The Canadian Peoples / Les populations canadiennes Project

Tables Digitized from Canadian Census Published Volumes

Version 2.0
June 2023

Geoff Cunfer, Laurent Richard and Marc St-Hilaire

Overview

One product of The Canadian Peoples / Les populations canadiennes Project (TCP/LPC) is machine-readable tables of aggregate data digitized from the published census volumes. Assembled from individual level data collected throughout the country, the Census Office published total, aggregated values for each Province, Territory, Census Division (CD), and Census Subdivision (CSD) in Canada. Hitherto available only in paper format, TCP/LPC has digitized selected tables for use with spreadsheet, database, statistical, or GIS software.

In order to create GIS maps for each census at the level of Census Subdivisions, it was first necessary to identify a complete list of CSDs in the entire nation. To do so, TCP/LPC specified one published census table for each year as its "Key Table," which was digitized into an Excel spreadsheet that established the definitive list of CSD places for that census year. Every CSD listed in the Key Table has a single corresponding polygon on the GIS map.

Each CSD in the Key Table was assigned a unique identifier (TCPUID) that concatenates a 2-letter Province Code, a 3-number Census Division Code, and a 3-number Census Subdivision Code. For example, in 1891, the Hawkesbury Village CSD (005) in the Prescott CD (112) of Ontario (ON) has a TCPUID coded as ON112005. This unique identifier appears in the digitized Key Table and also in the GIS. It can be used as a join field to link the polygons in the GIS to the rows in the Excel spreadsheet. Note that TCPUIDs are not necessarily consistent from one census year to another.

TCP/LPC then proceeded to digitize many additional tables from the various census years using the same methods and creating comparable TCPUIDs for each CSD. However, it is important to note that the list of CSDs included in other tables, even within the same census year, may not be identical to those included in the Key Table and corresponding GIS map. In some instances the Census Office established very different lists of CSDs from one table to another. An extreme example appears in the 1911 census, where Volume 1, Table 1 presents data for 10,004 CSDs, while Volume 2, Table 2 presents data for only 3,532 CSDs. For this reason, the Key Tables should join seamlessly with the GIS polygon layers, but other digitized tables may not. Users will need to verify for themselves any discrepancies.

Methods

To digitize these tables, the team downloaded scanned PDFs of the published volumes from the web (https://archive.org/ and https://publications.gc.ca) and used optical character

recognition (OCR) techniques with ABBYY Fine Reader 14 software. The transcribed information was carefully reviewed and summary formulas were used to validate data values. For each attribute field, sums of CSDs were compared to CD- and Province-level totals, and subdivided variables were compared to totaled variables. For example, the total population of Ontario, as published in the table, should match the sum of all of the CSDs in Ontario. Likewise, the total population of each CSD should match the sum of males and females. In these ways, TCP/LPC checked and corrected OCR errors and, in some instances, flagged mistakes that originated in the primary source. When data problems were identified in the original census source, the team corrected them whenever possible.

The tables assembled by the Census Office each year were published in a format suitable for paper pages in bound volumes. The originals sprawl across double leaves, extend through scores of pages, included extensive textual column headings, and intermixed data at multiple scales. Such a structure is not always suitable for machine-readable tables and computer analysis. For this reason, TCP/LPC not only digitized and corrected the tables, but also re-formatted them into a variety of structures designed to be more useful for users working with a variety of computer analysis software packages. Each table presented in this collection is available in four distinct versions: "OCR", "CD", "CSD", and "Pub Tab" as described below.

OCR format ex. 1891_V1T2_OCR_202306.xlsx

The OCR version is a spreadsheet representing the results of optical character recognition software that converted image values directly from published census volumes into digital numbers. This version is a machine readable duplicate of the primary source. It includes the full text of the column headings and place names as they appeared in the published volumes. Any errors present in the original census tables are duplicated here (in addition to any OCR errors not detected at first glance), but the data are now digital.

CD and CSD formats ex. 1891_V1T2_CD_202306.xlsx
1891_V1T2_CSD_202306.xlsx

These two versions separate CDs from CSDs, geographic scales that are intermingled in the original published census volumes. One spreadsheet includes only CDs, the other only CSDs. These data have been corrected and regularized, and thus may be somewhat different from the original primary source. Statistical errors in the published tables have been corrected, where possible. Shortened variable names substitute for the full text column headings in the previous version, and standardized identification fields have been added.

There is a separate documentation file for each census year's CD and CSD tables that explains full variable names. Variable availability across all census years is presented in a "master variable" list called TCP_CANADA_CD-CSD_Mastvar.xlsx.

Pub Tab format ex. 1891_V1T2_PUB_202306.xlsx

Pub Tab is a spreadsheet that closely mirrors the presentation of published tables, including mixed entries for Provinces, Census Divisions, and Census Subdivisions. However,

CSD names were regularized in systematic ways, including cardinal points, wards in cities, etc., and variable fields have shortened names.

Pub Tabs also correct errors in the original source to the extent possible. A newly created "Notes" field explains these corrections. If a Notes entry starts with "TCP", it generally indicates a change was made in the data. If the Notes entry does not start with "TCP", it means the information is about something written by the Census Office in the original printed volume. The Pub Tab format corresponds to that created by the Canadian Century Research Infrastructure / Infrastructure de recherche sur le Canada au 20^{e} siècle Project (CCRI/IRCS) (https://ccri.library.ualberta.ca/).

Important Information about TCP/LPC Tables

The error correction process and variable naming were done independently for the Pub Tab format compared to the CD and CSD formats. This means those versions will have different column headings and may, in rare instances, present different decisions in regard to error correction.

As indicated below, each table is available in all four formats except for those from 1911 and 1921, which are available only in Pub Tab format.

TCP/LPC delivers a specially transcribed table for Prince Edward Island in 1871, despite the fact that PEI joined the Dominion of Canada only in 1873. Users should be aware that the variables available for PEI were quite different from those available for the rest of Canada. Even though both tables are named "Volume 1, Table 1," they appeared in two different publications and are therefore not identical.

Crops and livestock reported in 1851_V2T6 contain many challenges, especially regarding land and crops units. This is a well-known issue; see R.M. McInnis, "Some Pitfalls in the 1851-1852 Census of Agriculture of Lower Canada", *Histoire sociale/Social history*, Vol. XIV, No 27, pp. 219-231, May 1981. The values were transcribed as they are; no attempt was made to solve these issues.

Sometimes, the numbering of tables was challenging, especially in the 1851 and 1861 censuses. In those years, tables for Ontario and Québec were numbered separately, even though they contained identical variables. Sometimes they were called "appendix" rather than "table." We attempted to regularize naming conventions across all years. Likewise, we applied uniform Province codes (ON, QC) in 1851 and 1861, even though the colonies were named Upper Canada/Canada West and Lower Canada/Canada East in those pre-confederation censuses.

Available Tables

Census	Volume	Table	Main Topic	Key	Pub Tab	OCR, CD, CSD
1921	1	16	Place of Birth	Y	X	
1921	1	27	Origin		X	
1921	1	38	Religion		X	
1921	3	3	Housing		X	
1911	1	1	Population	Y	X	
1911	1	2	Marital Status		X	
1911	2	2	Religion	Y	X	
1911	2	7	Origin		X	
1911	2	28	Literacy		X	
1901	1	7	Population	Y	X	X
1891	1	2	Population and Housing		X	X
1891	1	3	Marital Status	Y	X	X
1891	2	16	Land Occupation		X	X
1891	4	2	Crops		X	X
1891	4	3	Livestock		X	X
1881	1	1	Population	Y	X	X
1881	3	24	Crops		X	X
1881	3	27	Fisheries		X	X
1871	1	1	Population	Y	X	X
1871	1	1 (PEI)	Many topics	Y	X	X
1871	3	23	Crops		X	X
1861	1	1 and 2	Origin	Y	X	X
1861	1	5 and 6	Ages		X	X
1861	2	11 and 12	Crops and Livestock		X	X
1851	1	1	Origin		X	X
1851	1	3	Ages	Y	X	X
1851	2	6	Crops and Livestock		X	X
1851	2	7	Mills and Factories		X	X

Overall, the 28 digitized tables contain more than a 1,000 variables and more than 75,000 rows constituting a matrix of roughly 1,873,200 cells.