WARCreate - Create Wayback-Consumable WARC Files from Any Webpage

Mat Kelly
Department of Computer Science
Old Dominion University
Norfolk, Virginia
mkelly@cs.odu.edu

Michele C. Weigle
Department of Computer Science
Old Dominion University
Norfolk, Virginia
mweigle@cs.odu.edu

ABSTRACT

The Internet Archive's Wayback Machine is the most common way that typical users interact with web archives. The Internet Archive uses the Heritrix web crawler to transform pages on the publicly available web into Web ARChive (WARC) files, which can then be accessed using the Wayback Machine. Because Heritrix can only access the publicly available web, many personal pages (*e.g.*, password-protected pages, social media pages) cannot be easily archived into the standard WARC format. We have created a Google Chrome extension, WARCreate, that allows a user to create a WARC file from any webpage. Using this tool, content that might have been otherwise lost in time can be archived in a standard format by any user. This tool provides a way for casual users to easily create archives of personal online content. This is one of the first steps in resolving issues of "long term storage, maintenance, and access of personal digital assets that have emotional, intellectual, and historical value to individuals" [3].

Categories and Subject Descriptors

H.3.4 [**Information Storage and Retrieval**]: Systems and Software; H.3.7 [**Digital Libraries**]: Personal Web Archiving

General Terms

Design

Keywords

Personal Web Archiving, WARC, Browser, Wayback Machine, Internet Archive

1. INTRODUCTION

The Internet Archive, along with web archives at other libraries and institutions, has done a remarkable job at archiving the public web. But in recent years, the web has become a home for a significant amount of original user-generated content, such as that posted on social media sites. Users are becoming increasingly aware of the need for personal web archiving [4, 5]. Unfortunately, this content is largely unavailable to standard web archives because it lives behind the "walled garden" of authentication and is part of the "deep web" [1]. Our goal is to allow users, once past authentication, to generate their own archives that can be browse-able in a user-friendly manner.

The Internet Archive's Wayback Machine is the most well-known interface for accessing web archives. The archived pages are stored in the standard Web ARChive (WARC) format [2] and are generated by the Heritrix[1] crawler. Unfortunately, Heritrix is limited to crawling only publicly accessible pages, so many personal pages (*e.g.*, password-protected pages, social media pages) cannot be easily archived. In addition, for pages that are location or user-agent aware, the version archived at Internet Archive is the one that the Heritrix crawler (run from San Francisco) sees. For example, the most recently available version[2] of http://www.craigslist.org redirects to http://sfbay.craigslist.org.

In an effort to facilitate the use of the standard WARC format for personal web archives, we have developed a tool to allow a user to archive any page, edit its metadata, and submit it to an instance of the Wayback Machine (from here on referred to as Wayback).

2. WARCREATE

WARCreate[3] is an extension for the Google Chrome web browser that allows a user to generate a WARC file from the current webpage. In addition to creating a valid WARC that can be viewed in Wayback, the extension provides options that address privacy concerns (*e.g.*, a user might want the data encrypted in some way), potential bleed over (*e.g.*, two different users see different content at http://facebook.com), and other issues that may not be relevant to conventional web archiving as performed by Heritrix and the Internet Archive.

To create a WARC file from the current webpage, the user clicks on the browser extension's icon in the address bar and then presses the Generate WARC button (see Figure 1). The browser extension gathers the resources (including external scripts, CSS and images) and HTTP headers normally used by the web browser to generate a webpage and adds metadata (the *warcinfo* records) to generate a WARC file that conforms to the standard's specification (Figure 2). Adherence to the specification allows the WARC to be read by Wayback.

When the compilation of the WARC file is complete, the file is downloaded to the local file system. The browser ex-

JCDL'12, June 10–14, 2012, Washington, DC, USA.
ACM 978-1-4503-1154-0/12/06.

[1] https://webarchive.jira.com/wiki/display/Heritrix/Heritrix
[2] Archived on July 25, 2011
[3] http://matkelly.com/warcreate

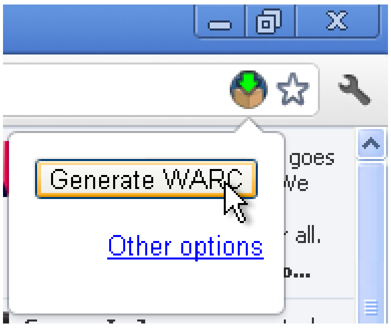
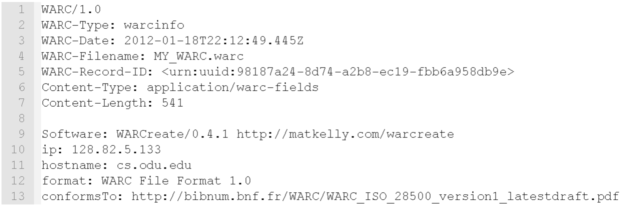


Figure 1: Creating a WARC is as simple as selecting the extension's icon and pressing the Generate WARC button. Other options can be selected for further customization of the generated file.

tension also allows a user to open an existing WARC file generated by the tool to make modifications. Any changes to the WARC file must be accounted for in the *content-length* WARC fields to ensure compatibility with the WARC standard. This overcomes the potential corruption of the WARC records that might occur if the file is edited with a simple text editor. Further functionality is being developed to allow the extension to read in, interpret and allow modification of any arbitrary WARC file. This would require comprehensive implementation of the WARC specification.

Unlike a web crawler, this tool archives only the single webpage that the user is visiting. In the future, we intend to add more of a crawling-like functionality. For now, the current extension can append the archive of the current page to an existing WARC file. This allows links to content from separate archiving sessions to be resolvable by Wayback.



```
WARC/1.0
WARC-Type: warcinfo
WARC-Date: 2012-01-18T22:12:49.445Z
WARC-Filename: MY_WARC.warc
WARC-Record-ID: <urn:uuid:98187a24-8d74-a2b8-ec19-fbb6a958db9e>
Content-Type: application/warc-fields
Content-Length: 541

Software: WARCreate/0.4.1 http://matkelly.com/warcreate
ip: 128.82.5.133
hostname: cs.odu.edu
format: WARC File Format 1.0
conformsTo: http://bibnum.bnf.fr/WARC/WARC_ISO_28500_version1_latestdraft.pdf
```

Figure 2: A generated WARC file is prefixed with a *warcinfo* record that conforms to a liberally attributed template while still maintaining minimum requirements.

3. OTHER CONSIDERATIONS

There are many facets of this type of personal web archiving that we leave for future exploration. The main issues we will investigate are ensuring privacy, maintaining archive integrity, preventing bleed over, and developing a service to consume the potentially private WARCs generated by the extension.

3.1 Privacy Concerns

Much of the data in personal web archives is of a sensitive or personal nature, thus the data placed in a WARC file must be protected. One way to achieve this is to encrypt the data so that only the recipient can interpret the archive. There are several issues to consider, such as ease of decryption, decentralization, increased space and time complexity, and loss of access when a central authority disappears. In the current Chrome extension, we have implemented a simple key-based encryption/decryption scheme for users that wish to test the reliability of this preliminary technique.

3.2 Archive Integrity

Heritrix subtly manipulates content in an archive for ease of navigation and to allow a webpage to be viewed in the modified context of the Wayback Machine user interface. Because data is manipulated, the integrity of the archive is compromised, as the page is not retained exactly in its original form. In future work, we hope to expand on the WARC format (and demonstrate this with an implementation) to provide WARC records containing the differential of how the webpage was manipulated. By utilizing this differential, the original content can be re-created much in the same way that software versioning uses differentials to store and revert to older versions of code.

3.3 Bleed Over

An issue that is not as common on the surface web as on the personalized deep web is that of a single URI referring to potentially different information. This can be seen when two users access a social media website while each is logged in (*e.g.*, http://www.facebook.com). The website uses authentication to provide customized information to each user. If two users were to create an archive of the same page, the resulting WARCs should be different and each user should only be able to access the WARC they created. This bleed over issue will be addressed in future work.

3.4 Service to Consume WARCs

Making WARC files accessible to the end user is just the first step. Unless a user has a personally deployed instance of Wayback, little can be done with the resulting file. We will investigate the feasibility of rapid deployment of a personal instance of Wayback or of finding an alternative for quick replay of small archives that were created with the Chrome extension. Initial approaches to accomplish this will attempt to decouple the code used to interpret WARC files from the open source Wayback and minimize the need for reinvention of these procedures while maintaining consistency of implementation with future versions of Wayback.

4. REFERENCES

[1] M. K. Bergman. The deep web: Surfacing the hidden value. *Journal of Electronic Publishing*, 2000.

[2] ISO. Information and documentation - WARC file format, 2009.

[3] C. C. Marshall. Rethinking Personal Digital Archiving, Part 1: Four Challenges from the Field. *DLib Magazine*, 14(3/4):2, 2008.

[4] C. C. Marshall, S. A. Bly, and F. Brun-Cottan. The long term fate of our digital belongings: Toward a service model for personal archives. In *Proceedings of IS&T Archiving*, pages 25–30, 2006.

[5] C. C. Marshall, F. McCown, and M. L. Nelson. Evaluating personal archiving strategies for internet-based information. In *Proceedings of IS&T Archiving*, pages 151–156, 2007.