Dodging the Memory Hole Project Report
Twitter Feed Monitoring and Automatic Archival Through WAIL

John Berlin
Department of Computer Science
Old Dominion University, Norfolk, VA - 23529 (USA)
jberlin@cs.odu.edu

1. Introduction

Current tools available to archive news content generated on Twitter require archivists to have some technical background in order to initiate the archival process using archival crawlers like Heritrix [1] or have access to institutional based subscription services like Archive-It[2]. This makes it difficult for archivists to preserve news stories as they evolve or as non-news affiliated users on Twitter begin to generate content pertaining to the story. Services such as Webrecorder[3] that do allow the archivists to preserve the evolution of the news story require them to both trust the service with their login credentials for Twitter and to manually drive the archival process.

This project seeks to address these issues by extending the Web Archival Integration Layer[4] (WAIL) with the ability to monitor a user's Twitter feed for automatic archival. Through user configuration the system looks for content in the feed, and when an entry meets the criteria it is archived.

2. WAIL

WAIL is a tool which integrates an archival crawler (Heritrix) and replay system (Wayback[5]) to facilitate an individual's preservation. WAIL also takes care of all the technical details required to use the packaged tools. This allows users of all technical backgrounds to archive what they see now. Originally created by Mat Kelly[6] and the Web Science and Digital Libraries (WS-DL) research group[7], I have taken over the responsibilities for the continual development of WAIL.

Apart of this effort WAIL has been to vastly revise WAIL by using modern Web technologies and introduced the concept of collection-based personal Web archiving which can be accomplished on a user's machine. Unlike subscription-based Web archiving services like Archive-It, WAIL provides an interoperable mechanism to accomplish this without reliance on an external service. WAIL has

[1] https://webarchive.jira.com/wiki/display/Heritrix
[2] https://archive-it.org
[3] https://webrecorder.io
[4] https://github.com/N0taN3rd/wail/wiki
[5] https://github.com/iipc/openwayback
[6] http://matkelly.com
[7] https://ws-dl.cs.odu.edu

been rebuilt using its core concept into an Electron-based[8] native application for a more consistent and accessible interface with better integration with Heritrix and Wayback. Figures 1 and 2 show the updated user interface of WAIL. This has also allowed the addition of this project into the new version of WAIL which is comprised of two parts: the monitoring of the Twitter feed and the automatic archival of the content.

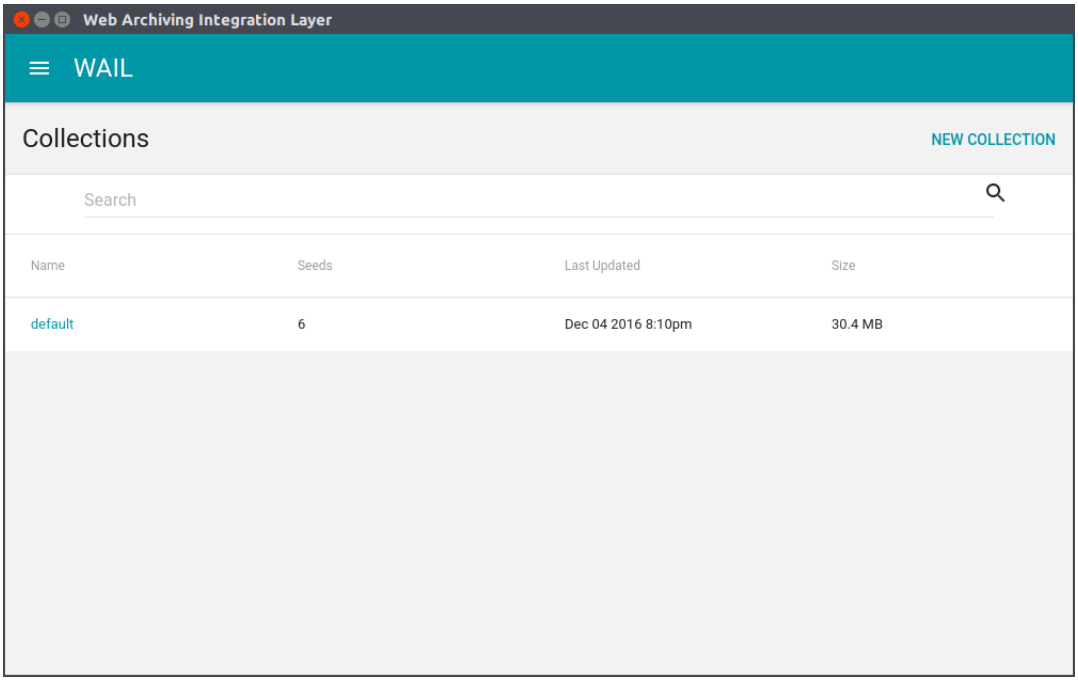

Figure 1: WAIL Main UI

[8]http://electron.atom.io

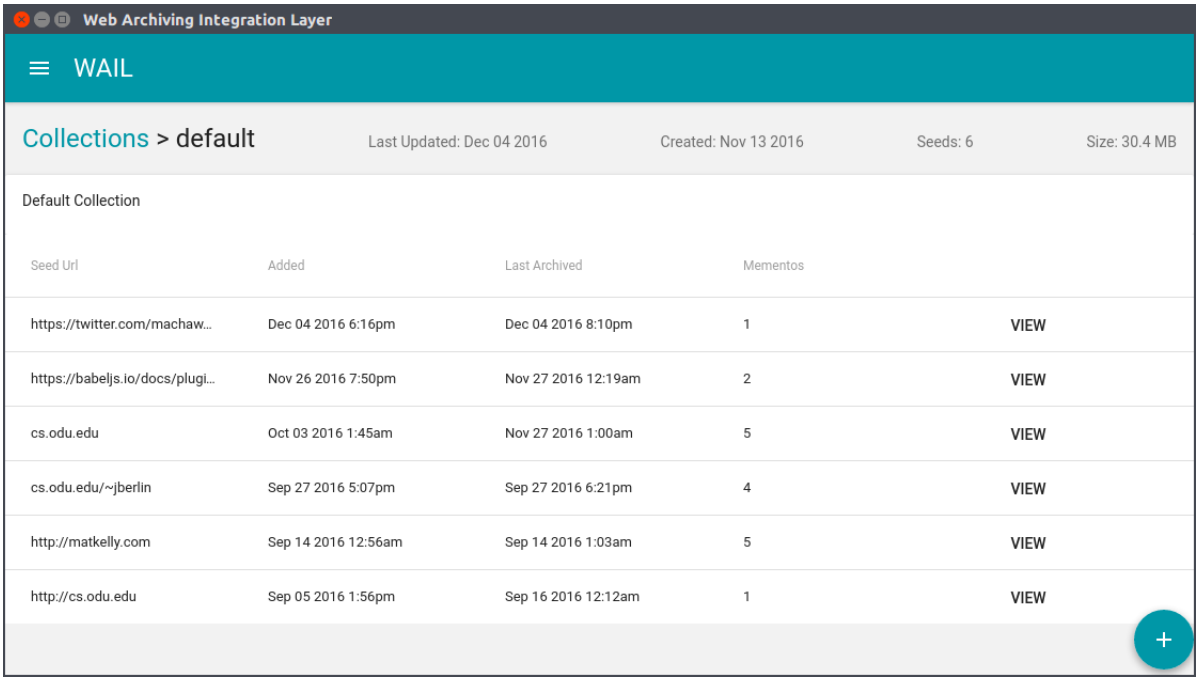

Figure 2: WAIL Collection View

3. Twitter Feed Monitoring

In order to discover Twitter content for preservation, WAIL has been registered as an application through Twitter. By registering WAIL with Twitter, WAIL can have access to the user's personal view of Twitter. This also allows WAIL to make requests to the Twitter API on behalf of the user while not compromising the user's login credentials [9]. Currently there are two options for Twitter archival feature implemented in WAIL. The first is monitoring a user's timeline for tweets which were tweeted after the monitoring has started with the option of selecting only the tweets containing hashtags specified by the user. The second, a slight variation of the first, will only archive tweets that have specific keywords in the tweet's body as specified by the user. Figure 3 shows the Twitter archival configuration interface available in WAIL which were combined into a single screen for this report.

[9]https://dev.twitter.com/web/sign-in

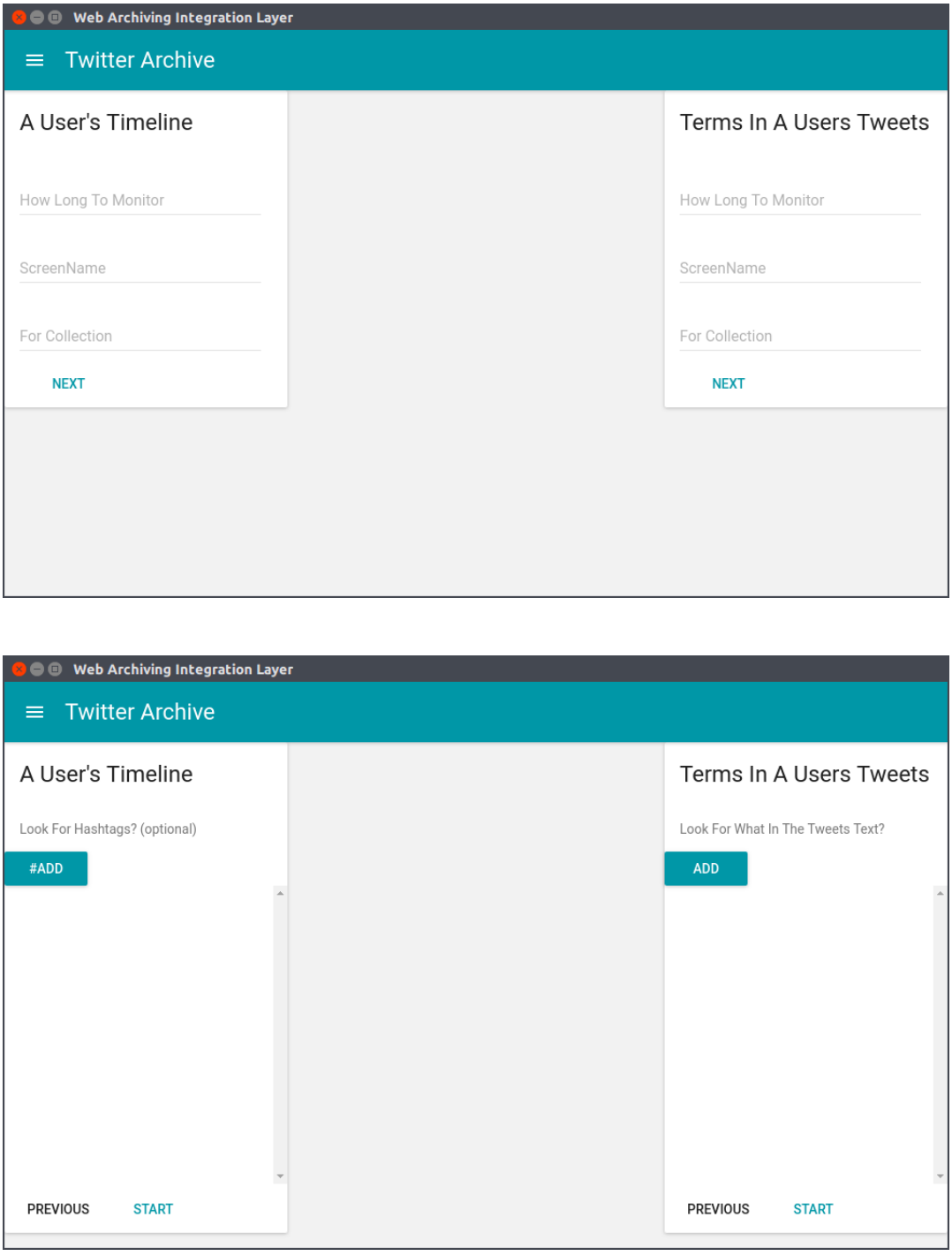

Figure 3: Twitter Archival Configuration User Interface

4. Archival Of Twitter Content

Before this project WAIL utilized the archival crawler Heritrix as the preservation means. Heritrix (the Internet Archive's archival crawler) executes HTTP GET requests to retrieve the target webpage and to archive the HTTP response headers and the content returned from the server. The embedded Javascript of the webpage is not executed potentially decreasing the fidelity of the capture[10]. This is problematic when archiving Twitter content since the rendering of tweets is done exclusively through client side Javascript. Addressing this is the utilization of a native Chromium browser via Electron in conjunction with the previously developed tool WARCreate [1].

WARCreate is a Google Chrome extension, also originally created by Mat Kelly, that allows a user to create a Web ARChive (WARC[11]) file from any browseable webpage. I made modifications to WARCreate in order to integrate it with WAIL to eliminate need for human intervention to decide when to generate the WARC and to work inside of the Electron provided Chromium browser. By integrating WARCreate into WAIL the archival process of Twitter content has been simplified to loading the URL of the tweet into the browser and waiting until the browser indicates that the page has been rendered in its entirety. Then the archival process through WARCreate is initiated. Once the WARC has been generated, it is added to a collection of web archives that was created by the user through WAIL.

5. Future Work

The Twitter monitoring implementation could make use of Twitter's streaming API[12] in order to archive Tweets created by users not specified in the initial configuration. Likewise the options for the features of what constitutes a desirable tweet for archiving could be expanded to include sentiment analysis of the tweet. The Electron version of WARCreate is still in development in order to generalize it for usage outside of WAIL. When completed it will be released as open source software.

6. Project Results

In this project I added to WAIL the ability to monitor the Twitter feed for a given user(s) and the automatic archival of the user(s) Tweets using a native browser. Additionally, I modified WARCreate in order to utilize it in the archival process of those Tweets thus ensuring the fidelity of the capture. The latest revision of WAIL which will include this project's implementation will be made available as open source software via github at https://github.com/N0taN3rd/wail, which is scheduled for release by the end of December 2016.

References

[1] M. Kelly and M. C. Weigle. Warcreate - Create Wayback-Consumable WARC Files from Any Webpage. In *Proceedings of the ACM/IEEE Joint Conference on Digital Libraries (JCDL)*, pages 437–438, Washington, DC, June 2012.

[10] http://ws-dl.blogspot.com/2013/11/2013-11-28-replaying-sopa-protest.html
[11] http://www.digitalpreservation.gov/formats/fdd/fdd000236.shtml
[12] https://dev.twitter.com/streaming/public