**Should we look to Gatekeepers or the Masses? Qualitative Content Differences in Professionally Curated versus Crowd-sourced Collections.**

Any preservation effort must begin with an assessment of what content to preserve and steward; web archiving is no different. Web archivists have adopted two main approaches in selecting material for future use. First, web archivists working for institutions such as the Internet Archive take a broad approach. They start a crawl with a very large computer-generated list of URLs (a seed list), such as the top million ranked websites in Alexa Internet's rank of trafficked websites, and expand from there. While providing invaluable data for researchers, these crawls are necessarily shallow; they do not follow many links into most domains, leaving deeper content unarchived. They are also prohibitively expensive for all but very large organizations. The second approach develops smaller topical collections curated by librarians and archivists. Using hand-selected seed lists, these collections assemble sites on topics (e.g. Canadian political parties or particular social protest movements). We can see both approaches to content selection and curation as being conducted by "gatekeepers," individuals making particular choices about what should be preserved.

We are now beginning to witness the rise of a third approach driven by "the masses": the idea of archiving pages contained in social media such as Twitter.[1] The question is how these approaches differ in the pages that we collect using these collection practices. Do crowd sourced-collections differ in content, tone, and subject coverage, as opposed to professionally curated ones?

For the fellowship, I will explore qualitative differences between web archives collected by "gatekeepers" and the "masses." Decisions around content selection for digital curation is a pressing problem facing historians and other practitioners: trying to weigh the differences between the institutional biases of professional curators versus the more ephemeral content that is perhaps favoured by social media users. This speaks to ongoing debates between different approaches to digital curation, as the "custodial" approach begins to widen out to more "pragmatic" approaches.[2]

Most importantly, these research questions are a starting point – an ability to begin to explore the conceptual foundation of web archiving, the role of content curators, selection, and the ongoing debates between Archival Science and web archiving as it has evolved. I hope that these can be conversations that we can have during my tenure at the Digital Curation Institute.

**Why now (and why a historian)?**

The stakes are high. If we do not diligently archive the web, the histories that we write will be fundamentally flawed. Imagine a history of the late 1990s or early 2000s that draws primarily on print newspapers, ignoring the revolution in communications technology that fundamentally affected how people share, interact, and leave historical traces behind. This is not an abstract concern. While there is no common rule for when a topic becomes 'history,' it took less than 30 years after the tumultuous year of 1968 for a varied, developed, and contentious North American historiography to appear on the topic of life in the 1960s [3]–[5]. Histories of the 1960s are now common. The history of the 1990s will be written soon. The year 2021 will mark the 30th anniversary of the creation of the first publicly accessible website. We are at the same distance from the 1990s and the birth of the Web that the first historians of the 1960s were.

**Building on Previous Quantitative Work**

This summer, my research team of Nick Ruest, Jimmy Lin, and myself will present our *Joint Conference on Digital Libraries* paper "Content Selection and Curation for Web Archiving: The Gatekeepers vs. the Masses."[6] Our team used three archives – a collection of 1,988,693 URLs tweeted by users on the #elxn42 Twitter hashtag, the August and November 2015 crawls of the University of Toronto Library's Canadian Political Parties and Political Interest Groups (CPP)

collection, and the holdings of the Internet Archive's Wayback Machine – and compared them to each other to discover the *amount* of overlap and difference. We were also curious to add to the conversation on how much of the Web is archived and to further develop earlier (halted) work done by the International Internet Preservation Consortium around using Twitter as a web archiving seed list generator.[7], [8]

The overlap was small between what users tweeted and the holdings of the Internet Archive and CPP. Only 0.269% of 902 of the URLs users tweeted would have been found in the Toronto collection; and only 10.05% would be found in the global Wayback Machine. Our reasons for the difference had to in some ways be speculative: as mentioned, the institutional bias of professional curation, as opposed to the more ephemeral sites and campaigns favoured by social media users.

**The Project: Qualitative Comparisons of Crowdsourced and "Gatekept" Content**

Based on this research, my hypothesis is that scholarly findings from a Twitter-based web archive would differ substantially from a professionally curated collection. The former is a laser-focused snapshot of collections of immediate interest from potentially millions of users, while the latter is a broader collection of a still relatively narrow band of domains selected by subject-matter experts.

To test this, however, we need to move beyond the quantitative and move towards the qualitative. During my McLuhan Fellowship, I will work with Dr. Christoph Becker (University of Toronto), Emily Maemura (his PhD candidate), and others to explore content differences between collections derived from Twitter URLs and professional curators. I hope to hold this fellowship over six months (between January and June 2017, with rough once-a-week presence in Toronto and a more intensive period in May to finish the project). Over this period we will do the following:

**Compare the web crawls of the University of Toronto CPP collection and the URLs from #elxn42:**

- **Topic analysis**: What topics do the CPP collection cover versus a Twitter collection? Building on work Maemura began at the Archives Unleashed hackathon I co-organized in March, as well as my own research, we will do this through text mining (word frequency, topic modelling, keywords used), URL analysis, and image analysis.
- **Metadata analysis**: How does linguistic diversity vary amongst websites crawled? Their geographic diversity? How does this compare to the data that we can extract from tweets?

**Run two workshops as part of the DCI lecture series:**

- **A Twitter archiving and analysis workshop**, taking users from the process of identifying a hashtag, to using the Streaming and Search APIs with Twitter, to basic analytics. This builds on my work in a forthcoming co-authored *Code4Lib* article on Twitter archiving.[9]
- **A Web Archiving analysis workshop** with the warcbase (Warcbase.org) platform.

**Give an invited lecture and assist the DCI with a keynote event:**

- One will be on findings, centering on the metaphor of "Gatekeepers vs. Masses" at the Coach House Institute.
- In collaboration with the DCI, host an invited speaker: Possibilities include my collaborators Niels Brügger (Aarhus University) or Megan Sapnar Ankerson (University of Michigan). Alternatively, the Internet Archive's Brewster Kahle or Jefferson Bailey may be interested.

As my work moves into this area, there is an opportunity for active collaboration and engagement with the iSchool's Digital Curation Institute. Building upon areas of importance to your organization, our collaboration would help further raise the profile of this exciting field.

**References**


[1] E. Summers, "A Ferguson Twitter Archive," *Inkdroid.org*, 30-Aug-2014. [Online]. Available: http://inkdroid.org/2014/08/30/a-ferguson-twitter-archive/?utm_source=rss&utm_medium=rss&utm_campaign=a-ferguson-twitter-archive. [Accessed: 10-Dec-2015].

[2] C. Dallas, "Digital curation beyond the 'wild frontier': a pragmatic approach," *Arch. Sci.*, pp. 1–37, Sep. 2015.

[3] T. Gitlin, *The Sixties: Years of Hope, Days of Rage*. Bantam Books, 1987.

[4] M. Isserman, *If I Had a Hammer: The Death of the Old Left and the Birth of the New Left*. Basic Books, 1987.

[5] C. Levitt, *Children of privilege: student revolt in the sixties : a study of student movements in Canada, the United States, and West Germany*. University of Toronto Press, 1984.

[6] I. Milligan, N. Ruest, and J. Lin, "Content Selection and Curation for Web Archiving: The Gatekeepers vs. the Masses," in *Processings of the Joint Conference on Digital Libraries*, Newark, New Jersey, 2016.

[7] S. G. Ainsworth, A. Alsum, H. SalahEldeen, M. C. Weigle, and M. L. Nelson, "How Much of the Web is Archived?," in *Proceedings of the 11th Annual International ACM/IEEE Joint Conference on Digital Libraries*, New York, NY, USA, 2011, pp. 133–136.

[8] H. Hockx-Yu and M. Pitt, "Evaluating Twittervane: Project Final Report." International Internet Preservation Consortium, 16-Jun-2013.

[9] N. Ruest and I. Milligan, "An Open-Source Strategy for Documenting Events: The Case Study of the 42nd Canadian Federal Election on Twitter," *Code4Lib J.*, no. 32, Forthcoming.