A Foundation Ontology for Global City Indicators

Mark S. Fox

August 2013

Global Cities Institute

UNIVERSITY OF TORONTO
JOHN H. DANIELS FACULTY OF ARCHITECTURE, LANDSCAPE, AND DESIGN

Daniels

170 Bloor Street West, Suite 1100
Toronto, Ontario M5S 1T9 Canada
Tel +1 416-966-2368
Fax +1 416-966-0478
gcidirector@daniels.utoronto.ca

A Foundation Ontology for Global City Indicators

Global Cities Institute Working Paper No. 03

Mark S. Fox

Department of Mechanical & Industrial Engineering, University of Toronto
5 King's College Circle, Toronto, M5S 3G8. Tel: 416-978-6823
Email: msf@eil.utoronto.ca

August 2013

KEYWORDS

city indicators; ontology; data standardization; semantic web

Global Cities Institute
170 Bloor Street West, Suite 1100
Toronto, Ontario M5S 1T9 Canada
Tel +1 416-966-2368
Fax +1 416-966-0478
gcidirector@daniels.utoronto.ca

A Foundation Ontology for Global City Indicators[1]

Mark S. Fox

Department of Mechanical and Industrial Engineering
5 King's College Road, Toronto ON M5S 3G8
University of Toronto
Email: msf@eil.utoronto.ca

Global Cities Institute Working Paper No. 3
August 2013

1. Introduction

Cities are moving towards policy-making based on data[2]. Increasingly recognized over the past decade by urban researchers, city professionals and political leaders, city level data is scarce and inconsistent. In 2007, it was recognized that "there are thousands of different sets of city (or urban) indicators and hundreds of agencies compiling and reviewing them. Most cities already have some degree of performance measurement in place. However, these indicators are usually not standardized, consistent or comparable (over time or across cities), nor do they have sufficient endorsement to be used as ongoing benchmarks." (Hoornweg et al., 2007).

In response to this challenge, in 2010 the Global City Indicators Facility (GCIF)[3] was created at the University of Toronto, to work with cities globally in identifying a common set of indicators and establishing standardized definitions and methodologies that can be consistently applied globally (Global City Indicators Facility, 2010a; McCarney, 2011). As McCarney (2012b) states, "Policy responses to today's most pressing challenges and opportunities for sustained prosperity are hindered by a set of core weaknesses in current research and information at the city level… no city data has conformed to a standardized

[1] "Global City Indicators©" is a term created by the Global City Indicators Facility in 2010 at the University of Toronto. All rights apply. GCI refers to the indicators created by the GCIF to establish a global standard of over 100 city indicators with a standardized definition and methodology, tested with over 250 cities globally since 2010. The GCIs are now in a draft international standard currently being voted upon by member countries with a view to publishing the GCIs in 2013.
[2] "Data driven decision making is one of the reasons New York City is the safest big city in America," said Mayor Bloomberg. "Just as data helps us reduce crime, prevent fire fatalities and keep incarceration levels low, we believe understanding data can help us work with judges and criminal justice agencies to further improve the effectiveness and efficiency of our criminal justice system." Press Release, New York City, PR-012-13, 7 January 2013.
[3] http://www.cityindicators.org/

methodology that can ensure sound global comparison and learning across cities, making globally comparative research and exchange impossible. The Global City Indicators Facility (GCIF), originally established to fill this gap, is now situated with over 250 participating cities globally, and is positioned as a global leader and center of excellence on globally standardized city metrics. The GCIF creates a global knowledge platform on cities and is positioned as a basis for scholarship and global leadership on cities."

The primary problem with indicator development is that definitions are people oriented; they are provided in natural language, i.e., English, and not in a more formal, possibly computer readable language. As the old joke goes, ask two lawyers a question and you will get three opinions. The same generally holds true for most attempts at defining terminology, such as indicators, using an imprecise language like English. The reader of the definition imposes their own interpretation based on their understanding of the language and the environment in which they live (i.e., how their own city may define some terms).

Consider the definition of a Student/teacher ratio as provided in Hoornweg et al. (2007, p. 45): “Student/teacher ratio”[4]. This has been expanded in World Bank (2008, p. 18) to: “Student/teacher ratio”, where the numerator is “Number of Students”, and the denominator is “Number of Teachers”. One problem is whether “student” refers to full time students, or part time students. Are they regular students or special needs students? Do they include kindergarten students or not? It is also difficult to compare an indicator for a single city across time if the definition of student changes. For example, today the educational system includes students with special needs, but 60 years ago they may not have been enrolled. Without a more precise definition of terms, it makes it difficult to compare an indicator across cities where each city interprets what a student is differently, or against itself where definitions change.

Obviously, the definition and documentation of indicators can be expanded, as has been done in the forthcoming ISO proposed standard for Global City Indicators (ISO, 2013)[5]. Following is the definition of student teacher ratio provided by the GCIF (McCarney 2012a; McCarney 2013):

- **Core Indicator Requirements:** "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). The result shall be expressed as the number of students per teacher. Private educational facilities shall not be included in the student/teacher ratio. One part-time student enrolment shall be counted as one full-time enrolment; in other words a student who attends school for half a day should be counted as a full-time enrolment. If a city reports full-time equivalent (FTE) enrolment (where two half day students equal one full student enrolment), this shall be noted. The number of classroom teachers and other instructional staff (e.g. teachers’ aides, guidance counselors), shall not include

[4] Yes, just three words ☺.

[5] This standardization of definition and methodology by the GCIF in partnership with cities worldwide since 2010 has led into current work by the Facility with the International Organization for Standardization (ISO) and the development of a new ISO draft international standard on city indicators has been prepared (GCIF 2010b; McCarney 2012a; McCarney 2013; ISO 2013).

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, for example, a teacher working one day per week should be counted as 0.2 teachers, and a teacher working three days per week should be counted as 0.6 teachers."

- **Data sources:** "The number of full-time equivalent primary school classroom teachers and the number of enrolled primary school students should be collected from the local public school system or Ministry of Education."
- **Data interpretation:** "The student/teacher ratio reflects teacher workload and the availability of teachers' services to their students. The lower the student/teacher ratio, the higher the availability of teacher services to students. The student/teacher ratio has implications not only for the cost of education, but also for the quality. Higher educational attainment is correlated with a lower student/teacher ratio."

The GCIF definition of student teacher ratio clearly addresses some of the issues raised above. Never the less, there will always be a disconnect between the actual value of a city's indicator and the data sources and processes used to measure it. While the indicator's value is recorded in a machine-readable form (e.g., database or semantic web), the sources and measurement processes are buried in datasets and documents that are inaccessible or only human readable. In the end, all we are left with is a record of indicator values without an understanding of what they actually measure and how they were measured. We have to rely on the good will of the people who reported to the data to adhere to the definitions.

Our goal is to formalize the definition of city indicators using the technology of Ontologies (Gruber, 1993; Grüninger & Fox, 1995) as implemented in the Semantic Web (Berners-Lee et al., 2001). By doing so we will:

- enable the creation of more precise definitions thereby reducing the ambiguity of interpretation, for example, defining what "instructional staff" means,
- take indicators out of the realm of humans and into the realm of computers where the world of Big Data, open source software, mobile apps, etc., can be applied to analyze and interpret the data, and
- achieve interoperability, namely the ability to access, understand, merge and use indicator data available from datasets spread across the web.

In the following, we review the requirements for a city indicator and work being done out side of the semantic web to achieve

2. What is an Ontology?

An ontology is an "explicit representation of shared understanding" (Gruber, 1993). It "consists of a representational vocabulary with precise definitions of the meanings of the terms of this vocabulary plus a set of formal axioms that constrain interpretation and well-formed use of these terms" (Campbell & Shapiro, 1995). The terms are classes and their properties. For example, both `Student` and `Teacher` are classes[6]. Classes are arranged

[6] In this paper, classes begin with an upper-case letter and properties with a lower-case letter. All terms appear in `Courier` font.

into taxonomies (Figure 1). For example, `Student` and `Teacher` are subclasses of `Person`. Classes have properties whose values can be numbers or strings. A `Person` can have a property `has_age` with a value that is a positive integer. If `Student` is a subclass of `Person`, then `Student` "inherits" the property `has_age` from `Person`. Properties may also have values that are restricted to instances of other classes. For example, `Teacher` may have a property `teaches_course` where the value is an instance of the class `Course` (e.g., the actual course `MIE1501_2013` is an instance of class `MIE1501` which is a subclass of `Course`).

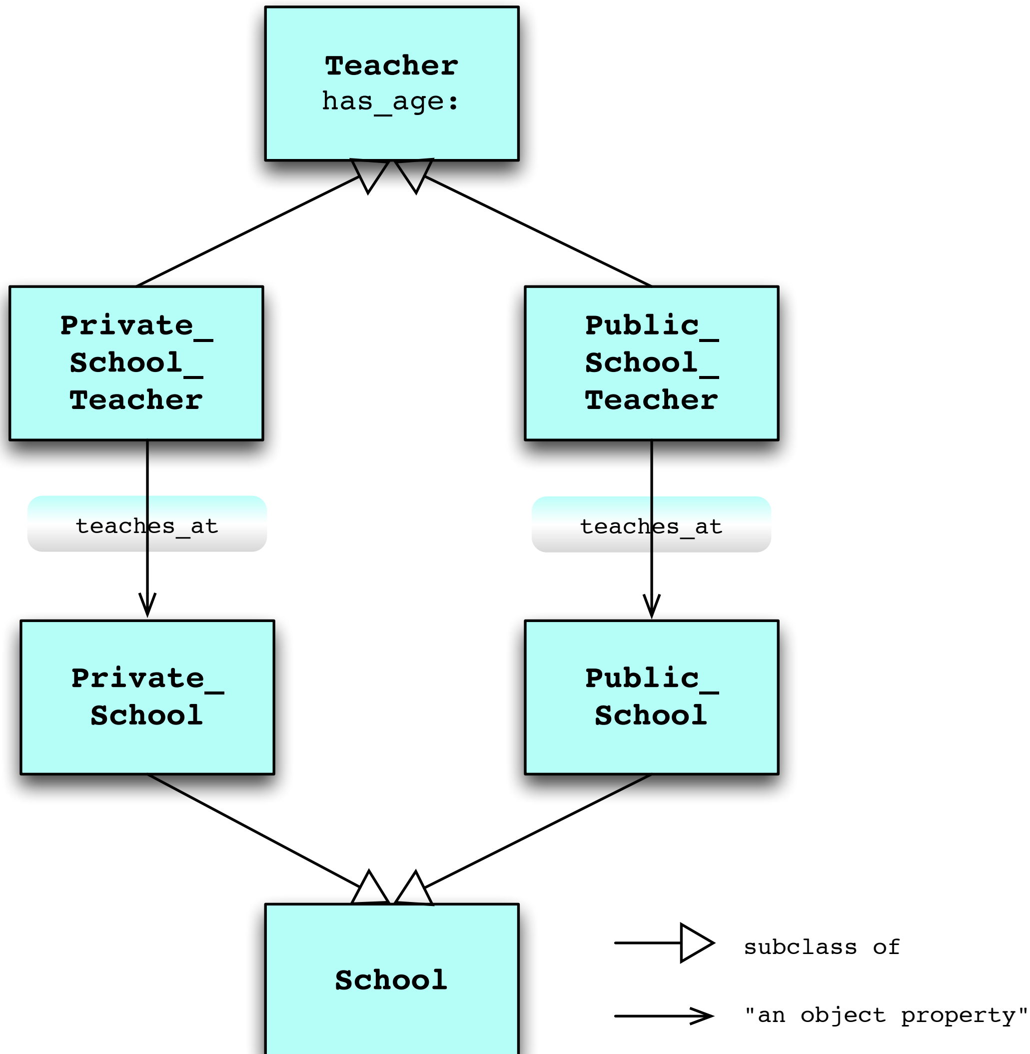

Figure 1

Axioms both define and constrain the interpretation of the terms. For example, a teacher is defined to have to teach in a public school. In Figure 1, there are four axioms with respect to a teacher. `Public_School_Teacher` is defined to be a sub class of `Teacher` and inherits all of its properties (e.g., `has_age`), and it is constrained to `teaches_at` a `Public_School`. A private school teacher has a similar definition but teaches at a private school.

A variety of methods exist for specifying axioms. The most common method on the Sematic Web is a version of Description Logics (Nardi & Brachman, 2002) implemented in the Semantic Web language OWL 2 (Hitzler et al., 2012). For each method a variety of logic theorem provers exist to evaluate the consistency of what is being represented based on the axioms. The specification of terms and axioms, and their consistency testing is facilitated using a graphical, interactive tool such as Protegé (Noy et al., 2001). Most ontologies found on the Semantic Web are little more than taxonomies of classes and properties without axioms.

The primary goal of ontology engineering is to develop a shareable representation of knowledge. The belief is that by engineering the terms and axioms properly, they will be reusable across a broad spectrum of applications. With reuse, we can achieve interoperability, namely the ability to access, analyse and merge data from many diverse sources across the web because they use the same ontology or specify a mapping between their ontology and other more broadly used ontologies.

Ontology engineering begins by determining competency requirements for the target ontology, which is defined by a set of questions that the ontology must be able to answer (Grüninger & Fox, 1995). Based on these competency questions, the terms and axioms are developed. Development takes a layered approach where application specific ontologies (e.g., manufacturing ontologies) are defined in terms more foundational ontologies such as time, activity, resource, location, etc. For example, a manufacturing operation would be defined in terms of more general classes such as activities and resources. Secondly, if an ontology already exists on the Semantic Web that satisfies the some or all of the competency requirements, then it will be reused.

In Section 4, we use the ontology engineering process by specifying examples of competency questions and defining city indicator specific classes and properties in terms of more generic ontologies found on the Semantic Web.

3. City Indicators

In this section we introduce the concept of City Indicators. The rapid growth of Asian cities led the Asian Development Bank to launch a city indicator project in 1999. The objectives of the project were to "to establish a policy-oriented urban indicators database for research, policy formulation, monitoring of the development impact of interventions in the urban sector, comparison of performance between cities, and improving the efficiency of urban service delivery." (Westfall and de Villa, 2001 p. x). The result of the project provides the motivation and detailed definition of indicators. It also anticipates an important role for the World Wide Web in the representation and interconnection of indicator data.

The Organization for Economic Co-operation and Development (OECD: www.oecd.org) "provides a forum in which governments can work together to share experiences and seek solutions to common problems." At the core of their work is a large number of indicators spanning topics such as health, education, environment and trade. The indicators are documented in detail, in English, and the results are published as spreadsheets. Definitions of the indicators using Semantic Web ontologies are not available. On the other hand, some OECD datasets have been the object of research in how to automatically transform statistical databases into linked data (Hausenblas et al., 2009; Capadisli et al., 2013).

As part of IBM's Smart Cities initiative, they have developed an Ontology representing various types of city knowledge, including city organization and services, flow of events and messages, and key performance indicators (Uceda-Sosa et al., 2011). OWL definitions of the classes and properties are provided. Axiomatization is limited and so its use of foundational ontologies.

In light of previous efforts to define city indicators, Hoornweg et al. (2007), identified the following aspects a good "indicator must possess to be accurate, timely and relevant for policy purposes:

- **Objective**: clear, well defined, precise and unambiguous, simple to understand.
- **Relevant**: directly related to the objectives.
- **Measurable and replicable:** easily quantifiable, systematically observable.
- **Auditable**: valid, subject to third-party verification, quality controlled data (legitimacy across users).
- **Statistically** representative at the city level.
- **Comparable/ Standardized** longitudinally (over time) and transversally (across cities).
- **Flexible**: can accommodate continuous improvements to what is measured and how. Have a formal mechanism for all cities and interested parties to comment on.
- **Potentially Predictive**: extrapolation over time and to other cities that share common environments.
- **Effective**: tool in decision making as well as in the planning for and management of the local system.
- **Economical**: easy to obtain/inexpensive to collect. Use of existing data.
- **Interrelated**: indicators should be constructed in an interconnected fashion (social, environmental and economics).
- **Consistent and sustainable over time:** frequently presented and independent of external capacity and funding support."

The *raison d'être* for the Global City Indicator Facility is to define city indicators that satisfy these aspects. The proposed ISO standard provides clearer, more precise and more detailed definitions for each indicator. The Global City Indicator Ontology translates these definitions into a form that is machine readable while enhancing many of the aspects above.

4. An Ontology for Global City Indicators

In this section we develop the Global City Indicators ontology (GCI). The GCI ontology is built on a set of "foundational" ontologies that provide many of the basic concepts with which to construct the GCI terminology. We illustrate the construction of the GCI ontology using a single city indicator: Student/Teacher Ratio (STR). A number of issues arise in representing its definition. These issues will be addressed as we build the GCI ontology "one brick at a time" using foundational ontologies.

4.1. Placename Ontology

- What is the city being measured?
- What area does it cover?
- What places does it contain?

The STR is computed over a geographic area. In the case of GCIs, it would be a city. Hence, a requirement of the GCI ontology is the ability to identify the geographic area over which the indicator has been calculated. That is, to associate a "placename" with a geographic area. Such placenames could conceivably be applied to areas larger than a city, such as a region, state or country, or smaller than a city, such as a neighborhood or postal code. For example, a reference to Toronto should cover the city of Toronto but a reference to the Greater Toronto Area should cover the larger area encompassing neighbouring cities. But it must be clear which each refers to. A second requirement is that when two indicators are supposed to be computed over the same geographic area, they are in fact the same area. This means that an area has to have a unique identifier.

There are a number of ontologies that represent geographic and place information. Schema.org[7] provides classes of placenames such as `sc:City`, `sc:Country`, and `sc:State`. It also provides classes for `sc:GeoCoordinates` (i.e., elevation, latitude, and longitude) and `sc:GeoShape` denoted by a polygon or circle. The Linkedgeodata.org ontology[8] extends what can have a placename by providing classes for `gd:neighborhood`, `gd:building`, `gd:bridge`, `gd:hospital`, `gd:airport`, `gd:prison`, etc.

The GeoNames project provides over eight million placenames spanning the world. It provides an International Resource Identifier (IRI) for every placename so that they can be uniquely referred to. The GeoNames' placenames are instantiations of the Geonames Ontology[9] that integrates a number of ontologies, including Schema.org and Linkedgeodata.org, to provide a broad set of classes that span almost every conceivable type of place. Geonames also provides a web interface that allows anyone to search for and/or add new placenames to its knowledge base.

At the core of the Geonames ontology is the `geo:Feature`. A `geo:Feature` contains the following properties:

- `name`: text name of the feature, e.g., "Toronto".
- `alternativeName` - a number of alternative names for the feature.
- `featureClass` – Class of feature such as Administrative (e.g., state, parish), Hydrographic (e.g., stream, lake), and Area (e.g., Parks).
- `featureCode` – Code for the feature within the class.

---

[7] The Schema.org ontology is available at: http://schema.org/. We will use the prefix "sc:" to identify classes and properties from the ontology.
[8] The Linkedgeodata.org Ontology is available at: http://www.linkedgeodata.org/ontology/. We will use the prefix "gd:" to identify classes and properties from the ontology.
[9] The Geonames Ontology is available at: http://www.geonames.org/ontology/ontology_v3.1.rdf#. We will use the prefix "geo:" to identify classes and properties from the ontology.

- population – Population of the feature.
- postalCode – One or more postal codes in which the feature resides.
- wgs84_pos:lat – Latitude of the feature.
- wgs84_pos:long – Longitude of the feature.
- nearbyFeatures – Features spatially located nearby.
- wikipediaArticle – One or more articles in Wikipedia about the feature.

A key component is the geo:featureCode which adapts and extends the feature codes developed by the United States National Geospatial-Intelligence Agency (NGA).

In Figure 2, we show how the Geonames ontology is related to the schema.org ontology.

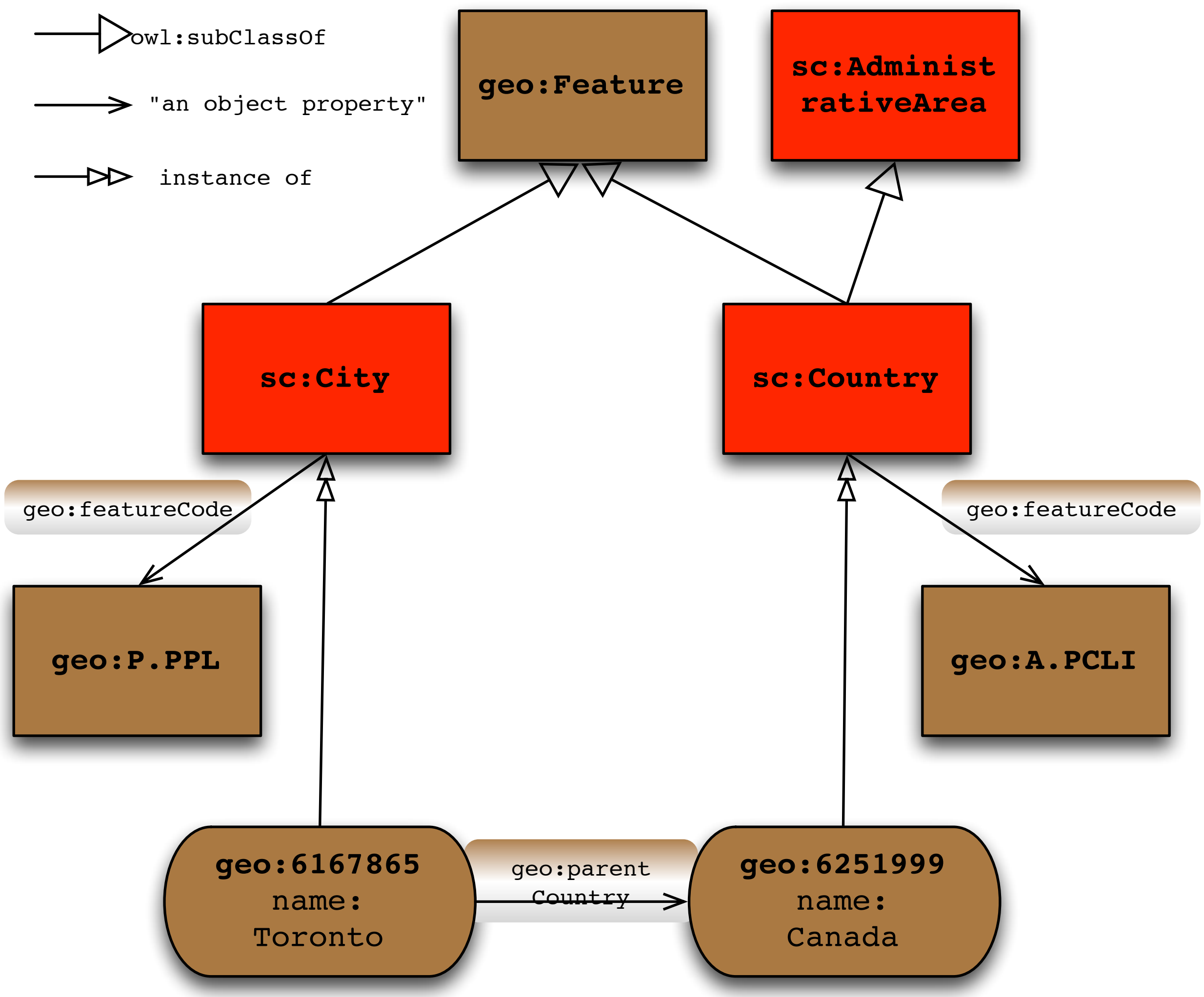

Figure 2

The unique IRI for the city of Toronto is: http://www.geonames.org/6167865. Being an instance of sc:City, it inherits a geo:featureCode of geo:P.PPL which denotes "a city, town, village, or other agglomeration of buildings where people live and work". It is asserted to have a geo:parentCountry of geo:6251999 which is the unique IRI for Canada.

4.2. Measurement Ontology

- How is the indicator constructed?
- What is it scale? Can one indicator be 2x another?
- What are its units? Mega, kilo?

A city indicator is a measure of some property of a city. At the core of an indicator lies a number. The question is what does that number represent? Of course, a written explanation of the indicator is provided, but that is for human consumption. The problem is how do we define an indicator in a way that the computers can understand?

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, and whether the number is the composition of dimensions such as velocity being composed of speed and direction, and whether it has a starting point such as absolute zero on the Kelvin scale.

In the case of the STR, the purpose of a measurement ontology is to provide the underlying semantics of the number, such as what is being measured and the unit of measurement. The importance of grounding an indicator in measurement ontology is to assure that the numbers are comparable, not that they are measuring the same thing (which is dealt with later), but the actual measures are of the same type, e.g., ratio of student and teacher population counts, or that the counts of the student and teacher populations are of the same magnitude (i.e., thousands vs millions).

Upper level ontologies such as SUMO (Niles and Pease, 2001) and CYC (Matuszek et al., 2006) provide classes for representing quantities, but the OM ontology[10] (Rijgersberg et al., 2011) provides a more rigorous ontology based on measurement theory. In the following, we review some of the basics:

Subset of the OM Measurement On tology		
Class	Definition	Examples
Quantity	Refers to what is being measured. It links the phenomenon (e.g., an object) being measured to the value of the measurement (Measure).	Length, diameter.
Unit_of_measure	"A unit of measure is a definite magnitude of a quantity, defined and adopted by convention and/or by law. It is used as a standard for measurement of the same quantity, where any other value of the quantity can be expressed as a simple multiple of the unit of measure."	"For example, length is a quantity; the meter is a unit of length that represents a definite predetermined length. When we say 10 meter (or 10 m), we actually mean 10 times the definite predetermined length

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

		called 'meter'."
Measure	"Combines a number to a unit of measure on an interval or ratio scale."	"3 metres", "10 kilograms"

Before we can represent the concept of a STR, there are several building blocks that need to be put in place. First, we need to represent the cardinality of a set. The STR is the ratio of Student to Teacher, which is the ratio of the number of students to the number of teachers. Both students and teachers represent sets, i.e., the set of all students within a placename and the set of teachers within the same placename. The size of each set is its cardinality.

Figure 3 depicts the new measurement classes required to represent the number of students and teachers. We start by defining a unit of measure: `gci:Cardinality_unit`. As the meter is the unit of measure for length, a `gci:Cardinality_unit` is the unit of measure for the size of a set. The `gci:Cardinality_unit` has a ratio scale: `gci:Cardinality_scale`, which is a subclass of `om:Ratio_scale` and is has a zero element (namely zero).

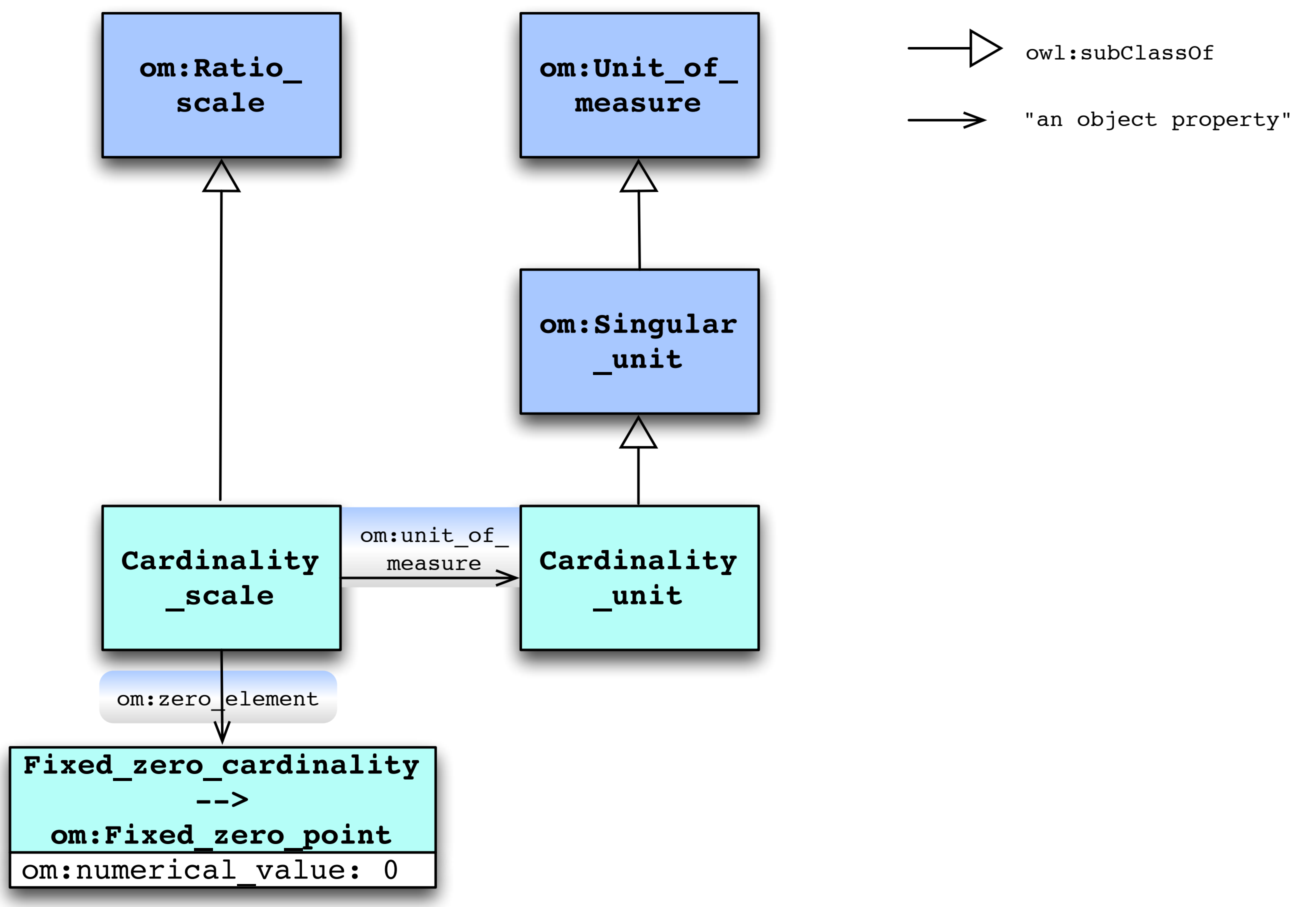

Figure 3

In Figure 4, we specialize the `gci:Cardinality_unit` to the class `gci:Population_cardinality_unit` which is the unit of measure for the cardinality of set defined by a Population (defined in the next section), and associate the symbol “pc” with it. For example, 1100pc represents a population cardinality (or size) of 1100. We can take full advantage of prefix notations available in OM to scale the numbers by defining units of measures: `gci:kilopc`, `gci:megapc` and `gci:gigapc` which are multiples of `gci:Population_cardinality_unit`. 1.1 kilopc represents 1100 pc.

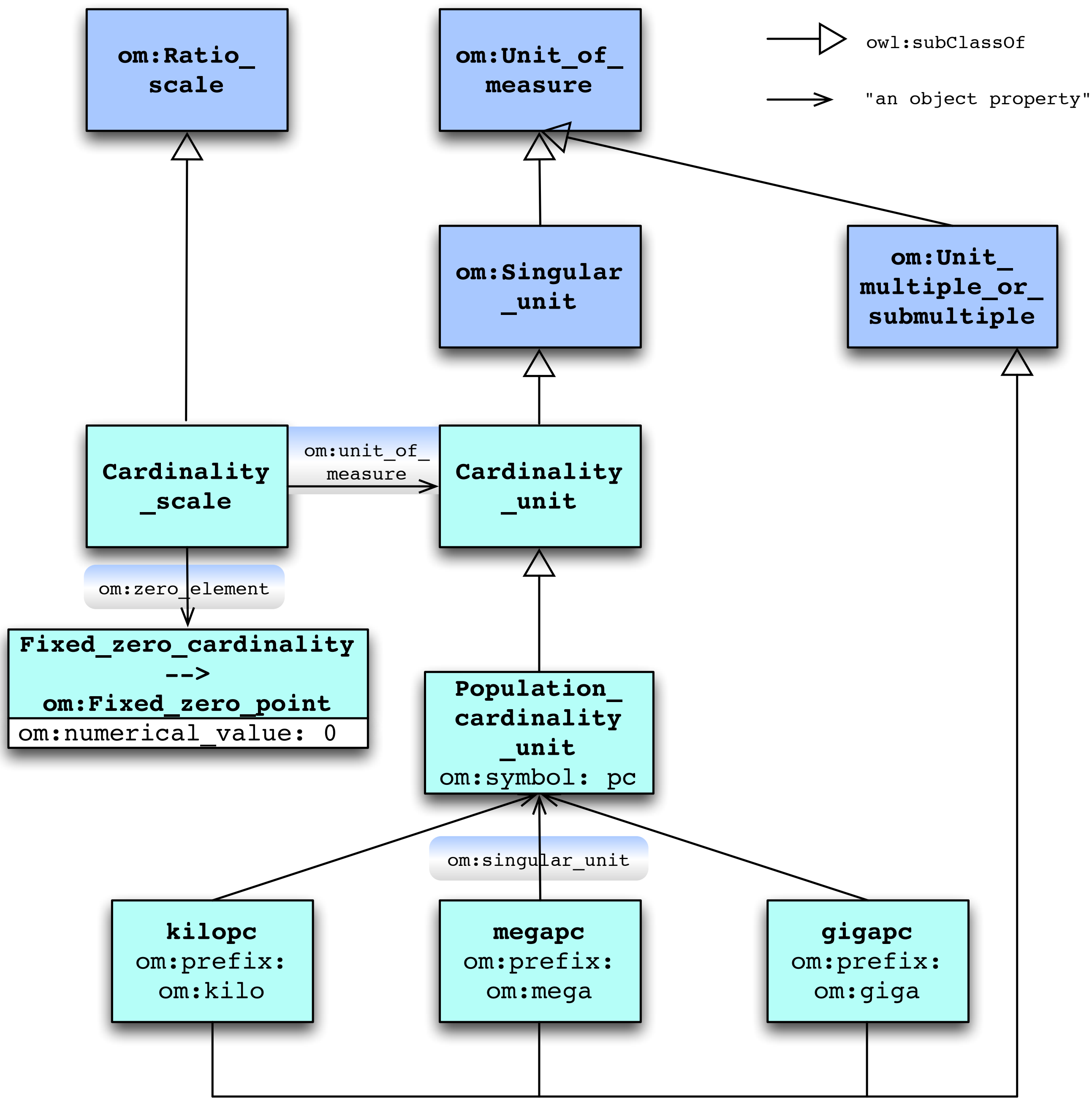

Figure 4

With the above defined, we can now introduce the unit of measure for measuring a population ratio such as STR. `gci:Population_ratio_unit` is defined to be a subclass of `om:Unit_division`. It has two properties:

- `om:numerator` whose range is restricted to being a `gci:Population_cardinality_unit`.
- `om:denominator` whose range is restricted to being a `gci:Population_cardinality_unit`.

In other words, a population ratio is derived from two population cardinalities.

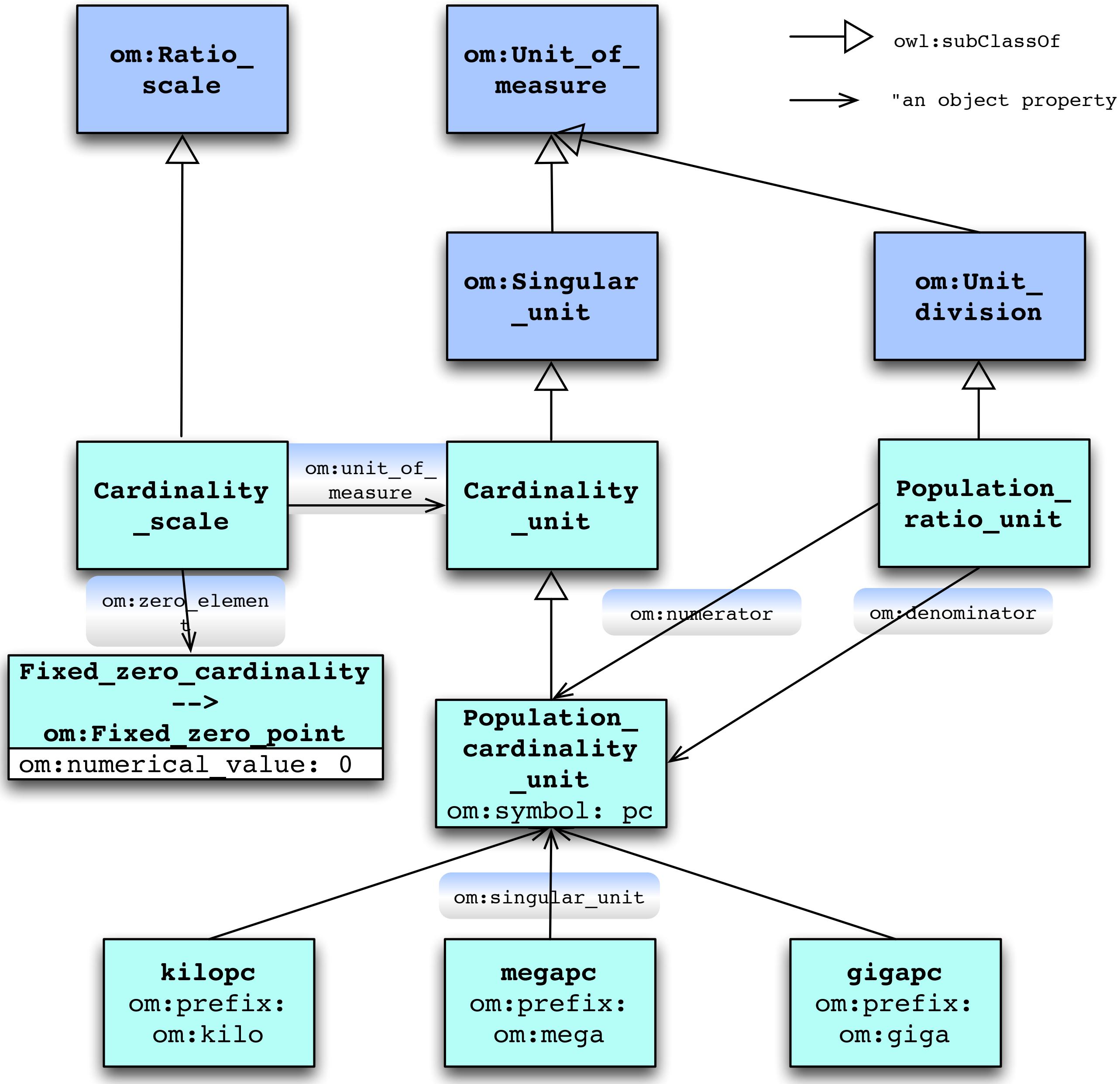

Figure 5

The above, provides the unit of measures for populations (pc) and population ratios (pc/pc) (the *how*). We now have to define *what* we are measuring which is referred to as a Quantity in the OM ontology. First, we need to define the `om:Quantity` for the size of the teacher and student populations from which the STR is derived. In Figure 6 we introduce `gci:Population_size` as a subclass of `om:Quantity`. Its `om:unit_of_measure` is the `gci:Population_cardinality_unit`.

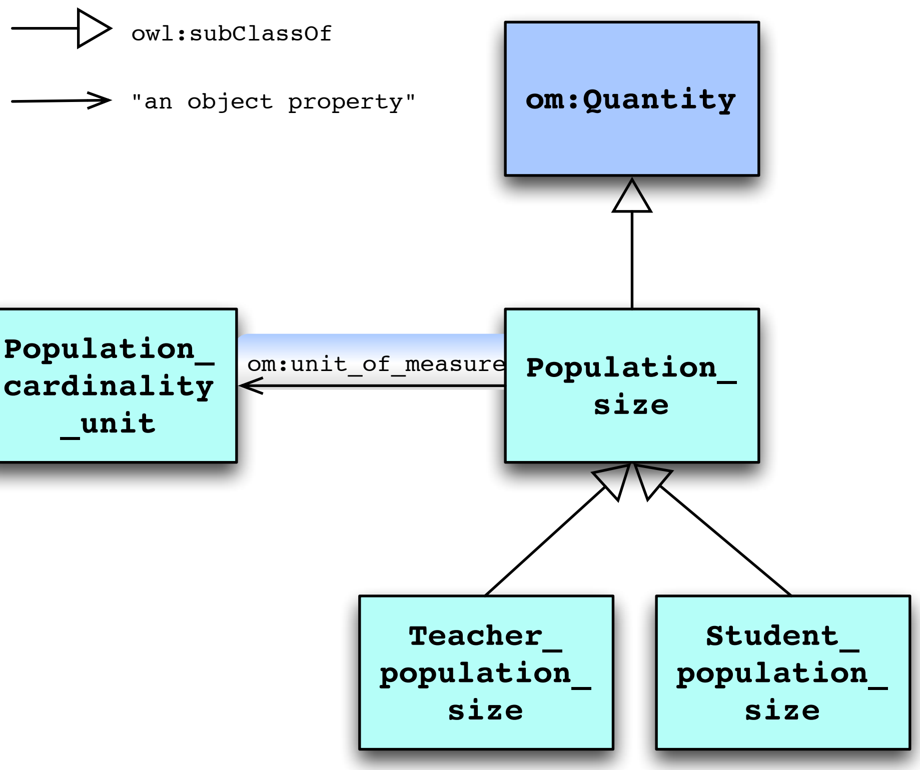

Figure 6

We now have the requisite infrastructure to define GCIs (Figure 7). First we define the class of GCIs, `gci:Global_city_indicator`, as a subclass of `om:Quantity`. All GCIs will be a subclass of `gci:Global_city_indicator`. `gci:Education_GCI` is introduced as a subclass of `gci:Global_city_indicator` with a property that it is a `gci:for_city_service` `gci:Education_city_service`. Simply, this denotes that this indicator is for the education city service.

The actual value for a city's STR will be an instance of the quantity `gci:Student_teacher_ratio_GCI` class, which is a subclass of `gci:Education_GCI`. It has the following properties:

- **`om:unit_of_measure`**, whose range is the `gci:Population_ratio_unit`. This signifies that the quantity is a ratio with a numerator and denominator that are restricted to being `gci:Population_cardinality_unit's`.
- **`gci:numerator & gci:denominator`**, whose ranges are `gci:Student_population_size` and `gci:Teacher_population_size` classes

respectively, which satisfy the `gci:Population_ratio_unit` numerator and denominator constraints.

- **`gci:city`**, whose range is a `geo:Feature` that uniquely identifies the city for which this is an indicator.
- **`gci:teacher_def & gci:student_def`**, whose range are a subclass of `Teacher` and `Student` respectively. These define the properties of the teachers and students that we are measuring. For example, all full time students in grades 1 through 12.

The `Quantity` instance would link the object being measured (i.e., City) with the actual measurement being an instance of a `Measure`. The instance of `Measure` then contains the measurement's numeric value and a link the Unit of measure.

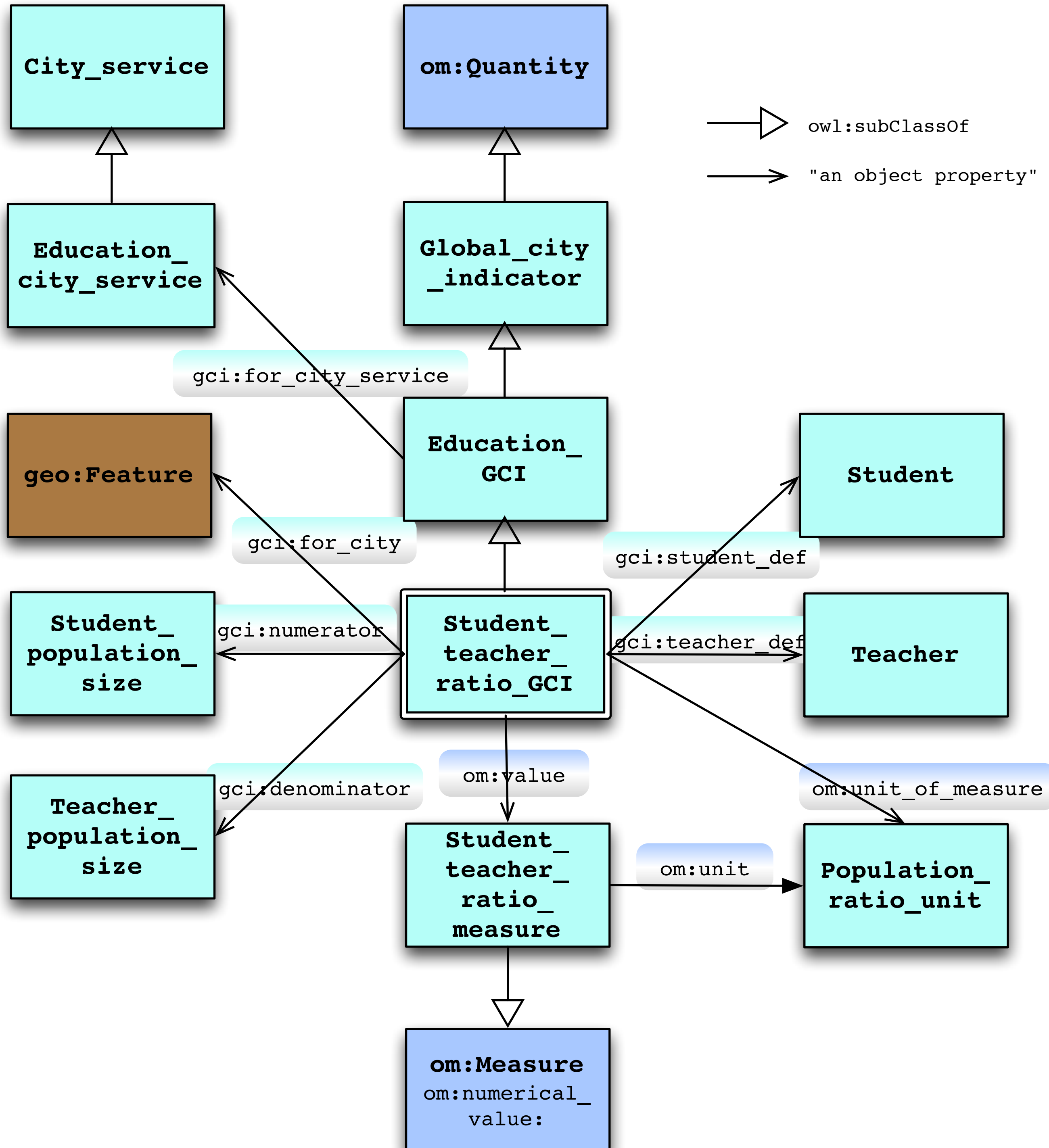
City_service
om:Quantity
owl:subClassOf
"an object property"
Education_ city_service
Global_city _indicator
gci:for_city_service
geo:Feature
Education_ GCI
Student
gci:for_city
gci:student_def
Student_ population_ size
gci:numerator
Student_ teacher_ ratio_GCI
gci:teacher_def
Teacher
Teacher_ population_ size
gci:denominator
om:value
om:unit_of_measure
Student_ teacher_ ratio_ measure
om:unit
Population_ ratio_unit
om:Measure
om:numerical_ value:

Figure 7

At this point you may have noticed that neither the `gci:Student_population_size` nor `gci:Teacher_population_size` have been linked to the students nor teachers within a city. We do so in the next section where we introduce the statistics ontology.

4.3. Statistics Ontology

- What defines the members of the population?
- What is its unit of measure?
- Where is the population being counted?

The STR indicator is based on a measure of the number of students and teachers within a population designated by a Placename, namely a city. One can view both sizes as a statistical measurement in the sense that there is a population that we want to perform a measurement of, namely a school population, and we are counting the number of members that satisfy a description of a Student and a Teacher, respectively. While the STR requires a count of the population, other measures would require statistical measures of mean, deviation, etc. of other characteristics of the population.

Anticipating the larger requirements of the Global City Ontology, we have adopted the GovStat[11] general statistics ontology (Pattuelli, 2009). Figure 8 depicts the main classes and properties of the GovStat ontology. The core class is the `gs:Population` to be measured. (A definition of the population is not provided and will be part of our extension to GovStat in Figure 9.) A `gs:Population` is linked to a parameter (e.g., mean, standard deviation) by the `gs:is_described_by` property, and the parameter is a sub class of `gs:Parameter`. In statistics it is almost always the case that only a portion of the population is measured. This portion is represented by the class `gs:Sample`, and the parameter being measured is represented as a subclass of `gs:Statistic`. Finally, the variable for which the parameter is being measured is defined by the class `gs:Observation` which `gs:Statistic` links to via the property `gs:is_composed_of`, and the actual variable which is a subclass of `gs:Variable` is linked to `gs:Observation` via the property `gs:is_a_characteristic_of`.

[11] The GovStat Ontology is not available online, but a version with the GCI extensions can be found at: http://ontology.eil.utoronto.ca/govstat#. We will use the prefix "gs:" to identify classes and properties from the ontology.

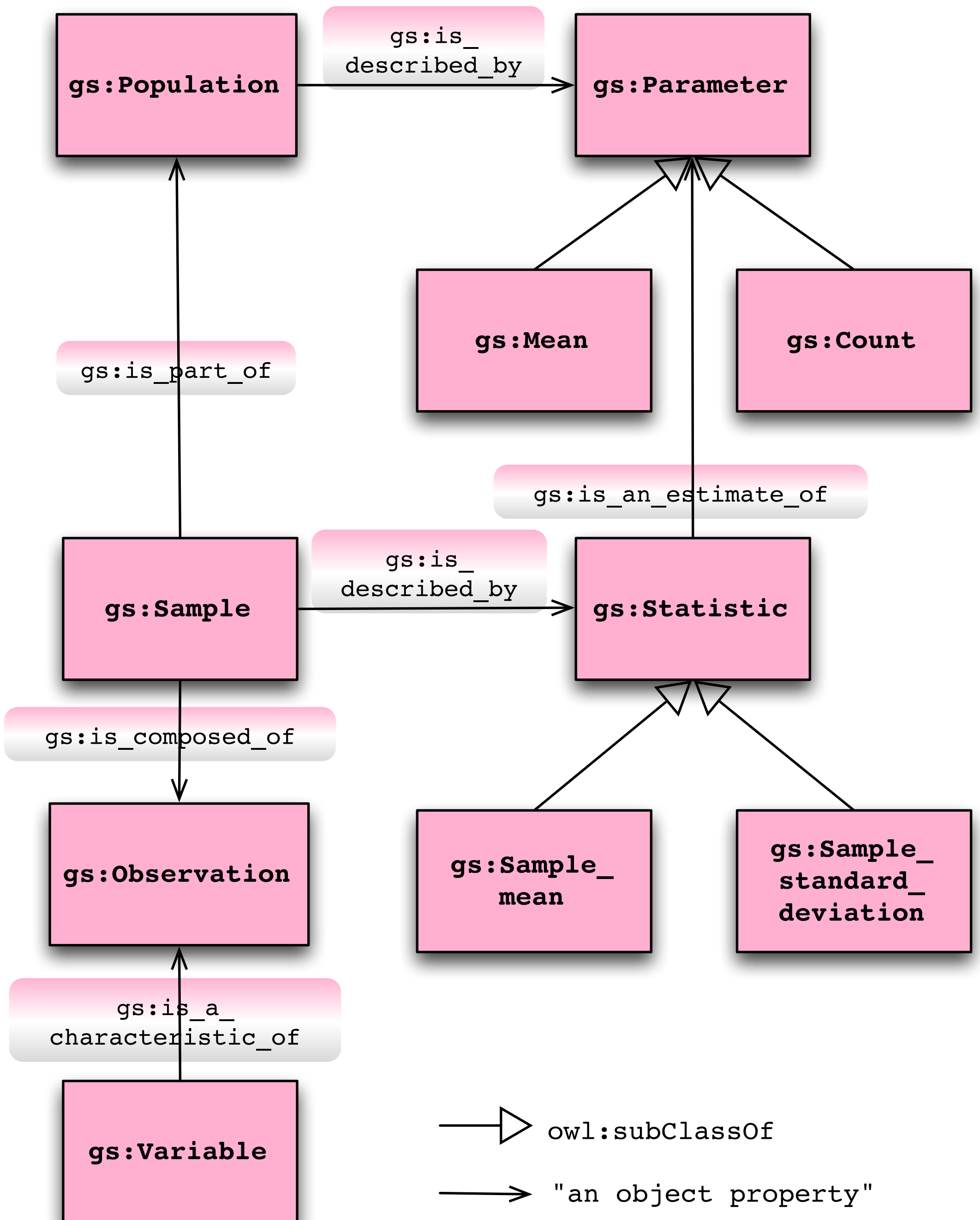

Figure 8

What we are missing at this point is a definition of the population that we are measuring or from which a sample is to be taken. For the STR indicator the `gs:Population` must identify the area in which the population resides, i.e., the city, and what characterizes a member of the population, namely the characteristics of a Student or Teacher. For example, the characteristics of a Teacher could be:

- Fulltime, defined as teaching 30 or more hours per week, and
- Teaches at the primary or secondary level, where primary spans grades 1 thru 8 and secondary spans 9 thru 12.

As depicted in Figure 9, we have extended the GovStat ontology as follows:

- Added a property to `gs:Population`, `gs:located_in`, that identifies the area that the Population is drawn from.
- Added a property to `gs:Population`, `gs:defined_by`, that identifies the class that all members of the Population are subsumed by.

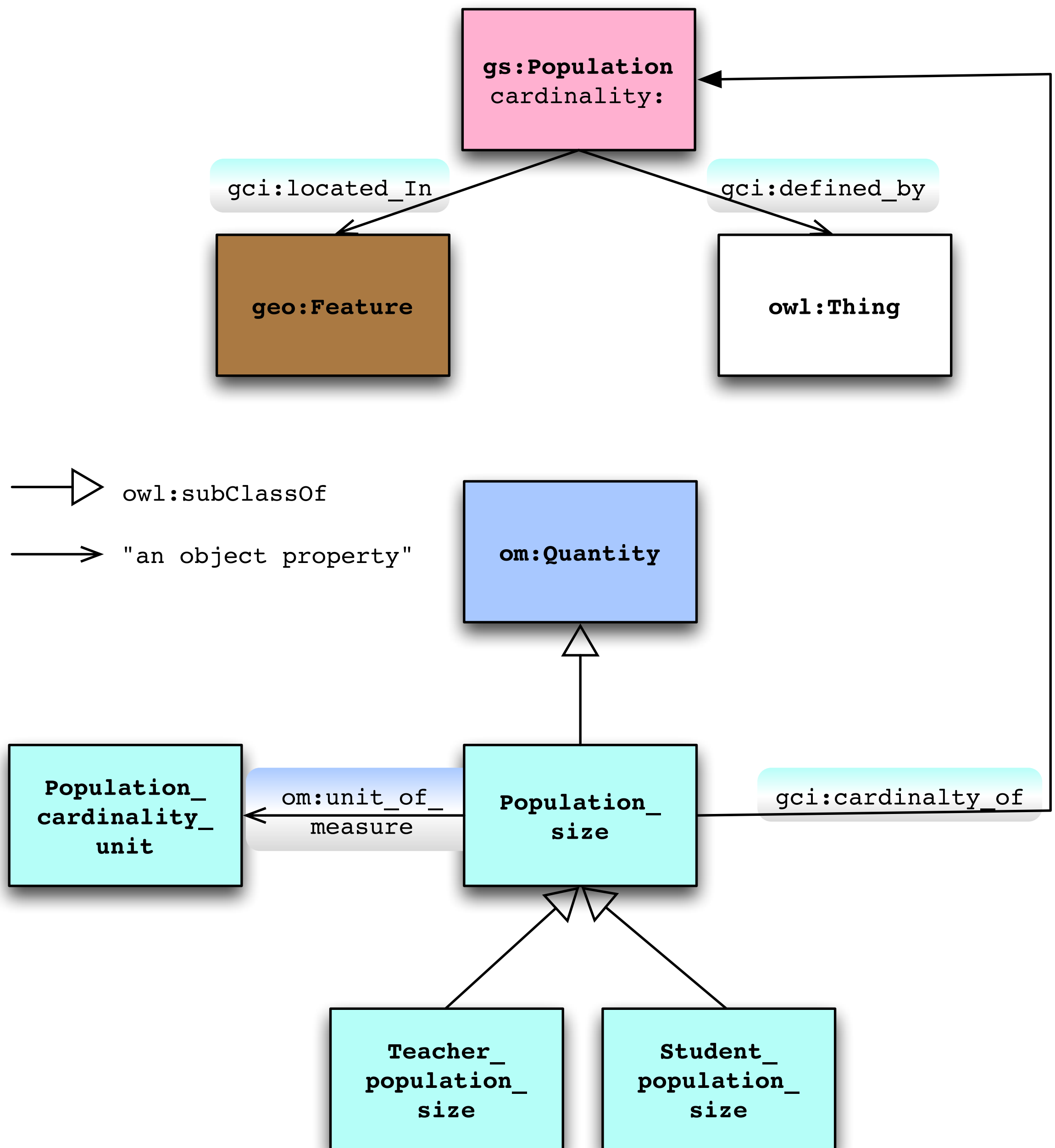

Figure 9

In order to complete definition of `gci:Population_size` pictured above, we need a further constraint. The property of (`gs:is_property_of`) the `gci:Population` must be a `gs:Count` parameter.

4.4. Provenance Ontology

- Who created the actual value of the GCI?
- When was it created?
- What process was used to create it?
- Has this GCI been revised?

Up to this point we have focused on the representation of indicator itself. But another important aspect of an indicator is its providence, namely where did it come from and how was it derived. Over the last decade, concerns around information validity, provenance and trust have grown. With the web now containing billions of documents authored by millions of people, the need to know whether the content is valid, where the content came from and whether to trust its creator has taken on an increasing importance. With respect to an indicator, we need to know:

- Who created it,
- What activities were performed to generate it,
- What datasets were used in its generation, and
- When was it generated?

Much of the research into provenance has grown out of workflow management where the focus has been the evolution of a document as it proceeds through a sequence of edits, perhaps by different people and/or systems. Tracking the various versions created, who did what and when has been the primary concern. This research has culminated in the proposed Semantic Web standard called the PROV Ontology[12] (Belhajjame et al., 2012), which has based on the work of Hartig & Zhao (2010) and Moreau et al. (2010). In the following we outline the basic concepts of the PROV ontology and indicate how it is incorporated into the GCI ontology.

At the heart of the PROV ontology are three classes:

- **`pr:Entity:`** represents any artifact for which we want to specify its provenance. In our case it would be an indicator or the data from which the indicator was directly or indirectly derived.
- **`pr:Activity`**: the action (or sequence of actions) that creates or transforms an entity. In our case it may be a computation performed over some data set such as census data.
- **`pr:Agent`**: the person, organization, or system that performs or plays some role in the activity that transforms an entity. In our case it may be a software application that mines a data set or a person who reviews a data set.

---

[12] The PROV Ontology can be found at: http://www.w3.org/ns/prov#. We will use the prefix “pr:” to identify classes and properties from the ontology.

Along with these classes are defined a set of properties that define the causal relationship among entities and activities:

- **`pr:wasGeneratedBy`**: It links an `pr:Entity` (domain) to a `pr:Activity` (range), identifying the activity that generated the entity.
- **`pr:used`**: It links an `pr:Activity` (domain) to an `pr:Entity` (range), identifying the entities used by an activity.
- **`pr:wasAssociatedWith`**: It links an `pr:Activity` (domain) to a `pr:Agent` (range), identifying the agents that play a role in the activity.
- **`pr:wasAttributedTo`**: It links an `pr:Entity` (domain) to an `pr:Agent` (range), identifying the agents that had a role in creating the entity.
- **`pr:wasRevisionOf`**: Links two `pr:Entity`'s where domain entity is a revision of the range entity.
- **`pr:wasDerivedFrom`**: Links two `pr:Entity`'s where domain entity was derived from the range entity (without indicating the method of derivation).

Finally, the PROV ontology provides a time property that specifies the time an entity was created.

- **`pr:generatedAtTime`**: It links a `pr:Entity` (domain) to a `pr:time` (range), identifying the time the entity was generated.

Figure 10 depicts the integration of the PROV ontology into the GCI ontology as follows. First, make `gci:Global_City_Indicator` a `owl:subClassOf pr:Entity`. Consequently, every indicator we create will be treated as a `pr:Entity` and inherit its properties, including `pr:generatedAtTime` which provides us with the time that the indicator was created, and `pr:wasRevisionOf` which allows us to track revisions to the value of the indicator. It also allows us to link the GCIs to a `pr:Activity` via a `pr:wasGeneratedBy` to show what activity generated the GCI, and to a `pr:Agent` via a `pr:wasAttributedTo` to show who the source of the GCI was. Finally, the `gci:numerator` and `gci:denominator` are made to be `owl:subPropertyOf pr:wasDerivedFrom` to show what entities were used to derive the GCI.

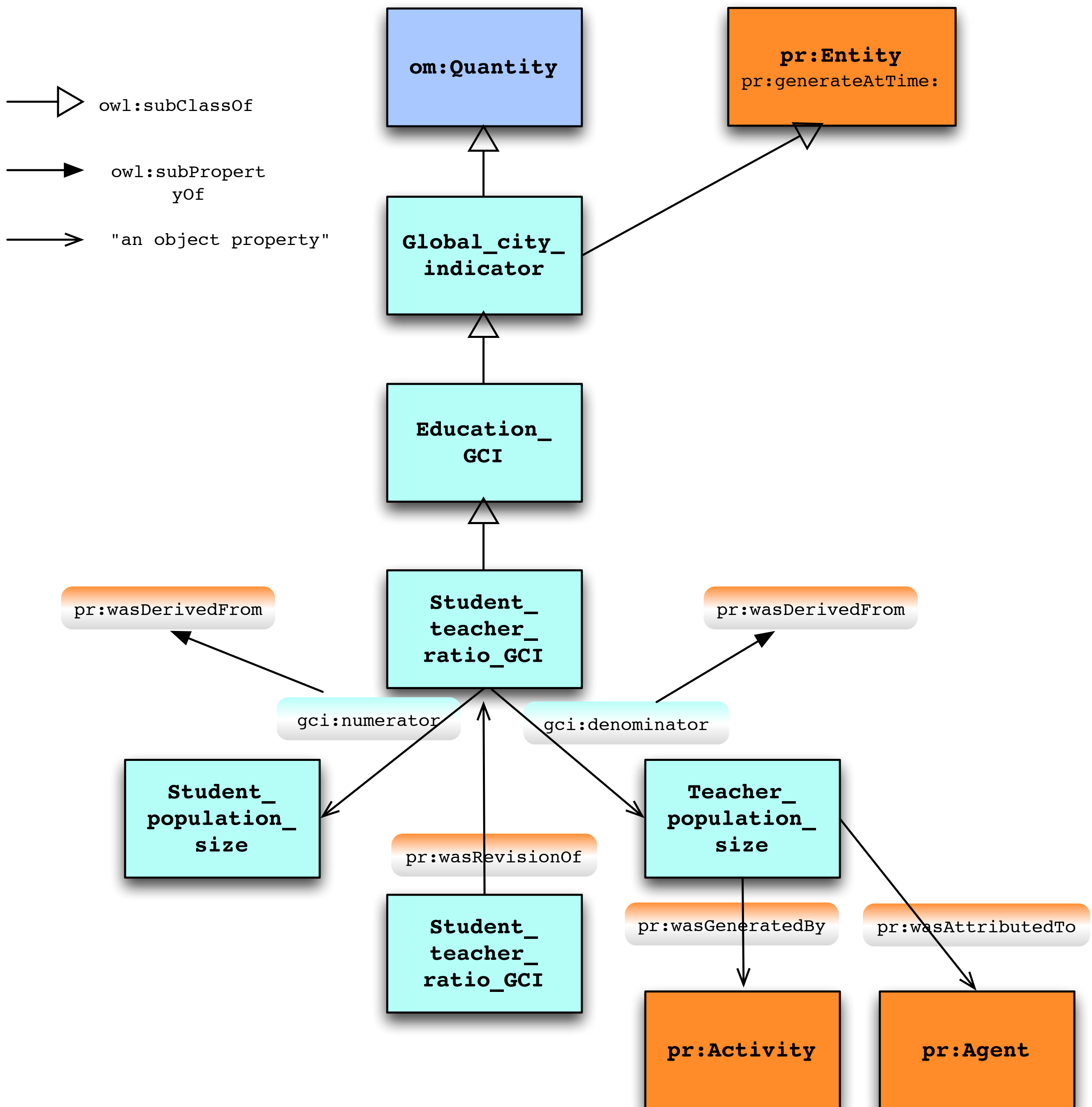

Figure 10

4.5. Time Ontology

- Over what period of time was this GCI constructed?
- How long did it take?
- Was the teacher population sizing done during the same time that the student population sizing was done?

Fundamental to the concept of provenance is the time at which measurements are taken, computed or derived. Questions may arise regarding the temporal relationship among indicators and among measurements. Not just at what time something occurred, but whether something occurred before, after or during some external event. For example, was "Total Employment" of New Orleans determined before or after Hurricane Katrina? Or did Katrina take place during the interval that the indicator was determined? To answer these questions, we need a much richer notion of time that supports reasoning about time points, time intervals and the relationships amongst them.

Many time ontologies have been developed. We have chosen OWL-Time[13] for its simplicity and ability to represent time as a point or interval. OWL-Time is based on the work of Allen & Ferguson (1997) and described in Hobbs & Pan (2006).

The root class for OWL-Time is the `ot:TemporalEntity`. It has two subclasses:

- **`ot:Instant`**: It represents a time point.
- **`ot:Interval`**: It represents a period of time with a beginning and an end. An `ot:ProperInterval` is an ot:Interval where the start time is less than the end time.

An **`ot:Interval`**'s starting point, ending point and duration are denoted by the following properties:

- **`ot:hasBeginning`**: links a `ot:TemporalEntity` (domain) to an `ot:Instant` (range) where the latter denotes the beginning of the `ot:TemporalEntity`.
- **`ot:hasEnd:`** links a `ot:TemporalEntity` (domain) to an `ot:Instant` (range) where the latter denotes the end of the `ot:TemporalEntity`.
- **`ot:hasDurationDescription`**: links a `ot:TemporalEntity` (domain) to an `ot:Interval` (range) where the latter denotes the duration of the `ot:DurationDescription`.

Two other classes of note are:

- **`ot:DateTimeDescription`**: A specification of a date plus time using a year, month, day, hour, etc. set of properties.
- **`ot:DurationDescription`**: is a class whose instance can combine multiple descriptions such as 2 days and 2 hours to specify a duration.

Finally, there is a set of properties that relate `ot:ProperInterval`'s, including `ot:inside`, `ot:intervalOverlaps`, `ot:intervalAfter`, `ot:intervalContains`, etc.

Figure 11 depicts the addition of the time ontology to our ontology. The integration of the time ontology occurs with the `pr:Entity`. We modify `pr:generatedAtTime` from being a data property to an object property whose range is an `ot:TemporalEntity`. With this change the time of an `pr:Entity` can be either a point or an interval. Secondly, it can take advantage of the relational reasoning (i.e., is the generation time of a `pr:Entity` before, during, or after some other `pr:Entity` or event) supported by the ontology.

---

[13] The OWL-Time Ontology can be found at: http://www.w3.org/2006/time. We will use the prefix "ot:" to identify classes and properties from the ontology.

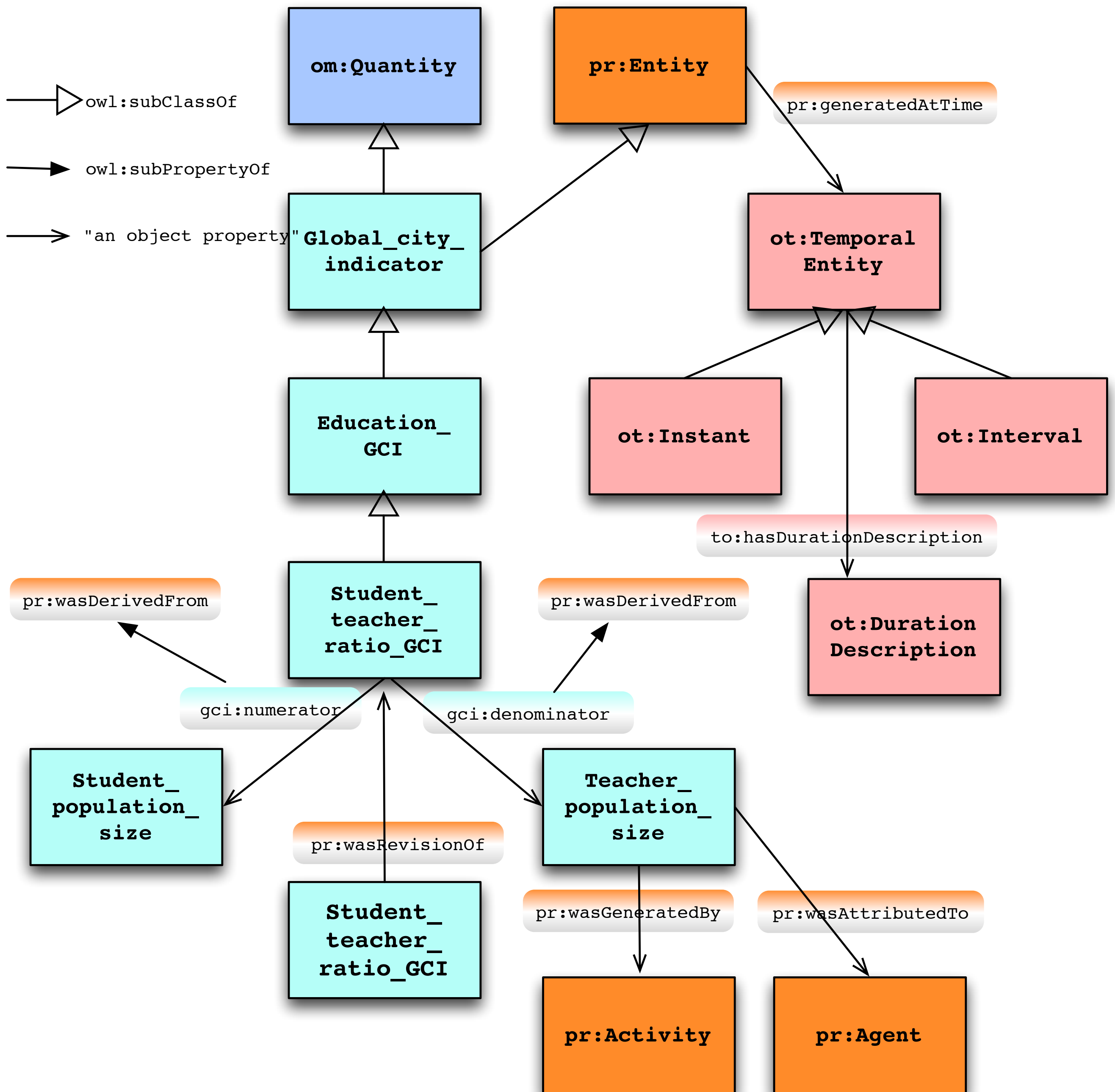

Figure 11

4.6. Validity Ontology

- Is the GCI believed to be an accurate measure by its creator?
- Over what time is it believed to be accurate?

An ongoing issue with the web is whether information/data found on a page is correct (true) or incorrect (false). Whether the creator of the information deliberately makes false statements, or unknowingly copies false information from another site, there is no way to discern what is correct from incorrect. The same holds with city indicators. Data and analyses that are believed to be true at the time they are gathered or computed, may be found over time to be incorrect. Or it may not be clear whether the information is true or not, especially if the indicator is based on a sampling of a population, but one can assign a degree of validity to the information. In addition, in the case where data is derived from other data, and the latter is no longer valid at some point of time, then the former becomes invalid for that same point of time. For example, `gci:Student_teacher_ratio` is derived from `gci:Student_population_size` and `gci:Teacher_population_size`, if `gci:Student_population_size` is valid only within an interval of time such as the year for which it is gathered, then outside of that interval, both `gci:Student_population_size`'s and its dependent `gci:Student_teacher_ratio`'s validity are unknown.

Fox & Huang (2005) provide an ontology, called the Knowledge Provenance Ontology[14] (KP), for representing the validity (certainty) of a proposition. It assigns to a "proposition" a validity between [0,1] or "unknown." This validity may be dynamic in that it changes over time. An example of the latter is any population count that is representative of the population only at a point of time or for an interval of time. The time interval during which the proposition's validity is known is called the "effective" time interval.

A set of axioms are defined in Huang & Fox (2004a; 2004b) that define how validity is propagated within a dependency network. In the simple case, if a GCI is assigned a validity of 1 (i.e., it is true) but it also has an effective time interval specified for it, then the GCI is valid during that time interval and unknown otherwise. If any GCIs' validity are unknown during a time interval then any GCIs' validity that dependent on it are also unknown during the same time interval.

At the core of KP is the `kp:KP_prop` class which identifies a proposition to which a validity, effective time interval and dependencies can be assigned. We add to the definition of `gci:Global_city_indicator` that it is a `owl:SubClassOf kp:KP_prop (`Figure 12`).` Hence any `gci:Global_city_indicator` is also a proposition to which we can assign a validity, effective time interval and dependencies.

[14] The Knowledge Provenance Ontology can be found at: http://ontology.eil.utoronto.ca/kp#. We will use the prefix "kp:" to identify classes and properties from the ontology.

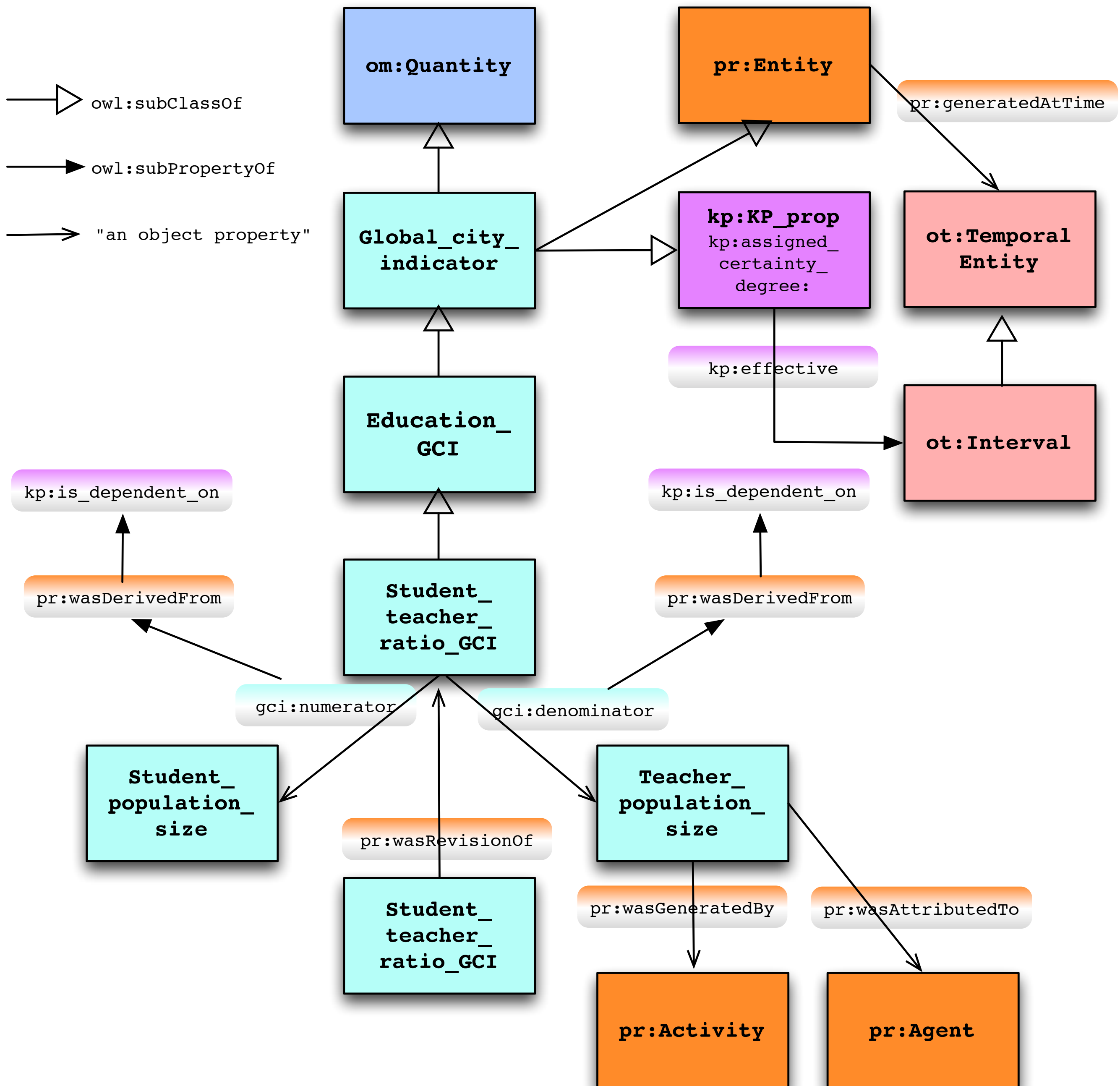

Figure 12

The following properties are associated with a `kp:KP_prop` and are inherited by all `gci:Global_city_indicator`'s:

- **`kp:assigned_certainty_degree`**: This is a data property that maps a `kp:KP_prop` (domain) onto a number [0,1] (range) or unknown. It is the degree of certainty that the proposition is valid (true) from the perspective of the creator of the `gci:Global_city_indicator` instance.
- **`kp:effective`**: This is a data property that maps a `kp:KP_prop` (domain) onto a time

interval (range). It is the time during which the `kp:assigned_certainy_degree` is valid for the `gci:Global_city_indicator` instance.

- **`kp:is_dependent_on`**: This is a object property that maps a `kp:KP_prop` (domain) onto another `kp:KP_prop` (range). It states that the `kp:assigned_certainty_degree` for the `gci:Global_city_indicator` instance is dependent upon one or more `kp:KP_prop`'s.

Given that `kp:is_dependent_on` is a generalization of `pr:wasDerivedFrom`, we add to the KP ontology that `pr:wasDerivedFrom` is an `owl:subPropertyOf` `pr:wasDerivedFrom`.

4.7. Dynamic Placenames

- Has the city's boundary changed during the time between two measures of an indicator?

Consider the unique placename for the City of Toronto. If we wish to do a longitudinal analysis of an indicator for Toronto, we run into a problem. The geographic definition of Toronto changed in 1998 after its amalgamation with five adjacent municipalities. Yet in the Geonames ontology there is a single Toronto; there is no representation for how placenames evolve over time. Kauppinen and Hyvönen (2007) have addressed this problem. They propose an ontology based on Spatial Temporal Regions. A placename has associated with it a spatial region, defined by a polygon, and a time interval over which the placename and the region do not change.

In the Global City Ontology we will refer to placenames whose spatial regions can change over time as Dynamic Placenames. Rather than adopt Kauppinen and Hyvönen's terminology directly, we adapt their ideas by reusing the provenance, time and validity ontologies to represent how place names change over time and the cause of their change.

Figure 13 depicts simplified example of how to represent a dynamic placename for the city of Toronto. First, the placename for each version of the City of Toronto will have to be unique. In the example we append the time period for each version to the name, though just having a unique number is sufficient. We link the [1998-] version of Toronto to the [1967-1998] version via a `pr:wasRevisionOf` property from the provenance ontology to show that the former is a revision of the latter. Secondly, for each placename we link it using the validity property `kp:effective` to a time interval over which the placename is valid.

From a longitudinal analysis perspective, when we compare indicators for a single city over time, we will know the extent to which these comparisons are valid as the city's composition may have changed over time.

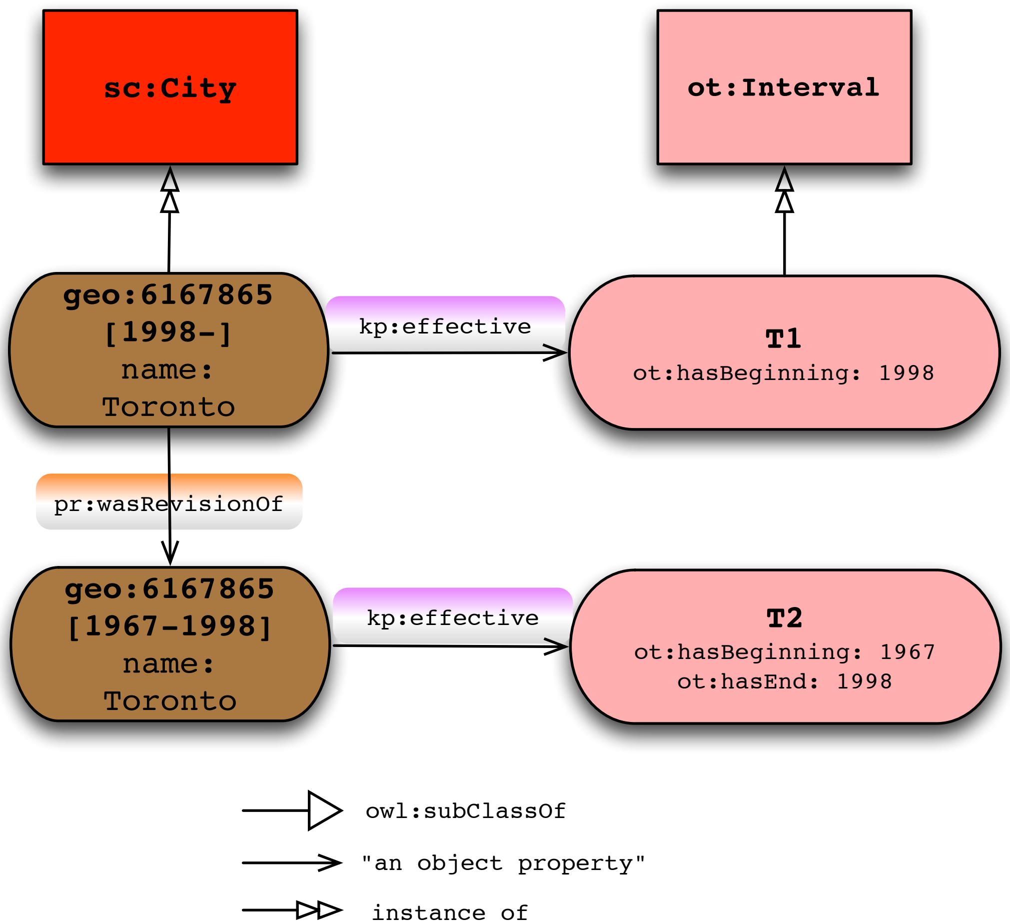

Figure 13

4.8. Trust Ontology

- Do you trust the creator of the GCI?
- Do you trust the process used to create the GCI?

The final piece of the GCI ontology "puzzle" is the representation of trust. The problem we wish to address is how to represent the degree of trust we have in the creator of indicator values and the data from which they are derived. Huang & Fox (2006) define trust as follows:

> *"Trust is the psychological state comprising (1) expectancy: the trustor expects a specific behavior of the trustee such as providing valid information or effectively performing cooperative actions; (2) belief: the trustor believes that expectancy is true, based on evidence of the trustee's competence and goodwill; (3) willingness to be vulnerable: the trustor is willing to be vulnerable to that belief in a specific context where the information is used or the actions are applied."*

This representation of trust differs from degree of validity as trust refers not to the degree of certainty in the data but our trust in the agent/organization that produced the data. The obvious example is how to represent the trust we have in an organization that has a history of “cooking the numbers.” The consequence of not having trust in the producer of data is that the validity one assigns to data or indicator will be reduced by this lack of trust.

Huang & Fox (2006) and Huang (2008) provide an ontology of trust[15]. The ontology views trust as occurring between two agents, where $agent_1$ has or has not trust in $agent_2$. Trust arises out of direct experience or the experience of others whom you may trust. Trust is also context dependent. For example, $agent_1$ may trust $agent_2$ in providing information on topics relevant to their expertise, such as a meteorologist characterizing the climate of a city, but lacks trust in $agent_2$ outside of their field of expertise. Finally, they identify two types of trust: 1) *trust in belief*, where $agent_1$ believes what $agent_2$ believes, and 2) *trust in performance*, where $agent_1$ believes that $agent_2$ will perform an activity properly.

[15] The Trust Ontology can be found at: http://ontology.eil.utoronto.ca/trust#. We will use the prefix “tr:” to identify classes and properties from the ontology.

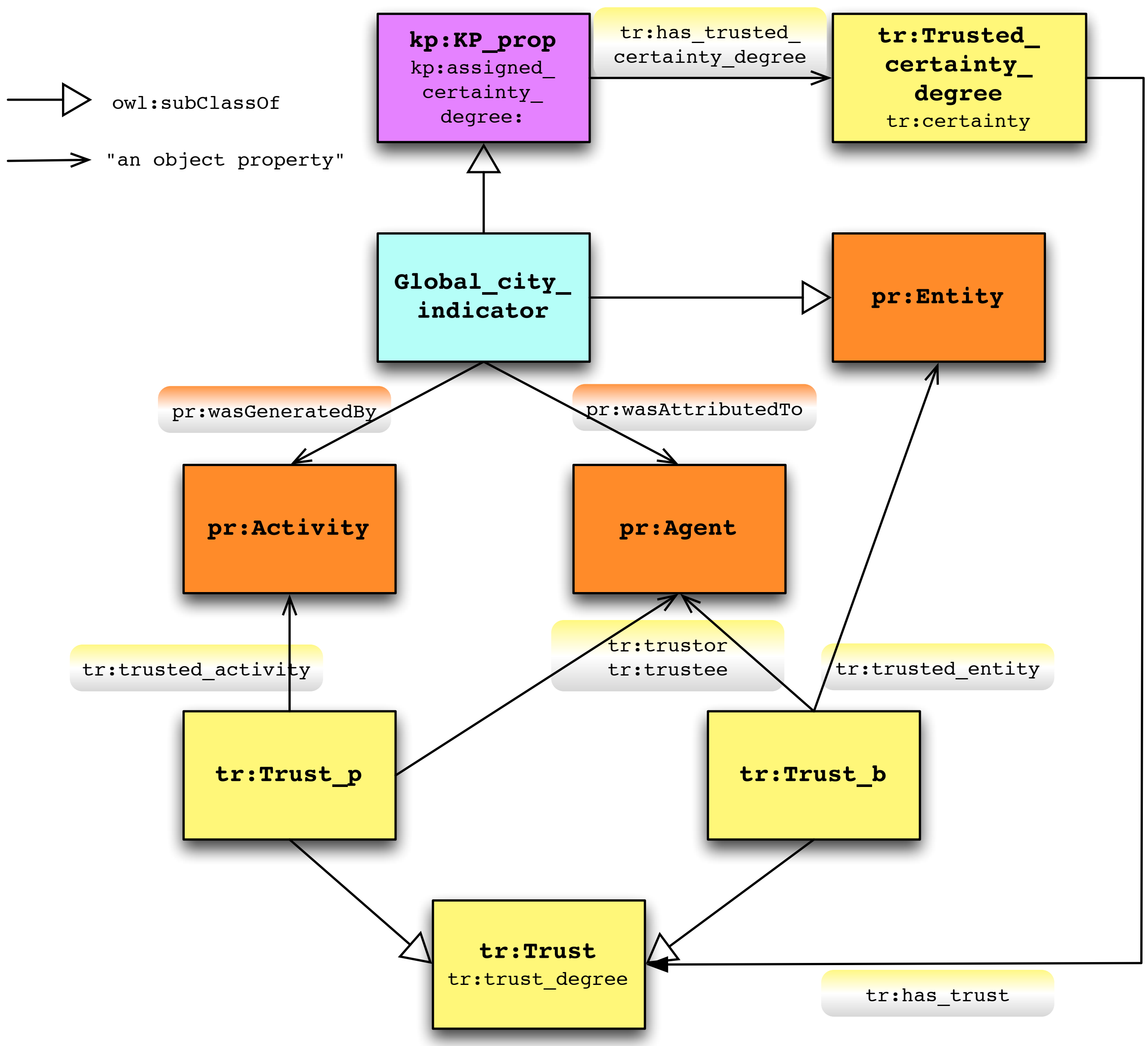

Figure 14

The Trust Ontology also addresses how the validity of an indicator or data changes by taking the original degree of validity, asserted by the creator (agent$_2$), and modifying it by the degree of trust the "user" (agent$_1$) has in the creator. This resultant validity is dependent on agent$_1$ and agent$_2$.

In Figure 14 we integrate the Trust Ontology by defining the class `tr:Trust` which has a data property `tr:trust_degree` which is the degree to which the `tr:trustor` trusts the `tr:trustee`. It is specialized into two classes:

- **`tr:Trust_p`**, which is trust in performance. It has an object property `tr:trusted_Activity` that links it to a `pr:Activity`, and

- **`tr:Trust_b`**, which is trust in belief. It has an object property `tr:trusted_Entity` that links it to a `pr:Entity`, which all `gci:Global_city_indicator`'s are a subclass of.

We then extend the KP Ontology by adding an object property, `tr:Trusted_certainty_degree`, that links a `kp:KP_prop` to a new class that represents the certainty degree computed by combining the GCI's `tr:assigned_certainty_degree` provided by its creator (who is the `tr:trustee`) and the trust that the user (who is the `tr:trustor`) has in the creator. The latter is represented by an object property `tr:has_Trust` that links the `tr:Trusted_certainty_degree` to the `tr:Trust` that represents the user's belief in the creator.

5. Consistency

Now that we have defined the foundational ontologies for placenames, measurement, statistics, provenance, validity and trust, and its extensions to represent a student/teacher ratio, we extend the ontology by introducing a set of consistency rules (axioms). These rules make sure the various parts of an indicator refer to the same places, have the same units, etc. They are obvious and simple extensions to the ontology, but necessary in that they can detect errors that commonly occur in datasets.

In the following we informally describe each rule. Each rule is implemented in prolog. The prolog implementation is available – see the Appendix.

5.1. Placename Rules

The purpose of these rules is to check that the City associated with the STR is consistent with the cities associated with each of the Teacher and Student populations. We want to assure that they are referring to and measuring the populations for the same geographic area.

> **Rule G1:** The city for the STR being measured is the same as the cities where its numerator and denominator are measured.

The city for the STR is defined by its `gci:city` property. The placename for the city must match the placename specified in the `gci:located_in` property attached to the `gs:Population` that is linked to the `gci:Population_size` class via a `gci:cardinality_of`. See Figure 15.

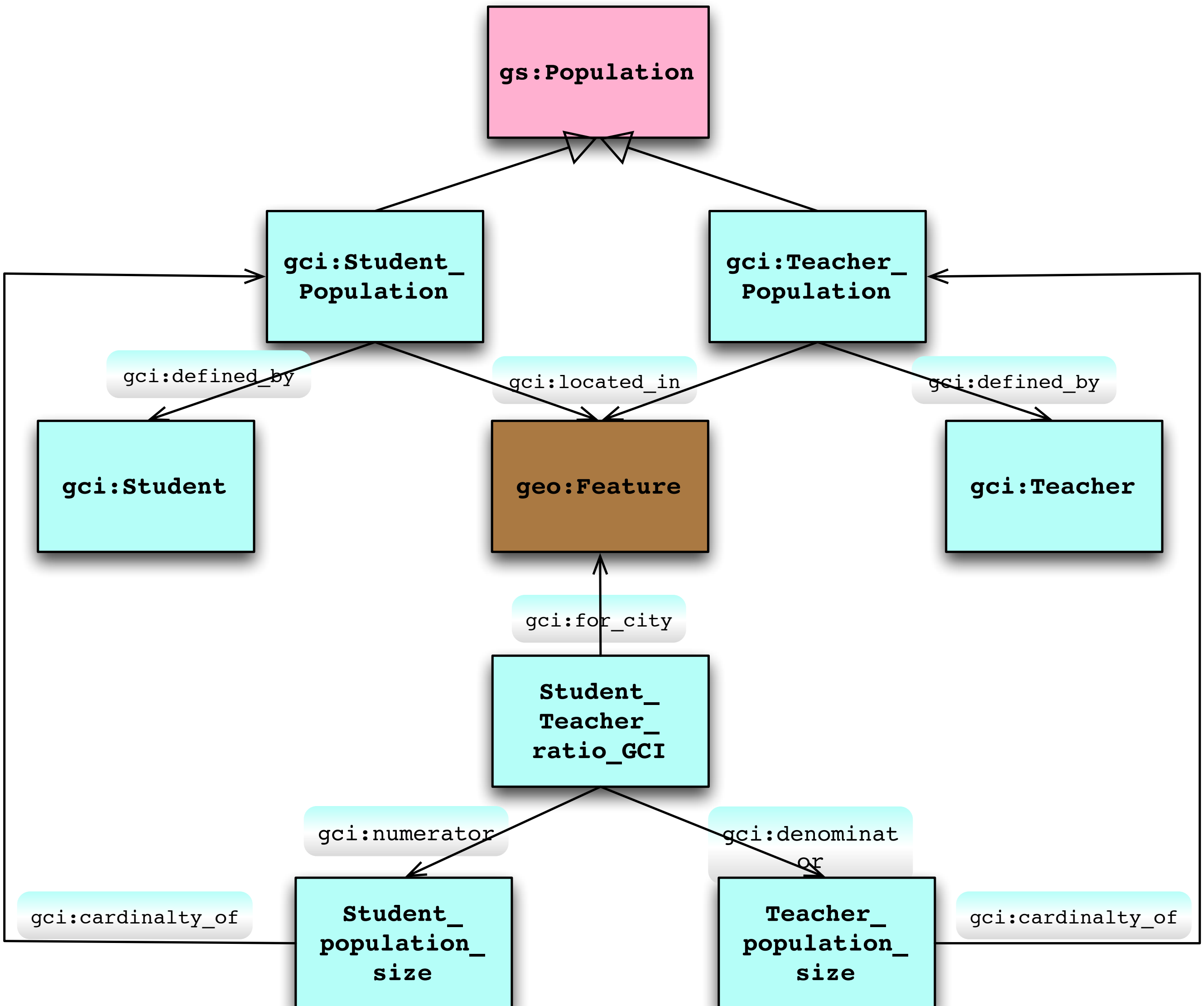

Figure 15

5.2. Measurement Rules

These rules define the consistency of measurements used in a STR.

> **Rule M1:** The numerator and denominator of a `gci:Student_teacher_ratio_GCI` are the correct type.

Rule M1 verifies that the numerator and denominator are of the types specified by the definition of the `gci:Student_teacher_ratio_GCI`, namely that the numerator is a `gci:Student_population_size` and the denominator is a `gci:Teacher_population_size`. The rule is generalized to apply to any GCI that is a ratio.

> **Rule M2:** The numerator and denominator of the `gci:Student_teacher_ratio_GCI` are consistent with the numerator and denominator of its unit of measure.

Rule M2 verifies that the unit of measures of the STR's numerator and denominator, are consistent with the units specified by the `gci:Population_ratio_unit`. As defined, the `gci:Population_ratio_unit`'s numerator and denominator have to be `gci:Population_cardinality_unit`'s. The rule is generalized to apply to any GCI that is a ratio.

> **Rule M3:** If the numerator and denominator of a `gci:Student_teacher_ratio_GCI` are the same type, then they should have the same unit of measure.

Since the numerator and denominator of the STR are population counts, then Rule M3 verifies that the numerator and denominator are of the same units, e.g., you cannot have `gci:Student_population_size` measured in `gci:kilopcs` and `gci:Teacher_population_size` measured in `gci:pcs`. The rule is generalized to apply to any GCI that is a ratio.

> **Rule M4:** The units of the actual measurement are the same as defined by GCI it is a measure of.

In the STR case, the unit of measure of the `gci:Student_teacher_ratio_GCI` and the unit for the `gci:Student_teacher_ratio_measure` must be the same, namely a `gci:Population_ratio_unit`.

> **Rule M5:** The value of the `gci:Student_teacher_ratio_measure` is equal to the value of the `gci:Student_teacher_ratio_GCI` numerator divided by the denominator.

Though rule M5 is obvious, it is still necessary to check that calculations are performed correctly.

5.3. Statistics Rules

The statistics rules assure that the populations being measured are consistent with the indicator in which they are being used.

> **Rule S1:** The definitions of student and teacher as specified by the `gci:Student_teacher_ratio_GCI` are the same as used by its numerator and denominator.

This rule checks to see that the `gci:Student_population` pointed to by the STR's numerator and the `gci:Teacher_population` pointed to by the STR's denominator have `gci:defined_by` ranges that are consistent with the STR's `gci:student_def` and `gci:teacher_def` respectively.

5.4. Provenance Rules

Provenance is used to document the participants and the means with which an indicator was generated. It is used for forensic purposes such as determining whether the same methods were used to generate different versions of the same indicator for the same city.

Rule P1: Two versions of the same indicator are inconsistent with each other if different methods were used to generate them.

Consider the situation where the same indicator is measured annually. Each version of the indicator is linked to the prior year's version via `pr:wasRevisionOf` property. In order to assure that the two versions are comparable, we have to assure that the same methodology was used for each. This is done by comparing the `pr:Activity` used to generate numerator for each version, and doing the same for the denominator.

Rule P2: Two versions of the same indicator are inconsistent with each other if the cities are not the same.

A difference of cities can arise because the wrong placenames have been used, or in a dynamic placename situation, the city itself has undergone a change, such as a merger, during the last year.

One could imagine a rule that relates the time the indicator was generated, as recorded by its provenance, with its effective period, but its effective period could either be prior to generation or after, depending on policy.

5.5. Validity Rules

The validity rules assure that the time period for which the STR is specified to be valid, is consistent with the data from which it is derived.

Rule V1: The effective time period for which the STR is valid is contained within the effective time periods of its numerator and denominator.

The numerator and denominator of the STR, namely the `gci:Student_population_size` and `gci:Teacher_population_size`, must have effective time periods that at least overlap, and the STR's effective time period must be contained within that overlap.

Rule V2: The `kp:assigned_certainty_degree` of the STR is less than or equal to the max of the `kp:assigned_certainty_degree` of its numerator and denominator, and greater than or equal to the min of its numerator and denominator.

The STR is a function of the Student and Teacher population sizes. Hence it's trust degree cannot be more/less than the max/min of the individual trusts of the measures it depends on.

5.6. Trust Rules

The trust rules assure that the agents in the trust relations are consistent and they refer to the same Entities.

Rule T1: The trustee in a trust relationship is the same as the `pr:wasAttributedTo` `pr:Agent` for an indicator.

The STR indicator must be linked to a `tr:Trusted_certainty_degree` which is in turn linked to a `tr:Trust_b` (i.e., trust in the belief of an `pr:Agent`), which in turn points to a `tr:trustee` `pr:Agent`, that trustee must also be the `pr:Agent` that created the indicator (which is pointed to by the `pr:wasAttributedTo` link of the STR indicator).

> **Rule T2:** The trusted certainty degree of an indicator should have a Trust instance that links to the indicator via the `tr:trusted_entity` property.

> **Rule T3:** The trusted certainty degree of an indicator is less than or equal to the indicator's certainty assigned by its creator.

The point here is that the trusted certainty degree cannot be greater than the certainty assigned by the indicator's creator. It can only be reduced.

6. Example

In this section we show how a specific instance of the STR used throughout this paper is represented as instances (individuals) of the ontology. The instances are represented as bifurcated rectangles where the top part identifies the name of the instance followed by the class it is an instance of (<instance name> -->> <class name>). The bottom part contains data properties and object properties for which, for brevity, we do not depict using a link.

Let's assume that we want to create a STR for the city of Toronto, all we have to do is the following (Figure 16):

- Create an instance (`ex:TO_str`) of quantity `gci:Student_teacher_ration_GCI`. This will be starting point of the value for the city of Toronto's student teacher ratio indicator. Set the `gci:for_city` object property to the URI of the Toronto placename.
- Create an instance of measure `gci:Student_teacher_ratio_measture` (`ex:TO_str_m`), fill the `om:numerical_value` property with the actual ratio (40 in this example). Link `ex:TO_str` to `ex:TO_str_m` using the `om:value` object property.

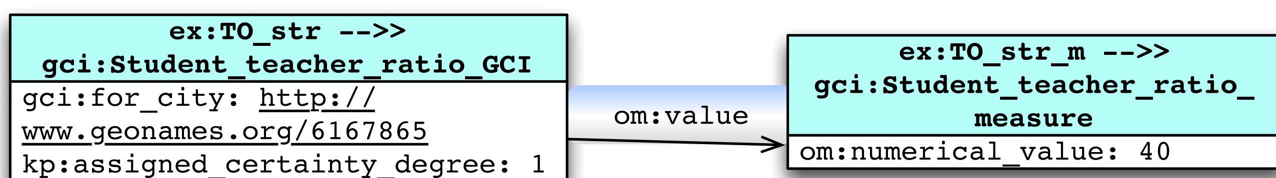

Figure 16

With this representation, based on the classes these are instances of, we know that `ex:TO_str` represents a Student-Teacher ratio for the Education city service for the city of Toronto. The unit of measure is the ratio of two Population sizes, where the populations are defined by the Student and Teacher classes, respectively.

Next we will add two types of provenance (Figure 17):

- The date/time this ratio was created (23 January 2013 at 10am), and
- The Agent who created it (Joe Smith),

by instantiating their respective classes.

We then add the validity of the indicator. Since `gci:Student_teacher_ratio_GCI` is a subclass of `kp:KP_prop`, it inherits the `kp:assigned_certainty_degree` data property which we set to 1, i.e., the creator believes the value of the indicator is true.

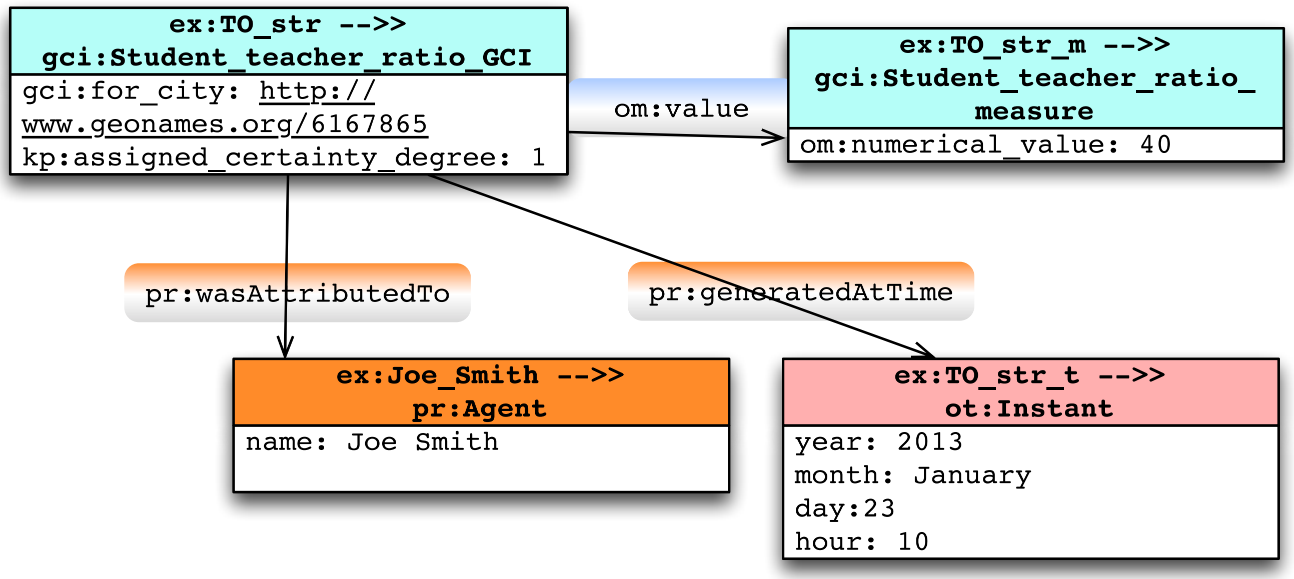

Figure 17

Based on the above, we know that the source of the indicator was Joe Smith, which he created on January 23rd, 2013 at 10am, and that Mark Fox believes the indicator to be absolutely correct.

Lastly (Figure 18), we add the degree we trust Joe Smith by doing the following:

- Adding the trustor, Mark Fox, and the degree to which he trusts Joe Smith by instantiating the trust in belief class (`tr:Trust_b`).
- Adding the trusted certainty degree to the Student Teacher ratio.

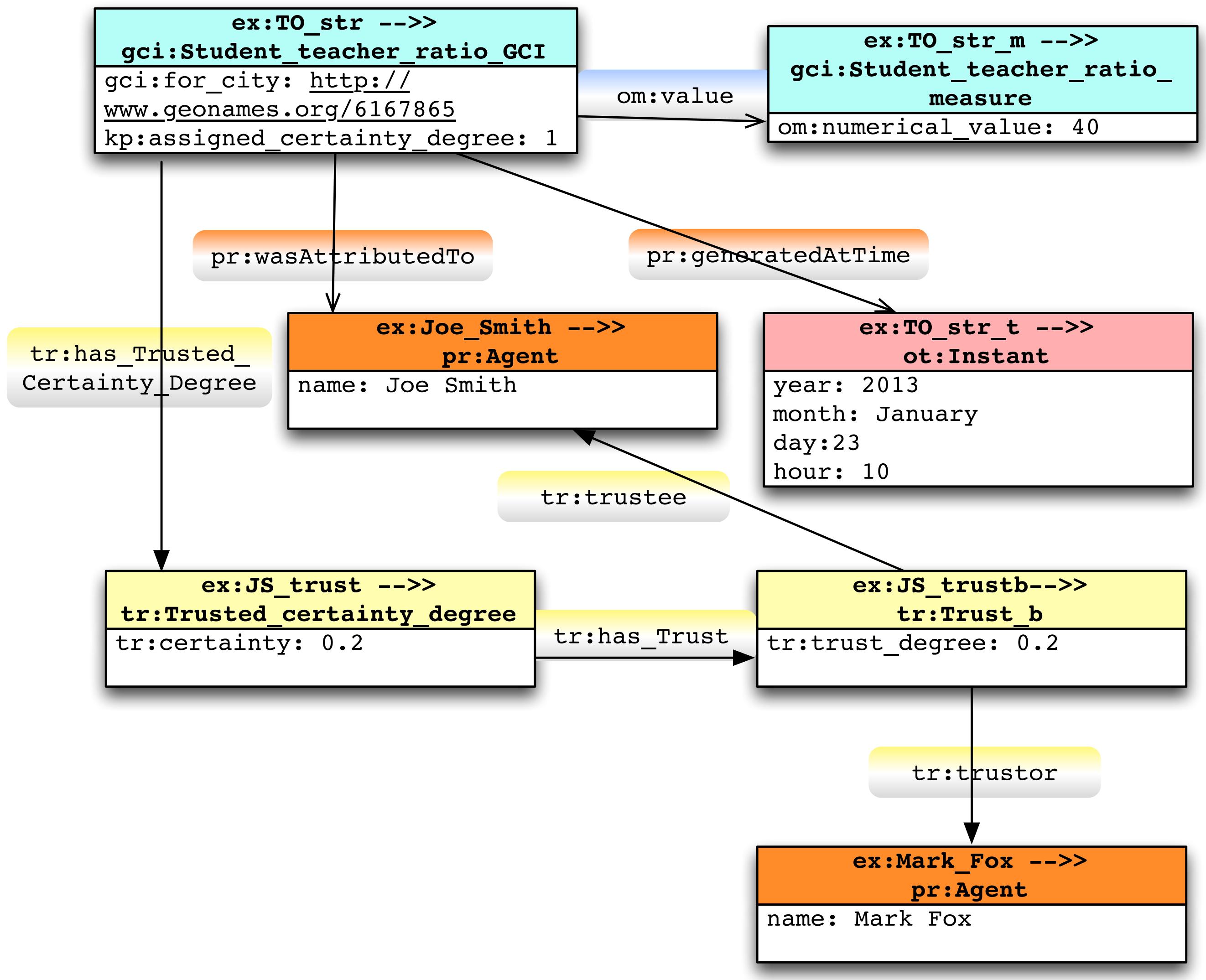

Figure 18

With this additional information, we know that although Joe Smith believes the STR to be correct, Mark Fox assigns a trusted certainty degree of 0.2 because he does not trust what Joe Smith believes.

To complete the example, we show how the quantity teacher population size is instantiated (Figure 19). `ex:TO_topsize` is an instance of `gci:Teacher_population_size` with the value specified by the `om:value` property linking it to a measure `ex:TO_tps_m`, which in turn specifies the teacher population size to be 1 `kilopcs`. The teacher population size is the cardinality of the population `ex:TO_tpop` linked to by `gci:cardinality`. This population is located in Toronto, and its membership is defined by the class of teachers who teach fulltime in primary school (this class definition has been abbreviated).

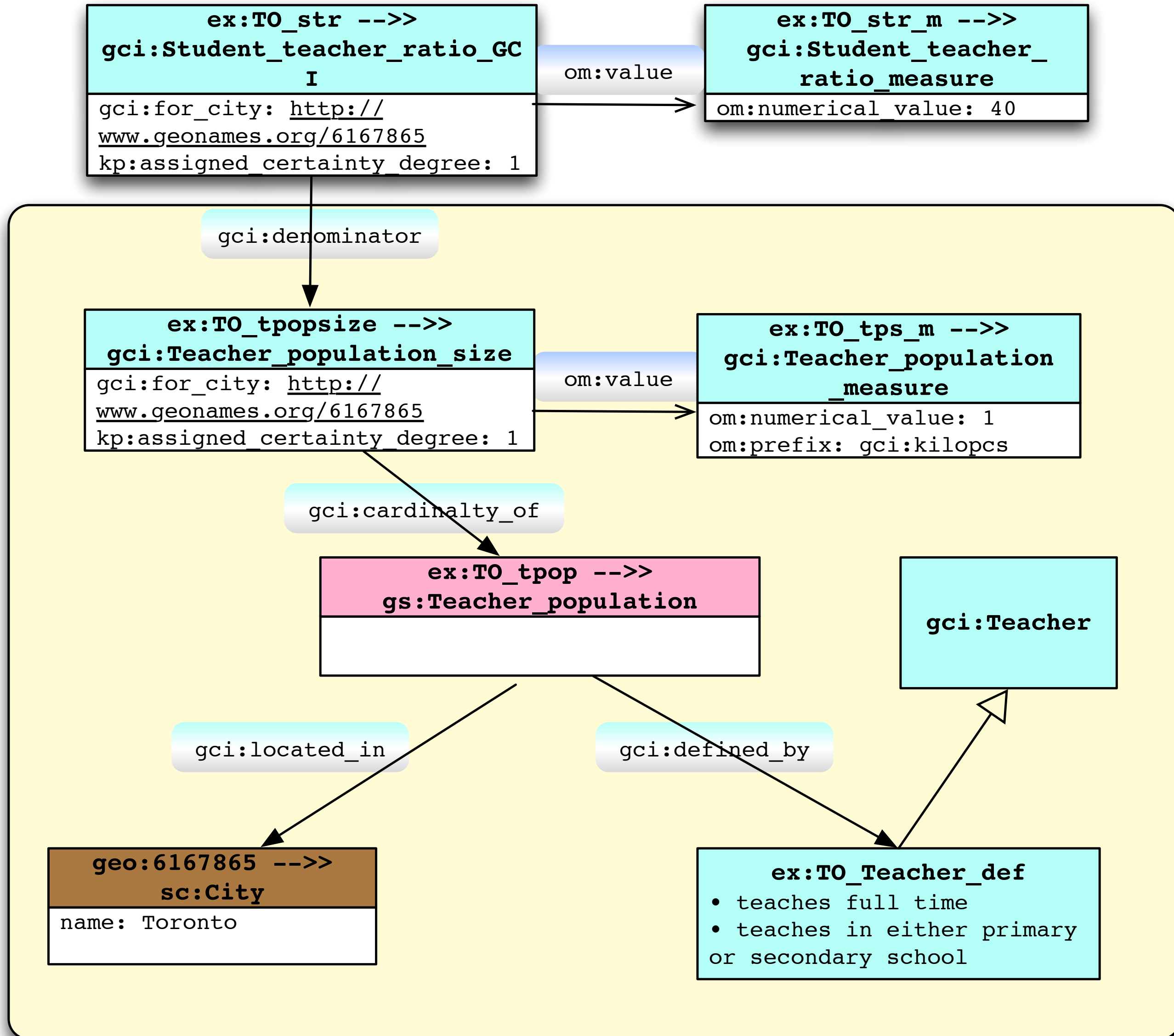

Figure 19

7. Evaluation

We approach the evaluation of Global City Indicator ontology from four perspectives:

1. Is the ontology Competent? In Grüninger & Fox (1995), the requirements of an ontology are defined by a set of competency questions. These questions define how the ontology is to be used by applications. In order for an ontology to be competent with respect to a set of questions, it must be able to correctly deduce answers assuming the model has been instantiated correctly.
2. Is the ontology Consistent? An OWL ontology is inconsistent if it contains a class that cannot possibly have any instances.
3. Is the ontology General? Are the classes general enough to represent other indicators and can it be easily extended where necessary?

4. Do we satisfy the aspects of a good indicator as identified in Hoornweg et al. (2007)?

7.1. Competency

The competence of an ontology is defined by a set of questions the ontology must be able to answer. These questions fall into three categories:

1. Questions that require a simple retrieval of the value of a property. For example, the city of a particular indicator.
2. Questions that require the following of one or more links (properties) in the network. For example, measurement consistency rule M1.
3. Questions that require some type of computation. For example, longitudinal or transversal analysis.

Regarding category 1, it is clear from the representation what data can be directly retrieved. Regarding category 2, the section on consistency defines a set of consistency competency questions in the form of rules. These rules have been implemented and tested, and are available for review (see Appendix). Regarding category 3, our future work will explore the types of analysis questions that need to be answered and any further extensions to the ontologies required.

7.2. Consistency

Each of the ontologies used herein were tested for consistency using RacerPro v2.0. The following table summarizes the ontologies tested and the result:

Ontology	URI	Status
Placenames	http://www.geonames.org/ontology/ontology_v3.1.rdf	Consistent
Measurement	http:// www.wurvoc.org/vocabularies/om-1.8/	Loading Error
Statistics	http://ontology.eil.utoronto.ca/govstat.owl	Consistent
Provenance	http://www.w3.org/ns/prov	Loading Error
Time	http://www.w3.org/2006/time	Consistent
Validity	http://ontology.eil.utoronto.ca/kp.owl	Consistent
Trust	http://ontology.eil.utoronto.ca/trust.owl	Consistent
Global City Indicator	http://ontology.eil.utoronto.ca/GCI-v1.owl	Unable to test

We were unable to test the consistency of the Global City Indicator ontology due to it's importation of two ontologies, Measurement and Provenance, which each generated an error when attempting to be loaded.

7.3. Generality

A major goal of the development of the Global City Indicator ontology for STR is to make it as general as possible so that it can be reused across the remaining indicators. Seven ontology modules were used in the STR example: Placename, Measurement, Statistics, Provenance, Time, Validity and Trust. Some of these modules were externally developed and used without change or extensions, like Time and Provenance, some were extended significantly, such as Measurement.

The following table lists, as rows, all of the Global City Indicators that have been defined and created by the Global City Indicators Facility (McCarney, 2012a; Global City Indicators Facility, 2010b). Columns 3 through 9 are the ontology modules discussed in this paper.

Columns 10 through 17 are additional modules that have to be developed in order to represent the remaining indicators. A cell that contains a light green colour signifies that the indicator can use the corresponding ontology module as is. A cell that contains a dark green colour signifies that the indicator requires the creation of the corresponding ontology module. 8 new modules have been identified:

- Government Finance
- Government Organization Structure
- Economics of Cities
- Census information
- Environment
- City Infrastructure
- Education, Technology and Innovation
- Heath, Safety and Emergency Response.

An examination of the table shows that all indicators would reuse the 7 modules described in this paper, while at most 3 new modules would have to be added for any one remaining indicator. The total reuse of the 7 modules introduced in this paper is a strong indication of the generality of the ontology.

Category	Indicator	3. Placename	4. Measurement	5. Statistics	6. Provenance	7. Time	8. Validity	9. Trust	10. Govt Finance	11. Govt Organization	12. Economics	13. Census	14. Environment	15. City Infrastructure	16. Edu, Tech, Innov	17. Health, Safety, Emer
Profile Indicators																
Government	Type of government (e.g. Local, Regional, County)															
	Gross Operating Budget (US$)															
	Gross Operating Budget per capita (US$)															
	Gross Capital Budget (US$)															
	Gross Capital Budget per capita (US$)															
Economy	Average household income [US$]															
	Annual inflation rate (avg. of last 5 years) [%]															
	Cost of living [US$]															
	Income distribution [GINI Coefficient]															
	Country's GDP [US$]															
	Country's GDP per capita [US$]															
	City Product per capita [US$]															
	City product as a % of country's GDP															
	Total employment															
	Employment % change based on the last 5 years															
	Number of businesses per 1000 population															
	Annual avg. unemployment rate															
	Commercial/Ind. assessment as % of total assess't															
People	Total population															
	Population density (per sq. kilometer)															
	% of country's population															
	% of population that are children (0-14)															

Category	Indicator	3. Placename	4. Measurement	5. Statistics	6. Provenance	7. Time	8. Validity	9. Trust	10. Govt Finance	11. Govt Organization	12. Economics	13. Census	14. Environment	15. City Infrastructure	16. Edu, Tech, Innov	17. Health, Safety, Emer
	% of population that are youth (15-24)															
	% of population that are adults (25-64)															
	% of population that are senior citizens (65+)															
	Male -Female ratio (# of males per 100 females)															
	Annual population change															
	Population Dependency Ratio															
	[%] of population that are new immigrants															
	% of pop. migrating from elsewhere in the country															
Housing	Total number of households															
	Total # occupied dwelling units (owned & rented)															
	Persons per unit															
	Dwelling density [per km2]															
Geography and Climate	Region															
	Climate type															
	Land Area (sq. kilometers)															
	% of non-residential area (sq. kilometers)															
	Average annual temperature [Celsius]															
	Average annual rainfall [mm]															
	Average annual snowfall [cm]															
City Services Indicators																
Energy	% of pop. with authorized electrical service															
	Total residential electrical use per capita [kW]															
	Total electrical use per capita [kW]															
	Electrical interruptions per customer [avg. #/yr]															
	Avg. length of electrical interruptions [hours															
Solid Waste	% of pop. with regular solid waste collection															
	% of solid waste that is recycled															
	% of solid waste disposed of in an incinerator															
	% of solid waste burned openly															
	% of solid waste disposed of in an open dump															
	% of solid waste disposed of in a sanitary landfill															
	% of solid waste disposed of by other means															
Waste Water	% of population served by wastewater collection															
	% of wastewater that has received no treatment															
	% of wastewater receiving primary treatment															
	% of wastewater receiving secondary treatment															
	% of wastewater receiving tertiary treatment															
Water	% of population with potable water supply service															
	Domestic water consumption per capita [litres/day]															
	% of pop. with sust. access to improved water source															
	% of population with sustainable access to an improved water source															
	Total water consumption per capita [litres/day]															
	% of water loss															
	Water service interruption per household [avg. hrs/yr]															
Transportation	Km of high capacity public transit per 100,000 pop.															

Category	Indicator	3. Placename	4. Measurement	5. Statistics	6. Provenance	7. Time	8. Validity	9. Trust	10. Govt Finance	11. Govt Organization	12. Economics	13. Census	14. Environment	15. City Infrastructure	16. Edu, Tech, Innov	17. Health, Safety, Emer
	Km of light passenger transit per 100,000 pop.															
	Number of personal automobiles per capita															
	Annual number of public transit trips per capita															
	Number of two-wheel motorized vehicles per capita															
	Number of non-stop commercial air destinations															
	Transportation fatalities per 100,000 population															
Education	Student/teacher ratio															
	% of students completing primary & secondary ed.															
	% of students completing primary education															
	% of students completing secondary education															
	% of school-aged population enrolled in schools															
	% of male population enrolled in schools															
	% of female population enrolled in schools															
Health	No. of in-patient hospital beds per 100,000 pop.															
	No. of physicians per 100,000 pop.															
	No. of nursing/midwifery personnel per 100,000 pop.															
	Average life expectancy															
	Under age five mortality per 1,000 live births															
Safety	No. of police officers per 100,000 population															
	No. of homicides per 100,000 population															
	Violent crime rate per 100,000 population															
Recreation	M^2 of public indoor recreation space per capita															
	M^2 of public outdoor recreation space per capita															
Fire & Emergency	No. of firefighters per 100,000 population															
	No. of fire related deaths per 100,000 population															
	Response time for fire department from initial call															
Finance	Debt service ratio															
	Tax collected as percentage of tax billed															
	Own-source revenue as a % of total revenues															
	Capital spending as a % of total expenditures															
Governance	% women employed in the city govt workforce															
Urban Planning	Jobs/Housing ratio															
	Areal size of informal settlements as % of city area															
	Green area (hectares) per 100,000 population															
Quality of Life Indicators																
Civic Engagement	Voter participation in last municipal election															
	Citizen's representation: # local officials elected to office per 100,000 population															
Culture	% of jobs in the cultural sector															
Economy	% of persons in full time employment															
Environment	PM10 concentration															
	Greenhouse gas emissions measured in tonnes per capita															
Shelter	% city population living in slums															
	% households that exist without registered legal titles															

Category	Indicator	3. Placename	4. Measurement	5. Statistics	6. Provenance	7. Time	8. Validity	9. Trust	10. Govt Finance	11. Govt Organization	12. Economics	13. Census	14. Environment	15. City Infrastructure	16. Edu, Tech, Innov	17. Health, Safety, Emer
	# homeless people per 100,000 population															
Social Equity	% city population living in poverty															
Technology & Innovation	No. of internet connections per 100,000 pop.															
	No. of new patents per 100,000 per year															
	% of jobs in the cultural sector															
	# higher education degrees per 100,000 pop.															
	# telephones (landlines & cell) per 100,000 pop.															
	# landline phone connections per 100,000 pop.															
	No. of cell phone connections per 100,000 pop.															

7.4. Aspects of a Good Indicator

In Section 3, we discussed the aspects of a good city indicator defined by Hoornweg et al. (2007). In this section we revisit these aspects from the perspective of what and how the Global City Indicator Ontology achieves these aspects.

- **Objective**: *clear, well defined, precise and unambiguous, simple to understand.*

  The ontology provides a clear, precise representation of an indicator that is grounded in more foundational ontologies such as measurement theory, statistics, etc., This reduces, if not removes in most cases, ambiguity in the interpretation of the indicator.

- **Measurable and replicable**: *easily quantifiable, systematically observable.*

  This aspect is not addressed by the ontology.
- **Auditable**: *valid, subject to third-party verification, quality controlled data (legitimacy across users).*

  With the inclusion of provenance, validity and trust information in the ontology, the ability to audit the information is greatly enhanced. Add to it the more detailed information on the populations from which the data is drawn from, the quality of the data can be further verified.

- **Statistically representative** *at the city level.*

  While this aspect is not addressed by the ontology, the detailed representation of the place and populations sampled enables the audit function determine whether the information is statistically representative.

- **Comparable/ Standardized** *longitudinally (over time) and transversally (across cities).*

  The incorporation of dynamic placenames, measurement, time, statistics, and provenance makes it possible to perform longitudinal and transversal analysis and to verify that the data being compared is consistent with each other.

- **Flexible:** *can accommodate continuous improvements to what is measured and how. Have a formal mechanism for all cities and interested parties to comment on.*

  The ontology can be easily extended to include other measures as demonstrated by the generality of the underlying modules (i.e., placenames, provenance, measurement, etc.).

- **Interrelated**: *indicators should be constructed in an interconnected fashion (social, environmental and economics).*

  The Semantic Web's network representation is fundamentally an integrated representation, and enables the integration of indicators and the information they are based on.

- **Consistent and sustainable over time**: *frequently presented and independent of external capacity and funding support.*

  An important aspect of publishing indicators and their supporting data on the Semantic Web is the universal access it provides and its availability over time.

8. Conclusions

Industrial Engineering and Management Science both share the view that you cannot manage what you do not measure. The ability to enhance the quality and efficiency of the operations and services of a city depends upon being able to measure them. The development of city metrics faces many challenges. The first challenge is the selection and definition of the metrics. The second challenge is the adoption and use of these metrics by a large number of cities. These first two challenges have been the focus of the Global City Indicator Facility for the last five years. The third challenge is to represent the indicators so that they can be published, linked, merged, mashed, and analyzed based on the principles of the Semantic Web. This work addresses this third challenge. It selects, merges and extends a number of ontologies in order to provide a semantic basis for the Global City Indicators, while at the same time making it possible to publish the data for use across the Semantic Web.

There are two directions that our current research is heading. The first direction is to complete the Global City Indicator Ontology to span all 100 City Indicators. This will require additional ontologies, such as census, environment, and city finances to be added. The second direction is to extend the competency of the ontology to support longitudinal and transversal analyses of city data.

9. Acknowledgements

I would like to thank Hajo Rijgersberg for his feedback on the use of the OM Ontology, and Patricia McCarney for her feedback on Global City Indicators.

10. References

Allen, J. F., and Ferguson, G., (1997), "Actions and events in interval temporal logic." In *Spatial and Temporal Reasoning*. O. Stock, ed., Kluwer, Dordrecht, Netherlands, pp. 205-245.

Belhajjame, K., Deus, H., Garijo, D., Klyne, G., Missier, P., Soiland-Reyes, S., and Zednik,S., (2012), "PROV Model Primer", http://www.w3.org/TR/prov-primer.

Campbell, A. E., and Shapiro, S. C., (1995). "Ontological Mediation: An Overview." *Proceedings of the IJCAI Workshop on Basic Ontological Issues in Knowledge Sharing*.

Berners-Lee, T., Hendler, J., and Lassila, O., (2001), "The Semantic Web", *Scientific American*, Vol. 284, No. 5, pp. 28-37.

Campbell, A.E., and Schapiro, S.C., (1995), "Ontologic Mediation: An Overview", *Proceedings of the IJCAI Workshop on Basic Ontological Issues in Knoweldge Sharing*, Menlo Park CA, USA: AAAI Press.

Capadisli, S., Auer, S., Ngonga Ngomo, A-C., (2013), "Linked SDMX Data: Path to high fidelity Statistical Linked Data for OECD, BFS, FAO and ECB", Submitted to *Semantic Web*, IOS Press.

Fox, M.S., and Huang, J., (2005a), "Knowledge Provenance in Enterprise Information", *International Journal of Production Research,* Vol. 43, No. 20., pp. 4471-4492. http://www.eil.utoronto.ca/km/papers/fox-ijpr05.pdf

Fox, M.S., and Huang, J., (2005b), "An Ontology for Static Knowledge Provenance", In *Knowledge Sharing in the Integrated Enterprise,* IFIP, Vol. 183/2005, pp. 203-213.

Global City Indicators Facility (2010a). "Overview Report of the Global City Indicators Facility", GCIF, University of Toronto.

Global City Indicators Facility (2010b). "Global City Indicators©: Definitions and Methodologies" September 2010, GCIF, University of Toronto.

Ghahremanloo, L., (2012), "An Integrated Knowledge Base for Sustainability Indicators", Australasian Computing Doctoral Consortium, RMIT Melbourne, http://www.cs.rmit.edu.au/acdc2012/. *Initial work on developing an ontology for sustainability indicators – starting with GRI and OECD*

Ghahremanloo, L., Thom, J.A., Magee, L., (2012), “An Ontology Derived from Heterogeneous Sustainability Indicator Set Documents”, *Proceedings of the Seventeenth Australasian Document Computing Symposium*, pp. 72-79.

Gruber, T. R., (1993), “Towards Principles for the Design of Ontologies used for Knowledge Sharing.” *Proceedings of the International Workshop on Formal Ontology*, Padova, Italy.

Grüninger, M., and Fox, M. S., (1995), “Methodology for the Design and Evaluation of Ontologies.” *Proceedings of the Workshop on Basic Ontological Issues in Knowledge Sharing*, IJCAI-95, Montreal, Canada.

Hartig, O., and Zhao, J., (2010), “Publishing and Consuming Provenance Metadata on the Web of Linked Data”, *Proceedings of the Third International Provenance and Annotation Workshop.*

Hausenblas, M., Halb, W., Raimond, Y., Feigenbaum, L., & Ayers, D. (2009). “Scovo: Using statistics on the web of data.” In *The Semantic Web: Research and Applications* (pp. 708-722). Springer Berlin Heidelberg.

Henson, C., Neuhaus, H., Sheth, A., Thirunarayan, K., and Buyya., R., (2009), “An ontological representation of time series observations on the semantic sensor web.” *Proceedings of the 1st International Workshop on the Semantic Sensor Web*.

Hobbs, J.R., and Pan, F., (2006), “Time Ontology in OWL”, http:www.w3.org/TR/owl-time/.

Hoornweg, D., Nunez, F., Freire, M., Palugyai, N., Herrera, E.W., and Villaveces, M., (2007), “City Indicators: Now to Nanjing”, World Bank Policy Research Working Paper 4114.

Hitzler, P., et al., (2012), “OWL 2 Web Ontology Language Primer (2nd Edition)”, http://www.w3.org/TR/owl-primer.

Huang, J., (2008). "Knowledge Provenance: An Approach to Modeling and Maintaining The Evolution and Validity of Knowledge", PhD Thesis, Dept. of Mechanical and Industrial Engineering, University of Toronto.

Huang, J., and Fox, M.S., (2004a). "Dynamic Knowledge Provenance", *Proceedings of Business Agents and Semantic Web Workshop*, pp. 372-387, National Research Council of Canada.

Huang, J., and Fox, M.S., (2004b), "Uncertainty in Knowledge Provenance", *Proceedings of the European Semantic Web Symposium*, Springer Lecture Notes in Computer Science.

Huang, J., and Fox, M.S, (2006), "An Ontology of Trust – Formal Semantics and Transitivity," *Proceedings of the International Conference on Electronic Commerce,* pp. 259-270. http://www.eil.utoronto.ca/km/papers/huang-ec06.pdf

Kauppinen, T., and Hyvönen, E., (2007), “Modeling and Reasoning About Changes in Ontology Time Series”, In *Ontologies: A Handbook of Principles, Concepts, and Applications to Information Systems*, Integrated Series in Information Systems, Vol. 14, pp. 319-338, Springer.

Kim, H.M., Sengupta, A., Fox, M.S., and Dalkilic, M., (2007), “A Mesurement Ontology Generalizable for Emerging Domain Applications on the Semantic Web”, *Journal of Database Management*, Vol. 18, No. 1.

Matuszek, C., et al., (2006), “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*.

McCarney, P. L., (2011) “Cities and climate change: The challenges for governance” Coordinating Lead Author with H. Blanco, J. Carmin and M. Colley, Ch 9 in *Climate Change and Cities: First Assessment Report of the Urban Climate Change Research Network*, Cambridge University Press.

McCarney, P. L., (2012a), “Global City Indicators (GCI)©: Standardization of City Services and Quality of Life Indicators, Methodologies and Definitions”, Draft submission to the International Organization for Standardization (ISO) for the standardization of the Global City Indicators and corresponding definitions and methodologies entitled: Global City Indicators (GCI): Standardization of City Services and Quality of Life Indicators, Methodologies and Definitions©.

McCarney, P. L., (2012b), “Cities and Prosperity”, GCIF Policy Snapshot Cities and Prosperity, Series No.1, Global City Indicators Facility, University of Toronto.

McCarney, P. L., (2012c), “City Indicators on Climate Change: Implications for Governance”, *Environment and Urbanization Journal*, 3:1-39, Sage.

McCarney, P. L., (2013), “Why Cities? Why Metrics? – Tracking Cities in the 21st Century” Global Cities Institute Working Papers, Vol. 1 No. 1, pp. 1-32, University of Toronto.

McCarney, P. L., and Stren, R.S., (2008), “Metropolitan Governance: Governing in a City of Cities in State of the World's Cities Report 2008”, (Nairobi: United Nations-HABITAT).

Moreau et al., (2010), “The Open Provenance Model Core Specification (v1.1)”, *Future* Generation Computer Systems. Also see openprovenance.org.

Nardi, D., and Brachman, R.J., (2002), “Introduction to Description Logics”, In *Description Logic Handbook*, edited by F. Baader, D. Calvanese, D.L. McGuinness, D. Nardi, P.F. Patel-Schneider, Cambridge University Press, pp. 5-44.

Niles, I. & Pease, A., (2001), “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.

Noy, N. F., Sintek, M., Decker, S., Crubézy, M., Fergerson, R. W., & Musen, M. A. (2001). Creating semantic web contents with protege-2000. *Intelligent Systems*, IEEE, Vol. 16, No. 2,, pp. 60-71.

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

Rijgersberg, H., Wigham, M., and Top, J.L., (2011), "How Semantics can Improve Engineering Processes: A Case of Units of Measure and Quantities", *Advanced Engineering Informatics*, Vol. 25, pp. 276-287.

Uceda-Sosa, R., Srivastava, B., and Schloss, B., (2012), "Building a Highly Consumable Semantic Model for Smarter Cities", *In Proceedings of the workshop on AI for an Intelligent Planet,* ACM.

Viljanen, L., (2005), "Towards an Ontology of Trust", *Proceedings of the Second International Conference on Trust, Privacy and Security in Digital Business*, pp. 175.184.

Westfall, M.S., and d Villa, V.A., (2001), *Urban Indicators for Managing Cities*, Asian Development Bank.

World Bank, (2008), "Global City Indicators Program Report: Preliminary Final Report", April 2008.

11. Appendix

The Global City Indicator Foundation ontology can be found in: http://ontology.eil.utoronto.ca/GCI-Foundation.owl.

The Global City Indicator Foundation ontology with the student/teacher ratio extension can be found in: http://ontology.eil.utoronto.ca/GCI-v1.owl.

The consistency axioms implemented in SWI-Prolog can be found in: http://ontology.eil.utoronto.ca/gci-axioms.pl along with supporting axioms for time and OWL.

The example used in this paper is implemented in SWI-Prolog and can be found in: http://ontology.eil.utoronto.ca/gci-example.pl.