https://warclight.archivesunleashed.org

Warclight: A Rails Engine for Web Archive Discovery

Nick Ruest (York University)
Ian Milligan & **Jimmy Lin** (Univ. of Waterloo)

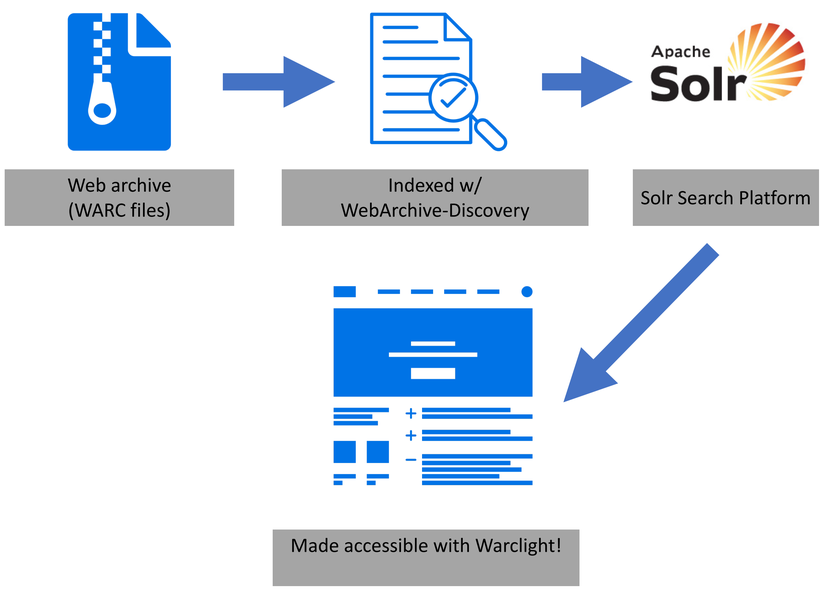
Introduction

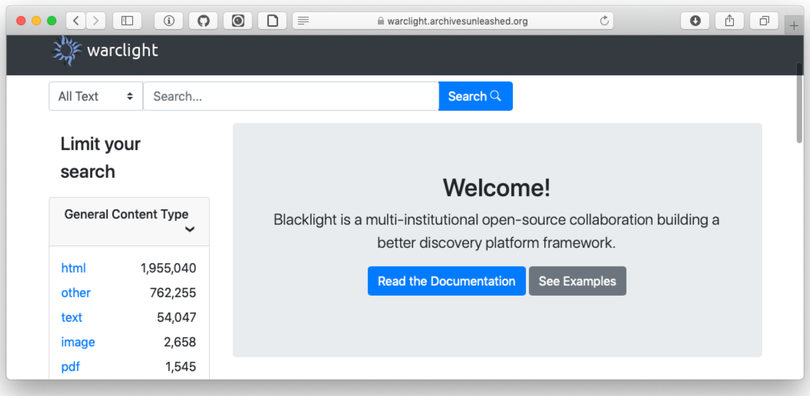
- Since 1996, organizations have been systematically crawling the web and building web archives.
- Researchers are faced with obstacles when using current methods to access collections (e.g. knowing specific URLs to search via a Wayback Machine, keyword searches, or full-text search across hundreds of thousands of documents).
- The standard search engine results page (SERP) is inadequate to support scholarly inquiries.
- Scholars thus need **exploratory interfaces** that support different methods of discovery in web archives.

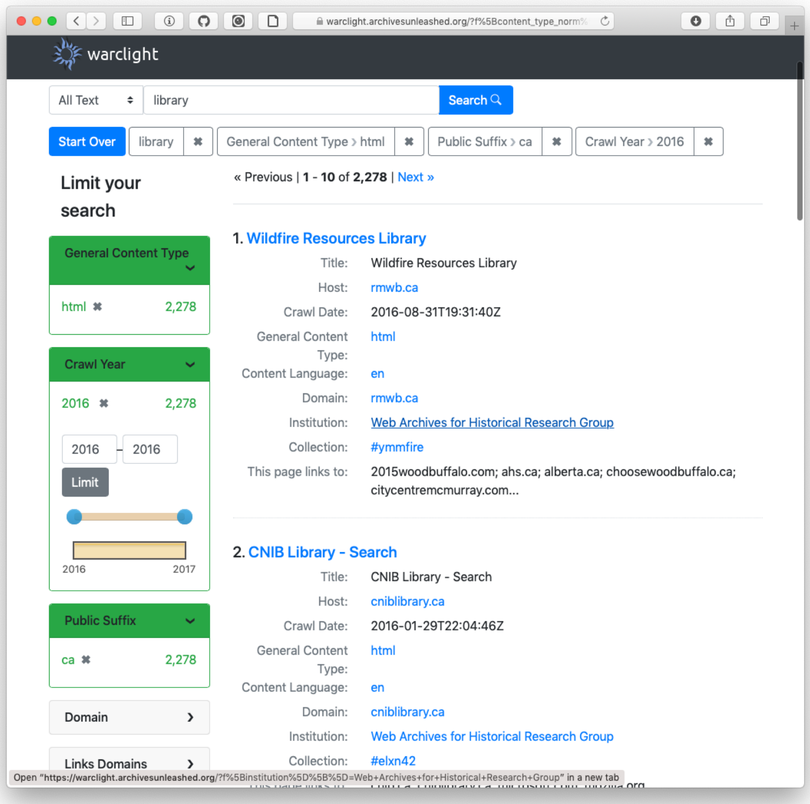
Inspiration and Design

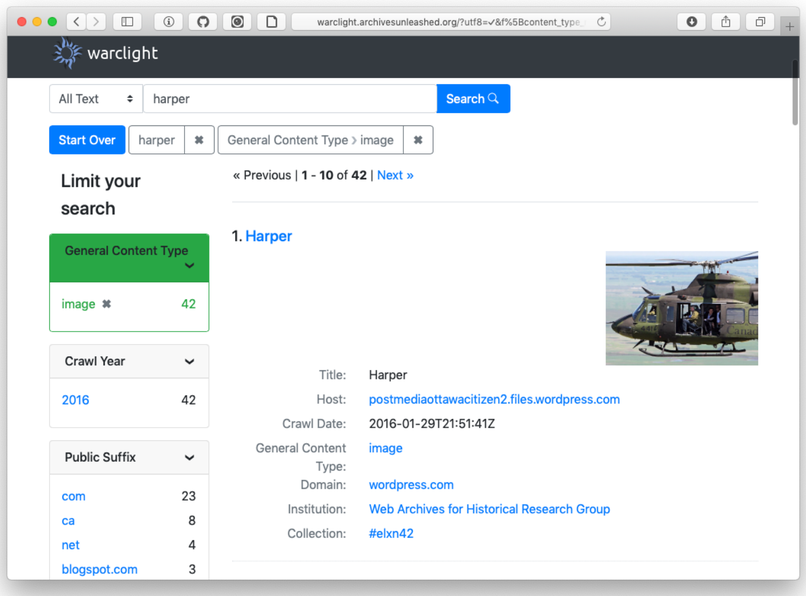
- Inspiration from **Shneiderman's mantra**: "overview first, zoom and filter, then details-on-demand." [1]
- Interface Design Requirements:
  - ✓ Provides context and devices for navigating the "information space";
  - ✓ Is transparent – every system action should be available for inspection and manipulation by the scholar; and
  - ✓ Incorporates mature and community accepted systems.
- **Faceted search and browsing meets requirement.**
- The UK Web Archive's Shine interface is an inspiration,[2] but difficult to implement in a library.
- **We thus thought to build our tool on top of Project Blacklight**[3]**, which enjoys widespread deployment and a vibrant community.**

Implementation





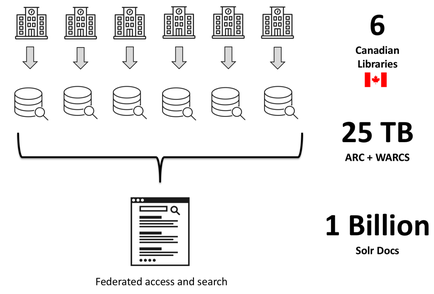




Warclight Features

- Landing page with overview of web archive content;
- Nine facets for discovery;
- Search entire full text of collection or specific fields;
- Collection and institution names added as facets to WebArchive-Discovery to work with Archive-It collections;
- Faceted search and browsing brings user to record view; and
- Each record has metadata displayed in interface and available as JSON object.

Scaling & Testing



- **Goal**: To provide federated access to Canadian web archival collections, combining holdings of six Canadian libraries.
- As indexes were too large for standalone Solr installation, we built an instance of Solr's distributed architecture: SolrCloud.
- Each partner institution's web archives was indexed into its own Solr collection.
- Provided a unified federated search experience over partner institution holdings with the SolrCloud Collections API.

SolrCloud appears to be a viable and relatively painless solution to scale out web archiving infrastructure.

Conclusions

- There is room in the web archive community for a standard discovery platform; we present Warclight as a starting point.
- Tapping into the existing digital libraries community provides us a solid foundation, but now **we are looking for a community who can drive future developments with use, feedback, and feature requests.**

References & Acknowledgments

[1] B. Shneiderman. 1996. The Eyes Have It: A Task by Data Type Taxonomy for Information Visualizations. In IEEE Symposium on Visual Languages.
[2] A. Jackson, J. Lin, I. Milligan, and N. Ruest. 2016. Desiderata for Exploratory Search Interfaces to Web Archives in Support of Scholarly Activities. In JCDL. 103–106.
[3] E. Sadler. 2009. Project Blacklight: A Next Generation Library Catalog at a First Generation University. Library Hi Tech 27, 1 (2009), 57–67.

THE ANDREW W. MELLON FOUNDATION
Social Sciences and Humanities Research Council of Canada
Conseil de recherches en sciences humaines du Canada
Canada
UNIVERSITY OF WATERLOO
YORK UNIVERSITÉ UNIVERSITY