On the Completeness of Open City Data for Measuring City Indicators

Mark S. Fox
Mechanical & Industrial Engineering Dept.
University of Toronto
Toronto, Canada
msf@eil.utoronto.ca

Christopher J. Pettit
City Futures, Faculty of Built Environment
University of New South Wales
Sydney, Australia
c.pettit@unsw.edu.au

Abstract—In an era of urbanization there is an increasing need to benchmark the performance of cities and do this against a robust set of measures that can be repeated and replicated over time. Open city data initiatives are on the increase and with the release of ISO37120 city indicators it is timely to critically appraise the synergies between open data and city indicators. Hence, in this paper we present a novel measure for assessing the completeness of open city data in the context of measuring city indicators. We apply this measure known as the “City Indicator Data Openness Measure” (CIDOM) in the context of two global cities, Toronto and Melbourne. Specifically we do this in the context of the ISO 37120 educational indicator 6.4: “Primary Education Student/Teacher Ratio”. This exercise has identified that there exists fundamental barriers in realizing the vision of open data if it is to support the reporting and analysis of city indicators. It raises further issues about sourcing data and also provides another argument towards developing a semantic approach for city indicators.

Keywords—City Indicators; Open Data; Ontologies; Quality of Life; ISO 37120; Semantic Web

I. INTRODUCTION

City indicators measure city services and functions, such as Education, Health, Safety, Transportation, Water and Energy. “There are thousands of different sets of city (or urban) indicators and hundreds of agencies compiling and reviewing them and using them for a myriad of purposes. However, these indicators are usually not standardized, consistent or comparable (over time or across cities) ...” (Horweeg et al., 2007). In response to this problem, a plethora of initiatives have been undertaken to identify and define global city indicators, including for the OECD Better Life Survey, Hassell’s LESS (Local-area Envisioning and Sustainability-scoring System) indicators, and the IBM Smart Cities Assessment Tool. Quality of Life is becoming a significant theme for a number of such indicators and much effort has been focused on measuring and mapping aspects of life in urban areas or the quality of urban life (Marans & Stimson 2013). However, much more limited attention has been placed on the underlying data and data structures which underpin these indicators.

Concurrent with the emergence of global city indicators, the Open Data movement has taken root. It is part of the broader Open Government movement where the belief is that making data publicly available will lead to more effective public oversight. Yet there are other benefits than simply oversight. With more “eyes on the data”, waste and inefficiencies can be detected, crowd-based solutions suggested, and the crowd harnessed to implement some of the solutions to the effective organization of data. But there are many problems with Open Data. Cities, and those agencies such as state government which record and publish data on cities, are publishing vast amounts of data, but the formats can range from spreadsheets, to documents in PDF format, to XML, to Semantic Web Resource Description Framework (RDF) formats. Secondly, a number of Cities are taking the approach of publishing the ‘low hanging fruit’, releasing those datasets that are easiest to release but not necessarily of high utility and therefore unsuitable to produce indicators. Thirdly, and more importantly, no two cities and often departments within a city use the same data models. A consequence of this “babel of open data” is that whilst there are a lot of datasets it is difficult to combine, analyze and compare city data.

Concurrent with the growth of Global City Indicators and Open Data, has been the development of the Semantic Web [1] and Linked Data [11]. The Semantic Web provides a solution to the “babel of open data”. From a format perspective, it provides a standard in the form of RDF triples [14], and from a data model perspective it provides a means by which vocabularies and ontologies can be defined, shared and reused across the web [5].

The convergence of global standards for city indicators, open city data and the semantic web provides a unique opportunity for cities and researchers to analyse and compare city performance. But there is another issue that needs to be addressed. If cities, and agencies with published data on cities, openly publish and link their indicators (using semantic web standards), can we truly understand and believe these indicators without the supporting data also being openly published? Sadly, each indicator relies on layers of supporting data, most of which is not published. Without the supporting data being publically available, we will not be able to determine whether the city’s version of an indicator conforms to the definition of the indicator, nor to what degree a city’s indicator reflects the underlying supporting data.
The goal of this paper is to provide a method, which we call the “City Indicator Data Openness Measure” (CIDOM), that can measure the “degree of openness” of a city indicator. By “degree of openness” we mean the extent to which the supporting data upon which a particular city indicator is based, is publically available. We acknowledge the open data movement is one that is gaining global momentum as a number of cities are continually making more data available. Therefore, the CIDOM is one that will need to be revisited over time to track the performance of a city in opening data that is suitable for measuring its performance, whether it be against a set of sustainability, livability, quality of life, well-being or other set of indicators.

II. GLOBAL 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.” [12]. 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 [10][2].

In light of previous efforts to define city indicators, [12], 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.
- Potential 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.”

Hornweg et al.’s analysis of the state of city indicators was the catalyst for the creation in 2010 of the Global City Indicators Facility (GCIF). GCIF’s goal was to work with cities in identifying a common set of indicators and establishing standardized definitions and methodologies that can be consistently applied globally [9][15]. The outcome of this effort is the international standard ISO 37120 “Sustainable development of communities — Indicators for city services and quality of life”.

There are a number of metrics being used to evaluate a city’s liveability. The Monocle annual quality of life survey purports to identify the “most liveable cities” using the following criteria: “safety/crime, international connectivity, climate/sunshine, quality of architecture, public transportation, tolerance, environmental issues and access to nature, urban design, business conditions, pro-active policy developments and medical care.” No further information is available as to how these are define nor measured.

The Mercer Quality of Life Survey ranks 221 cities on quality of life. This survey is conducted to help multinational employers pay their global work force fairly when they are on international assignments. More than 460 cities are evaluated against 39 factors and are benchmarked against New York, which is assigned a base score of 100. The quality of living reports are available for each city but need to be purchased. Subsequently there is not open data available relevant to this indicator product.

The Economist Intelligence Unit (http://www.eiu.com/home.aspx) produced a number of quality of life and liveability indicators combined these indicators are used to rank 140 international cities. The scores are calculated using approximately 30 qualitative and quantitative factors across five broad categories: stability, healthcare, culture and environment, education and infrastructure. The scores are compiled and weighted with a final score 1-100. Interestingly the ranking reports an overall decline of liveability across the world by 0.7% between 2009 and 2014 with the stability and safety factor driving this decline.
More recently, there have appeared repositories for city indicators. The Global Cities Institute (GCI), which led the development of ISO 37120, maintains a closed repository of ISO 37120 data for its member cities. The World Health Organization maintains the Global Health Observatory Data Repository, which contains a large number of health-related indicators but not at the city level.

### III. Openness Model

The goal of CIDOM is to signify the extent to which a city openly publishes indicator data. By extent we mean “how much” and in “what format” the supporting data used to derive the indicator is published in addition to the indicator value. Determining the supporting data that needs to be published is analogous to the process of verifying the process by which indicators are computed. This process has to verify that the definition of the indicator was satisfied in computing the final number. How can this be done? We begin with an indicator’s definition.

Consider the ISO 37120 educational indicator 6.4: “Primary Education 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). 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 administrators or other non-teaching staff. Kindergarten or pre-school 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.”

On the surface the definition is simple, being the ratio of the number of students to the number of teachers, but the definition reveals a lot more detail:

- The indicator is the ratio of two numbers whose units and scale must be the same (measurement theory).
- The number of students (numerator) and teachers (denominator) are cardinal measures of two different sets (measurement theory).
- The sets are based on a population defined within a geographic area (location/placename).

- The population being sampled is defined by a definition of a student or teacher (description logic).
- A student is defined as a full time student in primary and/or secondary school (description logic).
- Full time is defined as spending at least 1500 hours a year in school (description logic).
- Grade levels and Primary school has to be defined by each city.
- The definition of private educational facilities needs to be defined by each city, as they facilities are to be excluded.

The indicator “Student/Teacher Ratio” is the root of a dependency tree where the supporting definitions and data branch out below it. The tree is heterogeneous in that its nodes span various types of representations including analytical, statistical, spatial and logical. In addition the tree must represent meta-information such as the processes used to derive the data, its validity and trust.

CIDOM is based on the nodes of the dependency tree that have been openly published. The more nodes published in a more open format, the higher the CIDOM score. But how do we construct a dependency tree for an indicator? Fox [5][7] defines an ontology for representing ISO 37120 indicators. The following depicts a generic graph for representing indicators that are a ratio of two populations:

A ratio indicator has a unit of measure defined to be a ‘Population Ratio Unit’ that specifies that the indicator is the ratio of the sizes (cardinalities) of two populations. One population size is the numerator and the other the denominator. A ‘Population Size’ is defined as the cardinality of a ‘Population’, and ‘Population’ is defined by a ‘City’ that the population is located in, and by a description of a ‘Person’ within the ‘City’. For example, the ‘Person’ could be ‘Female Student’. Hence the ‘Population Size’ could be the number of ‘Female Student’s in a particular ‘City’.
If we can represent an indicator using this or another ontology, we can abstract the definition into a dependency graph. The following is an example of a dependency graph for the indicator 6.4, primary education student/teacher ratio. It reduces the definition found in Fox [7], (section 7.2) into a graph where the arcs represent simple computations such as aggregating numbers across multiple sources and selecting (i.e., filtering) out data. The yellow boxes represent the actual data while the blue boxes represent definitions provided by the ISO 37120 definition and are used to select that subset of data that satisfies the definition.

In the diagram above, the student/teacher ratio is dependent upon the number of students and teachers at public primary schools. The number of students is an aggregation of the number of students in each public primary school. The number of students in each school depends on a selection of a subset of students in public primary schools, based on the ISO 37120 6.4 definition of a student, including whether they are full or part time, where they are enrolled, etc. Public primary schools are selected from the set of all primary schools that in turn is selected from the set of all schools in the city.

As we can see, the dependency graph provides a representation of all of the supporting data that is to be used to derive the indicator. We can then determine for each node in the graph whether the data is openly available, whether it is on the city’s web site, or the organization that normally gathers such data and is the authoritative source (e.g., a school board).

In addition to determining whether the data corresponding to a node in the dependency graph is openly available, we also want to determine the format with which the node’s data is published. The following defines a sequence of successively more open publishing formats:

1. Published reports containing data, on a city web site. For example, PDF report.
2. Publish data openly on a city web site, can be any format such as csv, json, shp, geo json.
3. Data is made available as a service, rather than as a download.
4. Data contains metadata which conforms to a standard e.g Dublin Core, ISO19115.
5. Publish open data that conforms to linked data standards for the semantic web, namely RDF triples, but uses internal vocabulary.
6. Publish open data using commonly accepted vocabulary for city indicator data.
7. Publish open data using commonly accepted ontology for city indicator data.

Consequently, each node in the dependency graph would be annotated based on whether the data is openly available, and the format in which it is openly published. Based on the annotations, we can define metrics that represent the degree of openness of city indicator data. We define three versions of CIDOM as follows:

CIDOM-1: is the percentage of nodes in the dependency graph that are openly published by the city and related government agencies. It provides a general measure of the totality of the data published for an indicator.

CIDOM-2: is the number of levels of the dependency graph that are openly published by the city and related government agencies. More precisely:

- Level 0: Indicator value is not openly published.
- Level 1: Only the indicator value is openly published, e.g., on the student teacher ratio value is published without any supporting data.
- Level 2: All of the data that the indicator directly depends on is openly published, e.g., numerator and denominator of the student teacher ratio are published.
- Level 3: All of the data directly supporting level 2 is published.
- Level n: All of the data directly supporting level n-1 is published.

It provides a different view, namely the depth to which supporting data is completely published.

CIDOM-3: is the average of the format levels for each node in the dependency graph (i.e., pdf, json, service, etc.) at which the data is openly published by the city and related government agencies. It determines the dominating format used to publish.

The computation of CIDOM is dependent on the dependency graph. The dependency graph is dependent upon an indicator's definition. In order for CIDOM to be consistently applied, we have to have an agreement of the representation of the indicator, and based on it a representation of the dependency graph. Different dependency graphs will most likely vary below level 3, but be consistent from level 3 upward. Secondly, below level 3 there will be variations in how a city gathers and aggregates its data.

IV. APPLICATION

The next step is to apply the CIDOM to actual city data. Two cities have been selected on the basis they are undertaking to implement ISO 37120 and have policies and supporting portals for the release of open data. The two case
studies focus on the implementation of ISO 37120 6.4 “Primary Education Student/Teaching Ratio” as defined previously in this paper.

A. Melbourne Case Study

Define Metropolitan Melbourne has a population of approximately 4.14 million people and comprises 31 Local Government Areas (LGAs). In terms of liveability Melbourne has been rated as the most liveable city by the Economists Intelligence Unit’s survey for the past 4 years, with a current overall score of 97.5 out of 100. With the respect to schools there are approximately 1,500 schools in Metropolitan Melbourne.

In terms of open data The State Government of Victoria developed a policy for open release in 2012 making datasets freely available to the public the State’s default position. The Data.vic.gov.au portal (https://www.data.vic.gov.au) as of 19th Feb 2015 had 3,691 datasets with 76 available on the topic of education, contributed to by the Department of Education and Early Childhood Development. However, none of these include the teacher student ratios for primary schools that can be used to calculate this indicator for Metropolitan Melbourne. The Department of Education and Early Childhood Development does include a summary statistics for overall average class size and student number for Victoria and compare this against the OECD statistics but this is for secondary schools. The OECD also have data on the average class size for primary schools on a grade basis but this is for Victoria not for Metropolitan Melbourne. However, they state that the average class size is not the same as the student/teacher ratio as they did not include extra teaching support.

The City of Melbourne has been actively involved in city indicator projects they also have an open data portal https://data.melbourne.vic.gov.au/. However, there is limited available data on education or schools. Also The City of Melbourne is only 1 of the 31 LGAs comprising Metropolitan Melbourne so when trying to recreate the ISO 37120 core indicator 6.4 for Primary education student teacher ratio it does not have the appropriate geographical coverage.

At the national level the Australian Bureau of Statistics (ABS) compiles statistics on students, schools and staff through the National Schools Statistics Collection. This data is open and available to download. However, it only exists at the State and Territory level of geography so it not useful for creating the ISO 37120 6.4 indicator at the Metropolitan level. The Australian Urban Research Infrastructure Network (AURIN) portal, which is password protected and available for use by researchers through their university credentials and also to government offices, contains information on all the school locations, id and postcodes for the country but does not contain the attribute information to recreate the ISO 37120 indicator for Melbourne. There is also the My School website (http://www.myschool.edu.au/) which has detailed information on each School in Australia include the attributes required to recreate the ISO 37120 6.4 indicator. However, this data is available from the website as a report or map and there is not mechanism to aggregate the data to LGA or Metropolitan area. However, requests for data aggregations can be made through the Australian Curriculum, Assessment and Reporting Authority (ACARA). Requests typically take a month for request. Such a request has been made to obtain the ISO37120 6.4. indicator so that the indicator can be recreated for Metropolitan Melbourne. In summary there are many open data sources for data associated with Metropolitan Melbourne but after significant investigating all of these options Indicator 6.4 cannot be recreated from available open data unless a formal request is made through ACARA.

Based on the information available, the rating for Melbourne with respect to ISO 37120 indicator 6.4, is as follows:

- CIDOM-1: 0% as none of the 8 nodes in the dependency graph is openly available.
- CIDOM-2: is Level 0 as the indicator value is not available.
- CIDOM-3: is Level 0 as the indicator value is not available.

B. Toronto Case Study

The Greater Toronto Area (GTA) has a population of 5.5 Million. The City of Toronto has a population of 2.79 million. It is ranked as one of the most multicultural cities in the world with over 50% of the residents having been born outside of Canada, and over 140 languages and dialects spoken. In 2014 Greater Toronto has been ranked first in the following reports: Focus on Tax, Resilient Cities, Youthful Cities, and Intelligent Community Forum. It was ranked in the top 10 in reports such as: Scorecard on Prosperity, Transit score, Liveability Rankings, Cities of Opportunity, World’s Most Influential Cities and Global Talent Survey.

The City of Toronto approved its Open Data Policy in November 2011, resulting in the establishment of an Open Data Portal for the city. The open data site contains 179 datasets in various formats spanning Excel, PDF, ESRI, CSV, XML, and JSON, and divided into 15 categories such as Community Services, Culture and Tourism, Health, and Finance.

A search of the City of Toronto open data web site returned the following data sets:

- ISO 37120 Results for 2013. This document is a PDF of a slide presentation that provides the values for all of the ISO 37120 indicators and profile indicators. Information is not available on the source of the document nor the source of the data it is based on.
- School Locations – All Types. This dataset contains information on the type of school (public/private), school board which includes whether it is regular or catholic, language (e.g., French, English), name, address, longitude and latitude. It is in a proprietary format.

Information on school enrollment was found at www.city-data.com but there was no information on the source, accuracy or how recent the information is. The Toronto District School Board was contacted to find if there is additional information available. Discussions are still underway.
Based on the information available, the rating for Toronto with respect to ISO 37120 indicator 6.4, is as follows:

- CIDOM-1: 37% as 3 of 8 nodes in the dependency graph is openly available.
- CIDOM-2: is Level 2 as the indicator value along with the numerator and denominator are available.
- CIDOM-3: is Level 1 as the indicator value is available as a PDF only.

V. CONCLUSION

The Research Data Alliance Interest Group on urban quality of life indicators has been set up to contribute towards the challenge of setting up Quality of Urban Life (QuOL) indicators that can benchmark and enable comparative analysis between cities. In order to achieve this, the underlying data to create these indicators needs to be open and semantically enabled. This initial exercise by the authors reveals there is a lot of work to be done to realize the vision of meaningful open data that can support quality of life city indicator benchmarking and reporting.

The paper outlines a model for evaluating a city’s indicator open data (CIDOM) and also the rating of the release of the underlying data format. The authors provide examples of CIDOM using the Melbourne and Toronto Cities, which are both implementing ISO37120 and have open data policies and portals.

The authors have found that it is difficult to source the ISO37120 6.4 City Indicator Data for either city, whether it be open or not. For example the City of Melbourne is not the custodian of the underlying data to develop education, it is fact the Victorian Department of Education who is the custodian of such information, who in-turn provides this information to the Australian Curriculum, Assessment and Reporting Authority (ACARA), who are then responsible for handling such release of information and associated data requests. The same issue exists for the City of Toronto where it is the School District Boards that are the custodian of such information. However, for Toronto the 2013 ISO 37120 indicator values were available in a PDF document, but neither the city nor the school boards openly published the supporting data. This exercise undertaken by the authors has identified that there exists fundamental barriers in realizing the vision of open data if it is to support the reporting on city indicators standards such as ISO37120. It raises further issues about sourcing data and also provides another argument towards developing a semantic approach for city indicators.

In conclusion, the goal of CIDOM is to measure to degree to which cities openly publish both their indicator values and most importantly the supporting data used to derive them in a format that is computationally accessible. As such, it is meant to shed light on those barriers to openness of existing City Open Data efforts and provide motivation to move down the path of achieving more effective data release as part of the open data movement which is an important cornerstone to any smart city initiative.

ACKNOWLEDGMENT

This research was supported in part by the Natural Sciences and Engineering Research Council of Canada.

REFERENCES