Towards Semantically Annotated Data-Driven Methodologies for Composite Metric Development in Traffic Incidents

Daniel M. Mejia[1]

[1] The University of Texas at El Paso, El Paso, TX 79968, USA
dmmejia2@miners.utep.edu

Abstract. At the core of Smart Cities solutions, both in fundamental and applied research, is the use of technology to improve quality of life of city residents. Measuring the improvement that specific solutions bring normally requires data collection and analytics before and after the implementation of such solutions. This work involves traffic incidents, where data is available for use, but there is a lack of a comprehensive understanding on safety and efficiency metrics. We believe that methodologies for modeling and evaluating semantically annotated data can be a driving factor for further understanding real-world scenarios in a city. By using a data-driven ontology, it is expected that new information can be represented, manipulated and used. Data-driven ontologies can enable the creation of metrics for use by a wide variety of stakeholders, from domain experts to city residents. This work focuses on the creation of semantically annotated data-driven indicators that are maintainable, changeable, and transferable amongst cities with similar data.

Keywords: Smart Cities, Semantically Annotated Data-Driven Modeling, Ontology, Interdisciplinary Research, Linked Data

1 Problem Statement

Smart Cities encompasses the notion of incorporating technology into the everyday lives of residents within a city. The need for Smart Cities stems from rapid urbanization of both large metropolitan and smaller urban areas; it is expected that by 2050 over 70% of the world population will live in a city [1]. Throughout the world there have been many different types of indicators that have been created, but none of them are widely used beyond the city or country that initiates them. Many of the countries that have adapted any sort of metrics to measure city performance use the idea of a *triple bottom line*, which has a focus on environmental, economic, and social issues [2]. Since Smart Cities have more than one specific domain focus of research, there are indicators to represent transportation, energy consumption, water use, health, economy, land usage, and several more areas [3].

Since the late 1990s and early 2000s there has been a pattern of developing indicators and metrics for transportation systems. Based on work done by Mihyeon [4], the United States Department of Transportation (USDOT) has a mission to "serve the United States by ensuring a fast, safe, efficient, accessible and convenient transportation system… and enhance the quality of life of the American people." Through the enhancement of the quality of life, it is crucial for indicators to be made standard so that it can be measured beyond its initial context. Current metrics are captured through the collection of data such as number of crashes per year or number of injuries per year; however, there are no metrics that are driven by the data itself to determine what it actually means in practice.

In the area of Smart Mobility, safety is the cornerstone of a majority of its indicators. There are many factors that relate to indicators with respect to traffic incidents, including the severity of the incident. Severity indicators are shown by describing the effects of the incident that occurred, such as injury severity, fatalities, or damage [5]. Furthermore, the frequency of incidents on the roadway and its location also play a role in safety indicators [6]. There are additional indicators that can also be considered to be safety related including blood alcohol concentrations (BAC) of motorists as well as to consider if they were under the influence of drugs [7]. Safety indicators on roadways must also consider drivers age, speed, seatbelt use, weather indicators [8][9], vehicle types [10], and time of day.

2 Relevancy

In the city of El Paso, Texas there have been more than 60,000 traffic incidents since 2014 [11]. The traffic incidents range from minor fender benders to multiple fatalities in a single incident. Throughout the world, and especially the United States, departments of transportation are working towards providing safe roadways and travel for those who use its roads. The development of semantically annotated data-driven metrics is beneficial to providing a larger composite understanding of safety and efficiency in the roadways. Department of transportations throughout the United States as well as road users will be able to use the metric that is developed. The knowledge that is gained through the transformation from data will allow for policy makers to understand what is truly occurring on roadways and a way to use that information to make improvements with respect to safety and efficiency.

This work will enhance the way that data is cleaned, tracked, and used for improved interoperability and ad-hoc analysis. These improvements promote transferability amongst similar and different domains; it also provides a foundation to inform city residents, policy makers and other stakeholders on the advantages of using Linked Data for sharing and consuming city-generated data.

3 Related Work

Data relevant to every city activity we may be interested in can be found. Understanding and measuring the data that is found for comparison and growth is critical to the improvement of technology and society. Meadows [12] writes, "Indicators are natural, everywhere, part of everyone's life… Indicators arise from values (we measure what we care about), and they create values (we care about what they measure)".

I believe that indicators and measurements are key for fostering services that improve quality of life for city residents. Measuring indicators is a promising way to understand progress and makes it useful for comparisons [13][14]. Cheu [15] writes that some of the common indicators found that relate to traffic describe specific events, including, but not limited to:

- Average travel time (in minutes)
- Average Speed (MPH)
- Average Delay (in sec/person, sec/Vehicle)
- Travel time reliability (index)
- Number of crashes per year
- Number of injuries or fatalities per year

Of these indicators, none of them describe traffic in a way that would be relevant for everyday commuters. It does not describe what caused the incidents, nor if there is a possibility that the geographic location played a factor with it.

3.1 Sustained Indicators Over Time

Work being done by Lazaroiu et al. [16] aims to develop a sustainability indicator of Smart Cities by addressing the economy, mobility, environment, people, living conditions, and governance. Each of these areas have sub-areas that are associated to it, from which are given a set of different possible weighted indicators. These types of indicators may appear appropriate to determine some type of metric for a city, but since they are given based on the opinion of a set of individuals, it is not necessarily an accurate representation of the city. Similarly, one of the most accepted indicators in the world is the ISO 37120:2014 sustainable development of communities – indicators for city services and quality of life [17]. The ISO 37120:2014 is the first of its kind for city standards; particularly on a global scale. The standards presented by the ISO are significant advances in the way that Smart Cities are measured, but the issue lies in its overwhelming generalization including, but not limited to: number of public transportation trips and number of two-wheeled motorized vehicle per capita, which do not necessarily describe safety or efficiency [18].

As cities grow, policy makers are making decisions that are data driven [18]. The United States Department of Transportation has issued a five-year Research, Development and Technology Strategic Plan that describes their research and development

priorities. Throughout the world at least 1.2 million people are killed each year due to road incidents [19] and up to 50 million additional people suffer injuries [6]; of all the deaths, more than half of them are between 15-44 years old [19].

For USDOT, the four main focuses are to promote safety, improve mobility, improve infrastructure, and preserve the environment [20]. In the United States, the problems with rapid urbanization are developing quickly. It is predicted that over the next 30 years the population will increase by 70 million and the economy will double [20]. With rapid growth, it is critical to examine the way that safety and mobility can be improved and measured.

According to the work by Hoornweg et al. [21], there are 12 major characteristics that an indicator has: must have a clear objective, be relevant to the objectives, be measurable and replicable, statistically representative of the city, comparable and standardized, potential to predict, effective, economical, interrelated to society, consistent and sustainable.

3.2 Indicator Significance

The Federal Highway Administration (FHWA) is attempting to address national highway challenges over the next five years. The goals that the FHWA are looking toward improving based on [20] are:

- Highway safety
- Improving mobility of people and goods
- Maintaining infrastructure integrity
- Enhancing system performance
- Promoting environmental sustainability
- Preparing for the future

For indicators to be considered a true indicator, it must be measurable [18]. Indicators are derived from facts and reveal new information [7] by nature. The transformation of qualitative data into useful quantitative is the key way to make a standard way to understand what is truly going occurring on the roadways. *I believe that the key to transforming individual data sets into useful knowledge is developing a significant indicator.* Indicators without a significance is just another arbitrary number of some event or set of situations.

4 Research Questions

The proposed work aims to address the following research questions:

Q1. *What do semantically annotated data-driven models contribute to the understanding of traffic incidents by different groups of stakeholders?*

Q2. *How can methodologies for transformation of data to knowledge through semantic-based models, services, and practices instill trust into transportation indicators?*

Q3. *How can very large decoupled data-sets be used to create quantifiable, standard metrics that are both reusable and comparable to other geographic locations, with respect to traffic incidents? (i.e. compare one city to another)*

Q4. *How can cleaned data-sets and developed semantic-based metrics be used to make predictions about recurrent traffic incidents?*

5 Hypotheses

H1. By providing contextualization to a real-world scenario it expected that the semantic-based, data-driven models create more comprehensive metrics that will advance the understanding in traffic incidents from the perspective of different groups of stakeholders.

H2. The proposed methodologies will provide a standardization on the process for domain experts to explore alternative data sets. By introducing a standard method, domain experts will be ensured that the processing of data is accurate and may be useful for producing metrics necessary for them and society at large.

H3. Large decoupled data-sets is expected to provide the foundation of a data-driven semantic annotated focus of Smart City research. Through semantically annotated data and data that is not immediately recognized to be related to the domain, it is expected that weighted values can be incorporated to describe the importance of each data point and understand the effect of incidents occurring throughout the city.

H4. Given developed metrics created from data, new information is expected to be predicted by providing possible real-world scenarios to attempt to understand ways to prevent incidents and improve Smart Mobility. In addition, formally described metrics can be transferred between regions and allow side-to-side comparison.

6 Preliminary Results

I have developed a semantic-based framework for monitoring freight performance in the borderland area of El Paso. This area poses a unique challenge given its unique position as a border with Mexico [9]. By using an ontology as a high-level data model, my previous work showed that heterogeneous data sets have relationships that may be useful in the development in traffic safety and efficiency metrics. This idea leverages

the foundation of Linked Open Data that is critical for data-driven research. Fig. 1 shows the way linked-data is used in this research along with several data points that is used in developing an indicator. Fig. 1 is similar to other ontologies such as VEACON [22] which will be used to help expand this knowledge graph.

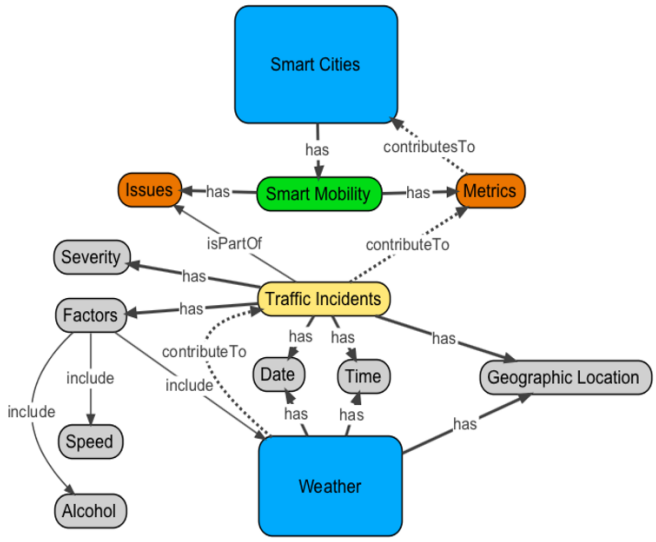

Fig. 1. Semantically annotated data-driven ontology model of Smart Cities and the Smart Mobility focus area [23].

From the implementation perspective, I have created multipurpose parsers that are able to handle different csv file inputs from multiple sources. The parsers clean and process the raw data into a numerical form (which will be used for future machine learning) then gives a JSON output of the cleaned data.

I have begun developing weights to the data points for the beginning stages of composite safety and efficiency indices for traffic incidents. The raw JSON files have been uploaded into NoSQL databases for preliminary competency questions. This research shows promise in taking large heterogeneous data sets, linking them together, then transforming it into data that can be queried, thus knowledge can be captured from it.

The critical composite indicator is computed in the following way: for every incident individual, take the summation of all of its attribute weights and divide by the maximum possible weight that can be obtained (the absolute worst case); then multiple the entire value by 100. This ensures that the absolute worst case possible will be equal to 100 and that the best case (no incident) is equal to 0. The range of possible values are between 0 and 100. Values 0-20 are considered to be a minor incident, 20.1 – 40 are considered to be major, 40.1 – 50 are considered to be severe. Values that are greater

than 50.1 are considered to be extreme because it will likely include physical injury or death. Incidents in general can quickly move from being minor to extreme based on circumstances of the event. Fig. 2 shows two incidents with various data points from the accident.

Crash ID	isFatal	isCMV	isSchoolBus	isRailroad	isSchoolZone	Crash Date	Crash Time	isThousandDamage	Primary Roadway ID	isIntersection	Weather Condition	Light Condition	Surface Condition ID	Traffic Control ID	Harmful Event	Surface Type	Serious Injury Count	Noninca pacitating Injury Count	Possible Injury Count	Non Injury Count	Unknown Injury Count	Death Count
144397						5/16/1																
14	1	0	0	0	0	5	18:56	1	2	1	11	1	1	8	2	4	4	1	2	1	0	2
149001																						
47	0	0	0	0	0	2/6/16	8:57	1	1	0	11	1	1	20	7	4	5	0	0	1	0	0


```
{
  "crashID" : "14439714",
  "compositeIndex" : "61.904761904761905",
  "sumIndex" : "780.0"
}
```

```
{
  "crashID" : "14900147",
  "compositeIndex" : "53.17460317460318",
  "sumIndex" : "670.0"
}
```


Fig. 2. Data set with computed critical composite index for each, respectively.

This work has led to early contributions in understanding incidents in the city of El Paso, but can be expanded to other cities throughout Texas and the United States. This work can be expanded by taking additional data sets and link them together for a refined indicator. The work contributes not only to the Smart Cities research but as well as Computer Science.

The methodology being used lays new ground in the way that semantic modeling can be used for linking heterogenous data as well as providing use for that data. Furthermore, the parsers that have been developed provide the state of the art in cleaning similar data sets for non-domain experts. The techniques used to create the parsers and JSON writers are independently created so that they can be run without needing the other, with exception to the data serving as an input. This methodology provides techniques to take theoretical knowledge graphs and models and transform them using relevant data into useful information that can be interpreted.

7 Approach

The proposed research is being done by a bottom-up data driven approach that focuses on data that has been collected from incidents and provided by official sources (e.g., government agencies). The data provides a historical reference as well as a provenance trace that enable users to trust the data. By establishing a way to connect heterogeneous data from the primary data model, I can begin establishing the effectiveness of this method compared to a top-down approach (where a model is made prior to data collection).

The proposed approach becomes iterative, where data, relationships, computations, and analysis can be made on what data is available. This allows for the answering of the research questions and hypotheses step by step and determine the quality of the approach at each stage of development for enhanced insight. Evaluation can

furthermore be expressed by understanding the way the critical composite index is developed and modified for improved understanding by domain experts and everyday users.

The proposed approach follows the idea that data drives the transformation of cities into Smart Cities. Smart Cities solutions come from observations, but this approach takes it to a deeper level where data drives the way we understand specific incidents in a city by providing a model to develop measurements from the Linked Data.

8 Evaluation Plan

Success will be measured two-fold: by developing a comprehensive methodology that can be analyzed iteratively and by developing metrics that is useful for domain experts as well as everyday road users. The ontologies and framework will be developed and evaluated iteratively first from the original data model until the final implementation that computes the metrics. The original data model comes directly from the data points that are being used and will be evaluated on the basis of correctness. Domain experts in the field will evaluate the usefulness by its ability to answer competency questions, willingness to adopt the metrics, and the trust that is instilled by the metrics.

Metrics will also be evaluated based on their ability to represent specific focuses such as safety or efficiency of road movement. The metrics developed will be compared to current metrics that have been already identified. Although a majority of the available metrics already developed are not representative of safety or efficiency at a composite level, they will be used to as a way to compare and analyze the validity of the developed metrics.

9 Reflections

The approach that is being taken by this work is data-driven. This idea stems from the foundation that data provides accurate facts. As a result of having data as the driving factor behind an entire methodology to develop a model representative of traffic incidents as well as the development of a metric it provides a foundation to enable access to knowledge bases by rendering the knowledge in different modalities (i.e. natural language text and raw data).

Acknowledgments
I would like to acknowledge my Ph.D. faculty advisor Dr. Natalia Villanueva-Rosales for her insight and support with this research. This work used resources from CyberShARE Center of Excellence, which is supported by National Science Foundation grant number HRD-0734825.

References

[1] M. Naphade, G. Banavar, C. Harrison, J. Paraszczak, and R. Morris, "Smarter Cities and Their Innovation Challenges," *Computer (Long. Beach. Calif).*, vol. 44, no. 6, pp. 32–39, Jun. 2011.

[2] K. Mori and T. Yamashita, "Methodological framework of sustainability assessment in City Sustainability Index (CSI): A concept of constraint and maximisation indicators," *Habitat Int.*, vol. 45, no. P1, pp. 10–14, 2015.

[3] S. K. McMahon, "The development of quality of life indicators—a case study from the City of Bristol, UK," *Ecol. Indic.*, vol. 2, no. 1–2, pp. 177–185, Nov. 2002.

[4] C. Mihyeon Jeon and A. Amekudzi, "Addressing Sustainability in Transportation Systems: Definitions, Indicators, and Metrics," *J. Infrastruct. Syst.*, vol. 11, no. 1, pp. 31–50, 2005.

[5] A. Laureshyn, Å. Svensson, and C. Hydén, "Evaluation of traffic safety, based on micro-level behavioural data: Theoretical framework and first implementation," *Accid. Anal. Prev.*, vol. 42, no. 6, pp. 1637–1646, 2010.

[6] E. Hermans, T. Brijs, G. Wets, and K. Vanhoof, "Benchmarking road safety: Lessons to learn from a data envelopment analysis," *Accid. Anal. Prev.*, vol. 41, no. 1, pp. 174–182, 2009.

[7] E. Hermans, F. Van den Bossche, and G. Wets, "Uncertainty assessment of the road safety index," *Reliab. Eng. Syst. Saf.*, vol. 94, no. 7, pp. 1220–1228, 2009.

[8] D. Mejia, E. Torres, N. Villanueva-Rosales, and K. Cheu, "Integrating Heterogeneous Freight Performance Data for Smart Mobility," *IEEE SCI 2017*, 2017.

[9] D. Mejia, "Integration of Heterogeneous Traffic Data to Address Mobility Challenges In the City of El Paso," The University of Texas at El Paso, 2017.

[10] E. J. Torres, "Ontology-Driven Integration of Data for Freight Performance Measures," The University of Texas at El Paso, 2016.

[11] TxDOT, "Crash Records Information System," 2018. [Online]. Available: https://cris.dot.state.tx.us. [Accessed: 15-Jan-2018].

[12] D. Meadows, "Indicators and information systems for sustainable development," *A Rep. to Balat. Gr.*, pp. 1–25, 1998.

[13] National Statistical Institute of Italy, "Enviornmental Sustainability Indicators in Urban Areas: An Italian Experience," *Jt. ECE/Eurostat Work Sess. Methodol. Issues Environ. Stat.*, no. 6 November 2006, pp. 1–15, 2001.

[14] R. B. Hiremath, P. Balachandra, B. Kumar, S. S. Bansode, and J. Murali, "Indicator-based urban sustainability-A review," *Energy Sustain. Dev.*, vol. 17, no. 6, pp. 555–563, 2013.

[15] R. L. Cheu and E. Balal, "Development of a Comprehensive Metric for Transportation , Environment , and Community Health," 2018.

[16] G. C. Lazaroiu and M. Roscia, "Definition methodology for the smart cities model," *Energy*, vol. 47, no. 1, pp. 326–332, 2012.

[17] R. Steele, "ISO 37120 standard on city indicators – how they help city leaders set tangible targets, including service quality and quality of life," no. October, 2014.

[18] M. S. Fox, "A Foundation Ontology for Global City Indicators," no. May 2015, pp. 1–37, 2015.

[19] World Health Organization, "World Report on Road Traffic Injury Prevention.," *Inj. Prev.*, vol. 10, no. 4, pp. 255–256, 2004.

[20] United States Department of Transportation (USDOT), "Research, Development, and Technology Strategic Plan FY 2017-2021," no. December, 2016.

[21] D. Hoornweg, F. Nunez, N. Palugyai, M. Villaveces, and H. W. Longfellow, "City Indicators: Now

to Nanjing," *World Bank Work. Pap.*, no. JANUARY 2007, pp. 1–71, 2007.

[22] J. Barrachina *et al.*, "VEACON: A Vehicular Accident Ontology designed to improve safety on the roads," *J. Netw. Comput. Appl.*, vol. 35, no. 6, pp. 1891–1900, 2012.

[23] D. Mejia and N. Villanueva-Rosales, "Semantically Annotated Data-Driven Models for Improved Smart Mobility Interoperability," *Int. Smart Cities Conf. Under Rev.*, 2018.