Catalogue no. 37260001
Issue no. 2024001

Harnessing Open Data

The Open Database of Educational Facilities (ODEF)

Metadata document: concepts, methodology and data quality

Version 3.0.1

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

Release date: 13 December 2024
Last updated: 17 November 2025

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: infostats@statcan.gc.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

This document details the process of collecting, compiling, and standardizing the individual datasets of the Open Database of Educational Facilities (ODEF), which is made available under the Open Government Licence – 2[1].

In its current version (version 3.0), the ODEF contains approximately 19,000 records. The scope of this dataset includes primary, secondary, and post-secondary educational institutions, both public and private. The database is expected to be updated periodically as new open datasets become available.

This database is one of several 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

1.1. Changes from version 2.1

Version 3 of the ODEF is based on version 2.1 of the ODEF. The number of schools on record decreased from 18,982 to 18,858. New schools were added, some schools were retired, and some schools changed their names.

In addition to the previous version, the ODEF now indicates whether a school offers full-time French Immersion courses, and if so, at which level:

- Early immersions, starting at Kindergarten on Grade 1
- Middle immersion, starting at Grade 4
- Late immersion, starting at Grade 6

However, Provinces and Territories have different definitions of what constitutes Early, Middle, and Late immersion. For more details, please consult the Provincial and Territorial governments.

1.2. Changes from version 3.0

The column “facility_type” was dropped by mistake and has been restored. Wrong information about official minority language schools in Manitoba has been fixed.

2. Data sources

Multiple data sources were used to create the ODEF – for all school categories and all provinces, 92 files and 14 unique data providers were collected. The data collected vary in coverage across provinces and territories for each dataset.

The data providers, which include multiple levels of government, are outlined in a supplementary CSV file provided with the downloaded data and include attribution statements 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.

3. Reference period

The supplementary CSV file of data providers, described above, lists the date each underlying dataset was last

[1] See: https://open.canada.ca/en/open-government-licence-canada

updated by the provider (when known). Data were gathered between July and November 2024. 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. Compilation methodology

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

- Conversion of original data files and fields to standard formats and fields names.
- Conflation with Census subdivision (CSD) data.
- Simple and spatial deduplication. Deduplication was performed in a conservative manner to avoid false positives (for more details, see Data standardization).
- Cleaning and data validation.

While effort was made to ensure that the data is correct, it is possible that the scripts used to process the data 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 represents what is available from the original sources without imputation. The exception to this is the imputation of CSD names, as discussed below.

4.1 Conflation with census subdivision (CSD)

ODEF records were conflated with the 2021 CSD[2] spatial boundaries. Using the location coordinates of each school facility, each record within the study area was intersected with CSD polygons through a spatial join operation using the Python package GeoPandas[3] . ODEF features were then assigned CSD name and CSD unique id values. For schools missing coordinates, a string match between the name of the municipality hosting the school and the CSD name was attempted. The procedure assigned a CSD name and CSD unique ID for 89.6% of the schools on record.

4.2 Scope of inclusion

Only records within the scope of the project were included in the final ODEF dataset, version 3.0. Records which indicated online learning facilities without a fixed address were filtered from source data. Records where the facility ostensibly had a physical location, but the address or geographic coordinates were missing, were kept in the dataset.

4.3 Removal of duplicates

Various deduplication methods were employed to reduce the presence of duplicate records in the ODEF version 3.0 dataset. These methods included: simple deduplication of identical records, the removal of duplicates with exact string matching of select record variables, deduplication based on proximity thresholds between features, and the use of fuzzy record matching using the python package recordlinkage[4], which compares records within a database and returns a similarity score for each comparison field. Duplicate records that were found were then dropped after

[2] '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/2021/ref/dict/az/Definition-eng.cfm?ID=geo012

[3] GeoPandas is a Python package for the manipulation of geospatial data: http://geopandas.org/index.html

[4] See: https://pypi.org/project/recordlinkage/

validation.

4.4 Cleaning and standardization

Due to the different standards adopted in the original sources, certain variables were deemed to be in scope for standardization and others were maintained as-is from the source data. The methodology and limitations of each standardization technique are described below. Trivial cleaning techniques, such as removal of whitespace characters and punctuation removal, are not discussed in this document.

General standardization
During standardization, province and territory values were standardized to adopt the 2-character code defined by ISO-3166[5], and region names which were provinces or territories had their name fully written out. When possible, record variables were also updated to use a single value when multiple values were used to describe the same value.

Imputation of International Standard Classification of Education (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[7]. Currently, only Ontario, Alberta, British Columbia, and the Northwest Territories indicate pre-kindergarten facilities as a separate level in their data. The conversion from a school level description to a grade range is shown in Table 2.

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	..

[5] https://www.iso.org/obp/ui/#iso:code:3166:CA
[6] The International Standard Classification of Education (ISCED) is a statistical framework for organizing information on education maintained by UNESCO. ISCED 0 indicates early childhood development education https://uis.unesco.org/en/topic/international-standard-classification-education-isced
[7] https://uis.unesco.org/en/glossary-term/isced-0-early-childhood-education-includes-isced-01-and-isced-02

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
Newfoundland and Labrador, Prince Edward Island, Nova Scotia, Nunavut	K	1-6	7-9	10-12
Alberta, Northwest Territories	Pre-K/K	1-6	7-9	10-12
New Brunswick	K	1-5	6-8	9-12
Quebec	K	1-6	7-11	
Ontario	Pre-K/K	1-8	9-12	
Manitoba	K	1-4	5-8	9-12
Saskatchewan	K	1-5	6-9	10-12
British Columbia	Pre-K/K	1-7	8-12	
Yukon	K	1-7	8-12	

General cleaning

Cleaning was performed with the goal of limiting the amount of altering of the original data when possible. Fields across all data files were consistently named, ordered, spaces replaced with underscore characters and made lowercase. Redundant spatial fields including WKT (well known text)[8], latitude and longitude were dropped and replaced by a single geometry column. Records which lacked geometries or had geometries outside the Canada extent were dropped. Lastly, minor cleaning including the removal leading and trailing whitespaces from cell values was also applied to some ODEF dataset variables.

5. Data dictionary

Each infrastructure type contains a set of common variables shared across categories, as well as some unique variables. Variables common across infrastructure types include ID, subtype, source ID, data provider, census subdivision name, census unique identifier, province name, and geometry. Descriptions for each of these variables, as well as variables unique to each infrastructure type can be found in the data provider CSV file accompanying the data.

6. Data accuracy

Infrastructure data in the ODI were collected from open data sources, mainly from open data portals or otherwise public webpages of authoritative sources (mainly municipalities or provinces). In general, other than the processing required to harmonize the different sources into one database, the underlying datasets were taken “as is.”

7. Contact Us

The LODE open databases are conceived for continuous 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.

[8] The "Well-Known Text" format, abbreviated as WKT, is a standard text-based format used to represent vector geometric objects from Geographic Information Systems (GIS).