RESEARCH STATEMENT – 1st draft

My research interests are in line with adding value to the existing isolated data of the civil infrastructure industry to reduce manual and duplicate effort in data manipulation, processing and creation by introducing new knowledge and computational intelligence infrastructure systems that can support automated data extraction, mining and information reasoning.

Due to the fragmented nature of the construction industry, digital models and datasets (used by designers, contractors, owners, etc.) are being developed and managed in a non-collaborative manner where data is repeatedly regenerated instead of fully reused. Moreover, most of the existing data, information and domain knowledge are presented in natural language which is non machine-readable formats such as texts, and potential information sources in visual media like videos, images.

- and improve efficiency in data manipulation and enhancing the efficiency of digital data utilization, reduce data collection effort, reusability of digital data from multiple isolated data sources throughout the life cycle of civil infrastructure projects, highway agencies, public and private sources.
- These data are stored and managed separately and stored in heterogeneous data format, images, videos, natural language text documents, and language.
- The adoption of various advanced digital technologies has enabled a large portion of the project life cycle data to become available in digital format. However, due to the fragmented nature of the highway industry sector, digital data sets are being archived, and managed separately. Due to this reason, the transportation assets as a whole have not yet fully benefited from the growing amount of digital data since digital data from distinct actors are yet to be linked and fully reused.
- The implementation of advanced, and computerized technologies such as Building Information Modeling (BIM), Geographic Information System (GIS), throughout the life cycle of a highway project has allowed a large portion of project data to be available in a digital format. The efficiency improvement in sharing these data between project participants and stages, will in turn, translate into increased productivity, efficiency in project delivery and accountability.
- However, due to the fragmented nature of the highway industry sector, digital data sets are being archived, and managed separately and in their own preferred formats. Due to this reason, the transportation assets as a whole have not yet fully benefited from the growing amount of digital data since digital data from distinct actors are yet to be linked and fully reused and external actors are not aware the existence or value of other data inventory. Since different project participants may use proprietary software platforms with different data structures, exchange of data becomes very challenging. Data exchange in a heterogeneous environment may lead to data loss, damage and requires time consuming processing in downstream phases.
- The major cost was time spent finding, verifying facility and project information, and transferring that information into a useful format.
- By seamlessly using electronic engineered files generated during planning, design and construction phases, a significant amount of efforts can be saved as assets are managed in order to provide superior results.
- In the last three years I have conducted ample research to 1) enable the linkage of life cycle digital data storage for decision making in asset management, 2) address the issue of data terminology discrepancy among across different data sources.
- Presently, I have been working on developing a natural language based data extraction engine for civil infrastructure and documenting the current data and information flow throughout the life-cycle for various transportation assets.
- In my future research, I plan to expand current and past research to new research areas to natural speech recognized infrastructure for manipulating multiple construction project digital models

Paragraph 2 – Past research 1 – **Life-cycle data space linkage**

- Asset management requires data that are generated from various upstream project phases from design, construction and condition survey.
- In the conventional practices, discipline divisions in a highway agency store and manage data in their own data inventory.
- To facilitate the interlinking of disparate and heterogeneous life-cycle data-spaces so that digital data generated in upstream phases can be fully reused in asset management.
- I developed a novel framework that enables the interconnection of life-cycle digital data sources (BIM models from designers, construction data from Project Management Systems, and asset condition tables form Asset management systems) and translate into meaningful information for decision maker in highway asset management.
- The platform includes several data translators that are able to convert data different formats into graph networks of data (Resources Description Framework) which is readable to both human and machine and linkable and mergerable to external data spaces.
- This proposed platform has been published on the Journal of Automation in Construction (Le & Jeong 2016).
- The results of this research are expected to provide an effective and efficient means to facilitate seamless digital data exchange throughout the life cycle of a highway project. The proposed mechanism can be a potential solution for digitally handing over as-designed and as-built data to the operation phase and eliminate the costly and time consuming paper-based process. This approach can also leverage the effective concurrent collaboration between multiple partners not only within the same project but also with other construction sectors such as city planning, and other civil infrastructure (pipeline, railway, water supply, etc.). Once local data sets can be instantly accessed by other related disciplines, better decision making with holistic and long-term benefits would be achieved.
- Future research, I plan to expand

Paragraph 3 - Past research 2 – Inconsistency of data name among different sources

- Data terminology discrepancy is a crucial issue and is a challenging roadblock to integrating, exchange and sharing data among multiple disciplines, partner, geographic regions.
- The inconsistency of data terminology due to the fragmented nature of the highway industry has imposed big challenges on integrating digital data from distinct sources. The issue of semantic heterogeneity may lead to the lack of common understanding of the same data between the sender and receiver. Explicit semantic relations among terms in digital dictionaries, such as ontologies can enable the meaning of a roadway concept name to be transparent and unambiguously understood by computer systems. However, due to the lack of an effective automated method, current practices of identifying these relations hardly rely on a manual process of knowledge acquisition from domain experts or text documents which is laborious and time-consuming.
- The lack of common understanding of the same data represented in different terms leads to the failure of data exchange or the extraction of wrong data. This proposed research will provide the civil infrastructure industry (specifically, the highway industry) with a powerful algorithm that will be based on the recent advancements in Natural Language Processing (NLP) to recognize user's intention from their natural language input. A semantic matching algorithm will be designed to enable the automated extraction of the entity names having meanings equivalent to the user's desired data.
- The diversity of data terminology is a big hurdle to the computer-to-computer communication when computer is not yet able to precisely understand data meaning; data integration in such a heterogeneous environment is therefore likely lead to failure or extraction of wrong data. This unresolved issue has turned into a big burden to end users who still play as a middleware in digital data exchange; and in many cases, it is impossible for them to deeply understand the structure and meaning of data labels stored in large and complex datasets to precisely extract subsets of data.

- To enable semantic transparency for commonly used technical terms among highway agencies across the United State, I have developed a computational infrastructure that supports automated development of a machine-readable dictionary of American-English civil engineering terms. The proposed platform leverage Natural Language Processing (NLP) techniques and machine learning to extract English-American roadway terms and their meanings from natural language technical documents for instance roadway design manuals and specifications.
- This NLP based methodology is expected to assist professionals in extracting roadway terms and their semantic relations from text documents. The present framework is not to completely eliminate human involvement, but is expected to significantly reduce manual efforts and become an enabling tool that can help researchers in the highway domain quickly develop supporting ontologies and other forms of semantic resources for their specific use cases. With respect to facilitating semantic interoperability for the civl infrastructure sector, the findings of this study would accelerate the process of removing the current bottleneck in extensive machine readable dictionaries which are required for an unambiguous data sharing, integration or exchange.
- In future research I plan to…Value of future research

Paragraph 4 – Current research – Natural language based data retrieval engine for highway projects.

- I have recently developed a successful NSF proposal (PI: Dr. Jeong) which is awarded for mostly $300,000 to develop a computational theory and its platform that can analyze users' plain English data requirements, and automatically match their intention to the data entities in the heterogeneous source datasets based on the semantic equivalence.
- Simple and easy extraction of desired data from large and complex digital infrastructure data critically decides the degree of reusability of up-stream digital models and their associated project data. However, the state-of-practices on digital data retrieval in the civil infrastructure domain shows that the traditional ad-hoc data query, which is manual and error-prone, has imposed big burdens on professionals. Users are required to have deep understanding of data structure, meanings behind each data label and a query language. This is especially challenging for the civil infrastructure domain since data are commonly gathered from multiple sources with possible conflict of data terminology.
- This research proposes a novel approach for a fast and unambiguous reuse of digital models for the civil infrastructure industry by developing an automated data retrieval engine which is capable of recognizing user intention from their natural language queries (e.g., words, phrases, questions) and extracting the desired data from heterogeneous digital datasets.
- In this NSF project, I currently the lead researcher in an interdisciplinary project team of Linguistics, Machine Learning and Construction Engineering and Management.
- In order to enable computer systems to understand user's data requirements in natural language, I have been translating domain knowledge in design manuals, guidelines and specification into an extensive machine-readable dictionary for the civil infrastructure using my recently developed NLP-based method as mentioned above.
- Upon finishing this digital dictionary, I will utilize recent advances in Natural Language Processing (NLP) techniques, machine-learning based semantic measure methods to develop the data retrieval system. NLP will be utilized to process and interpret users' natural language inputs. An un-supervised machine learning model will be applied to build up a knowledge base that will consist of formal definitions of technical terms in the civil sector, specifically the highway industry sector.
- The direct intellectual merit of this proposed research is to address the fundamental interoperability issue in digital data exchange and provide the civil infrastructure industry with a user-friendly and automatic data extraction platform.
- This research is expected to make transformative impacts on digital date exchange and sharing in the civil infrastructure industry, and promote and accelerate the industry's transition to the digital

project delivery as digital models and their associated data can be readily and seamlessly reused through the project life cycle.

- This research will fundamentally transform the way data users interact with and query digital modeling data and information in the civil infrastructure domain.
- The research outcomes will provide fundamental tools and resources for other researchers and industry professionals for various text mining and intelligence inference systems which are emerging research areas in the construction industry. Thus, this research will create significant synergistic and ripple effects to the construction industry.
- In order to fill that gap, this research proposes a novel approach for a fast and unambiguous reuse of digital models for the civil infrastructure industry by developing an automated data retrieval engine which is capable of recognizing user intention from their natural language queries (e.g., words, phrases, questions) and extracting the desired data from heterogeneous digital datasets. This research will employ the recent advances in Natural Language Processing (NLP) techniques, machine-learning based semantic measure methods to develop the data retrieval system. NLP will be utilized to process and interpret users' natural language inputs. An un-supervised machine learning model will be applied to build up a knowledge base that will consist of formal definitions of technical terms in the civil sector, specifically the highway industry sector. A knowledge base, also known as a digital dictionary, is a critical component for intelligence systems to perform such knowledge works as interpretation of meaning behind human-oriented inputs. A semantic data retrieval algorithm will be designed to match the user intention to the data entities in the heterogeneous source dataset. The matching will be based on the meaning equivalence rather than the string similarity to eliminate the potential of false extraction due to the issue of data terminology inconsistency between the data creator and the user.
- This research is envisioned to fundamentally transform the way in which professionals interact with complex and non-human-readable digital datasets in the civil infrastructure industry. This project will provide a novel computational infrastructure that will enable users to express data queries in plain English. If successful, the burdens currently imposed on users will be significantly eliminated. Once data extraction from digital models becomes straightforward, the 'bottle neck' of MVD availability is also expected to be removed. The research success will translate into the willingness of accepting digital datasets by all project stakeholders, and the seamless digital data exchange through the project life cycle can be truly achieved.
- I plan to extend research to speech language rather written format that allow user to communicate with digital models using natural speech language.
- Value of future research

Paragraph 4 - Current project 2 – Data and information flow across the life-cycle of transportation assets

- With respect to application research to facilitating digital project delivery for transportation agencies, Iowa Highway Research Board and Mid-West Transportation Center have collaboratively funded $180,000 for research proposal mainly contributed by myself.
- The purpose of this research is to enhance the understanding of data and information workflow during the life-cycle of transportation assets by capturing the industry knowledge and experience and developing a business process map and a data sharing map.
- To accomplish that objective, I have conducted series of working group discussions with various participant professionals from various divisions from different project phases, disciplines involved during the life-cycle of various transportation asset including signs and guardrails.
- Based on these discussions I have identified the workflow that require data sharing during the project life cycle and captured data exchange requirements specifying what data to be shared by whom and to whom.
- For each type of transportation assets, I have developed current practice and ideal process map and exchange requirement documents that show what data, who and when to be transferred to whom.

- The significant improvement of data and information sharing between project participants and across various project development stages is possible with a model based project delivery process, and electronic and digital data transfer systems, which will in turn translate into increased productivity, efficiency in project delivery, accountability, and asset management.
- The proposed research aims to develop a guide to help professionals of State Departments of Transportation (DOTs) understand the digital data and information flow during the project life-cycle for various type of transportation assets including pavements, bridges, culverts, signs, guardrails, etc. The guidebook will include but not limit to the following topics: (1) business use cases in which data sharing between project actors is needed, (2) business processes that define clear sequences of the activities to be performed for data and information sharing and exchange, and expected outcome; and (3) data requirements, data sources, levels of detail, software applications and tools involved in specific data exchange use cases.
- The objective of this study is to capture industry experts' knowledge and needs regarding digital data and information sharing during the life cycle of transportation assets. In order to achieve that goal, Literature review, benchmarking the vertical industry practices, and focus group discussions will be extensively used. A working group will be formed including domain industry professionals with various expertise from State DOTs, consultants, contractors, and software vendors. This focus group discussion will help identify and document the data exchange scenarios, data flows, data requirements, data format, and supporting software applications. Based on discussion results, a process map and a data map will be developed. The process map will show the data exchange processes throughout the project lifecycle, and the data map presents what data required to be shared by whom and to whom and when. The next section specifically describes required work tasks.
- This research will provide knowledge on data flow, value of data generated by different project stakeholders. This understanding will allow researchers and professionals to identify current roadblocks in digital data transferring, find a method for leverage existing data and reduce data-recreation.
- This project is expected to provide a better understanding on data and information flow throughout the lifecycle for various transportation assets. It will significantly enhance the process of data collecting and sharing between project participants. The major deliverables of this project include process maps, data maps and a guidebook for DOTs that can facilitate the implementation of data and information transfer for highway asset management.
- Value of future research

Paragraph 5 – Future research - intelligent digital data and information model and automated data

My central/ primary research question/interests - intelligence system for design alternative selection knowledge base system, reasoning to answer what if questions, nlp research related)

- My research in NLP related research provide profound impact the the construction industry where natural language documents are still play major role in data and information sharing and communication among project stakeholders including project contracts, project inspection reports. I'm enthusiastic to continue pursue my research career in line with the interests in developing both theoretical and applicable platform and computational infrastructure to assist researchers and professionals dig into information included in text documents, for instance translating federal and state design guidance, rules and requirements documented specifications in text documents into an extensive resources of digital constraints to support automated compliance checking, detection of closure statements in contractual documents that involves risk to contractors, process RFI (Request for Information).
- I also plan to expand my research areas to other forms of data presented in other form of natural language beside text, such as video, voice, images. One-call center data is just message transferring to participant operators, in urban cities where numerous underground facilities utilities such as electricity, water pipe and sewer, gas, internet, cables using natural language

speed, excavator, location, locates, excavating, digging. Constructing BIM information models using current method such as Lidar are still have low accuracy and time-processing for those is costly. One-call center receive request and send it to participant members who are operators and owners of facilities who is responsible for locate/marking their utilities by paints or flags. Research is need to allow for construction of a digital library of those utilities. Average speed to answer takes hours and days.

- Moreover, I'm also every interested in intelligence project delivery, virtual reality and intelligence system in design and reasoning system where user interact with components of project (design, planning, decision making, cost scheduling) using natural language. The system can answer what if questions in design, planning of the project. Images and video processing the keep track worker health condition and warning, alert when unusual.
- For educational research project, construction knowledge for instance, "construction activities in foundation construction?" for collecting for learning, where new knowledge and technology, these knowledge has been introduced and available online, a unique system that can, support self-studying construction engineering for both undergraduate and graduate students. This knowledge resources is important for practices in the preliminary design phase when historic design, construction data are value resources for AI (Artificial Intelligence) in learn and predict design, schedule, budget, risks and other aspects of project management.
- Potential sponsors: NSF, NIST, Federal Highway Agency, State DOTs. Benign involved in some project, work with dot, meeting with some other people from many state DOTs, their great interest in transferring to digital data project delivery.

Final paragraph – overall good expressions of my research