IMPROVING ACCESS TO DIGITAL HISTORICAL CENSUS BOUNDARIES IN CANADA

Jeff Allen & Amber Leahey
University of Toronto

Introduction

Historical census boundary datasets are invaluable resources for mapping and analyzing demographics over space and time. In Canada, finding and using historical census boundary data can be a little difficult. Statistics Canada makes tabular census data available online for the 2011, 2006, 2001, and 1996 Censuses, with some summary profile tables available back to 1991. For boundary files however, fewer censuses are accessible, with only 2011, 2006, and 2001 available online. Today, access to the older collections is typically mediated by Statistics Canada, or academic libraries who have access through the Data Liberation Initiative (DLI) program. Given that the data from these earlier years are not readily available online publically, it prevents researchers from easily accessing and using them. In addition, for some of the older censuses, the digital spatial data are stored in archaic data formats which present challenges for use in modern Geographic Information Systems (GIS).

In the fall of 2015, Scholars Portal and the University of Toronto Map and Data Library embarked on a project to bring together the dispersed collection of digital census geography datasets and make them available online so they can be easily accessed by researchers, students, and the general public. This project makes data and documentation available openly through the Ontario Council of University Libraries (OCUL) Scholars GeoPortal platform (http://geo.scholarsportal.info). In making the collection available online openly and all in one place, these datasets will be shared and reused more effectively, thus reducing barriers and duplication for researchers everywhere.

This paper outlines the current status of census boundary datasets in Canada and then details our work which includes collecting known datasets from a variety of sources, data conversion, composing a comprehensive set of metadata, and providing online access to the collection. We also compiled an extensive inventory of all known boundaries produced in order to keep track of the collection as well as assess any gaps to help plan future digitization projects. We hope that this work is utilized and shared with others so that more attention is given to this important historical GIS collection.

Overview of Census Geography in Canada

The Census of Canada program provides a statistical portrait of the country. It is administered by Statistics Canada who are mandated "to collect, compile, analyse, abstract and publish statistical information relating to the commercial, industrial, financial, social, economic and general activities and condition of the people" (Statistics Act, 1971). The Canadian Census dates back to 1666, when French colonial administrators collected information on the new settler populations of New France. There were a number of colonial and regional census projects that occurred during the 18th century and first half of the 19th century, which depending on historical circumstance, focused on collecting data on armaments and agricultural resources. The first post-confederation census was conducted in 1871 and the census was administered by the Ministry of Agriculture until 1912. The Statistics Act was passed in 1918 shifting the responsibility of the census to the new Dominion Bureau of Statistics who administered the census decennially up until 1951 (Statistics Act, 1918). The first mid-decade census was conducted in 1956 and censuses have been conducted quinquennially ever since. In 1971, the Statistics Act was amended, which resulted in the Dominion Bureau of Statistics being replaced by Statistics Canada, a full-fledged federal department (Statistics Act, 1971). This change also introduced new methodologies like self-enumeration instead of in-person interviews,

splitting the census into long-form and short-form questionnaires, and storing collected data in machine readable formats. The most recent census was distributed in May 2016 by Statistics Canada and the data for this census is being planned for staggered release in late 2016 and early 2017.

Census data is inherently linked to both *when* and *where* it was collected. Data is collected at the household level at specific addresses on specific dates. Census boundaries are delineated by Statistics Canada to enable the enumeration and aggregation of census data to designated areal units. Census boundaries range in area from those representing entire provinces and territories down to individual urban blocks. This allows for mapping and analyzing census data at different scales. Some census boundaries are designed for the enumeration of certain census variables. For example, crop reporting districts are delineated for the analysis of the Census of Agriculture. Larger census regions are typically composed of smaller regions to allow for the upward aggregation of census data (e.g. dissemination areas are composed of blocks, census tracts composed of dissemination areas, and so on).

Figure 1 shows the census boundaries and their hierarchical relationships for 2011 Census. Every census year has uniquely defined boundaries. They are redrawn because of changes in population distributions or enumeration methodologies. In some cases, the naming conventions of boundaries have changed as well. For example, enumeration areas were renamed dissemination areas in 2001.

Beyond areal boundaries for disseminating data, Statistics Canada also produces other types of spatial datasets for analyzing and visualizing census data. Road network files and block-faces are produced to connect census data to streets and address ranges; ecumenes are delineated for thematic cartography purposes; and geographic attribute files are generated for linking between boundary levels, coordinate data, and population and dwelling counts.

Aggregated tabular census data can be linked to boundary files in a geographic information system (GIS) for mapping and spatial analysis using unique geographic identifier codes. Common applications of mapping census data include choropleth and dot density maps for visualizing spatial patterns of social,

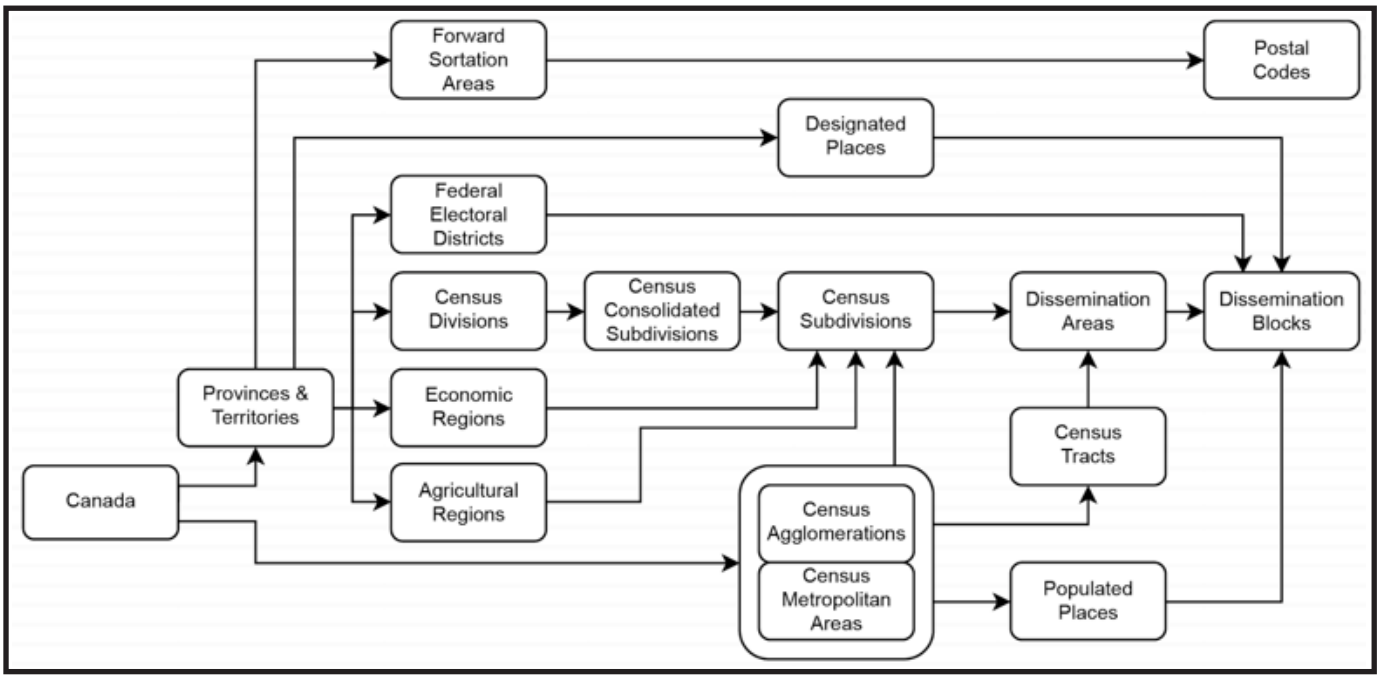

Figure 1 - Diagram of Hierarchical Relationships of Census Boundaries for the 2011 Census (adapted from Statistics Canada documentation)

economic, and demographic characteristics. Analysis of census data linked to boundaries are used to aid wide range of public planning and policy decisions (e.g. healthcare, education, transportation, etc.), for delineating electoral districts, and have countless research applications, particularly in the social sciences. Boundaries from older censuses can be used for mapping demographics at certain points in time and spatial comparison with other historical datasets. This is part of a larger increasing trend in using GIS to aid historical research (see, for example, Gregory & Ell, 2007 or Knowles & Hillier, 2008). Moreover, combining census data and boundaries from different census years can lead to insights on how places change over time. Canadian research in this area include analyzing spatio-temporal patterns of population density (e.g. Millward & Bunting 2008), urban growth (e.g. Burchfield & Kramer, 2015), and gentrification (e.g. Meligrana & Skaburskis, 2005).

Status of Census Spatial Datasets & Project Motivation

Today, Canadian census boundaries are typically produced and stored digitally, as vector datasets in a spatial data warehouse (e.g. representing boundaries using points, lines, and polygons). Together these form the national spatial data warehouse and provide mechanism for the enumeration, collection, and production of a variety of census data products. Boundaries are represented as features, and each feature (e.g. polygon) has associated attribute data including a unique identifier to link with aggregated census data for mapping and analysis.

Digital boundary files for the Canadian Census have been produced by Statistics Canada since 1971. Boundaries are available back to earlier pre-confederation censuses, thanks to the research and data creation of the Historical Atlas of Canada. Up until recently, most early digital spatial datasets were only available for purchase from Statistics Canada, or through the department's Data Liberation Initiative (DLI) program, a national consortium made up of universities that formed together in 1996 to pay for and access Statistics Canada data, namely Public-Use Microdata Files (PUMFs). Part of the DLI includes census data, and boundary files, including census tracts and dissemination/enumeration areas, with some boundary coverages back to 1971. Without the DLI, individual datasets would typically cost several hundreds of dollars, and these high costs severely limited who was able to acquire and use these datasets for research and analysis (Klinkenberg, 2003).

Access to the DLI collection, including boundary files, was typically mediated by the library at subscribing DLI institutions, some providing links to the data files online, and most only have access via a local connection FTP server. Given that the data for 1971 to 1996 are not available online publically, this prevents people from finding and using these census boundary files. The collection also has little metadata for the data files, which is required for description and indexing in repositories, such as in Scholars GeoPortal. The accompanying data documentation provide details about the data and source information, however, machine-actionable metadata is required for description and discovery on the web, and greatly enables data reuse by capturing important information about the original data, including coordinate systems, projections, collection period, purpose statements, feature counts, etc.

At the time of writing, Statistics Canada has made their spatial datasets for the 2001, 2006, and 2011 censuses freely available online. These will be joined in November 2016 by the boundary datasets that delineate the 2016 census. For the 2006 and 2011 censuses, Statistics Canada provides spatial data as Shapefiles, MapInfo TAB format, and Geography Markup Language (GML). Shapefiles are widely used across GIS applications today, and are largely considered the standard for sharing spatial vector datasets. GML in an open format that uses XML grammar to define geographical features, it is less frequently used by researchers, but it is an open standard supported by the Open Geospatial Consortium (OGC). MapInfo TAB is a lesser used today, and like ESRI's Shapefile format, is a proprietary vector data format designed for use in it's own software.

For digital data produced prior to 2006, boundaries were published and remain stored in spatial data formats that are currently out-of-date and can only be opened by specific, often proprietary, GIS software. For example, the ArcInfo Interchange format (E00) and the MapInfo TAB format, were widely used to store Statistics Canada digital spatial

boundaries. There are also older datasets that are only available as flat files containing ASCII text. They require a codebook to parse the data to provide any use. Some of these datasets come with SPSS syntax files, generated by the University of Toronto Map and Data Library, but again, these require SPSS or other statistical software packages. SPSS is an expensive, proprietary software that not everyone has access too. At the very least, some knowledge of programming is required to read the data, and this isn't considered accessible to the public.

National Infrastructure Projects and Other Digitization Initiatives

Prior to 1971, census boundaries were not produced digitally, only on paper. There have been several separate projects conducted by different academics, librarians, and cartographers, which have digitized historical census boundaries into vector datasets for use in GIS. Part of our project was to acquire these digitized historical boundaries and make them easily available in Scholars GeoPortal alongside digital boundaries from more recent censuses.

Probably the most substantial digitization project was conducted by the Canadian Century Research Infrastructure (CCRI). The CCRI created a harmonized database of census subdivisions boundaries from 1911 to 1951. This database also allowed for dissolving boundaries and associated data up to census divisions and the constructed framework enabled the location, selection, aggregation, and analysis of data for any census year from 1911 and 1951. Working from modern census boundaries as a reference, the CCRI generated a harmonized spatial database for this recreation of historical boundaries. The CCRI has been instrumental in providing a basis for historical census data mapping and analysis and it is well documented and often referenced by historians and GIS researchers (for more information on this project, see St. Hilaire et al., 2007).

There have been several other digitization initiatives conducted by cartographers and librarians across Canada as well. On such initiative was the Historical Atlas of Canada Online Learning Project (HACOLP), which included digitizing census divisions from 1851 to 1961. These boundaries are part of an online interactive cartographic application and are available for download as Shapefiles. Another project was conducted by librarians from the University of British Columbia who digitized urban Census Tracts and Census Metropolitan Areas for the 1951 Census across Canada (Brittnacher & Lesack, 2013). The University of Toronto Map and Data Library have also undertaken digitizing projects. They digitized 1981 census tracts to vastly improve accuracy over imprecise original data files and they have digitized 1961 census tracts for Toronto from paper maps, which were previously unavailable in any digital format. These projects typically used a technique in GIS of editing modern boundaries to align with the historical boundaries displayed on a georeferenced paper map. This technique allows for maintaining the precision of newer boundaries and saves time by not needing to digitize boundaries that have remained stable.

Data Migration Project (1971 to 2001)

We have been conducting a data migration project to convert census boundaries from 1971 to 2001 from their original, out-of-date, digital formats into Shapefiles to allow for easier usability and long term preservation. Mapping and geographic analysis of census data requires accurate and accessible census spatial datasets. Also, digital data is often more susceptible to obsolescence compared to material sources like paper maps. Over time, data becomes less accessible as file formats change and newer software offers less support for older formats. Data migration is the process of transferring data between storage types and is used as a form of digital preservation to make sure historical datasets, like census boundaries, can be used for people now and in the future. Moreover, since these datasets have become open as part of the Data Liberation Initiative, they should be freely and easily accessible across GIS applications. The Shapefile format was chosen as the output since it is widely used both in proprietary (e.g. ArcGIS, Global Mapper, FME, etc.) and open source GIS (e.g. QGIS, GRASS, PostGIS, etc.). Also, there are plenty of tools available to convert Shapefiles into other geospatial formats if needed (e.g. GDAL/OGR).

Beyond data format conversion, census boundary datasets are also being enhanced as part of the data migration process to further their spatial analysis capabilities in modern GIS applications. All

census boundary datasets are being transformed into North American Datum 1983 (NAD83), which is the datum that Statistics Canada currently uses for their datasets. Over the years, the projections and coordinate systems of census geography datasets varied from Lambert conformal conic, Universal Transverse Mercator, or unprojected NAD27. Conforming datasets data to a single geographic coordinate system allows for consistency when comparing between census years and boundary types. Moreover, features in census boundary datasets are then dissolved to their unique identifiers (e.g. CTUID for Census Tracts). Older datasets typically did not include multi-part features. For example, each island in a group of islands that were part of the same census area would have separate records in the dataset. Dissolving to unique identifiers combines all features with the same identifier into one multi-part feature. This allows for for one-to-one joins with associated tabular data. For some census boundary datasets, additional fields were generated to allow for easier relationships with associated tabular data and other census geography files. For example, in one dataset, existing identifier fields were converted from integers into strings with leading 0s (e.g. from '1' to '001') to allow for joins with tabular data that have the same structure. Also, the original datasets for some census boundaries, primarily prior to 1991, were divided by metropolitan area (e.g. there were separate datasets for Montreal, Vancouver, etc.). These have been appended into one Shapefile to provide a Canada wide coverage.

Much of the data migration process was automated through custom Python scripts with help from geospatial libraries like ArcPy and GDAL. ArcPy is the Python library for scripting geoprocessing tasks in ArcGIS while GDAL is an open source translation library for geospatial data formats. For this project, automated tasks include batch converting between file formats (e.g. from .e00 to .shp), dissolving and appending features, joining and updating attribute fields, defining coordinate systems, and parsing ASCII text files. Converted datasets are checked using Statistics Canada documentation to confirm their coverage and feature counts, and where possible, are compared to any datasets that were previously converted from different Canadian University libraries (University of Toronto, Waterloo, Western, and Queens).

Organization in Scholars GeoPortal

All acquired and converted datasets are being made available through Scholars GeoPortal as open content meaning that the datasets are available for anyone to access, regardless of affiliation. In Scholars GeoPortal, each census geography dataset can be viewed with reference to a base map, and if the user wants, in conjunction with other datasets. Datasets can be queried either by attribute or on map selection. Each dataset layer has unique symbology and labels identifying the names or unique identifier codes of individual boundaries. Data are available for downloaded as a zip package which includes the converted datasets, documentation, the original data, and any associated attribute tables (e.g. concordance tables).

Each individual dataset has detailed metadata describing its coverage, source, and notes on the data migration or digitization process. Metadata records have unique URIs, meaning datasets can be easily linked to, shared, and found in external search engines. For organization, individual metadata records are aggregated into series records by year, language, data collection category. For 1991 and onward, census boundaries are divided into two categories Digital Boundary Files (DBF) and Cartographic Boundary Files (CBF). DBFs depict the full extent of the geographical areas, including the coastal water area while CBFs depict the geographical areas by clipping to the shorelines of Canada and its coastal islands. CBFs are typically used for general map making as well as calculating population densities and other areal functions. There are also series records for special collections like road network files or health regions. French datasets and associated series records are also available for the 1996, 2001, 2006, and 2011 censuses. These have the same geographic data as their English counterparts, but include French fields in associated attribute tables.

Metadata is generated as part of the loading process into the Scholars GeoPortal. The metadata standard used for the portal is based on the ISO 19115 - North American Profile. A custom

metadata editor provides the form for the descriptive fields, and information and values are entered online in the editor. The metadata provides the rich descriptive information about the boundary files, as described above, and links to the web map service to provide access to the resources online. Users are able to search across dataset metadata, filter based on keyword, spatial coverage, and year of publication, allowing for improved access and discovery online. Metadata and data are provided openly for anyone to search, find, access, and download. The creation of rich, standard metadata enables easy access online, and provides a machine-actionable record (XML) of the dataset information that can be stored and preserved for the long-term.

Looking Forward

Figure 2 shows the digital census boundaries available and our progress (at the time of writing) collecting and converting census spatial datasets and uploading them to Scholars GeoPortal. However this table also indicates that there are a number of significant gaps within the collection. Overall, we are hopeful that this project will raise awareness for librarians, researchers, and cartographers to share any datasets that we are not aware of, and, moreover, spur future digitization projects to fill the gaps in the collection. There are already some ongoing efforts to fill in these gaps. For example, Statistics Canada is in the process of converting and digitizing Enumeration Area boundaries from 1971 and 1981. Since census geography is hierarchical, these datasets can be dissolved up to recreate other missing boundaries. Furthermore, our collection of census boundaries will be added to with spatial datasets from the 2016 census, which is planned for release in November, 2016 and will subsequently be uploaded in Scholars GeoPortal.

Another major issue and avenue for future work going forward is the current lack of harmonization and accurate concordance between spatial boundaries over time. This hinders the ability to conduct accurate spatio-temporal analysis, particularly from the early digital years of the census (1971 to 2001) where boundaries were defined with varying methodologies, precision, and coordinate systems. There are some concordance and correspondence tables for relating census data between years to the same boundaries. However, these existing tables are limited as they only exist for pairs of years and they do not indicate any percentage changes in area or population. This potentially leads to imprecise results when using them to examine how demographics at specific places change over time. There have been some attempts to rectify

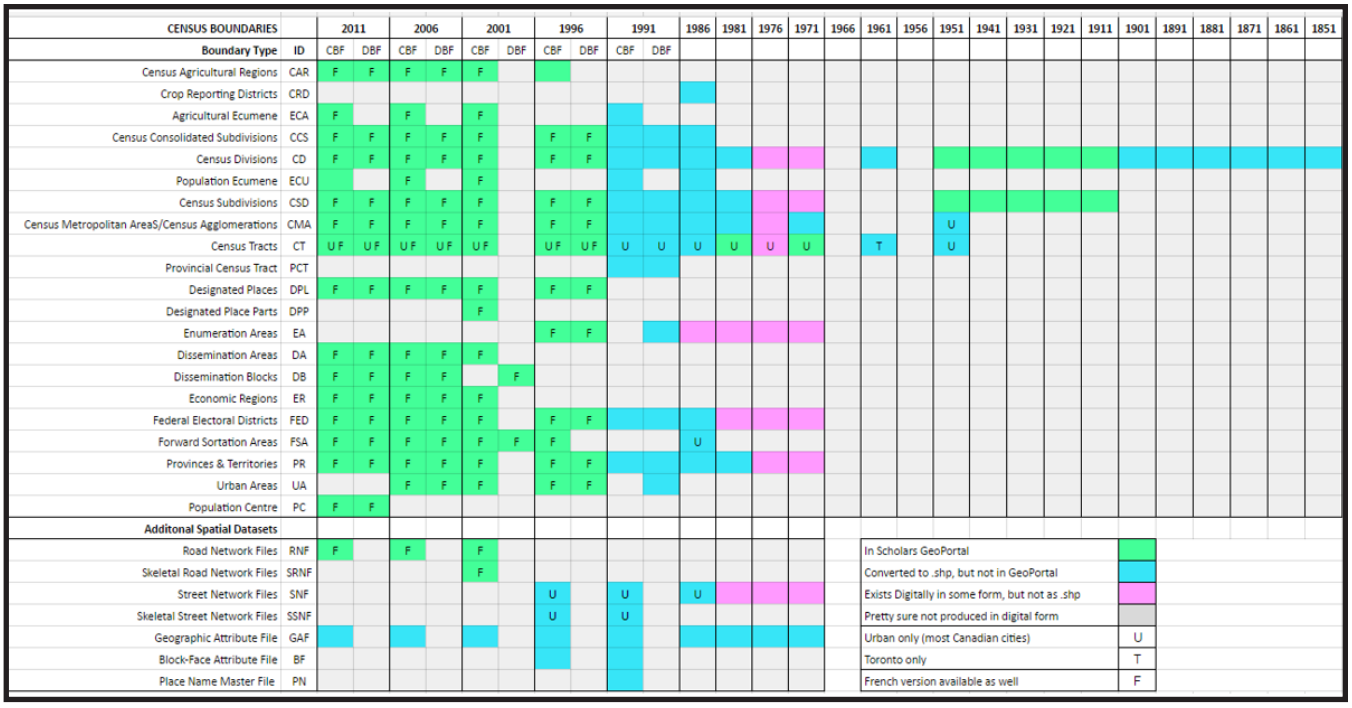

Figure 2 - Inventory of Datasets

these issues. For example, when the CCRI digitized census subdivisions from 1911 to 1951, they used consistent boundaries to allow for harmonized spatio-temporal analysis over this time period. In another project, Schuurman et al. (2006) looked at fixing spatial mismatch between the 1996 to 2001 census boundaries in Vancouver by conflating the road network files. More extensive longitudinal harmonization and concordance projects have been undertaken in Great Britain (Gregory & Ell, 2005) and the United States (Logan, Xu, & Stults 2014), but not in Canada at the same scale. The first author of this paper is currently working on a project to create a harmonized longitudinal spatial database of census tracts across Canada to allow for accurate spatio-temporal analysis of census data at the neighbourhood level.

Conclusion

In this paper, we overviewed the landscape of digital historical census boundaries in Canada and detailed our work collecting these datasets from a variety of sources, converting those in older formats into Shapefiles, and making them available online in Scholars GeoPortal. Our progress has made these datasets more accessible and easier to use for researchers, librarians, and the general public. Moreover, by consolidating and converting these datasets, we are enabling long term preservation to prevent them from becoming lost or obsolete. With the creation of an inventory, we plan to assess any gaps between digitally available boundaries and those only available in paper maps in order to spur future digitization projects. Improving the collection may also require further curation, harmonization, and collaboration between stakeholders (government, libraries, researchers, etc.). This will be explored as we move forward and work towards building a more comprehensive national historical census boundary database.

References

Brittnacher, T. & Lesack, P. (2013). Boundary Files, Census of Canada 1951. Accessed from http://hdl.handle.net/11272/10268

Burchfield, M., & Kramer, A. (2015). Growing Pains. Neptis Foundation.

Gregory, I. N., & Ell, P. S. (2005). Breaking the boundaries: Integrating 200 years of the Census using GIS. Journal of the Royal Statistical Society. Series A (General), 168(Part 2), 419–437.

Gregory, I. N., & Ell, P. S. (2007). Historical GIS: technologies, methodologies, and scholarship (Vol. 39). Cambridge University Press.

Logan, J. R., Xu, Z., & Stults, B. J. (2014). Interpolating U.S. Decennial Census Tract Data from as Early as 1970 to 2010: A Longitudinal Tract Database. The Professional Geographer, 66(3), 412–420.

Klinkenberg, B. (2003). The true cost of spatial data in Canada. Canadian Geographer, 47(1), 37–49.

Knowles, A. K., & Hillier, A. (2008). Placing history: how maps, spatial data, and GIS are changing historical scholarship. ESRI, Inc.

Meligrana, J. & Skaburskis, A. (2005). Extent, location and profiles of continuing gentrification in Canadian metropolitan areas, 1981–2001. Urban Studies, 42(9), 1569–1592.

Millward, H. & Bunting, T. (2008). Patterning in urban population densities: A spatiotemporal model compared with Toronto 1971-2001. Environment and Planning A, 40(2), 283–302.

Schuurman, N., Grund, D., Hayes, M., & Dragicevic, S. (2006). Spatial/temporal mismatch: A conflation protocol for Canada census spatial files. Canadian Geographer, 50(1), 74–84.

St-hilaire, M., Moldofsky, B., Richard, L., & Beaudry, M. (2007). Geocoding and Mapping Historical Census Data. Historical Methods, 40(2), 76–91.

Statistics Act (R.S.C., 1985, c. S-19). Accessed from http://laws-lois.justice.gc.ca/eng/acts/S-19/

Statistics Act (1918 vol. I 1918 vol. I 139 1918)

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