Web Semantics: Science, Services and Agents on the World Wide Web 48 (2018) 48–65

Contents lists available at ScienceDirect

Web Semantics: Science, Services and Agents on the World Wide Web

journal homepage: www.elsevier.com/locate/websem

The semantics of populations: A city indicator perspective

Mark S. Fox

Mechanical & Industrial Engineering, University of Toronto, 5 King's College Rd, Toronto ON M5S 3G8, Canada

ARTICLE INFO

Article history:
Received 9 February 2017
Received in revised form 24 October 2017
Accepted 8 January 2018
Available online 31 January 2018

Keywords:
Population semantics
Ontology
City indicators
Consistency checking

ABSTRACT

This paper addresses the question of how to represent the semantics of populations. 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 modeling perspective. The motivation for this work is the development of ontologies for the representation of city indicator definitions. A city indicator measures the performance of a city in areas such as education, transportation and the environment. The definitions of city indicators rely on definitions for populations of people, built form, events, activities, and sensor measurements. This paper provides a model for representing membership extent, temporal extent, spatial extent, and measurement of populations. It demonstrates the approach by representing the definitions of city indicators as defined by ISO 37120, the interpretation of these definitions by cities, and their comparison to ascertain whether a city's interpretation is consistent with the standard.

Crown Copyright © 2018 Published by Elsevier B.V. All rights reserved.

1. Introduction

Cities use a variety of metrics to evaluate and compare their performance. With the introduction of ISO 37120 (ISO37120, 2014), which defines 100 indicators for measuring a city's quality of life and sustainability, it is now possible to consistently measure and compare cities, assuming they adhere to the standard. The majority of the indicators in this standard are defined as a ratio of parameters of two populations. For example, the Primary Student Teach Ratio indicator is the ratio of the size of the population of students to the size of the population of teachers. By populations, we are not referring to people, but in a statistical sense to a finite collection of "things" under consideration. For education indicators, the populations include students and teachers. For environment indicators, the populations include observations generated by sensors at different points in time and different locations. For fire and safety indicators, the populations include 911 call events.

As the data used to derive indicators is made available on city open data sites, it enables the development of software applications that will aid in the analysis of city performance. In particular, it becomes possible to automate the longitudinal analysis (i.e., how and why a city's indicators change over time) and transversal analysis (i.e., how and why cities differ from each other), in order to discover the possible causes of differences.

But the assumption that cities will adhere to the standard is a strong one, as cities often interpret definitions differently [1,2]. Before any meaningful analysis can be performed, three questions with respect to consistency need to be answered: Is a city's interpretation of an indicator:

1. Definitionally consistent, e.g., is the definition of student and teacher populations reported by a city consistent with the indicator's definitions?
2. Intra-indicator consistent, e.g., are the student and teacher populations in the indicator from the same time and location?
3. Inter-indicator consistent, e.g., are the city's definitions of student and teachers, used to specify the populations, consistent across time?

In our analysis of city indicators, inconsistency in the interpretation of population definitions lies at the heart of many differences in performance and without the representation of the semantics of populations, detection of these inconsistencies remains an arduous, manual process.

This paper 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[1] but is essentially silent on the representation of population definitions from a data modeling perspective.

Existing survey, statistics, census and indicator vocabularies/ontologies do not address the issue of how to represent the definition of populations. At best they can represent individuals

E-mail address: msf@eil.utoronto.ca.

[1] "The population must be fully defined so that those to be included and excluded are clearly spelt out (inclusion and exclusion criteria)". [3].

https://doi.org/10.1016/j.websem.2018.01.001
1570-8268/Crown Copyright © 2018 Published by Elsevier B.V. All rights reserved.

that are members of populations, but do not define what are admissible members of a population.

We take an ontology engineering approach to representing population definitions where the expressiveness of Description Logic is used to define what are admissible members of a population being measured. The results of this work form the core of an ontology for representing indicator definitions [4], which is used to determine the consistency of a city's indicators.

A second potential use of population semantics is in the harvesting of city data that is required to calculate a particular indicator. Data harvesting focuses on the collecting, cleaning, integrating and enriching (e.g., inferring missing values) of open data.

The work reported here and in [4] is the basis of a new ISO standards project to create an ontology for representing the definitions of city indicators (ISO/IEC AWI Project 21972[2]). This is part of ISO/IEC Joint Technical Committee 1 Working Group 11 on Smart Cities.

In the following, we explore the semantic requirements for representing population definitions as found in city indicators. We then review approaches in the literature to defining statistical populations. Next we provide an overview to the PolisGnosis project (Fox, 2017), that provides the context in which this research was conducted, and its ontology design pattern for indicators. We then define our semantics for populations, followed by an example that demonstrates its application to the representation of ISO 37120 indicator populations. Finally, we evaluate and discuss the ontology.

2. City indicators: Measuring city performance

ISO 37120 "Sustainable development of communities – Indicators for city services and quality of life" defines 100 indicators divided into 17 themes, including Education, Energy, Health, Safety, Finance and Shelter. Each indicator contains a definition that reduces ambiguity of interpretation by cities, leading to greater consistency in measurement and comparability across cities. An example of an indicator definition is Education theme indicator "Primary Student Teacher Ratio":

> "The student/teacher ratio shall be expressed as the number of enrolled primary school students (numerator) divided by the number of full-time equivalent primary school classroom teachers (denominator). ...
> Private educational facilities shall not be included in the student/teacher ratio.
> One part-time student enrollment shall be counted as one full-time enrollment; ...
> The number of classroom teachers and other instructional staff (e.g. teachers' aides, guidance counselors), shall not include administrators or other non-teaching staff. Kindergarten or preschool teachers and staff shall not be included.
> The number of teachers shall be counted in fifth time increments, …". (ISO37120, 2014, p. 9–10)

Contained in this indicator is the definition of two populations: student and teacher. For each population there are strict constraints on who is to be included in the populations. For example, the teacher population does not include administrative staff, nor kindergarten/preschool teachers. Secondly, based on the context in which the definition appears, it is assumed (though not stated directly in the definition), that the populations are to be drawn from the same city and at the same time.

A second example is the Financial theme indicator "Debt Service Ratio (debt service expenditure as a percentage of a municipality's own-source revenue":

> "Debt service ratio is the ratio of debt service expenditures as a per cent of a municipality's own source revenue. Debt service ratio shall be calculated as the total long-term debt servicing costs including lease payments, temporary financing and other debt charges (numerator) divided by total own source revenue (denominator). .... Total own source revenue shall be calculated as the total revenue less transfers". (ISO37120, 2014, p. 21)

The two populations are long-term debt servicing costs and total own source revenue. In the latter case, own source revenues exclude transfers (e.g., from the province/state). Each population is composed of a variety of financial transactions/instruments that must conform to the indicator definition. From a temporal perspective, debt and revenue are aggregated over the year of the indicator.

A third example is the Shelter theme indicator "Number of Homeless per 100 000 Population":

> "The number of homeless per 100 000 population shall be calculated as the total number of homeless people (numerator) divided by one 100 000th of the city's total population (denominator). ....
> ... Absolute homelessness refers to those without any physical shelter, for example, those living outside, in parks, in doorways, in parked vehicles, or parking garages, as well as those in emergency shelters or in transition houses for women fleeing abuse". (ISO37120, 2014, p. 39)

The definition of the homeless population provides examples, but is not definitive, leading to differences in interpretation across cities. Secondly, homeless counts in cities are not conducted daily. Many cities may conduct it once a year or once a season.

In each of these examples, variations in the interpretations of the definitions of the numerator and denominator populations by cities can lead inconsistencies, which in turn leads to anomalous results. If we are to automate the forensic analysis of city performance, it is necessary to provide the automation with a precise and unambiguous representation of an indicator's definition, and how the definition was interpreted the city (or cities) being analyzed. Equally important is that cities provide the provenance of the data used to calculate an indicator.

3. PolisGnosis project and architecture

We contextualize our work on the semantics of populations within the PolisGnosis project (Fox, 2017). The goal of the PolisGnosis project is to construct an analysis engine that can diagnose a city's performance. It will automate the longitudinal analysis, i.e., how and why a city's indicators change over time, and transversal analysis, i.e., how and why cities differ from each other, in order to discover the possible root causes of differences.

We wish to create a "universal" analysis engine that is not tailored to specific indicators or cities. Therefore the design of the PolisGnosis analysis engine must satisfy the following requirements:

1. **Indicator Independence.** Since there are a vast number of indicators used by cities, beyond those defined in the ISO 37120 standard, and ISO standards evolve over time, we do not want our analysis engine to have any knowledge of indicator definitions "hardwired" into its code. An indicator's definition must be an input to the analysis engine.
2. **City Independence.** In order to achieve city independence, we need to know two things: (1) how did the city interpret an indicator definition? Do they define a teacher differently or a homeless shelter differently than the standard? We need the city's interpretation of an indicator as input for analysis. (2) Cities openly publish vast amounts of data that our analysis engine would like to use. But the data lacks any

[2] https://www.iso.org/standard/72325.html

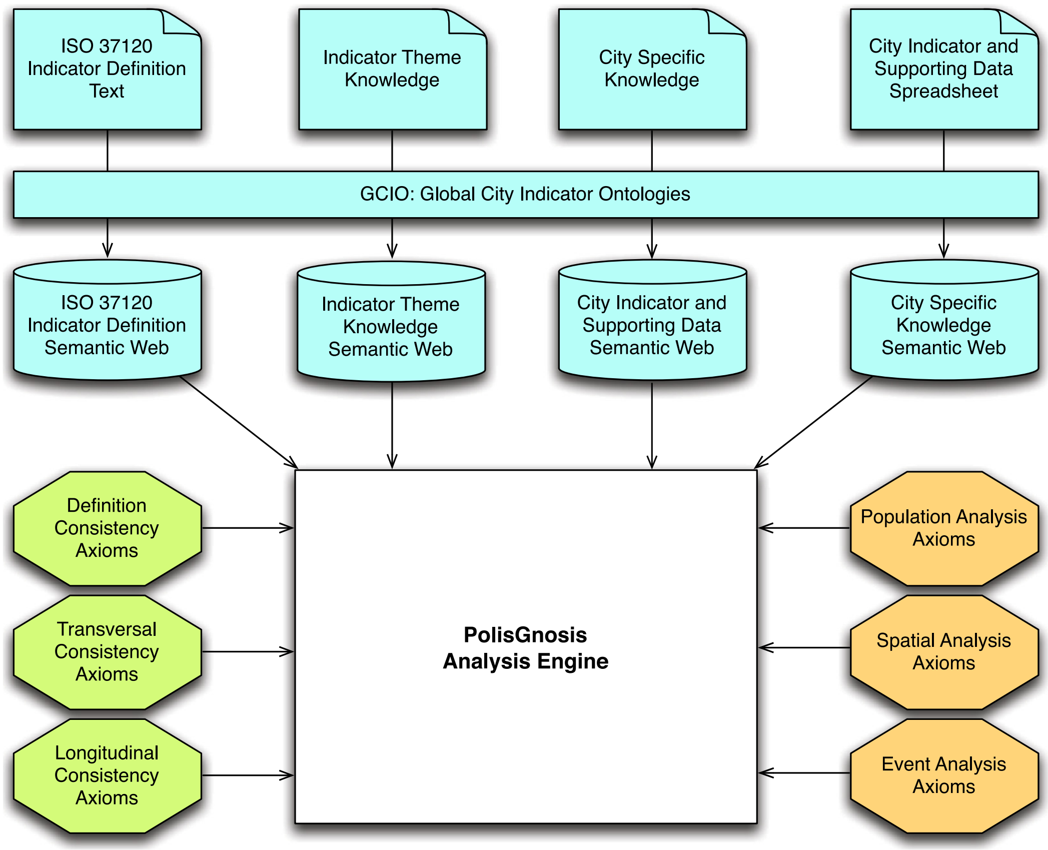

Fig. 1. PolisGnosis architecture.

standard models or vocabularies — every dataset differs in structure, attributes and content. It would be practically impossible to construct an analysis engine that can understand these datasets. Hence the analysis engine will assume that cities will adopt or translate the data used to compute their indicators into a standard for representing the information used to derive its indicators.

3. **Analysis Independence.** Given the variety of indicators and the ensuing variety of data used to derive them, a variety of methods of analysis may also be required. Rather than hardwire these methods into PolisGnosis, it would be better if they too were inputs to the analysis engine.

Due to these requirements, the PolisGnosis analysis engine must have as input:

- **Indicator Definition**: The definitions of city indicators as provided by some standard,
- **Indicator Theme Knowledge:** Theme specific common sense knowledge, required to interpret the indicator definitions,
- **City Specific Knowledge:** definitions, including how they define theme specific classes and the indicators that use them, for a particular year,
- **City Data:** used to derive the city's indicators, and
- **A set of analysis axioms**.

Fig. 1 depicts the architecture of PolisGnosis. The boxes above the PolisGnosis analysis engine focus on the representation of indicators and the data used to derive them. It is assumed that all definitions and relevant city data are transformed into a standard representation that can be published on the Semantic Web. Our analysis engine uses the Global City Indicator (GCI) ontologies [4].[3]

The boxes to the left of the analysis engine focus on consistency analysis of a city's interpretation of the indicators to verify that they are consistent with the definitions of the indicators, and that indicators for a city over time or comparing two cities are consistent. The boxes to the right of the analysis engine will focus on theories to diagnose the possible root causes of longitudinal and transversal differences.

3.1. GCI ontology

The GCI Foundation ontology [4–6] provides the ontology design patterns for representing indicators and indicator metadata. Fig. 2 depicts a portion of the generic indicator pattern. The ISO 37120 indicator (usually) has associated a unit of measure that is the ratio of two populations, and the year it was measured. It is composed of a numerator and denominator. Each is a quantity that is a measure of population. The population members are defined, in this example, by another class and city.

Fig. 3 depicts the hierarchy of ontology modules used to construct ISO 37120 indicator definitions. At the lowest level are the foundation ontology modules. The next level is the Enterprise level containing an Organization ontology. Next is the GCI Ontology level where each theme specific ontology is located, e.g., Education, Energy, Finance, etc. Finally the ISO 37120 Ontology level contains the modules in which the indicator definitions are specified. A separate module is defined for each theme of indicators.

In order to construct an indicator definition using an existing GCI theme ontology such as education, the theme ontology needs to be imported into the ontology that will define the new indicator.

[3] The latest version of the theme ontologies can be found at: http://ontology.eil.utoronto.ca.

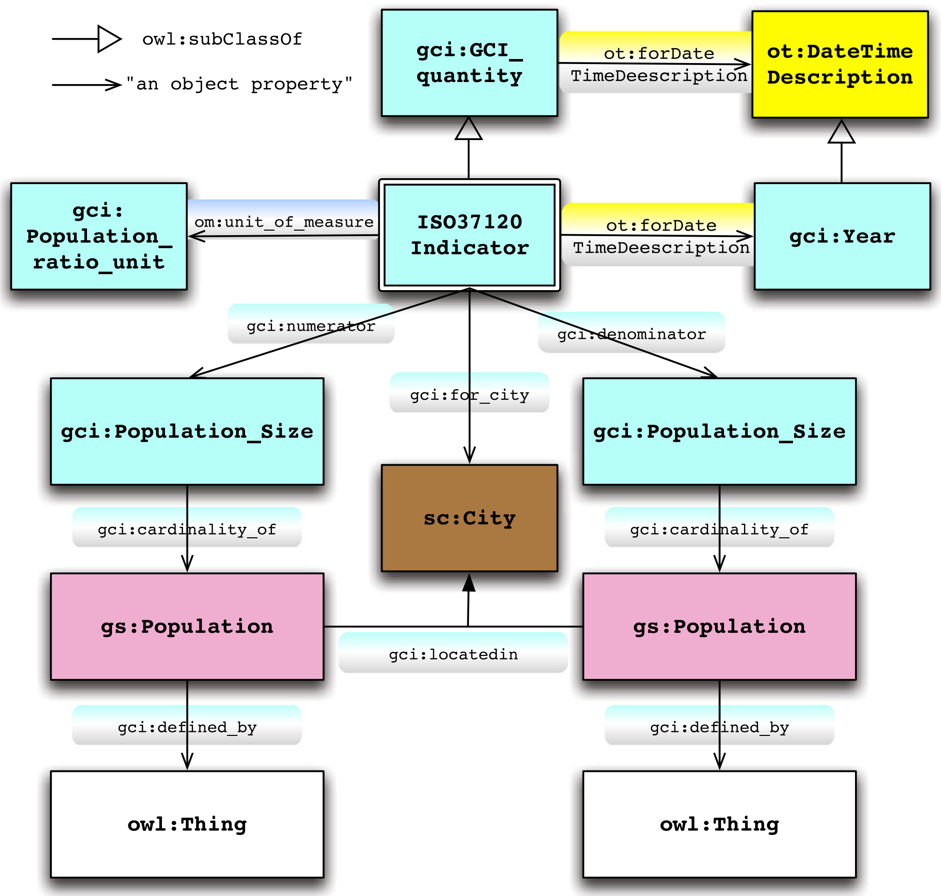

Fig. 2. Indicator design pattern.

Appendix B lists the URLs for the currently available GCI theme ontologies.[4] If indicators are to be constructed for a new theme, whether ISO 37120-based or not, then a new theme ontology needs to be defined, starting by importing the GCI Foundation ontology.[5] Once the new theme ontology is created, then the new indicators can be defined.

3.2. Analysis

The analysis we focus on in PolisGnosis is to determine whether inconsistencies exist between a standard indicator definition (e.g., ISO37120) and the interpretation of the definition used by the city, for a specific year. The city's version of how they interpret an indicator definition is provided in the City Specific Knowledge input. It may contain specializations of generic theme ontology classes (e.g., their definition of a teacher or shelter may differ from the standard), and how they construct their interpretation of the indicator definition using the generic and custom classes.

It is simple to detect a change or difference in an indicator's value, but the issue is what led to the change? Diagnosing the possible root cause of differences is a challenge because the definitions are composed of many types, including: quantitative (e.g., counts and ratios), statistical (e.g., Population sampling), logical (e.g., definitions of students, teachers), spatial (e.g., city boundaries) and temporal. Sources of change can occur anywhere.

Consistency analysis can provide insight into root causes. For example, definitional inconsistencies may exist between a city's definition of a population and the ISO definition (e.g., adding special needs students to the student population). Longitudinal inconsistencies may exist due to changes in a city over time (e.g., city boundary changes over time).

What remains is to determine whether these inconsistencies are the root causes of change. There may be other possible root causes of change not addressed by consistency analysis. Significant events can play a major role in changes in indicator values. For example, hurricanes can affect economic indicators, hosting a major international sporting event can affect recreation indicators, and immigration can affect shelter indicators. Identifying events that correlate with changes in indicator values is an important step towards determining root causes.

Given the standard indicator definition and a city's version of it, it is necessary for performing consistency and root cause analysis to have a precise definition of populations. Without a clear semantics for populations, we are unable to perform consistency analysis. Without determining consistency, we lack the ability to validate any comparisons based on indicators.

3.3. City Indicator Consistency Checker (CICC)

There are three categories of consistency analysis that PolisGnosis is able to perform based in part on the Population ontology.

The first category, definitional consistency, determines whether a city's published indicator data is consistent with the indicator's (e.g., ISO 37120) definition. Consider the primary student teacher ratio example. Is the definition published by the city consistent with the ISO definition of "public school"? Are the grades included consistent with "primary grades" definition? Are teachers who are included consistent with the ISO definition of teachers, e.g., they

[4] The latest versions of available theme ontologies are listed at: http://ontology.eil.utoronto.ca.

[5] http://ontology.eil.utoronto.ca/GCI/Foundation/Foundation-v2.owl.

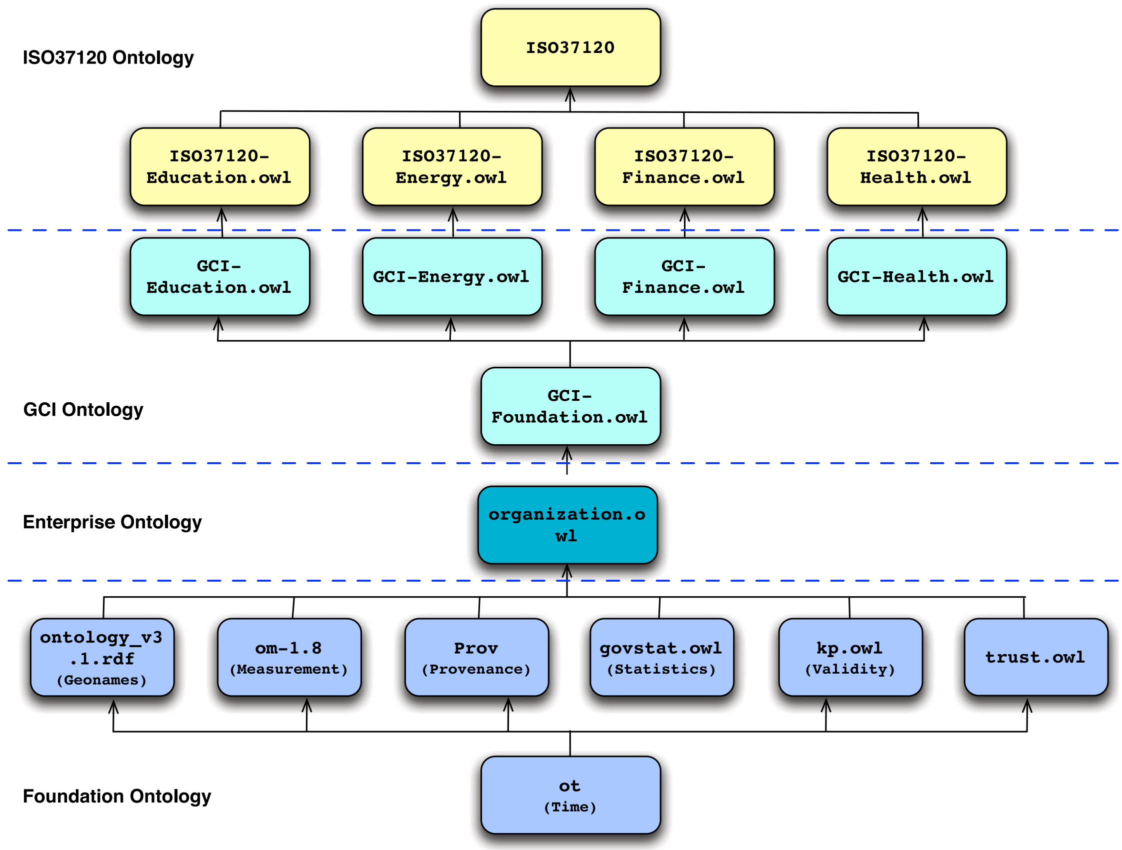

Fig. 3. Module interactions.

cannot be administrators? Existing DL reasoners can detect many types of definitional inconsistencies, but not explain the type of inconsistency [7].

The second category of consistency, intra-indicator consistency, determines whether the data a city used to derive an indicator is internally consistent? One type of intra-indicator consistency is temporal consistency. For example, was the count of the number of students for the same time period as the count of the number of teachers? Another type is spatial consistency. For example, were the counts for both students and teachers for the same geographic areas? A third type is measurement consistency. For example, are the units for both the count of students and teachers of the same units? Could one be in 100 s and the other 1000 s? Existing DL reasoners do not detect this category of inconsistencies.

The third category of consistency, inter-indicator consistency, determines whether the published indicator definitions for two cities being compared are consistent. If we assume that each city's indicator data are definitional and intra-indicator consistent, what types of inconsistency can arise? Consider longitudinal consistency analysis (comparing the same indicator for the same city at two different times). Temporal–spatial consistency problems arise when the geographic boundaries of the city change in the intervening time between the measuring of the indicators, e.g., merger of suburbs into the city. Temporal-definitional consistency problems arise when the definition of entities change over time, e.g., definition of student changes to include special needs students. Event consistency problems arise when a significant event occurs during the intervening time. For example, climatic events such as a hurricane may have an effect. Existing DL reasoners do not detect this category of inconsistencies.

Some inconsistencies may only be potential (vs actual). Inter- or intra-indicator spatial or temporal inconsistencies arise where there are spatial or temporal overlaps of the populations measured. But the existence of an overlap may be sufficient for the consistency of an indicator. For example, if the students are measured during the period October to November, and Teachers are measured from November to December, the overlap may be sufficient to provide acceptable results.

The PolisGnosis City Indicator Consistency Checker (CICC) is able to determine whether city indicator data is definitional, intra-indicator and inter-indicator consistent [7,8]. The CICC takes as input an indicator definition, the indicator's theme common sense knowledge, a city's interpretation of the indicator definition for a particular year, including modifications to theme knowledge, a city's indicator value and possibly the data used to derive it, all represented using the Population and GCI ontologies.

The CICC performs definitional consistency analysis by recursively comparing an indicator definition to the city's interpretation of the indicator definition for a specific year. Starting at the top of each indicator pattern, each node of the definition is compared to the corresponding node in the city interpretation. At each node it checks for type, temporal, place and measurement consistency. It determines whether the inconsistency is actual or potential. For example, in checking the homeless indicator, the root node (iso37120:15.2) of the ISO37120 definition in Fig. 4 is compared to the root node (15.2_trt_2013) of the city's definition in Fig. 5, to see if time, place, and units of measure are consistent. But when checking the numerator populations, it has to check that Homeless_Person, which defines the population in the ISO standard, subsumes trt_homeless_person_2013, which defines the population for Toronto.

Some types of definitional consistency can be detected using existing reasoners, such as subsumption analysis to detect differences in the definition of members of a population. We chose to

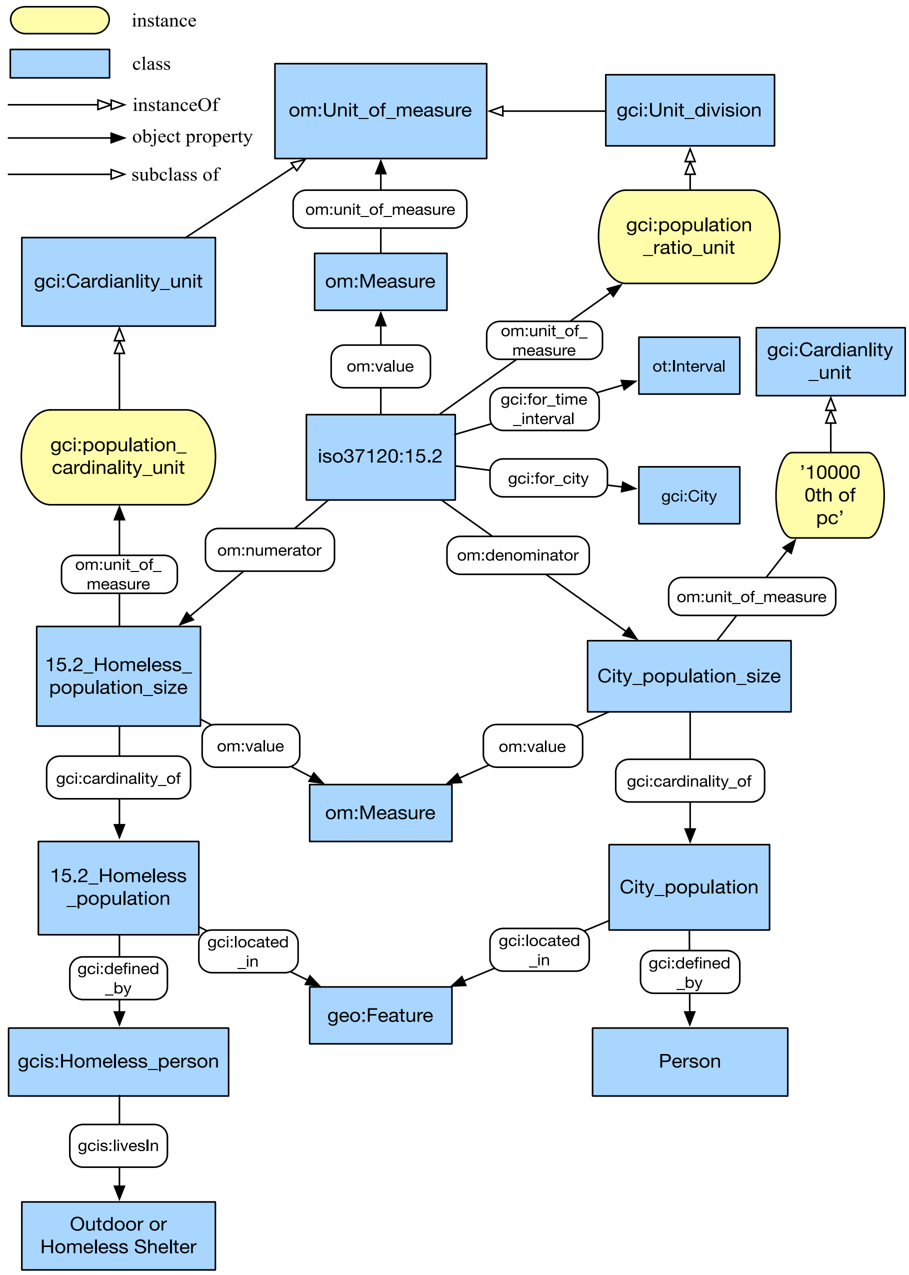

Fig. 4. Ratio of homeless to city population.

build our own consistency checker (in prolog) because existing reasoners do not provide adequate explanation of inconsistencies. Nor can they perform intra and inter-indicator consistency checking due limits in the expressiveness of Description Logic. Nor can they detect potential inconsistencies.

4. Related work

4.1. Surveys

Surveys are a major source of data used to compute indicators. The question is to what extent do they support the use of ontologies to specify the Population they are surveying and the members of the Population surveyed?

There have been a number of efforts to define vocabularies for representing survey data. DISCO [9] builds on the Data Documentation Initiative (DDI) standards for representing social sciences datasets (www.ddialliance.org). DISCO represents a survey variable as a class. Though it has the potential to link a variable to another class using the "concept" property, their example does not use this capability and is focused on text definitions. In addition, meta data, including temporal and spatial, can be defined for the dataset. Specification of the members of the population being surveyed (referred to as "Universe") is specified as text.

A draft specification of a subset DDI in RDF has been posted on the web [9]. It is a direct translation of a subset of DDI into RDF triples. The translation does not attempt to redesign the elements to take advantage of web semantics such as the OWL version of Description Logic, nor does it attempt to map variables and concepts to any domain ontologies.

Fox & Katsumi (2016) define a survey ontology that represents the content and logic of surveys, and provides a mapping of questions and answers on to domain specific ontologies, enabling their semantic interpretation.

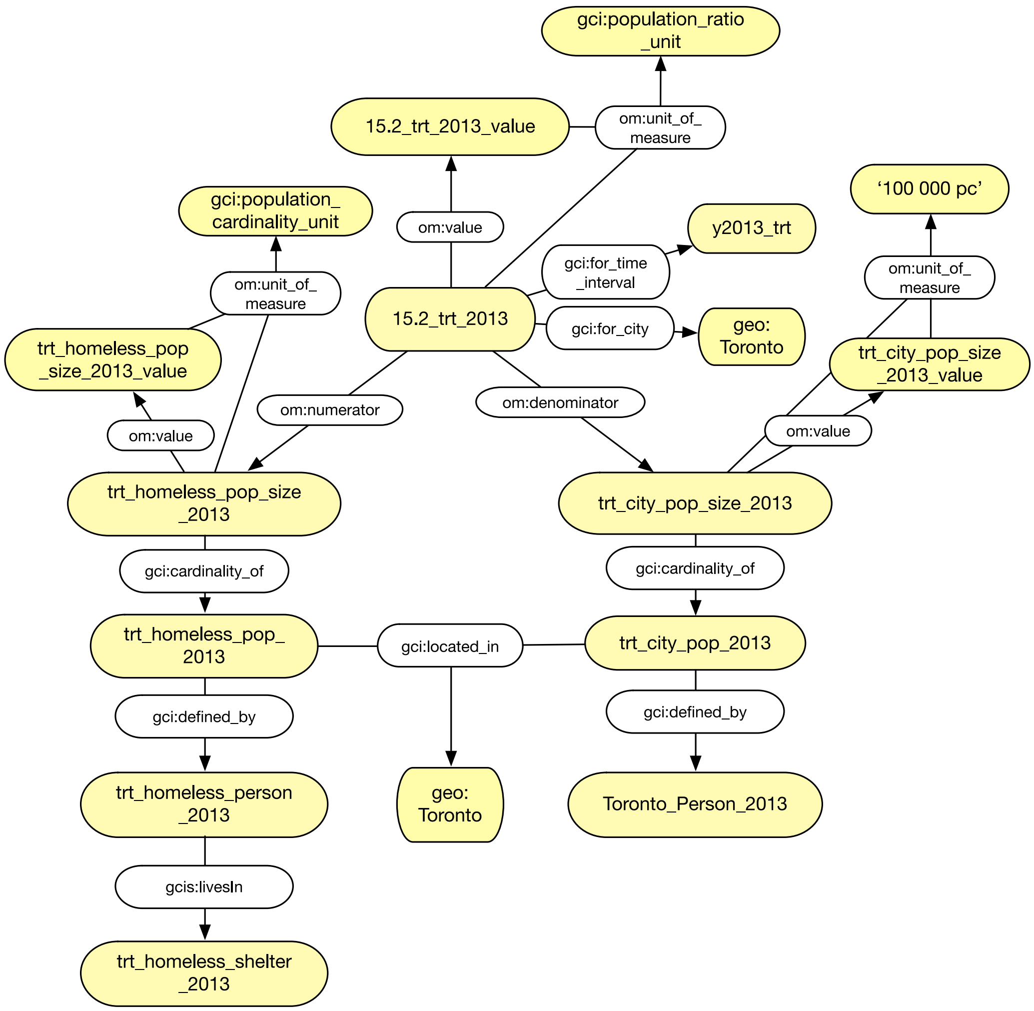

Fig. 5. Toronto homeless indicator pattern for 2013.

4.2. Statistical datasets

More generally, there exists vocabularies/ontologies for the representation and publishing of statistical datasets. SDMX [10] uses RDF triples to represent the observations that comprise the dataset. Its representation includes a mapping from a "ComponentProperty" to a Concept, thereby enabling a semantic interpretation of the property. Petrou et al. [11] define a vocabulary for publishing census data. Their core concept is an Observation that is linked to a dataset, geocode and value.

The RDF Data Cube Vocabulary [12] provides a RDF representation for representing datasets. It provides terms for specifying the dimensions of a dataset, what is being measured and its units of measure, and individual observations. It uses the Dublin Core terms [13] for specifying meta-data for a dataset.

4.3. Statistics

A different approach to a statistics ontology is the GovStat ontology [14]. GovStat provides the classes and properties for representing basic statistical concepts. Fig. 6 depicts a subset of the GovStat ontology relevant to our task.[6] The core class is the 'Population' to be measured. A 'Population' is linked to a 'Parameter' (e.g., mean, standard deviation) by the 'is_described_by' property.

Statistical sampling is represented by the 'Sample' class, and the parameter being measured is represented as a subclass of 'Statistic'. The variable for which the parameter is being measured is defined by the class 'Observation' which 'Statistic' links to via the property 'is_composed_of', and the actual variable which is a subclass of 'Variable' is linked to 'Observation' via the property 'is_a_characteristic_of'.

There is movement towards publishing city and country statistics as linked open data. Eurostats (ec.europa.eu/eurostat) recently convened a workshop[7] where it was suggested that the W3C Data Cube ontology be used.

4.4. Indicators

The Health Indicators Ontology[8] defines a taxonomy of indicators, such as 'MortalityRateIndicator'. It uses the PHOnt ontology,[9] which specifies that an indicator is a characteristic of a population whose members (persons) can have personal characteristics such as housing quality. In the PHOnt ontology, a population is a collection that identifies the persons who are members, versus defining population membership. The CitySDK Project [15,16] provides APIs for accessing city information, with tourism being a major focus.

6 In the figure, the "gs" prefix is used to denote the GovStat ontology defined at http://ontology.eil.utoronto.ca/govstat.owl

7 https://ec.europa.eu/eurostat/cros/ESS-LOD-2017_en

8 http://surveillance.mcgill.ca/km/Indicators/HealthIndicators.owl

9 http://surveillance.mcgill.ca/km/PublicHealth/phont.owl

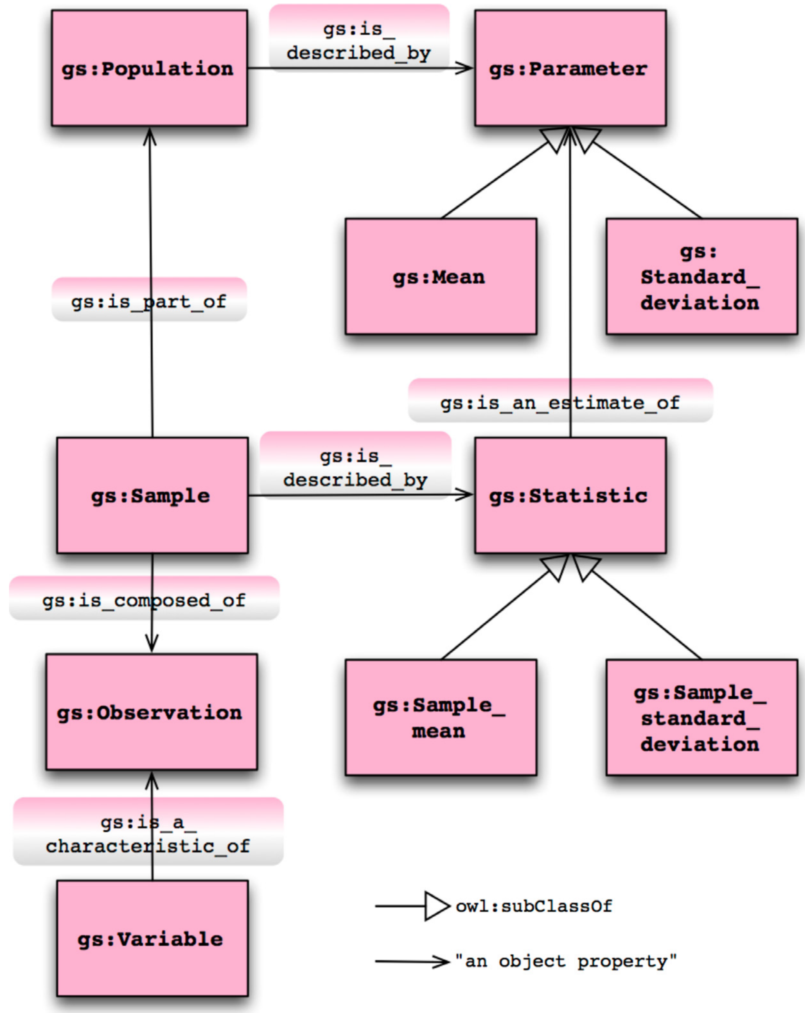

Fig. 6. GovStat statistics ontology.

The main concepts are Points of Interest, Events and Itineraries. It has been adopted by several cities in Europe.

The Global City Indicator Ontologies [4–6] (Fox, 2017) provide both a foundation ontology for representing indicator definitions, and a set of ISO 37120 theme specific ontologies (see Appendix B for a list) for representing a theme's common sense knowledge.

4.5. Harvesting city data

Cities are publishing vast amounts of data via their open data portals, for the majority of these datasets standard formats and vocabularies do not exist, the data is very incomplete, at least from an indicator perspective [17], it is often sparse and contains errors. Consequently, there is a gap between what is published and what can be used. Research that addresses this gap can be characterized as the *harvesting* of open data, which focuses on the collecting, cleaning, integrating and enriching (e.g., inferring missing values) open data.

The first stage of data harvesting is the collection of datasets. Country level efforts have been underway for several years to create government data repositories. Two examples are the Urban Big Data Centre open data repository at Glasgow University (ubdc.gla.ac.uk/dataset), and the Australian Urban Research and Infrastructure Network portal (aurin.org.au). An example of a commercial effort to catalog mostly municipal level open datasets is Namara (namara.io), which contains over 100 000 datasets from across North America.

The second stage is the transformation of the datasets into a common format. The Open City Data Pipeline Project [18,19] focuses on "collecting and integrating quantitative indicators about cities, including basic statistics such as demographics but also socio-economic and environmental information". It uses the RDF Data Cube Vocabulary to represent datasets and their observations, and is supplemented with the PROV ontology [20] for representing provenance.

The third stage focuses on the quality of the data. For one type of quality problem, i.e., missing data, the Open City Data Pipeline use machines learning techniques to infer missing values in a dataset. There exist other methods for addressing quality issues in the database literature, such as integrity constraints and data quality rules [21].

The fourth stage focuses on semantic mapping, namely mapping classes and properties, onto existing ontologies and vocabularies. A review of this literature can be found in [22].

Given an integrated database, it is then possible to perform queries that incorporate constraints on what is retrieved. For example, it is possible to retrieve only the data restricted to a city during a particular year, assuming the database has dimensions referring to time and place. For example, it is then possible to sift through open data that is relevant to a particular indicator.

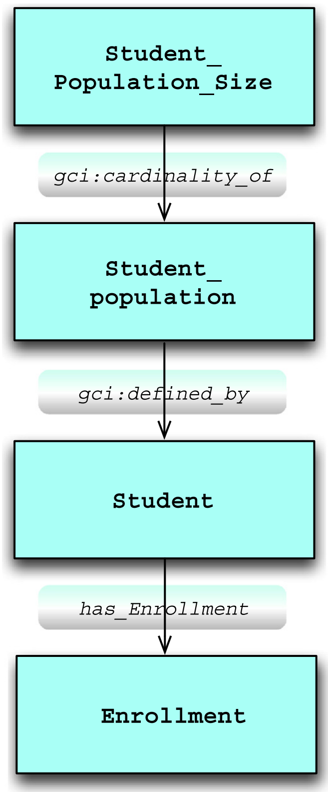

Fig. 7. Student population definition.

4.6. Summary

None of the survey, statistics, census, harvesting methods and indicator vocabularies/ontologies address the issue of how to explicitly define a population — except the GCI Foundation ontology. At best they can represent individuals that are members of populations, Or they can represent constraints on what is retrieved during the harvesting process. Neither provide an explicit representation of the definition of a population that can be used for consistency analysis. The next section describes the population definition portion of the GCI Foundation ontology.

5. Semantics of populations

Measuring parameters of populations lies at the heart of city indicators. By populations, we are not referring just to people, but in a statistical sense to a finite collection of "things" under consideration. For education indicators, the populations include students and teachers. For environment indicators, the populations include sensor readings taken at different points in time. For fire and safety indicators, the populations include 911 call events. In the following, we deconstruct the semantics of populations so that we can precisely model the definition of an indicator and its supporting data. We extend the GovStat ontology to represent the semantics of populations. A subset of the GovStat ontology is recreated directly into our ontology.[10] The GovStat classes and properties that are reused have the prefix of "gs".

[10] The original GovStat ontology does not have a publically available owl file.

5.1. Membership extent

The first question we have to answer is what defines the members of a population? Consider the population of teachers. The ISO 37120 definition states:

> "The number of classroom teachers and other instructional staff (e.g. teachers' aides, guidance counselors) shall not include administrators or other non-teaching staff. Kindergarten or pre-school teachers and staff shall not be included".

The authors of this definition assume that the reader understands the terms teacher, classroom, instructional staff, guidance counselors, etc. In other words, they assume that the reader has a "common sense" understanding of the domain of education. Hence, to model this population, we have to link the concept of 'Population' to the concepts in an education ontology (or at least the portions of the education domain relevant to the ISO 37120 Education theme) that defines its members.

To accommodate the definition of a population's member, we add the following property depicted in Fig. 8:

- defined_by: This property links a 'Population' to a class that defines the necessary and/or sufficient conditions to be a member of the population. Note that the range can be any subclass of 'owl:Thing'.

Continuing with our education example, if we want to define a population of students, we link the 'Population' class to the 'Student' class using 'defined_by'. The definition of 'Student' will then have to be rich enough to determine membership in the 'Population', as specified by the ISO 37120 Education theme. Fig. 7 shows a portion of the numerator of the indicator pattern specialized for the student population. The student population is 'defined_by' the 'Student' class. The Description Logic definition of the 'Student' class provides the necessary and/or sufficient conditions for membership in the student population. One of those conditions is that it has a 'has_enrollment' property linking to an 'Enrollment' class. Fig. 13 depicts the 'Student' class which is described in more detail in Section 6.

5.2. Spatial extent

A second semantic requirement for the representation of indicator populations is the specification of the physical area from which the population is drawn. In theory, it should be an entire city, but in practice it may be drawn from a larger or smaller area. There are two approaches to identifying a population's spatial extent. The first identifies a physical area by a place name, aka toponym, and the second by geo-spatial location.

Toponyms can not only identify cities, but are used to identify any type of place, including villages, neighborhoods, parks, buildings, hills, hospitals, and schools. But with the toponym approach, there is an additional requirement that we be able to uniquely identify a place. Names of cities are often duplicated throughout the world. How do we know that when we refer to Toronto, that it is the Toronto in Ontario, or the Toronto in Ohio or California? The semantic web provides a solution, namely each city or town of Toronto has an IRI that uniquely identifies it. Geonames[11] is both a web site and a namespace. It contains over ten million place names, each having a unique IRI.[12]

In parallel with the development of Geonames, the Schema.org taxonomy has been developed to tag web page content with

[11] Note that other toponym datasets exist such as LinkedGeodata (http://linkedgeodata.org), and can be substituted for Geonames.

[12] For example, the geoname IRI for the City of Toronto Ontario is http://www.geonames.org/6167865.

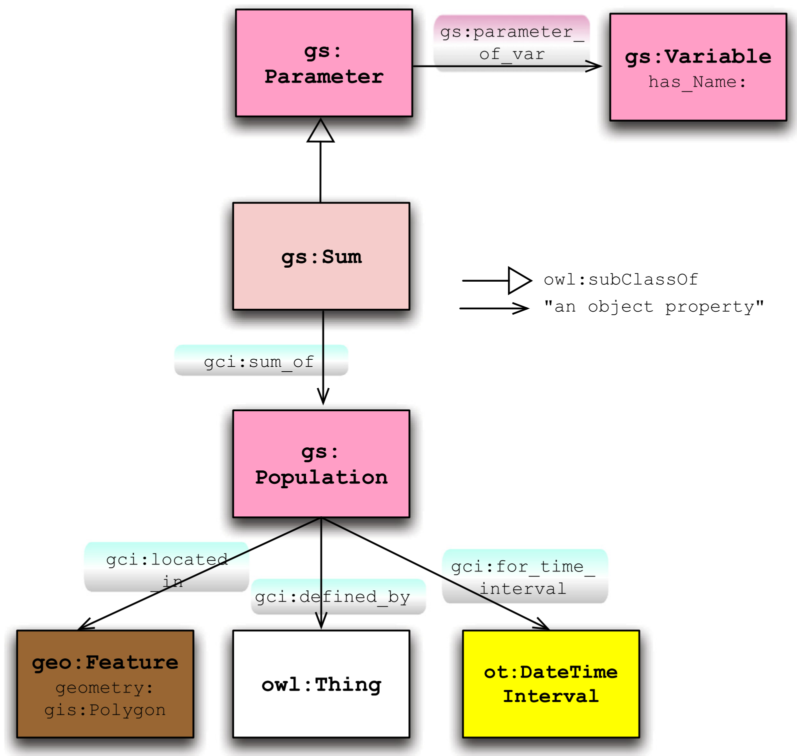

Fig. 8. Population properties and integration of statistics.

associated classes and properties, thereby making it easier for search engines to index content. Schema.org provides a rich set of place related classes, e.g., City, State, Airport, Mountain, that can be used to extend Geoname Features. This is accomplished by making a specific IRI for a place like Toronto an instance (type) of a schema.org City.

The second method of identifying the spatial extent of a population is to specify its geospatial location. This can be accomplished by specifying a polygon whose vertices are specified as latitude and longitude. A number of geometry ontologies exist, including OpenGIS (www.opengis.net) which provide for the representation of geometric classes such as Point, Line, Circle, and Polygon, and relating them using properties such as intersects, overlaps, contains, and disjoint. We can integrate place names with geospatial information by adding the 'geometry' property to 'geo:Feature' and whose range is a Polygon. We can associate with the place name the geospatial polygon that bounds the city, and relate it to other place names using geometry.

Fig. 8 depicts the integration of both place name and geo-spatial information with 'Population'.[13] The property 'located_in' is added to 'Population' to specify the IRI of the place name that includes the 'geometry' property that specifies its spatial geometry.

5.3. Temporal extent

A third semantic requirement for the representation of indicator populations is the specification of the time period over which the population is drawn. The periods over which populations are measured may vary by starting time, end time, duration and granularity. For example, in the student teacher ratio indicator, the student population may have been drawn in 2012 and the denominator, teacher population during 2013. Or more specific start and end times may be specified within the same year, such as February 1, 2013 to March 28, 2013 for the student population, and September 15 to November 30th for the teacher population. The ability to specify the time period at any level of granularity is important.

The OWL-Time ontology [23] satisfies these requirements. It provides for the representation of time points, intervals and durations, plus the specification of these at any level of granularity: year, month, week, day, hour, minutes and second. Secondly it provides a rich set of temporal relations, such as before, overlaps and during, that can be used to link and reason over the temporal points and intervals.

We extend the definition of a Population by adding the property 'for_time_interval'.[14] The range of the property is a 'DateTimeInterval' that specifies the start and end date/times for the population. Note that we use a 'DateTimeInterval' as opposed to 'Year', which is the range of the 'forDateTimeDescription' property of the 'ISO37120Indicator' class. Indicators are defined for a specific year, but the data used to derive may be sub or super intervals of that year. Fig. 8 depicts this addition.

5.4. Statistics

The fourth semantic requirement for populations is the representation of the parameter/statistic being measured. The student teacher ratio indicator is based on a measure of the number of students and teachers within a population designated by a city (Placename). One can view both as a statistical measurement in the sense that there is a population that we want to perform a measurement of, namely a city population, and we are counting the number of members that satisfy a description of a Student and

[13] Classes and properties from Geonames have the prefix "geo", and from OpenGIS have the prefix "gis".

[14] The three properties 'located_in', 'defined_by' and 'for_time_interval' could have been replaced with a single property such as 'constraint_to' whose range is a class that contains the three properties. The reason for having the three properties attached directly to Population is to highlight the importance of these constraints.

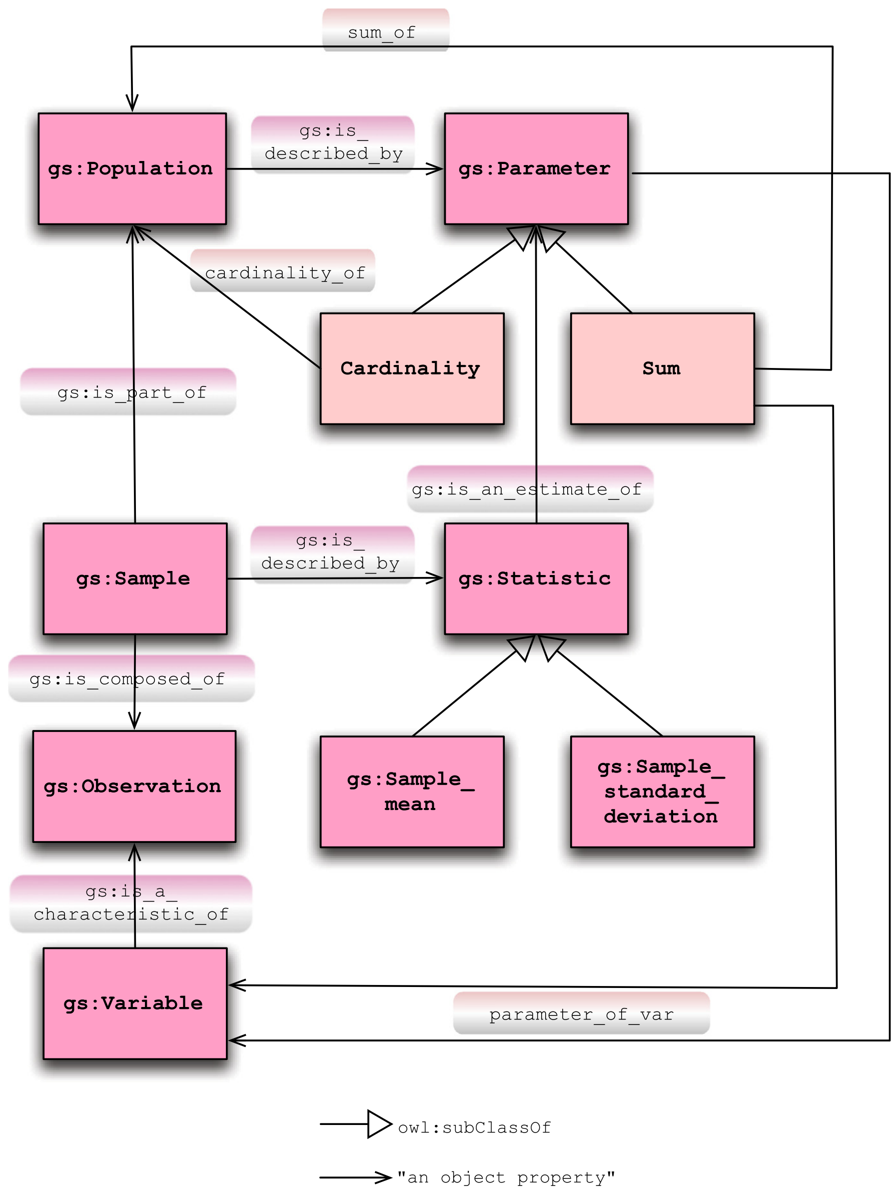

Fig. 9. Govstat extension.

a Teacher, respectively. While the Student Teacher Ratio requires a count of the population, other indicators may require statistical measures of mean, deviation, etc. or other characteristics of the population

The GovStat ontology provides the basic pattern for representing Populations and Parameters and their corresponding Samples and statistics (Section 4.3). Fig. 8 depicts the extensions made to the GovStat ontology to accommodate measuring the cardinality of the Population and the sum of a property over the members of the 'Population'. Both 'Sum' and 'Cardinality' are subclasses of 'Parameter'. Each has a reverse link back to 'Population':

- Cardinality parameter, links back to the population using 'cardinality_of',
- Sum parameter links back using the 'sum_of' property.
- Mean parameter links back using the 'mean_of' property.

'Parameter' links via the 'parameter_of_var' property to a 'Variable' which specifies the variable/data property that is to be summed, averaged, etc. across the members of the Population. 'Variable' has a data property that specifies the variable/property over which the parameter is computed. It is specified as a string since a property cannot be a value of another property in OWL 2.

Each subclass of 'Parameter', inherits the 'parameter_of_var' property that links it to the variable (data property) of the members of the population for which the calculation is being performed (i.e., sum, mean, etc.).

Fig. 9 depicts the integration of the parameter 'Sum' with 'Population', and the variable over which the sum is performed.

5.5. Quantities

The final semantic requirement is the specification of the quantity and unit of measure of the parameter. For example, does the parameter specify the sum or mean of the length measured in meters of the members of the population? Where length is the quantity and meters is the unit of measure.

Measurement ontologies provide the basic concepts that underlie numbers. They divide measurement into a quantity such as length (the what) and a unit of measure such as meters (the how). A Unit of measure has a scale classified as interval or ratio,

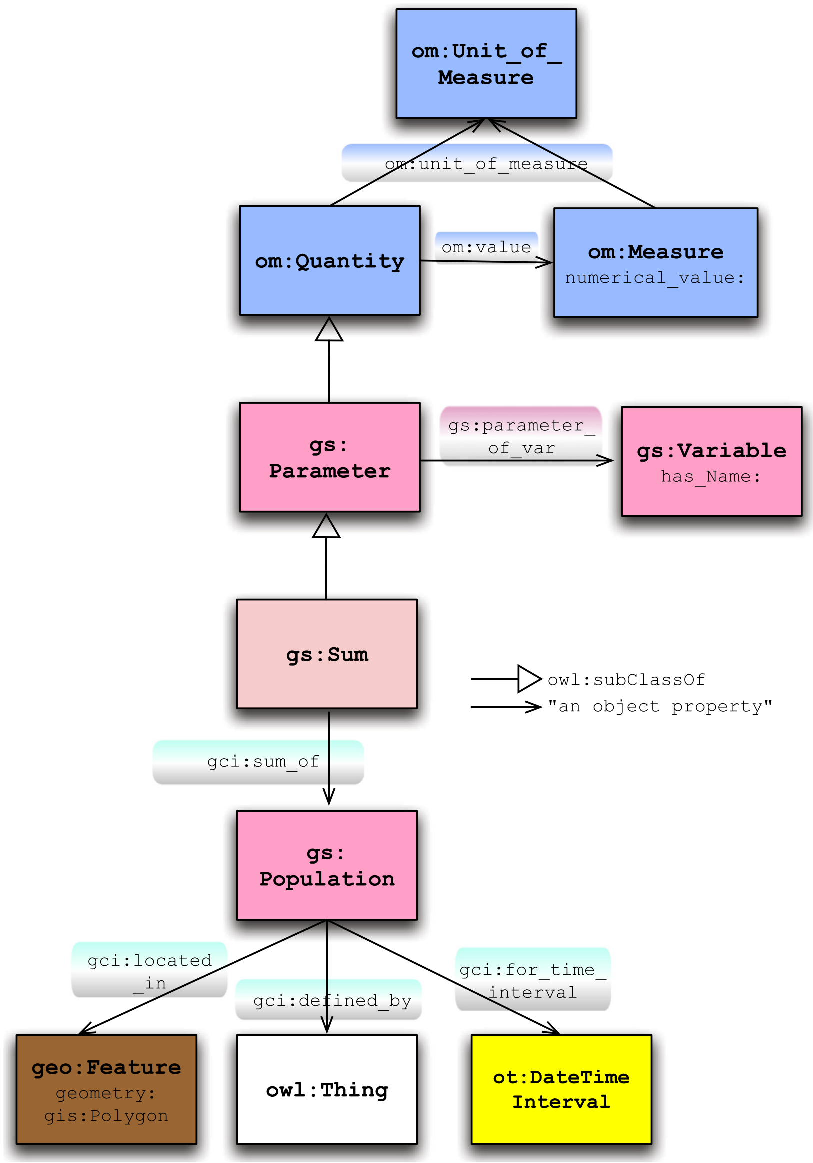

Fig. 10. Integration of measurement ontology.

and whether the number is the composition of dimensions, such as velocity composed of speed and direction, and whether it has a starting point such as absolute zero on the Kelvin scale. Upper level ontologies such as SUMO [24] and CYC [25] provide classes for representing quantities, but the OM ontology[15] [26] provides an ontology based on measurement theory.

Fig. 10 depicts the integration of the OM measurement ontology. A 'Parameter' is defined to be a subclass of Quantity, inheriting the value and 'unit_of_measure' properties. The value property links the 'Quantity' to a 'Measure' where the actual value of the quantity is specified. The 'unit_of_measure' property links the quantity to the 'Unit_of_Measure' class where the unit is specified.

6. Example

In this section, we demonstrate how the Semantics of Populations is used in the representation of the Student Teacher Ratio indicator definition. Additional detail on this and the rest of the ISO 37120 education indicators can be found in [27].

Fig. 11 depicts the "upper" structure of the Indicator. The 'Student_teacher_ratio_GCI' is ultimately a subclass of 'Quantity', and whose value is a 'Student_teacher_ratio_measure' with a 'unit_of_measure' of 'population_ratio_unit'. It has a numerator that is a 'Student_Population_size' and a denominator which is a 'Teacher_population_size'. Both are subclasses of 'Quantity' (not shown). It is the student and teacher population size quantities that rely on the semantics of populations.

Fig. 12 depicts the population ratio pattern that forms the basis of this indicator, and most other indicators in the ISO 37120 standard. The left side of the diagram depicts the numerator, 'Student_Population_Size', which is a 'Quantity' and a 'Sum', representing the cardinality of the Student_Population. The 'Student_Population' is 'defined_by' 'Student', and 'located_in' a 'City'. 'Student' is defined as having a property 'has_Enrollment'.

Fig. 13 depicts a subset of the classes and object properties that define a student, as required by the ISO 37120 definition. Briefly, a student is a person who has an enrollment. An enrollment identifies the educational program, educational facility, grade, courses

[15] The OM ontology can be found at: http://www.wurvoc.org/vocabularies/om-1.8/ We use the prefix "om:" to identify classes and properties from the ontology. Definitions and examples are taken directly from the ontology where quoted.

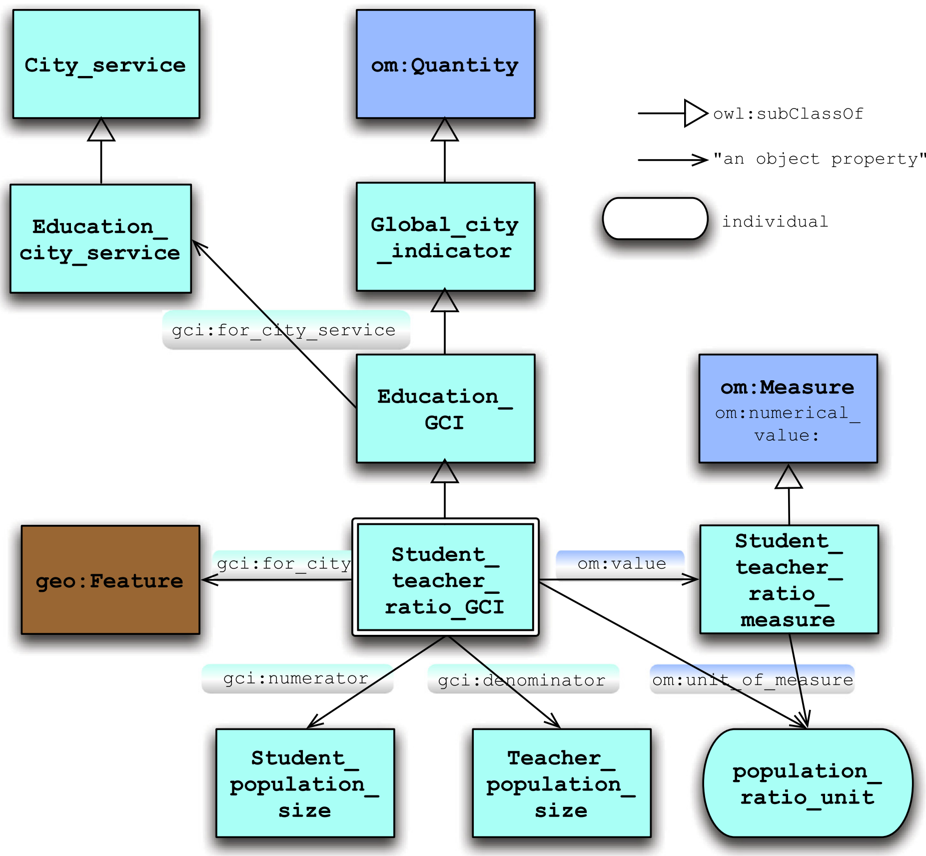

Fig. 11. Student teacher ratio upper level.

and their enrollment status. A student can have more than one enrollment, as they may be taking other types of courses in different programs (e.g., yoga). The 'Program' class is specialized to a 'SchoolProgram' and then to a 'GradeLevel' that enables the representation of primary and secondary school programs. This portion of the education ontology enables the modeling of a student enrolled in courses in a primary level program, and that the program is provided by a public education facility.

Appendix A provides a Description Logic definition of many of the classes in Fig. 13 using the Manchester syntax. The complete GCI Education ontology is defined in [27].

To complete our example, we provide a definition of Student and Teacher, for indicator 6.4 Primary School Student Teacher Ratio, using Description Logic syntax. The definition is built using the Education ontology defined in [27], of which the student related classes are provided in Appendix A.

6.4_Student $\sqsubseteq$ Student
  $\sqcap >$ 1 has_Enrollment 6.4_Enrollment
6.4_Enrollment $\sqsubseteq$ Enrollment
  $\sqcap =$ 1 attends PublicPrimarySchool
  $\sqcap \geq$1 enrolled_Grade PrimaryGrade
  $\sqcap =$ 1 enrolled_Program GradeLevelPrimary
  $\sqcap =$ 1 enrolled_Status (Full_Time $\sqcup$ Part_Time)
6.4_Teacher $\sqsubseteq$ Teacher
  $\sqcap =$ 1 has_Placement 6.4_Placement
6.4_Placement $\sqsubseteq$ Teaching_Placement
  $\sqcap =$ 1 educational_Staff_At PublicPrimarySchool
  $\sqcap =$ 1 for_SchoolYear (= 1 schoolYear_For 6.4)

7. Evaluation

In this section we evaluate the Population ontology along three dimensions: (1) its sufficiency in representing populations as part of the definitions of indicators; (2) its sufficiency in enabling the checking of consistency of standard indicator definitions against their interpretation by a city; and (3) in what ways it can be used to support the harvesting of city data.

7.1. Representation sufficiency

As described in Section 3, the PolisGnosis analysis engine is designed to be indicator independent. This imposes the constraint that indicator definitions be provided as inputs into the analysis engine, along with the city's version of the generic theme classes and their interpretation of the definitions. Our approach has been to define the Global City Indicator Foundation Ontology, which can be used to represent any type of indicator, so that they can be "understood" by the PolisGnosis analysis engine. Consequently, the first method of evaluation is to determine whether the ontology is sufficient to represent the population definitions in ISO 37120 city indicators.

At the time of writing, theme specific ontologies and the indicator definitions in the following nine ISO 37120 themes have been completed, spanning over 50% of the standard's indicators:

- Education [27]
- Energy [28]
- Environment [29]
- Finance [30]
- Fire & Emergency [31]
- Safety [32]
- Recreation [33]
- Shelter [34]
- Telecommunications and Innovation [35]

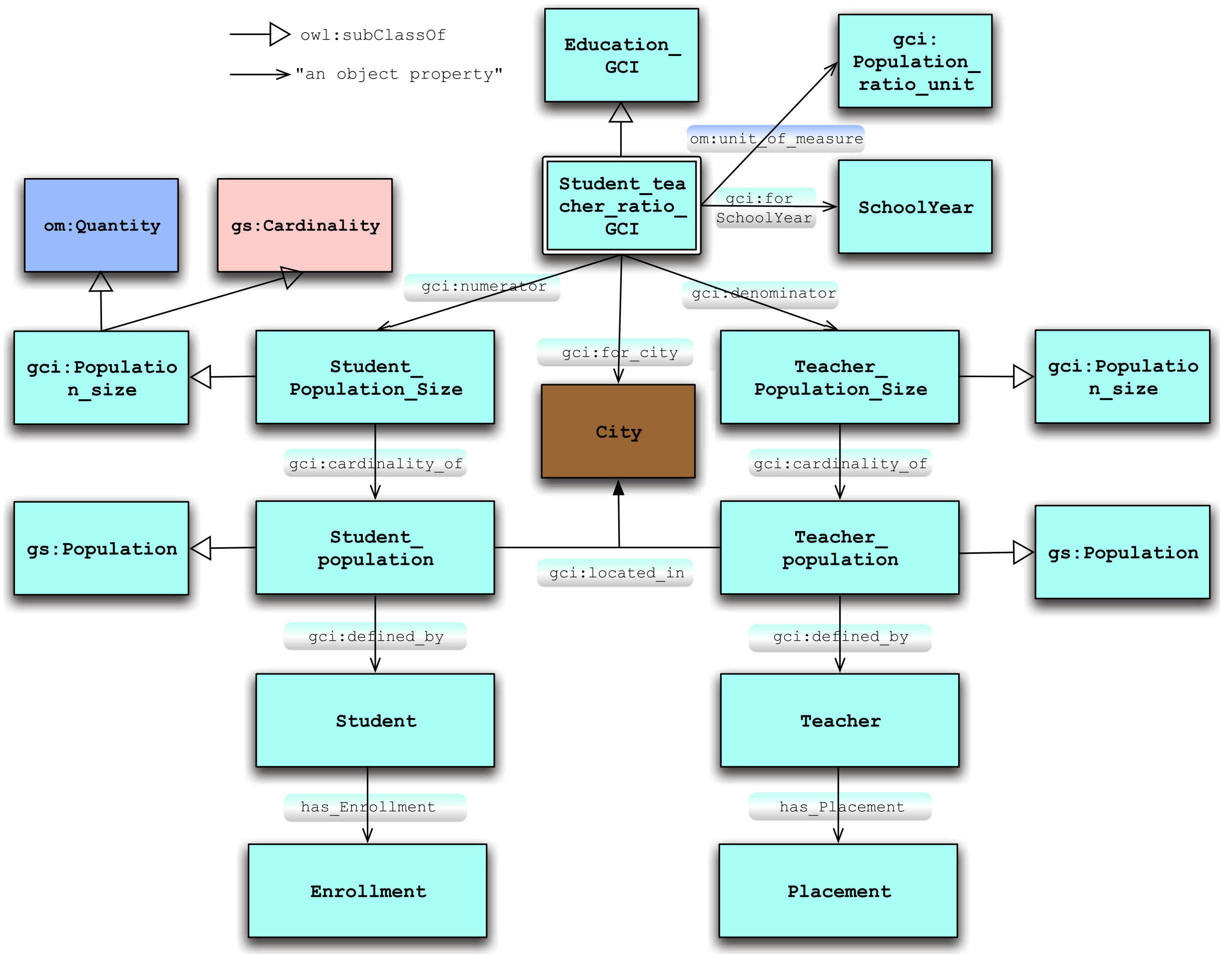

Fig. 12. Student teacher ratio pattern.

In addition to the above, definitions for the indicators in the following themes are under development: Governance, Health, Transportation and Urban Planning. Appendix B lists the indicators for each of the completed themes.

Each theme has two OWL files defined. A generic ontology for the theme, e.g., education ontology, containing the generic/common sense education classes required to represent an education indicator, and an indicator definition file that contains the definitions of the theme's indicators. The latter indicator definition file requires the generic theme ontology to define a Population. URIs for each theme's generic ontology and the definition of its indicators are provided in Appendix C.

With over 50% of the ISO 37120 indicators definitions represented using the Population ontology along with the GCI Foundation ontology and the theme specific ontologies, the Population ontology is sufficient for representing indicator-related populations.

Finally, Santos et al. [36] have extended the GCI Ontology, including the Population related concepts and properties, to support the generation of indicator visualizations.

7.2. Consistency checking sufficiency

The intended use of the Population ontology is to enable the checking of consistency between an indicator's standard definition, and the definition employed by a city. Inconsistencies can exist along three dimensions: (1) definitional, longitudinal and transversal; (2) actual vs potential; and (3) spatial, temporal, measurement and type. A generic reasoner, CICC has been created that can take as input the indicator's standard definition, the generic theme ontology, and the city's interpretation of the indicator definition, and determine consistency along all three dimensions.

CICC has been demonstrated on the ISO 37120 Shelter indicators [7,8]. A set of test cases were created for each type of inconsistency CICC was designed to detect. The inconsistencies were based on the limited shelter data available from the city of Toronto. A problem with analyzing city indicators is that the data required to perform the analysis is not openly available [17]. Consequently, the majority of the data had to be created. Never the less, the test cases were sufficient to demonstrate that the ontology is sufficient to enable this type of analysis.

7.3. Data harvesting sufficiency

Although harvesting city data was not the intended use of the ontology, we can explore its sufficiency. Consider the following two cases where we want to harvest all students in primary, public school within a city:

Case 1 — Data Properties: where the class linked to by the 'defined_by' property is composed entirely of data properties. For example, the definition of a student is composed solely of integer properties for age and grade with constraints on their values, and simple string values characterizing the type of school and whether it is public or private. This representation can easily be transformed into a database query with filters.

Case 2 — Object Properties: where the class linked to by the 'defined_by' property is composed of a mix of object and data

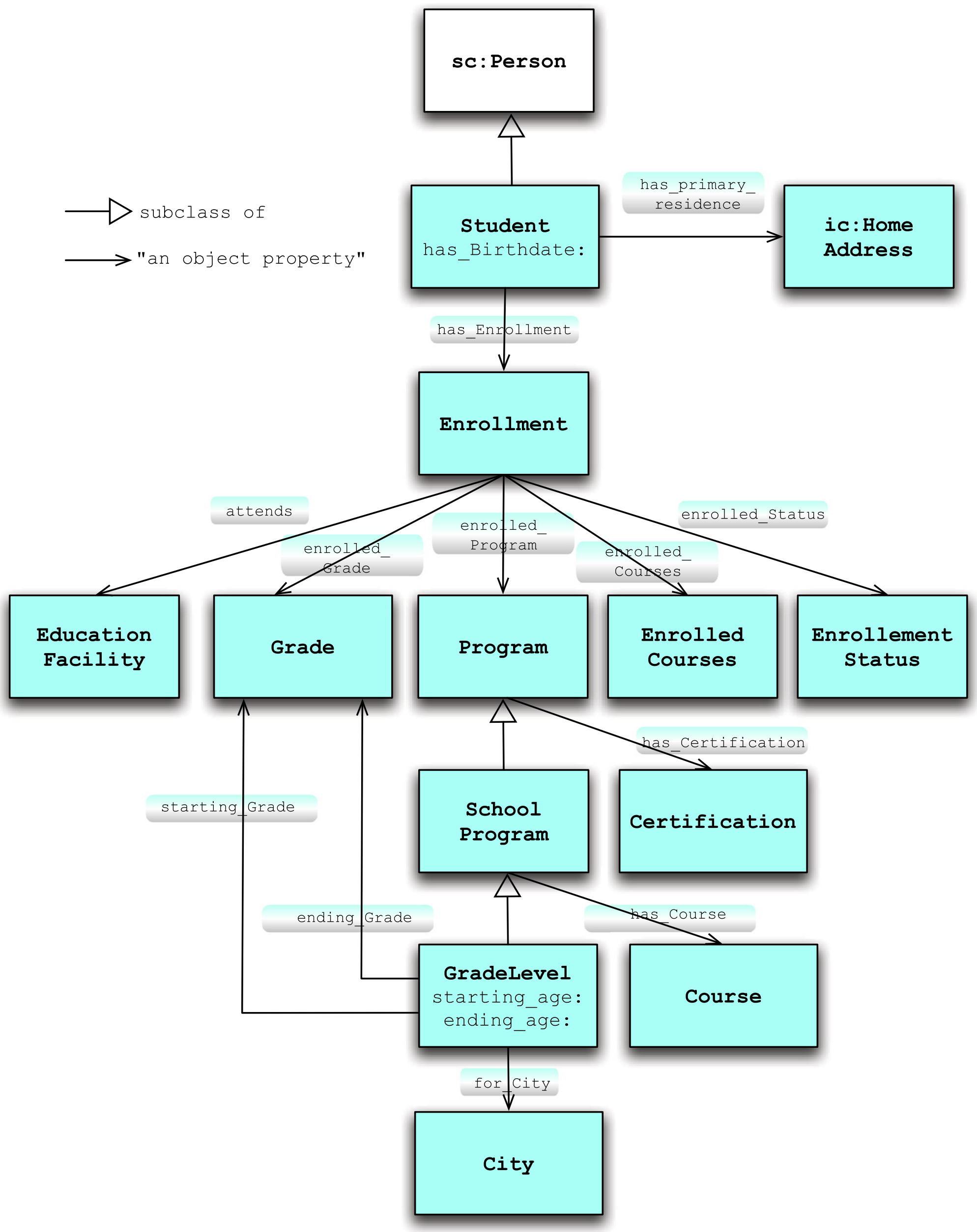

Fig. 13. Student class in the GCI education ontology.

properties. For example, the definition of a student as provided in Section 6. Translation of this into a database query will require a significant number of joins across multiple relations, assuming the data is available. Fig. 13 identifies at least ten relations (objects) that the query may have to search.

Realistically, most open city data falls into the first case. Richer representations of data using ontologies such as suggested here are beyond most datasets, education being an exception. Never the less, beyond case 1 harvesting, it may be possible to use the richer definitions of populations to evaluate the meta-data used to define a dataset to see if it is relevant to the indicator.

8. Discussion and conclusion

In this paper we defined the semantics of populations. The requirement for a precise, unambiguous, computer-understandable definition of statistical populations has been necessitated by the growth of applying data analytics to open city data [5]. Within the context of city indicators, where cities analyze city performance as a key component of evidence-based policy formation, the issue of indicator validity lurks beneath the surface. The most basic question that needs to be addressed is whether cities are measuring the same thing, and are their measurements consistent with the standards they purport to represent?

As described earlier, the basic indicator design pattern relies on measuring parameters of populations. If cities define their interpretation of the indicators such that they are not consistent with the indicator definition, or with each other, it may invalidate the analysis. By defining the semantics of populations using Description Logic, and implementing it in OWL on the Semantic Web, it is now possible to build reasoners that can perform definitional, intra-indicator and inter-indicator consistency analysis [8]. Hence it is possible to determine whether indicators are comparable across time or cities. This approach does place a burden on cities to publish their interpretation of the indicator definitions using these and other ontologies.

The population ontology provides a different perspective on what needs to be represented about populations, namely their definition. As discussed in the related work section, statistics ontologies have focused on the representation of datasets, and not on the definition of populations; they can represent the members of a population but not the definition of the population. This is a subtle but significant difference, as the population definition is critical to enabling consistency analysis.

The only other approach to defining populations is the use of the RDF Data Cube Vocabulary, which provides for the specification of the range of a cube's dimensions as rdfs classes. While useful, it does not provide the greater expressiveness of Description Logic for defining both the range of properties, but how these properties are combined into necessary and sufficient conditions.

The population definitions can be used in the harvesting of city data. It can be used to construct constraints on queries across merged city datasets, such as those constructed by the Open City Data Pipeline. But harvesting data without performing consistency checking serves to perpetuate the incomparability of indicators. In other words, computing indicator values based on the harvested data does not guarantee that the values are comparable. Differences in interpretation of indicator definitions can lead to inclusion of data that is not filtered by simple (Case 1) queries.

There are still many challenges ahead. First and foremost is the development of domain ontologies with which to define population members. Though the PolisGnosis project has defined domain ontologies for many of the ISO 37120 indicator themes, more work needs to be done. But the biggest challenge is adoption. The representation and publishing of city indicator definitions and the data used to derive them, requires a significant commitment by cities. The expertise required to do this is limited. How can we convince cities, and their enterprise software vendors to adopt these standards for representing and publishing their indicator-related data?

An important step towards the adoption of our approach is its incorporation into city modeling and data standards. The population and foundation ontologies are the basis for a new ISO standards project to create an ontology for representing the definitions of city indicators (ISO/IEC AWI Project 21972: An Upper Level Ontology for Smart City Indicators[16]). They are also incorporated as an extension to the City Anatomy Ontology defined in the ISO standards project ISO/NP 37105.[17]

Ultimately, it is up to the cities themselves to mandate the incorporation of standards in the development of their next generation enterprise systems.

Acknowledgments

This research was supported in part by the Natural Science and Engineering Research Council of Canada. I would also like to thank the reviewers who provided many important suggestions.

Appendix A

Class	Property	Value restriction
Student	owl:subClassOf has_Enrollment has_Birthdate owl:equivalentClass has_primary_residence	sc:Person min 1 Enrollment exactly 1 xsd:dateTime cyc:Student exactly 1 ic:HomeAddress
Enrollment	attends enrolled_Program for_SchoolYear enrolled_Courses enrolled_Grade enrolled_Status	exactly 1 EducationFacility exactly 1 Program exactly 1SchoolYear min 1 Enrolled_Course exactly 1 Grade exactly 1 Enrollment_Status
Enrolled_Course	for_Course has_Result has_Comment	exactly 1 Course exactly 1 xsd:string only xsd:string
Program	has_Certification has_Fulltime_Hours has_Fulltime_Period	some Certification exactly 1 positiveInteger exactly 1 TimePeriod
SchoolProgram	owl:subclassOf has_Course has_SP_Type	Program min 1 Course all SP_Type
GradeLevel	owl:subClassOf starting_Grade ending_Grade gci:for_City starting_age ending_age	SchoolProgram exactly 1Grade exactly 1 Grade exactly 1 City exactly 1 positiveInteger exactly 1 positiveInteger

Appendix B. GCI ontologies

The following are the nine themes and their corresponding indicators. The indicators span a broad variety of populations that we have successfully represented using our ontology:

- Education [27] - 7
  - Percentage of female school-aged population enrolled in schools
  - Percentage of Students Completing Primary Education: Survival Rate
  - Percentage of Students Completing Secondary Education: Survival rate
  - Primary Education Student/Teacher Ratio
  - Percentage of male school-aged population enrolled in schools
  - Percentage of school-aged population enrolled in schools
  - Number of higher education degrees per 100 000 population
- Energy [28] - 7
  - Total residential electrical energy use per capita (kWh/year)
  - Percentage of city population with authorized electrical service
  - Energy consumption of public buildings per year (kWh/m^2)
  - The percentage of total energy derived from renewable sources, as a share of the city's total energy consumption
  - Total electrical energy use per capita (kWh/year)
  - Average number of electrical interruptions per customer per year
  - Average Length of electrical interruption (hours)
- Environment [29] - 8
  - Fine particulate matter (PM2.5) concentration

16 https://www.iso.org/standard/72325.html
17 https://www.iso.org/standard/62064.html

- Particulate matter (PM10) concentration
- Greenhouse gas emissions measured in tonnes per capita
- NO2 (nitrogen dioxide) concentration
- SO2 (sulphur dioxide) concentration
- O3 (Ozone) concentration
- Noise Pollution
- Percentage change in number of native species

- Finance [30] - 4
    - Debt service ratio (debt service expenditure as a percentage of a municipality's own-source revenue)
    - Capital spending as a percentage of total expenditures
    - Own-source revenue as a percentage of total revenues
    - Tax collected as a percentage of tax billed
- Fire & Emergency [31] - 6
    - Number of firefighters per 100 000 population
    - Number of fire-related deaths per 100 000 population
    - Number of natural disaster-related deaths per 100 000 population
    - Number of volunteer and part-time firefighters per 100 000 population
    - Response time for emergency response service from initial call
    - Response time for fire department from initial call
- Safety [32] - 5
    - Number of police officers per 100 000 population
    - Number of homicides per 100 000 population
    - Crimes against property per 100 000 population
    - Response time for police department from initial call
    - Violent crime per 100 000 population
- Recreation [33] - 2
    - Square meters of public indoor recreation space per capita
    - Square meters of public outdoor recreation space per capita
- Shelter [34] - 3
    - Percentage of city population living in slums
    - Number of homeless per 100 000 population
    - Percentage of households that exist without registered legal titles
- Telecommunications and Innovation [35] - 3
    - Number of internet connections per 100 000 population
    - Number of cell phone connections per 100 000 population
    - Number of landline phone connections per 100 000 population

Appendix C. GCI theme ontologies

- Education [27]
    - Generic ontology: http://ontology.eil.utoronto.ca/GCI/Education/GCI-Education.owl
    - Indicator definitions: http://ontology.eil.utoronto.ca/GCI/ISO37120/Education.owl
- Energy [28]
    - Generic ontology:
        - http://ontology.eil.utoronto.ca/GCI/Energy/GCI-Service.owl,
        - http://ontology.eil.utoronto.ca/GCI/Energy/GCI-BuildingOccupancy.owl
    - Indicator definitions: http://ontology.eil.utoronto.ca/GCI/ISO37120/Education.owl
- Environment [29]
    - Generic ontology:
        - http://ontology.eil.utoronto.ca/GCI/Environment/Pollution.owl
        - http://ontology.eil.utoronto.ca/GCI/Environment/Species.owl
        - http://ontology.eil.utoronto.ca/GCI/Environment/Sensors.owl
    - Indicator definitions: http://ontology.eil.utoronto.ca/GCI/ISO37120/Environment.owl
- Finance [30]
    - Generic ontology: http://ontology.eil.utoronto.ca/GCI/Finance/GCI-Finance.owl
    - Indicator definitions: http://ontology.eil.utoronto.ca/GCI/ISO37120/Finance.owl
- Fire and Emergency [31]
    - Generic Ontology:
        - http://ontology.eil.utoronto.ca/GCI/FireEmergency/EResponse.owl
        - http://ontology.eil.utoronto.ca/GCI/FireEmergency/FBOrganization.owl
        - http://ontology.eil.utoronto.ca/GCI/FireEmergency/FDRDeath.owl
    - Indicator Definitions: http://ontology.eil.utoronto.ca/GCI/ISO37120/FireEmergency.owl
- Public Safety [32]
    - Generic ontology: http://ontology.eil.utoronto.ca/GCI/PublicSafety/GCI-PublicSafety.owl
    - Indicator definitions: http://ontology.eil.utoronto.ca/GCI/ISO37120/PublicSafety.owl
- Recreation [33]
    - Generic ontology: http://ontology.eil.utoronto.ca/GCI/Recreation/GCI-Recreation.owl
    - Indicator definitions: http://ontology.eil.utoronto.ca/GCI/ISO37120/Recreation.owl
- Shelter [34]
    - Generic ontology: http://ontology.eil.utoronto.ca/GCI/Shelters/GCI-Shelters.owl
    - Indicator definitions: http://ontology.eil.utoronto.ca/GCI/ISO37120/Shelters.owl
- Telecommunications and Innovation [35]
    - Generic ontology: http://ontology.eil.utoronto.ca/GCI/Innovation/GCI-Innovation.owl
    - Indicator definitions: http://ontology.eil.utoronto.ca/GCI/ISO37120/Innovation.owl

Appendix D. Ontology metrics

The Population ontology is an extension of a subset of the GovStat ontology. The following metrics cover the GovStat subset and the extensions.

Metric	Value
Axiom	88
Logical axiom count	53
Declaration axioms count	31
Class count	17
Object property count	12
Data property count	2
DL expressivity	ALCHIQ(D)

References

[1] D. Hoornweg, F. Nunez, M. Freire, N. Palugyai, E.W. Herrera, M. Villaveces, City Indicators: Now to Nanjing, World Bank Policy Research Working Paper 4114, 2007.

[2] E. Slack, How much local fiscal autonomy do cities have? A comparison of eight cities around the world, IMFG Perspectives 19 (2017).

[3] A. Banerjee, S. Chaudhury, Statistics without tears: Populations and samples, Industrial Psychiatry J. 19 (1) (2010) 60–65.

[4] M.S. Fox, The role of ontologies in publishing and analyzing city indicators, Computers, Environment and Urban Systems 54 (2015) 266–279 http://eil.mie.utoronto.ca/wp-content/uploads/2015/06/fox-ceus15.pdf.

[5] M.S. Fox, City Data: Big, Open and Linked, Enterprise Integration Laboratory, University of Toronto, 2013a. Working Paper. http://www.eil.utoronto.ca/wp-content/uploads/smartcities/papers/City-Data-v5.pdf.

[6] M.S. Fox, A Foundation Ontology for Global City Indicators, Global Cities Institute, University of Toronto, 2013b. Working Paper 3. http://eil.utoronto.ca/smartcities/papers/GCI-Foundation-Ontology.pdf.

[7] Y. Wang, City Indicator Consistency Analysis: An Approach to Inconsistency Detection (MASc thesis), Mechanical and Industrial Engineering, University of Toronto, 2017. http://eil.mie.utoronto.ca/wp-content/uploads/2015/06/Wang_Yetian_201706_MASc_thesis.pdf.

[8] Y. Wang, M.S. Fox, Consistency analysis of city indicator data, in: Planning Support Systems for Smarter Urban Futures, Springer, 2017 http://eil.mie.utoronto.ca/wpcontent/uploads/2015/06/CUPUM17_chapter_final.pdf.

[9] T. Bosch, R. Cyganiak, J. Wackerow, B. Zapilko, DDI-RDF Discovery Vocabulary, 2013, http://rdf-vocabulary.ddialliance.org/DDI-RDF_Discovery_Vocabulary.pdf.

[10] R. Cyganiak, S. Field, A. Gregory, W. Halb, J. Tennison, Semantic statistics: bringing together SDMX and SCOVO, LDOW 628 (2010).

[11] I. Petrou, G. Papastefanatos, T. Dalamagas, Publishing census as linked open data: a case study, in: Proceedings of the 2nd International Workshop on Open Data (P. 4), June, ACM, 2013.

[12] R. Cyganik, D. Reynolds, J. Tennison, The RDF data cube vocabulary, W3C Recommendation, W3C (Jan. 2014).

[13] Dublin Core metadata initiative, Dublin Core metadata element set, version 1.1. July 1999. Dublin Core recommendation. http://dublincore.org/documents/dcmi-terms/.

[14] M.C. Pattuelli, The GovStat Ontology: Technical Report, The GovStat Project, Integration Design Laboratory, School of Information and Library Science, University of North Carolina at Chapel Hill, 2003. http://ils.unc.edu/govstat/papers/govstatontology.doc.

[15] M. Groen, W. Meys, M. Veenstra, Creating Smart Information Services for Tourists by Means of Dynamic Open Data, Proceeding UbiComp '13 Adjunct Proceedings of the 2013 ACM conference on Pervasive and ubiquitous computing, pp. 1329–1330, 2013.

[16] R.L. Pereira, P.C. Sousa, R. Barata, A. Oliveira, G. Monsieur, CitySDK tourism API –building value around open data, Journal of Internet Services and Applications 6 (24) (2015).

[17] M.S. Fox, C. Pettit, On the Completeness of Open City Data for Measuring City Indicators, Proceedings of the IEEE Conference on Smart Cities, Guadalajara Mexico, 2015.

[18] S. Bischof, C. Martin, A. Polleres, P. Schneider, Collecting, integrating, enriching and republishing open city data as linked data, Proceedings of the14th International Semantic Web Conference, Bethlehem, PA, USA, 2015, pp. 57–75.

[19] S. Bischof, B. Kampgen, A. Harth, A. Polleres, P. Schneider, Open City Data Pipeline, Working Papers on Information Systems, Information Business and Operations, No. 01/2017, Department of Information Systems and Operations, Vienna University of Economics and Business, 2017. http://epub.wu.ac.at/view/p_series/S1/.

[20] T. Lebo, D. McGuinness, S. Sahoo, PROV-O: The PROV ontology, W3C recommendation, W3C, 2013, http://www.w3.org/TR/2013/REC-prov-o-20130430/ (Apr. 2013).

[21] F. Chiang, R.J. Miller, Discovering data quality rules,, Proc. VLDB Endow. (2008) 1166–1177.

[22] L. Otero-Cerdeira, F.r. Rodriquez-Martinez, A. Gomex-Rodriguez, Ontology matching: A literature review, Expert Syst. Appl. 42 (2) (2015) 949–971.

[23] J.R. Hobbs, F. Pan, Time Ontology in OWL, 2006. http://www.w3.org/TR/owl-time/.

[24] I. Niles, A. Pease, Towards a standard Upper Ontology, in: Proceedings of the 2nd International Conference on Formal Ontology in Information Systems FOIS-2001, Ogunquit, Maine, New York: ACM Press, 2001.

[25] C. Matuszek et al., An Introduction to the Syntax and Content of Cyc, in: Proceedings of the 2006 AAAI Spring Symposium on Formalizing and Compiling Background Knowledge and Its Applications to Knowledge Representation and Question Answering, 2006.

[26] H. Rijgersberg, M. Wigham, J.L. Top, How semantics can improve engineering processes: A case of units of measure and quantities, Adv. Eng. Inform. 25 (2011) 276–287.

[27] M.S. Fox, An Education Ontology for Global City Indicators, Enterprise Integration Lab, University of Toronto, 2014, Working Paper. http://eil.utoronto.ca/wp-content/uploads/smartcities/papers/GCI-Education.pdf.

[28] A. Komisar, M.S. Fox, A Energy Ontology for Global City Indicators (ISO 37120), Working Paper, Enterprise Integration Laboratory, University of Toronto, 2017.

[29] D. Dahleh, M.S. Fox, An Environment Ontology for Global City Indicators (ISO 37120), Enterprise Integration Laboratory, University of Toronto, 2016. Working Paper. http://eil.mie.utoronto.ca/wp-content/uploads/2015/06/An-Environmental-Ontology-for-Global-City-Indicators-12oct2016.pdf.

[30] Z. Wang, M.S. Fox, A Finance Ontology for Global City Indicators (ISO 37120), Enterprise Integration Laboratory, University of Toronto, 2016. Working Paper. http://eil.mie.utoronto.ca/wp-content/uploads/2015/06/GCI-Finance-20160606.pdf.

[31] N. Rauch, M.S. Fox, A Fire and Emergency Ontology for City Indicators, Enterprise Integration Lab, University of Toronto, 2017. Working Paper. http://eil.mie.utoronto.ca/wp-content/uploads/2015/06/Fire-and-Emergency-response-V10.pdf.

[32] K. Khazai, M.S. Fox, A Public Safety Ontology for Global City Indicators (ISO 37120), Enterprise Integration Laboratory, University of Toronto, 2017. Working Paper.

[33] T. Abdulai, M.S. Fox, Recreation Ontology for Global City Indicators (ISO 37120), Working Paper, Enterprise Integration Laboratory, University of Toronto, 2017.

[34] Y. Wang, M.S. Fox, A Shelter Ontology for Global City Indicators (ISO 37120), Enterprise Integration Laboratory, University of Toronto, Revised: 28 August 2015, 2015. Working Paper. http://eil.mie.utoronto.ca/wp-content/uploads/2015/06/GCI-Shelter-Ontology.pdf.

[35] A. Forde, M.S. Fox, A Telecommunication and Innovation Ontology for Global City Indicators (ISO 37120), Enterprise Integration Laboratory, University of Toronto, 2015. Working Paper. http://eil.mie.utoronto.ca/wp-content/uploads/2015/06/GCI-Innovation-Ontology.pdf.

[36] H. Santos, V. Dantas, V. Furtado, P. Pinheiro, D.L. McGuinness, From Data to City Indicators: A Knowledge Graph for Supporting Automatic Generation of Dashboards, in: Proceedings of European Semantic Web Conference - ESWC 2017 (May 28–June 1 2017, Portorož, Slovenia), 2017.