Catalogue no. 46260001
Issue no. 2021001

Linkable Open Data Environment

The Open Database of Addresses (ODA)

Metadata document: concepts, methodology and data quality

Version 1.0

Data Exploration and Integration Lab (DEIL)
Centre for Special Business Projects (CSBP)

Release date: April 29th, 2021

Statistics Canada Statistique Canada

Canada

How to obtain more information

For information about this product or the wide range of services and data available from Statistics Canada, visit our website, www.statcan.gc.ca.

You can also contact us by:

Email at STATCAN.infostats-infostats.STATCAN@canada.ca

Telephone, from Monday to Friday, 8:30 a.m. to 4:30 p.m., at the following numbers:

- Statistical Information Service 1-800-263-1136
- National telecommunications device for the hearing impaired 1-800-363-7629
- Fax line 1-514-283-9350

Depository Services Program

- Inquiries line 1-800-635-7943
- Fax line 1-800-565-7757

Standards of service to the public

Statistics Canada is committed to serving its clients in a prompt, reliable and courteous manner. To this end, Statistics Canada has developed standards of service that its employees observe. To obtain a copy of these service standards, please contact Statistics Canada toll-free at 1-800-263-1136. The service standards are also published on www.statcan.gc.ca under "Contact us" > "Standards of service to the public."

Note of appreciation

Canada owes the success of its statistical system to a long-standing partnership between Statistics Canada, the citizens of Canada, its businesses, governments and other institutions. Accurate and timely statistical information could not be produced without their continued co−operation and goodwill.

Published by authority of the Minister responsible for Statistics Canada

Cette publication est aussi disponible en français.

Table of Contents

Acknowledgments

This project benefited from a collaboration with OpenAddresses, specifically on the code for address compilation and processing. Their foundational work and insights on that are gratefully acknowledged.

1. Overview

For the purpose of exploring open data for official statistics and to support geospatial research across various domains, the Data Exploration and Integration Lab (DEIL) undertook a project to create an accessible and harmonized database of addresses released as open data by various levels of government within Canada[1]. This document details the process of collecting, compiling, and standardizing the individual datasets of address points that were used to create the experimental *Open Database of Addresses* (ODA) which is made available under the *Open Government Licence - Canada*[2].

Statistics Canada acknowledges the many local governments that produce public address listings, which are the source of the ODA. These addresses will also be integrated into a new National Address Register (NAR) of residential and non-residential addresses, to be made available later this year by Statistics Canada. Compiled from a multitude of sources, The NAR will be a comprehensive and standardized source of publicly-available addresses and related geographic coding. It is part of the Data Strategy for the Federal Public Service.

In its current version (version 1.0), the ODA contains over 10 million addresses. The database is expected to be updated periodically as new open datasets of address points from government sources become available, until full integration in a National Address Register. The ODA is provided as a zipped CSV file at the provincial or territorial level.

In addition, the compilation and processing codes used to generate the ODA are available at https://github.com/CSBP-CPSE/Canadian-Open-Address-Point-Processing. This allows for automated updates of the data, and in this way, a comprehensive civic address database can be refreshed in real time as municipalities and local governments update their open data files.

This dataset is one of a number of datasets created as part of the Linkable Open Data Environment (LODE). The LODE is an exploratory initiative that aims at enhancing the use and harmonization of open data from authoritative sources by providing a collection of datasets released under a single licence, as well as open-source code to link these datasets together. Access to the LODE datasets and code are available through the Statistics Canada website and can be found at:

https://www.statcan.gc.ca/eng/lode

2. Data Sources

Across Canada, local governments create and maintain civic addresses. The ODA derives its record directly from these authoritative sources, which made these records available under an open data license compatible with the Government of Canada Open Data License. Hence, multiple data sources were used to create the ODA. The compilation expanded on the work initiated by the organization OpenAddresses, which aggregates open address data globally on its GitHub page[3]. In total, address points from 99 data providers were used (though some sources overlap geographically).

The data providers, which include municipal, regional, and provincial levels of government, are listed in Supplemental Table 1, along with links to the original data sources. Attribution to each of these data sources is listed, as per the licence requirements. Where applicable, licence versions are also shown. For further information on the individual licenses, users should consult directly with the information provided on the open data portals for the various data providers.

[1] This includes municipal, regional, and provincial.
[2] See: https://open.canada.ca/en/open-government-licence-canada
[3] See: https://www.github.com/openaddresses/openaddresses

3. Reference Period

Ideally, the reference period would be the period for which the underlying address data refers to. Such information, however, was not always available from the open data portals. Refreshing frequency of the original databases varies as well across sources, with some reporting weekly updates and others semi-annual, annual, or irregular updates; Supplemental Table 1 provides the date each dataset used in the ODA was downloaded. Data were gathered from January to April of 2021. Users are cautioned that the download date should not be used to indicate the reference period of the data. If specific information concerning the reference period of data is required, users should contact the appropriate data providers shown in Supplemental Table 1.

4. Target Population

The goal of the ODA is to create a comprehensive and harmonized repository of civic addresses made available from local government open data sources across Canada. Addresses may identify residential buildings, commercial or institutional buildings, or simply refer to parcels. Furthermore, buildings and parcels may be assigned multiple addresses. The ODA includes all unduplicated civic addresses that were possible to compile from the local and provincial government sources listed in Supplemental Table 1.

5. Compilation Methodology

The compilation methodology of the ODA is nearly fully automated, to allow for potentially frequent update of the database. As more local governments move to high frequency updates of their open civic address databases, this will allow for a near real-time comprehensive ODA.[4]

The code used for the collection and pre-processing is based on a modified version of the processing pipeline developed by OpenAddresses. This process downloads the individual data files and standardises them to the same set of columns using a dictionary mapping described in JSON input files, and includes minor processing when necessary such as separating addresses into separate civic number and street name fields, or else combining fields from the original data where necessary. For every source, this process produces a CSV file with standardised address point data.

Users should note that, within the 99 datasets obtained, each data provider attached a different set of variables to the address point data. In some instances, the different fields making up the address (street number, street name, etc.) were already provided separately, while in other cases they needed to be parsed from more complete address fields. Likewise, in some cases street types and directions were standardized to common abbreviations, while in others they were provided in a fully expanded form. Finally, the data was also provided in a range of file formats, from simple comma separated value (CSV) files, to geographic file formats such as shapefiles or geojson, or else the data was accessed programmatically through an API.

The compilation codes account for these differences and harmonize sources into a standard format. Hence, the adoption or modification of formatting standards at the sources would require future adjustments in the processing codes.

Additional processing was done in four steps:

1. *Standardisation*: Street addresses were parsed and standardised into street name, street type, and street direction fields (e.g., to transform “MAIN STREET NORTH” into “MAIN”, “ST”, “N”). This process was done with a modified version of RASK (Road Attribute Search Key), a tool used at Statistics Canada for standardising addresses from administrative sources for record linking. Any sources missing a full address column had this information filled in by concatenating the unit, street number, and full street name. Sources missing city names had this imputed from the

[4] The compilation code is accessible at: https://github.com/CSBP-CPSE/Canadian-Open-Address-Point-Processing. This code is based on a modified version of the processing pipeline developed by OpenAddresses.

name of the source file (e.g., to transform “city_of_banff.csv” into “BANFF”). The processed and imputed columns in the database are those with the suffix “_pcs”.

2. *Cleaning*: Records with missing coordinates or street names were dropped. All coordinates were truncated to 5 decimal places (corresponding to metre-level precision). The files were deduplicated at the level of the original sources by dropping records with identical coordinates, units, street number, and standardised street name.

3. *Spatial join*: All records were spatially joined with the Statistics Canada 2016 Census Subdivision (CSD) boundary file to assign CSDUID, CSD name, and PRUID. A small number of records that could not be placed into CSDs were dropped.

4. *Final Merge*: All data sources were combined into a single Canada-wide dataset of address points. Duplicates were dropped a second time following the same criteria as in step 2. Because it occurs in the original data that the same street address may have multiple representative coordinates, a group identifier was computed by grouping together entries with the same CSDUID, street number, and processed address components, so that entries with the same group id will correspond to the same street address and can be further processed by the end user if necessary.

In Step 4, a unique identifier is computed and assigned to each record. This unique identifier is the result of a hash using the Blake2b algorithm in Python’s hashlib library, generated from a concatenation of the coordinates and processed address fields (the civic number, unit, and standardised street name). This means that a unique record for the purpose of the ODA is defined only by its coordinates and street address, and not by other fields such as provider or city.

In some cases it was necessary to download and pre-process the data before passing it through the initial collection pipeline (for example, to account for files formatted in such a way as to not be able to be read by the pipeline, encoding issues, or in the case of Montreal, to split address ranges into individual rows). The pre-processing scripts and a description are available on the project GitHub page.

6. Data Dictionary

This data dictionary below describes the variables contained within the exploratory ODA.

Variable - Latitude	
Name	latitude
Format	Double
Source	Provided as is from original data
Description	The latitude in decimal degrees of the address point truncated to 5 decimal places.

Variable - Longitude	
Name	longitude
Format	Double
Source	Provided as is from original data
Description	The longitude in decimal degrees of the address point truncated to 5 decimal places.

Variable - Source ID	
Name	id_source
Format	Alphanumeric
Source	Provided from original data
Description	Unique object or field ID assigned to the records as recorded in original data sources.

Variable - ODA ID	
Name	id
Format	Alphanumeric
Source	Internally generated during data processing
Description	Unique ID assigned to records derived from a hash computed from the coordinates and standardized address fields.

Variable - Group ID	
Name	id_group
Format	Alphanumeric
Source	Internally generated during data processing
Description	Field ID assigned to records sharing address information (civic number, street name, street type, street direction) but with different geocoordinates.

Variable – Civic Number	
Name	street_no
Format	String
Source	Provided from original data
Description	The street number of the address, either as provided or else parsed from the full address.

Variable – Full Street Name	
Name	street
Format	String
Source	Provided from original data
Description	The street name of the address, including street type and direction where applicable, either as provided or else parsed from the full address.

Variable – Street Name	
Name	str_name
Format	String
Source	Provided from original data
Description	The street name of the address, without type and direction, as provided.

Variable – Street Type	
Name	str_type
Format	String
Source	Provided from original data
Description	The street type of the address as provided.

Variable – Street Direction	
Name	str_dir
Format	String
Source	Provided from original data
Description	The street direction of the address as provided.

Variable – Unit	
Name	unit
Format	String
Source	Provided from original data
Description	The street unit of the address either as provided or else parsed from the full address.

Variable – Municipality	
Name	city
Format	String
Source	Provided as is from original data
Description	The name of the municipality.

Variable – Postal Code	
Name	postal_code
Format	String
Source	Provided as is from original data
Description	The postal code of the address.

Variable – Full address	
Name	full_addr
Format	String
Source	Provided as is from original data or imputed
Description	The full address, either as provided or else created from the concatenation of the other fields.

Variable – Processed City	
Name	city_pcs
Format	String
Source	Internally generated during data processing
Description	The name of the municipality, imputed from the file name of the original source if necessary.

Variable – Processed Street Name	
Name	str_name_pcs
Format	String
Source	Internally generated during data processing
Description	The standardised street name of the address, without type and direction.

Variable – Processed Street Type	
Name	str_type_pcs
Format	String
Source	Internally generated during data processing
Description	The standardised street type of the address.

Variable – Processed Street Direction	
Name	str_dir_pcs
Format	String
Source	Internally generated during data processing
Description	The standardised street direction of the address.

Variable – Census subdivision unique identifier	
Variable Name	csduid
Data Format	Integer
Source	Canadian census subdivision boundaries 2016 (Statistics Canada GeoSuite product)
Description	The census subdivision ID where the address is located.

Variable – Census subdivision name	
Name	csdname
Format	String
Source	Canadian census subdivision boundaries 2016 (Statistics Canada GeoSuite product)
Description	Name of census subdivision.

Variable – Province unique identifier	
Name	pruid
Format	Integer
Source	Canadian census subdivision boundaries 2016 (Statistics Canada GeoSuite product)
Description	The province ID where the address is located.

Variable – Data Provider	
Name	provider
Format	Text (String)
Source	Created based on origins of input dataset
Description	Name of the municipality, region, or province/territory that provided the dataset.

7. Data Accuracy

All addresses were collected from authoritative government sources, made available to the public as open data. In general, other than the processing required to harmonize the different sources into one database, the underlying datasets obtained from the various open data portals were taken “as-is”.

During the processing stage to create the ODA, several steps were taken to standardize the output, including the standardization of street types and a deduplication of entries. It is possible that the process used to standardize the addresses may have introduced some errors, but these are expected to be minimal. Likewise, it is possible that duplicate entries remain in the database. To control for possible processing inaccuracies, the full address column is also provided without standardization applied.

The ODA represents the government open data that was found at the time of compilation and should thus not be interpreted as a complete or objective “ground-truth” of what addresses actually exist in Canada. The current coverage across Canada of the ODA remains incomplete in some jurisdictions. The gaps in the database reflect areas where open data on addresses from government sources were not found. Some of these gaps may be closed as more civic addresses are release as open data by local governments.

8. Geographic Representation

The Open Database of Addresses is available on the Statistics Canada website with coordinates given in latitudes and longitudes using the WGS84 standard ellipsoid.