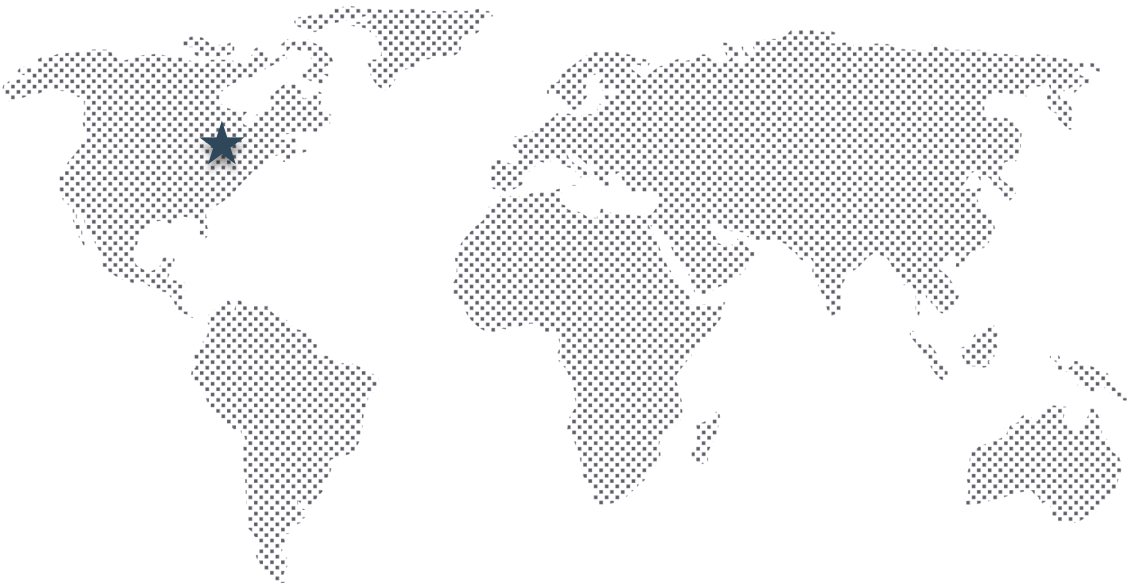
Proceedings *of the*



2019

Web Archiving & Digital Libraries

Workshop

June 6, 2019

Urbana-Champaign, Illinois

Martin Klein

Zhiwu Xie

Edward A. Fox

Editors

WADL 2019 homepage

Web Archiving and Digital Libraries
JCDL 2019 (http://2019.jcdl.org)
Urbana-Champaign, Illinois, USA

Please see the approved WADL 2019 workshop proposal.

Please also see last year's WADL 2018 homepage and the homepage of the 2017 edition of WADL. That workshop led in part to a special issue of International Journal on Digital Libraries.
We fully intend to publish a very similar journal issue based on WADL 2019 contributions.

Submissions (please provide contact and supporting info in <= 2 pages):

- EasyChair submission page: https://easychair.org/conferences/?conf=wadl2019
- Due: April 15, 2019
- Notifications: April 26, 2019
- Please use the ACM Proceedings template.
- **Categories:** (pick one of the three submission categories below and identify it in the submission)
  - 20 min. presentation + 10 min. Q&A
  - Poster/Demonstration + lightning talk
  - 30 min. panel with interactive plenary discussion

SCHEDULE:

Featured Talk: by Cathy Marshall

Thursday, June 6th, 8am-5pm

Time	Activity, Presenters/Authors	Title of Presentation
8:00	Welcome, Introductions	Everyone speaks!
8:30	Cathy Marshall	In The Reading Room: what we can learn about web archiving from historical research
9:30	Break	
10:15	Corinna Breitinger	Securing the integrity of time series data in open science projects using blockchain-based trusted timestamping
10:35	Sawood Alam	Impact of HTTP Cookie Violations in Web Archives
10:55	Liuqing Li, Ed Fox	Users, User Roles, and Topics in School Shooting Collections of Tweets
11:15	Ian Milligan	From Archiving to Analysis: Current Trends and Future Developments in Web Archive Use
12:00	Lunch	
13:00	Jasmine Mulliken	Web Archive as Scholarly Communication
13:30	Brenda Reyes Ayala	Using Image Similarity Metrics to Measure Visual Quality in Web Archives
13:50	Sergej Wildemann	A Collaborative Named Entity Focused URI Collection to Explore Web Archives

14:10	Sawood Alam	MementoMap: An Archive Profile Dissemination Framework
14:30	Break	
15:00	Workshop summary discussion	
16:00	End	

Description:

The 2019 edition of the Workshop on Web Archiving and Digital Libraries (WADL) will explore the integration of web archiving and digital libraries. The workshop aims at addressing aspects covering the entire life cycle of digital resources and will also explore areas such as archiving processes and tools for "non-traditional" resources such as scholarly and government data, 3D objects, and digital online art.

In addition, the chairs will initiate the workshop proceedings being published in a special issue of the JCDL Bulletin.

WADL 2019 will cover all topics of interest, including but not limited to:

Special Event Archiving	Collection Building	Crawling of Dynamic, Online Art, and Mobile Content
Social Media Archiving	Archival Standards, Protocols, Systems, and Tools	Archival Metadata, Description, Classification
Discovery of Archived Resources	Extraction and Analysis of Archival Records	Community Building
Diversity in Web Archives	Ethics in Web Archiving	Interoperability of Web Archiving Systems

Objectives:

- to continue to build the community of people integrating web archiving & digital libraries
- 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: Martin Klein, Los Alamos National Laboratory Research Library, mklein@lanl.gov,
- Co-chair: Zhiwu Xie, Professor, Director of Digital Library Development, Virginia Tech Libraries, zhiwuxie@vt.edu,
- Co-chair: Edward A. Fox, Professor and Director Digital Library Research Laboratory, Virginia Tech, fox@vt.edu http://fox.cs.vt.edu,

Program Committee:

- Justin F. Brunelle, The MITRE Corporation, jbrunelle@cs.odu.edu

- Sumitra Duncan, Frick Art Reference Library, duncan@frick.org
- Joshua Finnell, Colgate University, jfinnell@colgate.edu
- Bela Gipp, UC Berkeley, California and Universität Konstanz, bela@gipp.com
- Abbie Grotke, Library of Congress, abgr@loc.gov
- Helge Holzmann, Internet Archive, helge@archive.org
- Frank McCown, Harding University, fmccown@harding.edu
- Michael Nelson, Old Dominion University, mln@cs.odu.edu
- Nicholas Taylor, Stanford U. Libraries, ntay@stanford.edu
- Michele Weigle, Old Dominion Univiersity, mweigle@cs.odu.edu

Securing the integrity of time series data in open science projects using blockchain-based trusted timestamping

Patrick Wortner[1], Moritz Schubotz[1,2], Corinna Breitinger[3], Stephan Leible[1], Bela Gipp[1]
[1] University of Wuppertal, Germany {lastname@uni-wuppertal.de}
[2] FIZ Karlsruhe, Leibnitz Institute for Information Infrastructure, Germany
[3] Universit of Konstanz, Germany, corinna.breitinger@uni-kn.de

ABSTRACT

The open science movement has become a synonym for modern, digital, and inclusive science. At the same time, open science introduces new challenges for digital libraries, as well as the long-term preservation and the quality assurance of open science datasets. According to open science principles, not only researchers but also citizens should be able to contribute data, e.g. so-called 'citizen science projects'. For such democratized projects, securing the *integrity* and *longevity* of research data is a particular concern. We propose an approach capable of securing the integrity of time series data directly as it is generated. The data is automatically stored in a decentralized and tamper-proof manner while using blockchain technology to prevent any subsequent modification. Our prototype demonstrates how time series data recorded by sensors, e.g. temperature, current, and vibration sensors, can be transparently and immutably stored. By demonstrating an inexpensive modular hardware prototype in combination with open source software, we show that the entry barrier is low for implementing open science projects capable of securing data integrity and offering decentralized long-term data storage. This in turn, can increase the legitimacy of open science datasets and citizen science projects in particular.

1 Introduction

Traditionally, the workflow of researchers and scientists consisted of the following steps: (1) researching literature, (2) formulating a hypothesis, (3) performing experiments and recording measurements, (4) evaluating data, and (5) publishing the results. If the fifth step, however, is repeatedly unsuccessful, scientists cannot succeed in today's research environment. In part due to this pressure to publish, it is not uncommon for scientists to attempt to manipulate steps 3 and 4 to be able to publish [5, 6]. In contrast to the traditional academic research cycle and publishing process, today's *open-science movement* encourages and facilitates the publishing of raw research measurements early in the research cycle and even encourages the publishing of negative results. This, in turn, has introduced new challenges to storing and verifying research data throughout the research cycle.

Performing research according to open science principles [12] provides at least two significant advantages. First, data fabrication become more difficult if raw data and intermediate results are published. Second, additional insights can be obtained and published by other researchers without the time-consuming experimentation step. The open science movement also blurs the line between scientists and citizens by enabling inclusion of interested individuals, for example, in so-called 'citizen science projects'.

Currently, there is no agreement among scientists on a standardized procedure for reviewing open science datasets in a way that is analogous to today's literature-assessment process. Methods for researching and reviewing scientific literature using content-based and bibliometric measures are well established for traditional publications. While bibliometric measures could also be applied to datasets, a content-based assessment based on the raw data seems hardly feasible. Additionally, the quality of measurement data is influenced by factors, such as: (1) well-defined error estimates, (2) accurate meta-data, (3) high redundancy, (4) adequate sampling rate, (5) adherence to all known physical laws, etc. While such factors must be manually verified by readers and scientists, we believe that scientists could significantly benefit from an automated method to guaranteed the integrity and longevity of time series research data immediately as it is generated. Thus, we propose securing time-series research data using a technical solution to protect data against any subsequent changes or manipulation. To achieve this, we make use of decentralized trusted timestamping, which relies on cryptographic hashing and the tamper-proof characteristics of blockchain technology [8].

To demonstrate our proposed solution, we present a simple and inexpensive hardware prototype that can be used by citizens to record measurements for certain physical properties. The corresponding open source software makes use of blockchain technology and decentralized data storage to ensure the immutability of the sensors' time series data.

2 Using blockchain for securing sensor measurement data

The blockchain underlying cryptocurrencies, e.g. Bitcoin [11], offers unique characteristics, such as decentralization, immutability, and trusted timestamping [14, 17]. This makes blockchain technology valuable in developing novel applications [3]. Several projects have been proposed to support researchers, including managing academic reputation [15], protecting intellectual property in academic manuscripts submitted for peer

review [7], or tracking individual contributions in a collaborative research project [13]. From a technical point of view, there is much literature on blockchain technology and its strengths and weaknesses in different stress tests and use cases [14, 16, 17]. A blockchain can be viewed as a decentralized database without a central authority to manage the data it stores [1]. Data stored on a blockchain is immutable and permanent. We use these characteristics to ensure that each measurement value recorded by a sensor is made tamper-proof. With this contribution, we hope to support the viability of citizen science projects and today's open science movement by making data and entire datasets more trustworthy.

In the case of citizen science projects, manipulating many sensors in a decentralized network would require significant effort and would be difficult to achieve without being detected. The easier way is to manipulate or prune the data after it was measured or aggregated – which we are capable of preventing with our approach.

Once a hash has been included in the form of a transaction on a blockchain, one can verify that the data associated with the hash (for example, sensor measurement values) were not manipulated after they were collected [4]. We do not propose to store all raw measurement data directly on a blockchain. Instead, we only store the hash value of digital measurements or sets of measurement data. This is more efficient in terms of performance, cost, and scalability, yet it allows proving that the data existed in a certain format at a certain time, thus increasing data integrity and transparency.

3 Prototype for open science projects

Capturing time series data from sensors and verifiably securing all data with a trusted timestamp requires building an interface between the physical world and a blockchain. To demonstrate our idea, we implemented an inexpensive and easy to use Raspberry Pi prototype that captures vibration, electric current and temperature measurements. The modular design of the prototype allows for easy customization. For the basic module, we used a Raspberry Pi, a general-purpose input/output board, and an initial set of three sensors. To test the sensors of the model, we designed a testing module: a rotor with a battery power supply. The rotor from the testing module consumes power and generates vibrations, which produce signals that can be measured with the sensors from the basic module.

We also implemented an open source software to capture the incoming data stream from the sensors, partition the stream into data-chunks, timestamps each chunk, and finally store the chunks. The software and installation manual, as well as links to the first datasets, are available on our GitHub repository[1]. To install the software, the Raspberry needs to be connected to a computer. The software reads the data-stream from the sensors and appends it to an internal buffer. After a user-defined time, or when the data volume size exceeds the chunk size of about 256 kB, a new chunk is created. A hash of this data chunk is computed and uploaded to the trusted timestamping service Originstamp[2] [10]. From the second chunk onward, a reference to the previous chunk is included in the hash.

However, the hash is meaningless without the associated data. Many services exist to upload data to a central server. However, central storage can fail, cease to exist, or be censored and tampered with. We, therefore, use the Interplanetary Filesystem (IPFS) to be independent of a central authority for storage and to ensure data verifiability and longevity. This peer-to-peer network is organized in blocks and the block-address is the hash of the file content, which we already used to generate the trusted timestamp. Therefore, we installed IPFS on the Raspberry to upload the data chunk-by-chunk.

For redundancy, we additionally plan to copy the time series measurement data from IPFS to the long-time archival platform Zenodo[3].

4 Conclusion

Measurement values and data streams from sensors are not immune to manipulation or retrospective selective pruning. The ability to securely prove that research data was not manipulated or omitted is especially important for both open science and citizen science projects. In this paper, we proposed a method for independently and securely verifying the time of creation of sensor data. By storing the data in a decentralized manner using IPFS and by relying on a blockchain-backed solution for storing tamper-proof and decentralized trusted timestamps associated with discrete chunks of measurement values, we showed how sensor data can be made securely verifiable. Our prototype demonstrates how our proposed solution can be easily implemented and used with hardware sensors. We argue that enabling any researcher or interested citizen to verifiably trace the integrity of measurement values and their time of origin can significantly strengthen the open science movement and can increase the trustworthiness of citizen science projects.

REFERENCES

[1] Beck, R., Czepluch, J.S., Lollike, N., and Malone, S. Blockchain-the Gateway to Trust-Free Cryptographic Transactions. *ECIS*, (2016).

[2] Brambilla, G., Amoretti, M., and Zanichelli, F. Using Blockchain for Peer-to-Peer Proof-of-Location. *arXiv preprint arXiv:1607.00174*, (2016).

[3] Casino, F., Dasaklis, T.K., and Patsakis, C. A systematic literature review of blockchain-based applications: current status, classification and open issues. *Telematics and Informatics*, (2018).

[4] Catalini, C. and Gans, J.S. *Some simple economics of the blockchain.* 2016.

1 https://github.com/FellowsFreiesWissen/Blockchain_Pi

[2] www.originstamp.org

[3] www. zenodo.org

[5] Fanelli, D. How many scientists fabricate and falsify research? A systematic review and meta-analysis of survey data. *PloS one 4*, 5 (2009), e5738.

[6] Fanelli, D. Do pressures to publish increase scientists' bias? An empirical support from US States Data. *PloS one 5*, 4 (2010), e10271.

[7] Gipp, B., Breitinger, C., Meuschke, N., Beel, J., and Breitinger, C. CryptSubmit: Introducing Securely Timestamped Manuscript Submission and Peer Review Feedback using the Blockchain. *Proceedings of the ACM/IEEE-CS Joint Conference on Digital Libraries (JCDL)*, (2017).

[8] Gipp, B., Meuschke, N., and Gernandt, A. Decentralized Trusted Timestamping using the Crypto Currency Bitcoin. *Proceedings of the iConference 2015*, (2015).

[9] Grech, A. and Camilleri, A.F. Blockchain in education. 2017.

[10] Hepp, T., Schoenhals, A., Gondek, C., and Gipp, B. OriginStamp: A blockchain-backed system for decentralized trusted timestamping. *Information Technology 60*, 5–6 (2018), 273–281.

[11] Nakamoto, S. Bitcoin: A peer-to-peer electronic cash system. (2008).

[12] Nosek, B.A., Alter, G., Banks, G.C., et al. Promoting an open research culture. *Science 348*, 6242 (2015), 1422–1425.

[13] Schubotz, M., Breitinger, C., Hepp, T., and Gipp, B. Repurposing Open Source Tools for Open Science: a Practical Guide. 2018. https://doi.org/10.5281/zenodo.2453415.

[14] Seebacher, S. and Schüritz, R. Blockchain technology as an enabler of service systems: A structured literature review. *International Conference on Exploring Services Science*, (2017), 12–23.

[15] Sharples, M. and Domingue, J. The blockchain and kudos: A distributed system for educational record, reputation and reward. *European Conference on Technology Enhanced Learning*, (2016), 490–496.

[16] Yli-Huumo, J., Ko, D., Choi, S., Park, S., and Smolander, K. Where is current research on blockchain technology?—a systematic review. *PloS one 11*, 10 (2016), e0163477.

[17] Zheng, Z., Xie, S., Dai, H., Chen, X., and Wang, H. An overview of blockchain technology: Architecture, consensus, and future trends. *Big Data (BigData Congress), 2017 IEEE International Congress on*, (2017), 557–564.

A Collaborative Named Entity Focused URI Collection to Explore Web Archives

Sergej Wildemann
L3S Research Center
Hannover, Germany
wildemann@L3S.de

Helge Holzmann
Internet Archive
San Francisco, CA, USA
helge@archive.org

ABSTRACT

Vast amounts of data are stored by Web archives in order to preserve the history of digital mankind. But without ways to easily navigate and access resources of interest, their potential cannot be fully exploited. When full-text indexes seem unfeasible due to the size and additional temporal dimension, topic focused collections could provide structure and a starting point for many research questions. Here, we present a collaborative Web platform to collect and annotate URIs that characterize named entities over specific time frames. Initial data is provided by aggregating and evaluating multiple datasets and further enrichment with metadata.

1 INTRODUCTION

Navigating Web archives like the Internet Archive's *Wayback Machine*[1] can be difficult without knowing the exact URI and date of interest. To improve access, a site search[2] based on anchor texts was implemented there, which guides the users to relevant domains for the entered keywords while ignoring the exact path and date.

In the past, we explored several ways to emphasize the temporal dimension of these archives by providing improved retrieval methods as well as search interfaces. This included the usage of user generated tags from social bookmarking systems and the indexation of anchor texts as a surrogate of the target resources[1, 2].

Here, we present the Web platform *Tempurion* (see Fig. 1) which shifts the focus from a broad spectrum exploration tool for Web archives towards an annotated and topic related URI collection for named entities. The underlying dataset is based upon the integration of multiple sources such as entity classifications from DBpedia, URIs and tags from Wikipedia, Wikidata, Delicious and the German Web archive as well as temporal enrichments from the Internet Archive's CDX index. Potential users are encouraged to contribute to the collection by providing additional resources and metadata or influence the ranking of results by voting. Public access to the underlying dataset is further provided in a machine-readable way via a RESTful API. A live version is accessible under:

https://tempurion.l3s.uni-hannover.de

2 CONCLUSION

As we built this platform and collect initial data, we are interested in feedback from and collaboration with other researchers to enrich the collection, improve usability and integrate ideas to add more structure.

[1] https://web.archive.org
[2] https://blog.archive.org/2016/10/24/beta-wayback-machine-now-with-site-search/

WADL'19, June 2019, Urbana-Champaign, Illinois USA
2019. ACM ISBN 978-x-xxxx-xxxx-x/YY/MM…$15.00
https://doi.org/10.1145/nnnnnnn.nnnnnnn

REFERENCES

[1] Helge Holzmann and Avishek Anand. 2016. Tempas: Temporal Archive Search Based on Tags. In *WWW, Companion Volume.*
[2] Helge Holzmann, Wolfgang Nejdl, and Avishek Anand. 2016. On the Applicability of Delicious for Temporal Search on Web Archives. In *SIGIR.*

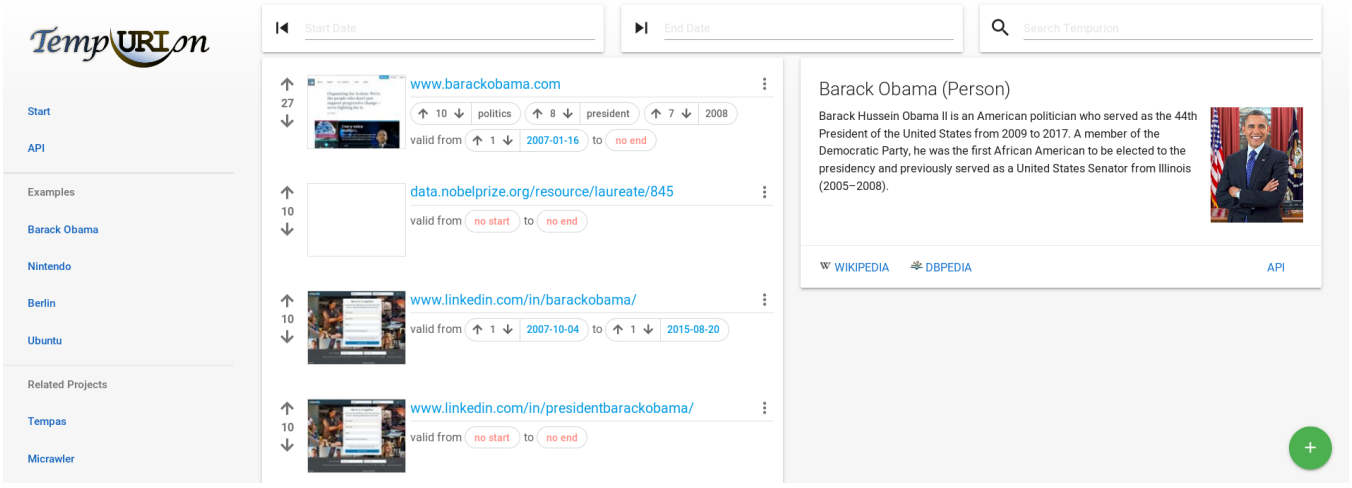


Figure 1: Entity result view in Tempurion

Users, User Roles, and Topics in School Shooting Collections of Tweets

Liuqing Li, Rishabh Anand, and Edward A. Fox
Department of Computer Science, Virginia Tech
Blacksburg, VA, USA
[liuqing,risha97,fox]@vt.edu

ABSTRACT

Especially since the 1999 Columbine High School massacre, a great deal of attention has focused on school shooting events occurring in the United States. As part of our ongoing collection and analysis of tweets about important events, we have selected five school shootings from the period 2012 - 2018. Beginning with a random sample of 10,000 tweeters from each of those tweet collections, we compared the behavior of different types of tweeters. We deployed TwiRole, our new tool for user classification (brand, female, male), and TweetLDA, for topic modeling. We uncovered interesting results about user engagement, topic distribution, and user role differentiation. These results suggest the value of further exploration with our hundreds of event-related collections, to identify patterns behind different types of disasters.

KEYWORDS

User classification, topic modeling, school shooting, Twitter

1 INTRODUCTION

More than 100 people were killed or injured in school shootings in the United States in 2018 [2]. These mass shootings have had considerable effect on survivors, families, schools, and the public. They have elicited voluminous discussions among users on social media. Accordingly, we consider this topic to be an important part of our research on Web archiving and digital libraries.

We treat Twitter as a barometer of users and topics, to investigate the patterns regarding school shootings. Our analysis is guided by three research questions: 1) Is there any pattern regarding user engagement in school shootings? 2) Is there any pattern regarding topics across different collections? 3) Do different users have different concerns regarding specific topics?

We first prepared five school shooting collections of tweets since 2012, deployed TwiRole to identify user roles (i.e., brand / organization, female, and male), and employed TweetLDA as an aid to generate topic categories from the entire corpus. Then, considering users, user roles, and topics, we conducted a comprehensive analysis of the five tweet collections. In this paper, we summarize our findings.

2 RELATED WORK

Most researchers studying school shootings have analyzed a single event. Few researchers analyzed school shootings across user types.

[1, 4–7, 14] analyzed sentiments/emotions (e.g., sadness and anxiety) and their patterns (e.g., shift, desensitization) during a disaster, while [3, 10, 13] focused on tweeting patterns (e.g., keywords, mentions, and hashtags). [15] applied keyword filtering and latent Dirichlet allocation (LDA) on a random sample of tweet data and discovered a set of topics related to school shooting. [12] built a domain network graph for mass shooting events by extracting embedded URLs from tweets. [8] analyzed users' emotion patterns about disasters on Twitter, but they were limited to three school shooting collections and not concerned about topics.

3 METHODOLOGY

Figure 1 gives a data flow diagram of our analysis approach. First, focusing on school shooting collections of tweets, we sampled users from clean tweets after preprocessing. Second, we applied TwiRole [9] to identify user roles. For this classification, TwiRole utilizes a hybrid model, leveraging basic features (e.g., name, description, profile image brightness), advanced features (e.g., k-top words in tweets), and deep features from profile images. We also developed and deployed a web interface[1] for online classification. Third, we extracted tweets with users having roles from cleaned tweets, applied TweetLDA [16] for topic modeling, and integrated our topics with other sources. Finally, we carried out a comprehensive analysis of school shooting events, considering users, user roles, and topics.

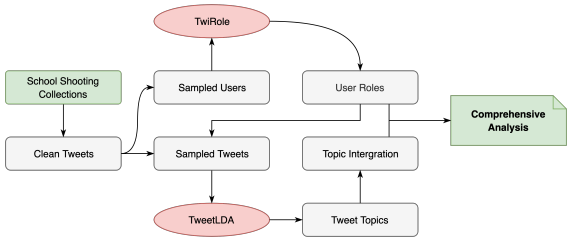


Figure 1: Data flow diagram of our analysis approach

4 PRELIMINARY ANALYSIS

4.1 Data Description

Though we already had hundreds of tweet collections dating from 2012, we applied GetOldTweets3 [11] to create five larger, more complete, tweet collections on school shooting. The time range of each collection covers up to about one month after the corresponding disaster. Later, retweets (RTs) and non-English tweets were filtered out during cleaning. Table 1 shows the details of our collections; we can share their tweet IDs upon request.

4.2 User Engagement

We randomly selected 10,000 users from each collection and applied TwiRole to predict their roles. We manually checked 100 users per

[1] http://vis.dlib.vt.edu:3001

Table 1: An overview of the five school shooting collections

Collections	Date	# of Tweets
Sandy Hook Elementary School shooting	12/14/2012	97,282
Santa Monica shooting	06/07/2013	23,093
Umpqua Community College shooting	10/01/2015	17,820
Stoneman Douglas High School shooting	02/14/2018	22,320
Santa Fe High School shooting	05/18/2018	47,708

role to evaluate TwiRole's performance. The overall accuracy is 82.7%. Then, we counted the number of tweets posted by each role and the average percentages are: 45.5% ± 2.5% (brand), 23.2% ± 3.1% (female), and 31.3% ± 1.6% (male); see Figure 2. Brand users are the primary group who might publish information or deliver messages, while male users participated more actively than female users in discussions of the selected school shootings.

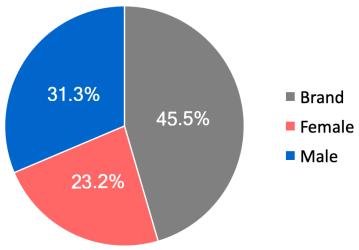


Figure 2: Average percentage of user engagement

4.3 Topic Distribution

We combined tweets posted by a labeled user into one document, assembled all such documents, removed shooting names and places, and utilized TweetLDA to identify 15 topics from our corpus. We created four categories and their corresponding words that come from Zhang et al.'s dictionaries [15] and our topics; see Table 2.

For each collection, we retrieved relevant tweets by taking typical words in each category as a query, and counted the number of tweets among the above four categories. Figure 3 shows the percentages of categories across the five collections. We found that "information" is the dominant topic in all school shootings, which is consistent with our expectation. Tweeters posted few tweets about mental health, but focused on the topics related to gun control and emotion. Especially, people were more concerned about gun control and expressed their feelings regarding the school shootings that had more people dead or injured (i.e., 2012 Sandy Hook, 2018 Douglas, and 2018 Santa Fe), compared with the other two school shootings.

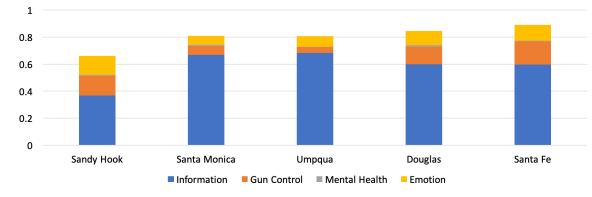


Figure 3: Percentages of different categories among collections

Table 2: Major categories and typical words in school shootings

Category	Typical Words
Information	news, shooter, accidental, gunman, arrest, victim, dead, report, kill, suspect, breaking, injured, custody, wounded, police, fatality, update, fire, campus, student, confirm
Gun Control	gun, control, nra, congress, senate, legislation, safety, violence, illegal, antigun, debate, governor, laws
Mental Health	mental, health, illness
Emotion	absurd, anger, angry, awful, crazy, disgust, frighten, heart, pray, rip, sad, tragic, dedicated, deep, silence

4.4 User Role Differentiation

Focusing on the "emotion" topic, we calculated the tweeting percentage of each role among collections; see Figure 4. The result indicates that female users were posting more tweets containing emotional words than male and brand users. Regarding the Sandy Hook Elementary School shooting, which was the most gruesome attack among the five shootings, a higher percentage of the tweets reflected emotion, for all of the groups of users.

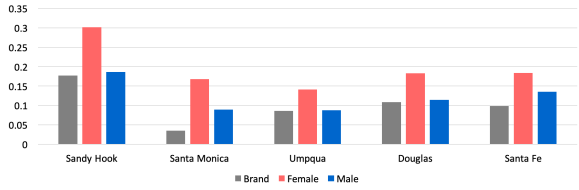


Figure 4: Role differentiation in the "emotion" topic

5 SUMMARY AND FUTURE WORK

We conducted a comprehensive analysis of five school shooting collections, considering users, user roles, and topics. We observed:

- The relative percentages of user engagement appear stable across collections ($P_{Brand} > P_{Male} > P_{Female}$);
- "Information" is the dominant topic in all school shootings while "mental health" was seldom discussed. Tweeters were more concerned about "gun control".
- Tweeters expressed their feelings most when the victim count was highest.
- Female users posted more emotional tweets than male and brand users.

In the future, we will carry out a temporal analysis, further explore different emotions across user roles, and deploy our approach on other types of events to help with further discovery of patterns.

6 ACKNOWLEDGMENTS

Thanks go to the US NSF for supporting the Global Event and Trend Archive Research (GETAR) project, through grant IIS-1619028 as well as IIS-1619371 to partner Internet Archive. We also thank the CS4624 student team for work on the TwiRole web interface.

REFERENCES

[1] Allameh, E. M. *Analyzing Emotions on Twitter during the 2014 Purdue University Shooting Crisis.* PhD thesis, Purdue University, 2015.

[2] BBC. 2018 'Worst Year for US School Shootings'. https://www.bbc.com/news/business-46507514. Accessed: 2019-03-28.

[3] Budenz, A., Purtle, J., Klassen, A., Yom-Tov, E., Yudell, M., and Massey, P. The Case of a Mass Shooting and Violence-related Mental Illness Stigma on Twitter. *Stigma and Health* (2018).

[4] Doré, B., Ort, L., Braverman, O., and Ochsner, K. N. Sadness Shifts to Anxiety over Time and Distance from the National Tragedy in Newtown, Connecticut. *Psychological science 26*, 4 (2015), 363–373.

[5] Glasgow, K., Vitak, J., Tausczik, Y., and Fink, C. "With Your Help... We Begin to Heal": Social Media Expressions of Gratitude in the Aftermath of Disaster. In *International Conference on Social Computing, Behavioral-Cultural Modeling and Prediction and Behavior Representation in Modeling and Simulation* (2016), Springer, pp. 226–236.

[6] Jones, N. M., Wojcik, S. P., Sweeting, J., and Silver, R. C. Tweeting Negative Emotion: An Investigation of Twitter Data in the Aftermath of Violence on College Campuses. *Psychological methods 21*, 4 (2016), 526.

[7] Li, J., Conathan, D., and Hughes, C. Rethinking Emotional Desensitization to Violence: Methodological and Theoretical Insights From Social Media Data. In *Proceedings of the 8th International Conference on Social Media & Society* (2017), ACM, p. 47.

[8] Li, L., and Fox, E. A. Understanding Patterns and Mood Changes through Tweets about Disasters. In *ISCRAM* (2019).

[9] Li, L., Song, Z., Zhang, X., and Fox, E. A. A hybrid model for role-related user classification on twitter. *CoRR abs/1811.10202* (2018).

[10] Mazer, J. P., Thompson, B., Cherry, J., Russell, M., Payne, H. J., Kirby, E. G., and Pfohl, W. Communication in the Face of a School Crisis: Examining the Volume and Content of Social Media Mentions during Active Shooter Incidents. *Computers in Human Behavior 53* (2015), 238–248.

[11] Mottl, D. GetOldTweets3. https://github.com/Mottl/GetOldTweets3, 2018. Accessed: 2019-03-28.

[12] Starbird, K. Examining the Alternative Media Ecosystem through the Production of Alternative Narratives of Mass Shooting Events on Twitter. In *Eleventh International AAAI Conference on Web and Social Media* (2017).

[13] Wachter, L. K. JK Rowling's Tweets following and regarding the Parkland School Shooting–A Critical Discourse Analysis. Master's thesis, Malmö universitet/Kultur och samhälle, 2018.

[14] Wang, N., Varghese, B., and Donnelly, P. D. A Machine Learning Analysis of Twitter Sentiment to the Sandy Hook Shootings. In *2016 IEEE 12th International Conference on e-Science (e-Science)* (2016), IEEE, pp. 303–312.

[15] Zhang, Y., Wang, Y., Foley, J., Suk, J., and Conathan, D. Tweeting Mass Shootings: The Dynamics of Issue Attention on Social Media. In *Proceedings of the 8th International Conference on Social Media & Society* (2017), ACM, p. 59.

[16] Zhao, W. X., Jiang, J., Weng, J., He, J., Lim, E.-P., Yan, H., and Li, X. Comparing Twitter and Traditional Media using Topic Models. In *European conference on information retrieval* (2011), Springer, pp. 338–349.

Using Image Similarity Metrics to Measure Visual Quality in Web Archives

Brenda Reyes Ayala
University of Alberta
Edmonton, Alberta, Canada
brenda.reyes@ualberta.ca

Ella Hitchcock
University of Alberta
Edmonton, Alberta, Canada
ehitchco@ualberta.ca

James Sun
University of Alberta
Edmonton, Alberta, Canada
csun@ualberta.ca

CCS CONCEPTS

• **Information systems → Digital libraries and archives**; • **Applied computing → Digital libraries and archives**; • **General and reference →** *Measurement*.

KEYWORDS

web archiving, quality assurance, quality, similarity, web archives

ACM Reference Format:
Brenda Reyes Ayala, Ella Hitchcock, and James Sun. 2018. Using Image Similarity Metrics to Measure Visual Quality in Web Archives. In *JCDL 2019: Web Archiving and Digital Libraries (WADL) workshop, June 06, 2019, Urbana-Champaign, IL.* ACM, New York, NY, USA, 3 pages. https://doi.org/10.1145/1122445.1122456

1 INTRODUCTION

Web archives are becoming increasingly important to our digital infrastructure, as is judging the quality of these archives. With the growing popularity of complex visual styling enabled by Javascript, AJAX, and cascading style sheets (CSS), visual representation of information on the web has become an important part of judging the quality of archived web pages [7], [5]. In the context of web archives, we define visual correspondence as "the similarity in appearance between the original website and the archived website", as initially defined by Reyes Ayala (2018) [4]. This paper examines how the visual correspondence of an archived website can be measured using popular image similarity measures. Using these measures we evaluate how visual correspondence can be used as an indication of overall archive quality. We are interested in answering the following research question: How effective are different similarity measures at measuring the visual correspondence between an archived website and its live counterpart?

2 METHODOLOGY

2.1 The Dataset

We chose three different web archives in order to apply the similarity metrics, two from the University of Alberta and one from the British Library's UK Web Archives: The "Idle No More" collection [10], the Western Canadian Arts collection [11], and the UK Web Archives Open Access (OA) collection [8]. "Idle No More" is a topical web archive created by the University of Alberta using the Archive-It service [2]. It aims to preserve websites related to "Idle No More", a Canadian political movement encompassing environmental concerns and the rights of indigenous communities. The Western Canadian Arts collection, also available on Archive-It, intends to collect and preserve the born digital resources created by filmmakers in Western Canada. The British Library's OA web archive is a more general collection encompassing UK websites that can be made available online according to British legal deposit laws.

2.2 Generating the screenshots

In order to measure the visual correspondence of an archived website to its live counterpart, we created a set of tools called "wa screenshot compare", currently freely available as a Github repository [3]. Written in Python, these tools take a seedlist as input and generate screenshots of the live websites using Pyppeter (a Python port of the Puppeteer screenshot sofware) and and a headless instance of the Chrome browser [9]. "wa screenshot compare" then generates a list of all archived versions of the live sites that are available from the University of Alberta's Archive-It collection. Screenshots are then taken of the archived websites. For the UK OA collection, we were unable to retrieve every archived capture of the seedlist, we therefore decided to take a screenshot of the oldest capture, since that was usually the capture which web archivists from the British Library analyzed for their QA process.

Despite our initial assumptions, this was not a trivial process. A significant issue was archiving institutions' use of banners to indicate to users that they are viewing an archived website. Usually these banners include the name of the institution that created the archived website, the name of the collection the site belongs to, and the time and date when the archived website was created. Our initial approach was to append the text "id_" to the url of the archived websites, as this displays the archived website without the banner, and thus would lead to more accurate screenshots. However, qualitative inspection of the screenshots across all three collections revealed that this approach often breaks the CSS styling of the archived site, resulting in a screenshot that was even farther from the actual appearance of the archived website. A decision was then made to take the screenshots with the banner included, with the assumption that, as the banner is fairly small, it would not impact similarity scores too much.

Link rot was another, more serious challenge to our approach, as many of archived websites are no longer online. As Reyes Ayala (2018) [4] pointed out, visual correspondence can only be measured

JCDL 2019, June 06, 2019, Urbana-Champaign, IL

ACM ISBN 978-1-4503-9999-9/18/06…$15.00
https://doi.org/10.1145/1122445.1122456

Table 1: Characteristics of Web Archive Collections Used for Similarity Judgments

Collection	No. Seeds	No. Seeds Still Available	% Collection Still Available
Idle No More	196	182	92.86
Western Canadian Arts	101	95	94.06
UK Open Access	659	516	78.30

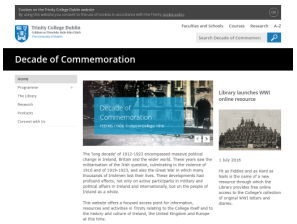
if the original website still exists, otherwise there can be no comparison. Table 1 describes the characteristics of the web archives used to generate screenshots. We categorized as "lost", those websites that returned an HTTP status code other than 200 and were not redirects. From the table, we can see that the UK OA collection has suffered from significant link rot, as almost 22% of the seeds are no longer reachable. The relationship between link rot and web archives has been studied in more detail in [1]. Our experience highlights the importance of conducting visual quality assessments early in the web archiving process, while the websites collected are still online and accessible for comparison.

2.3 Calculating similarity

The tool "wa screenshot compare" then puts the two sets of screenshots through a similarity analysis based on two popular image similarity measures: Structural Similarity Index (SSIM) and Mean Squared Error (MSE). We chose SSIM and MSE due to their popularity in the image comparison community and their accessibility. As mentioned above, both SSIM and MSE are easily available in larger Python libraries and there is a large amount of documentation available. MSE is the measurement of difference in pixels between two images without any reference to the vector position of the pixels, meaning that if for example two images contained the same number of measured pixel values the images would be deemed similar, no matter the position of the pixels in the image [6]. The Structural Similarity Index introduces a structural component to the comparison process by taking the pixel vector positions in both images into account, comparing pixel properties across both images one pixel at a time, and preserving the position of the pixels as they are compared [13]. We added a third measure, which we call “vector distance” [12]. Each screenshot is divided into its different pixels. A measure of similarity is then calculated, which produces the distance between the RGB values of each screenshot. The greater the distance, the greater the difference between the two images, and thus, the greater the difference between the two websites. We changed this metric slightly by subtracting every result from 100, thus giving us the percentage similarity between a pair of images. The scales for each measure are shown below.

- SSIM: calculates similarity on a scale of [-1,1]. 1 is perfect similarity
- MSE: calculates similarity on a scale of $[0, \infty]$. 0 is perfect similarity
- Vector distance: calculates similarity on a scale [0-1]. 1 is perfect similarity

Figures 1 and 2 illustrate how two websites are archived, with very different results. The archived website for Trinity College in Dublin, seen in Figure 1, could be classified as one of "medium quality." The intellectual content of the website has been preserved,



Screenshot of current, live website

Screenshot of archived website

Figure 1: Comparison of images for the website "Trinity College Dublin: Decade of Commemoration". SSIM = 0.51, MSE = 61536.53, Vector Distance = 59.87

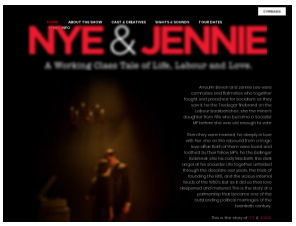
but its styling has been lost, which is reflected in its similarity scores. Interestingly, it yielded a SSIM score of 0.51 and a vector distance score of 59.87. This capture was flagged as having QA issues by web archivists from the British Library. Figure 2 presents an example of an archived website that is of low quality; the archived version is simply a blank page and all content has been lost. This is reflected in the very low SSIM and vector distance scores, and the archived site itself was categorized as a failed capture by web archivists. Throughout our analyses, the MSE measure proved to be the most difficult to interpret, as it has no proper upper bound. It seems MSE works best as a relative measure. We can see the score from the low quality website (169603.88), compare it to the score from the medium quality website (61536.53), and surmise that one capture is worse than the other; however, we will still be lacking an absolute scale.

2.4 Correlation Analysis

In order to determine if there were relationships between different similarity measures, we performed a correlation analysis on all our similarity scores for the three web archives collections. The results are shown in Table 2. The correlation coefficients indicate that there is a moderate negative correlation between SSIM and MSE score, a very strong negative correlation between MSE and Vector distance, and a moderate-to-strong (depending on the collection) correlation between SSIM and vector distance scores. Since the relationship between MSE and vector distance is an almost perfect negative correlation, this suggests that one measure might be easily substituted for another. Because we found MSE scores relatively difficult to interpret, we recommend the use of vector distance as a measure of similarity.

Table 2: Correlation between Different Similarity Measures in Web Archives

Collection	SSIM - MSE	MSE - Vector	SSIM - Vector
Idle No More	-0.61	-0.97	0.61
Western Canadian Arts	-0.72	-0.98	0.78
UK Open Access	-0.63	-0.97	0.69
All	-0.65	-0.97	0.71



Screenshot of current, live website

Screenshot of archived website

Figure 2: Comparison of images for the website of the play "Nye & Jennie". SSIM = 0.28, MSE = 169603.88, Vector Distance = 8.83

3 CONCLUSIONS AND FUTURE RESEARCH

Our experiments showed that image similarity metrics can be successfully applied in order to measure the visual correspondence (and thus visual quality) of archived websites. Our results indicated that these metrics were able to successfully distinguish between website captures of poor quality and those of higher quality. A natural next step is to conduct experiments to find which similarity measures most closely match up with human judgments of visual correspondence in a web archive. This research is only the first step in developing a comprehensive toolkit for automated or semi-automated quality assurance processes in web archives, which will in turn help web archivists create better web archives.

ACKNOWLEDGMENTS

We would like to thank Andy Jackson from the British Library for his help and support in providing access to our UK dataset.

REFERENCES

[1] Scott G. Ainsworth, Ahmed Alsum, Hany SalahEldeen, Michele C. Weigle, and Michael L. Nelson. 2011. How Much of the Web is Archived?. In *Proceedings of the 11th Annual International ACM/IEEE Joint Conference on Digital Libraries (JCDL '11)*. ACM, New York, NY, USA, 133–136. https://doi.org/10.1145/1998076.1998100

[2] Internet Archive. [n. d.]. Archive-It: Web Archiving Services for Libraries and Archives. https://archive-it.org

[3] Brenda Reyes Ayala. [n. d.]. wa screenshot compare. https://github.com/reyesayala/wa_screenshot_compare

[4] Brenda Reyes Ayala. 2018. *A grounded theory of information quality in web archives.* Ph.D. Dissertation. University of North Texas.

[5] Justin F. Brunelle, Mat Kelly, Michele C. Weigle, and Michael L. Nelson. 2016. The Impact of JavaScript on Archivability. *Int. J. Digit. Libr.* 17, 2 (June 2016), 95–117. https://doi.org/10.1007/s00799-015-0140-8

[6] Ahmet M. Eskicioglu, Paul S. Fisher, and Si-Yuan Chen. 1994. Image quality measures and their performance. *The 1994 Space and Earth Science Data Compression Workshop* (April 1994), 57–67. http://ntrs.nasa.gov/search.jsp?R=19940023754

[7] Karl Gyllstrom, Carsten Eickhoff, Arjen P. de Vries, and Marie-Francine Moens. 2012. The Downside of Markup: Examining the Harmful Effects of CSS and Javascript on Indexing Today's Web. In *Proceedings of the 21st ACM International Conference on Information and Knowledge Management (CIKM '12)*. ACM, New York, NY, USA, 1990–1994. https://doi.org/10.1145/2396761.2398558

[8] Andy Jackson. [n. d.]. UKWA Manual QA Dataset. https://github.com/iipc/qa2019/tree/master/ukwa-manual-qa-dataset

[9] miyakogi. [n. d.]. Pyppeteer: headless Chrome/Chromium automation library. https://github.com/miyakogi/pyppeteer

[10] University of Alberta. [n. d.]. Idle No More Collection. https://archive-it.org/collections/3490

[11] University of Alberta. [n. d.]. Western Canadian Arts Collection. https://archive-it.org/collections/6296

[12] Rosettacode.org. 2018. Percentage difference between images. https://rosettacode.org/wiki/Percentage_difference_between_images#Python

[13] Zhou Wang, Alan C Bovik, Hamid R Sheikh, Eero P Simoncelli, et al. 2004. Image quality assessment: from error visibility to structural similarity. *IEEE transactions on image processing* 13, 4 (2004), 600–612.

MementoMap: An Archive Profile Dissemination Framework

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

ABSTRACT

We introduce *MementoMap*, a framework to express and disseminate holdings of web archives (archive profiles) by themselves or third parties. The framework allows arbitrary, flexible, and dynamic levels of details in its entries that fit the needs of archives of different scales. This enables Memento aggregators to significantly reduce wasted traffic to web archives.

1 INTRODUCTION AND BACKGROUND

The Memento framework [12] introduced a uniform means for various web archives to interoperate when it comes to accessing archived representations of a given Uniform Resource Identifier (URI). This enabled easy aggregation of mementos from many web archives. However, there is no standard way to access holdings of web archives without providing a full or partial lookup URI. Without the prior knowledge of holdings of each web archive, a naive Memento aggregator might flood all aggregated archives with unnecessary broadcast lookup requests for which they might not have any good results to return. The log of a Memento aggregator service[1] running at ODU using MemGator [3] shows that in over three years of its service it performed about 62M lookup requests to 14 different web archives of which only 5.44% requests returned any mementos. By knowing the holdings of these web archives (i.e., archive profiles), remaining over 94% wasted requests could have been avoided, which would have resulted in quicker response to users and reduced overhead to archives.

Previous web archive profiling works were focused on identifying different means to discover holdings of an archive and create archive profiles that maximize efficiency of aggregator lookup routing while minimizing various associated costs [4, 5, 7–10]. In our previous work we proposed CDXJ [2] as a potential format to represent archive profiles. In this work we are focusing on standardization of a format called *MementoMap*, which allows an efficient and flexible way of representing and disseminating archive profiles. This is an improvement over the previously proposed CDXJ format as it allows simpler and more storage efficient syntax and arbitrary depth in keys of the same file by using wildcards. This format is evaluated against the complete index of Portuguese Web Archive called Arquivo.pt (PWA) with various levels of details to assess associated costs and accuracy [6].

2 MEMENTOMAP USING UKVS

MementoMap is a framework for profiling web archives and expressing their holdings in an adaptive and flexible way to easily scale by utilizing Unified Key Value Store (UKVS) [1]. It is inspired by the simplicity of widely used `robots.txt` and `sitemap.xml` formats, but for a purpose other than search engine optimization. Figure 1 illustrates a sample *MementoMap* file that starts with some

```
1  !context ["https://git.io/mementomap"]
2  !id      {uri: "https://archive.example.org/"}
3  !fields  {keys: ["surt"], values: ["frequency"]}
4  !meta    {name: "Example Archive", year: 1996}
5  !meta    {type: "MementoMap"}
6  !meta    {updated_at: "2018-09-03T13:27:52Z"}
7  *                    54321/20000
8  com,*                10000+
9  org,arxiv)/          100
10 org,arxiv)/*         2500~/900
11 org,arxiv)/pdf/*     0
12 uk,co,bbc)/images/* 300+/20-
```

Figure 1: A Sample MementoMap in UKVS Format

metadata headers. Header lines are prefixed with "`!`" sign to ensure they are separated from data lines and surfaced on top when the file is sorted. The "`!fields`" header tells that the first column is a *SURT* (i.e., Sort-friendly URI Reordering Transform) [11] and is used as a lookup key (there can be more than one key column such as *Datetime* or *Language*) that is followed by a value column which holds "`frequency`" information. Unlike the standard SURT, *MementoMap* allows wildcard-based partial *URI Keys* to enable flexibility in how detailed or concise one wants it to be depending on use cases, full or partial knowledge about the archive's holdings, and available resources. Each data line can optionally also contain a single-line JSON block, which is not illustrated here for simplicity sake. The `frequency` column is formatted as "`[URI-M Count]/[URI-R Count]`" where both counts are optional and the separator is also optional if only the `URI-M Count` is present. Moreover, these counts can have an optional suffix character `+`, `-`, or `~` to express that the numbers are not exact and represent a lower bound, an upper bound, and a rough estimate respectively. The first data line in the example means there are a total of exactly 54,321 mementos (*URI-Ms*) of exactly 20,000 *URI-Rs* in the archive and the next line suggests that there are at least 10,000 mementos from the "`.com`" *TLD* (i.e., Top Level Domain). The next two lines suggest that there are 100 mementos of the `arxiv.org` homepage and many more captures of pages with deeper paths. However, the next line illustrates an exclusion of a subtree by being more specific under `/pdf/*` that has zero mementos. For a detailed description of the format, more variations of representations of archive profiles, and some other use case refer to the UKVS.

A *MementoMap* can either be generated by the archives themselves or by third parties based on their external observations. We propose the "`mementomap`" link relation for its dissemination and discovery. We implemented a tool to generate *MementoMap* efficiently from the index of an archive (or other means of listing archival holdings) and open-sourced it under MIT license[2]. The tool allows configuration options to identify criteria of rolling multiple occurrences of a *URI Key* prefix into a shorter key with wildcard recursively. The tool also implements a binary search mechanism in *MementoMap* files.

[1] https://memgator.cs.odu.edu/api.html

[2] https://github.com/oduwsdl/MementoMap

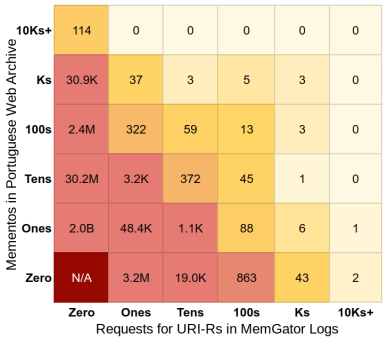


Figure 2: Overlap Between Archived and Accessed Resources

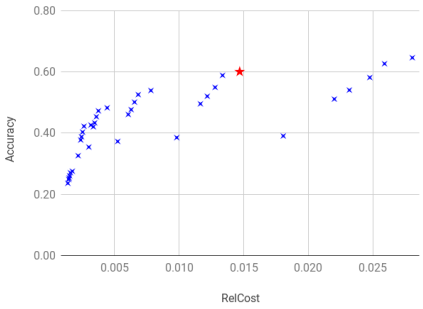


Figure 3: Relative Cost vs. Lookup Routing Accuracy

3 EVALUATION

For evaluation, we used the complete index of Arquivo.pt containing about 5B mementos of over 2B unique URI-Rs (i.e., original URIs) and 3.3M unique URIs in ODU's MemGator logs looked up over 5.2M times over a period of more than three years. Figure 2 shows a breakdown of what people are looking for in archives and what web archives hold. The 1.1K entry in the "Ones" row and "Tens" column shows that there are over a thousand *URI-Rs* that were requested 10–99 times in *MemGator* and each has 1–9 mementos in PWA. Large numbers in the "Zero" column show there are a lot of mementos that are never requested from *MemGator* (a usage-based profile might miss this data). Similarly, the "Zero" row shows there are a lot of requests that have zero mementos in PWA (a content-based profile might miss this data). The (Zero, Zero) corner suggests there are undetermined number of *URI-Rs* that were never archived or accessed. Irrespective of how the profile information was generated, *MementoMap* framework allows efficient means to express both what an archive holds and what it does not.

To evaluate the accuracy of *MementoMap*-based profiles, we first sampled all the unique *HTML* URI-Rs from the index that return `200` status code, which resulted in a dataset of about 1B unique entries (almost half of the original index). We then generated many *MementoMap* files from this using different configuration options to optimize the holdings representation and measured the accuracy of these when evaluated against MemGator logs. Figure 3 shows the relationship of relative cost (i.e., the ratio of number of *URI Keys* in the *MementoMap* and total number of unique *URI-Rs* in the archive) vs. routing accuracy (i.e., URIs who's presence or absence in the archive was correctly identified). The highlighted data point in the figure shows that there is a configuration that produces a *MementoMap* with only about 1.5% of the unique entries in the index and still correctly routes lookup requests with 60% accuracy with 100% recall. The accuracy can further be improved by 1) exploring other optimal configurations for subtree pruning, 2) generating profiles with the full index, not just a sample, and 3) including entries for absent resources from the "Zero" row of the Figure 2.

4 CONCLUSIONS AND FUTURE WORK

We introduced *MementoMap*, a serialization and dissemination format for archive profiles based on *UKVS*. The format allows arbitrary level of details for individual records and gives ability to dynamically roll well-populated subtrees up. We evaluated it against a large index of Arquivo.pt with billions of mementos and three years of a Memento aggregator log. The format is suitable for both small and large web archives and scales well. We open-sourced our implementation of *MementoMap* generation and binary search.

UK Web Archive, Arquivo.pt, National Records of Scotland, and National Library of Australia have recently shown interest in being able to express the summary of their holdings. As a pilot project these archives can be invited to generate *MementoMap* from their collections using our tool and advertise it using the "`mementomap`" link relation. Memento Aggregator services such as LANL's Time-Travel and ODU's MemGator can then leverage this information to better route lookup requests to these archives.

5 ACKNOWLEDGEMENTS

We thank Fernando Melo and Daniel Gomes from Arquivo.pt for the CDX dataset. Supported in part by NSF grant IIS-1526700.

REFERENCES

[1] Sawood Alam. 2019. Unified Key Value Store (UKVS). https://github.com/oduwsdl/ORS/blob/master/ukvs.md.
[2] Sawood Alam, Ilya Kreymer, and Michael L. Nelson. 2015. Object Resource Stream (ORS) and CDX-JSON (CDXJ). https://github.com/oduwsdl/ORS.
[3] Sawood Alam and Michael L. Nelson. 2016. MemGator - A Portable Concurrent Memento Aggregator. In *Proceedings of the 16th ACM/IEEE-CS Joint Conference on Digital Libraries (JCDL '16)*.
[4] Sawood Alam, Michael L. Nelson, Herbert Van de Sompel, Lyudmila L. Balakireva, Harihar Shankar, and David S. H. Rosenthal. 2016. Web Archive Profiling Through CDX Summarization. *International Journal on Digital Libraries* 17, 3 (2016), 223–238. https://doi.org/10.1007/s00799-016-0184-4
[5] Sawood Alam, Michael L. Nelson, Herbert Van de Sompel, and David S. H. Rosenthal. 2016. Web Archive Profiling Through Fulltext Search. In *Proceedings of 20th International Conference on Theory and Practice of Digital Libraries, TPDL 2016*. 121–132.
[6] Sawood Alam, Michele C. Weigle, Michael L. Nelson, Fernando Melo, Daniel Bicho, and Daniel Gomes. 2019. MementoMap Framework for Flexible and Adaptive Web Archive Profiling. In *Proceedings of the 19th ACM/IEEE-CS Joint Conference on Digital Libraries (JCDL '19)*.
[7] Ahmed AlSum, Michele C. Weigle, Michael L. Nelson, and Herbert Van de Sompel. 2014. Profiling Web Archive Coverage for Top-Level Domain and Content Language. *International Journal on Digital Libraries* 14, 3-4 (2014), 149–166.
[8] Nicolas Bornand, Lyudmila Balakireva, and Herbert Van de Sompel. 2016. Routing Memento Requests Using Binary Classifiers. In *Proceedings of the 16th ACM/IEEE-CS Joint Conference on Digital Libraries (JCDL '16)*. 63–72.
[9] Robert Sanderson. 2012. Global Web Archive Integration with Memento. In *Proceedings of the 12th ACM/IEEE Joint Conference on Digital Libraries*. 379–380.
[10] Robert Sanderson, Herbert Van de Sompel, and Michael L. Nelson. 2012. IIPC Memento Aggregator Experiment. http://www.netpreserve.org/sites/default/files/resources/Sanderson.pdf.
[11] Kristinn Sigurðsson, Michael Stack, and Igor Ranitovic. 2006. Heritrix User Manual: Sort-friendly URI Reordering Transform. http://crawler.archive.org/articles/user_manual/glossary.html#surt.
[12] Herbert Van de Sompel, Michael L. Nelson, and Robert Sanderson. 2013. HTTP Framework for Time-Based Access to Resource States – Memento, Internet RFC 7089. https://tools.ietf.org/html/rfc7089.

Impact of HTTP Cookie Violations in Web Archives

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

ABSTRACT

Certain *HTTP Cookies* on certain sites can be a source of content bias in archival crawls. Accommodating *Cookies* at crawl time, but not utilizing them at replay time may cause cookie violations, resulting in defaced composite mementos that never existed on the live web. To address these issues, we propose that crawlers store *Cookies* with short expiration time and archival replay systems account for values in the `Vary` header along with URIs.

1 INTRODUCTION AND BACKGROUND

For a long time we have been observing a strange behavior of various web archives when accessing mementos [11] of Twitter pages, some of the mementos would be replayed in non-English languages. This happens even if those Twitter timelines belong to English-speaking personalities, archived using crawlers in North America, and were not requested in any specific language explicitly as shown in Figure 1. After a thorough investigation we figured it out that it is happening due to the use of *HTTP Cookies* for content negotiation by Twitter [4, 9]. We found that almost half of the mementos of Barack Obama's Twitter timeline out of over 9,000 properly archived mementos in five different web archives were in non-English languages, of which, almost half were in Kannada (a regional Indian language) alone, and remaining in 45 other languages (as shown in Figure 2). While language diversity in web archives is generally a good thing, this non-uniform bias is disconcerting when a page is archived in a language not anticipated.

One day we were looking at a Twitter timeline's memento which should have been in English, but was primarily in Portuguese (for the reason described above), after a while we noticed that a notification appeared in Urdu, suggesting that there were 20 new tweets (as shown in Figure 3). On further inspection found that the page contained a sidebar block in English too. Apparently, we were seeing a defaced composite memento of a page that perhaps never existed on the live web. We knew that live-leakage (also known as Zombies) [6] and temporal violations [1] can cause such malformed memento reconstruction and we also knew their potential prevention techniques [3, 8]. However, this mixed-language Twitter timeline issue cannot be explained by zombies nor temporal violations. After a thorough investigation we found that *Cookies* were again the reason behind this replay issue [2, 10].

HTTP is stateless, but often applications need to maintain state information between a client and a server. This is often done with the help of *Cookies* [5]. Servers can send one or more `Set-Cookie` headers containing strings of name-value pairs along with scope (domain and path) and expiration information. Clients store them and send them back with each request in the scope using `Cookie` header until expired or removed. *Cookies* are used for session management, personalization, content-negotiation, tracking, and client-side key-value store. The latter is less common now after wide adoption of *LocalStorage* and other similar techniques in web browsers.

Figure 1: Barack Obama's Twitter Timeline is Archived in Urdu, Which Should be in English

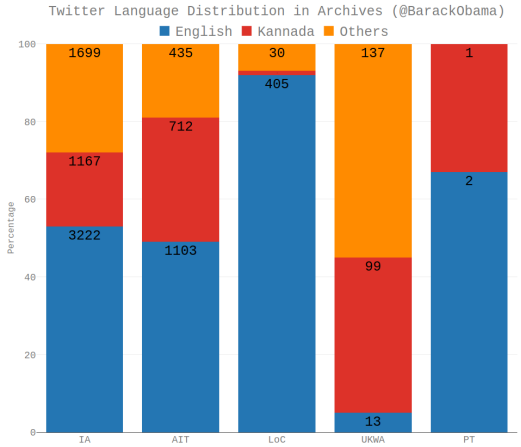


Figure 2: Language Distribution of Mementos of Barack Obama's Twitter Timeline in Different Web Archives

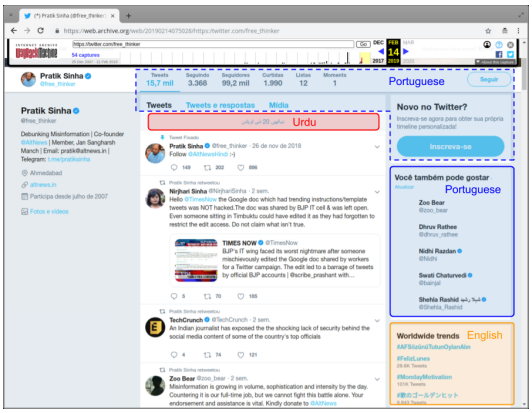


Figure 3: A Memento of a Twitter Timeline Simultaneously in Multiple Languages (Portuguese, English, and Urdu)

```
<link rel="alternate" hreflang="x-default"
                      href="https://twitter.com/">
<link rel="alternate" hreflang="fr"
                      href="https://twitter.com/?lang=fr">
... [45 LINKS TRUNCATED] ...
<link rel="alternate" hreflang="kn"
                      href="https://twitter.com/?lang=kn">
```

Figure 4: 47 Alternate Language Links in Twitter

```
$ curl -s -c /tmp/tt.cook https://twitter.com/?lang=ar \
>       | grep "<html"
<html lang="ar" data-scribe-reduced-action-queue="true">
$ grep lang /tmp/tt.cook
twitter.com FALSE / FALSE 0 lang ar
$ curl -s https://twitter.com/ | grep "<html"
<html lang="en" data-scribe-reduced-action-queue="true">
$ curl -s -H "Accept-Language: ur" https://twitter.com/ \
>       | grep "<html"
<html lang="ur" data-scribe-reduced-action-queue="true">
$ curl -s -b /tmp/tt.cook https://twitter.com/ | grep "<html"
<html lang="ar" data-scribe-reduced-action-queue="true">
```

Figure 5: Language Content Negotiation in Twitter Using Query Parameters, Accept-Language, and Cookies

2 INTERNATIONALIZATION IN TWITTER

Twitter uses standard method of internationalization in its publicly accessible pages by including `alternate` links in 47 supported languages and the `x-default` landing language (as illustrated in Figure 4) to help search engines point users with different locales to the correct language. This technique is utilized by many other multi-lingual sites such as Facebook and Instagram. However, unlike other popular multi-lingual sites, when accessing a language-specific URI (that contains a `lang` query parameter), Twitter sets a `lang` *Cookie* with the corresponding language (as illustrated in Figure 5). This *Cookie* sticks throughout the session and forces all subsequent pages to be served in that language until another language-specific URI overwrites the *Cookie*. This essentially means Twitter performs language content negotiation using *Cookie* header, though it does not acknowledge it in a `Vary` header.

3 COOKIE VIOLATIONS

Some websites insist that certain *Cookies* are present in a request before they return desired content otherwise they issue redirects and attempt to set those *Cookies*. Failure to send their desired *Cookies* in subsequent requests may turn such sites into crawler traps without any useful content. Web archiving crawlers such as Heritrix[1] have built-in support for cookies. However, the web surfing pattern of crawlers is generally breadth-first-style and comprehensive (not necessarily how human surf the web) for which they use frontier queue of URIs to be crawled. In case of the Twitter's example above, when one of the language-specific `alternate` link is crawled, it impacts all the subsequent non-language-specific URIs due to the `lang` sticky *Cookie*. Kannada being the last language in the list (in Figure 4) gets more exposure before it gets overwritten by another language, resulting in the disproportionate language bias.

Popular archival replay systems (such as OpenWayback[2] and PyWB[3]) utilize only the canonicalized *URI-R* and the datetime of the capture to select a memento to replay. Other request headers that might have been used for content negotiation (such as `Accept-Language` or `Geolocation` etc.) are ignored at replay. Traditional crawlers did not execute JavaScript, so the likelihood of a custom request header being utilized during crawling was minimal, but it is changing with headless browser-based crawlers. *Cookies*, however, have been supported even in traditional crawlers that are used by some sites for content negotiation (as is the case with Twitter). Moreover, aggregating private archives [7] containing authenticated resources without isolating them based on session *Cookies* has some privacy implications.

Based on our assessment we propose that *Cookies* in crawlers are kept short-lived and pruned frequently to minimize the impact of sticky *Cookies*. Accommodating *Cookies* (or other headers that affect the response) at capture/crawl time, but not utilizing them at replay time has this consequence of cookie violations, resulting in defaced composite mementos. On the contrary, blindly utilizing every *Cookie* as a filter at replay would result in many false negatives. Unfortunately, *Cookie* names are opaque strings and carry no agreed upon semantics to identify ones that affect the payload.

[1]https://github.com/internetarchive/heritrix3
[2]https://github.com/iipc/openwayback
[3]https://github.com/webrecorder/pywb

4 CONCLUSIONS AND FUTURE WORK

We identified that certain *Cookies* on certain sites can be a source of content bias in archival crawls. To address this issue we propose that crawlers store *Cookies* with short expiration time explicitly, irrespective of the original value. We also identified that *Cookie Violations* at replay time have the potential to deface composite mementos and reconstruct pages from web archives that never existed on the live web. Archival replay systems need to behave like HTTP proxies or cache servers that accommodate values in the `Vary` header along with URIs. Not every *Cookie* is created equal, those impacting the content need to be identified and accounted for at replay. This is a difficult problem which opens up the possibility for a more extensive research to fully address the issue.

5 ACKNOWLEDGEMENTS

We thank Plinio Vargas for his contribution in our early investigation. This work is supported in part by NSF grant IIS-1526700.

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 the 26th ACM Conference on Hypertext & Social Media*. 257–266.
[2] Sawood Alam. 2019. Cookie Violations Cause Archived Twitter Pages to Simultaneously Replay in Multiple Languages. https://ws-dl.blogspot.com/2019/03/2019-03-18-cookie-violations-cause.html.
[3] Sawood Alam, Mat Kelly, Michele 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)*. ACM, New York, NY, USA, 237–240. https://doi.org/10.1109/JCDL.2017.7991579
[4] Sawood Alam and Plinio Vargas. 2018. Cookies Are Why Your Archived Twitter Page Is Not in English. https://ws-dl.blogspot.com/2018/03/2018-03-21-cookies-are-why-your.html.
[5] Adam Barth. 2011. HTTP State Management Mechanism, Internet RFC 6265. https://tools.ietf.org/html/rfc6265.
[6] Justin F. Brunelle. 2012. Zombies in the Archives. https://ws-dl.blogspot.com/2012/10/2012-10-10-zombies-in-archives.html.
[7] Mat Kelly, Michael L. Nelson, and Michele C. Weigle. 2018. A Framework for Aggregating Private and Public Web Archives. In *Proceedings of the 18th ACM/IEEE-CS Joint Conference on Digital Libraries (JCDL '18)*. 273–282. https://doi.org/10.1145/3197026.3197045
[8] Ada Lerner, Tadayoshi Kohno, and Franziska Roesner. 2017. Rewriting History: Changing the Archived Web from the Present. In *Proceedings of the 2017 ACM SIGSAC Conference on Computer and Communications Security*. 1741–1755.
[9] David S. H. Rosenthal. 2018. All Your Tweets Are Belong To Kannada. https://blog.dshr.org/2018/04/all-your-tweets-are-belong-to-kannada.html.
[10] David S. H. Rosenthal. 2019. The 47 Links Mystery. https://blog.dshr.org/2019/03/the-47-links-mystery.html.
[11] Herbert Van de Sompel, Michael L. Nelson, and Robert Sanderson. 2013. HTTP Framework for Time-Based Access to Resource States – Memento, Internet RFC 7089. https://tools.ietf.org/html/rfc7089.