Working with WARCS: New Tools for Harvesting, Accessing, and Researching Web Archives

Jefferson Bailey
Internet Archive
300 Funston Avenue
San Francisco, CA, CA 94118, USA
jefferson@archive.org

Maria Praetzellis
Internet Archive
300 Funston Avenue
San Francisco, CA, CA 94118, USA
maria@archive.org

Vinay Goel
Internet Archive
300 Funston Avenue
San Francisco, CA, CA 94118, USA
vinay@archive.org

ABSTRACT

The process of collecting, managing, and serving archives of web published materials poses many challenges. Web archives tend to be quite voluminous, data-wise and number-of-items-wise, with even highly curated collections being many terabytes in size and tens, if not hundreds, of millions of URLs. Web archives also have longitudinal complexity, bearing a temporal aspect similar, but different, to serial publications, with frequent changes in content (and often state of existence) even at the same URL, gobs of metadata both content-based and transactional that has research or administrative value, and have overall characteristics that make them highly suitable for data mining and computational analysis. However, web archiving as a practice is still mostly fractionally-staffed, with few institutions able to contribute engineering resources to innovation and research and development.

The Internet Archive has been pioneering web archiving technologies for 20 years. A number of recent initiatives have sought to reinvigorate innovation around new tools designed to address improving content capture, reconceptualizing access and discovery, and facilitating the increasingly popular scholarly approaches of computational analysis of large born-digital collections. This workshop will provide hands-on demonstrations of, and, when possible, training in, new tools for capturing, analyzing, facilitating use of web archive collections. This includes training in browser based capture methods, curatorial tools, data derivation and extraction methods for research use, demonstration of APIs and their uses for studying archived web data, new search and aggregation tools, and frameworks for web archive processing.

Workshop Scope & Intended Content

Attendees will use a range a tools as part of the workshop, including the ability to run scripts and tools on web archive data that they collector prior to, or bring with them to, the workshop. In general the workshop will cover these areas of practice through a mix of hands-on work, demonstrations, and discussion:

- Participants will use and discuss new API based tools for enhanced access to, and research use of, web archive data. In-production and test APIs will be used for data transport, content analysis, metadata aggregation, and other collection management and discovery activities.
- Participants will be trained in how to generate datasets from web archive collections, including gaining familiarity with the suite of derivative formats, indexing tools, and how to facilitate and support researchers interested in using archived web data.
- Participants will be trained in the latest open-source crawling technologies including browser based capture and API harvesting. This includes both Brozzler, a distributed web crawler, and new other tools designed to harvest social media data. Participants will be given the opportunity to compare crawling mechanisms by using a variety crawling methods, and discuss how they can incorporate these tools in their own institutions.
- Demonstration of recent innovations in providing access to large web archive corpora, including user testing and technical review of new search technologies for querying large-scale web archives, including a review of methods, technologies, interfaces, and other cutting-edge approach to dealing with web archive content discovery.
- Review new dataset formats and tools designed to enable easy data processing, extraction and derivation of web archive data for computational analysis. This includes the development of a community space to share modules and recipes for analyzing data sets that can be reused among scholars as well as the exploration of derivative data formats and interfaces like Python-base Jupyter Notebooks for pre-scripted data analysis.

Requirements & Target Audience

The workshop requires a meeting room with wireless internet access and a projector with screen. Participants will bring their own laptop computers and there should be sufficient power outlets. The workshop hosts will coordinate preliminary activities over email and provide some basic technical support beforehand for workshop attendees.

The target audience for this workshop includes professionals working in digital library services that are collection, managing, or providing access to web archives, scholars using archived web data in their work, or digital library and preservation professionals working to provide computational access to digital collections. Some knowledge of or experience with web archives is helpful but not required and some familiarity or comfort with working at the command line interface is helpful.

Keywords

Web archiving; digital collections; data mining; research