Building a National Web Archiving Collaborative Platform: The Web Archives for Longitudinal Knowledge Project

Ian Milligan
University of Waterloo
Waterloo
i2millig@uwaterloo.ca

Nick Ruest
York University
Toronto
ruestn@yorku.ca

Ryan Deschamps
University of Waterloo
Waterloo
ryan.deschamps@uwaterloo.ca

KEYWORDS

Web archiving, Blacklight, Solr, Canada

ACM Reference format:
Ian Milligan, Nick Ruest, and Ryan Deschamps. 1997. Building a National Web Archiving Collaborative Platform: The Web Archives for Longitudinal Knowledge Project. In *Proceedings of Web Archiving and Digital Libraries 2017, Toronto, ON, June 2017 (JCDL 2017),* 1 pages.
https://doi.org/10.1145/nnnnnnn.nnnnnnn

In the absence of a national web archiving strategy, Canadian governments, universities, and cultural heritage institutions have pursued disparate web archival collecting strategies. Carried out generally through contracts with the Internet Archive's Archive-It services, these medium-sized collections amount to a significant portion of Canada's born-digital cultural heritage since 2005. While there has been some collaboration between institutions, notably via the Council of Prairie and Pacific University Libraries in western Canada, most web archiving collecting has been taking place in silos. Researchers seeking to use web archives in Canada are thus limited not only to the Archive-It search portal, but also to exploring on a silo-ed collection-by-collection basis. Given the growing importance of web archives for scholarly research, our project breaks down silos and generate a common search portal and derivative dataset provider.

Our Web Archiving for Longitudinal Knowledge (WALK) Project, housed at http://webarchives.ca/and with our main activity via our GitHub repo at https://github.com/web- archive-group/WALK, has been bringing together Canadian partners to integrate web archival collections. Co-directed by a historian and a librarian, the project brings together computer scientists working on the warcbase project, doctoral students working on governance issues, and students running tests and usability improvements. We currently have 20TB of web archival collections, aggregated from the Universities of Toronto, Alberta, Winnipeg, and Victoria, as well as Dalhousie and Simon Fraser University. Our workflow consists of:

- Signing Memorandum of Agreements (MOU) with partner institutions;
- Gathering WARCs from partner institutions into Compute-Canada infrastructure through the Research Portals and Projects (RPP) program;
- Using warcbase[1] to generate scholarly derivatives, such as domain counts, link graphs, and files that can be loaded into network analysis software[2];
- Adapting the Blacklight[2] front end to serve as a replacement for our current SHINE interface; this will allow built-in APIs, faceted search by institution, and inter-operability with university library catalogues[1];
- Using a team of research assistants to describe each collection using Python and R;
- Finally, using multiple correspondence analysis, generating profiles of each web archive with an eye towards assisting curators in finding collection overlap/gaps[3].

This presentation provides an overview of the WALK project, focusing specifically on questions of interdisciplinary collaboration, workflow, dataset creation and dissemination. As web archiving increasingly happens at the institutional level, the WALK project suggests one way forward towards collaboration, collection development, and researcher access.

REFERENCES

[1] Andrew Jackson, Jimmy Lin, Ian Milligan, and Nick Ruest. 2016. Desiderata for Exploratory Search Interfaces to Web Archives in Support of Scholarly Activities. In *Proceedings of the 16th ACM/IEEE-CS on Joint Conference on Digital Libraries (JCDL '16).* ACM, New York, NY, USA, 103–106. https://doi.org/10.1145/2910896.2910912

[2] Jimmy Lin, Ian Milligan, Jeremy Wiebe, and Alice Zhou. 2017. Warcbase: Scalable Analytics Infrastructure for Exploring Web Archives. *ACM Journal of Computing and Cultural Heritage* (2017).

[3] Ian Milligan, Nick Ruest, and Jimmy Lin. 2016. Content Selection and Curation for Web Archiving: The Gatekeepers vs. The Masses. In *Proceedings of the 16th ACM/IEEE-CS on Joint Conference on Digital Libraries (JCDL '16).* ACM, New York, NY, USA, 107–110. https://doi.org/10.1145/2910896.2910913

JCDL 2017, June 2017, Toronto, ON

ACM ISBN 978-x-xxxx-xxxx-x/YY/MM.
https://doi.org/10.1145/nnnnnnn.nnnnnnn

[1] http://warcbase.org
[2] http://projectblacklight.org/