2016

Leveraging Heritrix and the Wayback Machine on a Corporate Intranet: A Case Study on Improving Corporate Archives

Justin F. Brunelle
Old Dominion University

Krista Ferrante

Eliot Wilczek

Michele C. Weigle
Old Dominion University

Michael L. Nelson
Old Dominion University

Repository Citation

Brunelle, Justin F.; Ferrante, Krista; Wilczek, Eliot; Weigle, Michele C.; and Nelson, Michael L., "Leveraging Heritrix and the Wayback Machine on a Corporate Intranet: A Case Study on Improving Corporate Archives" (2016). *Computer Science Faculty Publications*. 10.
https://digitalcommons.odu.edu/computerscience_fac_pubs/10

Original Publication Citation

Brunelle, J. F., et al. (2016). "Leveraging Heritrix and the Wayback Machine on a corporate intranet: A case study on improving corporate archives." *D-Lib Magazine 22*(1/2): 1-11. 10.1045/january2016-brunelle

Search D-Lib:

D-Lib® Magazine

The Magazine of Digital Library Research

HOME | ABOUT D-LIB | CURRENT ISSUE | ARCHIVE | INDEXES | CALENDAR | AUTHOR GUIDELINES | SUBSCRIBE | CONTACT D-LIB

D-Lib Magazine

January/February 2016
Volume 22, Number 1/2
Table of Contents

Leveraging Heritrix and the Wayback Machine on a Corporate Intranet: A Case Study on Improving Corporate Archives

Justin F. Brunelle
The MITRE Corporation and Old Dominion University
jbrunelle@mitre.org

Krista Ferrante and Eliot Wilczek
The MITRE Corporation
{kferrante, ewilczek@mitre.org}

Michele C. Weigle and Michael L. Nelson
Old Dominion University
{mweigle, mln@cs.odu.edu}

DOI: 10.1045/january2016-brunelle

Printer-friendly Version

Abstract

In this work, we present a case study in which we investigate using open-source, web-scale web archiving tools (i.e., Heritrix and the Wayback Machine installed on the MITRE Intranet) to automatically archive a corporate Intranet. We use this case study to outline the challenges of Intranet web archiving, identify situations in which the open source tools are not well suited for the needs of the corporate archivists, and make recommendations for future corporate archivists wishing to use such tools. We performed a crawl of 143,268 URIs (125 GB and 25 hours) to demonstrate that the crawlers are easy to set up, efficiently crawl the Intranet, and improve archive management. However, challenges exist when the Intranet contains sensitive information, areas with potential archival value require user credentials, or archival targets make extensive use of internally developed and customized web services. We elaborate on and recommend approaches for overcoming these challenges.

1 Introduction

On the World Wide Web (WWW), web resources change and — unless archived — their prior versions are overwritten and lost. We refer to this as representations of resources existing in the perpetual *now*. The International Internet Preservation Consortium (IIPC) identifies several motivators for web archiving, including archiving web-native resources of cultural, political, and legal importance from sources such as art, political campaigns, and government documents [1].

To automatically archive such resources at web scale, web archives use crawlers to capture representations of web resources as they exist at a particular point in time. Historians, data scientists, robots, and general web users leverage the archives for historical trend analysis, revisiting now-missing pages, or reconstructing lost websites [2]. Corporate web archives can also hold a store of contextualized information about capabilities and development activities that shape how people think about the present and future [3].

Changing resources and users that require access to archived material are not unique to the public web. Resources within corporate Intranets change just as they do on the WWW. However, the Internet Archive [4] [5] and other public archives do not have the opportunity to archive Intranet-based resources [6]. As such, the responsibility for archiving corporate resources for institutional memory, legal compliance, and analysis falls on the corporate archivists.

In this work, we investigate the results, recommendations, and remaining challenges with using the Internet Archive's archival tools (Heritrix [7] [8] and the Wayback Machine) to archive the MITRE Information Infrastructure (MII). MITRE is a not-for-profit company that operates several Federally Funded Research and Development Centers (FFRDCs) with the US Federal government [9].

Throughout our discussion, we use Memento Framework terminology [10]. Memento is a framework that standardizes web archive access and terminology. Original (or live web) resources are identified by URI-Rs. Archived versions of URI-Rs are called mementos and are identified by URI-Ms.

2 Related Work

In our past research, we investigated the use of SiteStory, a transactional web archive, for helping to automatically archive the MII [11]. We showed that SiteStory was able to effectively archive all representations of resources observed by web users with minimal impact on server performance [12]. Other transactional web archives include ttApache [13] and pageVault [14]. However, a transactional web archive is not suitable for archiving the MII due to challenges with storing sensitive and personalized content and challenges with either installing the transactional archive on all relevant servers or routing traffic through an appropriate proxy.

Our past work has demonstrated that web pages' reliance on JavaScript to construct representations leads to a reduction in archivability [15] and, therefore, reduced memento quality [16]. Several resources within the MII are constructed via JavaScript to make them personalized, and are not archivable using Heritrix. Other web crawlers exist and have been evaluated on corporate Intranets [17] but are not readily available or as proven as Heritrix.

3 Background and Setup

The Internet Archive uses Heritrix and the Wayback Machine to archive web resources and replay mementos on the public web. These tools — as they exist in the public web — cannot reach into a corporate Intranet, but are available as open-source solutions. The Internet Archive's automatic, web-scale crawler — Heritrix — begins with a seed list of URI-R targets for archiving. This seed list becomes the initial frontier, or list of URI-Rs to crawl. Heritrix selects a URI-R from the frontier, dereferences[1] the URI-R, and stores the returned representation in a Web ARChive (WARC) file. The WARCs are indexed and ingested into an instance of the Wayback Machine which makes the mementos available for user access.

Our goal was to construct an architecture similar to the Internet Archive using an archival crawler and playback mechanism within our corporate Intranet. Because of their ease of use and effectiveness in public web environments, we opted to use Heritrix and the Wayback Machine to archive the MII and help improve corporate memory, expand the portion of the MII the corporate archives could capture, document more changes to the MII over time, and enable user access of the archived MII resources. We installed Heritrix and the Wayback machine on a server on the MII.

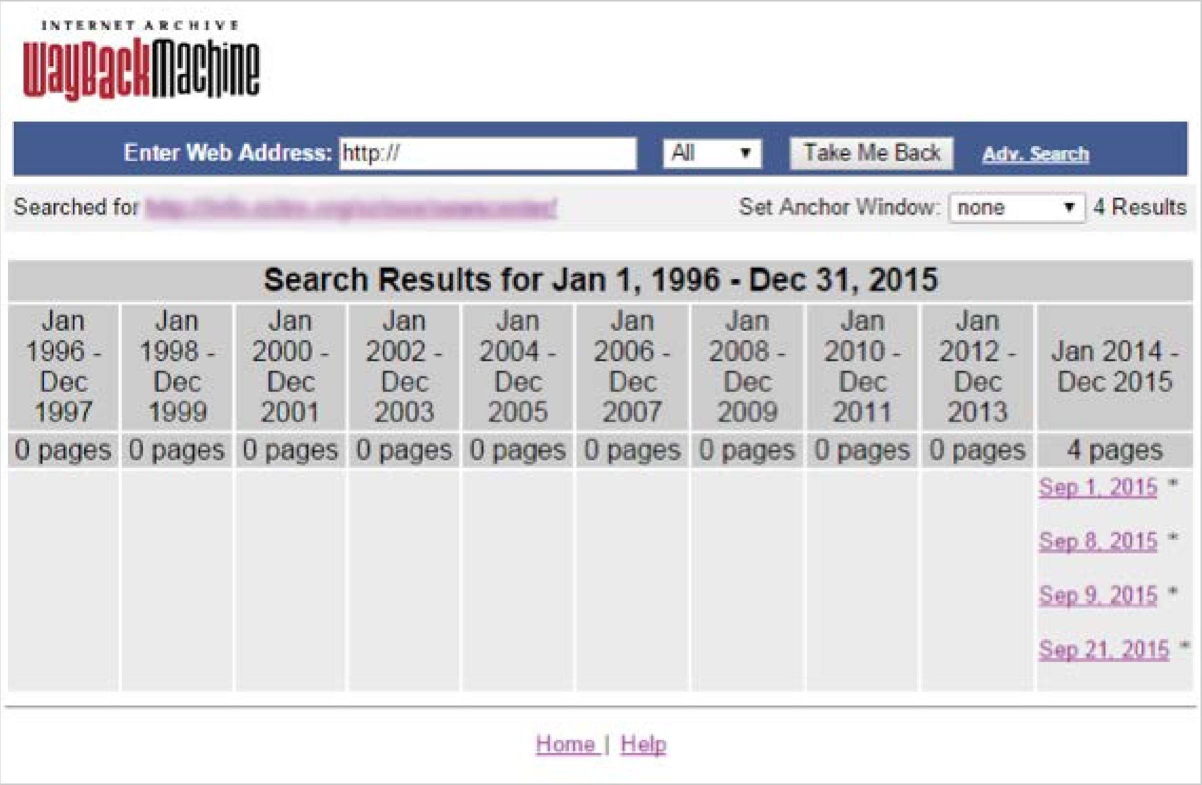
The installation and crawl setup of each tool took approximately 10 hours on a virtual machine hosted within the Intranet; this is a very minimal setup time for a production level crawling service. We undertook this work in a six-month exploratory project that we concluded in September 2015.

We configured the Heritrix crawler to only add URI-Rs within MITRE's Intranet to its frontier (i.e., those URI-Rs with a top-level domain (TLD) of `*.mitre.org`). We used a pre-selected set of 4,000 URI-Rs that are frequented by MITRE employees and are quickly accessible using keyword redirection (called "Fast Jumps") to MII resources.

Due to the nature of MITRE's work with the US federal government [9], the MII contains potentially sensitive resources that can only be hosted on servers or by services approved for such sensitive information. As such, these sensitive resources cannot be archived by an archival tool such as Heritrix and served by the Wayback Machine (the first of the archival challenges we discuss in this case study).

4 Crawl Results

We performed four crawls of our target resources at four times in September 2015 (Figure 1). From these mementos, we can observe changes to corporate information resources over time, and even recall information from past mementos of the resources (Figure 2).



Enter Web Address: http:// All

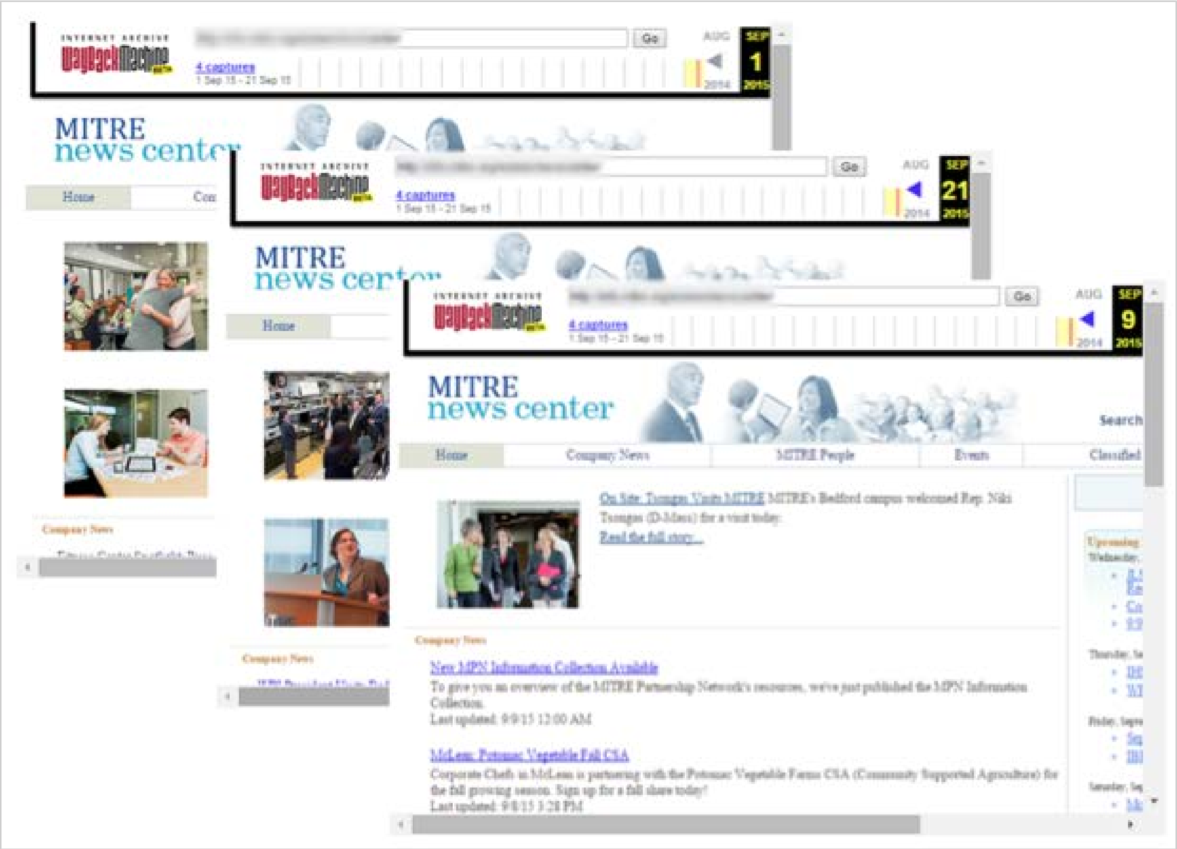
Searched for [redacted]

Set Anchor Window: none 4 Results

Search Results for Jan 1, 1996 - Dec 31, 2015

Jan 1996 - Dec 1997	Jan 1998 - Dec 1999	Jan 2000 - Dec 2001	Jan 2002 - Dec 2003	Jan 2004 - Dec 2005	Jan 2006 - Dec 2007	Jan 2008 - Dec 2009	Jan 2010 - Dec 2011	Jan 2012 - Dec 2013	Jan 2014 - Dec 2015
0 pages	0 pages	0 pages	0 pages	0 pages	0 pages	0 pages	0 pages	0 pages	4 pages
									Sep 1, 2015 * Sep 8, 2015 * Sep 9, 2015 * Sep 21, 2015 *

Figure 1: The MITRE Wayback Machine instance has four crawls from September 2015.



On Site: Tsongas Visits MITRE MITRE's Bedford campus welcomed Rep. Niki Tsongas (D-Mass) for a visit today.
Read the full story...

New MPN Information Collection Available
To give you an overview of the MITRE Partnership Network's resources, we've just published the MPN Information Collection.
Last updated: 9/9/15 12:00 AM

McLean: Potomac Vegetable Fall CSA
Corporate Chefs in McLean is partnering with the Potomac Vegetable Farms CSA (Community Supported Agriculture) for the fall growing season. Sign up for a fall share today!
Last updated: 9/8/15 3:28 PM

Figure 2: The mementos of the MITRE information resources allow users to navigate temporally within the Wayback Machine.

On our virtual machine (provisioned with 1 GB of memory, single core, and 125 GB of storage) the crawl that began from 4,000 URI-Rs took 25 hours to complete. At the time of completion, Heritrix had crawled 143,268 unique URI-Rs and occupies 34GB of storage. However, only 60% of the URI-R targets resulted in an HTTP 200 response code[2] (indicating the URI-R was successfully dereferenced and the representation archived). This is a lower success rate than expected for two reasons. First, the MII is closely curated, and second, the MII has a robust and high quality infrastructure. Both of these reasons would suggest that the MII would not have a high percentage of 400 and 500 class HTTP responses[3]. We omit the specific contributions to the low success rate due to security concerns, but outline two main reasons for the challenges in this section and discuss these challenges in further depth in Section 5 below.

First, much of the MII requires user credentials before the server will allow access to the resource. While Heritrix can be equipped with credentials, we omitted the credentials to avoid as much sensitive content as possible. Further, much of the personalized information that uses the credentials is built by JavaScript and, as a result, is not archivable [15].

Second, the MII includes several internally developed equivalents of WWW services, such as the MITRE versions of Wikipedia, YouTube, and GitHub. The Wikipedia and YouTube services had low archivability due to their reliance on JavaScript (and restricted access based on user credentials).

5 Challenges

We observed several challenges during our Intranet crawl. Some of these issues are well-known and pervasive across the archival community and the broader web community (e.g., reliance on JavaScript). However, others are unique to archiving corporate Intranets (e.g., user credentials and single sign on). In this section, we describe the challenges we observed during the crawl.

5.1 Accidental Crawl of Sensitive Information

MITRE is required to effectively and responsibly manage data — including sensitive data that is misclassified or misplaced within the Intranet [18] [19]. In the event sensitive information is misclassified or is not properly protected, clean-up is part of the corporate risk management plan and falls within MITRE's responsibilities. The clean-up procedure includes preventing future access to the sensitive information by MII users and, if an automatic archiving framework is actively crawling the MII, must also include clean-up of the archive.

In the event that a sensitive resource is crawled and archived by Heritrix, the data within the WARC must be properly wiped along with the index and database in the Wayback Machine[4]. The wiping process may result in the removal of other non-sensitive resources stored within the same WARC (which we refer to as *collateral damage*), or even destroying the device on which the WARC is stored.

The Internet Archive allows users to include a robots.txt file that prevents access to mementos as a mechanism for content owners to control access to mementos of their resources. The Internet Archive also maintains a blacklist of mementos that should not be available on their public web interface. While this is effective for a public archive that does not deal with sensitive content, it is not suitable for the MII. Sensitive information that is mistakenly crawled by Heritrix must be deleted in its entirety to ensure the proper control of the information. As such, simply blocking access to a memento from the web interface is not sufficient, and the memento must be completely destroyed.

5.2 User Credentials

Because MII users have credentials that are needed to access the MII (e.g., via single sign on), many servers expect to receive credentials before returning a representation. As such, the Heritrix crawler was not able to access some resources. Some of the URI-Rs redirected to login screens that Heritrix archived, but having user credentials would likely offer an opportunity to archive much more of the MII content; the login screens may be portals to entire subsections of the MII that are important to corporate memory.

During our proof-of-concept crawls, we opted to not provide Heritrix with user credentials. Because this was an exploratory investigation, we deemed the risk of accidentally crawling sensitive information and potentially losing all of our mementos as collateral damage of the cleanup process too great given the scope of our investigation.

5.3 Internally Developed Services & JavaScript

MITRE has developed its own equivalents for WWW services such as MITRE's YouTube, Wikipedia, GitHub, Delicious, and Facebook. Each of these services (with the exception of MITRE's internal GitHub equivalent) makes use of JavaScript to construct the representations. Because Heritrix does not execute JavaScript on the client, these services remained unarchived. Further, because these resources are developed internally and customized for MITRE, other archival tools that are specifically designed to archive their WWW counterparts (e.g., Pandora's YouTube archival process [20] and ArchiveFacebook [21]) may not be able to archive the MII-specific resources. Other services construct content for the user based on preferences using JavaScript, such as widget dashboards. These resources are entirely unarchivable without credentials and the ability to run client-side JavaScript. Alternatively, the GitHub equivalent within the MII was archived successfully 99% of the time because the URI-Rs added to the frontier by Heritrix do not require user credentials for access, and do not rely on JavaScript to load embedded resources.

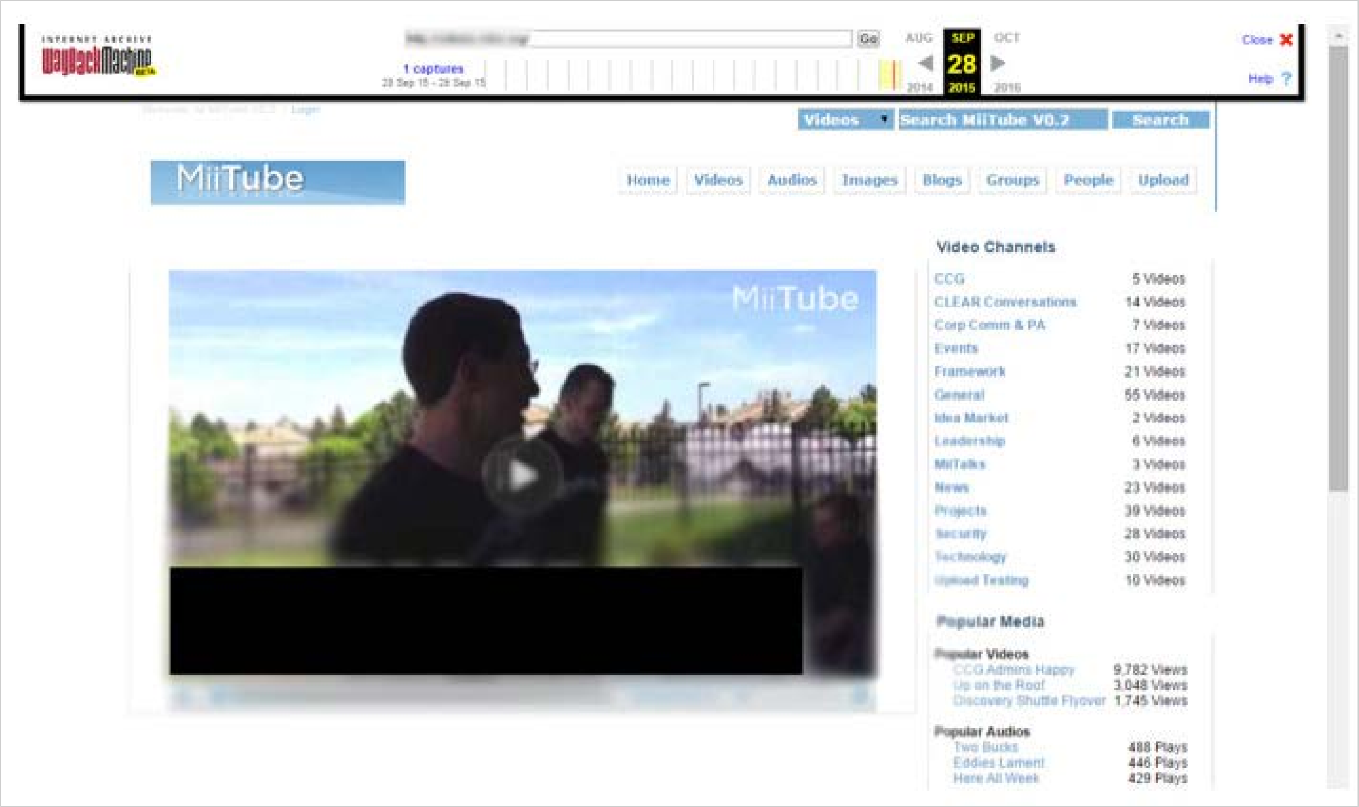
For example, we present MITRE's MIITube, a YouTube equivalent (Figure 3). MIITube uses JavaScript to load embedded images, which leads to leakage in the memento. The thumbnails of videos are all loaded by JavaScript in this memento, as shown in the HTTP GET and response, below.

```
HTTP GET [MII YOUTUBE]
Referer: http://waybackmachine.[MII Host]
/wayback/20150928131729/[MII YOUTUBE]
User-Agent: Mozilla/5.0 (Windows NT 6.1; WOW64)
AppleWebKit/537.36 (KHTML, like Gecko) Chrome/45.0.2454.99
Safari/537.36
Host: [MII YOUTUBE]
X-Requested-With:ShockwaveFlash/19.0.0.185

HTTP/1.1 200 OK
Date: Mon, 28 Sep 2015 13:41:21 GMT
Server: Apache/2.2.3 (Red Hat)
Last-Modified: Mon, 21 May 2012 23:22:39 GMT
ETag: "115801b-8c08-4c0942d90cdc0"
```

```
Accept-Ranges: bytes
Content-Length: 35848
Connection: close
Content-Type: image/jpeg
```

(Note that the observed request is to an image at `[MII YOUTUBE]`[5] rather than the MITRE-hosted Wayback Machine.)



[redacted]

Video Channels

CCG	5 Videos
CLEAR Conversations	14 Videos
Corp Comm & PA	7 Videos
Events	17 Videos
Framework	21 Videos
General	55 Videos
Idea Market	2 Videos
Leadership	6 Videos
MiiTalks	3 Videos
News	23 Videos
Projects	39 Videos
Security	28 Videos
Technology	30 Videos
Upload Testing	10 Videos

Popular Media

Popular Videos

CCG Admins Happy	9,782 Views
Up on the Roof	3,048 Views
Discovery Shuttle Flyover	1,745 Views

Popular Audios

Two Bucks	488 Plays
Eddies Lament	446 Plays
Here All Week	429 Plays

Figure 3: MIITube uses JavaScript to load embedded resources which leads to leakage.

6 Recommendations

From our experiences performing crawls of the MII, we make several recommendations that can be applied to the MII crawl effort as well as to other corporate and institutional Intranets, and identify strategies for overcoming challenges faced by many institutions, not just MITRE. We summarize these challenges and strategies in Section 6.3, Table 1.

6.1 Accidental Crawl of Sensitive Information

Because accidentally archiving sensitive information can result in collateral damage and loss of mementos within a WARC or storage device, we recommend the following:

- Use smaller storage devices to limit the collateral damage in the event that sensitive information is crawled;
- Develop a method to remove a single memento (e.g., a sensitive memento) from a WARC file to prevent collateral damage; and
- Identify high-risk vs. low-risk archival targets within the Intranet.

We also recommend content authors use robots.txt [22] and `noarchive` HTTP response headers [23] to help Heritrix avoid sensitive information. Examples of suitable `noarchive` HTTP response headers include `X-Robots-Tag: noarchive` and `X-No-Archive: Yes`. While crawlers in the WWW are not *required* to obey the `noarchive` headers, within a corporate Intranet we can assume the crawlers will be well-behaved and obey the `noarchive` headers and robots.txt files. Because sensitive material is required to be marked as such, it should follow that web-hosted sensitive content can be marked in the headers. We provide an example of the `noarchive` headers below from a test page located at an Old Dominion University server:

```
 $ curl -iv http://www.cs.odu.edu/~jbrunelle/secret.php
 GET /~jbrunelle/secret.php HTTP/1.1
 User-Agent: curl/7.35.0
 Host: www.cs.odu.edu
 Accept: */*

 HTTP/1.1 200 OK
 Server: nginx
 Date: Fri, 25 Sep 2015 11:58:22 GMT
 Content-Type: text/html
 Transfer-Encoding: chunked
 Connection: keep-alive
X-Robots-Tag: noarchive
 X-No-Archive: Yes
 Vary: Accept-Encoding

<html>
 ...
 </html>
```

6.2 User Credentials

To widen Heritrix's potential crawl frontier, Heritrix should be provided user credentials to access non-sensitive areas of the corporate Intranet that require user credentials. However, this may increase the opportunity to crawl sensitive content. Further, this will not mitigate all aspects of the challenges with personalized, JavaScript-built representations. For example, a set of user credentials that has no preferences or dashboard widgets will likely not improve the archival coverage of such personalized representations; such preferences are unreasonable to expect Heritrix to possess.

6.3 Internally Developed Services & JavaScript

Internally developed and customized WWW service equivalents will continue to cause archival challenges. We recommend using either open source equivalents for these services, leveraging hosted services, or maintaining internally developed services that have better archivability (e.g., not relying on JavaScript to build the representation) than their live web counterparts. However, this will not always result in high quality archives, particularly in the case of open-source resources that are built using JavaScript. To mitigate the impact of JavaScript on the archives, we recommend using a two-tiered crawling approach (as we present in our prior work [24]) using PhantomJS [25] or another headless browsing client to execute client-side JavaScript.

The current, single-tiered crawling approach is used by Heritrix in which the crawler issues an HTTP GET request for the URI-R of the crawl target and archives the response. From here, a two-tiered approach builds on the first tier by categorizing the returned representation as likely to be deferred or non-deferred, and using PhantomJS to load deferred representations and execute the client-side JavaScript to capture a more complete set of embedded resources. The result is a slower but more complete crawl of deferred representations.

Challenge	Elaboration	Potential Solution/Notes
Accidental archiving of sensitive information	Crawling sensitive information can potentially eliminate mementos from an entire WARC or even storage device due to the need to tightly control access to sensitive content.	Heritrix should not only avoid areas in which content may be sensitive, but also use smaller, individual storage devices to limit collateral damage. Additionally, the community would benefit from a WARC removal utility. Content authors can help prevent unauthorized crawling with robots.txt files and `X-Robots-Tag: noarchive` HTTP response

		headers.
User credentials	Services that require user credentials prevent Heritrix from accessing entire sub-sections of the MII.	Equipping Heritrix with credentials would remedy the challenge of access, but further investigation will identify whether this helps improve coverage of the MII.
User-specific widgets	The user-specific widgets within the MII are constructed by JavaScript (i.e., via widgets) and are personalized for the user. As such, the representation will not have rich information and embedded links that Heritrix can extract, resulting in a small frontier.	Incorporating PhantomJS or another JavaScript-enabled browser would enable the JavaScript-dependent representations. Further, Heritrix should be equipped with user credentials to properly access the resource.
Internally developed services	Internally developed services often construct the representation with JavaScript, and are also not archived with specialized archival tools developed for the WWW.	Equipping Heritrix with PhantomJS will remedy this issue, but a corporate Intranet should either invest in replicating the WWW archival tools for these services or maintain duplicate services of the WWW tools internally.

Table 1: Challenges identified during the MII crawl and recommendations for mitigating these challenges.

7 Conclusions

We performed an initial assessment of the suitability of the Internet Archive's open-source tools for archiving the MII (MITRE's corporate Intranet) finding them highly effective. We identified challenges with sensitive information, user credentials, and internally developed and JavaScript-dependent representations. We recommend mitigations to these challenges, and hope that our study of the MII helps initiate automatic corporate archiving projects in other Intranet environments. These automated approaches have the potential to save archival costs, improve corporate memory, and increase users' ability to leverage corporate archives [3].

With the completion of our exploratory project we will be looking to establish a production level service for archiving the MII. This will include working with MITRE's security office to set up crawl policies that identify high and low risk archival targets and then focusing on low risk targets in order to limit the risk for collateral damage from crawling sensitive information. We also plan to investigate single-memento WARC removal tools to further reduce the impact of crawling sensitive information. We will also examine the extent to which we can capture user-authenticated areas of the MII with user credential-enabled crawling.

More broadly, we will need to place the archiving of the MII within a larger documentation plan [26]. Capturing the Intranet needs to be undertaken within a framework of understanding what are the key resources that need to be preserved in order to sustain MITRE's corporate memory. Additionally, we need to understand the essential elements of the resources we are trying to archive. Cases where the presentation of an Intranet resource is an important component of its documentary value demonstrate a corporation's need for a web crawling archiving strategy. In situations where the Intranet presentation of a resource is not critical to its documentary value, it may make more sense to capture this resource in another manner. For example, it may make more sense for a corporate archives to preserve information about its corporation's projects that is tracked in a database and served to an Intranet through an export directly from the database rather than crawling the Intranet for the project data.

The case study we have presented and the next steps we proposed will help archive the MII for corporate memory, improved employee services, and improved information longevity. It also serves as a case study and brief explanation of archiving a corporate Intranet that can help prepare corporate archivists for implementing scalable web archiving strategies.

Notes

[1] The process of "visiting" a web resource involves dereferencing its URI-R, beginning with an HTTP GET request for the URI-R and receiving the representation.

[2] HTTP 200 response codes indicate that a URI-R was found and a representations was returned to the user-agent by the server.

[3] HTTP 400 class responses indicate that a URI-R is either missing or is unauthorized for viewing by the requesting user-agent. HTTP 500 classes indicate an error has occurred on the server.

[4] We worked closely with the MITRE security office to understand how sensitive resources might appear in a crawl target list, how the data must be cleaned in the event sensitive data is crawled, and the role of the security office during the archival process. For the purposes of this document and because of the sensitive nature of the details of these discussions, we omit the details of this process.

[5] We have used `[MII YOUTUBE]` instead of the full server URI for public release purposes.

Bibliography

[1] International Internet Preservation Consortium (IIPC), "Web Archiving," 2015.

[2] Y. AlNoamany, A. AlSum, M. C. Weigle and M. L. Nelson, "Who and what links to the Internet Archive," *International Journal on Digital Libraries*, vol. 14, no. 3, pp. 101-115, 2014.

[3] J. T. Seaman and G. D. Smith, "Your Company's HIstory as a Leadership Tool," *Harvard Business Review*, vol. December, no. R1212B, 2012.

[4] K. C. Negulescu, "Web Archiving @ the Internet Archive," *Presentation at the 2010 Digital Preservation*, 2010.

[5] B. Tofel, "'Wayback' for Accessing Web Archives," *Proceedings of the 7th International Web Archiving Workshop*, pp. 27-37, 2007.

[6] K. Hagedorn and J. Sentelli, "Google Still Not Indexing Hidden Web URLs," *DLib Magazine*, vol. 14, no. 7/8, 14(7), July/August 2008. http://doi.org/10.1045/july2008-hagedorn

[7] K. Sigurðsson, "Incremental crawling with Heritrix," *Proceedings of the 5th International Web Archiving Workshop*, September 2005.

[8] G. Mohr, "Introduction to Heritrix, an archival quality web crawler," *Proceedings of the 4th International Web Archiving Workshop*, 2004.

[9] The MITRE Corporation, "FFRDCs — A Primer," 2015.

[10] H. Van de Sompel, M. L. Nelson, R. Sanderson, "HTTP Framework for Time-Based Access to Resource States — Memento, Internet RFC 7089," December 2013.

[11] J. F. Brunelle, J. T. Morrison and G. Despres, "Installation and Experimentation of a Transactional Archive on a Corporate Intranet," The MITRE Corporation, MTR114406, 2011.

[12] J. F. Brunelle, M. L. Nelson, L. Balakireva, R. Sanderson and H. Van de Sompel, "Evaluating the SiteStory Transactional Web Archive With the ApacheBench Tool," in *Proceedings of the Third International Conference on Theory and Practice of Digital Libraries*, 2013.

[13] C. E. Dyreson, H. Lin and Y. Wang, "Managing versions of Web documents in a transaction-time Web server," in *Proceedings of the 13th International Conference on World Wide Web*, 2004.

[14] K. Fitch, "Web site archiving: an approach to recording every materially different response produced by a Website," in *9th Australasian World Wide Web Conference*, 2003.

[15] J. F. Brunelle, M. Kelly, M. C. Weigle and M. L. Nelson, "The impact of JavaScript on archivability," *International Journal on Digital Libraries*, pp. 283-301, 2015.

[16] J. F. Brunelle, M. Kelly, H. SalahEldeen, M. C. Weigle and M. L. Nelson, "Not all mementos are created equal: Measuring the impact of missing resources," *International Journal of Digital Libraries*, pp. 283-301, 2015.

[17] A. Heydon and M. Najork, "Mercator: A scalable, extensible Web crawler," *World Wide Web*, vol. 2, no. 4, pp. 219-229, 1999.

[18] Department of Defense, "National Industrial Security Program Operating Manual," DoD 5220.22-M, 2006.

[19] National Archives and Records Administration, "Executive Order 13526 of December 29, 2009," *Federal Register*, 2009.

[20] E. Crook, "Web archiving in a Web 2.0 world," in *Proceedings of the Australian Library and Information Association Biennial Conference*, 2008.

[21] M. Kelly, C. Northern, H. SalahEldeen, M. L. Nelson and F. McCown, "ArchiveFacebook," 2015.

[22] Robots.txt, "The Web Robots Page," 2015.

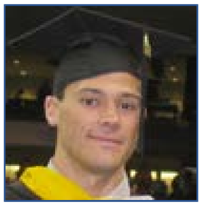
[23] The NoArchive Initiative, "The NoArchive Initiative," 2015.

[24] J. F. Brunelle, M. C. Weigle and M. L. Nelson, "Archiving Deferred Representations Using a Two-Tiered Crawling Approach," in *Proceedings of iPRES 2015*, 2015.

[25] PhantomJS, "PhantomJS," 2015.

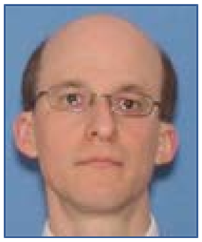
[26] H. W. Samuels, "Varisty Letters: Documenting modern colleges and universities," Scarecrow Press, Metuchen, NJ, 1992.

About the Authors



Justin F. Brunelle is a Computer Science doctoral candidate at Old Dominion University in the Web Science and Digital Libraries research group. His work involves the impact of JavaScript on the archives and approaches to better archive deferred representations. Justin is also a Lead Software Application Developer at The MITRE Corporation where he performs research on emerging technologies. More information on Justin can be found at http://www.justinfbrunelle.com/.

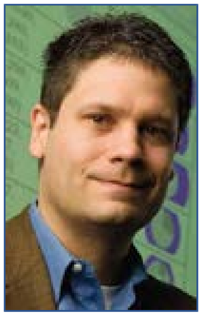
Krista Ferrante is the Corporate Archivist at The MITRE Corporation. She has previously worked as an archivists at MIT, Harvard University, Tufts University and the American Antiquarian Society. She received her MS in Library and Information Science at Simmons College in Boston.



Eliot Wilczek is the Corporate Records and Archives Manager at The MITRE Corporation. He has previously worked as a records manager and archivist at Tufts University, Brandeis University, and Bowdoin College. Eliot is also a doctoral candidate at the School of Library and Information Science, Simmons College.

Michele 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 a 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/.