Data Mining Web Archives

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

Lori Donovan
Internet Archive
300 Funston
San Francisco, CA, CA 94118, USA
lori@archive.org

ABSTRACT

Many institutions are now building rich, significant archives of web content. Though the number of web archiving programs has grown, access models for these collections have remained focused on URL-based discovery and traditional live-web-style browsing. Given the resources required to build and maintain web archives, finding new forms of access for these collection will help increase use and thus allow institutions to better advocate for the value of collecting and preserving web content.

Distant reading, text mining, digital humanities, and other data-driven forms of analysis have become increasingly popular methods of using digitized and digital collections. Web archives, being born-digital, of notable size and temporal breadth, having extensive metadata, and often created with a curated topical focus, are ideal resources for data mining and other forms of computational analysis.

This workshop will explore new methods of research use of web archives by giving attendees exposure to, and training in, the tools, methods, and types of analysis possible in working with datasets extracted from the entirety of curated web archive collections. Giving researchers datasets of specific extracted metadata elements, link graph data, named entities, and other post-processed data can help facilitate new uses and new types of visualization, inquiry, and analysis.

Workshop Objectives:

- Introduce attendees to web archives and the issues of provenance, formats, methods of collection, and the core tools and technologies involved in web archiving
- Give an overview of the types of derived datasets that can be created from web archives
- Provide sample datasets, scripts and tools, and outline research and use scenarios
- Explore methodological challenges and possibilities
- Lead attendees through a data analytic workflow that includes processing, publishing, and visualizing web archive data

Keywords

Web archiving, data mining, research, access