Catalogue no. 37260001
Issue no. 2021001

Exploring Open Data

The Open Database of Educational Facilities (ODEF)

Metadata document: concepts, methodology and data quality

Version 2.1

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

Release date: November 28, 2022

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

Acknowledgments

A first version of the database was realized with funding by Indigenous Services Canada (ISC) and Crown-Indigenous Relations and Northern Affairs Canada (CIRNAC). This updated version, with inclusion of Official Language Minority Schools, was realized with funding from Treasury Board Secretariat (TBS) and consultation from Canadian Heritage (PCH). Valuable feedback and comments were provided by these organizations and they 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 educational facilities 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 educational facilities that were used to create an update to the second version of the *Open Database of Educational Facilities* (ODEF), which is made available under the *Open Government Licence – Canada*[2].

In its current version (version 2.1), the ODEF contains 18,982 individual records. For this update to the database, information on public Official Language Minority Schools (OLMS) was added to the existing ODEF version 2.0. An OLMS is defined as an English-speaking school in Quebec, or a French-speaking school outside of Quebec. 967 existing records were identified as an OLMS and 38 new records were added for version 2.1. As the OLMS data were collected more recently than the ODEF data, some facilities had addresses updated to reflect changes. Additionally, latitude and longitude coordinates of OLMS facilities were updated for the matched ODEF records with missing data. CMA information was added with a spatial join using the sf package[3] in R for all records with available coordinate data to be consistent with the OLMS. The database is expected to be updated periodically as new open datasets become available. The ODEF is provided as a compressed comma separated values (CSV) file.

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

Multiple data sources were used to create the ODEF. The data providers, which include multiple levels of government, are provided in the Supplementary material as Table 1, including attribution to each data source as per the licence requirements. Where applicable, licence versions are also shown. 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. In addition to openly licensed databases, the ODEF also includes a set of publicly available listings of educational facilities for which permission to include was granted by the data providers. With the inclusion of the OLMS variable for Version 2.1 of the ODEF, all sources for OLMS information are included in Table 2 in the Supplemental material. For each province and territory where multiple data sources on OLMS status were found, one primary data source was chosen that had the greatest number of records and useful attributes such as grade levels and address information.

In addition to the primary sources listed in Table 2, validation was done by comparing lists to the webpages of official minority language school boards. This led to the addition of a small number of facilities that had been missing from the original data sources. The supplementary sources used are listed in Table 3 in the Supplemental material.

3. Reference period

The supplementary material lists either the update frequency or the date each underlying dataset was last updated by the provider (when known), as well as the date each dataset used in the ODEF was downloaded or provided by the data owner. Data were gathered between August 2019 and March 2021 for the ODEF data, and from November

[1] This includes municipal, regional, and provincial.
[2] See: https://open.canada.ca/en/open-government-licence-canada.
[3] SF is an R package for the manipulation of spatial data https://r-spatial.github.io/sf/

2021 to March 2022 for the OLMS status. 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

An education facility is a physical site at which the primary activity is imparting instruction to a body of students or participants. All education facilities in Canada are in scope for this dataset. These include all levels of education, private and public schools with no exclusions for funding arrangement, operator type, subject area, denomination, student type, location, etc.

As a result of this definition, the database covers facilities such as early childhood education, kindergarten, elementary, secondary, and post-secondary institutions, and specific vocational training centres (such as hairdressing schools). The database does not include virtual educational institutions.

For the OLMS status the target population is restricted to public K-12 official language minority schools. This may include both traditional schools and alternative schools if they are controlled by official language minority school boards or authorities.

Only minimal editing of the original datasets was performed. As work on the experimental ODEF 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.

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. A data dictionary of the variables used is provided in section 6. Data dictionary. To compile the data into a single database, the following was done:

- Concatenated address data were parsed and separated into their corresponding components (e.g. unit, street number and name, city name, etc.) using libpostal,[4] a natural language processing solution for address parsing.
- Deduplication using literal and fuzzy string matching. This was done in a conservative manner to avoid false positives (for more details, see Data standardization).

The original data files and fields were converted to standard formats and fields using the custom software OpenTabulate.[5] A limited number of entries were manually edited when it was clear that the parsing had not been done correctly. An example is addresses with hyphenated numbers such as “1035-55 street nw”, which may have been interpreted as having a civic number of “1035-55” and a street name of “street nw”, rather than a civic number of 1035, and a street name of “55 street nw”. 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 ODEF.

In general, the data included in the ODEF is 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 ISCED levels, discussed below.

In version 2 of the ODEF, the unique identifier has been changed from an integer to a hash computed from the facility name, address, and source id (if available) of the record.

[4] See: https://github.com/openvenues/libpostal.
[5] See: https://pypi.org/project/opentabulate/.

Geocoding

Records that did not include geocoordinates from the source were geocoded using the ESRI ArcGIS Online (AGOL) geocoder and the OpenStreetMap Nominatim geocoder. The AGOL geocoder returns coordinates, as well as a score and a geocoding type. Only records with a score above 90 and with address type indicating the coordinates were either an address, subaddress, point of interest, or intersection were retained for the final database. Records that could not be geocoded to the level of precision described above were then passed to the Nominatim geocoder. Schools were searched for using school names, city, and province, and were kept if the returned school name was a close match to the original school name. The Geo_Source column indicates if the coordinates of a record were provided by the original source or if they were geocoded.

Imputation of ISCED levels

The original data sources use a variety of standards, classifications and nomenclature to describe the education level or grade range. The ODEF uses the International Standard Classification of Education (ISCED)[6] to provide a standard definition of an education level. This required the conversion of a facility's grade range or education level to a corresponding ISCED level.

ISCED levels were derived from the grade range indicated by the data provider if available. Otherwise, education level was converted to a grade range, which was then mapped to ISCED levels. Entries in the original data that did not contain education level information were not assigned to ISCED levels and so these fields are blank in the ODEF.

Table 1 shows the direct mapping of ISCED levels from grade ranges and Table 2 shows the grade ranges in an education level by province and territory. It should be noted that the definition of "kindergarten" (K) as an education level label varies by providers as some of these schools support early childhood education. To avoid false positives, facilities that indicate support for pre-elementary students, as described by an education level string (not a grade range string), were not assigned values for the ISCED010 column. For example, Early Childcare Services in Alberta includes Kindergarten and may also include services for younger children, but was only mapped to ISCED020. Despite some of these facilities supporting childhood education, the notion of pre-elementary appears to vary between data providers and schools. This is shown in Table 2 with the assignment of "pre-elementary" to kindergarten when converted to a grade range.

Table 1: Data dictionary variables and their corresponding ISCED levels

Variable	Name	ISCED level	Grade range
Early childhood education	ISCED010	010	Pre-K
Kindergarten	ISCED020	020	K
Elementary	ISCED1	1	1-6
Junior secondary	ISCED2	2	7-9
Senior secondary	ISCED3	3	10-12
Post-secondary	ISCED4+	4+	-

[6] See: https://doi.org/10.1787/9789264228368-en.

Table 2: Education level conversion definition to grade ranges based on the province/territory

Province / Territory	Pre-elementary / kindergarten	Elementary / primary	Junior high / middle	Senior high
N.L., P.E.I., N.S., Alta., N.W.T., NvT.	K	1-6	7-9	10-12
N.B.	K	1-5	6-8	9-12
Que.	K	1-6	7-11	
Ont.	K	1-8	9-12	
Man.	K	1-4	5-8	9-12
Sask.	K	1-5	6-9	10-12
B.C., Y.T.	K	1-7	8-12	

Imputation of census subdivision (CSD) names

Census subdivision (CSD)[7] names were derived from geographic coordinates, namely latitude and longitude. 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.[8]

Institution type provided in source datasets

The provided institution type (e.g., public, private, etc.) was used as stated in the source data set without further reinterpretation, reassignment or mapping to a uniform classification. In comparison with the use of ISCED to standardize education levels, there is no known standard for institution type. When the data source did not have a type column but the data source itself was for a particular type (e.g., a file of public schools or a file of Private schools), then the facility type was set manually.

Data standardization

Due to the different standards adopted in the original data, steps taken to standardize the data were liable to 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.”

For OLMS entries where only a P.O. box address was provided, these addresses were removed and replaced with the civic addresses, which were found through manual web searches.

Finally, a limited number of entries that were not parsed correctly were identified by manual inspection and corrected.

[7] 'Census subdivision' is the general term for municipalities as determined by provincial or territorial legislation, or areas treated as municipal equivalents for statistical purposes. For a detailed definition see: https://www12.statcan.gc.ca/census-recensement/2016/ref/dict/geo012-eng.cfm.
[8] GeoPandas is a Python package for the manipulation of geospatial data: http://geopandas.org/index.html.

Removal of duplicates

The removal of duplicates was done using the Record Linkage Toolkit[9] package in Python, where Levenshtein and Cosine distances were computed on name and address fields for facilities within the same CSD. Record pairs with string similarity metrics above 0.9 were flagged for inspection and removed if they were determined to be duplicates.

For OLMS entries, record pairs were manually inspected to determine whether the matches indicated true or false duplicates. Using web searches to compare names and addresses between the matched pairs and in some cases, ground-truthing with mapping sites, most record pairs were identified as false duplicates. In addition, several pairs were identified as belonging to the same school but covering different grade ranges – these were indicated separately. In the end, only entries that seemed to be clear duplicates (very similar names, addresses, and equal grade information) were chosen for removal, or facilities with exact matches on names and address information.

6. Data dictionary

This data dictionary below describes the variables of the ODEF.

Variable – Record ID	
Name	Index
Format	String
Source	Internally generated during data processing
Description	Unique record ID automatically generated during data processing

Variable – Source ID	
Name	Source_ID
Format	String
Source	Provided as is from original data.
Description	The record's unique ID as in the original data source, if available.

Variable – Facility Name	
Name	Facility_Name
Format	String
Source	Provided as is from original data.
Description	Institution name.

Variable – Facility Type	
Name	Facility_Type
Format	String
Source	Provided as is from original data.
Description	Institution type (e.g. public, private, governmental, etc.).

Variable – Authority Name	
Name	Authority_Name
Format	String
Source	Provided as is from original data.
Description	Authority name.

Variable – Early Childhood Education	
Name	ISCED010
Format	Boolean
Source	Provided as is from original data or imputed from grade range data.
Description	Supports early childhood education students as defined by the ISCED level in Table 1.

[9] See: https://recordlinkage.readthedocs.io/en/latest/about.html

Variable – Kindergarten	
Name	ISCED020
Format	Boolean
Source	Provided as is from original data or imputed from grade range data.
Description	Supports kindergarten students as defined by the ISCED level in Table 1.

Variable – Elementary	
Name	ISCED1
Format	Boolean
Source	Provided as is from original data or imputed from grade range data.
Description	Supports elementary school students as defined by the ISCED level in Table 1.

Variable – Junior Secondary	
Name	ISCED2
Format	Boolean
Source	Provided as is from original data or imputed from grade range data.
Description	Supports lower secondary students as defined by the ISCED level in Table 1.

Variable – Senior Secondary	
Name	ISCED3
Format	Boolean
Source	Provided as is from original data or imputed from grade range data.
Description	Supports upper secondary students as defined by the ISCED level in Table 1.

Variable – Post-Secondary	
Name	ISCED4Plus
Format	Boolean
Source	Provided as is from original data or imputed from grade range data.
Description	Supports post-secondary students as defined by the ISCED level in Table 1.

Variable – Official Language Minority School Designation	
Name	OLMS_Status
Format	Boolean
Source	Matched records to a database of public K-12 official language minority schools.
Description	An official language minority school is an Anglophone school in Québec or a Francophone school in other provinces and territories. A value of 1 indicates the record is an OLMS.

Location Variables

Variable – Full Address	
Name	Full_Addr
Format	String
Source	A combination of address components or provided as is.
Description	Full address of facility.

Variable – Unit Number	
Name	Unit
Format	String
Source	Parsed from a full address string or provided as is.
Description	Civic unit or suite number.

Variable – Street Number	
Name	Street_No
Format	String
Source	Parsed from a full address string or provided as is.
Description	Civic street number.

Variable – Street Name	
Name	Street_Name
Format	String
Source	Parsed from a full address string or provided as is.
Description	Civic street name.

Variable – City	
Name	City
Format	String
Source	Parsed from a full address string or provided as is.
Description	Municipality name.

Variable – Province/Territory	
Name	Prov_Terr
Format	String
Source	Converted to two letter codes (internationally approved) after parsing from a full address string, or provided as is, or indicated by providers.
Description	Province or territory name.

Variable – Postal Code	
Name	Postal_Code
Format	String
Source	Parsed from a full address string or provided as is.
Description	Postal Code.

Variable – Province Unique Identifier	
Name	PRUID
Format	Integer
Source	Converted from province code.
Description	Province unique identifier.

Variable – CSD Name	
Name	CSDNAME
Format	String
Source	Imputed from geographic coordinates and city names using GeoSuite 2016.
Description	Census subdivision name.

Variable – CSD Unique Identifier	
Name	CSDUID
Format	String
Source	Imputed from either geographic coordinates or CSD name using GeoSuite 2016.
Description	Census subdivision unique identifier.

Variable – Longitude	
Name	Longitude
Format	Float
Source	Provided as is from original data.
Description	Longitude.

Variable – Latitude	
Name	Latitude
Format	Float
Source	Provided as is from original data.
Description	Latitude.

Variable – Geocoding source	
Name	Geo_Source
Format	String
Source	Created based on origins of geocoordinates.
Description	An indication of whether the latitude and longitude were provided in the original source, or if they were geocoded for the ODEF.

Variable – Data Provider	
Name	Provider
Format	String
Source	Created based on origins of input dataset.
Description	Name of the entity that provided the dataset.

Variable – CMA Name	
Name	CMANAME
Format	String
Source	Imputed from 2021 census boundary files based on spatial location.
Description	Census metropolitan area name.

Variable – CMA Unique Identifier	
Name	CMAUID
Format	String
Source	Imputed from 2021 census boundary files based on spatial location.
Description	Census metropolitan area unique identifier.

7. Data accuracy

All education facility data in the ODEF were collected from government 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.”

A few exceptions apply to OLMS entries. Some entries that did not appear in the original data sources were added after comparing them to the webpages of official language minority school boards. When schools were missing information, such as address or school board, this was filled in through manual searches.

Imputation of ISCED levels is done conservatively to avoid false positives. Consequently, the percentages of ISCED levels with a non-empty value differ by level.

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.

Finally, it should be noted that facility type, which discerns public, private, and other types of institutions, has different interpretations by province and data provider. For example, religious schools may be publicly funded in one jurisdiction but not in another.

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.