Open Projects at DEIL (Data Exploration and Integration Lab):

Open buildings, open businesses and a Linkable Open Data Environment

Alessandro Alasia
Statistics Canada

Presented to the Institute for Data Science
Carleton University, March 13 - 2019

Delivering insight through data, for a better Canada

Context

- Statistics Canada has a history of producing open data but the environment is evolving
- NSOs are increasingly becoming consumers and stakeholders of open microdata; there are new producers of open microdata
- This presentation is about the “discovery” of a new data mine and a novel approach to mine these data

1. **Crowdsourcing pilot**
2. **Open Databases**
3. ***Open Project approach***
4. ***Linkable Open Data Environment (LODE)***

Lessons learned from Crowdsourcing Pilot (2016 – 2018)

- Large volumes of open microdata are available from municipal and provincial sources
- High quality, reliable, accurate data provided with an open data license
- Enhanced potential of open source software and open development platforms to collaborate and reduce barriers to entry in the new data ecosystem

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

- **Microdata** – non-sensitive and non-personal information on buildings, businesses, addresses, property values, infrastructures, and much more
- **From authoritative sources** – municipal, regional, provincial governments and, increasingly, also private sector stakeholders
- **Released with an open data license** – encourages the use of the data
- **Rapidly expanding**

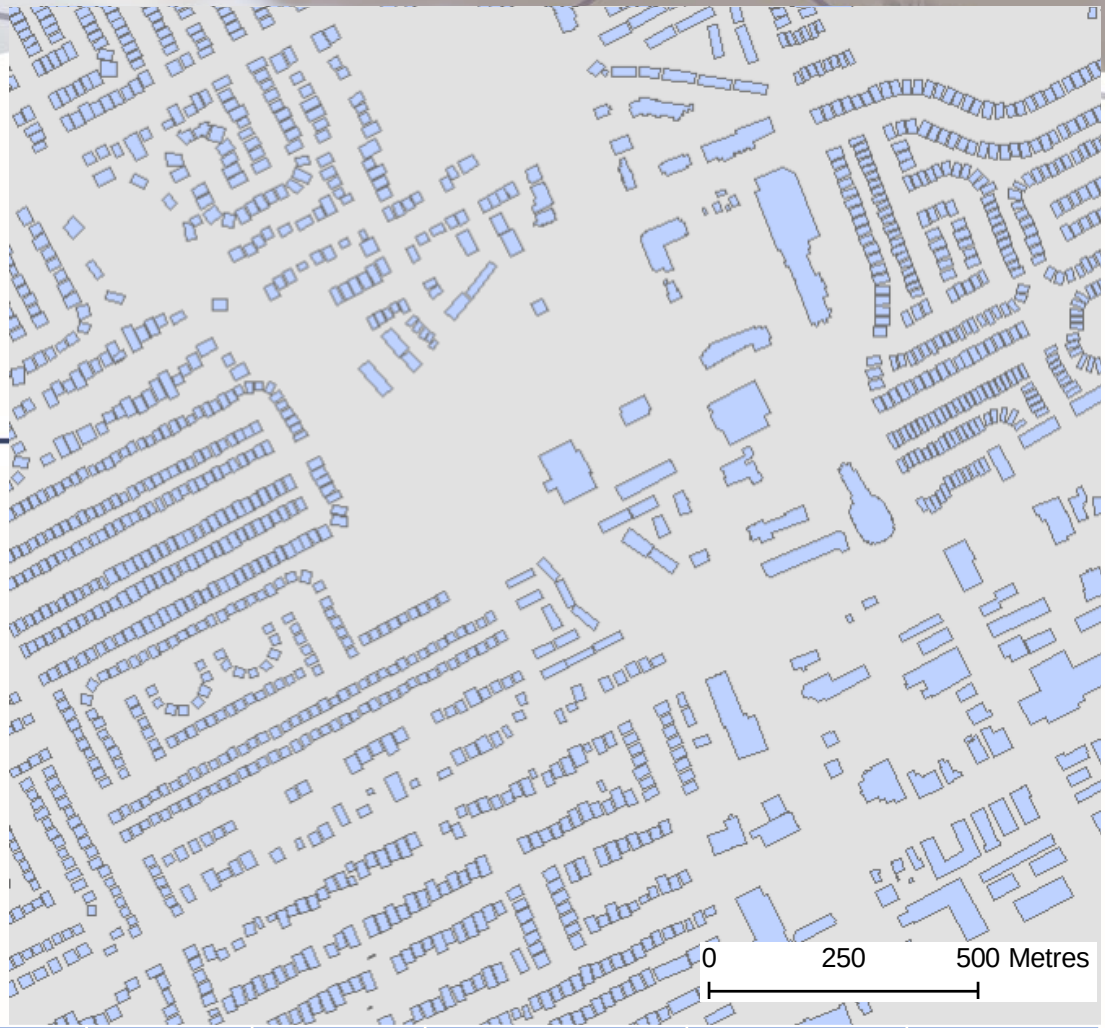
Open Database of Buildings – version 2, March 1st, 2019

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

- 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
- Harmonized and standardized 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

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 is available at:
  - https://github.com/Microsoft/CanadianBuildingFootprints
  - Microsoft blog post (link)
- Open data is a collaboration enabler and value multiplier

Is there a demand for harmonized open micro databases?

- **About 1,100 downloads** of the Open Database of Buildings (Nov 2018 – Jan 2019)
  - **Fleming College** and (possibly) UBC Master's of Data Science are using the Open Database of Buildings for analytical projects with students
  - **Digital Academy** (CSPS) is considering the use of the Open Database of Buildings and other open databases for the development of training modules
  - **Canadian Red Cross** have expressed substantial interest in this information for its operational work (disaster management and relief efforts)

- **There is enormous potential for use in Statistics Canada's programs** (building register, housing, environment, census, agriculture, etc.)

- There is need for **harmonization and standardizations of original data sources**

1

Crowdsourcing pilot

2

Open Databases

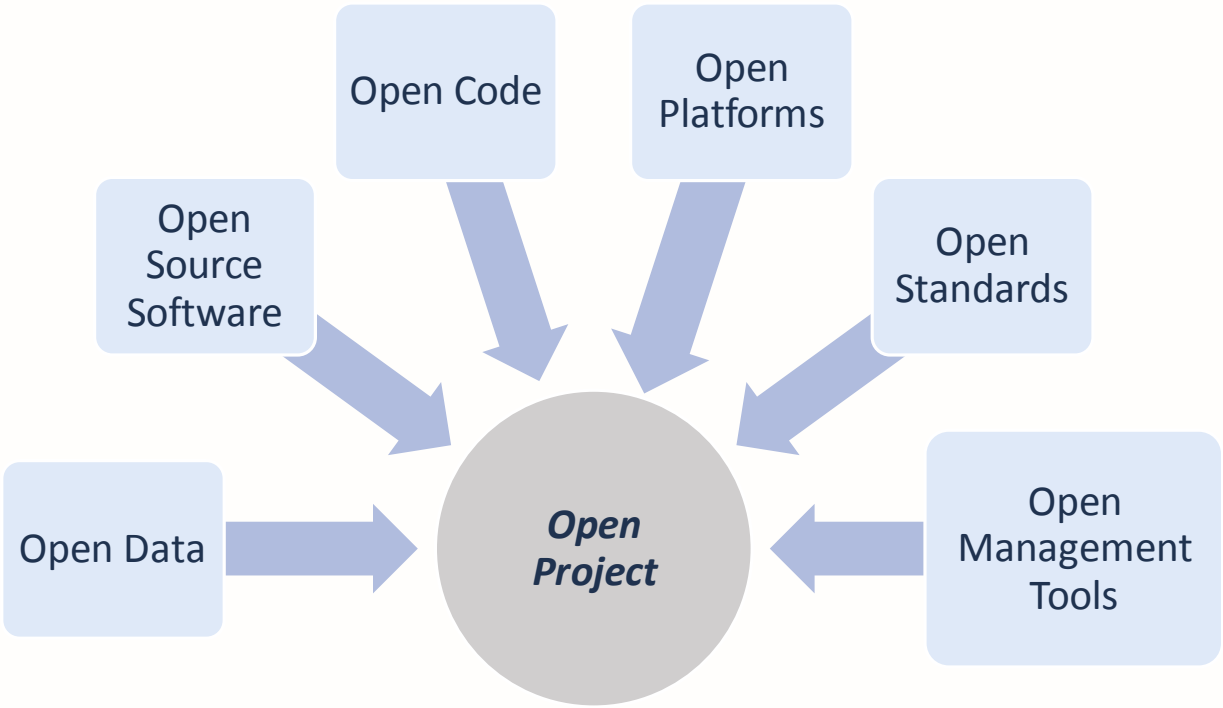
3

***Open Project* approach**

4

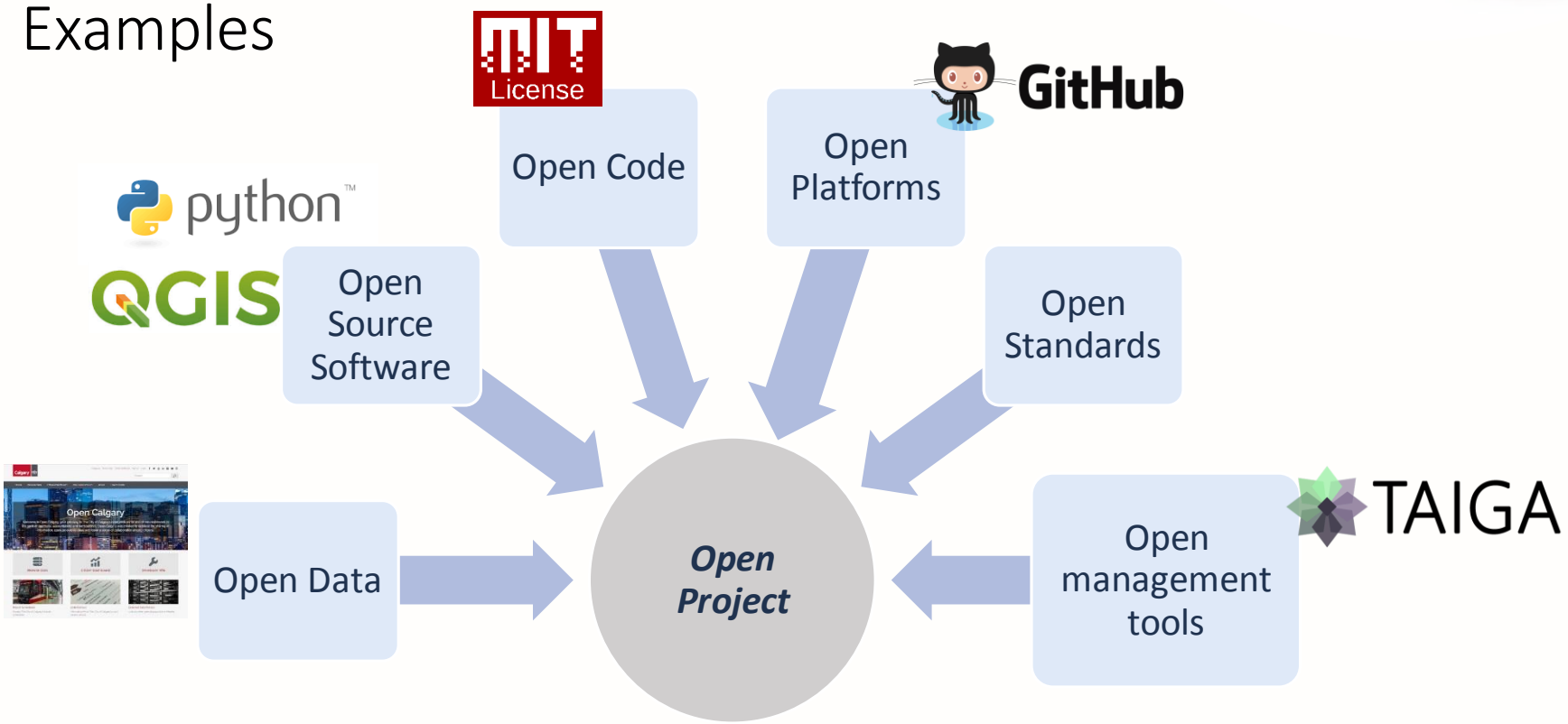
Linkable Open Data Environment (LODE)

The *Open Project* idea



Open Project = facilitates collaboration, reduces duplication, increases efficiency and transparency, accelerates knowledge sharing, amplifies value of data

Examples


License
GitHub
Open Code
Open Platforms
python
QGIS
Open Source Software
Open Standards
Open Calgary
Open Data
Open Project
Open management tools
TAIGA

Open project approach in data development:

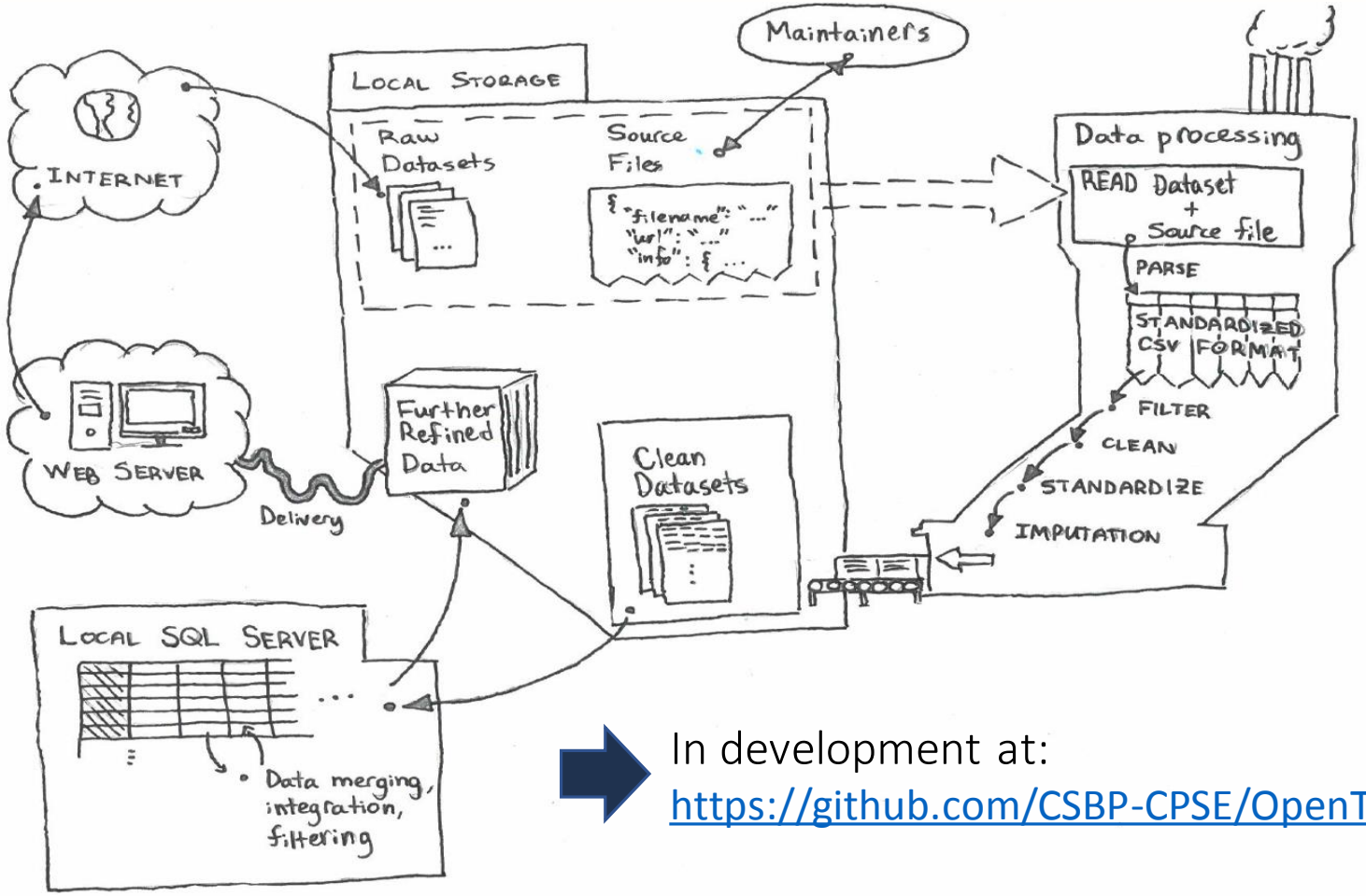
Open Database of Businesses

(work in progress)

The Open Database of Businesses (ODBZ)

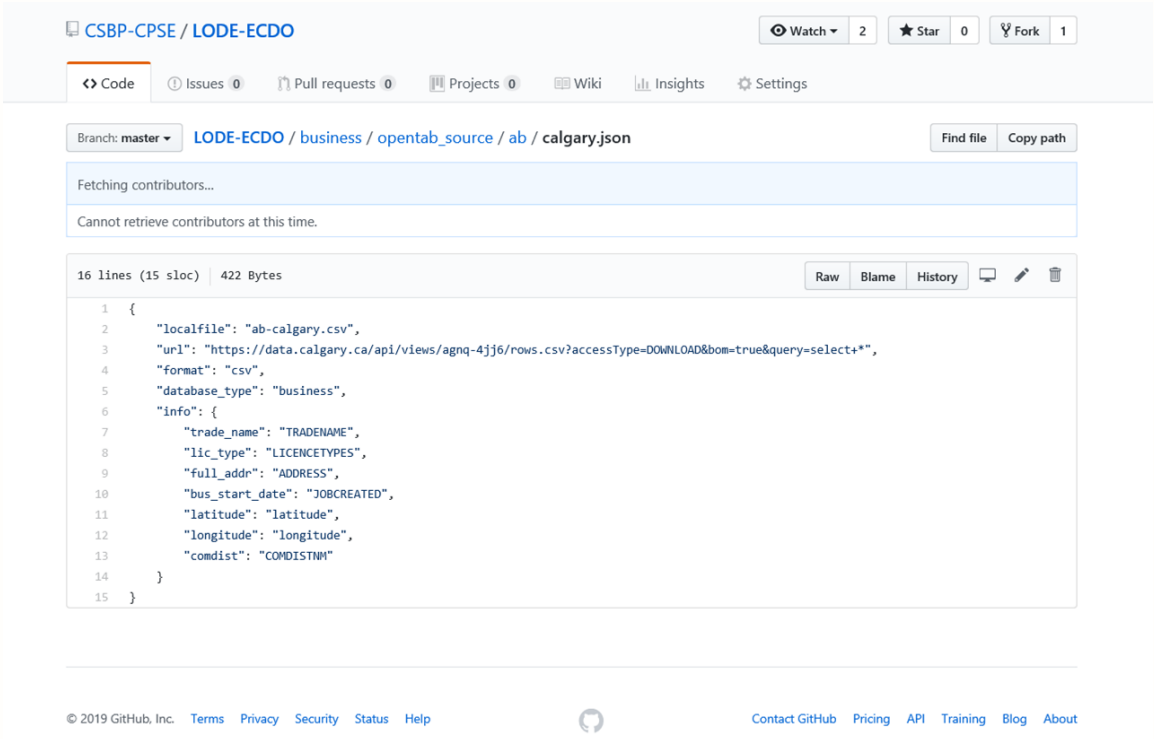
- Experimental open project initiated in the Summer 2018; work in progress
- Developed as open software solution ≠ database
- Uses government sources of open microdata on businesses
- Python script to compile, process, integrate; in development at:
  - https://github.com/CSBP-CPSE/OpenTabulate

A generalized workflow



In development at:
https://github.com/CSBP-CPSE/OpenTabulate

The Open Database of Businesses: example of a *source file - json*



Source data

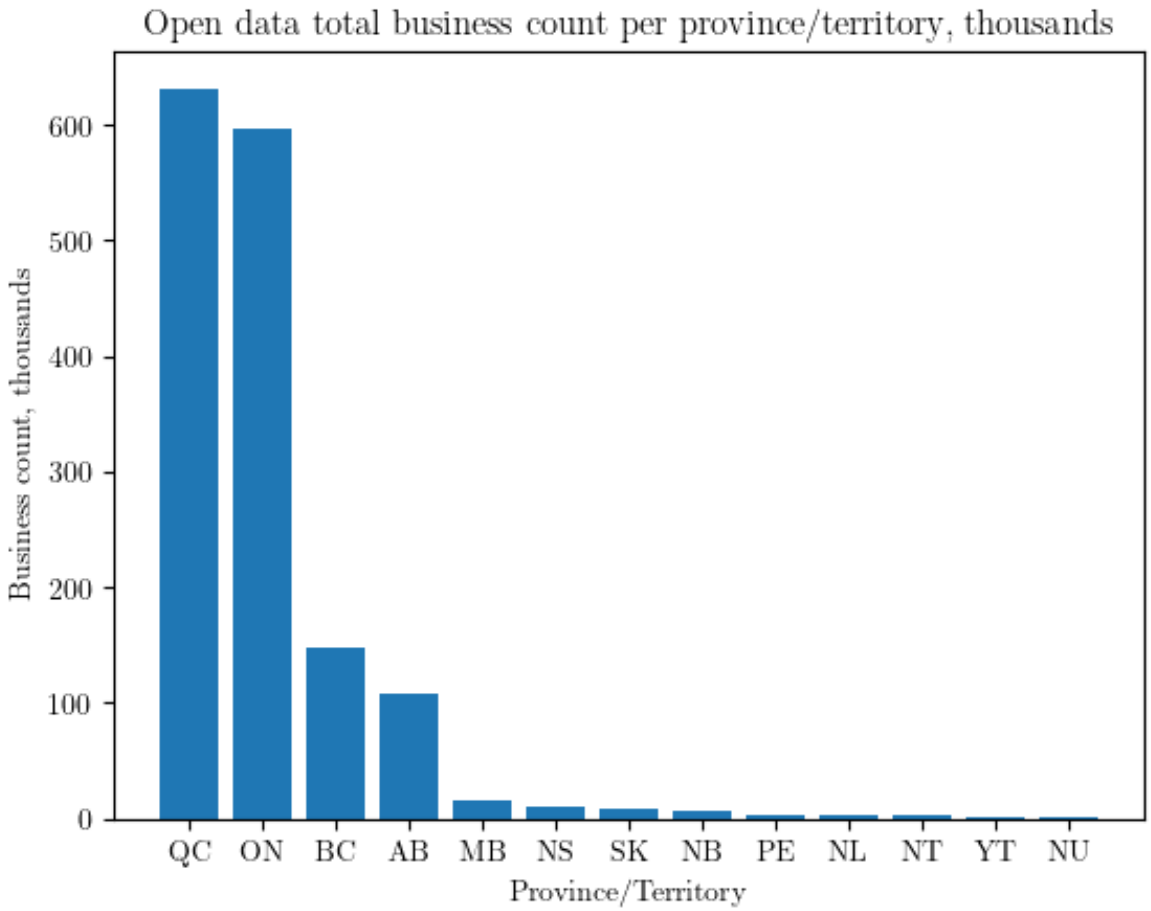
- 26 data providers (federal, provincial, municipal)
- Examples:

  - (Federal) Federal Corporations (GoC open data portal)

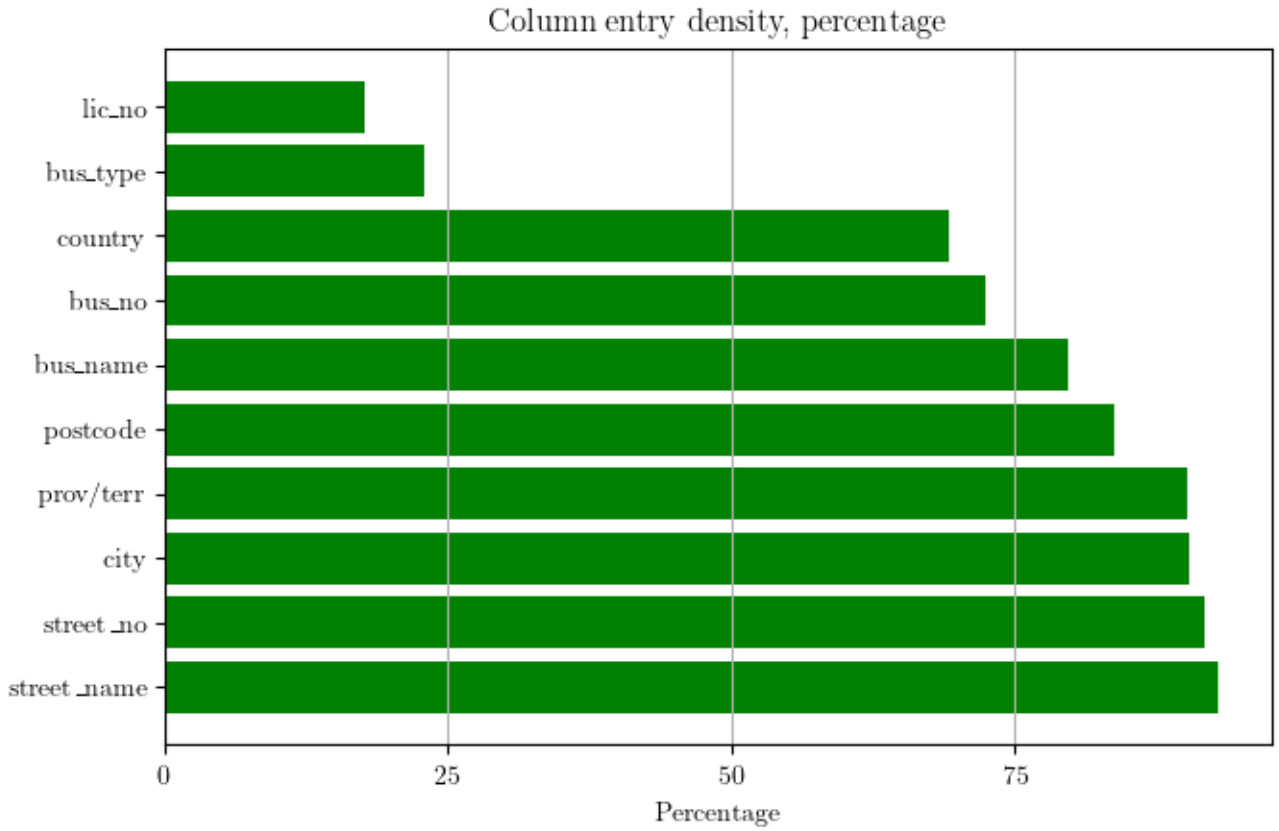
  - (Provincial) BC Indigenous Business Listings (BC Open data catalogue)
  - (Provincial) Registre des entreprises (données ouvertes Québec)

  - (Municipal) Toronto - Business Licences and Permits
  - (Municipal) Vancouver - Business licence
  - (Municipal) Durham - Business Directory

The Open Database of Businesses: *preliminary results*



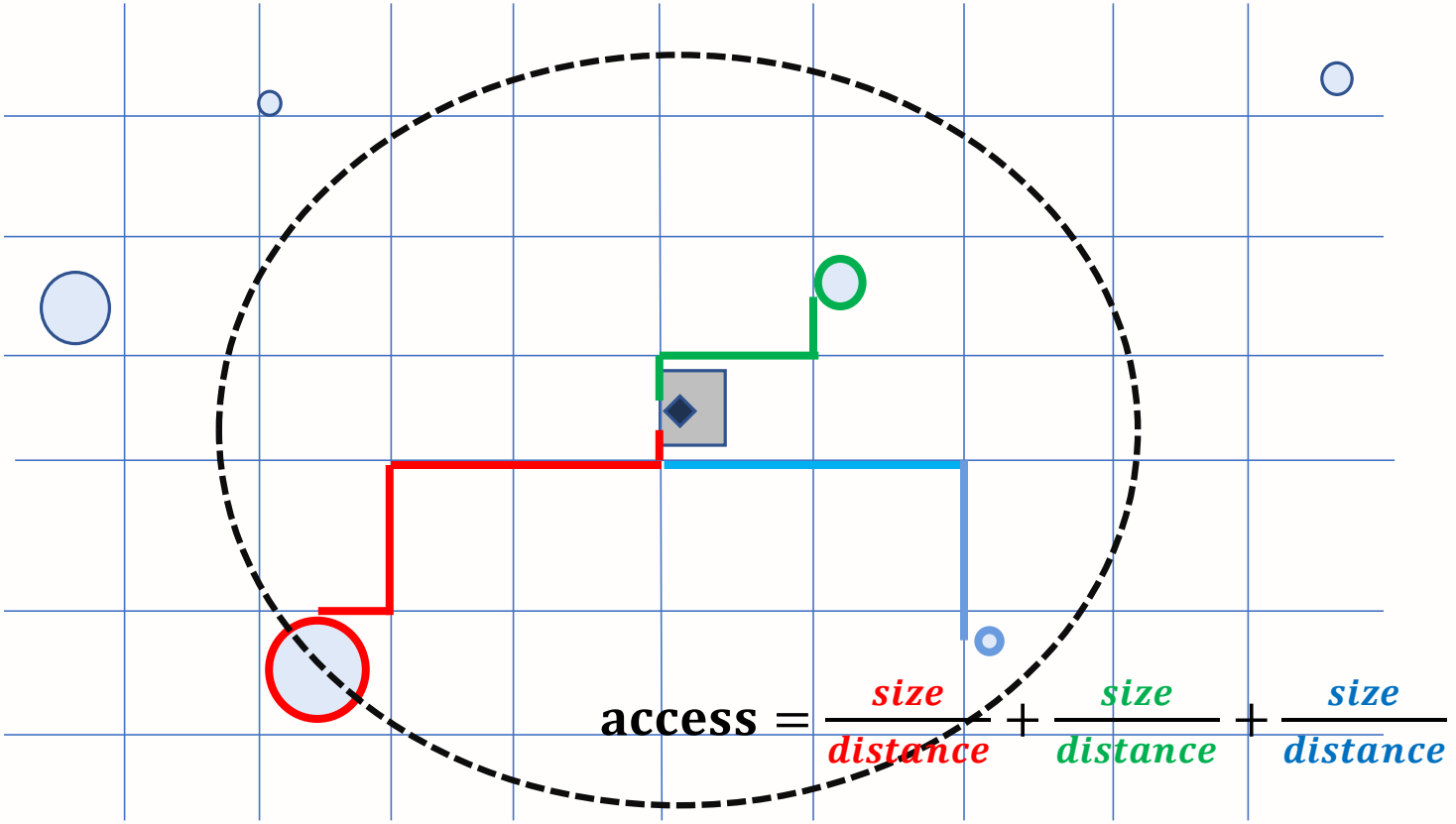
The Open Database of Businesses : *preliminary results*



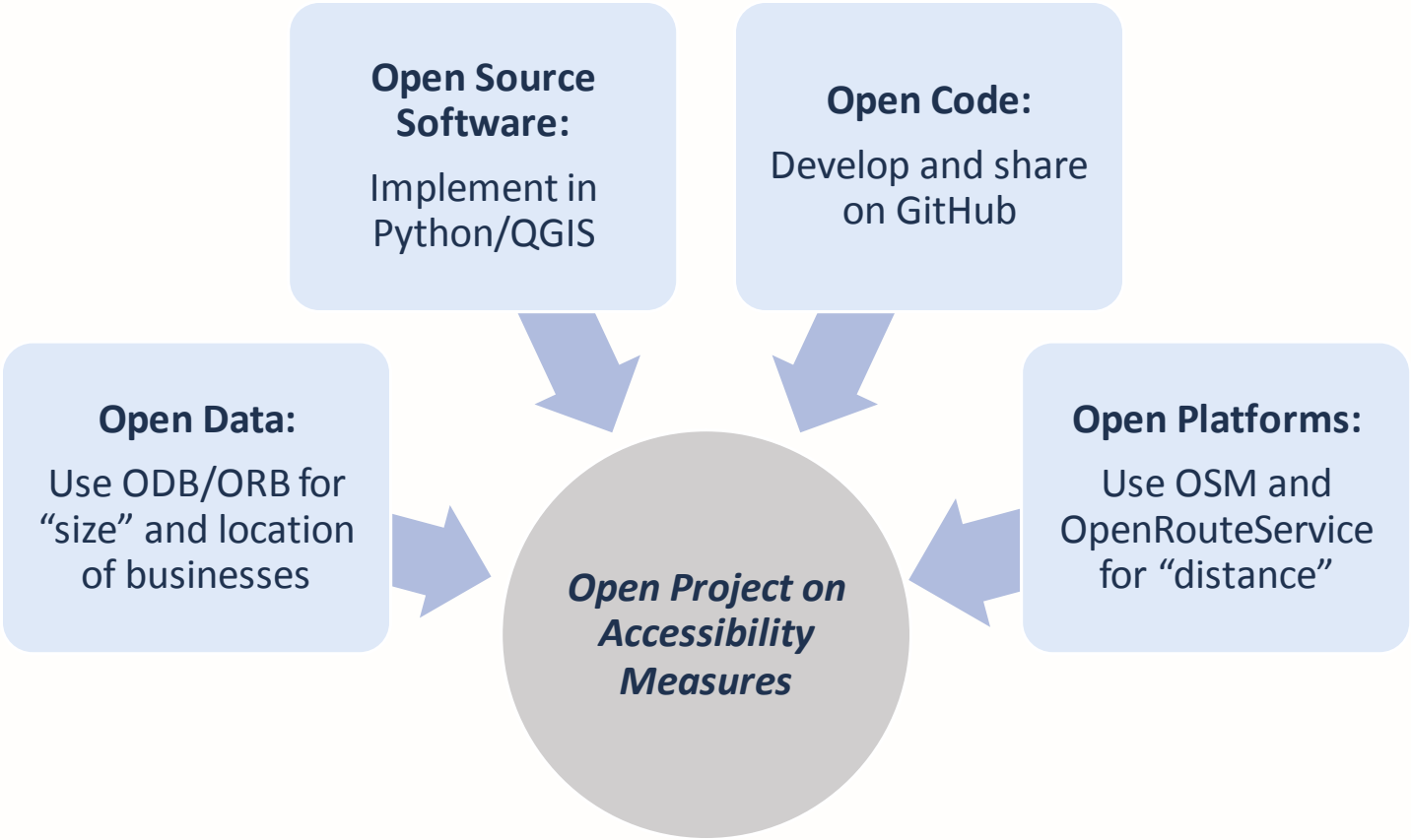
Open Project approach for analysis: Accessibility measures (a dimension of social inclusion)

What can we do with the data?

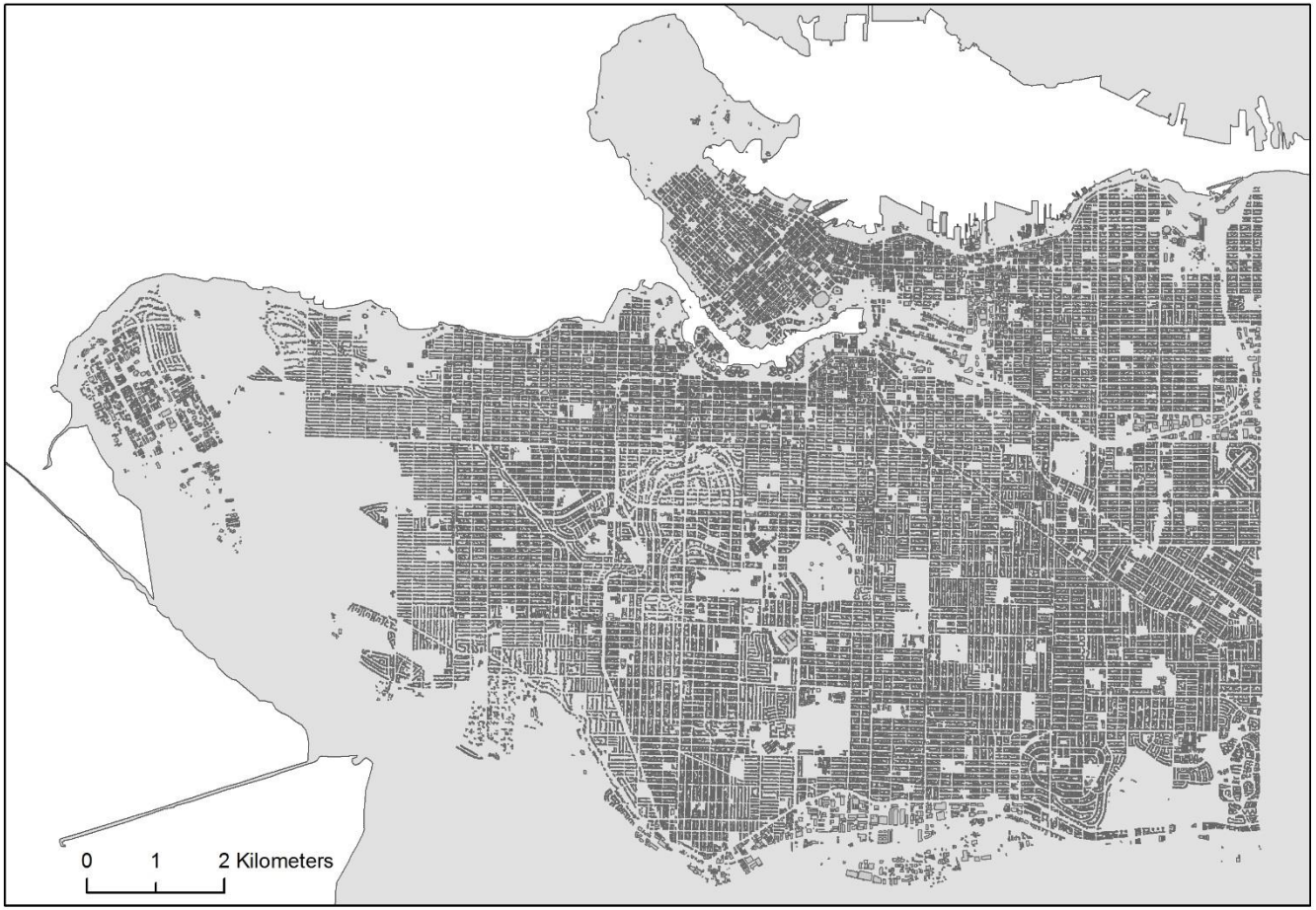
Example – Accessibility Measures: the theory



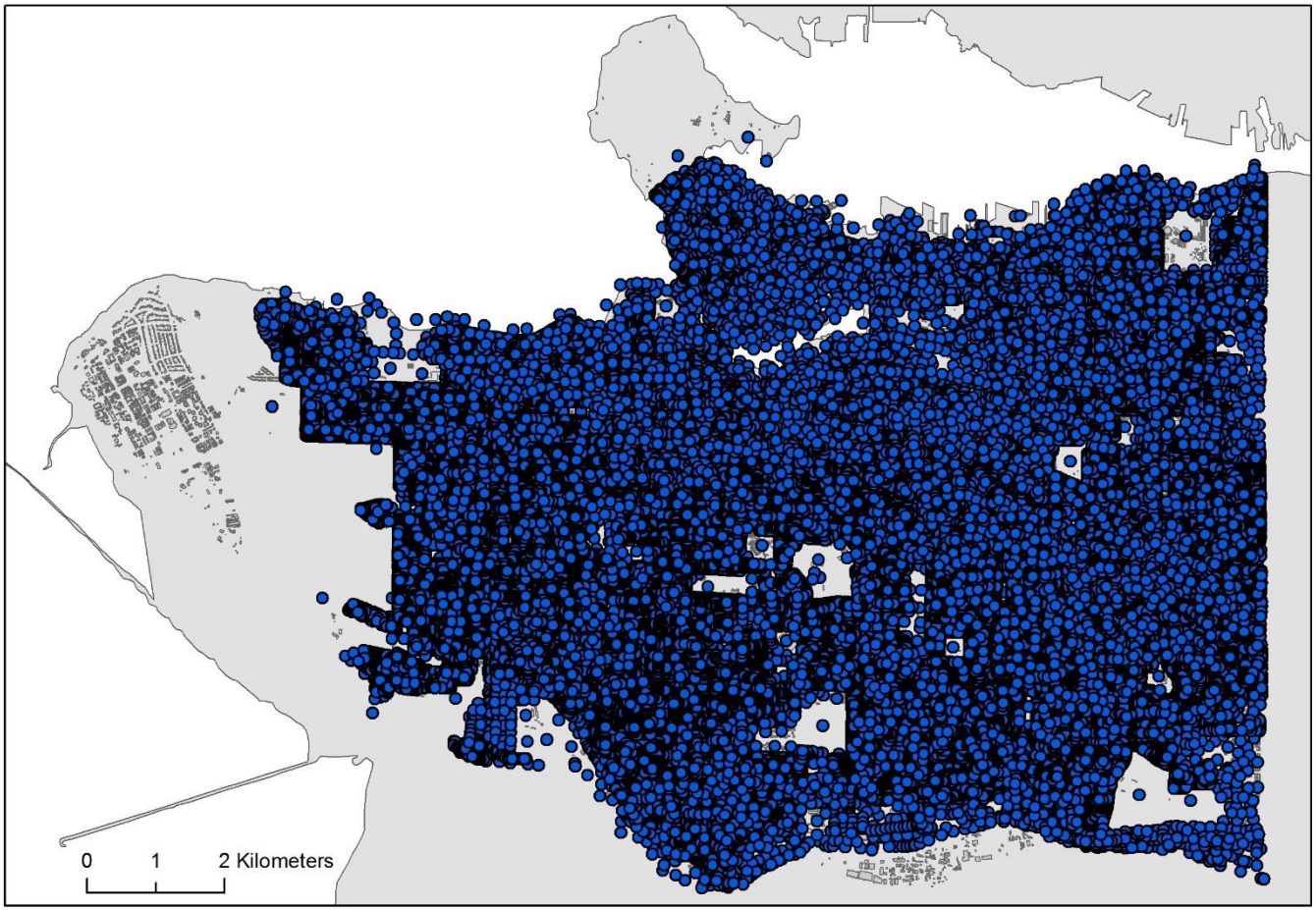
Is it possible with an open project approach?
Example: *Access to grocery stores*



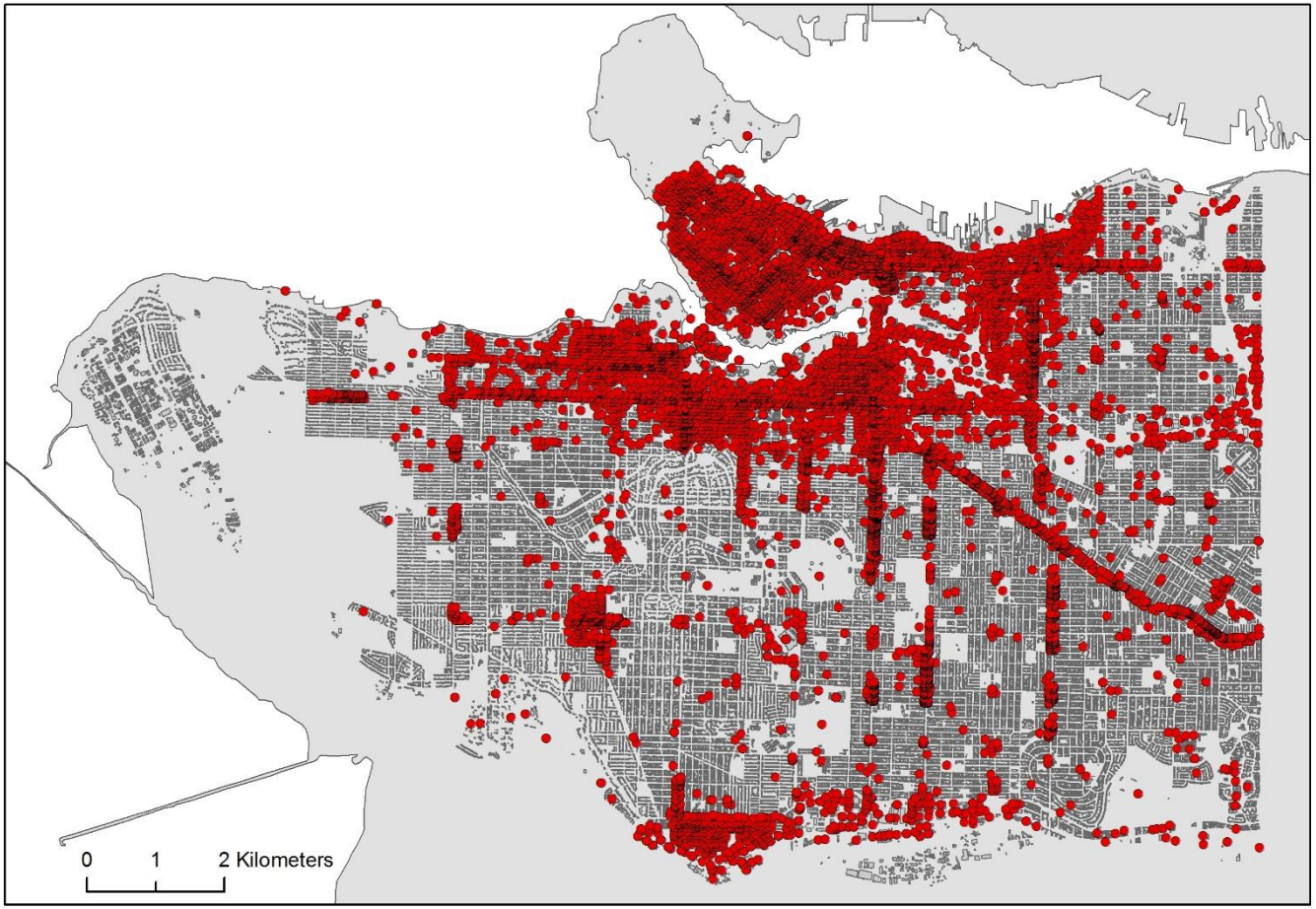
Example of Vancouver open data
buildings



Example of Vancouver open data *addresses*



Example of Vancouver open data *businesses*



Example of Vancouver open data

buildings + *addresses*

Example of Vancouver open data

buildings + *addresses* + *businesses*

Example of Vancouver open data

buildings + addresses + "food" (stars)

Only open data provide this Information with this level of accuracy

1

Crowdsourcing pilot

2

Open Databases

3

Open Project approach

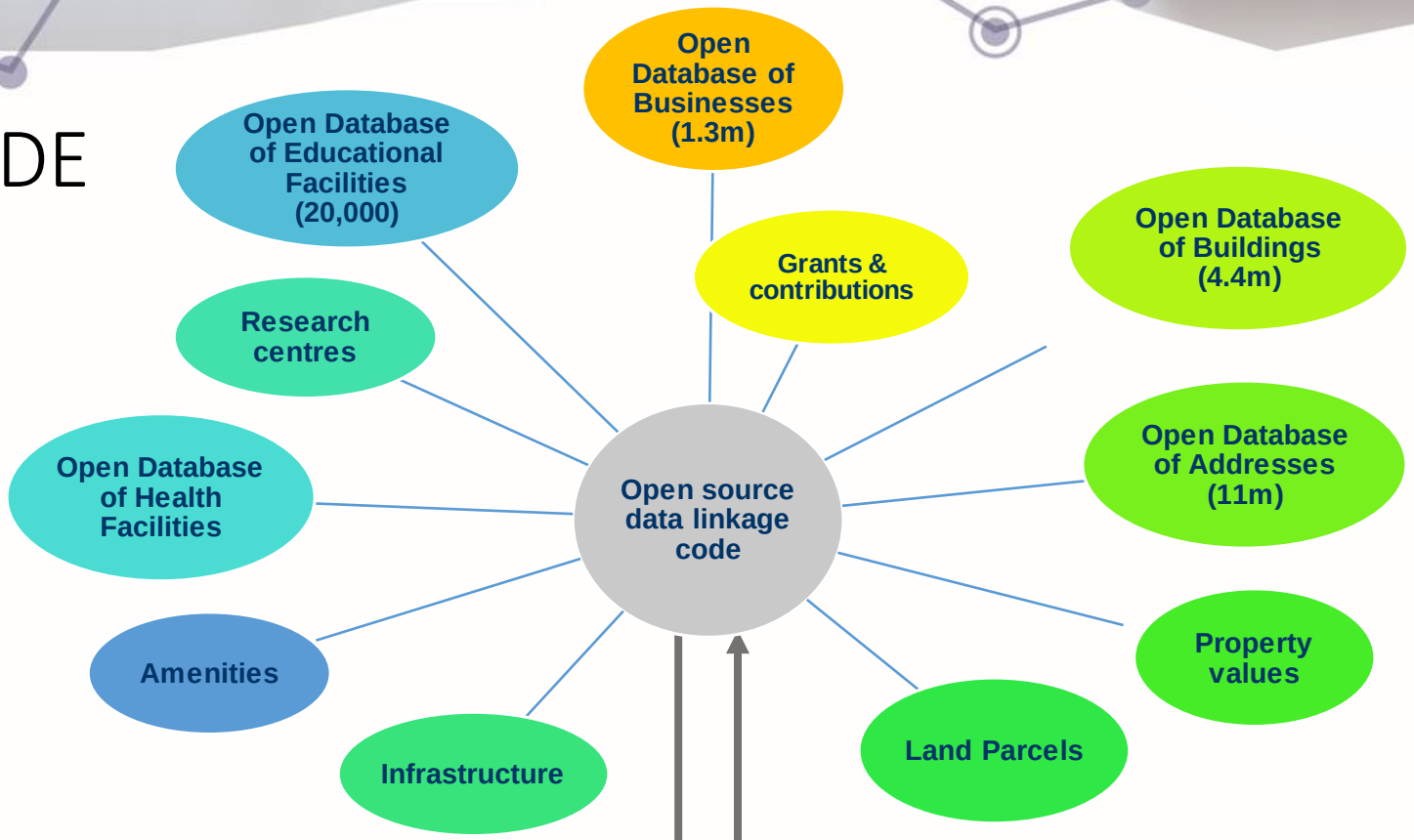
4

Linkable Open Data Environment (LODE)

Linkable Open Data Environment (LODE)

- Open microdata from authoritative sources that have been brought into an environment that is suitable for data linkage
- Vast majority of datasets are from governmental sources (municipal, regional, provincial or federal)
- In development and you can use/contribute to it:
  - https://github.com/CSBP-CPSE/LODE-ECDO

LODE



Statistics Canada microdata holdings (not open data)

Other major open data initiatives

- OpenStreetMap
- LandGIS
- Wikipedia

Where are we and what are the next possible development steps?

An aspirational timetable

	Assessment	Compilation	Cleaning	Dissemination	Expected coverage (%)	Completion of development phase
Buildings (with Microsoft)					100%	Completed
Education facilities					95%	3 months
Addresses (with OpenAddresses)					70-80%	3 months
Businesses					50-60%	6 months
Health facilities					95%	6 months
Public transit (GTFS)					95%	6 months
Parks & recreational					80%	6-12 months
Museum and culture					90%	6-12 months
Arenas & sports					90%	6-12 months
Infrastructure assets					TBD	TBD
Property values					20%	6-12 months

Note: coverage and completion timelines are preliminary estimates or aspirational.

Opportunity: Collaborative data creation

- *Open project* and open data as **enabler and data value multipliers**
- "Open" **reduces barriers** to data and knowledge sharing and costs (administrative, production, management)
- Open microdata provide **unique information** (e.g., building footprints, accurate geolocation of addresses)
- Enriching the open data ecosystems may be one way we can **unlock more data** and more of their value

Challenges

- Expand coverage of open microdata (working with all levels of government and other major open data providers)
- Governance and licensing (enabling common terminology and practices)
- Heterogeneity of formats, concepts, definitions (enabling common data definitions and standards)

Can we work together?

- Collaborate on project development (code, identification of municipal/provincial open data, etc.)
- Analytical tools for open databases (processing, cleaning, linkage)
- Analysis of open databases
- Other?

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