# Warclight: A Rails Engine for Web Archive Discovery

Nick Ruest,[1] Ian Milligan,[2] and Jimmy Lin[3]
[1] York University Libraries
[2] Department of History, University of Waterloo
[3] David R. Cheriton School of Computer Science, University of Waterloo

## ABSTRACT

This paper describes the development of Warclight, a portmanteau of the open-source Blacklight platform and the ISO-standard Web ARChive file format. Warclight allows users to explore web archives that have been indexed into Apache Solr using the UK Web Archive's Web Archive Discovery tool. Referencing previous work, we explain how the standard search engine results page is inadequate to support scholarly inquiries. Instead, Warclight provides full-text and faceted search, as well as faceted browsing, to enable exploration and discovery. Given the large sizes of many web archives, we share experiences with deploying our tool at scale using a federated architecture.

## 1 INTRODUCTION

Since 1996, the Internet Archive, various national libraries, and other organizations have been systematically crawling the web and building web archives. The most common access method to these collections is by URL and time, typically through a so-called "Wayback Machine": a user can view a particular version of a web page, move forward and backward in time to examine different versions, and follow links to contemporaneous pages. Of course, this mode of access is only useful if one already knows the exact URL of the desired content. Since this is often not the case, full-text search is an obvious solution and an often-requested feature by users. Although a number of interfaces do support this capability today, such deployments are by no means widespread. The challenges are more than technical, although provisioning sufficient computing resources to support full-text search is an ever-present obstacle as web archives continue to grow in size.

In previous work, we have argued that the obvious analogy of "Google for web archives" is actually a problematic notion [2]. In particular, the standard search engine results page (SERP), commonly dubbed "ten blue links", is inadequate to support scholarly inquiries. A Google-like search interface presupposes that the scholar has a well-formulated information need, which is often not the case. What scholars need, we argued, are exploratory interfaces that support different methods of discovery in web archives. In concrete terms, we advocated that systems be designed according to Shneiderman's mantra for visual information seeking: "overview first, zoom and filter, then details-on-demand" [5]. This means that instead of directly drilling down to search results, interfaces should provide context and devices for navigating the "information space". To Shneiderman's mantra, we proposed an addendum: "make everything transparent". In particular, tools in support of scholarship should not have "magic"; every system action should be available for inspection and manipulation by the scholar.

Of existing mature search techniques, we argued that faceted search and browsing capabilities were the closest to meeting these requirements. In 2016, we presented a concrete realization of this vision, based on the UK Web Archive's Shine interface [2]. Our prototype provided faceted search and trends visualization, allowing scholars to explore large web collections.

Several years later now, however, we've discovered several shortcomings with our Shine prototype. From the technical perspective, it was developed using the Play framework,[1] which does not have widespread support within the broader community. Thus, organizations were hesitant to deploy our prototype on their own collections. Without a virtuous cycle of broad support feeding widespread adoption that in turn attracts developer attention, enthusiasm for our Shine prototype faded.

As part of our long-term efforts to promote scholarly access to web archives [1, 3], Warclight, in essence, represents our second attempt at building a discovery tool for web archives. This time, we prioritized community and sustainability, deciding to build our tool on top of Project Blacklight, which already enjoys widespread deployment and a vibrant community.

## 2 IMPLEMENTATION

Apache Solr is a popular and mature search platform that powers search on many heavily-trafficked websites on the internet, including eBay, Comcast, Netflix, Disney, Bloomberg, as well as the Internet Archive. It makes sense to also build a discovery tool for web archives on this solid foundation. Warclight is designed to work with web archive content that has been indexed using the UK Web Archive's Web Archive Discovery tool,[2] which parses WARCs and ARCs (standard container formats for web archives) to build Apache Solr indexes using best practices and a set of fields agreed upon by the community. This in turn exposes a variety of searchable fields, including title, body, host, crawl date, and content type. The key here is that Web Archive Discovery provides a common Solr index schema that Warclight can build on, thereby allowing the community to coalesce around a shared standard.

Being a search platform, however, Solr does not prescribe a specific frontend interface to access its capabilities. This is where Project Blacklight [4] comes in: Blacklight is a Rails engine that provides a "basic discovery interface for searching an Apache Solr index".[3] Within the digital libraries community (which we note is larger than the web archiving community), this open-source project has gained significant traction. A glance at the project homepage will reveal pointers to scores of deployments, and browsing the project's code repositories will reveal an active and vibrant developer community. Warclight, being a portmanteau of Blacklight and WARC, aims to bring the former's rich set of features as well as its

[1] https://www.playframework.com/
[2] https://github.com/ukwa/webarchive-discovery
[3] https://github.com/projectblacklight/blacklight/wiki

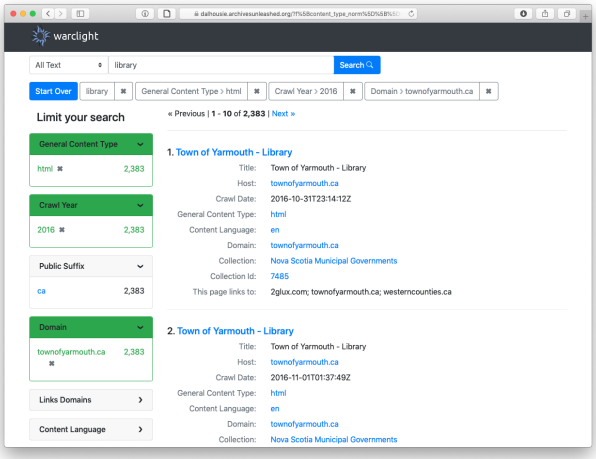


**Figure 1: The Warclight discovery interface, showing a search for Yarmouth, Nova Scotia libraries.**

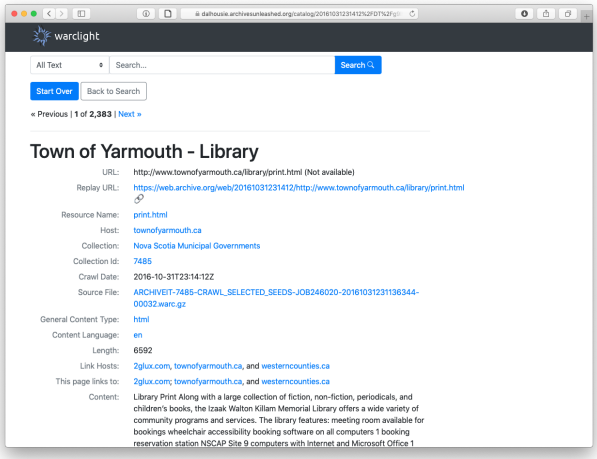


**Figure 2: Detailed record view for the Yarmouth Public Library website from 2016.**

sizeable community to web archiving. We are reminded of an old familiar adage: "If you want to go fast, go alone. If you want to go further, go together".

## 3 FEATURES

The starting point of a Warclight application is a landing page with introductory text and facets from an empty search. This also enables a user to begin with an overview of the contents of their web archive before drilling into detail. Default facets include content type (html, text, image, etc.), crawl year, top-level domain, domain, links that are present on the page, language, and resource name. To work with Internet Archive's Archive-It collections, we have also added the collection and institution names. A user can search the entire full text of the collection or specific fields (title, URL, and host). As an example, in Figure 1 we see that the user has selected 'html' documents, from 2016, from the `townofyarmouth.ca` domain, containing the keyword 'library'; note the facets running down the left hand side of the interface.

Faceted search and browsing will ultimately take the user to a record view (see Figure 2). The interface provides a link to the original URL if it is still available; otherwise, we provide an option to query the Memento TimeTravel API for a replay URL based on the crawled URL and the crawl date. Each record view provides numerous details, both displayed in the interface and also available as a JSON object that can be programmatically queried. We envision that some scholars might manually interact with our interface, while others might write scripts to manipulate web content.

## 4 SCALING

A major test of Warclight was our effort to provide federated access to Canadian web archival collections, bringing together the holdings of six Canadian libraries with federated search. With support from Compute Canada's Research Portals and Platforms funding stream, we were able to index over 25 TB of ARCs and WARCs from our partner institutions, resulting in over one billion Solr docs. Since the scale of the indexes were too large for a standalone Solr installation, we built an instance of Solr's distributed architecture, called SolrCloud. The infrastructure consists of four virtual servers: three SolrCloud nodes and one web server running nine production Warclight applications. Each partner institution's web archives were indexed into its own Solr collection, but we provide a unified federated search experience over our partner institutions' holdings with the SolrCloud Collections API. Based on our experiences thus far, SolrCloud appears to be a viable and relatively painless solution to scale out web archiving infrastructure.

## 5 NEXT STEPS

Our first version of Warclight is ready for use and an example site can be found at https://warclight.archivesunleashed.org/.

The next step, however, speaks to the importance of community that we articulated in the introduction. We believe that there is room in the web archive community for a standard discovery platform, and present Warclight as a starting point. Tapping into the existing digital libraries community provides us a solid foundation, but even more critical—we are looking for a community of *scholars* who can drive future developments with their use, feedback, and feature requests. Will we find it?

**Acknowledgments.** This work was primarily supported by the Andrew W. Mellon Foundation and Compute Canada's Research Platforms and Portals program. Additional funding for the project has come from Start Smart Labs and the Social Sciences and Humanities Research Council of Canada.

## REFERENCES

[1] R. Deschamps, S. Fritz, J. Lin, I. Milligan, and N. Ruest. 2019. The Cost of a WARC: Analyzing Web Archives in the Cloud. In *JCDL*.
[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] I. Milligan, N. Casemajor, S. Fritz, J. Lin, N. Ruest, M. Weber, and N. Worby. 2019. Building Community and Tools for Analyzing Web Archives through Datathons. In *JCDL*.
[4] E. Sadler. 2009. Project Blacklight: a next generation library catalog at a first generation university. *Library Hi Tech* 27, 1 (2009), 57–67.
[5] B. Shneiderman. 1996. The Eyes Have It: A Task by Data Type Taxonomy for Information Visualizations. In *IEEE Symposium on Visual Languages*.