Exploring Open Data and Open Projects:
The ODB - an experimental Open Database of Buildings

Alessandro Alasia, Jean Le Moullec, and Marina Smailes[1]

Abstract

This exploratory work aims at enhancing the use and harmonization of open building data from government sources for the purpose of contributing to the creation of a complete, comprehensive and open database of buildings in Canada. An outcome of this work is a first version of the Open Database of Buildings (ODB), a centralized and harmonized repository of building data made available under the Open Government License - Canada. In its current version (version 1.0), the ODB contains approximately 4.3 million building footprints with 10 variables. A key feature of this work is that it is being implemented via an *Open Project Approach*, which relates to the idea of favouring open options (e.g., open data and open source) when implementing projects, for the purpose of facilitating collaboration as well as increasing transparency and efficiency.

Key Words: Open Data; Municipal Microdata; Building Footprints; Open Projects.

1. Introduction

1.1 Exploring open data and open projects

Over the last two years, the Data Exploration and Integration Lab (DEIL) at Statistics Canada's centre for Special Business Projects has undertaken several projects on the topics of open data and open source. This work began in 2016 with the launch of a pilot project on assessing crowdsourcing as a means to collect data for official statistics. The data crowdsourced during this pilot related to building footprints and building attributes (e.g., addresses). During this pilot, as well as in subsequent projects, the numerous benefits and opportunities offered by open microdata and open source software solutions were brought to the forefront. It was also found that using these free and accessible options together, as opposed to their closed counterparts,[2] greatly facilitated data sharing and collaboration with not only other government departments but also the general public.

The lessons learned during and after the crowdsourcing pilot project laid the foundation that led to the development of the *Open Project* concept. The concept relates to the idea that favouring open options, where possible, when implementing projects can lead to many direct and indirect benefits, including increasing the degree of transparency and efficiency with which projects can be executed.

The work presented in this paper focuses on the development of the open database of buildings, which represents the first foray where the DEIL team implemented work in the spirit of the *Open Project* concept. The remaining sections that follow describe for potential users what the open database of buildings is and how it was compiled. The content presented, although not completely identical, can also be found in the metadata document that accompanies the ODB when it is downloaded.

1.2 The increasing importance of open data for National Statistical Offices

---

[1] Alessandro Alasia, Statistics Canada, 100 Tunney's Pasture Driveway, Ottawa, Ontario, Canada, K1A 0T6; Jean Le Moullec, Statistics Canada, 100 Tunney's Pasture Driveway, Ottawa, Ontario, Canada, K1A 0T6; Marina Smailes Statistics Canada, 100 Tunney's Pasture Driveway, Ottawa, Ontario, Canada, K1A 0T6
[2] For example, proprietary software or data that come with restrictions on usage.

Digital transformation is influencing every aspect of our social and economic life, bringing data and information flows to the core of virtually every business activity. This transformation is changing National Statistical Offices (NSOs). As public agencies with core mandates in data collection, production and management, several NSOs in recent years have embarked on a process of modernization, driven largely by the opportunities and the need to respond to the new realities of a data-rich environment.

Local actors, such as sub-national public agencies or private organisations, are producing increasing volumes of digital information, and this is an element of specific relevance for territorial statistics. Private sector companies and national governments are not the only ones going digital. Local governments are just as surely transitioning to a digital economy as well, with the result that an ever-growing volume of data is generated by municipal and provincial governments. Administrative and sensor data produced by cities are increasingly feeding national statistical systems and have growing potential for the production of territorial statistics. These entities have become active producers of local statistics through growing digitization of administrative records that they compile, manage and disseminate. The flourishing of local and municipal open data portals is only one indication of this trend.

According to the Canada Open Government website, open data is defined as structured data that is machine-readable, freely shared, used and built on without restrictions.[3] With these parameters set in place, open data becomes a source of information that can be creatively used, enhanced, and distributed freely across public and private entities. Federal, provincial and municipal open data initiatives are playing a part in encouraging data distribution and standardization within this broader ecosystem.

Open data and open source tools can be used to accelerate knowledge sharing. The use of open data from non-official sources is becoming increasingly common and is facilitated by the diffusion of these data on online platforms. Although the use of "open" will remain bounded by privacy constraints, there are a wide range of policy-relevant issues that can be addressed in an open space: measures of public infrastructure and facilities, recreational spaces, walkability, and similar indicators of regional and urban quality of life. The growing availability of georeferenced open micro-data offers NSOs unique opportunities for sub-national statistics.

1.3 Overview of the Open Database of Buildings

For the purpose of exploring open data for official statistics and to support geospatial research across various domains, work was undertaken to create an accessible and harmonized database of building footprints released as open data by various levels of government within Canada.[4] This document details the process of collecting, compiling, and standardizing the individual datasets of building footprints that were used to create the experimental Open Database of Buildings (ODB), which is made available under the Open Government License – Canada.[5]

In its current version (version 1.0), the ODB contains approximately 4.3 million building footprints. The database is expected to be updated periodically as new open datasets on buildings from government sources become available. This vector dataset is provided as a zipped shapefile at the provincial or territorial level.

Access to the dataset is available through the Statistics Canada website and can be found at:
https://www.statcan.gc.ca/eng/open-building-data/index

2. Data sources

Multiple data sources were used to create the ODB. The data providers, which include municipal, regional, and provincial levels of government, are shown in Table 2-1. Attribution to each of these data sources is listed, as per the license requirements, in Appendix A of the complete metadata document.[6] Where applicable, license versions are also

---

[3] See: Government of Canada, *Open Data.*
[4] This includes municipal, regional, and provincial.
[5] See: Government of Canada, *Open Government Licence – Canada.*
[6] See: Statistics Canada, *The ODB - an experimental Open Database of Buildings.*

shown in this appendix. For further information on the individual licenses, readers should consult directly with the information provided on the open data portals for the various data providers.

Table 2-1
Listing of data providers included in the current version of the Open Database of Buildings (ODB)

Alberta	British Columbia		New Brunswick	Northwest Territories
1. Banff 2. Canmore 3. Chestermere 4. Cochrane 5. Edmonton 6. Grande Prairie 7. Lethbridge 8. Strathcona County	1. Chilliwack 2. Kamloops 3. Kelowna 4. Nanaimo 5. New Westminster 6. North Vancouver	7. Prince George 8. Saanich 9. Squamish 10. Surrey 11. Vancouver 12. Victoria 13. White Rock	1. Fredericton 2. Moncton 3. Saint John	1. Yellowknife
Nova Scotia	Ontario		Quebec	Saskatchewan
1. Cape Breton 2. Halifax 3. Nova Scotia	1. Barrie 2. Brampton 3. Brantford 4. Burlington 5. Caledon 6. Cambridge 7. Durham Region 8. Guelph 9. Hamilton 10. Huron County 11. Kingston 12. Kitchener 13. Muskoka	14. Newmarket 15. Niagara Falls 16. Niagara Region 17. Norfolk County 18. Oakville 19. Ottawa 20. St. Catharines 21. Toronto 22. Waterloo 23. Waterloo Region 24. Welland 25. York Region	1. Longueuil 2. Montreal 3. Québec 4. Repentigny 5. Rimouski 6. Shawinigan 7. Sherbrooke	1. Regina

3. Data collection period

Data were gathered between January 2018 and August 2018 from the applicable open data portals. Appendix A of the complete metadata document[7] provides the date each municipal dataset used in the ODB was downloaded. Readers are cautioned that the download date should not be used to indicate the reference period of the data. Ideally, the reference period for this building data would be clearly defined, however, the composition of datasets sourced from different data providers does not easily lend itself to a concrete reference period.

The period in which the underlying building data was collected or updated by the municipality was not consistently provided through the open data portals. If specific information concerning the reference period of data is required, users should contact the appropriate data providers shown in Appendix A of the complete metadata document.[8]

4. Target population

---

[7] Ibid.
[8] Ibid.

The goal of the ODB is to create a centralised and harmonized repository of building footprints available from government open data sources across Canada. Emphasis was placed on compiling buildings where substantial activity (e.g., economic activity) may occur. For this reason, footprints smaller than $10m^2$ were removed as they were deemed too small to be of interest for most users. As work on the experimental ODB progresses, definitions and thresholds will evolve. Readers are reminded that unedited data can always be obtained directly from the open data portals of the various data providers.

5. Compilation methodology

Individual building datasets were downloaded from their respective government open data portals. Within the 61 datasets obtained, each data provider attached a different set of variables to the building footprints. Only variables common across all data sources were included in the final output. Prior to merging, datasets were standardised and cleaned to ensure consistency in measurements and units. In the few cases where geographical coverage of municipal datasets intersected each other, a five-meter buffer was applied to one dataset, and buildings within that buffer from the other datasets were removed.

6. Data dictionary

The data dictionary below describes the variables contained within the exploratory ODB.

Variable - Object ID	
Name	OBJECTID (alias FID)
Format	Object ID (specific to ArcGIS)
Source	Internally generated during data processing
Description	Unique object or field ID automatically generated during data processing

Variable - Shape	
Name	Shape
Format	Geometry (specific to ArcGIS)
Source	Automatically generated from the geometric properties of the polygon datasets
Description	Describes the type of vector data in the dataset

Variable - Longitude	
Name	Longitude
Format	Double
Source	Generated from the centroid of the building polygon
Description	The longitude in decimal degrees of the centroid of the building polygon

Variable - Latitude	
Name	Latitude
Format	Double
Source	Generated from the centroid of the building polygon
Description	The latitude in decimal degrees of the centroid of the building polygon

Variable - Census subdivision unique identifier	
Variable Name	CSDUID
Data Format	Integer
Source	Canadian census subdivision boundaries 2016 (Statistics Canada GeoSuite product)
Description	The census subdivision ID where the building is located

Variable - Census subdivision name	
Name	CSDNAME
Format	String
Source	Canadian census subdivision boundaries 2016 (Statistics Canada GeoSuite product)
Description	Name of census subdivision

Variable - Data Provider	
Name	Data_prov

Format	Text (String)
Source	Created based on origins of input dataset
Description	Name of the municipality, region, or province/territory that provided the dataset

Variable - Building Unique ID	
Name	Build_ID
Format	Text (String)
Source	Generated by concatenating the census subdivision unique identifier with a sequential 6-digit code that is unique to each building in that census subdivision
Description	A unique identification for each building based on the census subdivision in which it resides

Variable - Perimeter	
Name	Shape_Length
Format	Float
Source	Calculated from the geometric attributes of the building polygon
Description	The perimeter of the building in units of metres (name specific to ArcGIS)

Variable - Area	
Name	Shape_Area
Format	Float
Source	Calculated from the geometric attributes of the building polygon
Description	The area of the building in units of square metres (name specific to ArcGIS)

7. Record layout

An example of the record layout in Table 7-1 format is reported below. Items marked with an asterisk represent fields indexed automatically by the GIS software used during processing.

As open data holdings expand in Canada, the hope is that future updates will enrich the ODB by adding more variables and increasing the geographic coverage.

Table 7-1
Example record layout for the Open Database of Buildings (ODB)

OBJECTID*	Shape*	Longitude	Latitude	CSDUID	CSDNAME	Data_prov	Build_ID	Shape_Length	Shape_Area
1	Polygon	-115.561757	51.18907	4815035	Banff	Banff	48150350000001	16.560241	16.963528
2	Polygon	-115.569331	51.171372	4815035	Banff	Banff	48150350000002	87.531972	330.625531
3	Polygon	-115.569616	51.178173	4815035	Banff	Banff	48150350000003	104.044015	573.938947

8. Data accuracy

All building footprints in the ODB were collected from government sources of 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". Readers are advised that the individual methodologies used, as well as the amount of information describing these methodologies, varied by data provider. For more information on an individual data provider's methodologies, please follow the links to their website found in Appendix A of the complete metadata document.[9]

During the processing stage to create the ODB, several steps were taken to improve accuracy and to standardization of the output. A small number of building geometries that had improbable shapes were removed. To ensure consistency

---

[9] Ibid.

in measurements and units (e.g., the measures for building area and perimeter), some variables in the underlying datasets were systematically removed and recalculated based on a standardized projection.

Where geographical coverage of input datasets intersected each other, it is important to note that the choice of buffer size applied (five metres was chosen) and which overlapped segments of buildings to remove will impact the results. The choice of which segments to keep was based on several heuristic factors, specifically preference was given to datasets that appeared to have more detailed polygons and/or a higher number of building counts.

The experimental ODB represents the governmental open data that was found at the time of compilation and should thus not be interpreted as a complete or objective "ground-truth" of what buildings actually exist in Canada. The current coverage across Canada of the ODB is incomplete with data primarily existing for larger cities. The gaps in the database reflect areas where open data on buildings from government sources was not found. As more open data becomes available and added to the ODB, the expectation is that coverage will increase over time.

9. Conclusion

Open micro-data from municipal, provincial and federal sources present tremendous opportunities for data integration and Statistics Canada has a role to play. The ODB is an example, maybe the first example, of the potential to NSOs of working in this open micro-data space.

As work is being done to expand the coverage of the ODB, other statistical domains are being explored for the creation of more harmonized open databases of micro-data (for example, building addresses, business listings, public infrastructure, as well as others). These databases, along with the potential of linking them together, will help NSOs contribute to the creation and maintenance of a set of basic data infrastructure that is key to supporting the Canadian economy and society in the increasingly digital, data driven world.

Along with new opportunities comes new challenges, and most of the challenges facing NSOs as they work in the open micro-data space have yet to be fully unpacked. For instance, the use of georeferenced open micro-data is limited to non-personal and non-sensitive statistics, but the exact definition of what falls into these categories may vary across jurisdictions and is partly driven by public perception. Questions of data ownership, openness and sharing, to mention a few, will inevitably be part of this debate, which may take on unique connotations in the multilevel governance system that typically characterizes open data generated at the sub-national level.

References

Statistics Canada (2019), *The ODB - an experimental Open Database of Buildings, Metadata document: concepts, methodology and data quality.*