

Canadian Century Research Infrastructure (CCRI)

Introduction to Georeferencing for the CCRI

Byron Moldofsky, Project Coordinator, CCRI University of Toronto Team

The *CCRI* is a pan-Canadian, multi-disciplinary and multi-institutional effort to develop a set of interrelated databases centred on data from the 1911, 1921, 1931, 1941 and 1951 Canadian censuses. The long term plan is that these databases will be joined to other data that cover the periods from 1871 to 1901, and from 1961 to 2001. The result will be a new foundation for the study of social, economic, cultural, and political change.

The CCRI is an interdisciplinary project involving historians, sociologists, geographers and others interested in research using historical population databases. Partners also include Statistics Canada, the National Archives, and corporate partners. Funding is provided by CFI and OIT, and other provincial funding agencies. Please visit the CCRI website for an overview of the project, the group of seven university centres and their primary researchers, and the overall research agenda. Connections have been made with other Canadian projects, as well as international groups of historical population researchers, such as IPUMS and IMAG.

www.canada.uottawa.ca/ccri

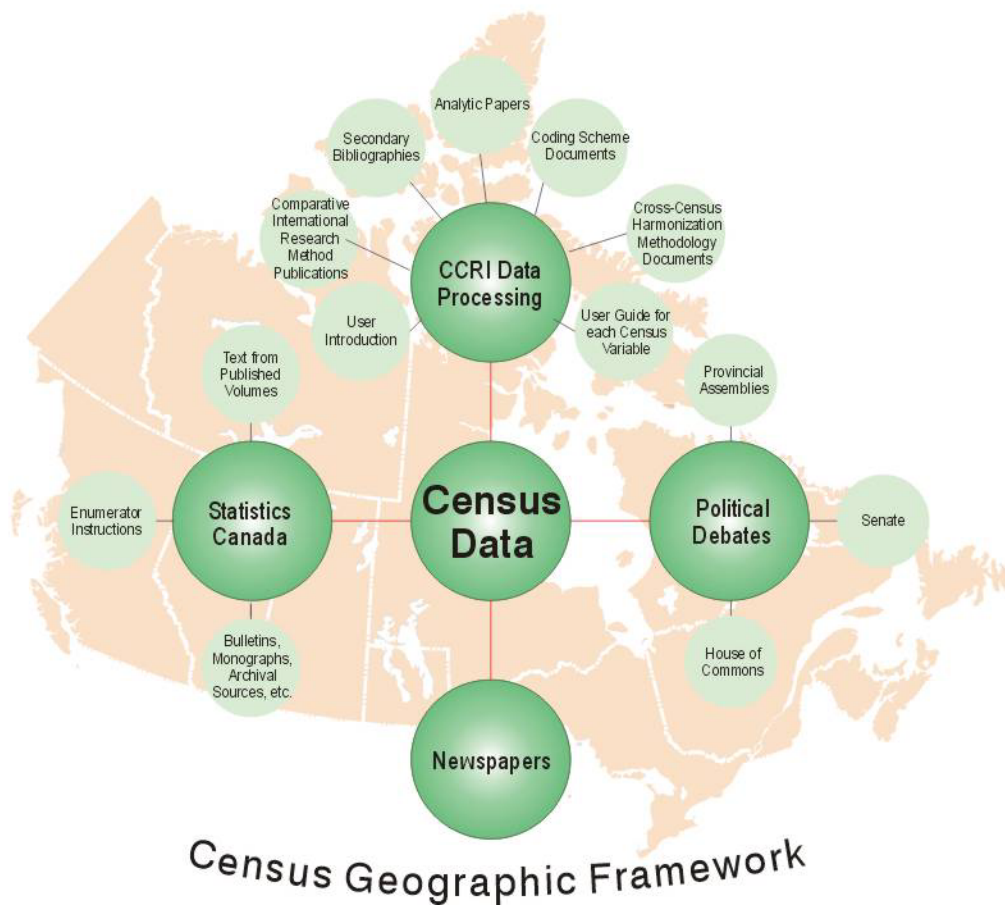
The core of the project is a database consisting of a microdata sample of manuscript census records, stratified generally by geography. Dwellings will be the unit of sampling, with every individual within a selected dwelling recorded, for all census variables. The samples will vary in size from 5% in 1911, to ~3% in 1951. After data entry is done, the records will be coded appropriately, by variable. In this way a "Microdata sample" similar to those created by Statscan for the modern censuses (1971-2001) will be created. Concurrently, research is being done on the "contextual data" around the census: documentary evidence such as government records and newspaper reports of the time, which have an impact on the meaning and interpretation of the census variables.

The conceptual structure of the project has been visualized using what has come to be known as the "**Green Balloons Diagram**" (see next page). This gives one an idea of the interrelated components of the project.

Integral to the entire project is the construction of a geographic framework for the historical census data, using Geographic Information System (GIS) technology. GIS map layers are being created to enable geographic location, selection, aggregation and analysis of sample data, as well as some mapping of generalized census data. This will allow researchers to ask questions of the database which are much more geographically specific than in the past. Interface tools to make these geographic queries and analysis as user-friendly as possible are also being developed.

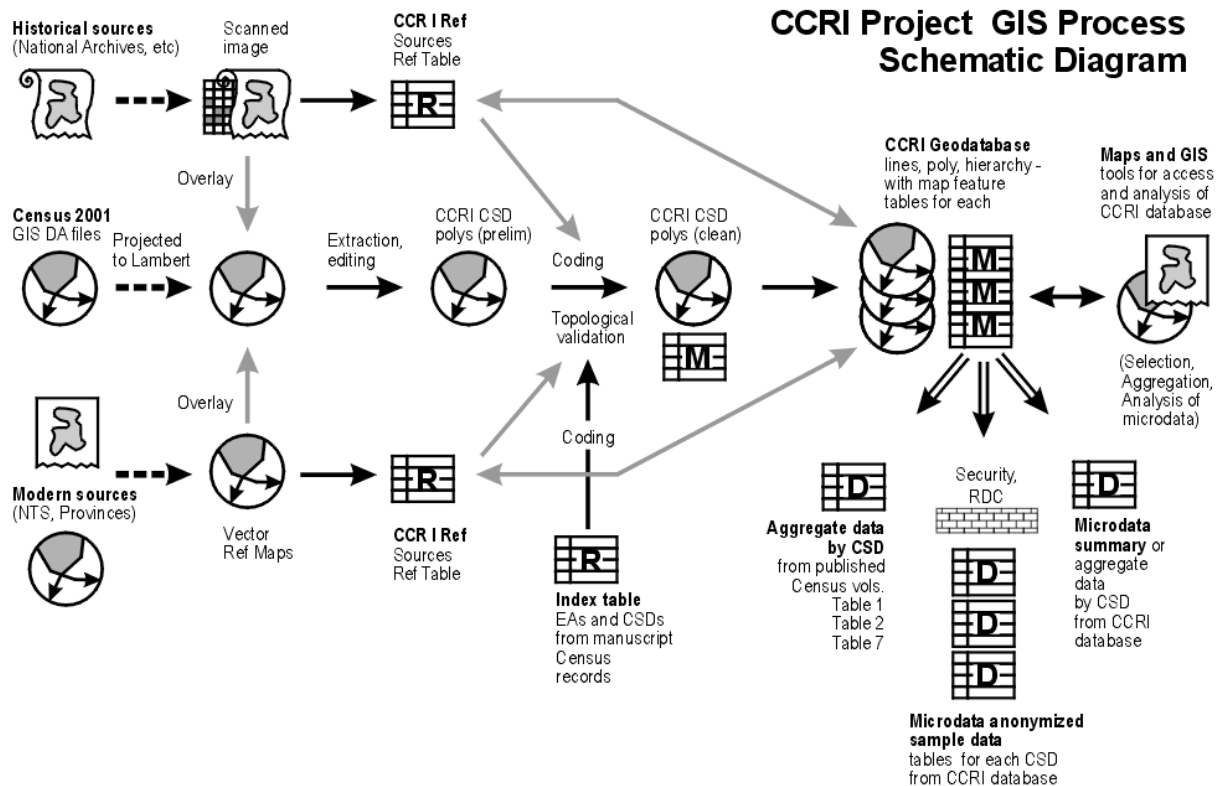
Figure 1: "Green balloons" diagram

CCRI Virtual Research Infrastructure Contextual Data



The diagram below, “CCRI Project GIS Process Schematic Diagram” outlines in a conceptual way the method we are using for production of the GIS framework. This is being done in tandem with our partners at Université Laval, lead by Prof. Marc St.-Hilaire. The 2001 Census geographic files produced by the StatsCan Geography Division is our primary reference base. Care is being taken to conform as much as possible to the precedents set in level of detail and coding in this modern reference source. Other modern GIS reference sources, such as the National Topographic Data Base, and Provincial base layers at appropriate scales, are being used to supplement this source. Historical sources, primarily maps from the National Archives of Canada, are being used to identify and guide the editing of the StatsCan geographic files to reconstruct the boundaries of the 1911 census. This is being done at the Census Subdivision level in the census geographic hierarchy. Other historical sources, maps and textual, will be used to refine these boundaries as much as possible, given the finite resources of the CCRI project. After 1911 is done, 1921-41 will be built, based on this template.

Figure 2.



To implement this approach, we are using ArcGIS software, from ESRI (www.esri.com). It provides the tools necessary to integrate the historical map images and modern data sources. For organizational and analysis purposes, our aim is to develop or adapt a **Geodatabase Data Model** appropriate for these datasets. We have been looking at that being developed for the US Census geography as a starting point, which have been documented in an ESRI book *Designing Geodatabases: Case studies in GIS Data modeling*.

To organize our efforts and efficiently plan the use of our resources, we have arrived at a broad set of goals, objectives and strategy. These are summarized in the graphic below.

Figure 3. Geo-coding subgroup: Goal, Objectives and Strategy

Geocoding Goal

Create GIS framework for mapping and analysis of historical census data for CCRI

Geocoding Objectives

- Use GIS as a tool to locate Census aggregated data, and the sample microdata, to the CSD level
- Respecting the limits of confidentiality, allow users represent cartographically, and analyse spatially, the available CCRI aggregate and sample microdata
- Allow users of the CCRI data to select microdata for analysis, by area or other geographic criteria

Geocoding Strategy

In designing and implementing plan, identify objectives and products of geocoding as “**Core**”, or “**Complementary**.” **Core** objectives **must** be fulfilled, and allocation of resources must ensure this; **complementary** objectives are highly desirable additional goals, and priorities should be set to attain as many of these as possible as well

Further to this, we have also defined our Core products, and Complementary products, tentatively, as follows:

Figure 4. Core and Complementary products

Core products

1. GIS layer of CDs and CSDs for each Census as identified in published census volumes , using placeholders or approximate boundaries as necessary
2. Statistical files associated with CDs and CSDs
 - Microdata tabulated by CD and CSD
 - Aggregate data from Published volumes tabulated by CD and CSD

Complementary products

3. GIS layers of CSDs for each Census as identified in Census manuscripts, using enhanced boundaries as possible
4. Proposed aggregations of CSDs based on a variety of criteria, for analysis purposes
 - Sampling/size criteria
 - Cultural, demographic, socio-economic criteria
 - Geographic (locational) criteria
5. Access interface and tools
6. Analysis interface and tools

Inherent in our definition of complementary products is the question of how research on historical populations, specifically in a geographic context, can be addressed. We have been exploring the various research approaches that may be enabled as a result of embedding the census microdata in a consistent, topologically correct GIS environment. In looking at these, limitations of the data, especially due to small sample sizes, must be recognized.

Data exploration in a geographic context will be the first, most obvious method. Investigating various methods of aggregating the data and our geographic units, both to maintain data confidentiality restrictions, and for mapping and analysis, is the next important area to address. This can be extended to consider the possibilities for structuring the data to enable comparisons over time, i.e. between census periods. Finally, looking at how spatial statistics may be useful in characterizing patterns within the data, and in interpretative analysis, is a necessary part of our role.

As the project proceeds, potential users will be consulted to address the needs of the research community. Please contact us with any comments or suggestions which may help us to build a better Canadian Century Research Infrastructure.