Archiving the Web

Working paper submitted to the CARL Committee
on Research Dissemination

September 8, 2014

Canadian Association of
Research Libraries

Association des bibliothèques
de recherche du Canada

Introduction

It is difficult to articulate the actual size of the web. It is vast and ever-changing as internet users the world over constantly add, change and remove content. Amidst all that material is a great deal of global cultural heritage being documented online. It behooves librarians, archivists and other information professionals to preserve the Web, at least as much of it as possible, for future generations.

In little over two decades, the Web has grown from being a relatively small service for scientists to an integral part of everyday life at an unprecedented rate. It began as a communication and research exchange hub for researchers and is now the pre-eminent global information medium everyone depends on. More than serving as a massive source of information, the web also constitutes a unique record of twenty-first century life of critical importance to current and future researchers. But the speed at which it develops, grows and transforms poses a definite threat to "our digital cultural memory, its technical legacy, evolution and our social history."[1]

Research libraries have a vital role in helping to preserve parts of our historical and cultural legacy on the web. In a published interview about the future directions of research libraries, University of Toronto Librarian, Larry Alford, remarked on the types of fundamental library values that will likely endure:

> Libraries are still very much about acquiring materials and preserving them, regardless of the format, so that they are still accessible hundreds of years from now. Many of the blogs and websites that led to the Arab Spring are now gone; they just disappeared. And yet pamphlets distributed in Paris by various factions during the French Revolution still exist, stored in libraries. We in libraries must begin to acquire and preserve the "pamphlets" of the 21st Century – blogs, websites and other digital commentary on the events of our time.[2]

Identifying, capturing, describing, preserving and making accessible parts of the Web which document that collective commentary on the "events of our time" fits libraries' mission to acquire, preserve and render visible and accessible our accumulated knowledge, history and culture. Libraries have long been in the business of preserving documentary heritage. That mission, which librarians have applied to both print and analogue audio visual media, equally applies to recorded knowledge and cultural expression recorded on the web. A key challenge is for libraries to nimbly adapt to constantly evolving web technology during a time of unprecedented change.

Background and other reasons for archiving the Web

The Internet Archive has been crawling and preserving the web since the late 1990s, but other organizations' involvement is crucial to the enterprise of web preservation. No single institution can realistically hope to collect and piece together an archival replica of the entire web at the frequency and depth that would be needed to effectively document entire societies', governments' and cultures' evolutions online. Only a hybrid approach, one that includes complimentary approaches involving the broad crawls the Internet Archive regularly conducts paired with heritage organizations' tackling of deep curated collections by theme or site, can ensure that a truly representative segment of the web is preserved for posterity.[3]

---

[1] Pennock, Maureen, *Web Archiving*, DPC Technology Watch Report, 13-01, March 2013, Digital Preservation Coalition, p. 3 http://dx.doi.org/10.7207/twr13-01

[2] Anderson, Scott, Search and Discovery, University of Toronto Magazine, June 19, 2013 http://www.magazine.utoronto.ca/life-on-campus/search-and-discovery-u-of-t-libraries-larry-alford/

[3] Grotke, Abbie, *Web Archiving at the Library of Congress*, Computers in Libraries, Vol. 31 No. 10, December 2011 http://www.infotoday.com/cilmag/dec11/Grotke.shtml

Approaches to web archiving vary from bulk or domain (e.g. all the web pages of a particular country's web domain, Iceland (.is) or France (.fr), selective, thematic to event-based projects. For example, the Library of Congress Web Archive includes over 250 terabytes of data comprising various event and thematic web collections.[4] Library and Archives Canada manages the Government of Canada Web Archive. By 2005, LAC had harvested the web domain of the Government of Canada. The GC WA contains over 170 digital objects comprising over 7 terabytes of data, though material harvested with crawls since 2007 are is not publicly available currently.[5] Government-maintained websites and web archives have high usage rates. Since 2009, *Bibliothèque et Archives Nationales du Québec* (BAnQ) has endeavoured to create deep curated thematic archived web collections. As May 2014, six collections containing web content about the 2012 provincial elections and the main political parties are accessible from BAnQ's website.[6] The UK Government Web Archive receives on average 100 million hits per month.[7] The Danish National Web Archive takes a snapshot of ".dk" websites four times per year. Researchers can observe how the Internet has developed as a whole, in Denmark. The archive preserves varied web material achieving a balance between text, images, and video content.[8]

Challenges to archiving the Web

There are other reasons for archiving the web aside from the imperative to preserve born digital documentary heritage of social, cultural and historical interest. According to various estimates, the average life spans of web pages range from 44, 75 to 100 days. Specific content vanishes, often as people update pages and move or delete content to make way for more up-to-date information. One can even view broken links and "404 page not found" messages as the modern-day equivalent of documents listed as "lost" in library catalogues but far more prevalent. Government agencies often have a legal obligation to preserve official records online. At this time the Canadian federal government is not legally obligated to preserve government websites nor are provincial governments. Canada's Depository Services Program (DSP) does collect PDFs from select agencies, there is not at present any organization conducting deep harvests of that particular type of content, and much of it has disappeared. LAC is taking steps to remedy that situation.[9]

Organizations face social, legal and technological challenges to web archiving. The very pervasiveness of the web presents a problem insofar as we just take it for granted. Information is simply readily available through a few keystrokes, inputting a query into a popular search engine, whenever we need it. Creators of web content do not necessarily create webpages and websites with preservation in mind. Frequent crawls to harvest web content are important to have snapshots of the web over time. It is largely up to librarians, archivists and other information professionals to proactively capture historically and culturally valuable content before it is lost.

Government information on the web presents a significant preservation challenge. With successive parties' governments coming into power and ceding it to their opponents or departments and agencies

[4] Library of Congress, Web Archive Collections http://www.loc.gov/webarchiving/collections.html#collect_05

[5] Library and Archives Canada, Government of Canada Web Archive http://www.collectionscanada.gc.ca/webarchives/index-e.html

[6] BAnQ, Archivage Web – curated collections: Coalition Avenir Québec (CAQ), Option nationale (ON), Parti libéral du Québec (PLQ), Parti québécois (PQ), Parti vert du Québec (PVQ), Québec Solidaire (QS) http://www.banq.qc.ca/collections/archives_web/?q=&i=&r=Parti+politique

[7] Pennock, Maureen, *Web Archiving*, DPC Technology Watch Report, 13-01, March 2013, Digital Preservation Coalition, p. 3-4 http://dx.doi.org/10.7207/twr13-01

[8] Kuchler, Hannah, "How to preserve the Web's past for the future," Financial Times, April 11, 2014 http://www.ft.com/cms/s/2/d87a33d8-c0a0-11e3-8578-00144feabdc0.html#axzz30ZWsVFEy

[9] Conversation with a member of the Canadian Government Information Private LOCKSS network (CGI-PLN)

being shut down, replaced or merging with other entities or organization with different mandates, the preservation of government information, reports and publications is by no means assured. Library staff at eleven institutions initiated the Canadian Government Information Private LOCKSS network October 2012.[10] The CGI-PLN's mission is to preserve digital collections of government information. Preservation entails that digital research materials remain accessible across geographically dispersed serves. Protection measures against data loss forward format migration are among the actions this group will perform to steward government information in Canada.

Four operating principles guide the CGI-PLN's work:

- Commitment to the long-term preservation of government information
- Application of the LOCKSS digital preservation software for preserving and replicating content in secure distributed servers
- Ongoing exploration of new digital preservation technologies and best practices
- Low-cost, sustainable preservation strategies that maintain sufficient capacity to accommodate large digital collections

A steering committee oversees the CGI-PLN's work. And a technical sub-committee advises the steering committee on technology and network capacity. Overall, it is a dispersed but highly motivated group. Members communicate frequently over e-mail and a collaborative group wiki. The steering committee meets on a quarterly basis. This initiative seeks to provide no-fee access to digital collections of government information that are "deemed to have enduring value" and include information from and about government, inter-governmental agencies, and non-government organizations. The targeted materials are publications or any digital content that federal or provincial departments or agencies produce. The CGI-PLN also intends to include digitized collections of government information and publications that member institutions host.[11]

Aside from problems owing to oversight or inertia, conflict also presents a significant threat the survival of archival, museum and library collections in areas of the world affected by war and political instability. The content on the Web is also vulnerable in such cases. A group of investigative reporters working in the Crimea, a contested area which has led to tense relations between the Ukraine and Russia in 2014, turned to the Internet Archive to safeguard their group's web pages and reports. The Crimean Centre for Investigative Journalism feared its reports could be taken off the internet at any moment given current civil unrest occurring in the Crimea. They entered the web addresses of their webpages and reports thereby "freezing the pages in time" and ensuring they collected new records each day.[12]

The legality of web archiving initiatives poses another significant non-technical challenge. Does an institution have the right, legally, to provide access to copies of web content independently of the original site and without the explicit consent of the owner or content creator? Can archiving web content in such cases constitute a breach of a site owner's copyright? That can depend on the country concerned or the collecting institution's remit. Use of Creative Commons licences and crown copyright can provide some clarity.

---

[10] The CGI-PLN's member institutions are: University of Alberta, Simon Fraser University, University of British Columbia, University of Calgary, University of Saskatchewan, University of Victoria, McGill University, Dalhousie University, Scholars Portal, University of Toronto, Stanford University, and the Legislative Assembly of Ontario.
[11] CGI-PLN http://plnwiki.lockss.org/wiki/index.php/CGI_network
[12] Kuchler, Hannah, op cit.

UK legal deposit legislation allows for selective, permissions-based web archiving coordinated at the British Library. The UK Archives' web preserving activity is smaller in scope; it has clear statutory permission to archive and provide access to crown-copyrighted material stemming from a legal mandate from the Public Records Act. In the US, the Library of Congress carries out much of its web archiving work on a permissions basis. The Internet Archive, without any explicit legislative mandate, operates largely on a "silence is consent" approach taking archived material down should the content owner request it. Other countries use legal deposit mechanisms but may opt to restrict access to reading rooms or even make use of "dark archives."[13] Whichever the approach organizations take to web archiving, they need to consider data protection and citizens' privacy rights as well.[14]

Different types of Web archiving

There are three types of large-scale web archiving: client-side archiving; transactional archiving; and server-side archiving.[15]

Client-side archiving is scalable and cost-effective with little input required by web content owners. Other characteristics:

- Involves use of web crawlers such as Heritrix or HTTrack which act like browsers using the HTTP protocol and gather content delivered from servers
- Crawler follows a seed instruction, crawls all links associated with the seed to a specified depth, captures copies of all available files

Transactional archiving is intended to capture client-side transactions rather than directly hosted content. Other characteristics:

- Supports growth of more comprehensive collections
- Enables user access recording
- Records client/server transactions over time
- Requires implementing code on the server hosting the content, and is more often used by content owners or hosts than external web content collecting agencies

Direct server-side archiving requires active participation from publishing organizations or the content owners. Other characteristics:

- Entails copying files from a server without using the HTTP protocol
- There are potential issues when it comes to generating working versions of websites – e.g. when trying to recreate a similar hosting environment to that of the original live site; this can be the case with database-driven websites
- This approach can be good for capturing content crawlers miss

[13] Dark archive: "(n.) In reference to data storage, an archive that cannot be accessed by any users. Access to the data is either limited to a set few individuals or completely restricted to all. The purpose of a dark archive is to function as a repository for information that can be used as a failsafe during disaster recovery." Webopedia http://www.webopedia.com/TERM/D/dark_archive.html

[14] Pennock, Maureen, *Web Archiving*, DPC Technology Watch Report, 13-01, March 2013, Digital Preservation Coalition, p. 10 http://dx.doi.org/10.7207/twr13-01

[15] Ibid, p. 7

It bears noting that web crawlers have their limitations. They can encounter difficulty in harvesting content from database-driven websites, streamed audio-visual files, certain java-scripted contents, and they cannot crawl the deep web's password-protected content.[16]

Quality control

Quality control is another important consideration for web archiving. Archivists', librarians' and other information professionals' curatorial role is needed to catch potential biases of web content, unfounded claims or even outright nonsense present in some content. Research and development may yield better web crawling technologies able to discern junk content from that which is valuable for research, but there will arguably always be a need for individuals to appraise web content in some capacity.

Web archiving in CARL and other libraries

Various CARL libraries have partnered with the Internet Archive, and are undertaking other web archiving initiatives. Some of the challenges they have to address have to do with staffing and resource limitations. An environmental scan the University of Alberta Libraries Digital Initiatives Unit conducted with a number of American university libraries and a few state libraries in 2011 is illustrative of the constraints most libraries must work within as they embark on web archiving projects.

The respondents reported that they did not have official procedures. Most employed "ad hoc" approaches for web archiving because of a shortage of full-time staff to do the work. In order to capture collections of time-sensitive and broad-in-scope web content, some institutions preferred to automate many of their processes. Most web archiving projects require the attention of 1 to 3 professional staff and 1 or 2 student assistants. State libraries mandated to preserve and provide access to government information and documents to the web can often allot more staff and time web archiving projects. The respondents of the University of Alberta Libraries' environmental scan relied mostly on Archive-It. They used Archive-It webinars for training purposes and sought personal assistance from Archive-It staff when needed. Some institutions can be limited in what they can realistically accomplish in this area having only one FTE whose attention is already divided among a number of priority areas in their work.[17]

In practical terms, at least one staff member with subject knowledge will select URLs but, depending on the size of the project, that individual might need to collaborate with other colleagues or hire student assistants to help carry out the work of building the collection. The same URL selectors assign metadata for pages in archived web collections independently or with in consultation with other colleagues possessing the required subject expertise. Student assistants sometimes receive some training to input metadata as well.[18]

University of Alberta Libraries, Simon Fraser University Library, UBC Library, University of Manitoba Libraries, University of Saskatchewan Library, University of Victoria Libraries and the University of Winnipeg Library participate in a COPPUL (Council of Pacific and Prairie University Libraries) Archive-It licensing deal. [19] Offered by the Internet Archive, Archive-It is a web-based application that enables users to build, manage, preserve and provide access to collections of web content. It is a fully hosted subscription service that also offers collection development tools for scoping, selection, and metadata

---

[16] Anthony, Adoghe, Web archiving: techniques, challenges, and solutions, International Journal of Management and Information Technology, Vol. 5 No. 3, September 2013, p. 602 http://www.cirworld.com/index.php/ijmit/article/view/532123

[17] Lau, Kelly E., *University of Alberta Born Digital Working Group Environmental Scan for Archive-It Partner Institutions Report*, September 2012, p. 4

[18] Lau, Kelly E., op cit., p. 6

[19] COPPUL, Archive-It Participants http://www.coppul.ca/dbs/view.php?dbid=214

input for cataloguing among others. Users are able to choose from 10 different web crawl frequencies. The archived content of collections using Archive-It includes HTML, PDFs, images and other document formats. Captured content is browsable within 24 hours after being archived, and it is possible to run full text searchers within 7 days.[20]

The way the software works is: Archive-It crawls websites and copies the information and the files embedded on selected websites. More specifically, it begins with specified seed URLs, it verifies that URLs are accessible and archives them. The crawler software checks the embedded contents (CSS, JavaScript, images, etc.). Archive-It searches for links to other webpages and archives them when they are in scope. A crawl continues until there are no more in-scope links and pages to capture or it reaches the maximum time allotted for the crawl or reaches a specified data limit.[21]

This service already supports metadata importing, restricted access to certain addresses and password protection if necessary, and in browsing quality assurance. And first quarter updates for 2014 support capture of media-rich web content including social media. The most recent update also includes visualization tools for analyzing archived collections.[22] For institutions with limited staffing and resources to bring to bear on web preservation projects, Archive-It is a reliable option developed by the Internet Archive – a non-profit (funded largely by research libraries and private donations) that focusses solely on preserving as much of the web as possible for future generations. Collaboration from Archive-It-subscribing libraries, however, is key complementing the Internet Archive's broad crawls and content captures with deeper curated collections. Both kinds of work will help ensure a more complete picture of the Web's legacy for future researchers.

The following two vignettes provide brief summaries of what the scope of a Canadian research library's web archiving activity might look like at this time.

University of Alberta Libraries

As stated on its website, University of Alberta Libraries uses Archive-It "to collect web content of importance to the U of A community that is at risk of being lost, deleted, or forgotten over time." The library currently provides access to 15 thematic collections of archived web content including the June 2013 Alberta floods, the Alberta Oil Sands, an energy/environment collection, the Idle No More movement, Canadian business grey literature collection, and Canadian health grey literature among others. The collections comprise web material of enduring value and cover important regional events or issues. For example, June 2013 saw some of the worst flooding in Alberta's history, with areas in the southern portion of the province being most affected. The *Alberta Floods June 2013* collection captured 165 websites that detail the events as they happened, their impact on communities, and the recovery efforts. The *Prairie Provinces Politics & Economics* collection captures sites related to the Canadian Prairie Provinces' politics, economics, society, and culture. The collection focuses on Alberta though it also covers of Saskatchewan and Manitoba. *La francophonie de l'ouest canadien / Western Canadian*

Francophonie is another Archive-It collection UAL is developing, it archives websites chronicling life in the four Prairie Provinces' francophone communities.[23]

---

[20] Reed, Scott, *Archive-It: a web archiving service of the Internet Archive since 2006*, presentation at the 2013 conference of the Association of Canadian Archivists

[21] Ibid

[22] Ibid

[23] University of Alberta Libraries, Archive-IT collections, https://archive-it.org/organizations/401

Bibliothèque et Archives Nationales du Québec (BAnQ)
Since 2009, BAnQ has been selectively archiving websites created in the province of Québec. As of September 2014, the library has preserved 39 collections covering the 2012 provincial election, 2013 municipal elections, political parties in the province, captures of various Quebec cities' and towns' websites, and two more governmental websites: the National Assembly, and the Ministry of Municipal and Regional Affairs. This particular collection of archived web content has grown fairly quickly, from just about a half dozen sites, in May 2014, to 39 just four months later. BAnQ makes every effort to maintain the original design and structure of the content as it was originally created. That being the case, the library still places a brief disclaimer to the effect that, although the essential design and information architectures of the preserved sites remain largely intact, the user may still encounter some anomalies when consulting these particular digital collections.[24]

Survey – Archiving the Web
Over the summer, the CARL office polled the membership with just one question: whether or not its member libraries have embarked on any kind of web archiving projects. Based on the information responses received, a subset of 14 libraries (including a non-CARL institution) was sent a follow-up survey comprising just five questions (See **Appendix 2**). The questionnaire[25] was intended to provide a general picture of the kinds of activities CARL libraries are pursuing in developing collections of content captured from the open web. Only four libraries responded, for a response rate of 28%. The resulting small sample, can only give an approximate impression of what Canadian research libraries are doing around web archiving. **Appendix 1** shows the aggregated survey responses.

All four responding libraries are currently undertaking web archiving work to support research and teaching at their institution. They capture, describe, and make the content discoverable in a few different ways. It is not possible to convey the full scope of the kinds of web material preserved. One participant indicated that their library is focusing its efforts on various Government of Canada websites. Three harvest content using Archive-It, and the other employs the Heritrix web crawler and Wget.

The number of staff who devote some of their time to web archiving projects is limited. Two libraries each have one staff member involved in this work, one has five, and the other respondent indicated that subject specialist librarians carry out the work with the assistance of support staff and technical services staff (both as needed), but did not specify how many staff members are involved and remarked that the web archiving-related activities "are only one small portion of the work of all these individuals".

One responding library is carrying out the work internally, another has engaged several faculty members "to help build subject-specific collections", another has collaborated with a few academic staff who have an interest in web archiving. The latter is also exploring opportunities to pursue further web preservation with the library school at their institution as components of one or more classes – e.g. collection development, digital preservation, etc. The fourth responded is involved in collaborative GoC website harvesting at the national level with several libraries. The nature of web archiving work typically requires that it be done collaboratively.

[24] Bibliothèque et Archives Nationales du Québec (BAnQ), Archivage Web http://www.banq.qc.ca/collections/collections_patrimoniales/archives_web/index.html?q=&r=Gouvernemental
[25] Survey questionnaire adapted from Abbie Grotke, *Web Archiving at the Library of Congress,* and revised with feedback from the University of Alberta Libraries (with CARL's appreciation).

In terms of cataloguing captured web content, approaches among the respondents seem to vary: from using the Islandora Web ARChive Solution Pack, applying Dublin Core metadata, exploring how the data might be rendered discoverable in a discovery layer using OAI-PMH, to having full catalog records created through the integrated library system (ILS) at the document, seed and collection level. The survey participants enable browsing and searching of web collections using features provided by Archive-It or the Islandora Web ARChive Solution Pack. One library's browsing and searching solution is still in development. These solutions also appear to integrate with other digital collections and various finding aids. One respondent said that their library has created a webpage that provides search functionality across all the web archive collections created so far.

It is still very much the early days for this kind of work, preserving select content from the open web. The COPPUL collaborative approach, using Archive-It, and the Canadian Government Information Private LOCKSS Network are two good collaborative models libraries can continue to work from to help preserve important research, historical and cultural information that is currently at risk of being lost.

Works cited

Anderson, Scott, "Search and Discovery", University of Toronto Magazine, June 19, 2013 http://www.magazine.utoronto.ca/life-on-campus/search-and-discovery-u-of-t-libraries-larry-alford/

Anthony, Adoghe, Web archiving: techniques, challenges, and solutions, International Journal of Management and Information Technology, Vol. 5, No. 3, September 2013 http://www.cirworld.com/index.php/ijmit/article/view/532123

Canadian Government Information Private LOCKSS network (CGI-PLN) http://plnwiki.lockss.org/wiki/index.php/CGI_network

Grotke, Abbie, *Web Archiving at the Library of Congress*, Computers in Libraries, Vol. 31 No. 10, December 2011 http://www.infotoday.com/cilmag/dec11/Grotke.shtml

Kuchler, Hannah, "How to preserve the Web's past for the future," Financial Times, April 11, 2014 http://www.ft.com/cms/s/2/d87a33d8-c0a0-11e3-8578-00144feabdc0.html#axzz30ZWsVFEy

Lau, Kelly E., *University of Alberta Born Digital Working Group Environmental Scan for Archive-It Partner Institutions Report*, September 2012

Pennock, Maureen, *Web Archiving*, DPC Technology Watch Report, 13-01, March 2013, Digital Preservation Coalition, p. 3 http://dx.doi.org/10.7207/twr13-01

Reed, Scott, *Archive-It: a web archiving service of the Internet Archive since 2006*, presentation at the 2013 conference of the Association of Canadian Archivists

Appendix 1 – Survey questionnaire: Archiving the web

Respondent	1	2	3	4
1.) Describe your organization's current web archiving activities. Include, at a minimum, the following information: **Summary of activities**	Capturing, preserving, and disseminating websites.	Building collections in Archive-it to support research and teaching. Most crawls are not yet publically available as we sort out permission issues, etc., with university lawyer	We have been involved in web archiving since 2009, seeing this as a component of our collection development activities. We archive relevant web content, provide descriptions, and make it discoverable in a variety of ways.	Government of Canada selected websites, institutes and conferences
What web archiving tools are used to harvest content?	Heritrix and/or Wget	Archive-it	Archive-it	Archive-it
How many of your staff members are involved in this work?	1	5	Work to develop and maintain collections is carried out by subject librarians with support staff assistance, descriptive work is carried out by librarians and support staff in bibliographic services, technical support and training for those involved is shared among a small group of support staff. These web archiving related activities are only	Currently one

			one small portion of the work of all of these individuals.	
Is the work collaborative with other groups on campus, other libraries or organizations (e.g. campus IT staff, an academic department, a library consortium, another organization)?	Just our local unit.	Not yet. At the present , this is a library-spefici initaitive, although we have engaged with several professors to help us build subject-specific collections.	We have worked in a preliminary way with one or two academic staff members who have an interest in web archiving for one reason or another, and have also explored opportunities with our Library School around web archiving as a component of one or more classes (e.g., collection development, digital preservation). We are also partnering very closely with our Humanities Computing department to examine web archiving for text mining and visualization, and have recently hired a CLIR postdoctoral fellow (joint appointment with Humanities Computing) to focus on this the research potential of web archiving.	GoC web harvesting is being conducted cooperatively with several libraries nationally. COPPUL was able to negotiate a consortium agreement on our behalf.
2.) What sorts of tools do	(No response)	N/A	Spreadsheets	Subject librarians appraise

you use for the following: **Web content appraisal?**				web content in the way they would any other materials for our collections. Overall approval for new web archive collections comes from our system-wide collection development committee and is based on our overall collections policy.
Web content selection?	(No response)	N/A	Spreadsheets	Subject librarians monitor web content within their areas and select as they feel appropriate. We have a web form for proposing new web archive collections which gathers details which are then reviewed by a collections manager and taken to our system-wide collection development committee for approval. We also have a web form that anyone (including the public) can use to recommend a website for inclusion within our web archive. The details are provided to the appropriate individuals who will determine

				whether or not the site is included in our archive.
Work flow management?	(No response)	N/A	Archive-it & email	We use Google docs for sharing planning and workflow documents and email lists for shared communications.
3.) Do you currently have any collection / selection policies that apply to web content? **Subjects / Provide a few examples:**	(No response)	Government of Canada Websites of institutional interest. This includes: DFO, CRTC, CBCA and the National Atlas of Canada.	(No response)	(No response)
Types of sites (e.g. - government, commercial, institutional, individual, etc.) / Provide a few examples:	(No response)	See above.	(No response)	(No response)
Any other criteria that focus the scope of your web archiving activity? Please specify:	(No response)	Common records schedule	Our web archiving policy can be found at http://bit.ly/1A5EVmS It adheres to the principles of our overall collection development policy which can be found at http://bit.ly/1rMylTU	Transitory -- in danger of disappearing

4.) Are there any policies at your organization that determine access rights and permissions required for content crawled? For example, do you have a lawyer or someone with legal and/or copyright expertise on your project team or at least who provides advice as needed?	Common records schedule goes through legal.	We have a draft policy based on that of another institution, but it is still under review.	Details on access and permissions can be found in our web archiving policy (see response to question #3). This was informed through discussions with our copyright office as well as campus legal counsel.	No; GoC websites have a fairly liberal license
5.) How is access for archived web content handled by your organization in terms of the criteria below? **Cataloguing?**	Islandora Web ARChive Solution Pack	DC metadata applied to collections, seeds, and documents as needed. We're looking at how we might expose this data via OAI-PMH in our discovery layer (Summon)	Some materials have had full catalogue records created and included in our ILS, mainly at the document level, though some at the seed and collection level. We also create metadata within the Archive-It tool at the various levels (document, seed, collection) as deemed most appropriate. We make our metadata in Archive-It available via OAI.	In development
Browsing and searching?	Islandora Web ARChive Solution Pack	Via Archive-it	Our collections are available through the Archive-It interface where	In development

			browsing and searching is available.	
Integration with existing digital collections, the library website, lib-guides and other finding aids?	Islandora Web ARChive Solution Pack	We have a page on our website dedicated to Archive-it	We have a page on our website that provides a search across all of our web archive collections, as well as direct links to individual collections. Many of our subject librarians include links to and/or search widgets for relevant web archive collections within their Libguides.	In development

Appendix 2 – Survey questionnaire: **Archiving the web**

Archiving the web

1.) Describe your organization's current web archiving activities. Include, at a minimum, the following information:

Name of institution:

Summary of activities:

What web archiving tools are used to harvest content?

How many of your staff members are involved in this work?

Is the work collaborative with other groups on campus, other libraries or organizations (e.g. campus IT staff, an academic department, a library consortium, another organization)?

2.) What sorts of tools do you use for the following?

Web content appraisal?

Web content selection?

Work flow management?

3.) Do you currently have any collection / selection policies that apply to web content? Check all that apply.

☐

Subjects / Provide a few examples:

☐

Types of sites (e.g. - government,commercial, institutional, individual, etc.) / Provide a few examples:

Any other criteria that focus the scope of your web archiving activity? Please specify:

4.) Are there any policies at your organization that determine access rights and permissions required for content crawled? For example, do you have a lawyer or someone with legal and/or copyright expertise on your project team or at least who provides advice as needed?

5.) How is access for archived web content handled by your organization in terms of the criteria below?

Check all that apply, and provide a short description for items checked.

Cataloguing?

Browsing and searching?

Integration with existing digital collections, the library website, lib-guides and other finding aids?

Submit