The Linkable Open Data Environment - LODE:
Overview of Data, Tools, and Collaborations

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

May 2, 2019

Delivering insight through data, for a better Canada

Open microdata: a vast, growing but still underutilized type of data

- Microdata – non sensitive and non-personal information on buildings, businesses, addresses, property values, infrastructure assets, and much more

- From authoritative sources – municipal, regional, provincial governments and, increasingly, also private sector stakeholders

- Released with an open data license – that encourages the use of the data

- Rapidly expanding

Linkable Open Data Environment (LODE)

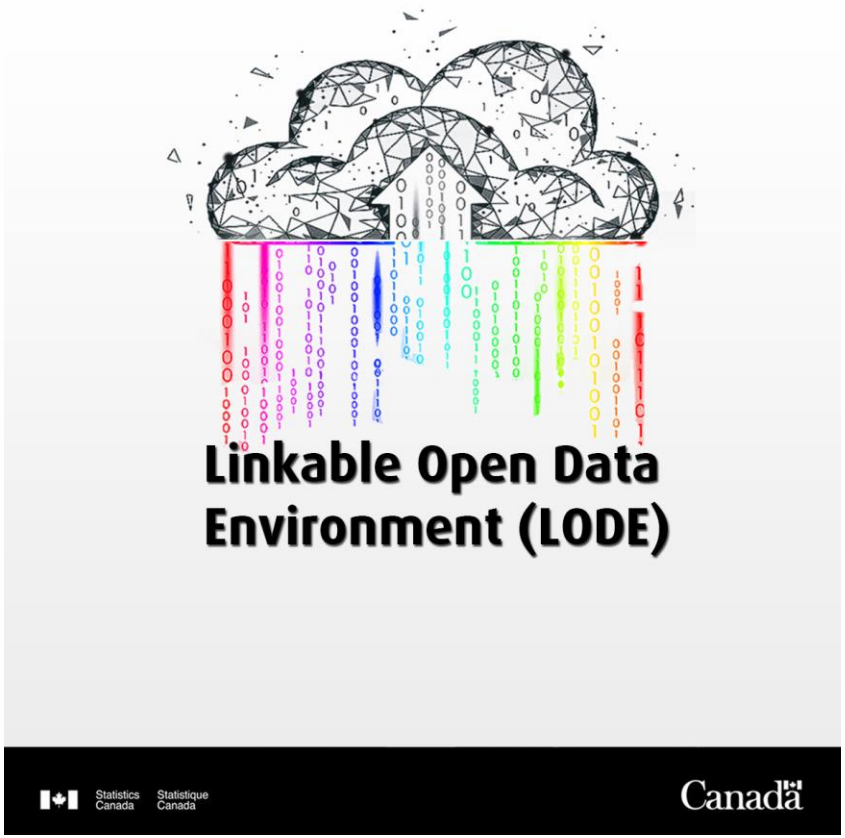
- Open microdata from authoritative sources that have been brought into an environment that is suitable for data linkage

- Goal: harmonised and standardised datasets made available under the Open Government License - Canada

- Vast majority of datasets are from governmental sources (municipal, regional, provincial or federal)

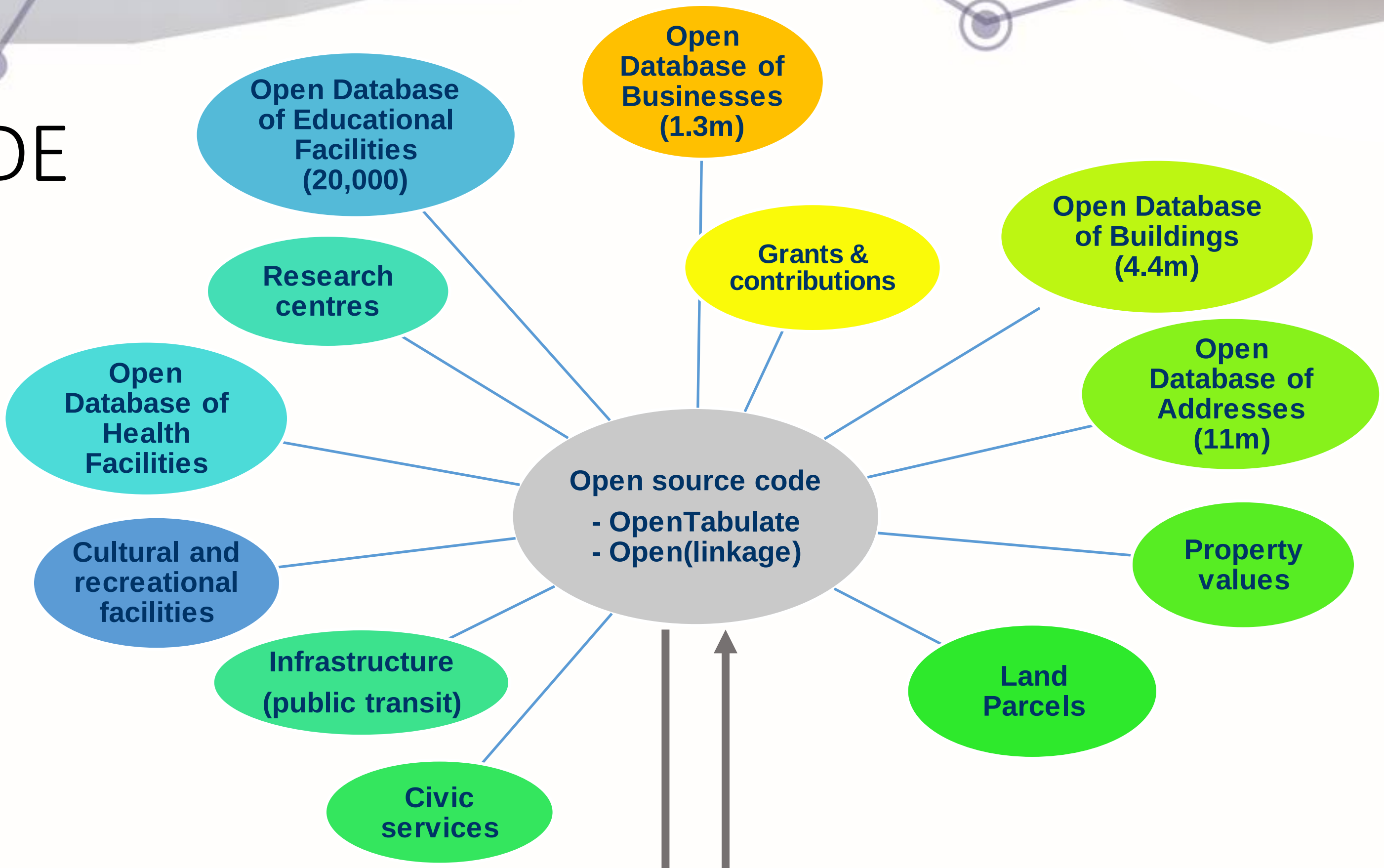
- LODE is in development and you can use/contribute to it

Draft logo


Linkable Open Data Environment (LODE)
Statistics Canada Statistique Canada
Canada

- Experimental and work in progress
- Much of LODE development is expected to be completed between April 2019 and March 2020
- Funded through external cost-recovery projects and internal projects

LODE


Open Database of Educational Facilities (20,000)
Open Database of Businesses (1.3m)
Grants & contributions
Open Database of Buildings (4.4m)
Research centres
Open Database of Addresses (11m)
Open Database of Health Facilities
Open source code
- OpenTabulate
- Open(linkage)
Property values
Cultural and recreational facilities
Infrastructure (public transit)
Land Parcels
Civic services

Statistics Canada microdata holdings (not open data)

Other major open data initiatives

- OpenStreetMap
- LandGIS
- Wikipedia

LODE Databases
open for you to be used and shared

Open Database of Buildings – version 2, March 1st, 2019

https://www.statcan.gc.ca/eng/open-building-data/index

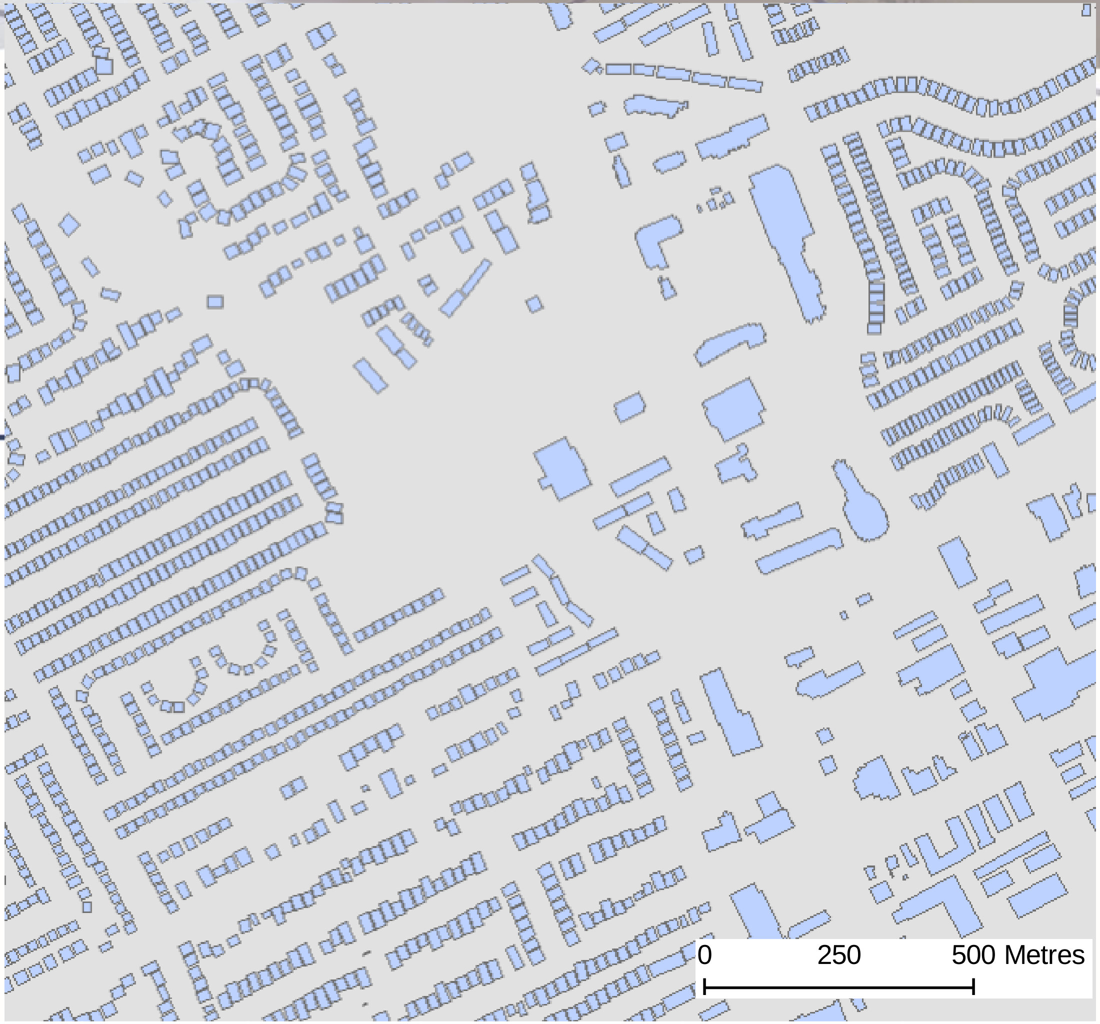
Open Database of Buildings (ODB)
Statistics Canada Statistique Canada
Canada

- A compilation of 65 datasets originating from various government sources of open data (provincial, municipal)
- 4.4 million records of building footprints and variables calculated and standardized across all data providers
- Harmonised and standardised dataset made available under the Open Government License - Canada

The ODB: example of the data

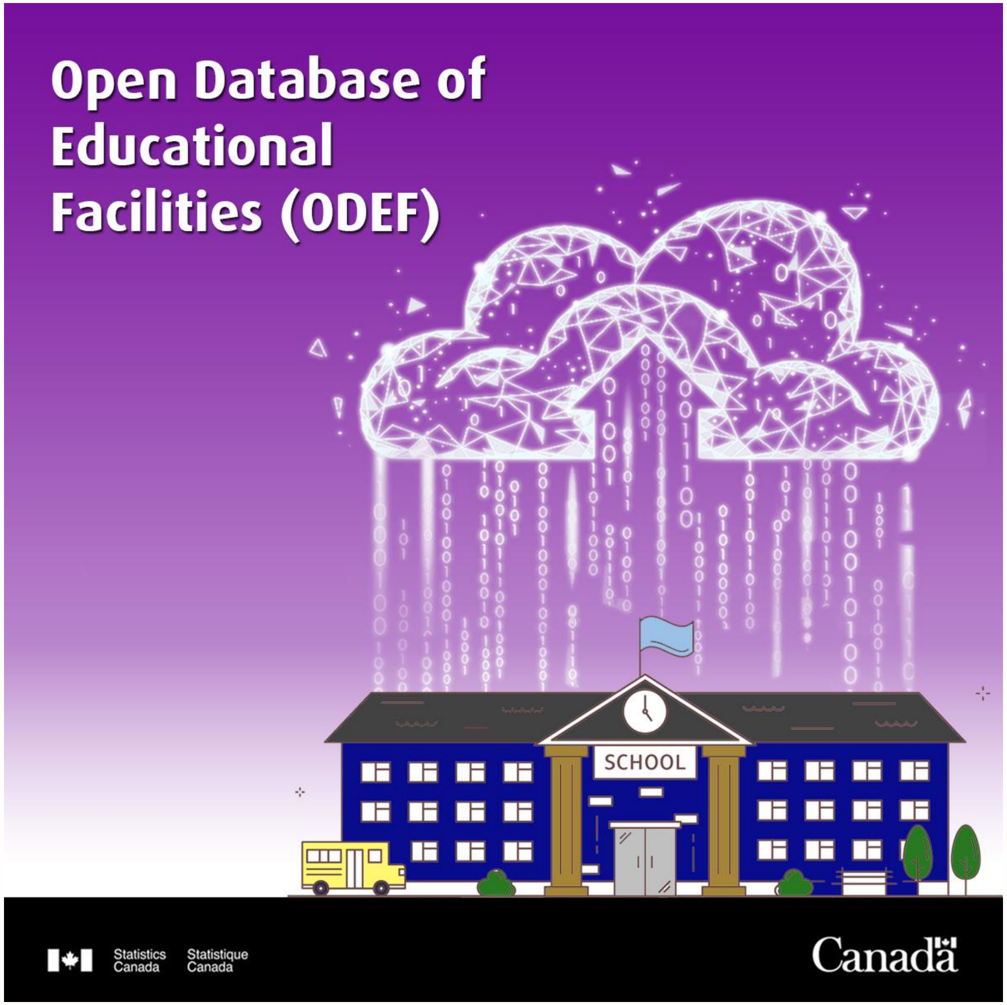
- Ex: Footprints for Richmond Hill, Toronto

- Quality is generally high, buildings are tightly knit



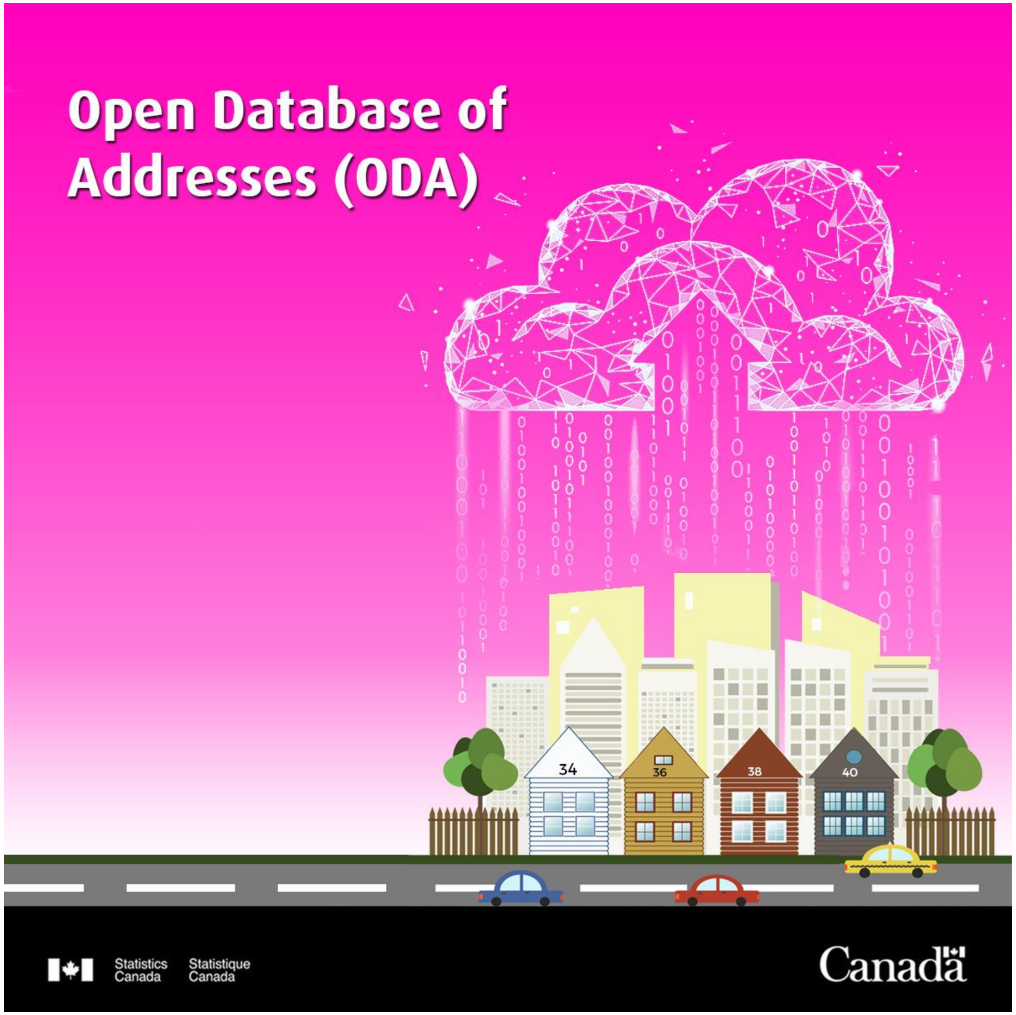
OBJECTID*	Shape*	Longitude	Latitude	CSDUID	CSDNAME	Data_prov	Build_ID	Shape_Length	Shape_Area
1	Polygon	-115.561757	51.18907	4815035	Banff	Banff	48150350000001	16.560241	16.963528
2	Polygon	-115.569331	51.171372	4815035	Banff	Banff	48150350000002	87.531972	330.625531
3	Polygon	-115.569616	51.178173	4815035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

Draft logo



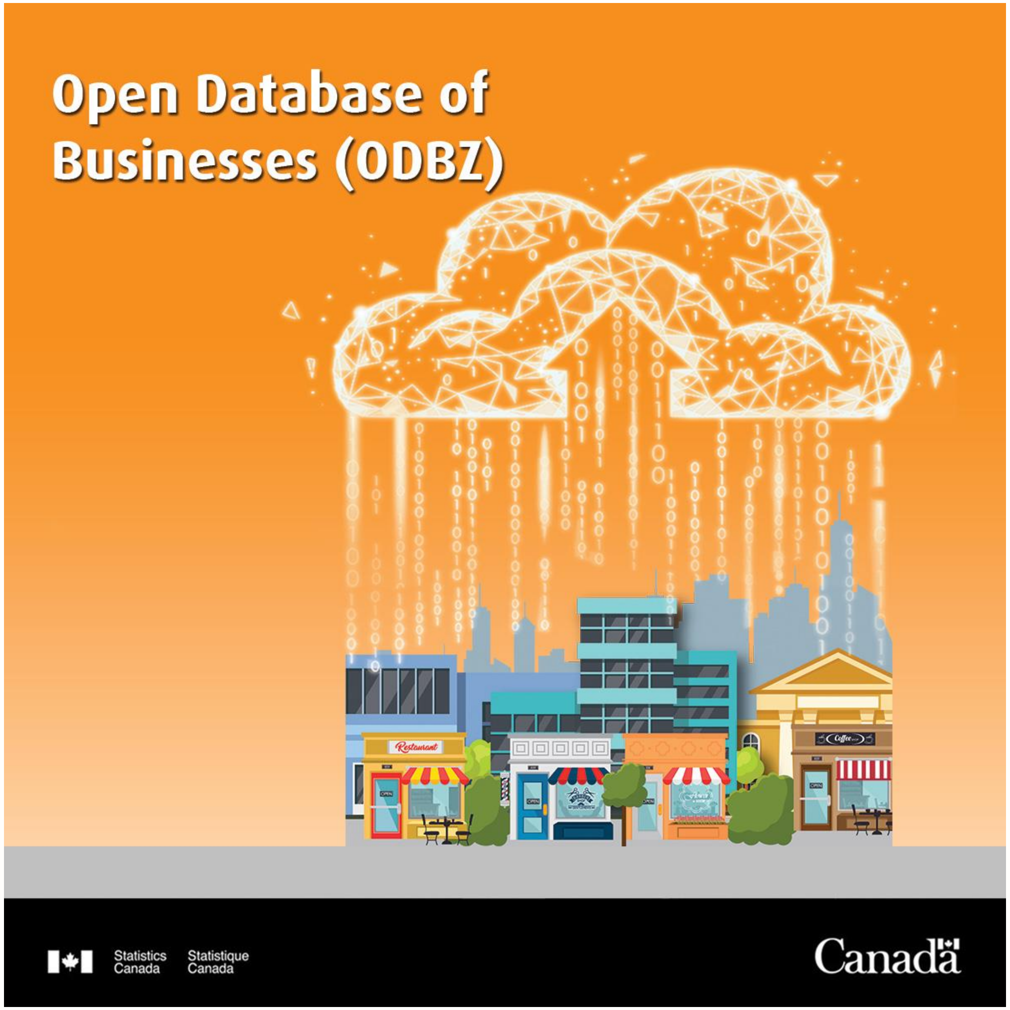
- Completed with cost-recovery project funded by Crown-Indigenous Relations and Northern Affairs Canada (CIRNAC)
- 20,000 records (open data + publicly available data)
- Currently finalizing the review of licenses and approvals to release (public data) with GoC open license
- Harmonised and standardised dataset will be made available under the Open Government License – Canada
- Expected release: May/June 2019 – CSV file

Draft logo


Open Database of Addresses (ODA)
34
36
38
40
Statistics Canada Statistique Canada
Canada

- Preliminary version is completed, work in collaboration with OpenAddresses
- About 11 million records from municipal and provincial open data sources
- Harmonised and standardised dataset will be made available under the [Open Government License – Canada]
- It will not include Postal Codes
- Expected release: June 2019

Draft logo


Open Database of Businesses (ODBZ)
Restaurant
Coffee
Statistics Canada Statistique Canada
Canada

- Preliminary version is available in LODE (GitHub)
- 26 data providers (federal, provincial, municipal)
- Approximately 1.4 million records (to be cleaned, de-duplicated, harmonised, etc.)

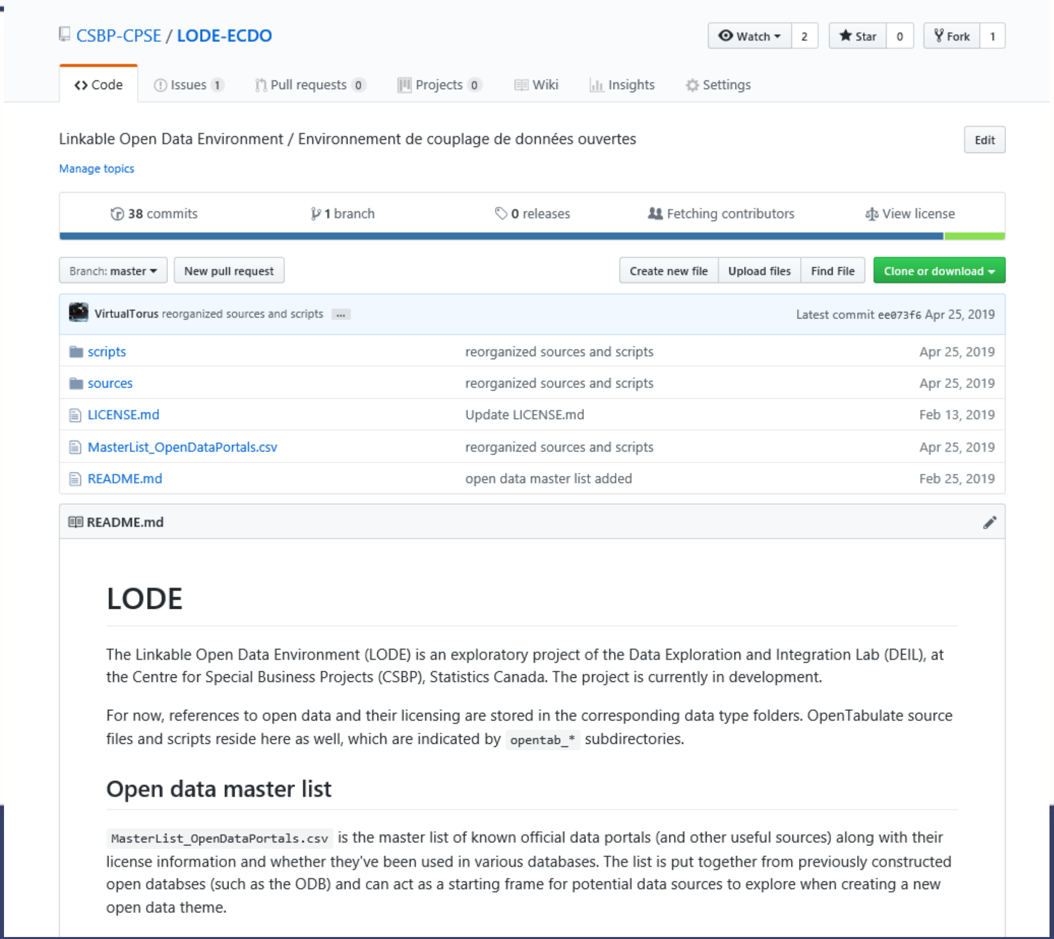
Draft logo

- Work in progress
- Preliminary compilation of datasets available in LODE (GitHub)

LODE resources and processing tools for open collaboration, use and development

LODE on GitHub: repository of open microdata data sources

See: https://github.com/CSBP-CPSE/LODE-ECDO



Linkable Open Data Environment / Environnement de couplage de données ouvertes

VirtualTorus reorganized sources and scripts … Latest commit ee073f6 Apr 25, 2019

scripts	reorganized sources and scripts	Apr 25, 2019
sources	reorganized sources and scripts	Apr 25, 2019
LICENSE.md	Update LICENSE.md	Feb 13, 2019
MasterList_OpenDataPortals.csv	reorganized sources and scripts	Apr 25, 2019
README.md	open data master list added	Feb 25, 2019

README.md

LODE

The Linkable Open Data Environment (LODE) is an exploratory project of the Data Exploration and Integration Lab (DEIL), at the Centre for Special Business Projects (CSBP), Statistics Canada. The project is currently in development.

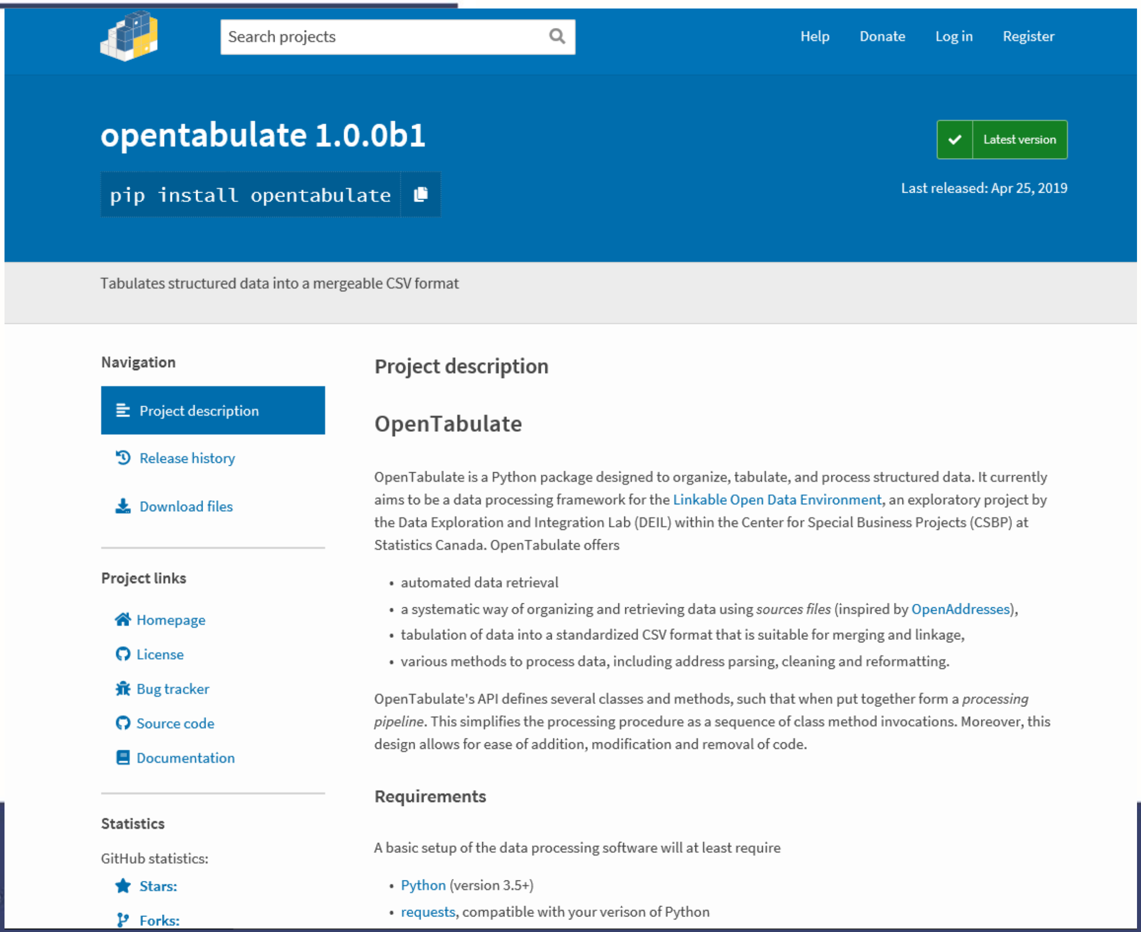
For now, references to open data and their licensing are stored in the corresponding data type folders. OpenTabulate source files and scripts reside here as well, which are indicated by `opentab_*` subdirectories.

Open data master list

`MasterList_OpenDataPortals.csv` is the master list of known official data portals (and other useful sources) along with their license information and whether they've been used in various databases. The list is put together from previously constructed open databses (such as the ODB) and can act as a starting frame for potential data sources to explore when creating a new open data theme.

OpenTabulate 1.0 - a Python package, on Pypi, for LODE compilation and data cleaning

[See: https://pypi.org/project/opentabulate/](https://pypi.org/project/opentabulate/)



opentabulate 1.0.0b1

```
pip install opentabulate
```

Last released: Apr 25, 2019

Tabulates structured data into a mergeable CSV format

Navigation

- Project description
- Release history
- Download files

Project links

- Homepage
- License
- Bug tracker
- Source code
- Documentation

Statistics

GitHub statistics:

- Stars:
- Forks:

Project description

OpenTabulate

OpenTabulate is a Python package designed to organize, tabulate, and process structured data. It currently aims to be a data processing framework for the Linkable Open Data Environment, an exploratory project by the Data Exploration and Integration Lab (DEIL) within the Center for Special Business Projects (CSBP) at Statistics Canada. OpenTabulate offers

- automated data retrieval
- a systematic way of organizing and retrieving data using *sources files* (inspired by OpenAddresses),
- tabulation of data into a standardized CSV format that is suitable for merging and linkage,
- various methods to process data, including address parsing, cleaning and reformatting.

OpenTabulate's API defines several classes and methods, such that when put together form a *processing pipeline*. This simplifies the processing procedure as a sequence of class method invocations. Moreover, this design allows for ease of addition, modification and removal of code.

Requirements

A basic setup of the data processing software will at least require

- Python (version 3.5+)
- requests, compatible with your verison of Python

Open microdata is enabling collaboration

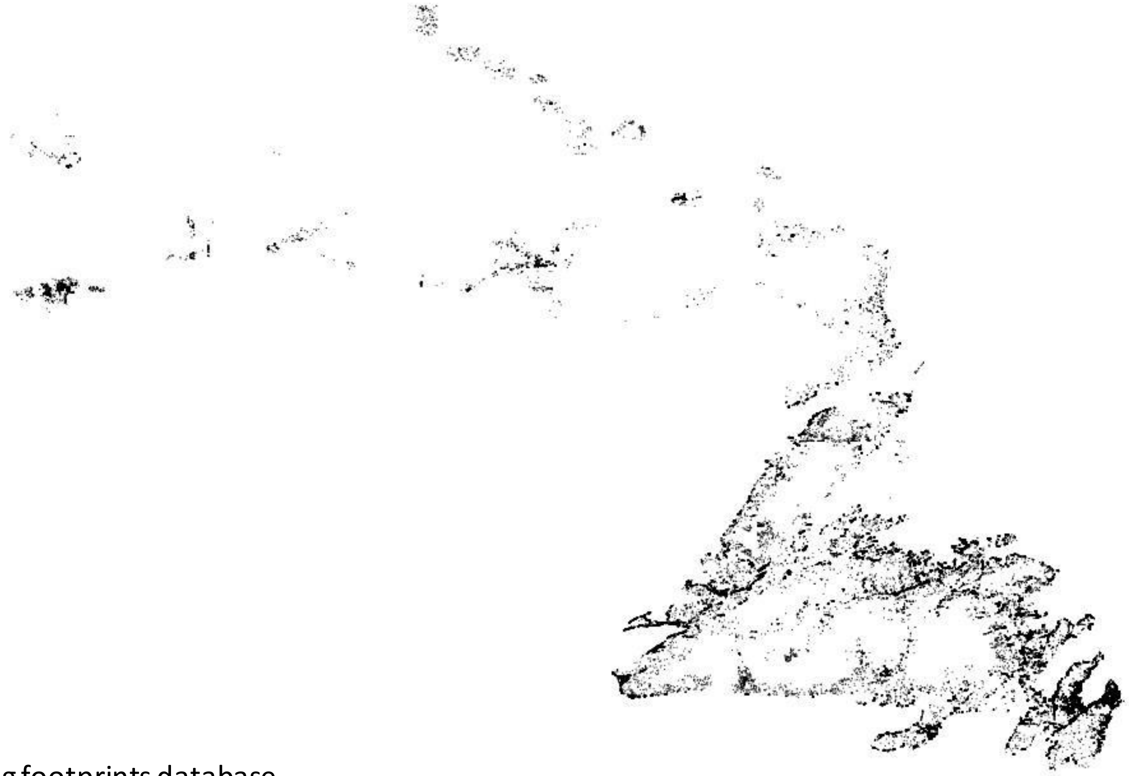
Opportunity: collaborative data creation and analysis with many stakeholders **outside Statistics Canada**

- Open data as enabler and data value multipliers
- Reduces barriers to data sharing and costs (administrative, production, management). No need to have complex agreements and administrative burden
- Enriching the open data ecosystems may be one way we can unlock more data and more of their value
- Examples:
  - **Fleming College** collaborative project (ODB analytics and web mapping)
  - **UBC Master's of Data Science** collaborative project (ODB analytics)
  - **Canadian Read Cross** and **Digital Academy (CSPS)** on data analytics
  - **OSM communities across Canada** data imports into OSM and data cleaning and improvements

Enabling collaboration: Microsoft building footprints

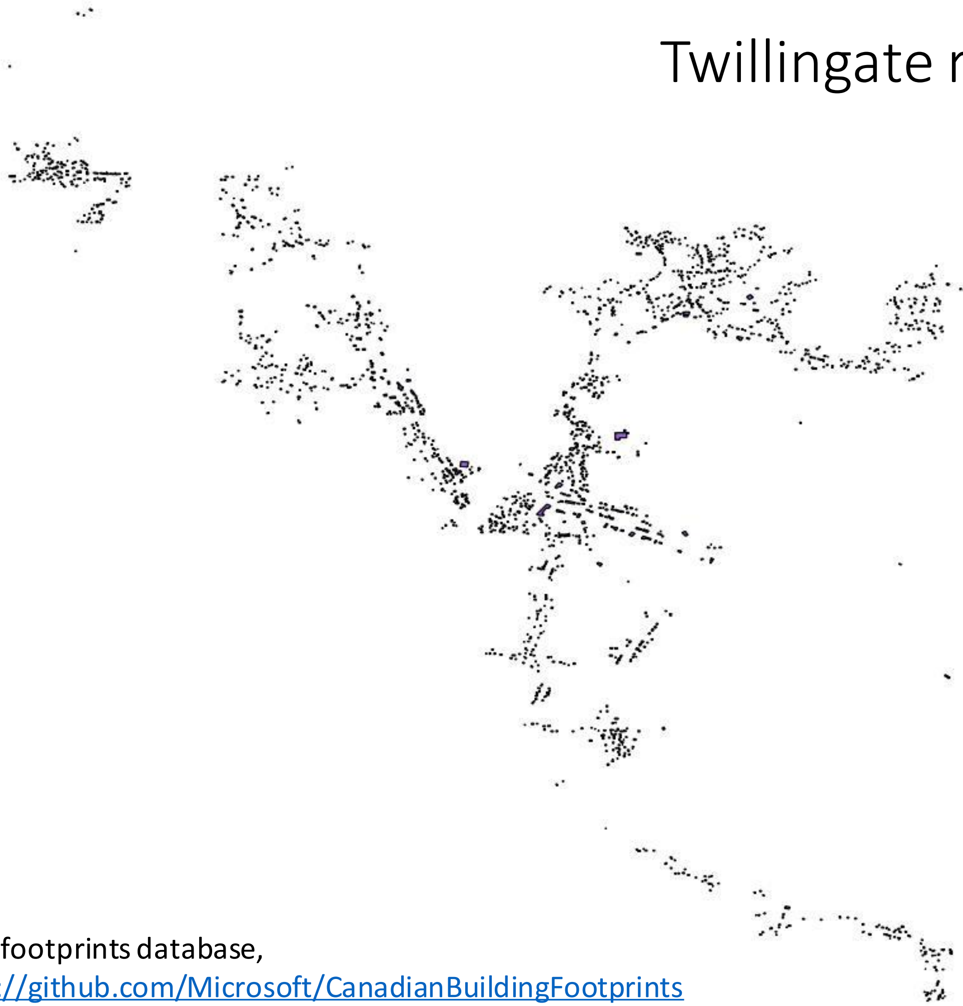
- July 2018, StatCan and Microsoft started a collaboration to complete the mapping of building footprints across Canada
- Microsoft had released an open database of 125 million footprints for the U.S. based on satellite imagery extraction
- Microsoft used the Open Database of Building (version 1.0) to train a neural network model to extract building footprints from satellite imagery. The Microsoft database (about 12 million footprints) is available at:
  - Microsoft blog post (link)
  - Bing blogs (link)
  - https://github.com/Microsoft/CanadianBuildingFootprints
- Open data is a collaboration enabler and value multiplier

Newfoundland and Labrador … as never seen before



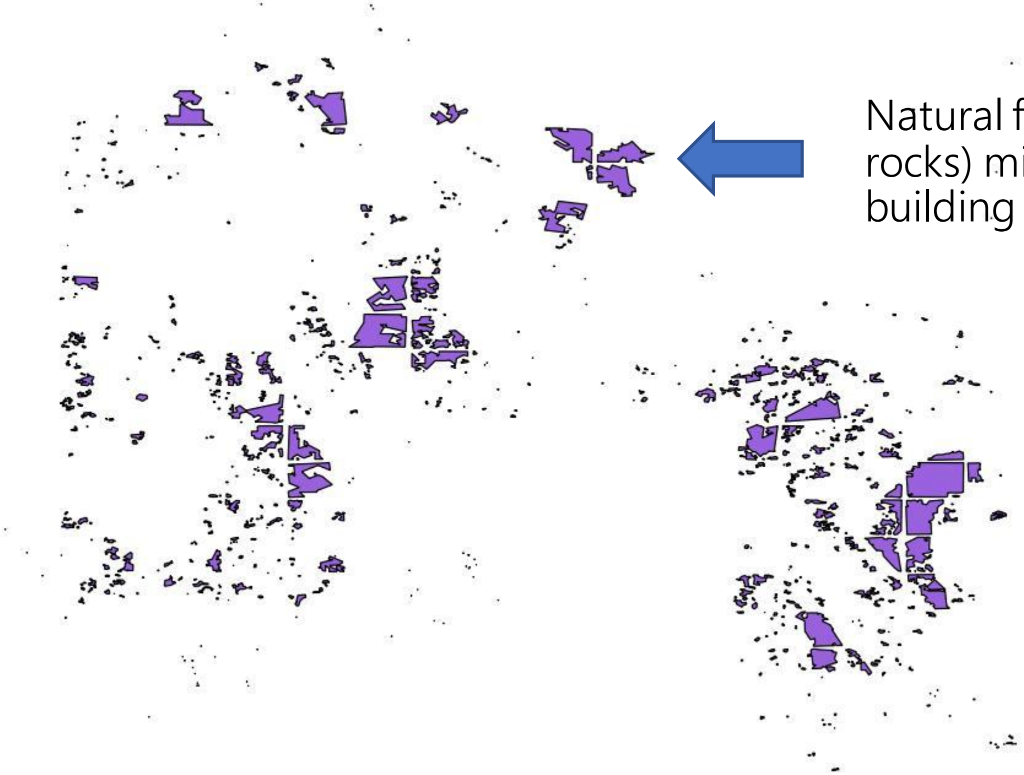
Source: Microsoft building footprints database,
available at: https://github.com/Microsoft/CanadianBuildingFootprints

Twillingate region (NFL)



Source: Microsoft building footprints database,
available at: https://github.com/Microsoft/CanadianBuildingFootprints

...but still a lot of false positives (mostly in remote regions)



Natural features (e.g., rocks) misclassified as building footprints

Source: Microsoft building footprints database, examples from Labrador region, available at: https://github.com/Microsoft/CanadianBuildingFootprints

Conclusions

- You can use LODE (data and code)
- You can contribute to data development
- You can contribute to code development
- LODE is open and is there to be used!

THANK YOU!

For more information,
alessandro.alasia@canada.ca
haaris.jafri@canada.ca

MERCI!

Pour de plus amples renseignements,
alessandro.alasia@canada.ca
haaris.jafri@canada.ca

#StatCan100