A Data Integration Framework for Urban Systems Analysis Based on Geo-Relationship Learning

Zheng Yang[1]; Karan Gupta[2]; Archana Gupta[3]; and Rishee K. Jain[4]

[1]Urban Informatics Lab, Dept. of Civil and Environmental Engineering, Stanford Univ., 473 Via Ortega, Stanford, CA 94305.
[2]Urban Informatics Lab, Dept. of Civil and Environmental Engineering, Stanford Univ., 473 Via Ortega, Stanford, CA 94305.
[3]City of Palo Alto, 250 Hamilton Ave., Palo Alto, CA 94301.
[4]Urban Informatics Lab, Dept. of Civil and Environmental Engineering, Stanford Univ., 473 Via Ortega, Stanford, CA 94305.

Abstract

The world is rapidly urbanizing, and for the first time in history over 50% of the world's population reside in urban areas. This rapid urbanization brings about tremendous challenges at the intersection of governance, infrastructure and the environment. Advanced sensing and data analytics techniques have been developed in the context of so called "smart cities" with the goal of providing insights on how urban systems could be designed and managed more effectively. However, the proliferation of data from heterogeneous sources makes interoperability and mining of such urban data streams difficult. Facilitating the extraction of insights that support data-informed policymaking and program recommendations will require frameworks to integrate such heterogeneous data streams. In this paper, we introduce a novel data integration framework that utilizes a RDF (resource description framework) model to integrate disparate urban data streams based on geo-relationships that are iteratively learned from semantic information and the structure of relational databases. The development of our framework was driven by interviews and observations of city officials responsible for managing and integrating urban data and a review of the various types of disparate datasets generated from sources like departmental databases, sensors, and crowdsourcing. Finally, we apply our proposed framework to an urban data scenario in order to demonstrate the applicability and usefulness of the framework.

1. INTRODUCTION

The world is rapidly urbanizing. Over 50% of world population now reside in cities and the number is expected to increase to 67% by 2050 (United Nations, 2014). Continuous growth of urban population increases the demands and consumption of limited resources, resulting in numerous challenges for city officials in respect to urban systems management and decision-making. Rapid development of new sensing technologies has led to an explosion of data streams in numerous urban domains (e.g. land, building, energy, and human). However, harnessing such data and translating it into insights for more effective and sustainable urban system management remain a major challenge. Doing so requires collecting, processing, integrating, sharing, and analyzing data that spans various urban domains and organizational departments within a municipal government. For the most part, city officials utilize data in an isolated manner as no framework exists for them to easily integrate and analyze data from heterogeneous sources. Most city officials rely on a manual process for data integration that is both time- and labor- consuming and lacks scalability as more and more data becomes available. Moreover, many urban datasets lack common features or keys to be directly linked with other datasets. This data integration

problem has prevented city officials from leveraging such data to enhance decision-making and improve urban systems management.

In this paper, we introduce a novel data integration framework that utilizes a RDF (Resource Description Framework) model to integrate disparate urban data streams based on geo-relationships that are iteratively learned from semantic information and the structure of relational databases. The goal of this framework is to provide a first-step towards enabling more data-driven analysis and management of urban systems. This paper is structured as follows: Section 2 describes the current city practices in respect to data management and integration based on interviews and the observations of our embedded research team at the City of Palo Alto, USA; Section 3 summarizes the common datasets from urban systems, and investigates the existing methods for urban data integration; Section 4 proposes a novel framework for defining metadata schema based on geo-relationship learning and demonstrates its merits on a simple example; Section 5 provides conclusions and avenues for future work.

2. CURRENT DATA INTEGRATION & MANAGEMENT PRATICES IN CITIES

Municipal governments are increasingly making their data publicly available in order to encourage citizen engagement and academic research on urban systems management. In order to understand the current practices of how data integration is conducted in cities, we interviewed several city officials responsible for managing urban data and utilized the observations of our embedded and integrated research team at the City of Palo Alto, USA.

We found that the current policies of data integration are generally proclaimed as a set of guidelines, executive orders and legislations, which establish basic governing expectations and rules and ensure an overall buy-in from various sources. The lifecycle of data integration begins at identification of data and ends with final publishing of the linked datasets. A robust data integration process is critical in ensuring the datasets adheres to the policy throughout its lifecycle. The *identification* step begins with either an external request for adding new data or the team responsible for managing the urban data program identifies the programmatic needs for such addition. *Evaluation* of such the identification involves rudimentary checks such as availability of data and whether it is outside the purview of U.S. data protection. The next step is to *ensure the quality* of data depending on the specific requirements of different departments. Departments have ad-hoc technologies that are built to suit their own needs that are not necessarily designed with a holistic citywide purview. As a result, this leads to heterogeneous data with different levels of granularity that exist in departmental silos. The qualified datasets are then *evaluated for legal, privacy or security issues*. This evaluation is vital to protect the legal and fiduciary responsibilities of city officials. Findings of this evaluation step are useful to improve and update current data policies. For example, integrating latitude/longitude of a location with its address was not allowed for public security reasons. However, with advances in data analytics, this poses a new challenge as latitude/longitude information can be extrapolated to an actual address. The final step in this lifecycle is to *publish* data in an internal system or open data portal.

Generally, for city departments responsible for managing different urban systems, sharing and linking data is still in its infancy. The current practices require the same formats of data across heterogeneous sources and face integration barriers if there are no shared features defined by

prerequisite standardization. As a result, there is no scalable and practical method that can be used by city officials to represent and analyze interconnected and interdependent urban data streams.

3. URBAN DATA & INTEGRATION METHODS

We conducted a literature review and onsite observations at the City of Palo Alto to identify and categorize possible datasets from heterogeneous urban sources. Examples of common urban systems and relevant data types, sources and characteristics are presented in Table 1. It can be seen that an unprecedented amount of static and dynamic data are already being generated from departmental databases, sensors, and crowdsourcing. These emerging data are generally dispersed and unstructured (Yuan et al., 2012; Zielstra et al., 2013; Khan et al., 2013; Balaji et al., 2016), demonstrating the need for an interpretable and scalable urban data integration framework.

Table 1. Data Examples Available for Urban System Management.

Urban System	*Data Type (e.g.)*	*Data Source (e.g.)*	*Characteristics*
Building	Architecture, System, Occupancy	Databases (e.g. BIM), Sensors (e.g., PIR), Crowdsourcing (e.g. Smartphone)	Static + Dynamic
Land	Geometry, Use	Databases (e.g., GIS), Sensors (e.g. RFID), Crowdsourcing (e.g. Social Media)	Static + Dynamic
Road	Geometry, Network	Databases (e.g. CAD)	Static
Transportation	Vehicle, Traffic, Parking	Databases (e.g. RDB), Sensors (e.g. Camera), Crowdsourcing (e.g. GPS)	Static + Dynamic
Vegetation	Property, Geospatial data	Databases (e.g. open data portal)	Static
Utility	Water, Electricity, Gas	Databases (e.g. SaaS), Sensors (e.g. Smart Meter), Crowdsourcing (e.g. IoT)	Static + Dynamic
Environment	Weather, Atmosphere	Database (e.g. record), Sensors (e.g. particles detector), Crowdsourcing (e.g. portable devices)	Static + Dynamic

In general, we found that data from urban systems have different representations and semantics and thus pose challenges for data integration. Organizations like OGC (Open Geospatial Consortium) have published standards for data interoperability such as CityGML (OGC, 2012), which are only available for integrating homogeneous data with the similar formats. Additionally, methods such as the information network have been utilized to facilitate the abstraction and connections of data in a specific domain (e.g., one type of urban system) but are limited in their ability to integrate data across different domains (Sun and Han, 2012). Relational databases with consistent schema could address the semantic disparity issue (Ziegler and Dittrich, 2004) but fail to integrate semi-structured or unstructured data that are common in the urban context. Automated systems have been then developed to convert tabular data to semantic web representations or object-oriented models by identifying and analyzing structure, content, and semantic attributes of local databases (Han et al., 2008; Venetis et al., 2011). However, such methods are limited in their ability to handle dynamic data streams (e.g., sensor data) and lack the

interpretability and operability necessary for augmenting decision-making of city officials. Other studies have tried to integrate fixed data, sensor data, and live social media data but such methods require manual efforts or well-defined schema that are highly domain-specific (Lopez et al., 2012; Bocconi et al., 2015). Thus far there is no complete ontology specifically designed for urban systems management (Zhu and Ferreira Jr, 2015). Some efforts have been made on methods to directly integrate knowledge behind data instead of the data itself (Zheng, 2016) but they are task-specific and not generalizable for supporting the wide array of applications required in urban system management. In summary, currently there is a lack of an interpretable and scalable data integration framework for data from urban systems (Sheridan and Tennison, 2010), especially when there are no shared features among datasets. A context-rich metadata schema to describe the entities, attributes, and relationships of data collected from heterogeneous sources is required such that smart city applications like energy optimization and transportation management can be easily implemented across a wide range of cities.

4. GEO-RELATIONSHIP LEARNING BASED METADATA SCHEMA

In order to address the challenges discussed above, we developed a data integration framework that derives a meta-data schema based on geo-relationship learning. A geo-relationship is defined as one kind of the ontology which defines the semantics, structures and representations of concepts in a particular domain. It has the potential to guide the integration of data from different sources. RDF (Resource Description Framework) (W3C, 2014) model is a data architectural model by World Wide Web Consortium to represent metadata schema for integrating data from different sources based on domain ontology. Entities and relationships in the RDF model are explicitly expressed in a human-logic-similar manner so that city officials can easily follow the schema and map their data from heterogeneous urban systems to the ontology for comprehensive understanding and systematic analysis. For example, it is possible to use a RDF model to define "buildings" in class A as an equivalent to "energy efficient infrastructure", which is difficult in traditional relational databases. As a result, we chose to utilize the RDF model as the core basis of our framework to represent entities, attributes, and relationships of data by *triples <subject-predicate-object>*. Each *triple* indicates the entity *subject* has the relationship *predicate* to another entity *object*. A RDF model is essentially a graph formed by nodes (*subjects* and *objects*) and directed edges (*predicates*). The simple structure of *triples* makes it possible to represent large interconnected urban data in an expressive way, especially when the structures of data are unknown or changing. Our framework was developed to define the metadata schema by identifying the geo-relationship between two different entities (i.e., adjacency). The learning of a geo-relationship follows the method of transitioning input from relational databases to a RDB model using RDBToOnto (Cerbah, 2008a), by which the strong expressive power of sematic web formalisms capture the underlying connections and hierarchy of geo-information. First the data are entered into a relational database, and each table is used as the source of an entity representing one urban system in the RDF model. The latitude/longitude of each entity is used to search the hierarchical distances with other entities. All distances are classified by the type of entities and represented as new features (e.g., located in or adjacent to) representing the geo-relationships among tables. After normalization of the new tables, key-based associations of tables then form the *predicate* of entities. Local constraints and dependencies are added to refine the populated metadata schema with predefined names assigned to instances. Entities are further categorized into subclasses by mining the patterns of relation attributes to reveal the additional structure hierarchy (Cerbah, 2008b).

A typical urban scenario is provided to demonstrate the typical procedure of implementing our framework and subsequently defining metadata schema. Figure 1 shows the original datasets for four urban systems within an urban area (Figure 2a), including roads, buildings, vegetation and in-situ sensors.

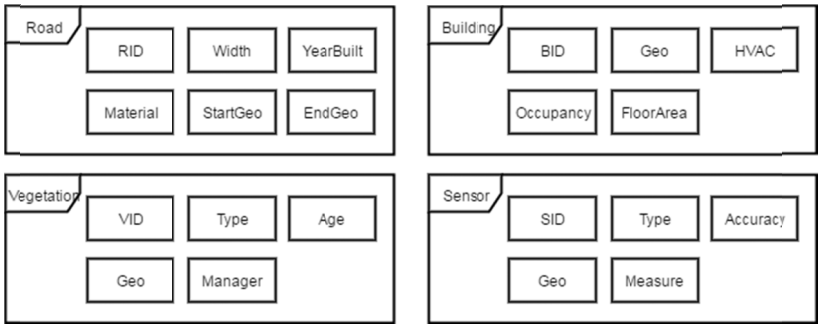

Figure 1. Original datasets for four urban systems ('Geo' represents latitude and longitude information).

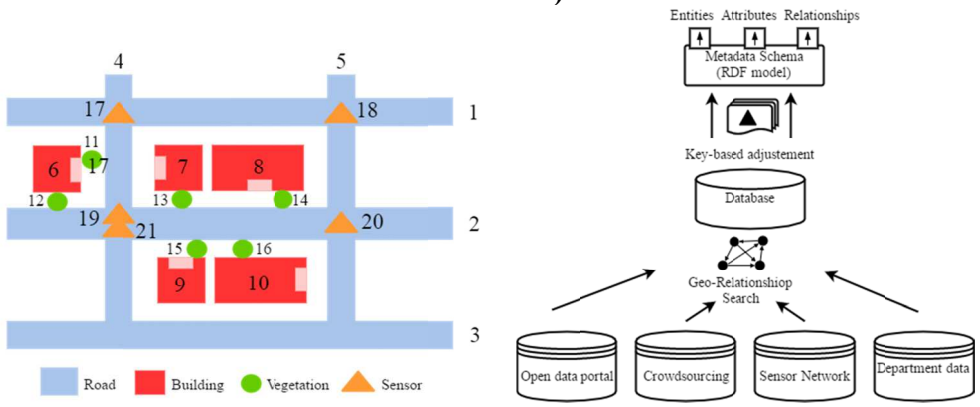

Figure 2a. Prototype for urban scenario. 2b. Process of framework implementation.

There are five roads in two directions. Five buildings are located on the lands encompassed by the roads with pink rectangles as the main entrances. Vegetation are planted between buildings and roads. In-situ sensors are installed at the intersections of two roads for periodically monitoring the temperature and counting pedestrians, and in the BMS (Building Management System) inside one building for recording building system energy use. XID is the globally unique identifier of each entity for the urban system X (X=road, building, vegetation, or sensor). The framework is implemented to the scenario, including three steps (Figure 2b): 1) search the possible geo-relationships using the hierarchical distances calculated by the latitude/latitude of each entity. The results are summarized in Figure 3a with the lines representing the geo-relationships found among different entities; 2) normalize the database and adjust the structure of datasets based on local constraints, dependencies and classification; 3) form metadata schema with entities, attributes, and relationships using *triples*.

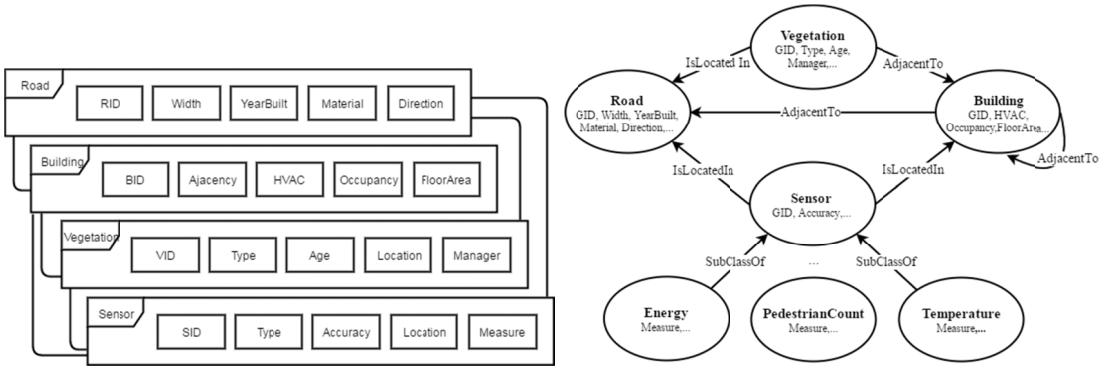

Figure 3a. Datasets with learned geo-relationships. Figure 3b. Output of metadata schema based on geo-relationship learning.

The results in Figure 3b show three types of *triples* (IsLocatedIn, AdjacentTo, SubClassOf) that are generated in order to integrate the data from the four urban systems. Dynamic and static data are equally processed with data representations, structures and semantics captured. The conversion from tables to a RDF model allows the categorization of sensor types and tagging of specific properties to each. Such categorization and tagging are generally not available through a traditional relational database. As a result, the output metadata schema captures all the information in Figure 2a and the interlinked relationships between various datasets. This ability is not influenced by different urban scenarios. For example, the framework can adapt to the change of "retrofitting two buildings", re-calculate hierarchical distances, and update the geo-relationships between buildings and other entities. Changes made to the data structure or the addition of new datasets as they become available can be easily integrated into the schema and thus allow it to be dynamic in nature.

Compared to a relational database, this graph schema facilitates easy analysis of coupled spatial-temporal dynamics between urban systems and easy interpretation of the results. For example, using the RDF model, queries of retrieving data under predefined conditions such as "select energy consumption data of one building when there is no pedestrian on its front street and the adjacent building is not occupied" can are easily implemented using common querying languages like SPARQL (W3C, 2008).

5. CONCLUSIONS & FUTURE WORK

With rise of low-cost in-situ sensors and the "smart city" movement, a tremendous amount of urban data are being generated each day. In order to translate such data into insights and augment decision-making, city officials must analyze disparate datasets from a wide variety of heterogeneous sources. In this paper, we introduced a novel data integration framework that utilizes a RDF (Resource Description Framework) model to integrate disparate urban data streams based on geo-relationships. Our framework is grounded in observational analysis of current data management and integration practices occurring in cities.

The primary contribution of this work is the development of an easily interpretable and applicable metadata schema that enables city officials to analyze disparate, heterogeneous and isolated urban data. The proposed framework is capable of continuously updating the metadata schema by learning changes in geo-relationships, allowing it to be dynamic in nature and reflect changing

data availability and conditions in cities. Overall, this work represents a crucial first-step towards enabling more data-driven analysis and management of urban systems. Future work aims to formalize the process of metadata schema formulation and test our framework on real data from a "living lab" in downtown Palo Alto, CA, USA.

AKNOWLEDGEMENTS

The authors would like to thank Zhaoxin Fu (MS student, Stanford University) and Jonathan Reichental (Chief Information Officer, City of Palo Alto) for their help in the development of this work. The material presented is based in part upon work supported by the US National Science Foundation under Grant No. 1642315. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the National Science Foundation.

REFERENCES

Balaji, B., Bhattacharya, A., Fierro, G., Gao, J., Gluck, J., Hong, D., Johansen, A., Koh, J., Ploennigs, J., Agarwal, Y., Berges, M., Culler, D., Gupta, R., Kjærgaard, M. B., Srivastava, M., and Whitehouse, K. (2016). "Brick: Towards a unified metadata schema for buildings." *Proceedings of the ACM International Conference on Embedded Systems for Energy-Efficient Built Environments*. ACM, Stanford, CA, 41-50.

Bocconi, S., Bozzon, A., Psyllidis, A., Bolivar, C. T., and Houben, G. J. (2015). "Social glass: A platform for urban analytics and decision-making through heterogeneous social data." *Proceedings of the 24th International Conference on World Wide Web*. Florence, Italy. 175-178.

Cerbah, F. (2008a). "Learning highly structured semantic repositories from relational databases." *Proceedings of the European Semantic Web Conference*. Tenerife, Spain. 777-781.

Cerbah, F. (2008b). "Mining the content of relational databases to learn ontologies with deeper taxonomies." *Proceedings of the IEEE/WIC/ACM International Conference on Web Intelligence and Intelligent Agent Technology*. Sydney, Australia. 553-537.

Han, L., Fini, T., Parr, C., Sachs, J., and Joshi, A. (2008). "RDF123: from Spreadsheets to RDF." *Proceedings of the International Semantic Web Conference*. Karlsruhe, Germany. 451-466.

Khan, Z., Anjum, A., and Kiani, S. L. (2013). "Cloud based big data analytics for smart future cities." *Proceedings of the IEEE/ACM 6th international conference on utility and cloud computing. IEEE Computer Society*. Dresden, Germany, 381-386.

Lopez, V., Kotoulas, S., Sbodio, M. L., Stephenson, M., Gkoulalas-Divanis, A., and Aonghusa, P. M. (2012). "QuerioCity: A linked data platform for urban information management." *Proceedings of the International Semantic Web Conference*. Boston, MA. 148-163.

OGC (2012). "Virtual 3D city models." <http://www.citygml.org/> (Apr. 24, 2012).

Sheridan, J., and Tennison, J. (2010). "Linking UK government data." *Proceedings of the Linked Data on the Web (LDOW)*. Raleigh, North Carolina.

Sun, Y., and Han, J. (2012). "Mining heterogeneous information networks: principles and methodologies." *Synthesis Lectures on Data Mining and Knowledge Discovery*. 3(2), 1-159.

United Nations (2014). "World's population increasingly urban with more than half living in urban areas." <http://www.un.org/en/development/desa/news/population/world-urbanization-prospects-2014.html> (Jul. 10, 2014).

Venetis, P., Halevy, A., Madhavan, J., Pasca, M., Shen, W., Wu, F., Miao, G., and Wu, C. (2011). "Recovering semantics of tables on the web." *Proceedings of the VLDB Endowment.* Seattle, WA. 4(9), 528-538.

W3C (2008). "SPARQL query language for RDF." <https://www.w3.org/TR/rdf-sparql-query/> (Jan. 15, 2008).

W3C (2014). "Resource description framework." <http//www.w3/org/RDF/> (Feb. 25, 2016).

Yuan, J., Zheng, Y., and Xie, X. (2012). "Discovering regions of different functions in a city using human mobility and POIs." *Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. ACM. Beijing, China. 186-194.

Zheng, Y. (2016). "Methodologies for cross-domain data fusion: An overview. IEEE transactions on big data." (2015). *IEEE Trans on Big Data*. 1(1), 16-34.

Zhu, Y., and Ferreira Jr, J. (2015). "Data integration to create large-scale spatially detailed synthetic populations." *Planning Support Systems and Smart Cities, Geoinformation and Cartography.* Springer. 121-141.

Ziegler, P., and Dittrich K. R. (2004). "Three decades of data intecration - all problems solved?" *Building the Information Society.* 3-12.

Zielstra, D., Hochmair, H. H., and Neis, P. (2013). "Assessing the effect of data imports on the completeness of openstreetmap - A united states case study." *Trans in GIS*. 17(3), 315-334.