Lost in the Infinite Archive: The Promise and Pitfalls of Web Archives

The Web is having a dramatic impact on how we research and understand the recent past. Historians, who have long laboured under conditions of source scarcity – we wish we had more information about the past, but it was not recorded or preserved – are now confronted with primary sources on a scale that defies both conventional methodologies and standard computational methods.[1] Web archives offer profound promise. Take a comparative example. The Old Bailey Online describes its holdings of 197,745 trials between 1674 and 1913 as the 'largest body of texts detailing the lives of non-elite people ever published'.[2] The web archive of GeoCities, a platform for web publishing that operated from the mid-1990s to the early 2000s, amounts to over 38 million pages. Eventually, historians will have access to billions of such sources written by people of various classes, genders, ethnicities, and ages. While the World Wide Web is not a perfect democracy, by any means and any of the categories listed above, it still represents a massive shift. As a result, web archives exemplify this conundrum and represent challenge as well as opportunity.

What information do we want to access? How was the information collected? How do national boundaries intersect with the realm of the Internet? What are the implications of working with such large archives, collected without the informed consent or even knowledge of the overwhelming majority of contributors? These are pressing concerns. For the most part, historians cannot write histories of the 1990s unless they use web archives: with them, military historians will have access to the voices of rank-and-file soldiers on discussion boards; political historians, to blogs, the cut and thrust of

websites, electoral commentary and beyond; and of course, social and cultural historians, to the voices of the people on a scale never before possible.

The stakes are high. If we do not come to grasps with web archives, the histories that we write will be fundamentally flawed. Imagine a history of the late 1990s or early 2000s that draws primarily on print newspapers, ignoring the revolution in communications technology that fundamentally affected how people share, interact, and leave historical traces behind. Yet even as we use web archives, we need to be cognizant of their functionalities, strengths, and weaknesses: we need to begin to theorize and educate ourselves about them, just as historians have been cognizant of analog archives since the cultural turn. As new discovery methods for finding information in web archives begin to appear, historians need to be ready to participate; otherwise we might not know why one particular response is number one, versus number one million.

The sheer amount of social, cultural, and political information generated and presented almost every day within the web archive since the Internet Archive began collecting in 1996 represents a complex data set that will fundamentally reshape the historical profession. We need to be ready.

On Complex Data Sets: Three Different Examples

This is not an abstract concern: the history of the 1990s will be written soon. While there is no common rule for when a topic becomes 'history,' it took less than 30 years after the tumultuous year of 1968 for a varied, developed, and contentious North American historiography to appear on the topic of life in the 1960s.[3] Carrying out 'recent histories,' be they of the 1970s or of events only a few years ago, brings with them a host of

methodological issues from a lack of historiography, historical participants who can 'talk back,' and issues of copyright and privacy.[4] The year 2021 will mark the 30th anniversary of the creation of the first publicly accessible website. Just as media, government, and business radically transformed their practices in the 1990s, historians must do so as well to analyze this information. 'New media' is not that new anymore.

Historians run very real risks if they are not prepared. Currently, the main way to access the archived Web is through the Wayback Machine, most notably associated with the Internet Archive. The Internet Archive emerged out of a concern around a 'digital dark age' in the mid-1990s, where rapid technological evolution led to fears around whether our heritage was being preserved. Responding to this, Internet entrepreneur Brewster Kahle founded the Internet Archive in June 1996, which began to rapidly grow their web archive collection. They did so by sending 'web crawlers,' automated software programs, out into the Web to download webpages that they found. This crawling process meant that depending on how the Web developed and the limits placed on a crawler, the crawler could indefinitely collect – generating an infinite archive.[5]

While the Internet Archive was collecting data from 1996 onwards, the next step was to make it accessible to researchers. In 2001, they launched the still-dominant form of interacting with web archives: the Wayback Machine. You can try it yourself at http://archive.org/web. It is limited. A user needs to know the exact Uniform Resource Locator (URL) that they are looking for: a website like http://www.geocities.com/enchantedforest/1008/index.html, for example. The page is then retrieved from the web archive and displayed. If you know the URL of the page you

are interested in, and only want to read a few, the Wayback Machine works by generating facsimiles of those pages. They are not perfect, as they may not collect embedded images, or might grab them at slightly different times (to avoid overloading any single server, the crawler might download the text of a website and then the image a few hours or even days later; this can lead to the storing of websites that never existed in the first place).[6] Beyond technical issues, it is difficult to find documents with the Wayback Machine unless you know the URL that you want to view.

This latter shortcoming disqualifies it as a serious research tool unless it is paired with a search engine of some kind. Historians are used to full-text search interfaces. However, imagine conducting research through date-ordered keyword search results, carried out on billions of sites. It would produce an outcome similar to the current methods by which historians search digitized newspapers.[7] In the absence of contextual information about the results found, they can be useless. It is possible to find almost anything you want within 38 million web pages. I can find evidence on any matter of topics that advances one particular argument or interpretation. Without the contextual information provided by the archive itself, we can be misled.

Three case studies can help us better understand the questions, possibilities, and challenges facing historians as we enter this archival territory. The first is the Wide Web Scrape, a compilation of billions of objects collected by the Internet Archive between 9 March and 23 December 2011. Next, I explore work that I have been doing with a collection of political websites created between 2005 and 2015. Finally, I explore the GeoCities end-of-life torrent, to get at the heart of ethical challenges.

Together, these studies suggest a path forward for historians. Those of us who use web archives do not need to become programmers, but do need to become aware of basic Web concepts: an understanding of what metadata is, how the Web works, what a hyperlink is, and basic definitional concepts such as URLs. Beyond this, however, is the crucial dimension of algorithmic awareness. When we query archives, we need to know why some results are coming to the top and others at the bottom. If we turn our research over to black boxes, the results that come from them can reaffirm biases: websites belonging to the powerful, for example, rather than the marginalized voices we might want to explore and consider. The decisions that we as historians make now will have profound effects as tools begin to be developed to access web archives.

Data is Bigger Than the Nation: The Wide Web Scrape

As a data set, the Wide Web Scrape is exhaustive, transcending national borders. The 2,713,676,341 item captures – websites, images, PDFs, Microsoft Word documents, and so forth – are stored across 85,570 WebARChive (WARC) files.[8] The WARC file format, which is certified by the International Standards Organization, preserves web-archived information in a concatenated form.[9] Generated by the Internet Archive, these files also serve as a good introduction to the geographic challenges of web archives: historians tend towards geographic boundaries, but these archives can transcend them. WARC files are an abundant resource, but that abundance is double edged.

As a Canadian historian looking for a relatively circumscribed corpus, I decided to focus on the Canadian Web, or *websphere*, as best I could. The 'Canadian Web', is however, intrinsically a misnomer. The Web does not work within national boundaries. It

is a global network, transcending traditional geopolitical barriers (local fissures still appear, as seen in 'this video is not available in your country' messages).[10] The Internet Archive exploits the Web's borderless nature in their global crawling of material in a way national domain crawls by national institutions cannot. From Denmark to Britain, researchers collecting and studying national webspheres have taken different approaches. Some, such as the Danish NetLab, have confined their studies to national top-level domains (.dk).[11] Others, such as the British Library's born-digital legal deposit scheme, use algorithms and human intervention to find British sites outside of the .uk domain.

What does the data collected along the lines of a national websphere – a top-level domain such as .ca – look like? While all archival records are only as useful as the discovery tools that accompany them – a misfiled box in a conventional archive might as well not exist – the size of these collections elude traditional curation. From the holdings of the Wide Web Scrape, we examined the CDX files (akin to archival finding aids which contain information about the records found within archival boxes), and which can be measured in gigabytes rather than terabytes. They contain millions of lines of text like:

```
ca,yorku,justlabour)/ 20110714073726 http://www.justlabour.yorku.ca/
text/html 302 3I42H3S6NNFQ2MSVX7XZKYAYSCX5QBYJ
http://www.justlabour.yorku.ca/index.php?page=toc&volume=16 - 462
880654831 WIDE-20110714062831-crawl416/WIDE-20110714070859-02373.warc.gz
```

From this, we can learn a few things: in this case, we learn that the record is justlabour.yorku.ca, collected on 14 July 2011 at 7:37 GMT. It redirected (HTML code 302) to the table of contents for volume 16. If you visit justlabour.yorku.ca today, you'll be redirected to a more recent issue. CDX files help us find specific records. Accordingly, I used them to download a sample of 622,365 .ca URLs.

Working with this data set was an interesting window into the choices historians need to make when they work with large data sets from the Web. Derived data – plain text, named entities (discussed later), extracted links, hyperlinks with anchor text – can be useful. Yet at every stage they present historians with questions. Some extracted hyperlinks will be relative – that is, /destination.html rather than http://www.history.ca/destination.html. Should they be reclassified if we want to make a chart of all the hyperlinks connecting different websites, and at what stage? To create plain text files, we use the warcbase platform.[12] I was able to run textual analysis, extract location data, postal codes, and names of people, and explore the topics people were discussing. This method had the downside, however, of removing images, backgrounds, and layouts, meaning that text is taken out of context. While the size of the data sets under discussion mitigates this to some extent, we are still profoundly altering sources.

There were three promising ways to query this data, each of which sheds light on various web archival challenges: keywords, named entity recognition (which finds entities like locations and names within text), and hyperlink structures. To search a large body of material with keywords, the Apache Solr search engine is ideal. It can index material and respond to queries from a number of front-ends that can run locally on a computer.[13] The United Kingdom's Web Archive, for example, uses a custom front-end Solr portal that provides full-text search access to their collections.[14] One view prompts you to enter a query, and to then subsequently see the relative frequency of that term rise and fall over time (how often was the word 'nationalize' used in 2006, for example, compared to 2012). With specific queries, this search approach works well. Yet on a broad scale, when looking for cultural trends, more context is necessary.

The most promising keyword approach to my data set was clustering, which takes a set of documents and groups them. If a web collection contained websites about cats, dogs, and pigs, the algorithm might cluster the cat sites together. Conversely, it might find another characteristic – the ages of the authors, perhaps – and cluster them that way. There are several different algorithms to choose from, although in my experience the Lingo clustering algorithm provides the best results (See Fig. 1).[15]

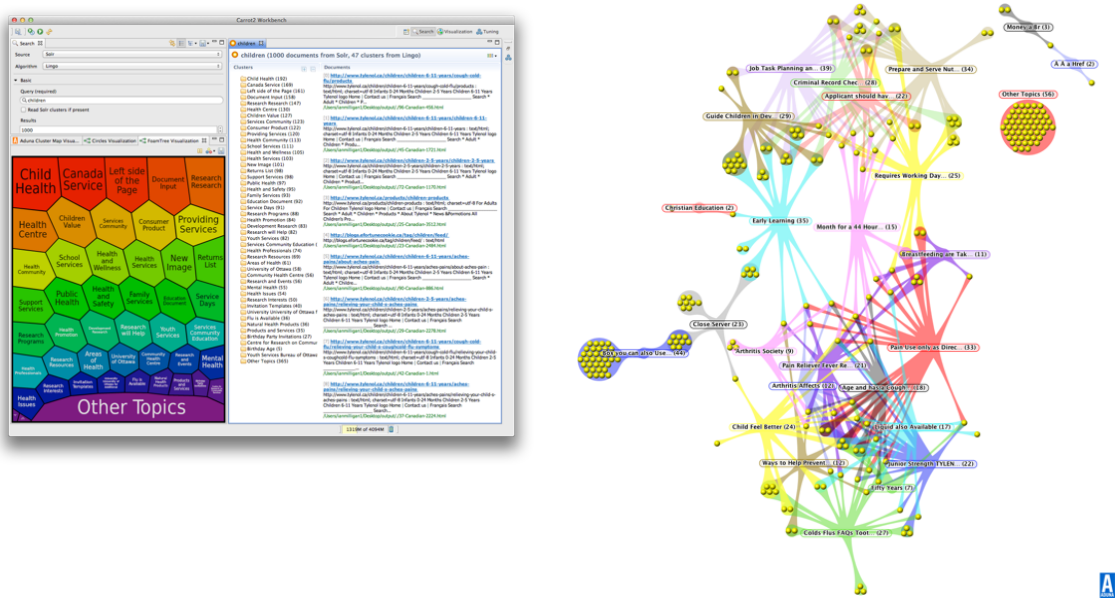

Fig. 1: Carrot2 clustering workbench results

The free Carrot2 front end (http://project.carrot2.org/), which interfaces easily with a Solr database, is the most useful. From a query for 'children', we see that this sample of 622,365 websites contains pages relating to child health, health centres, service providers, public health, educational services, and consumer products such as Tylenol. Clicking on the graphical representation brings the user to a list of documents, and another click brings up an individual document. The image on the right is the graphical representation of overlapping clusters, such as the simplified Fig. 2:

Fig. 2: Example of connected clusters

If a dot is connected to two clusters, it belongs to both. These connections can provide a rough sense of how representative things are: there are many websites about breastfeeding, for example, but not many about Christian early childhood education institutions. More importantly, it is possible to isolate a corpus to study. Used jointly, the Solr database and Carrot2 front end help transcend the Wayback Machine's limitations.

The main drawback with this approach is the need to know what you are looking for. Extracting commonly mentioned locations can be fruitful, as in Fig. 3:

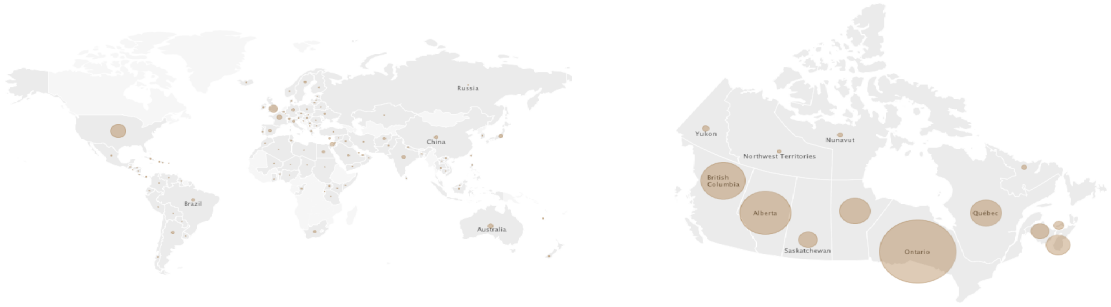

Fig. 3: Countries (other than Canada) mentioned in .ca top-level domain sample (left); Canadian provinces mentioned (right)

Extracted using a combination of Stanford Named Entity Recognition (NER), Google Maps API, and verification by student research assistants, this process found location names – for example, 'Toronto' or 'Johannesburg' – and geolocated them by assigning coordinates. While longitudinal data will be more useful, allowing us to see how various locations changed over time, at this point we can see the attention paid towards Canadian

trading partners and the complete absence of attention towards sub-Saharan Africa. Within Canada, Québec is overrepresented vis-à-vis the province of Ontario.

Web-wide scrapes represent the dream of social history: a massive documentary record of the lives of everyday people, their personal websites, small businesses, labour unions, community groups, and so forth. Yet the value of this information is balanced by the sheer size and complexity of these data sets. Web-wide scrapes represent one extreme of what we can do with web archives: exploring a massive record of human activity, collected on a previously unimaginable scale.

Archive-It Political Collections: An Ideal Size?

Web-wide scrapes are time consuming and expensive to work with. Recognizing this, web archivists have begun to move towards more accessible data sets that bridge the gap between the lightweight CDX file and the heavy-duty WARC file (both of which we have seen in the preceding section). In this section, I argue that while our first inclination, as with the Wide Web Scrape, might be to go right to the content, more fruitful historical information can be found within the metadata.

Archive-It, a web archiving subscription service provided by the Internet Archive for universities and other institutions, recently piloted their research services portal. It provides access to Web Archive Transformation, or WAT, files: a happy medium between CDXs and WARCs. These provide rich metadata: everything that a CDX has, plus metatext about the website, the title, and the links and anchor text from each site. They are essentially the WARCs sans content, making them much smaller.

Beginning a decade ago, the University of Toronto Library (UTL) has put together thematic web collections with Archive-It. One of their major collections is about Canadian political parties and political interest groups, collected quarterly since 2005. Canada has seen pivotal changes within its political sphere over the last ten years, between 2005 and 2015: an arguable militarization of Canadian society, the transition from the 'natural governing party' of the centrist Liberal Party of Canada to the Conservative Party of Canada (and back in late 2015), as well as major policy changes on foreign policy, science policy, and climate change.[16] Given these critical shifts, it is surprising on one level that UTL's collection was not used more – the collection, for example, has never been cited before we began to work with it. On another level, however, it is unsurprising: the current portal to work with the collection at https://archive-it.org/collections/227 has only a very basic search function. It was only by reaching out to librarians at UTL and the Internet Archive that I was able to get the files and begin to explore what we could actually do with them. Ultimately, it became clear that metadata was just as – and in many cases more – useful than the content itself (we ended up providing access to the content through http://webarchives.ca, an implementation of the British Library's Shine frontend).

By using either the Internet Archive's web analysis workshop or warcbase, a web archiving platform, we can extract links the WAT files in this collection by domain.[17] The results look similar to the example in Table 1.

Source	Target	Weight (number of links)

Conservative.ca	Liberal.ca	10
Liberal.ca	NDP.ca	10

Table 1: Hyperlink Example

In this case, we can see that among the sub-sites that make up the Conservative Party's website there are ten links to websites within the liberal.ca domain, and vice versa. This sort of data needs to be used with caution, however: one strategic, high-profile link to a website might have more impact than lots of smaller links. For example, a single link on the front page of a political party's website has far greater impact than hundreds of links contained in the footers of biographical statements. We call this the 'weight' because it dictates how much emphasis should be put on the lines that connect various nodes.

This data can be useful on a large scale. Consider Fig. 4, which visualizes the external links stemming from and between the websites of Canada's three main political parties. Each line, or edge, represents a hyperlink between domains (or nodes).

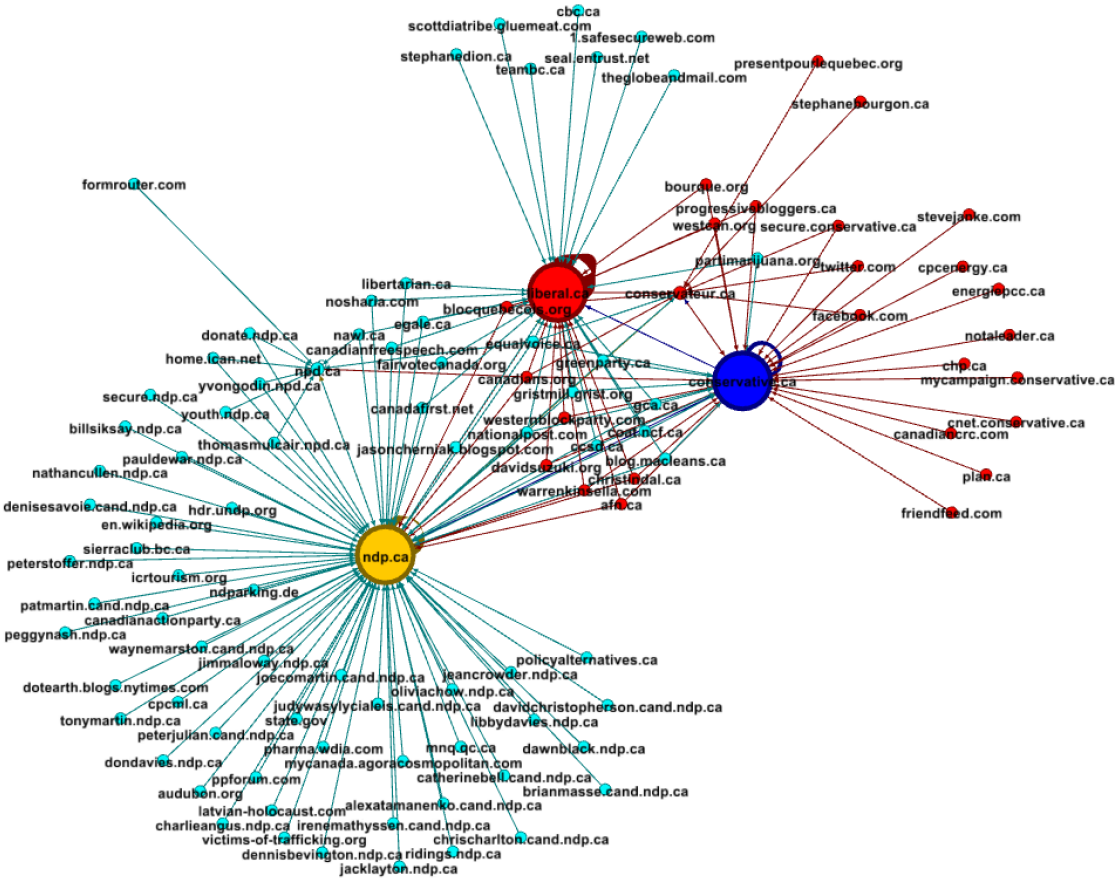

Fig. 4: Three major political parties in Canada, the NDP, Liberals, and Conservatives, 2005–2009.

Above, we can see which pages only link to the left-leaning New Democratic Party (NDP or 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. In the middle are the websites that either link to all three parties or to just two of the three (to the left and right of the Liberal node, respectively). Even from this graph we can see that while many groups link to only the Liberals and the NDP, or to the Liberals and the Conservatives, few link just to the NDP and the Conservatives.

By taking quarterly slices of the data, we can also use metadata to identify the broad contours of a narrative as in Fig. 5.

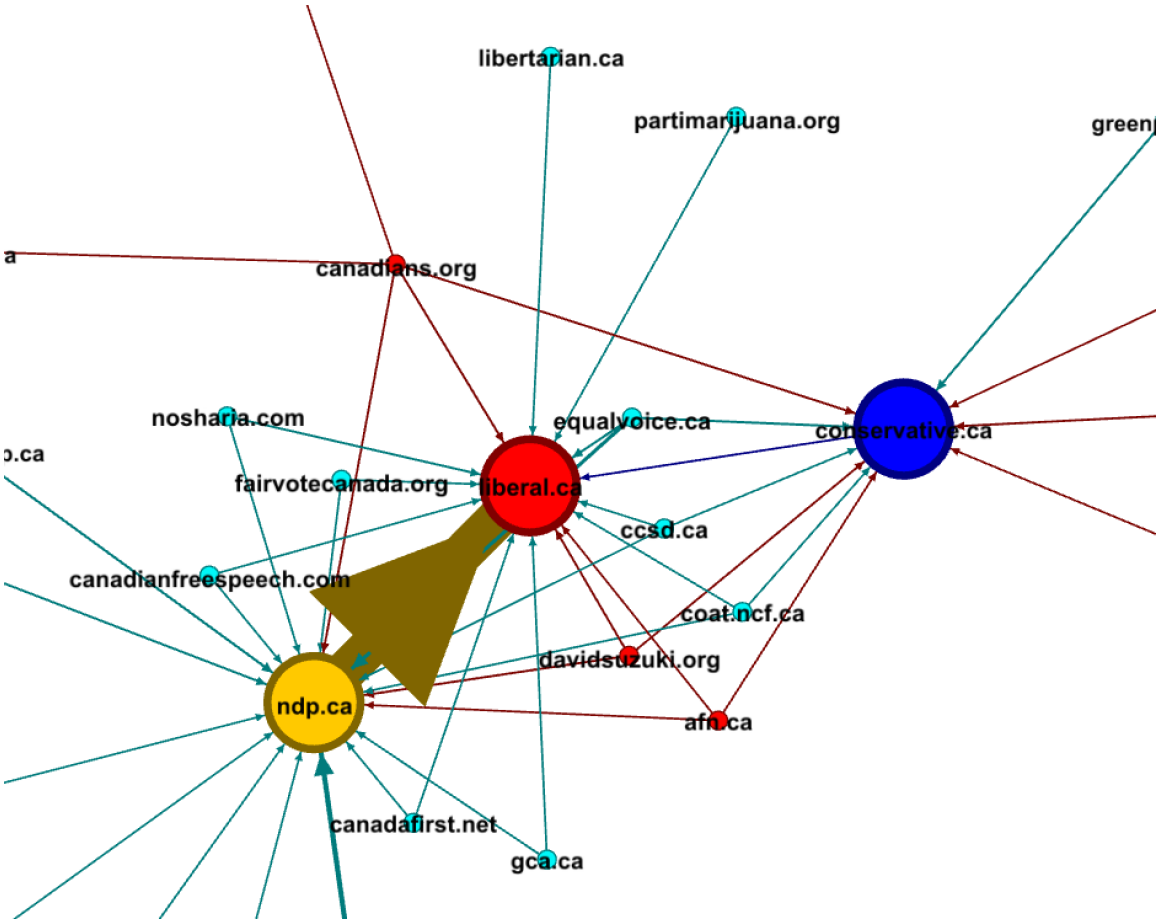

Fig. 5: Link structures during the lead-up to the 2006 federal election

We can see that several entities link to all three parties, such as the environmentalist davidsuzuki.org or the Assembly of First Nations (afn.ca), and we can also see how all of the organizations linked to each other. 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.

We can also find sections of this collection that link far more to themselves than to other parts. These divisions lend themselves well to specific extraction. Consulting the

UTL's entire collection via WARC files may be too difficult, but link analysis can tell us what to download. One experiment proved interesting. I took the two main political parties, the Liberals and Conservatives, over the period of study and (relying solely on links) found the communities that grew out of their party websites. The results were interesting: liberal.ca was in the same community as interest groups such as the National Association of Women and Law and media organizations such as *Maclean's* magazine and the Canadian Broadcasting Corporation. Most interestingly, the left-wing New Democratic Party of Canada appeared in the same community. For the Conservatives, they were grouped with many cabinet ministers' pages, but also with groups such as Consumers First, which fought for price parity between Canada and America.

By extracting some of these pages and topic modeling the results, we can confirm existing narratives and raise new questions. Topic modeling finds 'topics' in text. For example, imagine that I am writing about women in a male-dominated labour movement. When I write about the women, I use words like 'femininity', 'equity', 'differential', and 'women'. Men: masculinity', 'wildcat', or 'foremen'. In this thought experiment, imagine I am drawing these words from buckets full of slips of paper. Topic modeling reverses that process, putting those words back into the bucket and telling me what is in it. It is a quick way to get a sense of what might be happening in a large body of text.[18]

Taking the link community that appeared around political parties, we were able to find topics most closely connected to them. In December 2014, the Liberals were highlighting cuts to social programs, issues of mental health, municipal issues, housing, and their new leader, Justin Trudeau (now, as of October 2015, the new Prime Minister of Canada). The Conservatives: Ukraine, the economy, family and senior issues, and the

high-profile stimulus-based Economic Action Plan. For 2006, the results were surprising. The Liberals: community questions, electoral topics (given the federal election), universities, human rights, childcare support, and northern issues. The Conservatives: some education and governance topics, but notably, several relating to Canada's aboriginal population. While the Liberals had advanced a comprehensive piece of legislation designed to improve the conditions of Canada's aboriginal population, Conservative interest in the topic was surprising: perhaps it reflects the Conservative opposition to it? As one commenter on an earlier draft suggested, it may represent the influence of key advisors, one of whom was a leading Conservative scholar of native-newcomer relations. Questions are raised, suggesting promise in marrying content and metadata in such a manner.

A Place of Their Own: Exploring the Ethical Minefield of GeoCities

In general, the sheer scale of distantly reading millions of websites or exploring the public record of political parties has kept us in the previous cases removed from everyday individuals. As the Web became mainstream in the mid-to-late 1990s, GeoCities played a critical role. For the first time, users could create their own web pages without learning HTML or FTP. On sites like GeoCities, they could become part of part of virtual communities, held together by volunteers, neighbourhood watches, web rings, and guestbooks. Even though in 1999 GeoCities was perhaps the third most popular website in existence, Yahoo! deleted it in 2009. Dedicated teams of Internet archivists, such as Archive Team (http://archiveteam.org), created the web archive that we can use today. It is large: at its peak, GeoCities had around 38 million pages.

GeoCities began in late 1994 as a service predicated on geospatial metaphors and giving voices to those who 'had not had an equal voice in society'.[19] Users could easily create new sites within an existing GeoCities community, such as the Enchanted Forest for children or Area 51 for science fiction fans. They received an 'address' based on their neighbourhood: www.geocities.com/EnchantedForest/1005/index.html. In an era when the Web was understood as a new 'frontier', this claim to an actual address resonated.[20] User numbers skyrocketed, from 1,400 in July 1995 to 100,000 by August 1996 and a million by October 1997.

I have been exploring the question of how community was created and enacted there. A significant minority of users threw themselves into the site. When a user arrived to create their site, they had to choose where to live: a small 'cottage' in the Enchanted Forest, perhaps, or a 'tent' in Pentagon.[21] Reminders exhorted them to fit into the site's theme, reach out to neighbours, and crucially – in a move reminiscent of the American 1862 *Homestead Act* – 'move in' and improve their property within a week.[22] Some users became community leaders, welcoming new arrivals and teaching them the ropes. An awards economy boomed, with users creating their own awards and giving them to other sites. They visited each other's guestbooks. Messages are disproportionately from GeoCities users rather than visitors from outside. This community structure persisted until 1999, when Yahoo! bought GeoCities and turned it into a conventional web host.

Like in the previous section, we can explore neighbourhoods with topic modelling. We can see topics in the Enchanted Forest about parties, friends, soldiers and children's characters such as Pingu. In Heartland, topics relating to family, church, and genealogy appear, and in the LGBT-focused WestHollywood, the focus is on gender,

transgender issues, and fighting against hate crimes. Over time, the topics discussed in some neighbourhoods changed. Pentagon moved beyond being a hub for deployed and constantly moving service people towards serving as a forum for political discussions and military history. Heartland came to advance a vision of family focused on Christianity and genealogy. These findings demonstrate that neighbourhoods both shaped and were shaped by user contributions.

How did this come to be? By extracting links, we can begin to find the central nodes that dozens or even hundreds of other websites linked to, as well as the web of connections that held everybody together. This gives us a few hundred websites per neighbourhood to investigate: the community leaders who received kudos from their members, sites that accumulated awards, those with active guestbooks. These factors produced many hyperlinks, both in and out, making these sites critical nodes.

Websites like GeoCities raise ethical questions. Unlike in our previous case studies, which dealt with institutional websites, in GeoCities we are dealing with largely personal websites from over a decade ago. The majority of these people almost certainly did not create these sites with a future historian in mind, nor are they likely to be aware that their sites live on within the Internet Archive or the Archive Team torrent. They did not give consent to the archiving of their sites, nor did they have access to a robots.txt file that could have changed access parameters (see http://archive.org/about/exclude.php). Indeed, unless they remember their URL, users cannot see if their sites were archived in order to pursue their removal from the archive. Traditional archival collections often have restrictions: donor requests, privacy legislation, or the protection of personal information on medical, financial, or other grounds. While historians have ethical responsibilities at

all times, in many cases the onus of making a collection available and accessible lies with institutions. Oral historians, on the other hand, operate outside traditional institutions, instead working in the personal spaces of their interviewees. Institutional review boards, committees that oversee how human subjects are used in research within most North American contexts, govern their work. While none of the above is simple, it is well-travelled ground. Where do web archives fall between these poles?

Strictly speaking, as we generally treat websites as 'publications', it is legal to quote from tweets, blogs, websites, and so forth. Legal does not equal ethical, though. As Aaron Bady notes, 'The act of linking or quoting someone who does not regard their twitter as public is only ethically fine if we regard the law as trumping the ethics of consent.'[23] We need to consider user privacy expectations, which is at the heart of the distinction between a political candidate's site and a GeoCities homestead. This is not to treat users as dupes but to recognize that somebody posting a website in an obscure corner of GeoCities might have an expectation of privacy: many of these sites would not have been discovered by regular users but are easily discovered by web crawlers methodically crawling a community structure.

We can find guidance from web scholars. danah boyd, a web scholar, notes that students with open Facebook profiles regarded a teacher visiting their page as a breach of privacy, social norms, and etiquette.[24] The Association of Internet Researchers provides guidance that has researchers consider the public or private nature of the website and the differences between dealing with sources *en masse* versus individually.[25] Stine Lomberg has emphasized the importance of distance but also, when exploring content, of considering user expectations of privacy.[26]

Historians need to consider these factors when deciding how to appropriately use this material. Some GeoCities cases bring these questions into perspective. Memorial sites, by people who lost children or other loved ones, are both private and intimate but also have well-travelled guestbooks, often by people who lost loved ones of their own. Other searches bring up pages about suicide or depression. These can only be found thanks to today's modern discovery tools. If a 15-year old wrote to the government with a rant, privacy legislation would excise her or his name; if you find the rant in GeoCities, the name – or their pseudonym (which can sometimes be connected to real names) – would be there. These are resources that would never make it into a traditional archive.

We have power because we can access the blogs, ruminations, and personal moments of literally millions of people that would never before have been accessed – but we need to use this power responsibly. With the Wayback Machine, the lack of full-text search provides some privacy, but as we undertake more computational inquiries historians can uncover things forgotten since their creation. My own take on this question is twofold, drawing on earlier literature: we need to consider the scale at play. Mining a few thousand sites and dealing with – and writing about – people in aggregate presents few privacy concerns, whereas zooming in on a handful of websites and closely reading them does. A website many other sites connect to, a proud prominent view counter in the corner (or other equivalent markers of popularity that have supplanted this now dated approach), a well-travelled guestbook, signals a website of an owner who wanted to be read and encountered, and who conceived of themselves as part of a broader Web of documents. A smaller website addressed to an internal audience, written by a teenager and full of revealing messages and pictures, is a different thing altogether.

GeoCities represents a new kind of primary source: the largely non-commercialized, unfettered thoughts of millions of everyday people in the mid-to-late 1990s, left for historians today. We can learn invaluable things, from the forms online community took on the Web to the opinions and thoughts on a host of social, political, or cultural issues or topics.

Conclusions

These three disparate web archiving case studies all demonstrate the critical questions that lie at the heart of these new complex data sets. The technical challenges are clear: not enough processing power or computer memory, the need to find access to a computing cluster, and the variety of file formats and types that underlie them. Rather than a narrow-lens pedagogical approach that stresses say the WARC file, historians who want to use these sources – arguably a necessity when undertaking topics in the 1990s and beyond – need to have a flexible understanding of software and standards.

While this article has focused on the research process, further issues will emerge when scholars attempt to publish this type of work. Images, already a sticking point with many publishers, are borrowed, altered, shared, throughout the Web: can one publish a notable image found in a 1996-era web archive if this has no contactable author or even real name? How can we share our research data with each other if we need to worry about digital rights? How do we balance global copyright regimes with the local contexts of journals and academics? At the least, pedagogical training in copyright is needed, as well as advocacy around orphan works and strengthening fair dealing/use.

Despite these challenges and cautions, which need to be heeded as we move forward, I want to return to the original promise articulated at the beginning of this paper. Each of these case studies, from the Wide Web Scrape to the political movements archive to GeoCities, presents promise. They provide more voices from a more diverse body of people, furthering the goals of social historians to write their histories from the bottom up, to move our stories away from the elites and dominant players of society to the everyday. Web archives are not going to have a slight impact on the practice of history: they are going to force a profound shift. We will have more sources than ever before, by people who never could have conceivably reached large audiences or had their words recorded. We should be optimistic, but we need to be prepared.

---

[1] R. Rosenzweig, 'Scarcity or abundance? preserving the past in a digital era', *American Historical Review*, 108, no. 3 (2003), 735–62.

[2] Old Bailey Online, *The proceedings of the Old Bailey, 1674–1913*, http://www.oldbaileyonline.org/, last accessed 16 June 2015.

[3] For examples from the Canadian context, see C. Levitt, *Children of privilege: student revolt in the sixties : a study of student movements in Canada, the United States, and West Germany* (Toronto, 1984) or D. Owram, *Born at the right time: a history of the baby boom generation* (Toronto, 1997).

[4] An excellent and path-breaking anthology on this topic is C. Potter, R. Romano, eds., *Doing Recent History: On Privacy, Copyright, Video Games, Institutional Review Boards, Activist Scholarship, and History That Talks Back* (Athens, GA, 2012).

[5] No good history of the Internet Archive yet exists, but for more information see D. Gillmor, 'Future Historians Will Rely on Web,' *Philadelphia Inquirer*, September 22, 1996; Internet Archive, 'The Internet Archive: Building an 'Internet Library,'' 20 May 2000, http://web.archive.org/web/20000520003204/http://www.archive.org/; A. Brown, *Archiving Websites: A Practical Guide for Information Management Professionals* (London, 2006); S. Meloan, 'No Way to Run a Culture', *Wired*, February 13, 1998, http://web.archive.org/web/20000619001705/http://www.wired.com/news/culture/0,1284,10301,00.html.

[6] S. G. Ainsworth, M. L. Nelson, and H. Van de Sompel, 'Only One Out of Five Archived Web Pages Existed As Presented,' *Proceedings of the 26th ACM Conference on Hypertext & Social Media*, (New York, NY), 257–66.

[7] I. Milligan, 'Illusionary Order: Online Databases, Optical Character Recognition, and Canadian History, 1997–2010,' *Canadian Historical Review* 94, no. 4 (2013), 540–69.

[8] For more, see Internet Archive, *Wide Crawl started March 2011*, 2012, http://archive.org/details/wide00002, last accessed 16 June 2015; V. Goel, *2011 WIDE Crawl (wide00002),* 2012, http://archive.org/~vinay/wide/wide-00002.html, last accessed 16 June 2015.

[9] International Standards Organization, *ISO 28500:2009 – Information and documentation – WARC file format*, 2009, http://www.iso.org/iso/catalogue_detail.htm?csnumber=44717, last accessed 16 June 2015.

[10] For more on the history of the Internet, see J. Abbate, *Inventing the Internet* (Cambridge, Mass, 2000); T. Berners-Lee, *Weaving the Web: The Original Design and Ultimate Destiny of the World Wide Web* (San Francisco, 2000); J. Ryan, *A History of the Internet and the Digital Future* (London, 2011). A good overview of how the Internet works can be found in A. Blum, *Tubes: A Journey to the Center of the Internet* (New York, 2013).

[11] N. Brügger, D. Laursen, and J. Nielsen, 'Studying a Nation's Web Domain over Time: Analytical and Methodological Considerations', presented at the International Internet Preservation Consortium 2015, Palo Alto, California, April 27, 2015. http://netpreserve.org/sites/default/.../2015_IIPC-GA_Slides_02_Brugger.pptx, accessed July 27, 2015.

[12] Warcbase is documented at https://github.com/lintool/warcbase/wiki.

[13] Apache Software Foundation, *Apache Lucene – Apache Solr*, accessed 21 August 2013, http://lucene.apache.org/solr/, last accessed 16 June 2015.

[14] UK Web Archive, *Shine application*, http://www.webarchive.org.uk/shine, last accessed 16 June 2015.

[15] S. Osiński, J. Stefanowski, and D. Weiss, 'Lingo: search results clustering algorithm based on singular value decomposition', *Advances in soft computing, intelligent information processing and web mining: proceedings of the International IIS: IIPWM'04 Conference* (Zakopane, Poland, 2004), 359–68, http://www.cs.put.poznan.pl/dweiss/site/publications/download/iipwm-osinski-weiss-stefanowski-2004-lingo.pdf, accessed 27 July 2015.

[16] More information on these shifts can be found in I. McKay and J. Swift, *Warrior Nation: Rebranding Canada in an Age of Anxiety* (Toronto, 2012) and Y. Frenette, 'Conscripting Canada's Past: The Harper Government and the Politics of Memory,' *Canadian Journal of History* 49, no. 1 (2014): 50-65.

[17] V. Goel, *Web archive analysis workshop - Internet research - IA webteam confluence*, https://webarchive.jira.com/wiki/display/Iresearch/Web+Archive+Analysis+Workshop, last accessed 16 June 2015; J. Lin et al, *warcbase*, https://github.com/lintool/warcbase, last accessed 16 June 2015.

[18] This is a shorter version of the great M. Jockers, 'The LDA Buffet: A Topic Modeling Fable,' *matthewjockers.net*, http://www.matthewjockers.net/macroanalysisbook/lda/, last accessed 5 November 2015.

[19] S. Hansell, 'The neighbourhood business: GeoCities' cyberworld is vibrant, but can it make money?', *New York Times*, 13 July 1998.

[20] F. Turner, *From counterculture to cyberculture: Stewart Brand, the Whole Earth network, and the rise of digital utopianism* (Chicago, 2008).

[21] G. Graham, *The Internet: a philosophical inquiry* (London, 1999), 148.

[22] J. Logie, 'Homestead Acts: rhetoric and property in the American West, and on the World Wide Web', *Rhetoric Society Quarterly* 32, no. 3 (1 July 2002), 33–59.

[23] A. Bady, *#NotAllPublic, Heartburn, Twitter*, 10 June 2014, http://thenewinquiry.com/blogs/zunguzungu/notallpublic-heartburn-twitter/, last accessed 16 June 2015.

[24] d. boyd, *It's complicated: the social lives of networked teens* (New Haven, 2014), 58.

[25] A. Markham and E. Buchanan, *Ethical decision-making and Internet research: recommendations from the AOIR Ethics Working Committee (version 2.0)*, September 2012, http://aoir.org/reports/ethics.pdf, last accessed 16 June 2015.

[26] S. Lomborg, 'Personal Internet archives and ethics', *Research Ethics* 9, no. 1 (1 March 2013), 20–31.