Ensuring access to Canada's core historical data collections in the social sciences and humanities: possibilities emerging from the Canadian Census Discovery Partnership

Submitted by:

Leanne Trimble, Data & Statistics Librarian, University of Toronto (contact person - leanne.trimble@utoronto.ca)
Graeme Campbell, Open Government Librarian, Queen's University
Katie Cuyler, Public Services Librarian & Government Information Librarian, University of Alberta
Alex Guindon, GIS & Data Services Librarian, Concordia University
Tracey P. Lauriault, Associate Professor of Critical Media and Big Data, Communication and Media Studies, School of Journalism and Communication, Carleton University
Amber Leahey, Data & GIS Librarian, Scholars Portal
Susan Mowers, Research Librarian (Data), University of Ottawa
Sandra Sawchuk, Data Services Librarian & User Experience and Engagement Librarian, Mount Saint Vincent University
Zack Taylor, Assistant Professor, Department of Political Science & Director, Centre for Urban Policy and Local Governance, Western University

Current issues

For decades, Canada has been discussing the need for a national research data management strategy and for a national data archiving solution (Hackett, 2001; Humphrey, 2012). Despite the work of initiatives such as the Portage Network, this national strategy has yet to recognize the research importance of Canada's core historical data sources in the humanities and social sciences. In this case we are referring to foundational infrastructure for national data collections, instrumental to open government and democracy.

These "long tail" national data assets have been collected primarily by government, but are used heavily by researchers, the private sector, and multiple levels of government alike. These data are collected by the census, surveys, and administrative programs, and constitute a portrait of the development of Canadian society, economy, institutions, built and natural environments, infrastructure, and politics. Collectively, these data collections are an underexploited national patrimony. They form a foundation upon which research can build, and their long-term preservation and accessibility are essential to future research and governance.

While the custodians of these collections put a great deal of time, effort and funds into the collection and dissemination of the historical data, records and collections, their management has been inconsistent and fragmented over the years. Our national archive, Library and Archives Canada (LAC), has lacked the resources to systematically engage with data collections. As a result, there are significant gaps in access today, and greater concerns from the research community about preservation and access for the long term. This responsibility does not solely rest with LAC, rather, it requires concerted government-academic partnerships to be successful.

With initiatives like Federated Research Data Repository (FRDR) and Dataverse North, the Portage Network has developed some of the foundational infrastructure needed to make this a reality. The decision to adopt a distributed or federated model, rather than a formal national data archive housed at one institution, makes sense in the Canadian context (Humphrey et al., 2016). We applaud the work that has been done to date to create both an infrastructure and a community of practice within the library and research community, bringing together expertise around high performance computing, networked infrastructure, and research data management, curation and preservation.

However, these distributed national infrastructure projects have not yet generated deep partnerships between the research community and government organizations such as Library and Archives Canada, the Treasury Board of Canada Secretariat, Statistics Canada, Natural Resources Canada, and similar organizations at the provincial level. Infrastructure initiatives focused on capturing current scholarly research outputs do not address the gap that remains in protecting our digital heritage. Data from the pre-digital and early-digital era is highly at risk - either locked away in print publications, or digitized in image-based formats, with relatively few sophisticated, scalable tools for extracting the data. Additionally, a one-size-fits-all technical solution (such as Dataverse) does not acknowledge the interconnectedness of these data sources nor offer solutions to improve Canadians' ability to meaningfully use the data. In practice, local institutions will continue to build fragmented solutions for their own little corners of the data services landscape unless we can come together to support these core collections through NDRIO's national data infrastructure. Collaborations between researchers, libraries and government bodies, such as the Canadian Census Discovery Partnership described in the case study below, are an essential component of our federated model for research data management and stewardship in Canada.

Case study: The Canadian Census Discovery Partnership

The Canadian Census Discovery Partnership is a fledgling partnership between Statistics Canada, Library and Archives Canada, and the academic research community. The project involves census researchers, academic librarians, and government experts. The partnership's vision is to create an openly available, bilingual online discovery platform that will include all born-print and born-digital statistical tables, datasets, and mapping products, as well as all relevant documentation, going back to the earliest known pre-Confederation censuses. Each of the partners are building upon their historical and current role in the collection, dissemination, and preservation of census and other government data (Cook & Waiser, 2010; Watkins & Boyko, 2011; Worton, 1998).

Censuses, or population counts, have been conducted in the territory now known as Canada since 1665-66 in New France (Statistics Canada, 2015). Intermittent censuses were conducted by French and British colonial governments until Confederation in 1867, and the first official "Canadian" census was conducted in 1871. Canada's census is intimately tied to colonialism, power structures, political governance, and to the stories that we tell about our shared history (Curtis, 2001; Hamilton, 2007). It is also our most valuable primary economic, social, and cultural data collection, and is, despite its flaws, an essential research tool for the formation of new knowledge about the populations that lived here in the past and present. The sources of information that make up the censuses of Canada are rich, diverse, and complex. They consist of archival records, publications, and data files, containing a mix of data, expert analyses, and supporting documentation. These materials span the print and digital eras, come in a wide range of formats, and are accessed from a multiplicity of locations

(including government websites, library collections, and digital data repositories such as the Internet Archive).

There have already been some significant investments made to improve the accessibility of census data. For example, the Statistics Canada Library digitized its collection of historical Census of Canada print publications, and Library and Archives Canada created searchable databases of census archival records. There have also been large researcher-driven projects such as the Canadian Century Research Infrastructure (CCRI). The academic library community has been actively involved for many years in trying to improve census discovery, through consortia-led data portals such as <odesi> and Scholars GeoPortal, as well as a wide range of finding aids and bibliographies for specific areas of the collection. All of these efforts have produced essential research products and tools, but they lack coordination. There is wide variability in the effectiveness of search tools for these collections, and no single entry point for discovery of all types of materials. Even once identified, historical materials are often not easy to use - they may be print-only volumes, they may consist of scanned pages full of tabular data, or they may be in an obsolete digital data format. Finding, understanding, and using the available data is challenging and time-consuming. When a large proportion of the research project lifecycle is taken up by tasks related to identifying and extracting data out of print volumes and other formats, this has a significant impact on research productivity.

At the core of the Canadian Census Discovery Partnership's work is the completion of a detailed inventory of all data and documentation going back to the earliest censuses. The robust, interoperable metadata and taxonomies will capture the census materials as a complex set of interconnected social and intellectual artifacts that can be examined temporally, spatially, and topically. Structures such as linked open data (LOD) may present an opportunity to turn a large collection of dispersed research materials into a "web of knowledge" (Hart & Dolbear, 2018, p. 9). Statistics Canada is undertaking a similar exploratory linked open data project specifically for microdata, known as the Linkable Open Data Environment (LODE). To accomplish this kind of work for the historical census would require large-scale data digitization as a companion to the inventory metadata. The benefits would be significant, enabling deeper contextual understanding of the relationships between data sources, and mobilizing the data in new ways (such as training datasets in machine learning). The Canadian Census Discovery Partnership aims to investigate and engage with stakeholders working on such forward-looking projects, and through this process to decide upon a path towards a discovery platform/infrastructure for historical census data.

While the Canadian Census Discovery Partnership is focused on solving these problems for census data, we are also interested in how this could be scaled to encompass other components of Canada's core historical collections. We believe that NDRIO should be engaged in these conversations in the larger context of Canada's intellectual and knowledge assets.

Future DRI state

We envision a future in which:

- The Canadian government, together with data stewards, is committed to and invested in the long-term preservation of Canada's core historical data collections. Ideally, this is accomplished through collaboration between government agencies and the academic

community, with a recognition that simply placing data files into an open data portal is not sufficient.
- There is high-quality, bilingual metadata about individual historical data products - metadata that allows for ready discovery across time, place, topic, material type, and collection. This level of description can allow us to explore the deeper connections between our data by leveraging technological advances in data linking. No longer will Canada's historical collections be siloed.
- Individual researchers do not need to use their own research funds to digitize data from historical publications. Researchers will not duplicate effort digitizing materials that someone else has already digitized. Historical materials will be discoverable, searchable and available in machine readable formats.
- Free, open and bilingual discovery platforms bring together formerly disconnected datasets and allow for sophisticated research across long time periods. The search capabilities of these tools will be based on rich metadata and controlled vocabularies that will leverage the expertise of librarians, researchers and subject-experts.

We envision the Historical Census Discovery Partnership working to achieve these goals for Canada's historical census data, but also serving as a test case to flesh out an infrastructure that would be most powerful if it is used for all of Canada's core data collections.

Bridging the gap

NDRIO can play a role in bringing about this vision. We feel that NDRIO should:
- Commit to supporting the humanities and social sciences communities by recognizing the essential role of Canada's core historical data collections in supporting digital research.
- Recognize the need for data service providers within the research and library community to work directly with government data custodians. Canada's digital research infrastructure should not be a thing entirely separate from its government data distribution mechanisms (e.g. the federal open data portal). There are important connections between them.
- Play a key role in identifying important digital research infrastructure projects and in bringing together academic researchers, librarians and government experts to realize them. Creating joint DRI initiatives will vastly improve research efficiency, better leverage expertise and limit unnecessary duplication of effort.
- In the spirit of Open Data, prioritize projects that aim at making public datasets more widely available and better suited for research, and that have the potential to reach a large number of researchers across several disciplines and from different types of organizations (universities, governments, libraries, NGOs, community organizations, etc.).
- Support projects that build out infrastructures that will not only house data, but also make data more usable. Canada should support large scale projects to make our core data collections more useful as research inputs. Metadata and data digitization projects such as the Canadian Census Discovery Partnership, are key to achieving this for our core humanities and social sciences data. Yet there are very few sources of funding for such projects.

Through concerted and strategic efforts to identify and prioritize projects that will curate, preserve, and improve access to Canada's national historical data collections, NDRIO can facilitate digital research and knowledge creation in the humanities and social sciences.

References

Cook, T. & Waiser, B. (2010). The Laurier promise: Securing public access to historic census materials in Canada. In C. Avery & M. Holmlund (Eds.), *Better off forgetting?: Essays on archives, public policy, and collective memory* (pp. 71-108). University of Toronto Press.

Curtis, B. (2001). *The politics of population: State formation, statistics, and the census of Canada, 1840-1875*. University of Toronto Press.

Hackett, Y. (2001). A national research data management strategy for Canada: The work of the National Data Archive Consultation Working Group. *IASSIST Quarterly,* 25(3), 13-16.

Hamilton, M. (2007). “Anyone not on the list might as well be dead”: Aboriginal Peoples and the Censuses of Canada, 1851–1916. *Journal of the Canadian Historical Association/Revue de la Société historique du Canada, 18*(1), 57-79.

Hart, G. & Dolbear, C. (2018). *Linked data: a geographic perspective.* CRC Press.

Humphrey, C. (2012). From national institution to national infrastructure. Blog post on *Preserving Research Data In Canada: the Long Tale of Data.* https://preservingresearchdataincanada.net/category/national-data-infrastructure/

Humphrey, C., Shearer, K. & Whitehead, M. (2016). Towards a collaborative national research data management network. *International Journal of Digital Curation,* 11(1), 195-207

Statistics Canada. (2015, December 30). *History of the Census of Canada*. Statistics Canada. https://www12.statcan.gc.ca/census-recensement/2011/ref/about-apropos/history-histoire-eng.cfm

Watkins, W. & Boyko, E. (2011, November). *The Canadian Data Liberation Initiative: An idea worth considering?* Working Paper no. 006. International Household Survey Network. https://ihsn.org/sites/default/files/resources/IHSN-WP006.pdf

Worton, D. (1998). *The Dominion Bureau of Statistics: A history of Canada’s central statistical office and its antecedents, 1841-1972*. McGill-Queen’s University Press.