Smart Data Management of Urban Infrastructure using Geographic Information Systems

Booma Sowkarthiga Balasubramani
Ph.D. Candidate, University of Illinois at Chicago, USA, bbalas3@uic.edu

Mohamed Badhrudeen
Ph.D. Candidate, University of Illinois at Chicago, USA, mmoham55@uic.edu

Sybil Derrible (Corresponding Author)
Associate Professor, University of Illinois at Chicago, USA, derrible@uic.edu

Isabel Cruz
Professor, University of Illinois at Chicago, USA, ifcruz@uic.edu

ABSTRACT: Cities all over the world are converting maps of their infrastructure systems from legacy formats [such as paper maps and computer-aided design (CAD) drawings] to geographic information systems (GIS). Compared with CAD, GIS tend to offer more flexibility in terms of managing, updating, analyzing, and processing data. Nonetheless, the conversion process to GIS can be extremely challenging from a technical point of view. Moreover, the original data in a legacy format often contain errors, and pieces of infrastructure are often missing. What is more, even once the conversion process is complete, the maintenance of the data and the fusion of the data set with other data sets can be challenging. Leveraging recent technological advances (such as machine learning and semantic reasoning), this paper proposes a framework to better manage infrastructure data. More specifically, a smart data-management protocol is presented to successfully convert infrastructure maps from CAD to GIS that includes a data-cleaning procedure in CAD and machine-learning algorithmic solutions to validate or suggest edits of the infrastructure once converted to GIS. In addition, the protocol includes elements of version control to keep track of how urban infrastructure evolves over time as well as a procedure to combine GIS infrastructure maps with other data sets (such as socio- demographic data) that can be used for optimal scheduling of asset maintenance and repair

KEYWORDS: smart data, smart cities, infrastructure, data management

INTRODUCTION

The use of geographic information system (GIS) in managing information about infrastructure systems has garnered attention from both public and private organizations, mainly because of the ability of GIS to effectively manage and analyze spatial data. In particular, infrastructure data, both existing and new, collected by those organizations can be used to make informed decisions (Everett *et al.* 2015, Saeed *et al.* 2020) and help design smarter, more sustainable, and more resilient infrastructure systems (Derrible 2018, 2019, Mohareb *et al.* 2016). Despite the availability of a tremendous volume of urban infrastructure data, however, the lack of accurate data in a workable format still remains an issue (Boria *et al.* 2020). The need to improve the state of the existing infrastructure in conjunction with its dependent infrastructure(s) is a strong rationale to develop smart data–management systems. Simultaneously, significant opportunities for smarter data management of urban infrastructure systems are on the rise as many cities are moving towards the vision of "smart cities," employing recent advances in information technology (Beck et al. 2007, Cruz *et al.* 2013) and creating open data portals that enable city administrators and residents to explore urban data and perform predictive analyses (Ahmad *et al.* 2017, Lee and Derrible 2020). For example, New York City has over 11,000 km (approximately 6,800 mi) 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 maintenance to be effective, it is important to locate pipes that need to be replaced in priority while accounting for funding constraints.

Although infrastructure information can be created using GIS from scratch, it is preferable to convert existing infrastructure information into GIS, not only because the process should be easier, but also because the legacy data might have additional information that might not be evidently copied manually. In public and private organizations, the format that has been widely used for storing infrastructure information is computer aided design (CAD). The goal is therefore to convert CAD data into GIS without losing the important information. Rarely, does one find any data to be free of errors, however, and therefore it is necessary to clean the CAD data before converting into GIS. The errors can range from misplaced to missing features (e.g., a misplaced manhole). In addition to errors present in the initial CAD data, errors are also often introduced during the conversion process to GIS. Therefore, some corrections have to be made directly in CAD format to reduce the potential number of errors that would show up and be more complicated to address in GIS. Afterwards, CAD data can be converted to GIS, and effort can then be put into addressing the errors present in GIS. These errors in the GIS data have to be corrected before the data is added to the main database.

Finally, the error-free GIS data can be managed and updated regularly in a database. Sound data-management practices can then be put in place, for example by including elements of version control, to keep track of the temporal evolution of the infrastructure systems. Furthermore, the advantage of using a general database is that data from other sources can be accessed for analysis purposes—e.g., how many people get affected by a water main break— thus speeding the analysis and leading to efficient decision-making processes in line with the vision of "smart cities." Furthermore, since infrastructure systems are accessed and used by a wide variety of people with different use cases—including service providers, policymakers, researchers, and the general public—it is important to place an access control mechanism on the database. To that end, this study proposes a theoretical framework to efficiently integrate and manage multi-sourced infrastructure database.

In the remainder of the study, the proposed framework to effectively convert and manage the converted CAD data into GIS data is explained; a more detailed explanation of the CAD-to-GIS conversion framework has been given by Boria *et al.* (2020). In addition, the preprocessing procedure that ensures the accuracy of the information before the data is added to the database, and its subsequent management and version control, are also discussed, as is the access-control mechanism that facilitates various levels of access to different stakeholders.

PROPOSED FRAMEWORK

This section introduces the proposed framework for smart data management and operations of urban infrastructure systems illustrated in Fig. 1 and briefly describes its components.

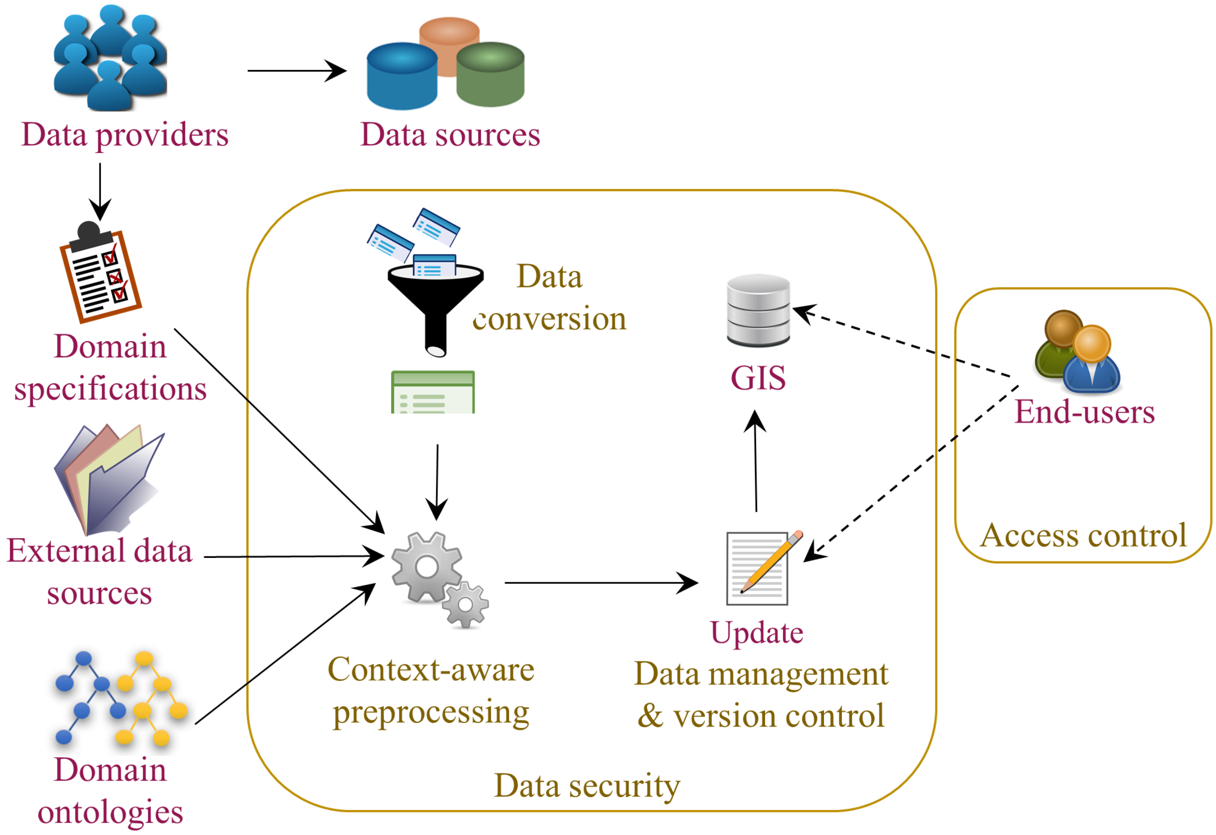


Figure 1: Proposed framework

Data Conversion

The proposed process of converting CAD to GIS consists of five steps: identification of needs, CAD data cleaning, conversion, georeferencing, and GIS data cleaning. Fig. 2 shows the detailed version of the five-step process, adapted from Boria *et. al.* (2020).

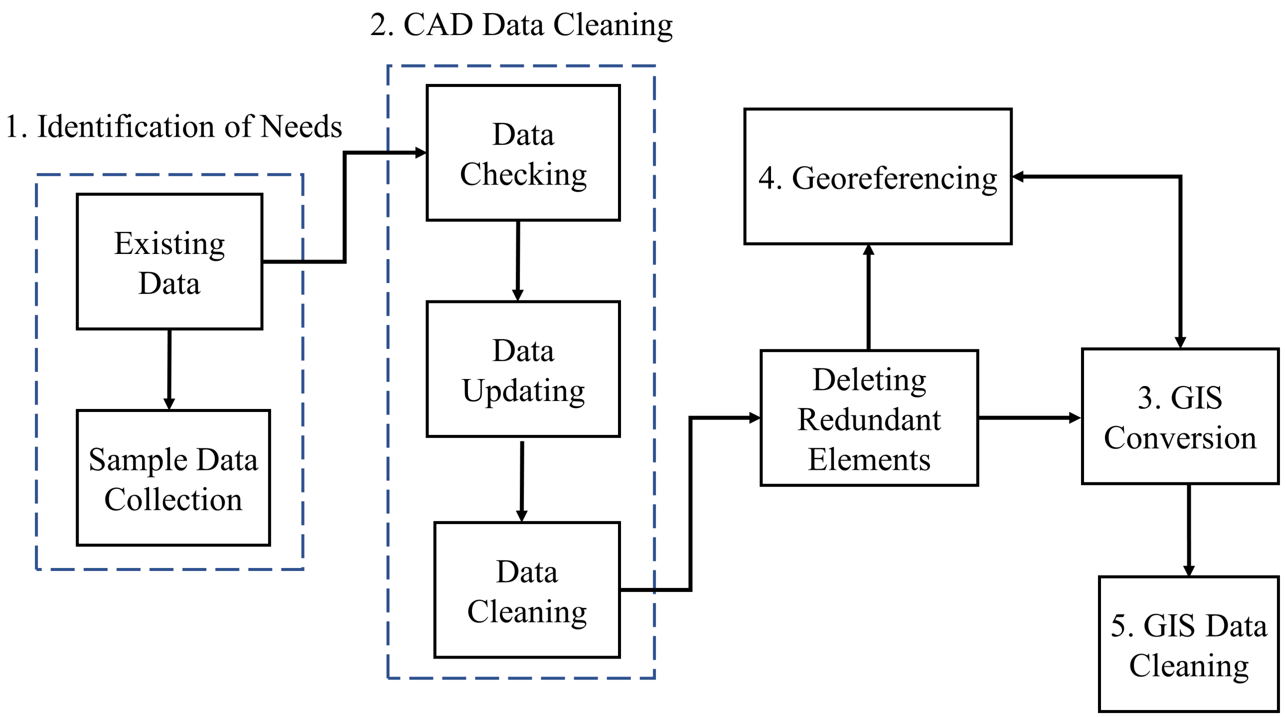


Figure 2: CAD to GIS conversion flow chart; adapted from Boria *et. al.* (2020)

For existing CAD data, the *identification of needs* step allows the user to verify the accuracy of the CAD information. It also helps to understand the nature of the infrastructure and its features. This step facilitates the planning of the actions that need to be taken in the next steps.

In the second step, *CAD data cleaning* is performed in conjunction with the details collected in the previous step. At this stage, data cleaning includes the identification of unwanted information, if any, and subsequently deleting them. Additionally, it enables the correction of errors that are easier to address in CAD rather than in GIS. For instance, labels in CAD provide vital information of the features and are placed near the appropriate feature. Nonetheless, sometimes due to lack of space in the drawing, an arrow is used to identify to which feature a label belongs. In Fig. 3, for example, the labels Valves 1 and Valves 2 are used here to identify valves in CAD. Although the labels serve a purpose in CAD, they often do not provide necessary information in GIS after the conversion. Because each feature (e.g., valves) is assigned a unique ID during the conversion, keeping the redundant information only complicates the process. Further, the arrow that is used for identification in CAD does not provide any information in GIS either. Hence, it is preferable to address this problem associated with the placement of labels directly in CAD.

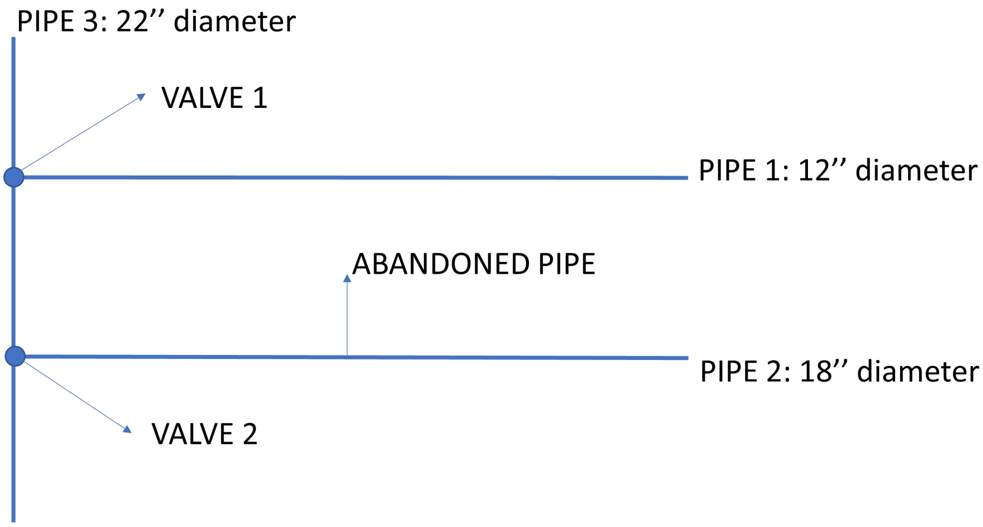


Figure 3: Label placement in CAD

The next two steps, *conversion* and *georeferencing*, are interchangeable. Conversion is a straightforward process in which CAD data are converted to a standardized shapefile format (i.e., .shp) readable by GIS, and georeferencing is the process of assigning a global coordinate system to the converted shapefiles. Georeferencing is an important step in the process because it geographically positions each feature where it is located in the world. In fact, this is one of the advantages of GIS over CAD. Additionally, georeferencing provides a means to perform spatial analyses [e.g., to identify vulnerable areas (Wisetjindawat 2017)], which is impossible in CAD. Put simply, georeferencing involves the use of a reference data source (e.g., buildings or roads) to align the newly converted data to this reference data source. Generally, more overlap (in the form of common points) between the newly converted data and the reference data sources results in more accurate georeferencing.

Finally, the *GIS data cleaning* step is applied to process converted files to both identify and correct errors. Here, identifying the errors serves two purposes: finding errors that (1) could not be corrected in CAD, and (2) might have been introduced during the conversion process. One of the errors that is often not corrected in CAD is labeling because labels hold vital information about the features that should be included in GIS data. Therefore, labels have to be separated and added to the attribute table in GIS. For example, if the labels are the diameter of the pipe, after conversion, a separate column can be created (e.g., with the header Diameter) in the GIS attribute table and this value is added to the corresponding pipe (Fig. 3).

Context-Aware Preprocessing:

As discussed, the conversion of CAD drawings to GIS format does not ensure accuracy of infrastructure data. Although public and private organizations deal with validating map accuracy through on-field inspection and ever-improving real-time sensor information, the accuracy of such infrastructure data still remains problematic. For example, the placement of texts in CAD drawings can be misinterpreted as an attribute to a different feature in the infrastructure. Inaccuracies in infrastructure data such as missing components (e.g., a water pipe) and incorrect location of components (e.g., water pipes running across the street rather than along the street) are also quite common. Therefore, in addition to using them for georeferencing, the use of external data sources, which are called *supporting layers*, such as roads and buildings, are used here to aid in the preprocessing of infrastructure layers to improve accuracy (Balasubramani *et al.* 2017). This context-aware preprocessing phase (Fig. 4) is envisioned as a two-step process: (1) identification of inconsistencies using ontologies based on domain specification, and (2) use of machine learning techniques to confirm/refute and fix errors based on information from supporting layers and/or domain expertise.

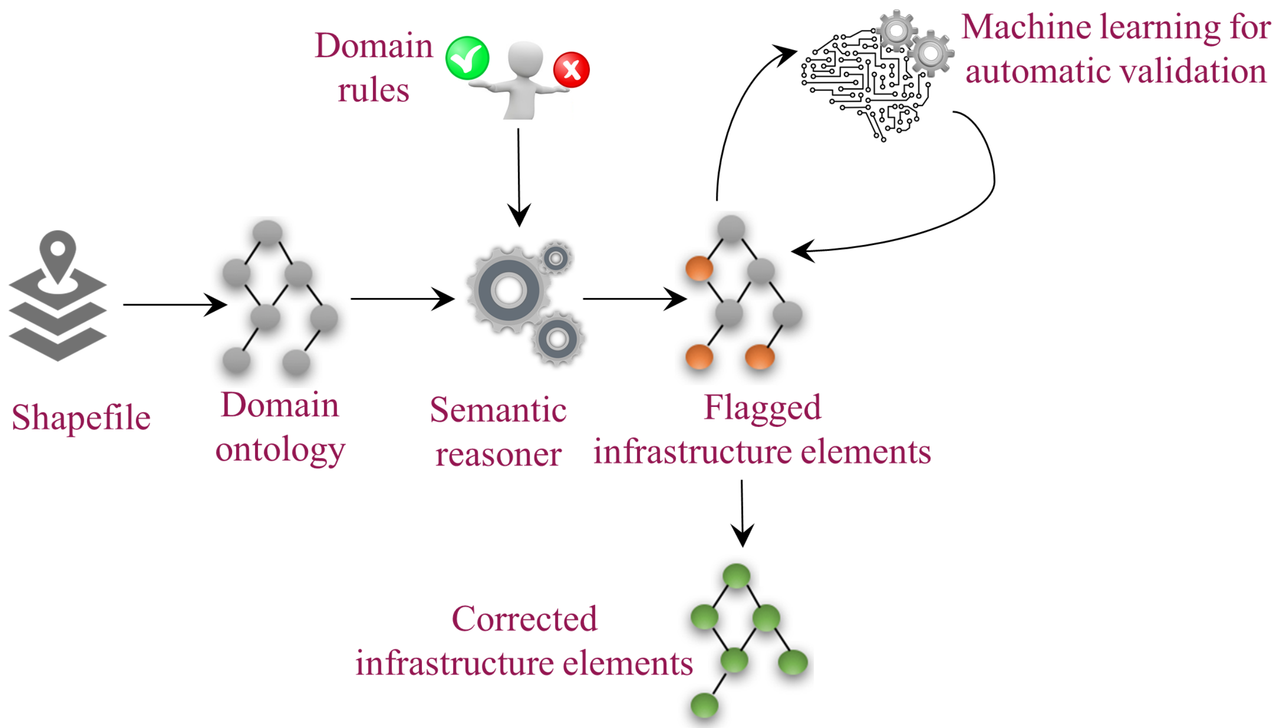
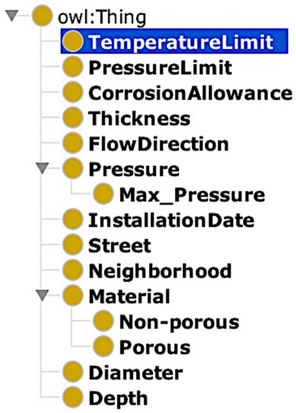

Domain rules
Machine learning for automatic validation
Shapefile
Domain ontology
Semantic reasoner
Flagged infrastructure elements
Corrected infrastructure elements

Figure 4: Context-aware preprocessing

Identification of Inconsistencies Using Ontologies

Ontologies are knowledge representation structures that help establish the relationship between various entities. The use of ontologies for urban data to generate patterns for decision making have already been explored (Balasubramani *et al.* 2016). The approach proposed here extends the use of ontologies to identify inconsistencies in infrastructure data. A specific type of ontologies called domain ontologies [Fig. 5(a)] are used, which are standardized ontologies with structured vocabularies that domain experts use to share and annotate information in their fields (Noy and McGuinness 2001). This step also involves a set of rules for each domain, which are nothing but asserted facts developed based on domain knowledge, to identify problem areas (e.g., a water pipe is connected to at least two other pipes) [Fig. 5(b)]. Tools to infer logical consequences from a set of asserted facts, called semantic reasoners are employed to validate whether the ontology for the given domain satisfies the set of rules. The instances of the ontology (e.g., a specific water pipe) that does not satisfy the constraints are then flagged for further investigation.



```
domain: Water_Pipe
     rdf:type owl:Class;
     rdfs:subClassOf   domain: Infrastructure;
     rdfs:subClassOf
        [
          rdf:type    owl:Restriction;
          owl:onProperty   domain:numOfConnections;
          owl:minCardinality    “2”^^xsd:integer
        ].
```

(a) Water Pipes ontology (partial)   (b) Example of a domain rule encoded in the Water Pipes ontology

Figure 5: Identification of inconsistencies using ontology

Machine Learning for Preprocessing

Machine learning (ML) for preprocessing can be deployed in both detection and correction of errors (Badhrudeen *et al.* 2020). Depending on the type of data being dealt with, techniques like classification are applied to detect errors. More often than not, the infrastructure data sets are voluminous, and it takes time to validate manually each error that is flagged by the semantic reasoner. In such cases, one may rely on *active learning* (Settles 2009), where candidate errors are identified and then validated based on domain expertise or supporting layers (if available). For example, an overlay of the land-use data on the water pipe network could help to pinpoint that the water pipe with only one connection is the one terminating in a dead-end street, and hence it is not an error. This result (that is, error/ not an error) is applied to infrastructure elements flagged with the same error, using ML.

Data Management and Version Control

Infrastructure data are quite complex in terms of the variety of data formats, volume of data, number and type of users (e.g., database administrators, policymakers, and general public), and the kind of analysis performed on the data (Le 2012). Therefore, storing and retrieving such data from a typical file system is insufficient. Also, to ensure accurate and complete digital information, the addition and removal of infrastructure elements should precisely be documented. This calls for a standard version control technique. However, most infrastructure systems are represented in a single file (for example, a water pipe network is a single CAD file), and changes to this system by different stakeholders makes it difficult to merge updates. Although maintaining data sets in a centralized system and versioning data sets have been explored vastly in the past, different permission levels assigned to different types of users complicate the system administration, among other challenges. The authors envision a system with an ability to store only changes in the data (such as addition/removal of elements) in each version of the data set. This enables keeping track of the changes in the infrastructure over time, while simultaneously minimizing the risk of data redundancy.

Access Control

Urban infrastructure data are being used by multiple stakeholders ranging from service providers, policymakers, and the general public to explore, query, and analyze those data to make intelligent decisions, and understand and be aware of the events around them (e.g., traffic conditions due to a water main break). Therefore, it is necessary to conceal sensitive information. The idea is to allow users to query location- and time-specific data based on their particular data needs. For example, the general public need not be aware of the underground infrastructure data along with its cross-thematic and cross-administrative boundaries collaborations. This is achieved by authorizing different levels of access to the data sets to different sets of users. To further ensure the security of infrastructure data, the access information and the actual data are stored in an encrypted format on separate servers. This makes it costefficient to secure the server containing the access information because the volume of data is relatively smaller compared with the actual data.

CONCLUSIONS

The successful management of “smart cities” relies heavily on robust data-management practices to collectively build solutions using predictive analytics or data visualization and the likes. This study proposes a framework for managing digital maps of infrastructure systems, which involves a conversion of non-standardized data formats to a single standardized format, an improvement and maintenance of the accuracy of infrastructure data sets using semantic reasoning and ML techniques, and an overview of the version control mechanism that stores only changes in the data to minimize data redundancy. The techniques discussed are more general and can easily be applied to various domains, providing a great potential for expansion of this system at/among various organizational levels (e.g., cross-jurisdictional or interorganizational collaborations). Overall, although much work remains to be done, smart data management is a key component of smart cities, and aided by recent technological advances (such as ML), smart data management has the potential to completely transform how urban infrastructure systems are planned and operated.

DATA AVAILABILITY STATEMENT

No data, models, or code were generated or used during the study.

ACKNOWLEDGEMENT

The authors thank the following students for the preliminary work related to this framework: Omar Belingheri, Guillemette Fonteix, Nalin Naranjo, and Eric Boria. The authors are also indebted to coinvestigators Michael Siciliano, Roberto Tamassia, and Goce Trajcevski for many helpful discussions.

This work was partially supported by NSF awards CNS-1646395, CAREER 155173, III-1618126, and CCF-1331800.

REFERENCES

Ahmad, N., Derrible, S. and Cabezas, H. (2017). “Using Fisher information to assess stability in the performance of public transportation systems.” *Royal Society Open Science*, 4(4):160920.

Badhrudeen M., Naranjo N., Mohavedi A. and Derrible S. (2020). “Machine learning based tool for identifying errors in CAD to GIS converted data.” In: *Ha-Minh C., Dao D., Benboudjema F., Derrible S., Huynh D., Tang A. (eds) CIGOS 2019, Innovation for Sustainable Infrastructure. Lecture Notes in Civil Engineering*, vol 54. Springer, Singapore.

Balasubramani, B.S., Shivaprabhu, V.R., Krishnamurthy, S., Cruz, I.F. and Malik, T. (2016). “Ontology-based urban data exploration.” *Proceedings of the 2nd ACM SIGSPATIAL Workshop on Smart Cities and Urban Analytics (p. 10)*. ACM.

Balasubramani, B.S., Belingheri, O., Boria, E.S., Cruz, I.F., Derrible, S. and Siciliano, M.D. (2017). “GUIDES: Geospatial Urban Infrastructure Data Engineering Solutions.” *Proceedings of the 25th ACM SIGSPATIAL International Conference on Advances in Geographic Information Systems (p. 90)*. ACM.

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

Boria, E.S., Badhrudeen, M., Fonteix, G., Derrible, S., and Siciliano, M. (2020). “A protocol to convert infrastructure data from Computer Aided Design (CAD) to Geographic Information Systems (GIS).” Available at https://arxiv.org/abs/2006.14112 (accessed July 8, 2020)

Cruz, I. F., Ganesh, V. R., Caletti, C., and Reddy, P. (2013). “GIVA: A Semantic Framework for Geospatial and Temporal Data Integration, Visualization, and Analytics.” *Proceedings of the 21st ACM*

SIGSPATIAL International Conference on Advances in Geographic Information Systems (pp. 544–547). ACM.

Derrible, S. (2018). “An Approach to Designing Sustainable Urban Infrastructure.” *MRS Energy and Sustainability*. 5 (E15).

Derrible, S. (2019). *Urban Engineering for Sustainability*. MIT Press, Cambridge, MA.

Everett, S., Athigakunagorn, N., Roshandeh, A. M., Labi, S., and Sinha, K. C. (2015). “Geographic Information System Tool for Enhancing Administration of Overweight-Vehicle Permits.” *Transportation Research Record*, 2478(1), 75–81.

Le, Y. (2012). “Challenges in data integration for spatiotemporal analysis.” Journal of Map & Geography Libraries, 8(1):58-67.

Lee, D., and Derrible, S. (2020). “Predicting Residential Water Demand with Machine-Based Statistical Learning.” ASCE Journal of Water Resources Planning and Management, 146(1): 04019067

Mohareb, E., Derrible, S., and Peiravian, F. (2016). “Intersections of Jane Jacob’s Conditions for Diversity and Low-Carbon Urban Systems: A Look at Four Global Cities.” *Journal of Urban Planning and Development*. 142(2).

Noy, N. F., and& McGuinness, D. L. (2001). “Ontology development 101: A guide to creating your first ontology.” Available at https://protege.stanford.edu/publications/ontology_development/ontology101.pdf (accessed July 8, 2020)

Saeed, T.U., Nateghi, R., Hall, T. and Waldorf, B.S. (2020). “Statistical Analysis of Area-wide Alcohol-related Driving Crashes: A Spatial Econometric Approach.” Geographical Analysis. doi:10.1111/gean.12216

Settles, B. (2009). “Active learning literature survey.” *Computer sciences technical report 1648, University of Wisconsin-Madison.*

Wisetjindawat, W., Kermanshah, A., Derrible, S., and Fujita, M. (2017). “Stochastic modeling of road system performance during multihazard events: Flash floods and earthquakes.” *Journal of Infrastructure Systems,* 23(4), 04017031.