White Paper Report

Report ID: 109070

Application Number: HD-51670-13

Project Director: Michele Weigle (mweigle@cs.odu.edu)

Institution: Old Dominion University Research Foundation

Reporting Period: 5/1/2013-12/31/2014

Report Due: 3/31/2015

Date Submitted: 3/27/2015

Final Performance Report

NEH HD-51670-13

Archive What I See Now

Michele C. Weigle

Old Dominion University Research Foundation

March 2015

"Archive What I See Now": Bringing Institutional Web Archiving Tools to the Individual Researcher

Michele C. Weigle, Michael L. Nelson
Department of Computer Science
Old Dominion University
Norfolk, VA 23529 USA
{mweigle, mln}@cs.odu.edu

Introduction

As the web has become the repository for much of our social culture, humanities scholars and social scientists have recognized the need for archiving web objects to support their research. The Internet Archive and others are trying to capture a broad range of web sites, but they cannot capture everything. So, individual researchers often develop their own narrow collections of web pages that are critical for their research. Unfortunately, current web archiving tools have been developed at the institutional level and not the personal level, making archiving by individuals (especially non-IT experts) difficult and sometimes even cost-prohibitive. This project has focuses on building open-source tools for personal-scale web archiving to support humanities scholars.

Web archiving typically consists of a web crawler (to gather large sets of web pages and their embedded resources), a standard archive format, and an access mechanism (to view the archived contents). In a typical institutional archiving setup, the web crawler is Heritrix, the archive format is the Web ARChive (WARC), and the indexing and access tool is the Wayback Machine (or, the open-source OpenWayback[1], both generally referred to as Wayback in the remainder of this document). Recently, enhanced access to archived web pages (including those accessible by Wayback and other archiving methods) has been provided through the Memento protocol and various browser extensions (e.g., Firefox add-on MementoFox and Google Chrome extension Memento Time Travel), which allow users to navigate the past web in their browser. Heritrix performs well at crawling large sets of web sites, but it is complex and often requires expensive technical personnel for setup and maintenance.

Our main goal was to enable users to generate files suitable for use by large-scale archives (i.e., WARC files) with tools as simple as the "bookmarking" or "save page as" approaches that they already know. Although simply generated, these files would then be in a format compatible with professional-grade archives. Our main contribution is in allowing individuals to "archive what I see now". The user can create a standard web archive file ("archive") of the content displayed in the browser ("what I see") at a particular time ("now"). During the course of this start-up grant, we have made significant progress in developing our tools to provide this service.

Project Activities

Our previous work produced WARCreate, a Google Chrome extension, that can create an archive of a single web page in the standard WARC format and save it to local disk. One advantage of WARCreate over Heritrix, besides its simplicity, is the ability to capture any web page from the perspective of the web browser. This means that web pages requiring

[1] http://netpreserve.org/openwayback

authentication, pages from social media sites, pages displayed after some user interaction, and page elements such as ads generated based on the context of the particular web browser can all be archived in a standard format. This tool allows users to "archive what I see now".

In addition to WARCreate, we had also developed WAIL (Web Archive Integration Layer) that packages Heritrix and Wayback into a user-friendly format. WAIL allows non-technical users to install and run these powerful tools with a one-click installation. In combination with WARCreate, WAIL allows users to view user-generated WARCs (created with WARCreate) on their local machine in Wayback.

The main tasks that we proposed in our start-up grant (performance period: May 2013 – December 2014) were as follows:

- Task 1 – Develop a Firefox version of WARCreate
- Task 2 – Add the ability to upload user-generated WARCs to a server
- Task 3 – Implement sequential archiving to allow users to save more than one web page at a time.

Below, we detail our project activities related to each of the three main tasks.

Task 1 - Develop a Firefox version of WARCreate

We had already developed a Google Chrome version of WARCreate (Figure 1), but had received several requests for a Firefox version. Our goal was to increase the use of this tool by providing the additional extension.

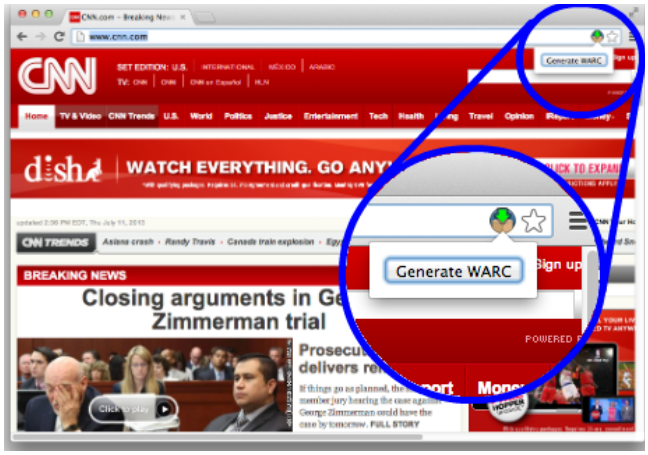

Figure 1: WARCreate Google Chrome extension

We completed a preliminary version of WARCreate for Firefox in November 2014. WARCreate requires capturing HTTP headers for all responses that come back to the browser. As the mechanism for doing this is different in Firefox than in Chrome, a significant amount of time was spent in accessing these headers and ensuring that they were written in the proper format to generate valid WARC files.

Figure 2 shows a screenshot of the WARCreate Firefox add-on icon in the top right corner of the browser. When the user clicks on the icon, the standard download dialog box appears to ask the user if they want to save the file. A development version of the WARCreate Firefox add-on is available at https://github.com/machawk1/ffwarcreate.

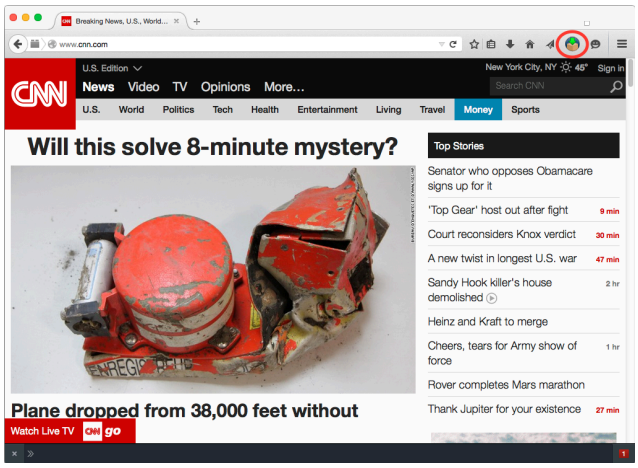

Figure 2: WARCreate Firefox add-on

While developing the Firefox version of WARCreate, we continued to respond to user feedback on the Google Chrome version. As the use of Google Chrome has increased (reported market share of close to 50% as of Feb 2015[2]), we have seen more interest in the Google Chrome version of WARCreate than in a Firefox version. Currently, the Google Chrome extension has 432 users and a rating of 5 stars in the Google Chrome Store. We are continuing to receive feedback and feature requests from users of the Chrome WARCreate.

Task 2 – Upload WARCs to a server

We implemented a mechanism in the Chrome version of WARCreate for a user to upload a user-generated WARC to a specified server. The specification of the server is available through WARCreate extension options page. The code required for the server is in the Appendix.

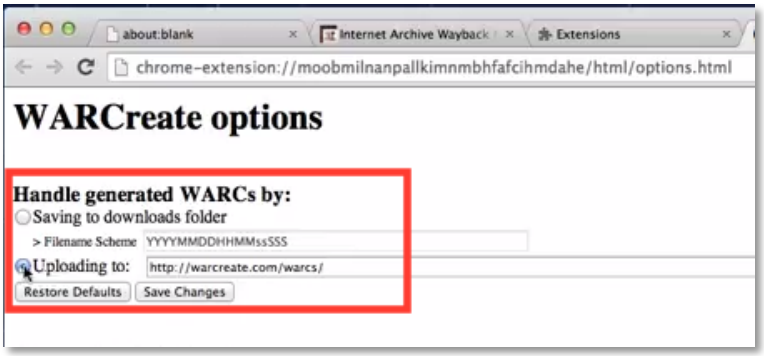

Figure 3: WARCreate options page for specifying upload server

After this option is set, WARCs generated by WARCreate are uploaded to the specified server and can then be accessed by an instance of Wayback running on the server. To demonstrate this, we use the Wayback instance included in our tool WAIL. A screencast of this operation is available at https://www.youtube.com/watch?v=b0DkhxDCdvg

Another of our goals was to allow users to upload generated WARCs to their existing Archive-It collections. Archive-It adds special headers to the WARC files that they create, so we have been

[2] http://gs.statcounter.com/ - browser-ww-monthly-201305-201502

working with Archive-It to ensure that the WARCs generated by WARCreate and destined for Archive-It either include the headers or that WARCreate has a mechanism to add those headers later. Figure 4 shows the Chrome extension options page that allows users to specify their Archive-It collection ID and collection name so that this information can be added to the WARC upon creation.

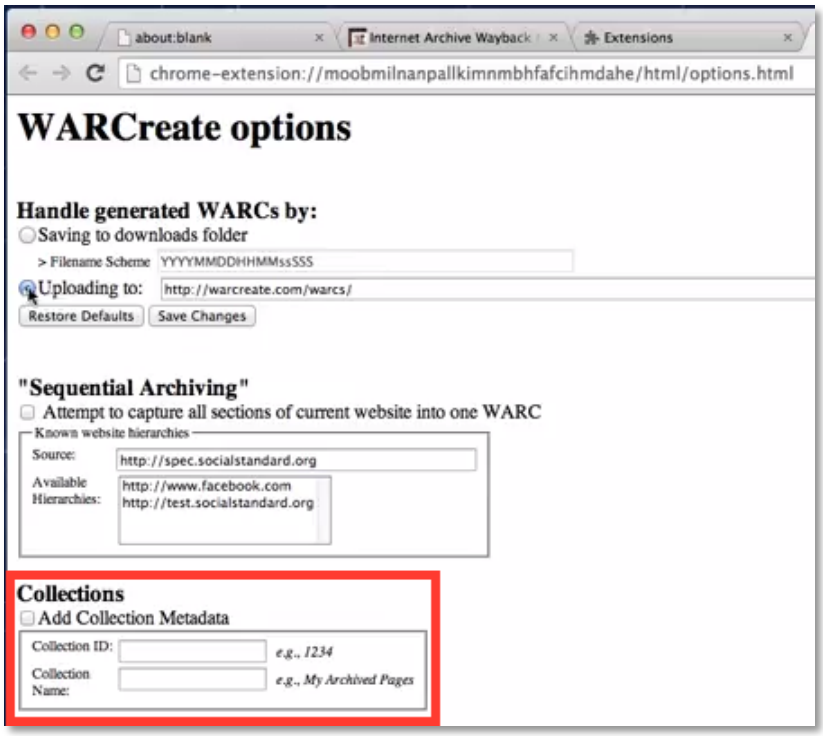

Figure 4: WARCreate options page for adding Archive-It collection metadata to WARC

There are several unresolved issues related to uploading user-generated WARCs to Archive-It. First, there should be some mechanism to ensure that the WARCs have not been tampered with before upload. Also, URI clashing needs to be resolved. For instance, the page that one user sees when logged into http://www.facebook.com/ is not the same page that anyone else would see at http://www.facebook.com/. Users also need the ability to review the contents of WARCs before upload to Archive-It, so that they can ensure that no private information is uploaded to a public collection. This may conflict with the need to ensure the authenticity of the generated WARCs. Many of these issues are policy-based rather than technical, so we leave them for future work. Mat Kelly, the PhD student who developed these tools, will be investigating some these issues as part of his dissertation work.

Task 3 – Sequential Archiving

The initial version of WARCreate could only archive a single page at a time. Sequential archiving is the process of archiving multiple web pages in sequence into a single WARC.

There are two main aspects to sequential archiving. First, we want to allow a user to archive multiple pages that are behind authentication. For instance, a user may want to generate a single WARC containing their Facebook presence. This would include the "wall" content at http://www.facebook.com/, the user's profile, list of friends, photos, messages, etc. This will

require archiving several pages sequentially. We have developed a method for specifying the components of a website, such as Facebook, and WARCreate can read the specification and archive the required pages. Figure 5 shows the options page for providing the location of such a specification. (For more information on the specification, see Mat Kelly's Masters' Thesis[3].)

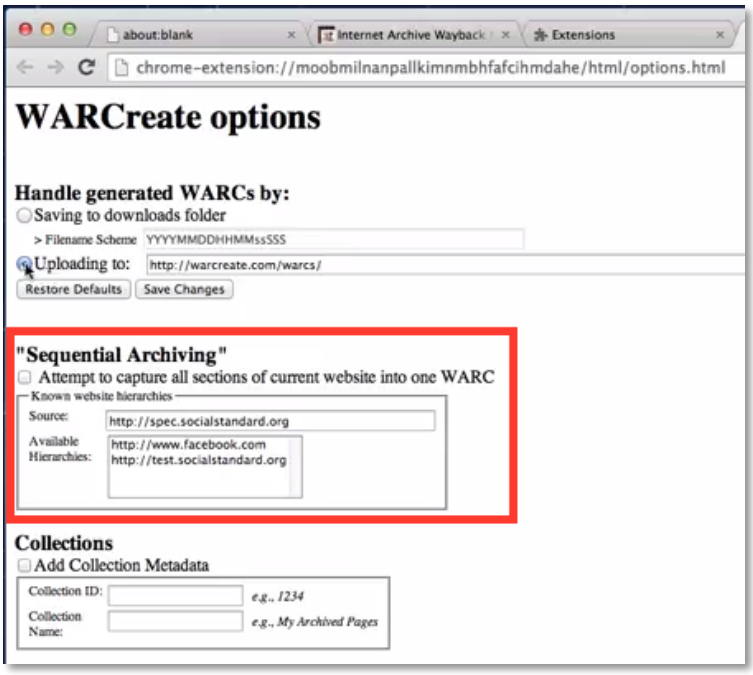

Figure 5: WARCreate options page for sequential archiving

Second, we want to allow users to archive the entire content of a particular website. For example, one of advisory board members has the use case of local churches individually archiving their local websites. Instead of navigating to each page and creating a separate WARC for each, the curators should be able to specify the domain of the site and have it be archived automatically. This is exactly what the Internet Archive's Heritrix web crawler can already do. Since Heritrix is packaged along with Wayback in WAIL, we decided against re-developing this functionality in WARCreate. We will provide easy options for non-technical users to perform this type of archiving through WAIL.

Publicizing Our Results

We have made several posts related to this project on our research group's blog (http://ws-dl.blogspot.com) to publicize this project. The 9 posts that mention WARCreate (both before and during this grant's performance period) have collectively accumulated over 10,000 views.

We have also presented this work in several venues:

- "WARCreate and WAIL: WARC, Wayback, and Heritrix Made Easy", Digital Preservation 2013, July 2013
- "Archive What I See Now", 2013 Archive-It Partners Meeting, November 2013
- "Tools for Managing the Past Web", 2014 Archive-It Partners Meeting, November 2014

[3] http://ws-dl.blogspot.com/2012/08/2012-08-20-ms-thesis-extensible.html

Continuation of the Project

This project will continue at least until 2017 with funding from an NEH Digital Humanities Implementation Grant (HK-50181-14). We are grateful for the NEH's continued support of this project.

This project afforded us the opportunity to connect with the larger digital humanities community and make contact with political scientists, art librarians, and communications researchers. Several of these individuals are serving as Advisory Board members on our current DHIG. This includes Dr. Avi Santo, Director of the Institute of Humanities at ODU.

Grant Products

Our grant products consist of software, blog posts, and presentations.

Software

- WARCreate - http://warcreate.com/
- WAIL - http://matkelly.com/wail/

This, along with much of our research group's other software, is also available at https://ws-dl.cs.odu.edu/Main/Software

Blog Posts

- "2013-07-10: WARCreate and WAIL: WARC, Wayback and Heritrix Made Easy", http://ws-dl.blogspot.com/2013/07/2013-07-10-warcreate-and-wail-warc.html,
  - 2114 views as of March 2015
- "2013-10-11: Archive What I See Now", http://ws-dl.blogspot.com/2013/10/2013-10-11-archive-what-i-see-now.html
  - 513 views as of March 2015

Presentations

- "WARCreate and WAIL: WARC, Wayback, and Heritrix Made Easy", Digital Preservation 2013, July 2013, http://www.slideshare.net/matkelly01/digital-preservation-2013
- "Archive What I See Now": Bringing Institutional Web Archiving Tools to the Individual Researcher, Slides from shutdown-cancelled NEH ODH Project Directors' Meeting (originally scheduled for Oct 4, 2013), http://www.slideshare.net/mweigle/archive-what-i-see-now
- "Archive What I See Now", 2013 Archive-It Partners Meeting, November 13, 2013, http://www.slideshare.net/matkelly01/archive-what-i-see-now-archiveit-partner-meeting-2013-2013
- "Tools for Managing the Past Web", 2014 Archive-It Partners Meeting, November 18, 2014, http://www.slideshare.net/mweigle/2014-weigleaitpublic

Appendix

The following PHP script is required to for a server to accept WARC uploads from the WARCreate extension. It should be named `index.php` and placed in the location specified in the WARCreate options (same location as where the WARCs will be uploaded to).

```
<?php
header('Access-Control-Allow-Origin: *');
$post_data = file_get_contents('php://input'); //read raw POST stream
if (!empty($post_data)) {
    $filename = time().".warc";
    $file = fopen(dirname(__FILE__)."/".$filename, 'w+');
    fwrite($file, $post_data);
    fclose($file);
    http_response_code(201);
    echo $filename;
} else {
      file_put_contents(dirname(__FILE__)."/".time().".warc","post data was
empty".$post_data."\r\n".implode("|",$_POST)." << End contents");
}
?>
```