Lies, damn lies, and fake news: toward a library science and digital curation-oriented methodology for web archiving national news and social media

Tom J. Smyth, MA, MIS, PMP (@smythbound)
Web and Social Media Preservation Program¹, Library and Archives Canada
Ottawa, Canada
tom.smyth@bac-lac.gc.ca

Abstract:

In recognition of their importance, the National Library and Archives Canada (LAC) has web archived Canadian news and social media since 2006. As part of our Web and Social Media Preservation Program collections development and methodological evolution, and to make advances in efficiency and in our ability to respond to and document unforeseen events of national historic or international significance, we started broadly web archiving front page news and social media from selected sources and in an automated manner in 2015.

Since the web now constitutes an invaluable primary source for 21st century digital society and for the writing of future history, managed action is necessary to preserve it while it is available. The news media provides many perspectives and chronological descriptions of how any major event unfolded; social media provides the full spectrum of political, regional, economic, and cultural first-hand perspectives. This paper discusses the web archival strategies and methodologies developed at LAC that drive our news and social media evaluation and acquisition, with particular attention to the politics of “fake news”, COVID-19, and the invasion of Ukraine. It also describes the details of media collection in association with large-scale thematic web archival (research) and data collections – all of which is informed by the guiding principles of digital curation and preservation.

Our deliberate targeting of 'fake news’ and 'extreme political viewpoints’ will be explored, with arguments for its collection as source to support future inquiries and research in its own right. Ethical and legal considerations (or potential consequences!) libraries could face for labelling or excluding such content, and strategies for addressing this, are also provided.

Keywords: libraries and fake news, collecting news media, web archiving, collection development, curating research collections.

¹ Library and Archives Canada 2023b.
“There are three kinds of lies: lies, damned lies, and statistics.”
-Mark Twain (possibly Benjamin Disraeli).

“Fallacies of statistical sampling occur in generalizations which rest upon an insufficient body of data – upon a “sample” which misrepresents the composition of the object in question…rarely have so many generalized so much from so little as have impressionistic historians of the past generation”.
- David Hackett Fischer, Historian's Fallacies.²

What does statistics have to do with this subject? The first passage has been oft quoted, but both are very relevant to the media of today as well (indeed, one might passably replace the word “statistics” with “the media” in the first maxim and escape unscathed). This is not to insult statistics! If one looked hard enough, it is likely that a news media source of any political stance, or with any possible opinion, could be found. As an amateur historian, these issues informs the author’s approach to web archives.

The second quote speaks to the possible outcome of the unwittingly use of statistics (or media). There is wisdom in Fischer’s warnings thought his pivotal book, and this is very pertinent to this paper, as libraries and information professionals are quite challenged to do their work in today’s age of normalized “fake news”. What can we do about this, as stewards of information veracity? (Predictably: it depends).

Introduction and Web Archiving Program History

The National Library and Archives of Canada (LAC) has been selecting electronic news media since 1994, when it began collecting such resources on a voluntary deposit basis under the-then National Library of Canada.

Web archiving also began in the National Library of Canada in 1994, and was linked directly to its participation as a founding member of the International Internet Preservation Consortium (IIPC).³ This collaboration for the stewardship and digital preservation of the Internet was influential in establishing the context and terminology for LAC’s legislated authority for the web archiving of Canadian web resources. Per the Library and Archives of Canada Act 2004, section 8 (2)⁴:

Powers of Librarian and Archivist

Sampling from Internet

(2) In exercising the powers referred to in paragraph (1)(a) and for the purpose of preservation, the Librarian and Archivist may take, at the times and in the manner that he or she considers appropriate, a representative sample of the documentary material of interest to Canada that is accessible to the public without restriction through the Internet or any similar medium.

Resources in formats for the web are recognized internationally as a critical aspect of any nation’s digital documentary heritage.⁵ Owing to their precariousness and sometimes-ephemeral context (especially for new media!), it is essential to capture web-based resources before they are taken permanently offline where they could be forever lost to the digital archive, and where they constitute primary sources for research on the 20th and 21st centuries.

² Hackett 2000, 104-5.
³ IIPC 2023.
⁵ Milligan 2019, 6-7, 24.
Documenting International Events, National Crises via the Media and Web Archiving

As early as 2006, we envisioned web archiving as a means of documenting and responding to significant and commemorative events in Canadian history. Our earliest, non-governmental and thematic collections covered state funerals, historic anniversaries, and e.g., British Royal visits.

We quickly evolved to covering federal and provincial or territorial elections. The importance of collecting news media as an essential component of web archival collections became evident when we began specifically targeted key pundits for the analysis of controversial issues in Canadian politics, and as a key perspective on any given political party’s espoused policy and election campaign. Further, collecting the web edition of a mainstream newspaper proved important, as it sometimes generated commentary and political content faster, or demonstrated exclusive content that did not appear in any other (mainly print) edition.

In 2009, we began the curation of a formal collection documenting the Vancouver 2010 Olympic and Paralympic games, the first on Canadian soil since the inception of the web archiving program at LAC. This project was also pivotal in terms of methodological and collections development. An Olympics can be a polarizing and political affair, particularly for the host nation. Opinions vary, naturally, on the wisdom (pros and cons) of hosting an Olympics due to the economic, infrastructural, regional, and environmental impacts. An essential perspective for documentation for us in 2009 was that of the indigenous nations on whose unceded land the games were taking place.

Thus, the Vancouver 2010 project had many things to teach; not only was this the genesis of our disciplined, library collection development-oriented approach to web archival curation (including documentation of sub-themes within a theme and depth indicators), but this also showed us the importance of targeting diverse and politically “leftist” and “rightist” news media. It also became essential, for posterity and future research, to acquire cultural news media sources that communities sometimes produce for themselves; to collect national, provincial or territorial, and regional media for their varied perspectives; but also to make a concerted effort to represent the full political spectrum of news media and opinion in the web archive. Only this could provide a balanced and nuanced interpretation of any particular issue that might become a topic of research inquiry many years later.

Politics and “Fake News” in Times of Crisis: Ethics and Legalities of Labelling Sources

This paper does not advocate for, or in engage in, labelling any particular source or content as “fake news”, but it also does not intend to argue for a relativistic or subjective perspective on what is and is not a (scientific) “fact”. It is important to state this, as not all library clients will have a similar philosophy.

We should consider this issue in the context of collecting news and social media, which themselves may or may not demonstrate various political perspectives. We should consider how one’s own political stance could influence the type of media anyone chooses to trust and consume, or how a preferred media source can influence one’s opinion on any particular matter. How should these sometimes charged and sensitive political and personal preferences affect library collection development strategies? As a curator, how does one maintain a neutral (dare I say, ‘objective’) approach, insofar as that as possible?

Taking it a step further, in the context of building research collections and in delivering library services, we should consider “for whom, or according to whom, is a given source or perspective indeed “fake news”? We should also endeavour to provide some context and precision on what is meant by the term. According to the UNESCO Handbook on Journalism, ‘Fake News’ and Disinformation:
Table 1: Disinformation, Misinformation, Malformation per UNESCO

| Disinformation: | Information that is false and deliberately created to harm a person, social group, organisation or country |
| Misinformation: | Information that is false but not created with the intention of causing harm |
| Mal-information | Information that is based on reality, used to inflict harm on a person, social group, organisation or country. |

Any or all of these meanings could be referenced when the term is utilized. This is vital to discuss as it raises important dimensions for libraries in e.g., the contexts of e.g., reference services and reader’s advisory.

From the outset, librarians are trained not to be judgemental when delivering reader’s advisory services, and our ethical approach is to provide topical recommendations aligned with a client’s tastes without being critical, or trying to influence a client’s preferences, or worse – steering them toward one’s personal tastes. (Herald and Wiegand 2006, 20, 25-8).

Similarly, while this might have been something of a more private matter in previous years, librarians today can encounter awkward situations where they are presented abruptly with a patron’s worldview (whatever that is). How do these principles of non-judgemental reader’s advisory (objectivity) apply when it becomes obvious that a client has deep belief in (what we may personally consider to be) “fake news?” I do not mean to engage in epistemological relativism. Rather, it is worth stating that some patrons will come into a library with an obvious bias, which would influence the types of resources they are asking for and/or their approach to a research or reference query. This adds not a new dimension, but new facets to the reference or reader’s advisory interviews.

In reader’s advisory contexts, should “fake news” be recommended when “fake news” is requested? Does the librarian have a role in challenging the worldview in which the client is clearly invested? Certainly not, unless the exchange has a connotation of critical source selection or academic research. Put another way, we should tend to respond with tactful critical thought and analysis where the patron’s interview permits a dialogue on source veracity.

Otherwise, potentially serious issues could arise. For example: the writer once overheard the request: “Can you recommend resources that prove COVID-19 vaccines cause autism”? (etc.) Directing the patron to the shelf in the library that is labelled “fake news” therefore could cause obvious issue and unintended insult. Legal issues could arise where the client in question is actually a publisher, and then finds their own works under “COVID 19 – Popular Sources – Fake News”.

But how should we as library managers deal with such issues as respects “fake news”? The answer surely lies in a more tactful approach and a disciplined collection development strategy. Let us turn now to a discussion of how this arose.

“Events-Based Collections” Web Archival Collections

From 2011, we began to react to, and reprioritize our resources for documenting unanticipated events with international or national historic significance. From this date, we called these “events-based collections”, or what the international web archiving community sometimes refers to today as “rapid response” collections or methodologies.

---

6 Cherilyn and Posetti 2021.
7 According to Elise Gravel, who my kids absolutely love, «...ça peut être difficile de décider qui croire. Si tu as des doutes sur une source, tu peux demander d’un bibliothécaire. Les bibliothécaires sont formés pour t’aider à trouver des informations fiable. » / If you have doubts about a source, you can ask a librarian. Librarians are trained to help you find reliable information. Gravel 2022, 73.
Table 2: Early Events-Based Collections at LAC:

<table>
<thead>
<tr>
<th>Event</th>
<th>Start Date</th>
</tr>
</thead>
<tbody>
<tr>
<td>Wildfires in Canada</td>
<td>May 2011</td>
</tr>
<tr>
<td>Floods in Alberta</td>
<td>Jun. 2013</td>
</tr>
<tr>
<td>Lac Megantic Train Derailment</td>
<td>Jul. 2013</td>
</tr>
<tr>
<td>Ottawa Bus Crash</td>
<td>Sept. 2013</td>
</tr>
<tr>
<td>War Memorial/Parliament Attack</td>
<td>Oct. 2013</td>
</tr>
<tr>
<td>Humboldt Broncos Bus Accident</td>
<td>Apr. 2018</td>
</tr>
</tbody>
</table>

While we do not like calling these our “disaster collections”, often these events-based projects document unpleasant events. The first of large collection scale documented the Lac-Mégantic rail disaster that took place in Quebec, Canada in July 2013. In examining the resultant news media, and owing to the nature the reporting on the disaster as it unfolded, we realized we were dealing with multiple, sometimes-conflicting interpretations or speculation on what had happened, before the facts were known. Who or what source was right, and how did this influence public opinion? This demonstrated the importance of preserving such varied news perspectives for future research; this was a dimension to source that would enable not only the reconstruction of what happened, but a comparison at a later date of known facts versus the reporting versus public impressions.

Thus, from 2013, we had not yet progressed to collecting daily news, and we were reacting in an ad hoc manner to document major events. However at this time in the program history, many simultaneous events arose that challenged our available resources, and we realized we needed a better strategy to document the basic timeline of any particular event.

From 2014-16, severe wildfires became widespread across Canada. As an issue with major impact on our environment and the economy and that of particular provinces or territories, LAC began curating a “wildfires” collection with particular attention to the impact on Fort McMurray, Alberta starting May 2016. Approaching the wildfires per our normal thematic curation methodology proved problematic. First, our investigations showed there were few non-media web resources covering the wildfires outside small sections of the provincial or territorial government domains – and therefore the news media coverage and social media became the primary sources on the wildfires. However, some new process was needed to ensure the timelier (and less labour-intensive) collection of the news.

Automating National and Regional News Media Collection and Selection Methodology

After our experiences with the collections under Table 2, and as part of our program evolution (in part to free resources for evaluation, quality control, and arrangement but also to address the growing volume of electronic news data that needed processing), LAC transitioned to an automated workflow for the daily acquisition of ‘front-page’ news media. While we had the legislated authority to request them, scraping the front page news of the day was simpler and prevented us from losing important articles that would be moved behind the news media’s “paywall” after a day or two – which would complicate web archival acquisition and quality control.

But how did we as a national library decide what web-based newspapers or news media should be collected? The thinking on this is worth relating:

1. Print Newspaper Collection and Policy

In 2015-16, we were fortunate that an internal study had been conducted for the purposes of revising the print newspaper collection development policy, which provided us with a critical piece of data: circulation statistics. This spoke to readership, or how widely a given mainstream newspaper was consulted, and in what region. This influenced what we targeted for web editions.

---

8 Public Safety Canada 2013.
9 Public Safety Canada, 2016.
2. Digital Curation and Preservation Orientation

The Web and Social Media Preservation Program at LAC is situated within the Digital Preservation Division. As such, principles of digital curation and preservation sustainability are always in mind (and underpin) our web archival acquisition. Resources in formats for the web are collected faithfully as preservation objects from their original context on the web, in their “as-published” state, for the purposes of posterity and future research. Resources that are untenable for preservation, or discovery and access for the public (for any valid reason, e.g., due to use of proprietary software) are not collected.

3. Philosophic Legal Deposit Orientation

As mentioned, the federal web archiving program in Canada had its origins in the emerging ‘electronic’ wing of the Legal Deposit area. In terms of curation philosophy, our web archival collections development is aligned with, and derives lessons from, those of a national legal deposit program.

As outlined in UNESCO’s Guidelines for Legal Deposit Legislation:

“...Legal deposit may be defined as the requirement, enforceable by law, to deposit with one or more specified agencies copies of publications ...Publications most widely claimed for legal deposit are library materials...A continuing objective, dating from the first legal deposit law, the Ordonnance de Montpellier of 1537, has been the amassing of a national collection of books and related materials...maintained by the state for the benefit of the citizens...[thus] the first objective of a legal deposit law should be...the accumulation of a national collection in order to preserve, transmit and develop the national culture.”

Legal Deposit often strives to be comprehensive, knowing this is likely not achievable. However, considerable effort is expended to collect what national libraries consider the most valuable and perhaps representative content for posterity, preservation, and future access. It is difficult to predict what kind of research questions scholars will ask in the future, therefore it is incumbent on us to collect primary sources (and data) with widely varying perspective, utility, and focus – in order to support diverse modes of future digital and research inquiry. This philosophy applies to our web archival collection development and curation.

4. Library Collection Development Policy Approach

At LAC, evaluation and selection of web-based resources (indeed our approach to thematic research collections) proceeds according to the same principles of other library collection development policies for other medium or format types. Briefly summarized, the following factors are forwards in the minds of staff during curation activities:

Table 3: Library Collection Development Criteria for Web Archival Curation:

1. **Degree of Canadian content;**
2. **Representativeness of Canadian society;**
3. **Significance in relation to current or historical events;**
4. **Diversity of perspective** (ranging across the full political, regional, cultural spectra);
5. **Authority** (quality, factual accuracy, bias, and integrity of the information presented);
6. **Currency** (how up-to-date the information is; how pertinent it remains);
7. **Presentation context** (readability, organization of content, and physical and technical quality);
8. **Uniqueness** (how unique the perspective/information conveyed is—potentially owing to its lack of availability / in any other medium);
9. **Research Value** (likelihood of future consultation; ability to complement similar collections);
10. **Suitability** (of the subject matter, style, and language to the intended or target audience);

---

Larivière 2000, 6-7.
11. **Official Languages** (we strive to curate quality resources in both French and English as official languages of Canada, but also to include multicultural and multilingual resources);

12. **Risk of Loss** (While all web content is ephemeral, some types are at greater risk of loss. Media may be important to the Canadian historical record but may only briefly occupy the public attention, and/or be rapidly replaced by new content. Such media often exists for short periods of time before disappearing from the web permanently).

13. **Pace of Content Generation and Web Archival Scheduling** (how quickly a given news media source generates new content determines when a new crawl job is warranted, which in turn influences crawl schedules and frequencies, etc).

5. **Multidisciplinary Approach.**

At LAC, the Web and Social Media Preservation Program\textsuperscript{11} staff is comprised of digital librarians and archivists with diverse academic backgrounds that bring their professional library and archival science perspectives to the curation of the web archival collections. Some staff have backgrounds in history and geography (and some have digital humanities experience), which assists in building robust and responsive research collections. Further, we draw upon the expertise of our colleagues within the institution and across Canada as topically relevant, which assists in ensuring inclusivity, but also in providing an archival voice to communities that might not otherwise have such opportunity.

We should now turn to how social media is evaluated, selected, and acquired at LAC.

**Thematic Social Media Collection Methodology**

Social media has become an everyday aspect of life for many people, which serves as a primary mode of communication and information consumption (many today learn about major events from Twitter or Reddit before the mainstream media). As such, social media are now an important component of contemporary digital cultural heritage, particularly as they serve as evidence for public opinion, and as dialogue between citizens on all aspects of society, culture, and politics.

While collecting the traditional news media provides an important narrative of historical events, social media collected on the same event can potentially provide eyewitness accounts, first-hand interpretations, and additional contemporary opinions and perspective. In many ways, social media presents the opposite of the historian’s traditional problem: whereas he or she might have been frustrated with a dearth of primary source on any distant historical event, the modern digital historian is now presented with as many first-hand or eyewitness perspectives or interpretations as they care to parse! This makes the collection of social media highly complementary to mainstream or traditional news media sources.

While social media at LAC had been acquired since the Vancouver 2010 Olympic and Paralympic Games (mostly in the form of Canadian athlete’s Facebook pages), in 2015 we began focusing more particularly on conversations Canadians were having on Twitter about the pending October 2015 federal election.

Our preference for Twitter deserves some special comment. It should be mentioned that, as a general policy, where the content of a given social media source is repetitive across social media platforms, we have tended to prefer the easiest to collect in terms of sustainability, effort required, and return on that investment (and this usually became Twitter via the API, for which we have preferred the tool Tware). The volume of traffic generated on Twitter for the 2015 Canadian federal election was so great that several Canadian political hastags have been collected in an ongoing manner ever since (particularly the hashtags #canpoli, #cdnpoli). Twitter’s (previous) ease of collection and ubiquity also resulted in the development of a methodology whereby we tracked the generation of topical hashtags on a given project’s curation subject (including their pace of content generation), and thereby identified which

\textsuperscript{11} Library and Archives Canada 2023b.
were the most popular. These were then targeted for ongoing collection throughout the timeline of the web archival collection.

Today at LAC, social media is collected broadly, but primarily as a resource type in association with thematic and curated web archival research and data collections. Facebook, Instagram (Flickr), Reddit, Twitter (until the recent changes), and YouTube are primarily collected.

Our developments were timely and prepared us well for the onset of the several major global incidents, with astonishing geo-political and economic implications.

COVID-19 and Subsequent Events, Invasion of the Ukraine\(^1\)\(^2\)

By March 2020, it was clear that the coronavirus was going to have a serious impact on Canada and perhaps the world. Thanks to our automated web archiving of front-page news and our social media approach, LAC was in an established position to acquire this material upon the outbreak of COVID-19 as early as December 2019.

With some thirty-four (34) newspapers and media sources collected on a daily basis, our resources were free to pivot from other tasks to the evaluation and curation of some 2000+ non-media web resources on COVID-19 – which became the largest collection we have yet curated to date. As can be seen in the image below, we curated some fifteen subthemes for the COVID-19 collection ranging from academic research perspectives to impact on professional sports events. Many important issues developed as a result of the coronavirus pandemic including new issues in the “Black Lives Matter” movement (on which we selected fifteen hashtags), but also the Truck Convoy (or “Freedom Convoy”) Protests against COVID-19 policies that arguably started in Canada and spread worldwide (and on which we selected nine hashtags).

To document the events in Ukraine, we relied mostly on our media crawls and supplemented with the curation of an additional 150+ web resources providing Canadian perspectives (these resources were then contributed to the International Internet Preservation Consortium’s ‘War in Ukraine’ Collection).\(^3\)

Table 4: Statistics on COVID-19 and Ukraine Collections

<table>
<thead>
<tr>
<th>Event</th>
<th>Date</th>
<th>Result</th>
</tr>
</thead>
<tbody>
<tr>
<td>War in Ukraine</td>
<td>Feb 2022-present</td>
<td>175+ non-media resources, 300 gigabytes of data</td>
</tr>
</tbody>
</table>

Of course, there has been a great deal of dis-, mis-, and perhaps mal-information swirling actively around these issues. How have we at LAC managed them?

\(^{1}\) As the IFLA WLIC 2023 Media Session will concentrate on these issues, this paper has concentrated on methodologies instead.

\(^{2}\) IIPC 2022.
Our Strategy for “Fake News” or Extreme or Fringe Political Media

Given our legal deposit philosophy and our intentional representation of the full political spectrum of opinion in our web archives, at LAC we deliberately target and collect sources that might be considered ‘fringe, fake news, or extreme political viewpoint’, for use as source to support future inquiries and research in its own right. This has long been our tradition.

Such resources are freely available in our web archive, but in terms of acquisition and curation, we do separate them and present them for public access under the topic header “alternative perspectives” (sometimes also grouped with “protests” as we see here in the context of our COVID-19 collection). This avoids labeling them as “fake news”, while making clear that they do not reflect the stance of e.g., the mainstream media.

Additionally, we also take care to provide a generalized web archiving program disclaimer: “…collection of a web resource does not confer or reflect any implicit or explicit endorsement, accuracy, or agreement of the information contained in the targeted resources by LAC or the Government of Canada.”

Conclusion, Outcomes, and Government of Canada Web Archive Public Access Portal

At present, LAC’s Web and Social Media Preservation Program and web archival collections contain a modest ~9.3 terabytes of just social media, which includes some ~90 million tweets from Twitter as a preferred source.

The news media collections comprise ~21.6 terabytes at the time of writing, and LAC currently collects thirty-four (34) newspapers or news media sites on a daily basis (in English, French, multilingual and indigenous languages), on the basis of their representativeness, authority, currency, and their federal, provincial, regional, political, and cultural perspectives.

Lastly, and in conclusion, we are pleased to announce that our public discovery and access portal for our web archival collections has been re-launched:

---

14 Library and Archives Canada 2023a.
15 Library and Archives Canada 2023b.
16 The overall archive is a modest 120+ terabytes and 3.2 billion documents.
17 Ibid.
At present, the media collections relevant to e.g., COVID-19 have been arranged and can be consulted within that collection. In the future, we will develop a separate section for the consultation of our media collections, and continue to curate and release new collections.

Tom J. Smyth, Ottawa 2023

Acknowledgments

This paper is dedicated to Karen Krzyzewski, who tragically passed in the Ottawa Bus Accident on September 18th, 2013. Karen was one of my staff members and one of the earliest to believe in, and dedicate themselves to, the work of the LAC Web Archiving Program. She was also an early advocate for the disciplined collection of the news media. I hope she would be pleased with what the program and the GCWA have become.

The author would like to thank the current and past staff members of the Web and Social Media Preservation Program that have worked very hard to curate web archival collections of fantastic quality, but also to realize the milestone of the GCWA. I remain honoured and humbled to work with and for you.

REFERENCES


Please note:

Before your paper can be added to the IFLA Repository the Author Permission form must be completed here: https://survey.alchemer.com/s3/7184923/WLIC-2023-Author-Permission-Form