2017-01-01

Integration Of Heterogeneous Traffic Data To Address Mobility Challenges In The City Of El Paso

Daniel Michael Mejia
University of Texas at El Paso, dmmejia2@miners.utep.edu

INTEGRATION OF HETEROGENEOUS TRAFFIC DATA TO ADDRESS MOBILITY CHALLENGES IN THE CITY OF EL PASO

DANIEL MICHAEL MEJIA

Master's Program in Computer Science

APPROVED:

Natalia Villanueva-Rosales, Ph.D., Chair

Omar Badreddin, Ph.D.

Kelvin Cheu, Ph.D.

Charles Ambler, Ph.D.
Dean of the Graduate School

Dedication

To my family and everyone who has supported me. For the Glory of God. TGBTG.

INTEGRATION OF HETEROGENEOUS TRAFFIC DATA TO ADDRESS MOBILITY CHALLENGES IN THE CITY OF EL PASO

by

DANIEL MICHAEL MEJIA, B.S.C.S.

THESIS

Presented to the Faculty of the Graduate School of

The University of Texas at El Paso

in Partial Fulfillment

of the Requirements

for the Degree of

MASTER OF SCIENCE

Department of Computer Science

THE UNIVERSITY OF TEXAS AT EL PASO

May 2017

Acknowledgements

During the summer of 2016 I began to take an interest in research because I was accepted in the U.S. – Mexico Study Abroad Program focusing on Smart Cities in Guadalajara, Jalisco, Mexico. During my experience studying abroad I had the opportunity to immerse myself in the life of a researcher and understand the core principles that research provides for all of society. I also had the opportunity to learn more about Smart Cities and the ground-breaking work that was being done to ensure that we will have a sustainable infrastructure for generations.

I would first like to thank God and give Him all the Glory for all the success that I have had. I would also like to thank my family who has given me the upmost support, guidance, and has been a blessing to my life and academic career. It is with their support that I am able to continue working towards my goals each day. I would like to thank all my friends who have seen me through my entire academic experience and have given me their support.

I would like to thank Dr. Natalia Villanueva-Rosales for all her support and efforts in guiding me towards research in Smart Cities and data integration. She has greatly shown me what it is to be a researcher, never settle for mediocre and excel at all that I do. I would also like to thank Dr. Kelvin Cheu who has been a critical source of expertise in this area of research. I would like to thank Dr. Omar Badreddin who has given invaluable feedback and expertise of Computer Science.

I would like to acknowledge Eduardo Torres for laying the groundwork of this project and allowing me to expand it using a Computer Science approach. I would also like to thank Eric Camacho, Jose Caballero, and Moinul Porag Chowdhury for their contributions made to this project.

This work was supported in part by the National Science Foundation under CREST Grant, HRD-1242122 for the Cyber-ShARE Center of Excellence. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the author and do not necessarily reflect the views of the National Science Foundation (NSF).

Abstract

Transportation performance measures are defined as quantitative and qualitative indicators that rely on data or information to explain mobility, traffic congestion, safety, environmental sustainability and other factors. Although performance measures have been used for freeways and other highways, not many have been specified and applied to the freight transportation system. Under the Fixing America's Surface Transportation Act (FAST Act), state transportation agencies and metropolitan planning organizations in the United States are implementing freight performance measurement systems for performance assessment. This research aims to expand the existing limited freight performance measures and organize them into a comprehensive framework that can be reused for Smart Mobility.

The proposed framework consists of four criteria: safety, mobility, traffic congestion, and environmental sustainability. Each criterion consists of several qualitative and quantitative indicators. To address the challenge of integrating freight data of different sources and formats, an ontology-based approach has been proposed, demonstrated and evaluated using ontology evaluation techniques. The ontology was created using a bottom-up, data-driven approach, that included an initial concept map used to verify the conceptualization with domain experts before formalizing it in an ontology. The proposed ontology framework addresses specific metrics for the U.S. – Mexico border in El Paso, TX. However, the proposed framework is sufficiently generic and can be extended to integrate and aggregate data on a larger scale such as a state or country. This work also expands on the idea that a generic framework can be implemented beyond a single year of measurement. Data spanning over three years will also be used to determine the potential use of expanded ontologies in Smart Cities, specifically in Smart Mobility. By increasing the data set and adding external non-related heterogeneous data, this

work proposes a generic ontology that can be reused in multiple environments and scenarios. Freight Performance is a foundational piece of Smart Mobility, and by integrating heterogeneous data, it provides a way to evaluate and improve services in Smart Cities.

Table of Contents

List of Figures

List of Formulas

Chapter 1: Introduction

As the world population grows, many people are moving from rural to urban living areas. More than 50% of all the population in the world live in a city, or urban area, and by 2050 it is projected that 70% of all people will live in an urbanized community (Naphade, Banavar, Harrison, Paraszczak, & Morris, 2011). With the increased projection of growth for future generations, many actions are being taken to understand the need to make cities smarter; that is to incorporate and develop technologies that will improve the productivity of services, sustainability of the environment and resources, as well as to maintain the efficiency of the city. Moreover, each of these focuses of Smart Cities are intended to focus on the people, whether citizens, visitors, or general users of the city to have an increased quality of life.

There are many facets of Smart Cities such as, Smart Health, Smart Living, Smart Governance, Smart Buildings, Smart Mobility among others (Naphade et al., 2011). Each of these areas make up a "System-of-Systems" (Naphade et al., 2011) that work independently of one another to improve the entire city as a whole. This research is based primarily on understanding Smart Mobility and how it can be applied to a city with a modern, yet growing infrastructure.

1.1 Motivation

Understanding Smart Cities, specifically Smart Mobility needed greatly to gain information for analysis and technology improvements (Priano & Guerra, 2016). One of the driving factors to understand Smart Mobility is data. Through representation and manipulation, data can be transformed into information that will help domain experts and members of society

to become more knowledgeable about what the data is conveying (Orendain, Neri, & Bechelani, 2015).

Smart Mobility, on its own, is a large area of research that is being conducted to help improve the way that people move around a city. Smart Mobility extends beyond traffic congestion; it depends on a large set of factors including traffic incidents, weather and foot traffic, for example.

Much of the work being done in Smart Cities focus on finding efficient ways of obtaining data. Many of the current ways of retrieving data for Smart Mobility are through historical data sets, the Internet of Things (IoT), and crowdsourced information.

Moreover, understanding Smart Mobility provides a practical understanding to real-world applications. In freight performance analysis, there is a large amount of heterogeneous data that is untapped that can be potentially used to better understand the movement of freight trucks. This interdisciplinary work between Computer Science and Civil Engineering attempts to develop a domain specific ontology that can be expanded not only for freight performance analysis, but in mobility performance in general. This research focuses on freight performance in El Paso, TX, primarily in 2014 then extended for multiple years and attempts to model Smart Mobility with respect to a developed city.

1.2 Goal & Objectives

The **goal** of this thesis is to create data-driven metrics that provide a more comprehensive insight for freight performance in El Paso, TX.

The **objectives** of this thesis are:

O1. Create a Smart Mobility ontology that describes Freight Performance and how it may be affected by external factors that have not been previously considered.

O2.Enable the answering of questions about freight performance that require the contextualization and formalization of data provided by domain experts.

O3.Evaluate and refine the `Freight Performance Ontology` with the following criteria: i) consistency evaluation ii) data-driven evaluation, and iii) question answering evaluation based on competency questions provided from domain experts.

O4. Instill trust in the users of the Freight Performance Metrics ontology by providing the provenance trace of data using the W3C's recommendation Provenance Ontology (PROV-O) (Lebo, Sahoo, & McGuinness, 2013).

1.3 Organization

This thesis will be separated into seven chapters. Chapter 1 will describe the motivation, goal and objectives that is intended by this work. Chapter 2 will describe the background of Smart Cities and the ontology design process. Chapter 3 will discuss the work related to data integration and freight performance. Chapter 4 will describe the methodology towards developing a generic model of freight performance for the enhancement of Smart Mobility. Chapter 5 will describe the evaluation techniques used to ensure a quality model was produced. Chapter 6 will discuss the conclusions and contributions of this research work. Chapter 7 will discuss future work related to freight performance and Smart Mobility as it pertains to El Paso, TX and cities throughout the United States.

Chapter 2: Background

This chapter will describe the background of Smart Cities, Semantics and Ontologies.

2.1 SMART CITIES

There are many cities throughout the world, however not all of them are considered *Smart Cities*. Some of the core ideas at the center of being a Smart City is being intelligent, digital, sustainable, high use of technology, and with a goal of improving the quality of life for its citizens. Intelligent cities produce data through some sort of digital infrastructure or reporting techniques based on emerging technologies. Sustainable cities use the information that they gather to implement ways to reduce its carbon footprint and make themselves more efficient with respect to energy consumption (Dameri, 2013). Although Smart Cities are continuously being developed, there is no one widely accepted definition of what Smart Cities are. For this thesis, Smart Cities provide technology and services to promote an increase of efficiency, sustainability and productivity of services as well as to improve the quality of life for all the users of the city.

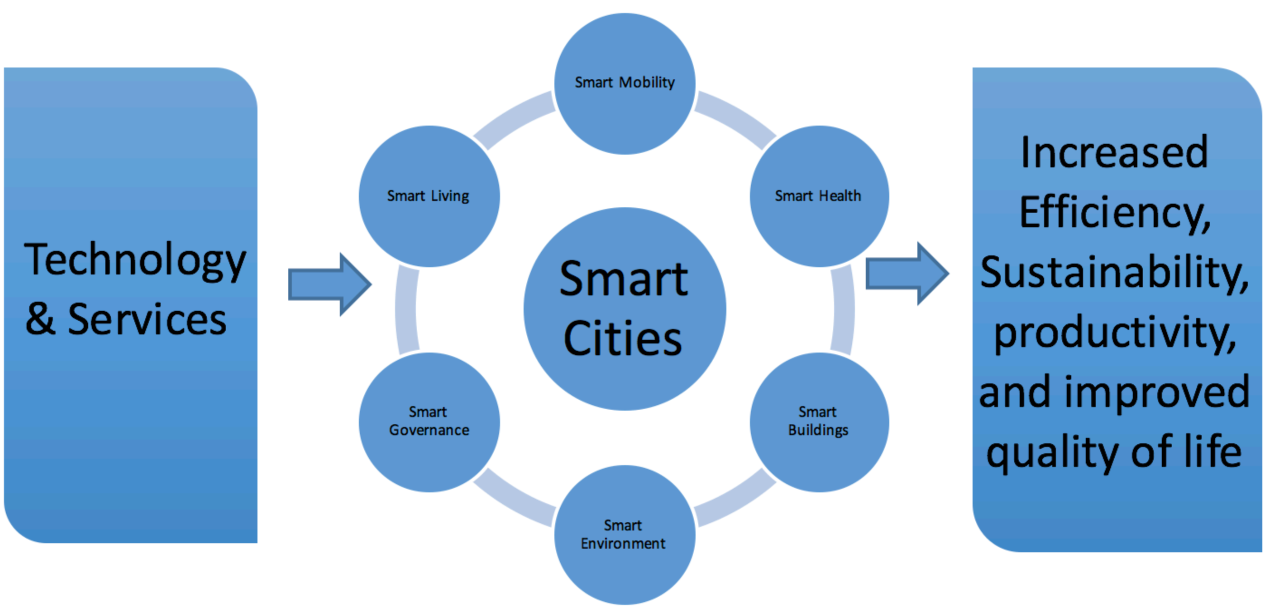

Figure 1. Integration of technology and Services into Smart Cities

Figure 1. describes the implementation and goals of Smart Cities. The incorporation of new technologies and services can be implemented to transform a city into a Smart City. There are many components of Smart Cities including, but not limited to: Smart Mobility, Smart Health, Smart Buildings, Smart Environment, Smart Governance and Smart Governance. These Smart City focus areas make up a "System-of-Systems" (Naphade et al., 2011) that work independently to build up an entire Smart City.

Many cities are attempting to transform into a Smart City because of the challenges that rapid urbanization has caused (Chourabi et al., 2012). Economic growth, technological progress, and environmental sustainability are also factors for cities to begin the transformation process (Naphade et al., 2011). It is expected that by 2030 many developing cities throughout the world will need approximately $40 trillion to support new urban infrastructure, which helps incorporate new technology directly into the infrastructure (Naphade et al., 2011). Investment into the infrastructure of cities promotes technological progress and environmental sustainability so that information can be taken and acted upon (Naphade et al., 2011).

Although there is not a completely agreed upon process to transform a city into a Smart City, there are several elements that are generally accepted. New technologies, a developed infrastructure to handle new technologies, proper governance (Chourabi et al., 2012) (Dameri, 2013) and policy (Nam & Pardo, 2011) are the building blocks to developing Smart Cities.

2.2 Ontology Design

Ontologies, high-level data models, provide means to describe concepts and their relationships in a domain. Regardless of the natural language (i.e. English, Spanish, French, etc.),

ontologies keep their defined structure (Chandrasekaran, Josephson, & Benjamins, 1999). Ontologies play a major role in interdisciplinary research; this is because it gives researchers the ability to see information at a high-level with a clear and formal vocabulary (Guarino, 1998). By introducing a common vocabulary, it is possible to develop a relationship between these classes and as a result link them together.

One of the key aspects of using ontologies in research is that it provides semantics – or meaning and context, to the data and relationships. This context provides researchers the ability to clearly see relationships between data that is not explicitly linked together, and thus link them together. Furthermore, it provides a vocabulary that is common amongst interdisciplinary researchers and allows them to share information. By describing concepts explicitly, ontologies provide a bridge to merge many areas of research (Noy & McGuinness, 2001). This research uses ontologies to describe freight performance and Smart Mobility in the city of El Paso, TX.

Figure 2. shows the relationship of some of the most used classes and subclasses in this research. The incidents class indicates all the data points that exist involving a freight truck. The figure shows that all of the individual incidents have an ID number that has been assigned to it by the state reporting agency (CRIS, 2015). Every individual is an incident, however, some incidents are more specific; if there was a fatality, then it becomes a fatal incident, if there was a weather-related incident, then it becomes a weather incident, with precedence being fatal incidents. This represents a part of how ontologies represent relationships and express how semantics are added to data sets.

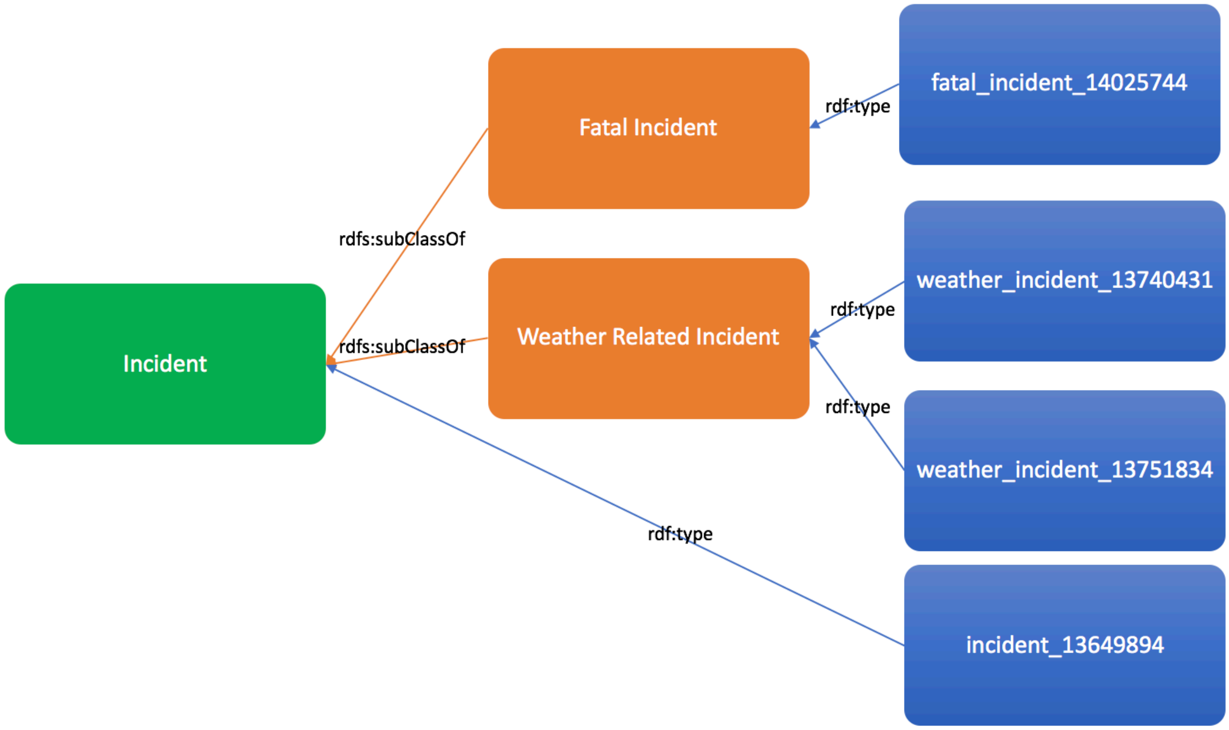

Figure 2. Representation of Incidents, its subclasses, and individuals of the class. The rdf and rdfs namespaces refers to the RDF language and RDF Schema respectively, part of the W3C recommendation languages used to describe ontologies.

Based on the guidelines presented in (Noy & McGuinness, 2001), the ontologies that are developed should follow a 7 step process. The process includes 1) determining a domain and scope of ontologies, 2) consider other ontologies for reuse, 3) develop a set of key terms, 4) define a hierarchy of classes, 5) define the properties of all the classes, 6) determine data types for populated data, 7) and create instances (or individuals) in the ontology.

Each step of the process should be carefully examined by the developer of the ontology. The developer should also be in constant communication with the domain expert to understand each domain specific parts that are involved with every step of the development process. In

general, ontologies will often begin as a conceptual taxonomy – or concept map, (Chandrasekaran et al., 1999).

2.2.1 RDF Triples

Ontologies are most often written using the Resource Description Framework (RDF) and the Web Ontology Language (OWL). These languages allows for machines to understand the semantics of the concepts and their relationships in a formal way (Dibowski & Kabitzsch, 2010).

RDF contains a structure that is based on a subject, predicate, and object (Cyganiak, Wood, & Lanthaler, 2014). According to the W3C (Cyganiak et al., 2014), asserting an RDF triple means that a relationship exists between the subject and an object with the relationship defined by the predicate. The subject is also considered to be the individual, the predicate the property between classes, and a value to be the object itself (Dibowski & Kabitzsch, 2010). The following is an RDF triple from the **`Freight Performance Ontology`**.

`incident_13649894 rdf:type fp:Incident`

Formula 1. Triple expression of incident individual and Incident class

In the example shown above, a relationship from the **`Freight Performance Ontology`** is shown as a triple. This triple shows how a subject is related to the object by the predicate rule. The subject is a specific individual, `incident_13649894`, that is a member of the object `fp:Incident`. As a result of being a member of the class, the incident (subject) is related to `fp:Incident` (object) by being of `rdf:type` (predicate).

Since everything represented using OWL can be represented as a triple, it creates special relationships that is similar to an object-oriented paradigm (Dibowski & Kabitzsch, 2010).

Furthermore, OWL follows the schema that is defined by RDF so relationships can be made between classes and the relationship that is between them.

2.2.2 SPARQL

According to W3C (Harris, Seaborne, Apache, Foundation, & Prud, 2013), SPARQL is commonly used to query ontologies; it is a query language for RDF. First, SPARQL is given its definition of what ontology to query by the namespace that is defined as a prefix of the query. Similar to the structure SQL, SPARQL uses keywords such as "SELECT" and "WHERE," amongst others. However, SPARQL searches ontologies based on the Triples that are defined in RDF. By searching for the subject, predicate, and object, SPARQL is able to access information similar to a graph database searching through all the nodes and bringing it to one place. A simple SPARQL query is the following:

```
PREFIX
fp:<http://ontology.cybershare.utep.edu/smartcities/FreightPerfo
rmance/#>
    SELECT ?incident
        WHERE {
            ?incident rdf:type fp:Incident;
        }
```

Formula 2. SPARQL Query example that retrieves all the incidents available in the **Freight Performance Ontology** with incident individual

The above query does the following: It will return all `?incident` values where the `?incident` is of `rdf:type` (relationship-to) `fp:Incident` (Ontology – **`Freight Performance Ontology`**, class – `Incident`). As a result, it will return all items in the class. As shown above, the query follows the triple format of subject, predicate, and object where the subject is what is being requested, the predicate is the relationship to what is being looked for and the object is where it is looking for the information.

2.3 DESCRIPTION LOGICS

Description logics, similar to first order logic, lays the foundation to describe relationships between classes and its members. Furthermore, it also gives insight to how the classes and members can relate to other classes and members. Since ontologies are unstructured, they can be difficult to understand. Description logics add a bridge to lightly structure an unstructured tool and make it understood providing semantic based logic (Baader, Horrocks, & Sattler, 2001). Everything related to ontologies is based on semantics, the meaning of the information. By providing a semantic based logic, a human understanding of the information is possible. For example, the description logics expression in Formula 3 describes that `WeatherRelatedIncident` is equivalent to an `Incident` and if in that incident there exists a relationship where `hasWeatherIncident` is true. . Description logics provide the rules that ontologies are based on, whether the relationships formed are true

$$\textbf{WeatherRelatedIncident} \equiv \textbf{Incident} \sqcap \exists.(\textbf{hasWeatherIncident})$$

Formula 3. Description Logic expression of the defined class `WeatherRelatedIncident` in the **`Freight Performance Ontology`**

The software used to understand and interpret rules described in description logics are called description logics reasoners. Reasoning services are used to determine consistency amongst ontology classes. Relationships use description logics as a foundation to logically prove that no logical rule is violated. Furthermore, by incorporating description logics into a reasoner, ontologies move towards linking data that is not explicitly related to each other by providing inferences (Baader et al., 2001). Since ontologies provide a semantic level meaning to data (Cruz & Xiao, 2005), when combined with description logic reasoners, they can make human understandable concepts understood by computers, thus they are able to make inferences on data and its relationships.

Along with determining consistency, they provide a way to make assertions about data. The assertions that are made must follow the logical rules that it is defined by. Since it is able to make logical assertions, it is then able to make inferences based on the information that it knows. Through reasoning, it is used to determine transitive relationships that are necessary for ontology development and queries.

Chapter 3: Related Work

This chapter will describe the relevant work that has been done in the area of freight performance and data integration.

3.1 Freight Performance

Based on the work of (Torres, 2016), freight transportation moves goods by water, land or freight. An estimated 20 thousand-million tons of freight was moved in the United States in 2013; which is approximately $49.3 billion/day equating to 55 million tons (Bureau of Transportation Statistics, 2015).

Understanding the way freight moves around a country, state, county, or city is important for the people who inhabit the area that the freight vehicles take up. This work will provide a foundation to bridge the gap between heterogeneous data sets and analysis that the data may play on freight movement. Four of the major categories that this work will focus on is Safety, Mobility, Congestion, and Environmental Sustainability (Torres, 2016). Furthermore, freight performance plays a major role in understanding Smart Mobility. As safety is the primary concern for this area of research, understanding the movement of large vehicles in the city of El Paso help researchers take into consideration other factors that may increase or decrease productivity and safety on local roads. Safety includes fatalities, injuries (Torres, 2016), and the possibility of weather factors affecting freight performance.

The Federal Highway Administration governed by the U.S. Department of transportation has published many highway and freight performance standards from states throughout the country ("Operations Performance Measurement Program," 2017). Throughout the United States, one of the most important aspects of freight performance that is being measured is the

movement of freight vehicles and the incidents that they are involved in. A lot of the work being done right now is understanding the role that freight movement plays on the economy (O'Rourke, Beshers, & Stock, 2015). In the state of Minnesota, calculations have been done to show the way that incidents, fatalities, and injuries affect the economy (O'Rourke et al., 2015). In effect, the research being done in this thesis can relate to the economics of incidents occurring in the state of Texas if a similar calculation table was developed. Moreover, there is work being done in addressing and modeling sustainability problems. This is being done through the Sustainability Performance Measurement framework (SPM) (Ramani, Potter, DeFlorio, Zietsman, & Reeder, 2011). This framework attempts to answer: "*What does a transportation agency need to be equipped with in order to successfully address sustainability issues through performance measurement?*" (Ramani et al., 2011) It follows a hierarchical structure to show their needs of understanding sustainability, development of goals, development of objectives, and implementation of transportation performance measures all with the intent of improving sustainability (Ramani et al., 2011). This work directly relates to the research that is being done in this thesis by first understanding the data that is provided to improve sustainability, efficiency, and productivity. Furthermore, the work being done in sustainability efforts based off performance measures acquires heterogeneous data from a large variety of sources and has many goals that are attainable by through data integration (Ramani et al., 2011).

In Missouri, a system has been developed so that all crash incidents must be reported to the Missouri State Highway Patrol to be added to a database. The incidents reported are added to the MoDOT system database so that it can be used for analysis (Shelton, 2016). The database being used primarily for statistical purposes as the information is not used in larger tools to analyze incidents regarding freight performance. Similarly, Texas has a database that reports all

of the traffic incidents that occur in the state; this can be used to retrieve information from both freight and consumer vehicle accidents (CRIS, 2015). This database is used repeatedly for data population of the **`Freight Performance Ontology`**.

In Minnesota, mode share is used to understand domestic freight movement. One of the most critical parts of the system is using information to prepare for the future. Since Minnesota is expected to have snowfall on a normal and nearly predictable basis, they use data to predict when to prepare for weather events; the weather events that they are most interested in is snow and ice. By having historical information regarding previous events, they are able to prepare the roads to handle the ice on the road. Furthermore, they also use data from the citizens to evaluate their satisfaction of services that are done by the freight mode share (Department of Import and Export Control, 2015). From the work being done in Minnesota, this thesis can grow to improve the way the understanding of historical data can be used to improve future responses, to reduce weather related incidents and ultimately improve safety of freight operators, and motorists on the road.

There is additional work being done with respect to urban freight transportation that models stakeholders of freight movement in an ontology. This model shows how each stakeholder is affected by the movement of goods (Anand, van Duin, & Tavasszy, 2014). This work relates to this thesis in which it could possibly be used to expand the **`Freight Performance Ontology`** that is being expanded to show how each class can affect potential stakeholders throughout the entire supply chain. An additional ontology models public transportation systems with respect to availability, comfort, and accessibility to the system. In this ontology the European Union provides a set of metrics that can be used to compare the ontology to (Mnif, Galoui, Elkosantini, Darmoul, & Ben Said, 2015).

There is a large amount of data that measure important freight performance indicators (finances, productivity, comparison to external performance, or resource allocation)(Board, 2011), much of which is difficult to determine what is meaningful (Torres, 2016). The data provided for this research has been determined meaningful by a domain expert. Furthermore, the data primarily focuses on the indicators described by the domain expert. This research will explore additional environmental data factors and attempt to link them together to better understand possible metrics that are related to freight performance.

3.2 Data Integration

Data integration is an area of research that has many different components involved in a successful application. Bringing the data together in one place is often a cause of concern for researchers to use the data. Data integration is thought to be the bringing of information into a single standardized place using a specific schema for all of the data being brought in (Dale et al., 1992). The idea of bringing data into a specific place is necessary for all data to be understood and interpreted in a clear way without losing the information that it presents. With many different data sources, researchers may find the need to take data from multiple sources and combine or share the information together (Heimbigner & Mcleod, 1985).

Many of the issues with many data sources is that they are often in different forms. The data may be in different forms because of the natural language it is in, data types, or the different technological systems it was originally presented in (hardware & software differences) (Sheth, 1999).

Large sets of data have been attempted to be gathered as a result of attempting to understand freight performance measures. Relational databases have been commonly used as a way to handle large sets of data from GPS (Global Positioning Systems) (Ma, 2011). Data

integration of freight performance has a critical success factor being the quality of data that is collected (Liao, 2009). This work has shown that it has been difficult to create queries because of the immense size of data that is populated into the relational database. Furthermore, the work being done with relational databases primarily focuses on one set of data (movement from place to place) based on information from commercial truck vehicles (McCormack & Hallenbeck, 2006).

Based on the work of (Martinez-Cruz, Blanco, & Vila, 2012), ontologies provide semantics to the data that it is representing by incorporating object and data properties that are relevant to the data point. Moreover, ontologies do not need to be normalized for relationships to be clearly seen. Ontologies provide the generic framework and flexibility needed for different types of data, whereas relational databases are often limited to the data type that it can handle. Lastly, the reasoners in ontologies provide a way to infer new information and relationships where relations databases require assertions to be made using primary and foreign keys. Using a join function in relational database does not compensate for the semantics that ontologies provide through its structure and reasoners.

Ensuring that the data is accurate allows for integration of provenance models to help understand its history and add a trust value to it. It is typical with freight data to have a large data set that needs to be handled. The research being done in understanding movement of freight trucks pertains primarily in understanding the efficiency of moving goods from place to place. Although the research done in this work is related to understanding efficiency of freight movement, it focuses on understanding freight performance based on integrating multiple data sources that provide additional information pertaining to the factors affecting freight performance, e.g., weather conditions.

Relational databases are commonly used to handle data (Liao, 2009); the handling of large data in relational databases are similar to what is being done in ontologies. However, the work being done in freight performance usually relates to one specific target area (usually movement of goods). The information that is extracted from the databases are necessary for research, however since there are areas beyond movement of goods, ontologies are better suited to handle the data. Since triple stores are able to handle billions of triples (Dibowski & Kabitzsch, 2010), there is not an issue to store or query the data set. According to the work done by (Rohloff, Kurt and Dean, Mike and Emmons, Ian and Ryder, Dorene and Sumner, 2007), adding 100 million triples (up to a billion triples) to a triple store at a time and testing the loading time appears to be nearly O(log n). However, the querying of the same data size appears to be nearly O(n) – linear. The complexity of the querying of SPARQL can furthermore be explored by using description logics. Introducing universal and existential qualifiers, a logical formula can be developed to show the relationship between computation time and individuals that are being evaluated.

Moreover, since data is coming from many heterogeneous sources, including historical data, ontologies are able to handle abstract relationships between data sources. Furthermore, the work done in this research utilizes an ontology so that it may use the generic reasoner to not only to ensure that relationships do not violate any logical rules, but as well as to gain new inferred axioms that are determined by the reasoner. Relational databases do not have the capabilities to ensure that relationships between tables are logically accurate, nor have inferred relationships.

The work being done in this research focuses on integrating more than one type of data category and integrate them into a single cohesive framework. This research is extending what a

relational database does by storing a dataset, and moves towards structuring unstructured heterogeneous data using ontologies.

Ontologies attempt to mitigate the issues that occur with data integration by developing a standard way of describing heterogeneous data. By building a vocabulary and mapping concepts to one another through a defined set of relationships, the data can be described clearly and can be seen how data is related to one another. Furthermore, these relationships are enhanced by providing semantic annotations to the data; by including semantics to the data, computers are able to understand the meaning of what has been defined by humans (Buccella, Cechich, & Rodriguez Brisaboa, 2003).

Introducing the concepts of ontologies to data integration also brings the idea of semantic data integration. The idea of semantic data integration stems from the idea that researchers can develop conceptually accurate representations of data and its relationship to other heterogeneous data so that the heterogeneities can be eliminated (Cruz & Xiao, 2005). Developing a way to conceptually understand the data that is being represented can provide a way to connect the data together cohesively.

The work being done in this research takes a data-driven approach to reduce heterogeneity between data sources through the development of an ontology. Additionally, the semantics that are introduced by the incorporation of an ontology to data provides a way to take data from different sources and merge them together (Gardner, 2005) in the same way that domain experts would manually merge them. This work provides semantics to the data that is being collected from multiple data sources, and provides a way to link them together without modifying the original sources of data and enhancing the data with context and constraints that

reflect domain expertise.

Chapter 4: Methodology

This chapter describes the process of designing and implementing the **`Freight Performance Ontology`**, which can be found at http://ontology.cybershare.utep.edu/smart-cities/FreightPerformance.owl and its corresponding documentation at http://ontology.cybershare.utep.edu/smart-cities/FreightPerformance#. A portion of the work regarding the **`Freight Performance Ontology`** is based off of (Mejia, Torres, Villanueva-Rosales, & Cheu, 2017).

4.1 Approach

The design of this research is focused on understanding freight performance, and Smart Mobility in general, through a data-driven bottom up approach. A data-driven approach uses current data and domain expertise to define what the main concepts that are of interest. Based on the data that is given, an ontology can be designed.

By approaching a design based on data, the research will focus on developing logical relationships between data sources guided by domain expertise. Building from specific concepts that describe data to less specific concepts defines a bottom-up design approach. The approach of this research was to modify a concept map based on the specifications of a civil engineer and transform it to a simplified version. The simplified concept map was then used to create and modify a **`Freight Performance Ontology`** that is for use in consumer vehicle mobility descriptions and analysis, as specified by the domain expert.

4.2 STEP 1. CREATING A CONCEPT MAP

Following the findings of (Pennington, 2010), the initial task was to build a concept map to illustrate the main concepts in the ontology and confirm with the domain experts (i.e., students and faculty in civil engineering, practicing engineers in state and local government agencies) of the initial model for freight performance metrics. The concept map was designed to represent the relationships between various concepts – or classes, and the data that is populated within it. The concept map organized sub-domain relationships between different concepts.

The concept map was built to describe freight performance in four categories, defined by a domain expert: Safety, Traffic Congestion, Environmental Sustainability, and Mobility (Torres, 2016). Each of the major categories of the ontology have additional characteristics to further describe the sources of data and metrics for each category. The safety category is where most of the data will be populated; one of the major indicators of safety are incidents. There are two subclasses of incidents: weather-related incidents and fatal-incidents; these specific concepts were also defined by the domain expert based on the data provided by the TxDOT Crash Records Information System (C.R.I.S) (CRIS, 2015). Based on the data that will be populated into the ontology, there is information that shows if weather was determined to be a factor in the incident. If someone was deceased as a result of the incident, then the fatal incident class takes precedence (Torres, 2016). All other incidents that were determined to not be the result of weather or had no fatalities were an ordinary incident data point.

Additionally, the domain expert determined the need to know notification and arrival times of emergency personnel to the scene. This is represented in the concept map as response time to incidents; furthermore, the average response time is a function that is determined when the ontology is populated.

Lastly, the safety category has a `Daily Weather Metrics` that is related to the indicators of safety given the implicit relationship of weather with weather related incidents. This class has data properties that include temperature, wind, dust, and precipitation.

To illustrate the concept map, Figure 3 shows the Safety category; this category describes the relationships between incidents, their locations, and additional factors related to the incident. The nodes in green represent the classes in which the data source is from Excel files, the purple nodes represent information that is computed based off the data in real-time, and the remaining nodes must continue to be researched.

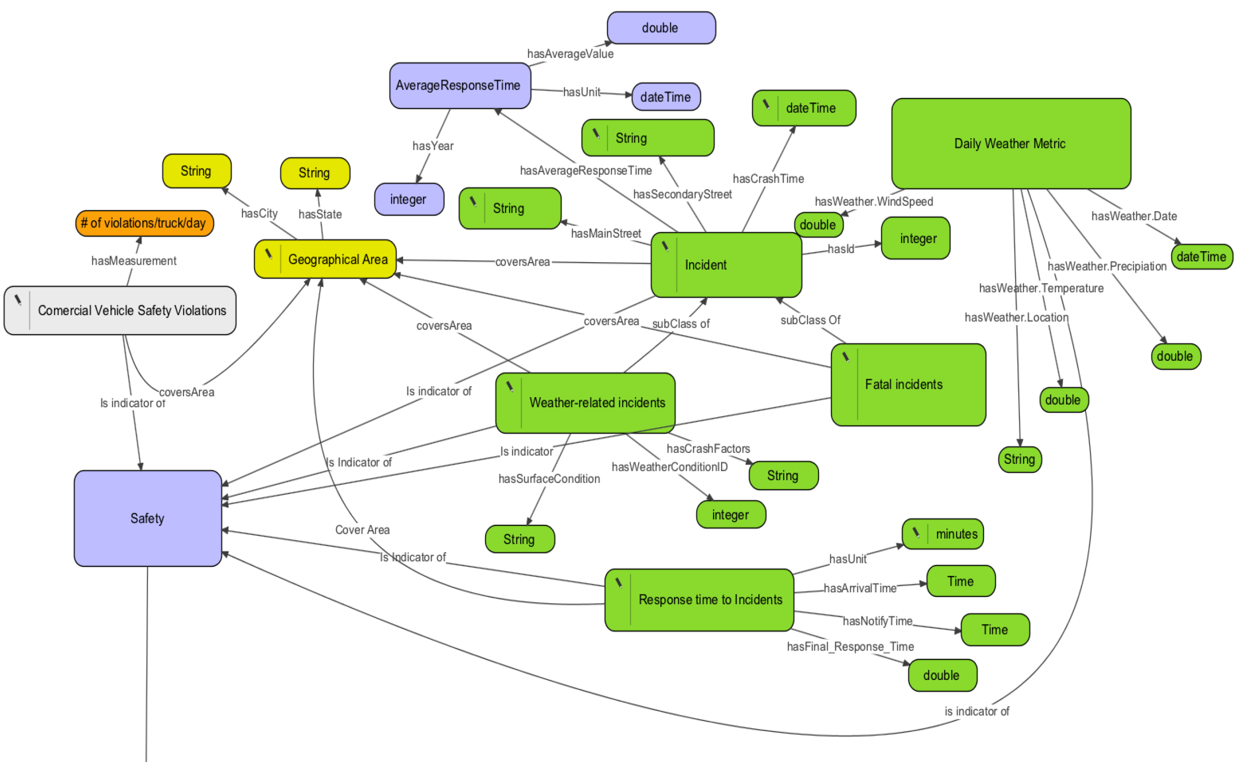

Figure 3. `Safety` Category of the concept map for Freight Performance Metrics (Torres, 2016) extended with weather metrics information.

4.3 STEP 2. DEFINING COMPETENCY QUESTIONS

Competency questions were created to identify the main concepts of the ontology, how they are related to each other and evaluate the ontology after the design phase. The competency questions were provided by a civil engineer, an expert in freight performance. The competency questions are designed to gather information that experts in the domain area are most interested in. The competency questions were driven by the data that is available. Moreover, the questions that are asked are designed to get information that is not explicitly stated in the data. The competency questions initially provided for this research are listed below:

1. *Which Port of Entry has the shortest average waiting time in 2014?*
2. *Does El Paso have congestion/day of more than or equal to 8 hours and 50% of the system congested?*
3. *What is the Average Daily Traffic volume of the I-10 segment at the Americas Interchange?*
4. *How many accidents are there at I-10 at Hawkins interchange?*
5. *What is the Toll Revenue per car? (Toll Revenue/ADT Volume at Cesar Chavez Hwy at Fonseca)*
6. *How many accidents occurred during morning peak hours (7am to 9am) in 2014?*
7. *What were the weather conditions when weather-related incidents occurred?*

The competency questions and concept map were created based on the data already available and how that data could be integrated to create categorized metrics by domain experts. The ontology was created starting with the most specific concepts (e.g., **`WeatherRelatedIncident`**) building to higher level ideas (e.g.,

WeatherRelatedIncident isIndicatorOf Safety) from data values, while respecting the relationships indicated by domain experts.

4.4 STEP 3. ONTOLOGY IMPLEMENTATION

The **Freight Performance Ontology** was built using the following process: 1) classes using an ontology editor, 2) use the classes to use as a starting point for the Java-based OWL API (Horridge & Bechhofer, 2011), 3) develop parsers for data sets and merge them together, and 4) populate the ontology with the data and run a reasoner. The implementation of the ontology was done by a team of computer scientists (Mejia, Chowdhury, et al., 2016).

4.4.1 Concepts to Classes

The first step of building the ontology was mapping the classes and relationships from the concept map to ontology terms. Following the steps suggested in (Noy & McGuinness, 2001) and (Villanueva-Rosales & Dumontier, 2008) the concepts and concept hierarchy of the **Freight Performance Ontology** were created using the ontology editor Protégé (Musen, 2015) and the Web Ontology Language (OWL) (W3C, 2012). OWL reuses terminology from RDF (Cyganiak et al., 2014) such as Class and `subClassOf` that are listed in the queries in Chapter 5.

The classes were manually created using the ontology editor, and all the relationships were described according to the specification of the concept map, developed in coordination with the domain expert and Computer Scientists. This process created an is-A hierarchy of all the classes. The schema of all the classes (also called TBox in description logics) provided the structure required to populate the ontology with. Furthermore, the hierarchy provides a generic framework to measure other instances of incidents outside of the scope of freight performance.

Based on logical understanding of relationships of classes and individuals, some of the classes were changed from the original concept map. The classes were changed to represent more accurately what the relationships represented based on a hierarchical design. This process helped ensure that would-be data members of the class would have a logical representation to the real-world.

The class hierarchy OWL file was imported into the population program that our the team developed. The program took in the file of classes that we needed wanted to be in the final ontology product; by doing this there was no need to create each class programmatically using the OWL API.

The file was loaded into the program and associated to be a prefix that would be associated to all future individuals that would be created. The prefix associated each class needed from the file, and thus each individual could be associated directly to the class hierarchy defined in the program.

As new individuals are created, the program will associate them to the existing classes that is in the hierarchy. A new ontology file is written with all the classes from the class hierarchy and the eventually populated individuals.

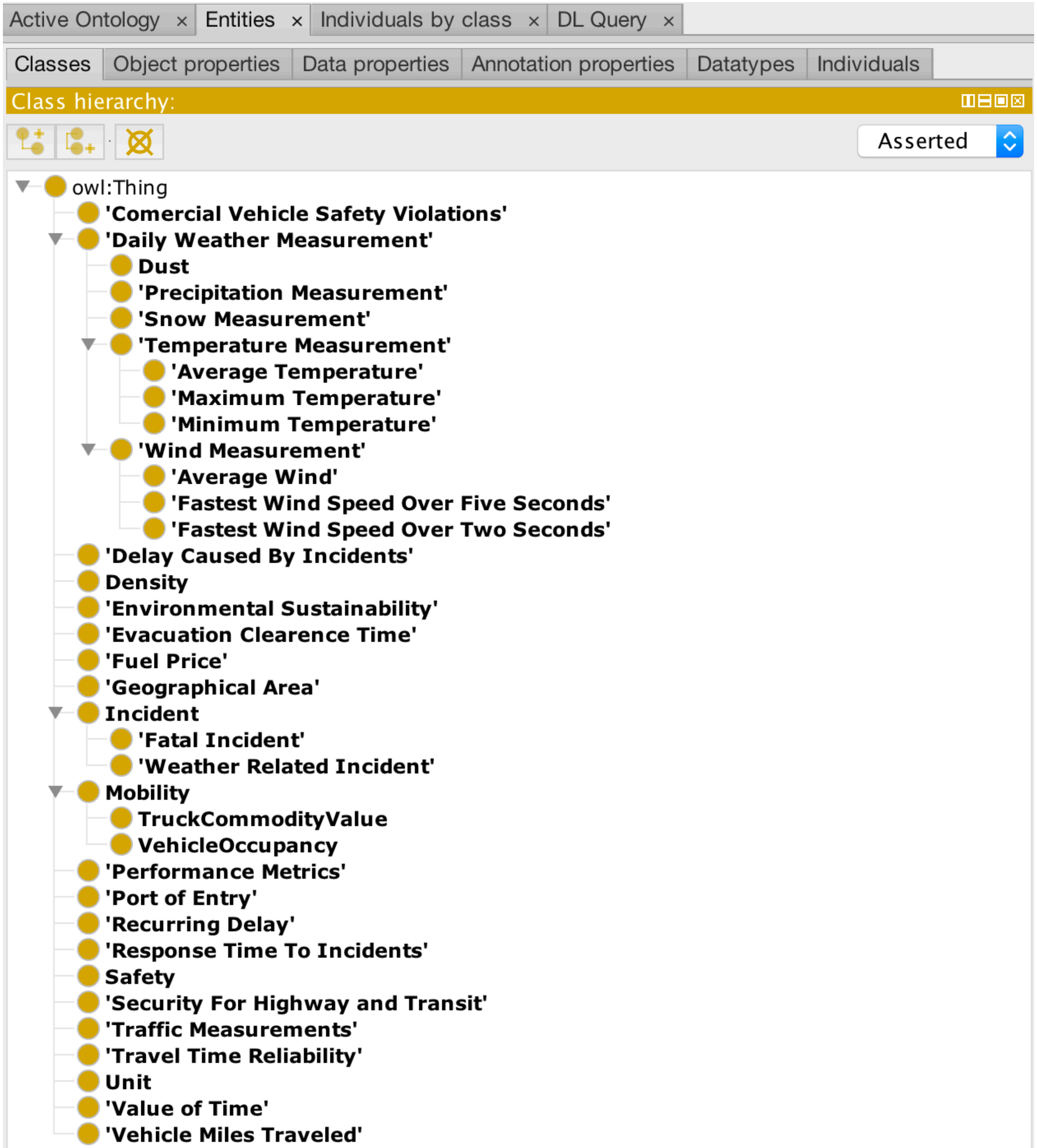

Figure 4. Concept hierarchy (Classes) of the Freight Performance Ontology using Protégé 5.1 (Musen, 2015)

4.4.2 Ontology Population (Data Processing)

One of the main advantages of using ontologies is the ability to semantically annotate data. To annotate heterogeneous data with the concepts developed in the ontology, *parsers* were developed by the team of computer scientists to format all data into RDF using the OWL API.

4.4.3.1 Processing Data from Excel Files

The ontology population process consisted of ingesting data provided in Excel files of over 860 incidents that occurred in El Paso County in 2014. The file contained information of all the data formatted by Comma Separated Values (CSV). Only specific information was requested by the domain expert and as a result not all information was retrieved for each incident. The file contained a column heading that was used identify the information contained in each one of the columns and how it would be mapped to the ontology. The columns that were requested by the domain expert were programmatically and dynamically searched and indexed; by assigning the columns an index, it provides a way to keep track of where the values for each incident will be. When the parser is executed, it maps the indexed values to the values of each individual incident to an instance of the ontology; by doing this it ensures that the values in each column match with the information that are of interest.

A similar approach was taken for retrieving weather data provided by the National Oceanic and Atmospheric Administration (NOAA) (NCEI, 2017). The weather data provided additional heterogeneity between concepts and explored additional relationships that could be useful to integrate the data.

4.4.3.2 Processing Data from PDF Files

Another source of data for the **`Freight Performance Ontology`** were PDF files. The PDF used contained specific information on toll revenue (CRRMA, 2014) that was identified to be relevant by a civil engineer. A software library was used to search through the text and retrieve the necessary values of toll revenue (Lowagie, 2010); this was initially done as part of a course project by by a team of computer scientists (Mejia, Camacho, Caballero, & Chowdhurry, 2016).

4.4.3.3 Processing Online Data

A website was used to find the values of international ports of entry crossing times on the U.S. – Mexico border (Bridge of the Americas, El Paso, TX and Ysleta Port of Entry, El Paso, TX) ("Border Crossing Information System Commercial Vehicles," 2016). Through web services, relevant values were retrieved. The values taken were for the period of January 1, 2014 and December 31, 2014, inclusively.

4.4.3 Ontology Population

After parsing each of the Excel files, PDF files and online data, data was added into the ontology using the OWL API library. All the individual data points that were used from the Excel files, PDF files, and online data became individual instances of ontology concepts.

Initially, the data parsers and the OWL API methods were separate as to ensure that each works properly individually. Each individual parser written for the files were placed into a single file that would be used populating the data files. Since the parsers were written with high modularity, they could easily be transferred to a new file that would populate the ontology without causing errors to the other parser functions.

The population of an ontology took each individual data points from the parsed file and created instances of concepts in the ontology; a process called ontology population. In this research, the OWL API was used to programmatically create instances from: Excel files, PDF files and websites. Each data parser contained functions that made a distinction between different data fields that would be populated into the ontology. Each data field was associated to the class hierarchy by using the predefined prefix; the individuals would be then associated to the class that it was intended to be a member of. Furthermore, the individual would be assigned a data and object property that was defined in the class hierarchy.

When the functions are called in the program, the parser will contain the functions to add each individual directly to a new ontology file that is being written. The parser will loop through each individual data point, associate it to the proper class it is going to be a member of, assign its data and object property, then add it to the ontology.

Combining the parsers with the functions provided from the OWL API provides accessibility to the data and keeps each data set separate from the others. It gives a layer of protection to the other data points that are also being run on the program.

The information taken from the Excel files (CRIS, 2015) were mapped to ontology concepts such as: incident ID, crash time (and date), fatality information, crash location, and emergency response time.

Additional weather data was also used integrate weather conditions for each day of the year. Weather conditions included daily average temperature, daily maximum temperature, daily minimum temperature, daily precipitation/snowfall, daily wind averages, and daily maximum recorded wind speeds over two seconds and over five minutes, respectively. Each of these individuals were also associated with a date and populated into the ontology.

Once the ontology was populated with all the data, a generic reasoner, HermiT (Information Systems Group Oxford University, 2016), was used to verify the consistency of the populated ontology and validate that all the instances created had the appropriate syntactic format infer new information. Inferred information refers to information that is implicitly defined by asserted axioms, but not explicitly defined in the knowledge base.

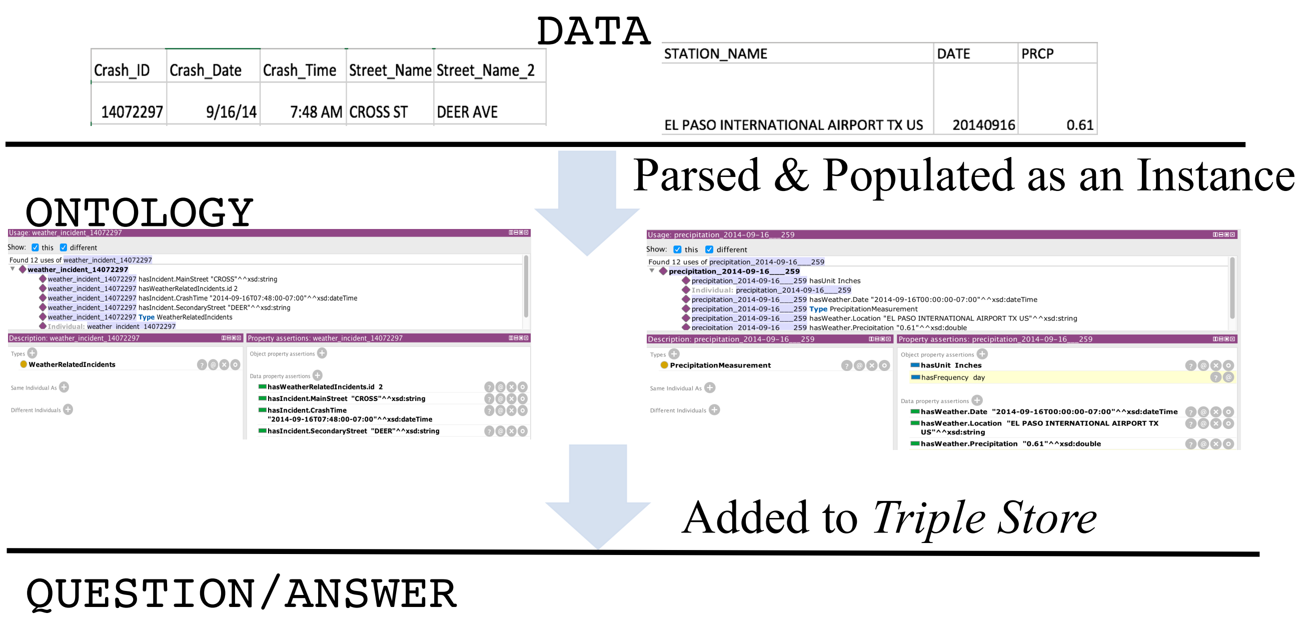

Figure 5. Flow of raw data from data sources to the **`Freight Performance Ontology`** and to the question/answer component.

The raw data shown in figure 5 is obtained from CMV Incidents (CRIS, 2015) as well as the weather (NCEI, 2017) for that particular day (Mejia et al., 2017).

4.4.4 Data Provenance

Data provenance provides the ability to understand where, when, and how data was collected so that users may understand the usefulness of data and most importantly its

trustworthiness (Moreau & Groth, 2013). In this research, provenance annotations were added to keep track of where data points originated from; by understanding the source of the data, a final user can evaluate whether the source is trustworthy.

To create a trace of provenance, additional concepts in the ontology were added to describe data sources that are relevant to the **`Freight Performance Ontology`**. The classes that describe the data sources have values that correspond to the file names that are being used to populate the ontology. Furthermore, all the classes that hold instances related to data have an asserted axiom that describes the data source. Provenance annotations were asserted using the Provenance Ontology (PROV-O), a W3C recommendation (Lebo et al., 2013). Incorporating data provenance fosters further interoperability with other ontologies and tools that also use PROV-O.

4.5 Ontology Knowledgebase

As described in section 2.2.1, concepts and instances in an ontology, can be represented as RDF triples (subject, predicate and object) that relate each individual to the class it is in. Triples are managed using a *Triple Store* – a database-like server that can handle billions of triples (Dibowski & Kabitzsch, 2010).

The populated **`Freight Performance Ontology`** was uploaded in to the *Triple Store.* In this research, Jena Fuseki (Apache, 2016) was used as a *Triple Store*, to store the triples generated from the **`Freight Performance Ontology`** and queried the ontology, similar to the way relational databases are queried. The *Triple Store* was locally hosted and contained 31,000 triples regarding incidents, weather, provenance, and other freight performance data.

Chapter 5: Evaluation

This chapter describes how the **`Freight Performance Ontology`** was evaluated.

5.1 Ontology Evaluation

The process of evaluating an ontology was a three-fold: 1) consistency evaluation, 2) data-driven evaluation, and 3) question-answering evaluation.

5.1.1 Consistency evaluation

One evaluation criteria of an ontology is whether it is consistent to ensure that the concepts and axioms do not contradict each other and that no logical rules are violated. The **`Freight Performance Ontology`** was evaluated against consistency using the HermiT (Information Systems Group Oxford University, 2016) reasoner provided through both the OWL API and through the Protégé interface; both of which were used.

5.1.2 Data-driven evaluation

Data-driven evaluation (Brank, Grobelnik, & Mladenić, 2005) is performed by comparing an ontology with the data referred to in the ontology. Given that the **`Freight Performance Ontology`** was created using a data-driven, bottom-up approach, a manual evaluation was performed to compare the data from the different sources and how this data was represented in the populated ontology. In this comparison, verification was done to ensure that all the data was properly mapped (i.e., no data loss) and that it was properly classified in the expected classes. This evaluation was made by a team of computer scientists creating the ontology population and the civil engineers as domain experts. A complimentary evaluation to the data-driven evaluation is the question answering evaluation, explained in the next subsection.

5.1.3 Question answering evaluation

An additional evaluation criteria of an ontology (Vrande et al., 2010) is its ability to answer the competency questions (i.e., queries) defined early on by domain experts. This application-based evaluation (Brank et al., 2005) helps understand if the ontology produces the expected results and provides additional information through inferences made by the reasoner. Ontologies provide context to data, which enable experts the ability to ask queries that involve such a context as well as inferences from a reasoner. This task is referred to as question answering.

The **`Freight Performance Ontology`** was evaluated by its ability to answer the competency questions. This thesis only contains a portion of the evaluated competency question to highlight specific functionality of the **`Freight Performance Ontology`**. Initially, each competency question was written in plain English. Each question was translated into a query that can be processed by an ontology-based system. Given the structure of the questions, the research team opted to translate them into SPARQL queries (Harris et al., 2013). The namespace **`fp:`** is used to identify terms in the **`Freight Performance Ontology`**.

Each of the following queries represent the human readable version of the competency questions and the SPARQL query executed against the *Triple Store.* Note that competency questions 2 and 3 needed additional data beyond what it is currently in the knowledge base. However, these questions have a similar structure to questions 1 and 5 respectively and they will be able to be queried once data is available.

<u>Query 1 (Competency question 1)</u>

The following query represents the competency question 1: *Which Port of Entry has the shortest average waiting time in 2014?* Query 1 is constructed to find the average waiting

time for two of the International Port of Entries in El Paso, TX. This query takes the average of each bridge by accessing them by **`rdf:type fp:PortOfEntry`**, where **`fp:hasPOE.Minutes`** returns the average waiting times for each individual month in 2014. The bridges are accessed individually by **`fp:hasPOE.Location "Americas Port of Entry"^^xsd:string`** and **`fp:hasPOE.Location "Ysleta Port of Entry"^^xsd:string`**. An aggregation function AVG will find the average of all bridge waiting time for each of the bridges, **`?adelay`** (Americas delay) and **`?ydelay`** (Ysleta delay), respectively. It will then check if the "Americas Port of Entry" delay is greater than the "Ysleta Port of Entry" delay by comparing the two bridge wait times. If **`?adelay`** is greater than **`?ydelay`** it will return the value contained in **`?YsletaPOE`**, if it is false it will return the value in **`?AmericasPOE`**. The result of the query answering competency question 1 is **"Ysleta Port of Entry"** with an average waiting time of **41.6** minutes. This is less than the result of **"Americas Port of Entry"** with an average waiting time of **44.75** minutes. The SPARQL representation of Query 1 is listed below:

```
SELECT IF((?adelay >?ydelay), ?YsletaPOE, ?AmericasPOE)
WHERE{
   SELECT ?AmericasPOE (AVG(?americasdelay) as ?adelay)
?YsletaPOE (AVG(?ysletadelay) as ?ydelay)
   WHERE{
     ?poe1 rdf:type fp:PortOfEntry;
         fp:hasPOE.Location "Americas Port of Entry"^^xsd:string;
          fp:hasPOE.Location ?AmericasPOE;
          fp:hasPOE.Minutes ?americasdelay.

     ?poe2 rdf:type fp:PortOfEntry;
          fp:hasPOE.Location "Ysleta Port of Entry"^^xsd:string;
```

```
          fp:hasPOE.Location ?YsletaPOE;
          fp:hasPOE.Minutes ?ysletadelay.
    }
    GROUP BY ?AmericasPOE ?YsletaPOE
}
```

Formula 4. SPARQL query to determine the lowest bridge waiting time

Results:

```
============================
| "Ysleta Port of Entry" |
----------------------------
```

Figure 6. Results from Shortest Port of Entry Query

The following complimentary query shows that indeed, "Ysleta Port of Entry" has a shorter average wait time for 2014. This query is similar to the above query; however, it returns the name of the Port of Entry along with its average waiting time for each of the two respective locations that the domain expert is interested in.

```
SELECT  ?location1 (AVG(?americasdelay) as ?adelay) ?location2
(AVG(?ysletadelay) as ?ydelay)
  WHERE{
     ?poe1 rdf:type fp:PortOfEntry;
        fp:hasPOE.Location "Americas Port of Entry"^^xsd:string;
        fp:hasPOE.Location ?location1;
        fp:hasPOE.Minutes ?americasdelay.

     ?poe2 rdf:type fp:PortOfEntry;
        fp:hasPOE.Location "Ysleta Port of Entry"^^xsd:string;
        fp:hasPOE.Location ?location2;
        fp:hasPOE.Minutes ?ysletadelay.
     }
```

```
GROUP BY ?location1 ?location2
```

Formula 5. SPARQL query to determine the average waiting time for all Port of Entries

Results:

```
---------------------------------------------------------------------------------------------
| location1                 | adelay | location2              | ydelay                      |
=============================================================================================
| "Americas Port of Entry" | 44.75  | "Ysleta Port of Entry" | 41.666666666666666666666666 |
---------------------------------------------------------------------------------------------
```

Figure 7. Results of the average waiting time for each Port of Entry

As expected, this query confirms competency question 1 by showing the exact waiting time over the year 2014 for each Port of Entry.

Query 2 (Competency Question 4)

The following query demonstrates how ontologies queried with SPARQL can be used to aggregate data. Query 2 corresponds to competency question 4: *How many accidents are there at I-10 at Hawkins interchange?* The first part of the query retrieves all of the incidents that of **`rdf:type fp:Incident`** (and its subclasses). Then it looks for all of the incidents that have **`fp:hasIncident.MainStreet`** reported to be **`I 10`** and its **`fp:hasIncident.SecondaryStreet`** reported as **`Hawkins`**. All distinct **`?individual`** are and counted and a **`?count`** is returned.

```
SELECT (COUNT(DISTINCT ?individual) AS ?count)
  WHERE {
     ?individual rdf:type fp:Incident;
          fp:hasIncident.MainStreet "I 10"^^xsd:string;
          fp:hasIncident.SecondaryStreet "HAWKINS"^^xsd:string;
}
```

Formula 6. SPARQL retrieving a total count of incidents at the I 10 and Hawkins interchange

Results:

```
---------
| count |
=========
| 2     |
---------
```

Figure 8. Total number of Incidents at I-10 and Hawkins

The results returned from the query is **2**, which means that of all the incidents that were reported in 2014, 2 of them happened at I-10 and Hawkins interchange.

<u>Query 3 (Competency Question 5)</u>

The following query describes competency question 5: *What is the Toll Revenue per car? (Toll Revenue/ADT Volume at Cesar Chavez Hwy at Fonseca*). The data sources involved with this query is toll revenue from a PDF file and traffic measurements from an Excel file (Torres, 2016).

```
SELECT  ?individual ?value ?primaryStreet ?intStreet
?trafficVolume
      WHERE {
          ?individual rdf:type fp:TrafficMeasurements;
               fp:hasTollRevenue.Value ?value.



          ?adtIndividual rdf:type fp:TrafficVolume;
                  fp:hasTrafficVolume.PrimaryStreet "Cesar
     Chavez Border Hwy"^^xsd:string;
                  fp:hasTrafficVolume.PrimaryStreet
     ?primaryStreet;
                  fp:hasTrafficVolume.IntersectionStreet
     "Fonseca Dr"^^xsd:string;
```

```
                    fp:hasTrafficVolume.IntersectionStreet
    ?intersectionStreet;
                    fp:hasTrafficVolume.Value ?trafficVolume.

}
```

Formula 7. SPARQL Query to retrieve traffic all toll revenue and traffic volume at Cesar Chavez Border Hwy and Fonseca Dr.

Results:

individual	value
<http://ontology.cybershare.utep.edu/smart-cities/FreightPerformance/TollRevenue>	"81644.0"^^xsd:double

primaryStreet	intersectionStreet	trafficVolume
"Cesar Chavez Border Hwy"	"Fonseca Dr"	35520

Figure 9. Results of total revenue and traffic volume at Cesar Chavez Border Hwy and Fonseca Dr.

The query in formula 7 is a two-part query that answers competency question 5. The query first retrieves the `?individual` of `fp:TrafficMeasurments` and retrieves the ?value of the `fp:hasTollRevenue.Value`. The next part of the query retrieves all `?adtIndividual` that have `fp:hasTrafficVolume.PrimaryStreet "Cesar Chavez Border Hwy"` of type `^^xsd:string` and `fp:hasTrafficVolume.IntersectionStreet "Fonseca Dr"` of type `^^xsd:string`. This query returns the `?individual` instance and ?value of traffic tolls, the `?primaryStreet` and `?trafficVolume`.

Query 4 (Competency Question 6)

The following query describes competency question 6: *How many accidents occurred during morning peak hours (7am to 9am) in 2014?* The data sources for this query is the CMV2014 Excel file (CRIS, 2015) which only includes the incidents for 2014. When data for other years is integrated in the knowledge base, an additional filter for the year will be needed.

```
SELECT (COUNT(DISTINCT ?individual) AS ?totalPeakIncidents)
WHERE {
     ?individual rdf:type fp:Incident;
                       fp:hasIncident.CrashTime ?date.
     BIND (hours(?date) as ?hour)
     BIND (minutes(?date) as ?minute)
     FILTER(?hour >= 7 && ?hour<=8)
}
```

Formula 8. SPARQL Query to count all of the incidents between 7 and 9 am in 2014

Results:

```
-----------------------
| totalPeakIncidents |
=======================
| 114                |
-----------------------
```

Figure 10. Results of total number of distinct incidents between 7am and 9am

This query counts every `?individual` that is returned and defines it as a `?totalPeakIncidents`. The query retrieves every `?individual` instance from `fp:Incident` and retrieves the incident ?date through the data property `fp:hasIncident.CrashTime`. The `?date` value is then parsed by using built in SPARQL functions to retrieve `hours` and `minutes`. The query then filters all the individuals that have an hour greater than or equal to 7 and less than or equal to 8. By doing this we assume that the time we are interested in is between 7 and up to the minute before 9 am. By running this

query, it is found that a total of 114 incidents involving commercial motor vehicles occurred between 7 and 9 am.

Query 5 (Competency Question 7)

The following query leverages the ontology process illustrated in Figure 5. The query demonstrates how ontologies and SPARQL automatically integrate data of two different data sources. Query 5 corresponds to competency question 7: *What were the weather conditions when weather-related incidents happened?* The data sources used in this query are: incident (CRIS, 2015) and weather conditions (NCEI, 2017). The first part of the query retrieves an individual of **`rdf:type fp:WeatherRelatedIncident`** (incident), and the specific time of the incident through the property **`fp:hasIncident.CrashTime`** with a specific date. The second part of the query retrieves a metric of **`rdf:type fp:hasWeather.Precipitation`**, where the metric has values **`fp:hasWeather.Date`**. These two values are represented by **`?incidentDate`** and **`?weatherDate`**, respectively. The query filters all the dates that have two dates equal (day, month, and year). This query allows a user to understand the weather conditions that may have contributed to the incident occurrence. The SPARQL representation of Query 2 is listed below:

```
SELECT ?weatherDate ?incidentDate ?rain ?snow
WHERE {

     ?incident rdf:type fp:WeatherRelatedIncident;
               fp:hasIncident.CrashTime ?incidentDate.

     ?precipitation rdf:type fp:PrecipitationMeasurement;
                    fp:hasWeather.Precipitation ?rain;
                    fp:hasWeather.Snow ?snow
                    fp:hasWeather.Date ?weatherDate.
```

```
	BIND(year(?incidentDate) as ?incidentYear)
	BIND(year(?weatherDate) as ?weatherYear)
	BIND(month(?incidentDate) as ?incidentMonth)
	BIND(month(?weatherDate) as ?weatherMonth)
	BIND(day(?incidentDate) as ?incidentDay)
	BIND(day(?weatherDate) as ?weatherDay)
	FILTER(?incidentYear = ?weatherYear && ?incidentMonth =
	?weatherMonth && ?incidentDay = ?weatherDay)
}
```

Formula 9. SPARQL query to retrieve all rain events when there is a weather-related incident

Results:

weatherDate	incidentDate	rain
"2014-12-09T00:00:00-07:00"^^xsd:dateTime	"2014-12-09T17:20:00-07:00"^^xsd:dateTime	"0.0"^^xsd:double
"2014-08-26T00:00:00-07:00"^^xsd:dateTime	"2014-08-26T22:44:00-07:00"^^xsd:dateTime	"0.06"^^xsd:double
"2014-08-22T00:00:00-07:00"^^xsd:dateTime	"2014-08-22T04:11:00-07:00"^^xsd:dateTime	"0.67"^^xsd:double
"2014-03-02T00:00:00-07:00"^^xsd:dateTime	"2014-03-02T05:53:00-07:00"^^xsd:dateTime	"0.16"^^xsd:double
"2014-08-02T00:00:00-07:00"^^xsd:dateTime	"2014-08-02T09:09:00-07:00"^^xsd:dateTime	"0.37"^^xsd:double
"2014-03-19T00:00:00-07:00"^^xsd:dateTime	"2014-03-19T10:09:00-07:00"^^xsd:dateTime	"0.0"^^xsd:double
"2014-09-16T00:00:00-07:00"^^xsd:dateTime	"2014-09-16T05:07:00-07:00"^^xsd:dateTime	"0.61"^^xsd:double
"2014-09-18T00:00:00-07:00"^^xsd:dateTime	"2014-09-18T11:27:00-07:00"^^xsd:dateTime	"0.5"^^xsd:double
"2014-09-16T00:00:00-07:00"^^xsd:dateTime	"2014-09-16T07:48:00-07:00"^^xsd:dateTime	"0.61"^^xsd:double
"2014-03-14T00:00:00-07:00"^^xsd:dateTime	"2014-03-14T19:43:00-07:00"^^xsd:dateTime	"0.02"^^xsd:double
"2014-07-01T00:00:00-07:00"^^xsd:dateTime	"2014-07-01T22:18:00-07:00"^^xsd:dateTime	"0.08"^^xsd:double
"2014-09-18T00:00:00-07:00"^^xsd:dateTime	"2014-09-18T13:41:00-07:00"^^xsd:dateTime	"0.5"^^xsd:double
"2014-08-25T00:00:00-07:00"^^xsd:dateTime	"2014-08-25T11:13:00-07:00"^^xsd:dateTime	"0.01"^^xsd:double
"2014-09-17T00:00:00-07:00"^^xsd:dateTime	"2014-09-17T22:45:00-07:00"^^xsd:dateTime	"1.86"^^xsd:double
"2014-08-12T00:00:00-07:00"^^xsd:dateTime	"2014-08-12T17:08:00-07:00"^^xsd:dateTime	"0.05"^^xsd:double
"2014-09-22T00:00:00-07:00"^^xsd:dateTime	"2014-09-22T14:30:00-07:00"^^xsd:dateTime	"0.52"^^xsd:double
"2014-03-27T00:00:00-07:00"^^xsd:dateTime	"2014-03-27T03:49:00-07:00"^^xsd:dateTime	"0.0"^^xsd:double
"2014-08-01T00:00:00-07:00"^^xsd:dateTime	"2014-08-01T14:40:00-07:00"^^xsd:dateTime	"0.16"^^xsd:double
"2014-09-12T00:00:00-07:00"^^xsd:dateTime	"2014-09-12T09:03:00-07:00"^^xsd:dateTime	"0.22"^^xsd:double

Figure 10. Results where there was rain and an incident on the same day

The results retrieved from the query include all the weather incidents and the precipitation on that day. For example, one result of **`?weatherDate`** is **`"2014-09-17T00:00:00-07:00"^^xsd:dateTime`** and **`?incidentDate`** is **`"2014-09-17T22:45:00"`** and **`?rain`** was **`"1.86"^^xsd:double`** (inches) for that date. A total precipitation of **`1.86`** inches in the El Paso area is not very common and a user can further

investigate if this was the main reason of having weather-related incidents that day. Although each of the queries done in this research is specific to the competency question provided by the domain expert, they can be generalized further or made more specific based on what is necessary.

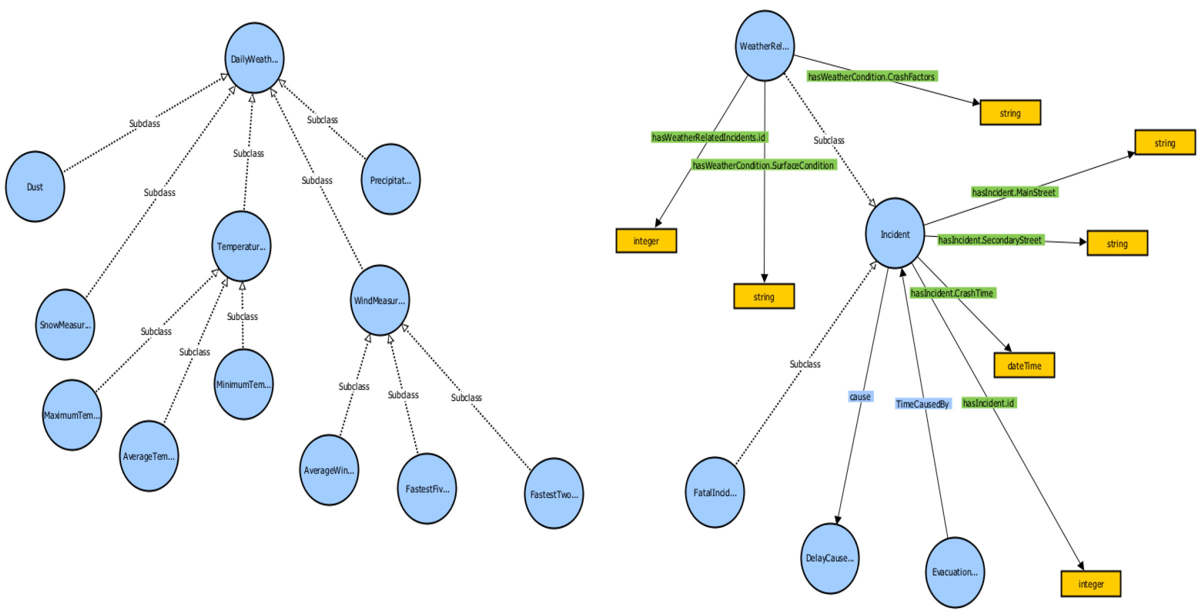

Figure 11. Visualization of defined classes in the **`Freight Performance Ontology`** using VOWL (Lohmann, Negru, Haag, & Ertl, 2016)

Figure 11. shows the how the classes “`DailyWeatherMetric`” and “`Incidents`” are disjoint. This model was produced by VOWL (Lohmann et al., 2016) as a Protégé plugin. The model shows the relationship between classes, and its subclasses and is only a portion of the entire ontology. As a result, it can be understood that there is no direct association to these two classes; however, comparison between the two can be provided by linking date values together and providing a semantic relationship beyond a simple join as it is done in relational databases.

Moreover, the query can be represented in the ontology as a defined class of a more specific type of incidents – those whose weather conditions has been made available. In comparison to Figure 3. The concept map and the ontology represent the same information. The two classes are not directly connected to each other, however only connected through a common link, which is the date in the safety category.

Moreover, provenance models have been added to the ontology by using PROV-O (Lebo et al., 2013). By adding PROV-O to the ontology, classes are mapped in the **`Freight Performance Ontology`** to understand the history of data sources and the ontology. The provenance model added was done by creating an additional concept called **`DataSources`** which have individuals that related to the data sources used. Figure 12., and Figure 13. Show that there is an asserted provenance axiom that an **`Incident`** has an **`Influencer`** of **`some Activity`**; and that all **`Incidents wasInfluencedBy CMV2014-2016`**, which is an individual representing the true data source.

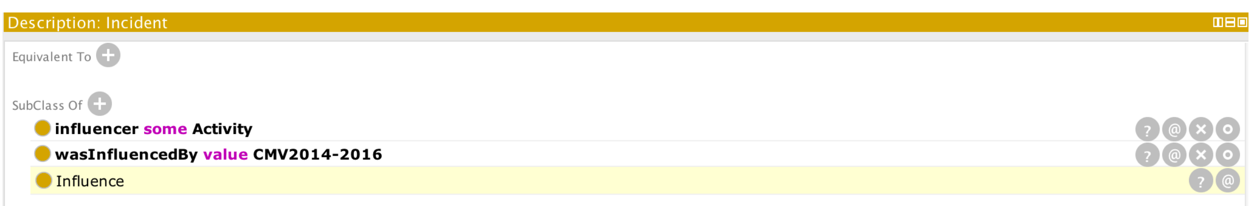

Figure 12. Asserted axiom that the **`Incident`** class has an influencer that is some **`Activity`**; and **`wasInfluencedBy CMV2014-2016`** (Data Source)

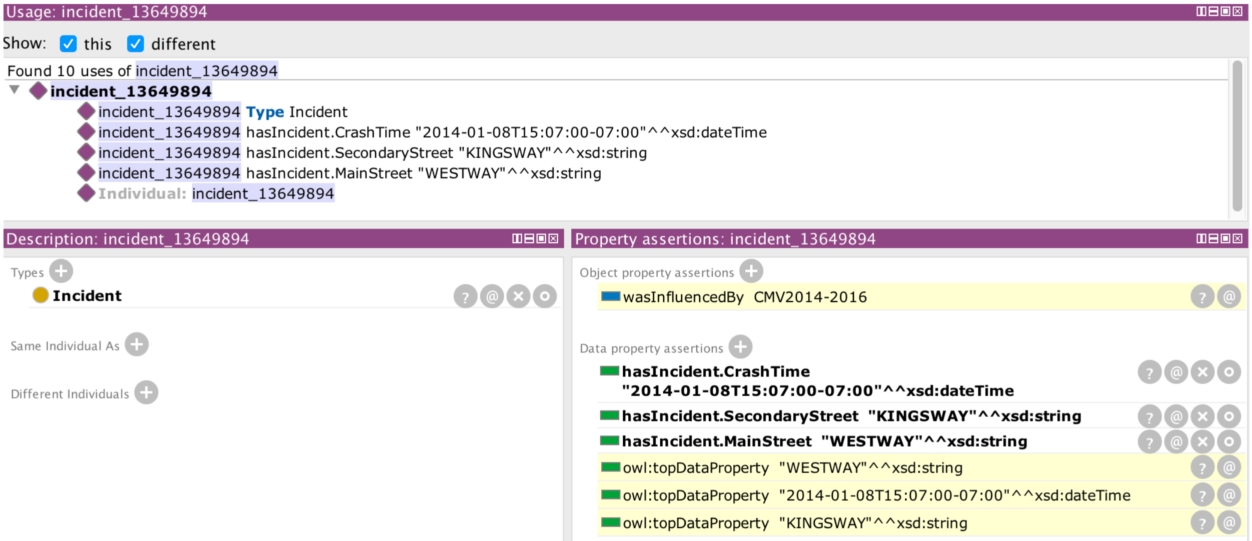

Figure 13. Example of an incident with an inferred object property **`wasInfluencedBy`** (from PROV-O

By introducing the provenance model all the data can be traced back to an original data source. It is asserted that the data source that is shown in each individual, is the actual data source for that respective individual.

```
SELECT ?predicate ?object
   WHERE {
     fp:Incident ?predicate ?object.
    }
```

Formula 10. SPARQL query that will retrieve all the information related to a specific ontology class.

Results:

```
-------------------------------------------------------------
| predicate        | object                                  |
=============================================================
| rdf:type         | owl:Class                               |
| rdfs:subClassOf  | prov:Influence                          |
| rdfs:subClassOf  | _:b0                                    |
| rdfs:subClassOf  | _:b1                                    |
| rdfs:comment     | "Each Incident (Accident) that occurs"  |
| rdfs:label       | "Incident"                              |
-------------------------------------------------------------
```

Figure 14. Properties related to the class `Incident`

Formula 10. is a query that retrieves all the relationships and properties from the relationships with respect the **`Incident`** class. The object shows that **`prov:Influence`** is a property of the class **`Incident`**; as a result, the class has an **`Influencer`**. Limitations to Jena Fuseki include not supporting queries to obtain the source of the **`influencer`** directly through SPARQL. However, listed in the appendix is a portion of the OWL file that shows the relationship between the **`Incident`** class and the data source **`Activity`** that is its **`influencer`**. By looking at the class **`Incident`** it can be seen that the entire class is asserted to be influenced by `CMV2014-2016`, and through the reasoner, all of the individual incidents are also influenced by the same data source.

5.1.4 Validation with domain experts.

Through a context-level ontology evaluation (Brank et al., 2005) the **`Freight Performance Ontology`** can be evaluated using user and domain experts requirements and indicators. Furthermore, potential users have been validated with 1) a transportation planner in

the Texas Department of Transportation El Paso District; 2) a senior planner at El Paso Metropolitan Planning Organization; and 3) a freight operator (Torres, 2016).

In addition, a set of factors which are important for freight transportation has been developed by a stakeholder. The metrics that they are most interested in have been considered by the ontology design and the queries requested. Although safety is the focus of this research, all the relevant categories have also been evaluated, based on the data available.

Furthermore, the stakeholder also determines that Port of Entry crossing time is of great importance. Web services have been developed that will automatically retrieve information from the Texas A&M University Transportation Institute ("Border Crossing Information System Commercial Vehicles," 2016) and make queries on that information. Both documents that the stakeholder has given are provided in the appendix.

5.2 Supplementary Evaluation

In addition to the individual year (2014) that has been evaluated using the competency questions, a larger year range has also been evaluated. These evaluations will show in detail how the expansion of additional data can provide insight to Freight Performance Analysis. Additional comparisons of the results for competency question 4 using a single year (2014) to 3 years (2014-2016): *How many accidents are there at I-10 at Hawkins interchange?* It is found that the results increase dramatically. Previously the total number of incidents were 2 (in 2014), and increase to 14 over the years of 2015-2016, thus showing that 12 additional incidents happened during the next two years.

Furthermore, the addition of 2 years of data provides a larger data set that can be analyzed by domain experts; a larger data set will likely give experts the ability to find patterns between incidents, weather conditions, and the locations of all their incidents.

By providing an increase of data it is shown that the ontology is able to handle over 100,000 triples – or relationships between individuals.

Chapter 6: Conclusion

This chapter will evaluate the goals and objectives with respect to the results obtained in this research.

6.1 Summary

The proposed ontology-based framework provides a foundation to understand heterogeneous data relevant to Smart Mobility in El Paso County. Freight performance assessment facilitates the understanding how the movement of freight trucks affect Safety, Traffic Congestion, Environmental Sustainability, and Mobility in this area. The proposed ontology framework addresses specific metrics for the U.S. – Mexico border in El Paso such as the average time waiting at the Port of Entry. However, this framework is generic and can be extended to integrate and aggregate data on a larger scale such as a state or country. This work can be easily transferred to integrate data across multiple domains (e.g. Smart Transit, Smart Buildings and Smart Healthcare) to increase the efficiency and productivity of analyzing and understanding domain specific data.

The **`Freight Performance Ontology`** increases the ability to access data in different formats, organize it and query it in a way that facilitates human decision making. As a result, this helps improve the productivity of government officials, freight companies, and most importantly the residents of cities. Not only can it be used in academic interdisciplinary research, it allows for practical uses in everyday large-scale data-driven systems.

6.2 CONTRIBUTIONS

This work illustrates how to integrate heterogeneous data from external factors, especially those that are not directly related to one another without domain expertise. Since weather data is not directly related to Freight Performance, the importance of developing generic transportation ontologies that can handle data from various unrelated sources to better understand Smart Mobility in a city is shown. Moreover, the information that has been gathered provided inferences between unlinked classes and related the information based on a single common variable being a date. The linking of data has provided new insight on the fundamental need to obtain reliable data and integrate it to find additional insight for analysis.

This work took an initial ontology that was specified for freight performance and extended it beyond the scope freight performance indicators. The external data that was incorporated into the ontology has provided additional representation of what may have been causing factors for various incidents within the city. By adding multiple years of data, a larger comparison set for domain experts to view and make analysis on is developed.

The approach that has been used to evaluate the completeness and success of the ontology was done using the same data that was driven to build it. This has shown that in fact the ontology has all the characteristics that was intended based on the specifications of the concept map and competency questions created by the domain experts. As a result of the data-driven evaluation it is shown that adding several years of information beyond a single year can only provide extensive understanding of what is occurring on the roads in the city of El Paso. All external unrelated information has provided domain experts the ability to move beyond general freight and mobility data and extend their search to other data sources.

This work has shown how incorporating data provenance can increase the understanding of the data. Moreover, the data that has a history can provide an insight to understand if it is trustworthy; through trustworthy data, domain experts are able to determine if the results that are retrieved from the queries are adequate and reliable. The provenance model added to the **Freight Performance Ontology** has also laid the foundation for integrating data-driven research in Smart Mobility by providing accuracy to results gathered. The advantage of adding provenance to an ontology is to ensure that the users will trust the data that is being returned to them. This cannot be done by a simple relational database, and is a key feature of using ontologies for ensuring trustworthy and reliable data sources.

Finally, the framework that has been developed by the implementation of the **Freight Performance Ontology** opens new platforms to continue adding data that it is populated with. Moreover, it extends applicable use for Smart Mobility to begin understanding the consumer motor vehicles and how their movement around the city compares or **influences/is influenced** by freight Vehicles. By providing a general framework, it is possible to understand problems and key areas where safety issues may be throughout the city of El Paso. By improving the understanding of what occurs on the roads of El Paso; experts, commuters, and government leaders will be able to interpret more closely the needs of its citizens. By providing a framework, the city of El Paso can improve its mobility indicators for all citizens, and promote a Smarter way of being mobile in the city.

Chapter 7: Future Work

By developing an ontology for Freight Performance, it can be extended to all vehicle movement in a city. The movement in a city can be evaluated on a larger scale to understand the city needs and incidents. In the future, this work can continue to be expanded beyond the three years of data that has been populated into the **`Freight Performance Ontology`**. In addition, this ontology will be further refined to integrate additional data required by stakeholders (e.g., gas prices) and mapping to other relevant ontologies for interoperability purposes.

Furthermore, this work will also be expanded by obtaining real-time data through either web-services or the Internet of Things (IoT). The understanding of a modern city that has the potential to begin transformation into a Smart City must begin producing its own data in real time. The data that will be produced in real-time will be stored and maintained as additional triples, so that it can be queried and immediately evaluated. Furthermore, as an extension of real-time systems, this work will be used as a basis to use as a predictive model with the incorporation of machine learning concepts.

The implementation of IoT systems will look towards using both sensor and user-centric data. This work will begin by placing sensors locally at The University of Texas at El Paso campus so that initial tests can be done to understand the movement of students. As research progresses, users will be able to provide relevant data to be used and analyzed. With the incorporation of user-centered data research will be done begin to introduce provenance models to track the history of data and determine its trust value. This future work aims to question understand several additional concepts that relate not only to freight performance, but Smart Mobility in general:

Q1. How can user-centric information be used to add analysis to mobility metrics?

Q2. How can provenance models be added to instill trust into user-centric information?

Q3. How can historical data be combined with real-time data to understand mobility and enable the production of a standard metric?

Q4. How can the Freight Performance Ontology be enhanced to incorporate additional unlinked data or concepts that may also provide insight to concepts outside of the freight performance domain (i.e. Consumer vehicles)?

Q5. How can data from Smart Buildings, Smart Environment, Smart Campus etc. be used to understand and obtain additional metrics for Smart Mobility?

Through the research that is being done on a local campus a framework will be developed to see how to best implement these services in a larger scale, the city of El Paso. By doing this future work can move from solely relying on data that is produced on external entities. By having access to the raw data, it can provide a means to take measurements specific to relevant areas of interest. This process will provide us the understanding of how generic models can be used beyond the city of El Paso and become implemented throughout cities throughout the United States. As society moves towards a technological advanced environment, new techniques intended to pursue techniques to gather and link heterogeneous data can be done; this goal will help us become closer to a Smart City that focuses on the efficiency, productivity, and sustainability of its resources. Most importantly, the future work will help improve the quality of life of all users of the city.

References

Anand, N., van Duin, R., & Tavasszy, L. (2014). Ontology-based multi-agent system for urban freight transportation. *International Journal of Urban Sciences*, *18*(2), 133–153. https://doi.org/10.1080/12265934.2014.920696

Apache. (2016). Fuseki: serving RDF data over HTTP.

Baader, F., Horrocks, I., & Sattler, U. (2001). Description Logics and the Semantic Web. *Image Rochester NY*, *22*(3), 1–7. https://doi.org/10.1007/978-3-540-32254-2_14

Board, T. R. (2011). *Performance Measures for Freight Transportation*. (N. C. F. R. Program, Ed.). United States Department of Transportation.

Border Crossing Information System Commercial Vehicles. (2016). Retrieved April 1, 2016, from http://bcis.tamu.edu/Commercial/en-US/queryArchivedData.aspx

Brank, J., Grobelnik, M., & Mladenić, D. (2005). A survey of ontology evaluation techniques. *Proceedings of the Conference on Data Mining and Data Warehouses*, 166–170. https://doi.org/10.1.1.101.4788

Buccella, A., Cechich, A., & Rodriguez Brisaboa, N. (2003). An ontology approach to data integration. *Journal of Computer Science & Technology*, *3*(2), 62–68.

Bureau of Transportation Statistics. (2015). *Freight Facts and Figures*. Washington D.C.

Chandrasekaran, B., Josephson, J. R., & Benjamins, V. R. (1999). What are ontologies, and why do we need them? *IEEE Intelligent Systems*, *14*(1), 20–26. https://doi.org/10.1109/5254.747902

Chourabi, H., Nam, T., Walker, S., Gil-Garcia, J. R., Mellouli, S., Nahon, K., … Scholl, H. J. (2012). Understanding Smart Cities: An Integrative Framework. In *2012 45th Hawaii International Conference on System Sciences* (pp. 2289–2297). IEEE.

https://doi.org/10.1109/HICSS.2012.615

CRIS. (2015). *CMV 2014-2016 incidents*. Texas Department of Transportation.

CRRMA. (2014). *Fiscal Year 2014 Financial Report*.

Cruz, I. F., & Xiao, H. (2005). The Role of Ontologies in Data Integration. *Journal of Engineering Intelligent Systems*, *13*(4), 1–18. https://doi.org/10.1.1.60.4933

Cyganiak, R., Wood, D., & Lanthaler, M. (2014). RDF 1.1 Concepts and Abstract Syntax. *W3C Recommendation 25 February 2014*, (February), 263–270. https://doi.org/10.1007/s13398-014-0173-7.2

Dale, L., Michael, D., Laurie, J., Goodhue, B. D. L., Wybo, M. D., G, C. H. a, & Kirsch, L. J. (1992). The Impact of Data Integration on the Costs and Benefits of The Impact of Data Integration on the Costs and Benefits of Information Systems Introduction. *Misq*, *16*(September), 293–311. https://doi.org/10.2307/249530

Dameri, R. P. (2013). Searching for Smart City definition: a comprehensive proposal. *International Journal of Computers & Technology*, *11*(5), 2544–2551. https://doi.org/10.1007/s13132-012-0084-9

Department of Import and Export Control. (2015). Annual Performance Report 2015, 1–20. Retrieved from https://www.parliament.lk/.../performance-report-department-of-imports-and-exports-...%5CnCached

Dibowski, H., & Kabitzsch, K. (2010). Ontology-based Device Descriptions and triple store based device repository for automation devices. In *2010 IEEE 15th Conference on Emerging Technologies & Factory Automation (ETFA 2010)* (pp. 1–9). IEEE. https://doi.org/10.1109/ETFA.2010.5641257

Gardner, S. (2005). Ontologies and semantic data integration. *Drug Discovery Today*,

10(60761002), 2440–2443. Retrieved from http://www.sciencedirect.com/science/article/pii/S135964460503504X

Guarino, N. (1998). Formal Ontology and Information Systems. *Information and Software Technology*, *98*(1998), 81–97. https://doi.org/10.1.1.29.1776

Harris, S., Seaborne, A., Apache, T., Foundation, S., & Prud, E. (2013). SPARQL 1.1 Query Language, *11*(March), 1–98. https://doi.org/citeulike-article-id:2620569

Heimbigner, D., & Mcleod, D. (1985). A Federated Architecture for Information Management. *ACM Transactions on Office Information Systems*, *3*(3), 253–278. https://doi.org/10.1145/4229.4233

Horridge, M., & Bechhofer, S. (2011). The OWLAPI: A Java API for OWL ontologies. *Semantic Web*, *2*(1), 11–21. https://doi.org/10.3233/SW-2011-0025

Information Systems Group Oxford University. (2016). HermiT OWL Reasoner The New Kid on the OWL Block. Retrieved April 1, 2016, from http://www.hermit-reasoner.com

Lebo, T., Sahoo, S., & McGuinness, D. (2013). PROV-O: The PROV Ontology. *W3C Recommendation*, (April), 1–80. Retrieved from http://www.w3.org/TR/2013/REC-prov-o-20130430/

Liao, C.-F. (2009). Using Archived Truck GPS Data for Freight Performance Analysis on I-94/I-90 from the Twin Cities to Chicago.

Lohmann, S., Negru, S., Haag, F., & Ertl, T. (2016). Visualizing ontologies with VOWL. *Semantic Web*, *7*(4), 399–419. https://doi.org/10.3233/SW-150200

Lowagie, B. (2010). *iText in Action* (2nd ed.). Manning Publications.

Ma, X. (2011). Processing Commercial GPS Data to Develop a Web- Based Truck Performance Measures Program, (206). https://doi.org/10.1.1.700.9692

Martinez-Cruz, C., Blanco, I. J., & Vila, M. A. (2012). Ontologies versus relational databases: Are they so different? A comparison. *Artificial Intelligence Review*, *38*(4), 271–290. https://doi.org/10.1007/s10462-011-9251-9

McCormack, E., & Hallenbeck, M. (2006). ITS Devices Used to Collect Truck Data for Performance Benchmarks. *Transportation Research Record*, *1957*(1), 43–50. https://doi.org/10.3141/1957-07

Mejia, D., Camacho, E., Caballero, J., & Chowdhurry, M. (2016). Cyberinfrastructure Applications Final Report. El Paso.

Mejia, D., Chowdhury, M. P., Caballero, J., Camacho, E., Torres, E., & Villanueva-Rosales, N. (2016). *Smart Mobility: Ontology-Based Data Integration for Freight Performance Analysis*.

Mejia, D., Torres, E., Villanueva-Rosales, N., & Cheu, K. (2017). Integrating Heterogeneous Freight Performance Data for Smart Mobility. *Unpublished*.

Mnif, S., Galoui, S., Elkosantini, S., Darmoul, S., & Ben Said, L. (2015). Ontology based performance evaluation of public transport systems. *2015 4th International Conference on Advanced Logistics and Transport (ICALT)*, 205–210. https://doi.org/10.1109/ICAdLT.2015.7136622

Moreau, L., & Groth, P. (2013). PROV-Overview: An Overview of the PROV Family of Documents. *W3C Note*, (April), 1–9. Retrieved from http://www.w3.org/TR/2013/NOTE-prov-overview-20130430/

Musen, M. A. (2015). The protégé project. *AI Matters*, *1*(4), 4–12. https://doi.org/10.1145/2757001.2757003

Nam, T., & Pardo, T. A. (2011). Smart city as urban innovation: Focusing on management,

policy, and context. In *5th International Conference on Theory and Practice of Electronic Governance (ICEGOV2011)*. https://doi.org/10.1145/2072069.2072100

Naphade, M., Banavar, G., Harrison, C., Paraszczak, J., & Morris, R. (2011). Smarter Cities and Their Innovation Challenges. *Computer*, *44*(6), 32–39. https://doi.org/10.1109/MC.2011.187

NCEI. (2017). *WeatherData*. National Oceanic and Atmospheric Administration.

Noy, N. F., & McGuinness, D. L. (2001). Ontology Development 101: A Guide to Creating Your First Ontology. *Stanford Knowledge Systems Laboratory*, 25. https://doi.org/10.1016/j.artmed.2004.01.014

O'Rourke, L., Beshers, E., & Stock, D. (2015). *Measuring the Impacts of Freight Transportation Improvements on the Economy and Competitiveness September 2015*.

Operations Performance Measurement Program. (2017). Retrieved from https://ops.fhwa.dot.gov/perf_measurement/example_programs.htm

Orendain, D. O., Neri, L., & Bechelani, D. (2015). The role of the Data Scientist within Smart Cities. *IEEE-GDL CCD Smart Cities*, 1–8.

Pennington, D. D. (2010). The dynamics of material artifacts in collaborative research teams. *Computer Supported Cooperative Work*, *19*(2), 175–199. https://doi.org/10.1007/s10606-010-9108-9

Priano, F. H., & Guerra, C. F. (2016). Fully smart cities: Analysis of the city to Smart City transformation process. In *2016 IEEE International Smart Cities Conference (ISC2)* (pp. 1–8). IEEE. https://doi.org/10.1109/ISC2.2016.7580745

Ramani, T., Potter, J., DeFlorio, J., Zietsman, J., & Reeder, V. (2011). *A Guidebook for Sustainability Performance Measurement for Transportation Agencies*.

https://doi.org/10.17226/14598

Rohloff, Kurt and Dean, Mike and Emmons, Ian and Ryder, Dorene and Sumner, J. (2007). An Evaluation of Triple-Store Technologies for Large Data Stores. *On the Move to Meaningful Internet Systems 2007: OTM 2007 Workshops*, *4806*, 1105–1114. https://doi.org/10.1007/978-3-540-76890-6_38

Shelton, M. (2016). *Keep Customers and Ourselves Safe*. *Measures of Departemental Performance, MoDOT*.

Sheth, A. (1999). Interoperating Geographic Information Systems. *Interoperating Geographic Information Systems*, 5–30. https://doi.org/10.1007/978-1-4615-5189-8

Torres, E. J. (2016). *Ontology-Driven Integration of Data for Freight Performance Measures*. The University of Texas at El Paso.

Villanueva-Rosales, N., & Dumontier, M. (2008). yOWL: An ontology-driven knowledge base for yeast biologists. *Journal of Biomedical Informatics*, *41*(5), 779–789. https://doi.org/10.1016/j.jbi.2008.05.001

Vrande, D., Studer, R., Koreferent, E., Hendler, J. A., Koreferent, Z., Vorsitzende, C. W., & Werner, U. (2010). Ontology Evaluation. *Handbook on Ontologies*, *16*(June), 251–274. https://doi.org/10.1007/978-3-540-92673-3

W3C. (2012). Web Ontology Language (OWL). Retrieved from https://www.w3.org/OWL/

Appendix

APPENDIX A: FREIGHT PERFORMANCE CONCEPT MAP

The Freight Performance concept map is based on the ideas of a civil engineer, a domain expert in Freight Performance. The focus of this work is to establish a clearer understanding of the relationships between classes and external data sources, especially **`Daily Weather Metrics`**. This concept map can be found at (Torres, 2016).

APPENDIX B: CODE SNIPPET

```
//Named individual
OWLNamedIndividual weatherIncidentIndividual = factory.getOWLNamedIndividual(IRI.create("weather_incident_"+allIncidents.get(i).Crash_ID));
OWLClassAssertionAxiom WeatherIncidentClassAssertion = factory.getOWLClassAssertionAxiom(WeatherRelatedIncidentClass, weatherIncidentIndividual);
manager.applyChange(new AddAxiom(ontology,WeatherIncidentClassAssertion));

//Gives the data properties and adds them to the ontology
OWLAxiom hasMainStreetAxiom = factory.getOWLDataPropertyAssertionAxiom(hasIncidentMainStreet, weatherIncidentIndividual, allIncidents.get(i).Rpt_Street_Name);
manager.applyChange(new AddAxiom(ontology, hasMainStreetAxiom));

OWLAxiom hasSecStreetAxiom = factory.getOWLDataPropertyAssertionAxiom(hasIncidentSecondaryStreet, weatherIncidentIndividual, allIncidents.get(i).Rpt_Sec_Street_Name);
manager.applyChange(new AddAxiom(ontology, hasSecStreetAxiom));

OWLNamedIndividual incidentResponseIndividual = factory.getOWLNamedIndividual(IRI.create("weather_incidentResponseTime_"+allIncidents.get(i).Crash_ID));
OWLClassAssertionAxiom incidentResponseClassAssertion = factory.getOWLClassAssertionAxiom(ResponseTimeToIncidentsClass, incidentResponseIndividual);
manager.applyChange(new AddAxiom(ontology,incidentResponseClassAssertion));

OWLLiteral dateLiteral = factory.getOWLTypedLiteral(dateString, OWL2Datatype.XSD_DATE_TIME);
OWLAxiom hasCrashTimeAxiom = factory.getOWLDataPropertyAssertionAxiom(hasIncidentCrashTime, weatherIncidentIndividual, dateLiteral);
manager.applyChange(new AddAxiom(ontology, hasCrashTimeAxiom));

OWLLiteral notifyLiteral = factory.getOWLTypedLiteral(notifyString, OWL2Datatype.XSD_DATE_TIME);
OWLAxiom responseNotifyAxiom = factory.getOWLDataPropertyAssertionAxiom(hasResponseTimeToIncidentsNotifyTime, incidentResponseIndividual, notifyLiteral);
manager.applyChange(new AddAxiom(ontology, responseNotifyAxiom));

OWLLiteral arriveLiteral = factory.getOWLTypedLiteral(arriveString, OWL2Datatype.XSD_DATE_TIME);
OWLAxiom responseArriveAxiom = factory.getOWLDataPropertyAssertionAxiom(hasResponseTimeToIncidentsArrivalTime, incidentResponseIndividual, arriveLiteral);
manager.applyChange(new AddAxiom(ontology, responseArriveAxiom));
```

APPENDIX C: INCIDENT PROVENANCE

```xml
<!-- http://ontology.cybershare.utep.edu/smart-cities/FreightPerformance/#Incident -->

<owl:Class rdf:about="http://ontology.cybershare.utep.edu/smart-cities/FreightPerformance/#Incident">
    <rdfs:subClassOf>
        <owl:Restriction>
            <owl:onProperty rdf:resource="http://www.w3.org/ns/prov#influencer"/>
            <owl:someValuesFrom rdf:resource="http://www.w3.org/ns/prov#Activity"/>
        </owl:Restriction>
    </rdfs:subClassOf>
    <rdfs:subClassOf>
        <owl:Restriction>
            <owl:onProperty rdf:resource="http://www.w3.org/ns/prov#wasInfluencedBy"/>
            <owl:hasValue rdf:resource="http://ontology.cybershare.utep.edu/smart-cities/FreightPerformance/CMV2014-2016"/>
        </owl:Restriction>
    </rdfs:subClassOf>
    <rdfs:comment rdf:datatype="http://www.w3.org/2001/XMLSchema#string">Each Incident (Accident) that occurs</rdfs:comment>
    <rdfs:label rdf:datatype="http://www.w3.org/2001/XMLSchema#string">Incident</rdfs:label>
</owl:Class>
```

APPENDIX D: PERMISSION OF USE

The following correspondence represents approval to use parts of a submitted paper in this thesis.

From: **M.E. Brennan** me.brennan@ieee.org
Subject: Re: Permission to Use Document for Thesis
Date: April 25, 2017 at 9:33 AM
To: Mejia, Daniel M dmmejia2@miners.utep.edu

MB

Hi Daniel,

I have two responses for you. If you have to submit before the paper is accepted you would need to cite as unpublished:

J. K. Author, "Title of paper," unpublished.

Once the paper has been accepted and you have signed the copyright transfer agreement, do as described below:

The IEEE does not require individuals working on a thesis to obtain a formal reuse license however, you must follow the requirements listed below:

Textual Material

Using short quotes or referring to the work within these papers) users must give full credit to the original

source (author, paper, publication) followed by the IEEE copyright line © [Year of publication] IEEE.

In the case of illustrations or tabular material, we require that the copyright line © [Year of original

publication] IEEE appear prominently with each reprinted figure and/or table.

If a substantial portion of the original paper is to be used, and if you are not the senior author, also obtain

the senior author's approval.

Full-Text Article

If you are using the entire IEEE copyright owned article, the following IEEE copyright/ credit notice should be placed prominently in the references: © [year of original publication] IEEE. Reprinted, with

permission, from [author names, paper title, IEEE publication title, and month/year of publication]

Only the **accepted** version of an IEEE copyrighted paper can be used when posting the paper or your thesis on-line. You may not use the **final published** version

In placing the thesis on the author's university website, please display the following message in a

prominent place on the website: In reference to IEEE copyrighted material which is used with permission

in this thesis, the IEEE does not endorse any of [university/educational entity's name goes here]'s

In this thesis, the IEEE does not endorse any of [university/educational entity's name goes here]'s

products or services. Internal or personal use of this material is permitted. If interested in

reprinting/republishing IEEE copyrighted material for advertising or promotional purposes or for creating

new collective works for resale or redistribution, please go to

http://www.ieee.org/publications_standards/publications/rights/rights_link.html to learn how to obtain

a License from RightsLink.

If applicable, University Microfilms and/or ProQuest Library, or the Archives of Canada may supply single copies of the dissertation.

Kind regards,

M.E. Brennan

Ms M.E. Brennan
IEEE
501 Hoes Lane
Piscataway, NJ 08854-4141 USA
me.brennan@ieee.org

+1 (732) 562-2660

Vita

Daniel Mejia was born in in El Paso, Texas where he was born and raised. He is the youngest of three children and graduated from Andress High School in 2012 with honors. He began attending UTEP in August 2012 where he graduated *cum laude* with a Bachelor of Science degree in Computer Science in May 2015.

Daniel Mejia participated in the U.S. – Mexico Study abroad program focusing on Smart Cities in the summer of 2016. While participating in this program he developed a large interest in understanding how Computer Science can play a larger role in understanding and transforming Smart Cities. It was on this program where he discovered his desire to do research in data integration and Smart Cities. He began doing research with Dr. Natalia Villanueva-Rosales in August 2016.

Daniel has submitted a co-authored paper to the 2017 IEEE Conference on Smart City innovations where his work is under review at time of this thesis publication. During his time as a student, Daniel has been a guest lecturer in several classes discussing various Computer Science topics.

Contact Information: dmmejia2@gmail.com

This thesis/dissertation was typed by Daniel Mejia.