WADL 2016 Panels: Worldwide activities on Web archiving; Social media, Web archiving, and digital libraries

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

Daniel Kerchner
George Washington University Libraries
Washington, DC 20052 USA
kerchner@gwu.edu

ABSTRACT

In addition to presentations based around scholarly papers, the Web Archiving and Digital Libraries 2016 workshop also featured two panel discussion sessions. Each panel centered on a theme and featured short presentations by the panelists, followed by a moderated discussion and interaction with workshop participants; the panels are described herein.

1. PANEL 1: WORLDWIDE ACTIVITIES ON WEB ARCHIVING

This panel focused on showcasing worldwide activities related to Web archiving. The presented topics ranged from crawling technologies, collaborative collection-building strategies, access methodologies, and on-going efforts to partner with and support researchers in their use of Web archives.

Vinay Goel, Senior Data Engineer, Internet Archive began by giving an update on some of the Web archiving projects that Internet Archive was involved with: global and domain scale crawling, collaborative crawling around spontaneous events and orphan domains, topic, curated crawling using Archive-It, and research partnerships with a number of institutions around the world. The talk highlighted the value of expanding access models by providing examples of search interfaces and researcher datasets, and describing scalable tools and expressive frameworks like ArchiveSpark that hide the complexities of cluster computing and of data formats and Web collection indices. The talk stressed the need for engaging with user communities for collaborative technology development with a focus on fault tolerant content acquisition and distribution models, and building scalable and interoperable crawling and access systems.

Ian Milligan, Assistant Professor, Department of History, University of Waterloo, shared details on Archives Unleashed, a series of Web archive hackathons, sponsored by Rutgers University, Internet Archive, and the University of Waterloo, and supported by the US National Science Foundation and the Social Sciences and Humanities Research Council of Canada. The goal of these datathons was to bring together multi-disciplinary researchers and to have them collaborate on exploring and developing cutting-edge research tools, and to build consensus on the methodologies of analyzing Web archives.

2. PANEL 2: SOCIAL MEDIA, WEB ARCHIVING, AND DIGITAL LIBRARIES

Panelists presented on the topics below, followed by an engaging discussion on a wide range of topics including but not limited to issues of providing access to social media collections, outstanding challenges in collecting social media content, documenting provenance of social media collecting, and key differences between Web archiving and social media collecting:

Ian Milligan, Assistant Professor, Department of History, University of Waterloo, spoke on collecting and analyzing Canadian Federal elections tweets as a case study for documenting events. The case study collected tweets using the #elxn42 hashtag and analyzed the collection for tweet frequency over the course of the election timeframe, word frequency, top users, most tweeted URLs, and other metrics. The talk touched on considerations of legality, ethics, and the lowering of technical barriers to empower more individuals to build collections.

Mark Phillips, Associate Dean for Digital Libraries, University of North Texas, described the University of North Texas Libraries' experience providing access to social media collections by packaging social media archive content for their repository. The talk highlighted how UNT Libraries has defined a set of collection-level metadata to describe a collection of Twitter data; and how UNT Libraries is providing varying levels of access to the collection data; some are provided as lists of tweet identifiers which a researcher could then "rehydrate," whereas others are available as full JSON Twitter data sets.

Laura Wrubel, Software Development Librarian, George Washington University Libraries, presented an introduction to the Social Feed Manager[1], version 1.0 of which had just been released. Social Feed Manager (SFM) is an open source Web application that allows users to create collections of data from social media platforms including Twitter, Tumblr, Flickr, and Weibo. The talk highlighted how SFM is innovative in the way that it stores social media platform API responses, typically JSON, in WARC[2] files, so that the entire request and response are recorded; and how SFM is innovative in recording the history of the user's creation and modification of the collection. The talk closed with questions and an invitation for feedback on the topic of appropriate and useful provenance metadata as pertains to collections of social media data.

[1] http://gwu-libraries.github.io/sfm-ui/

[2] http://www.digitalpreservation.gov/formats/fdd/fdd000236.shtml