GUIDES – Geospatial Urban Infrastructure Data Engineering Solutions (Demo Paper)

Booma Sowkarthiga Balasubramani[1], Omar Belingheri[2], Eric S. Boria[1], Isabel F. Cruz[1], Sybil Derrible[1], Michael D. Siciliano[1]
[1]University of Illinois at Chicago, USA
[2]University of Milan Bicocca, Italy
bbalas3@uic.edu,o.belingheri@campus.unimib.it,{eboria2,ifcruz,derrible,siciliano}@uic.edu

ABSTRACT

The digitization of legacy infrastructure constitutes an important component of smart cities. While most cities worldwide possess digital maps of their transportation infrastructure, few have accurate digital information on their electric, natural gas, telecom, water, wastewater, and district heating and cooling systems. Digitizing data on legacy infrastructure systems comes with several challenges such as missing data, data conversion issues, data inconsistency, differences in the data format, spatio-temporal resolutions, structure, semantics and syntax, difficulty in providing controlled access to the datasets, and so on. Therefore, we introduce *GUIDES*, a new data conversion and management framework for urban infrastructure systems, which is comprised of big data analytics, efficient data management techniques, semantic web technologies, methods to ensure information security, and tools that aid visual analytics. The proposed framework facilitates: (i) mapping of urban infrastructure systems; (ii) integration of heterogeneous geospatial data; (iii) a secured way of storing, analyzing and querying data while preserving the semantics; (iv) qualitative and quantitative analysis over several spatio-temporal resolutions; and (v) visualization of static (e.g., land use) and dynamic (e.g., road traffic) information.

KEYWORDS

GIS, Geospatial data, Data integration, Ontology, Data analytics, Visualization, Data security, Data management, Urban infrastructure

1 INTRODUCTION AND MOTIVATION

Complex geo-analytical applications require integration of two or more cross-domain geospatial datasets such as soil types, underground water pipes, traffic conditions, weather forecasts and other events that change with respect to time and space for effective spatial decision-making. The integration of datasets from multiple different sources involve matching of the datasets geometrically and topologically, as well as having some correspondence of attributes.

Significant opportunities for smarter data management of urban infrastructure systems are on the rise, as many US cities are moving towards the vision of "smart cities", employing recent advances in information technology and creating open data portals that enable city administrators and residents to explore urban data and perform predictive analyses. Despite the availability of a tremendous volume of available data on cities, the lack of accurate geospatial data for underground infrastructure systems remains a problem. The need to address the poor state of the existing infrastructure is a strong rationale to develop such data management systems. For example, New York City has over 6,800 miles of water mains whose average age is 69 years. Over two thirds of them are made of materials susceptible to internal corrosion and prone to leakage, causing over 400 water main breaks in 2013 alone. For an effective infrastructure maintenance, it is important to locate pipes that need to be replaced in order of priority while accounting for funding constraints. Moreover, effort could be put into coordinating across agencies so that road excavation and resurfacing be done at the same time. Such scenarios also reiterate the need for integrating multiple heterogeneous geospatial datasets.

The ultimate objective is to enable a wide variety of people, such as service providers, policy makers, and the general public to explore, query, and analyze urban data to make intelligent decisions, and understand and be aware of the events around them, while simultaneously concealing the sensitive data. However, there are several technical challenges associated with achieving this vision. Data come from various sources, they possess differences in format, representation, context, tools, traits, structure, events, data models, spatio-temporal resolution, data collection and storage techniques, and the relationship between various system properties in a given region. Also, data are most often erroneous, incomplete, and inconsistent, leading to uncertainty. All these factors affect a data management framework, leading to imprecise results when the data is analyzed.

In this paper, we describe *GUIDES*, a novel framework to map and query urban underground infrastructure systems, and show how heterogeneous geospatial datasets are mapped and integrated. This framework involves big data analytics, methodologies for efficient data management, algorithms to impart information security, semantic web-based techniques to enable context-aware decisions, and state-of-the-art visualization tools. We also present two scenarios which demonstrate the capabilities of *GUIDES* in terms of pre-processing and integration of heterogeneous data.

2 FRAMEWORK

This section introduces the *GUIDES* framework (Figure 1) for data management and briefly describes its components.

2.1 Mapping

2.1.1 Data Sources and Data Providers. Big-data driven decision making in smart city applications require the integration of diverse map-based data sources, many of which are non-standardized. Standardization of data sources and coordination among *data providers*, such as municipalities and service providers can improve the accuracy of data that is being centrally integrated. Some municipalities further verify map accuracy through on-field inspection and ever-improving real-time sensor information. However, the accuracy of such geospatial data still remains problematic. *GUIDES* also employs relevant external data sources such as census data, economic

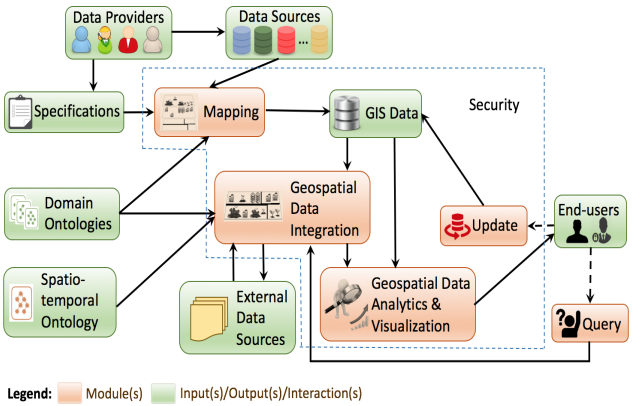

Figure 1: The GUIDES Framework.

data, among others. One of the initial challenges of this framework is to create accurate GIS-based representations of underground infrastructure systems from existing legacy data sources, and on-field verification to enable the mapping of multiple thematic layers of such data.

2.1.2 Mapping & Pre-processing. Mapping deals with the conversion of data from one or more non-standardized sources into a single standardized format. Legacy data formats lack geographical information and often contain all the relevant information in one single source. Dimensions, for example, are often shown directly on the CAD drawing as opposed to being an attribute of a piece of infrastructure. Pre-processing algorithms that can automatically detect and solve these types of issues are critical. *GUIDES* follows a three-step approach for pre-processing. First, a set of rules were developed based on domain knowledge to identify problem areas. For example, a manhole in a water distribution system represented as a circle with multiple nodes as opposed to a single node with an attribute "manhole". The second step leverages machine learning to train and apply classifiers.To increase the accuracy of the classifiers, new variables based on the network properties (e.g. number of connections) are first generated from the converted data. We also incorporate GIS features to test whether a point is located within a polygon or not. After having highlighted misplaced or missing elements, the third step is to "suggest" the correct configuration, for which, we train other machine learning classifiers and leverage the information present in other infrastructure systems. For instance, given that most underground infrastructure systems are buried under roads, road data can be used to suggest where missing infrastructure should be located. Once complete, all errors and added infrastructure elements can be flagged until they are validated manually, during maintenance or new construction.

2.2 Geospatial Data Integration

Data may be collected with different spatial and temporal resolutions, update frequencies, and geometry types [10, 12] with heterogeneity across dimension, location, scale and source. To address these challenges, *GUIDES* uses two kinds of ontologies: (i) a set of domain ontologies; and (ii) a spatio-temporal ontology. The domain ontology deals with instances related to a specific domain (e.g., water distribution) obtained from the GIS database or external data sources, whereas the spatio-temporal ontology consists only of the spatial and temporal hierarchies (e.g., urban spatial hierarchy) and their corresponding instances. To handle the differences in data formats between the datasets, and to address several forms of heterogeneities, we perform *instance matching* based on the spatio-temporal components, which identifies the relationships between each domain ontology and the spatio-temporal ontology. The differences in spatial and temporal resolutions are handled using spatio-temporal algebraic functions such as *within*, *intersects*, *overlaps*, *contains*, and so on.

2.3 Geospatial Data Analytics & Visualization

The analytics component incorporates geostatistical models that enable data exploration, precise predictions of values for geospatial entities, evaluation and comparison of various geostatistical models, quantification of uncertainty, and geostatistical simulations. This component also implements models for descriptive, exploratory, and explanatory statistics. It is important to display multivariate data in a way that facilitates better decision making and data exploration with focus (e.g., details on an area where a water leakage is being repaired) and context (e.g., sketch of other infrastructure elements around the focus area) at the same time. Therefore, the visualization component consists of geometric, glyph- or icon-based, pixel-oriented, and hierarchical techniques, along with an interactive map-based interface to support data exploration.

2.4 Query & Update

GUIDES supports a wide range of geospatial queries using an interactive map-based interface that incorporates the semantics of different domains in the query processing algorithms, along with aggregated data for use by various users (e.g. administrators, residents, maintenance crews). The interface facilitates queries using an API for the *SPARQL* (SPARQL Protocol and RDF Query Language) engine with spatio-temporal extension, to support multiple geometry types, spatio-temporal functions, geospatial aggregate functions and update statements. Once a query is issued, the super classes (parent classes) and sub-classes (child classes) of the spatio-temporal components are retrieved from the spatio-temporal ontology, and geospatial data integration is carried out. The set of functions that needs to be considered for a specific query depends on which dataset's instances are sub/super/equivalent classes (in the spatio-temporal ontology) of the spatio-temporal components in the query. Users can query location- and time-specific data based on their particular data needs and authorized level of access. As required by their work, some users would have access to infrastructure data that is deemed secure, but would not be searchable within public access platforms.

3 DEMONSTRATION SCENARIOS

This section demonstrates how the *GUIDES* framework enables pre-processing and ontology-based data integration mechanisms for urban infrastructure data, using the *Water Pipes* and *Buildings* maps from the University of Illinois at Chicago campus. These maps were initially in dwg format and they were converted to shapefile format. The original data contained several errors and inconsistencies, and the conversion process generated several errors as well. We used QGIS[1] for visualizing the maps. The maps are transformed into a list of nodes and edges using a Python script [11], splitting the edges at intersections with nodes.

[1] http://qgis.org/

3.1 Pre-processing the Water Pipes Map

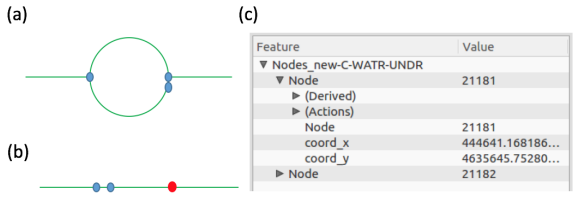

Figure 2: Errors in Water Pipes Map.

The screenshot in Figure 2a is part of the *Water Pipes* dataset, which is converted into shapefile from AutoCAD drawing. This figure represents a manhole and should have been represented as a single node as opposed to a circle (composed of several polylines) with 3 nodes. The nodes (in the right side) are also disconnected. This subsection deals with identifying and correcting such errors.

3.1.1 Fixing Duplicate Nodes. In Figure 2b, though the feature highlighted in red appears to be a single node, the corresponding feature table in Figure 2c shows that it is in fact two nodes with two different IDs. That is, the two nodes are separate features within the layer and there is no edge connecting them. These types of scenarios are common and pose obvious issues, even when the most basic operations on the network are performed. For example, trying to find a path that goes through the edges in Figure 2b will fail, simply because the overlapping nodes are not connected. To resolve this, a Python script which involves the *osgeo* and *networkx* libraries is run to remove one of those nodes for each pair of such duplicate nodes, and connect its edges to the other copy of the node.

3.1.2 Differentiating Infrastructure Elements and CAD symbols. Figure 2a shows an example of a loop representing a manhole. Zooming into this figure, we can see that only one of the three nodes is actually connected to the loop edge. The loop was deleted and replaced with a new node in its center, with proper connections to the others nodes on either side of it. A field named *Is_drain* is also added to this node and is set to 1 in the attribute table of the map, so that the information is kept intact, though the loop is removed.

3.1.3 Context-aware Pre-processing. To identify further errors in the *Water Pipes* map, we used the *Buildings* layer. For example, intuitively, a water pipe should either end in a building, or be connected to other water pipes. If a water pipe is either disconnected from the rest of the network or ends abruptly, it is reasonable to assume that there is an error, and should be flagged for correction. Such cases can be identified by finding the end nodes in the *Water Pipes* layer, that is, nodes with degree 1. This hypothesis has been confirmed by our experiments with synthetic map layers for *Water Pipes* and *Streets*. After the random removal of water pipes, we validated whether our logic suggests proper corrections to restore the initial map. In doing so, using the constraints enforced by the *Streets* layer (say, a water pipe would normally run underneath a street, without crossing its borders), has proven to be fundamental in reducing the number of false positives (pipes we would have incorrectly added), raising the precision from 59% to 93%. Applying this hypothesis to the UIC datasets, we should also ensure that the end nodes that are placed within the perimeter of a building should not be flagged. Therefore, this turns out to be a *point-in-polygon* problem [13]. To resolve this, we use the *osgeo* Python library, which, given a point (a node in the water pipes) and a polygon (a building), checks whether the the point falls within the area of the polygon.

The *osgeo* library allows for the creation of multipolygons, which are objects that can contain several polygons. We made use of this feature to have one object containing all the polygons of the buildings, instead of having one object (a polygon) for each building. The point-in-polygon check was then performed with the multipolygon in one iteration over the nodes, instead of using two iterations to check if any of the nodes (1st iteration) is in any of the buildings/polygons (2nd iteration). Although the idea of having a single iteration seems to be intuitively more computationally efficient, it is not. The solution with two iterations results in a much faster computation and was therefore chosen for the final implementation.

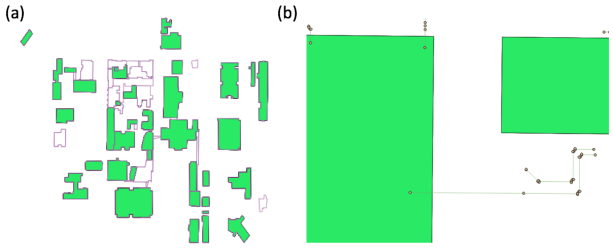

Figure 3: (a) Buildings (Purple Lines) and Corresponding Polygons (Green Areas); (b) UIC Water Pipes and Buildings Layer - Partial.

From Figure 3a, we can see that the polygons (green areas) do not cover all of the buildings (purple lines) that the map contains. This is again because, the map contains inconsistencies such as broken edges, detached nodes, and so on, which makes it impossible for *GUIDES* to build all the polygons properly. Therefore, this layer needs to be pre-processed to remove impurities and connect nodes that define the boundaries of a building, which we do by testing whether a *building* node is on the edge of a full polygon or not, and if not, it can be connected to the closest node and flagged. Figure 3b shows a section of the maps of the water pipes, the buildings, and the polygons. All the nodes that are not inside the green areas and that have degree 1 will be flagged.

3.2 Ontologies in Geospatial Data Integration

This step encompasses ontology matching. More specifically, we intend to match the instances associated with different concepts in the ontologies. To do this, a domain ontology for each dataset (e.g. crime as in Figure 4a), and a generic spatio-temporal ontology (Figure 4b) are constructed using Protègè[2], a widely used ontology editor. For urban underground infrastructure systems, the local ontologies should be generic enough to accommodate minor changes to the datasets. For example, the *Water Pipes* dataset of New York and that of Chicago might have different structures, but they refer to the same domain (Water Mains). Therefore, the ontology should not only consider the concepts based on the attributes in the dataset, but also the generalization and specialization of each of these concepts. The future goal of the integration module is to

[2]http://protege.stanford.edu/

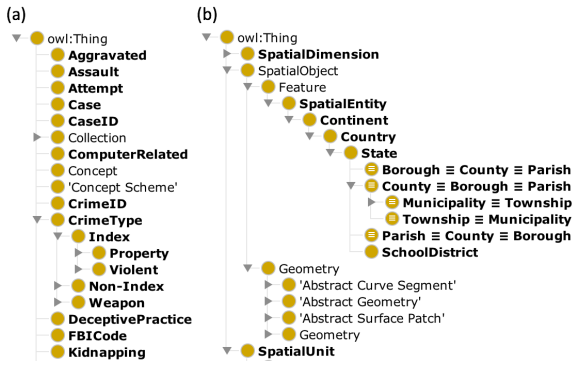

Figure 4: (a) Domain Ontology (Example: Crime) - Partial; (b) Spatio-temporal Ontology - Partial.

match the instances in each local ontology, with the instances of the global ontology, based only on the spatio-temporal components, using AgreementMakerLight (AML) [8], one of the top ontology matching tools [7].

4 RELATED WORK

GIVA [6], an interactive map-based application for Geospatial and temporal data Integration, Visualization, and Analytics, enables mapping several datasets with each other, for a given region and a time interval. GUIDES adds on to the capabilities of GIVA, in terms of pre-processing, categories of users, use of external data sources, and the mechanism used for data integration in the urban infrastructure domain. OpenGrid [14], a map-based open-source platform, was developed by the City of Chicago, to support advanced queries to identify and monitor incidents across the city. It is built on top of Plenario [4], which is a geospatial data warehouse that allows data from many sources to be registered into a common spatial and temporal frame. OpenGrid only accepts queries that are based on point data (e.g., location of potholes) and only on certain datasets. Many of the vital infrastructure systems, such as gas, electricity, water, sewer, transportation, and telecommunication systems are not included in OpenGrid, and does not support data integration, though it can be extended to perform predictive analytics on urban data [1].

There are also several existing works on performing urban data exploration and urban data analytics. Chang et al. [5] proposed a model that enables visualization of urban relationships using data aggregation techniques. However, it does not support geospatial data integration, and does not address potential issues such as the differences in data models or the spatial and temporal resolutions. Urbane [9] is a 3D multi-resolution framework that involves a data-driven approach for decision making in urban settings. The main idea of this framework is to integrate various datasets and perform impact analysis to evaluate the relationships between multiple attributes and analyze how it impacts the neighborhood. One of the main drawbacks of Urbane is that it supports data exploration only on three different scales (between neighborhoods, within a neighborhood, and indivdual building). Beck et al. [2, 3] have proposed an integrated framework for utility data that uses light-weight ontologies to support knowledge and data integration. They took into account various forms of heterogeneities, but the setup requires major changes when a new dataset is to be integrated. Real-time scenarios are more complex than these frameworks could handle. For example, the existing urban data analytics platforms like OpenGrid [14] do not support cross-domain querying mechanisms, thus reinforcing the need for a new framework like *GUIDES*.

5 CONCLUSIONS AND FUTURE WORK

In this paper, we have introduced *GUIDES*, a semantic web-based data management framework, which supports integration, querying, analytics, and visualization of heterogeneous geospatial datasets, focusing on the urban underground infrastructure domain. The framework also supports several types of users such as administrator, planner, maintenance crew, general public, and so on, with various levels of access. We highlighted the key architectural elements and their capabilities to handle several challenges associated with geospatial data. Given the novelty of the proposed framework and complexity of the problems this framework intends to address, there is a great potential for expansion of this framework in all possible directions. Opportunities for integration of the *GUIDES* framework with open data exploration platforms such as OpenGrid, will also be explored.

ACKNOWLEDGMENTS

GUIDES is being partially supported by NSF Award CNS-1646395 and NSF CAREER 1551731.

REFERENCES

[1] Balasubramani, B. S., Shivaprabhu, V. R., Krishnamurthy, S., Cruz, I. F., and Malik, T. Ontology-based Urban Data Exploration. In *Proceedings of the 2nd ACM SIGSPATIAL Workshop on Smart Cities and Urban Analytics* (2016), p. 10.

[2] Beck, A. R., Cohn, A. G., Sanderson, M., Ramage, S., Tagg, C., Fu, G., Bennett, B., and Stell, J. G. UK Utility Data Integration: Overcoming Schematic Heterogeneity. In *Sixth International Conference on Advanced Optical Materials and Devices* (2008), International Society for Optics and Photonics, pp. 71431Z–71431Z.

[3] Beck, A. R., Fu, G., Cohn, A. G., Bennett, B., and Stell, J. G. A Framework for Utility Data Integration in the UK. In *Proceedings of the Urban Data Management Society Symposium* (2007), Taylor & Francis, pp. 261–276.

[4] Catlett, C., Malik, T., Goldstein, B., Giuffrida, J., Shao, Y., Panella, A., Eder, D., van Zanten, E., Mitchum, R., Thaler, S., et al. Plenario: An Open Data Discovery and Exploration Platform for Urban Science. *IEEE Computer Society Technical Committee on Data Engineering 37*, 4 (2014), 27–42.

[5] Chang, R., Wessel, G., Kosara, R., Sauda, E., and Ribarsky, W. Legible Cities: Focus–dependent Multi–resolution Visualization of Urban Relationships. *IEEE Transactions on Visualization and Computer Graphics 13*, 6 (2007), 1169–1175.

[6] Cruz, I. F., Ganesh, V. R., Caletti, C., and Reddy, P. GIVA: A Semantic Framework for Geospatial and Temporal Data Integration, Visualization, and Analytics. In *Proceedings of the 21st ACM SIGSPATIAL International Conference on Advances in Geographic Information Systems* (2013), ACM, pp. 544–547.

[7] Faria, D., Pesquita, C., Balasubramani, B. S., Martins, C., Cardoso, J., Curado, H., Couto, F. M., and Cruz, I. F. OAEI 2016 Results of AML. *ISWC International Workshop on Ontology Matching (OM)* (2016), 138.

[8] Faria, D., Pesquita, C., Santos, E., Cruz, I. F., and Couto, F. M. AgreementMakerLight: A Scalable Automated Ontology Matching System. In *Proceedings of the 10th International Conference on Data Integration in the Life Sciences (DILS)* (2014), pp. 29–32.

[9] Ferreira, N., Lage, M., Doraiswamy, H., Vo, H., Wilson, L., Werner, H., Park, M., and Silva, C. Urbane: A 3D Framework to Support Data Driven Decision Making in Urban Development. In *IEEE Conference on Visual Analytics Science and Technology (VAST)* (2015), pp. 97–104.

[10] Gotway, C. A., and Young, L. J. Combining Incompatible Spatial Data. *Journal of the American Statistical Association 97*, 458 (2002), 632–648.

[11] Karduni, A., Kermanshah, A., and Derrible, S. A Protocol to Convert Spatial Polyline Data to Network Formats and Applications to World Urban Road Networks. *Scientific Data 3* (2016).

[12] Le, Y. Challenges in Data Integration for Spatiotemporal Analysis. *Journal of Map & Geography Libraries 8*, 1 (2012), 58–67.

[13] Sharma, A. K., and Gill, S. K. Methods to Define a Single Point in the Polygon.

[14] Tuecke, S., Foster, I., and Kesselman, C. The OpenGrid Services Architecture. *The Grid: Blueprint for a New Computing Infrastructure* (2004), 215–242.