ARTICLE IN PRESS

Web Semantics: Science, Services and Agents on the World Wide Web ▮ (▮▮▮▮) ▮▮▮–▮▮▮

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 xxxx

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.

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/

Please cite this article in press as: M.S. Fox, The semantics of populations: A city indicator perspective, Web Semantics: Science, Services and Agents on the World Wide Web (2018), https://doi.org/10.1016/j.websem.2018.01.001.

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

Please cite this article in press as: M.S. Fox, The semantics of populations: A city indicator perspective, Web Semantics: Science, Services and Agents on the World Wide Web (2018), https://doi.org/10.1016/j.websem.2018.01.001.