

Data Management Plan

Expected Data

The main data that will be produced in this project is the source code of the WARCreate tool. The code will likely be a mix of JavaScript, Python, and other languages as the environment dictates (e.g., XML User Interface Language for developing Firefox add-ons).

By-products of testing the WARCreate tool will be standard WARC (Web ARChive) files. These files will be generated from publicly available web sites and from websites behind authentication (e.g., pages behind login prompts, social media profiles).

As this project's purpose is to provide a tool to allow researchers to generate their own data (i.e., create their own WARC files), we will not be generating WARC file datasets for the public ourselves. But we will create a set of example WARCs generated by the tool to be distributed with the software.

Period of Data Retention

ODU has 120 TB of mass storage available for faculty research data with a default retention time of 5 years. As a backup we will store all generated source code and example datasets on this mass storage system for at least 5 years.

Data Formats and Dissemination

The source code and example datasets will be stored in local and then public revision control systems (e.g, Google Code or SourceForge), which will handle the dissemination. We will make the source code available to the public under a Creative Commons CC BY-SA-3.0 license (<http://creativecommons.org/licenses/by-sa/3.0/>) as soon as it is mature enough to share. We will place the source code and example datasets into the community as soon as possible in order to receive feedback and get other developers involved with the project. We will not have an explicit embargo period for the data or software.

The example datasets will be sample WARC files generated by the tool. For detailed information about the WARC format, see

<http://www.digitalpreservation.gov/formats/fdd/fdd000236.shtml>

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 project director's web page, <http://www.cs.odu.edu/~mweigle>.

Data Storage and Preservation of Access

We will store our source code in our local ODU-CS SVN repository, in ODU mass storage, and at a public open-source software distribution site, such as Google Code or SourceForge.