Catalogue no. 21260002
Issue no. 2020001

Linkable Open Data Environment

The Open Database of Recreational and Sport Facilities (ODRSF)
Metadata document: concepts, methodology and data quality

Version 1.0

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

Release date: September 28, 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

1. Overview

The Open Database of Recreational and Sport Facilities (ODRSF) is a database of recreational and sport 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 used to create the ODRSF.

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

The ODRSF is made available under the Open Government Licence – Canada[2]. In its current version (Version 1.0), the ODRSF contains approximately 182,000 individual records. The database is expected to be updated periodically as new open datasets become available. The ODRSF is provided as a compressed comma separated values (CSV) file.

2. Data Sources

A total of 452 data sources were used to create the ODRSF. The sources used are detailed in a 'Data Sources' CSV file provided together with the data file on the ODRSF 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. All recreational and sport facility data in the ODRSF were collected from government data sources, either from open data portals or from publicly-available web pages.

The distinction between open and other publicly available data is based on the licensing terms 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. In general, no warranty is expressed and there are minor conditions stipulated by the provider.

Publicly available data, that are not associated with an open data license, are generally provided with terms of use that may restrict some of the aspects that would otherwise be permitted under open data licensing.

For further information on the individual licences or terms of use, users should consult the information provided on the open data portals or web sites of the various data providers, as reported in the Data Sources CSV file.

3. Reference Period

The Data Sources CSV reports, 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 that each dataset used in the ODRSF was downloaded or provided by the organization that is the source of the data. Data were gathered in 2020 and 2021. Users are cautioned that the download date should not be used as an indication of the reference date of the data.

4. Target Population

For the purposes of the ODRSF database, recreational and sport facilities are facilities wherein the primary activity

[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/odrsf

is concerned with either recreation or sport. The target population includes brick and mortar recreational and sport facilities that offer programs or services to the general public as well as those such as trails for hiking or skiing, sports fields, and other types of facilities that may be located outside of brick and mortar structures.

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

7112 – Spectator sports

7131 – Amusement parks and arcades

7139 – Other amusement and recreation industries

Facilities are included when their primary activities are related to recreation or sports, 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 are not included. It should be noted that the focus of the ODRSF is on the facility (point of service). This may or may not correspond to a business entity, as some facilities such as trails or beaches may not be associated with any business entity while others, for example a multi-sport complex, may be related to a number of discrete entities.

5. Compilation Methodology

Data Standardization and Cleaning

The first processing component for compiling the ODRSF database was comprised of 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[4] software developed by the LODE team. A data dictionary of the variables used is provided in section 7. The methodology and limitations of the techniques used in each step used in the data cleaning process are described below.

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.

When address information was available, addresses were parsed using the same methodology applied to other LODE databases such as the Open Database of Education Facilities[5] and the Open Database of Cultural and Arts Facilities[6]. The libpostal[7] 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[8].

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

[5] See Open Database of Education Facilities: https://www150.statcan.gc.ca/n1/en/catalogue/37260001

[6] See Open Database of Cultural and Art Facilities: https://www150.statcan.gc.ca/n1/en/catalogue/21260001

[7] See: https://github.com/openvenues/libpostal

[8] 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

For instance, 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.

Removal of Duplicates

As data were sourced from entities with geographically overlapping jurisdictions (e.g., a province, municipality and a private sector organization), the same record can appear in more than one source dataset. 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 Levenshtein distances calculated through the Python package FuzzyWuzzy[9], 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. An entry is marked as a duplicate when that score meets a given threshold of similarity.

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, 5,937 duplicates were removed.

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.

Identification of Invalid Entries and Other Data Cleaning Steps

Identifying erroneous entries was done both programmatically and manually. Data entries that could not be correctly processed by automated techniques were filtered and stored in a separate file and manually corrected later. Data entries were formatted through the removal of excess whitespace and punctuation, standardization of fields such as postal code, and province/territory names.

Classification and Assignment of Recreational and Sport Facility Type

The original data sources use a variety of standards, classifications and nomenclature to describe the various types of recreational and sport facility. With no classification for recreational and sport facilities that is broadly adopted and recognized in Canada, one of the main challenges in the implementation of the ODRSF was the harmonization of records into comparable groups. 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.

The following classification of recreational and sport facilities is used for Version 1.0 of the ODRSF. While most of the class names are self-explanatory, further clarifications are provided below. In addition, and where available, the facility types as provided in the data sources (e.g., outdoor pool, tennis court, sports field, etc.) are also included in the ODRSF without any modification, reassignment, or mapping to a uniform classification.

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.

[9] FuzzyWuzzy is a Python package that can be configured to use the Levenshtein distance to compute similarity measures between strings, see: https://github.com/seatgeek/fuzzywuzzy.

- trails: urban and rural trails or pathways for walking, hiking, or biking.
- sports fields: fields on which sports can be played.
- arenas: facilities where sports and/or recreational activities take place.
- athletic parks: recreation areas focused on athletic activity.
- beaches: waterfront beach areas.
- casinos: casino or gambling facilities.
- community centres: community centres and leisure facilities.
- gyms: both public and private gym facilities.
- marinas: marina facilities.
- parks: parks and greenspaces, including both city and national parks.
- playgrounds: play spaces which are distinct from parks in that they have specifically been classified as such by the publisher of the data. Often includes playground equipment.
- pools: indoor and outdoor swimming pools.
- race tracks: tracks for racing.
- rinks: most commonly ice rinks.
- skate parks: parks used for skateboarding.
- splash pads: urban areas for water play.
- stadiums: facilities where sports and/or recreational activities take place.
- miscellaneous: facilities that do not fall into any of the above categories.

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.

The sports field classification category combines multiple types of sports fields such as baseball fields, soccer fields, and others. Where available, the detailed information on the type of sports field is preserved in the Source Facility Type variable.

Geocoding and Determination of Census Subdivision

In general, the data included in the ODRSF 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)[10] 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 operation using the Python package GeoPandas.[11]

The second attribute is the city name, where literal string matching was done with each recreational and sport facility municipality name and a list of CSD names.[12]

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.

[10] '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

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

[12] See: https://geosuite.statcan.gc.ca/geosuite/en/index

6. Database Coverage

The current version of the ODRSF (Version 1.0) database as provided contains approximately 182,000 recreational and sport facilities.

As the total number of all recreational and sport facilities in the country is not known with a reasonable degree of certainty, the coverage obtained with the sources used was not able to be thoroughly quantitatively assessed. Looking at the individual category of golf courses, however, shows that the ODRSF contains 592 golf courses, approximately 25% of all the 2,182 golf courses estimated to be in Canada[13]. Likewise, there were 1,303 rinks and arenas located in the ODRSF, approximately 60% of the 2,183 rinks[14] and arenas estimated to be in Canada. The distribution of the latter category showed similar coverage trends across geographies, with 82% to 87% of arenas and rinks respectively being located in Ontario and the Prairie provinces compared to an estimated two-thirds for these types of facilities overall.

From the above results, it is clear that the ODSRF is not a comprehensive listing of facilities within Canada. This is to be expected since not all jurisdictions publish data on recreational and sport facilities or categorize them in the same ways. The exception to this is when sources do purport to list all facilities of a certain type within a jurisdiction so for those particular facility type categories and jurisdictions; for those sources, coverage would be expected to be fairly complete. However, if facilities of a certain category were omitted by a source, then those facilities might be missing from the database unless they were obtained from a different source.

7. Data Dictionary

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	Recreational or sport 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 – ODRSF Facility Type	
Name	ODRSF_Facility_Type
Format	String
Source	Generated from source data or metadata
Description	Facility type assigned from ODRSF categories

[13] Canadian Industry Statistics, 2020. https://www.ic.gc.ca/app/scr/app/cis/summary-sommaire/71391
[14] Survey of Energy Consumption of Arenas, 2014. https://www150.statcan.gc.ca/n1/daily-quotidien/160830/dq160830d-eng.htm

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 – Street Type	
Name	Street_Type
Format	String
Source	Parsed from a full address string or provided as is
Description	Civic street type

Variable – Street Direction	
Name	Street_Direction
Format	String
Source	Parsed from a full address string or provided as is
Description	Civic street direction

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

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 or added by geolocation
Description	Longitude

Variable – Latitude	
Name	Latitude
Format	Float
Source	Provided as is from original data or added by geolocation
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

8. 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 ODRSF, 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.

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.