Leveraging Web Archiving Tools for Research and Long-Term Access

Lori Donovan
Internet Archive
300 Funston Ave
San Francisco, CA 94118
1-415-561-6799 x4
lori@archive.org

ABSTRACT

This workshop will introduce participants to web archiving concepts and challenges, including creating web archives and providing for access and research use.

General Terms

Preservation strategies and workflows, specialist content types, training and education.

Keywords

Web archiving, research services, access

1. INTRODUCTION

Web archiving is an important part of the digital preservation field. While most are familiar with the Wayback Machine available at archive.org, less are aware that there are a number of tools and services developed for organizations and individuals to create their own web archives, including the capability to search and analyze large data sets built around the WARC file format, an ISO standard for web archiving. In addition, web archives provide permanent URLs for citation and can show how a website has changed over time at a single URL, even if no longer available on the live web. In short, web archives can provide very necessary preservation tools for researchers and archivists to manage content that is only posted on the web.

This workshop will introduce participants (15--20) to basic web archiving concepts and challenges. Using the Archive--It (www.archive--it.org) web application, participants will have a hands--on opportunity to build a collection of content archived from the web, which can include their own organization's web presence, social media, digital exhibitions, data sets, or topical content publicly available on the web. Following the workshop participants will have a searchable archive available to them, including the option of downloading WARC files for long-term preservation or research.

The target audience for this workshop includes interested scholars researching the web and professionals responsible for digital library services or digital archives. No prerequisite knowledge of or experience with web archives is necessary, and the session does not require any programming or advanced technical knowledge of the web. The workshop will not be oriented towards those with deep knowledge of web archives or the WARC format, although there could be time allotted to a demonstration of another web archiving tool or project related to web archiving and this should be specified in the CFP (see below).

2. PARTICIPANT INFORMATION

In order to make the most of the workshop and ensure that the curriculum is tailored toward participant interest, some additional information about participants would be helpful. It should include:
--Description of participant interest areas and/or professional projects.
--Description of prior experience with using web archives or their own web archiving (if applicable).
--5 to 10 websites to be archived as part of 1 or more collections of content, and links to the Robots.txt files if applicable. More information is here:
https://webarchive.jira.com/wiki/display/ARIH/Robots+Exclusion+Protocol
With permission from participants, URLs will be crawled as a test (no data archived) prior to the workshop so post crawl reports can be analyzed as part of the workshop curriculum.
If possible, the instructor should receive this information at least 1-2 weeks before the workshop.

3. ABOUT THE INSTRUCTOR

Lori Donovan is a Partner Specialist at the Internet Archive helping libraries, museums and other cultural institutions archive web content. Over the past four years, Lori has given more than 25 presentations on web archiving at library, archives and digital preservation conferences both in the United States and internationally. Lori has a Masters of Science in Information from the University of Michigan specializing in Archives and Digital Preservation.