9-21-2022

UNI·CEN Documentation Report 2: Standardized Census Data Tables

Zack Taylor
Dept. of Political Science, Western University, zack.taylor@uwo.ca

unicen

Unified Infrastructure for Canadian Census Research

Documentation Report 2

Standardized Census Data Tables

Dr. Zack Taylor

September 21, 2022
Version 1.0

Western SocialScience | Network for Economic and Social Trends (NEST)

Contents

The UNI·CEN Project

Analysis of historical and contemporary Census data is an active area of research. Sociologists, historians, geographers, urban and regional planners, and political scientists have used these data to study the historical development of, and change in, international and domestic migration, urban settlement patterns, inter-group relations, economic change, and political representation. In the United States, United Kingdom, and other countries, projects increased the accessibility of historical Census data by compiling existing digital datasets, digitizing those that exist only in print, and creating modern systems to disseminate them to users.

The **UNI·CEN** (Unified Infrastructure for Canadian Census Research) project follows the example of these international projects. We compiled available aggregate Census data at several commonly used levels of geography for the 1851–2021 period and converted it to a standardized table format. We digitized mapped boundaries, data tables, and geographic coding schemes pertaining to census tracts for the 1951–66 period. We developed a standardized variable naming system and coded the compiled data with it, enabling analysis and visualization of change over time. We also developed a set of geographic linkage tables that enable comparison of places across time despite inconsistent naming and coding. Finally, we have assembled available corresponding digital boundary files and reformatted them to join to the data.

Undertaken between 2018 and 2022, **UNI·CEN** is a project of Western University's **Network for Economic and Social Trends** (NEST). **UNI·CEN** parallels a companion project, the **Canadian Communities Policy Observatory** (https://observatory.uwo.ca), a portal that enables visualization, analysis, and retrieval of place-based data. Both projects are funded by Western University's Faculty of Social Science.

Project Team

Investigators, UNI·CEN Project

- Dr. Zack Taylor, Project Leader and Associate Professor, Department of Political Science, Western University
- Dr. Victoria Esses, Professor, Department of Psychology, Western University
- Dr. Dave Armstrong, Associate Professor, Department of Political Science, Western University

Postdoctoral Fellow (Mitacs – Esri Canada)

- Dr. Christopher Macdonald Hewitt, Western University

Research Assistants

- Moira Benedict
- Brittany Bouteiller
- Kyle Hendricks
- Allan Hsu
- Alissa McInnis

Overview

The **UNI·CEN Data Tables** series contains reformatted versions of all publicly available digital Census data. The original files come from a variety of sources created at different times and for different purposes. As a result, they are stored in a variety of file formats and use inconsistent naming schemes for variables and geographic units.

The goal of the this project is to create a series of data tables that:

- Standardize the table format for ease of manipulation
- Standardize variable names and codes to enable longitudinal analysis
- Make explicit the denominator required to calculate percentages of variables
- Make explicit the hierarchy of variables to facilitate the management of subtotaling and percentaging
- Are stored in modern file formats
- Join with the **UNI·CEN Digital Boundary File** series (see **Documentation Report #3**)

File series

Two series of data tables are created, differentiated by their tabular format and file format:

1. **Long Format**: Intended for use in statistical environments, these files include variable names, data quality information, and information about variable hierarchy and type.
2. **Wide Format:** Intended for use in GIS environments, where each table row joins to a discrete polygon. Information about variable hierarchy and type are abbreviated in the variable name. Variable names and data quality information are not present.

Dissemination location

UNI·CEN Data Tables are stored on the open-access Borealis Dataverse repository at: https://borealisdata.ca/dataverse/unicen_aggregate.

Disclaimer

We have made every effort to verify the accuracy of the datasets produced by this project. We ask that users notify us of any errors they discover so that they can be corrected in future versions.

Acknowledgments

We are grateful for the generous advice of Amber Leahey, Data and GIS Librarian, Scholars Portal, and Leanne Trimble, Data Librarian, University of Toronto.

Levels of geography

The six levels of geography selected for inclusion in the UNI·CEN database are relatively stable from one year to next and are well established geographic concepts that will continue into the future:

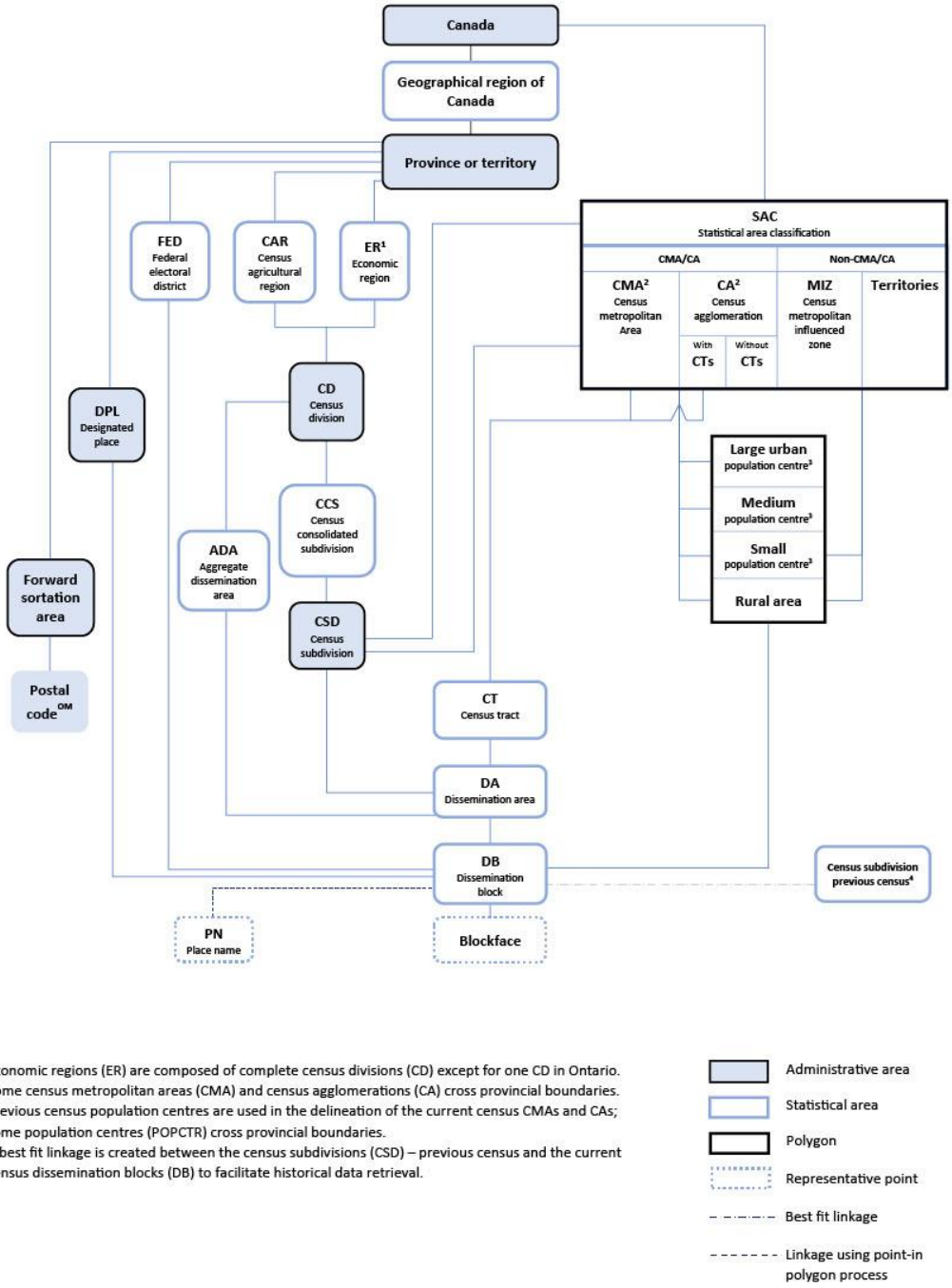
- Census Tract (CT) – Compact, neighbourhood-scale areas within CMAs and larger CAs that contain an average population of 5,000. CTs are delineated to maximize internal socio-economic homogeneity. Statistics Canada initiated the census tract program in 1951.
- Census Metropolitan Area/Census Agglomeration (CMA/CA) – Larger contiguous urban areas comprising multiple CSDs. CMAs have a population of at least 100,000; CAs have a population of at least 10,000. CMAs and CAs larger than 50,000 are tracted.
- Census Subdivision (CSD) – Municipalities, Indian Reserves, Indigenous settlements, or unincorporated areas equivalent to municipalities. CSDs nest within CDs and have national coverage.
- Census Division (CD) – Counties or equivalent entities. CDs nest with PRs and have national coverage.
- Province (PR) – Provinces and territories, which collectively comprise Canada.
- Canada (CA) – The country as a whole.

In addition, we have compiled a special package containing basic data for Federal Electoral Districts (FEDs) from the 1871 Census to the present – see **Report 5** for details.

Please consult Statistics Canada's Census Dictionary for precise definitions of geographic concepts and the criteria used to delineate them. As these have changed over time, it is important for users to be aware of their meaning.

The geographical levels are hierarchically organized with smaller units nesting within larger units (see **Figure 1.1**).

Figure 1.1: The hierarchy of geographic concepts.



Source: Census Dictionary, 2021, https://www12.statcan.gc.ca/census-recensement/2021/ref/dict/fig/index-eng.cfm?ID=f1_1

Statistics Canada disseminates data for other geographies as well:

- Forward Sortation Areas (FSA) – Areas defined by the first three digits of the postal code.
- Enumeration Areas (EA) and Dissemination Areas (DA) – Units with an average population of between 400 and 700. EAs were disseminated between 1961 and 1996, and nest within FEDs. DAs replaced EAs as a dissemination geography in 2001, and nest within CSDs. Both have national coverage.
- Aggregate Dissemination Areas (ADA) – Units with between 5,000 and 15,000 population with national coverage. ADAs nest within CDs, CMAs, and PRs. ADAs were introduced in 2016.
- Economic Regions (ERs) – Aggregations of CDs defined by agreement with provincial governments. Equivalent areas were introduced as “subprovincial regions” in 1971 and renamed ERs in 1996.
- Designated Place (DPL) – Submunicipal communities defined by agreement with provincial governments. DPLs were introduced in 1996.
- Census Consolidated Subdivision (CCS) – Aggregations of CSDs within the same CD. CCSs were introduced in 1966.
- Population Centres (POPCTR) – Areas with a population of at least 1,000 and a population density of at least 400 persons per hectare. POPCTRs were introduced in 2011.

We have elected not to incorporate these units within the UNI·CEN database for several reasons. The DPL, POPCTR, and ADA geographies were introduced relatively recently, limiting the ability to compare data across time. The EA and DA geographies are sufficiently small that their boundaries are less stable than larger units and their small populations lead to frequent suppression to preserve privacy. Although FSAs are frequently used in survey research, they do not have stable geographic boundaries and crosscut standard statistical units. The CCS and ER geographies are less commonly used, and can be retrieved from Statistics Canada or aggregated from CSDs or CDs if required.

Temporal coverage

Table 1.1 summarizes the availability of tables at these geographic levels by Census year. In summary:

- Data pertaining to CMA/CAs and CTs are available since 1951, when Statistics Canada first disseminated data at that level of aggregation.
- Data pertaining to CSDs and CDs are available for each decennial Census since 1851, and for quinquennial Censuses since 1976, with the exception of 1956, 1966, and 1971.
- Data pertaining to PRs and Canada as a whole are available for all years except 1956 and 1971.

The remaining gaps may be remedied by future digitization projects. CSD, CD, and PR data for 1956 are published in printed volumes and could be transcribed. CSD and CD data for 1961 and 1966 exist in a raw digital form but official lists of codes pertaining to these geographic units would have to be transcribed for the data to become usable. Available digital data files for 1971 are in poor condition. With additional investigation, it may be possible to extract more variables and additional geographic levels from them; alternatively, there is the possibility of digitizing tables from printed volumes.

Table 1.1: Data Table Coverage

	Level of Geography					
	CT	**CMA/CA**	**CSD**	**CD**	**PR**	**CA**
2016	X	X	X	X	X	X
2011	X	X	X	X	X	X
2006	X	X	X	X	X	X
2001	X	X	X	X	X	X
1996	X	X	X	X	X	X
1991	X	X	X	X	X	X
1986	X	X	X	X	X	X
1981	X	X	X	X	X	X
1976	X	X	X	X	X	X
1971	X	X	n/a	n/a	n/a	n/a
1966	X	X	n/a	n/a	X	X
1961	X	X	X	X	X	X
1956	X	X	n/a	n/a	n/a	n/a
1951	X	X	X	X	X	X
1941	n/a		X	X	X	X
1931			X	X	X	X
1921			X	X	X	X
1911			X	X	X	X
1901			X	X	X	X
1891			X	X	X	X
1881			X	X	X	X
1871			X	X	X	X
1861			X	X	X	X
1851			X	X	X	X

Data table structures: Long and wide formats

The data used to create the **UNI·CEN** data tables are held by several libraries and government bodies. The data are also stored in a variety of tabular layouts and legacy file formats. We processed the original data to produce tables with standard tabular layouts and modern file formats. Each file contains nationwide data for one Census year at a single level of geography.

The data tables are provided in two formats:

- **Long Format Tables** are intended for use in statistics environments. Variables are identified using standardized variable codes and descriptions. “Long” tables are organized such that observations on a variable are stored in a single column, meaning that geographic unit information and variable names repeat. These tables include variable names, data quality and suppression information, and notes. The Long Format data table structure is adapted from Statistics Canada’s 2016 Census Profiles.[1]
- **Wide Format Tables** are intended for use in GIS environments. Each row corresponds to a single geographic unit and each column to a unique variable. Single-digit denominator, sex, and location (urban/rural, farm/non-farm) flags and a year suffix are appended to the variable names. Files are exported in both CSV and DBF formats. In the latter case, variable names are shortened to accommodate the DBF file format limit of 10 characters, and the tables are split into multiple parts to accommodate the DBF file format’s 254-row limit. A codebook in Excel format links variable names to their descriptions and, for DBF files, shortened variable names.

Users performing statistics will normally use the Long Format tables. GIS users will use the Wide Format tables. **Figure 1.2** illustrates the distinction between long and wide formats.

[1] https://www12.statcan.gc.ca/census-recensement/2016/dp-pd/prof/details/download-telecharger/comp/page_dl-tc.cfm?Lang=E

Table 1.2: Long versus wide format – an example

Long						
geoname	**varname**	**varcode**	**dq**	**total**	**male**	**female**
Toronto	Age 0–19	age0019	0	35	18	17
Toronto	Age 20–64	age2064	1	40	20	20
Toronto	Age 65+	age065p	1	25	12	13
Montreal	Age 0–19	age0019	1	30	14	16
Montreal	Age 20–64	age2064	1	45	20	25
Montreal	Age 65+	age065p	1	25	13	12
Vancouver	Age 0–19	age0019	1	30	15	15
Vancouver	Age 20–64	age2064	0	50	25	25
Vancouver	Age 65+	age065p	1	20	11	9

Wide						
geoname	**age0019_t**	**age0019_m**	**age0019_f**	**age2064_t**	**age2064_m**	**…**
Toronto	35	18	17	40	20	
Montreal	30	14	16	45	20	
Vancouver	30	15	15	50	25	

File formats

The Long Format data tables are stored in Stata binary format. We use this format for several reasons. Unlike comma-separated or tab-separated value (CSV or TAB) files, Stata files distinguish between string and numeric formats. They also have no practical limit on the number of rows or columns in the table. Finally, Stata files are easily imported into other statistical applications, including R, SPSS, and SAS, and are read natively into the Borealis repository, which in turn enables users to download the data in delimited formats.

The Wide Format data tables intended for use in GIS environments are available in DBF and CSV formats. The CSV files are exported so that the *geosid* and *geopart* are quoted, meaning that they should be interpreted as strings. The variable columns are exported as numeric.

NOTE: Testing indicates that QGIS and ArcGIS discard formatting in the CSV files and import all columns as numeric, thereby dropping leading and trailing zeros in the *geosid*. There are several solutions:

- For Esri products, see this useful guide to creating schema.ini files: https://sites.temple.edu/introgis/2021/08/31/loading-numeric-data-as-text-in-arcgis-pro/
- For QGIS, see this guide to creating CSVT files: https://bnhr.xyz/2018/08/07/specifying-csv-data-types-using-a-csvt-file.html
- A third solution is to create a new field after import that pads the leading and trailing zeroes as needed.

NOTE: All NA (null) values are recorded as –1 in the wide-format CSV files. This is to avoid automatic formatting of numeric variables as strings on export, as would occur if empty cells were exported as "" or "na".

Long format table layout

Table 1.3 summarizes the column definitions in the Long Format tables; detailed descriptions follow.

Table 1.3: Long Format Table Layout

Column	Definition	Format
time	Date to which data pertains (e.g., the Census year)	String
level	Level of geography: 1 = CT, 2 = CSD, 3 = CD, 4 = CMA, 5 = PR	Numeric
src	Source of data: cen = Census; nhs = NHS (2011 only)	String
geosid	Geographic unit identification codes used in the source files.	String
geopart	geouid of province in which CMA part is located if a CMA is split between two provinces; otherwise "00"	String
loc	Location code: t = all locations, includes: u = urban locations rt = rural locations, includes: rf = rural farm locations rn = rural non-farm locations *** Currently only available for 1961**	String
t_code	Variable code. Consistent across spatial units and years, formatted to become the variable name in wide-reshaped tables	String

Column	Definition	Format
t_level	Variable level within theme. Indicates the level of hierarchy if counts within a theme sum to subtotals. Note that these are standardized across time, so subtotals are not always present. -1 = noncount variable; 0 = total; 1 = variables that sum to total; 2 = variables that sum to 1, etc.	Numeric
t_theme	Theme to which variable belongs; the first four letters of t_code	String
t_themedescr	Textual description of the theme	String
t_name	Textual description of the variable. Standardized across time.	String
t_parent	If the variables are hierarchically organized (see t_level), contains the t_code to which the variable sums. If the subtotal variable is not present and t_level > 1, use t_denom. If t_level = 0, t_parent = “total”. If t_level = -1, t_parent = “noncount”	String
t_denom	The theme’s denominator for calculating percentages with a count variable as the numerator. If t_level = 0, t_parent = “total”. If t_level = -1, t_parent = “noncount”	String
val_t	Value – total population total	Numeric
val_m	Value – male total (if available)	Numeric
val_f	Value – female total (if available)	Numeric
flag_t	Code indicating data suppression or unavailability (total) extracted from source file	String
flag_m	Code indicating data suppression or unavailability (m)	String
flag_f	Code indicating data suppression or unavailability (f)	String
gnr/gnr_lf	Global non-response rates, etc., which are reported differently in different sources	Numeric
dq	Data quality code. See below for details.	String
notes	Numbered notes – extracted from source files	Varies

Detailed explanation of table columns: long format

Time (time) – The date with which the data are associated. For the aggregate Census and National Household Survey data we use a four-digit year.

Source (src) – The source of the data is noted here. At present, there are only two sources: the Census (cen) and, in 2011 only, the National Household Survey (nhs). Other sources may be noted as other data are incorporated into this project.

Geographic identification codes (geosid, level) – Each geographic area has a unique code within each Census year source dataset, the *geosid*. Starting in 1981, StatCan standardized its geographic coding scheme, so that geographic identification codes for "like" units are sustained across time. Where possible, the UNI·CEN project uses post-1981 standard codes for CMAs and provinces when constructing pre-1981 geosids. A longer-term project is underway that will link "like" CSD and CD codes across time using time-invariant "universal" codes: *geouid*s. The construction of the geosids is shown in **Table 1.4**.

Note that in the case of CSDs, CMAs, and PRs, "like" does not mean geographically identical. A municipality that annexes surrounding territory between Census years retains the same code even though its boundaries change. When municipalities are amalgamated or new municipalities are incorporated, however, new CSD geosids are created. Similarly, CMAs retain the same code when CSDs are added or removed from them. The boundaries of provinces and territories have also changed over time, however they retain the same identifiers across time. The CD geography is unstable, with boundaries being redrawn in some provinces from time to time. Their spatial cross-linkage across time is being analyzed as part of this project and will be the subject of a future report. CT geosids have been consistent since 1971. When tracts are merged, subdivided, or redrawn, they are given new codes; unchanged tracts retain the same codes.

Table 1.4: Construction of geographic identifiers, 1981-present

geou|sid	level	Unit	Digits	Format	Example
ctuid	1	Census Tract	10	cmauid (3) + ct identifier (7)	525|0001.00
csduid	2	Census Subdivision	5	cduid (4) + csd identifier (3)	3520|005 City of Toronto
cduid	3	Census Division	7	pruid (2) + cd identifier	35|20 City of Toronto
cmauid	4	CMA/CA	3	region (1) + CMA/CA (2)	5|35 Toronto
pruid	5	Province or Territory	2	region (1) + province identifier (1)	5|9 British Columbia

CMA part (geopart) – In cases where CMAs or CAs cut across provincial boundaries, the *pruid* of the province in which the part is located is indicated in the *geopart* column. For example, the Ottawa-Gatineau CMA is split between Ontario and Québec.

Geographic Name (geoname) – With the exception of census tracts, which do not have non-numeric names, the textual names of geographic units are stored in the *geoname* column. The prefix varies depending on the geographic unit. For example, in a CSD table, the geoname is *csdname*.

Location (loc) – In 1961, data are reported separately for urban and rural locations, for farm and non-farm locations, and for combinations of the two categorizations.

Value (val_t, val_m, val_f) – The observations associated with each variable are stored in the *val_** columns. Where available, variables pertaining to individuals not only include values for the total population, but also separate values for the male and female population.

Data suppression or unavailability flags (flag_t, flag_m, flag_f) – If present in the original data, variable-specific codes indicating data suppression or unavailability are stored in sex-specific *flag* columns.

Global non-response rate (gnr) – If present in the original data, the global non-response rate percentage is stored in the *gnr* column. A smaller number indicates lower risk of non-response bias and therefore inaccuracy. GNR rates pertain to geographic units and were disseminated in the 2011 Census and National Household Survey and 2016 Census.

Data quality flag (dq) – In some Census years, data quality flags are reported indicating incomplete enumeration. These are reproduced in the *dq* column.

Notes (notes) – In some Census years, the original tables contain numbered or lettered footnotes. These are recorded in the *notes* column.

Theme (t_theme) – Each variable code (*varcode*) has a three-character prefix that indicates its theme. For example, age cohort counts have the prefix “agec” and dwelling period of construction “dwpc”. The theme column can be used to quickly filter out unwanted variables from your working dataset.

Standardized variable code (t_code) – Every observation in the original datasets is assigned a standardized variable code, or *t_code*. To accommodate column name length limitations in commonly used statistics software when reshaping the data, these have a maximum length of 27 characters.

Standardized variable name (t_name) – Every *t_code* has a corresponding standardized plain-language variable name.

Denominator (t_denom) and “parent” variable for subtotal (t_parent) – The type of variable (count, non-count, and denominator) is indicated by *t_denom*. Non-count variables such as averages, medians, and dollar values are not additive, and are given a *t_denom* value of –1. Denominators for calculating percentages within the theme are given a *t_denom* value of 0. Count variable *t_denom*s range from 1 to 9 depending on whether they are hierarchically organized; that is, if some are subtotals of component rows. For example, theme *flcs*, “Families – Lone-Parent – Number of Children – Sex of Parent” is organized as follows in 1991:

Variable	t_level	t_parent	t_denom
Total Lone Parent Families	0	total	total
↳ With Female Parent	1	Total Lone Parent Families	Total Lone Parent Families
↳ 1 child	2	With Female Parent	Total Lone Parent Families
↳ 2 children	2	With Female Parent	Total Lone Parent Families
↳ 3 children	2	With Female Parent	Total Lone Parent Families
↳ 4 children	2	With Female Parent	Total Lone Parent Families
↳ 5 or more children	2	With Female Parent	Total Lone Parent Families
↳ With Male Parent	1	Total Lone Parent Families	Total Lone Parent Families
↳ 1 child	2	With Male Parent	Total Lone Parent Families
↳ 2 children	2	With Male Parent	Total Lone Parent Families
↳ 3 children	2	With Male Parent	Total Lone Parent Families
↳ 4 children	2	With Male Parent	Total Lone Parent Families
↳ 5 or more children	2	With Male Parent	Total Lone Parent Families

In this case, the by-child counts sum to the total by sex of the parent, which in turn sum to the total by both sexes. Within each theme, all *t_level* = 1 variables sum to the total in *t_level* = 0, all *t_level* = 2 variables sum to the *t_parent* variable, and so on. This coding strategy enables the calculation of percentages not only for the top-level total, but also for subtotals within the theme.

Wide format table layout

Table 1.5 summarizes the column definitions in the Wide Format tables; detailed descriptions follow.

Table 1.5: Wide Format Table Layout

Column	Definition	Format
geosid	Geographic unit identification codes used in the source files.	String
geopart	geouid of province in which CMA part is located if a CMA is split between two provinces; otherwise “00”	String
variables	Each t_code (see above) is its own column with loc and sex suffixes	Numeric

Detailed explanation of table columns: wide format

Geographic identification codes (geosid) – Each geographic area has a unique code within each Census year source dataset, the *geosid*.

CMA part (geopart) – In cases where CMAs or CAs cut across provincial boundaries, the *pruid* of the province in which the part is located is indicated in the *geopart* column. For example, the Ottawa-Gatineau CMA is split between Ontario and Québec.

Variables – Variables in the wide-format CSV tables are named using the *t_code* (see above) plus the year (*time*) and additional flags indicating the information found in the topic format tables' *sex*, *loc*, and *t_level* columns, constructed as follows with examples. Note that multi-character *loc* and *t_level* indicators are abbreviated to a single character.

Table 1.6: Construction of wide-format variable names

Variable name	t_code	time	sex	loc (abbreviated)	t_level (abbreviated)
		yyyy	t = total m = male f = female	t = total u = urban r = rural total f = rural – farm n = rural – non-farm	d = denominator n = non-count variable 1…9 = level in hierarchy
agec05091956mt1	agec0509	1956	m	t	1
mart_tot1961fud	mart_tot	1961	f	u	d

Variables in the wide-format DBF tables are truncated to fit within the DBF format's 10-character limit, according to the following rules:

- *t_theme* = first four characters of *t_code* (positions 1-4)
- three-digit number, sequential within each theme (positions 5-7)
- single-digit abbreviated *t_level* code (position 8)
- two digits of year (positions 9-10)

Thus, the t_code **abanabor1abofina** where year = 2001, sex = t, loc = t, and t_level = 2 is converted into the Wide Format variable name **abanabor1abofina2001tt2**. In the DBF file, it is truncated to **aban009201**, whereby the "2" indicates t_level = 2 and the "01" indicates year = 2001.

NOTE: The abbreviated variable names in the DBF files are not consistent across Census years and do not in themselves indicate the *sex* and *loc* information. Consult the codebook for this information. Also, take care when comparing data from different centuries where the two-digit year suffix is the same, for example 1901 and 2001.

Validation

We used several procedures to validate the standardized coding of the source variables and the construction of the data tables.

1. **Check for theme total variable:** Does every *t_theme* have a total variable (*t_level* = 0)?
2. **Variable checksum:** For all count variables, total all *t_level* = 1 variables within each *t_theme* and compare the result to the *t_level* = 0 value.
3. **Geographic checksum:** Sum the population counts for all nested geographic units and compare to population count of the aggregate unit. For example, we can sum the populations of all CSDs within a given CD and compare the result to the population of the CD.

Data sources

The reformatted data come from a wide range of sources. Some of these sources are primary, in that they are derived from “born digital” datasets originally created and disseminated by Statistics Canada. Others come from secondary sources: projects which have digitized printed Census volumes. **Table 1.7** summarizes the data sources and their original formats.

Table 1.7: Data Source Summary

Year	Geography	Source	Format
1851	CSD, CD, PR	The Canadian Peoples Project	Excel
1861	CSD, CD, PR	The Canadian Peoples Project	Excel
1871	CSD, CD, PR	The Canadian Peoples Project	Excel
1881	CSD, CD, PR	The Canadian Peoples Project	Excel
1891	CSD, CD, PR	The Canadian Peoples Project	Excel
1901	CSD, CD, PR	The Canadian Peoples Project	Excel
1911	CSD, CD, PR	The Canadian Peoples Project	Excel
1921	CSD, CD, PR	The Canadian Peoples Project	Excel
1931	CSD, CD, PR	Canadian Century Research Infrastructure	Excel
1941	CSD, CD, PR	Canadian Century Research Infrastructure	Excel
1951	CSD, CD, PR	Canadian Century Research Infrastructure	Excel
1951	CT, CMA	Western Early Postwar Census Tract Digitization Project	Excel
1956	CT, CMA	Western Early Postwar Census Tract Digitization Project	Excel
1961	CT, CMA, PR	University of Toronto Map and Data Library	SPSS
1966	CT, CMA, PR	University of Toronto Map and Data Library	SPSS
1971	CT, CMA	University of Toronto Map and Data Library	SPSS
1976	CT, CMA, CSD, CD, PR	University of Toronto Map and Data Library	SPSS
1981	CT, CMA, CSD, CD, PR	University of Toronto Map and Data Library	SPSS
1986	CT, CMA, CSD, CD, PR	University of Toronto Map and Data Library	SPSS
1991	CT, CMA, CSD, CD, PR	University of Toronto Map and Data Library	SPSS
1996	CT, CMA, CSD, CD, PR	Statistics Canada Census Profile Series	IVT
2001	CT, CMA, CSD, CD, PR	Statistics Canada Census Profile Series	IVT
2006	CT, CMA, CSD, CD, PR	Statistics Canada Census Profile Series	IVT
2011	CT, CMA, CSD, CD, PR	Statistics Canada Census Profile Series	CSV
2016	CT, CMA, CSD, CD, PR	Statistics Canada Census Profile Series	CSV
2021	CT, CMA, CSD, CD, PR	Statistics Canada Census Profile Series	CSV

The Canadian Peoples Project

Funded by a Canadian Foundation for Innovation grant, the TCP has assembled 100% microdata. As a byproduct of this project, they also digitized selected aggregate data tables from Census volumes. All aggregate data tables were provided by the project team directly. As of this writing they have not been disseminated elsewhere. The project will be completed in 2024.
URL: https://thecanadianpeoples.com/
Years: 1851–1921, decennial
Geography: CSD, CD, PR
File Format: Excel

Canadian Century Research Infrastructure

Also funded by a CFI grant, the CCRI project ran from 2003 to 2009. The principal goal was to create microdata samples for the 1911–1951 period, however they also digitized selected aggregate tables from Census volumes. Where the CCRI and TCP aggregate data table coverage overlaps (1911, 1921), the newer TCP files are favoured.
URL: https://ccri.library.ualberta.ca/enindex.html
URL: http://web5.uottawa.ca/ccri/CCRI/Home.html
Years: 1911–1951, decennial
Geography: CSD, CD, PR
File Format: Excel

Western Early Postwar Census Tract Digitization Project

See **UNI·CEN Documentation Report #4** for details.
Years: 1951, 1956
Geography: CT, CMA
File Format: Excel

University of Toronto Map and Data Library

As a result of past data preservation projects, the University of Toronto library retains conversions in SPSS format of Statistics Canada data originally disseminated in punch card or magnetic tape formats. These datasets are Basic Summary Tabulations. As the consistency of multivariate crosstabulated variables is uneven over time, only sex crosstabs are retained, as appropriate. The 1961 and 1966 files are for enumeration areas only and contain no data suppression or random rounding. This being the case, higher levels of geography were constructed through aggregation. This was facilitated in 1961 by the manual reconstruction of the “official lists” of CSD and CD names for each geographic code. This task was not performed for 1966 due to resource constraints. See **UNI·CEN Documentation Report #4** for details. In later years, separate tabulations are available for different levels of geography. Some of the

source files were corrupted and could not be read; we are investigating whether these gaps can be remedied using alternative sources.
URL: https://mdl.library.utoronto.ca/census-of-canada
Years: 1961–1991, quinquennial
Geography: CT, CMA, CSD, CD, PR for all years *except* no CSD and CD in 1966 and 1971 and no PR in 1971.
File Format: SPSS

Statistics Canada Profile Series (Beyond2020 IVT)

For 1996, 2001, and 2006 we used the Profile Series disseminated in Beyond2020 IVT format. The data were exported as CSV files from the proprietary IVT format.
URL: http://odesi2.scholarsportal.info/
Years: 1996–2006, quinquennial
Geography: CT, CMA, CSD, CD, PR
File Format: IVT (Beyond2020)

Statistics Canada Profile Series (CSV)

Starting in 2011 we make use of the downloadable, CSV-format profiles available directly from Statistics Canada.
URL: Census 2011
https://www12.statcan.gc.ca/census-recensement/2011/dp-pd/prof/details/download-telecharger/comprehensive/comp-csv-tab-dwnld-tlchrgr.cfm?Lang=E#tabs2011
URL: National Household Survey 2011
https://www12.statcan.gc.ca/nhs-enm/2011/dp-pd/prof/details/download-telecharger/comprehensive/comp-csv-tab-nhs-enm.cfm?Lang=E
URL: Census 2016
https://www12.statcan.gc.ca/census-recensement/2016/dp-pd/prof/details/download-telecharger/comp/page_dl-tc.cfm?Lang=E
URL: Census 2021
https://www12.statcan.gc.ca/census-recensement/2021/dp-pd/prof/details/download-telecharger.cfm?Lang=E
Years: 2011–2021, quinquennial
Geography: CT, CMA, CSD, CD, PR
File Format: CSV

NOTES:

- The TCP and CCRI files are under review and will be made available later in 2022.
- The 2021 Census will be incorporated into UNI·CEN following the final Census Profile release on November 30, 2022.