Exploring and Discovering Archive-It Collections with Warcbase

1. Introduction

Big Data is reshaping the historical profession in ways we are only now beginning to grasp. The growth of digital sources since the advent of the World Wide Web in 1990-91 presents new opportunities for social and cultural historians. Large web archives contain billions of webpages, from personal homepages to professional or academic websites, and now make it possible for us to develop large-scale reconstructions of the recent web. Yet the sheer number of these sources presents significant challenges: if the norm until the digital era was to have human information vanish, "now expectations have inverted. Everything may be recorded and preserved, at least potentially" (Gleick, 2012).

While the Internet Archive makes archived web content available to the general public and mainstream scholarly community through its "Wayback Machine," (at http://archive.org/web) which allows visitors to enter a Uniform Resource Locator (URL) to visit archived web versions of a particular page, this system is limited: not only do visitors need to know the URL in the first place, but they are limited to individual readings of single webpages.

By unlocking the Wayback Machine's underlying system of specialized files, primarily WebARChive (ARC and WARC) files, we can develop new ways to systematically track, visualize, and analyze change occurring over time within web archives. Warcbase, an open-source platform for managing web archives built on Hadoop and HBase, provides a flexible data model for storing and managing raw content as well as metadata and extracted knowledge. Tight integration with Hadoop provides powerful tools for analytics and data processing. Using a case study of one collection, this paper introduces the work that we have been doing to facilitate web archive access with warcbase.

2. Project Rationale and Case Study

In 1996, the Internet Archive launched a complementary research services company, Archive-It, which offers subscription-based web archiving to collecting institutions.

The University of Toronto Library (UTL) began collecting a quarterly crawl in 2005 of Canadian political parties and political interest groups (the collections were separate in 2005, merging in 2006) (University of Toronto, 2015). The collection itself has a murky history: UTL had been part of a broader project that would have collected political websites. It fell through, but UTL opted to carry out their crawl on their own and the librarian was responsible for selecting the seed list herself (faculty and other librarians did not respond for calls for engagement). While formal political parties are robustly covered, the "political interest groups" collection was a bit more nebulous: sites were discovered through keyword searches, and some were excluded due to robots.txt exclusion requests. Beyond this brief sketch, we have little information about the decisions made in 2005 to create this collection. This lack of documentation is a

shortcoming of this collection model, as if a historian was to use this material in a peer-reviewed paper, questions would be raised about its representativeness.

If a user wants to use the Canadian Political Parties and Interest Groups Collection (CPP) through Archive-It today, they visit the collection page at https://archive-it.org/collections/227 and enter full-text search queries. In August 2015, our group also launched http://webarchives.ca, based on the British Library's SHINE front end for web archives; this was a way to facilitate a different form of more casual user access, aimed at the general public (we discuss this in a separate paper).

The Archive-It portal is limited. There are no readily-available metrics of how many pages have been collected, how they break down by domain and date, and the portal undoubtedly provides skewed results unless the search phrase is dramatically narrowed down.

Consider the search for "Stephen Harper," Canada's Prime Minister between 2006 and 2015 in Figure 1.

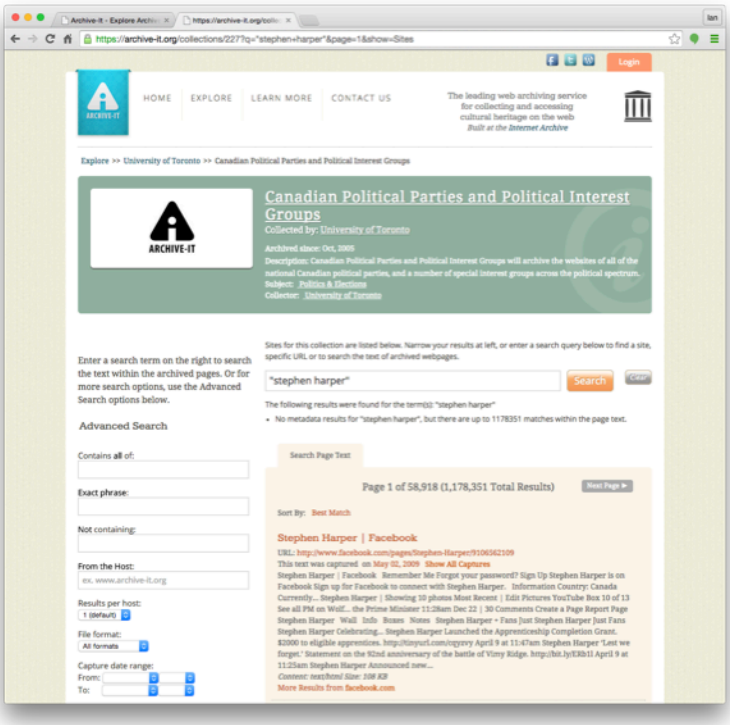

Figure 1: Archive-It Search Portal

The results are decent: Harper's Facebook page from 2009, a Twitter snapshot from 2010, and some long-form journalism articles and opposition press releases. But amidst the 1,178,351 results, there is no indication as to how the ranking took place, what facets are available, and how things may have changed over the last ten years of the crawl.

The data is there, but the problem is access.

3. Warcbase: A Platform for Web Archive Analysis

Warcbase is a web archive platform, not a single program. Its capabilities comprise two main categories:

1. Analysis of web archives using the Pig or Spark programming languages, and assorted helper scripts and utilities
2. Web archive database management, with support for the HBase distributed data store, and OpenWayback integration providing a friendly web interface to view stored websites

One can take advantage of the analysis tools (1) without bothering with the database management aspect of Warcbase – in fact, most digital humanities researchers will probably find the former more useful. This paper focuses on the former capabilities, showing how we can use the warcbase platform to carry out text and network analyses.

4. Using Warcbase on Web Archival Collections: Text Analysis

We have begun to document all warcbase commands on a GitHub wiki, found at https://github.com/lintool/warcbase/wiki. We begin with installation instructions, and then provide simple scripts written in Apache Pig or Apache Spark to run the commands.

While possible to generate a plain text version of the entire collection, a more fruitful approach has been to generate date-ordered text for particular domains. If a researcher is interested in say, the Green Party of Canada's evolution between 2005 and 2015, they can extract the plain text for greenparty.ca by running the following script:

```
import org.warcbase.spark.matchbox.ArcRecords
import org.warcbase.spark.matchbox.ArcRecords._

val r = ArcRecords.load("/path/to/input", sc)
  .keepMimeTypes(Set("text/html"))
  .discardDate(null)
  .keepDomains(Set("greenparty.ca"))
  .extractCrawldateDomainUrlBody()
r.saveAsTextFile("/path/to/output/")
```

All they would need to change would be the `path/to/input` to the directory with their web archive files, the `path/to/output` for where they want to save the resulting plain-text files, and the `greenparty.ca` value to whatever domain they might be interested in researching.

They then receive a date-ordered output of all plain text for that domain (as per the `extractCrawldateDomainUrlBody` command). It can then be sorted and used in other research avenues. For example, this plain text could be loaded into a text analysis suite such as http://voyant-tools.org/ or other digital humanities environments.

We have also been experimenting with other visualizations based on the extracted plain text. Computationally intensive textual analysis can be carried out using warcbase itself. Using the Stanford NER package in parallel, we have a script that extracts entities, counts

them, and then visualizes them using D3.js to help see overall changes in a web archival collection. The instructions can be found at https://github.com/lintool/warcbase/wiki/Pig:-Named-Entity-Recognition-(on-a-cluster). Figure 2 below shows the output of the NER visualizer.

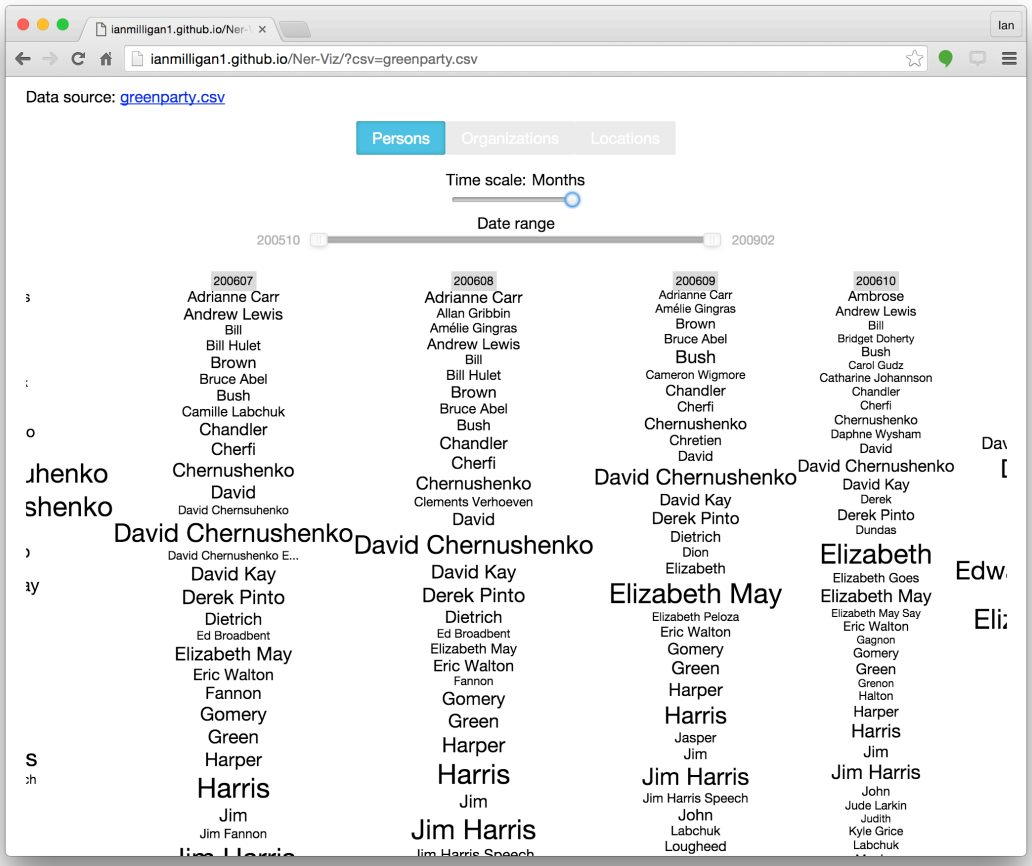

Figure 2: Named Entity Visualization within Warcbase

Finally, another text approach is topic modelling (Blei et al., 2003). LDA works by finding topics in unstructured text. To visualize topic models, we elected to use the Termite Data Server, which is a visual analysis tool for exploring the output of statistical topic models ("uwdata/termite-data-server," n.d.). As Figure 3 demonstrates, the visualization allows you to get a top-down view at the topics found in a web archive.

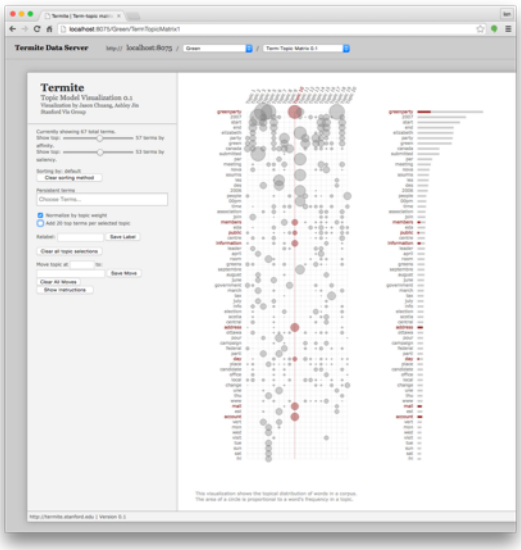

Figure 3: Termite Topic Model

Warcbase presents versatile opportunities to extract plain text and move it into other environments for analysis. Unlike the keyword-based Archive-It portal, we now have data that can be inquired in many fruitful ways.

5. Using Warcbase on Web Archival Collections: Hyperlink Analysis

Warcbase can also extract hyperlinks. While text can be very important, these sorts of metadata can often be more important: allowing us to see changes in how groups link to each other, what articles and issues were important, and how relationships changed over time.

Consider Figure 4, which visualizes the links stemming from and between the websites of Canada's three main political parties.

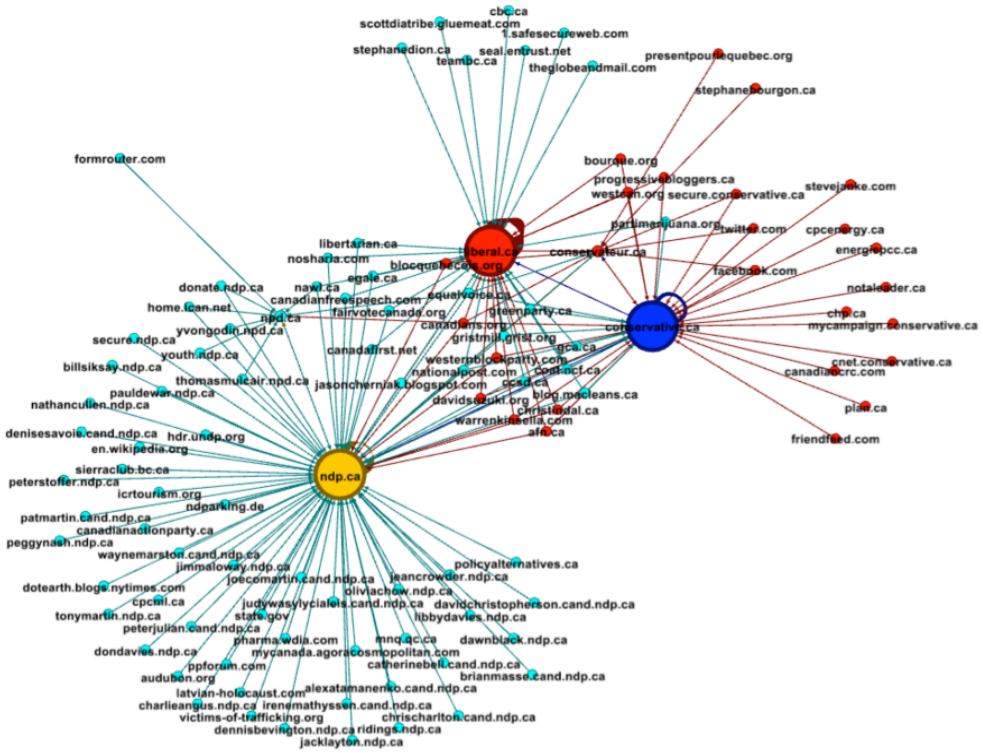

Figure 4: Three major political parties in Canada

Above, we can see which pages only link to the left-leaning New Democratic Party (ndp.ca), those that link only to the centrist Liberals (liberal.ca) in the top, and those that only connect to and from the right-wing Conservative Party at right. We can use it to find further information, such as in Figure 5.

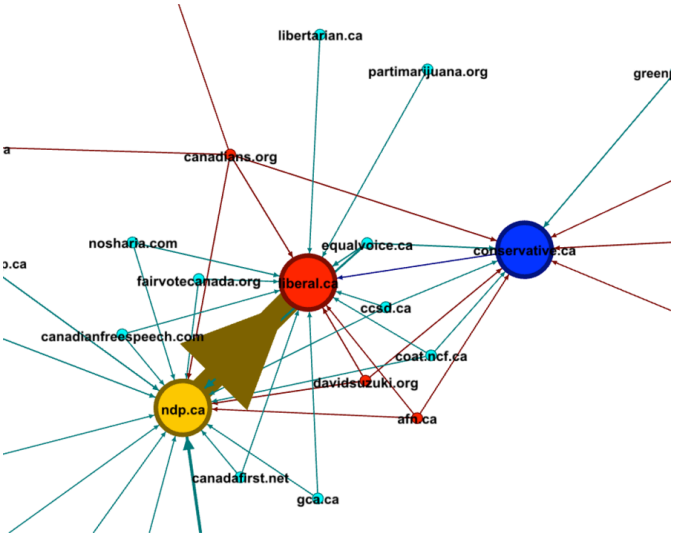

Figure 5: NDP attack

The above links are from the 2006 Canadian federal election. The Liberal Party was then in power and was under attack by both the opposition parties. In particular, the left-

leaning NDP linked hundreds of times to their ideologically close cousins, the centrist Liberals, as part of their electoral attacks, ignoring the right-leaning Conservative Party in the process. Link metadata illuminates more than a close reading of an individual website would. It contextualizes and tells stories itself.

While we have traditionally used Gephi to do analysis, importing material into Gephi from warcbase required many manual steps as documented at https://github.com/lintool/warcbase/wiki/Gephi:-Converting-Site-Link-Structure-into-Dynamic-Visualization. We have been prototyping a link analysis visualization in D3.js, which can run in browser (Figure 6).

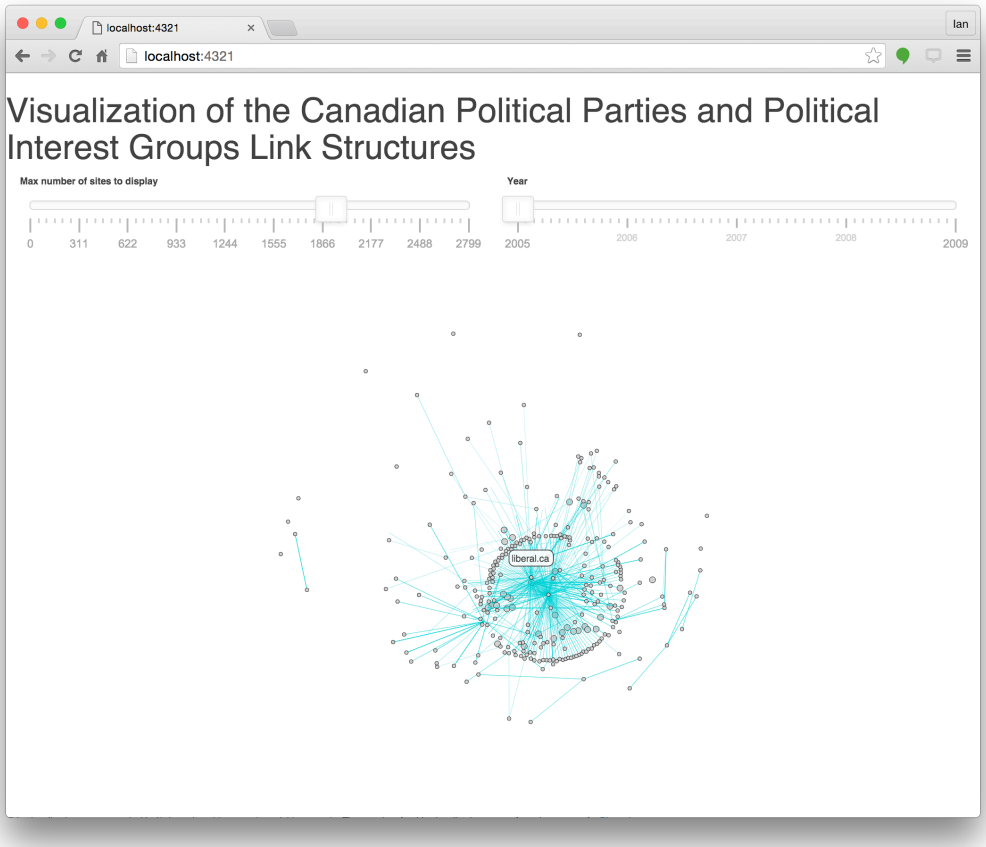

Figure 6: Link Visualization

6. Conclusions

With the increasingly widespread availability of large web archives, historians and Internet scholars are now in a position to find new ways to track, explore, and visualize changes that have taken place within the first two decades of the Web. Warcbase will allow them to do so. This project is among the first attempts to harness data in ways that will enable present and future historians to usefully access, interpret, and curate the masses of born-digital primary sources that document our recent past.

References

Blei, D.M., Ng, A.Y., Jordan, Michael I., 2003. Latent Dirichlet Allocation. J. Mach. Learn. Res. 3, 993–1022.

Brügger, Niels. "The Archived Website and Website Philology: A New Type of Historical Document." Nordicom Review 29.2 (2008): 155–75.

Brügger, Niels, and Niels Ole Finnemann. "The Web and Digital Humanities: Theoretical and Methodological Concerns." Journal of Broadcasting & Electronic Media 57, no. 1 (2013): 66–80.

Gleick, James. *The Information: A History, a Theory, a Flood* (London: Vintage, 2012).

Lin, Jimmy, Milad Gholami, and Jinfeng Rao. "Infrastructure for Supporting Exploration and Discovery in Web Archives." In Proceedings of the 23rd International Conference on World Wide Web, 851–56. WWW '14 Companion. Republic and Canton of Geneva, Switzerland: International World Wide Web Conferences Steering Committee, 2014. doi:10.1145/2567948.2579045.

University of Toronto, 2015. Archive-It - Canadian Political Parties and Political Interest Groups [WWW Document]. URL https://archive-it.org/collections/227 (accessed 7.24.15).

uwdata/termite-data-server [WWW Document], n.d. . GitHub. URL https://github.com/uwdata/termite-data-server (accessed 7.24.15).