Catalogue no. 21260003
Issue no. 2023001

Exploring Open Data

The Open Database of Businesses (ODBus)

Metadata document: concepts, methodology and data quality

Version 1.0

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

Release date: November 28, 2023

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.

Table of Contents

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 a harmonized database of businesses released as open data by various levels of government and other entities within Canada. This document details the process of collecting, compiling, and standardizing the individual datasets of the Open Database of Businesses (ODBus), which is made available under the Open Government Licence – Canada.[1]

In its current version (version 1.0), the ODBus contains approximately 450,000 records. As data collection is from available open sources, and many business micro datasets are derived from business licence registration, this version of the ODBus focuses on identifying individual licences. Businesses may be duplicated if they hold multiple business licences. This is further detailed in section 4, Target Population. The database is expected to be updated periodically as new open datasets become available.

This dataset is one of several datasets created as part of the Linkable Open Data Environment (LODE). The LODE is an initiative that aims to enhance 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

The ODBus is comprised of data from 70 sources. The data providers, which include multiple levels of government and other entities, are outlined in a supplementary CSV file of data sources accompanying the data, including attribution to each data source as per the licence requirements. For further information on the individual licences, users should consult directly with the information provided on the open data portals of the various data providers.

While the province of Quebec also provides their entire business registry[2] as open data, this was not included in the ODBus due to licencing incompatibility.

3. Reference period

The supplementary CSV file on data sources lists either the update frequency or the date each underlying dataset was last updated by the provider (when known). Data were gathered between May 2022 and December 2022. 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.

4. Target population

The Open Database of Businesses targets businesses across Canada that are provided within open business directories and business license datasets. The scope of businesses collected relies upon the availability of open data provided from business directories and municipal, provincial, and federal sources. Therefore, businesses that require a licence to operate are more likely to be included than other types of businesses, although this varies by the data source. Depending on the data provider, if businesses are registered separately for different licences

[1] https://open.canada.ca/en/open-government-licence-canada
[2] https://www.donneesquebec.ca/recherche/fr/dataset/registre-des-entreprises

based on business activity, then the same business may appear in multiple records due to their unique licences.

This database does not define or identify hierarchical structures of businesses and may contain single operating locations where goods or services are provided as well as head offices and regional offices.

Businesses with and without employees are both in scope for this database, however, it is not possible to determine within which category a business falls unless the data provider listed an employee count within the source dataset. The ODBus is meant to enhance access to open data on businesses across Canada and is not a complete listing of businesses or representative sample of business activity in Canada. Users may consult the list of data sources to assess the current coverage of the ODBus in the Supplemental table provided with the data download.

Only minimal editing of the original datasets was performed. As work on the experimental ODBus progresses, definitions and thresholds will evolve. Users are reminded that unedited data can be obtained directly from the open data portals or from the various data providers, as listed in the Supplemental table of sources mentioned above.

Differentiation from the Business Register

The Business Register (BR) is Statistics Canada's continuously maintained central repository of information on businesses and institutions operating in Canada[3]. The ODBus database is separate from the Business Register as well as other business data collected at Statistics Canada through surveys and other administrative sources. The BR was not used to validate any business entries and cannot be compared to the ODBus as the data sources, processing methods and maintenance are different.

There are 446,573 records in the ODBus, however this does not cover all businesses in Canada. This count also does not include the Enterprise Register of Québec[4], which contains over 2.6 million business records. As previously mentioned, these records were not included due to license incompatibility. As of December 2022, the official release based on the Business Register reports that there were 1,336,336 employer businesses in Canada and 3,021,567 non-employer businesses with annual revenues greater than $30,000.[5]

5. Compilation methodology

The primary processing component for the database comprised reformatting the source data to CSV format and mapping the original dataset attributes to standard variable (column) names. To compile the data into a single database, the following steps were taken:

- The original data files and fields were converted to standard formats and fields using the custom software OpenTabulate[6].
- Concatenated address data were parsed and separated into their corresponding components (e.g., unit, street number and name, city name, etc.) using libpostal[7] a natural language processing solution for address parsing.
- Entries missing latitude and longitude information were geocoded by matching parsed addresses against the Open Database of Addresses.
- Deduplication using literal string matching. This was done in a conservative manner to avoid false positives (for more details, see Data standardization).

[3] Business Register (BR) (statcan.gc.ca)
[4] Registre des entreprises - Registre des entreprises - Données Québec (donneesquebec.ca)
[5] The Daily — Canadian business counts, December 2022 (statcan.gc.ca)
[6] https://pypi.org/project/opentabulate/
[7] https://github.com/openvenues/libpostal

- Cleaning and standardization (for more details, see Data standardization).

While effort was made to ensure that the data is correct, it is possible that the scripts used to process and parse the addresses may unintentionally cause other, undetected, errors. Should any such errors be reported, they will be corrected in future versions of the ODBus.

In general, the data included in the ODBus represents what is available from the original sources without imputation. The exception to this is the geocoding of entries missing coordinates, and the imputation of CSD names and NAICS codes, as discussed below.

Geocoding

Records that did not include geocoordinates from the source were geocoded by matching entries against the Open Database of Addresses (ODA)[8].

Fuzzy matching was used to compare parsed addresses (street number, street name, city) to corresponding columns in the ODA. Records that scored above a conservatively set threshold were taken to be valid matches. The geo_source column indicates whether the coordinates of a record were provided by the original source or if they were geocoded.

Imputation of NAICS codes

The original data sources use a variety of standards, classifications, and nomenclature to describe the business type. This database retains all the original descriptions from the data sources, described in section 6 Data dictionary.

However, with the goal of standardizing the enterprise classification, the ODBus uses Statistics Canada's business classification standard, the North American Industry Classification System (NAICS)[9] to provide a standard definition of business type.

Based on the NAICS sector definitions given in Table 1, information found in the source business descriptions and business sectors were used to match 86% of business to their corresponding two-digit NAICS codes. Of the NAICS codes available in the Open Database of Businesses, 25% were present in the original source material and 61% were deduced using keywords found in the business description that were matching the sector definitions. Imputation of NAICS codes is done conservatively to avoid false positives.

Table 1: North American Industry Classification System (NAICS) Canada 2022 Version 1.0

Code	Sector
11	Agriculture, forestry, fishing and hunting
21	Mining, quarrying, and oil and gas extraction
22	Utilities
23	Construction
31-33	Manufacturing
41	Wholesale trade
44-45	Retail trade
48-49	Transportation and warehousing
51	Information and cultural industries
52	Finance and insurance
53	Real estate and rental and leasing
54	Professional, scientific and technical services

[8] https://www.statcan.gc.ca/en/lode/databases/oda
[9] https://www.statcan.gc.ca/en/subjects/standard/naics/2022/v1/index

55	Management of companies and enterprises
56	Administrative and support, waste management and remediation services
61	Educational services
62	Health care and social assistance
71	Arts, entertainment and recreation
72	Accommodation and food services
81	Other services (except public administration)
91	Public administration

Imputation of census subdivision (CSD) names

Census subdivision (CSD)[10] names were derived from latitude and longitude coordinates. These are placed into the corresponding CSDs by linking the coordinate points to the CSD polygons through a spatial join operation using the Python package GeoPandas[11].

Data standardization

Due to the different standards adopted in the original sources, steps that were taken to standardize the data could possibly produce errors. The key principles of the methodology used were the avoidance of false positives and of significant alterations to the data. The methodology and limitations of each technique are described below. Trivial cleaning techniques, such as removal of whitespace characters and punctuation removal, are omitted from discussion.

Address Parsing

The libpostal address parser, an open-source natural language processing solution to parsing addresses, was used to split concatenated address strings into strings corresponding to address variables, such as street name and street number. Occasionally, addresses were split incorrectly due to unconventional formatting of the original address. While effort was made to identify and correct these entries in the final database, some incorrectly parsed entries may have remained undetected. Exceptions are entries with street numbers of the form of two numbers separated by a hyphen or space. Entries of this form usually indicate that the address parser incorrectly parsed a numbered street name (e.g., “123 100 ave” is parsed into the street number “123 100” and the street name “ave”), or else that a unit has not been identified correctly (as in “3-100 main st”). Numbers of this form are automatically separated, where the right most number is prepended to the street name if the street name is a variant of the word “street” or “avenue.”

Removal of duplicates

Potential duplicate results were identified by searching for exact matches between variables. Entries were marked as duplicates if they matched across the following variables: business names, licence numbers, street numbers, street names, postal code, business sector, business description, licence type and primary NAICS code. In cases where one record had a null value and the other record had a value for a particular variable, they would be considered a match if the remaining variables all had matching values.

In total, 10,330 records were identified as duplicates. The majority of these occurred between datasets which contained businesses from the same geographical area.

Cleaning and Standardization

Business records that did not contain a business name were removed from the final database due to limited identification possibilities. Business status was changed to a unified labeling namely Active, Not Active, Pending and

[10] https://www12.statcan.gc.ca/census-recensement/2021/ref/dict/az/Definition-eng.cfm?ID=geo012
[11] GeoPandas is a Python package for the manipulation of geospatial data: http://geopandas.org/index.html.

Closed, where Closed was removed from the final dataset. The standardisation was as follows:

> The following codes and labels were change to “Active”: '1', 'OPEN', 'Licensed', 'Approved', 'APPROVED', 'Issued', and 'ISSUED’
>
> The following codes and labels were changed to “Not Active”: 'Move in Progress', 'Invalid Status Code', and 'Close in Progress'
>
> The following codes and labels were changed to “Pending”: 'Pending', 'Pending Renewal', 'Renewal Licensed', 'Renewal Notification Sent', 'renewal notice', and 'RENEWAL NOTICE'
>
> The following codes and labels were changed to “Closed”: ‘Out of Business', 'Inactive', and 'Cancelled'

The latitudes and longitudes of businesses were rounded to 5 decimal points; however, certain records may provide less precision if they were received as such from the source material. All personal information such as mailing address, phone number, and fax were removed from the final database.

While the scope of this database covers businesses in Canada and obvious out-of-country businesses were removed, it is possible that a few businesses outside of Canada may remain. Some businesses were classified by the data providers as “out of town.” Some investigation suggested these businesses did reside in the same location as the data provider; however, they conduct business elsewhere and therefore were classified as such. They were left as is in the final dataset.

6. Data dictionary

The data dictionary describing the variables of the ODBus is available in a supplementary CSV file that accompanies the data download, titled “ODBus_record_layout”.

7. Data accuracy

All business data in the ODBus were collected from open data sources, either from open data portals or otherwise public webpages. In general, other than the processing required to harmonize the different sources into one database, the underlying datasets were taken “as is.”

Natural language processing methods are used to do the parsing and separation of address strings into address variables, such as street number and postal code. The methods are reputable for performance and accuracy, but as with all statistical learning methods, they have limitations as well. Poor or unconventional formatting of addresses may result in incorrect parsing. At this stage, no further integration with other address sources was attempted; hence, although address records are generally expected to be correct, residual errors may be present in the current version of the database.

8. Contact Us

The LODE open databases are modelled on ongoing improvement. To provide information on additions, updates, corrections, or omissions, or for more information, please contact us at statcan.lode-ecdo.statcan@statcan.gc.ca. Please include the title of the open database in the subject line of the email.