Research Statement - Tuyen T. Le

My research interests are in line with data and information technologies, and intelligent systems to enhance life-cycle data utilization in construction and infrastructure management. The adoption of advanced digital technologies such as Building Information Modeling (BIM), Geographic Information System (GIS), or Lidar throughout the project life-cycle has enabled a large portion of asset data to become available in digital format. The improved efficiency in sharing and utilizing of these complex machine-readable datasets, will in turn, translate into increased productivity, cost savings, and efficiency in project delivery and accountability. In the last three years of my doctoral program, I have conducted ample research on reducing manual and duplicate efforts in data collection, processing, and utilization by creating new knowledge and computational infrastructure with regard to 1) life-cycle data linkage, 2) data terminology discrepancy, 3) natural language based data extraction, and 4) data and information flow through transportation assets' life cycle. In my future research, I plan to continue to work on facilitating seamless digital project delivery and asset management and broaden my research focus to the areas of intelligent systems and mining texts, sounds, visuals and other unstructured data.

Life-cycle data space linkage. In the conventional practice, life-cycle data are archived and managed individually by different departments in isolated and heterogeneous data warehouses within a highway agency. To enable decision makers in asset management to effectively reuse digital data created by upstream design and construction partners, I developed a data linking platform that interconnects disparate digital data sources (BIM models, construction and project and asset management systems). The system includes several translators to convert data into graph-based network where relevant data items are linked to one another across the data inventory. I plan to expand this work from management of a single asset to integrated urban and infrastructure management to allow for concurrent collaboration between construction sectors (building, pipeline, railway, water supply, etc.). Once local datasets become accessible to other related disciplines, better decision making with holistic and long-term benefits would be achieved.

Inconsistency of data terminology. Data terminology discrepancy is a big hurdle to integration or sharing of digital data among multiple disciplines, partners, or geographic regions. The lack of common understanding to the same or similar data presented in different terms can lead to the extraction of wrong data or misinterpretation. 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 leverages Natural Language Processing (NLP) techniques and Artificial Intelligence (AI) to extract roadway terms and learn their meanings from roadway design manuals, guidelines and specifications. I have developed an algorithm to classify heterogeneous technical terms into distinct semantic groups that are synonyms, hyponyms, and attributes. In future research, I'm interested in implementing the proposed method in developing a national map of term synonyms among highway agencies which is crucial to avoid mismatches when integrating state historic project data (e.g., cost index, asset condition) to support data-driven infrastructure management.

Natural language based data retrieval engine. Regarding to data acquisition, I have recently developed a successful proposal (PI: Dr. Jeong) which is awarded for mostly $300,000 by National Science Foundation (NSF) to develop a computational partial model extraction engine that allows users to use plain English data requirements to query civil infrastructure digital data. Simplicity in acquisistion of desired data from large and complex machine-readable digital infrastructure data critically decides the degree of reusability of up-stream digital models and their associated project data. However, in the traditional ad-hoc data retrieval approach, extracting digital data, which is a big burdens on professionals, is manual, time-consuming and error-prone. Users are required to have deep understanding of data structure, meanings behind each data label and a query language. This research aims to develope 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. I'm currently a lead researcher in an interdisciplinary project team of Linguistics, Machine Learning and Civil Engineering. In order to enable computer systems to understand users 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 completion of this digital dictionary, I will utilize NLP to develop an algorithm for interpreting and translating users' natural language inputs into machine-readable query codes. I plan to develop a NSF proposal expanding this data retrieval system to speech recognition that allow users to communicate with BIM/CIM models using their voice.

Life-cycle Data and Information Flow. In the spectrum of implementation research, a research proposal mainly contributed by myself is funded by the Iowa Highway Research Board and Mid-West Transportation Center for $180,000 to enhance the understanding of data and information workflow during the life-cycle of transportation assets. To accomplish this objective, I have been conducting a series of working group discussions with professionals from different divisions and disciplines involved during the life-cycle for various type of transportation assets including signs, guardrails, pavements, etc. Using the knowledge captured from these discussions, for each type of asset, I have developed several process maps visualizing the current data workflow and a data exchange matrix specifying when and what data required to be shared by whom and to whom. This understanding of the current life-cycle data flow will allow researchers and professionals to identify current roadblocks in digital data transferring, and provide recommendations for leverage existing data to reduce waste in data-recreation. In my future research agenda, I plan to develop a research proposal targeting NCHRP that aims to develop a national guide that can be used by highway agencies to evaluate their maturity in digital based project delivery, identify requirements and tools that need to help them facilitate seamless digital data transferring throughout the project life-cycle.

Future research. In addition to continuing my research on data sharing and retrieval as discussed above, my long term goals are smart building and civil information model (smart BIM/CIM) and big data analytics. Smart BIM/CIM is my vision for the next generation of digital models which does not contain only data and information but is also integrated with domain knowledge which can perform self-reasoning and response to the *what if* question to assist in design, project planning and in other decision making. I plan to utilize NLP to derive and translate domain knowledge in text documents such as specifications into machine-readable rules, and integrated these digital knowledge into BIM/CIM models to create smart digital models. In addition, I'm also enthusiastic with pursuing my research career goals in the area of data mining from texts, visuals or other types of unstructured data. Unstructured data, particularly text documents such as project contracts, RFIs (Requests for Information), project inspection reports are still play a major role in data and information sharing and communication among project stakeholders. Manually reading, reviewing and analyzing these texts are laborious, time-consuming and error-prone. I'm interested in developing computer-assisted text reader systems to assist researchers and professionals in reviewing project related text documents and digging for further value information. My long-term research focus also include alternative data and information source such as visuals (images, videos) or sounds to reduce data collection efforts for decision making in construction job site and infrastructure management. The success from my previous NSF proposal related to NLP implementation in CIM illustrates promising funding opportunity from national research programs. NSF and NIST (National Institute of Standards and Technology) are some examples of primary potential sponsors for my future research. Besides, being involved in several DOT projects, I have recognized that highway agencies show high interests in transitioning to digital data project delivery. I plan to secure funding from Federal Highway Administration (FHA), National Cooperative Highway Research Program (NCHRP) and State DOTs.