Assessing the loss of Western Canadian digital heritage

Tasbire Saiyera[1], Dr. Brenda Reyes Ayala[2], Qiufeng Du[3]

[1] Department of Computer Science, [2] Department of Library and Information Studies, [3] Department of Computer Science

{saiyera, reyesaya, qiufeng}@ualberta.ca

Conseil de recherches en sciences humaines du Canada | Social Sciences and Humanities Research Council of Canada

Introduction

Our privilege of accessibility to the web has made many of us reliant on online sources as our primary source of information. As convenient as it may be to post information online, it is just as easy to lose our digital footprint, whether for accidental or for purposeful reasons. During Stephen Harper's tenure as Prime Minister of Canada, many Canadian websites, such as Aboriginal Canada, were taken down (Kingston, 2015). More recently, University of Alberta librarians rushed to preserve online studies on health, climate change policy, and poverty reduction after the United Conservative Party won the Alberta elections (Derworiz, 2019). Evidently, our digital heritage is not only vulnerable to technological issues, but also to the people in power.

The practice of capturing websites using web archives to preserve them as legal, historical, or informational records has been in effect since 2009 for the University of Alberta (University of Alberta, 2020). Using Archive-It (Archive-It, 2014), the university libraries classified websites as relevant to Western Canadian history if they resonated with the themes of their 19 collections. The themes include western Canadian art, important events in Alberta history, and Canadian indigenous movements. Using some of the collections relevant to Western Canadian digital history, this study addresses the following research questions:

1. Of the websites classified as being relevant to Western Canadian cultural heritage, how many of them have disappeared from the live web?
2. How extensively are these websites preserved/archived by institutions (including the University of Alberta Libraries)?

Literature Review

In 2003, the United Nations Educational, Scientific and Cultural Organization (UNESCO) established its Charter on the Preservation of the Digital Heritage, where they recognized that “[cultural] resources of information and creative expression are increasingly produced, distributed, accessed and maintained in digital form, creating a new legacy - the digital heritage” (UNESCO, 2003). The problem of disappearing web resources, called link rot, was first studied in depth by (Koehler, 2002), who monitored the status of a random set of URIs over four years. His results showed that approximately 67% of URIs became inaccessible after the four-year period. In their research, McNally, Wakaruk, and Davoodi (2015) examined the extensive removal of Canadian government web content and its impact on researchers, who would no longer have access to historical Canadian government web content essential for scrutinizing government policy and activities. They stated that web archiving programs were performing a crucial role in maintaining their role as stewards of government information.

Collection Name	Age of Collection in Years	Description	No. of Collection Unavailable
Idle No More	8	A Canadian political movement encompassing environmental concerns and the rights of indigenous communities	91 (46.43%)
Western Canadian Arts	6	Born digital resources created by filmmakers in Western Canada	14 (13.87%)
Fort McMurray Wildfire 2016	5	Websites related to the Fort McMurray Wildfire of 2016	13 (25%)

Table 1: Web archive collections and their link rot percentages

Collection Name	No. of Archive-It Captures (mostly UofA)	No. of Internet Archive Captures	Total No. of Captures
Idle No More	7005 (32.3%)	14049 (64.8%)	21664
Western Canadian Arts	289 (8.6%)	2930 (87.7%)	3342
Fort McMurray Wildfire 2016	4693 (34.3%)	8706 (63.7%)	13677

Table 2: Web archive collections and their preservation status

Methodology

We created a Python program that checked to see if a specific URL is still available by checking its HTTP status code. We categorized "lost" websites as those that returned an HTTP status code other than 200. Table 1 lists some of the collections and their link rot percentages. Furthermore, we deployed MemGator to determine if copies of the websites were also present in web archives around the world. MemGator is able to search the web archives of many institutions and return a list of archived copies of a website, the time and date they were captured, and the institution that captured them (Alam & Nelson, 2016). Table 2 lists the number of captures held by institutions explored by MemGator.

Conclusion

Considering that Archive-It belongs to the Internet Archive, potential corruption of the Internet Archive’s servers pose the threat of Single Point of Failure (SPOF): the failure of a single component leads directly to the failure of a preservation system (Rosenthal et. al, 2005). Our efforts to preserve our digital heritage could prove futile if our systems of preservation fail. More importantly, our digital heritage should not be compromisable. The practical next steps would be to conduct further studies using collections from different institutions in order to assess the extensiveness of the loss of our digital Canadian heritage.

Results

Although 13% - 46% of the websites from the Western Canadian heritage collections disappeared from the live web, these percentages are lower than Koehler's (2002) 67% percentage disappearance rate. Furthermore, it is promising to see the thousands of captures by the University of Alberta and the Internet Archive (unfortunately, we found that the Library and Archives of Canada did not preserve any of the websites from the three collections). For instance, the Idle No More collection, comprised of 196 websites, was found to have 21,664 captures amounting to about 110 captures per website (we have not individually looked into how many captures there are per website; this percentage is merely an estimate).

References

Alam, S. & Nelson, M. (2016). MemGator - A portable concurrent memento aggregator: Cross-Platform CLI and server Binaries in go. In Proceedings of the 16th ACM/IEEE-CS on Joint Conference on Digital Libraries, JCDL 2016 (pp. 243-244). New York, NY, USA: Association for Computing Machinery. doi: 10.1145/2910896.2925452

Archive-It. (2014). Archive-It: University of Alberta. Retrieved from https://archive-it.org/organizations/401

Derworiz, C. (2019, April). Librarians archiving alberta’s scientific reports before change in government. Global News. Retrieved from https://globalnews.ca/news/5182789/alberta-election-librarians-scientific-reports/?bcmt=1

Kingston, A. (2015, September). Vanishing Canada: Why we're all losers in Ottawa's war on data. MacLeans. Retrieved from https://www.macleans.ca/news/canada/vanishing-canada-why-were-all-losers-in-ottawas-war-on-data/

Koehler, W. (2002). Web page change and persistence—a four-year longitudinal study. Journal of the American Society for Information Science and Technology, 53(2), 162-171. Retrieved from https://onlinelibrary.wiley.com/doi/abs/10.1002/asi.10018 doi:10.1002/asi.10018

McNally, M.B, Wakaruk, A., & Davoodi, D. (2015). Rotten by design: Shortened expiry dates for government of Canada web content. Proceedings of the Annual Conference of CAIS, Canada. doi: https://doi.org/10.29173/cais909

Rosenthal, D.H, Robertson, T., Lipkis, T., Reich, V., & Morabito, S. (2005). Requirements for digital preservation systems: A bottom-up approach, D-Lib Magazine, 11(11). Retrieved from: http://www.dlib.org/dlib/november05/rosenthal/11rosenthal.html

United Nations Educational Scientific and Cultural Organization. (2003). Charter on the preservation of the digital heritage. Retrieved from http://portal.unesco.org/en/ev.php-URL_ID=17721&URL_DO=DO_TOPIC&URL_SECTION=201.html

University of Alberta. (2020). Digital Preservation Services. Retrieved from https://www.library.ualberta.ca/digital-initiatives/preservation

Acknowledgement

This project is supported in part by funding from the Social Sciences and Humanities Research Council.