2013

A Method for Identifying Personalized Representations in Web Archives

Mat Kelly
Old Dominion University

Justin F. Brunelle
Old Dominion University

Michele C. Weigle
Old Dominion University

Michael L. Nelson
Old Dominion University

Original Publication Citation

Kelly, M., Brunelle, J.F., Weigle, M.C., & Nelson, M.L. (2013). A method for identifying personalized representations in Web archives. *D-Lib Magazine, 19*(11/12), 1-11. doi: 10.1045/november2013-kelly

Search D-Lib:

D-Lib Magazine

The Magazine of Digital Library Research

D-Lib Magazine

November/December 2013
Volume 19, Number 11/12
Table of Contents

A Method for Identifying Personalized Representations in Web Archives

Mat Kelly, Justin F. Brunelle, Michele C. Weigle, and Michael L. Nelson
Old Dominion University
{mkelly,jbrunelle,mweigle,mln}@cs.odu.edu

doi:10.1045/november2013-kelly

Printer-friendly Version

Abstract

Web resources are becoming increasingly personalized — two different users clicking on the same link at the same time can see content customized for each individual user. These changes result in multiple representations of a resource that cannot be canonicalized in Web archives. We identify characteristics of this problem by presenting a potential solution to generalize personalized representations in archives. We also present our proof-of-concept prototype that analyzes WARC (Web ARChive) format files, inserts metadata establishing relationships, and provides archive users the ability to navigate on the additional dimension of environment variables in a modified Wayback Machine.

Introduction

Personalized Web resources offer different representations [8] to different users based on the `user-agent` string and other values in the HTTP request headers, GeoIP, and other environmental factors. This means Web crawlers capturing content for archives may receive representations based on the crawl environment which will differ from the representations returned to the interactive users. In summary, what we archive is increasingly different from what we as interactive users experience.

Web servers often provide lighter-weight representations to mobile browsers and the larger, full-feature representations to desktop browsers. Content viewed from a mobile browser is often different than content viewed from a traditional desktop browser [9, 28]. This allows mobile devices to more easily and more quickly browse the Web. With the increasing prevalence of mobile browsers on the Web (50% - 68% of sites have mobile representations [25]), it is becoming important to capture these mobile representations of resources.

Mobile pages often contain links to additional resources instead of embedded text and often reduce the number of images embedded in the page [9]. For example, the mobile representation of http://espn.go.com/ contains a section on ESPN Videos, while the desktop representation does not. When http://espn.go.com (the "original resource", identified by URI-R), is accessed, it redirects to http://m.espn.go.com, effectively giving two separate but related URI-R values that go into the archive. Subsequently, the URIs of their archived versions ("mementos", identified by URI-M) are indexed separately and the relationships

between URI-Rs and URI-Ms is not presented to the user. To quantify the differences, the desktop representation contains 201 links, while the mobile representation contains only 58 links. These link sets are mutually exclusive, with the mobile representation linking to specific resources (such as box-scores and gamecasts) while the desktop representation links to higher-level resources (such as narratives that include box-scores and may have links to gamecasts). A user may review news articles or other content on a mobile device and be unable to recall the article in an archive. To capture and record the complete set of content at http://espn.go.com, each of these different representations, both mobile and desktop, need to be stored in Web archives.

Heritrix [18], the Internet Archive's crawler, offers archivists the ability to modify the `user-agent` string to simulate a variety of browsers during archiving. Heritrix can crawl the mobile Web by setting its `user-agent` string to a mobile browser. This can potentially lead to multiple representations of the same content being captured. If a desktop and mobile representation of the same resource are captured simultaneously, they will potentially collide in an archive.

As archiving expands into the mobile and other dynamic domains, archives will contain representations generated with a variety of environmental influences. Therefore, it is no longer sufficient to only navigate archives in the temporal dimension; archives must also provide users the opportunity to understand how a representation was crawled and navigate representations by their environmental influences. The factors influencing the representations seen by a crawler or user need to be recorded and presented to the user viewing the captured representation.

In this work, we explore the issue of personalized representations in Web archives, propose a framework to solve this problem, and present a proof-of-concept prototype that integrates personalized representations. We study live resources (identified by URI-Rs) seen by users and proposed methods for mapping archived representations called mementos (identified by URI-Ms) to a canonical representation. This prototype extends the description of mementos from only "when" they were archived (temporal dimension) to "where" and "how" (GeoIP and browser environments). Users can then browse between mementos based on temporal or environmental dimensions.

Personalized, Anonymous Representations

Dynamic and personalized representations of Web 2.0 resources that are generated by technologies such as JavaScript can differ greatly depending on several factors. For example, some sites attempt to provide alternate representations by interpreting the `user-agent` portion of the HTTP GET headers and use content negotiation to determine which representation to return.

We ran a pair of limited crawls of the cnn.com front page with Heritrix 3.1 and then accessed the mementos captured by Heritrix with a desktop Mac and an Android phone. The first crawl captured the cnn.com front page and specified a desktop version of the Mozilla browser as the `user-agent` in the header string, as seen in Figure 1. The resulting Web ARChive (WARC) file [26] is viewed in a local installation of the Wayback Machine [29] and is shown in Figures 3(a) and 3(c).

The second crawl captured the cnn.com front page and specified an iPhone version of the Mozilla browser as the `user-agent` string in the header, as seen in Figure 2. The resulting WARC, as viewed in the Wayback Machine, is shown in Figures 3(b) and 3(d). The mobile and desktop representations differ in archives, but their relationship as permutations of each other is neither recorded nor seen by users; a user of the Wayback Machine may not understand how these representations are generated since they are identified by the same URI-R. We refer to these differing representations of the same URI-R built with differing environments as *personalized representations* of the resource R.

The headers in Figures 1 and 2 reference the `user-agent` string with http://yourdomain.com, which is a place holder for the URI for whom the crawl is being executed. For example, a crawl originating from Old Dominion University's Computer Science department would read http://www.cs.odu.edu/.

```
WARC/1.0
WARC-Type: request
WARC-Target-URI: http://www.cnn.com/
WARC-Date: 2013-03-05T16:57:00Z
WARC-Concurrent-To: <urn:uuid:d338e6e5-6854-329b-adbb-de70a62e11f0>
WARC-Record-ID: <urn:uuid:a3e3e726-97bc-3cca-af5e-d83dd1827e05>
Content-Type: application/http; msgtype=request
Content-Length: 266

GET / HTTP/1.0
```

```
User-Agent: Mozilla/5.0 (compatible; heritrix/3.1.0 +http://yourdomain.com)
Connection: close
Referer: http://cnn.com/
Accept: text/html,application/xhtml+xml,application/xml;q=0.9,*/*;q=0.8
Host: www.cnn.com
Cookie: CG=US:--:--; CG=US:--:--
```

Figure 1. HTTP GET request from Heritrix with the desktop Mozilla `user-agent`.

```
WARC/1.0
WARC-Type: request
WARC-Target-URI: http://www.cnn.com/
WARC-Date: 2013-03-05T16:38:08Z
WARC-Concurrent-To: <urn:uuid:cc7f75cc-fbaa-352a-8939-7cf5dd7792c7>
WARC-Record-ID: <urn:uuid:fcc902ba-b327-3f43-a5a8-a29861c4fa7e>
Content-Type: application/http; msgtype=request
Content-Length: 400

GET / HTTP/1.0
User-Agent: Mozilla/5.0 (iPhone; U; CPU iPhone OS 4_0 like Mac OS X; en-us)
AppleWebKit/532.9 (KHTML, like Gecko) Version/4.0.5 Mobile/8A293 Safari/
6531.22.7 (compatible; heritrix/3.1.0 +http://yourdomain.com)
Connection: close
Referer: http://cnn.com/
Accept: text/html,application/xhtml+xml,application/xml;q=0.9,*/*;q=0.8
Host: www.cnn.com
Cookie: CG=US:--:--; CG=US:--:--
```

Figure 2. HTTP GET request from Heritrix with the iPhone Mozilla `user-agent`.

These examples illustrate the potential for collisions of personalized representations with the same URI within Web archives. The potential exists for a mobile and a desktop representation of a page (or constituent and embedded resources) to be captured simultaneously, and therefore be indistinguishable. The live Web version of cnn.com is identified by http://www.cnn.com/ regardless of the `user-agent` string and resulting representation. While the distinction between representations could be accomplished with the `Vary` HTTP headers [8] which would alert caches and clients that the representation is personalized, CNN does not employ this header. The result is both the mobile and desktop versions use http://www.cnn.com/ for the URI-R values.

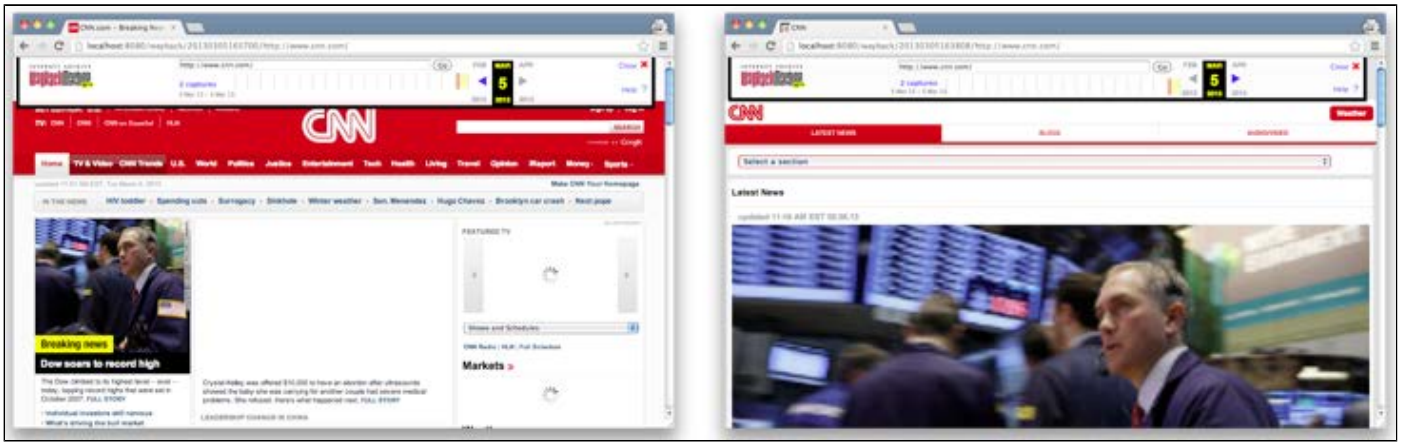

Figure a: The cnn.com memento when crawled by Heritrix with a desktop Mozilla `user-agent` accessed from a

Figure b: The cnn.com memento when crawled by Heritrix with an iPhone mozilla `user-agent` accessed from a

Mac.

Mac.

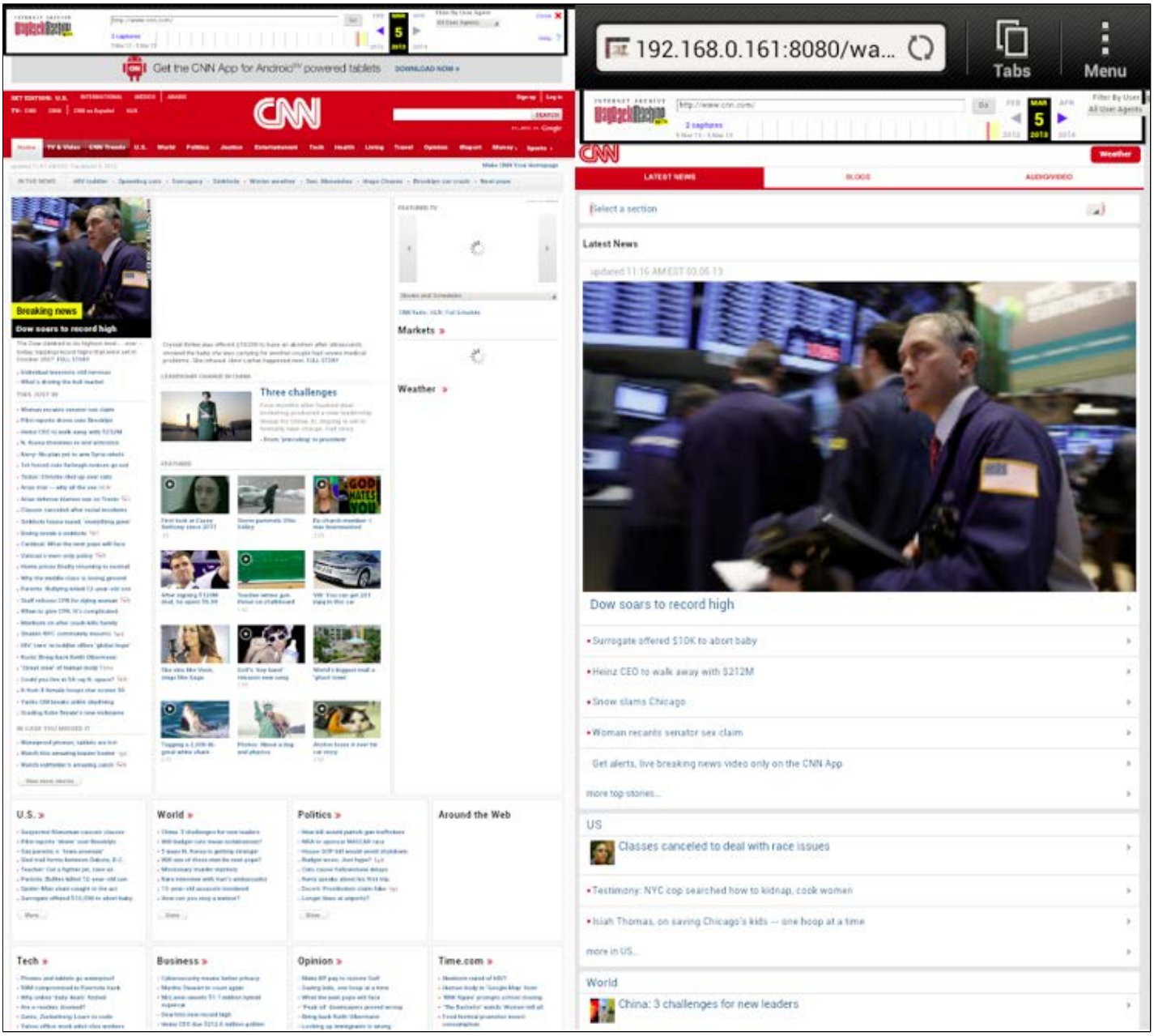

Figure c: The cnn.com memento when crawled by Heritrix with a desktop Mozilla `user-agent` accessed from an Android Phone.

Figure d: The cnn.com memento when crawled by Heritrix with an iPhone Mozilla `user-agent` accessed from an Android Phone.

Figure 3: Mementos differ based on the parameters influencing the representations at crawl/capture time and the devices used to access the mementos.

Some sites provide local news and weather content depending on the GeoIP of the requester. For example, a user requesting http://www.nbcnews.com/ without an IP (via anonymous browser) will get news and weather from New York, NY with a request for the user to enter a local zip code (Figure 4b). Alternatively, a user requesting the page from Suffolk, Virginia will receive news and weather from neighboring Portsmouth, Virginia (Figure 4a).

Figure a: When browsing from Suffolk, VA, nbcnews.com shows headlines from Suffolk and Portsmouth, VA.

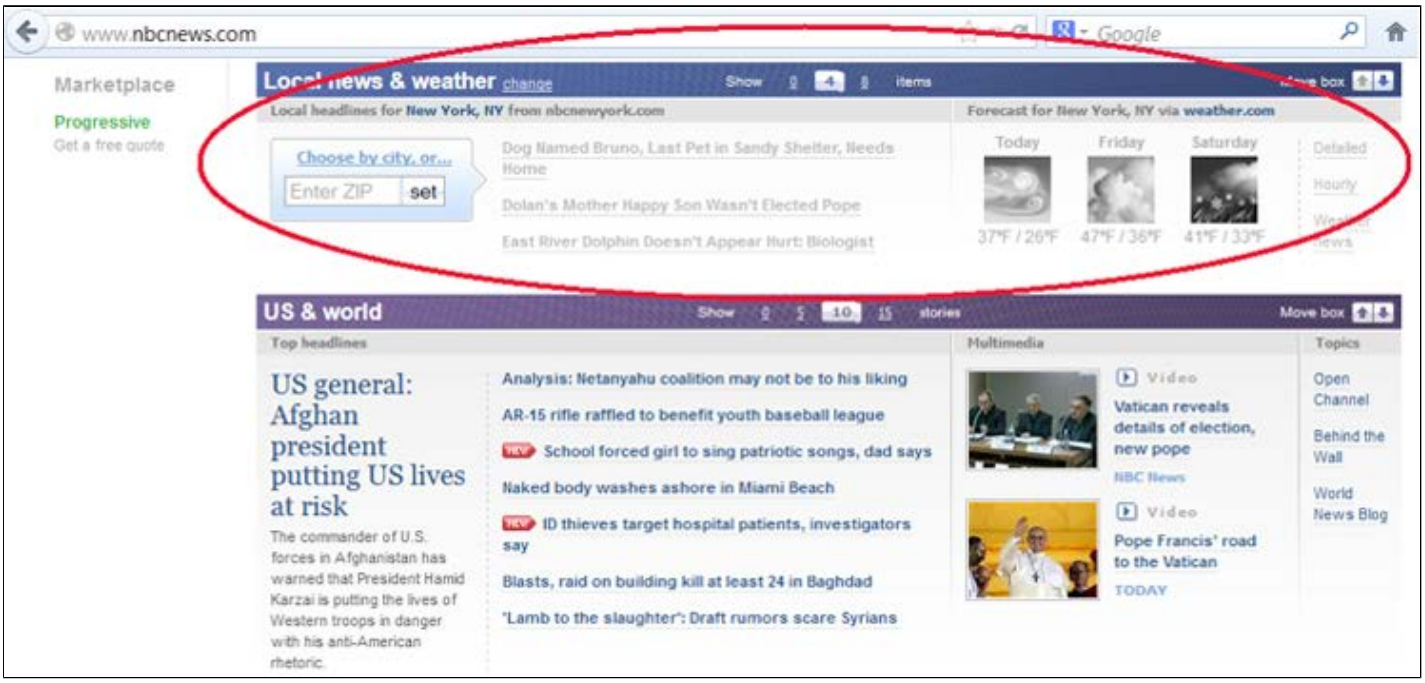

Figure b. When browsing anonymously, nbcnews.com shows headlines from New York City, NY.

Figure 4: The live versions of nbcnews.com differ based on the ability to interpret the GeoIP of the requester.

Figure 4 contains examples of personalized representations that pose problems for archives. The factors that influence the representations need to be documented, and users of archives should be able to not only browse mementos by time, but also by which representations are available.

The representations in Figure 4 demonstrate that environment variables do not have an impact limited to the *look and feel* or stylesheets of a resource. The content in the page (specifically, the local news stories and weather shown in the representations in Figure 4) changes depending on the environment, in this case the GeoIP. For this reason, we need to link to the original

request, which is not referenced in HTTP response headers.

Related Work

The prevalence of dynamic Web 2.0 content has increased on the Web [19, 22]. Several papers have identified the roles of JavaScript and other Web 2.0 technologies in a social and Web environment that is becoming increasingly unarchivable [4, 6, 15, 23, 35]. The mobile Web is also growing, with mobile devices contributing 18% of all Web traffic. [13] Additionally, the iPhone has a version of the Memento client [30], which has the potential to highlight the problems identified in this paper.

Several efforts have investigated how to capture dynamic content, although not with archival intent. These efforts, however, contribute greatly to the understanding of how dynamic resources can change on the client. Livshits et al. at Microsoft Research have performed extensive research in the monitoring and capture of client-side JavaScript [10—12, 16]. Most of this work has targeted debugging and security monitoring. Other works have included the interpretation of client-side state for crawling purposes [1, 5, 14]. Several efforts have focused on monitoring Web 2.0 resources on the client [2, 3, 10, 11, 17, 21]. Other work has studied the capture of dynamic content for search engine optimization or search engine indexing [24].

Identification Models

It is important to understand the potential methods of identifying a memento that has personalized representations. There are three approaches that we considered: identification in the URI, content negotiation, and modifications to the client.

The first option would modify the URI that identifies an archived version of a resource to include information about the representation using HashBang (#!) URIs [7, 27, 34]. The information in the HashBang portion of the URI will identify the specific representation state, such as a mobile or desktop version. However, this URI does not generalize and does not integrate with existing archives which index resources by URI.

The second option using content negotiation, similar to that done by the Memento Framework's Memento-Datetime [20, 31—33] would specify a URI-M to access and also specify the environment variables — like `user-agent` or GeoIP — to use when picking a representation of the resource. However, this shifts the responsibility away from the server, requiring a specialized client.

The third option starts with a post-processing of the content captured by the crawler. The post-processor needs to provide metadata embedded in the memento that specifies the environment variables that went into creating the captured representation. The client that replays the representation needs to specify which representation is being shown to archive users. There are also navigation controls to allow for migration between representations captured with different environment variables, effectively giving the user the ability to navigate mementos based on the environment as well as traditional temporal parameters.

This is the most complex of the options provided, but is also the most extensible and effective. This method also requires the most modification of existing tools, and is likely only effective for a subset of archives until widespread adoption takes place. However, this provides the highest degree of information to be provided to and controlled by the users.

We selected this third option for implementation. We created a script to identify representations of resources that were generated with different environment variables, and modify WARC files to identify these different related representations as personalized representations. Then we modified a local installation of the Wayback Machine to allow for navigation between these personalized representations. The details of these efforts are explained in the remainder of this paper.

During consideration of the impact these changes will have on services, we kept the Memento Framework in mind [32]. We can modify Memento TimeMaps to include the relationships identified by Web archives. This effectively utilizes the services that already consume information from the archives to carry related representations from the archives to the users. We propose a multipart HTTP response that returns multiple related TimeMaps as part of a single request. When a user requests a TimeMap for http://www.example.com/, the archive discovers that there is a mobile representation, http://m.example.com/ of the resource and returns the TimeMap for the original representation and the mobile representation of example.com as a multipart message, with each separate TimeMap being sent as a separate part.

Identifying Personalized Representations

We identified related representations by analyzing the WARC records' metadata and finding URIs that are derivations of one another (such as http://m.example.com/ and http://www.example.com/). We read the WARCs and from the `warcinfo` record

we extracted the `user-agent` string and IP of the crawler for the WARC file's records. Then, we extracted each URI-R from the `response` records and wrote the metadata records.

We wrote a post-processor script to discover personalized representations given a directory of WARC files, simulating the analysis of an entire local archive. The script reads the individual WARC records and extracts information about the mementos such as the `user-agent` used, the crawler settings, and the URN of each record. We then used URI guessing and `user-agent` analysis [25] to find related WARC records based on the extracted information. A URI such as http://m.example.com/ is treated as a personalized representation of http://example.com/. Similarly, http://www.cnn.com/ when accessed with a mobile browser (and associated `user-agent` string) is treated as a personalized representation of http://www.cnn.com/ accessed with a desktop browser.

Once we identified the personalized representations of the resources, we created a metadata record for each personalized representation. In the record (shown in Figure 5), we used values to identify the `user-agent` (`http-header-user-agent`), submitting user's (or crawler's) IP address (`contributor-ip`), the URN of the personalized representation of this resource (`hasVersion`), and the geographic location of the contributor (`geo-location`).

```
WARC/1.0
WARC-Type: metadata
WARC-Target-URI: http://www.cnn.com/
WARC-Date: 2013-03-05T16:57:00Z
WARC-Concurrent-To: <urn:uuid:d338e6e5-6854-329b-adbb-de70a62e11f0>
WARC-Record-ID: <urn:uuid:d32e3c28-c5d0-3df2-b1f8-1ffb93a38865>
Content-Type: application/warc-fields
Content-Length: 216

http-header-user-agent: Mozilla/5.0 (compatible; heritrix/3.1.0
+http://yourdomain.com)
contributor-ip: 192.168.1.7
hasVersion: <urn:uuid:cc7f75cc-fbaa-352a-8939-7cf5dd7792c7>
geo-location: multicast
description:
```

Figure 5. A sample metadata record that creates a link to the mobile representation of the desktop cnn.com resource.

The metadata records are inserted into the WARC file containing the personalized representation to identify all other representations of the resource available in the local archive.

The metadata record in Figure 5 identifies the WARC record for the desktop representation of cnn.com — identified by `urn:uuid:d338e6e5-6854-329b-adbb-de70a62e11f0` — as a personalized representation of the mobile WARC record of cnn.com — identified by `urn:uuid:cc7f75cc-fbaa-352a-8939-7cf5dd7792c7` — via the hasVersion value. It also records the `user-agent` for the concurrent record and the IP address of the contributor — in this case, the host running the Heritrix crawler — via the `contributor-ip` value. The geographic location of the IP address is also identified via the `geo-location` value. In this record, the contributor uses a multicast address that cannot be mapped to a geographic location. This `metadata` record can be expanded to include additional information about the representation being described through the `description` value.

A complementing metadata record for the mobile representation of cnn.com identifies a link back to the desktop representation. This creates a bidirectional link for personalized representations so that each personalized representation identifies its counterpart. The examples in this paper have focused on 1-to-1 links between personalized representations (a single desktop and a single mobile representation). However, it is possible to have one-to-n relationships identified in the metadata (multiple desktop representations and multiple mobile representations). This would identify multiple personalized representations for a single resource. The post-processing of WARCs will add time and complexity to the process of ingesting WARCs into a repository. Since records in WARC files are not sorted, our post-processing script uses linear match between two sets of mementos, which runs in $O(n^2)$ time. The metadata records that we add only increase the WARC file size by approximately 1 KB per personalized representation. This is a small size increase for WARC files that range from hundreds of MB to two GB.

Recognizing personalized representations

To replay the WARCs and provide a means to experience the relationship that we established between the mobile and desktop captures, we proposed an implementation consisting of a modification of Wayback Machine with an additional user interface element that allows the user to quickly toggle between the representations of the page captured. In this prototype (Figure 6), we provide a drop-down menu that displays all versions available for the current URI as established by the `metadata` record (Figure 5) that we added to the WARC during post-processing.

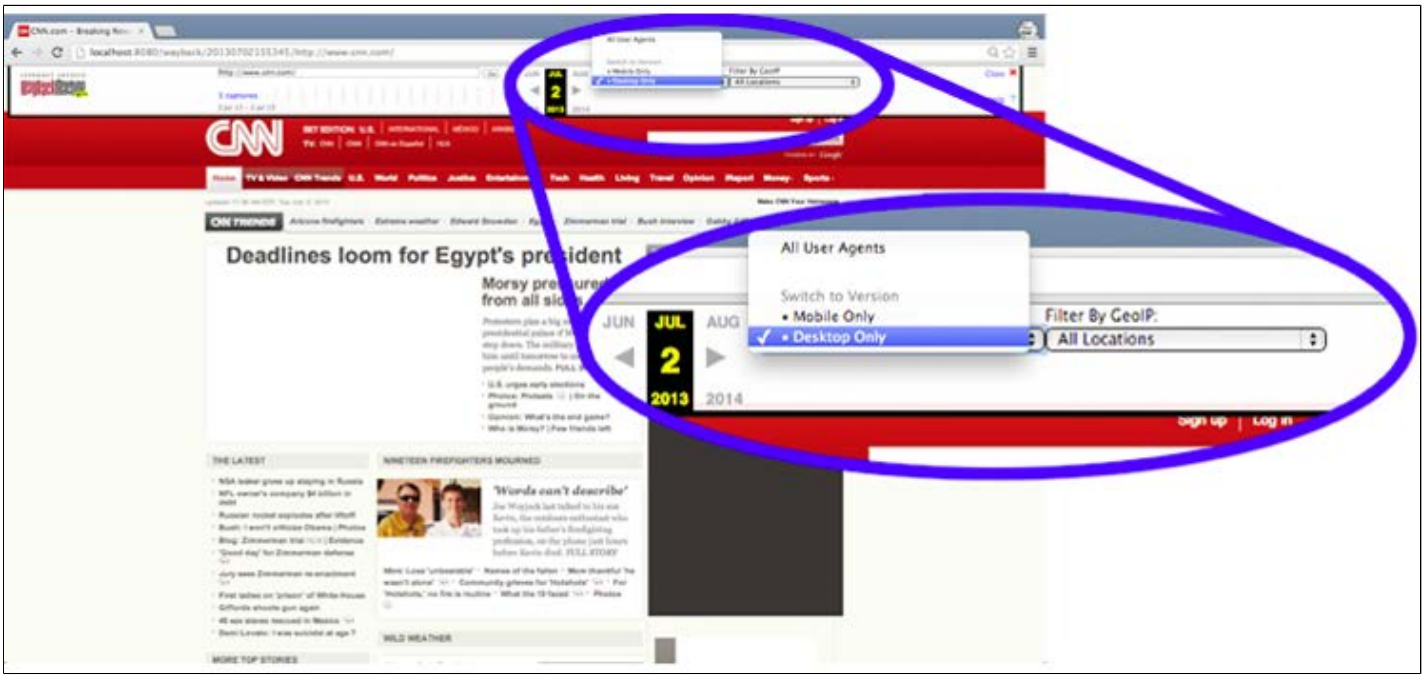

Figure a. Mobile and desktop versions of the cnn.com front page exist in archives. The user has the option of which version to access.

Figure b. Two source locations of the cnn.com front page are available from Washington, DC and Portsmouth, VA. The user has the option of which version to access.

Figure 6: Modifications to local Wayback allow for additional indexing and thus further reference for the URI currently being replayed to additional representations of the memento in the archive. In the

Wayback toolbar we have provided a means for a user to quickly view other representations available for the page being replayed. The dropdowns are conditionally displayed based on whether additional representations are available.

By comparing the various `user-agent` strings available in combination with the IP of the crawler, we can allow the user to choose which dimension (e.g., GeoIP and `user-agent`) is most important (only the `user-agent` dimension is shown) and retain that setting in traversing the available mementos. When a user selects an option in our modified UI, this setting is set in the user's browser as a cookie. This cookie is retrieved and the settings read to get a temporally adjacent memento from the TimeMap for the URI. Whereas replaying the archive is usually limited to the dimensions of *what* (URI) and *when* (Memento-Datetime), the UI extension adds the additional dimensions of *where* (GeoIP) and *how* (`user-agent`).

Conclusion

Current technologies introduce the opportunity for collisions in Web archives. The granularity of URI-Ms makes differentiating between representations impossible with the current configurations in archives. Temporal data alone is no longer sufficient to describe a memento; environmental variables must also be recorded and presented to the user, allowing the user to navigate between multiple dimensions of a representation. Users can then decide whether browsing on the temporal dimension or the environmental parameters is more suitable for their goals.

We present examples of the representation collision problem as well as a proof-of-concept solution. We use a post-processor to analyze WARCs and insert metadata identifying related representations of mementos. Through a modified Wayback Machine, we allow users to navigate personalized representations of mementos through their environmental parameters as well as on the temporal dimension. Our future work will focus on allowing users to nominate candidates for merging as related personalized representations, effectively providing a tool for crowd-sourcing the problem with personalized representations.

Acknowledgements

This work is supported in part by NSF grant 1009392 and the Library of Congress.

References

[1] K. Benjamin, G. von Bochmann, M. Dincturk, G.-V. Jourdan, and I. Onut. A Strategy for Efficient Crawling of Rich Internet Applications. In Web Engineering, Lecture Notes in Computer Science, pages 74—89, 2011. http://doi.org/10.1007/978-3-642-22233-7_6

[2] E. Benson, A. Marcus, D. Karger, and S. Madden. Sync Kit: A Persistent Client-Side Database Caching Toolkit for Data Intensive Websites. In Proceedings of the 19th international conference on World Wide Web, WWW '10, 2010. http://doi.org/10.1145/1772690.1772704

[3] S. Chakrabarti, S. Srivastava, M. Subramanyam, and M. Tiwari. Memex: A Browsing Assistant for Collaborative Archiving and Mining of Surf Trails. In Proceedings of the 26th VLDB Conference, 26th VLDB, 2000. http://dl.acm.org/citation.cfm?id=758378

[4] E. Crook. Web Archiving in a Web 2.0 World. In Proceedings of the Australian Library and Information Association Biennial Conference, pages 1—9, 2008. http://doi.org/10.1108/02640470910998542

[5] C. Duda, G. Frey, D. Kossmann, and C. Zhou. AjaxSearch: Crawling, Indexing and Searching Web 2.0 Applications. Proc. VLDB Endow., 1:1440—1443, August 2008. http://dl.acm.org/citation.cfm?id=1454195

[6] B. Fleiss. SEO in the Web 2.0 Era: The Evolution of Search Engine Optimization, 2007.

[7] Google. AJAX crawling: Guide for webmasters and developers, 2013.

[8] I. Jacobs and N. Walsh. Architecture of the World Wide Web, Volume One. Technical Report W3C Recommendation 15 December 2004, W3C, 2004.

[9] A. Jindal, C. Crutchfield, S. Goel, R. Kolluri, and R. Jain. The Mobile Web is Structurally Different. In INFOCOM Workshops 2008, IEEE, pages 1—6, 2008. http://doi.org/10.1109/INFOCOM.2008.4544648

[10] E. Kiciman and B. Livshits. AjaxScope: A Platform for Remotely Monitoring the Client-Side Behavior of Web 2.0 Applications. In The 21st ACM Symposium on Operating Systems Principles, SOSP '07, 2007. http://doi.org/10.1145/1294261.1294264

[11] K.Vikram, A. Prateek, and B. Livshits. Ripley: Automatically Securing Web 2.0 Applications Through Replicated Execution. In Proceedings of the Conference on Computer and Communications Security, November 2009.

[12] B. Livshits and S. Guarnieri. Gulfstream: Incremental Static Analysis for Streaming JavaScript Applications. Technical Report MSR-TR-2010-4, Microsoft, January 2010.

[13] T. Macchi. 2012 Mobile Traffic Report: How Much Did Mobile Traffic Increase in 2013?, January 2013.

[14] A. Mesbah, E. Bozdag, and A. van Deursen. Crawling Ajax by Inferring User Interface State Changes. In Web Engineering, 2008. ICWE '08. Eighth International Conference on, pages 122 —134, July 2008. http://doi.org/10.1109/ICWE.2008.24

[15] A. Mesbah and A. van Deursen. An Architectural Style for Ajax. Software Architecture, Working IEEE/IFIP Conference on, pages 1—9, 2007. http://doi.org/10.1109/WICSA.2007.7

[16] L. A. Meyerovich and B. Livshits. ConScript: Specifying and Enforcing Fine-Grained Security Policies for JavaScript in the Browser. IEEE Symposium on Security and Privacy, 0:481—496, 2010. http://doi.org/10.1109/SP.2010.36

[17] J. Mickens, J. Elson, and J. Howell. Mugshot: Deterministic Capture and Replay for JavaScript Applications. In Proceedings of the 7th USENIX conference on Networked systems design and implementation, NSDI'10, 2010.

[18] G. Mohr, M. Kimpton, M. Stack, and I. Ranitovic. Introduction to Heritrix, an archival quality web crawler. In Proceedings of the 4th International Web Archiving Workshop (IWAW04), September 2004.

[19] K. C. Negulescu. Web Archiving @ the Internet Archive. Presentation at the 2010 Digital Preservation Partners Meeting, 2010.

[20] M. L. Nelson. Memento-Datetime is not Last-Modified, 2011.

[21] @NesbittBrian. Play framework sample application with JWebUnit and synchronous Ajax.

[22] NetPreserver.org. IIPC Future of the Web Workshop — Introduction & Overview, 2012.

[23] M. E. Pierce, G. Fox, H. Yuan, and Y. Deng. Cyberinfrastructure and Web 2.0. In High Performance Computing and Grids in Action, pages 265—287, 2008.

[24] S. Raghavan and H. Garcia-Molina. Crawling the Hidden Web. Technical Report 2000-36, Stanford InfoLab, 2000.

[25] R. Schneider and F. McCown. First Steps in Archiving the Mobile Web: Automated Discovery of Mobile Websites. In JCDL '13: Proceedings of the 13th ACM/IEEE-CS Joint Conference on Digital Libraries. http://doi.org/10.1145/2467696.2467735

[26] Technical Committee ISO/TC 46. The WARC File Format (ISO 28500), 2008.

[27] J. Tennison. Hash uris, 2011.

[28] P. Timmins, S. McCormick, E. Agu, and C. Wills. Characteristics of Mobile Web Content. In 1st IEEE Workshop on Hot Topics in Web Systems and Technologies, 2006, HOTWEB '06, pages 1—10, 2006. http://doi.org/10.1109/HOTWEB.2006.355263

[29] B. Tofel. 'Wayback' for Accessing Web Archives. In Proceedings of the 7th International Web Archiving Workshop (IWAW07), 2007.

[30] H. Tweedy, F. McCown, and M. L. Nelson. A Memento Web Browser for iOS. In JCDL '13: Proceedings of the 13th ACM/IEEE-CS Joint Conference on Digital Libraries. http://doi.org/10.1145/2467696.2467764

[31] H. Van de Sompel, M. L. Nelson, and R. Sanderson. HTTP framework for time-based access to resource states — Memento draft-vandesompel-memento-06, 2013.

[32] H. Van de Sompel, M. L. Nelson, R. Sanderson, L. L. Balakireva, S. Ainsworth, and H. Shankar. Memento: Time Travel for the Web. Technical Report arXiv:0911.1112, 2009.

[33] H. Van de Sompel, R. Sanderson, M. L. Nelson, L. L. Balakireva, H. Shankar, and S. Ainsworth. An HTTP-Based Versioning

Mechanism for Linked Data. In Proceedings of the Linked Data on the Web Workshop (LDOW 2010), 2010. (Also available as arXiv:1003.3661).

[34] W3C staff and working group participants. Hash URIs, December 2011.

[35] D. F. Zucker. What Does Ajax Mean for You? Interactions, 14:10—12, September 2007. http://doi.org/10.1145/1288515.1288523

About the Authors

Mat Kelly is a PhD student of Computer Science at Old Dominion University. He is employed by NASA Langley Research Center through Science Systems and Application, Incorporated (SSAI) of Hampton, Virginia; Blade Agency of Gainesville, Florida and Old Dominion University Research Foundation of Norfolk, Virginia. His expertise lies in finding new ways to reinvent the wheel, introducing needless complication and overcoming data protection schemes. He is a Scorpio and is happily married with two dogs and resides in Portsmouth, Virginia.

Justin F. Brunelle is a Computer Science graduate student at Old Dominion University. His work involves the impact of multi-state, client-side representations on the archives and how it is preserved in the archives. Justin is also a Senior Application Developer at The MITRE Corporation where he performs research in the cloud computing and big data domains. More information on Justin can be found at http://www.justinfbrunelle.com/.

Michele C. Weigle is an Associate Professor of Computer Science at Old Dominion University. Her research interests include digital preservation, web science, information visualization, and mobile networking. She received her PhD in computer science from the University of North Carolina at Chapel Hill.

Michael L. Nelson is an associate professor of computer science at Old Dominion University. Prior to joining ODU, he worked at NASA Langley Research Center from 1991-2002. He is a co-editor of the OAI-PMH, OAI-ORE, Memento, and ResourceSync specifications. His research interests include repository-object interaction and alternative approaches to digital preservation. More information about Dr. Nelson can be found at: http://www.cs.odu.edu/~mln/.