Analytics for Monitoring Usage and Users of Archive-It Collections

Todd Suomela
University of Alberta
Edmonton, AB, T6G 2R3, Canada
todd.suomela@ualberta.ca

The University of Alberta (UA) web archive collections are managed by a distributed group of subject and specialist librarians across the university. Librarians, faculty, and the public can propose collections; proposals are reviewed by the Collections Development Committee to fit with the overall collection policies of the library. Once approved, individual collections are managed by subject librarians who propose seeds, manage crawl schedules, and monitor crawl activity for quality.

One challenge for a distributed management structure is maintaining standard operational procedures across multiple collections. Subject librarians have many different demands on their time, so web archiving may be reduced because of other demands on their time. In January 2015 I was appointed web archiving coordinator to streamline the operational procedures for web archiving. Part of this activity has been analyzing Google analytic tracking of users and developing management reports for all collections at the University of Alberta.

The initial user analytics show 30,000-50,000 page views per month and 250-500 sessions per day. Activity is unevenly distributed between different collections. The most trafficked collection, the Heritage Community Foundation, receives 300-4000 page views per day. Other collections may receive zero page views for many days and then have spikes of activity of several hundred views per day.

The R statistics for analyzing crawl reports show a wide range of crawl sizes and frequencies. UA has crawls running with different periods of repetition: daily, weekly, monthly, bimonthly, quarterly, semi-annually, annually, or one-time. Crawl sizes during 2014 ranged from a few megabytes to 900 GB. The reports created with these scripts are being used to manage the annual data budget for Archive-It (https://www.archive-it.org/) and are available for use by other people at https://github.com/tsuomela/ai-crawl-reports .