Catalogue no. 21260001
Issue no. 2020001

Exploring Open Data

The Open Database of Cultural and Art Facilities (ODCAF)

Metadata document: concepts, methodology and data quality

Version 1.0

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

Release date: October 2, 2020

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

1. Overview

This experimental Open Database of Cultural and Art Facilities (ODCAF) is one of a number of datasets being created as part of the Linkable Open Data Environment (LODE). The LODE is an exploratory initiative of the Data Exploration and Integration Lab (DEIL) at Statistics Canada. It aims at enhancing the use, accessibility 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. This initiative is also meant to explore open data for official statistics and to support geospatial research across various domains. The LODE datasets and code are available through the Statistics Canada website and can be found at: https://www.statcan.gc.ca/eng/lode

The ODCAF is a database of cultural and art facilities released as open data. Data sources include various levels of government within Canada[1] and professional organizations. This document details the process of collecting, compiling, and standardizing the individual datasets of cultural and art facilities that were used to create the ODCAF. The ODCAF is made available under the *Open Government Licence – Canada*[2].

In its current version (Version 1.0), the ODCAF contains approximately 8,000 individual records. The database is expected to be updated periodically as new open datasets become available. The ODCAF is provided as a compressed comma separated values (CSV) file.

2. Data Sources

Multiple data sources were used to create the ODCAF. The sources used are detailed in a 'Data Sources' CSV file located within the zipped data folder available for download on the ODCAF webpage[3]. The links to the original datasets, licenses or terms of use, attribution statements and additional notes are also included in the Data Sources CSV file. For further information on the individual licences, users should consult directly the information provided on the open data portals of the various data providers. In addition to openly licensed databases, the ODCAF also includes a publicly available listing of cultural and art facilities.

The distinction between open and other publicly available data is based on the licensing terms (explicit or implicit) attached to each source dataset used. Open data licenses permit, in varying degrees, usability for any lawful purpose, redistribution (re-sharing) and modification and re-packaging of the data. However, open data licenses can impose some restrictions, such as attribution of original source, share-alike (re-sharing only with like conditions), and no commercial use. Examples of open data licenses are Creative Commons, MIT, GPLv3, and Canada's Open Government License. In general, no warranty is expressed and there are very minor conditions stipulated by the provider.

Publicly available data that are not open data might be associated with proprietary licensing or terms of use that may restrict some of the aspects that would otherwise be permitted under open data licensing.

3. Reference Period

The Data Sources CSV provides, when this is known, either the update frequency or the date each underlying dataset was last updated by the provider (this information is collected at the time the dataset was accessed for this project). Additionally, the Data Sources CSV provides the date each dataset used in the ODCAF was downloaded or provided by the organization that is the source of the data. Data were gathered between January 2020 and July 2020. Users are cautioned that the download date should not be used as an indication of the reference date of the data. To obtain specific information concerning the

[1] This includes municipal, regional, and provincial governments.
[2] See: https://open.canada.ca/en/open-government-licence-canada
[3] See: https://www.statcan.gc.ca/eng/lode/databases/odcaf

reference dates of the source datasets, users might contact the relevant data providers directly.

4. Target Population

For the purposes of the ODCAF database, cultural and art facilities are facilities wherein the primary activity is of a cultural nature or is related to the arts. The target population includes only brick and mortar cultural and art facilities that offer programs or services to the general public.

In terms of the North American Industry Classification System (NAICS)[4], the facilities in the ODCAF are primarily in the following sub-sectors:

711 - Performing arts, spectator sports and related industries

712 - Heritage institutions

Facilities are included when their primary activities have a cultural or arts character, regardless of the source of funding, private or public status, operator type, location or other attributes. However, facilities that are not open to the general public and those that are primarily commercial in nature are not included. Thus, a theatre that offered ballet performances would be in scope, while a ballet school that offered training and performances only to paying students would not.

5. Compilation Methodology

This section provides an overview of the processing done to compile the ODCAF.

Data Standardization and Cleaning

The first processing component for compiling the ODCAF database comprised reformatting the source data to CSV format and mapping the original dataset attributes to standard variable (field) names. This was done using a version of the custom OpenTabulate[5] software developed by the LODE team. A data dictionary of the variables used is provided in section 8.

Owing to the different classification systems and data attributes used in the source datasets and the need to standardize through application of several processing steps, the potential exists for the introduction of errors.

The methodology and limitations of the techniques used in each step used in the data cleaning process are described below. Trivial cleaning techniques, such as removal of whitespace characters and punctuation removal, are omitted from discussion.

Address Parsing

The libpostal[6] 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

[4] North American Industry Classification System (NAICS) Canada 2017 Version 3.0 (https://www.statcan.gc.ca/eng/subjects/standard/naics/2017/v3/index)
[5] See: https://pypi.org/project/opentabulate/.
[6] See: https://github.com/openvenues/libpostal

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.” Otherwise, the left most number is appended to the unit column.

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 results are 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 to or detected by the LODE team subsequently, they will be corrected in future versions of the ODCAF.

Removal of Duplicates

The removal of duplicates was done using both literal and fuzzy string matching on the facility name and street name, conditioned on the street number and province; by “conditioned,” it is meant that a fuzzy comparison between two facilities is made provided that the street numbers and provinces agree. The fuzzy comparison is done using the Python package FuzzyWuzzy[7], which returns a similarity score between 0 and 100 for two strings, where a score of 100 indicates that the shorter string is a sub-string of the larger string. A threshold value for the returned score of the comparison is chosen empirically, indicating when an entry is marked as a duplicate.

If two entries contained identical street number and province information, then their street names and facility names were compared. When these were nearly identical (defined as having the sum of the similarity scores for the facility names and street names to be at least 195 out of a possible 200), then the entries were marked as duplicates. Recognized duplicates were deleted without manual intervention. The chosen threshold was selected close to the maximum score, which minimized any removal of false positives. When duplicates were found, whichever record contained more non-empty fields was retained. In total, 2,435 duplicates were removed.

Identification of Invalid Entries

A pair of filters was used to process the data after the address parsing stage. This captured entries with invalid postal code or province code information and wrote them to a file separate from the database for further processing. Most of these entries were manually corrected and added back into the database. The choice of these two filters is based on their capabilities in detecting potential errors in postal codes and province codes.

Other Data Cleaning Steps

- Data entry formatting (removal of excess whitespace and punctuation), removal of postal code, province/territory names.
- During processing, separation of entries with incorrect postal code or 2-letter province/territory code format from the cleaned data and their manual editing.

Selection of Record to Retain in Case of Duplicates

In some instances, a facility was present in more than one source. In such cases, the record with the most information available was retained. Where information between sources did not match, validation tools were used to decide which to retain.

[7] FuzzyWuzzy is a Python package that uses the Levenshtein distance to compute similarity measures between strings, see: https://github.com/seatgeek/fuzzywuzzy.

Classification Used and Assignment of Cultural and Art Facility Type

The original data sources use a variety of standards, classifications and nomenclature to describe the type of cultural and art facility. Unfortunately, there is no classification for cultural and art facilities in Canada that is used universally. The following classification of cultural and art facilities is used for Version 1.0 of the ODCAF:

- **Arts or cultural centre:** Establishments primarily engaged in promoting culture and arts
- **Artist:** Individual artists engaged in creating artistic works
- **Festival site:** Sites on which arts or cultural festivals are held
- **Gallery:** Establishments primarily engaged in the display of artistic works
- **Heritage or historic site:** Sites of cultural, artistic, or historic significance
- **Library or archive:** Establishments primarily engaged in the display, curation, and sharing of primarily written material such as manuscripts, periodicals, and other items such as maps or images
- **Miscellaneous:** Establishments associated in some way with promoting or providing culture or arts that do not fall into any of the above categories
- **Museum:** Establishments primarily engaged in the display, curation, and sharing of collections of artifacts, fine arts, and other objects of artistic, cultural, or historical importance
- **Theatre/performance and concert hall:** Establishments primarily engaged in the public performance of artistic or cultural works

The classification is intended to have broad categories that are helpful in distinguishing major types of facilities and yet enable accuracy in mapping source-specific facility types. Facility types are determined from source-specific facility types and source coverage metadata information. Assignments are made using keywords and validated afterwards, with changes made manually whenever needed. When classifying facilities based on source metadata information, this was done analytically on a case by case basis.

Geocoding and Determination of Census Subdivision

In general, the data included in the ODCAF are what is available from the original sources without imputation. The exception to this is the geocoding and the imputation of CSD names and categories, discussed below.

Census subdivision (CSD)[8] names were derived from two different attributes in the data.

The first attribute comprises the 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

[8] '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.

operation using the Python package GeoPandas.[9]

The second attribute is the city name, where literal string matching was done with each cultural and art facility municipality name and a list of CSD names. The city names with at least ten entries that did not receive a CSD name through this process were manually assigned a CSD name by using Place Names in GeoSuite.[10]

Geocoding was carried out for some sources that provide address data but no geo-coordinates. Latitude and longitude were determined and validated using tools on the internet. A subset of the source-provided geo-coordinates were also validated using the internet. Some coordinates have also been removed from the original sources when it was determined they were derived from postal codes or other aggregate geographic areas as opposed to street address.

While efforts have been made to ensure the accuracy of geo-coordinates, no guarantees are implied, and errors and inaccuracies are possible.

Inclusion in the ODCAF of Facility Type Provided in Source Datasets

The facility types as provided in the data sources (e.g., exhibition or cultural centre, community library, centre d'art, etc.) are also included in the ODCAF without any modification, reassignment, or mapping to a uniform classification.

6. Database Coverage

The ODCAF current version (Version 1.0) database as provided contains approximately 8,000 cultural and art facilities.

As the total number of all cultural and art facilities in the country is not known with a reasonable degree of certainty, the coverage obtained with the sources used was not quantitatively assessed. However, many of the sources purport to list all facilities of a certain type within a jurisdiction. Thus, within these facility type categories and jurisdictions, coverage would be expected to be fairly complete. However, if facilities of a certain category were omitted in a source, then these might be missing from the database, unless they were obtained from a different source.

7. Data Quality

All cultural and art facility data in the ODCAF were collected from government data sources, either from open data portals or publicly-available webpages. In general, other than the processing required to harmonize the different sources into one database, the underlying datasets were taken "as is." The accuracy and completeness of the information is in general a function of the source datasets used.

Classifying facilities

> Assignment of facility type was largely based on facility types provided by source datasets. In instances where facility type was either unclear or not defined by the source, facility type was classified based on further research or using meta-information, such as name of dataset.

Removing duplicates

[9] GeoPandas is a Python package for the manipulation of geospatial data: http://geopandas.org/index.html.
[10] See: https://geosuite.statcan.gc.ca/geosuite/en/index.

Some source datasets do overlap; datasets which cover only a particular type of arts or cultural facility for an entire province, for example, may overlap with data provided only for specific towns. Although deduplication techniques are used, not all duplicates might have been removed. Modifying the deduplication methods to seek out the remaining duplicates would generate numerous false positives, which would require additional manual intervention. Further details are available in the sub-section Removal of Duplicates above.

Correcting invalid entries

A few entries with erroneous province/territory names and postal codes were detected and manually corrected. Further details on the identification of erroneous entries are also reported in the sub-section Identification of Invalid Entries above.

Address parsing

Natural language processing methods were used for parsing and separation of address strings into address variables, such as street number and postal code (which is removed from the final released database). The methods are reputable in the field 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. Data Dictionary

This data dictionary below describes the variables of the ODCAF.

Variable – Index	
Name	Index
Format	String
Source	Internally generated during data processing
Description	Unique number automatically generated during data processing

Variable – Facility Name	
Name	Facility_Name
Format	String
Source	Provided as is from original data
Description	Cultural or arts facility name

Variable – Source Facility Type	
Name	Source_Facility_Type
Format	String
Source	Provided as is from original data
Description	Facility type chosen by data provider

Variable – ODCAF Facility Type	
Name	ODCAF_Facility_Type
Format	String
Source	Imputed from source data or metadata
Description	Facility type assigned from nine ODCAF categories

Location Variables

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	City or municipality name (certain records may list the neighbourhood 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 – Province Unique Identifier	
Name	PRUID
Format	Integer
Source	Converted from province code
Description	Province unique identifier

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

Variable – CSD Unique Identifier	
Name	CSDUID
Format	Integer
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 – Data Provider	
Name	Provider
Format	String
Source	Created based on origins of input dataset
Description	Name of the entity that provided the dataset

9. 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@canada.ca. Please include the title of the open database in the subject line of the email.