Open Database of Health Facilities (ODHF)

Metadata document: concepts, methodology and data quality

Version 1.0

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

April 17, 2020

Statistics Canada Statistique Canada

Canada

Table of Contents

1. Overview

The Open Database of Health Facilities (ODHF) is a Canada-wide health facilities database. It has been compiled by the Centre for Special Business Projects (CSBP) at Statistics Canada. This document discusses the methodology used to create the ODHF. This document pertains to the first release of the ODHF (Version 1.0) in April 2020.

The database uses both open data as well as publicly available data (a dataset being designated open depending on whether or not the data are distributed under an open data license). Most of the data are sourced from municipal, regional and provincial/territorial governments, federal agencies, or independent not-for-profit organizations specializing in the health information field. The data have been either web-scraped, downloaded, or obtained directly from the data sources.

The main objective of producing the ODHF is the dissemination of this information through the harmonization and integration of, and, to a limited extent, the addition of geolocation information to the data assembled from the various sources used.

Version 1.0 of the ODHF contains approximately 9,000 individual records. The ODHF is provided as a compressed comma separated values (CSV) file. The database is expected to be updated periodically as new datasets become available or as other improvements are made.

The ODHF is one of a number of datasets created as part of the Linkable Open Data Environment (LODE), an initiative at CSBP. The LODE is an exploratory initiative that aims at enhancing the use and harmonization of open and publicly available data from authoritative sources by providing a collection of datasets released under a single licence. LODE also provides 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 the link above.

2. Target Population

A health facility is a physical site at which the primary activity is the provision of healthcare. Health facilities in Canada that provide healthcare services are in scope for this dataset. Specifically, in terms of the North American Industry Classification System (NAICS)[1], the following industries are in scope:

621 - Ambulatory health care services

622 - Hospitals

623 - Nursing and residential care facilities

Facilities are included when their primary activities relate to healthcare, regardless of the source of funding, private or public status, operator type, location or other attributes not listed here. However, alternative medicine (e.g., herbalists) and specialist areas (e.g., chiropractors, dentists, mental health specialists, etc.) are not included in the current ODHF version (Version 1.0). However, when the sources used contained these out-of-scope facilities, some of these might still be present in the ODHF database.

Facilities that are in areas indirectly related to overall healthcare delivery, e.g., pharmacies, social assistance, etc.,

[1] North American Industry Classification System (NAICS) Canada 2017 Version 3.0 (https://www.statcan.gc.ca/eng/subjects/standard/naics/2017/v3/index)

are also not in scope of the current version of the ODHF.

3. Data Sources

The sources used are detailed in Appendix A for open data sources and in Appendix B for publicly available data sources. The links to the original datasets, licenses or terms of use, attribution statements and additional notes are also included in Appendix A and Appendix B.

Nearly all data sources used to create this database are publicly available sources, such as municipal governments, provincial/territorial governments and health authorities and agencies, and independent not-for-profit organizations specializing in the health information field. The data were obtained either from open data portals located on websites or web-scraped, or were provided directly by the source. In most cases, sources were discovered using major search engines or through professional contacts. Sources were sought in all Canadian provinces and territories.

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 data 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 generally restrict several of the aspects that would otherwise be permitted under open data licensing. The sources are detailed in Appendix A for open data sources and in Appendix B for other publicly available data sources.

The links to original datasets used for the current version of the ODHF (Version 1.0), licenses or terms of use, attribution statements and additional notes are also included in appendices A and B. For further information on the individual licenses, users should consult directly with the information provided on the data portals for the data providers.

4. Reference Period and Last Update Dates

In principle, the reference date of the database would be the date on which all health facilities in existence on that date would be included in the dataset. Ideally, this would be the same date for all datasets used. However, this is not the case and the reference dates on which a particular source was current or was updated would vary by provider. In some cases, such detail was not present in the information made available by data providers.

Appendix A and Appendix B provide the date when each source dataset was last updated by the provider (this information is collected at the time the dataset was accessed for this project). As all data sources only had one version available, this is what has been used and taken to be the most current available.

Users are cautioned that the last update date should not be interpreted as the reference date of the data. If specific information concerning the reference period of data is required, users should contact the appropriate data providers shown in Appendix A: Open Data Sources and Appendix B: Other Publicly Available Data Sources.

5. Compilation Methodology

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

Data Cleaning

The primary processing component for the database comprised reformatting the source data to CSV format and mapping the original dataset attributes to the variable (column) names defined for this project. A data dictionary of the variables used for this project is provided in section 8 Data Dictionary. To clean the data, the following was done:

- Address parsing and normalization
    - Concatenated address data were parsed and separated into the respective location variables using libpostal,[2] a state-of-the-art natural language processing solution for address parsing. A small number of addresses were parsed incorrectly and were manually corrected.
    - Data entry formatting (removal of excess whitespace and punctuation), normalization of postal codes and addresses, province/territory names.
    - Some data entries that were filtered out by automated cleaning methods were manually corrected. See section 8 for more details.

- Removal of Duplicates

    - The removal of duplicates is done using fuzzy string matching based on a criteria involving the facility name, street name, street number and geo-coordinates. The criteria was derived empirically and with the intent of avoiding false positives.

- Identification of Erroneous Entries

    - 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.

- 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.

For more details on the software used to process the data, please refer to CSBP's GitHub page.[3]

Determination of Health Facility Types

The original data sources use a variety of standards, classifications and nomenclatures to describe the type of health facility. Unfortunately, there is no classification for health facilities in Canada that is used universally. Health authorities classify their facilities independently using different classifications systems. The following classification

[2] See https://github.com/openvenues/libpostal.
[3] See https://github.com/CSBP-CPSE/OpenTabulate.

of health facilities is used currently for the database:

- **Ambulatory health care services**: Establishments primarily engaged in providing health care services, directly or indirectly, to ambulatory patients. *Example: medical clinic, mental health center.*

- **Hospitals**: Establishments, licensed as hospitals, primarily engaged in providing diagnostic and medical treatment services, and specialized accommodation services to in-patients. *Example: emergency department, general hospital.*

- **Nursing and residential care facilities**: Establishments primarily engaged in providing residential care combined with either nursing, supervisory or other types of care as required by the residents *Example: nursing home.*

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 (e.g., cancer treatment centers are classified as 'Hospitals') and source coverage metadata information. Assignments are done 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.

Table 1 illustrates the use of keywords to assign type categories to the health facilities based on the classification used for the ODHF.

Table 1 Health facility type assignment criteria examples (based on keywords)

Variable	*Condition*	*Value*	*Classification*
Facility type	contains the keywords	'clinic', 'home care', 'primary care'	Ambulatory health care services
Facility type	contains the keywords	'hospital', 'hôpital', 'general', 'centre hospitalier'	Hospitals
Facility type	contains the keywords	'longterm care', 'nursing home', 'rehabilitation centre', 'soins prolongés', 'réadaptation'	Nursing and residential care facilities

Geocoding and Determination of Census Subdivision (CSD or Municipality)

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.

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

Census subdivision (CSD)[4] (or municipality) was derived from the geographic coordinates by linking to the CSD

[4] 'Census subdivision' is the general term for municipalities as determined by provincial or territorial legislation, or areas treated as municipal

polygons through a spatial join operation using the Python package GeoPandas.[5] or by using the city name available in the record's address field using GeoSuite.[6]

6. Database Coverage

The ODHF current version (Version 1.0) database as provided contains approximately 9,000 health facilities.

As the total number of all health 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 institutions of a certain type (e.g. acute care hospital, residential care) within a jurisdiction. Thus, within these institution type categories and jurisdictions, coverage would be expected to be fairly complete. However, if facilities of a certain category were omitted in a source, e.g., outpatient medical clinics, then these might be missing from the database, unless they were obtained from a different source.

7. Data Quality

The accuracy and completeness of the information is in general a function of the source datasets used. Except as noted, the underlying datasets are taken "as-is".

Classifying facilities. Assignment of facility type was done conservatively to guard against misclassification. In some cases, the source did not specify a facility type and this could not be determined using other methods either. Consequently, some records have an empty value for facility type.

Duplicates. Some datasets provide data where the rows do not represent unique facilities. Although deduplication techniques are used, it is expected that there are some duplicates remaining.

Address parsing. Natural language processing methods were used to do the parsing and separation of address strings into address variables, such as postal code and street number. The methods are reputable for state-of-the-art performance and accuracy, but as with all statistical learning methods, they have limitations as well. Poor or unconventional formatting of addresses might result in incorrect parsing. Upon manual review of the database, no incorrect parses were identified. At this stage, address records in the database are expected to be correctly parsed.

Geo-coordinates. Some facilities that did not have geo-coordinates were geocoded using OpenStreetMap's Nominatim API. The accuracy of the geocoding was manually validated by using proprietary mapping services available on the internet. In some cases, facility coordinates were also manually determined from online map services.

---

equivalents for statistical purposes. For a detailed definition see: https://www12.statcan.gc.ca/census-recensement/2016/ref/dict/geo012-eng.cfm.

[5] GeoPandas is a geospatial package for Python available at: https://geopandas.org/

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

8. Data Dictionary

This data dictionary describes the variables contained within the ODHF. The database is provided in a CSV format. Each facility is listed per row and its attributes provided in columns. The corresponding column variables are described in the data dictionary below.

Health Facility Variables

Variable - Index	
Name	index
Format	Integer
Source	Assigned serially.
Description	Unique serial number for each facility.

Variable - Facility Name	
Name	facility_name
Format	String
Source	Provided as is from original data.
Description	Health facility name.

Variable – Source Facility Type	
Name	source_facility_type
Format	String
Source	Provided as is from original data.
Description	Regional health authority assigned health facility type.

Variable – ODHF Facility Type	
Name	odhf_facility_type
Format	String
Source	Imputed from source data or metadata.
Description	Value determined using the classification criteria used (see section 5)

Variable – Provider	
Name	provider
Format	String
Source	Assigned based on the provider's identity.
Description	The identity or name of the data provider.

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 (type and direction).

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

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

Variable – Province/Territory	
Name	province
Format	String
Source	Converted to two letter codes after parsing from a full address string, or provided as is, or indicated by the provider.
Description	Province or territory name.

Variable – Source-Format Street Address	
Name	source_format_str_address
Format	String
Source	Street address from the data source provided as is.
Description	Street address in the source data.

Variable – CSD Name	
Name	CSDname
Format	String
Source	Imputed from geographic coordinates and city names.
Description	Census subdivision name.

Variable – CSD Unique Identifier	
Name	CSDuid
Format	Integer
Source	Imputed from CSD name using GeoSuite 2016.
Description	Census subdivision unique identifier.

Variable – Province or Territory Unique Identifier	
Name	PRuid
Format	Integer
Source	Imputed from CSD unique identifier by taking the first two digits.
Description	Province or territory unique identifier.

Variable – Latitude	
Name	latitude
Format	Float
Source	Provided as is from original data or corrected value if source value found inaccurate during validation
Description	Latitude

Variable – Longitude	
Name	longitude
Format	Float
Source	Provided as is from original data or corrected value if source value found inaccurate during validation
Description	Longitude

Appendix A: Open Data Sources

Data Provider	Province/ Territory	Link	License/ Terms of Use Link	Last Updated by Provider	Description
British-Columbia (Province)	B.C./C.-B.	https://catalogue.data.gov.bc.ca/dataset/emergency-rooms-in-bc	https://www2.gov.bc.ca/gov/content/data/open-data/open-government-licence-bc	12/24/2019	Emergency services in British-Columbia
British-Columbia (Province)	B.C./C.-B.	https://catalogue.data.gov.bc.ca/dataset/hospitals-in-bc	https://www2.gov.bc.ca/gov/content/data/open-data/open-government-licence-bc	12/25/2019	Hospitals in British-Columbia
British-Columbia (Province)	B.C./C.-B.	https://catalogue.data.gov.bc.ca/dataset/residential-care-facilities	https://www2.gov.bc.ca/gov/content/data/open-data/open-government-licence-bc	12/26/2019	Res. Care in British-Columbia
British-Columbia (Province)	B.C./C.-B.	https://catalogue.data.gov.bc.ca/dataset/walk-in-clinic-in-bc	https://www2.gov.bc.ca/gov/content/data/open-data/open-government-licence-bc	12/27/2019	Walk-ins in British-Columbia
Nova-Scotia (Province)	N.S./N.-É.	https://data.novascotia.ca/Health-and-Wellness/Hospitals/2kxr-ajui	https://novascotia.ca/opendata/licence.asp	2/15/2019	Hospitals in Nova-Scotia
Prince Edward Island (Province)	P.E.I./Î.-P.-É.	https://data.princeedwardisland.ca/Health-and-Home/OD0050-Health-PEI-Facility-Locations/dfge-zd27	https://www.princeedwardisland.ca/en/information/finance/open-government-licence-prince-edward-island	8/8/2019	Health Care Facilities in Prince Edward Island
Québec City, Québec (Municipality)	Que./Qc	https://www.donneesquebec.ca/recherche/fr/dataset/daa10606-5fdd-4c9b-b5ef-235081690b6e	https://creativecommons.org/licenses/by/4.0/deed.fr	2/24/2020	Hospitals in Québec City, Québec
Gatineau, Québec (Municipality)	Que./Qc	https://www.donneesquebec.ca/recherche/fr/dataset/vgat_1267315911	https://creativecommons.org/licenses/by/4.0/deed.fr	2/25/2019	Hospitals in Gatineau, Québec
Nova Scotia (Province)	N.S./N.-É.	https://data.novascotia.ca/Health-and-Wellness/Long-Term-Care-and-Residential-Care-Facilities/x76a-axw2	https://novascotia.ca/opendata/licence.asp	2/15/2019	Res. Care in Nova Scotia
Ontario (Province)	Ont./Ont.	https://geohub.lio.gov.on.ca/datasets/ministry-of-health-service-provider-locations *(via: https://data.ontario.ca/dataset/hospital-locations)*	https://www.ontario.ca/page/open-government-licence-ontario	10/15/2019	Health Care Facilities in Ontario
Horizon Regional Health Authority (New Brunswick)	N.B./N.-B.	https://gnb.socrata.com/Health-and-Wellness/Hospitals-in-New-Brunswick-Operated-by-Horizon-Hea/9bqr-479n	http://www.snb.ca/e/2000/data-E.html	3/18/2020	Hospitals in New Brunswick operated by Horizon
Vitalité Regional Health Authority (New Brunswick)	N.B./N.-B.	https://gnb.socrata.com/Health-and-Wellness/Hospitals-in-New-Brunswick-Operated-by-Vitalit-Hea/vc6s-v9py	http://www.snb.ca/e/2000/data-E.html	3/18/2020	Hospitals in New Brunswick operated by Vitalité

Alberta (Province)	Alta./Alb.	https://open.alberta.ca/publications/hospital-services-in-alberta	https://open.alberta.ca/licence	7/1/2018	Hospitals and Health Care Facilities in Alberta
Manitoba (Province)	Man./Man.	https://services.arcgis.com/mMUesHYPkXjaFGfS/ArcGIS/rest/services/Rural_Health_Care_Facilities_in_Manitoba/FeatureServer/0	(Waived)	6/30/2017	Health Care Facilities in Manitoba

Appendix B: Other Publicly Available Data Sources or Sources of Directly-Provided Data

Data Provider	Province/ Territory	Link	License/ Terms of Use Link	Last Updated by Provider	Description
Canadian Institute for Health Information	Canada	Provided directly via email	(Waived)	not available	Health Care Facilities in Canada
Manitoba (Province)	Man./Man.	https://www.gov.mb.ca/health/waittime/map.html	https://www.gov.mb.ca/legal/copyright.html (Waived)	not available	Hospitals in Manitoba
Manitoba - Winnipeg Regional Health Authority	Man./Man	https://wrha.mb.ca/locations-services/	https://wrha.mb.ca/terms-of-use/	not available	Locations of facilities managed by the Winnipeg Regional Health Authority
Manitoba - Interlake-Eastern Regional Health Authority	Man./Man	https://www.ierha.ca/default.aspx?cid=6147&lang=1	N/A	not available	Locations of facilities managed by the Interlake-Eastern Regional Health Authority
Manitoba - Northern Health Region	Man./Man	https://northernhealthregion.com/	N/A	not available	Locations of facilities managed by the Northern Health Region
Manitoba - Prairie Mountain Health	Man./Man	https://www.prairiemountainhealth.ca/our-locations	https://www.prairiemountainhealth.ca/disclaimer	not available	Locations of facilities managed by the Prairie Mountain Health Authority
Manitoba - Southern Health Region	Man./Man	https://www.southernhealth.ca/en/finding-care/	https://www.southernhealth.ca/en/disclaimer/#terms	not available	Locations of facilities managed by the Southern Health Authority
Nunavut (Territory)	Nvt./Nt	https://www.gov.nu.ca/health/information/qikiqtani-general-hospital	N/A	not available	Single Hospital in Nunavut
Public Health Agency of Canada	Canada	Provided directly via email	(Waived)	not available	Hospitals in Canada
Newfoundland and Labrador (Province)	N.L./T.-N.-L.	https://www.health.gov.nl.ca/health/findhealthservices/in_your_community.html	https://www.gov.nl.ca/disclaimer/	not available	Health Care Facilities in Newfoundland and Labrador
North-West Territories (Territory)	N.W.T./T.N.-O.	https://www.hss.gov.nt.ca/en/hospitals-and-health-centres	https://www.gov.nt.ca/terms (Waived)	not available	Health Care Facilities in North-West Territories

Manitoba (Province)	Man./Man.	https://www.ierha.ca/	N/A	not available	Health Care Facilities in Manitoba
Yukon (Territory)	Y.T/Yn.	Provided directly to CSBP via email	(Waived)	not available	Health Care Facilities in Yukon Territories
Saskatchewan (Province)	Sask./Sask.	https://www.saskhealthauthority.ca/Services-Locations/Pages/Home.aspx	N/A	not available	Health Care Facilities in Saskatchewan
Québec (Province)	Que./Qc	http://sante.gouv.qc.ca/en/repertoire-ressources/recherche/	N/A	not available	Health Care Facilities in Québec