Content Selection and Curation for Web Archiving: The Gatekeepers vs. the Masses

Ian Milligan[1], Nick Ruest[2], and Jimmy Lin[1]
[1] University of Waterloo [2] York University
{i2milligan,jimmylin}@uwaterloo.ca, ruestn@yorku.ca

ABSTRACT

Any preservation effort must begin with an assessment of what content to preserve, and web archiving is no different. There have historically been two answers to the question "what should we archive?" The Internet Archive's broad entire-web crawls have been supplemented by narrower domain- or topic-specific collections gathered by numerous libraries. We can characterize this as content selection and curation by "gatekeepers". In contrast, we have witnessed the emergence of another approach driven by "the masses"—we can archive pages that are contained in social media streams such as Twitter. The interesting question, of course, is how these approaches differ. We provide an answer to this question in the context of a case study about the 2015 Canadian federal elections. Based on our analysis, we recommend a hybrid approach that combines an effort driven by social media and more traditional curatorial methods.

1. INTRODUCTION

Any preservation effort must begin with an assessment of what content to preserve: archivists refer to this as appraisal, which is related to what librarians call collection development. This process remains inescapable, even in the digital context. Even if there are no technical barriers (e.g., storage capacity) to simply "keep everything" (and inevitably, there are—in most cases, available budget), such a strategy is not feasible for a variety of other reasons. This is especially true for web archiving, which refers to the systematic collection and preservation of web content.

The web has become an integral part of our daily lives and captures our "collective memory", recording everything from major world events to the rhythm of commerce. Even personal minutiae are valuable in that they offer a snapshot of our society, much in the same way that a diary from the 17th century provides insight into what the world was like then. It would not be an exaggeration to say that the web has become an important part of our cultural heritage worthy of preservation. As web pages are ephemeral and disappear with great regularity [7], the only sure way of preserving web content for posterity is to proactively crawl and store portions of the web.

Any web archiving effort must begin with the following question: which sites should we crawl and how frequently? Historically, there have been two answers to this question, which has been supplemented by a third more recently. The Internet Archive has been collecting and storing web content since 1996, and to date has amassed hundreds of billions of pages totaling tens of petabytes. The Internet Archive's actual crawl strategy is opaque, but the organization aims to periodically gather a broad snapshot of the web as a whole—this thus serves as the first possible answer to our question: broad across-the-web scrapes. The second answer is supplied by a loosely-organized network of national, academic, and other libraries who adopt a strategy that is similar to the development of special collections. Based on some mandate, librarians scope their crawls—in the case of national libraries the mandate might be preserving pages in their country's domain; in many academic libraries, special web archive collections are created because they capture events of interest. The librarians who undertake such collection development essentially serve as information gatekeepers. Finally, the third, and most recent development, is to drive web archiving efforts based on social media—for example, archive those pages that are linked to from tweets. In contrast to librarians, we might think of this approach as content selection and curation by the masses. The question is how these approaches differ.

The contribution of this paper is an answer to this question in the context of a case study. We compare the contents of three different web archive collections with respect to the 2015 Canadian federal elections: a professionally curated collection by the University of Toronto, a collection formed by gathering pages linked from Twitter, and the general collection in the Internet Archive's Wayback machine. Based on our analysis, we recommend a hybrid approach that combines an effort driven by social media and more traditional curatorial methods. A manually-curated collection provides a robust foundation—site infrastructure, unpopular parties, marginal candidates—to layer the selection of the "masses" upon. On their own, popularly-curated web crawls are insufficient. Yet with contextual data, they can be very powerful.

2. BACKGROUND AND RELATED WORK

To begin, why is this an important issue to explore? The answer is simple: the content selection and curatorial decisions that we are making today define the source record of

JCDL '16, June 19 - 23, 2016, Newark, NJ, USA

ISBN 978-1-4503-4229-2/16/06…$15.00

DOI: http://dx.doi.org/10.1145/2910896.2910913

tomorrow. Thirty years from now, when historians study contemporary society, we do not want them to have an unnecessarily warped vision of the world today.

We are not the first to note the limitations of manually- or algorithmically-curated web archives. Farag and Fox [3] noted that while manual curation can render "high quality–time consuming" web archives, social media-based curation leads to "low quality–time saving" collections. Using tweets from high-profile events, including the Ebola outbreak and American Thanksgiving, they collected and compared URLs. The emphasis of that work, however, is the construction of an event model, whereas we are focused on providing a historical corpus for researchers; thus, the "low quality" of a citizen-created archive might be an advantage in our case. Similarly, the work of Georgescu et al. [4] in event-model extraction used Wikipedia edits for event detection, concluding that the citizen-generated approach is promising.

A common question, however, is: How comprehensive is archiving coverage? One study has found that between 35% and 90% of the web "has at least one archived copy." [1] This roughly lines up with earlier findings by Payne and Thelwall [8] and Russell and Kane [10]. We add to that literature by comparing three different collections around one event, examining content overlap and potential biases.

Others have explored the power of social media in providing seed lists. In the aftermath of the shooting of Mike Brown in Ferguson, Missouri and the ensuing protests, the Internet Archive's subscription archiving service, Archive-It, announced that they were "accepting URL nominations for web archive collection on Ferguson" to generate their seed list. Ed Summers extracted URLs from the #Ferguson Twitter hashtag and submitted those as seeds [11].

The most significant undertaking in this area was the IIPC-funded Twittervane project, developed by the British Library with the aim of "monitoring and analysing Twitter traffic relevant to a given theme and generate a list of most frequently shared web resources." While they brought the program to a prototype stage, curatorial feedback was lukewarm. The IIPC report found that only 20-30% of the URLs tweeted could be considered valid archival selections. For example, of the top seven URLs found by the Library of Congress test user, only one was relevant to their overall collecting approach [5]. Our project updates this work.

3. A TALE OF THREE COLLECTIONS

Using a study of a recent Canadian election, we compare:

- A collection of 1,988,693 URLs tweeted by users on the #elxn42 Twitter hashtag;
- The holdings of the Internet Archive's Wayback Machine;
- The August and November 2015 crawls of the Canadian Political Parties and Political Interest Groups (CPP) web archive collection.

These represent different collecting paradigms. The first represents the curatorial decisions of the "masses," or of the 318,176 unique users who used the #elxn42 hashtag. The second, the broad yet shallow crawls conducted by the Internet Archive. And finally, the curated collection gathered by the University of Toronto between 2005 and 2015.

3.1 Twitter

For archivists interested in user-generated corpora, Twitter shows promise. It provides insights into the opinions, beliefs, and sentiments of everyday people. This comes in both the form of the 140-character limited tweet content itself, as well as the links shared to tweets, websites, and documents. While Twitter is not a representative sample of broader society—skewing young, college-educated, and affluent (above $50,000 US household income) [2]—it represents a dramatic increase in the amount of information generated, retained, and preserved from ordinary citizens.

Twitter		**CPP (Aug./Nov.)**	
twitter.com	615421	liberal.ca	55536
cbc.ca	143941	greenparty.ca	45788
youtube.com	66886	policyalternatives.ca	37810
huffingtonpost.ca	66758	socialist.ca	26856
theglobeandmail.com	63401	davidsuzuki.org	25487
thestar.com	53051	canadians.org	24424
ctvnews.ca	49295	ccrweb.ca	19521
globalnews.ca	46488	afn.ca	15879
twimg.com	39989	blocquebecois.org	10899
macleans.ca	35280	egale.ca	7837

Table 1: Top tweeted domains (left) and top CPP domains from the Aug./Nov. 2015 crawls (right).

A Canadian federal election was called on 3 August 2015, presenting a case study to compare user-tweeted URLs versus the seed list in our more conventional CPP collection (more details below). We carried out harvesting of the #elxn42 hashtag (the 2015 Canadian federal election hashtag) to compare what voters tweeted about with the formal seed list from the CPP collection. In total, we collected 3,918,932 tweets [9]; these tweets and the URLs contained in them form a foundation for the web archive.

To create the social media collection, our team began capturing tweets with the #elxn42 hashtag on 3 August 2015 using `twarc` [13], a command line tool and Python library for archiving Twitter JSON data, using Twitter's streaming and search APIs. We stopped collecting on 5 November 2015, the day after Justin Trudeau was sworn in as the 42nd Prime Minister of Canada. Using the `twarc` analysis library, we extracted tweeted URLs. As Twitter uses automatic link shortening, we also unshortened every URL in the dataset so that we would be able to create a canonical list of URLs tweeted for further analysis. We were able to create this using a combination of open-source tools: `unshorten.py` and `unshrtn` [12]. A total of 1,988,693 URLs were tweeted (50.9% of all tweets contained a URL), 334,841 of which were unique. By aggregating and sorting the URLs, we could see the domains that were tweeted the most in Table 1 (left). We find that twitter.com is the top tweeted domain largely due to "quoted" tweets, a form of retweeting, commenting upon, and endorsing other content.

3.2 Canadian Political Parties

To compare the Twitter-based web archive with another collection, we used the Canadian Political Parties and Political Interest Groups (CPP) collection. We have been using this for an analysis of Canadian politics between 2005 and 2015, and have provided public access to it through our http://webarchives.ca/ portal. The CPP collection is the product of a quarterly crawl, beginning in 2005, by the University of Toronto Libraries using Archive-It, the Internet Archive's web archiving subscription service. It includes all major Canadian federal parties, many minor ones, as well as a nebulous group of "political interest groups," ranging

	CPP	Twitter	Wayback
CPP	-	0.341%	74.30%
Twitter	0.269%	-	10.06%
Wayback	N/A	N/A	-

Table 2: Intersection analysis. Read as percentage of *row* found in *column*, e.g., 0.341% of URLs from CPP were in the Twitter #elxn42 collection.

from groups advocating for marriage equality, the banning of land mines, environmental issues, and Canada's First Nations. With over fifteen million documents crawled, it is an unparalleled collection of recent Canadian political history.

The collection has a significant downside, in that it has opaque seed list selection criteria. The librarian responsible for scoping this collection in 2005 has retired. Curatorial choices were not documented. While political parties are well covered, the interest groups were largely discovered through keyword searches, some were excluded due to robots.txt exclusions, and the seed list was largely developed by one person.

By aggregating and sorting the URLs, we can see the domains that are most represented in the CPP collection in Table 1 (right).

3.3 Internet Archive

The Internet Archive engages in broad crawling. For example, in the March–December 2011 Wide Web Scrape, they began with the top million URLs based on the Alexa Internet rankings, and crawled from there. These crawls capture many sites, but to a limited depth.

4. INTERSECTION ANALYSIS

To query the Internet Archive's Wayback Machine, we used their Wayback CDX Server API.[1] This takes a URL and determines whether there is an archived, accessible copy in the Wayback Machine. We ran lists of all the unique URLs in the CPP collection and the #elxn42 Twitter collection through the API, which provides a list of all timestamps of available captures. We then checked to see if the Wayback Machine had a copy of the webpage within the August–December period. Results are shown in Table 2.

Of the 1,988,963 URLs that were tweeted (334,841 unique URLs), there was low coverage in the CPP collection (drawing only on the August and November 2015 crawl URLs): of the URLs in CPP, only 0.341% are found in the Twitter collection. We thus have very different collections: the library gatekeepers have captured a very different picture of Canadian politics than the "masses" on Twitter.

To add to this understanding, we subsequently carried out an investigation of what tweeted URLs from #elxn42 would be included in the Internet Archive or the CPP collection. To do so, we took our list of 334,841 unique URLs and submitted them to the Wayback CDX Server API.

Of these URLs, 33,685 were present in the Wayback Machine with a snapshot between August and December 2015. This gives the Wayback Machine a coverage, within our time period, of 10.05%. If we were to remove the time period limit, 68,112 of the URLs (or 20.34%) had at least one snapshot dating back to 1996, but not necessarily within our time period. While both values are below the lower bound of the 35-90% coverage from previous work, this reflects the changing nature of websites—more social media—as well as the early timing of our inquiry.

[1] https://github.com/internetarchive/wayback/

Included		**Excluded**	
cbc.ca	3035	twitter.com	173931
youtube.com	2639	linkis.com	11071
thestar.com	1665	youtube.com	6026
theglobeandmail.com	1644	instagram.com	5302
huffingtonpost.ca	1561	globalnews.ca	4709
twitter.com	1550	cbc.ca	4529
ctvnews.ca	1423	facebook.com	4282
nationalpost.com	1262	rabble.ca	3859
globalnews.ca	1062	huffingtonpost.ca	3762
ottawacitizen.com	836	fw.to	3284

Table 3: Top #elxn42 Twitter domains included (left) and excluded (right) in the Wayback (August–December crawls).

What #elxn42 URLs were and were not included in the Wayback Machine was fascinating. Table 3 (left) shows the top ten domains that were found in the Wayback Machine. The top ten domains that were *not* present in the Wayback Machine included significant overlap with these same domains, as seen in Table 3 (right). Some of these are social media websites (Facebook, Instagram), and others are Canadian media outlets that are likely not crawled much, such as the left-wing news site rabble.ca, some Canadian Broadcasting Corporation pages, and Global News—in the broad global scope of the Internet Archive, they may receive little attention. However, there was quite a bit of overlap on major traditional print newspapers: The *Toronto Star*, the *Globe and Mail*, and the *National Post*, Canada's three highest-circulation newspapers, also had their websites included in both the #elxn42 corpus and the Wayback Machine. A few omissions were technical. One is a link shortener (fw.to) that is not supported by our link unshortening package. The other, linkis, is a platform that personalizes shared sites (most of these tweets were shared using the linkis client).

Which of the URLs tweeted on the #elxn42 hashtag would have been included—or would be potentially included—in the CPP collection? The actual inclusion coverage is low, amounting to 902 or 0.269%. This does not tell the full story, however. Comparing the domains tweeted with the CPP's fifty seed domains, we found that 59,576, or 17.79%, were part of the fifty domains. While slightly lower than the global Wayback Machine, this is roughly comparable. This suggests that the CPP collection does indeed capture websites of significant public interest.

Finally, we were curious about what URLs found within the CPP collection—drawing on the two most recent crawls in August and November 2015—would be found in the global Wayback. We discovered that 74.3% of CPP URLs were found there with snapshots within the last six months; removing the time limit, we observe 83.94% coverage for CPP URLs in the Wayback dating back to 1996.

While Archive-It and the Wayback Machine are similar, largely due to the former being routinely piped into the latter, they are not identical. The differences were largely driven by changes to crawl scope: the CPP collection included RSS feeds, forms, calendars (often crawler traps), and more discussion forum content. CPP also contained a

few hundred YouTube videos that were out of scope in the Wayback. As Archive-It crawl operators have considerable crawl discretion, including the ability to ignore robots.txt, the slight variation is unsurprising.

5. DISCUSSION

The three crawl paradigms discussed in this paper offer relative advantages and disadvantages. The CPP collection provides a broader documentary overview of Canadian politics than the Twitter corpus, as reflected in the low 0.341% coverage. This is due to three reasons beyond the reality of the CPP collection spanning ten years (December 2005 to present) versus the few months of the federal election.

First, several websites that were collected as part of the CPP collection were not tweeted at all. These included unpopular fringe parties who have largely faded from the public eye. While not commanding popular support, they provide useful historical information about the extremes of the political spectrum. Manual curation, done with sensitivity, can ensure the inclusion of more minority viewpoints.

Second, the CPP collection includes entire websites: from calendars, to menus, archived pages, to terms of service, and beyond. Users do not tweet this important content.

Third, the CPP collection has an institutional bias in it. Comparing the top ten domains tweeted in the #elxn42 dataset versus the top ten in the CPP collection (see Table 1) reveals that the only overlap is the Conservative Party of Canada's website. Curation by professionals, performed over a long period of time, tends to focus on stable institutions (understandably). On the other hand, Twitter users tweeted more ephemeral sites and social media: issues of popular discussion and controversy, such as political platforms, controversial press releases, and popular events (which all rank highly in the correlation between the CPP and the Twitter corpus).

We cannot rely on the Internet Archive as a replacement for either professionally-curated collections or Twitter-based crawls. The Internet Archive's main collection is necessarily broad but shallow: websites are only crawled, in some cases, a few times a year, and only to a certain depth. Crawlers may not reach deep into large domains. Without input from Archive-It (which represent the efforts of professional curators), we anticipate that the coverage of the Wayback machine would be even spottier on topics of scholarly interest.

Access is also more easily enabled with smaller, focused collections in a way that providing access to broad crawls has so far been elusive. To use Internet Archive or most national library collections, users must know the exact URL of the resource they are looking for as an entryway; in other cases, such as the British Library, full-text search exists but is severely hamstrung by access and content rendering restrictions [6]. There is room for smaller, more circumscribed collections, as the popular and media success of http://webarchives.ca/ demonstrates: a subject-focused collection can appeal to both scholars and the general public, more importantly.

6. CONCLUSION

A hybrid approach between Twitter-based and traditional curatorial methods is recommended. The Archive-It collections provide a foundation to lay the more specific Twitter-focused collections upon. Curators could be encouraged to collect event-based hashtags alongside traditional methods, perhaps in consultation with researchers. Tweeted URLs have an innate demographic and partisan bias within them, drawing on profiling information about Twitter users, but so do curated collections, which can suffer from a lack of documentation and transparency about how they are collected. For researchers, Twitter-based collections are at least documentable: the parameters of the hashtags chosen, streaming method used, and the rich metadata embedded in the tweet JSON itself can help contextualize further studies. In addition, our work significantly builds upon the earlier IIPC-funded Twittervane project, with more positive research outcomes. While one limitation of this short paper is that we could not explore the quality of the preserved content—instead, focusing more on quantities—future research will explore actual content differences.

We believe that scholarly findings from a Twitter-based web archive would differ substantially from a professionally curated collection. The former is a laser-focused snapshot of collections of immediate interest from potentially millions of users, while the latter is a broader collection of a still relatively narrow band of domains selected by subject-matter experts. They are apples and oranges, but complement each other very well. Most importantly, we need both.

Acknowledgments. This work was supported by the Social Sciences and Humanities Research Council of Canada under Insight grant 435-2015-0011 and the U.S. National Science Foundation under awards IIS-1218043 and CNS-1405688. Any opinions, findings, conclusions, or recommendations expressed are those of the authors and do not necessarily reflect the views of the sponsors.

7. REFERENCES

[1] S. G. Ainsworth, A. AlSum, H. SalahEldeen, M. C. Weigle, and M. L. Nelson. How much of the web is archived? *arXiv:1212.6177*, Dec. 2012.

[2] M. Duggan. The demographics of social media users, Aug. 2015.

[3] M. M. Farag and E. A. Fox. Building and archiving event web collections: A focused crawler approach. *Bulletin of IEEE Technical Committee on Digital Libraries*, 11(2), 2015.

[4] M. Georgescu, N. Kanhabua, D. Krause, W. Nejdl, and S. Siersdorfer. Extracting event-related information from article updates in Wikipedia. *ECIR*, 2013.

[5] H. Hockx-Yu and M. Pitt. Evaluating Twittervane: Project final report, June 2013.

[6] I. Milligan. Web archive legal deposit: A double-edged sword, July 2015.

[7] A. Ntoulas, J. Cho, and C. Olston. What's new on the web? the evolution of the web from a search engine perspective. *WWW*, 2004.

[8] N. Payne and M. Thelwall. A longitudinal study of academic webs: Growth and stabilisation. *Scientometrics*, 71(3):523–539, June 2007.

[9] N. Ruest and R. White. #elxn42 tweets (42nd Canadian Federal Election), Dec. 2015.

[10] E. Russell and J. Kane. The missing link: Assessing the reliability of internet citations in history journals. *Technology and Culture*, 49(2):420–429, 2008.

[11] E. Summers. A Ferguson Twitter archive, Aug. 2014.

[12] E. Summers and D. Krech. unshrtn, Dec. 2015.

[13] E. Summers, H. van Kemenadem, P. Binkley, N. Ruest, recrm, S. Costa, E. Phetteplace, T. G. Badger, M. A. Matienzo, L. Blakk, D. Chudnov, and C. Nelson. twarc: v0.3.3, Aug. 2015.