Available online at www.sciencedirect.com

ScienceDirect

Procedia Computer Science 108C (2017) 2059–2068

Procedia Computer Science

International Conference on Computational Science, ICCS 2017, 12-14 June 2017, Zurich, Switzerland

An ontological approach to dynamic fine-grained Urban Indicators

Salvatore F. Pileggi and Jane Hunter

School of Information Technology and Electrical Engineering (ITEE)
The University of Queensland
Brisbane, QLD, Australia
s.pileggi@uq.edu.au , jane@itee.uq.edu.au

Abstract

Urban indicators provide a unique multi-disciplinary data framework which social scientists, planners and policy makers employ to understand and analyze the complex dynamics of metropolitan regions. Indicators provide an independent, quantitative measure or benchmark of an aspect of an urban environment, by combining different metrics for a given region. While the current approach to urban indicators involves the systematic accurate collection of the raw data required to produce reliable indicators and the standardization of well-known commonly accepted or widely adopted indicators, the next generation of indicators is expected to support a more dynamic, customizable, fine-grained approach to indicators, via a context of interoperability and linked open data. Within this paper, we address these emerging requirements through an ontological approach aimed at (i) establishing interoperability among heterogeneous data sets, (ii) expressing the high-level semantics of the indicators, (iii) supporting indicator adaptability and dynamic composition for specific applications and (iv) representing properly the uncertainties of the resulting ecosystem.

Peer-review under responsibility of the scientific committee of the International Conference on Computational Science

Keywords: Ontology Design, Semantic Technology, Urban Indicators, Uncertain Knowledge

1 Introduction

The urban indicator, as a practical, objective, comparable measure of urban regions, is definitely not a novel concept in itself. In fact, over the past ten years, urban indicators have increasingly gained popularity across a broad range of different domains and disciplines, in which they are employed to assess different aspects of urban regions (from social and economic development to sustainability) and to achieve different goals.

Many of the major challenges facing human development in the future rely on complex indicators; the quality of life is often evaluated in terms of indicators; the design of governance strategies is based on that sensitive trade-off between the optimization of a set of target

Peer-review under responsibility of the scientific committee of the International Conference on Computational Science
10.1016/j.procs.2017.05.003

indicators and economic factors; understanding social phenomena (e.g. crime) requires indicators; both expert-based and citizen-based recommendations come from the observation of key regional indicators; the evolution of our cities (e.g. urban planning) and of its services (e.g. public transport systems) depend on the capability of defining, building, analyzing and interpreting indicators. In practice, urban indicators play a significant role in multi-disciplinary contexts due to their capability of synthesizing metrics to facilitate knowledge building and monitor the impact of synergistic strategies.

The application-level perception of indicators defines a simple key rule: better indicators imply better analysis capabilities. In technical terms however, the use of a certain indicator depends first of all on the availability of the raw data required to calculate or estimate that indicator. Moreover, the quality of a given indicator is directly proportional to the accuracy of the raw data that underpin it. The current approach to urban indicators involves the systematic and accurate collection of the raw data required to produce reliable indicators and the standardization of well-known, commonly accepted and widely adopted indicators. In addition, mainly because of the intrinsic difficulty and cost to build granular data sets, urban indicators are very often considered at the city level only and indeed they are referred to as global indicators because of their application to comparing big geographic regions globally. Unfortunately, most potentialities of urban indicators reside in the fine-grained analysis concerning smaller areas (e.g., suburbs within cities) which are very rarely addressed in the current platforms.

Within this paper, we are considering a novel scenario in which users are progressively leaving the classic static context characterizing urban indicators to move towards a more flexible and open model where indicators can be dynamically defined and changed according to the available data and the scope of the target analysis. Thus, we are considering a technological context, including the data ecosystem and the software environment, which is intrinsically more complex than the common spatial data frameworks. To enable extensibility, flexibility and interoperability, we propose an ontological approach to assure an enriched semantic data structure supporting different platforms and applications in the domain.

Structure of the paper The next section 2 provides an overview of previous, related work in this area; section 3 proposes a detailed description of the ontological framework that we have developed; section 4 discusses the uncertainties and some current related limitations within the current ontological model; finally, section 5 provides a summary of the achievements, future directions and some concluding remarks.

2 Related Work

Over the past five years, semantic technologies have been extensively used in the context of city data to address different issues, including the aggregation, the management and the processing of data, as well as the integration of data from across disparate domains through ontological approaches.

For instance, [1] focuses mainly on the collection of open city data which can be integrated, enriched and eventually republished as linked data. More generally, the problem of exchanging and sharing statistical data and metadata among organizations according to a semantic approach has produced some notable solutions (e.g. the RDF Data Cube [2]). Adding a spatial dimension to the Web of Data [3] represents a consolidated approach, as well as standard vocabularies may be extented to address specific requiremets [4]. Whilst, the QuerioCity platform employs Semantic Web technologies to catalog, index and query heterogeneous information generated by cities [5].

Ontologies are being extensively used in different contexts to achieve different goals different from data aggregation. For example, upper-level ontologies for global city indicators [6] model the city data domain from a theoretical perspective: as upper-level vocabularies, they facilitate and push semantic interoperability among indicators; those vocabularies are playing a key role in standardization. Although the general concept of Semantic GIS has been extensively addressed, as has the semantic representation of space [7] and complex reasoning over it, many potential applications of the semantic technology to urban data models appear to remain largely unexplored.

The work described in this paper, reflects a more pragmatic approach which aims to develop platforms and tools that assume a dynamically evolving concept of indicator. Indeed, in our understanding, an indicator is not something that is pre-defined and pre-calculated. On the contrary, it can be defined or re-defined along the way in response to different needs, requirements and scopes. Furthermore, we are considering a fine-grained model of indicator which is designed to support accurate and detailed analysis. In addition, we assume that the indicators are derived from distributed, pervasive and growing data sources on the cloud, that are programmatically accessible via APIs. We are not aiming at iterative extensions to the current platforms and tools dealing with urban indicators; rather, we are targeting a new generation of urban indicator services.

3 Ontological Framework

Semantic technology can contribute to the development of rich and effective environments for managing, processing and sharing urban data sets and indicators, at multiple levels to achieve different goals. Our approach promotes integrated solutions whereby the ontological representation of the information provides a semantic layer, namely enriched data structures, to support the development of novel platforms in open contexts.

The ontological framework objective of this paper has been designed to cater for a wide range of applications. From a knowledge representation perspective, it can be characterized by the following key principles:

- **Conceptualization and semantic support.** The ontological framework addresses indicators from a practical perspective. This doesn't assume a model where the conceptualization is not relevant or not playing a primary role. It is, on the contrary, is still a primary concern. Indeed our high-level profiles for indicators support multiple points of view including the structure, the description in a natural language, the provenance, categorization and classification, all based on open vocabularies. In other words, we are not just saying *what* an indicator is, but also *how it is composed*, *how it is calculated*, *for which purpose* in a given context, *how it is related* to other indicators or entities. Therefore, our approach is not alternative but rather complementary with the emerging open domain vocabularies (e.g.[6])

- **Functional support to software platforms.** Different applications of urban indicators may present very different requirements. But we anticipate that most systems will converge on a set of key functionalities. The use cases that we are currently considering are increasingly demanding the dynamic composition and run-time computation of indicators from distributed, raw, timely datasets. Our ontology is designed to provide a solid support for such functionalities in the context of dynamic interoperability.

- **Focus on data interoperability.** Indicators are the result of some logical or mathematical processing of raw data. Therefore, dealing with dynamic indicators means addressing dynamic data aggregation in federated distributed environments. Because different architectures adopt different perspectives to data and data repositories, a flexible mapping model is one of the critical design issues: it should either allow datasets to remain in their legacy repositories or enable them to be imported in a semantic format to become part of the system. The former class of solutions is, a priori, more suitable to address scalability and consistency requirements but implies a more sophisticated aggregation mechanism and, of course, a more complex semantic support; the latter is a much simpler solution in terms of architecture design and semantic support. But the latter approach requires specific mapping mechanisms to ensure data consistency and, at least in theory, additional computational resources to manage large sets of semantically enriched data.

- **Agnostic approach to the geographic space model.** The geographic space model is a critical asset for urban indicators as they are the result of computations on spatial data. However, there are a lot of models for the geographic space currently in use and the unification of those models within a unique universally accepted model is unrealistic. It is also important to take into account that specific studies could require ad-hoc models for the geographic space. For these reasons, our approach does not assume a well-defined model for the geographic space but rather provides a set of concepts and properties to represent different models and standards depending on the final user needs and on the available data. The relations among different models will also be significant in applications that integrate data related to more than one data model; however such integration (or spatial correlation) is also likely to introduce uncertainties (see section 4).

In the next sub-sections we first provide a conceptualized overview of the framework; than we move towards the description of its implementation as an OWL Ontology; finally we will discuss the ontological support provided in the context of a reference architecture.

3.1 Conceptual Framework

The concept map of fig. 1 represents a simplified view of our ecosystem in which the indicator concept co-exists with the methods adopted to produce the indicator, as well as with the data sets underpinning it. Therefore, within the proposed model, an indicator i is an extensible tuple composed of the following sets of concepts:

(a) A *profile* (P_i) composed of a number of *characterizations* and *user-level annotations*. While the characterizations define classifications on the base of open vocabularies and the association of i with virtual structures (e.g. layers), the user-level descriptors specify user-level descriptions and metadata, commonly expressed in a natural language.

(b) The *structure* (S_i) provides a set of indicator types. We distinguish among the common *application-level indicator*, *supporting indicator* and *structured indicators*. Unlike the application-level indicator, a supporting indicator is not normally used directly at an application level but it is defined to support the definition or the computation of other indicators. Structured indicators include the *composite indicator* and the *subindicator*: as the names suggest, the composite indicator is composed of sub-indicators. The type of an indicator can be specified according to many different perspective. By relating the indicators at a structural level, S_i facilitates knowledge inference, as well as the implementation of high-level mechanisms (e.g browsing, searching, discovering).

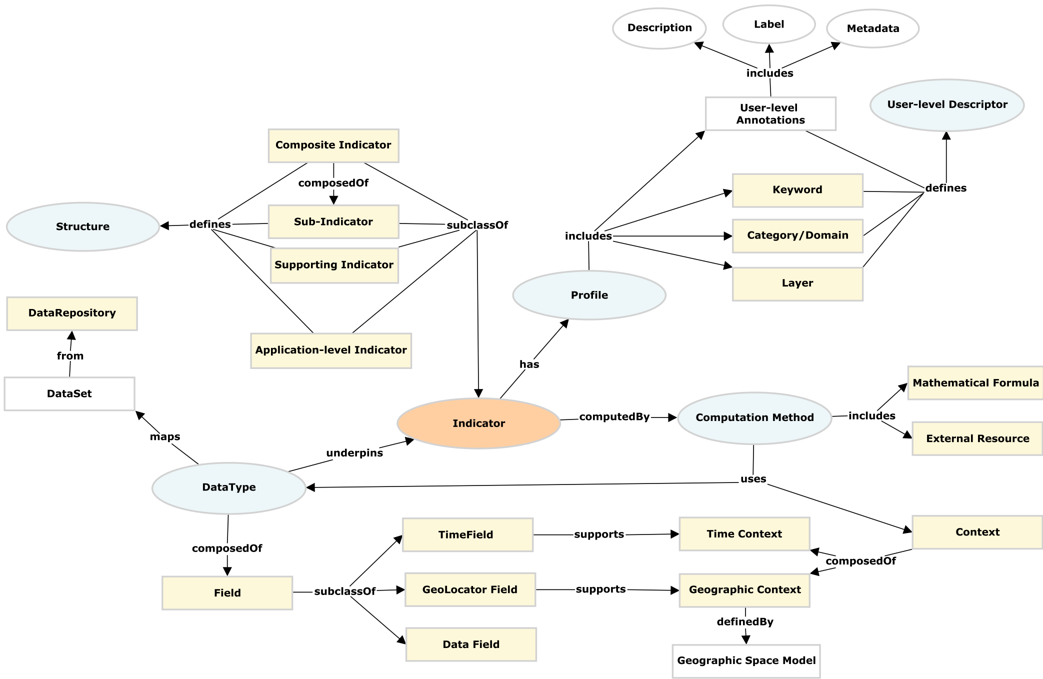

Figure 1: Conceptual Framework.

(c) A non-empty *data set* (D_i) underpinning i. Depending on the context,D_i can be either the data set used to compute the values of i or the values of i themselves.

(d) At least one *computation method* (M_i) providing a function $value_i$ reflecting the formula or the process to compute the values of i from D_i. If D_i represents the values of i, then $values_i$ is a simple relation of equivalence among sets.

Our model provides an intelligible perspective at the indicator ecosystem which is very close to the human common understanding of it. That is intrinsically extensible, meaning the cardinality of the tuple may be further increased or enriched to address additional representation requirements or features.

3.2 Indicator Ontology

We have implemented the conceptual framework as an OWL-DL ontology [8]. In order to simplify the description and the understanding of the data structure, as well as to facilitate its usability, we have partitioned the ontology in five sub-ontologies (groups of classes and properties) as the function of the scope within the model: *Indicator*, *Data*, *Profiling*, *Computations* and *Geographic Context*. In addition, those semantic structures are integrated with a number of generic-purpose annotation properties to provide the typical user-level annotations (such as descriptions in a natural language and labels), extended metadata (e.g. the publisher, the

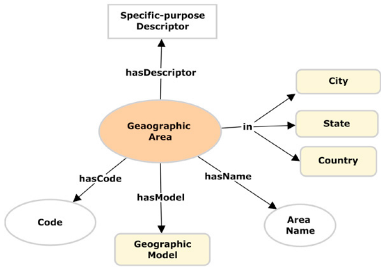

Figure 2: Geographic space representation.

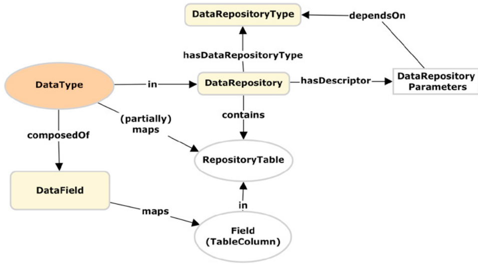

Figure 3: Mapping of a relational database.

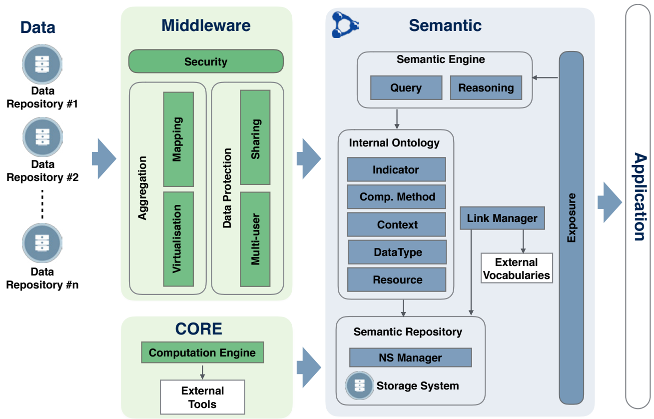

Figure 4: Reference architecture to enable the semantic infrastructure into real platforms.

owner and the data provider of a dataset) and specific-purpose descriptors (e.g. information related to the geographic areas).

The *Geographic Context* includes those concepts and properties needed to define the geographic space. The current implementation adopts a straightforward approach which assumes the different geographic areas uniquely identified by an ID withing the geographic model they are associated with (fig. 2). Generally speaking, a crucial aspect for the data aggregation is the *mapping* of our semantic representation with external assets, including data and computational resources. As an example, the concept map in fig. 3 shows the mapping between a *DataType* and a relational database. Such a mapping provides the semantic representation (DataType) of the physical data. The main classes, ordered by groups, are reported in table 1, as well as the main properties are listed in table 2. An exhaustive description of the ontology implementations and its use in practice is out of the scope of the paper.

3.3 Reference Architecture

An abstracted representation of the reference architecture which enables in fact our ontological support is shown in fig. 4. It is ideally structured into four functional layers (*Data*, *Middleware*, *Semantic*, *CORE* and *Application*) aimed at establishing a semantic ecosystem from distributed data environments scattered on the cloud. This implies the need for dynamic data retrieval, aggregation and processing mechanisms. Our architectural approach prioritizes:

Class	subclassOf	Description
	SubSet: Indicator	
Indicator	—	**Application-level Indicator**
CompositeIndicator	Indicator	Indicator composed of other indicators
SubIndicator	Indicator	Sub-indicator composing structured indicators
SupportingIndicator	Indicator	Indicator to support other indicators (e.g. application-level)
IndicatorType	—	**Type of an Indicator**
StructuredIndicatorType	IndicatorType	Type for structured Indicators
SupportingIndicatorType	IndicatorType	Type for supporting Indicators
	SubSet: Data	
Data	—	**Generic data**
DataType	—	**Data type**
DataRepository	—	**Data Repository**
DataRepositoryType	—	**Type of DataRepository**
SQLDatabase	DataRepositoryType	Type for SQL repository
RDFRepository	DataRepositoryType	Type for RDF repository
Field	—	**Generic field**
DataField	Field	Field associated with data
GeoLocatorField	Field	Field for geographic context specification
TimeField	Field	Field for time context specification
	SubSet: Profiling	
Domain	—	**Domain of a given data/indicator**
Layer	—	**Data or indicators layer**
SemanticDescriptor	—	**Semantic descriptor for a given concept**
	SubSet: Computations	
Formula	—	**Formula to calculate an indicator**
	SubSet: Geographic Context	
Country/State/City	—	Geographic filters
Area	—	**Geographic area**
GeographicModel	—	**Geographic model adopted**

Table 1: Main classes composing the ontology.

- *Semantics on a large scale.* It implies the accurate design of all the structural components which could propose significant trade-off between capability and performance. We apply a simple but effective logic: *load only what is needed, when you it is needed.* We will deal with the evaluation of our technological environment in a separate paper.

- *Semantic data aggregation.* It is the mechanism that enables the whole ecosystem in fact. A middleware layer that assures the use of a sophisticated semantics in a secure and reliable context is the real key factor for a successful application.

- *Dynamic access to external computational resources.* In order to assure a dynamic environment, our computation engine assumes the pervasive access to external computational resources and services. As for data, the performance of such mechanisms have a huge impact on the whole architecture.

4 Uncertainty and Open Issues

Our presentation of this work suggests a deterministic data ecosystem, namely the application of formal semantics to systematically and unambiguously specify the key concepts and critical relations associated with the urban indicator domain. However, despite the semantic support

Property	PropertyType	Domain	Intersection	Range
		SubSet: Indicator		
composedOf*	ObjectProperty	Indicator	—	—
computedBy	ObjectProperty	**Indicator**	Formula	—
hasIndicatorType	ObjectProperty	**Indicator**	IndicatorType	—
subIndicatorOf*	ObjectProperty	**Indicator**	Indicator	—
		SubSet: Data		
dataType	ObjectProperty	**Data**	DataType	—
hasDataField	ObjectProperty	**DataType**	DataField	—
hasDataRepository^{-}	ObjectProperty	**DataType**	DataRepository	—
hasDataRepositoryType^{-}	ObjectProperty	**DataRepository**	DataRepositoryType	—
isOutputOf^{-}	ObjectProperty	**Data**	Indicator	—
underpins^{+}	ObjectProperty	**DataType**	Formula	—
DBName^{-}	DataProperty	DataRepository	—	xsd:string
url^{-}	DataProperty	DataRepository	—	xsd:string
hasValue^{-}	DataProperty	Data	—	—
		SubSet: Profiling		
hasSemanticDescriptor	ObjectProperty	—	SemanticDescriptor	—
hasDomain	ObjectProperty	**Indicator** ∪ **Data**	Domain	—
hasLayer	ObjectProperty	**Indicator** ∪ **Data**	Layer	—
		SubSet: Computations		
hasInputParameter^{+}	ObjectProperty	**Formula**	DataType	—
hasMathExpression^{-}	DataProperty	Formula	—	xsd:string
inRepositoryTable^{-}	DataProperty	DataType	—	xsd:string
variableInRepository^{-}	DataProperty	Field	—	xsd:string
hasTimeStamp	DataProperty	Data	—	xsd:long
		SubSet: Geographic Context		
hasArea	ObjectProperty	—	Area	—
hasGeographicModel^{-}	ObjectProperty	**Area**	GeographicModel	—
hasCountry^{-}	ObjectProperty	**Area**	Country	—
hasState^{-}	ObjectProperty	**Area**	State	—
hasCity^{-}	ObjectProperty	**Area**	City	—
areaName	DataProperty	Area	—	xsd:string

$^{*}/^{+}$ pair of inverse properties $^{-}$ functional property

Table 2: Main properties in the ontological framework.

provided, uncertainty remains a major issue that requires specific attention to ensure reliable solutions. In applying the proposed ontology and dynamic data retrieval approach to the urban indicator domain, we have identified a number of significant causes of uncertainty, as follows:

"Bad" Data The philosophy behind open data is that data should be freely available to everyone to use as they wish without restrictions. Open Data is definitely a huge opportunity for a wide set of domains, from science to politics, especially when considered in a context of enhanced interoperability [9]. On the other hand, open data is also a risk, especially because the quality and the accuracy of the data cannot be always guaranteed. It applies to most open environments (e.g. Semantic Sensor Web [10]). Within this context, raw data play a critical role with regard to potentially generating uncertainties. Most publishers of indicators, only rely on a reduced number of reputable data sources (e.g. Government Agencies). This limits potential indicators to a small number of well-known official indicators. Other systems simply assume all data is reliable, and make no attempt to assess accuracy and reputability. Quantifying uncertainty and its propagation (e.g., in indicators derived by others) is complex and requires, apart from extensive semantics, specific mechanisms for analysis and risk assessment (e.g.[11]).

Concerning our semantic framework, we reduce significantly the uncertainty and push a completely open data philosophy (i) by providing formal relations between the indicators and the data underpinning them and (ii) by including descriptors for data providers, meaning the reliability of a given data set is associated mostly with the reputation of the provider. Thus, assuming the formal relation between the indicators and the data underpinning them, the data providers can be grouped and ranked according to their category. For example, the data provided by a government agency would be considered more reliable than the data retrieved by using Google APIs. The main advantage of this approach is that reliability is understood as an open concept, meaning there are no restrictions on the way in which that trust is estimated or evaluated; on the other hand, establishing largely accepted criteria, different from the generic obvious ones, to measure trust could be very unrealistic.

Similarity and Approximation The semantic equivalences among different indicators can be supported via upper-level domain vocabularies. However, complex cross-disciplinary environments are likely to require more complex semantic relationships, such as different indicators which are not completely equivalent in terms of semantics but merely similar in a given context (similarity). Establishing high-level semantic relations among indicators is one of the key feature for the evolution of the current platforms and environments; although such relations may introduce uncertainties. An example of relation which does not introduce uncertainties in the system is the similarity among indicators which have the same scope and, although produced in a different way from different data sets, have a similar "meaning" in a given context. At least in theory, there is no uncertainty as the similarity is recognized and formally defined at a semantic level. A different situation is the use of the similarity relation as an approximation of the target semantic. For instance, in the analysis of transportation systems, the average distance commuted to get an area a from an area b is often approximated by the distance between the centroids of the two areas. Depending on the context of the analysis, that approximation may be absolutely reasonable, as well as completely wrong, along a wide range of middle cases. If such approximation is not properly modelled and computed inside the ecosystem, the resulting uncertainty can be systematically propagated in the system producing results which can be inaccurate in the best case, completely wrong in the worst. Therefore, an indicator should be described and formally specified according to its real semantic and eventual similarities or approximations need a specific representation and processing. Depending on the relations considered and the application context, this kind of uncertainty can be modeled according to generic quantitative models (e.g. Probabilistic Semantics [12]) or by applying concept-specific frameworks (e.g. the Web of Similarity [13]).

Lacks in data sets The lack of relevant data to calculate indicators often leads to the use of next best data available. That establishes a kind of best-effort environment that relies on assumptions and approximations that are not properly modeled at a semantic level. This scenario intrinsically pushes uncertainties which are very similar to the ones discussed in the previous point. Any assumptions or approximations that underpin a given indicator, as well as the resulting uncertainties, should be modeled and taken into consideration throughout the production process and, of course, at an application level.

5 Conclusions

Addressing the next generation of urban indicators in an open technological context is an exciting objective which presents multiple convergent research issues. Indeed, while on one hand the use of rich semantics operating on a large scale enables in fact the target ecosystem, on the other hand it proposes significant challenges in terms of performance and management of the uncertainty.

This paper has mainly focused on the latter aspect: through an integrated semantic structure, which formally relates the high-level specification of the indicators with the methods to produce those indicators as well as with the data sets underpinning them, the uncertainty is easier to be identified and properly treated. However, a comprehensive representation and processing of the uncertainty in the indicator domain seems to be beyond the capabilities of the current Semantic Web technology. Our future work in this area will explore the application of novel approaches to model the uncertainty in the Semantic Web (e.g. [13]), as well as more sophisticated techniques to dynamically define and compose indicators from registered data sets through semantic inference. Moreover, we aims at the evaluation of the performance of our framework on a relatively large scale.

Acknowledgments

This research was supported by the Australian Government through the Australia Research Councils (ARC) Large Infrastructure, Equipment and Facilities (LIEF) funding scheme (project LE160100174).

References

[1] Stefan Bischof, Christoph Martin, Axel Polleres, and Patrik Schneider. Collecting, integrating, enriching and republishing open city data as linked data. In *International Semantic Web Conference*, pages 57–75. Springer, 2015.

[2] The RDF Data Cube Vocabulary. `https://www.w3.org/TR/vocab-data-cube/`. Accessed: 2016-11-30.

[3] Sören Auer, Jens Lehmann, and Sebastian Hellmann. Linkedgeodata: Adding a spatial dimension to the web of data. In *International Semantic Web Conference*, pages 731–746. Springer, 2009.

[4] Jovan Varga, Alejandro A Vaisman, Oscar Romero, Lorena Etcheverry, Torben Bach Pedersen, and Christian Thomsen. Dimensional enrichment of statistical linked open data. *Web Semantics: Science, Services and Agents on the World Wide Web*, 40:22–51, 2016.

[5] Vanessa Lopez, Spyros Kotoulas, Marco Luca Sbodio, Martin Stephenson, Aris Gkoulalas-Divanis, and Pól Mac Aonghusa. Queriocity: A linked data platform for urban information management. In *International Semantic Web Conference*, pages 148–163. Springer, 2012.

[6] Mark S Fox. A foundation ontology for global city indicators. *University of Toronto, Toronto, Global Cities Institute*, 2013.

[7] Salvatore F. Pileggi and Robert Amor. Mansion-gs: semantics as the n-th dimension for geographic space. In *International Conference on Information Resource Management, Conf-IRM*, 2014.

[8] Sean Bechhofer. Owl: Web ontology language. In *Encyclopedia of Database Systems*, pages 2008–2009. Springer, 2009.

[9] Prateek Jain, Pascal Hitzler, Amit P Sheth, Kunal Verma, and Peter Z Yeh. Ontology alignment for linked open data. In *International Semantic Web Conference*, pages 402–417. Springer, 2010.

[10] Salvatore F Pileggi, Carlos E Palau, and Manuel Esteve. Building semantic sensor web: Knowledge and interoperability. In *SSW*, pages 15–22, 2010.

[11] Andrea Saltelli. Sensitivity analysis for importance assessment. *Risk Analysis*, 22(3):579–590, 2002.

[12] Salvatore F Pileggi. Probabilistic semantics. *Procedia Computer Science*, 80:1834–1845, 2016.

[13] Salvatore F Pileggi. Web of similarity. *Journal of Computational Science*, 2016.