Revue COSSI

N° 11 | 2022
Préservation de la mémoire du Web en temps de crise

La création d’une archive web pour faciliter la recherche future sur la COVID-19 au Canada

Préservation numérique des sites web documentant les événements historiques nationaux à Bibliothèque et Archives Canada

Tom J. SMYTH

Édition électronique :
URL :
https://revue-cossi.numerev.com/articles/revue-11/38-la-creation-d-une-archive-web-pour-faciliter-la-recherche-future-sur-la-covid-19-au-canada
DOI : numerev_1834
Date de publication : 10/01/2022

Cette publication est sous licence **CC BY-NC-ND** (Attribution - No commercial - No derivatives).

Pour **citer cette publication** : J. SMYTH, T. (2022) La création d’une archive web pour faciliter la recherche future sur la COVID-19 au Canada . *Revue COSSI*, (11). https://doi.org/10.34745/numerev_1834

Cet article traitera de l'autorité législative du Programme de préservation du web et des médias sociaux à Bibliothèque et Archives Canada et présentera les stratégies d'acquisition et de projet pour la préservation des sites web documentant les événements historiques nationaux, en accordant une attention particulière à la création d'une archive web sur la COVID-19 et son impact sur le Canada.

Il expliquera comment et pourquoi l'archivage web est devenu une méthode importante et essentielle pour les organisations pour documenter les événements historiques nationaux, et discutera brièvement de la façon dont nous avons tenté de répondre aux attentes et aux exigences des principaux intervenants du programme. Cet article traitera ensuite de l'importance de méthodologies rigoureuses pour établir un ensemble de données de recherche, y compris le rôle potentiel du biais de sélection dans la continuité informationnelle et la préservation numérique, et comment cela peut être amélioré en définissant les objectifs et la portée des archives web thématiques au moyen de méthodes de développement des collections de bibliothèques. L'article discutera ensuite de la façon dont notre expérience à ce jour nous a préparés à documenter la COVID-19 et fera état de nos progrès à ce jour.

Enfin, l'article abordera la façon dont un programme d'archivage web peut aider les chercheurs à utiliser les archives web pour répondre à leurs questions de recherche.

Mots-clés :
Archivage web pour les événements historiques, Stratégies pour gestion de projets et d'acquisitions de collections web, Utilisation des archives web pour faire de la recherche, Préservation numérique du web

---

Le contexte législatif et la méthodologie d'acquisition des ressources web à Bibliothèque et Archives Canada

Sous quelle autorité Bibliothèque et Archives Canada (BAC) recueille-t-elle l'Internet canadien ?

Selon la Loi sur la Bibliothèque et les Archives du Canada (L.C. 2004), section 8(2) :

> **Attributions de l'administrateur général**

> 8 (1) L’administrateur général peut prendre toute mesure qui concourt à la réalisation de la mission de Bibliothèque et Archives du Canada et, notamment :
>
> **Réalisation d’échantillons à partir d’Internet**
>
> (2) Pour l’application de l’alinéa (1)a), l’administrateur général peut, à des fins de préservation, constituer des échantillons représentatifs, selon les modalités de temps ou autres qu’il détermine, des éléments d’information présentant un intérêt pour le Canada et accessibles au public sans restriction dans Internet ou par tout autre média similaire.

La loi était visionnaire à l’époque et dérivait du rôle de BAC en tant que membre fondateur du Consortium international pour la préservation d’Internet (IIPC) en 2003. Selon l’IIPC, l’archivage web est le processus qui consiste à acquérir des parties du web, à effectuer la préservation numérique de ces données et à les rendre accessibles et utilisables. Ainsi, l’objectif de l’archivage du web à BAC est de préserver et d’assurer l’accès futur aux ressources web qui sont un aspect essentiel du patrimoine documentaire du Canada.

Le programme de Préservation du web et des médias sociaux est la méthode principale pour réaliser le mandat de BAC en lien avec l'article 8(2) de la loi. Pour l’acquisition du patrimoine documentaire numérique du Canada publié sur le web, nous avons élaboré un programme avec cinq activités stratégiques principales :

1. Collection complète de la présence web du Gouvernement du Canada (depuis 2005-)
2. Collections de recherche thématique sur le web et les médias sociaux (2009-)
3. Documentation des événements imprévus d'importance historique nationale (2013-)
4. Récolte de sauvetage ou de conservation (2005-)
5. Acquisition des ressources nominées (2005-)

La création d’archives web sur la COVID-19 et ses répercussions sur le Canada fait appel aux stratégies et aux méthodologies inhérentes aux activités numéros 2 et 3.

Pourquoi l’archivage web est-il important pour les événements historiques nationaux ? Pour qui organisons-nous une archive web sur la COVID-19 ?

La COVID-19 a démontré que l’archivage web est l’une des rares mesures que les professionnels de l’information peuvent prendre immédiatement pour préserver un historique complet et ses ressources numériques clés.

Cependant, l’archivage du web est rarement un simple geste de collecte de données. Il existe des préjugés humains (conscients et inconscients) dans le choix des ressources à

inclure dans toute collection de recherche, et les archives web ne font pas exception à cette règle. Comment le caractère et la facilité d'utilisation de l'ensemble de données recueillies sont-ils influencés par cela? De quelle façon la sélection des ressources à inclure dans la collection influe-t-elle sur les sujets qu'elle peut étudier ? (Milligan, 2019).

Il est important de savoir que ces biais ont une incidence directe sur la sélection des données et, par conséquent, sur la continuité de l'information. Lorsque nous considérons la nature dynamique et précaire du web, nous devons être conscients du fait que notre sélection d'une ressource web pour l'inclusion dans une archive web pourrait être la seule action de préservation numérique que la ressource reçoit. Cela pourrait donc déterminer si la ressource survit et si elle est accessible aux chercheurs de l'avenir, ou si elle est perdue ou oubliée. Ainsi, les archives web et l'action de leur conservation numérique peuvent fonctionner comme une sorte de « bateau de sauvetage numérique » dans l'océan des mégadonnées. Ce bateau pourrait déterminer quelles voix survivent pour témoigner des perspectives et de l'histoire d'une telle communauté.

Peu importe à quel point nous prétendons être objectifs, ces préjugés demeurent. En plus de la nécessité de lutter contre les préjugés, nous sommes également limités par les ressources dont nous disposons et nous devons parfois prendre des décisions difficiles au sujet de nos collections. Cependant, il y a des stratégies que nous pouvons élaborer et déployer pour essayer de créer des archives web qui sont aussi inclusives que possibles.

Depuis 2005, les principaux partenaires et intervenants du programme de préservation du web et des médias sociaux de BAC sont les bibliothèques universitaires qui s'intéressent aux données fédérales et aux documents officiels, les chercheurs professionnels et le public. Qu'est-ce que les chercheurs espèrent trouver pour faciliter leur recherche à travers une masse de données telle que les archives web ? Autrement dit, dans 20 ans, lorsqu'un historien rédigera l'histoire de la COVID-19 et de ses répercussions sur le Canada, quelles sources et données principales aimerait-il avoir ?

Nous devons nous poser cette question lorsque nous voulons organiser une collection spéciale de ressources web pour documenter quelque chose de précis, comme un événement historique. Il est également important d'examiner comment les ressources web pourraient être utilisées dans les futurs contextes de données de recherche (Smyth, 2022a). Nous pourrions explorer textuellement (textually mine) le contenu des archives web, des visualisations pourraient être produites, des sujets obscurs réécrits, des hyperliens analysés pour leurs associations à d'autres sites, des sentiments textuels et des images pourraient être étudiés.

Les chercheurs sont très intéressés à avoir tout le domaine de premier niveau national canadien (c.â□□àâ□□d. *.ca) pour la recherche (Milligan et Smyth, 2019). Lorsque cela n'est pas possible, en raison d'un manque de ressources ou de capacité et qu'un certain sous-ensemble du domaine doit être sélectionné et acquis, nous devons nous assurer de

saisir un ensemble de données aussi vaste et diversifié que possible – un peu comme les principes et les méthodologies régissant le dépôt légal à une bibliothèque nationale. Mais en même temps, il est également important de noter que ce ne sont pas tous les créateurs de contenu canadien, par exemple, qui souhaitent ou ressentent le besoin d'enregistrer leurs sites web utilisant un domaine .ca. Par conséquent, le contenu web canadien ou le contenu d'intérêt pour le Canada ne font pas tous partie du domaine .ca (Milligan et Smyth, 2019; Webster, 2019). Cela montre qu'une stratégie hybride de collecte de domaine et de ciblage de contenu pertinent à l'extérieur du domaine est idéale et nécessaire (Milligan et Smyth, 2019; Webster, 2019).

Stratégies pour documenter les événements historiques nationaux : comment notre expérience, à ce jour, nous a-t-elle préparés à documenter la COVID-19 ?

Pendant la pandémie, il était important de gérer le travail du projet d'archivage web avec plus de précision pour nous assurer que nous concentrions nos ressources limitées sur les sujets et les activités prioritaires, afin d'obtenir un rendement maximal. Pour ce faire, nous avons développé notre méthodologie de gestion de projets d'archivage web, basés sur les principes de développement des collections de bibliothèques, pour l'organisation et la conservation des collections d'archives web thématiques.

Au début de 2016, après avoir documenté plusieurs événements historiques nationaux (p. ex., les feux de forêts catastrophiques au Canada, le mouvement « Idle No More » en 2012, la catastrophe ferroviaire du Lac-Mégantic en 2014, l'attaque sur la Colline du Parlement en 2014 et la mort de Leonard Cohen en 2016), le programme a décidé d'élaborer une méthodologie pour évaluer les sites web des médias et des journaux, dans les deux langues officielles et représentant différentes perspectives politiques et régionales. Nous avons ensuite choisi des sites web ayant un volume élevé de production de contenu, qui ont démontré une architecture web qui convient à la récolte et à la préservation numérique, et qui ont également la plus grande distribution, et nous avons commencé à cibler les pages de couverture pour la collecte quotidienne :

Acadie Nouvelle	Global News	La Presse	Toronto Star
Bloomberg Canada	Globe and Mails	Le Devoir	Vancouver Sun
CBC	Huffington Post Canada	Le Droit	Whitehorse Daily Star
City News Toronto	iPolitics	National Post	Winnipeg Free Press
CTV News	Journal de Montreal	Radio-Canada	Macleans - COVID-19
Financial Post	Journal de Quebec	Rabble.ca	

Cette méthode a été adoptée pour veiller à ce que la chronologie historique de tout événement historique national imprévu soit immédiatement documentée à mesure que les détails apparaissent dans les médias grand public partout au Canada. Cela a libéré du temps et des efforts au sein du programme, qui peut maintenant être consacré à la recherche, à l'évaluation, à la sélection et à l'acquisition de ressources web plus précieuses au lieu de répondre de façon ad hoc à la collection de médias. Au moment de la rédaction du présent rapport, nous avons mené environ 4 535 opérations techniques de collecte et d'analyse web (web archival crawls), recueillant environ 656 millions d'objets représentant environ 18 téraoctets de données, soit 25 % de nos archives totales – et en croissance quotidienne.

En conséquence directe, les collections d'événements historiques nationaux depuis 2016 sont beaucoup plus riches et comprennent des ressources plus spécialisées qui n'auraient pas été incluses et préservées auparavant.

Le programme a commencé à recueillir passivement des médias aux premiers stades de la pandémie de COVID-19 dès qu'elle a été signalée en décembre 2019. Cela nous a permis de définir de façon beaucoup plus précise les objectifs de conservation des archives web sur la COVID-19 et de réfléchir plus largement aux sujets que nous voulions documenter.

Définir les objectifs et la portée d'une archive web sur la COVID-19

Comme toute autre collection de bibliothèques, il est essentiel d'avoir une politique de développement de la collection qui définit les objectifs, les sous-thèmes qu'elle contient, l'ordre ou la priorité du développement et le degré d'exhaustivité prévu. Pour l'archivage web et selon notre méthodologie, nous définissons la portée (et par définition ce qui est délibérément hors de portée) de chaque sous-thème, métadonnées spécialisées ou vocabulaire contrôlé appliqué, et un aperçu du degré de contrôle de la qualité à effectuer par sous-thème.

En définissant, en surveillant et en contrôlant l'effort maximal à investir dans chaque sous-thème de collecte en matière de nombre total de ressources web à acquérir, d'activités techniques à réaliser et du nombre maximal d'équivalents temps plein (ETP) à investir en fonction du niveau de contrôle de la qualité assignée, les progrès peuvent être surveillés et ajustés au besoin. Ceci est particulièrement important pour les opérations de contrôle de la qualité pour les grandes collections puisque, comme décrit dans une présentation pour la conférence annuelle d'IIPC, le contrôle de la qualité est comme un trou noir et peut absorber des ressources infinies (Smyth 2021) ! En même temps, il faut demeurer flexible et être en mesure de s'adapter et d'accueillir de nouveaux sous-thèmes à mesure que l'événement national évolue et que de nouveaux contenus qui justifient la collecte sont générés.

En termes d'acquisition de médias sociaux et de Twitter en particulier, que nous

préférons collecter en utilisant l'API et le logiciel ouvert « Twarc », nous réalisons également des gains d'efficacité en analysant des « hashtags » topiques et canadiens de premier plan, puis en les recueillant par thème en fonction du taux le plus élevé de production de contenu, plutôt que de cibler les comptes. Cette méthode était particulièrement importante pour documenter le « Convoi de la liberté » (« Convoi de camionneurs ») pendant la pandémie, où le dialogue principal sur ces questions a eu lieu sur les médias sociaux (Smyth, 2022b).

Donc, les sujets généraux suivants ont été définis et documentés par l'acquisition et la préservation de ressources web pertinentes (tableau 1) :

Tableau 1. Résumé des thèmes documentés dans la collection d’archives web de BAC sur la COVID-19

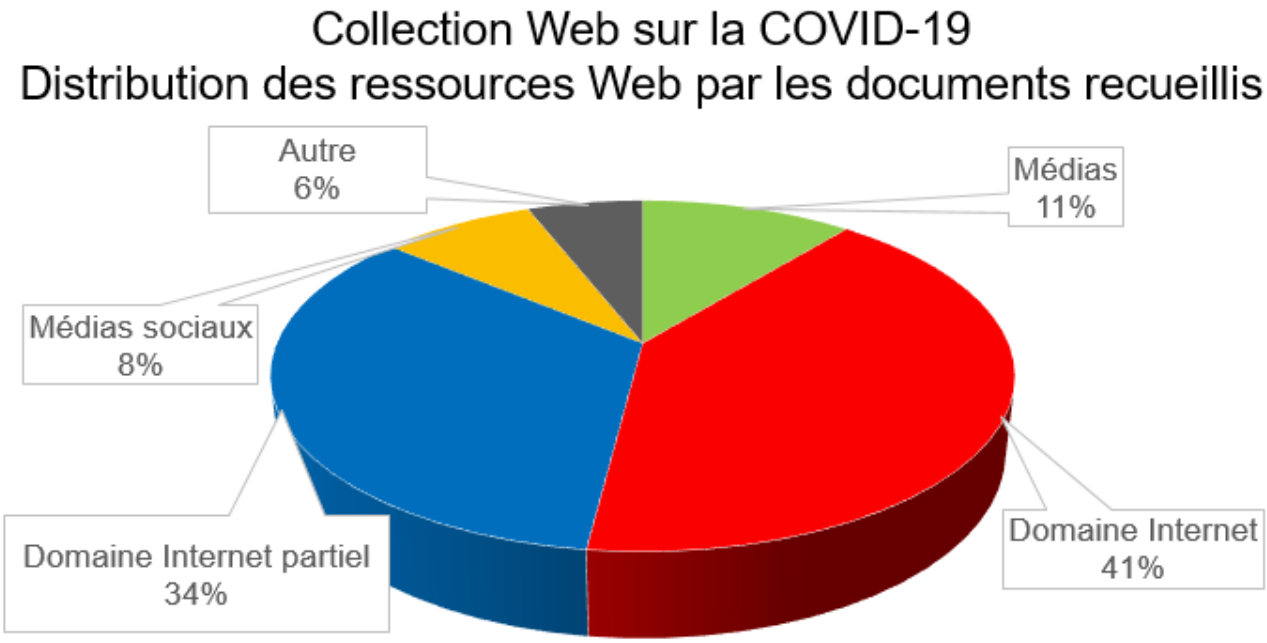
Grands sous-thèmes de la collection de la pandémie :	# Ressources :
Sciences de la santé	587
Arts et culture	302
Gouvernement fédéral, provincial et territorial	257
Organismes de bienfaisance	209
Affaires et économie	195
Protestations et autres points de vue (p. ex., convois de camionneurs)	141
Santé publique régionale	119
Familles et éducation	115
Religion	88
Perspectives autochthones	35
Total	2048

En ce qui concerne les progrès et les résultats à ce jour, et au moment de la rédaction, notre collection comprend actuellement (tableau 2 et figure 1) :

Tableau 2. Statistiques d’acquisition pour la collection d’archives web de BAC sur la COVID-19

Activités d’acquisition :	Total :
Nombre de sites de médias acquis quotidiennement	34
Total des ressources web sélectionnées	~2,048
Total des documents numériques recueillis	~478 millions
Données totales recueillies	~16 TO
Tweets saisis pour la collecte liée à la COVID-19	~3,83 millions

Figure 1. Distribution de ressources d'archives web par documents recueillis pour la collection BAC COVID-19



Aider les chercheurs à utiliser les archives web : vers un instrument de recherche d'archives web pour les collections complexes (comme la COVID-19)

A surefire way for a historian to recognize the value of the archival or library profession is to suddenly be confronted with the vast data of a web archive. Many of the problems confronting a web archive researcher result from suddenly not having the professional framework and infrastructure from which historians studying earlier time periods benefitted (Milligan, 2019, p. 213).

Étant donné notre rôle historique dans l'organisation, la compréhension et l'utilisation des ressources de recherche, comment les bibliothécaires et archivistes numériques travaillant comme spécialistes de la préservation et conservation numérique peuvent-ils aider les chercheurs de la prochaine génération à accéder aux données brutes et aux archives web, c.-à-d. aux données qui deviendront inévitablement la principale source pour les historiens des XXe et XXIe siècles ?

Comme réponse préliminaire, et comme évolution de la politique de développement de la collection thématique web, nous avons voulu transformer ce document interne sur la gestion du projet en un outil de recherche d'archives web pour le chercheur, et le publier avec la collection web par l'entremise du portail d'accès public des Archives web du gouvernement du Canada.

Ce court document pourrait alors servir de guide à la collection d'archives web et aux données thématiques pour tout chercheur potentiel. Il fournirait une définition des

thèmes que nous essayons de documenter, toutes les métadonnées spécialisées qui pourraient être utilisées pour faciliter la recherche en texte intégral (p. ex., par Library of Congress Subject Headings (LCSH) et par la terminologie descriptive sensible et faisant partie intégrante des ressources autochtones dans le cas des archives web de la Commission de vérité et réconciliation), un niveau de contrôle de la qualité attribué par thème est à prévoir, la liste complète des « graines d'archivage web » (web archiving seeds, c.-à-d. les adresses des sites web à collecter), ainsi que les tableaux de distribution des ressources par langue.

L'instrument de recherche des archives web devrait éclairer le chercheur en sciences humaines numériques et en histoire numérique d'un coup d'œil et répondre à la question suivante : « Cet ensemble de données ou cette collection d'archives Web seront-ils utilisés comme principale source historique pour ma question de recherche? » -- sans avoir à faire de recherche sur la source pour déterminer quels types de ressources ont été saisies, et/ou pour déterminer dans quelle mesure l'ensemble de données est complet ou représentatif. Nous pourrions ensuite faire évoluer le projet et ce document au fil du temps, en ajuster la portée à mesure que de nouvelles ressources sont générées et que de nouvelles questions pertinentes se posent, et documenter la collection et ce qu'elle offre aux chercheurs.

Il convient également de souligner que cette même méthodologie est actuellement appliquée à la documentation et aux réactions du Canada aux enjeux actuels en Ukraine.

Conclusion

De nombreux efforts sont en cours pour éliminer les dépendances techniques à l'utilisation informatique des archives web comme données et comme ressources textuelles pour la recherche.

Comme premier exemple, l'intégration des outils logiciels du projet « Archives Unleashed » et le partenariat entre ce groupe et Internet Archive, permet désormais d'introduire :

ARCH (Archives Research Compute Hub), the first cloud-based system designed from scratch to meet all of these six key principles [of Archive, Big Data, Concurrent, Distributed, Efficient, Flexible]. ARCH is an interactive interface, closely connected with Archive-It, engineered to provide analytical actions, specifically generating datasets and in-browser visualizations. It efficiently streamlines research workflows while eliminating the burden of computing requirements [for researchers]. Building off past work by both the Internet Archive (Archive-It Research Services) and the Archives Unleashed Project (the Archives Unleashed Cloud), this merged platform achieves a scalable processing pipeline for web archive research…ARCH's interface consists of four levels. These guide users to interact with their collections by generating datasets for analysis and engaging

with in-browser features. The goal of ARCH is to provide an efficient, streamlined workflow without burdening users with computing requirements or actions...ARCH has been designed as an integrated component of Archive-It (Holzmann et al., 2022, p. 1; 6-9).

Ce développement passionnant fera à son tour partie de la prestation de services qui sera désormais disponible pour les institutions qui utilisent la plateforme Archive-IT.

Bien que les méthodologies de « réponse rapide » existaient déjà, pendant la pandémie de 2019, l'archivage Web est devenu une priorité stratégique et une méthodologie pour la saisie, l'organisation et la conservation numérique, et la découverte et l'accès à des ressources du web qui constituent la principale source et la preuve de l'impact historique national de la COVID-19 sur chaque nation. C'est devenu une priorité d'action en reconnaissance du fait que les documents d'archives et les publications pourraient autrement prendre un temps indéterminé à produire et à devenir accessibles aux chercheurs.

Remerciements

L'auteur tient à remercier les membres de l'équipe du Programme de préservation du web et des médias sociaux et la section d'Intégration numérique pour leur professionnalisme, leur dévouement et leur travail acharné dans l'élaboration du programme d'archivage web à BAC pendant la pandémie et depuis 2009.

Références bibliographiques

Holzmann, Helge, Nick Ruest, Jefferson Bailey, Alex Dempsey, Samantha Fritz, Peggy Lee et Ian Milligan (2022). *ABCDEF - The 6 key features behind scalable, multi-tenant web archive processing with ARCH: Archive, Big Data, Concurrent, Distributed, Efficient, Flexible. JCDL '22: Proceedings of the 22nd ACM/IEEE Joint Conference on Digital Libraries, June 20-24 2022, Cologne, Germany, article no 13.* DOI: https://doi.org/10.1145/3529372.3530916

Justice Canada (2022). *Loi sur la Bibliothèque et les Archives du Canada (L.C. 2004, ch. 11).*

https://laws-lois.justice.gc.ca/fra/lois/l-7.7/index.html

Milligan, Ian (2019). *History in the age of abundance? How the Web is transforming historical research.* Montréal et Kingston, McGill-Queen's University Press.

Milligan, Ian et Tom J. Smyth (2019). Studying the web in the shadow of Uncle Sam: The case of the .ca domain. In: Niels Brügger et Ditte Laursen (dir.), *The historical web and digital humanities: The case of national web domains*. London, Routledge, 45-63.

Smyth, Tom J. (2022a). *Rapid response methodologies and projects: Documenting national historic events at Library and Archives Canada.* International Internet Preservation Consortium Annual Web Archiving Conference 2022.

Smyth, Tom J. (2022b). *Program policy and methodology for the acquisition of social media at Library and Archives Canada*. International Internet Preservation Consortium Workshop: Archiving Social Media 2022.

Smyth, Tom J. (2021). *The black hole of quality control: Toward a framework for managing QC effort to ensure value.* International Internet Preservation Consortium Annual Web Archiving Conference 2021.

Webster, Peter (2019). Understanding the limitations of the ccTLD as a proxy for the national web: Lessons from cross-border religion in the northern Irish web sphere. In: Niels Brügger et Ditte Laursen (dir.), *The historical web and digital humanities: The case of national web domains*. London, Routledge, 110-123.