

International Journal of Population Data Science

Journal Website: www.ijpds.org



Swansea University
Prifysgol Abertawe

Semantically interoperable census data: unlocking the semantics of census data using ontologies and linked data

Anderson Wong^{1,*}, Mark Fox¹, and Megan Katsumi¹

Submission History

Submitted:	21/12/2023
Accepted:	21/04/2024
Published:	20/05/2024

¹Urban Data Centre, School of Cities, University of Toronto, Canada

Abstract

The Canadian Census of Population is a survey that collects statistical information on the Canadian population. These censuses contain valuable socioeconomic data that is often used by both the public and private sectors for project planning and decision-making. However, there are a few issues that may arise when using census data. Firstly, data wrangling, which is often a time-consuming process, needs to be conducted in order to clean and prepare the data for integration and use. Secondly, different datasets across different census years may be using different terms to describe the same concept/entity, hence creating a problem of referential equivalence (i.e., how do we know whether two different datasets are referring to the same concepts/entities?). Lastly, the data found in a census is often described using natural language that isn't easily interpreted by machines and can be difficult to break down or deconstruct. In this paper, we develop and propose the use of an ontology for representing the data from the Canadian Census of Population as linked data in order to address the aforementioned issues, evaluate the ontology using competency questions based on real world use cases, and discuss the advantages of census linked data for integration and visualisation uses.

Keywords

linked data; knowledge graphs; census; census data

*Corresponding Author:

Email Address: anderson.wong@mail.utoronto.ca (Anderson Wong)



Introduction

The Canadian Census of Population is a nationally administered questionnaire that is designed to collect socioeconomic data on the Canadian population. This census is mandated by law in the Constitution Act of 1867 to be conducted once every five years [1]. In the 2016 Census of Population, most households received the short form questionnaire that contains basic questions about the age, sex, marital status, etc. of members of a given household while approximately 25% of Canadian households received a long form questionnaire that also includes more detailed questions about education, employment, etc. The data that is collected from the census can be accessed and downloaded from Statistics Canada's official website.¹

For the Canadian government, census data is necessary and valuable as the information is used for a wide variety of applications including calculating federal and provincial transfer payments, setting electoral boundaries, and supporting federal, provincial, and municipal government administration, planning, and policy development [2]. Census data also provides private sector organisations demographic insights into a geographic location of interest which could be useful for gauging consumer and market interest for various products and services.

The usefulness of census data can be greatly enhanced when integrated with other sources of data. For example, combining census data with transportation data, building data, zoning data, health data, etc. leads to significant insights, better policy, and operational decisions, as has been demonstrated by the plethora of data science applications developed over the last 10 years.² Yet it is known that integrating data from multiple sources requires significant effort. Data Wrangling, which includes the identification, cleansing and integration of data, takes approximately 70% of the time of a Data Science project [3]. Our goal is to facilitate the integration of census data by transforming it into semantically interoperable data. Semantic interoperability is the ability to exchange data with unambiguous, shared meaning. In doing so, we have to address two issues: referential equivalence and machine interpretable definitions of census characteristics/indicators.

The first issue is determining *referential equivalence*. How do we know that attribute names or the values of an attribute in two different datasets refer to the same thing? For example, if the name of a neighbourhood (or Ward, Census Tract) in two different datasets are the same or similar (e.g., minor spelling difference), how can we be sure they refer to the same neighbourhood in the same city without access to meta-data that may specify the city covered by the data? Linked Open Data (LOD) standards address referential equivalence directly. Uniform Resource Identifiers (URIs) are unique, global identifiers that point to some specific information. If two different datasets contain the same URI, then they are referring to the same thing. For example, if dataset A contains information on census tract 5350001 and dataset B contains information on census tract 5350001.00, are both datasets referring to same geographic area? If both datasets use the

same URI to represent the census tract, they are referring to the same geographic area.

The second issue is the lack of *machine interpretable characteristic definitions*. In the Canadian Census, a Characteristic, something to be measured in a census tract, is defined using natural language. For example, the characteristic "Total - Occupied private dwellings by structural type of dwelling - 100% data. To support reasoning about the characteristic, the definition would need to be deconstructed into its constituent concepts such "dwelling configuration," "occupied," "structural type" and structure types such as "private," "dwelling," "single-detached," "house," etc. By deconstructing the definition, questions can be asked spanning characteristics that refer to private dwellings or occupied dwellings across multiple characteristics. In other words, it is possible to answer questions that combine data from multiple characteristics, regardless of time and geography.

In the remainder of this paper, we will provide a brief introduction to the Canadian Census, a literature review of census ontologies and linked open data, our methodology and competency questions, the design of the Canadian Census Ontology, and example SPARQL³ queries and data visualisation applications.

Canadian census

The Canadian Census of Population is a survey that collects statistical information on the Canadian population and includes questions about age, sex, marital status, language, income, ethnicity, education level, dwellings, etc. The data from the census is anonymised and aggregated on various geographic levels (Figure 1) such as the national level (i.e., all of Canada), the provincial/territorial level (e.g., Ontario), the census metropolitan area level (e.g., Toronto), and the census tract level (i.e., small geographic areas that usually have a population between 2500 and 8000 persons).

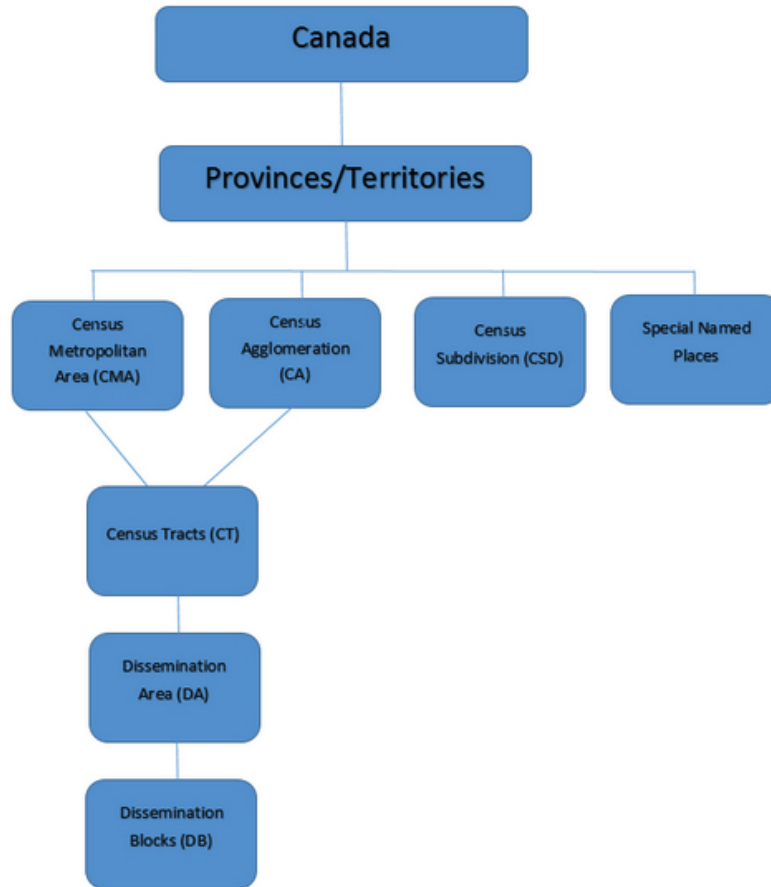
Canadian Census data is published as Census Profiles for various geographic areas. For instance, there are Census Profiles for provinces (e.g., Ontario), cities (e.g., Toronto), and census tracts (e.g., census tract 5350001.00). Each of these Census Profiles consists of characteristics that are individual indicators describing the population that is being surveyed. For example, the "Total private dwellings" characteristic describes the total number of private dwellings that are located within a given geographic area. Some characteristics may also have additional numerical values for describing the male and female populations, in addition to the total population. For instance, the "In low income based on the Low-income measure, after tax (LIM-AT)" characteristic has a value for representing the total number of people who are in low income, a value for the number of males who are in low income, and a value for the number of females who are in low income.

Some characteristics that are found in the Canadian census can have one or more sub-characteristics that provide more specific or detailed information related to the parent characteristic. For instance, there is a characteristic for age groups in a given geographic area (Table 1). It contains separate sub-characteristics for age ranges 0-14, 15-64, 65+.

¹<https://www12.statcan.gc.ca/census-recensement/index-eng.cfm>

²Example case studies can be found at: <http://ontology.eil.utoronto.ca/cem1002/>

³<https://www.w3.org/TR/rdf-sparql-query/>

Figure 1: Canadian Census Geographic Levels (from <https://libguides.tru.ca/censuscanada/censustract>).

The sub-characteristic 0-14 can be further subdivided into sub-characteristics 0-4, 5-9 and 10-14.

Table 1 depicts “Population and Dwellings” and “Age Characteristics” quantities as found in the StatsCan website.⁴

Ajani [4] advocates the standardisation of data using RDF⁵ and OWL,⁶ and the use of SPARQL for querying large datasets in order to address the issues of Volume, Variety, and Velocity associated with Big Data. Using census data as an example, Ajani explains that adapting semantic web technology could help integrate and standardise data from past, present, and future censuses, provide machine-readable metadata for census datasets, and enable convenient querying and extraction of desired data points from large datasets. Ajani believes that these benefits provided by this semantic web approach to distributing and disseminating census data would be helpful to governments and other organisations that use census data to make data-driven decisions.

⁴<https://www12.statcan.gc.ca/census-recensement/2016/dp-pd/prof/details/page.cfm?Lang=E&Geo1=CT&Code1=5350092.00&Geo2=CD&Code2=3520&SearchText=M5R3B2&SearchType=Begins&SearchPR=01&B1=All&TABID=2&type=0>.

⁵Resource Description Framework (RDF) is a standard for representing knowledge graphs in the form of triples: subject, property, value. In the context of Linked Data, the subject and property are defined using URIs. In the case that values refer to other entities (not literals), the value too is a URI. For further information see <https://www.w3.org/RDF/>.

⁶Web Ontology Language (OWL) extends RDF by providing a logical language (i.e., Description Logic) for defining the semantics of both entities (i.e., concepts) and attributes (i.e., properties). Concepts are defined by the logical combination of their properties. For further information see <https://www.w3.org/OWL/>.

To evaluate the quality of an open data system, Tim Berners-Lee [5], developed a five-star rating system for representing the core design principles for LOD. Open data with a one-star rating is defined as any data, in any format, that is available on the web with an open license (e.g., a scanned image of a data table). Open data that is in a machine-readable format (e.g., an Excel spreadsheet) is given an additional star for a total of a two-star rating. Open data can achieve a three-star rating if it additionally does not require the use of proprietary software to access (e.g., using comma-separated values (CSV) instead of Excel). Open data that uses W3C open standards, such as RDF and SPARQL, in addition to the above, get a four-star rating. Lastly, open data with a five-star rating contains all of the above while being linked to other LOD sources such as GeoNames or DBpedia.

Currently, the census data that is published on Statistics Canada’s website would be rated three stars under Tim Berners-Lee’s LOD rating system as it is available on the web with an open license and is also published in CSV format which is a non-proprietary format. From a referential equivalence perspective, StatsCan employs a Dissemination Geography Unique Identifier (DGUID) to identify geographic areas in the census datasets. The DGUID does not conform to URI standards, but could form a part of a URI. Similarly, census questions are numbered, but do not conform to URI standards.

In terms of machine-interpretable definitions, the CSV file lacks many things, including:

Table 1: Canadian census example table

Characteristic	5350092.00 Ontario [Census tract]			Toronto, CDR Ontario [Census division]		
	Data quality Map Change geography			Data quality Map Change geography		
	Total	Male	Female	Total	Male	Female
	Counts (unless otherwise specified)					
Population and dwellings						
Population, 2016 1	7,278	xxx	xxx	2,731,571	xxx	xxx
Population, 2011 1	7,151	xxx	xxx	2,615,060	xxx	xxx
Population percentage change, 2011 to 2016	1.8	xxx	xxx	4.5	xxx	xxx
Total private dwellings 2	4,113	xxx	xxx	1,179,057	xxx	xxx
Private dwellings occupied by usual residents 3	3,718	xxx	xxx	1,112,929	xxx	xxx
Population density per square kilometre	11,576.3	xxx	xxx	4,334.4	xxx	xxx
Land area in square kilometres	0.63	xxx	xxx	630.20	xxx	xxx
Age characteristics						
Total - Age groups and average age of the population - 100% data 4	7,280	3,530	3,750	2,731,570	1,313,590	1,417,985
0 to 14 years	605	305	305	398,135	204,190	193,945
0 to 4 years	240	115	125	136,000	69,895	66,105
5 to 9 years	220	110	105	135,025	69,350	65,680
10 to 14 years	150	80	70	127,110	64,945	62,165

- The sub-characteristic hierarchy is represented by the number of spaces in front of the Characteristic title in the Characteristics column. No explicit definition of the hierarchy exists in the CSV file.
- The units of measure of a cell value are not defined. Only by reading the definition of a characteristic can this be inferred. For example, is it a monetary value, population count, ratio?
- Information as to the provenance of a value is also missing, but can be inferred from the context. For example, from the context we know it is from the 2016 census long form for a particular census tract.
- A machine interpretable description of the characteristic being measured is missing. For example, if we wanted to find information about people in the age range of 5-9 years, we could do a string search in the CSV, but we would not be able to know the context of this particular characteristic other than the text description.

Review of census ontologies and linked open data

Ontologies and linked data can help enable the integration of census data with other data sources, allow for complex querying and visualisations, and improve the accessibility of the data [6]. For these reasons, there have been a number of efforts toward publishing census data as Linked Open Data (LOD).

Census data is often used with other data sources for analytical and statistical research. For instance, census data combined with energy use data can be used to estimate energy access in different regions of a country in order to track progress towards the United Nations Sustainable Development Goal 7 [7]. Combining census data with mortality rates

across socio-economic classes could elucidate possible trends in health and social inequalities between individuals of different socio-economic status [8]. Data from child hospitalisations due to abuse can also be combined with census data to estimate the incidence rate of physical abuse among children of different ages [9]. However, as censuses are generally conducted only once every 5 or 10 years [10], the time periods for the census data may be different from the time periods for the other data sources of interest. Oftentimes intercensal estimations are used to approximate data values for years between censuses, as seen in the three examples mentioned above. National statistical agencies may publish their own intercensal/postcensal estimates but other times, researchers may have to conduct their own estimates for their populations of interest. Assuming a constant rate of change between census years (linear interpolation) may be sufficient for approximating an intercensal data point although these estimates may be inaccurate when populations experience sudden changes (e.g., natural disasters) [11] or when the population of interest is too small [12].

Intercensal estimates may be adequate when combined with an additional dataset from a different time period. But when several datasets, each from different time periods, are being combined, then an approach unique to the data being combined, is required. In such a situation it is important to document the assumptions made about the method and validity of the inter-temporal estimates.

Statistics Canada provides a Linkable Open Data Environment (LODE) for public access of municipal, provincial, and federal data [13]. "Linkable" in the context of Statistics Canada's LODE refers to how the databases in the LODE are published under a single open data license and the data can all be processed and visualised using the same set of open-source tools⁷ provided by Statistics Canada. This does not mean that the data in the LODE is published as Linked Open Data. Currently, Statistics Canada's

⁷<https://www.statcan.gc.ca/en/lode/tools>

LODE consists of six open databases: The Open Database of Buildings (ODB), The Open Database of Educational Facilities (ODEF), The Open Database of Healthcare Facilities (ODHF), The Open Database of Cultural and Art Facilities (ODCAF), The Open Database of Addresses (ODA), and The Open Database of Recreational and Sport Facilities (ODRSF). LODE datasets use libpostal, a natural language processing solution, to separate address information into its individual components (e.g., street name, street number, city name, etc.) and addresses that did not have geocoordinates are geocoded using geocoders like ESRI ArcGIS Online (AGOL) and OpenStreetMap Nominatim geocoder. The ODB is published in Geographic Information System (GIS) format while the other five databases are published in CSV format. Although Statistics Canada has invested significant efforts into geocoding their data, they currently do not have a formal ontology for the semantic representation of geospatial information. Statistics Canada's LODE also includes their LODE Viewer tool which uses information from the above databases to provide the locations of healthcare facilities, cultural/art facilities, educational facilities, and recreational/sport facilities along with a map overlay of building footprints. Although Statistics Canada claims that they aim to harmonise data from different sources, their LODE initiative does not incorporate census data at this time [13]. Furthermore, Statistics Canada does not currently provide an ontology for representing the statistical data nor URIs in their LODE.

While Statistics Canada's LODE is an important initiative towards publishing and disseminating open government data, it is not "linked" in a semantic web context as entities, attributes, and values (where appropriate) are not specified using URIs, nor are the textual definitions machine interpretable, and they are not published using semantic web standards such as RDF and OWL.

Bukhari and Baker [14] provided a SPARQL endpoint for querying purposes in an effort to encourage the use of new and up-to-date data for critical, health-related decision-making. For this project, Bukhari and Baker opted to use D2RQ,⁸ a system for accessing relational databases as RDF graphs, to transform the data from the Canadian health census into LOD. With this approach, the authors provided a semantic mapping of Cancer Survival Patients by mapping the entities in the dataset to analogous concepts from well-known semantic ontologies using D2RQ. For example, the authors used the foaf:age property to represent the age of cancer survival patients as an integer and the foaf:gender property to represent the patients' gender as a string, as seen in their semantic mapping (Figure 2). Bukhari and Baker explain that publishing the data from the Canadian health census as LOD could help enhance timely decision-making processes and optimise the provision of health services by identifying geographic regions that experience higher than average cases of treatable-disease deaths and could also be used to correlate data across datasets (e.g., finding a negative correlation between breast feeding and breast cancer).

Census data is often published as aggregated data across geographic areas of interest in order to protect the privacy of individuals. Nevertheless, some census organisations also

provide census microdata (i.e., data on the individual level) for academic uses. This data is usually closed and available with restrictions. For example, Statistics Canada's Research Data Centres⁹ provide access to microdata only within their physical centre (i.e., no online access), and all analyses must be performed on their computers. Only the results of the analysis may be removed from the facility. As such, most of the existing academic work is focused on using aggregated census data while other academic projects might focus on census microdata instead. For example, the IPUMS project has been collecting and distributing census microdata for over 100 countries across 547 censuses and surveys for qualified researchers to use [15]. In the following, we review existing approaches to representing both aggregated census data and microdata as LOD.

Petrou et al. [16], defined a framework (Figure 3) for publishing the 2011 Greek aggregated census data as LOD, using the Resource Description Framework (RDF) data model, in order to improve the accessibility of statistical information. In their project, they mapped the Greek census data, which contained demographic data about the Greece population (e.g., sex, marital status, education level, etc.) and data about households and dwellings (e.g., building type, number of bedrooms, etc.), onto the Data Cube Vocabulary.¹⁰ To do this, the dataset as a whole was mapped to the qb:DataSet class and the columns in the dataset were mapped to the appropriate concepts in the Data Cube schema. They identified the column with the regional divisions (which are represented using geocodes) to be their dimension and accordingly mapped this concept to the qb:DimensionProperty class. The population column was identified as the measure, so this column was mapped to the qb:MeasureProperty class. Lastly, the unit of measurement (number of habitants, in this case) was mapped to the qb:AttributeProperty class. This approach allowed them to link an indicator (e.g., population of geocode 0102) to its numerical value (e.g., 16577), its geocode (e.g., 0102), and the dataset that the indicator belongs to (e.g., Permanent residence population census 2011).

Similarly, Aracri et al. [17] also worked on a similar project where they provided an approach for publishing the 15th Italian Population and Housing Census using LOD as a part of the Istat's (Italian National Institute of Statistics) Linked Data dissemination strategy. Like Petrou et al. [16], they were also interested in representing aggregated census data as LOD and also adopted the RDF Data Cube Vocabulary for translating census data to LOD. With the Data Cube Vocabulary, Aracri et al. provided a framework for representing a census indicator (e.g., "number of resident male age between 45 and 49 years") and linking it to its corresponding dataset, census section (i.e., location), census year, indicator value, etc. (Figure 4). However, unlike Petrou et al., they map the census indicators to qb:Observation while using their own attributes to describe the indicator. These attributes include census:Year (to describe the year of the census indicator), census:Sex (to describe the sex of the population that is being represented), census:AgeClass (to describe the age range of the population that is being represented), census:Citizenship (to describe the citizenship status of the population that is being represented),

⁹<https://www.statcan.gc.ca/en/microdata/data-centres>.

¹⁰<https://www.w3.org/TR/vocab-data-cube/>.

⁸<http://d2rq.org/>.

Figure 2: Semantic mapping of cancer survival patients' data from Bukhari and Baker (2013)

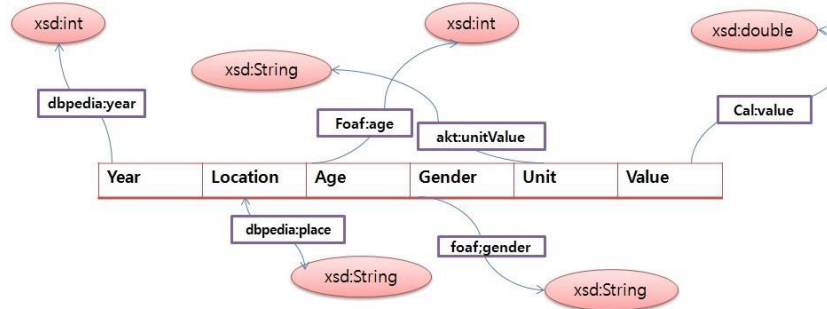


Figure 3: RDF representation for the population of a division from Petrou et al. (2013)

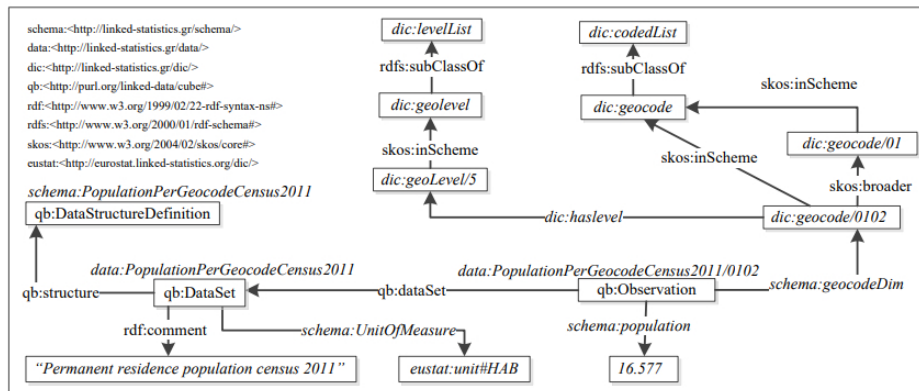
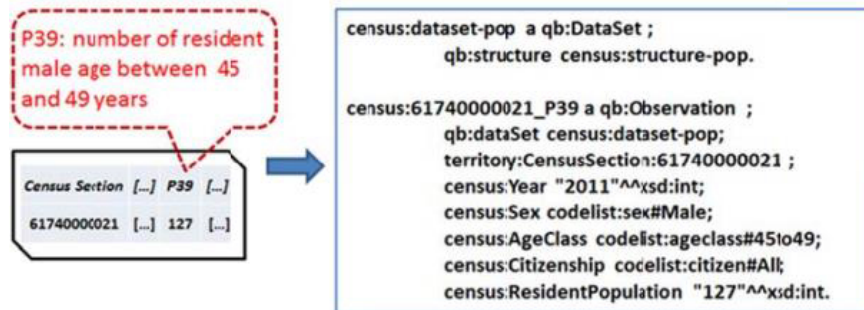


Figure 4: Example data cube observation from Aracri et al. (2014)



census:ResidentPopulation (to describe the numerical value of the indicator), and territory:CensusSection (to describe the geographic area that the indicator represents). With these attributes, Aracri et al. were able to represent and describe Italian census population data according to territory, sex, age class, etc. while linking each indicator back to the census dataset that they can be found in. Their ontological representation does not deconstruct and represent the definition of the indicator nor capture the units of measure for the indicator (i.e., population counts in this case).

To extend the Istat's LOD initiative, Aracri et al. [18] created an ontology for modeling data from the Istat's Base Statistical Registers (BSRs) which include Persons, Families, and Cohabitations. Unlike the previous two projects that were focused on the representation of aggregated census data, they developed a microdata approach to represent BSR data as it contains information about individuals. Their ontology is expressed in OWL2 and provides a standard for

representing the relationships between people (e.g., parentOf, son/daughterOf) and their residences. The main classes in this ontology are Person (represents a single person), Family (represents a group of persons bound by marriage, kinship, etc.), Nuclear Family (represents a group of persons forming a couple relationship or parent-child type), and Cohabitation (represents a group of persons who live together, without being bound by marriage, kinship, etc.). Aracri et al. expressed that they aim to integrate this ontology with the census ontology outlined in the Aracri et al. [17] report in the future in order to effectively integrate and disseminate Istat data from multiple different sources.

Fernández et al. [6] provide another example of modeling microdata as they developed an RDF scheme for representing the microdata from the 2001 Spanish Census. Using this approach, Fernández et al. were able to represent the characteristics of an individual (e.g., birth year, birthplace), familial relations between individuals in a family nucleus (e.g.,

Figure 5: Spanish Census Data Model from Fernández et al. (2011)

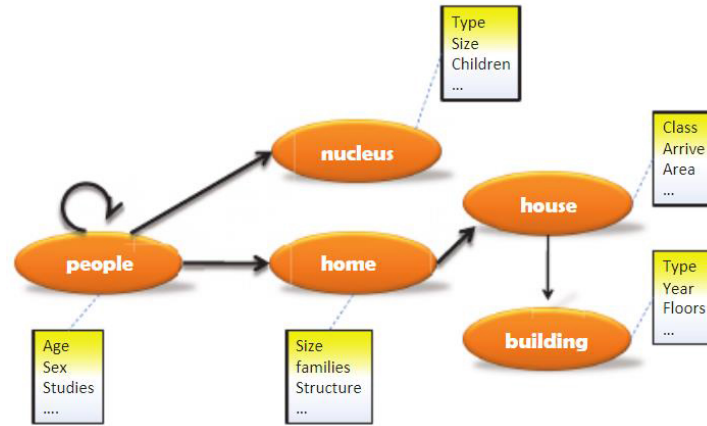
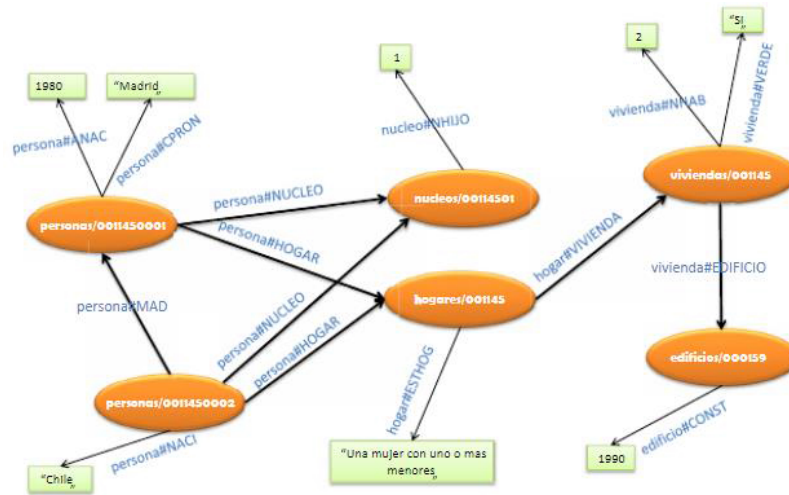


Figure 6: Instances using the Spanish Data Model from Fernández et al. (2011)



mother-child relationship), and identify the home/building that these individuals lived in (Figures 5, 6). To demonstrate the practicality of their work, the authors also created SPARQL queries that aggregate the microdata to show the total number of Spanish and foreign people per age and visualised the results as graphs and figures using the Google Visualisation API.

This review demonstrates that there is an increasing need and interest in representing and publishing census data using ontologies and LOD. While there have been several different approaches to modeling census data in European countries such as Greece, Italy, and Spain, there is little existing work on modeling the data found in the Canadian Census of Population.

There are four notable issues with translating Canadian census data. Firstly, there is a lack of unique identifiers for census characteristics at aggregate and tract level and there needs to be a unique name for each characteristic in order to ensure that each characteristic has its own unique URI.¹¹ Secondly, there is a lack of a hierarchical representation of spatial areas and there needs to be some consistent way of representing the relations between geographic areas. Thirdly,

¹¹Taylor [19] defines a method (p. 17) for constructing unique names for Canadian census characteristics that can be used to construct a URI.

there is a lack of a semantic representation of the definitions of census characteristics. Finally, there needs to be a standardised method for representing these characteristics on the semantic web. The existing census ontologies do not provide a simple way to aggregate census data from smaller geographic regions (e.g., census tracts) into data for larger geographic areas of interest (e.g., city, neighbourhood, ward). This is particularly important as governments and other organisations may need to make decisions and plan their operational work on a city/neighbourhood/ward level instead of the census tract level. The third issue has also not been addressed by existing work, namely how to represent the definition/meaning of a characteristic/indicator. The aim of this work is to fill the gap in existing literature in order to address these issues and translate the information in the Census of Population into RDF format and publish Canadian census data as ontology-based LOD.

Methodology and competency questions

We use the methodology defined in Grüninger & Fox [20] to engineer the Canadian Census Ontology. The process begins

by defining a set of usage scenarios. Based on the scenarios, we identify a set of competency questions that the ontology must answer. These are the requirements for what is to be represented and the deductions to be performed. We analyse these questions to determine groupings (categories), frequency of occurrence, and importance to the city's operation. The usage scenarios and competency questions are derived from these.

The next step of the process is to review existing ontologies to ascertain the extent they satisfy some or all of the competency questions. Relevant concepts and properties are candidates for inclusion in the final ontology. Next, the terminology (i.e., concepts and properties) required to answer the competency questions are defined. The semantics of the terminology are defined by constructing a set of axioms that define and/or constrain their interpretation. The axioms are important as they precisely define the terms, and can determine whether the data that underlies the terms are consistent. As part of this step, we identify and create ontology modules (aka microtheories¹²) that are foundational to the operations of a city. These microtheories form the building blocks for more specific city data. For example, many city applications rely upon the recipient of a service to be a resident of the city. What does it mean to be a resident? What are the necessary and sufficient conditions for residency? What are the core concepts and properties that should be used to construct a theory of residency? How are these concepts and properties defined (i.e., axiomatised)? Residency is just one of many other foundational ontologies that are expected to emerge during our research.

Finally, the ontology is evaluated based on the extent to which the ontology is able to answer the competency questions. The evaluation of the axioms with respect to the competency questions is supported through the use of automated theorem provers, following the work of Katsumi and Gruninger [22]. We can therefore claim and formally verify the correctness and completeness of an ontology with respect to its requirements.

In order to develop an ontology that is capable of representing census data in a useful and meaningful way, we must first understand how census data might be used by governments or other organisations. Census data can be used for a wide variety of academic, commercial, or governmental purposes. For instance, census data is often used by government organisations for evaluating and improving their social service programs [2]. As managing and improving community services is a crucial part of the government's operations, we have identified the following use cases and competency questions in the context of improving local neighbourhoods and their social services for the municipal government of Toronto.

In Toronto, the municipal government uses multiple sources of data (including census data) to evaluate the social and economic well-being of its neighbourhoods. In 2020, the Toronto municipal government had identified 33 Neighbourhood Improvement Areas (NIAs) which are vulnerable districts within the City that require additional social and economic support. To elaborate, the City uses

the Urban Health Equity Assessment and Response Tool (HEART) to score neighbourhoods under five domains of neighbourhood well-being [23] and NIAs are the neighbourhoods with scores that fall below the designated Neighbourhood Equity Benchmark. These five domains are Economic Opportunities, Social Development, Participation in Decision Making, Healthy Lives, and Physical Surroundings. Economic Opportunities refers to the economic status of the neighbourhood and the income levels and job opportunities that its residents have access to. Social Development refers to providing neighbourhood residents with access to social, recreational, and cultural services as well as resources for residents to improve their education and literacy. Participation in Decision Making refers to providing residents with opportunities to get involved with local decisions such as voting in elections. Healthy Lives refers to the physical and mental health of neighbourhood residents and their access to medical care resources. Lastly, Physical Surroundings refers to the natural and built environment of the neighbourhood, including public spaces, transportation infrastructure, and air quality.

After the NIAs are identified by the Urban HEART @Toronto Project Team, Neighbourhood Planning Tables are formed to represent and manage the social programs of the NIAs. These Neighbourhood Planning Tables consist of local residents, businesses, community agencies, City Councilors, and City staff who coordinate and cooperate to develop Neighbourhood Action Plans that are best suited for the social and economic conditions specific to each NIA. Data from the Canadian census could be used to help these Neighbourhood Planning Tables identify high priority issues in the NIA and help them design and deploy social programs that better meet the needs of the neighbourhood's demographics. In the following, we formulate competency questions (CQ) for retrieving census data relevant to the HEART domains that could be useful for the Neighbourhood Planning Tables' planning and implementation of action items. These competency questions can also help us evaluate the capability and usefulness of the Canadian Census Ontology in the example context of improving and managing NIAs in Toronto.

- **Domain:** Economic Opportunities

- **Indicator:** Unemployment

- **CQ1:** Which NIAs have the highest number of unemployed residents?

- This information can help identify NIAs that might be having difficulty with providing and promoting employment opportunities

- **Indicator:** Occupation

- **CQ2:** What industries do NIA residents work in?

- This information can help municipal governments identify the professional development resources that are most relevant to the NIA

¹²Microtheories, as defined by Lenat & Guha [21] (1991), define a context for reasoning, including facts and axioms.

- For example, a NIA where many residents work in the manufacturing sector might benefit from IPC¹³ training and certification

○ **Indicator:** Low-Income

- **CQ3:** Which NIAs have the most low-income residents?
 - This information can help the City identify the NIAs that may need more funding for services targeted towards low-income individuals

○ **Indicator:** Shelter Cost

- **CQ4:** Which NIAs have the most residents with a high shelter-cost-to-income ratio (30+%)?
 - This information can help the City identify NIAs that may need more funding for subsidised housing (e.g., Rent-Geared-to-Income Subsidy)

● **Domain:** Social Development

○ **Indicator:** Postsecondary Completion

- **CQ5:** Which NIAs have the most residents who do not have a postsecondary degree nor high school diploma?
 - This information can help identify NIAs that may need more education opportunities and connections

○ **Indicator:** Knowledge of Official Languages

- **CQ6:** Which NIAs have the most residents who know neither English or French?
 - This information can help identify NIAs that may need access to literacy training programs (e.g., ESL courses)

○ **Indicator:** Visible Minority Population

- **CQ7:** What are the largest visible minority groups in NIAs?
 - This information can help provide cultural insights about the NIAs which can be useful for developing recreational and cultural programs that meet the needs and wants of the residents

● **Domain:** Healthy Lives

○ **Indicator:** Senior Population

- **CQ8:** Which NIAs have the largest senior populations?
 - This information can help identify NIAs that may require more senior health services in the future

○ **Indicator:** Child Population

- **CQ9:** Which NIAs have the largest child populations?
 - This information can help identify NIAs that may require more child health services in the future

● **Domain:** Physical Surroundings

○ **Indicator:** Commuting Duration

- **CQ10:** Which NIAs have the longest commute times?
 - This information can help identify NIAs that may benefit from improved transportation infrastructure

○ **Indicator:** Public Transit Use

- **CQ11:** Which NIAs use public transit as the main mode of commuting?
 - This information can help identify NIAs that may need more public transit investment in the future

○ **Indicator:** Housing Suitability

- **CQ12:** Which NIAs have the most households living in unsuitable dwellings (according to NOS¹⁴)?
 - This information can help identify NIAs that may be experiencing problems with overcrowding

○ **Indicator:** Dwelling Age

- **CQ13:** Which NIAs have the highest number old dwellings?
 - This information can help identify NIAs that may have building safety issues (e.g., asbestos and lead paint in older dwellings constructed before 1970)

Additionally, NIAs may need to analyse and manage geographic areas that partially overlap with multiple census tract areas. As such, they may need to aggregate census data in a geographic area that cannot be simply expressed as a sum of census tracts. In these cases, it would be useful to know how much the designated area overlaps with the census tract boundaries in order to calculate a weighted average for the desired census data.

- **CQ14:** How much of a given geometry overlaps with the census tract boundaries?

Canadian census ontology design

This section defines an ontology representing census data and related information. Ontologies allow us to model information by providing a formal way to define and represent entities,

¹³<https://www.ipc.org/ipc-certifications>.

¹⁴<https://www23.statcan.gc.ca/imdb/p3Var.pl?Function=DEC&Id=100731>.

their attributes, and the relations among entities. Ontologies are composed of two main elements: classes and properties. A Class defines, by means of a formal, logical language, a concept in relation to other concepts in the domain. Classes can also have zero or more properties that are used to link classes to other classes (i.e., object properties) or literal values (i.e., data properties). Ontologies also include axioms that are used to define classes as logical combinations of their properties. In addition, ontologies may contain instances. Instances that satisfy the definition of the class may be inferred to be members of that class, and conversely members of the class are required to satisfy the definition of the class. For example, a CensusTract class would be the class that represents the set of all individual census tracts (e.g., census tract 5350001.00). While there are multiple different languages for expressing ontologies, the Canadian Census Ontology described in this paper is defined using Web Ontology Language (OWL 2) is detailed in this section and can be accessed using the following URL: <https://github.com/EnterpriseIntegrationLab/CKGN/tree/main/UniversityOfToronto/Census/Ontologies>. Furthermore, a table of key classes and properties can be found in Appendix I.

City data ontologies and standards

The Canadian Census ontology is defined in terms of ISO/IEC 21972 [24] that is an ontology-based standard for city indicators, and the ISO/IEC 5087 series [25–27] of ontology-based standards for the representation of City Data.

ISO/IEC 21972:2020¹⁵ is a machine-readable data standard for the representation and exchange of indicators/metrics. It provides design patterns representing indicators, indicator definitions, and their meta-data. The design patterns define concepts and properties for indicator types, indicator constituents that comprise its definition, units of measure, statistics, and populations.

Figure 7 depicts a difference indicator pattern that is defined as the difference of two terms, each measuring a statistic (in this case size) of a Population. Membership in a population is defined by a class [28]. In this example, the indicator takes the difference between the mean number of skills homeless youth have before and after participating in some skill training activity. Classes and properties from ISO/IEC 21972 are coloured in turquoise. ISO/IEC 21972 is based on the Global City Indicator ontology [28–30].

ISO/IEC 21972 addresses the question of how to represent the definition of populations, which lies at the heart of representing the definitions of indicators. This question is unusual in the sense that statistics is directly concerned with the definition of populations, but is essentially silent on the representation of population definitions from a data modelling perspective. In Figure 7, “Homeless youth Group – post intervention” is defined to be a subclass of i72:Population. The Population is defined by three types of information (Figure 8):

1. Membership Extent: The i72:defines_by property specifies a class that provides a prototypical description

¹⁵ISO/IEC 21972 Information technology — Upper level ontology for smart city indicators (ISO/IEC21972, 2020). <https://www.iso.org/standard/72325.html>.

that a member of the Population must satisfy. In Figure 8 the default value is owl:Thing. For a specific population, such as in Figure 7, the value is a class that defines a homeless youth.

2. Spatial Extent: the i72:located_in property specifies the physical area from which the population is drawn. The default value is geo:Feature which is the top class for the Geonames ontology.¹⁶
3. Temporal Extent: The i72:for_time_interval property specifies the time period over which the population is drawn. The default value is a TimeInterval defined in the OWL-Time ontology.¹⁷

The goal of ISO/IEC 5087 is to enable semantic interoperability by identifying and formalising the concepts that are shared across cities. To motivate the need for a standard city data model, consider the evolution of cities. Cities deliver physical and social services that traditionally have operated as silos. If during the process of becoming smarter, transportation, social services, utilities, etc. were to develop their own data models, then we would have smarter silos. To create truly smart cities data must be shared across these silos, which can only be accomplished through the use of a common data model. For example, “Household” is a concept that is commonly used by city services. Members of Households are the source of transportation, housing, education, and recreation demand. It represents who occupies a home, age, occupations, where they work, abilities, etc. Though each city service may gather and/or use different data about a Household, much of the data needs to be shared with each other. The concepts fall into two categories:

1. concepts whose instances are both produced (i.e., instantiated) and consumed (i.e., used) across multiple city services (e.g., Household, Service, Resident), and
2. concepts that are produced by one city service (e.g., transportation) but used by other city services (e.g., Vehicle, Transportation network).

This standard provides definitions in a machine-readable form using the Semantic Web Ontology language OWL. This enables the development of software tools that can consume, verify, and make inferences about city data. It ensures that data can be easily combined and shared in order to effectively support city planning and operations.

The city data model is stratified into three levels of abstraction. The Foundation Level covers very general concepts such as Time, Location, and Activity. The City Level covers concepts that are general to cities and span most services such as Households, Services, and Residents. The Service Level spans concepts commonly associated with a particular service but still shared with other services, such as Housing, Vehicles and Transportation network. Figure 9 depicts the three levels and the requirements for a concept’s inclusion at a particular level.

Figure 10 depicts the concept patterns in each of the 4 standards and proposed standards in the ISO/IEC 5087 city data standards. ISO/IEC 5087 is based in part on the iCity

¹⁶See <https://www.geonames.org/ontology/documentation.html>.

¹⁷See <https://www.w3.org/TR/owl-time/>.

Figure 7: ISO/IEC 21972 depiction of the definition of indicator “Average number of skills each homeless youth gained”

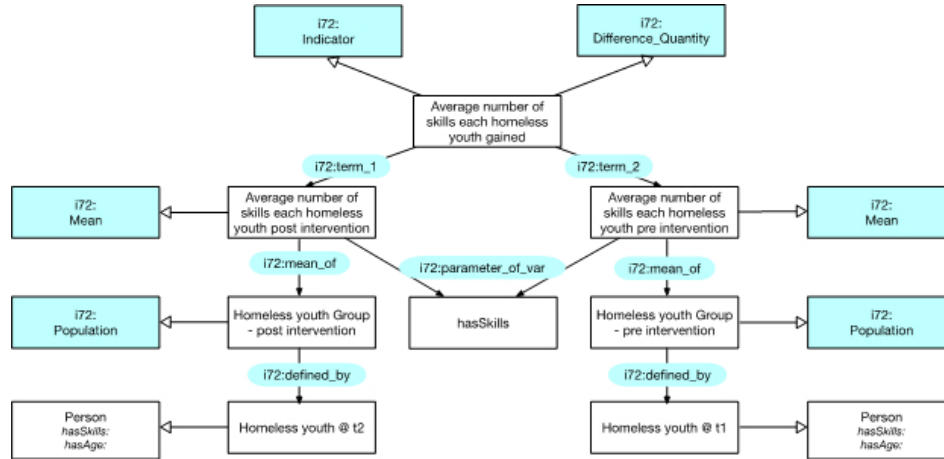


Figure 8: Population class properties

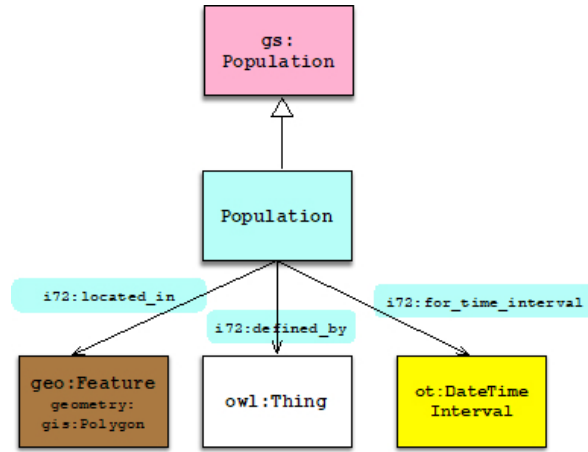
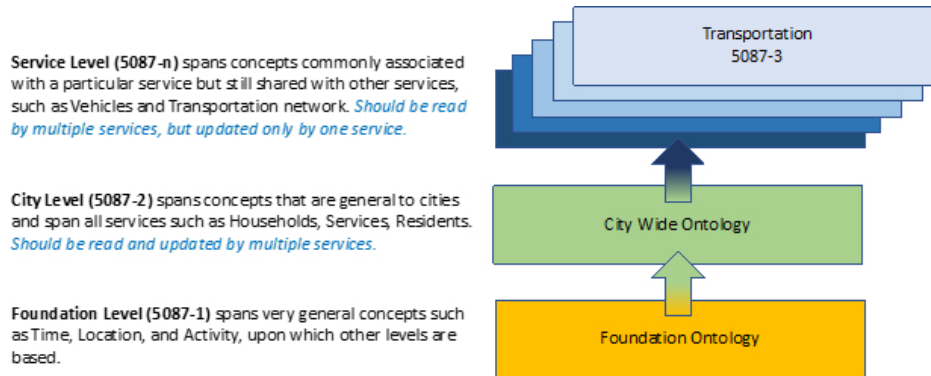


Figure 9: ISO/IEC 5087 stratification framework



Transportation ontology [31, 32] and the Global City Indicator ontologies [33].

Classes and properties from these standards are used to represent the definitions of Canadian Census characteristics.

Administrative areas

In this section we define the classes for representing the administrative areas of a city, and how they relate to census tracts. These administrative areas often contain more

than one census tract whose data can be aggregated, and different types of administrative areas can overlap spatially. For example, the boundary of a Ward may overlap with multiple neighbourhoods. The same may be true of other administrative areas such as school districts, health districts, police, and fire station areas of service.

As seen in Figure 11, each class can have instances that represent the individuals such as specific census tracts (e.g., ct-5350001-00), neighbourhoods (e.g., neighbourhood70), wards (e.g., ward14), and Canadian cities (e.g., Toronto).

Figure 10: ISO/IEC concept patterns

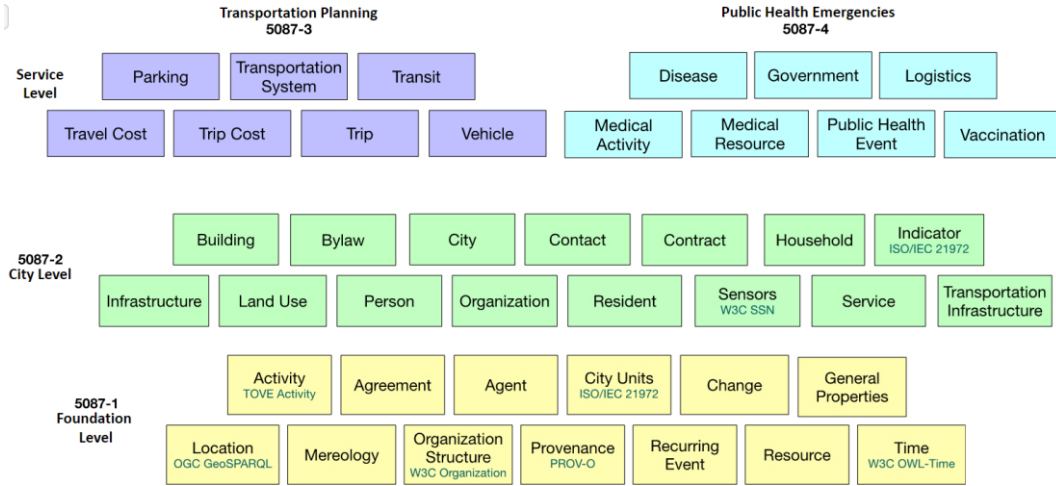
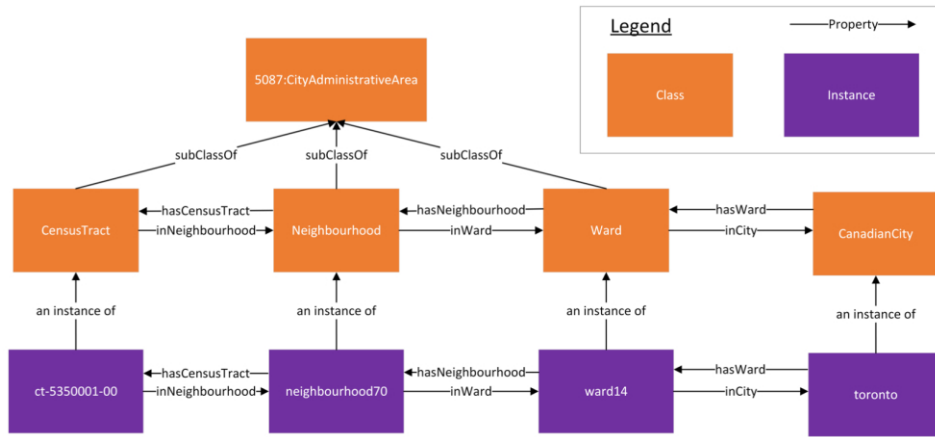


Figure 11: Representation of Canadian administrative areas



The `hasCensusTract`, `hasNeighbourhood`, and `hasWard` properties depicted above are all sub-properties of the `hasProperPart` property from ISO/IEC 5087-1 while the `inCity`, `inNeighbourhood`, and `inWard` properties are sub-properties of `properPartOf` from ISO/IEC 5087-1.

Census profiles and characteristics

The following section explains how Census Profiles and characteristics are represented using the Census Ontology.

Characteristics in the census are represented using the `Characteristic` class, which is a subclass of the `Indicator` class, which is a subclass of the `Quantity` class, as seen in Figure 12. Both the `Indicator` class and `Quantity` class are from the ISO/IEC 21972. Census characteristics are found in Census Profiles, described previously, datasets containing all the census information for a given geographic area (e.g., province, territory, census tract). Census Profiles are represented using the `CensusProfile` class and are linked to their respective characteristics using the `hasCharacteristic` property. Furthermore, an instance of `CensusProfile` represents the Census Profile dataset for a given geographic area. For example, the “ct-5350001-00CensusProfile2016” instance in Figure 12 would represent the Census Profile for the census tract 5350001.00 and would be related to individual

characteristics such as “ct-5350001-000-14Years2016Male”. As organisations, such as the Neighbourhood Planning Tables identified in our use case, may be interested in analysing census data across multiple years, a subclass of `Characteristic` and `CensusProfile` was created for the year 2016 as there is a separate set of characteristics and Census Profiles for each census year. The Census Profiles are also linked to the `DateTimeInterval` class from the OWL Time Ontology¹⁸ using the `hasTime` property, in order to represent the year of the Census Profile.

Each characteristic also has a class that is used to represent it and this class is also a subclass of the `Characteristic` class. For example, the `0-14Years2016` class in Figure 12 represents the characteristic for the number of people who were between the ages 0 to 14 in 2016. An instance of this class represents the characteristic for a given geographic area. For example, the “ct-5350001-000-14Years2016” instance in Figure 12 represents the characteristic for the number of people in the census tract 5350001.00 who were between the ages 0 to 14 in 2016. Some characteristics also have values for male and female population numbers and these male/female characteristics will be represented using instances with male/female at the end of the instance name. For

¹⁸<https://www.w3.org/TR/owl-time/#time:DateTimeInterval>.

example, the “ct-5350001-000-14Years2016Male” instance in Figure 12 represents the characteristic for the number of males in the census tract 5350001.00 who were between the ages 0 to 14 in 2016. Instances of characteristics are linked to an instance of the Measure class (which is also from the ISO/IEC 21972 Ontology) using the 21972:value property, which is used to represent the value of the characteristic. These Measure instances have a numerical_value property which expresses the value of the indicator using an integer or decimal value and a hasUnit property that links the Measure instance to the corresponding unit of measure. In the context of our use case about improving NIAs in Toronto, this representation of characteristics can help us identify and address possible socioeconomic inequities between the male and female populations (e.g., analysing income data might reveal a gender wage gap problem).

Sub-characteristics

In order to represent the sub-characteristic relation in the Canadian Census Ontology, the classes of the sub-characteristics were created as subclasses of the parent characteristic, as seen in Figure 14. Furthermore, instances of the characteristics are linked to their sub-characteristics using the hasPart property (e.g., ct-5350001-00TotalOccupiedPrivateDwellings2016 hasPart ct-5350001-00SingleDetachedHouse2016).

Although it is possible to define the parent characteristic as the sum of its sub-characteristics and use SHACL (Shapes Constraint Language) to constrain and validate the census data, the Canadian Census Ontology does not explicitly represent this relationship as, in practice, the numerical value of the parent characteristic is often not equal to the sum of its sub-characteristics. For instance, in Figure 13, we can see that there is a total of 250 occupied dwellings in the area but the sum of its four sub-characteristics (“Single-detached house,” “Apartment in a building that has five or more storeys,” “Other attached dwelling,” “Movable dwelling”) is equal to 245. This is due to the fact that some of the numerical values in the census profiles for smaller geographic areas were intentionally adjusted by Statistics Canada in order to protect the privacy of the citizens, although, these adjusted values will always be within 5 of the actual values [34]. Consequently, the Canadian Census Ontology does not explicitly constrain a characteristic to be the sum of its sub-characteristics.

In the context of our use case about improving NIAs in Toronto, this sub-characteristic representation can help provide Neighbourhood Planning Tables with more convenient access to both high- and low-level information when it is needed. For example, a Neighbourhood Planning Table might want to provide more bilingual ESL classes for new Chinese immigrants, and they could use the sub-characteristics to figure out which Chinese language (e.g., Mandarin, Cantonese) is most commonly used in the neighbourhood so they can schedule extra bilingual ESL classes that use the more popular language.

Populations

A population is the set of all the individuals that are being described by a given characteristic. Using the Canadian

Census Ontology, we can link a characteristic entity to its corresponding population entity using the cardinality_ of property from ISO/IEC 21972. For example, in Figure 15 we can see that the 0-14Years2016 characteristic is a cardinality of the 0-14YearsPopulation class (which represents the population of people who are between the ages 0 to 14), which is a subclass of Population class from the ISO/IEC 21972 Ontology and is defined_ by the Person0-14Years class (i.e., the 0-14YearsPopulation is comprised of people who are between the ages 0 to 14). Instances of a population can also be linked to the geographic area where the population is located using the located_ in property. For instance, in Figure 15, we can see that the ct-5350001-000-14Years2016Population instance (which represents the population of people who are between the ages 0 to 14 and live in the 5350001.00 census tract) is located_ in ct-5350001-00 (the 5350001.00 census tract).

Having this precise representation of a population allows us to connect the Characteristic (and its corresponding numerical value) with a Population entity and with its corresponding definition of a Person defines membership in the population. For example, a Neighbourhood Planning Table from our NIA use case might want to confirm whether an art gallery curator is classified as an Information and Cultural Industry job or an Arts, Entertainment and Recreation Industry job according to the Canadian census. By following the defined_ by links, they would be able to find the definitions for these jobs and conclude that that an art gallery curator is classified as an Arts, Entertainment and Recreation Industry job in the census. Thus, this representation allows us to unambiguously define how characteristics are defined while providing a way for connecting microdata (i.e., individual level data) to the Canadian Census Ontology.

Geometry

Neighbourhood Planning Tables might also be interested in aggregating census data in geographic areas that partially overlap with multiple census tracts. For instance, a Neighbourhood Planning Table might want to know the number of low-income residents that live within 1km of a certain food bank in order to help ensure that they allocate enough resources to that food bank. It may be difficult to aggregate census data for a geographic area like this as it may partially overlap with multiple census tracts which means that simply adding the numerical values for each census tract would likely result in an inaccurate total. Instead, by using the geometries of the census tracts, we could add a percentage of the numerical values for each census tract based on how much the designated area overlaps with each census tract and this would likely yield more accurate totals. More precise allocations of people to a neighbourhood could be achieved with the addition of housing locations.

To address this, the Canadian Census Ontology also includes a representation for describing the geometry of a census tract. As seen in Figure 16 below, census tract instances are linked to their corresponding location instance using the hasLocation property from ISO/IEC 5087-1 (e.g., ct-5350001-00Feature hasLocation ct-5350001-00Location), where the geometry of the census tract is described using

Figure 12: Representation of Canadian census characteristics

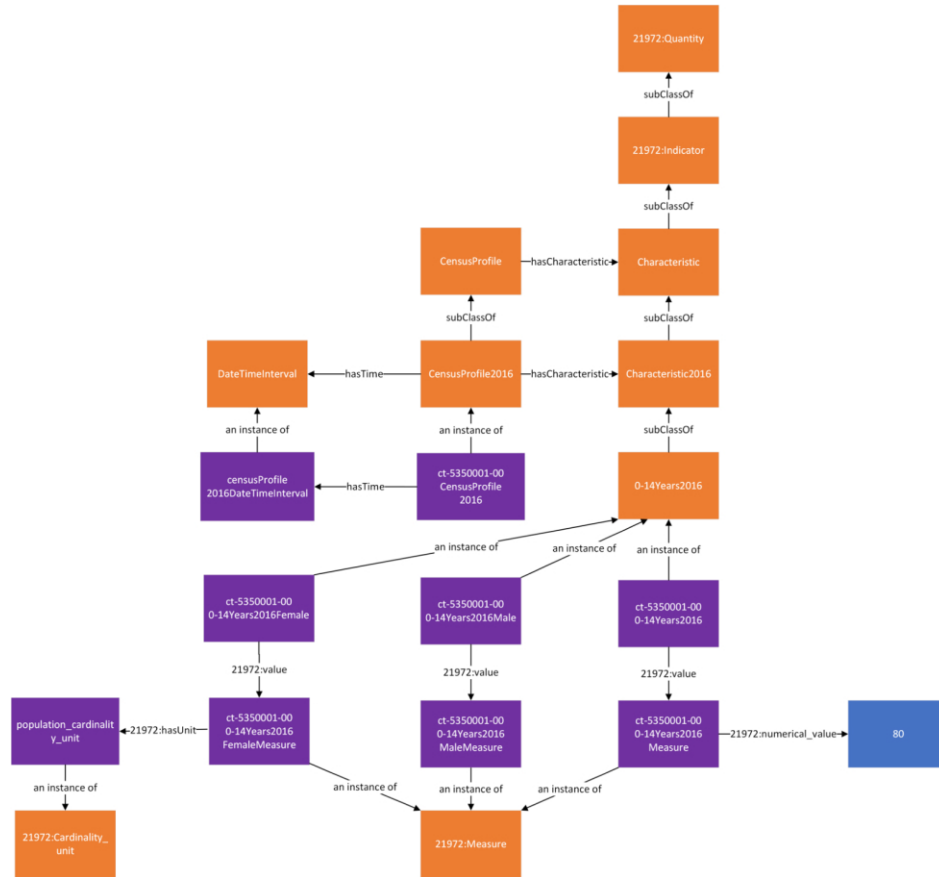


Figure 13: Private dwelling characteristic and sub-characteristics in the Canadian census (Statistics Canada, 2021)

Total - Occupied private dwellings by structural type of dwelling - 100% data	
Single-detached house	15.0
Apartment in a building that has five or more storeys	0.0
Other attached dwelling	23.0
Semi-detached house	8.0
Row house	9.0
Apartment or flat in a duplex	1.0
Apartment in a building that has fewer than five storeys	5.0
Other single-attached house	0.0
Movable dwelling	2.0

Figure 14: Private dwelling characteristic class with its subclasses in the Canadian census ontology



WKT (Well-Known Text) via the `asWKT` property from the GeoSPARQL ontology.¹⁹

¹⁹<https://opengeospatial.github.io/ogc-geosparql/geosparql11/index.html>.

Comparable characteristics

Census characteristics may change over time, which can make it difficult to link and compare census characteristics

Figure 15: Representation of population

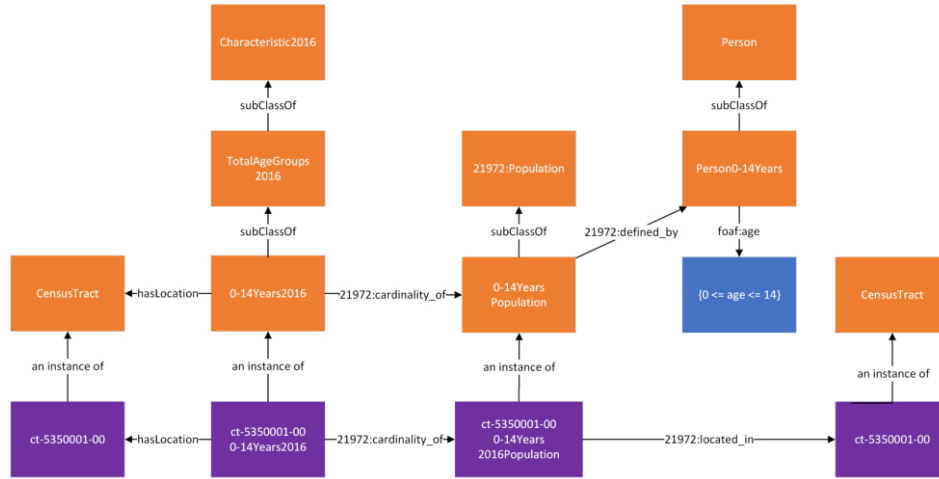
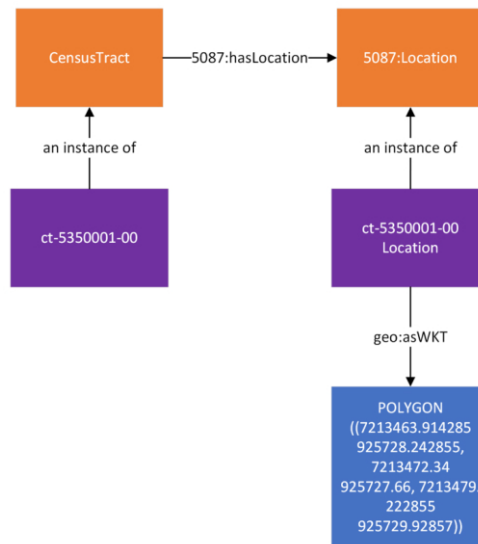


Figure 16: Representation of geometry



across different years. For example, the 2016 census uses the 2016 National Occupational Classification (NOC 2016) for categorising occupations while the 2021 census uses the 2021 National Occupational Classification (NOC 2021). While there are many similarities between NOC 2016 and NOC 2021, there are also some differences between the two classifications. For example, the Management Occupations category in NOC 2016 was replaced with the Legislative and Senior Management Occupations category in NOC 2021 with slight changes in how the jobs are classified [35]. At first glance, it is unclear whether these two characteristics are comparable to one another since they have very different names and use different versions of the NOC. To address this, we have included a comparableCharacteristic property which can be used to link characteristic classes from different census years in order to show that they are comparable to each other, even though they may appear to be very different (Figure 17). Using ISO/IEC 21972 to represent the definitions of these characteristics, it is possible to automate the analysis of how these two differ.

SPARQL queries for answering competency questions

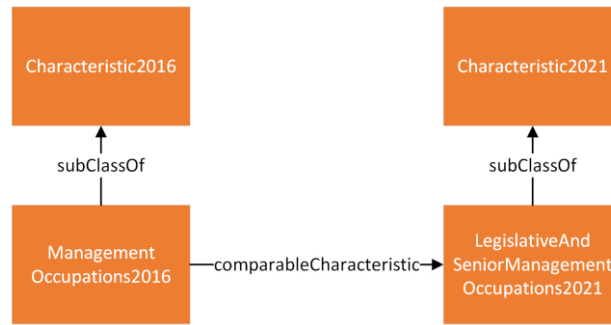
The Canadian Census ontology is defined using OWL2.²⁰ The ontology, and the census data instances of the ontology, are represented as a knowledge graph using RDF. The knowledge graph can be queried using SPARQL (SPARQL Protocol and RDF Query Language), a RDF query language.²¹ In this section, we translate competency question CQ3, defined in Section 4, into a SPARQL query to demonstrate the competency of the census ontology. The remainder of the competency questions' translations can be found in Appendix II.

We use the Southeast Scarborough Planning Table and the 2016 Canadian Census data for its four neighbourhoods it is responsible for (Neighbourhoods 135 (Morningside), 136

²⁰The OWL files can be found on our GitHub page: <https://github.com/EnterpriseIntegrationLab/CKGN/tree/main/UniversityOfToronto/Census/Ontologies>.

²¹<https://www.w3.org/TR/rdf11-concepts/>.

Figure 17: Using the comparable characteristic property to show that management occupations 2016 is comparable to legislative and senior management occupations 2021



(West Hill), 137 (Woburn), 139 (Scarborough Village)) as an example for the queries in this section.

CQ3: Which NIAs have the most low-income residents?

The following query finds the value of the characteristic “In low income based on the Low-income measure, after tax (LIM-AT)” (depicted in Figure 18) for the census tracts in the designated neighbourhoods and outputs a table that displays the neighbourhood in one column and the summed value of the characteristic across all the census tracts in that neighbourhood in another column. This LIM-AT characteristic is represented by the “LowIncomeMeasureAfterTaxPercent2016” class which is a subclass of Characteristic. This query could help Neighbourhood Planning Tables identify the neighbourhoods that have the highest number of low-income citizens. The diagram below illustrates some of the classes, properties, and instances used in this query.

```

PREFIX uoft: <http://ontology.eil.utoronto.ca/tove/cacensus#>
PREFIX toronto: <http://ontology.eil.utoronto.ca/Toronto/Toronto#>
PREFIX iso21972: <http://ontology.eil.utoronto.ca/ISO21972/iso21972#>
SELECT ?neighbourhood(sum(?value) as ?sumvalue)
WHERE{
  ?neighbourhood toronto:hasCensusTract ?censustract.
  FILTER (?neighbourhood IN (toronto:neighbourhood135,
    toronto:neighbourhood136, toronto:neighbourhood137,
    toronto:neighbourhood139
  ?limat a uoft:LowIncomeMeasureAfterTax2016;
  uoft:hasLocation ?censustract;
  iso21972:cardinality_of ?population;
  iso21972:value ?measure.
  ?measure iso21972:numerical_value ?value.
  ?population a ?populationclass.
  ?populationclass iso21972:defined_by
  uoft:PersonLowIncomeMeasureAfterTax2016
}
GROUP BY ?neighbourhood
ORDER BY DESC(?sumvalue)
  
```

By running the SPARQL query, we can see (Figure 19) that out of all the NIAs managed by the Southeast Scarborough Planning Table, Neighbourhood 137 (Woburn) has the highest number of low-income residents (14410) while Neighbourhood 135 (Morningside) has the lowest number of low-income residents (4085).

Similar queries can be constructed to answer the remaining competency questions, which can be found in Appendix II.

Integration and visualisations

Building the Canadian Census Ontology on top of the ISO/IEC 5087 and ISO/IEC 21972 standards allows us to integrate the data from the Canadian Census with other datasets that have been mapped onto these standards. For example, we have mapped Toronto’s neighbourhood crime data published by the Toronto Police Service²² onto the ISO/IEC 21972 indicator ontology and the ISO/IEC 5087 ontologies (i.e., police precinct areas are represented as administrative areas), allowing the dataset to be integrated with the Canadian Census data. This integration supports analyses exploring links between crime and socioeconomic factors. We have also mapped subsets of OpenStreetMap²³ data including roads, buildings, stores, etc. onto ISO/IEC 5087-1, -2 and -3. Integrating OpenStreetMap data with the census data supports the analysis of infrastructure, retail, commercial, etc. environments in relation to the socioeconomic data found in the census. Furthermore, the integration with other datasets allows users to use SPARQL queries to access data from multiple different sources without needing to deal with the different file formats that the datasets were originally published in.

Another advantage of publishing census data as linked data is that it enables researchers and analysts to build their own data analysis and/or data visualisation tools that can process the data in the knowledge graph. In fact, the Canadian Census Ontology supports a general approach for visualising metrics for any administrative area as it includes representation of geospatial data that can be linked to census data. For example, a Python program that utilises existing data visualisation packages (such as Folium) can be used to

²²<https://data.torontopolice.on.ca/datasets/TorontoPS::neighbourhood-crime-rates-open-data/explore>.

²³<https://www.openstreetmap.org/>.

Figure 18: Representation of the low-income query

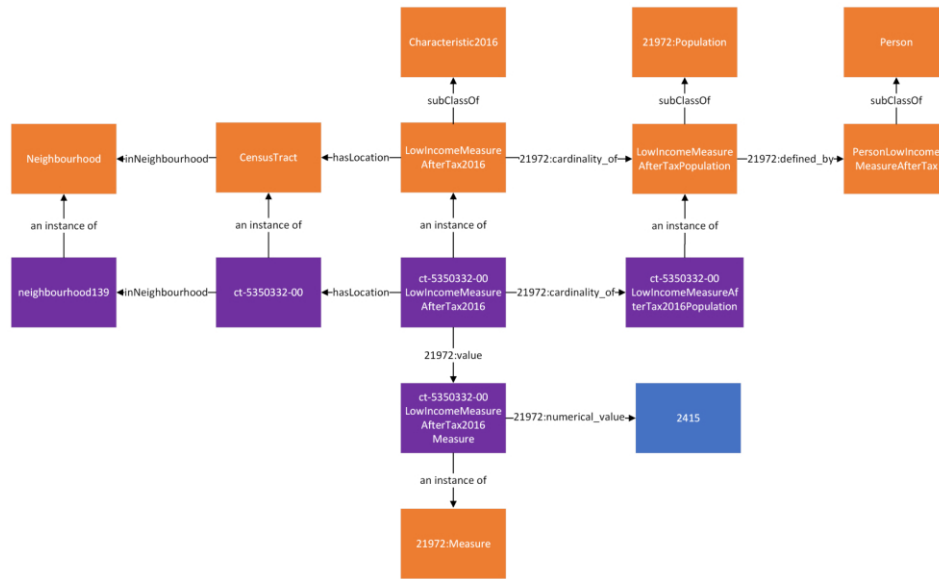
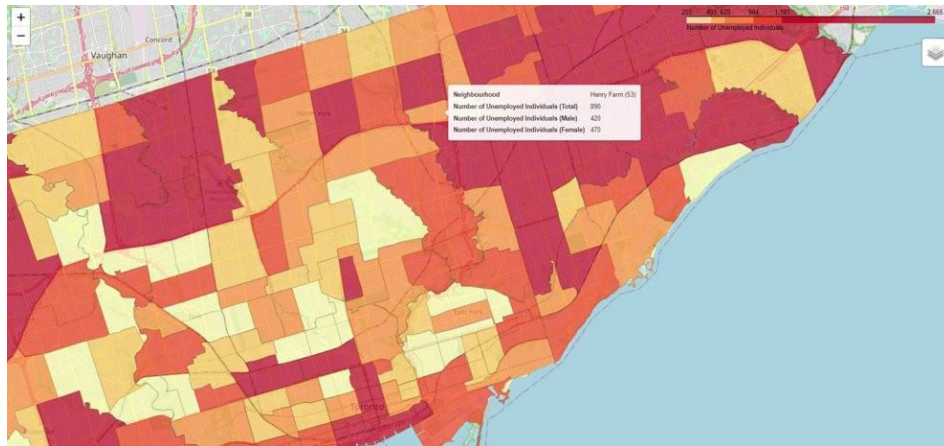


Figure 19: Results of the low-income query

	neighborhood	sumvalue
1	#neighborhood137	"14410"^^xsd:integer
2	#neighborhood136	"6780"^^xsd:integer
3	#neighborhood139	"5325"^^xsd:integer
4	#neighborhood135	"4085"^^xsd:integer

Figure 20: Choropleth map of unemployed individuals in the City of Toronto



easily generate interactive choropleth maps using census linked data. An example of such visualisation is shown in Figure 20 below:

This choropleth visualisation shows a map of Toronto, and its neighbourhoods which are colored based on the number of unemployed individuals in that neighbourhood. Yellow represents fewer unemployed individuals while red represents more unemployed individuals. A popup box that shows the neighbourhood name and the number of unemployed individuals also appears on mouse hover. This map was created using the CensusVis program²⁴ that was developed as a project under the University of Toronto's Enterprise

²⁴<https://github.com/andw2/CensusVis>.

Integration Laboratory. This program is capable of generating choropleth maps using linked data that is queried (using SPARQL queries like the ones seen in Section 7.0) from the knowledge graph and supports the visualisation of census data in various types of administrative areas. CensusVis enables easy creation of data visualisations as the user only needs to input the type of administrative area to be visualised (e.g., wards, neighbourhoods, etc.) and the class name of the census characteristic to be visualised (e.g., "Unemployed2016" for the number of unemployed individuals) in order to generate a choropleth map of the selected administrative area and census characteristic. Similar visualisations could also be created for other cities and administrative areas if they

are published as linked data using the Canadian Census Ontology.

Conclusion

The goal of our research is to “open up” census data by reducing the complexity and ambiguity inherent in integrating census data with other data sources. Complexity arises out of having to deconstruct census characteristics (metrics) definitions into their constituent concepts, in order to understand how it may relate to other data sources. Ambiguity arises in the imprecision in the definition of potentially relevant data (both attributes and values) found in other sources. Our approach is based on three technologies:

1. Ontologies for the precise definition of concepts and properties found in census and other sources of data. Ontologies use logic (e.g., Description Logic) to define concepts and properties, which are represented as graphs.
2. Linked Data, which introduces globally unique identifiers (i.e., URIs) for concepts, properties, and their instances. URIs are the basis for linking data across multiple sources.
3. Graph databases, which provide an implementation platform where ontologies define the types of concepts and properties instantiated in the graph, using the unique identifiers provided by linked data URIs. The graph database provides support for queries, e.g., SPARQL, visualization, browsing, etc.

Our approach to deconstructing census characteristic natural language definitions is to map the concepts embedded in a definition onto the ISO/IEC 21972 Indicator ontology that provides semantics for defining statistical populations. These populations are, in turn, defined by prototypical descriptions of members of the population. For example, prototypical description of a household, person, dwelling, etc. The representation of these prototypical descriptions is based on ontologies such as the ISO/IEC 5087 series of city data standards, which includes concepts and properties for the representation of households, person, dwellings, municipal administrative areas, etc.

The resulting knowledge graph is an integration of the multiple data sources, providing direct and indirect connections across the data/concepts contained therein. Using query languages such as SPARQL, complex relationships amongst the data can be explored and visualised without the cost of additional data wrangling.

With the completion of the ontology, we are pursuing the use of AI Large Language Models (LLM) for mapping census characteristic definitions onto our ontology and using LLMs as a natural language query interface for generating SPARQL queries for the knowledge graph.

Acknowledgements

This research was funded, in part, by Tata Consultancy Services and the Natural Sciences and Engineering Research Council of Canada.

Statement on conflicts of interest

The authors declare that there are no known conflicts of interest.

Ethics statement

This article did not require an ethics approval because it involves publicly available information that is not personally identifiable.

Data availability statement

The data for the 2016 Canadian Census of Population that was used for this article can be found on Statistics Canada's website here: <https://www12.statcan.gc.ca/census-recensement/2016/dp-pd/prof/index.cfm?Lang=E>.

References

1. Statistics Canada. Census of Population [Internet]. 2020. Available from: <https://www23.statcan.gc.ca/imdb/p2SV.pl?Function=getSurvey>.
2. Statistics Canada. How census data are used [Internet]. 2009. Available from: <https://www12.statcan.gc.ca/census-recensement/srvmsg/srvmsg404.html>.
3. Chessell M, Scheepers F, Nguyen N, van Kessel R, van der Starre R. Governing and managing big data for analytics and decision makers. IBM Redguides for Business Leaders. 2014 Aug 26;252.
4. Ajani S. An ontology and semantic metadata based semantic search technique for census domain in a big data context. *Int. J. Eng. Res. Technol.* 2014;3(2):1–5.
5. Berners-Lee T. Linked Data - design issues [Internet]. 2010. Available from: <https://www.w3.org/DesignIssues/LinkedData.html>.
6. Fernández JD, Martínez-Prieto MA, Gutiérrez C. Publishing open statistical data: the Spanish census. *Proceedings of the 12th Annual International Digital Government Research Conference: Digital Government Innovation in Challenging Times* 2011 Jun 12 (pp. 20–25).
7. Pokhriyal N, Letouzé E, Vosoughi S. Accurate intercensal estimates of energy access to track Sustainable Development Goal 7. *EPJ Data Science.* 2022 Dec 1;11(1):60. <https://doi.org/10.1140/epjds/s13688-022-00371-5>
8. Langford A, Johnson B. Trends in social inequalities in male mortality, 2001–08. *Intercensal estimates for England and Wales. Health Statistics Quarterly.* 2010 Sep 1;47:5–32. <https://doi.org/10.1057/hsq.2010.14>

9. Leventhal JM, Martin KD, Gaither JR. Using US data to estimate the incidence of serious physical abuse in children. *Pediatrics*. 2012 Mar 1;129(3):458-64. <https://doi.org/10.1542/peds.2011-1277>
10. United Nations Statistics Division - Demographic and Social Statistics [Internet]. Available from: <https://unstats.un.org/unsd/demographic/sources/census/alternativecensusesdesigns.htm>.
11. Wang Y, Zhang X, Lu H, Matthews KA, Greenlund KJ. Intercensal and postcensal estimation of population size for small geographic areas in the United States. *International Journal of Population Data Science*. 2020;5(1). <https://doi.org/10.23889/ijpds.v5i1.1160>
12. Weden MM, Peterson CE, Miles JN, Shih RA. Evaluating linearly interpolated intercensal estimates of demographic and socioeconomic characteristics of US counties and census tracts 2001–2009. *Population research and policy review*. 2015 Aug;34:541-59. <https://doi.org/10.1007/s11113-015-9359-8>
13. Statistics Canada. The linkable open data environment [Internet]. 2020. Available from: <https://www.statcan.gc.ca/en/lode>.
14. Bukhari AC, Baker CJ. The Canadian health census as Linked Open Data: towards policy making in public health. *Data integration in the life sciences*. 2013 Jul 11.
15. University of Minnesota. IPUMS International [Internet]. Available from: <https://international.ipums.org/international/>.
16. Petrou I, Papastefanatos G, Dalamagas T. Publishing census as linked open data: a case study. *Proceedings of the 2nd International Workshop on Open Data 2013 Jun 3* (pp. 1–3). <https://doi.org/10.1145/2500410.2500412>
17. Aracri RM, De Francisci S, Pagano A, Scannapieco M, Tosco L, Valentino L. Publishing the 15th Italian Population and Housing Census as Linked Open Data. *SemStats@ ISWC 2014*.
18. Aracri RM, Radini R, Scannapieco M, Tosco L. Using ontologies for official statistics: the ISTAT experience. *Current Trends in Web Engineering: ICWE 2017 International Workshops*. 2018. https://doi.org/10.1007/978-3-319-74433-9_15
19. Taylor Z. UNI•CEN Documentation Report 2: Standardized Census Data Tables. UNI-CEN documentation. 2022
20. Gruninger M, Fox M. Methodology for the design and evaluation of ontologies. In *Proc. IJCAI'95, Workshop on Basic Ontological Issues in Knowledge Sharing 1995*.
21. Lenat DB, Guha RV. The evolution of CycL, the Cyc representation language. *ACM SIGART Bulletin*. 1991 Jun 1;2(3):84–7. <https://doi.org/10.1145/122296.122308>
22. Katsumi M, Grüninger M. Theorem proving in the ontology lifecycle. *International Conference on Knowledge Engineering and Ontology Development*. 2010 Oct 25 (Vol. 2, pp. 37-49). <https://doi.org/10.5220/0003076400370049>
23. Toronto Strong Neighbourhoods Strategy [Internet]. City of Toronto. 2023. Available from: <https://www.toronto.ca/city-government/accountability-operations-customer-service/long-term-vision-plans-and-strategies/toronto-strong-neighbourhoods-strategy-2020/>.
24. International Organization for Standardization. ISO/IEC 21972:2020. Information technology Upper level ontology for smart city indicators. ISO; 2020.
25. International Organization for Standardization. ISO/IEC 5087-1:2023. Information technology City data model Part 1: Foundation level concepts. ISO; 2023.
26. International Organization for Standardization. ISO/IEC DIS 5087-2. Information technology City data model Part 2: City level concepts. ISO; 2023.
27. International Organization for Standardization. ISO/IEC AWI 5087-3. Information technology City data model Part 3: Service level concepts -Transportation planning. ISO; 2020.
28. Fox MS. The semantics of populations: A city indicator perspective. *Journal of Web Semantics*. 2018 Jan 1;48:48-65. <https://doi.org/10.1016/j.websem.2018.01.001>
29. Fox MS. A foundation ontology for global city indicators. University of Toronto, Toronto, Global Cities Institute. 2013 Aug. <https://doi.org/10.13140/RG.2.2.17487.10404>
30. Fox MS. The role of ontologies in publishing and analyzing city indicators. *Computers, Environment and Urban Systems*. 2015 Nov 1;54:266-79. <https://doi.org/10.1016/j.compenvurbsys.2015.09.009>
31. Katsumi M, Fox M. An Ontology-Based Standard for Transportation Planning. *JOWO*. 2019 Sep.
32. Katsumi M, Fox M. iCity Transportation Planning Suite of Ontologies. University of Toronto. 2020.
33. Fox MS. The PolisGnosis project enabling the computational analysis of city performance. In *IIIE Annual Conference. Proceedings 2017* (pp. 2009-2014). Institute of Industrial and Systems Engineers (IISE).
34. Statistics Canada. Census Profile, 2016 Census [Internet]. 2021. Available from: https://www12.statcan.gc.ca/census-recensement/2016/dp-pd/prof/index.cfm?Lang=_E.
35. Government of Canada, Statistics Canada. Census Profile, 2016 Census – Canada [Country] and Canada [Country] [Internet]. 2021. Available from: https://www12.statcan.gc.ca/census-recensement/2016/dp-pd/prof/details/page.cfm?Lang=_E.

Abbreviations

AGOL:	ArcGIS Online	LODE:	Linkable Open Data Environment
API:	Application Programming Interface	NIA:	Neighbourhood Improvement Area
BSR:	Base Statistical Registers	NOC:	National Occupational Classification
CQ:	Competency question	NOS:	National Occupancy Standard
CSV:	Comma-separated values	ODA:	Open Database of Addresses
DGUID:	Dissemination Geography Unique Identifier	ODB:	Open Database of Buildings
ESL:	English as a Second Language	ODCAF:	Open Database of Cultural and Art Facilities
ESRI:	Environmental Systems Research Institute	ODEF:	Open Database of Educational Facilities
FOAF:	Friend of a friend	ODHF:	Open Database of Healthcare Facilities
GIS:	Geographic Information System	ODRSF:	Open Database of Recreational and Sport Facilities
GML:	Geography Markup Language	OWL:	Web Ontology Language
HEART:	Health Equity Assessment and Response Tool	RDF:	Resource Description Framework
IEC:	International Electrotechnical Commission	SHP:	Shapefile
IPC:	Institute of Printed Circuits	SPARQL:	SPARQL Protocol and RDF Query Language
ISO:	International Organization for Standardization	URI:	Uniform Resource Identifier
LIM-AT:	Low-income measure, after tax	W3C:	World Wide Web Consortium
LOD:	Linked Open Data	WKT:	Well-Known Text



Appendix I

This appendix contains a table of key classes and properties of the Canadian Census Ontology (the OWL file can be found here at our GitHub page: <https://github.com/EnterpriseIntegrationLab/CKGN/tree/main/UniversityOfToronto/Census/Ontologies>)

Class	Property	Value restriction
CensusProfile	hasCharacteristic	only Characteristic
	hasTime	only time:DateTimeInterval
Characteristic	rdfs:subClassOf	21972:Indicator
	hasLocation	5087:CityOntologyThing
	hasPart	only Characteristic
	cardinality_of	only 21972:Population
	value	only 21972:Measure
CensusTract	subClassOf	5087:CityAdministrativeArea
	5087:properPartOf	only 5087:CityAdministrativeArea
	5087:hasProperPart	only 5087:CityAdministrativeArea
	5087:hasLocation	only 5087:Location

Appendix II Additional SPARQL query examples

This section contains the remainder of the competency questions' translations into SPARQL and the results returned by the knowledge graph.

CQ1: Which NIAs have the highest number of unemployed individuals?

The following query finds the value of the characteristic "Unemployed" for the census tracts in the designated neighbourhoods and outputs a table that displays the neighbourhood in one column and the summed value of the characteristic across all the census tracts in that neighbourhood in another column.

```

PREFIX uoft: <http://ontology.eil.utoronto.ca/tove/cacensus#>
PREFIX toronto: <http://ontology.eil.utoronto.ca/Toronto/Toronto#>
PREFIX iso21972: <http://ontology.eil.utoronto.ca/ISO21972/iso21972#>
SELECT ?neighbourhood (sum(?value) as ?sumvalue)
WHERE{
  ?neighbourhood toronto:hasCensusTract ?censustract.
  FILTER (?neighbourhood IN (toronto:neighbourhood135,
    toronto:neighbourhood136, toronto:neighbourhood137,
    toronto:neighbourhood139 ) )
  ?unemployed a uoft:Unemployed2016;
    uoft:hasLocation ?censustract;
    iso21972:cardinality_of ?population;
    iso21972:value ?measure.
  ?measure iso21972:numerical_value ?value.
  ?population a ?populationclass.
  ?populationclass iso21972:defined_by uoft:PersonUnemployed2016
}
GROUP BY ?neighbourhood
ORDER BY DESC(?sumvalue)

```

By running the SPARQL query, we can see (Supplementary Figure 1) that out of all the NIAs managed by the Southeast Scarborough Planning Table, Neighbourhood 137 (Woburn) has the highest number of unemployed residents (2665) while Neighbourhood 139 (Scarborough Village) has the lowest number of unemployed residents (980).

CQ2: What industries do NIA residents work in?

Supplementary Figure 1: Results of the unemployed query

	neighborhood	sumvalue
1	#neighborhood137	"2665" ^{xs:integer}
2	#neighborhood136	"1515" ^{xs:integer}
3	#neighborhood135	"995" ^{xs:integer}
4	#neighborhood139	"980" ^{xs:integer}

Supplementary Figure 2: Results of the industry query

	neighborhood	sumvalue
1	.#neighborhood137	"2800""xsd:integer
2	.#neighborhood136	"1025""xsd:integer
3	.#neighborhood135	"895""xsd:integer
4	.#neighborhood139	"620""xsd:integer

- For example, a NIA where many residents work in the manufacturing sector might benefit from IPC training and certification.

The following query finds the value of the characteristic “31-33 Manufacturing” (which represents the number of people who work in the manufacturing sector according to NAICS codes 31-33) for the census tracts in the designated neighbourhoods and outputs a table that displays the neighbourhood in one column and the summed value of the characteristic across all the census tracts in that neighbourhood in another column.

```

PREFIX uoft: <http://ontology.eil.utoronto.ca/tove/cacensus#>
PREFIX toronto: <http://ontology.eil.utoronto.ca/Toronto/Toronto#>
PREFIX iso21972: <http://ontology.eil.utoronto.ca/ISO21972/iso21972#>
SELECT ?neighbourhood(sum(?value) as ?sumvalue)
WHERE{
  ?neighbourhood toronto:hasCensusTract ?censustract.
  FILTER (?neighbourhood IN (toronto:neighbourhood135,
    toronto:neighbourhood136, toronto:neighbourhood137,
    toronto:neighbourhood139 ) )
  ?manufacturing a uoft:31-33Manufacturing2016;
    uoft:hasLocation ?censustract;
    iso21972:cardinality_of ?population;
    iso21972:value ?measure.
  ?measure iso21972:numerical_value ?value.
  ?population a ?populationclass.
  ?populationclass iso21972:defined_by uoft:Person31-33Manufacturing2016
}
GROUP BY ?neighbourhood
ORDER BY DESC(?sumvalue)

```

By running the SPARQL query, we can see (Supplementary Figure 2) that out of all the NIAs managed by the Southeast Scarborough Planning Table, Neighbourhood 137 (Woburn) has the highest number of residents (2800) who work in the manufacturing industry while Neighbourhood 139 (Scarborough Village) has the lowest number of residents (620) who work in the manufacturing industry.

CQ4: Which NIAs have the most residents with a high shelter-cost-to-income ratio (30+ %)?

The following query finds the value of the characteristic “Spending 30 Percent or More of Income on Shelter Costs” for the census tracts in the designated neighbourhoods and outputs a table that displays the neighbourhood in one column and the summed value of the characteristic across all the census tracts in that neighbourhood in another column.

```

PREFIX uoft: <http://ontology.eil.utoronto.ca/tove/cacensus#>
PREFIX toronto: <http://ontology.eil.utoronto.ca/Toronto/Toronto#>
PREFIX iso21972: <http://ontology.eil.utoronto.ca/ISO21972/iso21972#>
SELECT ?neighbourhood(sum(?value) as ?sumvalue)
WHERE{
  ?neighbourhood toronto:hasCensusTract ?censustract.
  FILTER (?neighbourhood IN (toronto:neighbourhood135,
    toronto:neighbourhood136, toronto:neighbourhood137,
    toronto:neighbourhood139 ) )
  ?highsheltercost a uoft:Spending30PercentOrMoreOfIncomeOnShelterCosts2016;
    uoft:hasLocation ?censustract;
    iso21972:value ?measure.
  ?measure iso21972:numerical_value ?value.
}
GROUP BY ?neighbourhood

```



Supplementary Figure 3: Results of the high shelter cost query

	neighborhood	sumvalue
1	.#neighborhood137	*6605**xsd:integer
2	.#neighborhood136	*3230**xsd:integer
3	.#neighborhood139	*2460**xsd:integer
4	.#neighborhood135	*2090**xsd:integer

Supplementary Figure 4: Results of the no degree/diploma query

	neighborhood	sumvalue
1	.#neighborhood137	*8850**xsd:integer
2	.#neighborhood136	*4990**xsd:integer
3	.#neighborhood139	*2955**xsd:integer
4	.#neighborhood135	*2465**xsd:integer

ORDER BY DESC(?sumvalue)

By running the SPARQL query, we can see (Supplementary Figure 3) that out of all the NIAs managed by the Southeast Scarborough Planning Table, Neighbourhood 137 (Woburn) has the highest number of residents with a high shelter-cost-to-income ratio (6605) while Neighbourhood 135 (Morningside) has the lowest number of residents with a high shelter-cost-to-income ratio (2090).

CQ5: Which NIAs have the most residents who do not have a postsecondary degree nor high school diploma?

The following query finds the value of the characteristic “No Certificate Diploma or Degree” for the census tracts in the designated neighbourhoods and outputs a table that displays the neighbourhood in one column and the summed value of the characteristic across all the census tracts in that neighbourhood in another column.

```

PREFIX uoft: <http://ontology.eil.utoronto.ca/tove/cacensus#>
PREFIX toronto: <http://ontology.eil.utoronto.ca/Toronto/Toronto#>
PREFIX iso21972: <http://ontology.eil.utoronto.ca/ISO21972/iso21972#>
SELECT ?neighbourhood (sum(?value) as ?sumvalue)
WHERE{
    ?neighbourhood toronto:hasCensusTract ?censustract.
    FILTER (?neighbourhood IN (toronto:neighbourhood135,
        toronto:neighbourhood136, toronto:neighbourhood137,
        toronto:neighbourhood139 ) )
    ?nodegree a uoft:NoCertificateDiplomaOrDegree2016;
        uoft:hasLocation ?censustract;
        iso21972:cardinality_of ?population;
        iso21972:value ?measure.
    ?measure iso21972:numerical_value ?value.
    ?population a ?populationclass.
    ?populationclass iso21972:defined_by
        uoft:PersonNoCertificateDiplomaOrDegree2016.
}
GROUP BY ?neighbourhood
ORDER BY DESC(?sumvalue)

```

By running the SPARQL query, we can see (Supplementary Figure 4) that out of all the NIAs managed by the Southeast Scarborough Planning Table, Neighbourhood 137 (Woburn) has the highest number of residents who do not have a postsecondary degree (8850) while Neighbourhood 135 (Morningside) has the lowest number of residents who do not have a postsecondary degree (2465).

CQ6: Which NIAs have the most residents who know neither English nor French?

The following query finds the value of the characteristic “Neither English Nor French” for the census tracts in the designated neighbourhoods and outputs a table that displays the neighbourhood in one column and the summed value of the characteristic across all the census tracts in that neighbourhood in another column.

```

PREFIX uoft: <http://ontology.eil.utoronto.ca/tove/cacensus#>
PREFIX toronto: <http://ontology.eil.utoronto.ca/Toronto/Toronto#>
PREFIX iso21972: <http://ontology.eil.utoronto.ca/ISO21972/iso21972#>

```

Supplementary Figure 5: Results of the neither English nor French query

	neighborhood	sumvalue
1	.#neighborhood137	"2575"^^xsd:integer
2	.#neighborhood136	"600"^^xsd:integer
3	.#neighborhood139	"525"^^xsd:integer
4	.#neighborhood135	"485"^^xsd:integer

```

SELECT ?neighbourhood(sum(?value) as ?sumvalue)
WHERE{
    ?neighbourhood toronto:hasCensusTract ?censustract.
    FILTER (?neighbourhood IN (toronto:neighbourhood135,
        toronto:neighbourhood136, toronto:neighbourhood137,
        toronto:neighbourhood139 ) )
    ?noenglishfrench a uoft:NeitherEnglishNorFrench2016;
    uoft:hasLocation ?censustract;
    iso21972:cardinality_of ?population;
    iso21972:value ?measure.
    ?measure iso21972:numerical_value ?value.
    ?population a ?populationclass.
    ?populationclass iso21972:defined_by
    uoft:PersonNeitherEnglishNorFrench2016
}
GROUP BY ?neighbourhood
ORDER BY DESC(?sumvalue)

```

By running the SPARQL query, we can see (Supplementary Figure 5) that out of all the NIAs managed by the Southeast Scarborough Planning Table, Neighbourhood 137 (Woburn) has the highest number of residents who know neither English nor French (2575) while Neighbourhood 135 (Morningside) has the lowest number of residents who know neither English nor French (485).

CQ7: What are the largest visible minority groups in NIAs?

The following query finds the value of each visible minority characteristic for the census tracts in the designated neighbourhoods and outputs a table that displays the visible minority in one column and the summed value of the characteristic across all the census tracts in all four neighbourhoods managed by the Southeast Scarborough Planning Table in another column.

```

PREFIX uoft: <http://ontology.eil.utoronto.ca/tove/cacensus#>
PREFIX toronto: <http://ontology.eil.utoronto.ca/Toronto/Toronto#>
PREFIX iso21972: <http://ontology.eil.utoronto.ca/ISO21972/iso21972#>
SELECT ?visibleminority (sum(?value) as ?sumvalue)
WHERE{
    ?neighbourhood toronto:hasCensusTract ?censustract.
    FILTER (?neighbourhood IN (toronto:neighbourhood135,
        toronto:neighbourhood136, toronto:neighbourhood137,
        toronto:neighbourhood139 ) )
    ?visibleminority rdfs:subClassOf
    uoft:TotalVisibleMinorityPopulation2016.
    ?visibleminoritygroup a ?visibleminority;
    uoft:hasLocation ?censustract;
    iso21972:cardinality_of ?population;
    iso21972:value ?measure.
    ?measure iso21972:numerical_value ?value.
    ?population a ?populationclass.
    ?populationclass iso21972:defined_by ?person
    FILTER NOT EXISTS { ?person rdfs:subClassOf uoft:Male.}
    FILTER NOT EXISTS { ?person rdfs:subClassOf uoft:Female.}
}
GROUP BY ?visibleminority

```



Supplementary Figure 6: Results of the visible minority query

	visibleminority	sumvalue
1	uoft:SouthAsianVisibleMinority2016	*37280**xsd:integer
2	uoft:BlackVisibleMinority2016	*18065**xsd:integer
3	uoft:FilipinoVisibleMinority2016	*9235**xsd:integer
4	uoft:ChineseVisibleMinority2016	*5670**xsd:integer
5	uoft:VisibleMinorityNIE2016	*2750**xsd:integer
6	uoft:WestAsianVisibleMinority2016	*2670**xsd:integer
7	uoft:MultipleVisibleMinorities2016	*2475**xsd:integer
8	uoft:LatinAmericanVisibleMinority2016	*1925**xsd:integer
9	uoft:ArabVisibleMinority2016	*1000**xsd:integer
10	uoft:SoutheastAsianVisibleMinority2016	*810**xsd:integer
11	uoft:KoreanVisibleMinority2016	*375**xsd:integer
12	uoft:JapaneseVisibleMinority2016	*240**xsd:integer

Supplementary Figure 7: Results of the senior population query

	neighborhood	sumvalue
1	:#neighborhood137	*8010**xsd:integer
2	:#neighborhood136	*4240**xsd:integer
3	:#neighborhood135	*2695**xsd:integer
4	:#neighborhood139	*2225**xsd:integer

ORDER BY DESC(?sumvalue)

By running the SPARQL query, we can see (Supplementary Figure 6) that the top three largest visible minority groups for the NIA's managed by the Southeast Scarborough Planning Table are South Asian with 37280 residents, Black with 18065 residents, and Filipino with 9235 residents.

CQ8: Which NIAs have the largest senior populations?

The following query finds the value of the characteristic “65+ Years” for the census tracts in the designated neighbourhoods and outputs a table that displays the neighbourhood in one column and the summed value of the characteristic across all the census tracts in that neighbourhood in another column.

```

PREFIX uoft: <http://ontology.eil.utoronto.ca/tove/cacensus#>
PREFIX toronto: <http://ontology.eil.utoronto.ca/Toronto/Toronto#>
PREFIX iso21972: <http://ontology.eil.utoronto.ca/ISO21972/iso21972#>
SELECT ?neighbourhood
(sum(?value) as ?sumvalue)
WHERE{
  ?neighbourhood oronto:hasCensusTract ?censustract.
  FILTER (?neighbourhood IN (toronto:neighbourhood135,
    toronto:neighbourhood136, toronto:neighbourhood137,
    toronto:neighbourhood139 ) )
  ?senior a uoft:65+Years2016;
    uoft:hasLocation ?censustract;
    iso21972:cardinality_of ?population;
    iso21972:value ?measure.
  ?measure iso21972:numerical_value ?value.
  ?population a ?populationclass.
  ?populationclass iso21972:defined_by
    uoft:Person65+Years2016
}
GROUP BY ?neighbourhood
ORDER BY DESC(?sumvalue)

```

By running the SPARQL query, we can see (Supplementary Figure 7) that out of all the NIAs managed by the Southeast Scarborough Planning Table, Neighbourhood 137 (Woburn) has the highest number of senior (age 65+) residents (8010) while Neighbourhood 139 (Scarborough Village) has the lowest number of senior residents (2225).

Supplementary Figure 8: Results of the child population query

	neighborhood	sumvalue
1	.#neighborhood137	*9625**xsd:integer
2	.#neighborhood136	*4635**xsd:integer
3	.#neighborhood139	*3365**xsd:integer
4	.#neighborhood135	*2880**xsd:integer

CQ9: Which NIAs have the largest child populations?

The following query finds the value of the characteristic “0-14 Years” for the census tracts in the designated neighbourhoods and outputs a table that displays the neighbourhood in one column and the summed value of the characteristic across all the census tracts in that neighbourhood in another column.

```

PREFIX uoft: <http://ontology.eil.utoronto.ca/tove/cacensus#>
PREFIX toronto: <http://ontology.eil.utoronto.ca/Toronto/Toronto#>
PREFIX iso21972: <http://ontology.eil.utoronto.ca/ISO21972/iso21972#>
SELECT ?neighbourhood (sum(?value) as ?sumvalue)
WHERE{
    ?neighbourhood toronto:hasCensusTract ?censustract.
    FILTER (?neighbourhood IN (toronto:neighbourhood135,
        toronto:neighbourhood136, toronto:neighbourhood137, toronto:neighbourhood139 ) )
    ?child a uoft:0-14Years2016;
    uoft:hasLocation ?censustract;
    iso21972:cardinality_of ?population;
    iso21972:value ?measure.
    ?measure iso21972:numerical_value ?value.
    ?population a ?populationclass.
    ?populationclass iso21972:defined_by uoft:Person0-14Years2016
}
GROUP BY ?neighbourhood
ORDER BY DESC(?sumvalue)

```

By running the SPARQL query, we can see (Supplementary Figure 8) that out of all the NIAs managed by the Southeast Scarborough Planning Table, Neighbourhood 137 (Woburn) has the highest number of child (age 0-14) residents (9625) while Neighbourhood 135 (Morningside) has the lowest number of child residents (2880).

CQ10: Which NIAs have the longest commute times?

The following query finds the value of the characteristic “commuting duration 60 minutes and over” for the census tracts in the designated neighbourhoods and outputs a table that displays the neighbourhood in one column and the summed value of the characteristic across all the census tracts in that neighbourhood in another column.

```

PREFIX uoft: <http://ontology.eil.utoronto.ca/tove/cacensus#>
PREFIX toronto: <http://ontology.eil.utoronto.ca/Toronto/Toronto#>
PREFIX iso21972: <http://ontology.eil.utoronto.ca/ISO21972/iso21972#>
SELECT ?neighbourhood(sum(?value) as ?sumvalue)
WHERE{
    ?neighbourhood toronto:hasCensusTract ?censustract.
    FILTER (?neighbourhood IN (toronto:neighbourhood135,
        toronto:neighbourhood136, toronto:neighbourhood137,
        toronto:neighbourhood139 ) )
    ?commute a uoft:60MinutesAndOver2016;
    uoft:hasLocation ?censustract;
    iso21972:cardinality_of ?population;
    iso21972:value ?measure.
    ?measure iso21972:numerical_value ?value.
    ?population a ?populationclass.
    ?populationclass iso21972:defined_by
        uoft:Person60MinutesAndOver2016
}
GROUP BY ?neighbourhood

```



Supplementary Figure 9: Results of the commute time query

	neighborhood	sumvalue
1	:#neighborhood137	"5320""xsd:integer
2	:#neighborhood136	"3275""xsd:integer
3	:#neighborhood135	"2010""xsd:integer
4	:#neighborhood139	"1775""xsd:integer

Supplementary Figure 10: Results of the public transit query

	neighborhood	sumvalue
1	:#neighborhood137	"7635""xsd:integer
2	:#neighborhood136	"3990""xsd:integer
3	:#neighborhood139	"2570""xsd:integer
4	:#neighborhood135	"2485""xsd:integer

ORDER BY DESC(?sumvalue)

By running the SPARQL query, we can see (Supplementary Figure 9) that out of all the NIAs managed by the Southeast Scarborough Planning Table, Neighbourhood 137 (Woburn) has the highest number of residents (5320) with a commute time of 60+ minutes while Neighbourhood 139 (Scarborough Village) has the lowest number of residents (1775) with a commute time of 60+ minutes.

CQ11: Which NIAs use public transit as the main mode of commuting?

The following query finds the value of the characteristic “public transit” for the census tracts in the designated neighbourhoods and outputs a table that displays the neighbourhood in one column and the summed value of the characteristic across all the census tracts in that neighbourhood in another column.

```

PREFIX uoft: <http://ontology.eil.utoronto.ca/tove/cacensus#>
PREFIX toronto: <http://ontology.eil.utoronto.ca/Toronto/Toronto#>
PREFIX iso21972: <http://ontology.eil.utoronto.ca/ISO21972/iso21972#>
SELECT ?neighbourhood (sum(?value) as ?sumvalue)
WHERE{
    ?neighbourhood toronto:hasCensusTract ?censustract.
    FILTER (?neighbourhood IN (toronto:neighbourhood135,
        toronto:neighbourhood136, toronto:neighbourhood137,
        toronto:neighbourhood139 ) )
    ?transit a uoft:PublicTransit2016;
    uoft:hasLocation ?censustract;
    iso21972:cardinality_of ?population;
    iso21972:value ?measure.
    ?measure iso21972:numerical_value ?value.
    ?population a ?populationclass.
    ?populationclass iso21972:defined_by
    uoft:PersonPublicTransit2016
}
GROUP BY ?neighbourhood
ORDER BY DESC(?sumvalue)

```

By running the SPARQL query, we can see (Supplementary Figure 10) that out of all the NIAs managed by the Southeast Scarborough Planning Table, Neighbourhood 137 (Woburn) has the highest number of residents (7635) who use public transit as their main mode of commuting while Neighbourhood 135 (Morningside) has the lowest number of residents (2485) who use public transit as their main mode of commuting.

CQ12: Which NIAs have the most households living in unsuitable dwellings (according to NOS)?

The following query finds the value of the characteristic “not suitable housing” for the census tracts in the designated neighbourhoods and outputs a table that displays the neighbourhood in one column and the summed value of the characteristic across all the census tracts in that neighbourhood in another column.

```

PREFIX uoft: <http://ontology.eil.utoronto.ca/tove/cacensus#>
PREFIX toronto: <http://ontology.eil.utoronto.ca/Toronto/Toronto#>
PREFIX iso21972: <http://ontology.eil.utoronto.ca/ISO21972/iso21972#>

```

Supplementary Figure 11: Results of the unsuitable dwelling query

	neighborhood	sumvalue
1	:#neighborhood137	"3535"^^xsd:integer
2	:#neighborhood139	"1320"^^xsd:integer
3	:#neighborhood136	"1220"^^xsd:integer
4	:#neighborhood135	"990"^^xsd:integer

Supplementary Figure 12: Results of the old dwelling query

	neighborhood	sumvalue
1	:#neighborhood137	"6010"^^xsd:integer
2	:#neighborhood136	"2255"^^xsd:integer
3	:#neighborhood139	"1430"^^xsd:integer
4	:#neighborhood135	"1180"^^xsd:integer

```

SELECT ?neighbourhood(sum(?value) as ?sumvalue)
WHERE{
  ?neighbourhood toronto:hasCensusTract ?censustract.
  FILTER (?neighbourhood IN (toronto:neighbourhood135,
    toronto:neighbourhood136, toronto:neighbourhood137,
    toronto:neighbourhood139 ) )
  ?notsuitable a uoft:NotSuitable2016;
  uoft:hasLocation ?censustract;
  iso21972:value ?measure.
  ?measure iso21972:numerical_value ?value.
}
GROUP BY ?neighbourhood
ORDER BY DESC(?sumvalue)

```

By running the SPARQL query, we can see (Supplementary Figure 11) that out of all the NIAs managed by the Southeast Scarborough Planning Table, Neighbourhood 137 (Woburn) has the highest number of households (3535) who live in unsuitable dwellings while Neighbourhood 135 (Morningside) has the lowest number of households (990) who live in unsuitable dwellings.

CQ13: Which NIAs have the highest number old dwellings?

The following query finds the value of the characteristic "1960 or before construction" for the census tracts in the designated neighbourhoods and outputs a table that displays the neighbourhood in one column and the summed value of the characteristic across all the census tracts in that neighbourhood in another column.

```

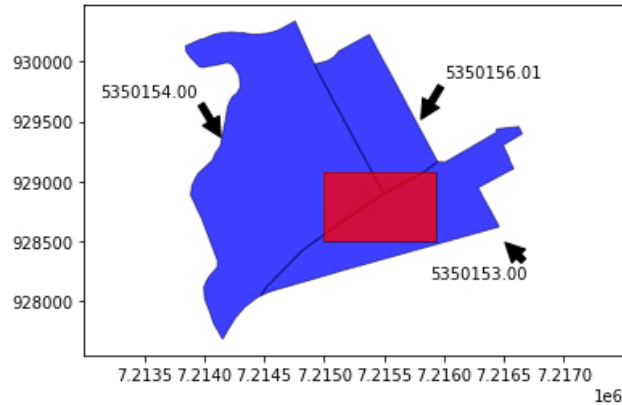
PREFIX uoft: <http://ontology.eil.utoronto.ca/tove/cacensus#>
PREFIX toronto: <http://ontology.eil.utoronto.ca/Toronto/Toronto#>
PREFIX iso21972: <http://ontology.eil.utoronto.ca/ISO21972/iso21972#>
SELECT ?neighbourhood(sum(?value) as ?sumvalue)
WHERE{
  ?neighbourhood toronto:hasCensusTract ?censustract.
  FILTER (?neighbourhood IN (toronto:neighbourhood135,
    toronto:neighbourhood136, toronto:neighbourhood137,
    toronto:neighbourhood139 ) )
  ?olddwelling a uoft:1960OrBeforeConstruction2016;
  uoft:hasLocation ?censustract;
  iso21972:value ?measure.
  ?measure iso21972:numerical_value ?value.
}
GROUP BY ?neighbourhood
ORDER BY DESC(?sumvalue)

```

By running the SPARQL query, we can see (Supplementary Figure 12) that out of all the NIAs managed by the Southeast Scarborough Planning Table, Neighbourhood 137 (Woburn) has the highest number of dwellings (6010) that were constructed before 1960 while Neighbourhood 135 (Morningside) has the lowest number of dwellings (1180) that were constructed before 1960.



Supplementary Figure 13: Matplotlib visualisation of three census tract geometries (blue) and one non-census geometry (red)



As the Canadian Census Ontology also includes geospatial information, we can use this information to approximate census characteristic values for non-census geometries. Statistics Canada publishes²⁵ cartographic boundary files for census areas in various file formats (e.g., SHP, GML) and these files contain the polygon shapes for representing the geometry of census areas. We can then use functions from Python packages, such as Shapely, to calculate how much an administrative area (e.g., Ward, Neighbourhood) overlaps with the census tract areas.

CQ14: How much of a given geometry overlaps with the census tract boundaries?

Although the SPARQL query language does not have any built-in support for calculating overlapping geometries, programming languages, such as Python, can be used for handling geospatial work. Although there has been increased interest and use of grids and gridded data for aggregating new/future census data into small and uniform land areas, much of the historical census data is published for administrative areas so there may be instances where there is a need for calculating intersecting areas. In Supplementary Figure 13 shown above, a red, non-census geometry polygon overlaps with three different blue census tract polygons. Using the intersection function from the Shapely Python package, we can calculate the area of each overlapping section and divide them by the area of the corresponding census tract polygons in order to obtain the percentages. In this example, the red polygon occupies 34% of the total area of the 5350153.00 census tract, 7% of the total area of the 5350154.00 census tract, and 6% of the total area of the 5350156.01 census tract. With this, we can then multiply the numerical value of the census characteristics for these three census tracts by the percentages in order to obtain an approximate value of the characteristic for the red polygon. For instance, the 5350153.00 census tract has a population of 4462 people, the 5350154.00 census tract has a population of 6157 people, and the 5350156.01 census tract has a population of 5688 people. Multiplying these numbers by their corresponding percentages would give us an approximate population of 2289 for the area represented by the red polygon. Thus, this approach would allow us to approximate values of census characteristics for non-census areas.

²⁵<https://www12.statcan.gc.ca/census-recensement/2011/geo/bound-limit/bound-limit-2016-eng.cfm>.