Visualizing Webpage Changes Over Time

Michele C. Weigle, Michael L. Nelson, Deborah Kempe, Pamela Graham, Alex Thurman

Data Management Plan

Expected Data and Outputs
The main product of this project will be software, specifically for generating thumbnail representations of TimeMaps. The software will be written in either Java (for the Wayback Machine), or in Python (for "pywb" -- the open source Python implementation of the Wayback Machine that is gaining traction in the web archiving community). The software that will write to process third-party TimeMaps (i.e., creating thumbnails of TimeMaps in arbitrary web archives that we do not have root access to) will likely be written in Python. Software written for embedding thumbnail representations in web pages will be written in Javascript and HTML.

ODU has a dark archive of most of Archive-It's collections (totaling 230+TB and > 5.3M mementos, representing Archive-It through April 2013, but updates are in the process of being obtained), which are stored in WARC (Web ARChive) files, which are the official (ISO 28500:2009) and most popular format for storing the results of web crawling activities. For detailed information about the WARC format, see http://www.digitalpreservation.gov/formats/fdd/fdd000236.shtml

We will also work directly with the WARC files of specific interest to NYARC and CUL that may lie outside of our collection.

We expect the output of generating thumbnails to be stored in a new, yet-to-be-determined format, but it will be serialized in JavaScript Object Notation Format (JSON), the de facto standard format for web processing. The images themselves will likely be Portable Network Graphics (PNG) format, a popular and standard (IETF RFC 2083) format for storing images on the web.

Period of Data Retention
ODU has 120 TB of mass storage available for faculty research data. As a backup we will store all generated source code and example datasets for demonstrating the code on this mass storage system for at least 5 years, though there is no standard time limit on data storage for this mass storage system.

The Archive-It WARC files themselves are part of an ongoing institutional commitment by ODU to being a dark archive for Archive-It, so those files will be maintained and internally available for at least the next 5 years, most likely much longer.

Our research group at ODU now uses GitHub for our code, documentation, and code management. We commit to maintaining public access to our developed code for at least 5 years through GitHub, or other public source code system, if for some reason GitHub is no longer available.

NYARC and CUL are clients of Archive-It, who will be responsible for the long-term archiving of their WARC files.

Data Formats and Dissemination
As stated above, we now use Github for code dissemination; our existing projects can be explored at https://github.com/oduwsdl/

As we have done in the past, we will continue to use this method to make our code publicly available to the web archiving community and the public at large. We use the MIT Open Source license, see for example: https://github.com/oduwsdl/ipwb/blob/master/LICENSE

The WARC files themselves are not publicly available (as per our agreement with Archive-It, and consistent with us being a dark archive). However, there are several commonly used, open-source tools available that create WARC files (e.g., Heritrix, wget, webrecoder) so there will not be a problem in applying our code in new scenarios. Similarly, the Wayback Machine software and pywb are both open source as well.

Publications resulting from this project will be stored as PDF documents in public repositories such as the ACM Digital Library and the arXiv eprint server. These will also be made available from the ODU WS-DL research group's web page, http://ws-dl.cs.odu.edu, promoted on our blog http://ws-dl.blogspot.com, and on our Twitter page http://twitter.com/WebSciDL.

Data Storage and Preservation of Access
ODU has 120 TB of mass storage available for faculty research data. As a backup we will store all generated source code and example datasets for demonstrating the code on this mass storage system.

Our source code and documentation will be stored and maintained at Github, with local copies of the code repositories on research servers at ODU (with several TBs of storage space currently available).

The Archive-It WARC files themselves are part of an ongoing institutional commitment by ODU to being a dark archive for Archive-It.