archived)</p> <p>The ambush began with gunshots that killed five officers and sent screaming crowds scrambling for cover. It ended ...</p>	<p>WTKR.COM CNN WIRE</p>  <p>Hundreds arrested in protests over shootings by police WTKR.com (Jul 11, 2016)</p> <p>Exclude (saved/archived)</p> <p>After a weekend of confrontation, sporadic violence and arrests during protests against the police shootings of black men, the mother of one of ...</p>	<p>WTKR.COM TRIBUNE MEDIA WIRE</p>  <p>The Baton Rouge protest photograph everyone is talking about ... (Jul 11, 2016)</p> <p>Exclude (saved/archived)</p> <p>Will this be the photograph that symbolizes this past week's protests? An image of what appears to be a woman's peaceful resistance to police ...</p>
--	---	---	---	---

Figure 4: A subset of a Local collection for query: "Protesters and Police" for Norfolk Virginia.

Table 1: List of News sources for $l = "23529"$ (Norfolk Virginia, USA)

Index	Miles away	Name	Type	City/County
1	2.98	Hampton Roads Messenger	Newspaper	Norfolk (VA, USA)
2	2.98	Inside Business	Newspaper	Norfolk (VA, USA)
3	2.98	Virginian Pilot	Newspaper	Norfolk (VA, USA)
4	2.98	WHRO	TV	Norfolk (VA, USA)
5	2.98	WTKR	TV	Norfolk (VA, USA)
6	2.98	WVEC	TV	Norfolk (VA, USA)
7	4.58	Mace & Crown	College Newspaper	Old Dominion Univ (VA, USA)
8	5.84	Spartan Echo	College Newspaper	Norfolk State University (VA, USA)
9	10.95	WAVY	TV	Portsmouth (VA, USA)
10	12.82	Daily Press	Newspaper	Hampton (VA, USA)

in order to generate the HTML representation. Second, we send URIs to just one archive - Archive.is. We plan to fully implement the ideal archiving architecture once the infrastructure to support such implementation is available.

3.4 Community Collection Building

We believe there is value when multiple users contribute to the same collection. This is similar in spirit to the Internet Archive’s request to the public to contribute³ URIs for the 2016 Orlando Nightclub Shooting Web Archive [9]. The Local Stories Collection Generator enables users to tag a collection with a hashtag. The hashtag provides a means for thematically-related collections to be organized. Collections built by communities could be a valuable asset to research which seeks to juxtapose different perspectives of local reports about the same subject. It could also provide a means of creating high quality datasets to be consumed by machine learning processes.

4 EVALUATION

To evaluate LMP’s Local Stories Collection Generator, we measure the degree of exposure Local collections have compared to non-Local collections as well as some properties of both collections. We claim that Local collections have less exposure compared to non-Local collections. Hence, through collection building, archiving, and sharing, LMP could facilitate the increase of exposure of Local news sources. Our evaluation dataset comprised of 20 pairs (Local and non-Local) of collections corresponding to 20 different stories. Furthermore, each collection (Local and non-Local) was further split into two classes: G - extracted from the default Google SERP, and NV - extracted from the Google News vertical SERP. This means a given story x from our evaluation dataset is represented by two collections - Local and non-Local, and x ’s Local and non-Local collections are further split into G and NV . In total, the story x has four sub-collections (Table 2). The evaluation dataset is publicly available [15].

For each collection we measured archival coverage, tweet index rate, temporal range, precision, and sub-collection overlap.

4.1 Evaluation Dataset

Each story of the evaluation dataset (Table 2) is comprised of two collections - Local and non-Local. The Local collection was generated by our Local Stories Collection Generator (Algorithm 1), by applying the “site:” parameter to the Collection query and the local news sources discovered by Geo. The non-Local collection does not apply the “site:” parameter to the Collection query.

³<https://archive.is/eGuKh>

4.2 Evaluation Metrics

We evaluated the 20 pairs of collections across five dimensions. The first two metrics, archival coverage and tweet index rate, were used to approximate collection exposure. For each metric, let us consider the definition of the metric, our claim about Local and non-Local collections for the metric, and how the metric was extracted from our evaluation dataset:

4.2.1 Archival coverage.

- (a) **Definition:** Given a collection C , the archival coverage is the fraction of C which is archived. For example, if we found 10 archived stories from C (where $|C| = 50$), the archival coverage or archive rate of C is $10/50 = 0.2$. It is important to note that for a URI, we only considered whether or not the URI was archived and not the quality [5] of the Memento (archived copy).
- (b) **Claim:** We claim that non-Local collections possess higher archive rates than Local collections. This may be partly due to the idea that Web archives favor more popular websites. For example, a story from CNN might be more likely to be archived than a story from a small town’s newspaper website.
- (c) **Extraction:** For a story S , represented by a URI in a collection C , the binary archived state ($S \in \{Archived, NotArchived\}$) was extracted by utilizing the MemGator [2] utility.

4.2.2 Tweet index rate.

- (a) **Definition:** Similar to the archive rate, given a collection C , the tweet index rate is the fraction of C which could also be found embedded in a tweet. For example, if we found 40 URIs from C (where $|C| = 50$) embedded in tweets, the tweet index rate of C is $40/50 = 0.8$.
- (b) **Claim:** We claim that non-Local collections possess higher tweet index rates than Local collections. The reason for this may be due to the user-search behavior. Many users find stories from the search engines, and may share these discovered stories on their social media accounts. Consequently, since non-Local collections are created directly from SERPs without modifying the collection query, non-Local collection should enjoy higher tweet index rates.
- (c) **Extraction:** For a story S , represented by a URI in a collection C , the binary tweet index state, $S \in \{Found, NotFound\}$ was extracted by searching Twitter.

The last three metrics, precision, temporal range, and sub-collection overlap were employed for experimentation: We measured the precision of the resulting collections to see if the focus on Local news

Table 2: Evaluation Dataset comprised of 20 pairs (Local and non-Local) of Collections. Local & non-Local Collection are further split into two sub-collections: G - Collection extracted from the Default Google SERP & NV - News Vertical SERP.

Index	Collection	Location	Local story count		non-Local story count	
			G	NV	G	NV
1	dakota access pipeline protest	58538 (Cannon Ball, North Dakota)	43	50	49	50
2	Hurricane Matthew	29925 (Hilton Head Island, South Carolina)	48	50	51	50
3	Cajun Navy Flood	70801 (Baton Rouge, Louisiana)	44	59	54	50
4	drunk governor susana martinez hotel party police	87501 (Santa Fe, New Mexico)	53	19	50	50
5	albuquerque road rage 4 year old	87101 (Albuquerque, New Mexico)	64	55	50	50
6	albuquerque homeless police shooting	87101 (Albuquerque, New Mexico)	59	50	50	50
7	labelle amoeba	33935 (LaBelle, Florida)	47	18	50	36
8	UF taser	32601 (Gainesville, Florida)	40	40	53	50
9	Pizza Gate Comet Ping Pong	20008 (Washington, DC)	52	35	50	50
10	Ghost Ship fire	94601 (Oakland, California)	49	47	51	50
11	Abortion law	43017 (Dublin, Ohio)	43	51	55	50
12	Music City Bowl	37219 (Nashville, Tennessee)	53	56	47	40
13	Tornado	36547 (Gulf Shores, Alabama)	47	56	51	50
14	US Customs delays	33126 (Miami, Florida)	60	54	49	50
15	housing prices housing crisis	94115 (San Francisco, California)	52	51	53	50
16	housing prices housing crisis	94539 (Freemont, California)	50	50	53	50
17	Trump election	02138 (Cambridge, Massachusetts)	50	44	51	50
18	Trump election	71613 (Pine Bluff, Arkansas)	42	44	51	50
19	Marijuana legalization	89127 (Las Vegas, Nevada)	57	52	51	50
20	Marijuana legalization	75202 (Dallas, Texas)	52	41	51	50
Subcollections Total			1,005	922	1,020	976
Collections Total			1,927		1,996	
Total			3,923			

sources impacted the relevance of the SERPs. We considered the temporal range metric in order to see if some sub-collections are temporally biased toward older or newer documents. Finally, we considered the sub-collection overlap to gauge the level of agreement (common stories) between Local and non-Local collections. A low degree of overlap between Local and non-Local collections indicates a disparity in the sampling of news sources of Local and non-Local collections, thereby justifying the need for exposing Local collections, since non-Local collections are presumed to be more exposed. For each metric, let us consider the definition of the metric, our claim about Local and non-Local

collections for the metric, and how the metric was extracted from our evaluation dataset:

4.2.3 Temporal range.

- (a) **Definition:** Given a collection C , the temporal range of C is the distribution of the creation timestamps of the stories in C .
- (b) **Claim:** We claim that non-Local collections are temporally biased to produce newer stories than Local collections. The reason for this may be that the non-Local collections are produced directly from SERPs without modifying the collection

query, and Search Engines are optimized to produce more recent documents.

- (c) **Extraction:** Most news stories have creation timestamps. We extracted these timestamps from the SERPs.

4.2.4 Precision.

- (a) **Definition:** Given a collection C , the precision of C is the fraction of C that are relevant to the collection query. A story S in C is relevant if the story is on topic with respect to the query used for creating the collection. For example, our evaluation collection no.7 (“labelle amoeba”) is about a 12-year-old’s battle with a brain-eating amoeba in Labelle Florida, USA. A story in this collection about the singer Patti Labelle on Ameoba Music is considered non-relevant.
- (b) **Claim:** We claim that non-Local collections possess a higher precision than Local collections. This may be partly due to the fact that since non-Local collections are built without modifying the query, the SERP benefits from being populated from an unrestricted number of high quality sources. But a non-Local collection may suffer if a query applies to multiple localities. For example, a non-Local collection for the “flood” query may include definitions or general information pages, which are not inappropriate. However, if the intent was localized, because the query “flood” is applicable to many areas, a localized search such as LMP’s Local Stories Collection Generator is more appropriate. Also, a localized search may be appropriate for building collections for the early stages of an event.
- (c) **Extraction:** The evaluation dataset was manually evaluated by a group of 14 evaluators. Each evaluator was presented with multiple collections. For each story in a collection, an evaluator scored the story as relevant if the story was on topic with respect to the collection query, and non-relevant otherwise. Additionally, we considered a story relevant or non-relevant only if the score was by a margin of 2 votes or more. Stories that did not fulfill this criteria (e.g., a score of 3-2) were not included in our precision calculation. The evaluation results are publicly available [15].

4.2.5 Sub-collection overlap.

- (a) **Definition:** Given a collection set C populated from the evaluation dataset, let sub-collection sets L_G and L_{NV} define sets populated from Local-G and Local-NV, respectively. Similarly, let sub-collection sets NL_G and NL_{NV} define sets populated from non-Local-G and non-Local-NV, respectively. The overlap set of Local and non-Local collections, is given by $(L_G \cap L_{NV})$, and $(NL_G \cap NL_{NV})$, respectively. The Jaccard index of the overlap sets of Local and non-Local collections may indicate the degree to which collections have both Local and non-Local interests (Eqn. 1).

$$\frac{|(L_G \cap L_{NV}) \cap (NL_G \cap NL_{NV})|}{|(L_G \cap L_{NV}) \cup (NL_G \cap NL_{NV})|} \quad (1)$$

- (b) **Claim:** We claim Local sub-collections L_G and L_{NV} have more in common (more overlap) than non-Local sub-collections NL_G and NL_{NV} , due to the site directive restriction imposed by the Local Stories Collection Generator (Algorithm 1).
- (c) **Extraction:** For each collection, we calculated the overlap of Local and non-Local collections and the Jaccard index of both overlap sets as described in the definition.

5 EVALUATION RESULTS DISCUSSION

Non-Local collections G and NV produced archive rates of 0.83 and 0.80, respectively, while Local collections G and NV produced archive rates of 0.52 and 0.63, respectively, confirming our claim that non-Local collections possess higher archive rates than Local Collections (Fig. 5a). We took the archival coverage test further by testing if a random page of a Local collection top-level site is less likely to be archived than a random page of a non-Local collection top-level site. Our experiments did not confirm this claim (Fig. 6 a & b). This may be due to the sampling technique we employed. For each top-level site, we crawled from the top pages to a maximum depth of four to extract random links. Since top-level sites are more likely to be archived, the descendants of top-level sites are also likely to be archived. Also one-third of the Local domains set were also available in the non-Local domains set. This is not surprising since there is not always a clear distinction between a Local source and a non-Local source. For example, the *Washington Post* newspaper is a national newspaper, but it is also the Local newspaper for the residents of Washington, DC. However, top-level sites and URIs from non-Local collections possess a higher magnitude of Mementos than their Local counterparts (Fig. 6c). Consequently, LMP’s Local Stories Collection Generator’s archiving capability can provide additional exposure to Local collections.

Our claim about tweet index rates was confirmed, with non-Local collections (G and NV) producing higher tweet index rates of 0.71 and 0.80, respectively, compared to their local counterparts with tweet index rates 0.44 and 0.59, for G and NV , respectively (Fig. 5b). Consequently, LMP’s collection generator provides a means for users to share locally built collections on Twitter, thus providing additional exposure to Local collections.

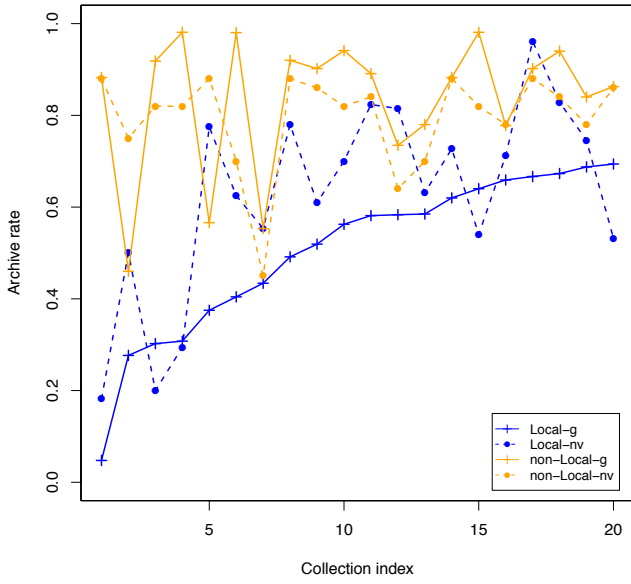
Our claim about precision was partially confirmed with the following precision rank ordering from highest to lowest (based on a relevance margin of 2 votes or more): Local-G: 0.84, non-Local-G: 0.72, Local-NV: 0.71, and non-Local-NV: 0.68. Relaxing the threshold of relevance to a margin of 1 vote or more resulted in the following precision rank ordering from highest to lowest: non-Local-G: 0.84, Local-G: 0.79, non-Local-NV: 0.71, and Local-NV: 0.70 (Fig. 7). These results show that type-G collections produce documents at a higher precision than NV .

Our claim about the temporal range was confirmed: Non-Local-NV collections possessed the highest probability of producing the newest document with a probability of 0.75. On the other hand, Local-G collections produce the oldest documents with a probability of 0.7. The consequences of these probabilities are crucial: One must sample Local-G collections in order to maximize the chances of finding the first reports about a story or event. Let the probability $P_{new}(sub-collection)$, define the probability of the event that $sub-collection \in \{Local-G, Local-NV, non-Local-G, non-Local-NV\}$ produces the newest document. Similarly, let the probability $P_{old}(sub-collection)$ define the probability of the event that $sub-collection$ produces the oldest document. Table 3 outlines both probabilities for each event.

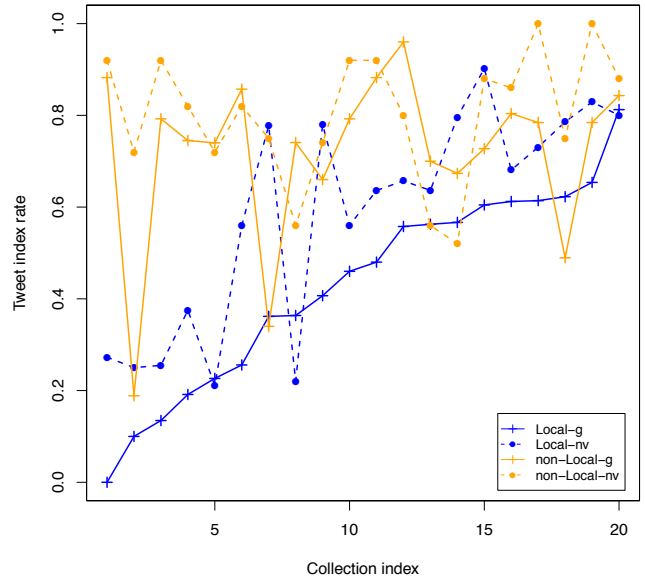
Finally, Local collections showed a higher overlap rate than non-Local collection, confirming our claim.

6 FUTURE WORKS

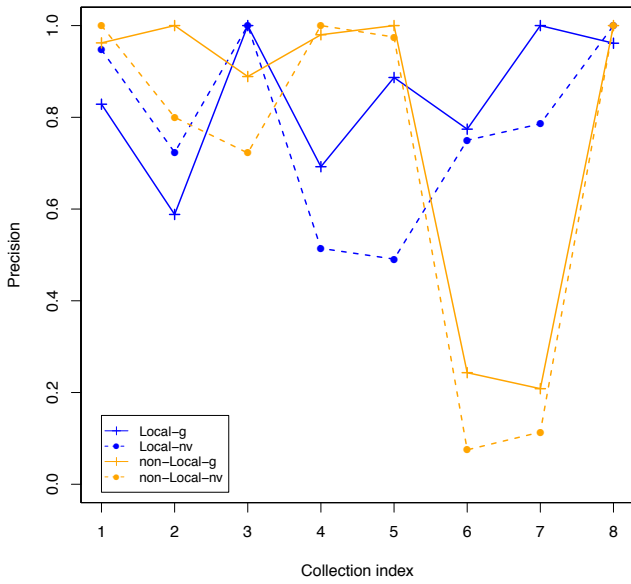
Geo relies on a local new media repository which houses information about local news media. For example, for a single newspaper, this repository contains the name of the newspaper, website, and the geo-coordinates (latitude and longitude) of the city in which the newspaper organization is located. We rely on third party sources which provide



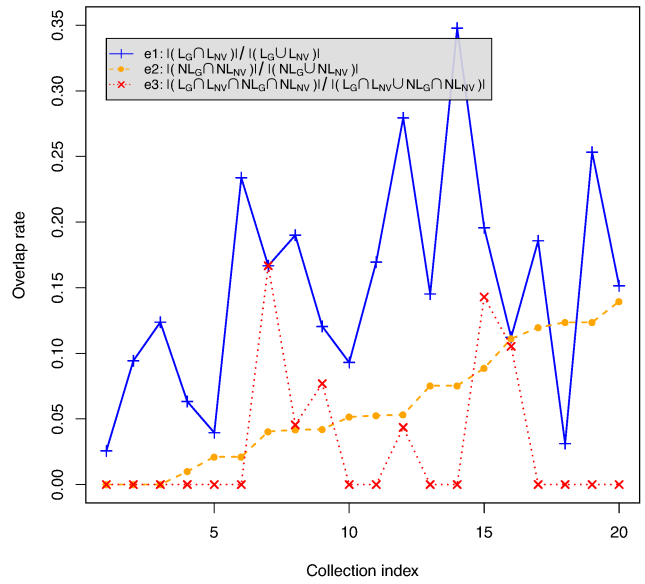
(a) Distribution of archive rates for Local & non-Local collections. Non-Local-G & non-Local-NV produced higher archive rates (0.83 & 0.80, respectively) than Local-G & Local-NV sub-collections (0.52 & 0.63, respectively.)



(b) Distribution of tweet index rates for Local & non-Local collections. Non-Local-G & non-Local-NV produced higher tweet index rates (0.71 & 0.80, respectively) than Local-G and Local-NV sub-collections (0.44 & 0.59, respectively.)

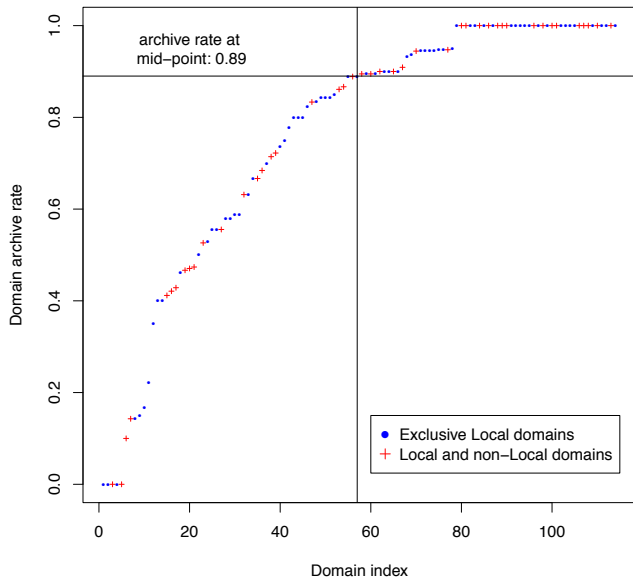


(c) Distribution of precision for Local & non-Local collections. Local-G & Non-Local-G produced higher precision values (0.84 & 0.72, respectively) than Local-NV & Non-Local-NV sub-collections (0.71 & 0.68, respectively.)

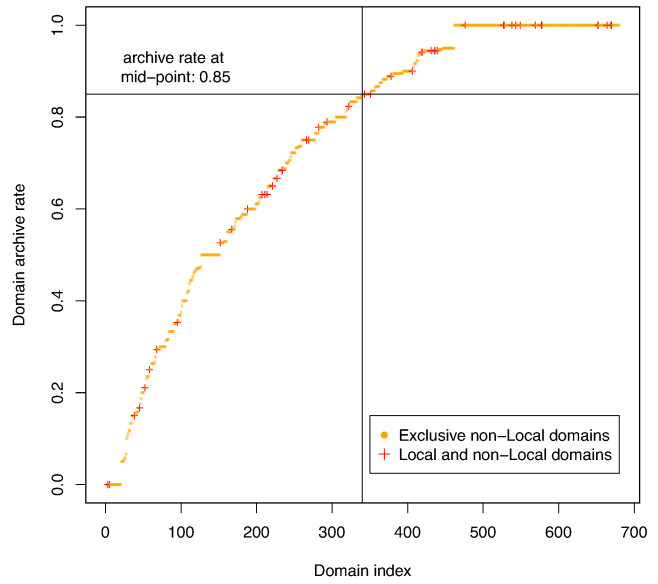


(d) Distribution of Overlap for Local (e1), non-Local (e2), & the overlap between Local and non-Local overlaps (e3). Local sub-collections showed higher overlap than non-Local sub-collections.

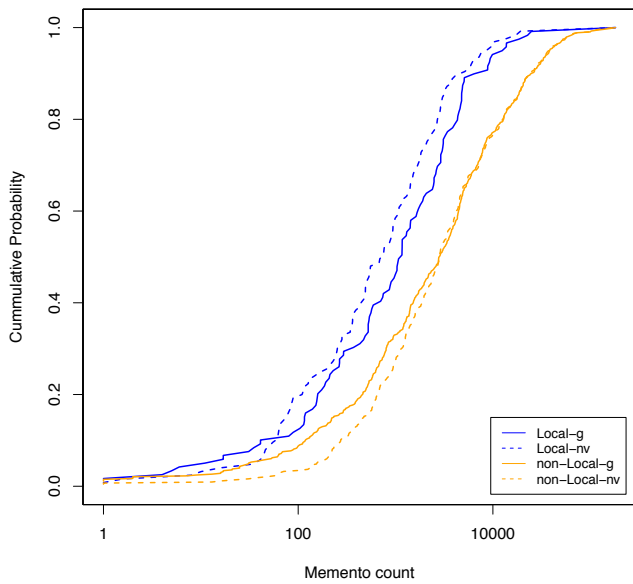
Figure 5: Distribution of Archive rates (a), Tweet index rates (b), Precision (c), & Overlap (d)



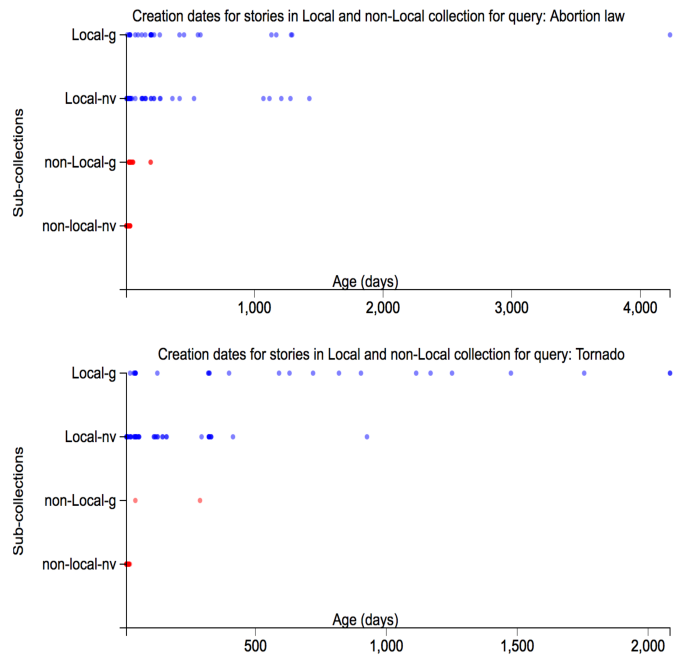
(a) Distribution of archive rates for random collections of pages from Local websites. Common domains in red. Random Local pages are as likely to be archived as Non-Local random pages.



(b) Distribution of archive rates for random collections of pages from non-Local websites. Common domains in red. Random non-Local pages are as likely to be archived as random Local pages



(c) CDF of Memento count for top-level sites of Local (G & NV) & non-Local (G & NV) collections. Non-Local collections with a higher Probability of Mementos



(d) Creation date distributions for “Abortion law” (top) and “Tornado” (bottom), showing Local-G sub-collection with a highest probability of producing the oldest documents (P_{old}) (Table 3).

Figure 6: Distribution of Archive rates for random pages (a & b), CDF of Memento count (c), & Creation dates (d).

7 CONCLUSIONS

We introduced Geo - a tool which returns a list of local media sources in order of proximity to a user-supplied zip code. We also introduced our Local Stories Collection Generator. This tool helps users discover, build, archive and share collections of stories about an event expressed by a user-supplied query and zip code pair. Our evaluation results confirmed our claims that non-Local collections produce higher archive, tweet index, and precision rates than Local collections, thereby justifying the need to further expose Local collections. We also learned from our evaluation results that Local-G collections have the highest probability of reporting the first account of a story, and there is more overlap between Local collections (G and NV). Our tools, local news repository, and evaluation results are publicly available [15]. In recent decades that have seen the decline of Local media due to various forces such as the influence of social media, measures such as the Local Memory Project which strive to help increase the exposure of Local media content are pertinent.

ACKNOWLEDGEMENTS

This work was made possible in part by IMLS LG-71-15-0077-15 and support from the Harvard Law School Library. We are grateful for the support.

REFERENCES

- [1] Scott G Ainsworth, Michael L Nelson, and Herbert Van de Sompel. 2015. Only One Out of Five Archived Web Pages Existed as Presented. In *Proceedings of HT 2015*. ACM, 257–266.
- [2] Sawood Alam and Michael L Nelson. 2016. MemGator-A portable concurrent memo-to- aggregator: Cross-platform CLI and server binaries in Go. In *Proceedings of JCDL 2016*. IEEE, 243–244.
- [3] Donna Bergmark. 2002. Collection synthesis. In *Proceedings of JCDL 2002*. ACM, 253–262.
- [4] Jeremy Blackburn and Adriana Iamnitchi. 2013. An architecture for collecting longitudinal social data. In *2013 IEEE ICC Workshops*. IEEE, 184–188.
- [5] Justin F Brunelle, Mat Kelly, Michele C Weigle, and Michael L Nelson. 2016. The impact of JavaScript on archivability. *IJDL* 17, 2 (2016), 95–117.
- [6] Mohamed MG Farag, Sunshin Lee, and Edward A Fox. 2017. Focused crawler for events. *IJDL* (2017), 1–17.
- [7] 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.
- [8] 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*. ACM, 75–84.
- [9] Internet Archive Global Events. 2016. 2016 Pulse Nightclub Shooting Web Archive. <https://archive-it.org/collections/7570>.
- [10] LMP. 2016. Local Memory Project - Geo. <http://www.localmemory.org/geo/>.
- [11] LMP. 2016. Local Memory Project - Local Stories Collection Generator. <https://chrome.google.com/webstore/detail/local-memory-project/khineeknpgfcholchjihmhoficfp>.
- [12] LMP. 2016. Local Memory Project - Vis. <http://www.localmemory.org/vis/>.
- [13] Rasmus Kleis Nielsen. 2015. *Local journalism: the decline of newspapers and the rise of digital media*. IB Tauris.
- [14] Alexander Nwala. 2017. Local Memory Project - going global. <http://ws-dl.blogspot.com/2017/04/2017-04-18-local-memory-project-going.html>.
- [15] Alexander Nwala. 2017. Local Memory Project - going global. <https://github.com/harvard-lil/local-memory>.
- [16] Alexander Nwala and Michael Nelson. 2016. A Supervised Learning Algorithm for Binary Domain Classification of Web Queries using SERPs. *arXiv:1605.00184* (2016).
- [17] Jialun Qin, Yilu Zhou, and Michael Chau. 2004. Building domain-specific web collections for scientific digital libraries: a meta-search enhanced focused crawling method. In *Proceedings of JCDL 2004*. IEEE, 135–141.
- [18] 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>.
- [19] USNPL. 2016. US Newspaper List. <http://www.usnpl.com/>.
- [20] James Warren. 2016. How the media blew Flint. <https://www.poynter.org/2016/how-the-media-blew-flint/392662/>.
- [21] Seungwon Yang, Kiran Chitturi, Gregory Wilson, Mohamed Magdy, and Edward A Fox. 2012. A study of automation from seed URL generation to focused web archive development: the CTRnet context. In *Proceedings of JCDL 2012*. ACM, 341–342.

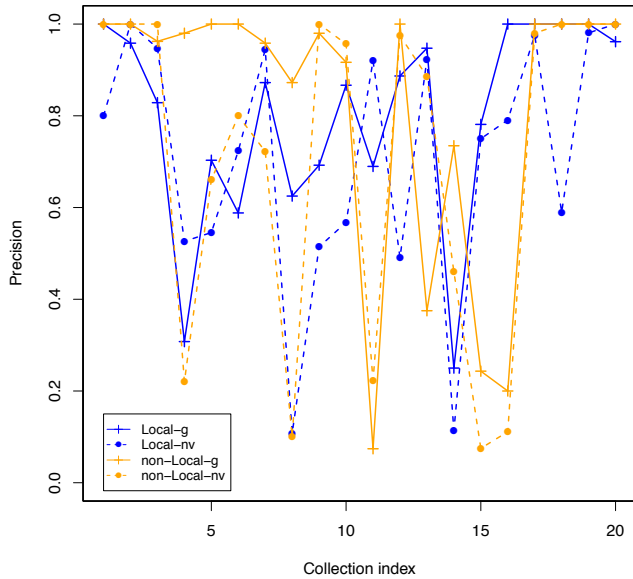


Figure 7: Distribution of precision for Local & non-Local collections for Relaxed Relevance Margin of 1 Vote or more. Non-Local-G & Local-G produced higher precision values (0.84 & 0.79, respectively) than Non-Local-NV & Local-NV sub-collections (0.71 & 0.70, respectively.)

Table 3: Probabilities of $P_{new}(sub-collection)$ & $P_{old}(sub-collection)$ events. $P_{new}(sub-collection)$ & $P_{old}(sub-collection)$ define probabilities of the events that the given sub-collections produce the newest & the oldest documents, respectively.

Sub-collections (events)	P_{new}	P_{old}
Local-G	0.1	0.7
Local-NV	0.17	0.2
non-Local-NV	0.75	0.1

subsets of this information. Thus, we are limited by their sizes, frequency of update, and biases. If for some reason, a local news media does not make it into these third party lists, the media might be excluded from our local news repository. We believe this discovery problem justifies the need for further research to explore the automatic or semi-automatic discovery of local news media and their respective geographical information.

We are entirely dependent upon Google to discover news stories as opposed to interacting with the various media sources directly with interfaces such as OpenSearch. This means we are at the mercy of the Google index and its preferences. Further research may be needed to compare both interactions.

Finally, Geo relies on third party lists and does not check for fake news websites. The current political climate has shown the need for content managers to take an active role in restricting the spread of fake news content. The fake news problem and the various ways to tackle it are still being debated. We envision benefiting from this discourse to help implement a filter that prevents including fake news websites in our local news repository.