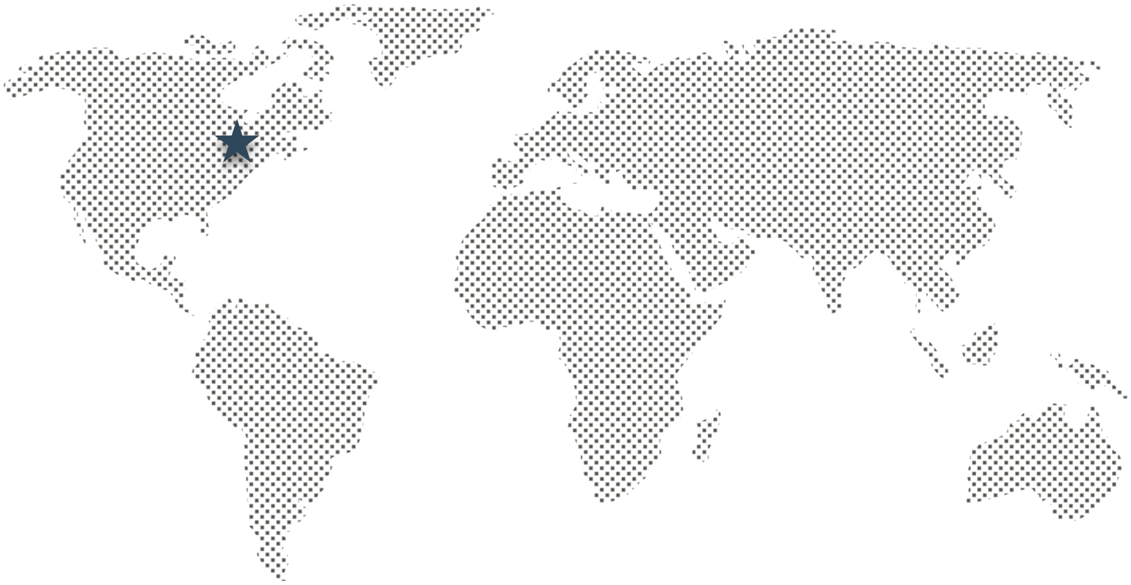
Proceedings *of the*



2017

Web Archiving & Digital Libraries

Workshop

June 22-23, 2017

Toronto, Canada

Edward A. Fox

Zhiwu Xie

Martin Klein

Editors

Preliminary Proceedings of
Web Archiving and Digital Libraries 2017

WADL 2017

A Workshop of the ACM/IEEE Joint Conference on Digital Libraries (JCDL 2017)
Toronto, CANADA
June 22-23, 2017

Table of Contents:

- WADL 2017 webpage (4 pages, including agenda, schedule)
- Attendees list
- Submissions, listed and appearing in order of presentation, as follows:

Presenters/Authors	**Title of Presentation**
Ross Spencer	Poster 1: HTTPreserve – Web Preservation in Documentary Heritage
Muhammad Umar Qasim	Poster 2: WARC-Portal: A Tool for Exploring the Past
Mat Kelly, Lulwah Alkwai, Sawood Alam, Michael Nelson, Michele Weigle, Herbert Van De Sompel	Poster 3: Impact of URI Canonicalization on Memento Count
Saket Vishwasrao, Zhiwu Xie, Edward Fox	Poster 4: Web Archiving Through In-Memory Page Cache
Ian Milligan, Nick Ruest, Ryan Deschamps	Paper 1: Building a National Web Archiving Collaborative Platform: The Web Archives for Longitudinal Knowledge Project
Sawood Alam, Mat Kelly, Michele Weigle, Michael Nelson	Paper 2: Avoiding Zombies in Archival Replay Using ServiceWorker
Ziquan Wang, Borui Lin, Ian Milligan, Jimmy Lin	Paper 3: Topic Shifts Between Two US Presidential Administrations
Brenda Reyes Ayala	Paper 4: Web archives: A preliminary exploration of user expectations vs. reality
Emily Maemura, Dawn Walker, Matt Price, Maya Anjur-Dietrich	PANEL: Challenges for Grassroots Web Archiving of Environmental Data
Tom J. Smyth	Paper 5: Legal Deposit, Collection Development, Preservation, and Web Archiving at Library and Archives Canada
Muhammad Umar Qasim and Sam-Chin Li	Paper 6: Working Together Toward a Shared Vision : Canadian Government Information Digital Preservation Network (CGI DPN)
Nick Ruest	Paper 7: Strategies for Collecting, Processing, and Analyzing Tweets from Large Newsworthy Events
Saurabh Chakravarty, Eric Williamson, Edward Fox	Paper 8: Classification of Tweets using Augmented Training

WADL 2017 webpage

Web Archiving and Digital Libraries
JCDL 2017 (http://2017.jcdl.org), Toronto, Canada; 6/22-6/23 Workshop, Room 205, Faculty of Information, 140 St. George Street, U. Toronto

Please see approved workshop proposal at: http://fox.cs.vt.edu/WADLjcdl17.pdf

Please see last year's WADL 2016 webpage. That workshop led in part to a special issue of International Journal on Digital Libraries. There might be another similar issue partially based on WADL 2017.

SCHEDULE:

Featured Talk: Ashley Sands, Senior Library Program Officer, Institute of Museum and Library Services (IMLS), USA: National Digital Platform (NDP), funding opportunities, and examples of currently funded projects

Thursday, June 22, 2-5pm

Time	Activity, Presenters/Authors	Title of Presentation
2:00	Welcome, Introductions	Everyone speaks!
2:10	Ashley Sands, IMLS	FEATURED TALK: National Digital Platform (NDP)
2:40	Lightening Talks	5 minutes each
	Ross Spencer	Poster 1: HTTPreserve – Web Preservation in Documentary Heritage
	Muhammad Umar Qasim	Poster 2: WARC-Portal: A Tool for Exploring the Past
	Mat Kelly, Lulwah Alkwai, Sawood Alam, Michael Nelson, Michele Weigle, Herbert Van De Sompel	Poster 3: Impact of URI Canonicalization on Memento Count
	Saket Vishwasrao, Zhiwu Xie, Edward Fox	Poster 4: Web Archiving Through In-Memory Page Cache
3:00	BREAK	Refreshments, viewing posters and demos
3:20	Ian Milligan, Nick Ruest, Ryan Deschamps	Paper 1: Building a National Web Archiving Collaborative Platform: The Web Archives for Longitudinal Knowledge Project

3:45	Sawood Alam, Mat Kelly, Michele Weigle, Michael Nelson	Paper 2: Avoiding Zombies in Archival Replay Using ServiceWorker
4:10	Ziquan Wang, Borui Lin, Ian Milligan, Jimmy Lin	Paper 3: Topic Shifts Between Two US Presidential Administrations
4:35	Brenda Reyes Ayala	Paper 4: Web archives: A preliminary exploration of user expectations vs. reality
5:00	Close of Session 1, Dinner optional	WADL attendees are encouraged to gather informally for dinner to continue discussions.

Friday, June 23, 9-12pm

Time	Activity, Presenters/Authors	Title of Presentation
9:00	Emily Maemura, Dawn Walker, Matt Price, Maya Anjur-Dietrich	PANEL: Challenges for Grassroots Web Archiving of Environmental Data
9:30	Tom J. Smyth	Paper 5: Legal Deposit, Collection Development, Preservation, and Web Archiving at Library and Archives Canada
9:55	Muhammad Umar Qasim and Sam-Chin Li	Paper 6: Working Together Toward a Shared Vision : Canadian Government Information Digital Preservation Network (CGI DPN)
10:20	BREAK	Discussions, Networking
10:40	Nick Ruest	Paper 7: Strategies for Collecting, Processing, and Analyzing Tweets from Large Newsworthy Events
11:05	Saurabh Chakravarty, Eric Williamson, Edward Fox	Paper 8: Classification of Tweets using Augmented Training
11:30	Closing discussion	Plans for future activities and collaborations

Description:

- Selected works will likely be published in a special issue of IEEE TCDL Bulletin.
- This will explore the integration of web archiving and digital libraries, over the complete life cycle: creation/authoring, uploading/publishing in the Web, …
- It will cover all topics of interest, including but not limited to:

Archiving (events)	Big data	Classification, clustering
Client/proxy/server side	Crawling (focused)	Curation, quality control

collecting		
Databases / collections (of webpages)	Discovery	Extraction & analysis
Filling gaps	Globalization, languages	Social sciences
Linking archives	Metadata	Mobile devices
Network science	Preservation	Resource description
Standards, protocols	Systems, tools	Tweet collections and connections

Objectives:

- to continue to build the community of people integrating web archiving & DLs
- to help attendees learn about useful methods, systems, and software in this area
- to help chart future research and improved practice in this area
- to promote synergistic efforts including collaborative projects and proposals
- to produce an archival publication that will help advance technology and practice

Workshop Co-chairs:

- Chair, Edward A. Fox, Professor and Director Digital Library Research Laboratory, Virginia Tech, fox@vt.edu http://fox.cs.vt.edu, +1-540-231-5113
- Co-chair, Zhiwu Xie, Professor, Director of Digital Library Development, Virginia Tech Libraries, zhiwuxie@vt.edu, +1-540-231-4453
- Co-chair, Martin Klein, Los Alamos National Laboratory Research Library, mklein@lanl.gov, +1-505-667-5809

Program Committee:

- Jefferson Bailey, Internet Archive, jefferson@archive.org
- Mohamed Magdy Farag, Arab Academy for Science and Technology, mmagdy@aast.edu
- Joshua Finnell, Los Alamos National Laboratory, joshfinnell@lanl.gov
- Vinay Goel, Internet Archive, vinay@archive.org
- Andrea Goethals, Harvard Library, andrea_goethals@harvard.edu

- Gina Jones, Library of Congress, gjon@loc.gov
- Deborah Kempe, Frick Art Reference Library, kempe@frick.org
- Lauren Ko, University of North Texas Libraries, lauren.ko@unt.edu
- Frank McCown, Harding University, fmccown@harding.edu
- Michael Nelson, Old Dominion Univ., mln@cs.odu.edu
- Thomas Risse, L3S Research Center, Leibniz Universitat Hannover, risse@L3S.de
- Nicholas Taylor, Stanford U. Libraries, ntay@stanford.edu
- Matthew Weber, Rutgers U., matthew.weber@rutgers.edu

Closely related events and results:

- Web Archiving and Digital Libraries (WADL'15), 24 June, at JCDL 2015, see website and proceedings in a special issue of the IEEE TCDL Bulletin, V. 11, Issue 2, Oct. 2015
- Working with Internet Archives for Research (WIRE 2014) NSF workshop, 17-18 June 2014, Cambridge, MA – see http://wp.comminfo.rutgers.edu/nsfia/
- Web Archiving and Digital Libraries (WADL'13), 25-26 July, at JCDL 2013, see http://www.ctrnet.net/sites/default/files/JCDL2013WorkshopWebArchiving20130603.pdf and report in SIGIR Forum http://sigir.org/files/forum/2013D/p128.pdf
- Web Archive Globalization Workshop, WAG 2011 – see http://cs.harding.edu/wag2011/ , with 4 organizers plus 5 presenters and about 20 participants, held in Ottawa after JCDL 2011 (June 16-17)
- Ongoing work by attendees in this area, growth in collaborative activity involving the Internet Archive, and specific community building successes like the Web Archive Cooperative – see http://infolab.stanford.edu/wac/
- Annual meetings of the International Internet Preservation Consortium (IIPC), partner meetings of the Internet Archive (Archive-It), and ten workshops held with ECDL/TPDL: International Web Archiving Workshop (IWAW), 2001-2010

OLD: Submissions (please provide contact and supporting info in <= 1 page):

- EasyChair submission page: https://easychair.org/conferences/?conf=wadl2017
- Due: April 24, 2017
- Notifications: May 13, 2017
- **Categories** (pick one of the 3) are:
- Poster/Demonstration + lightening talk
- 20 min. presentation + 10 min. Q&A
- 30 min. panel with interactive plenary discussion

WADL Submission 2017

Author Name: Ross Spencer

Job Title: Digital Preservation Analyst, Archives New Zealand

Submission Type: Poster/Demo + Lightning talk

Poster Title: HTTPreserve – Web Preservation in Documentary Heritage

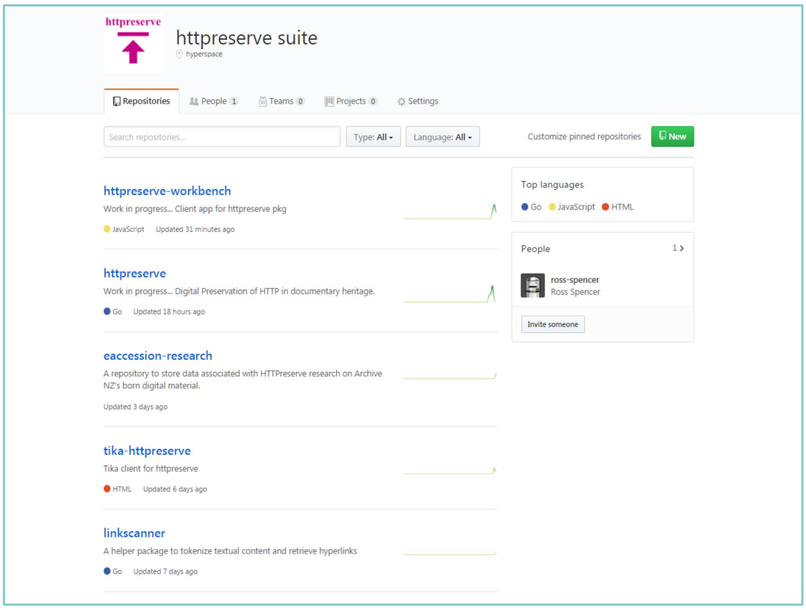
Summary:

The paper paradigm is still prevalent in the government archive and the government agency. It will be for a good number of years yet as they wrestle with appropriate time-scales for the disposal of records (destruction, or transfer to an archive). Most records in most jurisdictions will be transferred between 20 and 30 years.

The web, much older than these timescales and will exist in one form or another in the records either end.

This raises questions as to whether a government agency can preserve the public record in its completeness. A hyperlink that provides evidence on how a decision was made may not be transferred through normal practices and will instead need to be archived and monitored through mechanisms developed by the web-archiving community. Some links may already be lost. Some may only exist in places like The Internet Archive. We do not have adequate tools or workflows to understand the extent of bit-rot that may exist in agencies records.

This poster and demo will demonstrate a newly developed suite of tools called HTTPreserve (https://github.com/httpreserve) that may fill the gap that helps government agencies, and other archives to develop the appropriate processes for preservation of links in documentary heritage.



httpreserve-workbench
Work in progress... Client app for httpreserve pkg
● JavaScript Updated 31 minutes ago

httpreserve
Work in progress... Digital Preservation of HTTP in documentary heritage.
● Go Updated 18 hours ago

eaccession-research
A repository to store data associated with HTTPreserve research on Archive NZ's born digital material.
Updated 3 days ago

tika-httpreserve
Tika client for httpreserve
● HTML Updated 6 days ago

linkscanner
A helper package to tokenize textual content and retrieve hyperlinks
● Go Updated 7 days ago

Figure 1: HTTPreserve Suite on GitHub.com

WARC-Portal: A Tool for Exploring the Past

Muhammad Umar Qasim
University of Alberta Libraries
Edmonton, Alberta Canada
umar.qasim@ualberta.ca

Abstract— The World Wide Web is a significant source of information for researchers and information seekers. This valuable information medium, however, is highly volatile in nature and any existing content can easily be updated, altered or removed. Web archiving allows for taking snapshots of web sources and archiving them for later use when the original may no longer be available. Information professionals and researchers are increasingly involved with creating web archiving collections. Numerous open source tools are available that support web archiving capture and preview functions, but provide very limited support for analyzing the captured content. The WARC-Portal project aims to deal with extracting, searching and analyzing Web ARChive (WARC) files. The project provides intuitive and easy access for researchers to browse and search through custom collections, and provides tools for analyzing these collections. These tools include an array of search features and filters, as well as helpful visualizations based on the selected content. The tool presents the web archiving data in an intuitive way that will help researchers and information professionals to find patterns and trends. WARC-Portal is intended to support digital humanists, information professionals and social science researchers in their web archiving activities.

Keywords—**Web Archiving; Visualization; World Wide Web;**

INTRODUCTION

Web archiving is the process of taking a snapshot from the live web, storing it, and making the collected content available for future use.

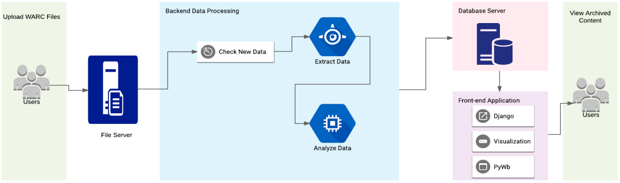


Figure 1: WARC-Portal Architecture

Besides many other benefits, web archiving presents a huge potential for researchers to mine and extract important information. Techniques such as, text mining, link analysis, trends analysis and geographical analysis & mapping can be used to identify or confirm important facts from archived content [1]. The most commonly and widely used file format for web archiving is warc[2]. The WARC (Web ARChive) format is a method for combining multiple digital resources into an aggregate archival file together with related information [3]. There are numerous tools available to create warc files and to view the contents of these files. However, these available tools provide limited support for researchers to do an in-depth analysis of any specific warc files. There is a need for a platform which can help individual researchers create collections, ingest warc files and then analyze these warc files with limited to no knowledge of coding.

The WARC-Portal project aims to deal with extracting, searching and analyzing warc files. It is intended to provide a platform that helps researchers to create collections, ingest warc files and then to do an in-depth analysis of selected files. Figure 1 shows an architectural view of the portal with backend processing, database and front-end components. In addition to platforms's own functionality, the portal utilizes some of the already available tools such as warcbase and IBM Watson analytics to support an extended set of features. This poster presentation will provide a sneak preview of this platform.

ACKNOWLEDGMENT

This work would not have been possible without the development work done by Adriano Marini, Mate Verunica, Cheng Chen, and Kevin Tang. I would also like to thank Todd Suomela for helping with defining the requirements of this project. Special thanks to the warcbase team for the development of warcbase tool and making it available as an open source platform.

REFERENCES

[1] Emily Reynolds. A review of contemporary research use cases for web archives. Library of Congress, UMSI, ASB13, 2013. http://netpreserve.org/sites/default/files/resources/UseCases_Final_1.pdf

[2] Marta Teperek and Danny Kingsley. Archiving Webpages – securing the digital discource. Unlocking Research, University of Cambridge. 2015 https://unlockingresearch.blog.lib.cam.ac.uk/?tag=warc

[3] Library of Congress. WARC, Web Archive file format. Sustainablity of Digital Formats: Planning for Library of Congress Collections. https://www.loc.gov/preservation/digital/formats/fdd/fdd000236.shtml

Impact of URI Canonicalization on Memento Count

Mat Kelly, Lulwah M. Alkwai, Sawood Alam,
Michael L. Nelson, and Michele C. Weigle
Old Dominion University
Department of Computer Science
Norfolk, Virginia, USA
{mkelly,lalkwai,salam,mln,mweigle}@cs.odu.edu

Herbert Van de Sompel
Los Alamos National Laboratory
Los Alamos, New Mexico, USA
herbertv@lanl.gov

Web archives return Memento TimeMaps with a list of URI-Ms for the HTTP transactions observed at archival time. TimeMaps have generally been used as a count of the number of representations of a URI-R present in an archive. However, TimeMaps may include URI-Ms for archived representations, redirections, and errors (See RFC7231). For example, 57% of the URI-Ms for http://vimeo.com produce an HTTP Redirect to another URI-M in the TimeMap that returns an HTTP Status OK. TimeMaps do not explicitly return a "count" value to indicate the number of mementos listed in the TimeMap that produce a non-redirecting HTTP status code when dereferenced. The heuristic of determining how many captures are represented by URI-Ms in a TimeMap cannot be completed without dereferencing.

Redirection in a Web archive can be attributed to a variety of canonicalization rules [1]. Preserving and replaying these redirects allows an archive to accurately reproduce the HTTP transactions that would have occurred when the URI being accessed resided on the live Web. Because of the potential for redirection, the heuristic of counting URI-Ms with Link relation values of "memento" is an inaccurate means of determining the number of unique representations inferred from a TimeMap. We further emphasize the distinction per the Memento specification that the identifiers for mementos (URI-Ms) in a TimeMap are identifiers for archived HTTP transactions (e.g., transmission of HTTP 2XX, 3XX, 4XX, etc.) rather than identifiers for representations.

Based on the number of URI-Ms in a TimeMap not necessarily resolving to unique mementos when archival redirects are followed, we examined the mementos from contemporarily large TimeMaps to evaluate the patterns and schemes used in Memento canonicalization. Through this, we identify the difference between the number of mementos available as reported by the TimeMap through naive "rel" counting heuristics to the temporally unique mementos identified once these mementos are dereferenced.

Table 1: Google over time (abbreviated), bucketed by year, based on IA mementos extracted from the TimeMap. M_{TM} is the memento count based solely on the data in the TimeMap, M_{RC} is the count based on exclusion of redirects when dereferenced, and DI is the ratio of non-redirecting mementos to redirecting mementos.

year	M_{TM}	M_{RC}	DI
2006	735	483	1.917
2007	1,055	842	3.953
2008	1,376	894	1.855
2009	6,074	4,335	2.493
2010	9,326	6,530	2.335
2011	20,634	9,279	0.817
2012	102,533	16,240	0.188
2013	228,405	25,203	0.124
2014	164,865	22,738	0.160
2015	17,978	11,286	1.686
2016	139,520	5,805	0.043

URI canonicalization associates differently formatted URIs (See RFC6596) and allows after-the-fact clustering of URIs that likely reference the same resource. As URI schemes from a Web site change over time, canonicalization is critical for retaining a cohesive, comprehensive listing of the mementos available for a Web page.

Table 2: Scheme and subdomain for redirects when dereferencing URI-Ms for google.com.

	dest	http			https		
orig		none	www	other	none	www	other
http	none	1,279	68,837	55	12	20,825	27
	www	8,934	490,836	204	32	77,610	16
	other	0	224	22	0	26	2
https	none	14	731	0	0	296	1
	www	1,117	72,874	27	15	18,525	2,101
	other	0	0	0	0	0	0

Table 3: Dereferencing 7 other large Web sites' TimeMaps from Internet Archive produces the above distribution of status codes for each site.

host	% 3XX	% 200	M_{TM}	DI
google	84.89	15.11	695,525	0.178
yahoo	88.16	11.83	418,896	0.134
sourceforge	73.34	26.63	31,408	0.363
instagram	67.32	32.65	55,228	0.485
vimeo	57.04	42.94	199,262	0.752
cnn	49.97	50.01	87,148	1.001
wikipedia	44.62	55.19	25,973	1.240
whitehouse	44.57	55.24	26,006	1.243

We acquired HTTP response headers for all URI-Ms accumulated from multiple Web archives for various URI-Rs. We obtained a TimeMap for google.com from our locally deployed Memento aggregator containing 714,470 URI-Ms from 8 different Memento-compliant archives. 89.1% of the URI-Ms returned were from Internet Archive.

Equation 1 excludes mementos that resolve to HTTP 3XX status codes. $|TM|_D$ represents the count of mementos that result in non-3XX statuses based on the URI-Ms in a TimeMap.

$$|TM|_D = \sum_{m=1}^{len(M)} \begin{cases} 0 & 300 \geq httpStatus(m) < 400, \\ 1 & \text{otherwise.} \end{cases} \tag{1}$$

$$|TM|_I = |TM|_{rel} - |TM|_D \tag{2}$$

We quantify the ratio of mementos with non-redirecting HTTP status codes (Equation 1) to those with redirects (Equation 2) in Equation 3 as DI.

$$DI = \begin{cases} \frac{|TM|_D}{|TM|_I} & |TM|_I > 0, \\ \infty & \text{otherwise.} \end{cases} \tag{3}$$

If an archive reports revisit records as an HTTP redirect based on the CDX listing, and this redirect is propagated to the archive's Memento endpoint thus producing a unique URI-M, the DI's value for the URI-R decreases. Requesting the URI-M using the Accept-Datetime HTTP header then observing the Memento-Datetime response header's presence often reveals this nuance, but by relying on the TimeMap data without requesting each URI-M, the DI for the URI-R is unknown. Table 2 shows the scheme and subdomain breakdown for google.com URI-Ms where a redirect is encountered upon dereferencing the URI-M. Table 3 shows the DI values based on the TimeMaps of 8 large Web sites.

REFERENCES

[1] Mat Kelly, Lulwah M. Alkwai, Michael L. Nelson, Michele C. Weigle, and Herbert Van de Sompel. 2017. Impact of URI Canonicalization on Memento Count. (March 2017). arXiv:1703.03302

Web Archiving Through In-Memory Page Cache

Saket Vishwasrao[1], Zhiwu Xie[2], and Edward A. Fox[3]
[1]Department of Electrical & Computer Engineering, [2]University Libraries, [3]Department of Computer Science
Virginia Polytechnic Institute and State University
Blacksburg, VA 24061
{saketv02, zhiwuxie, fox}@vt.edu

1. INTRODUCTION

Traditional web archiving acquisition methods include client-side, transactional, and server-side archiving [1]. Different from these approaches, nearline archiving [2] acquires content from web cache, therefore can run asynchronously from the HTTP request - response rhythms and bypass the server load peaks.

In this paper we introduce a new implementation of nearline archiving. Instead of assuming a web server will always use a file cache to store cached content in the file system, our approach leverages the widely used in-memory cache.

2. IN-MEMORY PAGE CACHE

Many in-memory cache products exist. memcached[1] and Redis[2] are perhaps the two most widely adopted ones. They may be used in various stages of the web stack, but we assume a page cache exists to store the full HTTP responses to be sent back to the clients from the origin server. Acquiring web content from the page cache therefore leads to similar effects as transactional web archiving.

However, reading cache content from the memory poses new challenges. Unlike in the case of the file system, most operating systems designate separate address spaces for each process and discourage mixing up. It is of course possible to hack the operating system then dump another process's memory, but this is generally frowned upon. Accessing the address space from the same process as the cache writing requires code modifications of either the web server or the application server. For in-memory cache products that allow API access, e.g., memcached, it is also possible to leverage their lenient security settings to allow multiple client processes to access each other's address spaces. We therefore can build a small application that constantly reads out the web cache. All these approaches, nevertheless, raise security concerns.

Learning the lessons from transactional web archiving, we intend to minimize the code modification of existing software. This is because this type of content acquisition requires the cooperation of the website owner and operator, yet it is difficult to convince them to install modified code that voids the "official support", whatever that means.

If we are not to modify existing code, all we can leverage are the existing features and their configurations. Fortunately we discovered a Redis feature that could be used for our purpose.

Redis has a built in persistence mechanism. This feature is added to Redis in order to ensure that the data stored in memory is not lost during failure. To implement persistence, Redis forks a low priority child process that writes stored data to the disk, and it is possible to configure Redis in the way that it writes Append Only File (AOF) to disk. The AOF file contains all cache modification since the last persistence. We can also configure the system such that the persistence happens on each cache update, each second, or without a deadline. The last option usually falls back to the default Linux kernel configuration, and will write every 30 seconds if no high priority process prevents it from happening.

3. IMPLEMENTATION STEPS

We implemented this idea on a Wikipedia mirror. Figure 1 give the system architecture. We use Nginx as the webserver, redis as both the page cache and the object cache, mediawiki for rendering dynamic pages, and MySQL as the backend database.

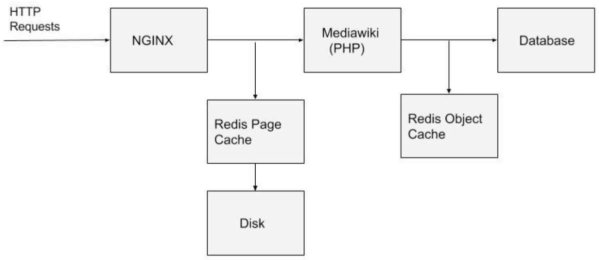


Figure 1. System Architecture

With a moderate page cache size, our initial results show moderate performance degradation in terms of both the latency and the highest number of requests the server can sustain. If we increase the size of the in-memory cache, the performance degradation is reduced, and eventually diminishes.

4. REFERENCES

[1] Masanès, J. 2006. *Web Archiving*. Berlin: Springer.

[2] Xie, Z., Nayyar, K., Fox, E.A. 2017. Nearline Web Archiving, *Bulletin of IEEE Technical Committee on Digital Libraries*, 13(1).

5. ACKNOWLEDGMENTS

This work was partially supported by Institute of Museum and Library Services award LG-71-16-0037-16.

[1] http:// memcached.org

[2] http://redis.io

Building a National Web Archiving Collaborative Platform: The Web Archives for Longitudinal Knowledge Project

Ian Milligan (Waterloo), Nick Ruest (York), and Ryan Deschamps (Waterloo)

In the absence of a national web archiving strategy, Canadian governments, universities, and cultural heritage institutions have pursued disparate web archival collecting strategies. Carried out generally through contracts with the Internet Archive's Archive-It services, these medium-sized collections amount to a significant portion of Canada's born-digital cultural heritage since 2005. While there has been some collaboration between institutions, notably via the Council of Prairie and Pacific University Libraries in western Canada, most web archiving collecting has been taking place in silos. Researchers seeking to use web archives in Canada are thus limited not only to the Archive-It search portal, but also to exploring on a silo-ed collection-by-collection basis. Given the growing importance of web archives for scholarly research, our project breaks down silos and generate a common search portal and derivative dataset provider.

Our Web Archiving for Longitudinal Knowledge (WALK) Project, housed at http://webarchives.ca/and with our main activity via our GitHub repo at https://github.com/web-archive-group/WALK, has been bringing together Canadian partners to integrate web archival collections. Co-directed by a historian and a librarian, the project brings together computer scientists working on the warcbase project, doctoral students working on governance issues, and students running tests and usability improvements. We currently have ~20TB of web archival collections, aggregated from the Universities of Toronto, Alberta, Winnipeg, and Victoria, as well as Dalhousie and Simon Fraser University. Our workflow consists of:

- Signing Memorandum of Agreements (MOU) with partner institutions;
- Gathering WARCs from partner institutions into ComputeCanada infrastructure through the Research Portals and Projects (RPP) program;
- Using warcbase (http://warcbase.org/) to generate scholarly derivatives, such as domain counts, link graphs, and files that can be loaded into network analysis software.[1]
- Adapting the Blacklight front end (http://projectblacklight.org/) to serve as a replacement for our current SHINE interface; this will allow built-in APIs, faceted search by institution, and inter-operability with university library catalogues.[2]
- Using a team of research assistants to describe each collection using Python and R.
- Finally, using multiple correspondence analysis, generating profiles of each web archive with an eye towards assisting curators in finding collection overlap/gaps.[3]

This presentation provides an overview of the WALK project, focusing specifically on questions of interdisciplinary collaboration, workflow, dataset creation and dissemination. As web archiving increasingly happens at the institutional level, the WALK project suggests one way forward towards collaboration, collection development, and researcher access.

References

[1] Lin, Milligan, Wiebe, and Zhou. "Warcbase: Scalable Analytics Infrastructure for Exploring Web Archives," ACM Journal of Computing and Cultural Heritage. Accepted w/ minor revisions.

[2] Jackson, Lin, Milligan, and Ruest, "Desiderata for Exploratory Search Interfaces to Web Archives in Support of Scholarly Activities," Proceedings of the ACM/IEEE Joint Conference on Digital Libraries 2016, vol. 16: 103-106.

[3] Milligan, Ruest, and Lin, "Content Selection and Curation for Web Archiving: The Gatekeepers vs. the Masses," Proceedings of the ACM/IEEE JCDL 2016, vol. 16: 107-110.

Avoiding Zombies in Archival Replay Using ServiceWorker

Sawood Alam, Mat Kelly, Michele C. Weigle, and Michael L. Nelson
Department of Computer Science, Old Dominion University
Norfolk, Virginia, USA - 23529
{salam,mkelly,mweigle,mln}@cs.odu.edu

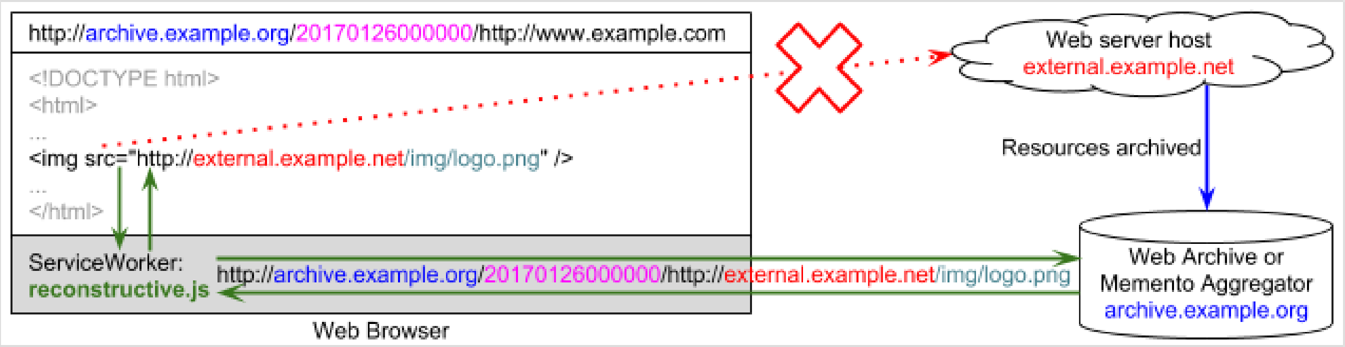

http://archive.example.org/20170126000000/http://www.example.com
<!DOCTYPE html>
<html>
...
<img src="http://external.example.net/img/logo.png" />
...
</html>
ServiceWorker:
reconstructive.js
http://archive.example.org/20170126000000/http://external.example.net/img/logo.png
Web Browser
Web server host
external.example.net
Resources archived
Web Archive or
Memento Aggregator
archive.example.org

Figure 1: ServiceWorker reconstructive.js intercepts a zombie resource and reroutes to its archived copy.

A *Composite Memento* is an archived representation of a web page with all the page requisites such as images and stylesheets. All embedded resources have their own URIs, hence, they are archived independently. For a meaningful archival replay, it is important to load all the page requisites from the archive within the temporal neighborhood of the base HTML page. To achieve this goal, archival replay systems try to rewrite all the resource references to appropriate archived versions before serving HTML, CSS, or JS. However, an effective server-side URL rewriting is difficult when URLs are generated dynamically using JavaScript. A failure of correct URL rewriting might yield an invalid/unintended URI or resolve to a live resource. Such live resources, leaking in a composite memento, are called "*zombies*".

ServiceWorker (SW) is a new client-side web API [2] that can be used to intercept all the network requests for embedded resources originating from web pages in its scope. We use SW API to reconstruct composite mementos from the originally captured data without any URL rewriting. By intercepting requests on the client-side, we are essentially rerouting instead of rewriting. Rerouting is an effective mechanism to block zombies, as it takes effect when the user-agent resolves a reference. Figure 1 illustrates how our SW implementation, `reconstructive.js`[1], intercepts a live leakage and reroutes it correctly to the corresponding archived copy.

To evaluate the archival replay reconstruction quality we created the Archival Capture Replay Test Suite (ACRTS)[2] with different scenarios of how a web page might initiate a network request. We archived ACRTS and saved the resulting Web ARChive (WARC) file. We then changed the live ACRTS site in a way that all the resource references remained the same, but their content was changed. Using various replay systems we loaded the archived ACRTS from the stored WARC file. Depending on how effective the replay system is, it might load resources from the archive (☑), leak from the live site (☒), or not load at all (□). The latter might happen either because the requested resource was not present in the archive or the replay system resolved the location incorrectly. Table 1 shows how well each of the listed archival replay systems reconstructs a composite memento when resource requests are originated from different conditions. A more extensive description and evaluation of this work is published in Alam et al. [1].

Table 1: URL Rewriting/Rerouting Results in Different Archival Replay Systems (A: OpenWayback, B: PyWB, C: Memento Reconstruct, D: Memento for Chrome, and E: Reconstructive)

Resource Loading Scenarios	A	B	C	D	E
Relative path	☑	☑	☑	☑	☑
Absolute rooted path	☑	☑	☑	☑	☑
Absolute local URL	☑	☑	☑	☑	☑
Absolute external URL	☑	☑	☑	☑	☑
External resource from an external iframe	☑	☑	☑	☑	☑
Loaded by an inline CSS	☑	☑	☑	☑	☑
Loaded by a CSS file	☑	☑	☑	☑	☑
Loaded by CSS @font-face	☑	☑	□	☑	☑
Loaded by image srcset	☑	☑	☑	☑	☑
Added by an inline JS on page load	☒	☑	☑	☒	☑
Added by an inline JS on page scroll	☒	☑	□	☒	☑
Added by an inline JS on click	☒	☑	☑	☒	☑
Added by a JS file	☒	☑	☑	☒	☑
Added by an Ajax request	☒	☑	□	☒	☑

REFERENCES

[1] Sawood Alam, Mat Kelly, Michele C. Weigle, and Michael L. Nelson. 2017. Client-side Reconstruction of Composite Mementos Using ServiceWorker. In *Proceedings of the 17th ACM/IEEE-CS Joint Conference on Digital Libraries (JCDL '17)*.

[2] Alex Russell, Jungkee Song, and Jake Archibald. 2015. Service Workers. https://www.w3.org/TR/service-workers/

[1] https://github.com/oduwsdl/reconstructive

[2] https://ibnesayeed.github.io/acrts/

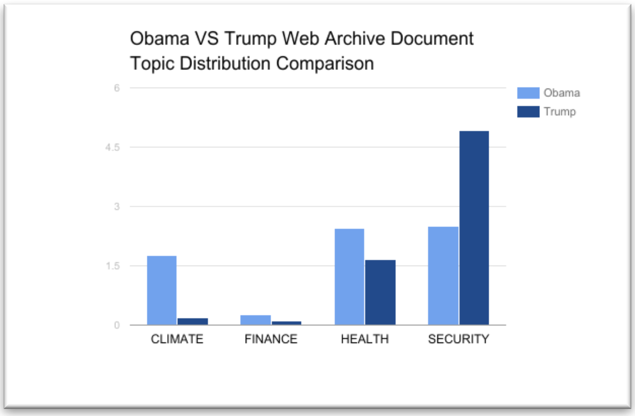
Topic Shifts Between Two US Presidential Administrations

Ziquan Wang, Borui Lin, Ian Milligan, Jimmy Lin

Historians cannot write the history of the 1990s without the Web.[1] Accordingly, web archives such as those collected by the Internet Archive will become a critically important source for learning, understanding, and analyzing recent history. However, the scope of web archives means that historians will not be able to read pages one at a time – instead, they will need to turn to distant reading techniques in order to explore the information within a large number of web pages in a timely manner. A pivotal research area within the web archival field is topic extraction.[2], [3] In this presentation, we discuss our work with text classification on two crawls: the first from President Barack Obama's end-of-term *whitehouse.gov* site, and the second from President Donald Trump's beginning-of-term *whitehouse.gov*.

The two datasets present an interesting contrast, especially in that the Obama administration's site is considerably larger than Trump's (319.6GB vs 12.9GB). We downloaded each crawl from the Internet Archive. For each page within each *whitehouse.gov* domain, we removed boilerplate text such as HTTP headers and HTML tags and extracted the main text content to form a document. Each document was then split into sentences using a pre-built PTBTokenizer from the Stanford CoreNLP package. This generated a large set of unlabeled sentences. Each sentence was then categorized into either CLIMATE, FINANCE, SECURITY, HEALTH, or UNCATEGORIZED. We summarized a list of keywords for each of the four categorized topics by reading sample articles from *whitehouse.gov* and used Spark to examine each sentence and assign a label in parallel based on majority votes of keyword occurrences from each class. We subsequently used this as a seed in our work to create a model to predict further sentences that belong to those topics but did not possess the keyword. If a given page contained no sentences that were categorized as above, the document was more likely to be uncategorized; if it possessed a hit, more sentences are likely to be categorized. We then created a dataset with 306,676 labelled sentences with 18% being categorized and 82% being uncategorized, splitting it into 80% and 20% parts for training and testing. We then used a bag-of-words model to convert sentences into feature vectors, and experimented with Linear Classifier and Naïve Bayes Classifier, finding that the former clearly outperformed the latter.

We then applied this training model to predict the topic of documents in parallel. As seen in the chart, there is an overwhelmingly high percentage of security-related web pages in the Trump web archive. Besides security, percentage of health-related web pages is also very high. This can be attributed to President Trump's desire to replace the *Affordable Care Act* with his *American Health Care* Act. Similarly, the number of climate-related pages in the Trump web archive has fallen.



This paper presents our research of using a Big Data framework such as MapReduce/Spark and Machine Learning classifier library such as the Stanford CoreNLP library to do text classification on *whitehouse.gov* web archives. We have successfully built a model that can well predict the topic (from a pre-defined topic list) of a page.

[1] I. Milligan, "Lost in the Infinite Archive: The Promise and Pitfalls of Web Archives," *Int. J. Humanit. Arts Comput.*, vol. 10, no. 1, pp. 78–94, Mar. 2016.

[2] Y. AlNoamany, M. C. Weigle, and M. L. Nelson, "Detecting off-topic pages within TimeMaps in Web archives," *Int. J. Digit. Libr.*, vol. 17, no. 3, pp. 203–221, Sep. 2016.

[3] G. Gossen, E. Demidova, and T. Risse, "Analyzing Web Archives Through Topic and Event Focused Sub-collections," in *Proceedings of the 8th ACM Conference on Web Science*, New York, NY, USA, 2016, pp. 291–295.

Web archives: A preliminary exploration of user expectations vs. reality

Brenda Reyes Ayala
University of North Texas, Department of Information Science
3940 North Elm, Suite C232
Denton, Texas 76203
Brenda.Reyes@unt.edu

Keywords

Web archiving; Digital Libraries; Grounded Theory; User Experience; User Expectations; Mental models;

1. INTRODUCTION

In the field of Human-Computer Interaction, humans are thought to construct *mental models* in order to understand the world around them. Mental models are internal representations of the world that help humans describe and understand their surroundings[2]. When humans interact with technology, it is common for them to form a mental model of the target system, that is, the system the user is learning or using [4]. Users' mental models often do not match the structure and features of an actual information system; this mismatch often leads to confusion and often unmet expectations. This paper examines how users perceive the process of web archiving and attempts to answer the following questions: What expectations do users and creators of web archives bring to web archives? How does the reality of web archiving differ from their expectations? The purpose of this paper is to present the results of a preliminary exploration of these topics.

2. METHODOLOGY

Data gathering and collection for this project came about as part of my dissertation work, which examines how users and creators of web archives perceive the notion of information quality. In order to carry it out, I collected and analyzed support tickets that had been submitted to the Internet Archive's Archive-It (AIT) service. The tickets collected were Level 1 support tickets that had been submitted by AIT clients over the course of several years. Since these tickets belonged to the Internet Archive, I negotiated a researcher agreement with the organization in order to obtain a cache of tickets.

The tickets received comprised 4,281 tickets from the years 2012 through 2016. In order to better analyze their content, they were put through extensive pre-processing in the form of several Python programs and Linux command-line scripts written by the researcher. After the tickets were cleaned, they were imported into the NVivo software package, a popular software for qualitative data analysis [3]. This research focused on tickets in which the client discusses a perceived flaw in an individual archived website or an entire web archive. At the time of writing, I have classified 128 AIT tickets.

The AIT tickets were analyzed using the Grounded Theory (GT) methodology. GT, created by Barney Glaser and Anselm Strauss, is an inductive methodology create explicitly for generating theory in the Social Sciences: working closely from the data, the researcher begins the work of generating a theory [1]. During data analysis, the researcher engages in coding. Coding allows the researcher to discover what is happening in the data and to grapple with what it means. Coding data in GT allows for *categories*, or conceptual elements to emerge.

3. FINDINGS AND DISCUSSION

As I analyzed the support tickets to find users' notion of information quality, another separate issue rose to the surface again and again: a continuing mismatch between the way the users perceived the process of web archiving and the concept of a web archive itself, and the ways both of these actually worked. The categories dealing with the mismatch between user expectations of how web archives worked and the reality are "Size of the crawl, archived website, or web archive", which appears 44 times in 29 tickets, and "Relevance", which appears 42 times across 25 tickets. The findings can be summarized using the following themes:

3.1 Organized vs. unorganized

Archive-It users often assumed that a website had specific size which would also be reflected in the archived site. Since the original website had X number of documents, it would also follow that the archived website also had X number of documents. However, the tickets analyzed showed that the reality did not reflect their mental model. Often, an archived website was much larger or smaller than the user had expected. This is illustrated in some of the following tickets submitted by users:

- "The crawl took 12 hours and returned 103,173 documents and 3.1GB of data. This can not be correct. Crawling the whole —.edu domain with my constraints yields 20,300 +- docs."
- "There are only 170 photos on this site but I ended up

JCDL 2017 June 19–23, 2017, Toronto, ON, Canada
© 2017 ACM. ISBN 978-1-4503-2138-9.
DOI: 10.1145/1235

with 15K new URLs."

- "We know there are many, maybe in the hundreds of .pdf files on the site and not one was captured."

As can be seen from the examples, the users, would often compare the actual size of the archived website to their mental model of the website's size. Any disparity was often cause for concern.

3.2 Self-contained vs. connected

Archive-It users implicitly assumed that a web archive would only include content that was closely related to that of the larger web archive. In reality, due to crawler settings, scoping rules, and the nature of the web, web archives often include content that is not topic-specific. This was specially the case with social media sites. Users saw the presence of this content as being of little relevance and superfluous:

- "The problem is, that a lot of unrelated content is being displayed: sites we are not supposed to have in our collection, social network pages like xing and facebook, porn- and dating sites, some of them even with illegal content, and so on."

- "Noticed that we captured a message board that has a lot of unwanted garbage posted on it."

In some cases, users would flag content as irrelevant or unwanted when it was actually necessary to preserve the functionality of archived pages. A website contains pages, or elements that are not obviously important but help "behind the scenes" to make other elements or pages render correctly or function properly. This is knowledge that is known by the partner specialist, but usually unknown or invisible to the user or creator of an archive. Partner specialists often had to explain the true nature of this seemingly irrelevant content:

- "The host *ytimg.com* serves code that effects the layout and client-side functionality of YouTube content."

- "It looks like there are a fair number of URLs for different sizes of the same image. If you aren't interested in those different sized images."

Closely related to this idea was the users' mental models of how web domains work. Users did not draw fine distinctions (or any distinction at all) between the concepts of domain and sub-domain in URLs, and how this would affect the capture of a website. For example, users did not differentiate between a website address such as http://www.stateu.edu (part of the *stateu.edu* domain) and a subdomain such as http://admissions.stateu.edu or http://libraryu.state.edu. Accordingly, they expressed surprise when they did not find this content had been captured by the crawler, as they assumed it would be automatically included. The AIT partner specialists had to explain these differences often:

- "It looks like your State University collection has one seed, www.stateu.edu. Since english.stateu.edu is a subdomain of www.stateu.edu it is not automatically in scope for that seed, but you can easily capture those links by adding english.stateu.edu as a seed and crawling it along with www.stateu.edu."

- "First, there were two crawls of this site on August 20 and one on August 22, so that accounts for three of the capture dates. Also, for each of those three crawls the crawler found links to three slightly different urls: http://www.library.une.edu/, http://www.library.une.edu/# and http://library.une.edu/ (no www). Technically speaking these are different URLs so each version was captured."

As the data shows, AIT users form mental models of websites and web archives that are often at odds with the realities of web archiving. Many users regard websites as organized, self-contained entities, not realizing the ad-hoc and interconnected nature of the web. They make the implicit assumption, which is not generally articulated, that the archived website should also match their mental model of the original website. The process of creating a web archive force people to deal in technicalities they may never have thought of before. Their misconceptions about what a web archive is also reveal users' misconceptions about what a website is and how it works.

4. ACKNOWLEDGMENTS

I would like to thank Lori Donovan and Jefferson Bailey of the Internet Archive, without whose help this research would not have been possible.

5. REFERENCES

[1] B. Glaser and A. Strauss. *The Discovery of Grounded Theory: Strategies for Qualitative Research.* Aldine Transaction, 1967.

[2] P. N. Johnson-Laird. Mental models and human reasoning. *Proceedings of the National Academy of Sciences of the United States of America*, 107(43):18243–18250, 2010.

[3] QSR International. Nvivo product range. http://www.qsrinternational.com/nvivo-product, 2016.

[4] N. Staggers and A. Norcio. Mental models: concepts for human-computer interaction research. *International Journal of Man-Machine Studies*, 38(4):587 – 605, 1993.

Challenges for Grassroots Web Archiving of Environmental Data

Communities and individuals outside of institutional digital library settings are increasingly concerned with preserving contemporary knowledge and culture accessed through networked infrastructures. New challenges arise for the digital libraries and web archiving communities to connect and collaborate with groups and individuals adopting open-source archiving tools for uses like data-driven forms of activism and the civic technology movement that seeks to understand and address civic challenges with technology.

In the wake of the US presidential election in November, more than 30 DataRescue and Guerilla Archiving events have sprung up across the United States and Canada. One activity at these events was nominating 'deep' federal agency web pages to Internet Archive's (IA) "End-of-Term" harvest, contributing to a ten times increase in seed nominations for that crawl. In addition, specific web pages and datasets were investigated and those seen as difficult to preserve through crawling were scraped and downloaded through automated scripts. These datasets were then uploaded to an online repository maintained at the University of Pennsylvania by the DataRefuge project, complete with descriptive and technical metadata about how they were created.

The Environmental Data and Governance Initiative (EDGI) was one of the facilitators for these events, and an example of a community provoked to undertake grassroots web archiving to support their ongoing data needs. EDGI began on Nov 11, 2016 as an email sent to 14 researchers, and has grown rapidly to an international network of some 65 people from 21 academic and non-profit institutions. EDGI seeks to address potential political threats to federal policy and scientific research infrastructure around environment and energy. In order to do so, over the last four months EDGI has built online tools, research networks, and hosted public events to proactively archive public environmental data and ensure its continued public availability.

We propose a panel discussion to address the cross-disciplinary issues of EDGI's work, engaging the web archiving community to address challenges to capture of specific data sets and formats, managing distributed web archival infrastructures, coordinating data and metadata cross diverse volunteer initiatives, and advocating to data publishers about important considerations for data preservation. We will have specific questions to engage the audience driven by EDGI member participation in the Archives Unleashed 4.0 Datathon.

Panel

- Facilitator: Emily Maemura, PhD Student, Digital Curation Institute
  Faculty of Information, University of Toronto
- Dawn Walker, PhD Student, Digital Curation Institute, EDGI Archiving Working Group
  Faculty of Information, University of Toronto
- Matt Price, EDGI Technical Lead
  Department of History, University of Toronto
- Maya Anjur-Dietrich, PhD Student, EDGI Archiving and Website Monitoring Working Groups
  Harvard University

Joint Conference on Digital Libraries 2017

Web Archives and Digital Libraries 2017

Abstract from: Tom Smyth, Library and Archives Canada (tom.smyth@canada.ca).
Suggested format: 20 min. presentation + 10 min. Q&A.

Tentative title:
Legal Deposit, Collection Development, Preservation, and Web Archiving at Library and Archives Canada

Library and Archives Canada (LAC) has conducted web archiving under two separate instruments within our federal legislation since 2005: within the context of national legal deposit, and under a separate power which enables LAC to collect the Web 'for [digital] preservation purposes'.

Owing to this library-centric approach to the Web at LAC, considerable legal deposit and digital library collection development thinking has been applied to our Web Archiving Program since 2013. This paper will provide an overview of LAC's web archival curation methodology, including our six primary web archiving activities; the current state of our program framework (including how we define web archival collection scopes; capturing social media relevant to larger collection themes; practical tools to govern quality control and creation of specialized metadata schemas for description); along with some thoughts on curating web archives as datasets for future digital humanities-oriented use -- all drawing on practical examples of web archival projects undertaken at LAC in the last two years. Highlighted collections will include our collaborative efforts at documenting the Truth and Reconciliation Commission in Canada; perspectives on First World War Commemoration and Vimy Ridge 100; the Rio 2016 Olympics; and celebrating the 150th anniversary of Canadian confederation ("Canada 150"). Adventures in web archival digital preservation and data management will be touched on to generate conversation.

This session would be of interest to web archiving practitioners; digital librarians, archivists and systems librarians; digital preservationists and those thinking about sharing web archival data; digital humanists and researchers that use the Web as primary historical source; or anyone considering the development of a web archiving program or library-oriented collection development policies.

(P.S.: I would be very pleased to supply a paper for later publication based on this talk).

Bio:

Tom J. Smyth is a senior librarian and the Manager of the Digital Integration unit within the Digital Preservation and Migration Division at Library and Archives Canada (LAC). Tom’s work involves managing special library and archival digital collections and programs, in digital preservation contexts. He has led LAC’s Web Archiving Program since 2009. Tom has a background in the Digital Humanities, and holds Master’s degrees in the Social Sciences and in Library and Information Studies from the University of Toronto.

Working Together Toward a Shared Vision

Canadian Government Information Digital Preservation Network (CGI DPN)

Muhammad Umar Qasim
Digital Preservation Officer
University of Alberta Libraries

Sam-Chin Li
Government Information Librarian
University of Toronto Libraries

ABSTRACT

Capturing, preserving and providing access to large amount of valuable content is a huge challenge which is very hard for a single institution to deal with. Dividing the bigger challenge into smaller issues, working collaboratively and using appropriate tools can help to address a difficult task. CGI-DPN members employed these approaches to capture government information for future generations to use. The group created workflows to collaboratively identify and archive valuable information sources, and used tools and technologies to preserve this content for long-term access in a distributed environment. In this presentation, we will talk about the background and the status of this project.

DETAILS

Access to government information is the foundation of a functioning democracy. Any such informational sources enhance informed citizen engagement and enables citizens to assess the governing bodies to keep them accountable for their actions. Therefore, public access to government information must not be restricted by the ability to pay, or by format, geography or any other administrative barriers. It should also be preserved as this primary source provides a collective understanding of our country. Appropriate preservation strategies must be employed to ensure the authenticity of this information since this information may be used to make evidence-based decisions and to assess the policies and actions of any government.

As of April 2014, Government of Canada stopped producing tangible-format publications and transitioned to a digital-only model. Government information in html format is available on the departmental and agencies' web sites. In addition, the Depository Services Program (DSP) collects government publications from many departments and agencies and keep an electronic collection of government publications in pdf format. The fact that digital information is at a higher risk of being inaccessible due to obsolescence, and easiness to be deleted or altered introduces a special challenge to the long-term access to authentic government information. In the absence of a comprehensive national digital preservation plan and an effective compliance audit program for government information publishing, the long-term access to this information is at risk.

In recent years, digital content got lost in an alarming rate, threatening the existence of our digital cultural memory [1]. In recognition of this threat, information professionals at various academic institutions across Canada formed a network named, Canadian Government Information Digital Preservation Network (CGI-DPN) [2]. The main objective of this network is to preserve digital government information of enduring value. The challenges for the network include how to identify valuable sources of information from a huge array of content and how to ensure the long-term accessibility of these sources.

With more than 10 institutions as part of this network, a collaborative approach deemed necessary so that no single institution is taking the huge responsibility of collecting, preserving and providing access to the content. The group members attempted to find a solution which facilitates collaboration, enables perpetual access to the archived content, is decentralized and affordable. Evidently, cultural heritage sector is heavily involved in web archiving activities and this approach seemed to be a best match to capture online content for future use [3]. A need for a central portal was found to be necessary to collaboratively work on this project. Archive-It tool provides several necessary features that are well suited for such a collaborative work [4]. However, it lacks any built-in support for the long-term preservation of the archived content. With its unique decentralized mechanism, LOCKSS provides an affordable and a well-suited option for the CGI-DPN to initiate a new PLN (Private LOCKSS Network) to preserve the content [5].

REFERENCES

[1] Pennock, Maureen, "Web Archiving", DPC Technology Watch Report, 13-01, March 2013, Digital Preservation Coalition, p. 3 http://dx.doi.org/10.7207/twr13-01

[2] Canadian Government Information Private LOCKSS network(CGI-PLN). http://plnwiki.lockss.org/wiki/index.php/CGI_network

[3] CARL, "Archiving the Web". Working paper submitted to the CARL Committee on Research Dissemination. September 2014. http://www.carl-abrc.ca/doc/Archiving_the_web.pdf

[4] Kuchler, Hannah, "How to preserve the Web's past for the future," Financial Times, April 11, 2014

[5] Victoria Reich and David S.H. Rosenthal. "Distributed Digital Preservation: Private LOCKSS Networks as Business, Social, and Technical Frameworks", Library Trends, vol. 57, no. 3, Winter 2009, pp. 461-475. doi:10.1353/lib.0.0047

Strategies for Collecting, Processing, and Analyzing Tweets from Large Newsworthy Events

Nick Ruest, York University, ruestn@yorku.ca

#WomensMarch, #Aleppo, #paris, #bataclan, #parisattacks, #porteouverte, #jesuischarlie, #jesuisahmed, #jesuisjuif, #charliehebdo, #panamanpapers, and #exln42 are all different hashtags, but they share several things in common. They are all large newsworthy events. They are datasets that each contain over a million tweets. Most importantly these collections raise some interesting insights in collecting, processing, and analyzing large newsworthy events[1].

Collecting tweets from these events can be challenging because of timing. Tweets can be collected from the Filter API[2] and Search API[3]. Both having their own caveats. The Filter API only captures the current Twitter stream, and is limited to collecting up to 1% of the overall Twitter stream[4]. The Search API allows you to collect more than 1% of the overall Twitter stream, but one can only collect up to 18,000 every 15 minutes, and is limited to a 7 day window. Generally, using a strategy of using the Filter and Search API to capture a given event is the best.

DocNow's `twarc`[5] includes a number of utilities to process a dataset after collection. These tools allow a researcher, librarian, or archivist to filter their dataset(s) down to what is needed for appraisal, and then accession. Noteworthy tools include; deduplication, source, retweets, date/times, users, and hashtags.

DocNow's utilities can be further used to curate related collections. One can extract all the urls of a dataset, unshorten them, and extract the unique urls to use as a seed list for a web crawler to capture websites related to a given event. One can also extract all of the image urls, and download all images associated with a dataset, which then can be used for image analysis[6], presentation, and/or preservation.

In conclusion, this presentation will provide an overview of collection strategy, insights from processing and analysis, ensuing web crawls, and image presentation from each collection.

[1] Ruest, N. and Milligan, I, An Open-Source Strategy for Documenting Events: The Case Study of the 42nd Canadian Federal Election on Twitter, Code4Lib Journal, Issue 32, 2016-04-25

[2] https://dev.twitter.com/streaming/public

[3] https://dev.twitter.com/rest/public/search

[4] Driscoll, K. and S. Walker, *Big Data, Big Questions| Working Within a Black Box: Transparency in the Collection and Production of Big Twitter Data*, International Journal of Communication, vol. 8, p. 20, Jun. 2014.

[5] https://github.com/docnow/twarc

[6] http://ruebot.net/post/1203867-elxn42-images

Classification of Tweets using Augmented Training

Saurabh Chakravarty
Virginia Tech
Department of Computer Science
saurabc@vt.edu

Eric Williamson
Virginia Tech
Department of Computer Science
ericrw96@vt.edu

Edward Fox
Virginia Tech
Department of Computer Science
fox@vt.edu

ABSTRACT

Classifying tweets in archived collections requires human annotation effort to create training data. Another challenge is that words absent from initial training data cannot aid in classification. In response, we present a way to classify archived collections of tweets using minimal training data. We propose augmenting the training data from an auxiliary source. Our approach shows significant improvements, using a Logistic Regression classifier.

1. INTRODUCTION

With over 1.5 billion tweets archived, covering over 1000 important events, we aim to aid researchers to more effectively use this resource by suitably classifying each tweet. Target classes are: real-world events, event categories, trends, and broad topics. A tweet can be classified into one or more classes. We need labeled training data to train each classifier. This process is tedious and requires a human actor to manually label each tweet with a list of categories. Hence, we must start with very little training data and generate training data on the fly from an auxiliary data source.

2. Approach

We use the Word2Vec [1] based feature selection and transformation technique. We use auxiliary data sets and the Google News corpus to train the word vectors on. We use Logistic Regression (LR) for classifying the tweets [2]. We use an iterative process based on [3] where we augment the training process by providing additional training examples from an auxiliary data source. We perform validation using a held-out validation set. We start training a classifier using the initial training data. We perform 20 iterations, providing 20 training examples randomly sampled from the auxiliary source at each iteration, and perform classification on the validation set at the end of each iteration. If the F1-score on the validation set is more than 0.98, we exit out of the loop. If the training examples for a given iteration improve the accuracy over the best observed accuracy, we add these examples to the training set. Otherwise they are discarded.

We use a cosine-similarity [4] based technique to label tweets in the auxiliary data source. Tweets that have a high mean cosine-similarity score with all positively labeled training tweets are labeled as positive. A low mean cosine-similarity score yields a negative label. Other cases typically lead to discarding tweets from the set to be added for improved training. We add the training examples generated as part of this process to the initial training set and again train a classifier.

Table 1. Positive and Negative Training Examples Distribution

Dataset	Training	Validation
Winter Storm	13 (7+,6-)	88 (52+, 36-)
Severe Weather	16 (7+, 9-)	91 (40+, 51-)
Obesity	13 (7+,6-)	184(79+,105-)
Ebola Outbreak	16 (7+, 9-)	51(25+, 26-)
Egyptian Revolution	22 (8+, 14-)	38 (23+, 15-)
Greece	17 (10+,7-)	57 (18+, 39-)
Environment	20 (10+, 10-)	42 (19+, 23-)

Table 2. F1-scores for the Classifiers

Dataset	Metrics	LR-Local	LR-Google
Winter Storm	Initial F1	0.58	0.96
	Max F1	0.94	0.96
Severe Weather	Initial F1	0.76	0.93
	Max F1	0.78	0.98
Obesity	Initial F1	0.9	0.92
	Max F1	0.93	0.95
Ebola Outbreak	Initial F1	0.86	0.86
	Max F1	0.9	0.9
Egyptian Revolution	Initial F1	0.6	0.91
	Max F1	0.83	0.97
Greece	Initial F1	0.42	0.68
	Max F1	0.46	0.77
Environment	Initial F1	0.58	0.78
	Max F1	0.67	0.83

3. Experimental setup

We performed experiments using the Google News (LR-Google) and auxiliary data source (LR-Local) word vectors with LR. Table 1 shows the training data distributions for the 7 datasets; Table 2 shows F1-scores for the two classification methods.

4. Acknowledgements

This material is based upon work supported by the US National Science Foundation through grants IIS-1319578 and IIS-1619028.

5. REFERENCES

[1] Google Inc., "Word2Vec Implementation," Google, 29 July 2013. [Online]. Available: https://code.google.com/archive/p/word2vec/. [Accessed 18 April 2017].

[2] E. P. S. Castro, S. Chakravarty and E. Williamson, "Classifying Short Unstructured Data using the Apache Spark Platform," in *JCDL*, 2017.

[3] Z. Lu, Z. Yin, S. J. Pan, E. W. Xiang, Y. Wang and Q. Yang., "Source Free Transfer Learning for Text Classification," 2014.

[4] Christian S. Perone, "Cosine Similarity for Vector Space Models," http://blog.christianperone.com/2013/09/machine-learning-cosine-similarity-for-vector-space-models-part-iii/.