# Natural Language Processing using Deep Learning for Classifying Water Infrastructure Procurement Records and Calculating Unit Costs

by

Milad Khaki

A thesis
presented to the University of Waterloo
in fulfillment of the
thesis requirement for the degree of
Doctor of Philosophy
in
Civil Engineering

Waterloo, Ontario, Canada, 2024

## Examining Committee Membership

The following served on the Examining Committee for this thesis. The decision of the Examining Committee is by majority vote.

| | |
|---|---|
| Supervisors: | Dr. Mark Knight,<br>Professor<br>Dept. of Civil and Environmental Engineering |
| | Dr. Andre Unger,<br>Associate Professor<br>Dept. of Earth and Environmental Sciences |
| Internal Committee Member | Dr. Chris Bachmann<br>Associate Professor<br>Dept. of Civil and Environmental Engineering |
| Internal Committee Member | Dr. Tarek Hegazy<br>Professor,<br>Dept. of Civil and Environmental Engineering |
| Internal-External Committee Member: | Dr. Neil Brisley<br>Associate Professor, School of Accounting and Finance |
| External Committee Member: | Dr. Edward Keedwell<br>Professor, Department of Computer Science<br>University of Exeter, United Kingdom |

## Author's Declaration

I hereby declare that I am the sole author of this thesis. This is a true copy of the thesis, including any required final revisions, as accepted by my examiners.

I understand that my thesis may be made electronically available to the public.

## Abstract

This thesis introduces a novel ontology-based deep learning classification model specifically tailored for civil engineering applications, focusing on automating the extraction and classification of water infrastructure capital works tenders and progress certificates. Utilizing ontology for standardizing tender-bid data and employing Named Entity Recognition (NERC) for item categorization, the model adeptly addresses the challenges posed by the diversity in document styles and formats.

Incorporating Long Short-Term Memory (LSTM) structures within the model enables the learning of both linear and non-linear dependencies between words. This aspect is particularly significant in tackling the unique language constructs present in tender-bid document records. The model's effectiveness is underscored by its impressive classification accuracy, achieving 92.6% for testing data and 98.7% for training data, thereby marking a significant advancement in the field.

The practical application of this model through a web server highlights its adaptability and efficiency in real-world scenarios. The model's role in tasks such as unit cost calculation establishes a new benchmark in the industry, showcasing the thesis's innovative contributions in areas like ontology-based data structuring and LSTM-driven automated unit cost computation.

Looking beyond its current scope, this research holds potential for broader applications and adaptations in different civil engineering domains. It lays the groundwork for future enhancements, including exploring multilingual extensions and specialized approaches aligned with evolving industry trends. This thesis amalgamates data preprocessing, deep learning, and engineering expertise to boost efficiency and accuracy significantly. The unique methodology fosters continuous improvement and broad applicability across different regions. The practical integration of this technology in civil engineering tasks, demonstrated through the web server, opens avenues for further development to encompass a wider array of applications.

Future research directions include refining the framework to cater to the dynamic needs of various civil engineering domains and extending the web server's capabilities for real-time data processing and analysis. Investigating the applicability of this methodology in other engineering or interdisciplinary contexts could also provide valuable insights, extending the utility of this research. This thesis lays a solid foundation for ongoing enhancements in capital work planning and tender contract assessment within the civil engineering industry.

**Acknowledgements**

My profound appreciation and heartfelt gratitude go to my esteemed supervisors, with special regard to Dr. Andre Unger, who has been instrumental throughout my Ph.D. program. His unfaltering guidance, remarkable mentorship, and continuous encouragement of my ideas have been paramount to the success of this research.

In addition, I extend my deepest thanks to my committee members, Dr. Chris Bachmann, Dr. Tarek Hegazi, and Dr. Neil Brisley, whose meticulous evaluation of my research proposal and thesis, along with their perceptive remarks, have greatly enriched my work. I express my sincere gratitude to Dr. Rizwan Younis and Mr. Brendan Shapton, whose constant support has been a crucial pillar throughout this project.

I would also like to acknowledge the cooperation and assistance from three Canadian cities (anonymized), whose invaluable contributions cannot be overstated. Their generous provision of the data required for this research was key in making this project a reality.

Lastly, I wish to extend my sincere appreciation to the external examiner, Dr. Edward Keedwell, whose careful scrutiny and insightful feedback of my work were of immense value. His expert evaluation not only refined my thesis but also deepened my understanding of the subject.

## Dedication

With heartfelt gratitude, I dedicate this thesis to my cherished wife, Nasim. Her unfaltering kindness, devotion, and steadfast support have been my guiding light throughout the course of my Ph.D. journey. Her enduring presence beside me has not only made my dreams achievable but also transformed them into reality.

I also dedicate this to my parents, Mohammad Hossein and Farahnaz. Their boundless love, unwavering support, and persistent encouragement have been instrumental in moulding me into the individual I am today. Their unwavering belief in me serves as the foundational bedrock upon which I've constructed my aspirations and accomplished my goals.

# Table of Contents

# List of Figures

# List of Tables

Heart bathed in wisdom's light, هرگز دل من ز علم محروم نشد

Secrets dwell, out of sight. کم ماند ز اسرار که مفهوم نشد

Fifty-three cycles, quest not done, پنجاه و سه وقت فکر کردم شب و روز

Unknowns many, knowledge won. معلوم شدم که هیچ معلوم نشد

# Chapter 1

# Introduction

Water utilities serve as authoritative entities in every municipality, responsible for water preservation, treatment, distribution, billing, and tasks that ensure residents receive clean and sustainable water. In this context, maintaining and improving existing water infrastructure becomes a long-term objective for each municipality. To achieve this goal, municipalities continuously engage in the planning, maintenance, management, and expansion of both water distribution and wastewater collection systems. As a result, a significant portion of their budget is typically allocated to these tasks.

These responsibilities include annual watermain and sanitary sewer capital works programs, which are the current focus of this thesis. To perform annual planning, the city must prioritize projects requiring immediate action. Once high-priority projects are selected, tenders are issued, and bids may be received from contractors. The bids must be analyzed, and the contract is generally awarded to the lowest reliable bidder. During this process, the engineer's estimate of the project's cost is a significant piece of information.

Engineers encounter several problems during this process. The primary problem is estimating the project price based on historically awarded project prices and inflation in these project costs. However, normalizing historical project information using unit costs has yet to be achieved. Therefore, engineers would only have access to the most recent project information, which is available in electronic format and follows the same formatting. As this task is not standardized, the project value estimation could be biased toward the engineer's judgment.

Once the bids are received, the engineer must perform a bid analysis and decide which contractor should be awarded the project. The bidding process is currently done online, and a certain level of consistency in the bid structure is enforced. However, the breakdown of tasks and assigning different records to different contract parts still depend on the contractor's preference. Therefore, another task assigned to the engineer is to import the submitted bids from various contractors, match different fields that correspond to each other, and create consistent formatting. Unfortunately, this task has no standardized or automated process and must be done manually.

After completing the project, the progress certificate must be cleaned, verified for its integrity, and archived. The archived document is then ready for future processes, such as annual unit cost analysis (for operational planning) or long-term unit cost analysis (required for tactical or strategic planning). However, based on feedback from industrial partners and to the best of our knowledge, this final step is often incomplete, or no trained staff is available to perform it regularly.

## 1.1 Problem State of the Art

In the current landscape, cities are adequately equipped with tools and personnel to effectively manage short- to mid-term planning. However, they encounter difficulties with large-scale analysis due to the enormous quantity of tender documents required and the extensive duration of examination. An increase in complexity, including extra contracts from previous years or contracts with inconsistent terms, could exceed municipal capabilities. Therefore, conducting an analysis spanning several years would require a significant augmentation of employee hours, making long-term planning in this scenario impractical without simplification strategies.

A consistent and standardized data source in municipalities is necessary to accurately calculate the correct inflation rate. Instead, municipalities often resort to financial indices like the consumer price index, designed for consumer goods, which inadequately captures the inflation trend in specific projects such as watermain and sanitary sewer works. Even when the inflation rate for current projects with consistent layouts and electronic format is calculated, referring back to over a decade's worth of information is necessary. For a precise inflation rate, it still needs to be attainable. Given the diverse origins of contract records, inconsistencies can arise during the import process. Such inconsistencies primarily stem from format differences in bidder contracts and alterations in required contract formats instituted by municipalities.

In summary, the issue of record inconsistency in either short-term (operational) or long-term (tactical and strategic) financial planning finds its roots in tender-bid documents. Records can exist in various forms and degrees of accessibility, escalating their challenge. For instance, some records may be archived and only available in paper format or scanned copies that are not tabulated, making them unsuitable for direct use in any analysis. Others might be archived but in electronic form, albeit with inconsistent formatting either in tabulation or arrangement of the records. Lastly, there can be archived records in an electronic format that, despite having consistent formatting, could be more consistent in disaggregating information and assigning the records to their corresponding parts.

Each of these issues presents a significant hurdle in the financial information process. As delineated in the following section, the proposed solution will address each of these issues, illuminating the novel contributions and objectives of this thesis and its contribution to the field of civil engineering.

## 1.2 Proposed Solution and Research Objective

The primary objective of this thesis is to introduce a decision support platform designed to empower municipalities with the capacity to conduct insightful "what if" financial scenarios. This platform envisioned as a tool grounded in historical data, leverages historical watermain and sanitary sewer capital works records, progress certificates, and contract summaries.

The platform harnesses information from historical, tender-bid documents and progress summaries. The system's backbone is an ontology-based deep learning classification model specifically designed to extract and categorize vital data from tenders and progress certificates related to water infrastructure capital works. This model addresses the diversity of styles and formats in these documents by standardizing the data using an ontology. This ontology integrates data and relationships from various contract tables, streamlining the data analysis.

The deep learning model parses tender item descriptions, simplifying the data and enhancing the accuracy in identifying and standardizing tender items. The model's outcome expedites the extracting and consolidating essential information from water infrastructure capital work documents.

The thesis argues that the unit cost calculation process, an integral part of tendering and bid analysis for every city, is well-suited for machine learning applications. This process is prone to errors and lacks standardization, requiring a labour-intensive approach. The integration of machine learning is expected to improve the process, making it more streamlined and standardized. This approach should reduce manual effort and potential errors while enhancing accuracy, consistency, and efficiency in the tendering process.

The central objective is to automate and standardize the calculation of unit costs for watermain and sanitary sewer capital works. The proposed solution offers a web-based front end seamlessly linked to a standardized database and associated tools. This interface ensures consistent access to historical contracts and facilitates the download of standardized revisions of all imported documents.

In summary, this thesis seeks to extend data analysis capabilities within the current industry practices. By accessing information across different times and geographical locations, it aims to enhance the utilization and analysis of project records. The objectives of the proposed solution, which facilitate the realization of the presented concept, are summarized in Figure 1.2.

## 1.3 Literature Review

This literature review provides an overview of the roles of ontology and data provenance within the architecture, engineering, and construction (AEC) industry. The review begins by discussing ontology and its relevance in organizing data from various sources, followed by an overview of the critical role that data provenance plays in ensuring data reliability and trustworthiness.

### 1.3.1 Ontology

The application of semantic web technologies, including ontologies, is vital in the AEC industry, specifically for enhancing interoperability. Ontology languages like OWL, grounded in Description Logic (DL), allow computers to understand and process data, promoting targeted semantic interoperability and efficient data exchange within the industry [Yang and Zhang, 2006], [Abdul-Ghafour et al., 2007], [Pauwels et al., 2011], [Venugopal et al., 2015], [Le and Jeong, 2016], [Hitzler et al., 2012], and [Baader and Nutt, 2003]. The Resource Description Framework (RDF) plays a crucial role in enabling the representation and combination of information from diverse knowledge domains [Schreiber and Raimond, 2014], [Hitzler et al., 2012], [Brickley and Guha, 2014], [Berners-Lee, 2003], [Horrocks et al., 2005], [W3C OWL Working Group, 2012].

This thesis takes the application of ontology in the AEC industry a step further by focusing on its use for standardizing data related to watermain and sanitary sewer systems capital works from multiple municipalities [Abdalla et al., 2015]. It demonstrates how ontology captures the "structure" of information in a standard format, facilitating the assimilation of data from diverse sources that vary in data storage format and granularity levels [Abdalla et al., 2015].

Despite differences in the construction of municipal tender documents, professional engineers can evaluate them and generate estimates, emphasizing the flexibility of ontologies [Zhou et al., 2016]. While this diversity poses challenges for data reuse, ontologies bridge the interoperability gap, facilitating the efficient reuse of previously generated data.

In civil engineering, ontologies, like the ifcOWL ontology, represent knowledge within a specific domain, including building data models, geometries, semantics, relationships, and properties [Rischmoller et al., 2000a], [Schevers and Drogemuller, 2006], [Beetz et al., 2005], [Agostinho et al., 2007], [Zhao and Liu, 2008], [Krima et al., 2009], [Beetz et al., 2009], [Barbau et al., 2012], and [Pauwels et al., 2015]. Extensions to these ontologies encapsulate additional rules and improve type information representation [Terkaj and Sojic, 2015], [Borgo et al., 2015], and [de Farias et al., 2015], demonstrating the role of ontologies in collaborative information management, building performance analysis, and energy management [Shah et al., 2011], [El-Diraby, 2013a], [Ruikar et al., 2007], [Anumba et al., 2008], [Lima et al., 2002], [Lima et al., 2003], [Lima et al., 2005], [Anumba et al., 2008], [Ricquebourg et al., 2007], and [Wicaksono et al., 2010].

The literature provides several examples of successful ontology applications in managing data from different sources [Rahm and Do, 2000], [Costin et al., 2017], [Musen, 1998], and [Yin et al., 2012], underscoring ontology's practicality for integrating multiple data sources and maintaining record quality and integrity.

This thesis fills a significant gap in the literature by focusing on ontologies for specific use cases in civil engineering, such as the rule definitions and knowledge particular to the "Water Systems, Civil Engineering field" [Bilgin et al., 2018] and [Shvaiko and Euzenat, 2005]. This application deviates from traditional ontology construction applications in construction management, emphasizing the ongoing evolution of civil engineering's approach to data standardization.

Semantic web technologies allow multiple ontologies to co-exist and link, often representing the same physical elements [Abdul-Ghafour et al., 2007], enabling efficient integration with systems outside the AEC domain like Geographical Information Systems (GIS) [Metral et al., 2009], [Pileggi and Amor, 2013], and [Metral et al., 2010]. Ontology in civil engineering, as evidenced by the ifcOWL ontology's use in construction and building information management [Kadolsky et al., 2014], [Baumgartel et al., 2014], [Kim and Grobler, 2009], has become essential in the industry.

Ontology rules formally represent domain knowledge critical to automated reasoning and inference, data quality assurance, and decision-making systems support [Stuckenschmidt, 2009]. The advent of semantic web technologies in the AEC domain acknowledges the importance of accurately modelling existing conditions rather than forcing them into a single predefined model [Pauwels et al., 2017], [Rezgui et al., 2011], and [El-Diraby, 2013c].

In conclusion, ontologies have become integral to civil engineering, facilitating semantic interoperability, efficient data exchange, collaborative information management, and standardization. Notwithstanding these developments, the inherent challenges in balancing expressive power and reasoning efficiency, ontologies are providing solutions to complex problems within the industry, underscoring their central role in the evolution of civil engineering [Abdalla et al., 2015], [Li et al., 2015], [Zhou et al., 2016], [Bilgin et al., 2018], and [Shvaiko and Euzenat, 2005].

### 1.3.2 Data Provenance and Quality Management in AEC

The scientific research community places considerable emphasis on data provenance, which traces the origin, lineage, and history of data [Moreau et al., 2013]. Ensuring data quality is crucial for decision-making systems, as inaccurate or incomplete data can lead to erroneous analyses and outcomes, a concern reflected in various studies [Fisher and Kingma, 2001], [Pipino et al., 2002], and [Sadiq et al., 2011]. Although the paper by Khaki [Khaki, 2021] focuses on aspects of data provenance, it is cited here for its broader relevance to the field, despite not aligning directly with the specific aims of this thesis.

Researchers employ an extended set of ontology rules and provenance records to address data errors and ensure data provenance. Ontology rules formally represent domain knowledge, enabling automated reasoning and inference [Stuckenschmidt, 2009]. By enforcing consistency and facilitating automated reasoning, these rules enhance data quality. Provenance records track the origin and lineage of data, aiding in error identification and correction [Moreau et al., 2013]. They provide valuable information about data sources and transformations, enabling researchers to trace data history and ensure its reliability and traceability.

The extended set of ontology rules encompasses a broader collection of rules utilized in ontology-based data management. These rules enforce consistency, enable automated reasoning, and improve data quality. Provenance records complement ontology rules by capturing data origin and lineage, supporting error identification, and facilitating error correction processes. Together, these techniques promote reliable and traceable data management.

A significant concern in the field relates to errors that could emerge while converting hard-copy documents into electronic format, causing a decline in data quality [Kim et al., 2003]. It is, therefore, essential to ensure accurate data provenance via thorough checking and error correction procedures to maintain the integrity of the original document's content [Kim et al., 2003].

In the context of correction of Optical Character Recognition (OCR) errors, identification of errors, data cleaning, and the construction of relational databases, the application of semantic web technologies within the AEC industries could have an indirect impact [Abanda et al., 2013a]. Semantic web technologies could enhance data comprehension, support identifying and correcting OCR errors, and aid in data cleaning and database construction.

However, the challenge of achieving data interoperability introduces potential semantic errors, particularly evident during the heterogeneous Information Delivery Manual (IFC) translation and binding processes across various Building Information Modelling (BIM) authoring tools [Lee et al., 2016]. Although not directly tied to data cleaning processes and error identification or correction, these challenges underscore the importance of maintaining the integrity of the IFC data model to ensure data quality.

In conclusion, data provenance, along with the application of ontology rules and semantic web technologies, presents a promising approach to tackle the challenges of error correction, data cleaning, and relational database construction. However, it is crucial to remain vigilant about potential errors introduced during data conversion processes and consider interoperability challenges within specific domains like AEC.

### 1.3.3 Literature Review of Relational Databases in AEC

One of the persistent challenges within the AEC industry, notably in dealing with water infrastructure, is the effective utilization of heterogeneous data. Historically, such data have often been stored in proprietary formats, limiting the capacity for comprehensive exploitation and analysis [Loffredo, 1998] and [Solihin et al., 2017]. Numerous attempts have been made to circumvent this constraint by making the data more accessible. Still, these efforts often limit the scope of available data and confine the capability for ad-hoc queries [Loffredo, 1998].

The industry is gravitating towards a more user-centric approach to tackle this issue. This new direction draws parallels to the data warehouse concept used in general database management systems [Adamson, 2010] and [Kimball and Ross, 2011]. Considerable advancements have been made in developing databases that allow data access beyond vendor-specific Application Programming Interfaces (APIs) [Loffredo, 1998]. The primary focus has been on the IFC model server and query-based systems for BIM data, with the main foundation being relational databases [You et al., 2004], [Beetz et al., 2010], [Mazairac and Beetz, 2013], [Jotne Co., 2014], [Liu et al., 2016], [Khalili and Chua, 2015], [Jiang et al., 2015], and [Li et al., 2016].

However, such systems, despite their innovation, pose significant performance concerns, particularly for complex queries, attributable to the complexity of the STEP model [Ghang et al., 2014], [Jeong et al., 2010], and [Solihin et al., 2017]. In response, semantic web technologies like RDF and OWL are increasingly adopted for information representation and creation of relational databases [Berners-Lee et al., 2001], [Berners-Lee, 2006], [Hausenblas and Kim, 2012], [Abanda et al., 2013b], [Schmachtenberg et al., 2014], and [Auer et al., 2015]. These technologies leverage the power of ontologies to consolidate data from diverse sources, thus simplifying data integration [Musen, 1998], [Yin et al., 2012], [Abdalla et al., 2015], and [Costin et al., 2017].

Ontology usage also assists in maintaining data quality and enhancing decision support systems by providing data provenance, a critical aspect in large data systems [Fisher and Kingma, 2001], [Pipino et al., 2002], [Sadiq et al., 2011], [Moreau et al., 2013], and [Khaki, 2021].

Despite the advances, there remains a need for a more streamlined model for efficient data management, particularly as the industry moves towards data-driven design and maintenance. The existing practices, which can sometimes be inefficient, do not always foster seamless data sharing, creating a significant barrier to disseminating data and findings across various scales [Traver and Ebrahimian, 2017], [Abdallah and Rosenberg, 2019], and [Smith et al., 2023].

Existing data repositories like the EPA's Storage and Retrieval System (STORET) and USGS's National Weather Information System (NWIS) illustrate the hurdles in facilitating smooth interaction between modern water infrastructure data repositories [Chen and Han, 2016] and [Choat et al., 2022]. While efficient in gathering large quantities of data, these systems overlook the need for a controlled language in the collected data, leading to overlaps and ambiguous variable codes that decelerate data querying [Chen and Han, 2016] and [Choat et al., 2022].

Efforts to resolve these issues, such as the development of the Observations Data Model (ODM) and Hydrologic Information System, have proven to be overly complex and computationally expensive for typical water monitoring and management tasks [Horsburgh et al., 2008], [Maidment, 2008], and [Horsburgh et al., 2016].

Recognizing these limitations, there has been a move towards developing a more streamlined model, like the water infrastructure data model, which simplifies data management processes, including data loading, querying, and exporting [Connolly and Beg, 2005]. This model, designed as a multidimensional data cube, organizes metadata through relational tables, making it more suitable for handling vast amounts of generated data.

In conclusion, despite advancements, existing relational database practices in the AEC industry sometimes lack efficiency and speed, highlighting the need for more streamlined models that can handle large data volumes effectively.

### 1.3.4 Contract Processing in AEC

In the recent literature, the role of semantic web technologies in contract processing has gained noticeable traction. While the primary focus of this research is not limited to the Civil Engineering domain, the established principles and techniques offer promising avenues to optimize processes in this field.

The implications of employing an extended set of ontology rules and provenance records extend to the field of building data management. The Linked Building Data (LBD) Community Group, operating within the World Wide Web Consortium (W3C), puts forth a vision of a comprehensive web that interconnects building data [W3C Report, 2014].

Researchers can formally represent domain knowledge specific to building data by leveraging ontology rules, enabling automated reasoning and inference. This enhances data consistency and quality, improving the reliability and usefulness of interconnected building data. In conjunction with ontology rules, provenance records provide valuable insights into the origin and lineage of building data, facilitating error identification and correction processes. The utilization of these techniques supports the establishment of a robust and traceable network of interconnected building data.

In the context of this thesis, understanding and leveraging the concepts of ontology rules and provenance records within the domain of building data management can significantly contribute to achieving the research objectives. Consistency can be ensured by incorporating an extended set of ontology rules, improving the quality of used data. Moreover, the utilization of provenance records allows for tracing the sources and transformations of data, thereby enhancing its reliability and facilitating error identification and correction.

Principles of knowledge representation and reasoning, particularly the Closed World Assumption (CWA) and Open World Assumption (OWA), significantly influence contract processing discussions [Tao et al., 2010], [Perez-Urbina et al., 2012], and [Terkaj and Sojic, 2015]. While not directly relating to contract processing, these assumptions impact the handling of undefined or ambiguous information. The CWA assumes any information not presently known or accessible to be false, a concept prevalent in traditional relational databases. Conversely, the OWA posits that a lack of knowledge does not inherently imply falsity, which is applicable in distributed systems like the World Wide Web, where the entirety of relevant information may not be locally available or explicitly stated. Therefore, representing these assumptions with Web Ontology Language (OWL) might substantially enrich the automatic processing of contracts by augmenting inference capabilities, handling incomplete or implicit information, and hence facilitating more comprehensive contract analysis.

In the context of automatic contract processing, particularly within the architecture, engineering, and construction (AEC) domains, data interoperability, defined as the ability for data from varied sources to function together effectively, plays a pivotal role. Semantic interoperability, a shared understanding of data definitions and meanings, is the industry's objective [Rischmoller et al., 2000b] and [Veltman, 2001]. BuildingSMART International's application of the Industry Foundation Classes (IFC) data model makes strides toward this goal, providing a framework for data exchange across various Building Information Modeling (BIM) authoring tools [International Organization for Standardization, 1994] and [Lee et al., 2016]. However, the IFC model presents challenges, specifically surrounding binding, adaptability, and extensibility [Lee et al., 2016]. Binding refers to linking data to its representative concept or object. Adaptability denotes the data model's ability to evolve in response to industry changes, while extensibility reflects the ease of adding new elements or features to the data model. These elements are essential to maintain up-to-date systems and data coherence [Whyte and Donaldson, 2015] and [Wang et al., 2020].

The development of domain ontologies offers considerable insights into contract processing. For instance, El-Gohary and El-Diraby designed a domain ontology to support knowledge-enabled process management and coordination across various urban infrastructure stakeholders and projects [El-Gohary and El-Diraby, 2010]. Likewise, El-Diraby and Osman developed a domain ontology for construction concepts in urban infrastructure projects [El-Diraby, 2013b]. Such studies indicate a move towards knowledge conceptualization in civil infrastructure, with potential implications for automated contract processing.

In conclusion, while automatic contract processing in Civil Engineering is not yet fully established, the principles and technologies discussed in allied domains provide a foundation for future research and practical applications in this field.

### 1.3.5 Text Classification in AEC

Extracting historical project cost information from municipal tender-bid documents is a complex and time-consuming task. This task becomes more challenging due to diverse project characteristics, expert biases, and unique material and service values [Younis et al., 2016]. Notably, the identification and categorization of input tender items for unit cost calculations or rescaling of historical projects can introduce inconsistencies [Rehan et al., 2016]. The inherent difficulty and complexity of these tasks emphasize the need for automated systems to ensure accuracy and efficiency.

In this regard, Text Categorization (TC) has emerged as a promising solution [Sebastiani, 2002]. Powered by natural language processing algorithms, TC has diverse applications ranging from document organization to classifying newspaper articles by theme [Lindén et al., 2018], text filtering, target audience evaluation [Magdy and Elsayed, 2016], word sense disambiguation [Raganato et al., 2017] and [Navigli, 2009], hierarchical webpage categorization [Qi and Davison, 2009], and sentiment analysis [Dang et al., 2020]. TC algorithms can achieve accuracies of 70%

Tender-bid document records present a unique application of TC, where records need to be classified into standardized categories, a problem known as Named Entity Recognition and Classification (NERC) [Paliouras et al., 2000] and [Isozaki, 2001]. However, conventional machine learning approaches struggle to handle complex sentences, sparse words for classification, and a vast pool of entities requiring classification despite meeting minimum accuracy requirements [Isozaki, 2001] and [Wu et al., 2006].

In the civil engineering domain, the adoption of TC is still in its infancy and has yet to be extensively quantitatively evaluated [Costin et al., 2017]. Current TC algorithms struggle with insufficient training data and fail to achieve the necessary accuracy [Wang and El-Gohary, 2021]. Moreover, the impact of text identification and classification accuracy on analysis results has yet to be thoroughly examined, suggesting that current text classification methods may not be entirely suitable for industry and municipal applications [Zhou and El-Gohary, 2016].

To overcome these shortcomings, this thesis proposes a combination of machine learning and mathematical models. Specifically, it employs deep learning techniques, notably long short-term memory (LSTM) models, for classifying items in watermain and sanitary sewer capital works based on tender-bid documents [Siami-Namini et al., 2019]. This thesis proposes leveraging deep learning techniques, particularly LSTM models, to enhance the classification of items in tender-bid documents, aiming to improve accuracy and efficiency across municipalities.

LSTM models are capable of learning hierarchical representations of data and comprehending complex linguistic structures [Siami-Namini et al., 2019]. These models directly learn from raw text data, eliminating the need for extensive manual feature engineering and improving their performance with increased data. Such features are advantageous when dealing with the voluminous nature of municipal tender documents, enhancing the approach's suitability for identifying tender item types from their descriptions in the context of civil engineering and water system infrastructure capital works.

Although the potential of text mining and machine learning is evident, challenges persist, such as missing data, data inconsistency, and the need for advanced data handling methods [Mohanta and Das, 2016], [Yang and Bayapu, 2020], [Gao and Pishdad-Bozorgi, 2020], and [Christopher Pereira, 2020]. Studies have demonstrated the potential of text classification in addressing these issues. Thereby improving efficiency, rapidly identifying urgent complaints, and enhancing customer satisfaction [Bosch et al., 2005], [Coussement and Van den Poel, 2008], [Pyon et al., 2011], [Hartmann et al., 2019], and [Hong et al., 2022]. These findings underscore the need for a more advanced, automated text classification system, particularly in civil engineering and water infrastructure capital works.

This thesis aims to fill the research gap with a deep learning approach that addresses the need for automated, accurate, and scalable text classification solutions in civil engineering.

Despite the promising applications of text classification in various domains, its potential has not been fully realized, mainly in civil engineering. Considering the immense volume of data in municipal tender documents and the inefficiencies of current manual processes, there is an urgent need for automated, accurate, and scalable text classification solutions. This research gap, coupled with the promising capabilities of machine learning and deep learning techniques, offers a unique opportunity for advancing text classification techniques in civil engineering and water infrastructure capital works. The deep learning approach proposed in this thesis is set to address these needs, paving the way for more efficient and effective management of municipal tender-bid documents.

### 1.3.6 Literature Review Summary

In this dissertation, we delve into the crucial methodologies and technologies utilized within the Architecture, Engineering, and Construction (AEC) sector, mainly concentrating on their application for data standardization and developing systems conducive to efficient data management and processing. The critical areas explored encompass ontology, data provenance, relational databases, contract processing, and text classification.

Ontologies have become pivotal within the AEC domain, facilitating semantic interoperability, effective data exchange, cooperative information management, and data standardization. Their employment in harmonizing data concerning watermain and sanitary sewer systems capital works across various municipalities corroborates these benefits. Such semantic web technologies allow multiple ontologies to function concurrently, facilitating integration with systems beyond the confines of the AEC domain, such as Geographic Information Systems (GIS).

Data provenance, essentially tracking data origin, lineage, and history, fortifies data quality and bolsters decision-making systems. Ontological rules coupled with provenance records significantly overcome hurdles linked to error correction, data cleaning, and the creation of relational databases, particularly in the AEC industry's context. However, the caveat of potential semantic errors arising during data interoperability underscores the need for persistent data quality checks.

Relational databases in the AEC industry, specifically those concerning water infrastructure, have struggled to effectively utilize heterogeneous data due to proprietary storage formats and insufficient comprehensive analysis capabilities. An observable trend toward user-centric databases reflecting the data warehouse concept implies ontologies' critical role in data consolidation and quality assurance. Despite this progress, the pursuit of efficiency and handling capacity remains fraught with challenges, highlighting the necessity for continual research.

The utility of semantic web technologies is also evident in contract processing within the AEC field. The capacity to integrate and interpret varied contract-related data holds the promise of streamlining contract management processes. The implementation of principles like the Closed World Assumption (CWA) and Open World Assumption (OWA) in tandem with the Web Ontology Language (OWL) may facilitate more efficient automatic processing of contracts by adeptly managing incomplete or implicit information.

The final focal point is the growing demand for powerful text classification techniques for extracting and interpreting historical project cost data from municipal tender-bid documents. Machine learning and deep learning techniques, specifically long-short-term memory (LSTM) models, can advance text classification techniques in civil engineering and water infrastructure capital works.

Considering the industry-wide challenges, including disorganized and non-standardized data, the following chapter introduces a methodology designed to tackle these issues, considering the distinct characteristics pertinent to each field. The foundation laid in this chapter offers a foundation for applying the proposed methodology, setting the stage for a comprehensive understanding of the contemporary landscape of data management in the AEC industry.

## 1.4 Methodology

The methodology of the proposed solution starts with the city engineer overseeing a watermain and sanitary sewer capital works project description and seeking to calculate unit costs while preparing to tender the project and receive bids. Next, the engineer transfers the received bids to the proposed system's interface (tablet as the front end). The server side deals with standardization and unit cost analysis. Ultimately, the engineer can review the results through the front-end interface. This methodology contrasts conventional engineering judgment, and the machine learning approach introduced. It illustrates how the proposed solution can simplify the intricate process of unit cost calculation and bid analysis.

Figure 1.1 provides a sample representation of the proposed solution concept, illustrating the functional blocks and the flow of information from project initiation to the final analysis.

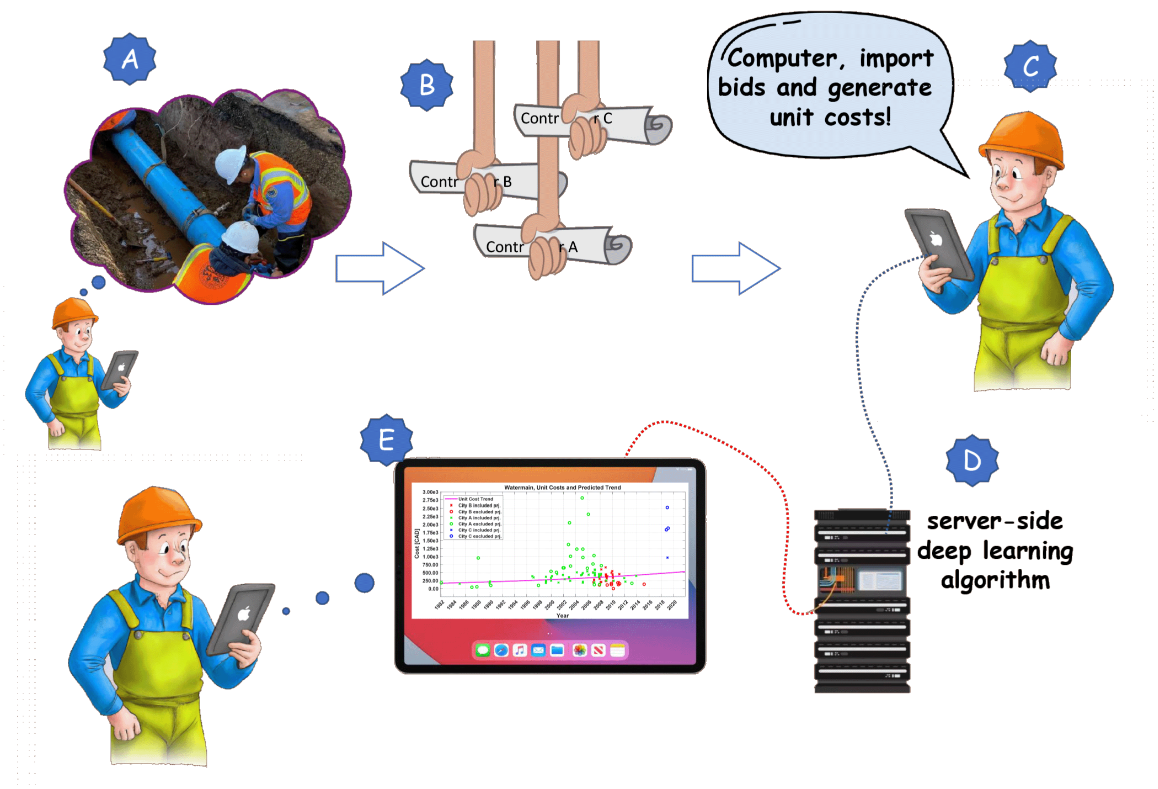



Figure 1.1: Sample representation of the proposed solution concept.

Predecessors of this project developed a unique set of item descriptions to ensure that all similar items have matching descriptions and identifying features (description, unit, and types of items, such as the specific diameter size of copper pipes) [Shapton, 2017]. This standardized rule is a crucial ontology component that safeguards the core database and transforms inconsistent incoming data into standardized revisions.

Data Provenance is another critical aspect of the methodology. The system should be capable of keeping track of changes or corrections made to data so data provenance records are maintained. Provenance records (in the form of metadata) can significantly contribute to cleaning and preserving records when integrating different data sources [Buneman et al., 2001] and [Dai et al., 2008]. In the current project, this essential concept ensures that tracking errors back to their sources is possible. The provenance records correspond to Block C in Figure 1.1.

Furthermore, Block B in Figure 1.1 represents how the two components of standardizing data, the ontology and automatic classification module, collaborate to maintain the integrity of the core database. Ontology rules keep the core dataset compatible with the standard format and layout required by the system. Also, they provide structure and filtering for the newly imported standard tenders. Any incoming new contract must first pass through the ontology rules, and the layout and consistency of the fields need to be checked. Once this step is completed, the classification module performs the subsequent step. The classification module ensures that items are accurately categorized according to the standardized model in this thesis.

Previous studies demonstrate that the results of unit cost index depend on correctly categorizing each item in a tender-bid document [Rehan et al., 2016]. Therefore, the classification module's accuracy is paramount, as it directly determines the reliability of all future analyses. The automatic classification module is responsible for identifying and classifying items. For instance, a sample watermain item can be classified as either a watermain-pipe or a watermain-hydrant item. Only the automatic classification module or a field expert can determine the item's correct mapping. This functionality is implemented using a deep learning-based classifier that leverages long short-term memory (LSTM) blocks, which are specialized for learning patterns in data sequences. The non-linear characteristics of the deep-learning approach, combined with the LSTM, allow the classifier to capture non-linear language constructs available in training data and use them to classify incoming tender-bid documents into their corresponding categories.

Catching to the user's needs is a straightforward task once the data resides in the core database, represented in Block C of Figure 1.1. It requires a simple interface to receive instructions and utilize the available tools for analysis. Steps D, E, and F in Figure 1.1 show this part of the proposed approach. In other words, the main bottleneck preventing municipalities from expanding the analysis and using the wealth of information buried in their historical records is the inability to combine all data sources to perform the required analysis.

## 1.5 Outcomes and novel contributions

The main deliverable of the proposed solution is a standardized database and methodology. This system is designed to import disaggregated data, process both existing and future electronic contracts, and incorporate scanned facsimiles of paper-based historical data. Considering the potential inconsistency and subjective nature of the contract item categorization and analysis, this thesis employs a deep learning-based classification method to allocate each record to its standard category accurately. This approach aims to mitigate the challenges associated with contract item categorization in the civil engineering domain.

Additionally, this project yields the standardization and categorization of tender-bid items: the integration and consolidation of project/contract records. A primary advantage of the standardized database is its ability to adhere to a uniform style and format, regardless of the city or contractor. This consistency simplifies the data processing tasks for the operator, who must only deal with a single data format. Another contribution of this thesis also includes the import and standardization of projects/contract records.

The thesis utilizes the unit cost index calculation method to establish a foundational confidence level for subsequent AI applications. This method is effective when dealing with financial data spanning multiple periods, as it helps adjust for price fluctuations caused by inflation. By mathematically rescaling the financial data, the influence of inflation is minimized, enhancing the accuracy of results in more complex studies.

The unit cost index estimates costs incurred during specified periods within the historical data [Rehan et al., 2016]. This normalization process provides vital insights for cost and price inflation calculations related to watermain and sanitary sewer capital works. Notably, the unit cost results play a crucial role in calculating inflation. A detailed examination of the inflation analysis procedure following the unit cost calculation is expounded in the work of Rehan et al. [Rehan et al., 2016].

Another outcome of the proposed approach is its scalability and extensibility. As the process is automated and requires limited human intervention, it is scalable and can efficiently handle larger volumes of documents from diverse municipalities or regions. Similarly, the proposed solution can be adapted to other industry fields by employing a customized basket of goods and services to develop a bespoke cost index. Thus, the current proposed solution for watermain and sanitary sewer capital works can serve as a model applicable to similar fields, such as building construction or roads.

The subsequent outcome of this thesis is an ensemble of toolboxes specifically designed to allow each municipality to utilize the standardized database. These tools yield valuable insights and recommendations derived from historical data analysis, enabling informed decision-making in watermain and sanitary sewer capital works. The composition and interplay of these toolboxes constitute a crucial part of the capital works plan for water utilities.

The proposed decision support system includes several toolboxes that enhance the efficiency and effectiveness of watermain and sanitary sewer capital works planning and execution. The toolboxes are as follows: The unit cost calculation toolbox, an evolved version of its predecessor proposed by Shapton [Shapton, 2017], enables straightforward computation of unit cost. *The bidding analysis and awarding toolbox* allows the operator to import project bids, evaluate them, and rank them based on selected criteria. *The standard tender summary toolbox* imports a contract and transforms it into the standardized format of the primary database. It enables the operator to download a consistent and revised update. *The contractor profiling toolbox* allows the operator to conduct a meta-analysis on bidders using historical data. It provides insight into project risk, geographical relationships, and bidder behaviours, potentially revealing collusion among frequent bidders. These toolboxes are integral to the proposed decision support system, aiming to improve efficiency in watermain and sanitary sewer capital works planning and execution.

## 1.6 Thesis Organization

Figure 1.2 summarizes the five chapters presented in this thesis and highlights the contributions of the main three chapters (Chapters Two, Three, and Four). Chapter One outlines municipalities' current data management practices for dealing with the volume of data and inconsistencies in their databases. It explains why this situation can escalate into a problem and outlines the implications of such a problem. Chapter One describes the proposed solution and explains how it addresses the identified issue.

Chapter Two begins by discussing municipalities' data challenges and current approaches to addressing them. It then presents the proposed data analysis solution and explains how it can tackle the existing issues. Chapter Three focuses on the next aspect of the problem


Figure 1.2: Thesis chapters objectives, innovations, and contributions

and its solution: identifying and classifying the records in each contract. This chapter introduces the deep learning approach, assesses the details of the method's implementation, and demonstrates its effectiveness in solving the problem based on the obtained results.

Chapter Four illustrates the toolbox generated based on the proposed solution, describes the different components of the solution, and explains how each component addresses a specific aspect of the problem. This chapter summarizes the information presented in the previous chapters by showcasing a sample case of the identified problem in the three reference cities. Chapter Five concludes this thesis by outlining the advantages and disadvantages of the proposed solution and suggesting possible paths to address them in future research.

# Chapter 2

# Record Standardization

## 2.1 Introduction

In this project, three anonymized Canadian cities serve as industrial partners and have provided awarded tender bid documents for watermain and sanitary sewer capital works. These documents, sourced from various departments within each city, offer a wealth of data for analysis. A typical tender consists of items describing combinations of labour, material, and equipment activities associated with watermain and sanitary sewer capital works pertinent to the design drawings of the capital works project. Contractors bidding on the tender must provide a unit cost for each item, with the total bid cost being calculated by summing the products of unit costs and their respective quantities. Despite differences in tender document formatting across the cities, professional engineers must evaluate these documents to produce market-efficient, legally binding estimates. Hence, standardizing and organizing these documents into a database is essential for municipalities. This structure facilitates creating engineering estimates and using historically awarded bid unit costs to calculate inflation in labour, material, and equipment costs.

Municipalities often lack a standardized data format in tender document construction, leading to a 'lack of information interoperability' [Zhou et al., 2016]. Despite this, contracts are structured to enable professional engineers to derive market-efficient, legally binding estimates from design drawings. The central goal of this thesis is to overcome these challenges by standardizing and organizing these documents into a comprehensive database that will allow municipalities to generate more precise engineering estimates and historical cost inflation analyses.

This chapter focuses on the objective of importing existing and future documents governed by the rules of ontology, which represent awarded tender bids for watermain and sanitary sewer capital works in compliance with municipal civil engineering best practices. The importing process must address two essential tasks: (1) integrating multiple data sources and (2) ensuring the quality and integrity of the imported records [Rahm and Do, 2000] and [Batini et al., 2021]. While several methods can be employed for the second task, ontology is particularly well-suited to the current application.

The initial analytical step requires importing and unifying heterogeneous data sources, such as scanned PDFs and electronic spreadsheets. The import and field unification processes carry a significant potential for errors [Devlin and Cote, 1996]. The unified data must be stored in the core database for future use and in-depth analysis. The ontology ensures the preservation of its structure. However, the effectiveness and control of ontology can be application-dependent. In the current case, the ontology consists of a set of rules, restrictions, tables, patterns, and styles defining the data format (both physical and conceptual). For instance, the data being analyzed includes two styles of documents from three cities. One city has an in-house standardized document format encompassing all contracts and bidding documents. Another city outsources the task of issuing tenders, meaning these tenders do not adhere to a standard style. The third city incorporates a combination of both approaches.

Although these documents are valuable in determining construction delays, they lack suitability for precise statistical, financial, or numerical analysis. In contrast, this thesis proposes a solution centred on tender summary documents containing detailed information on water systems capital work, allowing for the evaluation of accuracy and quality through comparisons of numerical results with engineering best practices.

To address this problem, the thesis develops a methodology that leverages natural language processing to standardize a lexicon - i.e., a vocabulary of frequently used terms in the field. This chapter aims to create and implement this lexicon structured within an ontology framework. This lexicon represents civil engineers' professional knowledge when tendering or bidding on watermain and sanitary sewer capital works projects in conjunction with engineering design drawings. Ontology rules, relations, definitions, and constraints underpin the lexicon, ensuring the correctness and integrity of the data structure.

The ontology outlines constraints and rules governing the data, safeguarding its integrity against errors. Furthermore, it allows for the unique recording of the description and cost of each item in a tender or bid document in a database. The resulting description then facilitates a machine learning algorithm to classify each tender item into standard-parts and standard-sub-parts related to watermain and sanitary sewer capital works. It enables the automation of engineer-estimated unit costs and inflation calculations. The primary contribution of the proposed method is the amalgamation of a pre-processing data methodology and a deep learning model. This combination is designed to capture, replicate, and automate professional engineers' expert knowledge in interpreting contracts for watermain and sanitary sewer capital works projects.

This thesis leverages ontology for data standardization and quality assurance in the context of civil engineering, specifically focusing on watermain and sanitary sewer capital works tender documents. The work contributes in three significant ways: firstly, by establishing a unique lexicon specific to these documents; secondly, by formulating an ontology using tailored filters for contextual error detection; and thirdly, by curating a set of common items, initially copied from RS-Means by Rehan et al. [Rehan et al., 2016], pertinent to watermain and sanitary sewer capital works. These items are integral to training an LSTM-based deep learning classifier, detailed in the following chapter, which facilitates the conversion of tender-bid documents from three cities into a standardized database. The classifier's application aims to consistently map contract items to pre-existing classification schemas, offering a solution to the time-consuming task of manual mapping while maintaining or even surpassing its accuracy.

The proposed methodology aims to integrate inconsistent data sources into a unified, standardized core dataset. The proposed methodology is designed to support importing new electronic documents from previously known sources with minimal involvement from the engineer. Furthermore, the methodology can extend to incorporate data from new entities, such as different cities, municipalities, and contractors, ensuring correct storage for future access. The proposed method can deal with the diversity and inconsistency of data formats coming from different cities and contractors. Also, it is resilient to errors occurring in the contract items' descriptions, units, and prices due to using an ontology to clean up errors and deep learning to rectify mistakes in categorization.

Additionally, the proposed approach is flexible enough to accommodate shifts in the style of item descriptions and representations over time, which may result from changes in policy or staffing within a municipality. This important feature can be achieved by retraining the classifiers with updated training samples of new data format (e.g. when a city changes its contract styles in the future). An essential aspect of the proposed method is the implementation of provenance records. These records trace the origins of each piece of data, adding an additional layer of reliability and accuracy to the data set. The provenance records help ensure that the municipality's system remains relevant and functional over time. New records' format, style, and contents will change over time, and error correction methodology has to modify the records accordingly. Data provenance records and the ontology structure can capture the abstract nature of these changes to keep the system from losing functionality for new records and contracts.

### 2.1.1 Flowchart for Importing Tender-Bid Documents

The flowchart of the proposed approach (Figure 2.1) outlines the critical steps and components involved in importing heterogeneous tender-bid documents, their standardization, and storage of the resulting data in a core database. This flowchart illustrates the engineering decision-making process that transforms tender-bid documents into organized, standardized tables apt for inclusion in the core database. These standardized records and documents represent cost estimations for watermain and sanitary sewer capital works. Nevertheless, standardization enables machine learning algorithms to analyze the data, discern patterns, and emulate civil engineering expert knowledge and expertise when classifying each item in a given tender-bid document. Thus, it has potential applications such as computing unit costs for watermain and sanitary sewer capital works.

Essential elements of the flowchart include the data import, data standardization, data storage, item classification (elaborated in Chapter Three), and future adaptability. Figure 2.1 not only represents the standardization process for imported documents and tenders but also differentiates the work detailed in Chapter Three of this thesis, particularly in blocks 2A.9 and 2A.10 (item classification routine). Despite standardizing the imported records (outputs of Chapter Two), further processing via machine learning analysis is necessary.

The following contribution of this thesis involves categorizing and sub-categorizing records according to a standard protocol determined by available training data and a summarized list of RS-Means items, to be examined in the following chapter. The suggested machine learning-based method offers flexibility to adapt to future changes in the item description and representation styles by retraining classifiers using updated training samples of the new data format. In such instances, the entire process envisioned in the flowchart is revisited; however, as the process is automated, it does not impose additional burdens on the engineer in charge.

The central feature of the proposed approach (handling diverse and inconsistent data sources while maintaining resilience to errors in contract item descriptions, units, and prices) relies on the methodology for processing spreadsheets generated by optical character

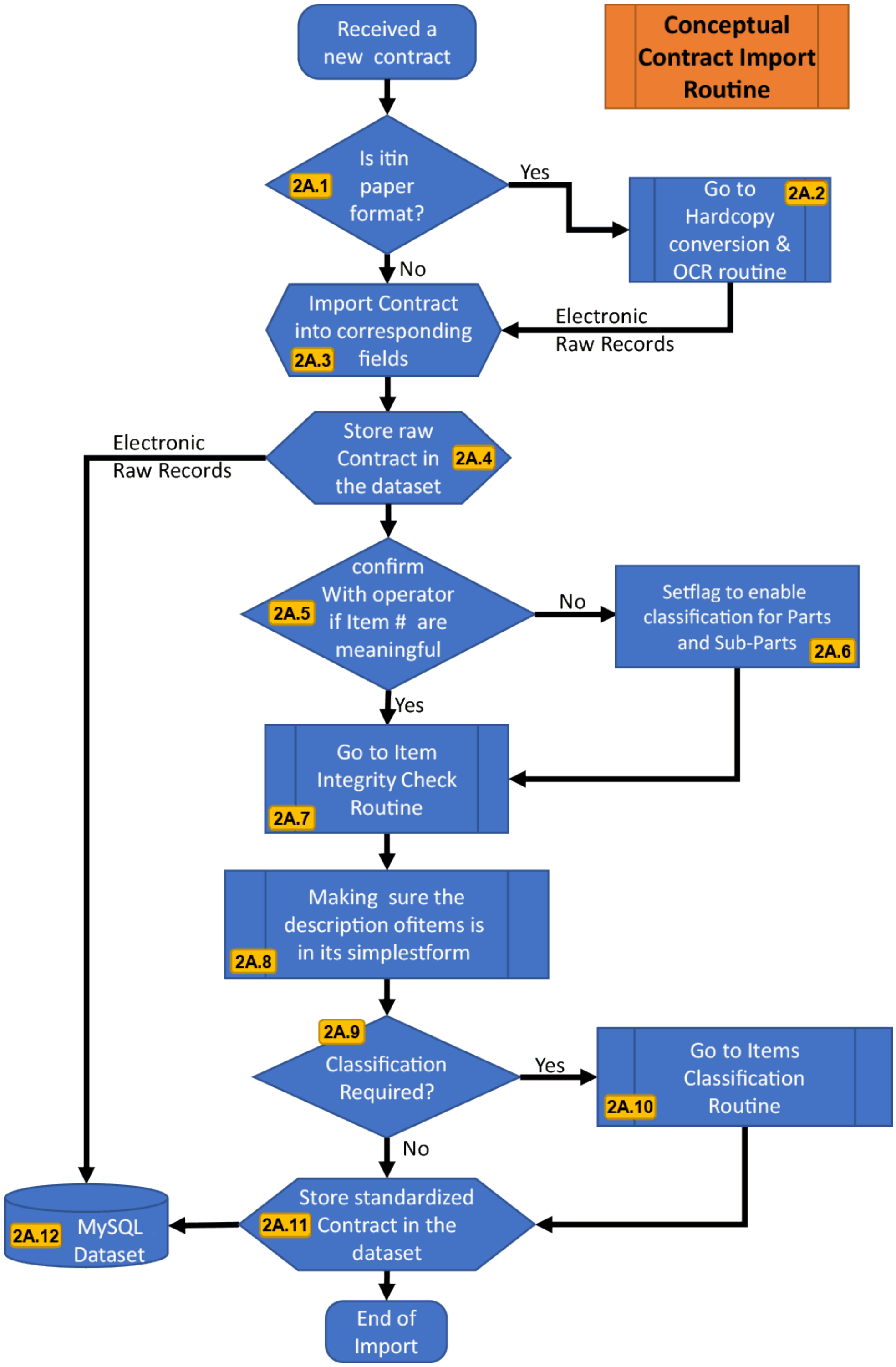



Figure 2.1: flowchart of the process involving the main routine starts with receiving the contract and ends with storing the processed data, ready for analysis in the core database.

recognition (OCR) from paper-based documents. Due to their sub-optimal quality, OCR algorithms may produce errors when converting hard-copy historical documents into importable spreadsheets. The OCR-based tender-bid document handling functionality is represented in blocks 2A.1, 2A.2, and 2A.3. Further details are provided in Appendix A.1, as this is not the main focus of the presented methodology. Another contribution of the proposed method is the implementation of provenance records, which are essential for data curation tasks. Blocks 2A.4 to 2A.10 include assigning and incorporating provenance records within the presented flowchart.

The remainder of this chapter details how the proposed methodology achieves the objectives. The following section discusses importing tender-bid documents and addresses the challenges encountered during this step. Subsequently, the concept of an ontology is introduced, with implementation aspects such as construction, rules, filters, tables, validation techniques, and natural language processing methods. The final section focuses on standardized data storage and access within the core database through the MySQL server. This section also clarifies the nature of provenance records and flags used in the pipeline and their contribution to the safety and auditability of error correction methods. The chapter concludes with a summary, discussion, and recommendations for future steps.

## 2.2 Data and Methodology

This section introduces three anonymized contracts as running examples to illustrate the methodology employed in this thesis. These contracts, labelled Contract A, Contract B, and Contract C, were obtained from three reference cities and provide valuable contract data for watermain and sanitary sewer capital works projects. The original forms of these contracts are presented in raw tables in Table 2.1 for Contract A, Table 2.2 for Contract B, and Table 2.3 for Contract C (located on Pages 28, 30, and 32 respectively). The process of progressively modifying these contracts throughout the pipeline is demonstrated in this and the following chapter. The process transforms the data's raw state into a format suitable for subsequent algorithmic or analysis-based steps.

Each contract represents standard features common to most watermain and sanitary sewer capital works contracts of one anonymized city. Professional engineers participating in or evaluating the tender bidding process readily understand the information presented in these contracts, thanks to their adherence to engineering best practices. Individual cost items are typically organized into rows, each with a unique description and bid price.

However, spreadsheets and items' representation styles and formatting vary across cities, as illustrated by the three representative contracts.

For instance, Contract B presents two sets of prices and quantities, while Contract A only includes one final price and quantity set. Only the final price and quantity values for Contract B are retained to standardize tender-bid documents. The other price fields are omitted as they are irrelevant to the current project and context. Due to non-standardized data presentation, the style and arrangement of information fields were verified with the engineer during import. The specific steps of this "import process" are implemented in the WaterIAM-Khaki system and detailed in block 2A.3 of the flowchart shown in Figure 2.1.

| Item # | DESCRIPTION | Part | Qty | UNIT | UNIT PRICE | TOTAL PRICE |
|---|---|---|---|---|---|---|
| A1 | Bonding | General | 1 | L.S. | 9,700.00 | 9,700.00 |
| A2 | Pre-condition survey | General | 1 | L.S. | 2,000.00 | 2,000.00 |
| | Construction layout & record information | | | | | |
| A5.a | a) layout | General | 1 | L.S. | 3,200.00 | 3,200.00 |
| A5.c | b) progress & final record photography | General | 1 | L.S. | 1,000.00 | 1,000.00 |
| A5.d | c) record survey & drawings | General | 1 | L.S. | 1,300.00 | 1,300.00 |
| F14.a | Clearing & grubbing, a) Remove existing trees & gardens | General | 1 | ea. | 749.00 | 749.00 |
| A11.b | Install, maintain & remove silt control devices - Light duty silt fence barrier, OPSD 219.110 | General | 35 | m | 15.00 | 525.00 |
| A7.c | Construction signs, traffic control & traffic management plan | General | 1 | L.S. | 29,000.00 | 29,000.00 |
| F4 | Trench or road sub-excavation 50 mm crusher-run stone (Prov.) | General | 10 | m3 | 34.00 | 340.00 |
| F5.b | 30Mpa concrete (Prov.) | General | 5 | m3 | 150.00 | 750.00 |
| F6 | Shoring & bracing left in place (Prov.) | General | 10 | m2 | 1.00 | 10.00 |
| F11 | Rock excavation hoe-ramming (Prov.) | General | 10 | m3 | 1.00 | 10.00 |
| F5.b | Unshrinkable fill (Prov.) | General | 5 | m3 | 150.00 | 750.00 |
| A9.a | 19 mm Clear Stone (Prov.) | General | 10 | Tonnes | 26.00 | 260.00 |
| Total General | | | | | | |
| | Test holes to verify depth & location ofinfrastructure | Road | | | | |
| F8.c | a) depth up to 2.0 m | Road | 3 | ea. | 100.00 | 300.00 |
| F8.d | b) depth up to 4.0 m | Road | 3 | ea. | 150.00 | 450.00 |
| F8.f | c) Via Hydro Vac. any depth | Road | 5 | Hrs | 175.00 | 875.00 |
| | Road excavation, removals, and disposal | | | | | |
| E1.b | a) asphalt material to an approved site | Road | 1580 | m2 | 2.00 | 3,160.00 |
| E1.c | b) concrete driveways & sidewalk to an approved site | Road | 260 | m2 | 4.00 | 1,040.00 |
| A3.b | Granular material 'A' | Road | 5600 | Tonnes | 13.00 | 72,800.00 |
| E10.b | Granular material 'M' | Road | 80 | Tonnes | 24.00 | 1,920.00 |
| E5.b | Construct concrete sidewalk any width, OPSD 310.010 & OPSD 310.020 - a) ordinary sidewalk | Road | 280 | m2 | 36.00 | 10,080.00 |
| E9.d | Asphalt milling - a) up to 75 mm depth, including tapers at limits & butt joints | Road | 430 | m2 | 1.00 | 430.00 |
| E4.a.1 | Supply & place hot mix asphalt - a) HL8 HS roadway base asphalt on Flynn Court | Road | 285 | Tonnes | 115.00 | 32,775.00 |
| E6.c | Granular driveway restoration | Road | 80 | m2 | 9.00 | 720.00 |
| E26 | Regrading of ditches & swales | Road | 400 | m | 6.00 | 2,400.00 |
| E7.e | Boulevard grading & sodding - a) grading & sodding | Road | 1035 | m2 | 5.00 | 5,175.00 |
| E7.c | Boulevard grading & sodding - b) supply & placement of 100mm topsoil | Road | 1035 | m2 | 4.00 | 4,140.00 |
| F2 | Supply & apply calcium chloride (Prov.) | Road | 0.5 | Tonnes | 1,500.00 | 750.00 |
| F3 | Application of water (Prov.) | Road | 10 | m3 | 20.00 | 200.00 |
| E33.a | Removal of existing items - a) pipes & culverts | Road | 10 | m | 13.00 | 130.00 |
| E33.c | Removal of existing items - b) fences | Road | 15 | m | 25.00 | 375.00 |
| Total Road | | | | | | |

Table 2.1: Sample running example of a contract from Contract A (City A), Part 1.

| Item # | DESCRIPTION | Part | Qty | UNIT | UNIT PRICE | TOTAL PRICE |
|---|---|---|---|---|---|---|
| C5.a.5 | Sanitary sewer laterals PVC DR-28 building sewer pipe with Class 'B' bedding & Granular 'A' cover & backfill, a) 100mm diameter | SanitarySewer | 30 | m | 94.00 | 2,820.00 |
| C20.a.1 | Reconnect existing sewer services, a) 100mm diameter sanitary lateral | SanitarySewer | 6 | ea. | 332.00 | 1,992.00 |
| C7.c | Flush & TV inspection (a) new sewer pipes | SanitarySewer | 56 | m | 16.00 | 896.00 |
| C7.e | Flush & TV inspection (b) existing sewer service laterals | SanitarySewer | 31 | m | 16.00 | 496.00 |
| C4.a | Maintenance holes standard 1200mm diameter circular precast concrete maintenance hole complete as per OPSD 701.010 a) manhole #SA- 2.0m deep | SanitarySewer | 1 | L.S. | 3,700.00 | 3,700.00 |
| C9 | Cleanouts | SanitarySewer | 6 | ea. | 335.00 | 2,010.00 |
| C6.d | Remove existing sanitary, combined sewer & casing a) maintenance holes | SanitarySewer | 2 | ea. | 341.00 | 682.00 |
| C6.a.1 | Remove existing sanitary, combined sewer & casing b) sewer pipe - 150mm diameter | SanitarySewer | 60 | m | 3.00 | 180.00 |
| | Total Sanitary Sewer | | | | | |
| D16.a.15 | Storm sewers PVC SDR 35 with Granular 'A' bedding, cover & backfill, a) 250mm diameter catchbasin lead | Storm Sewer | 5 | m | 152.00 | 760.00 |
| D16.a.16 | Storm sewers PVC SDR 35 with Granular 'A' bedding, cover & backfill, b) 300mm diameter catchbasin lead | Storm Sewer | 40 | m | 164.00 | 6,560.00 |
| D6.b.1 | Precast concrete catchbasin, b) double, OPSD 705.020 | Storm Sewer | 2 | ea. | 2,237.00 | 4,474.00 |
| | Total Storm Sewer | | | | | |
| | Watermain & large water services PVC Class 150 DR-18 pipe by open-cut | | | | | |
| B1.a.2 | b) 150 mm | Watermain | 144 | m | 110.00 | 15,840.00 |
| B1.a.3 | c) 200 mm | Watermain | 390 | m | 106.00 | 41,340.00 |
| B1.a.6 | d) 50 mm | Watermain | 120 | m | 58.00 | 6,960.00 |
| | Water valves, tapping valves & valve boxes | | | | | |
| B2.b | a) 150 mm diameter valve & box | Watermain | 3 | ea. | 1,200.00 | 3,600.00 |
| B2.c | b) 200 mm diameter valve & box | Watermain | 5 | ea. | 1,700.00 | 8,500.00 |
| B3.a | c) Hydrant sets, OPSD 1105.01 | Watermain | 4 | ea. | 4,600.00 | 18,400.00 |
| | Replace or install new water service with Type 'K' soft copper by open cut, OPSD 1104.01, | | | | | |
| B4.a.1 | a) 20 mm | Watermain | 135 | m | 57.00 | 7,695.00 |
| B17.c | Replace & install new water service with HDPE Series 160 tubing - Flynn Court, a) 20 mm | Watermain | 25 | m | 57.00 | 1,425.00 |
| B7.b | Main stops, a) 20 mm | Watermain | 16 | ea. | 249.00 | 3,984.00 |
| B6.a.2 | Curb stops, a) 20 mm | Watermain | 16 | ea. | 213.00 | 3,408.00 |
| | Reconnection of existing copper services to new main using up to 2 m Type 'K' soft copper pipe, OPSD 1104.01 | | | | | |
| B5.a.1 | a) 20 mm | Watermain | 16 | ea. | 312.00 | 4,992.00 |
| B12 | Watermain disinfection & testing | Watermain | 3 | L.S. | 1,263.00 | 3,789.00 |
| B8 | Remove & replace curb stop & box to property line as required | Watermain | 2 | ea. | 415.00 | 830.00 |
| B29 | Water shut down delays (Prov.) | Watermain | 5 | Hrs | 400.00 | 2,000.00 |
| B22 | Replace water service from property line into building excluding copper piping (Prov.) | Watermain | 4 | ea. | 415.00 | 1,660.00 |
| | Total Watermain | | | | | |

Table 2.1 continued, sample running example of a contract from Contract A (City A), Part 2.

| ITEM # | DESCRIPTION | UNIT | UNIT PRICE | FINAL QNTY | FINAL PRICE | TENDER QNTY | TENDER PRICE |
|---|---|---|---|---|---|---|---|
| PART 1 - ROADS – BEGIN | | | | | | | |
| 1.1 | Remove & dispose of existing concrete, curb & gutter | m | 3.28 | 1230 | 4,03440 | 1000 | 3,280.00 |
| 1.2 | Remove & dispose of existing asphalt driveway/walkway | m2 | 2.22 | 580 | 1,287.60 | 367.2 | 815.18 |
| 1.3 | Remove & store existing interlocking stone driveway/sidewalk | m2 | 25.00 | 110 | 2,750.00 | 20 | 500.00 |
| 1.4 | Remove & store existing wood/stone/brick driveway curbs | m | 10.00 | 40 | 400.00 | 10 | 100.00 |
| 1.6 | Remove & dispose of existing sanitary sewer 200mm diameter | m | 17.10 | 375 | 6,412.50 | 275 | 4,702.50 |
| 1.8 | Remove & dispose of existing watermain, 100mm diameter | m | 12.45 | 150 | 1,867.50 | 150 | 1,867.50 |
| 1.9 | Remove & dispose of existing fire hydrant, catchbasin, storm sewer (375mm, 300mm) | ea. | 311.35 | 3 | 934.05 | 2 | 622.70 |
| 1.10 | Remove & dispose of existing storm manhole | ea. | 156.00 | 4 | 624.00 | 4 | 624.00 |
| TOTAL PART 1 - ROADS – END | | | | | 18,310.05 | | 12,511.88 |
| PART 2 - SANITARY SEWERS | | | | | | | |
| 2.1 | Supply & install 200mm diameter SDR 35 PVC c/w Type 2 bedding, SANMH 104 to 103 | m | 113.75 | 73 | 8,303.75 | 70 | 7,962.50 |
| 2.2 | Supply & install 1200mm diameter precast concrete manhole, SANMH 103 | L.S. | 3,375.00 | 1 | 3,375.00 | 1 | 3,375.00 |
| 2.3 | Supply & install sanitary private drain connection, Connect to 200mm sanitary sewer | ea. | 749.00 | 90 | 67,410.00 | 70 | 52,430.00 |
| 2.4 | Connectto existing sanitary manhole & rebench at Cathcart Street | L.S. | 1,810.00 | 1 | 1,810.00 | 0.9 | 1,629.00 |
| 2.5 | Connectto existing sanitary manhole & rebench at Street | L.S. | 1,810.00 | 1 | 1,810.00 | 1 | 1,810.00 |
| 2.6 | Imported Granular C backfill (Provisional) | tonnes | 9.06 | 2000 | 18,120.00 | 2278.63 | 20,644.40 |
| 2.7 | Video Inspection | m | 3.85 | 533.4 | 2,053.59 | 246 | 947.10 |
| TOTAL PART 2 - SANITARY SEWERS - END | | | | | 102,882.34 | | 88,798.00 |
| PART 3 - STORM SEWERS | | | | | | | |
| 3.1 | All Pipes (375mm, 450mm, 525mm, 675mm) | m | 557.00 | 61 | 33,977.50 | 61.87 | 34,462.80 |
| 3.2 | Supply & install 675mm DIA cone CL100D pipe c/w Class B-1 bedding, STMH 4 to 5 | m | 424.00 | 76 | 32,224.00 | 70 | 29,680.00 |
| 3.3 | Supply & Install 825mm diameter cone CL65D pipe c/w Class B-1 bedding | m | 438.00 | 57.2 | 25,053.60 | 55 | 24,090.00 |
| 3.4 | Supply & install 1200mm diameter precast concrete manhole, STMH 7 | L.S. | 2,565.00 | 1 | 2,565.00 | 1 | 2,565.00 |
| 3.5 | Supply & Install 1500mm diameter precast concrete manhole, STMH 5 | L.S. | 6,610.00 | 1 | 6,610.00 | 0.9 | 5,949.00 |
| 3.6 | Connectto 300 -525mm diameter storm sewer | ea. | 746.00 | 15 | 11,190.00 | 20 | 14,920.00 |
| 3.7 | 150mm diameter SDR 28 PVC pipe from storm sewer to 2.0 m behind curb | m | 97.10 | 250 | 24,275.00 | 120 | 11,652.00 |
| 3.7 | Cleanout (in driveway) | ea. | 142.75 | 10 | 1,427.50 | 2 | 285.50 |
| TOTAL PART 3 - STORM SEWERS | | | | | 137,322.60 | | 123,604.30 |
| PART 4 - WATERMAIN | | | | | | | |
| 4.1 | Supply & install 150mm diameter watermain | m | 80.70 | 600 | 48,420.00 | 323 | 26,066.10 |
| 4.2 | Supply & install 150 off 450 tapping sleeve & valve | ea. | 3,800.00 | 1 | 3,800.00 | 1 | 3,800.00 |
| 4.3 | Supply & install 3-way hydrant c/w 150 x 150 tee, lead, 150 water valve & Storz connection | ea. | 5,400.00 | 2 | 10,800.00 | 1 | 5,400.00 |
| 4.4 | Hydrant extension (Provisional), 300mm | ea. | 565.00 | 1 | 565.00 | 1 | 565.00 |
| 4.5 | Remove & replace existing water service connection (from new 150mm to property line) | | | | | | |
| | 19mm diameter water service (open cut) | m | 87.20 | 700 | 61,040.00 | 350 | 30,520.00 |
| | 19mm main cock | ea. | 273.50 | 70 | 19,145.00 | 37 | 10,119.50 |

Table 2.2: Sample running example of a contract from Contract B (City B), Part 1.

| ITEM # | DESCRIPTION | UNIT | UNIT PRICE | FINAL QNTY | FINAL PRICE | TENDER QNTY | TENDER PRICE |
|---|---|---|---|---|---|---|---|
| 4.6 | Connectto existing water service at property line using vacuum excavation | ea. | 265.00 | 20 | 5,300.00 | 4 | 1,060.00 |
| 4.7 | Cut & cap existing watermain, 150mm | ea. | 1,075.00 | 1 | 1,075.00 | 1 | 1,075.00 |
| 4.8 | Swabbing, flushing, disinfection, test, etc., Wortley Road to Street | L.S. | 2,500.00 | 1 | 2,500.00 | 1 | 2,500.00 |
| 4.9 | Temporary overland water system | | | | | | |
| | connectto existing fire hydrant c/w backflow preventer | ea. | 3,320.00 | 2 | 6,640.00 | 2 | 6,640.00 |
| | temporary water service connection (25mm) | ea. | 61.60 | 75 | 4,620.00 | 74 | 4,558.40 |
| | temporary 100mm connection to Wortley Public school | ea. | 1,190.00 | 1 | 1,190.00 | 1 | 1,190.00 |
| 4.10 | Imported Granular C backfill (Provisional) | tonnes | 9.08 | 1000 | 9,080.00 | 1320.7 | 11,992.00 |
| TOTAL PART 4 – WATERMAINS - END | | | | | 174,175.00 | | 105,486.0 |
| PART 5 - ROADWORKS | | | | | | | |
| 5.1 | Excavation (normal disposal) | m3 | 9.11 | 4300 | 39,173.00 | 1800 | 16,398.00 |
| 5.2 | Subexcavation | m3 | 9.55 | 450 | 4,297.50 | 20 | 191.00 |
| 5.3 | Supply, place & compact granular subbase material, Granular B, Granular A | tonnes | 24.79 | 560.91 | 13,905.00 | 2383.1 | 59,077.10 |
| 5.4 | Supply, place & compact asphalt, HL-3 (fine) driveway asphalt (2010) | tonnes | 153.83 | 120 | 18,459.60 | 60.8 | 9,352.86 |
| 5.5 | Supply & Install concrete curb & gutter, OPSD 600.01 | m | 36.40 | 605 | 22,022.00 | 540 | 19,656.00 |
| 5.6 | Reinstall existing Interlocking stone driveway/sidewalk | m2 | 25.70 | 110 | 2,827.00 | 20 | 514.00 |
| 5.7 | Reinstall existing wood/stone/brick driveway curbs | m | 17.80 | 40 | 712.00 | 10 | 178.00 |
| 5.8 | Supply & place imported topsoil on boulevards | m2 | 4.32 | 3,400.00 | 14,688.00 | 1200 | 5,184.00 |
| 5.9 | Dust control, calcium chloride flakes (40kg) | ea. | 28.5 | 200.00 | 5,700.00 | 72 | 2,052.00 |
| 5.10 | Tree protection fencing | m | 2.38 | 1,500.00 | 3,570.00 | 1450 | 3,451.00 |
| TOTAL PART 5 – ROADWORKS - END | | | | | 12,354.10 | | 116,053.96 |
| PART 6 - MISCELLANEOUS | | | | | | | |
| 6.1 | 50% Labour & material | L.S. | 5,900.00 | 1 | 5,900.00 | 1 | 5,900.00 |
| | 50% Performance | L.S. | 5,900.00 | 1 | 5,900.00 | 1 | 5,900.00 |
| 6.2 | Engineers site trailer | L.S. | 2,450.00 | 1 | 2,450.00 | 0.5 | 1,225.00 |
| 6.3 | Traffic control plan & implementation | L.S. | 35,750.00 | 1 | 35,750.00 | 0.5 | 17,875.00 |
| TOTAL PART 6 – MISCELLANEOUS - END | | | | | 50,000.00 | | 30,900.00 |

Table 2.2 continued, sample running example of a contract from Contract B (City B), Part 2.

| Item No. | Item Description | Unit | Est. Quantity | Est. Rate / Unit | Estimated Total |
|---|---|---|---|---|---|
| | **Part A General - BEGIN** | | | | |
| 1 | tree removal | each | 18 | $ 1,250.00 | $ 22,500.00 |
| 2 | Payment for bonds | L.S. | 1 | $ 17,500.00 | $ 17,500.00 |
| 3 | Payment for all insurance | L.S. | 1 | $ 15,000.00 | $ 15,000.00 |
| 4 | Mobilization and demobilization | L.S. | 1 | $ 35,000.00 | $ 35,000.00 |
| 5 | Field office | each | 1 | $ 10,000.00 | $ 10,000.00 |
| 6 | Traffic Control | L.S. | 1 | $ 40,000.00 | $ 40,000.00 |
| 7 | Capital improvement project construction signs | each | 4 | $ 400.00 | $ 1,600.00 |
| 8 | Construction banners | each | 4 | $ 500.00 | $ 2,000.00 |
| 9 | Prepare hot mix asphalt mix trial batches, all mix types - Superpave mix design method | each | 2 | $ 600.00 | $ 1,200.00 |
| 10 | Pre-construction photos and videos | L.S. | 1 | $ 15,000.00 | $ 15,000.00 |
| 11 | Pre-construction and post-construction condition surveys | L.S. | 1 | $ 22,500.00 | $ 22,500.00 |
| 12 | Test pits as directed backfill with 50 mm crushed ggregate | each | 15 | $ 500.00 | $ 7,500.00 |
| 13 | Test pits as directed - backfill with unshrinkable fill | each | 15 | $ 550.00 | $ 8,250.00 |
| 14 | Large tree removal | each | 7 | $ 1,500.00 | $ 10,500.00 |
| 15 | Small tree removal | each | 12 | $ 1,000.00 | $ 12,000.00 |
| 16 | Provision of as-constructed survey and as-built drawings | L.S. | 1 | $ 25,000.00 | $ 25,000.00 |
| | **TOTAL Part A General - END** | | | | $ 245,50.00 |
| | **Part B Local Roads** | | | | |
| | **Section 1 Sewer** | | | | |
| 1 | Clean out existing catch basins and sumps | each | 30 | $ 85.00 | $ 2,550.00 |
| 2 | Clean, flush and video sanitary and storm sewers and maintenance holes -before construction | m | 1,950.00 | $ 10.00 | $ 19,500.00 |
| 3 | High-pressure flushing for excessive debris in sewers | m | 195 | $ 25.00 | $ 4,875.00 |
| 4 | Temporary class 1 non-woven geotextile fabric silt control for catch basins | each | 30 | $ 35.00 | $ 1,050.00 |
| 5 | Remove and replace cast iron catchbasin frame and grate to raised frame and circular grate | each | 21 | $ 800.00 | $ 16,800.00 |
| 6 | Remove and replace cast iron maintenance hole frame and cover - Type A/B cover and square frame | each | 15 | $ 550.00 | $ 8,250.00 |
| 7 | Remove single catch basins - full depth | each | 2 | $ 1,050.00 | $ 2,100.00 |
| 8 | Supply and install catch basin, ditch inlet catch basin lead and connection to catch basin and sewer | each | 2 | $ 5,100.00 | $ 10,200.00 |
| 9 | Clean, flush and video sanitary and storm sewers and maintenance holes -after construction | m | 1,950.00 | $ 10.00 | $ 19,500.00 |
| 10 | Clean out existing catch basins and sumps | each | 30 | $ 85.00 | $ 2,550.00 |
| 11 | Clean, flush and video sanitary and storm sewers and maintenance holes -before construction | m | 1,950.00 | $ 10.00 | $ 19,500.00 |
| 12 | High-pressure flushing for excessive debris in sewers | m | 195 | $ 25.00 | $ 4,875.00 |
| 13 | Temporary class 1 non-woven geotextile fabric silt control for catch basins | each | 30 | $ 35.00 | $ 1,050.00 |
| 14 | Remove and replace cast iron catchbasin frame and grate to raised frame and circular grate | each | 21 | $ 800.00 | $ 16,800.00 |
| 15 | Remove and replace cast iron maintenance hole frame and cover - Type A/B cover and square frame | each | 15 | $ 550.00 | $ 8,250.00 |
| 16 | Remove single catch basins - full depth | each | 2 | $ 1,050.00 | $ 2,100.00 |
| 17 | Supply and install catch basin, ditch inlet catch basin lead and connection to catch basin and sewer | each | 2 | $ 5,100.00 | $ 10,200.00 |
| 18 | Clean, flush and video sanitary and storm sewers and maintenance holes -after construction | m | 1,950.00 | $ 10.00 | $ 19,500.00 |
| | **Total - Part B Local Roads - END** | | | | $ 169,650.00 |

Table 2.3: Sample running example of a contract from Contract C (City C), Part 1.

| Item No. | Item Description | Unit | Est. Quantity | Est. Rate / Unit | Estimated Total |
|---|---|---|---|---|---|
| | **Part C Watermains - BEGIN** | | | | |
| | **Section II Water** | | | | |
| 1 | 150 mm PVC watermain, CL 235, DR18, within roadway | m | 10 | $ 900.00 | $ 9,000.00 |
| 2 | 200 mm PVC watermain, CL 235, DR18, within roadway | m | 420 | $ 950.00 | $ 399,000.00 |
| 3 | Looping of proposed watermain / water service / fire hydrant lead to avoid conflict with utility or service not shown on drawings -150 mm diameter | each | 2 | $ 1,500.00 | $ 3,000.00 |
| 4 | Looping of proposed watermain / water service to avoid conflict with utility or service not shown on drawings - 200 mm diameter pipe | each | 3 | $ 1,750.00 | $ 5,250.00 |
| 5 | 150 mm gate valve and valve box | each | 2 | $ 2,500.00 | $ 5,000.00 |
| 6 | 200 mm gate valve and valve box | each | 5 | $ 3,000.00 | $ 15,000.00 |
| 7 | New hydrant, complete | each | 3 | $ 12,500.00 | $ 37,500.00 |
| 8 | Connect new watermain to existing watermain, (all sizes) complete | each | 3 | $ 10,000.00 | $ 30,000.00 |
| 9 | Cut and cap the existing watermain ends (all sizes) | each | 9 | $ 850.00 | $ 7,650.00 |
| 10 | Remove existing tee / cross connection from existing watermain and replace with filler piece (all types and sizes) | each | 2 | $ 7,500.00 | $ 15,000.00 |
| 11 | Remove fire hydrant including valve box and capping end | each | 3 | $ 850.00 | $ 2,550.00 |
| | **Subsection 1 Water Services** | | | | |
| 1 | Test pit to investigate condition of water service | each | 40 | $ 500.00 | $ 20,000.00 |
| 2 | Remove and replace non-operational curb stops (all sizes) at streetline including all connections | each | 12 | $ 330.00 | $ 3,960.00 |
| 3 | Cut, extend and reconnect existing 19 mm diameter copper pipe service to new watermain, including any necessary copper pipe to complete | each | 5 | $ 2,700.00 | $ 13,500.00 |
| 4 | Cut, extend and reconnect existing 25 mm diameter copper pipe service to new watermain, including any necessary copper pipe to complete | each | 5 | $ 2,900.00 | $ 14,500.00 |
| 5 | 19 mm diameter copper water service connections to the property line, up to 8 m in length, complete | each | 15 | $ 3,200.00 | $ 48,000.00 |
| 6 | 19 mm diameter copper water service connections to the property line, greater than 8 m in length, complete | each | 15 | $ 3,500.00 | $ 52,500.00 |
| 7 | 25 mm diameter copper water service connections to the property line, up to 8 m in length, complete | each | 9 | $ 3,700.00 | $ 33,300.00 |
| 8 | 25 mm diameter copper water service connections to the property line, greater than 8 m in length, complete | each | 9 | $ 3,900.00 | $ 35,100.00 |
| | **Total - Part C Watermains - END** | | | | $ 749,810.00 |

Table 2.3 continued, sample running example of a contract from Contract C (City C), Part 2.

### 2.2.1 Importing Tenders

Figure 2.1 on Page 25 illustrates the contract importing step of the WaterIAM-Khaki data standardization pipeline. Various issues arise during importing, as demonstrated by examples in Tables 2.1 to 2.3 (on Pages 28 to 32), and are addressed within the WaterIAM-Khaki system. Each case is analyzed, and the corresponding solutions are actively implemented within the WaterIAM-Khaki, although alternative solutions may exist in the literature. To facilitate discussion, each flowchart block is named and numbered.

#### Tender Item Mapping

Industrial partners supplied tender-bid documents from diverse sources (e.g., contractors, companies, municipal engineers), which were not required to follow a standard formatting protocol. They were accepted as long as the tenders contained valid descriptions, quantities, and units. However, this led to variations in tender documents, resulting from factors such as column order, aggregation level preferences, and item categorization. Importing a tender into the core database entails mapping parsed and validated items to a standard set of items represented and curated in the ontology as part of the WaterIAM-Khaki server implementation.

Various approaches can be employed for the mapping process, including element-level integration (language-based, constraint-based, upper-level formal ontologies) and structure-level integration (graph-based, taxonomy-based, model-based) [Ratinov and Roth, 2009]. The methodology used in this project is a combination of element-level and structure-level integration. It offers a new approach within the civil engineering domain and aims to facilitate effective tender item mapping, enhancing data standardization and interoperability. After detailing the implemented methodology for tender item mapping in WaterIAM-Khaki, it is crucial to note and address common challenges encountered in this process.

#### Import issue, item number inconsistency

In the analysis of the three running examples, discrepancies in the "item number" field (Figure 2.2) have been rectified before the import process is complete. In Contract A, the field represents item types determined by municipal engineers' expert knowledge, whereas in Contract B and Contract C, it merely denotes the order of items and parts. The latter cases offer limited information, while the former provides specific details relevant to contractors and municipalities.

To standardize these records, the "item number" column in Contract A is retained for its informative content (e.g., Watermain standard-part and PVC Pipes standard-sub-part can be identified by "B1.a.x" item number). In contrast, the item number can be omitted for Contract B and Contract C as it solely indicates local order in tender documents. Suppose this field is absent from a contract item. In that case, the classification process determines the standard-part and standard-sub-part, with an appropriate provenance tag assigned to indicate post-import classification. Figure 2.1 on Page 25 visualizes this process and its related blocks in block 2A.5. The operator is responsible for clarifying the item number's meaning and flags it accordingly.

| ITEM # | DESCRIPTION |
|---|---|
| PART 1 - ROADS – BEGIN | |
| 1.1 | Remove & dispose of existing concrete, curb & gutter |
| 1.2 | Remove & dispose of existing asphalt driveway/walkway |
| 1.3 | Remove & store existing interlocking stone driveway/sidewalk |
| 1.4 | Remove & store existing wood/stone/brick driveway curbs |
| 1.6 | Remove & dispose of existing sanitary sewer 200mm diameter |
| 1.8 | Remove & dispose of existing watermain, 100mm diameter |
| 1.9 | Remove & dispose of existing fire hydrant, catchbasin, storm sewer (375mm, 300mm) |
| 1.10 | Remove & dispose of existing storm manhole |
| TOTAL PART 1 - ROADS – END | |
| PART 2 - SANITARY SEWERS | |
| 2.1 | Supply & install 200mm diameter SDR 35 PVC c/w Type 2 bedding, SANMH 104 to 103 |
| 2.2 | Supply & install 1200mm diameter precast concrete manhole, SANMH 103 |
| 2.3 | Supply & install sanitary private drain connection, Connect to 200mm sanitary sewer |
| 2.4 | Connect to existing sanitary manhole & rebench at Cathcart Street |

Contract A

| Item # | DESCRIPTION |
|---|---|
| A1 | Bonding |
| A2 | Pre-condition survey |
| | Construction layout & record information |
| A5.a | a) layout |
| A5.c | b) progress & final record photography |
| A5.d | c) record survey & drawings |
| F14.a | Clearing & grubbing, a) Remove existing trees & gardens |
| A11.b | Install, maintain & remove silt control devices - Light duty fence barrier, OPSD 219.110 |
| A7.c | Construction signs, traffic control & traffic management pl |
| F4 | Trench or road sub-excavation 50 mm crusher-run stone ( |
| F5.b | 30Mpa concrete (Prov.) |
| F6 | Shoring & bracing left in place (Prov.) |
| F11 | Rock excavation hoe-ramming (Prov.) |
| F5.b | Unshrinkable fill (Prov.) |
| A9.a | 19 mm Clear Stone (Prov.) |
| Total General | |
| | Test holes to verify depth & location of infrastructure |
| F8.c | a) depth up to 2.0 m |
| F8.d | b) depth up to 4.0 m |
| F8.f | c) Via Hydro Vac. any depth |

Contract B

| Item No. | Item Description |
|---|---|
| | **Part C Watermains - BEGIN** |
| | **Section II Water** |
| 1 | 150 mm PVC watermain, CL 235, DR18, within roadwa |
| 2 | 200 mm PVC watermain, CL 235, DR18, within roadwa |
| 3 | Looping of proposed watermain / water service / fire hydrant lead to avoid conflict with utility or service not shown on drawings -150 mm diameter |
| 4 | Looping of proposed watermain / water service to avoi conflict with utility or service not shown on drawings - 2 mm diameter pipe |
| 5 | 150 mm gate valve and valve box |
| 6 | 200 mm gate valve and valve box |
| 7 | New hydrant, complete |
| 8 | Connect new watermain to existing watermain, (all size complete |
| 9 | Cut and cap the existing watermain ends (all sizes) |
| 10 | Remove existing tee / cross connection from existing watermain and replace with filler piece (all types and sizes) |
| 11 | Remove fire hydrant including valve box and capping e |
| | **Subsection 1 Water Services** |
| 1 | Test pit to investigate condition of water service |
| 2 | Remove and replace non-operational curb stops (all sizes) at streetline including all connections |
| | Cut, extend and reconnect existing 19 mm diameter |

Contract C

Figure 2.2: Comparison of item numbers in three sample contracts.

| ITEM # | DESCRIPTION | UNIT | UNIT PRICE | FINAL QNTY | FINAL PRICE | TENDER QNTY | TENDER PRICE |
|---|---|---|---|---|---|---|---|
| PART 1 - ROADS | | | | | | | |
| 1.9 | Remove & dispose of existing fire hydrant, catchbasin, storm sewer (375mm, 300mm) | ea. | 311.35 | 3 | 934.05 | 2 | 622.70 |

Contract A

| Item # | DESCRIPTION | Part | Qty | UNIT | UNIT PRICE | TOTAL PRICE |
|---|---|---|---|---|---|---|
| B3.a | c) Hydrant sets, OPSD 1105.01 | Watermain | 4 | ea. | 4,600.00 | 18,400.00 |
| | Replace or install new water service with Type 'K' soft copper by open cut, OPSD 1104.01, | | | | | |

Contract B

| Item No. | Item Description | Unit | Est. Quantity | Est. Rate / Unit | Estimated Total |
|---|---|---|---|---|---|
| | **Part C Watermains - BEGIN** | | | | |
| | **Section II Water** | | | | |
| **11** | Remove fire hydrant including valve box and capping end | each | 3 | $ 850.00 | $ 2,550.00 |

Contract C

Figure 2.3: Example of assigning similar items to different categories.

### Import issue, item standard-part identification

Inconsistencies in the import process can arise while mapping the "Part" column. The "Part" field indicates an item's category and may be identified individually (e.g., Contract A) or by the section where the item is defined (e.g., Contract B and Contract C). This field may also be referred to as "Category" or "section" in tender-bid documents, as shown in Figure 2.3.

The system detects a notable inconsistency due to the non-standard naming of the part in Contract B (i.e., both "Road" and "Roadworks" parts, while only "Road" is the standard-part name). Additionally, while Contract A and Contract C categorize services related to hydrants under the "Watermain" part, Contract B misclassifies it in the "Road" part.

The first issue is addressed by mapping both parts to a single standard-part labelled "Road". The system identifies the standard-part for an item based on its description and other field values, leveraging expert knowledge incorporated into its algorithms. The classification algorithm presented in the next chapter is proposed and implemented to automate this process. The part of the WaterIAM-Khaki import routine handling standard-part identification is depicted in blocks 2A.9 and 2A.10 of the flowchart in Figure 2.1. The main objective of the next chapter is to address the issue of missing or standardizing standard-part in records.

In this case, the ontology actively identifies items with inconsistent "Part" fields and rectifies the detected issues. However, standardization of the fields is a task that is not the responsibility of the ontology and is done for almost all imported items by the classification algorithm. The only cases exempted from standardizing are those that the operator flagged as standardized previously. The imported items may follow a standard categorization of items (such as the case of Contract A and City A); however, there is no direct way of checking the compatibility of their standard with the one presented in this work. Thus, the ontology marks all imported items with "not standardized", corresponding to the "Pink Flag" in data provenance records (refer to Section 2.2.5 on Page 65).

#### Import issue, item description inconsistency

When converted from hard copies to electronic format, historical contracts often contain typos and errors arising from limitations in optical character recognition (OCR). For instance, Table 2.2 item 1.1 has an incorrect final price of 4,03440 CAD for concrete disposal in the "Road" part, as shown in Figure 2.4. The correct final price should be 4,034.40 CAD.

| ITEM # | DESCRIPTION | UNIT | UNIT PRICE | FINAL QNTY | FINAL PRICE | TENDER QNTY | TENDER PRICE |
|---|---|---|---|---|---|---|---|
| PART 1 - ROADS – BEGIN | | | | | | | |
| 1.1 | Remove & dispose of existing concrete, curb & gutter | m | 3.28 | 1230 | 4,03440 | 1000 | 3,280.00 |

Figure 2.4: Item 1.1 of Contract B shows an incorrect Final Price because of an OCR error.

The item description is another problem applicable to all three running examples of the contracts. In Table 2.2, item (4.5) of the Watermain part has the description: "remove & replace existing water service connection (from new 150mm to property line)". The original item is shown in Figure 2.5.

| ITEM # | DESCRIPTION | UNIT | UNIT PRICE | FINAL QNTY | FINAL PRICE | TENDER QNTY | TENDER PRICE |
|---|---|---|---|---|---|---|---|
| PART 4 - WATERMAIN | | | | | | | |
| 4.5 | Remove & replace existing water service connection (from new 150mm to property line) | | | | | | |
| | 19mm diameter water service (open cut) | m | 87.20 | 700 | 61,040.00 | 350 | 30,520.00 |
| | 19mm main cock | ea. | 273.50 | 70 | 19,145.00 | 37 | 10,119.50 |

Figure 2.5: example of the non-standard item description.

There are several issues with the current format of the description text that would prevent further analysis:

1. It comes from a table without meaningful item numbers; it requires further processing by the classifier to identify its standard-part and standard-sub-part.

2. As a pre-requisite for automatic classification, the description words should be in their most simplified form to enhance classification accuracy. After using natural language processing rules implemented in the ontology, the text is converted to the following:

> *remove* ***and*** *replace* ***exist*** *water service* ***connect*** *(from new* ***150 mm*** *to property line)*

3. The prepositions, auxiliary words, and punctuation characters are not acceptable for the classifier as they do not contribute to the meaning or classification of the item. Therefore, the description is updated to:

> *remove* ***and*** *replace* ***exist*** *water service* ***connect*** ~~*(from*~~ *new* ***150 mm*** ~~*to*~~ *property line*~~*)*~~

Figure 2.6 presents a flowchart illustrating the data integrity check process for imported items. The initial step (block 2C.1) verifies that the sum of tender document parts matches the total sum. Subsequent steps include validating unit cost and quantity fields (block 2C.2) and ensuring the consistency of the total cost of items with their existing counterparts (block 2C.3). The system includes a process to verify the integrity of an item's description and then updates the text using implemented natural language processing techniques.

We have implemented various filters that programmatically identify and correct issues, as demonstrated in Figures 2.4 and 2.5. These filters include:

- A sum check after importing items ensures the final price matches the provided contract sum if detected by the algorithm.
- Ontology rules that define price ranges for different standard-parts, generated through investigation of similar items and domain expert consultation.
- Rules that cross-check item descriptions against a water infrastructure systems lexicon, flagging terms that do not exist in the dictionary with "Yellow flags." The identified anomalies, once flagged, are addressed to ensure clarity in subsequent data processing steps.
- Rules ensuring correct currency item presentation by checking for "," ".", and "$" characters.

While some steps appear redundant, our analyses show their importance in ensuring data integrity. Ensuring the validity of items passed through OCR or other communication channels is vital for accurate post-processing. Our work with scanned documents from an industrial partner showed that the filters identified several errors that, if unnoticed, would impact the accuracy of the standardized records.

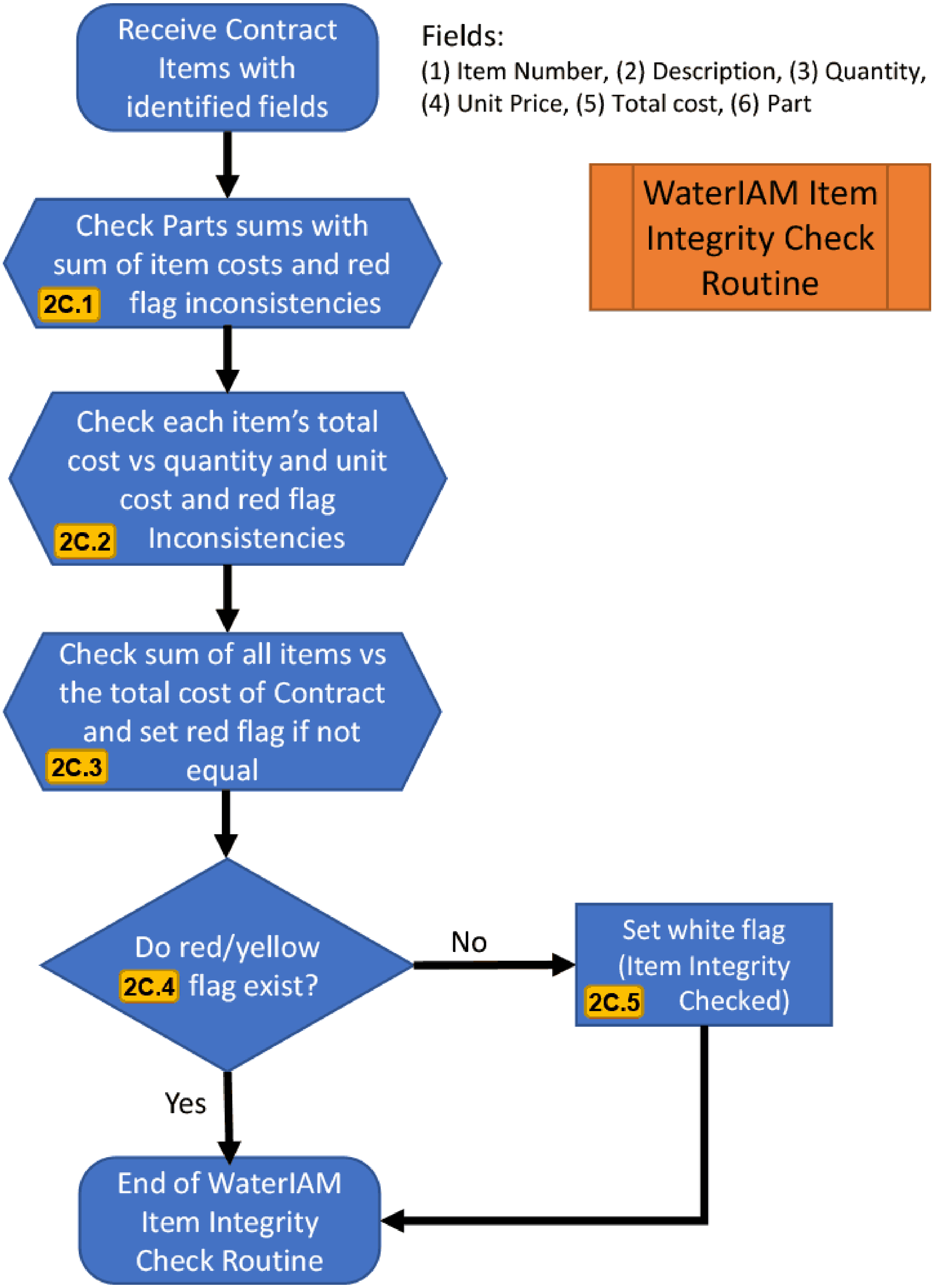



Figure 2.6: Flowchart of the WaterIAM Item Integrity Check Routine, converting paper format documents into tabular form for pre-processing.

Data provenance records are incorporated to address potential errors and track the changes applied to each record at each stage. Despite taking precautions, errors may still be present in the final items, adjusted fields may not reflect correct values, and additional errors may be introduced during correction and import. Therefore, each error correction step is documented for record integrity and auditability. Raw contract items are stored with a "Black flag" in the core database for reference, and items with provenance Meta-Data are marked with a "Violet flag".

### 2.2.2 Ontology

#### Introduction and Objectives

The ontology is designed to bring uniformity to civil engineering records, particularly focusing on tender bids. It is pivotal in identifying discrepancies in imported documents and aligning records with established standards. The ontology supports a hybrid analysis approach, crucial for enhancing the tender updating process, ensuring unit compatibility, and refining the machine learning classification system. Figure 2.7 illustrates the ontology's role in this context.

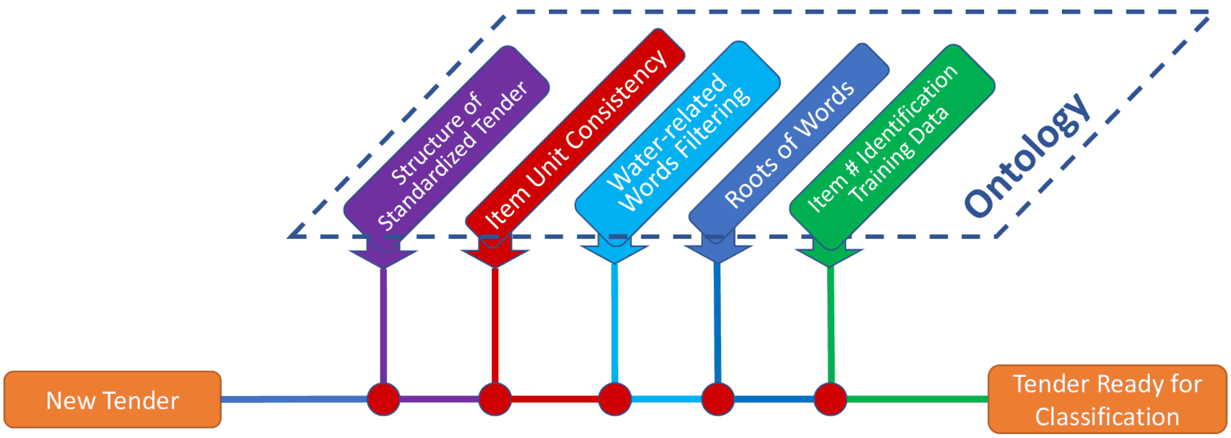



Figure 2.7: The ontology's role in filtering non-compliant terms, simplifying word classification for machine learning analysis, and assisting in the assignment of item numbers or classifications.

#### Construction Methodology

The construction of the ontology involves several sources, including formalized existing knowledge, historical project records, and analytic tools and decision-support systems. The development of the ontology model begins with this formalized knowledge and is validated and expanded through six iterative cycles using tabulated historical data.

#### Structural Components of the Ontology

The ontology is structured hierarchically, comprising classes, subclasses, object properties, and data properties. It includes classes such as Item, Unit, Equipment, and

ClassificationOutput, which are further detailed into subclasses to enhance specificity. Relationships between classes are defined using object properties like hasUnit, hasStndPart, and hasStndSubPart, while attributes of items are captured using data properties such as hasSize, hasDepth, hasCity, and hasContract. Cardinality rules are strictly enforced to maintain data integrity.

### Ontology Implementation and Applications

The ontology aids in standardizing item parts and categories, including normalizing dimensions like diameter and depth. It is instrumental in detecting errors in document imports, a process visualized in Figure 2.8.

Additionally, Figure 2.9 on Page 46 provides an overview of the ontology's data hierarchy, detailing categorizing words and sentences into classes based on their role in knowledge representation.

### Data Pre-processing and Tokenization

Data pre-processing is initiated with thorough cleaning and tokenization to ensure quality, adhering to rules that standardize word forms, and addressing pluralization and irregularities in English. Additionally, a surcharge calculation function is incorporated to adjust the unit price of items.

### Standardized Items and Named Entity Recognition

Standardized items and their categorization under standard-parts and standard-sub-parts are presented in Table 2.4, employing a many-to-one matching strategy.

The structure of the ontology for categorizing various elements related to Sanitary Sewers and Watermain construction projects is outlined in Listing 2.1 on Page 47. The ontology provides a structured and standardized way to represent and classify the various components involved in these projects, ensuring consistency and clarity in the management and documentation of these elements.

| Standard Part | Standard Sub-Part | Ontology Item Count | Sample Item Description | Sample Quantity | Sample Unit | Sample Unit Price |
|---|---|---|---|---|---|---|
| Sanitary Sewer | SS_Manhole | 10 | Standard 1200 mm dia. circular precast concrete maintenance hole complete opsd 701.010 including kor-n-seal equal connections. Depth 3.8 m | 49.6 | Each | $ 199.76 |
| Sanitary Sewer | SS_Lateral | 138 | Grout sewers to be abandoned with unshrinkable fill (maximum 600 mm diameter) | 1 | Lump Sum | $ 1200.00 |
| Sanitary Sewer | SS_Pipe | 96 | STA 5+905 Sanitary sewer laterals - 125mm diameter PVC DR-28 | 8 | m | $ 300.00 |
| Watermain | WM_Pipe | 87 | Watermain PVC Class 150 DR-18 pipe by open cut including bends, fittings, thrust restraints and side street piping (up to connection to existing watermain) 150 mm | 650 | m | $ 3,784.91 |
| Watermain | WM_Hydrant | 13 | Water Valve and Box , Hydrant sets, complete with anchor tee, OPSD 1105.01 | 3 | each | $ 108.02 |
| Watermain | WM_Service | 102 | Any water service not made of copper or less than 20mm dia. copper to be replaced 20 mm | 10 | m | $ 95.53 |
| Watermain | WM_Valve | 77 | Water Valve and Box - 150 mm diameter | 4 | Each | $ 935.76 |
| General | No sub-part | 37 | Install maintain and remove silt control devices | 1 | Lump Sum | $ 3000.00 |
| Provisional Item | No sub-part | 73 | Test holes as directed by the engineer to verify depth and location of infrastructure depth up to 0.5 m | 2 | each | $ 435.00 |
| Road | No sub-part | 265 | concrete any thickness, driveways and sidewalk Road excavation and removals including disposal to an approved site | 500 | m2 | $ 2.38 |
| Storm Sewer | No sub-part | 274 | Precast catchbasins, Single as per OPSD 705.010 | 10 | each | $ 1351.29 |

Table 2.4: Examples of standardized parts and sub-parts with their respective classifications.

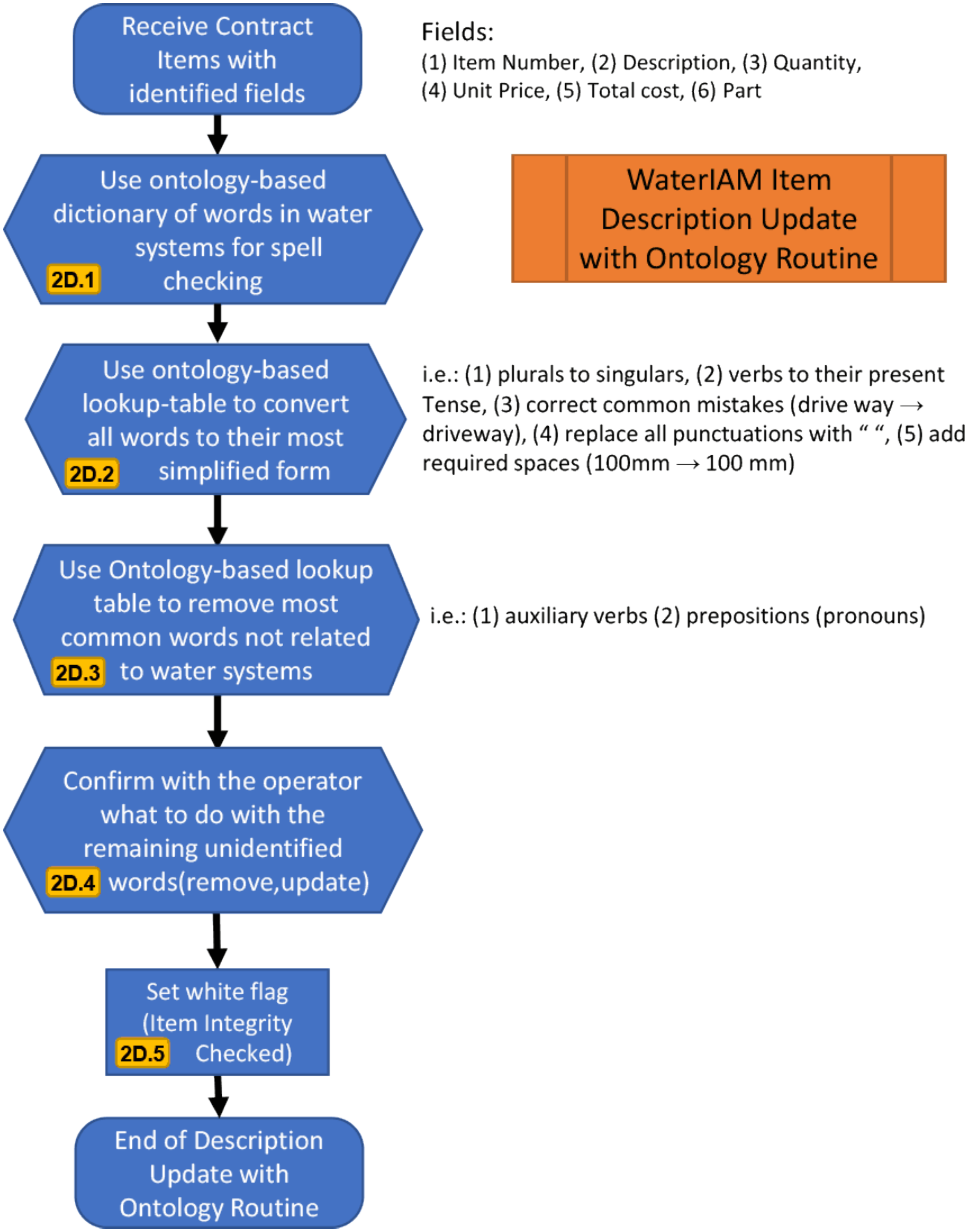



Figure 2.8: Process flowchart for updating WaterIAM item descriptions using the ontology routine.

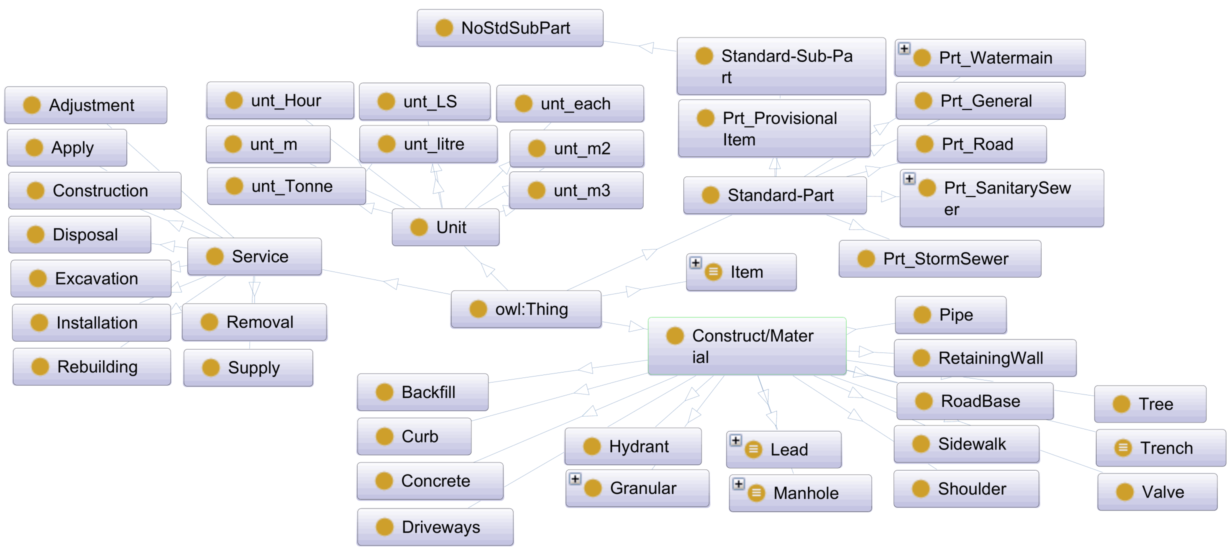



Figure 2.9: Overview of the ontology's data hierarchy, detailing the categorization of words and sentences into classes based on their role in knowledge representation.

Listing 2.1: Ontology representation of sanitary sewers and watermain items and their properties

```
owl:Thing

# Material Class
Class: Material
    SubClass: Backfill
    SubClass: Concrete
    SubClass: Curb
    SubClass: Driveways
    SubClass: Granular
        GranularA
        GranularB
        GranularM
    SubClass: Hydrant
    SubClass: Lead
    SubClass: Manhole
    SubClass: Pipe
    SubClass: RetainingWall
    SubClass: RoadBase
    SubClass: Shoulder
    SubClass: Sidewalk
    SubClass: Tree
    SubClass: Trench
    SubClass: Valve
    ObjectProperty: hasUnit
        Range: Unit
        Cardinality: exactly 1

# Service Class
Class: Service
    SubClass: Adjustment
    SubClass: Apply
    SubClass: Construction
    SubClass: Disposal
    SubClass: Excavation
    SubClass: Installation
    SubClass: Rebuilding
    SubClass: Removal
    SubClass: Supply
    ObjectProperty: hasUnit
        Range: Unit
        Cardinality: exactly 1

# Standard-Part Class
Class: Standard-Part
    SubClass: NoStdSubPart
    SubClass: Prt_General
    SubClass: Prt_ProvisionalItem
    SubClass: Prt_Road
    SubClass: Prt_SanitarySewer
    SubClass: Prt_StormSewer
```

```
    SubClass: Prt_Watermain
    ObjectProperty: hasStndSubPart
        Range: Standard-Sub-Part
        Cardinality: exactly 1

# Standard-Sub-Part Class
Class: Standard-Sub-Part
    SubClass: SS_Lateral
    SubClass: SS_Manhole
    SubClass: SS_Pipe
    SubClass: WM_Hydrant
    SubClass: WM_Pipe
    SubClass: WM_Service
    SubClass: WM_Valve

# Unit Class
Class: Unit
    SubClass: unt_Hour
    SubClass: unt_LS
    SubClass: unt_Tonne
    SubClass: unt_each
    SubClass: unt_litre
    SubClass: unt_m
    SubClass: unt_m2
    SubClass: unt_m3

# Equipment Class
Class: Equipment
    SubClass: SawCutter
    SubClass: Excavator
    SubClass: Bulldozer
    SubClass: Crane
    SubClass: Backhoe

# Item Class
Class: Item
    ObjectProperty: hasStndPart
        Range: Standard-Part
        Cardinality: exactly 1
    ObjectProperty: hasStndSubPart
        Range: Standard-Sub-Part
        Cardinality: exactly 1
    ObjectProperty: hasMaterial
        Range: Material
        Cardinality: exactly 1
    ObjectProperty: hasService
        Range: Service
        Cardinality: exactly 1
    ObjectProperty: hasUnit
        Range: Unit
        Cardinality: exactly 1
    ObjectProperty: hasEquipment
        Range: Equipment
        Cardinality: zero or more
```

```
    ObjectProperty: hasClassificationOutput
        Range: ClassificationOutput
        Cardinality: exactly 1

# Object Properties
ObjectProperty: hasConstruct
    Domain: Item
    Range: Material
    Cardinality: exactly 1
ObjectProperty: hasMaterial
    Domain: Item
    Range: Material
    Cardinality: exactly 1
ObjectProperty: hasService
    Domain: Item
    Range: Service
    Cardinality: exactly 1
ObjectProperty: hasStndPart
    Domain: Item
    Range: Standard-Part
    Cardinality: exactly 1
ObjectProperty: hasStndSubPart
    Domain: Standard-Part
    Range: Standard-Sub-Part
    Cardinality: exactly 1
ObjectProperty: hasUnit
    Domain: {Material, Service, Item}
    Range: Unit
    Cardinality: exactly 1
ObjectProperty: hasEquipment
    Domain: Item
    Range: Equipment
    Cardinality: zero or more
ObjectProperty: hasClassificationOutput
    Domain: Item
    Range: ClassificationOutput
    Cardinality: exactly 1

# Data Properties
DataProperty: hasDepth
DataProperty: hasSize

# Datatypes
Datatypes: integer

# ClassificationOutput Class
Class: ClassificationOutput
    SubClassOf: Thing
    Comment: This class is used to represent the possible classification outputs
     for items in Chapter 3.
    SubClass: Prt_General
    SubClass: Prt_ProvisionalItem
    SubClass: Prt_Road
    SubClass: Prt_SanitarySewer
```

```
        SubClass: SS_Lateral
        SubClass: SS_Manhole
        SubClass: SS_Pipe
    SubClass: Prt_StormSewer
    SubClass: Prt_Watermain
        SubClass: WM_Hydrant
        SubClass: WM_Pipe
        SubClass: WM_Service
        SubClass: WM_Valve
    ObjectProperty: hasClassificationOutput
        Domain: Item
        Range: ClassificationOutput
        Cardinality: exactly 1

Individuals:
  # Maintenance holes standard 1200mm diameter 2.2m deep
  Individual: MaintenanceHole_2_2m
    Type: SS_Manhole
    DataProperty: hasSize
      Value: 1200
      Unit: unt_mm
    DataProperty: hasDepth
      Value: 2.2
      Unit: unt_m
    DataProperty: hasCity
      Value: CityA
    DataProperty: hasContract
      Value: ContractY
    ObjectProperty: hasStndPart
      Value: Prt_SanitarySewer
    ObjectProperty: hasStndSubPart
      Value: SS_Manhole
    ObjectProperty: hasClassificationOutput
      Value: Prt_SanitarySewer

  # Remove existing sanitary sewer pipe 225mm diameter
  Individual: RemoveSewerPipe_225mm
    Type: SS_Lateral
    DataProperty: hasSize
      Value: 225
      Unit: unt_mm
    DataProperty: hasCity
      Value: CityB
    DataProperty: hasContract
      Value: ContractX
    ObjectProperty: hasStndPart
      Value: Prt_SanitarySewer
    ObjectProperty: hasStndSubPart
      Value: SS_Lateral
    ObjectProperty: hasClassificationOutput
      Value: Prt_SanitarySewer

  # Supply and place temporary 150mm diameter bypass waterline
  Individual: TempBypassWaterline_150mm
```

```
  Type: WM_Pipe
  DataProperty: hasSize
    Value: 150
    Unit: mm
  DataProperty: hasCity
    Value: CityC
  DataProperty: hasContract
    Value: ContractZ
  ObjectProperty: hasStndPart
    Value: Prt_Watermain
  ObjectProperty: hasStndSubPart
    Value: WM_Pipe
  ObjectProperty: hasClassificationOutput
    Value: Prt_Watermain

# Supply and install gate valve including valve box and rod 300mm diameter
Individual: GateValve_300mm
  Type: WM_Valve
  DataProperty: hasSize
    Value: 300
    Unit: mm
  DataProperty: hasCity
    Value: CityA
  DataProperty: hasContract
    Value: ContractY
  ObjectProperty: hasStndPart
    Value: Prt_Watermain
  ObjectProperty: hasStndSubPart
    Value: WM_Valve
  ObjectProperty: hasClassificationOutput
    Value: Prt_Watermain

# Supply, place and compact granular subbase materials Granular A
Individual: GranularA_Subbase
  Type: NoSubPart
  DataProperty: hasSize
    Value: GranularA
  DataProperty: hasCity
    Value: CityB
  DataProperty: hasContract
    Value: ContractX
  ObjectProperty: hasStndPart
    Value: Prt_Road
  ObjectProperty: hasStndSubPart
    Value: NoSubPart
  ObjectProperty: hasClassificationOutput
    Value: Prt_Road
```

### Heterogeneous data filtering

Ontologies provide flexibility in specific application contexts and ensure structural consistency. Each ontology revision evolves over time, capturing a particular domain's formalized knowledge. This encapsulation takes the form of a collection of concepts or entities and the relationships that connect these concepts. Using ontologies allows researchers to filter out inconsistent data and ensure only relevant information is analyzed. Ontologies enable researchers to filter out inconsistent data, ensuring the analysis focuses on pertinent information.

Consider the example of water systems: the ontology for this domain would feature entities such as pipes and valves. Our prior knowledge of these systems informs us that the size of pipes and valves in a project should generally correspond unless stated otherwise. Therefore, size becomes an attribute associated with the pipe and valve entities, with the established relationship stipulating that in a connected pipeline, their size attributes ought to align.

Such structured knowledge can generate a data model, creating a knowledge graph. In this graph, the entities and their attributes become nodes and sub-nodes, while the relationships take on the role of edges that link related concepts and attributes together. An ontology, whether in the form of text-based rules, graphical representations, RDF schemas, or as part of a data standardization pipeline, can capture and standardize information structures, facilitating maintenance and updates [Abdalla et al., 2015].

In practical applications, a well-crafted ontology can enable the integration and import of records from various data sources. This feature is leveraged to organize information drawn from multiple municipalities in the current project context. The core information remains consistent even if municipalities present records in various formats with different granularity. By providing a uniform organizational tool, ontology proves invaluable for systematizing data from these heterogeneous sources. This theme, previously introduced in our literature review, gains prominence in the implementation phase.

#### Word-frequency table

A key aspect of ontology is the lexicon, which represents frequently used words in the field of watermain and sanitary sewer systems capital works. Table A.1 in Appendix A.2 on Page 167 displays words with a frequency of 2 or more times in the available documents. A concise version is generated using the running examples from Table 2.1 on Page 28, Table 2.2 on Page 30, and Table 2.3 on Page 32 , represented by Table 2.5 on Page 54. This lexicon combines item descriptions from three running examples (Contract A, Contract B, and Contract C). To maintain consistency, words are converted to their root form (e.g., "connected" or "connection" → "connect") as part of the ontology's standardization framework. Li et al. [Li et al., 2015] present a method for automatically creating domain-oriented term taxonomy using ontology.

Units are standardized (e.g., square meter, m2, sq.m. → m2), and punctuation marks are removed according to ontology guidelines. The resulting table filters prospective documents during importing. Constructing this table necessitates a comprehensive survey of contracts across multiple cities. This approach captures prevalent field-specific words and minimizes the omission of informative terms. The main table (in the Appendix) uses all raw contracts from three cities for the primary lexicon, which is approximately equivalent to 300 tender documents (each tender containing 150 items). The available pool contained about 95,000 words, which, after removing redundancies and overly specific place names, was reduced to 910 words. These words form the corpus of the water system infrastructure description table presented in Appendix A.2.

| Column 1 | Frq1 | Column 2 | Frq2 | Column 3 | Frq3 | Column 4 | Frq4 | Column 5 | Frq5 | Column 6 | Frq6 | Column 7 | Frq7 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| and | 113 | mm | 58 | to | 39 | exist | 32 | remove | 29 | diameter | 27 | service | 23 |
| water | 22 | connect | 20 | supply | 19 | watermain | 19 | catchbasin | 18 | part | 18 | of | 17 |
| install | 16 | pipe | 16 | 150 | 15 | construct | 13 | sanitary | 13 | total | 13 | copper | 12 |
| manhole | 12 | sewer | 12 | storm sew. | 12 | valve | 12 | granular | 11 | replace | 11 | concrete | 10 |
| end | 10 | hole | 10 | maintain | 10 | provincial | 10 | with | 10 | complete | 9 | driveway | 9 |
| flush | 9 | new | 9 | opsd | 9 | 1 | 8 | 200 | 8 | all | 8 | asphalt | 8 |
| clean | 8 | curb | 8 | dispose | 8 | frame | 8 | hydrant | 8 | in | 8 | line | 8 |
| property | 8 | pvc | 8 | road | 8 | 2 | 7 | backfill | 7 | box | 7 | control | 7 |
| cover | 7 | cut | 7 | depth | 7 | Type | 7 | 19 | 6 | as | 6 | bed | 6 |
| class | 6 | for | 6 | lead | 6 | sanit.sew. | 6 | stone | 6 | test | 6 | up | 6 |
| video | 6 | 100 | 5 | 20 | 5 | 50 | 5 | at | 5 | include | 5 | mix | 5 |
| place | 5 | precast | 5 | sidewalk | 5 | temporary | 5 | the | 5 | tree | 5 | 1 | 4 |
| 25 | 4 | 300 | 4 | 4 | 4 | 5 | 4 | 8 | 4 | any | 4 | cast | 4 |
| draw | 4 | excavate | 4 | fire | 4 | from | 4 | grate | 4 | iron | 4 | length | 4 |
| material | 4 | on | 4 | or | 4 | out | 4 | provision | 4 | reconnect | 4 | sdr | 4 |
| silt | 4 | size | 4 | stop | 4 | storm | 4 | survey | 4 | traffic | 4 | 1200 | 3 |
| 3 | 3 | 35 | 3 | begin | 3 | boulevard | 3 | build | 3 | cap | 3 | circular | 3 |
| condition | 3 | ditch | 3 | dr18 | 3 | fence | 3 | general | 3 | grade | 3 | import | 3 |
| inspect | 3 | lateral | 3 | main | 3 | non | 3 | open | 3 | pit | 3 | pre | 3 |
| record | 3 | roadway | 3 | site | 3 | sod | 3 | street | 3 | 10 | 2 | 20 | 2 |
| 103 | 2 | 1104 | 2 | 235 | 2 | 310 | 2 | 375 | 2 | 450 | 2 | 525 | 2 |
| 6 | 2 | 675 | 2 | after | 2 | an | 2 | approve | 2 | avoid | 2 | before | 2 |
| bond | 2 | brick | 2 | by | 2 | calcium | 2 | case | 2 | chloride | 2 | cl | 2 |
| clear | 2 | combine | 2 | compact | 2 | cone | 2 | conflict | 2 | court | 2 | debry | 2 |
| direct | 2 | disinfect | 2 | excessive | 2 | extend | 2 | fabric | 2 | fill | 2 | flynn | 2 |
| full | 2 | gate | 2 | geotextile | 2 | great | 2 | gutter | 2 | high | 2 | hot | 2 |
| inlet | 2 | interlock | 2 | item | 2 | large | 2 | layout | 2 | local | 2 | loop | 2 |
| Misc. | 2 | necessary | 2 | not | 2 | payment | 2 | plan | 2 | pressure | 2 | propose | 2 |
| raise | 2 | rebench | 2 | reinstall | 2 | roadwork | 2 | section | 2 | show | 2 | sign | 2 |
| single | 2 | soft | 2 | square | 2 | store | 2 | subexcavat | 2 | sump | 2 | tap | 2 |
| tee | 2 | than | 2 | topsoil | 2 | tv | 2 | unshrink | 2 | use | 2 | utility | 2 |
| vacuum | 2 | weave | 2 | within | 2 | wood | 2 | wortley | 2 | 0 | 1 | 104 | 1 |
| 110 | 1 | 1105 | 1 | 1500 | 1 | 160 | 1 | 2010 | 1 | 219 | 1 | 250 | 1 |
| 28 | 1 | 30 | 1 | 40 | 1 | 600 | 1 | 7 | 1 | 701 | 1 | 705 | 1 |
| 75 | 1 | 825 | 1 | aggregate | 1 | application | 1 | apply | 1 | backflow | 1 | banner | 1 |
| barrier | 1 | base | 1 | batch | 1 | behind | 1 | brace | 1 | butt | 1 | capital | 1 |
| cathcart | 1 | cl100d | 1 | cl65d | 1 | cock | 1 | cross | 1 | crush | 1 | crusher | 1 |
| culvert | 1 | deep | 1 | delay | 1 | demobilize | 1 | design | 1 | device | 1 | dla | 1 |
| double | 1 | down | 1 | dr28 | 1 | drain | 1 | dust | 1 | duty | 1 | engineer | 1 |
| etc | 1 | exclude | 1 | extension | 1 | field | 1 | filler | 1 | final | 1 | fine | 1 |
| flake | 1 | garden | 1 | grub | 1 | hdpe | 1 | hl3 | 1 | hl8 | 1 | hoe | 1 |
| hs | 1 | hydro | 1 | implement | 1 | improve | 1 | info | 1 | infrastruct | 1 | insurance | 1 |
| into | 1 | investigate | 1 | joint | 1 | kg | 1 | leave | 1 | light | 1 | limit | 1 |
| location | 1 | manage | 1 | method | 1 | mill | 1 | mobilize | 1 | mpa | 1 | normal | 1 |
| number | 1 | off | 1 | office | 1 | operational | 1 | ordinary | 1 | overland | 1 | per | 1 |
| photo | 1 | photograph | 1 | piece | 1 | placement | 1 | post | 1 | prepare | 1 | preventer | 1 |
| private | 1 | progress | 1 | project | 1 | protect | 1 | public | 1 | ram | 1 | regrade | 1 |
| require | 1 | restore | 1 | rock | 1 | run | 1 | school | 1 | series | 1 | set | 1 |
| shore | 1 | shut | 1 | sleeve | 1 | small | 1 | standard | 1 | storz | 1 | streetline | 1 |
| subbase | 1 | subsection | 1 | superpave | 1 | swab | 1 | swale | 1 | system | 1 | taper | 1 |
| trailer | 1 | trench | 1 | trial | 1 | tube | 1 | verify | 1 | via | 1 | walkway | 1 |
| way | 1 | width | 1 | | | | | | | | | | |

Table 2.5: Word frequency table generated from the running example.

#### Field value verification

Field value verification involves using the ontology's training data to identify specific pipe and valve sizes for watermain items, as well as types and dimensions of components for sanitary sewer items. Applying such restrictions during contract item imports allows for detecting contextual errors, rectifying them, or raising error flags for operator review and revision.

For example, a sanitary sewer item with a pipe material (SS_Pipe standard-sub-part) and a pipe size of 160 mm. The ontology-based rules will detect inconsistency (since valid sanitary sewer pipe sizes in this range include 100, 125, 150, and 200 mm) and flag the record as potentially erroneous for operator investigation. We seek expert intervention in such cases, and the item's record receives a "Red flag".

In another scenario, an item belonging to SS_Pipe standard-sub-part is not identified as such. The item will be entered into the database as "standard-parts not detected" until the automatic classification system determines its standard-sub-part. Before classification, the item receives a "Pink flag" for not having a standard-part/standard-sub-part assigned. Once classified, a "Green flag" is assigned, indicating the automatic classification mechanism has determined the standard-part/standard-sub-part, and the record is safe for further processing.

#### Natural Language Processing

Natural Language Processing is employed to standardize and simplify contract language for water systems, compensating for the absence of a unified standard for material and service descriptions. Occasionally, descriptions may include extraneous information, such as street, contractor, or supervisor names, which is irrelevant to the item standard-part and can be excluded from the standardized dataset. Constructing a lexicon for watermain and sanitary sewer capital works prevents unnecessary data from entering the dataset while keeping a copy in the core dataset as raw reference data. The available data can also simplify descriptions by eliminating grammatical tenses and converting all nouns to singular forms. It facilitates the decoding and classification of descriptions for subsequent algorithms.

Different descriptions of the same word are merged (e.g., dozer, bulldozer, and bull dozer become "bulldoze"; driveway, drive way, driveways, drwy, dvwy, dwy are all converted to driveway), and this approach is applied to units consistency. Figure 2.10 displays word clouds based on each field's most frequently used words (watermain on the right and sanitary sewer on the left). Furthermore, Figure 2.8 on Page 45 presents a flowchart of the import routine using ontology to standardize item descriptions. Note that the word clouds serve illustrative purposes and do not possess computational significance.

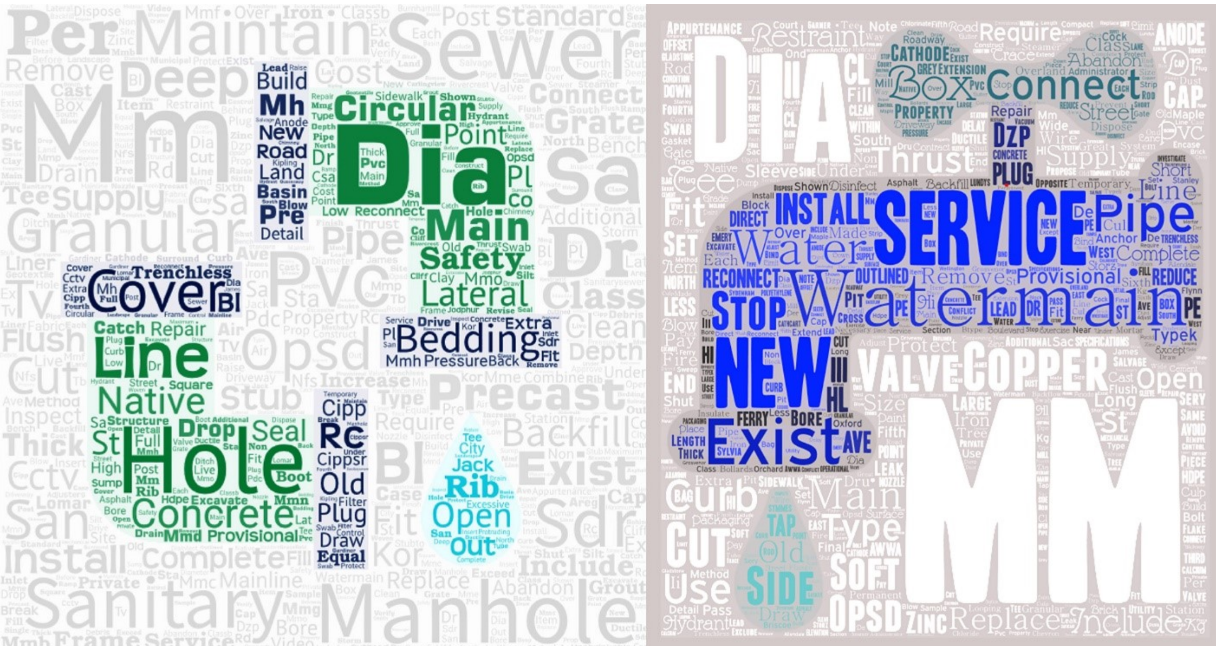



Figure 2.10: Word cloud representations of the ontology tables: Watermain (Right) and Sanitary Sewer (Left). The size of the words is directly proportional to their frequency of occurrence in their respective standard-part.

### 2.2.3 Standardized Data

Once the data is processed through the ontology standardization pipeline discussed in this chapter, it is expected to be of a quality suitable for machine learning classification. The only remaining step is to determine the standard-part and standard-sub-part of the data, which will be addressed in the next chapter.

Tables 2.6, 2.7, and 2.8 display running examples that have undergone all pipeline stages mentioned in this chapter. These tables demonstrate that both item descriptions have been simplified, and units have been made consistent. The part column in these tables is based on information reported in the source document, so the part values at this stage are unreliable for data analysis due to a lack of standardization. Moreover, during our analysis, we encountered parts from new cities not yet included in the standard-part categorization, such as "storm and road", necessitating operator review.

| Item # | Description | Part | Qty | Unit | Unit Price | Total Price | UID # |
|---|---|---|---|---|---|---|---|
| A1 | bond | General | 1 | L.S. | 9,700.00 | 9,700.00 | B_B2 |
| A2 | condition pre survey | General | 1 | L.S. | 2,000.00 | 2,000.00 | B_B3 |
| A5.a | constructinformation layout record | General | 1 | L.S. | 3,200.00 | 3,200.00 | B_B4 |
| A5.c | constructfinal information layout photography progress record | General | 1 | L.S. | 1,000.00 | 1,000.00 | B_B5 |
| A5.d | construct draw information layout record survey | General | 1 | L.S. | 1,300.00 | 1,300.00 | B_B6 |
| F14.a | clear exist garden grub remove tree | General | 1 | ea. | 749.00 | 749.00 | B_B7 |
| A11.b | barrier ctrl dev duty fence install light maintain opsd remove silt | General | 35 | m | 15.00 | 525.00 | B_B8 |
| A7.c | construct ctrl manage plan sign traffic | General | 1 | L.S. | 29,000.00 | 29,000.00 | B_B9 |
| F4 | crush excavate mm prov rd run stone sub trench | General | 10 | m3 | 34.00 | 340.00 | B_B10 |
| F5.b | 30mpa concrete prov | General | 5 | m3 | 150.00 | 750.00 | B_B11 |
| F6 | brace left place prov shore | General | 10 | m2 | 1.00 | 10.00 | B_B12 |
| F11 | excavate hoe prov ramming rock | General | 10 | m3 | 1.00 | 10.00 | B_B13 |
| F5.b | fill prov unshrinkable | General | 5 | m3 | 150.00 | 750.00 | B_B14 |
| A9.a | clear mm prov stone | General | 10 | Tonnes | 26.00 | 260.00 | B_B15 |
| F8.c | depth hole infrastructure locate m test up verify | Road | 3 | ea. | 100.00 | 300.00 | B_B16 |
| F8.d | depth hole infrastructure locate m test up verify | Road | 3 | ea. | 150.00 | 450.00 | B_B 17 |
| F8.f | depth hole hydro infrastructure locate test vac verify via | Road | 5 | Hrs | 175.00 | 875.00 | B_B 18 |
| E1.b | approve asphalt dispose excavate material remove rd site | Road | 1580 | m2 | 2.00 | 3,160.00 | B_B 19 |
| E1.c | concrete dispose drwy excavate remove rd sidewalk site | Road | 260 | m2 | 4.00 | 1,040.00 | B_B 20 |
| A3.b | a granular material | Road | 5600 | Tonnes | 13.00 | 72,800.00 | B_B21 |
| E10.b | granular m material | Road | 80 | Tonnes | 24.00 | 1,920.00 | B_B22 |
| E5.b | concrete construct opsd ordinary sidewalk width | Road | 280 | m2 | 36.00 | 10,080.00 | B_B23 |
| E9.d | asphalt butt depth include joint limit mill mm tapers up | Road | 430 | m2 | 1.00 | 430.00 | B_B25 |
| E4.a.1 | asphalt base hl8 hot hs mix place rdway supply | Road | 285 | Tonnes | 115.00 | 32,775.00 | B_B26 |
| E6.c | drwy granular restore | Road | 80 | m2 | 9.00 | 720.00 | B_B27 |
| E26 | ditch regrade swale | Road | 400 | m | 6.00 | 2,400.00 | B_B28 |
| E7.e | boulevard grade sod | Road | 1035 | m2 | 5.00 | 5,175.00 | B_B29 |
| E7.c | 100 boulevard grade mm place sod supply topsoil | Road | 1035 | m2 | 4.00 | 4,140.00 | B_B30 |
| F2 | apply calcium chloride prov supply | Road | 0.5 | Tonnes | 1,500.00 | 750.00 | B_B31 |
| F3 | application prov water | Road | 10 | m3 | 20.00 | 200.00 | B_B32 |
| E33.a | culvert existitem pipe remove | Road | 10 | m | 13.00 | 130.00 | B_B33 |
| E33.c | existfence item remove | Road | 15 | m | 25.00 | 375.00 | B_B34 |
| C5.a.5 | 100 backfill bedding cls cover dia dr grnlr lateral mm pipe pvc | SanitarySewer | 30 | m | 94.00 | 2,820.00 | B_B35 |
| C20.a.1 | 100 dia exist lateral mm reconnect sanitary srv sewer | SanitarySewer | 6 | ea. | 332.00 | 1,992.00 | B_B36 |
| C7.c | a flush inspect new pipe sewer tv | SanitarySewer | 56 | m | 16.00 | 896.00 | B_B37 |
| C7.e | b existflush inspect lateral srv sewer tv | SanitarySewer | 31 | m | 16.00 | 496.00 | B_B38 |
| C4.a | 1200 circular complete concrete dia hole mm precast standard | SanitarySewer | 1 | L.S. | 3,700.00 | 3,700.00 | B_B39 |
| C9 | # deep m mh opsd per sa | SanitarySewer | 6 | ea. | 335.00 | 2,010.00 | B_B40 |
| C6.d | clean out | SanitarySewer | 2 | ea. | 341.00 | 682.00 | B_B41 |
| C6.a.1 | a case combine exist hole maintain remove sanitary sewer | SanitarySewer | 60 | m | 3.00 | 180.00 | B_B42 |
| D16.a.15 | 150 b case combine dia exist mm pipe remove sanitary sewer | StormSewer | 5 | m | 152.00 | 760.00 | B_B43 |
| D16.a.16 | 250 a backfill cb bedding cover dia grnlr lead mm pvc sdr | StormSewer | 40 | m | 164.00 | 6,560.00 | B_B44 |
| D6.b.1 | cb concrete opsd precast single | StormSewer | 2 | ea. | 2,237.00 | 4,474.00 | B_B46 |
| B1.a.2 | cb concrete double opsd precast | Watermain | 144 | m | 110.00 | 15,840.00 | B_B47 |
| B1.a.3 | cls cut dr large mm open pipe pvc srv water watermain | Watermain | 390 | m | 106.00 | 41,340.00 | B_B48 |
| B1.a.6 | cls cut dr large mm open pipe pvc srv water watermain | Watermain | 120 | m | 58.00 | 6,960.00 | B_B49 |
| B2.b | cls cut dr large mm open pipe pvc srv water watermain | Watermain | 3 | ea. | 1,200.00 | 3,600.00 | B_B50 |
| B2.c | box dia mm tap valve water | Watermain | 5 | ea. | 1,700.00 | 8,500.00 | B_B51 |
| B3.a | box dia mm tap valve water | Watermain | 4 | ea. | 4,600.00 | 18,400.00 | B_B52 |
| B4.a.1 | box hydrant opsd settap valve water | Watermain | 135 | m | 57.00 | 7,695.00 | B_B53 |
| B17.c | copper cut k mm new open opsd replace srv soft type water | Watermain | 25 | m | 57.00 | 1,425.00 | B_B56 |
| B7.b | hdpe install mm new replace srv series tube water | Watermain | 16 | ea. | 249.00 | 3,984.00 | B_B57 |
| B6.a.2 | main mm stop | Watermain | 16 | ea. | 213.00 | 3,408.00 | B_B58 |
| B5.a.1 | curb mm stop | Watermain | 16 | ea. | 312.00 | 4,992.00 | B_B59 |
| B12 | copper exist k m mm new opsd pipe reconnect srv soft type use | Watermain | 3 | L.S. | 1,263.00 | 3,789.00 | B_B60 |
| B8 | disinfecttest watermain | Watermain | 2 | ea. | 415.00 | 830.00 | B_B61 |
| B29 | box curb line property remove replace require stop | Watermain | 5 | Hrs | 400.00 | 2,000.00 | B_B62 |
| B22 | prov delay down shut water | Watermain | 4 | ea. | 415.00 | 1,660.00 | B_B63 |

Table 2.6: Sample running example of a contract (Contract A) City A.

| Item # | Description | Part | Qty | Unit | Unit Price | Total Price | UID # |
|---|---|---|---|---|---|---|---|
| 1.1 | concrete curb dispose exist gutter remove | Road | 1230 | m | 3.28 | 4,034.40 | A_A2 |
| 1.2 | asphalt dispose drwy exist remove walkway | Road | 580 | m2 | 2.22 | 1,287.60 | A_A3 |
| 1.3 | drwy exist interlock remove sidewalk stone store | Road | 110 | m2 | 25.00 | 2,750.00 | A_A4 |
| 1.4 | brick curb drwy exist remove stone store wood | Road | 40 | m | 10.00 | 400.00 | A_A5 |
| 1.6 | 200 dia dispose exist mm remove sanitary sewer | Road | 375 | m | 17.10 | 6,412.50 | A_A6 |
| 1.8 | 100 dia dispose exist mm remove watermain | Road | 150 | m | 12.45 | 1,867.50 | A_A7 |
| 1.9 | 300 375 cb dispose exist fire hydrant mm remove sewer storm | Road | 3 | ea. | 311.35 | 934.05 | A_A8 |
| 1.10 | dispose exist mh remove storm | Road | 4 | ea. | 156.00 | 624.00 | A_A9 |
| 2.1 | 200 bedding c dia install mh mm pvc sdr supply type w | SanitarySewer | 73 | m | 113.75 | 8,303.75 | A_A10 |
| 2.2 | 1200 concrete dia install mh mm precast sanitary supply | SanitarySewer | 1 | L.S. | 3,375.00 | 3,375.00 | A_A11 |
| 2.3 | 200 connect drain install mm private sanitary sewer supply | SanitarySewer | 90 | ea. | 749.00 | 67,410.00 | A_A12 |
| 2.4 | Cathcart connect exist mh rebench sanitary street | SanitarySewer | 1 | L.S. | 1,810.00 | 1,810.00 | A_A13 |
| 2.5 | connect exist mh rebench sanitary street | SanitarySewer | 1 | L.S. | 1,810.00 | 1,810.00 | A_A14 |
| 2.6 | backfill c granular import provisional | SanitarySewer | 2000 | tonnes | 9.06 | 18,120.00 | A_A15 |
| 2.7 | inspect video | SanitarySewer | 533.4 | m | 3.85 | 2,053.59 | A_A16 |
| 3.1 | 375 450 525 675 all mm pipe | StormSewer | 61 | m | 557.00 | 33,977.50 | A_A17 |
| 3.2 | 675 b bedding c cl100d cls cone dla mh mm pipe supply w | StormSewer | 76 | m | 424.00 | 32,224.00 | A_A18 |
| 3.3 | 825 b bedding c cl65d cls cone dia install mm pipe supply w | StormSewer | 57.2 | m | 438.00 | 25,053.60 | A_A19 |
| 3.4 | 1200 concrete dia install mh mm precast storm supply | StormSewer | 1 | L.S. | 2,565.00 | 2,565.00 | A_A20 |
| 3.5 | 1500 concrete dia install mh mm precast storm supply | StormSewer | 1 | L.S. | 6,610.00 | 6,610.00 | A_A21 |
| 3.6 | 525 connect dia mm sewer storm | StormSewer | 15 | ea. | 746.00 | 11,190.00 | A_A22 |
| 3.7 | 150 behind curb dia m mm pipe pvc sdr sewer storm | StormSewer | 250 | m | 97.10 | 24,275.00 | A_A23 |
| 3.7 | clean drwy out | StormSewer | 10 | ea. | 142.75 | 1,427.50 | A_A24 |
| 4.1 | 150 dia install mm supply watermain | Watermain | 600 | m | 80.70 | 48,420.00 | A_A25 |
| 4.2 | install off sleeve supply tap valve | Watermain | 1 | ea. | 3,800.00 | 3,800.00 | A_A26 |
| 4.3 | c connect hydrant install lead storz supply tee valve w water way | Watermain | 2 | ea. | 5,400.00 | 10,800.00 | A_A27 |
| 4.4 | 300 extension hydrant mm provisional | Watermain | 1 | ea. | 565.00 | 565.00 | A_A28 |
| 4.5 | 150 connect exist mm new property remove replace srv water | Watermain | 700 | m | 87.20 | 61,040.00 | A_A29 |
| 4.5 | 19 cut dia line mm open srv water | Watermain | 70 | ea. | 273.50 | 19,145.00 | A_A30 |
| 4.6 | 150 connect exist mm new property remove replace srv water | Watermain | 20 | ea. | 265.00 | 5,300.00 | A_A31 |
| 4.7 | 19 cock line main mm | Watermain | 1 | ea. | 1,075.00 | 1,075.00 | A_A32 |
| 4.8 | connect excavate exist line property srv use vacuum water | Watermain | 1 | L.S. | 2,500.00 | 2,500.00 | A_A33 |
| 4.9 | 150 cap cut exist mm watermain | Watermain | 2 | ea. | 3,320.00 | 6,640.00 | A_A34 |
| 4.9 | disinfect flush street swab test | Watermain | 75 | ea. | 61.60 | 4,620.00 | A_A35 |
| 4.9 | backflow c connect exist fire hydrant overland prevent system temporary w water | Watermain | 1 | ea. | 1,190.00 | 1,190.00 | A_A36 |
| 4.10 | 25 connect mm overland srv system temporary water | Watermain | 1000 | tonnes | 9.08 | 9,080.00 | A_A37 |
| 5.1 | 100 connect mm overland system temporary water | Road | 4300 | m3 | 9.11 | 39,173.00 | A_A38 |
| 5.2 | backfill c granular import provisional | Road | 450 | m3 | 9.55 | 4,297.50 | A_A39 |
| 5.3 | dispose excavate normal | Road | 560.91 | tonnes | 24.79 | 13,905.00 | A_A40 |
| 5.4 | sub excavate | Road | 120 | tonnes | 153.83 | 18,459.60 | A_A41 |
| 5.5 | a b compact granular material place subbase supply | Road | 605 | m | 36.40 | 22,022.00 | A_A42 |
| 5.6 | asphalt compact drwy fine hl place supply | Road | 110 | m2 | 25.70 | 2,827.00 | A_A43 |
| 5.7 | concrete curb gutter install opsd supply | Road | 40 | m | 17.80 | 712.00 | A_A44 |
| 5.8 | drwy exist interlock reinstall sidewalk stone | Road | 3400 | m2 | 4.32 | 14,688.00 | A_A45 |
| 5.9 | brick curb drwy exist reinstall stone wood | Road | 200 | ea. | 28.50 | 5,700.00 | A_A46 |
| 5.10 | boulevard import place supply topsoil | Road | 1500 | m | 2.38 | 3,570.00 | A_A47 |
| 6.1 | 40kg calcium chloride ctrl dust flake | Prov | 1 | L.S. | 5,900.00 | 5,900.00 | A_A48 |
| 6.1 | fence protect tree | Prov | 1 | L.S. | 5,900.00 | 5,900.00 | A_A49 |
| 6.2 | 50% labour material | Prov | 1 | L.S. | 2,450.00 | 2,450.00 | A_A50 |
| 6.3 | 50% performance | Prov | 1 | L.S. | 35,750.00 | 35,750.00 | A_A51 |

Table 2.7: Sample running example of a contract (Contract B) City B.

| Item # | Description | Part | Qnty | Unit | Unit Price | Total Price | U_ID # |
|---|---|---|---|---|---|---|---|
| 1 | remove tree | General | 18 | each | 1,250.00 | 22500.00 | C_C1 |
| 2 | bond for payment | General | 1 | L.S. | 17500.00 | 17500.00 | C_C2 |
| 3 | all for insurance payment | General | 1 | L.S. | 15,000.00 | 15,000.00 | C_C3 |
| 4 | demobilize mobilize | General | 1 | L.S. | 35,000.00 | 35,000.00 | C_C4 |
| 5 | field office | General | 1 | each | 10,000.00 | 10,000.00 | C_C5 |
| 6 | control traffic | General | 1 | L.S. | 40,000.00 | 40,000.00 | C_C6 |
| 7 | capital constructimprovement project sign | General | 4 | each | 400.00 | 1600.00 | C_C7 |
| 8 | banners construct | General | 4 | each | 500.00 | 2,000.00 | C_C8 |
| 9 | all asphalt batches design hot method mix prepare superpave trial type | General | 2 | each | 600.00 | 1,200.00 | C_C9 |
| 10 | construct photos pre videos | General | 1 | L.S. | 15,000.00 | 15,000.00 | C_C10 |
| 11 | condition construct post pre surveys | General | 1 | L.S. | 22,500.00 | 22,500.00 | C_C11 |
| 12 | as backfill crush direct ggregate mm pittest with | General | 15 | each | 500.00 | 7,500.00 | C_C12 |
| 13 | as backfill directfill pittest unshrinkable with | General | 15 | each | 550.00 | 8250.00 | C_C13 |
| 14 | large remove tree | General | 7 | each | 1,500.00 | 10,500.00 | C_C14 |
| 15 | remove small tree | General | 12 | each | 1,000.00 | 12,000.00 | C_C15 |
| 16 | as built construct draw of provision survey | General | 1 | L.S. | 25,000.00 | 25,000.00 | C_C16 |
| 1 | basin catch clean exist out sump | SanitarySewer | 30 | each | 85.00 | 2,550.00 | C_C17 |
| 2 | before clean constructflush hole maintain sanitary sewer storm video | SanitarySewer | 1950 | m | 10.00 | 19,500.00 | C_C18 |
| 3 | debris excessive flush for high in pressure sewer | SanitarySewer | 195 | m | 25.00 | 4,875.00 | C_C19 |
| 4 | basin catch class control fabric for geotextile non silttemporary woven | SanitarySewer | 30 | each | 35.00 | 1,050.00 | C_C20 |
| 5 | basin cast catch circular frame grate iron raise remove replace to | SanitarySewer | 21 | each | 800.00 | 16,800.00 | C_C21 |
| 6 | a b cast cover frame hole iron maintain remove replace square type | SanitarySewer | 15 | each | 550.00 | 8,250.00 | C_C22 |
| 7 | basin catch depth full remove single | SanitarySewer | 2 | each | 1,050.00 | 2,100.00 | C_C23 |
| 8 | basin catch connect ditch inletinstall lead sewer supply to | SanitarySewer | 2 | each | 5,100.00 | 10,200.00 | C_C24 |
| 9 | after clean constructflush hole maintain sanitary sewer storm video | SanitarySewer | 1950 | m | 10.00 | 19,500.00 | C_C25 |
| 10 | basin catch clean exist out sump | SanitarySewer | 30 | each | 85.00 | 2,550.00 | C_C26 |
| 11 | before clean constructflush hole maintain sanitary sewer storm video | SanitarySewer | 1950 | m | 10.00 | 19,500.00 | C_C27 |
| 12 | debris excessive flush for high in pressure sewer | SanitarySewer | 195 | m | 25.00 | 4,875.00 | C_C28 |
| 13 | basin catch class control fabric geotextile non silttemporary woven | SanitarySewer | 30 | each | 35 | 1,050.00 | C_C29 |
| 14 | basin cast catch circular frame grate iron raise remove replace to | SanitarySewer | 21 | each | 800.00 | 16800.00 | C_C30 |
| 15 | a b cast cover frame hole iron maintain remove replace square type | SanitarySewer | 15 | each | 550.00 | 8,250.00 | C_C31 |
| 16 | basin catch depth full remove single | SanitarySewer | 2 | each | 1,050.00 | 2,100.00 | C_C32 |
| 17 | basin catch connect ditch inletinstall lead sewer supply to | SanitarySewer | 2 | each | 5,100.00 | 10,200.00 | C_C33 |
| 18 | after clean constructflush hole maintain sanitary sewer storm video | SanitarySewer | 1950 | m | 10.00 | 19,500.00 | C_C34 |
| 1 | cl dr18 mm pvc roadway watermain within | Watermain | 10 | m | 900.00 | 9,000.00 | C_C35 |
| 2 | cl dr18 mm pvc roadway watermain within | Watermain | 420 | m | 950.00 | 399,000.00 | C_C36 |
| 3 | avoid conflict dia draw fire hydrant lead looping mm not of on or propose service shown to utility water watermain with | Watermain | 2 | each | 1,500.00 | 3,000.00 | C_C37 |
| 4 | avoid conflict dia draw looping mm not pipe propose service shown utility water watermain with | Watermain | 3 | each | 1,750.00 | 5,250.00 | C_C38 |
| 5 | box gate mm valve | Watermain | 2 | each | 2,500.00 | 5,000.00 | C_C39 |
| 6 | box gate mm valve | Watermain | 5 | each | 3,000.00 | 15,000.00 | C_C40 |
| 7 | complete hydrant new | Watermain | 3 | each | 12500.00 | 37500.00 | C_C41 |
| 8 | all complete connect exist new size to watermain | Watermain | 3 | each | 10,000.00 | 30,000.00 | C_C42 |
| 9 | all cap cut ends exist size the watermain | Watermain | 9 | each | 850.00 | 7,650.00 | C_C43 |
| 10 | all connect cross existfill piece remove replace size tee type watermain | Watermain | 2 | each | 7,500.00 | 15,000.00 | C_C44 |
| 11 | box cap end fire hydrantinclude remove valve | Watermain | 3 | each | 850.00 | 2,550.00 | C_C45 |
| 1 | condition investigate of pit service testto water | Watermain | 40 | each | 500.00 | 20,000.00 | C_C46 |
| 2 | all connect curb include non operational remove replace size stop streetline | Watermain | 12 | each | 330.00 | 3,960.00 | C_C47 |
| 3 | any complete copper cut dia exist extend include mm necessary new pipe reconnect service watermain | Watermain | 5 | each | 2,700.00 | 13,500.00 | C_C48 |
| 4 | any complete copper cut dia exist extend include mm necessary new pipe reconnect service to watermain | Watermain | 5 | each | 2,900.00 | 14,500.00 | C_C49 |
| 5 | complete connect copper dia length line m mm property service water | Watermain | 15 | each | 3,200.00 | 48,000.00 | C_C50 |
| 6 | complete connect copper dia length line m mm property service water | Watermain | 15 | each | 3,500.00 | 52,500.00 | C_C51 |
| 7 | complete connect copper dia length line m mm property service water | Watermain | 9 | each | 3,700.00 | 33,300.00 | C_C52 |
| 8 | complete connect copper dia length line m mm property service water | Watermain | 9 | each | 3,900.00 | 35,100.00 | C_C53 |

Table 2.8: Sample running example of a contract (Contract C) City C.

### Data Provenance

Data provenance refers to the documentation and tracking of the origin, lineage, and history of data [Moreau et al., 2013]. It enables identifying and correcting errors, ensuring the data is trustworthy and reliable. In scientific research, data provenance is essential for reproducibility, accountability, and transparency [Garijo et al., 2014, Missier et al., 2013]. Moreover, data provenance is critical for decision support systems, where using incorrect or incomplete data can lead to erroneous analyses and unpredictable outcomes [Fisher and Kingma, 2001], [Khaki, 2021], [Pipino et al., 2002], and [Sadiq et al., 2011].

In this project, we developed a decision support tool that analyzes tender/contract documents to evaluate contractors' bids and behaviours. The outcome of this tool is the conversion of tender-bid as presented in Table 2.1 on Page 28, Table 2.2 on Page 30, Table 2.3 on Page 32 being used as input and converted to the output tables presented in Table 2.6 on Page 57, Table 2.7 on Page 58, and Table 2.8 on Page 59. Common errors in the current application were previously discussed in this chapter. We implemented a systematic approach for detecting and analyzing records, including potential errors, ensuring the tool's output accuracy and reliability. This approach includes investigating sample cases to identify error sources, finding systematic methods to address error types, and analyzing the resulting sensitivity to errors [Reeder and David, 2016].

Our system used an extended set of ontology rules and provenance records to address data errors and ensure data provenance. Ontology rules formally represent the domain knowledge, allowing for automated reasoning and inference [Stuckenschmidt, 2009]. Provenance records, on the other hand, document the origin and lineage of data, facilitating error identification and correction [Moreau et al., 2013].

When converted to electronic format, hard copies of archived documents require sanity checks before error correction to guarantee accurate data provenance. Using OCR technology to digitize printed documents can introduce errors, leading to data quality deterioration [Kim et al., 2003]. Therefore, ensuring that the digitized data accurately reflects the original document's content is essential.

In summary, data provenance is critical for ensuring the accuracy and reliability of decision support systems. It involves documenting and tracking the origin, lineage, and history of data, enabling error identification and correction. A systematic approach combining ontology rules and provenance records is necessary to address errors and ensure trustworthy data.

### 2.2.4 Image Pre-processing for Hard Copies

Analyzing scanned tenders and employing Optical Character Recognition (OCR) poses several challenges. The quality of scans is inconsistent, with pages often rotated or skewed due to varying scanning methodologies. Furthermore, table formatting varies; some tables showcase visible column borders, others only display header borders, while actual table cells are visually separated solely by whitespace. Such variation might require operator feedback. Consequently, automated data extraction using OCR tools, such as ABBYY, without pre-processing is often insufficient.

Due to significant variations in scanning quality and table layouts, a universal approach was ineffective. In response, we developed a suite of image analysis tools to refine table layouts in OCR-scanned PDFs. Once the document scan quality is sufficiently enhanced, OCR software like ABBYY FineReader can extract tabular data. This section overviews these tools and the steps required for tabular data extraction from a sample page.

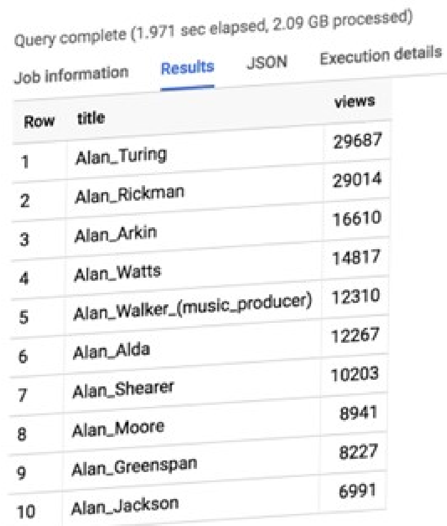
For illustrative purposes, we use an example of a single table from a single page detailing the process of enhancing scan quality. A simple measure of scan quality assessment is applying OCR software to the raw data, followed by a result examination. Further steps are unnecessary if the tabulation scheme aligns (i.e., the arrangement of data in cells) and maintains a minimum text conversion accuracy. However, frequently, pages are skewed and need realignment, along with improvements in image brightness and contrast. Any present marks (checkmarks, handwritten notes) should be removed or noted for final result adjustments.

The image de-skewing routine utilizes either the table's four external corners or one vertical and one horizontal line (both user-provided) to ascertain the level of skewness or rotation requiring correction. Another method for detecting lines in table rows and columns involves the Hough Transform [Aggarwal and Karl, 2006].

The table's external corners are identified using a manual or semi-automated process (involving the Hough Transform). This information helps create the transformation matrix needed to de-skew the image, applying scale or rotation adjustments as necessary. Even post-Hough Transform application and line detection, user feedback is critical to ensure accurate parameter and line detection.

As illustrated in Figure 2.11, there are non-aligned horizontal lines, even though the vertical lines are aligned, barring the overall image rotation. The line slope signifies page skewness, with left text boxes slightly shifted downward compared to those on the same row's right. Such a skewed image would yield low-quality results from OCR software, necessitating correction.

In conclusion, extracting data from tables in scanned documents is a complex and multi-faceted challenge. Issues such as varying table layouts, missing visual markers (e.g., table cell borders), and page rotation or skewness contribute to this complexity. No single solution can address all these cases, as evidenced by this project's experiences. Combining manual and automated methods, including OCR technology and image analysis tools, is vital to enhance scan quality and facilitate accurate data extraction.



Query complete (1.971 sec elapsed, 2.09 GB processed)

Job information | Results | JSON | Execution details

| Row | title | views |
|---|---|---|
| 1 | Alan_Turing | 29687 |
| 2 | Alan_Rickman | 29014 |
| 3 | Alan_Arkin | 16610 |
| 4 | Alan_Watts | 14817 |
| 5 | Alan_Walker_(music_producer) | 12310 |
| 6 | Alan_Alda | 12267 |
| 7 | Alan_Shearer | 10203 |
| 8 | Alan_Moore | 8941 |
| 9 | Alan_Greenspan | 8227 |
| 10 | Alan_Jackson | 6991 |

Figure 2.11: An example table requiring de-skewing (right side 3.5 Degrees higher than left) and counterclockwise rotation (0.75 Degrees) for accurate text recognition in cells.

### 2.2.5 Data Quality and Integrity Management

In this project, we have adopted a detailed approach to maintain data quality and integrity to identify errors, manage noise, and address omissions based on a deep understanding of the data's nuances. The identification of errors often hinges on predefined data requirements and typologies. For instance, when considering a dataset on watermains, the items involving "pipe material" contains standard entries such as "cast iron or CI," "PVC," or "ductile iron or DI." Any divergence from these accepted materials prompts an error flag. Similarly, in the case of sanitary sewer datasets, numeric entries indicating the pipe diameter in millimetres are expected. In their absence, potential errors are flagged.

Another common challenge is the noise introduced while translating physical records to digital data. This issue is frequently encountered when tender documents, initially in hard copy, are scanned and converted into electronic tables. During this process, elements such as handwritten annotations or checkmarks, originally designed to provide clarity, often introduce noise and disrupt the Optical Character Recognition (OCR) process.

Data omissions are notably challenging to identify and can significantly impact the dataset's integrity. For instance, watermain items contain information regarding each pipe diameter. An error flag is raised in cases where this information is absent, signalling a critical omission that can impact subsequent calculations and assessments. Similarly, missing data on the depth of the sanitary sewer maintenance holes, a factor vital for determining the cost while resizing the depth and diameter to the specifics of the unit cost, constitutes a significant omission.

The process of rectifying these issues requires a multi-faceted approach. Missing data, such as the diameter of a watermain pipe, are addressed with operator intervention and consultation with original engineering drawings or other copies of the bid document. On the other hand, issues such as a missing pipe diameter in a sanitary sewer dataset can be resolved by treating the item as a lateral sewer pipe. Other correction mechanisms rely on utilizing information from different fields or items. For instance, if the unit cost of a watermain pipe segment is missing, an estimation can be made by dividing the total cost by the quantity, assuming these fields are available. In situations where this approach is not viable, alternative strategies might be employed, such as using cost data from similar pipe segments or consulting industry-standard cost databases.

The origin of the data significantly influences its propensity for errors. Electronic table datasets derived directly from bid submission websites, services, or other digital platforms are typically less susceptible to errors. The absence of OCR processing, which often introduces transcription errors and other discrepancies, contributes to this reduced error propensity. Furthermore, these digital datasets undergo software validation checks, ensuring data completeness and integrity.

Despite these advantages, electronic table datasets are not entirely exempt from errors. Issues often arise due to human error during data entry, such as typographical errors, inconsistent terminology, or incorrect unit assignments. For instance, inconsistencies in a watermain dataset can occur when engineers interchangeably use the terms "PVC" and "polyvinyl chloride," or "DI" and "Ductile Iron." This inconsistency necessitates cleaning procedures to standardize the terminology. Likewise, a standard error in sanitary sewer datasets could be the inconsistent assignment of units for pipe length in feet or meters. It leads to issues during unit cost analysis and necessitates correction during data cleaning.

Even digital datasets are not immune to data omissions. A missing pipe length or diameter could go unnoticed if the system is not configured to enforce compulsory data entry for these fields. Moreover, logical errors could arise, such as inconsistencies between a watermain pipe diameter and the valves used to connect the pipe to the existing infrastructure. In summary, although automated error detection and correction mechanisms provide substantial support, human intervention is indispensable in certain situations. Depending on the nature and severity of the encountered issues, the level of intervention can vary. Regardless of the datasets' origins and complexities, maintaining vigilant quality control and adhering to robust data validation protocols remain essential to ensure the data's accuracy and reliability.

The function of ontology, especially its predefined rules or standards, is pivotal in structuring the data. Ontologies define relationships, establish hierarchies, and set constraints on valid data, eliminating redundancy and ensuring data coherence. For example, ontology rules standardize the names of pipe materials, flagging entries that deviate from the established terminology. Additionally, ontologies are crucial in managing logical errors and missing data. Rules can enforce the consistency of pipe and valve diameters in sequential items or require an associated depth for each maintenance hole entry, ensuring data completeness. By implementing ontology rules, the data cleaning process can be partially automated, minimizing the need for manual review and intervention and optimizing the time and effort required.

In conclusion, managing data quality and integrity in this project involves a complex interplay of error detection, noise management, and data omission handling. Utilizing both manual interventions and automated methods, along with robust ontological rules, ensures the data's consistency, accuracy, and reliability. The strategies outlined here form a comprehensive framework for maintaining the high standards required in the data's lifecycle.

### Provenance Flags

This section introduces flags, part of the ontology, to facilitate data import and maintenance and streamline the data analysis process. These flags facilitate efficient communication between the different components and stages of the data analysis process, promoting a more streamlined flow of information. Table 2.9 lists the flags and their descriptions.

**Record Cleaning Status**; Every record indicates whether a cleaning procedure has been applied. Reasons for labelling a record as "clean" include error correction, updating existing entries with new values, or amendments to contract payments. Cleaned records will carry the "Violet flag."

**Nature of the Data Quality Issue**; if a record is identified as faulty, the type of error, whether corrected or still present, is specified. Errors include missing data, outliers, incorrect formats (e.g., numeric instead of a string or litres instead of gallons), typos, spikes or abnormalities in data trends, noisy records or measurements, duplicate records, field data overload, and incorrect timestamps. Records with OCR-related errors should have the "Brown flag."

**Employed Cleaning Approach**; the cleaning approach used is explicitly mentioned in the provenance records. These approaches include interpolation and extrapolation for missing records, unit modification for inconsistencies, ignoring the error, filtering and removal of outliers, re-acquisition from redundancies, storage format change, manual correction and override, duplicate elimination, filling by a constant value from rules in Ontology, using the most probable value, replacing by central tendency value, and replacing by a value acquired through binning or clustering neighbouring records.

**Cleaning Revision**; records might require multiple cleaning iterations or further error identification post-cleaning. Hence, a field indicating the date of the cleaning process, the reason for (re)cleaning, and the revision number (in cases of multiple revisions) are included.

**Detection Method**; the detection method employed by the operator or cleaning software. Examples include operator-based, expert monitoring and flagging, watchdog programs, data mining transformation-caused outlier, filtered outlier, and dictionary-based or lookup table-based detection.

| Color | Flag Description | Permanent or removable | Item analyzable? |
|---|---|---|---|
| Red | error, requires manual handling | removable | NO |
| Violet | error removed, has provenance record | removable | YES |
| Yellow | error, minor, show warning to operator | removable | Maybe |
| Blue | Item has pre-determined Part & Sub-Part | permanent | YES |
| Pink | no Part/Sub-Part, requires classification | removable | NO |
| Green | Item has classified Part and Sub-Part | permanent | YES |
| Brown | Original item is hard-copy | permanent | YES |
| Black | Raw item with no change | permanent | NO |
| White | Item passed WaterIAM integrity check | permanent | YES |
| Grey | Item descriptions passed Ontology Check | permanent | Maybe |

Table 2.9: description of flags used in the record standardization process.

**Error Source;** determining the origin of the error is crucial for future ontology revisions. Possible error sources include meter-based errors, operator-based errors, integration-schema mismatch errors, records import algorithm issues, OCR algorithm errors, or unknown sources.

**Cleaning Tool Used;** if a cleaning tool was utilized, it is documented in the provenance records. Available data cleaning tools include locally programmed code, data wrangler, Drake, open refine, Winpure, Patnab, cleaning scope, Alteryx, and local lookup table, among other available tools.

**Additional Provenance Record Fields;** other fields that are included in the provenance records might specify whether the update's scope was local or global and whether the update is incorporated as an ontology rule for future imports. These flexible guidelines can be adapted to accommodate future expansions or address unforeseen issues.

The data quality and integrity management process dramatically benefit from the structuring power of ontology, particularly when it comes to defining and enforcing rules or standards. By defining relationships, establishing hierarchies, and setting parameters for valid data, Ontology facilitates the systematic organization and standardization of data. For example, an ontological rule might enforce uniformity in the terminology of pipe materials, enabling the system to identify and flag any entries that stray from this norm. Similarly, ontologies can help manage logical inconsistencies and address missing data, as rules can be designed to enforce certain conditions - such as matching diameters for pipes and valves in a sequence or ensuring each maintenance hole entry includes a depth attribute. With these rules in play, aspects of the data cleaning process can be automated, saving significant amounts of time and manual effort.

### 2.2.6 Relational Database Schema

The system's data model, realized through a relational database, is a critical component of the overall data pipeline. It serves as the foundation for the processing, standardization, and analyzing the tender data. It facilitates data storage and retrieval and fosters consistency, data integrity, and extensibility, which are fundamental characteristics of a robust data system. Furthermore, it reinforces adherence to the ontology's semantic rules, guaranteeing data conformity to established formats, relationships, and constraints.

In the '`contract_bid_item_tab`', the selection of varchar data type for the majority of the fields lends versatility to the data it can store. The variable character limits, based on the expected input size, contribute to space efficiency while maintaining a degree of flexibility. More importantly, the choice of using unique identifiers ('`contract_id,`' '`item_uid,`' '`unt_uid,`' '`ref_itm_uid,`' and '`doc_uid`') for associating entities among different tables enables the normalization of data, thereby reducing redundancy and inconsistency.

The '`item_reference_table`' plays a vital role in data standardization by providing centralized storage for standardized parts, subparts, descriptions, and possible alternatives of items. This standardization ensures that every item in the system can be uniquely identified and referenced, allowing a uniform interpretation and comparison of items across various contracts. The option of storing possible units of measurement for each item facilitates the handling of diverse units that might appear in the contracts, strengthening the system's adaptability.

In the '`bid_docs`' table, the decision to include fields for city and contractor information reflects the multi-dimensionality of the data, acknowledging that a contract is not just a list of items but also involves contextual information. By capturing this, the system provides a more comprehensive view of the tender process, thus facilitating more nuanced and context-specific analysis.

Lastly, the '`units`' table plays a pivotal role in harmonizing the units of measurement across different contracts. Including standard conversion methods and ratios enables seamless conversion of various units into their standard forms, ensuring comparability of items irrespective of their original units. The boolean field '`unt_standard`' provides an efficient way to quickly determine whether a unit is standard, thereby streamlining the standardization process.

The schemas' relationships between tables illustrate the system's ability to capture complex interdependencies among data elements. Using foreign keys, the data model supports joins between tables, facilitating comprehensive queries and detailed data analysis.

Overall, choosing a relational database for this data system provides a structured framework for efficient data management and ensures adherence to the ontology's semantic rules. This combination of efficiency, flexibility, and consistency makes it ideal for handling tender data. By catering to the varied needs of standardization, storage, retrieval, and analysis, the database serves as the backbone of the data pipeline, ultimately driving the goal of delivering accurate and insightful results.

### Schema Details

The primary goal of the data pre-processing chapter is to prepare the data for analysis, involving determining the type of items available in tender documents (data) for classification in the subsequent chapter. A crucial decision in this process is selecting a suitable storage format for the data.

In this case, the simplified standard data consists of a list of contract items for each city, containing a description, quantity, unit, and unit price fields. An SQL database has been implemented for data storage, as shown in Figure 2.12. The chosen relational database ensures accuracy and flexibility and is designed for future modifications with efficient storage consumption.

The implemented ontology, which includes rules, filters, definitions, and tables, ensures the data's consistency and standardization and guarantees consistency, standardization, understandability, and error-free content before entering the core database. However, the primary storage format remains a relational table in the SQL database management system.

The first table in the schema is the `‘contract_bid_item_tab’`, which serves as a catalogue of all the items involved in various contracts. Each entry in this table represents a distinct item, with fields capturing a wide range of data points. For instance, the ‘contract_id’ field assigns a unique identifier for each contract, and the `‘item_uid’` provides a unique identifier for the item. The `‘item_number’`, `‘item_quantity’`, and `‘item_unit_cost’` fields hold the item number, quantity, and cost per unit, respectively. Other fields like `‘item_parent_desc’` and `‘item_org_section’` offer narrative context. The table also interlinks with other tables through `‘unt_uid’`, `‘ref_itm_uid’`, and `‘doc_uid’`, which connect to the units table, item reference table, and bid documents table, respectively. The `‘ref_std_part’` and `‘ref_std_sub_part’` fields contain references to the standardized part and subpart of the item. Notably, the `‘contract_id,’ ‘item_uid,’ item_number’, ‘item_parent_desc,’ ‘unt_uid,’ ‘ref_itm_uid,’ ‘doc_uid,’ ‘ref_std_part,’ ‘ref_std_sub_part’`, and `‘item_org_section’` fields

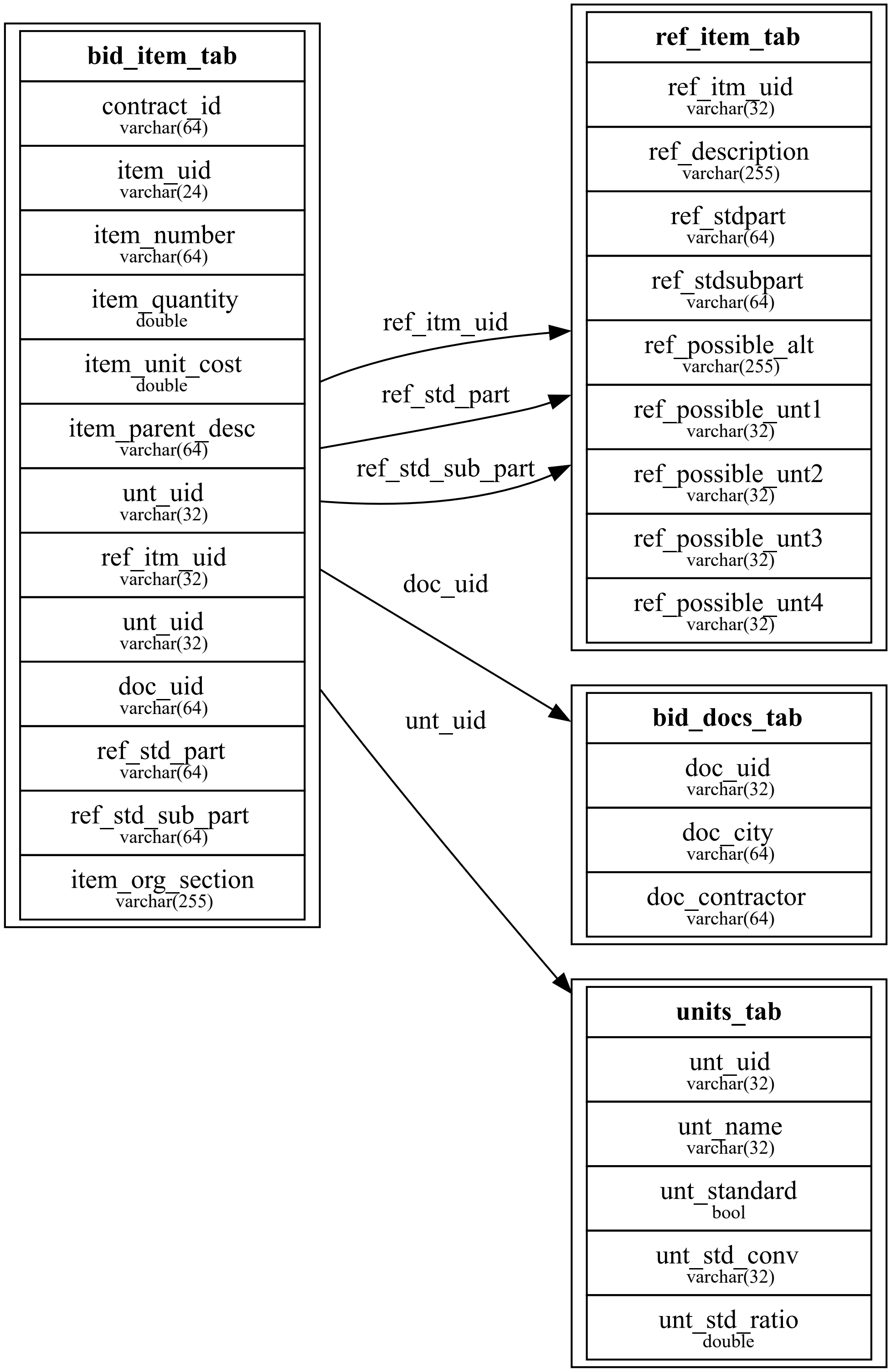



Figure 2.12: The enhanced entity-relationship diagram of the core database.

are all varchar types, with a maximum length ranging from 24 to 255 characters. On the other hand, the `‘item_quantity’` and `‘item_unit_cost’` fields are of double types, allowing for a high degree of precision in representing quantities and costs.

Next, the `‘item_reference_table’` serves as a repository for standardization information related to the items. Each row represents a distinct reference item, identified by ‘ref_itm_uid.’ The `‘ref_description’` field provides a fuller description of the reference item, and the `‘ref_std_part’` and `‘ref_std_sub_part’` fields delineate the standardized part and subpart of the item. The `‘ref_possible_alt’` field captures possible alternative references for the item, and the `‘ref_possible_unit1’, ‘ref_possible_unit2’, ‘ref_possible_unit3’`, and `‘ref_possible_unit4’` fields indicate potential units of measurement for the item. This table includes varchar fields such as `‘ref_itm_uid,’ ‘ref_description,’ ‘ref_std_part,’ ‘ref_std_sub_part’`, and `‘ref_possible_alt’`, each with different character limits from 32 to 255, offering flexibility for capturing a broad array of standardized parts, subparts, descriptions, and possible alternatives. The `‘ref_possible_unit1’, ‘ref_possible_unit2’, ‘ref_possible_unit3’`, and `‘ref_possible_unit4’` fields are also varchar types, each with a maximum of 32 characters.

The `‘bid_docs’` table provides an overview of each bid document, including the associated city and contractor. Each row corresponds to a separate document, designated by `‘doc_uid’`. The `‘doc_city’` field records the associated city, and the ‘doc_contractor’ field tracks the involved contractor. Notably, all the fields in this table (`‘doc_uid,’ ‘doc_city’`, and `‘doc_contractor‘`) are varchar types, with a maximum length of 64 characters, providing ample room for unique document identifiers, city names, and contractor names.

Finally, the `‘units’` table provides a directory of potential and acceptable units of measurement according to the ontology. Each unit is uniquely identified by `‘unt_uid’`, with `‘unt_name’` holding the unit name and `‘unt_standard’` indicating whether the unit is standard. The `‘unt_std_conv’` and `‘unt_std_ratio’` fields detail the standard conversion method and ratio for each unit. Here, `‘unt_uid’, ‘unt_name’`, and `‘unt_std_conv’` are varchar fields, each with a limit of 32 characters, appropriate for unique unit identifiers, unit names, and standard conversions. The `‘unt_standard’` field is a bool type, capable of holding a boolean value to show whether the unit is a standard one, while the `‘unt_std_ratio’` field is a double type for precise representation of standard conversion ratios.

In summary, the size of each field is designed based on the nature of the data it is expected to store, balancing storage efficiency with the flexibility to accommodate a wide range of values. This schema provides a structured and interlinked framework for capturing and retrieving detailed information about items, contracts, bid documents, and measurement units.

## 2.3 Conclusion

In this chapter, the complex challenges associated with standardizing and organizing tender bid documents for watermain and sanitary sewer capital works were discussed. These records are primarily sourced from three anonymized Canadian cities. The initiative is marked by its focus on converting a diverse array of documents, each with its unique formatting and structure, into a coherent and unified database. This pivotal transformation is not just a technical exercise but a strategic move to enhance the accuracy and efficacy of engineering estimates and inflation calculations in municipal projects, addressing the long-standing issue of information interoperability.

At the crux of this endeavor is the innovative integration of ontology and natural language processing techniques, which proved instrumental in ensuring the precision and integrity of data. These methodologies underpinned the data transformation, facilitating its standardization and making it conducive to advanced analysis and application. The methodology's standout feature is its adaptability, ensuring that the system remained relevant and robust amidst potential changes in data formats, styles, and contents. The integration of provenance records further bolstered this framework, providing essential traceability and accountability in the data handling processes.

In parallel, the chapter detailed the meticulous design and structure of the relational database schema, a cornerstone for the data's processing, standardization, and analytical processing. The schema is crafted to support consistency, data integrity, and extensibility, ensuring close alignment with the ontology's semantic rules. Each component of the schema is carefully designed to play a specific role in data storage and standardization, facilitating comprehensive queries and detailed data analysis. The thoughtful balance between storage efficiency and the ability to accommodate diverse data types is a key consideration in the schema's design.

The chapter also illuminated the broader implications of ontology in civil engineering, demonstrating its effectiveness in improving data quality and organization. The relational database format, employed for storing the organized data in the core database, exemplifies the ease and efficiency brought about by ontology in database management. The concept of data provenance was highlighted as a critical element, allowing for efficient error correction and audits. This feature is especially crucial given the evolving nature of new records and contracts, which often exhibit variations in format, style, and content.

In conclusion, this chapter has not only addressed the immediate challenges of standardizing tender bid documents in civil engineering but has also set a precedent for efficient, informed decision-making in municipal engineering projects. The methodologies and systems developed herein offer a blueprint for other cities and municipalities to enhance their data management and analysis capabilities. The integration of advanced techniques like ontology, data pre-processing, error detection and correction, and the incorporation of provenance flags have established a sophisticated and effective strategy for managing complex datasets in municipal engineering. This work significantly contributes to the field of civil engineering, promising applications beyond water systems and into other domains where chronological document management and contextual consistency are crucial.

# Chapter 3

# Automatic Record Classification

## 3.1 Introduction

Water utilities are crucial in providing residents with a reliable and clean water supply and managing water conservation, treatment, distribution, billing, and other essential tasks. At the core of these responsibilities lies the necessity for municipalities to construct in-house engineering cost estimates, primarily derived from historical tender-bid documents' unit cost indices. These indices are vital for designing and tendering new capital works projects concerning watermain and sanitary sewer systems. However, the critical role of water utilities in ensuring efficient and cost-effective water management faces significant challenges, particularly in the meticulous and complex process of extracting and analyzing historical project cost information.

The extraction of historical project cost information is a labour-intensive manual process. Also, the accuracy can be inconsistent, depending mainly on municipal experts' expertise and their personal preferences. In addition, the challenge of inconsistency arises during the process of rescaling or calculating the unit cost across different historical projects. Factors such as the unique values of materials and services, the size of projects, and inflation over time further compound this complexity. A recognized solution to this problem involves normalizing projects using a unit cost index [Younis et al., 2016]. This process requires carefully disaggregating project components and rescaling to a standardized unit project. Ensuring the correct identification and categorization of imported tender items across all available projects from contractors is vital for accurate and consistent results.

However, the variability in individual engineers' preferences and item categorizations may lead to disparate price estimates. The preceding chapter highlighted the lack of standardized data sources for watermain and sanitary sewer capital work projects. It renders current methodologies for standardizing unit cost estimates within a municipality both insufficient and highly dependent on individual cost-estimating engineers' expertise and practices. The problem becomes even more intricate when expanding the scope to multiple municipalities across regional, provincial, or national scales.

To deal with this issue, engineers usually limit their historical unit cost calculations to the most recent tender-bid documents to mitigate these challenges. They avoid rescaling to account for inflation and mostly adhere to similar tender-bid document style guidelines [Rehan et al., 2016]. However, inconsistencies can still arise during data import due to divergent contract records from various sources. Thus, creating a structured and homogeneous dataset is paramount for ensuring systematic access to historical records.

The preceding chapters have identified that existing methodologies. While those are helpful, they fall short of providing a universally applicable, automated solution. They lack the ability to effectively standardize and classify tender-bid items on a large scale while accounting for the nuances and complexities inherent in these documents. This gap in methodology is particularly evident when considering the challenges of rescaling or recalculating unit costs across various historical projects, further complicated by factors like material values, project sizes, and inflation.

Furthermore, the reliance on recent tender-bid documents and the avoidance of inflation rescaling introduces another layer of inconsistency. The variability in contract records and engineers' subjective nature of data importation lead to a fragmented approach to dataset creation. This inconsistent approach hinders the development of a structured, homogenous dataset essential for systematic access to historical records and accurate unit cost analysis.

The objective of this chapter is to explore the application of artificial intelligence (AI) models for the purpose of automating unit cost computation using historical watermain and sanitary sewer capital works projects. This automation and consistent classification of historical tender-bid documents aim to improve the accuracy of unit costs and develop more reliable engineering estimates for capital work projects. An essential part of this process involves adopting and evaluating various classification methodologies to ensure the required accuracy and performance are met.

In this chapter, we delve into developing an automated AI model by first leveraging machine learning and then subsequently, artificial intelligence to address the complex relationships inherent in tender-bid data. The chapter commences by examining initial classification methodologies, highlighting the limitations of distance metrics and ontology-based approaches. It then transitions to a detailed discussion on feature extraction, notably implementing the "Bag-of-Words" model, which is pivotal in natural language processing.

The next section of the chapter is dedicated to exploring the decision tree and its extension random forest (RF) classifiers, where it is introduced to enhance accuracy. This segment also examines the RF's superiority over other classifiers like Naive Bayes and k-nearest neighbours in pattern recognition within the dataset. Following RF, the data quality and accuracy of the model are still inadequate, and essential transition to deep learning is discussed. This transition, necessitated by the limitations of RF and its unsuitability with sequential data representation, emphasizes the adoption of deep learning methodologies and, particularly, the Long Short-Term Memory (LSTM) structure. At this point, the decision to favour LSTM models over Generative Pretrained Transformers (GPT) is examined, considering factors such as computational efficiency, training dataset size requirement, and interpretability.

To reconnect with this chapter's main objective, the methodology's innovation and contribution converge to create a sophisticated and adaptable AI model, aligning with clarifying items in tender-bid documents pertaining to watermain and sanitary sewer capital works. This model is accurate and versatile in its current form and shows promise for efficient processing of unsupervised data in the future. Therefore, the model can replicate the services that a professional engineer would provide when estimating unit costs.

The chapter concludes by demonstrating the DL model's classification capabilities of tender-bid items, as evidenced through a confusion matrix, RMSE, and R-squared values, showing strong predictive performance, particularly for watermain-related predictions. The Results section compares the performance of bidirectional and unidirectional LSTM models, with no marked advantage for BiLSTM. Emphasizing the importance of continuous improvement, the progressive improvement of the training data subsection highlights iterative refinement of the training-validation dataset, facilitated by collaborative efforts between the expert and the DL model, leading to increased dataset precision.

## 3.2 Methodology

This section describes the methodology employed in this study, systematically designed to cover all steps of data preparation, classification, and model development. The method is designed for high precision and reliability in classifying large volumes of tender and bid

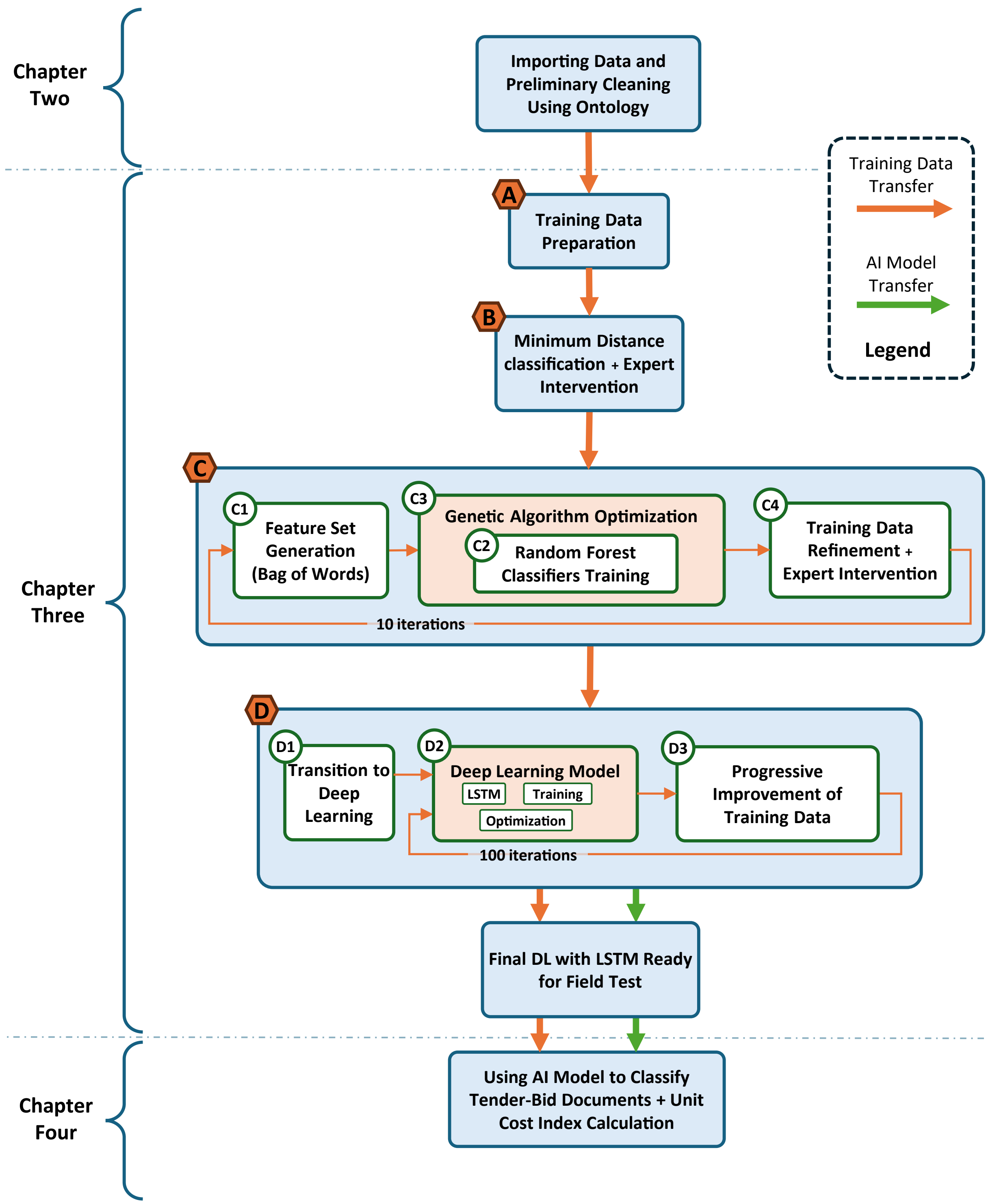



Figure 3.1: Comprehensive Methodology Flowchart for Data Preparation and AI Model Development in Water System Infrastructure Projects

data specific to watermain and sanitary sewer capital works projects. The process involves several stages of evolution, each designed to progressively refine the model and dataset, addressing the limitations encountered in the previous stages. Figure 3.1 is provided as a procedure guideline to show the transformation of both data and AI model to achieve the objective of this chapter.

- **Importing data and preliminary cleaning using ontology:** The methodology commences with the importation and preliminary data cleanup using ontology-based relations and restrictions, as elaborated in Chapter 2. This initial step is pivotal for standardizing the dataset, thereby providing a consistent foundational framework for the subsequent methodological stages. The ontology's role is crucial in ensuring a basic level of consistency and accuracy in the initial treatment of the data.

- **Step A, Training data preparation:** This step entails the data preparation methodology for classifying tender-bid documents in water infrastructure projects. Over 250 documents from three Canadian cities are analyzed, with each item or "record" comprising various fields like description, unit, and cost. The approach addresses categorization inconsistencies by standardizing item parts and sub-parts. The data is divided into training-validation and testing sets and are prepared for a 5-fold cross-validation, and the testing set, comprising different contracts, assesses the classifier's performance. This meticulous data preparation is essential for developing an accurate and efficient classification model for item classification.

- **Step B, Minimum distance classification with expert intervention:** The first stage involved employing a minimum distance method to calculate the proximity between each new item's description and existing items in the ontology. This process is augmented by matching the ontology requirements as additional constraints are added to the distance value, facilitating, and limiting the generation of potential class lists for each item. However, this method's reliance on human intervention, specifically requiring an engineering expert to manually oversee and select the most appropriate classification for each item, introduces subjective biases. Despite these challenges, this stage is essential in establishing a baseline dataset with preliminary class labels for the purpose of supervised learning, even though these classified training items are considered imperfect and occasionally misidentified.

- **Step C, Machine learning analysis and data enhancement:**

- **Step C1, Feature set generation using "Bag-of-Words":** In response to the initial method's inadequacies mentioned in Step B, the methodology is transitioned to utilizing a Bag of Words (BoW) representation for feature vector generation. BoW is selected due to its simplicity and effectiveness in capturing the nuances of textual data.
- **Step C2, Random forest classifier training:** At this point, the random forest (RF) classifier is utilized. This shift represents a significant methodological advancement, utilizing the RF's interpretability and ease of implementation. Despite a notable improvement in accuracy (reaching up to 90% in certain classes), the model's performance is uneven across different classes, indicating the need for further refinement.
- **Step C3, Genetic algorithm optimization:** Addresses the model's uneven performance through the integration of genetic algorithms aimed at refining feature selection by selecting the most suitable words for classification. Positioned as the third segment in this step, it effectively encompasses Step C2, illustrating the logical progression of data flow and procedures to include genetic algorithm (GA) optimization. This strategy significantly improves the random forest classifier's performance, with an average accuracy of 80.12% and peaks of 95% in some classes. The genetic algorithm's efficiency in navigating large feature spaces and its ability to avoid overfitting are critical factors in this improvement.
- **Step C4, Training data refinement and expert intervention:** This step encapsulates the critical methodology of iterative refinement of the training data for tender-bid item classification in water infrastructure projects. The process, underpinned by meticulous analysis of misclassified instances indicated by confusion matrices and expert collaboration, leads to the detection, re-evaluation, and correction of mislabeled data points. Spanning ten iterations and involving various methodological parts (C3, C2, and C4), this approach resulted in updating approximately 6% of the training data labels, which consequently improved the average accuracy of the model from 80.12% to 85.96%. However, these iterations also highlighted the limitations of the Random Forest model in handling the dataset's complexity, particularly its inability to enhance accuracy further or consistently identify error patterns beyond the achieved accuracy level.

- **Step D1, Transition to deep learning:** This part details the shift from a random forest classifier to a more advanced Long Short-Term Memory (LSTM) model, a significant step in handling a large dataset from various industrial partners. Initially, manual classification assisted by minimum distance calculations was used but proved impractical for the dataset's volume. The decision tree method, while initially useful, reached an average accuracy limit of 85.96% due to its inability to process sequential text information effectively. This limitation led to the adoption of LSTM models, chosen over the other alternative Generative Pretrained Transformers (GPT) due to LSTM's computational efficiency, suitability for the dataset's size, capability in capturing patterns in sequential data, and customization potential. The LSTM's implementation marks a strategic evolution in the project's approach to data classification, setting a foundation for future integration of more complex models like GPT.

- **Step D2, Deep learning model:** The development and configuration of the Deep Learning (DL) artificial neural network model is described in this Step. It includes its architecture and component functionalities, supported by a practical example demonstrating the model's data transformation process. Equipped with LSTM, the DL model significantly enhances the pattern detection capabilities through the process of training the LSTM model.

- **Step D3, Progressive improvement of training data:** In this step, an additional 5% of misclassifications within the dataset were identified and corrected. This improvement is achieved through a rigorous process of over 100 iterations involving the training of the deep learning model and subsequent review of misclassified items. This meticulous refinement enhances the model's precision, leading to an increase in classification accuracy beyond the initial target of 92%. Following these 100 iterations, the performance of the deep learning model is evaluated using test data. The final iteration of the model, demonstrating the most effective classification accuracy, is selected as the definitive version for practical application in the field.

- **Chapter 4,** This section signifies the practical application of the developed deep learning model. Aligned with the project's objectives, the model is tailored to automate or consistently classify historical tender-bid items. The remainder of this chapter delves into the application of the model's classification outputs for calculating unit costs in tender-bid documents. This step represents the real-world implementation of the DL model, demonstrating its utility in the field.

An important consideration at this stage is the potential issue of overfitting, a common challenge when a model and data are overly optimized in tandem. Nonetheless, in this specific context, the risk of overfitting is substantially lessened due to the ground truth established by the engineering expert. Contrary to situations with ambiguous or unidentified target classes, like a cancer prediction model, our model functions within a clearly defined and expert-validated classification system. Unlike observational labels that are typically accurate but not infallible, the initial data classes in our model were derived through a semi-automatic process involving human labelling, which can inadvertently introduce label noise. As such, the engineering expert plays a critical role in verifying the accuracy of each item's classification. Ultimately, this expert-guided verification of the training data serves as a protective measure, ensuring that the model remains precise and in line with practical, real-world standards, thus effectively mitigating the risks commonly associated with overfitting.

### 3.2.1 Step A, Training Data Preparation

Data preparation, a crucial phase in the proposed methodology, directly influences the efficiency and precision of the ensuing classification and prediction models. This subsection elucidates the processes involved in this stage. It focuses on elements such as input data, record categories, record inconsistencies, and data segregation for training, validation, and testing purposes.

#### Input Data

The previous chapter detailed the procurement of data for this project, encompassing over 250 water system infrastructure tender/bid documents from three major Canadian cities. The data is structured in lists comprising individual item sets, as illustrated in Table 2.4 on Page 44. Each element called a "record", has a distinctive description and cost value. Following the process of data importation and cleanup, each record is disaggregated into the subsequent fields (table columns):

a) Description (char [512], for instance, "supply and install of 150mm diameter PVC pipe"),

b) Unit (char [32], for example, "meter"),

c) Unit price (double, for instance, "80.70 CAD"),

d) Quantity (double, for example, "600"),

e) Contract (char [32], for instance, "redacted.name"),

f) City (char [32], for example, "redacted.city"),

g) Original category (char [32], for instance, "sanitarysewer").

The following fields are absent and will be incorporated during the classification process:

1. standard-part (for instance, "SanitarySewer"), and
2. standard-sub-part (for instance, "SS_Pipe").

The original section, determined by the municipality or contractor during tender issuance, is unconstrained and can vary (e.g., "Roads", "Road", "Road Works"), unlike the standard-part. Conversely, the standard-part is confined to the items outlined in Eq. 3.7 on Page 103. Different municipalities or contractors might organize items differently, leading to inconsistencies across tender-bid documents. For instance, in the scenario provided above, the original section is "SanitarySewer", while the correct one (according to the standardized definition) is "Watermain". Table 3.1 furnishes examples of each standardized part and sub-part of items, as defined by the contractor for the Watermain and SanitarySewer categories.

#### Record Categories

A widely accepted classification standard for an item's "part" comprises categories such as Road, General, Sanitary Sewer, Storm Sewer, Watermain, Provisional Items, and Miscellaneous, as illustrated in Table 2.4 on Page 44. The most prominent standard-parts engaged in this thesis are "Watermain" and "Sanitary Sewer", each of which is further dissected into standard-sub-parts. To streamline the design of the automatic classification and assemble more training input data, a specific set of item categories with analogous characteristics are consolidated into four primary standard-sub-parts for Watermain and three for Sanitary Sewer. The standard-sub-parts designated for Watermain include: WM_Services, WM_Pipe, WM_Valve, and WM_Hydrant; for Sanitary Sewers, these are: SS_Pipe, SS_Lateral, and SS_Manholes (refer to Figure 3.1).

From a machine learning standpoint, the availability of supervised input data is paramount. Therefore, securing classification by engineering expert on the standard-parts and sub-parts of a testing and validation set of contracts is crucial.

| Standard Part | Standard Sub-Part | General Item Description |
|---|---|---|
| Sanitary Sewer | Manhole | any item related to constructing a new or removing a manhole (maintenance hole) acceptable diameter range (1200mm to 3000mm) |
| Sanitary Sewer | Lateral | sanitary sewer items related to laterals, including (not limited to): (PVC, cast iron, asbestos cement, concrete, steel case pipes), (jack) bore, stub, break to the main line, direction drill, open cut, grouting, dye test, tv inspection, trenchless, cleanout, Inspection |
| Sanitary Sewer | Pipe | sanitary sewer items related to pipes (PVC, reinforced concrete, CIPP) (open cut, various sizes), new pipe, new connection, connection to existing |
| Watermain | Pipe | watermain items related to pipes, including (not limited to): new pipe installation (PVC), bore jack, direction drill, open cut, plugging, trenchless, copper pipe service installation, tapping sleeve, concrete pressure, casing jack and bore, |
| Watermain | Hydrant | watermain items related to hydrants, including new hydrant, bend tee fittings, reconnection of an existing hydrant |
| Watermain | Service | watermain items related to water services, including (not limited to): cathodic protection abandoning old watermain, removing/disposing of valve boxes/hydrants/pipes, installing new water service with type K copper, trenchless installation, disconnection and cap existing watermain, all appurtenances to connect to existing watermain, protection of existing watermain with concrete, leak repair, replacement of service, remove and replace of curb stop and box at the property line |
| Watermain | Valve | watermain items related to valves, including (not limited to): tapping sleeve valve, water valve and box, curb stop and box, shutdown delay, valve cleaning, curb stops, curb boxes, main stops |
| General | Not Applicable | general items, including (not limited to): bonding, fences, wooden barriers, maintaining and removing silt control devices, excavated soil retaining, pre-condition survey, site office, construction layout, unshrinkable fill, traffic control, clear stone, control monument |
| Provisional Item | Not Applicable | provisional items, including (not limited to): removing/replace of trees/stumps, pavement markings, crossing line painting, valve cleaning, contingency allowance, providing bulkheads at the concrete box, cleaning and grubbing, supply and installing calcium chloride incidental time and rates, lean mix concrete, dewatering, application of water, shoring and bracing, test holes, |
| Road | Not Applicable | road items, including (not limited to): granular materials, road excavation and disposal road base material, cold mix/recycled asphalt, temporary barriers, saw cuttings, speed bumps, dowel supply and installation, concrete curb gutter, building and adjustment of water valve chamber, repairing cracked sealing, salvaging road materials, relocation and repair of culverts, HDPE culverts, dead-end barricade OPSD, hot mix asphalt/cement, driveway restoration, boulevard grading, CSP culverts, asphalt milling, |
| Storm Sewer | Not Applicable | storm sewer items, including (not limited to): insulation or service, granular bedding backfill, concrete storm box, manufacture plug, catchbasins, adjusting storm manholes flush and tv inspection of storm sewers, abandoning old storm sewers, PVC pipes, culvert repair and restoration, and cleaning of silts, (reinforced) concrete storm sewers, plugging pressure grouts, precast chamber of storm manholes, catchbasin leads, perforated subdrains, supply/install/repair of catchbasin frame and grades |

Table 3.1: Breakdown of standard parts and sub-parts.

| Item # | Description | Part | Qty | Unit | Unit Price | Total Price | UID # | City | Contract |
|---|---|---|---|---|---|---|---|---|---|
| B4.a.1 | box hydrant opsd set tap valve water | Watermain | 135 | m | 57.00 | 7,695.00 | A_A53 | City A | Contract A |
| 1.9 | 300 375 cb dispose exist fire hydrant mm remove sewer storm | Road | 3 | ea. | 311.35 | 934.05 | B_B8 | City B | Contract B |
| 11 | box cap end fire hydrant include remove valve | Watermain | 3 | each | 850.00 | 2,550.00 | C_C45 | City C | Contract C |
| B3.b.2 | Hydrants complete with anchor tee 150mm diameter valve boxes and anodes according to opsd 1105.010 (provisional) | Roads | 1 | ea. | 556 | 556 | A_A99 | City A | Contract A |

Table 3.2: Representative items from contracts A, B, and C emphasize hydrants, while standard part and sub-part categorizations are missing.

### Records Inconsistency

A considerable source of inconsistency lies within each item's "Original-Part" field, as demonstrated in Table 3.2. For instance, while Contracts B and C categorize "services performed related to hydrants" under the "Watermain" part (in line with the most common assumption and the standard), Contract B classifies it under the "Road" part. Furthermore, Contract B uses both "Road" and "Roadworks" parts, even though only "Road" is an acceptable standard-part name. In cases involving labels like "Roads" and "Road works", the solution involves renaming both parts to "Road" and merging them into a single standard-part.

Addressing the discrepancy between "Watermain" and "Road" necessitates comprehending the context of water system contracts. For human operators, determining the appropriate standard-part can be challenging, necessitating expert knowledge. The introduced automatic classification method (DL) learns the pattern of item descriptions for all standard-parts and standard-sub-parts from the provided training data. As a result, the classifier can precisely determine the corresponding standard-part for an item with an unknown or incorrect part. Table 3.2 presents an example: although all four items describe "services, installation, or removals concerning hydrants", the second item is misclassified under the "Road" part. The classification model aims to detect and correct such errors by accurately assigning the item to its corresponding class (e.g., "Watermain" standard-part and "WM_Hydrant" standard-sub-part in this case).

### Training, Validation, and Testing Data

The data are divided into two main segments for developing the DL model: training-validation and testing contract data. The training and validation dataset comprises over 250 tender-bid documents from three anonymized cities' archived contracts/tenders. This dataset is utilized to construct the DL classifier, and any alterations could affect the performance of the DL model. Conversely, the testing data is exclusively used to assess the DL model's performance.

The training and validation of the DL model are conducted using 5-fold cross-validation. A separate set of contracts from the three cities, with no overlap with the training data, is used as testing data.

| City | Total # of Records | # of contracts | Watermain # | Sanitary Sewer # | Other Cat. # | Used for |
|---|---|---|---|---|---|---|
| Reference Items (manual generation) | 1161 | - | 281 | 242 | 637 | Training / Validation |
| City A | 736 | 3 | 165 | 112 | 459 | Training/Validation |
| City B | 1526 | 13 | 444 | 301 | 781 | Training/Validation |
| City C | 403 | 2 | 30 | 227 | 146 | Training/Validation |
| City A | 589 | 3 | 119 | 83 | 387 | Testing |
| City B | 265 | 2 | 83 | 44 | 138 | Testing |
| City C | 336 | 2 | 38 | 148 | 150 | Testing |

Table 3.3: Details of items utilized for training, validation, and testing to construct and evaluate the performance of the proposed DL classification model

### 3.2.2 Step B, Minimum Distance-Based Classifier Using Ontology

In the initial stages of our research, we employed a distance-based ontology algorithm for classifying items within our dataset. This early methodology, utilizing the RS-Means dataset as a classification benchmark, was instrumental in laying the foundation for more advanced classification techniques.

The approach involved measuring the similarity between item descriptions in our dataset and those in the RS-Means ontology, taking into account parameters such as "item standard-part," "item unit," and "item unit price." However, post-improvement of the dataset and addressing label noise, the method achieved an accuracy of 80.12%. It became clear that the semi-automatic nature of this approach was insufficient for the complete automation of the classification process.

A key feature of this method was its independence from pre-classified data, which is crucial for supervised learning. The min-distance calculation aided experts in manual record classification, preparing the data for subsequent model training. Nevertheless, the approach proved impractical and time-consuming for the vast volume and complexity of the dataset.

The ontology's role in this method involved aligning similar words from the dataset for accurate matching, extracting word roots from descriptions, and ensuring precise unit matching. Items were arranged alphabetically based on word roots to simplify edit distance calculations, with operators presented with the top ten closest matches for decision-making. Each item was weighted and prioritized to assist in this process.

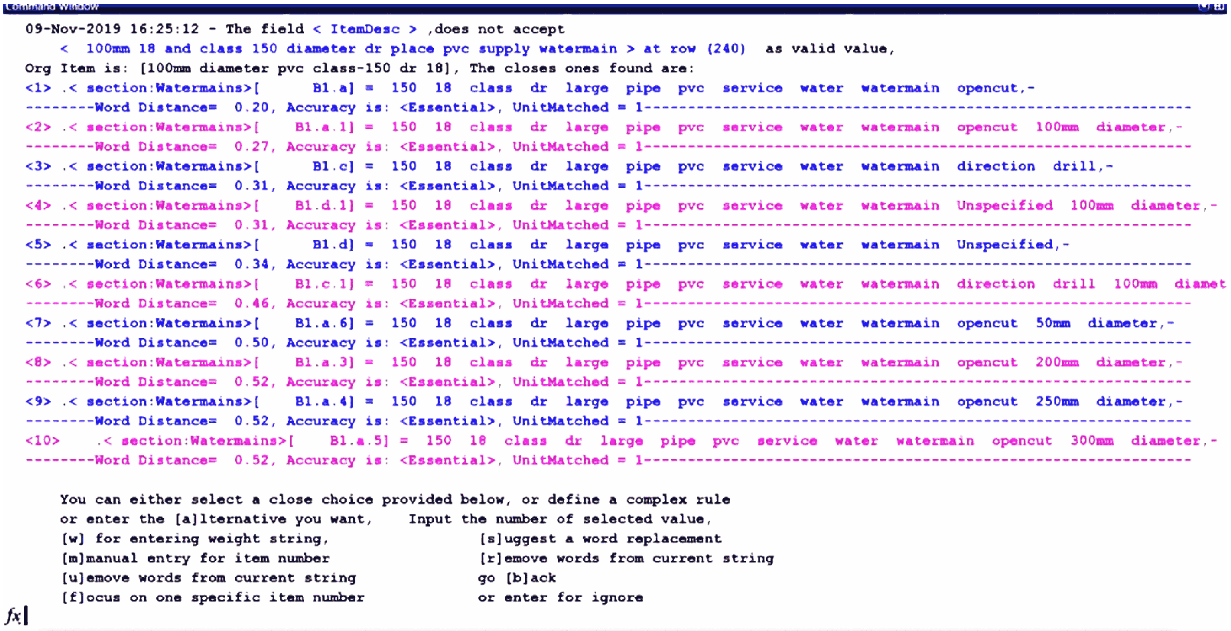



Figure 3.2: Implementing Ontology-based item detection and matching, combines manual and automated processes.

Despite these measures, the ontology-based distance approach had significant limitations. It depended heavily on operator input, with a 40% decrease in accuracy without human involvement. Consequently, this method was not chosen as the primary tool for importing new items but was used for classification and as a sanity check for mapping results, as depicted in Figure 3.2. These insights guided the transition to more automated and sophisticated classification techniques in subsequent stages of our research.

### 3.2.3 Step C1, Feature Set Generation using "Bag-of-Words"

Classifiers serve as an effective mechanism for identifying standard parts and sub-parts. In contrast to previous methods that primarily relied on a reference list of pre-classified items, developing a classifier necessitates appropriate training and testing data. The training dataset comprises contract or tender summaries, which experts have thoroughly examined to verify the correct assignment of items to standard parts. The primary advantage of a classifier is its ability to discern patterns and relationships, such as syntactic and semantic associations within item descriptions, thereby enabling the automation of future item classification. However, the efficacy of a classifier depends on consistent data sources. Retraining the classifier through a semi-automated process becomes necessary if the contract's content changes or new data emerges.

Pre-processing of item descriptions is required to classify them into standard parts and sub-parts. This pre-processing is facilitated by ontology, which breaks down the contract description into individual words, removes stop words, and applies rules to eliminate redundant words or numbered items. The feature extraction process includes word stemming and embedding. Terms are reduced to their root forms, and the ontology ensures consistency by standardizing variations of words to this root form. Words with low-information content are deemed irrelevant are discarded, leaving only significant ones. In word embedding, weights are assigned according to the frequency of words within the description.

In natural language processing and information retrieval, the "Bag-of-Words" (BoW) model serves as a fundamental approach for representing text data [Kim et al., 2005]. The BoW model, which is both straightforward and efficient, represents text (be it a sentence or document) as a bag or multiset of its words, disregarding the order and grammar but preserving the frequency or presence of words. The construction of a BoW representation begins with the compilation of the vocabulary, which encompasses all unique words discovered across the entire dataset. Each word within the vocabulary is assigned a unique index value. After that, for each text within the dataset, a vector is created where each entry corresponds to the frequency or presence (depending on the BoW variant used) of a word from the vocabulary in the text.

For instance, if we consider a vocabulary of ["tree," "this," "wind," "house," and "is"], the BoW representation of the sentence "this is a tree" would be [1, 1, 0, 0, 1] relative to this vocabulary. Each vector entry indicates the corresponding word's presence in the sentence's vocabulary. The semantics of the sentence are captured solely by the presence or absence of words, regardless of their relative order.

While the BoW model is useful, it presents certain limitations. Notably, it ignores word order, which can sometimes carry significant semantic information. Furthermore, it treats all words equally, even though some words may have a more significant semantic impact. Despite these limitations, the BoW model remains a foundational technique in numerous natural language processing tasks. For applications requiring greater semantic complexity, alternative text representation techniques, such as Term Frequency-Inverse Document Frequency (TF-IDF) vectors or word embeddings like Word2Vec or GloVe, can be used.

Figure 3.3 on Page 88 shows the results of stemming and applying the bag-of-words method to the running examples. The attribute vector contains over 128 words after removing redundancies and irrelevant words. Only words with a frequency of five or more are shown in the figures. The discrepancies among similar contracts from various cities underscore the classification problem's complexity.

### 3.2.4 Step C2, Random Forest Classifier

The Random Forest technique effectively mitigates overfitting and enhances the accuracy of decision trees [Biau and Scornet, 2016]. This approach involves generating multiple decision trees with varying parameters, leading to diverse classification outcomes. A key advantage of the Random Forest is its ability to provide predictions with confidence levels. For instance, if a hundred trees classify a contract item's standard part, with ninety-eight indicating Watermain and one each for Sanitary Sewer and Road, the item is classified as Watermain based on majority voting. This method generally surpasses the reliability of a single decision tree.

Consider a scenario with a minority winner: out of a hundred trees, forty-nine vote for Watermain, forty-eight for Sanitary Sewer, and three for Road. Though Watermain is selected, the close vote suggests a less confident classification. Instances with minority votes are documented for label noise investigation and expert review.

Figure 3.4 exemplifies a decision tree classifying items into the sanitary sewer standard part, demonstrating how word patterns influence classification. The training data's accuracy is crucial for the classifier's effectiveness. Random Forest's ensemble strategy, aggregating predictions, offers nuanced classification, accommodating data variability and complexity.

The resilience of Random Forest to errors or data contamination is another benefit [Dietterich, 2000]. This resilience is vital given the susceptibility of the dataset to errors. Testing other classifiers like Naive Bayes, k-nearest neighbour, and linear discriminant analysis revealed suboptimal performance compared to Random Forest. Specifically, Naive Bayes underperformed due to its feature independence assumption, which is not applicable in this context where item description words are correlated.

The core of this research involves using decision trees and Random Forests for classification. Decision trees act as structured flowcharts, dividing datasets based on attributes. Configured with a maximum depth of ten and using the Gini index for splitting nodes, they achieved an average accuracy of 85.96%. Random Forests, synthesizing outcomes from multiple trees, address individual tree limitations, especially in large, complex datasets [Breiman, 2001, Quinlan, 1986].

| | 15 | asphalt | b | bedding | c | class | concrete | connect | curb | cut | dia | dispose | driveway | excavate | exist | granular | install | m | manhole | mm | new | open | opsd | pipe | place | precast | pvc | remove | replace | sanitary | service | sewer | stone | storm | supply | valve | water | watermain |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| B_B2 | | | | | | | 1 | | 1 | | | 1 | | | 1 | | | | | | | | | | | | | 1 | | | | | | | | | | |
| B_B3 | | 1 | | | | | | | | | | 1 | 1 | | 1 | | | | | | | | | | | | | 1 | | | | | | | | | | |
| B_B4 | | | | | | | | | | | | | 1 | | 1 | | | | | | | | | | | | | 1 | | | | | 1 | | | | | |
| B_B5 | | | | | | | | | 1 | | | | 1 | | 1 | | | | | | | | | | | | | 1 | | | | | 1 | | | | | |
| B_B6 | | | | | | | | | | | 1 | 1 | | | 1 | | | | | 1 | | | | | | | | 1 | | 1 | | 1 | | | | | | |
| B_B7 | | | | | | | | | | | 1 | 1 | | | 1 | | | | | 1 | | | | | | | | 1 | | | | | | | | | | 1 |
| B_B8 | | | | | | | | | | | | 1 | | | 1 | | | | | 2 | | | | | | | | 1 | | | | 1 | | 1 | | | | |
| B_B9 | | | | | | | | | | | | 1 | | | 1 | | | | 1 | | | | | | | | | 1 | | | | | | 1 | | | | |
| B_B1 | | | | 1 | 1 | | | | | | 1 | | | | | | 1 | | 1 | 1 | | | | | | | 1 | | | 1 | | | | | 1 | | | |
| B_B11 | | | | | | | 1 | | | | 1 | | | | | | 1 | | 2 | 1 | | | | | | 1 | | | | 1 | | | | | 1 | | | |
| B_B12 | | | | | | | | 2 | | | | | | | | | 1 | | | 1 | | | | | | | | | | 2 | | 1 | | | 1 | | | |
| B_B13 | | | | | | | | 1 | | | | | | | 1 | | | | 1 | | | | | | | | | | | 1 | | | | | | | | |
| B_B14 | | | | | | | | 1 | | | | | | | 1 | | | | 1 | | | | | | | | | | | 1 | | | | | | | | |
| B_B15 | | | | | 1 | | | | | | | | | | | 1 | | | | | | | | | | | | 1 | | | | | | | | | | |
| B_B16 | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | |
| B_B17 | | | | | | | | | | | | | | | | | | | | 4 | | | | 1 | | | | | | | | | | | | | | |
| B_B18 | | | 1 | 1 | 1 | 1 | | | | | | | | | | | 1 | | 1 | 1 | | | | 1 | | | | | | | | | | 1 | 1 | | | |
| B_B19 | | | 1 | 1 | 1 | 1 | | | | | 1 | | | | | | 1 | | | 1 | | | | 1 | | | | | | | | | | | 1 | | | |
| B_B2 | | | | | | | 1 | | | | 1 | | | | | | 1 | | 2 | 1 | | | | | | 1 | | | | | | | | 1 | 1 | | | |
| B_B21 | | | | | | | 1 | | | | 1 | | | | | | 1 | | 2 | 1 | | | | | | 1 | | | | | | | | 1 | 1 | | | |
| B_B22 | | | | | | | | 1 | | | 1 | | | | | | | | | 1 | | | | | | | | | | | | 1 | | 1 | | | | |
| B_B23 | 1 | | | | | | | | 1 | | 1 | | | | | | | 1 | | 1 | | | | 1 | | | 1 | | | | | 1 | | 1 | | | | |
| B_B24 | | | | | | | | | | | | | 1 | | | | | | | | | | | | | | | | | | | | | | | | | |
| B_B25 | 1 | | | | | | | | | | 1 | | | | | | 1 | | | 1 | | | | | | | | | | | | | | | 1 | | | 1 |
| B_B26 | | | | | | | | | | | | | | | | | 1 | | | | | | | | | | | | | | | | | | 1 | 1 | | |
| B_B27 | | | | | 1 | | | 1 | | | | | | | | | 1 | | | | | | | | | | | | | | | | | | 1 | 1 | 1 | |
| B_B28 | | | | | | | | | | | | | | | | | | | | 1 | | | | | | | | | | | | | | | | | | |
| B_B29 | 1 | | | | | | | 1 | | 1 | 1 | | | | 1 | | | | | 2 | 1 | 1 | | | | | | 1 | 1 | | 2 | | | | | | 2 | |
| B_B3 | 1 | | | | | | | 1 | | | | | | | 1 | | | | | 2 | 1 | | | | | | | 1 | 1 | | 1 | | | | | | 1 | |
| B_B31 | | | | | | | | 1 | | | | | | 1 | 1 | | | | | | | | | | | | | | | | 1 | | | | | | 1 | |
| B_B32 | 1 | | | | | | | | | 1 | | | | | 1 | | | | | 1 | | | | | | | | | | | | | | | | | | 1 |
| B_B33 | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | |
| B_B34 | | | | | 1 | | | 1 | | | | | | | 1 | | | | | | | | | | | | | | | | | | | | | | 1 | |
| B_B35 | | | | | | | | 1 | | | | | | | | | | | | 1 | | | | | | | | | | | 1 | | | | | | 2 | |
| B_B36 | | | | | | | | 1 | | | | | | | | | | | | 1 | | | | | | | | | | | | | | | | | 1 | |
| B_B37 | | | | | 1 | | | | | | | | | | | 1 | | | | | | | | | | | | 1 | | | | | | | | | | |
| B_B38 | | | | | | | | | | | | 1 | | 1 | | | | | | | | | | | | | | | | | | | | | | | | |
| B_B39 | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | |
| B_B4 | | | 1 | | | | | | | | | | | | | 3 | | | | | | | | | 1 | | | 1 | | | | | | | 1 | | | |
| B_B41 | | 2 | | | | | | | | | | | 1 | | | | | | | | | | | | 1 | | | 1 | | | | | | | 1 | | | |
| B_B42 | | | | | | | 1 | | 1 | | | | | | | | 1 | | | | | | 1 | | | | | | | | | | | | 1 | | | |
| B_B43 | | | | | | | | | | | | | 1 | | 1 | | | | | | | | | | | | | | | | | | 1 | | | | | |
| B_B44 | | | | | | | | | 1 | | | | 1 | | 1 | | | | | | | | | | | | | | | | | | 1 | | | | | |
| B_B45 | | | | | | | | | | | | | | | | | | | | | | | | | 1 | | | | | | | | | | 1 | | | |
| B_B46 | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | |
| B_B47 | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | |
| B_B48 | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | |
| B_B49 | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | |
| B_B5 | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | |
| B_B51 | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | |

Figure 3.3: The stemming and bag-of-words representations of a sample tender-bid document, where only the highest frequency words are displayed (words with an occurrence of five times or more).

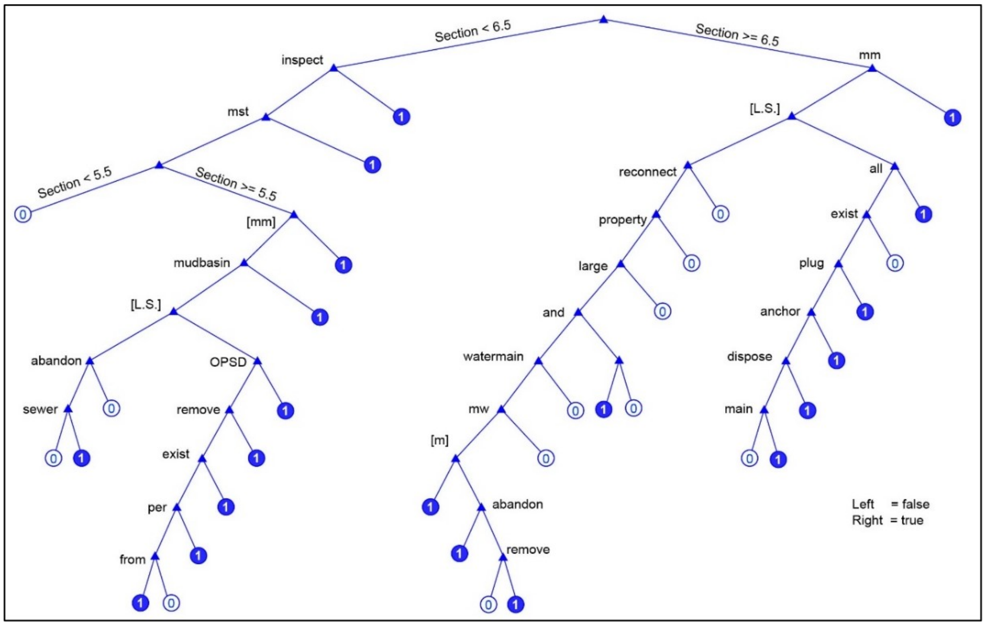



Figure 3.4: A sample of a decision tree classifier generated by the classification module determining if an item belongs to the sanitary sewer standard part.

### 3.2.5 Step C3, Enhancing Classifier Performance through Genetic Algorithm-Driven Feature Selection

In classifier optimization, the crux lies in the judicious selection of features. The genetic algorithm (GA), renowned for its ability to navigate large search spaces efficiently, was utilized. The GA commenced with a binary representation of a 736-word feature set extracted from an ontology dictionary. This set forms the foundation of the feature selection process. The GA's fitness function assessed each feature combination's efficacy, with specific emphasis on maintaining genetic diversity and optimizing classification accuracy. The GA's configuration included strict adherence to predefined constraints and a cap of five hundred generations to prevent overfitting. Additionally, the mutation rate was set at 1/N, dynamically adjusting as the number of features decreased. The results of the GA optimizations are feature vectors of 85 to 150 words.

To further enhance the understanding of this optimization process, it is crucial to delve into the specific role of the GA in classifier enhancement. The GA's prowess in feature selection is instrumental in distilling the essential elements from a vast pool of data, thereby facilitating the classifier's ability to discern and interpret complex patterns with greater accuracy and efficiency.

The meticulous optimization of the feature selection process, facilitated by the genetic algorithm (GA), significantly enhanced the random forest classifier's effectiveness. This process involved identifying the most potent features for precise classification, enabling the classifier to interpret complex relationships between words in contract item descriptions with improved precision and efficiency. The RF classifier's performance, as detailed in Figure 3.6 on Page 93 provided here and the Table 3.6 provided in the results section of this chapter on Page 110, demonstrated variability across different classes. For instance, in categories like provisional items, wm_hydrant, and ss_lateral, the accuracy rates were recorded at 34.12%, 66.67%, and 70.45%, respectively. These figures, though moderate, substantially exceed the baseline chance accuracy of 8.33%, indicating the classifier's relative effectiveness in these more challenging categories. In contrast, the classifier achieved exceptional performance in the majority of the classes, with accuracy rates surpassing 93.48%, thereby reflecting its overall robustness.

The study also uncovered limitations in the random forest classification approach, particularly in specific scenarios. For example, the wm_hydrant category, characterized by a limited number of samples, highlighted the challenges associated with insufficient data. Conversely, categories with an adequate number of samples, such as provisional items and ss_lateral, still struggled to achieve high accuracy, pointing to the intrinsic constraints of the random forest method in dealing with high input data variability and a large array of output classes.

A significant issue identified with the bag-of-words data representation, which disregards word order and perceives input text as an assortment of individual words. This assumption results in the loss of crucial sequential information, particularly relevant in the context of item descriptions. Despite carefully selecting the most relevant words via the genetic algorithm, the decision tree classifier's accuracy averaged at 85.96% and could not be increased further.

Figure 3.5 on Page 92 illustrates the classification design mechanism, encompassing stages like Ontology and Importing, Data Pre-processing, Classifier Training, Optimizing by Genetic Algorithm, and the Final Phase, outputting optimized classification. Each segment is vital to the solution's ability to identify standard parts and sub-parts in each contract item accurately.

### 3.2.6 Step C4, Enhanced Training Data Refinement and Expert Intervention

Training Data Refinement is a critical aspect of our methodology, focused on iteratively enhancing the accuracy and reliability of the Random Forest (RF) classifier in tender-bid item classification. This section delves into the details of the iterative refinement process, highlighting how each cycle contributes to refining the model.

The RF model undergoes dynamic training across multiple iterations. Initially, the process involves identifying and correcting misclassified instances in the training data with the help of an engineering expert. This step leads to progressive improvements in data quality and classification accuracy. The confusion matrix for the RF model, as shown in Figure 3.6, shows the results of several training iterations, where misclassifications were continually investigated and rectified. With each iteration, the RF model's performance improved, reflecting enhanced data classification and reduced confusion caused by inconsistencies in the training data. The confusion matrix also points out areas requiring improvement, marked by false negatives and positives, especially among closely related sub-parts.

Initially, the RF model played a pivotal role in classifying unclean data and refining the training dataset. Its user-friendliness and relative insensitivity to data errors made it a suitable initial tool for classification. Iterative enhancements in both data quality and RF hyperparameters suggested the potential of achieving performance comparable to more complex systems, as evidenced by the improved accuracy seen in Figure 3.6.

Furthermore, the interpretability of the RF model is one of its key strengths. It provides transparency in the decision-making process, which is essential for engineering experts. This clarity is invaluable when addressing discrepancies in tender items, facilitating intuitive understanding, and rectifying potential data or reasoning errors. The iterative refinement, coupled with the expert's input, ensures that the RF model not only becomes more accurate over time but also remains aligned with practical engineering standards.

The insights from this iterative refinement process highlight the importance of selecting appropriate machine-learning models and techniques for complex datasets. While the RF classifier and genetic algorithms brought significant improvements, their limitations underscore the potential need for exploring alternative machine learning algorithms or hybrid models.

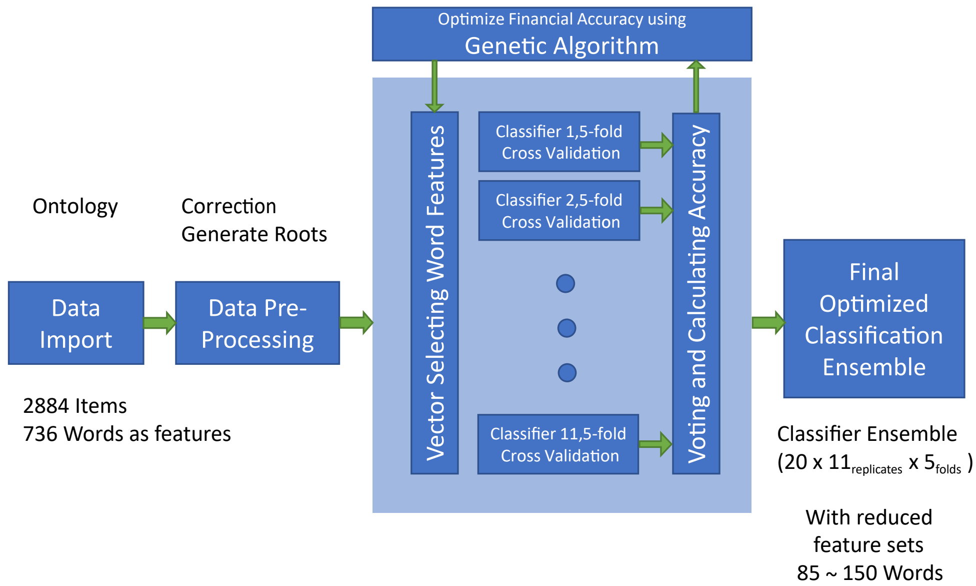



Figure 3.5: Block diagram of the decision tree classifier training and genetic algorithm optimization mechanism. Each block shows representative numbers of different records used from each city to populate the training.

### 3.2.7 Step D1, Transition to Deep Learning

The transition to Deep Learning represents a significant advancement in the methodology, signifying the transition from traditional decision tree classifiers to more advanced Long Short-Term Memory (LSTM) models. This section discusses the factors prompting this transition and the steps taken to adapt the vast, unstructured dataset for Deep Learning.

The need for a reliable and efficient classification method for the extensive dataset obtained from industrial partners across three cities drove this research. The unclassified state of the initial dataset deemed it unfit for immediate application with supervised learning methodologies. Consequently, a minimum distance (min-distance) calculation was initially suggested to assist the engineering expert in manually classifying records by measuring similarities between data points, thus preparing the data for subsequent model training. However, given the large dataset volume and the number of records needing classification, the proposed min-distance ontology method became impractical and time-consuming. In addition, the semi-automated process's susceptibility to human error introduced further complexity.

A random forest classifier was employed under engineering expert supervision to

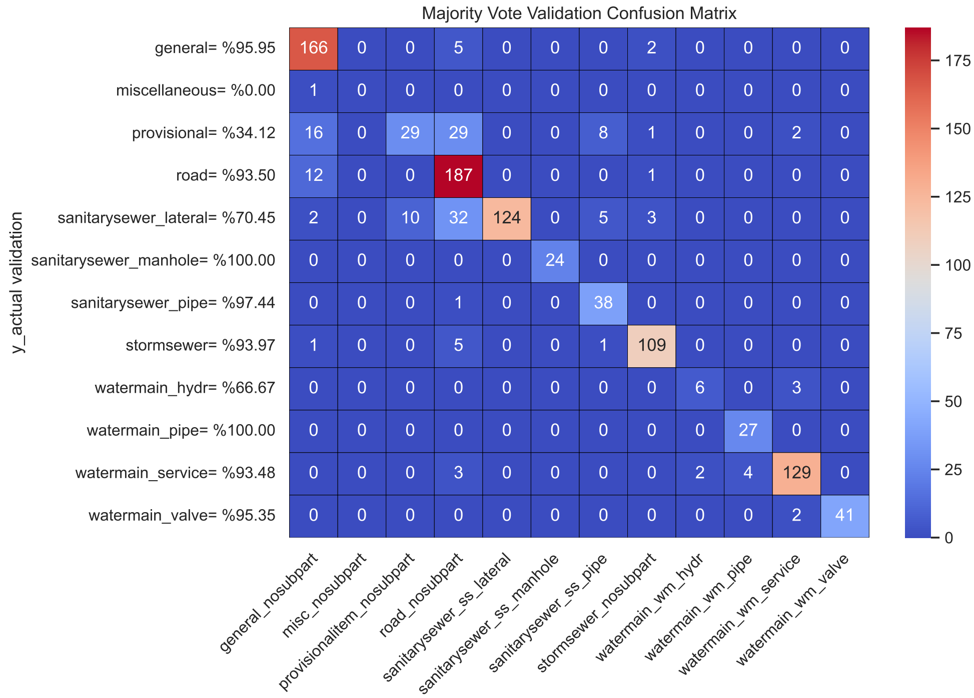



Figure 3.6: Confusion matrix from the classification of testing records of sample tenders using Random Forest only.

overcome these challenges and to achieve an acceptable level of accuracy. This precision level was vital for the deep learning model to discern the intricate patterns and relationships within the dataset. The decision tree algorithm initiated the automated process, paving the way for the LSTM deep learning model. It acted as a critical bridge, transitioning the dataset from an unclassified state to a structured format conducive to supervised learning. It resulted in a deep learning model adept at extracting valuable insights from an extensive and complex dataset.

While the decision tree algorithm made progress, it had significant limitations. One crucial challenge was compatibility issues with the bag-of-words representation of data, which ignores word order and treats input text as a collection of individual words. This led to the loss of sequential information, a crucial aspect in the context of item descriptions. Even after carefully selecting the most relevant words through a genetic algorithm, the accuracy of the decision tree classifier as shown on Figure 3.6 averaged 85.96%. This performance underscored that the traditional approach was promising but did not satisfy the application's stringent accuracy requirements when using the classified data for computing unit costs.

In response, the focus shifted toward deep learning methods, specifically the Long Short-Term Memory (LSTM) model. Known for its efficacy in natural language processing tasks, the LSTM model yielded promising results. A critical insight gained from this transition was the advantage of word-to-vector representation in enhancing classification accuracy. Building on this finding, the LSTM model was implemented.

#### LSTM vs. Generative Pretrained Transformers

The decision to employ Long Short-Term Memory (LSTM) models over Generative Pretrained Transformers (GPT) was informed by several considerations, both practical and theoretical:

- **Computational Efficiency**: LSTMs are generally more computationally efficient than large-scale transformers like GPT. Training GPT models, especially the larger variants, requires significant computational resources, which might not be readily available or cost-effective for every research project [Vaswani et al., 2017].
- **Dataset Size**: Transformers thrive on large datasets, especially models like GPT. Given the limited size of the dataset in this research (10-20 contracts), an LSTM was deemed more appropriate. Overfitting can concern transformers when data is limited [Wang et al., 2019].
- **Interpretability**: LSTMs provide greater interpretability due to their more straightforward structure than transformers. This is crucial in academic settings where understanding the model's decisions and being able to explain them is as important as the accuracy of the model itself.
- **Task Specificity**: While GPT models are designed to be generalists and perform a wide range of tasks, LSTMs can be tailored more specifically to a particular task. The specificity of the classification task in this research did not necessitate the broad capabilities of GPT.
- **Training Time**: Training an LSTM, especially on a smaller dataset, can be faster than training a large transformer model. This is crucial for iterative experimentation and rapid prototyping.
- **Memory Footprint**: LSTMs have a smaller memory footprint compared to large transformer models. This is advantageous when there are constraints in terms of available RAM or GPU memory.

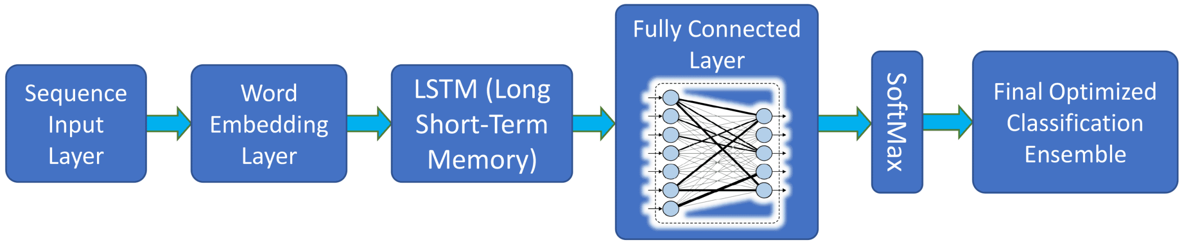



Figure 3.7: The block diagram of the implemented solution.

- **Maturity and Stability**: LSTMs have been around for a longer time compared to transformer models like GPT. They have a proven track record, and their behaviour is well-understood in the deep-learning community.
- **Customization**: LSTMs can be more easily customized or adapted to specific requirements. With GPT or other transformer models, making changes can be more complex due to the intricacies of the model architecture.

In addition to the above reasons, it is worth noting that the current LSTM model can serve as a foundation for future work. While decision trees were used as a stepping stone for the LSTM mechanism and generated ground truth data, the current LSTM model can similarly be used to generate input for the GPT engine in future endeavours.

## 3.3 Deep Learning Model

This section details the development and configuration of the Deep Learning (DL) artificial neural network model, showcased in Figure 3.7. The function of each component within this architecture is elaborated in subsequent subsections. To aid in understanding the DL model's process, a practical example is provided (Figure 3.8). This example demonstrates how an item, initially without standard-part and standard-sub-part identifiers (as mentioned in Table 3.2 on Page 82), is transformed into a numerical array suitable for DL classification (see Figure 3.9). This step in the methodology represents a crucial phase in developing high-accuracy classifiers for the DL model. A significant aspect of this phase is the iterative improvement of the dataset, which allows for more effective training of the DL model.

The running example examines an item from Table 3.2 on Page 82, which lacks standard-part and standard-sub-part identifiers. The transformation process begins by leveraging its description, as denoted in *Step A* of Figures 3.8 and 3.9. The item's section, mislabelled as "Roads", is correctly assigned to "Watermain" based on the ontology's patterns and descriptions. Consequently, the section's name should be adjusted to "Road", the item's unit supplanted by "Each", and the unit price annotated with a dollar sign "$" and a decimal point.

Subsequently, the item undergoes filtration, utilizing the ontology's filters and rules. This step aims to preserve consistency and maintainability, as displayed in *Step B* of Figures 3.8 and 3.9. The filters and rules eradicate superfluous words while the Natural Language Processing (NLP) library morphs complex words into their root form.

The ensuing transformations are:

- "hydrants": a plural form, is converted to singular,
- "with": a non-informative preposition, as per previous classification system training, is removed,
- "150mm" is parsed to two identifiable words: "150" and "mm",
- "boxes" and "anodes": both in plural form, are converted into singular,
- "and": coordinating conjunction lacking additional information is removed,
- "(provisional)": including unacceptable punctuation characters, is removed. Furthermore, "provisional", derived from the root "provide", is reduced to its first five characters, "provi".

The final description aligns with the ontology's requirements. For the running example, the resultant description is portrayed in *Step B* of Figure 3.9 as: "hydrant complete anchor tee 150 mm diameter valve box anode opsd 1105.010 provi".

After *Step B*, the item meets all ontology constraints and is subsequently conserved in the central, standardized dataset. However, the item is still devoid of the standard-part and standard sub-part. This step is where the DL model is crucial. As it emulates the expert's manual classification approach learned during the training phase, it effectively predicts these missing values based on the processed item descriptions.

To address this, the DL model includes the original category from the tender document in the revised description. The item's unit is also included as an additional input word to the description sentence. The DL model necessitates the conversion of the item's description

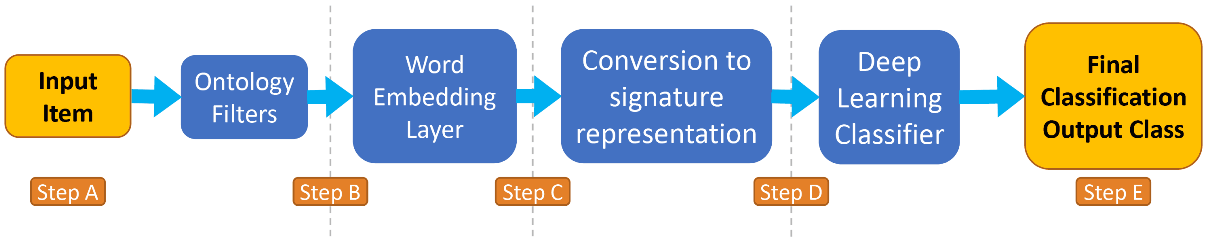



Figure 3.8: Block diagram of a record's transformation to determine its standard-part and standard-sub-part.

sentence into a numerical array. This transformation is facilitated by the Word Embedding Block (WEB), as indicated in *Step C* of Figures 3.8 and 3.9. The WEB purges words exclusive to the contract, city, time, or contractor, courtesy of the ontology's lexicon. The lexicon, fashioned by assessing over four hundred contracts and 90,000 words, comprises 2019 unique words. The ensuing descriptions for each item range between 60 and 350 words. The word count follows a Poisson distribution. Thus, a fixed encoding sequence length of two hundred words is sufficient to capture most of the information in each description. For sentences exceeding this limit, the surplus words are eliminated. Trials and observations have determined that a maximum of 200 words optimally preserves the majority of information since only 0.1% of the items necessitate the omission of words beyond the 200-word limit.

The resulting numerical array is presented in *Step C* of Figure 3.9, wherein each number signifies distinct word in the library. For instance, the words "anode" and "hydrant" correspond to numbers 37 and 24, respectively. These numbers are randomly assigned during initialization but remain immutable after that. The "word2vec" algorithm, elucidated by Goldberg et al. in [Goldberg and Levy, 2014], is then employed to convert each unique number into a corresponding vector. This conversion is denoted as *Step D* in Figures 3.8 and 3.9. The algorithm allows vectors to symbolize different words while encapsulating their similarities and differences. Figure 3.11 showcases vectors for five exemplar words: "valve", "hydrant", "excavate", "manhole", and "lateral". The first two words pertain to the Watermain standard-part and thus exhibit a high correlation in their vector representations, influencing the DL state similarly. Analogously, "manhole" and "lateral", related to the sanitary sewer standard-part, exhibit analogous behaviour. Conversely, "excavate", not directly associated with either Watermain or sanitary sewer standard parts, yields a significantly disparate vector representation. Figure 3.10 reveals correlation values between vectors of the sample words, corroborating these observations, and indicates a negligible

Step A sample item in original format

| Reference ID | Contract ID | Item Description | Quantity | Unit | Unit Price | Section | City |
|---|---|---|---|---|---|---|---|
| B3.b.2 | REDACTED | hydrants complete with anchor tee 150mm diameter valve boxes and anodes according to opsd 1105.010 (provisional) | 1 | ea. | 556 | Roads | REDACTED |

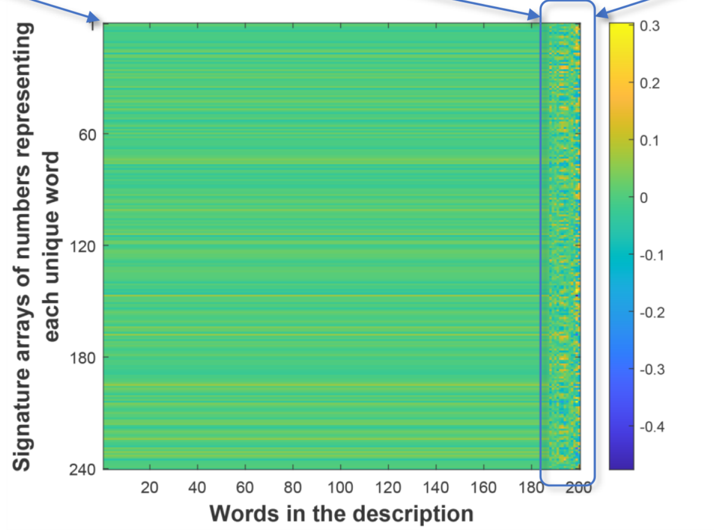
Step B sample item after passing ontology filters

| Reference ID | Contract ID | Modified Item Description | Quantity | Unit | Unit Price | Section | City | Standard Part | Standard Sub-Part | Original Contract Row # |
|---|---|---|---|---|---|---|---|---|---|---|
| B3.b.2 | REDACTED | hydrant complete anchor tee 150 mm diameter valve box anode opsd 1105.010 provi | 1 | Each | $ 556.00 | Road | REDACTED | | | 112 |

Step C representation of input item description after applying word embedding and zero padding

| 1st | 2nd | ... | ... | 187th | 188th | 189th | 190th | 191st | 192nd | 193rd | 194th | 195th | 196th | 197th | 198th | 199th | 200th |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | | | | provi | 1105.010 | opsd | anode | box | valve | diameter | mm | 150 | tee | anchor | complete | hydrant |
| 0 | 0 | 0 | 0 | 0 | 13 | 874 | 286 | 37 | 867 | 494 | 220 | 875 | 202 | 5 | 106 | 24 | 25 |

Step D




Step E sample item after passing ontology filters

| Reference ID | Contract ID | Modified Item Description | Quantity | Unit | Unit Price | Section | City | Standard Part | Standard Sub-Part | Raw Contract Row # |
|---|---|---|---|---|---|---|---|---|---|---|
| B3.b.2 | REDACTED | hydrant complete anchor tee 150 mm diameter valve box anode opsd 1105.010 provi | 1 | Each | $ 556.00 | Road | REDACTED | Watermain | Hydrant | 112 |

Figure 3.9: Visual representations of a sample record going through each transformation step to determine its standard-part and standard-sub-part.

correlation of "excavate" with terms related to the sanitary sewer or watermain standard-parts.

### 3.3.1 Step D2, Model Design

Constructing a robust deep-learning model requires careful decisions pertaining to architecture, input features, activation functions, and optimization methodology. For the DL model, the adopted architecture incorporates word embedding, LSTM, and Dense layers. Word Embedding is employed to convert words into dense vectors of fixed dimensions. The LSTM layer captures these vectors' temporal dependencies, making it suitable for handling sequences of words in our dataset. Lastly, dense layers are deployed for classification. Activation functions are integral for instilling non-linearity into the model, with ReLU and Softmax selected for the hidden and output layers, respectively. An analysis, in conjunction with iterative testing, was used to determine the optimal layer sizes and additional hyperparameters, thereby ensuring the model's robustness and performance.

### Long Short-Term Memory

As displayed in Figure 3.11, the vectors of size 240 are created for the 2019 most frequently occurring unique words in the current dataset. The figure represents the embedding dimension (width of WEB) and matches the number of LSTM blocks available in the DL model. Consequently, in this instance, the output from the embedding layer will be a matrix of 200 x 240 floating-point numbers ranging from -0.5 to 0.5. This output is designated as $X_t$ in Equations 3.1 on Page 103. As Figure 3.11 illustrates, some vertical vectors are highly correlated. The degree of correlation between two vectors directly corresponds to how their related words are associated within the context of the training data.

Figure 3.12 presents a block diagram of the unidirectional Long Short-Term Memory (LSTM) module used in this research. Each LSTM block accepts an input sequence, denoted by $X_t$, at every timestep of $t$. As the vector is serially fed into the LSTM module, each LSTM block receives a single value from the numbers array at every time instance.

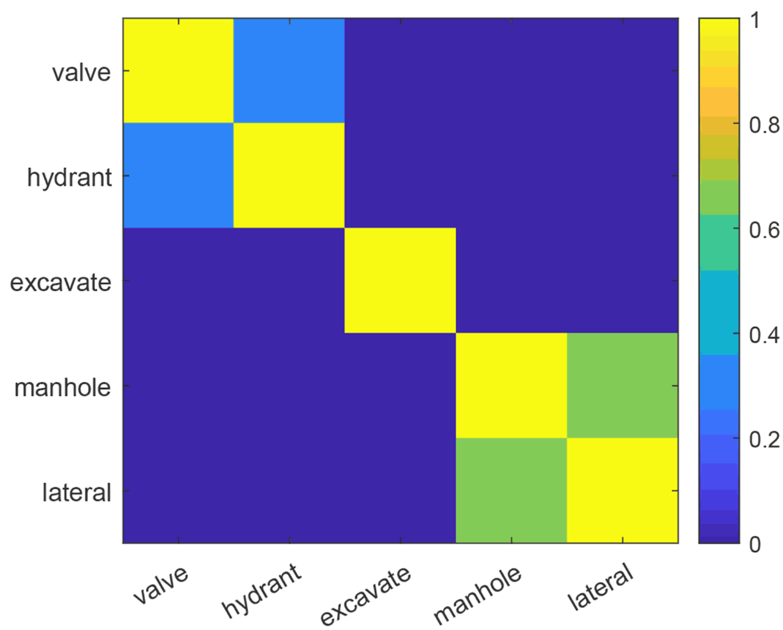



Figure 3.10: Comparison of correlation values among word encodings for different sample words.

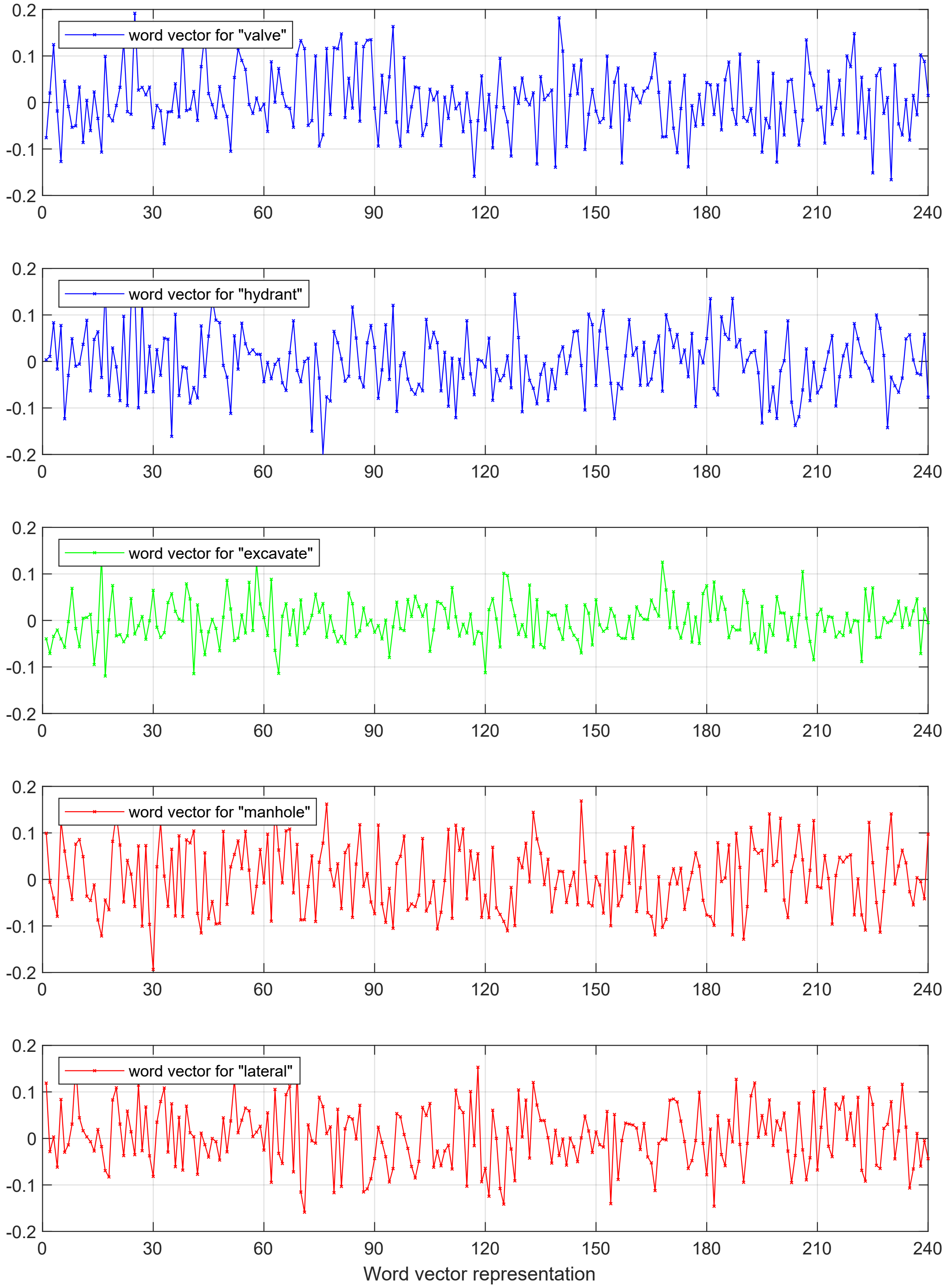



Figure 3.11: Visual representation illustrating how two correlated words ("root" and "fertilize") are encoded with highly correlated vectors and how a random word ("park") is encoded with significant variation.

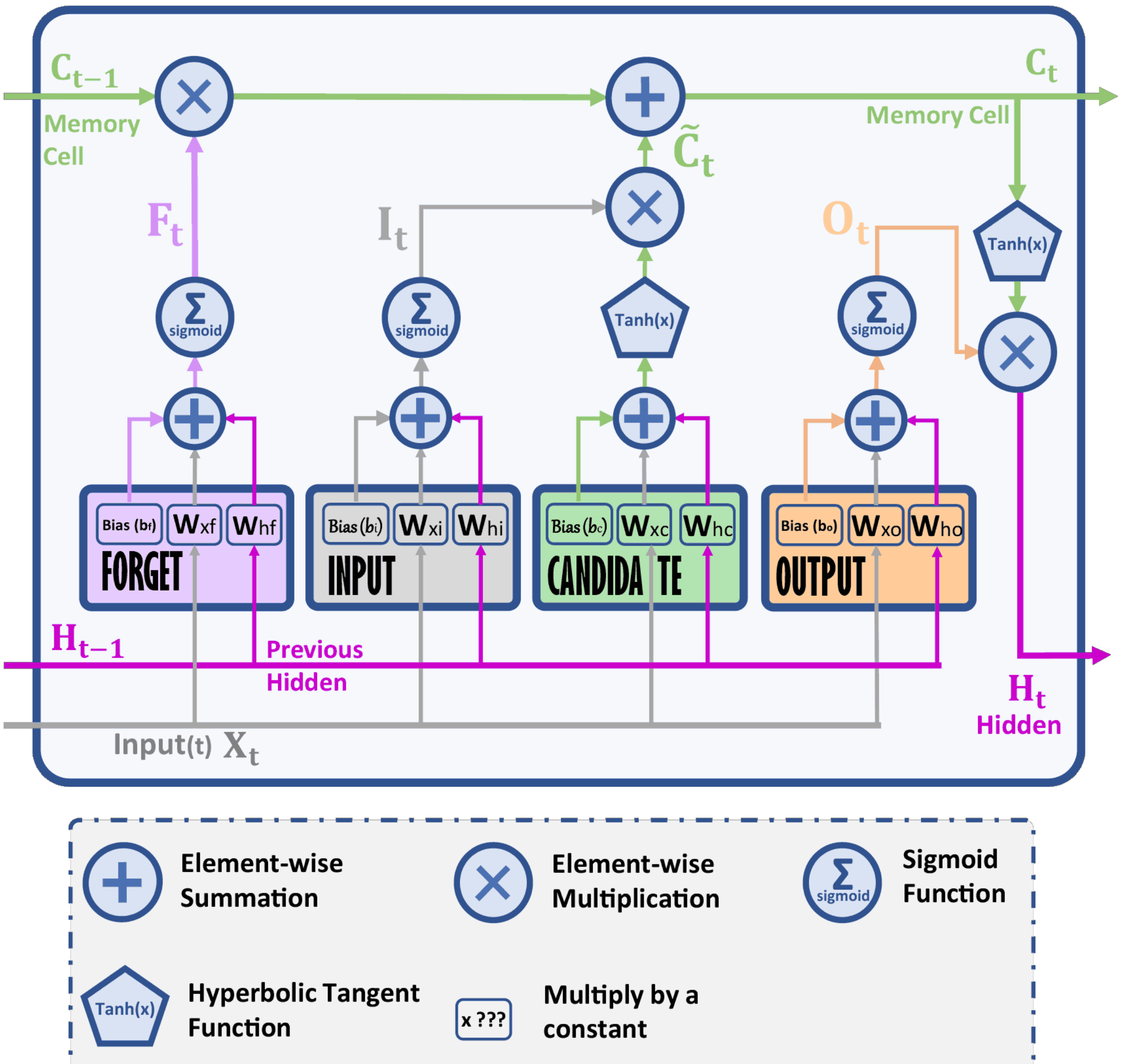



Figure 3.12: The graphical depiction of data flow within a single LSTM unit.

### Deep Learning Output Classes

The implementation and formulation of the deep learning model are based on the model presented by Van Houdt et al. [Van Houdt et al., 2020]. The LSTM model employs the activation function $activF() = Tanh()$ for state information updating and the $\sigma() = sigmoid()$ function for gate information updating. Each block's weights are represented by $W_{x*}$, recurrent weights by $W_{h*}$, and bias by $b$". In this context, $t$ indicates a variable's current value, while $t-1$ denotes the variable's previous value. The following equations for the input, forget, output, hidden, and candidate blocks at timestep $t$ calculate the output, which serves as the input for the subsequent stage of deep learning:

$$
\begin{aligned}
I_t &= \sigma(\mathbf{X_t}\ \mathbf{W_{xi}} + \mathbf{H_{t-1}}\mathbf{W_{hi}} + \mathbf{b_i}), && (3.1)\\
F_t &= \sigma(\mathbf{X_t}\ \mathbf{W_{xf}} + \mathbf{H_{t-1}}\mathbf{W_{hf}} + \mathbf{b_f}), && (3.2)\\
O_t &= \sigma(\mathbf{X_t}\ \mathbf{W_{xo}} + \mathbf{H_{t-1}}\mathbf{W_{ho}} + \mathbf{b_o}), && (3.3)\\
\tilde{C}_t &= tanh(\mathbf{X_t}\ \mathbf{W_{xc}} + \mathbf{H_{t-1}}\mathbf{W_{hc}} + \mathbf{b_c}), && (3.4)\\
C_t &= \mathbf{F_t}\ \odot\ \mathbf{C_{t-1}} + \mathbf{I_t}\ \odot \tilde{\mathbf{C_t}}, && (3.5)\\
H_t &= \mathbf{O_t} \odot \mathbf{tanh}(\mathbf{C_t}) && (3.6)
\end{aligned}
$$

The standard parts utilized for each input record's classification are defined as follows:

$$
\begin{aligned}
Y_v \in \quad & \{General,\ Miscellaneous,\ ProvisionalItem, && (3.7)\\
& Road,\ SanitarySewer,\ StormSewer,\ Watermain\}.
\end{aligned}
$$

Each item is organized into the standard contract's standard-part corresponding to overarching attributes of the watermain and sanitary sewer capital works project. If an item is classified as "Watermain" or "SanitarySewer", it requires further standard-sub-part classification. For Watermain and SanitarySewer items, the standard sub-parts are defined as:

$$
\begin{aligned}
Y_{v_{watermain}} &\in \{WM\_Pipe, WM\_Valve, WM\_Hydrant, WM\_Service\} && (3.8)\\
Y_{v_{sanitarysewer}} &\in \{SS\_Pipe, SS\_Lateral, SS\_Manhole\} && (3.9)
\end{aligned}
$$

At this juncture, the formal definition and application-specific deep learning structures with the LSTM Layer are delineated.

### Parameters and Hyper-Parameters of the Deep Learning Model

The LSTM network architecture begins with a sequence input layer of size one, accommodating numeric sequences. It is succeeded by a word embedding layer with an embedding dimension of 240, which maps words into vectors in a 240-dimensional space. The architecture consists of an LSTM layer with two*embeddingDimension (480) hidden units, allowing us to learn the dependencies between sequences. The LSTM layer's output is channelled into a fully connected layer featuring several nodes equal to the classes in our data. It is succeeded by a Softmax layer that assigns probabilities for each category and, finally, a classification layer that selects the class with the maximum probability.

For the LSTM model training, the Adam optimization algorithm is used. The optimal minibatch size, determined through trial and error, is between 52 and 98. A training gradient threshold of 2 was chosen to prevent gradient explosion, a common issue in training RNNs, and the training data are shuffled every epoch to enhance the model's robustness and generalizability [Bengio et al., 1994].

The model is trained using a holdout validation strategy, reserving 20% of the data for validation purposes. This ensures our model's performance is not overestimated and can be generalized effectively to unseen data. Regarding computational settings, the model uses a parallel execution environment for training, accelerating the process relative to training on a CPU.

All these considerations and decisions contribute to the robustness and effectiveness of the proposed LSTM model in tackling the text classification task. By fine-tuning these hyperparameters and architectural decisions, we developed a model that demonstrates strong performance in the given text classification task.

## 3.3.2 Step D3, Progressive Improvement of Training Data

This step involves a systematic, iterative refinement process for the training data, which is crucial for the effective learning and classification accuracy of the Deep Learning (DL) model. Key to this process is the detailed examination and correction of the dataset to ensure its precision and reliability. This iterative approach, widely recognized in machine learning, addresses label noise-a frequent challenge in dataset preparation [Karimi et al., 2020].

During each DL training session, a thorough analysis of the training results is undertaken to assess and enhance the model's robustness. Particular attention is paid to instances of misclassification, identifying, and correcting any inaccuracies introduced during the learning phase. This aspect is similar to the refinement done in the Random Forest training (Step C4), where the DL model, too, learns to identify patterns in the training data, including pseudo false positives and negatives.

The training process undergoes multiple cycles of refinement until reaching a state where no further improvements are discernable, ultimately yielding a dataset with minimal errors. This intensive process involved up to 100 iterations to reach the desired level of data stability and model precision.

In line with neural network characteristics, the LSTM network is initialized with random values, introducing an element of variability. To ensure consistent and dependable performance, the LSTM model undergoes multiple training sessions with different initializations. This method guarantees that the model's effectiveness is not coincidental but replicable across various training scenarios.

This refinement process embodies a mutual learning dynamic. Initially, the DL model (Step D2) learns from the data. Subsequently, the engineering expert evaluates the model's output (Step D3), leading to data updates based on these insights, and the cycle repeats (returning to Step D2). This continuous feedback mechanism ceases once the DL Classifier's errors no longer contribute to identifying new errors in the training and validation data. This method of progressive data cleaning aligns with approaches documented in existing literature [Gao et al., 2018, Khaki, 2021, Karimi et al., 2020].

## 3.4 Results

The principal aim of devising and training the Deep Learning (DL) artificial neural network model revolves around leveraging the classification output as essential additional data fields (standard part and sub-part) within tender-bid items. These enriched items can be used in financial and engineering estimates for prospective projects. The precision of the standard-part and sub-part assignment thus emerges as a crucial aspect. To comprehend how classes may intermingle during classification, the confusion matrix is deployed, with results encapsulated in Figure 3.13.

All experiments were executed on a high-performance computer outfitted with an Intel (R) Xeon E-2186G CPU operating at 3.80 GHz, 128 GB DDR4 RAM, a NVidia GeForce RTX 3060 Ti GPU, and NVME 2TB storage with a 3000 MBps read and write capability. The preponderance of the project coding was carried out in Python, complemented by marginal instances scripted in Visual Basic and Python. Barring explicit indication, the models were implemented using the available Python toolboxes.

### 3.4.1 LSTM-Based Deep Learning Results

Recent studies have highlighted the potential advantages of bidirectional LSTM in unveiling hidden relationships and dependencies among specific words more effectively [Siami-Namini et al., 2019]. However, in the context of this project, employing this LSTM structure did not result in a notable improvement in classification accuracy. Table 3.4 compares classification accuracy for identifying each standard-part and standard-sub-part using BiLSTM and unidirectional LSTM. The data indicates no statistically significant advantage in utilizing a BiLSTM over a conventional, single-directional LSTM model. In certain cases (e.g., Watermain Hydrant), the simpler LSTM model outperformed the BiLSTM.

| **Standard Part** | **Standard Sub-Part** | **Item Count** | **Testing Accuracy BiLSTM** | **Testing Accuracy LSTM** |
|---|---|---|---|---|
| *Sanitary Sewer* | *Lateral* | 176 | 98.3% | 96.6% |
| *Sanitary Sewer* | *Pipe* | 39 | 97.4% | 84.6% |
| *Sanitary Sewer* | *Manhole* | 24 | 87.5% | 100.0% |
| *Watermain* | *Pipe* | 27 | 88.9% | 96.3% |
| *Watermain* | *Valve* | 43 | 88.4% | 95.4% |
| *Watermain* | *Service* | 138 | 93.5% | 93.5% |
| *Watermain* | *Hydrant* | 9 | 77.8% | 100.0% |
| *Provisional Item* | *No Sub-Part* | 85 | 90.6% | 80.0% |
| *General* | *No Sub-Part* | 173 | 97.7% | 97.11% |
| *Miscellaneous* | *No Sub-Part* | 1 | 100.0% | 100.0% |
| *Road* | *No Sub-Part* | 200 | 94.5% | 93.0% |
| *Storm Sewer* | *No Sub-Part* | 116 | 88.8% | 91.38% |

Table 3.4: Classification results of validation items after training both the LSTM and BiLSTM-based DL Classifiers.

### 3.4.2 Model Performance Metrics

The unit cost approach employed in this study for estimating RMSE measures of unit costs compared to engineer estimates (Root Mean Square Error) and R-squared values is derived from the methodology adopted by my predecessor in the group, Rehan et al. This approach is thoroughly documented in their work, specifically in [Rehan et al., 2016], within Table 1, titled "Cost allocation procedure for the pipe component of watermain and sanitary sewer projects," which details the standard components of watermain and sanitary sewer projects.

The efficacy of the DL model is quantified through its performance metrics, which include RMSE and R-squared values, as detailed in Table 3.5. The RMSE values serve as the standard deviation of the residuals (prediction errors), where lower values denote a more accurate model. The DL model demonstrates strong performance with RMSE values ranging from 0.041 to 0.096. These low RMSE values indicate the model's robustness, with the most precise predictions observed for "watermain unit cost" (0.041 RMSE) and the least precise for "watermain unit hydrants" (0.096 RMSE).

R-squared, or the coefficient of determination, denotes the proportion of variance for a dependent variable that an independent variable or variables in a regression model can explain. High R-squared values suggest a good fit of the model to the data. The observed R-squared values range from 0.907 to 0.995, which are notably high and indicate that the model explains a substantial fraction of the variance in the data.

The model demonstrates a robust predictive capacity for "watermain unit cost" and "watermain unit pipes" with an R-squared value of 0.995. Conversely, its predictive capacity is slightly lower for "sanitary sewers unit pipes" with an R-squared of 0.907. Nonetheless, an R-squared of 0.907 is still considered satisfactory, indicating that the model captures a considerable portion of the variance in the data.

In essence, the DL model performs admirably in cost estimation. It has a strong fit for most of the data, particularly for "watermain unit cost" and "watermain unit pipes" regarding RMSE and R-squared. Though the performance is slightly weaker in predicting "sanitary sewers unit pipes" and "watermain unit valves," these predictions are still within acceptable limits.

### 3.4.3 Confusion Matrix Analysis

The confusion matrix for the LSTM model is presented in Figure 3.13, which shows the high accuracy achieved in mapping all standard-part and standard-sub-part items. Most

| Standard Part / SubPart | Watermain Unit Cost | Watermain Unit Pipes | Watermain Unit Valves | Watermain Unit Hydrants | Sanitary Sewers Unit Cost | Sanitary Sewers Unit Pipes | Sanitary Sewers Unit Manholes |
|---|---|---|---|---|---|---|---|
| **RMSE** | 21.536 | 8.964 | 28.397 | 0.000 | 11.541 | 5.630 | 0.000 |
| **R-Squared** | 0.9779 | 0.9957 | 0.9997 | 1.000 | 0.9986 | 0.9999 | 1.000 |

Table 3.5: Comparison of unit cost values regarding RMSE, and r-squared correlation computed for various Standard-Parts, contrasting the ground truth (engineer estimate) and the DL model classification outcomes.

false negatives and false positives are seen between standard-sub-parts of a standard-part, indicat

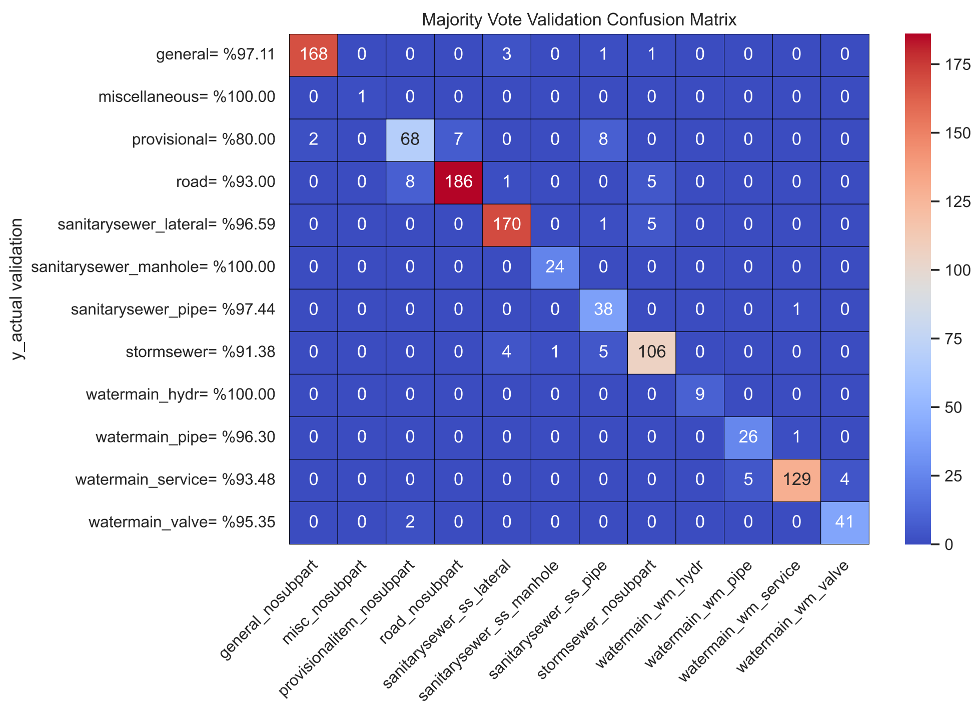



Figure 3.13: Confusion matrix from the classification of testing records of sample tenders using deep learning (LSTM only).

The RF model's ease of use and lower sensitivity to data errors made it an excellent initial tool. At the same time, the LSTM's capability to uncover complex patterns provides a deeper level of analysis, albeit with higher computational costs. The iterative improvements in the data quality and RF hyperparameters could potentially yield a model with performance comparable to that of the LSTM.

The comparison of unit cost values calculated using the ground truth classification (validated by expert classifications) against those derived from the model's classifications is illustrated in Figure 3.14 on Page 109 and Figure 3.15 on Page 112. These comparisons,

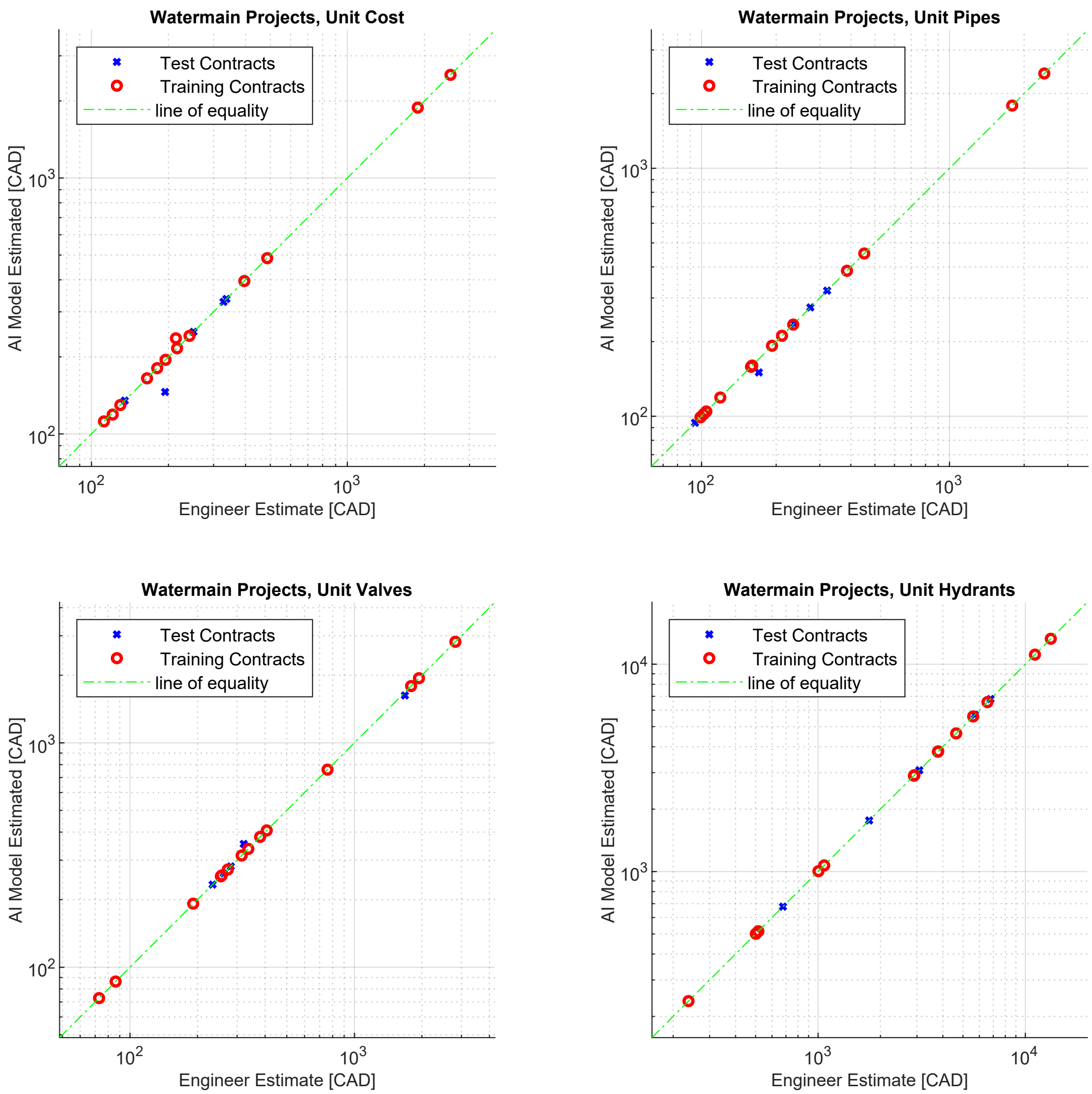



Figure 3.14: Comparison of unit cost values computed for different aspects of the watermain, contrasting the ground truth (actual costs) and the DLANN model classification outcomes.

Table 3.6 presents the classification results for both RF and LSTM models, showing the accuracy for each class and the number of items. The table complements the confusion matrices by providing a numerical representation of the classification performance, which offers a comprehensive view of the models' capabilities compared to the figures.

| Class | Accuracy (%) | | # Items |
|---|---|---|---|
| | Random Forest | LSTM Deep Learning | |
| general_nosubpart | 95.95% | 97.11% | 168 |
| misc_nosubpart | 0.00% | 100.00% | 1 |
| provisionalitem_nosubpart | 34.12% | 80.00% | 68 |
| road_nosubpart | 93.50% | 93.00% | 186 |
| sanitarysewer_ss_lateral | 70.45% | 96.59% | 170 |
| sanitarysewer_ss_manhole | 100.00% | 100.00% | 24 |
| sanitarysewer_ss_pipe | 97.44% | 97.44% | 38 |
| stormsewer_nosubpart | 93.97% | 91.38% | 106 |
| watermain_wm_hydr | 66.67% | 100.00% | 9 |
| watermain_wm_pipe | 100.00% | 96.30% | 26 |
| watermain_wm_service | 93.48% | 93.48% | 129 |
| watermain_wm_valve | 95.35% | 95.35% | 41 |

Table 3.6: Classification Results for Random Forest and LSTM Deep Learning

## 3.5 Conclusion

In this chapter, we developed and refined an automated system for classifying historical tender-bid documents in watermain and sanitary sewer capital works projects using Artificial Intelligence (AI) and Machine Learning (ML) techniques. The primary challenges addressed were the lack of standardized data sources and the variability in unit cost estimates due to individual engineers' preferences. These challenges were magnified when considering multiple municipalities across different scales.

The methodology employed involved multiple stages, beginning with the import and preliminary cleaning of data from over 250 documents across three Canadian cities. This data was then prepared for a 5-fold cross-validation process. The initial classification utilized a minimum distance method with expert intervention. However, the approach was shifted to feature set generation using the Bag of Words (BoW) model due to its limitations. Subsequently, a Random Forest (RF) classifier was trained, demonstrating improved accuracy but uneven performance across classes. To refine this, a genetic algorithm was integrated for feature selection optimization, enhancing the RF classifier's performance.

The next phase involved iterative refinement of the training data, which was done over ten iterations with expert collaboration. This refinement led to a significant improvement in the model's accuracy. Recognizing the limitations of the RF model in handling complex datasets, the methodology transitioned to deep learning.

The development of the Deep Learning (DL) model, specifically a Long Short-Term Memory (LSTM) network, marked a significant advancement. The DL model was chosen for its computational efficiency, suitability for sequential data, and customization potential. Over 100 iterations, the model's precision was further refined, surpassing the initial accuracy target of 92%. The final iteration of the model, selected for its optimal classification accuracy, was evaluated using test data.

The practical application of the DL model is emphasized in automating the classification of historical tender-bid items for unit cost calculation. An important consideration in this process is mitigating overfitting risks, achieved through expert-validated classification systems and a semi-automatic process involving human labeling.

The DL model's architecture is carefully designed to include components for transforming item descriptions into a format suitable for classification. This transformation process is vital for the DL model to emulate the expert's manual classification approach, allowing the accurate prediction of missing standard-part and sub-part values based on processed item descriptions.

In conclusion, the LSTM-based DL model demonstrated strong performance in the classification of tender-bid items, validated through rigorous testing and progressive refinement. The model's precision in standard-part and sub-part assignments is crucial for providing accurate financial and engineering estimates for infrastructure projects. This automated classification system represents a significant advancement in the field of AI and ML applied to civil engineering, offering a solution to the challenges of standardizing data and reducing variability in unit cost estimates across multiple municipalities.

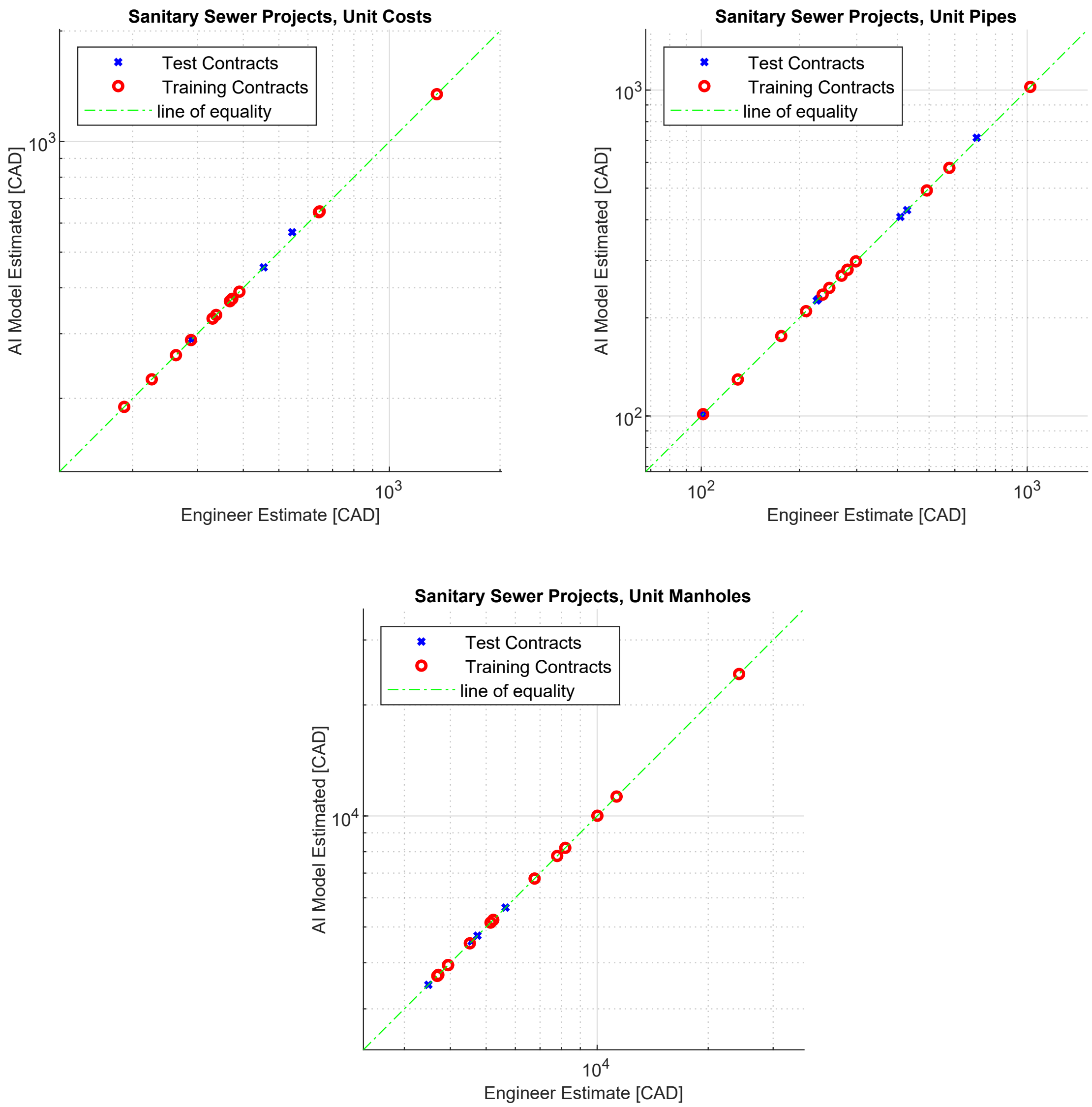



Figure 3.15: Comparison of unit cost values computed for various aspects of the sanitary sewer, contrasting the ground truth (actual costs) and the DL model classification outcomes. The bottom left plot illustrates watermain and sanitary sewer project unit cost indices.

# Chapter 4

# Application of the AI Model

The current chapter delineates the results of implementing the proposed methodology within the context of this research project. This implementation is organized into three main sections. The first section discusses the specific challenges and solutions concerning implementing the proposed methodology. The components to be addressed in this section include (I) natural language processing and the utilization of ontology, (II) the deep learning classification system for standardizing the imported data, and (III) the implementation of unit cost optimization. The subsequent section focuses on the results derived from applying the methodology to tenders available from the three cities described previously. Particular emphasis is placed on the numerical precision of the data and the used methods, ensuring the accurate implementation of the methodology.

The final section of the chapter unveils the outcomes of designing a user-friendly interface to cater to the end-user targets of the proposed method. The interface is crafted to be simple, intuitive, and responsive to user requirements. The most fitting solution, in this case, involved transitioning from Matlab/Python to Visual C# (read as "C Sharp") to achieve a streamlined project implementation. The proof-of-concept webserver and its implementation results are presented later in this chapter.

## 4.1 Implementation

The main code of this project is authored using the Matlab and Python scripting languages. While both languages do not compile code and run scripts and functions, these tools have been chosen for their experimental nature, enabling the utilization of various toolboxes and facilitating visualization. Nonetheless, the code is formulated in alignment with object-oriented programming principles, ensuring that migration to C++ or C# requires minimal effort.

The implementation is conducted in layers, facilitating the separation of architectural layers or components of the code. This layered approach enhances the understanding and execution of the code. The layers encompass:

- Layer A, imported raw table (tender)
- Layer B and C, Raw Table Cleanup and Ontology-Updated Processing
- Layer D, Deep Learning Input Preparation
- Layer E, Standardized Table with Classification Results
- Layer F, Normalized Table Ready for UCI Calculation
- Layer G, Results of Unit Cost Optimization
- Layer H, visualization, performance enhancement, and evaluation table

### 4.1.1 Layer A, imported raw table

This layer accepts input in Excel format, accommodating variations in style and formatting across cities and contractors. The arrangement of fields is not fixed, and the process relies on specific restrictions to validate the input table. If the following required fields are absent, the program will return an error and terminate:

- *"WaterIAM Contract", "WaterIAM Description", "WaterIAM Section", "WaterIAM Unit", "WaterIAM Quantity", "WaterIAM Unit Price", "WaterIAM City"*. The interpretation of these fields is largely self-evident, but certain entries that necessitate additional explanation are elaborated upon below:
  - *"WaterIAM Contract"*, is the name of the contract or tender is similar for all items in one tender document.
  - *"WaterIAM Description"*, a description of an item is usually limited to the description field of the item. However, in some cases, several items have one main description and short additional descriptions individually ('i.e. item 1 description

= "trenchless pipes", item 2 description = "200 mm", item 3 description = "300 mm" → item 1 WaterIAM Description = "trenchless pipes", item 2 WaterIAM Description = "trenchless pipes 200 mm", item 3 WaterIAM Description = "trenchless pipes 300 mm".

  - *WaterIAM Section*, the section that item is presented in, for example, items that come after a line indicating "Watermain Section" belong to the "watermain" section. The input file does not need to adhere to a standard set of descriptions. The rules in the ontology will update the section name to its corresponding standard one.
  - *WaterIAM Unit*, the unit items are limited to: "ea." (for "each" unit), "Hour", "L.S." (for lump sum unit), "m" (for linear meter), "m2" (for area in square meter), "m3" (for volume in cubic meter), and "Tonnes" (for mass in 1000 kilograms or tonnes).

    Any other item with a non-compatible unit should be converted to one acceptable in the above list. (i.e. "ft" (for linear length in feet) should be converted to "m", and the quantity and unit price should be updated accordingly).

- The optional acceptable fields include: "WaterIAM Item Number", "Org Description", "Org Sheet Name", "WaterIAM PSP Override", "WaterIAM Org Section", "WaterIAM Standard Part", "WaterIAM Standard Sub Part", "WaterIAM Total Price", "WaterIAM Multiple", "WaterIAM Depth". The absence of these fields will not hinder the execution, and further descriptions are provided below:

  - *"WaterIAM Item Number"*, a field that indicates the item number in the original tender document. Its value can be a number for the order and does not have a significant meaning, or in some cases (i.e. City A), the item number would indicate the section and details of the nature of the item.
  - *"WaterIAM PSP Override"*, as mentioned in Chapter 3, the output of the deep learning classification module would determine the standard-PSP. However, if the operator overrides this classification, it is possible to define the desired standard-PSP in this column.
  - *"WaterIAM Org Section"*, this column merely defines the original categorization of the tender item. This field provides essential information for the DLC and is recommended to be included in the input table if possible.
  - *"WaterIAM Standard Part"*, is a field that identifies a predetermined standard-part for each item. The value of the standard-part can result from the previous

classification mechanism or the feedback from an expert on particular items with acceptable standard-part. Note that although this field defines the standard part for the item, it does not override the classification of the DLC.

- *"WaterIAM Standard Sub Part"*, is a field that identifies a predetermined standard-sub-part for each item. The value of standard-sub-part can result from the previous classification mechanism or the feedback from an expert on the particular items with acceptable standard-sub-part. Note that although this field defines the standard part for the item, it does not override the classification of the DLC.

- *"WaterIAM Multiple"* this field is an essential step for the DLC training process. Some items belong to a particular standard-part, and the value of the original section should not affect them. However, due to the lack of diversity in data, the classifier defines a relationship between the original section and the standard-part. The "WaterIAM Multiple" field allows the operator to generate several similar pseudo items with the same standard-part and random original section to destroy this unwanted relationship effectively.

  An example of this situation is when the tender item description is: "concrete any curb piece private repair type Concrete storm sewers CSA A257 with Class 'B' bedding and Granular 'A' cover and backfill". In this case, the original section is identified as "StormSewer", and the standard-part is also recognized as "StormSewer". However, regardless of the original section, based on the description (and the fact that Storm Sewers is specifically mentioned), the item belongs to the "StormSewer" standard-part. Therefore, assigning the value of 10 to 20 to the "WaterIAM Multiple" field allows the system to rectify this confusion.

- *"WaterIAM Depth* is a field whose value is meaningful for a limited set of items (i.e. SanitarySewer_SS_Manhole). If the natural language processing module fails to identify the depth of the item, it is possible to override the value using this field specifically.

Layer A yields a raw table standardizing the inputs procured from the user, setting the stage for further analysis and the subsequent processing embodied in Layer B.

### 4.1.2 Layer B and C: Raw Table Cleanup and Ont-Updated Processing

In this combined layer, the raw table received from the previous stage is prepared through an extensive cleaning process with the aid of natural language processing (NLP), style adjustments, and formatting. This cleaned table is then refined to align with ontology requirements and generate the classes map table and functions for subsequent stages. The details of this combined layer are presented below.

- *Removing the Item Number*: (Description as in Layer B).
- *Updating Spelling According to the Ontology Lookup Table*: (Description as in Layer B, including handling of typos, backward entries, etc.).
- *Removal of Specific Words and Characters*: The function filters out predefined words and characters such as brackets, hyphens, etc. It includes removing quotations, filtering out characters in the `Ont_RemChars` array, and further custom handling as detailed in Layer C.
- *Lemmatization of Words*: Lemmatization is performed to reduce words to their root form. It helps unify terms with similar concepts, aiding natural language understanding and consistency with ontology. (Details from both Layers B and C).
- *Update of Words and Characters*: (Description as in Layer C).
- *Special Handling of Numbers and Symbols*: Specific conditions related to numbers and symbols are handled. (Details as in Layer C).
- *Word Splitting and Standardizing Spacing*: The text is split into individual words, and spacing is standardized. It includes conversions like "copper-pipe␣25mm␣␣diameter" to "copper␣pipe␣25␣mm␣dia". (Details from both Layers B and C).
- *Frequency Table Update*: (Description as in Layer C).
- *Final Formatting*: Any extra spaces between words are reduced to single spaces, ensuring consistent formatting throughout the table.

The resulting output of this combined layer is a refined table that has undergone initial cleanup and ontology-updated processing, ready for further analysis and manipulation. The table is saved in `PLLR.L03_OntolTbl.table`, and robust error handling is maintained throughout the process.

### 4.1.3 Layer D, Deep Learning Input Preparation

The fourth layer of the WTM system, implemented in the function `WTM_Layer_04_DeepLearning_PrepareInputs__v5p0`, is responsible for preparing the data for deep learning processing. It involves a series of steps:

- *Initialization and Warning Suppression*: The code starts by initializing an empty array for the organized data and turns off warning notifications to avoid unnecessary alerts during the process.
- *Record Importing and Validation*: The function iteratively processes each record in the ontology table from the previous layer. If any mandatory field like `Contract`, `UnitPrice`, or `Quantity` is missing, the record is skipped.
- *Calculation of Total Price*: If the `TotalPrice` field is missing or empty, it is computed by multiplying the `UnitPrice` and `Quantity` fields.
- *Record Structure Formation*: A new structure is formed for each record, populating fields like `Description`, `Unit`, `FinalPrice`, `StandardPart`, `StandardSubPart`, etc., with necessary transformations. These transformations include class-type extraction and specific string manipulations.
- *Handling Missing Fields*: The code handles various scenarios where fields might be missing or unclassified, assigning default values such as "Unclassified" or -1 where appropriate.
- *Standardization of Parts and Subparts*: The code ensures that standard parts and subparts are categorized correctly and named uniformly. It includes specific replacements for certain terms and removing records that don't meet the criteria.
- *Warnings and Classification Validation*: A warning is issued if the part categories do not match the expected number, and the code turns warning notifications back on after processing.
- *Final Data Structuring*: The code finalizes the standard parts and subparts, converting them into categorical variables and associating them with the organized data.
- *Data Assignment*: The final organized data table is assigned to `PLLR.L04_DLCTable.table` for further processing.

The output of Layer D is a structured table that has been prepared and cleaned specifically for deep learning applications, with all necessary fields appropriately transformed and categorized. This layer ensures that the data is in a suitable format to be utilized effectively in subsequent machine learning or deep learning tasks.

### 4.1.4 Layer E, Standardized Table with Classification Results

In this layer, the deep learning model undertakes the classification task for a given dataset, considering standardized parts and subparts within the system. The functionality of this layer can be outlined as follows:

- **Preparation:** The layer begins by ordering the fields of the deep learning classification table. If a classification table already exists, an error is raised to avoid overwriting.
- **Initialization:** Variables for tracking predictions, correct and wrong count, included contracts, and other necessary parameters are initialized.
- **Iterative Analysis:** The main body of the function iterates through the current data, performing deep-learning classification on each record. Process updates are printed to the console if verbose mode is enabled.
- **Issue Handling:** A specific check is conducted for issues regarding unsupported subparts, and appropriate warning messages are displayed if such an issue is found.
- **Prediction:** Depending on whether classification override is enabled, the function either uses a trained LSTM model to predict the subpart or employs the override value from the data. If the prediction is incorrect, related details are printed, the extent of which depends on the verbosity level.
- **Accuracy Tracking:** The function keeps track of correct and wrong predictions, updating the counters accordingly.
- **Post-Processing:** The predicted values are standardized into a string format, and the final classification results are stored in the `OutPredictions` structure.
- **Error Handling:** Any exceptions within the layer are caught and rethrown, allowing for appropriate error handling at a higher level.

The output of Layer E comprises a standardized table with classification results, encapsulating both predictions and reference data. It contributes to a more accurate and comprehensive understanding of the overall system by providing deep learning-based insights into the given data, thus offering a vital step in the data processing pipeline.

### 4.1.5 Layer F, Normalized Table Ready for UCI Calculation

Layer F of the process focuses on preparing a normalized table to calculate the Unit Cost Index (UCI). This layer carefully analyzes the contracts and corresponding dates, performs unit cost calculations, and organizes the results into a structured table. The primary functionalities of Layer F are described below:

- **Initialization:** All necessary variables, such as contracts, contract dates, and unit cost structure, are initialized. Existing data is cleared if verbose mode is enabled.
- **Contract Analysis:** The function iterates through each contract, identifying the corresponding date and ensuring it exists within the parameters.
- **Unit Cost Calculation:** For each contract, the layer invokes a Unit Cost Calculator that computes the actual unit costs for specific categories such as project costs, pipe costs, and other related fields.
- **Statistical Analysis:** Various statistical parameters, such as the minimum, maximum, mean, and product of unit costs, are analyzed, and the results are printed to the console.
- **Data Normalization:** The processed data is normalized and organized into a structured format, creating a table that includes fields like contracts, cities, dates, and different types of unit costs.
- **Box Plot Considerations:** The code includes provisions for handling box plot data if required, although this functionality appears to be reserved for future use.
- **Error Handling:** Proper error handling is implemented to catch any exceptions during the execution, ensuring that issues are promptly identified and addressed.
- **Finalization:** The normalized data is stored in the `L06_UCIStrct` structure, preparing it for subsequent UCI calculations.

The efforts in this layer to calculate and normalize the unit costs according to various criteria represent a vital step in understanding the financial dimensions of the system. The meticulous handling of contracts and the corresponding unit cost calculations demonstrate a robust approach to preparing the data for UCI calculation, a critical component in the overall analysis.

### 4.1.6 Layer G: Results of Unit Cost Optimization

Layer G focuses on optimizing unit costs, implementing the unit cost inflation model, and quantifying the results of this optimization using various parameters. This optimization step is crucial for reducing cost uncertainties and enhancing the predictability of cost estimates.

- **Initialization of Genetic Algorithm Parameters**: The optimization process involves setting up parameters for a Genetic Algorithm (GA). The parameters include the number of variables, step size, and lower and upper bounds. The GA operates within these defined limits to search for optimal solutions.

- **Analysis Type Determination**: For optimization, the method uses Geometric Brownian Motion (GBM), a popular stochastic process used in various financial and engineering applications. GBM is favoured for its mathematical tractability and the ability to model various real-world phenomena. Here, the GBM helps find the optimized unit cost calculation parameters.

  A key aspect of Layer G is calculating unit costs based on the selected method; currently, the code only supports the "Geometric Mean Value" method, but there is scope for other types of analyses to be incorporated.

- **Sorting Input Data**: All data from Layer F, sorted by date, is input into Layer G. Then, each Optimization Detail is iteratively processed. The optimization process displays iterations if the 'Verbose' option is enabled.

- **Optimization Loop**: The primary function in Layer G is 'local_OptimizationFunction,' which serves as the objective function for the GA. It takes an input vector, representing the GBM parameters for optimization, and returns the cost residuals. It performs various calculations related to GBM, such as calculating the log returns (SofT), the expected log return (EofT), and the variance of log returns (VarOfS), and it also computes the Z1 and Z2 scores, excluding outliers based on specified limits.

- **Cost Analysis**: The unit cost analysis uses the parameters obtained from the local optimization function. The analysis involves computing the geometric mean values and applying exponential functions to obtain the final unit cost.

- **Residual Evaluation**: A comprehensive residual evaluation is carried out to ensure that the optimization converges to a feasible solution. Conditions and limits are applied to control the residual.

- **Results and Summary**: The optimization results are stored in a struct, which includes all details related to the optimization, unit cost summary, and other analytical parameters.

This layer presents an optimization strategy involving various mathematical and statistical operations, targeting unit cost optimization. By employing both GBM and GA, the optimization aims to be suitable for diverse scenarios. However, further testing and validation might be necessary to confirm its effectiveness across all possible applications. The detailed recording and structuring of the results make Layer G a critical component within the model, contributing to the comprehension of unit cost dynamics without overextending its capabilities.

### 4.1.7 Layer H, Visualization, Improved Performance, and Eval. Table

Layer H in the system is devoted to visualizing unit costs and various other components related to water treatment management components. Below are the key functionalities and processes of this layer:

- **Definition of Analysis Types:**
  - Two analysis types are defined, focusing on specific components and statistical measurements.
  - A Comprehensive list of analysis types is formulated for detailed inspection.

- **Organization of Data:**

- Data is organized into structures based on date and attributes, facilitating easy access and manipulation.
- Specific curve data structure is formed to store box plot data.

- **City Categorization:**
  - Cities are categorized into three regions, ensuring data is appropriately classified.
  - Unknown city data triggers an error message, maintaining the integrity of the categorization.
- **Data Collection:**
  - Iterates through existing and new data sets to populate corresponding arrays.
  - Aligns data with respective city categories and marker types for analysis.
- **Visualization of Data:**
  - Utilizes error bar graphs to represent the data visually.
  - Different colours and markers distinguish regions and data sets.
  - Visualizes the mean and variation of unit costs for comprehensive insight.
- **Storage of Curve Data:**
  - Stores the curve data within the structure for subsequent utilization.
  - Encapsulates essential insights for further analysis or reporting.
- **Summary:**
  - Layer H serves as a robust tool for understanding cost dynamics within water treatment management.
  - Integrates statistical analysis with visualization, assisting in informed decision-making.

## 4.2 Results

This section elucidates the results derived from analyzing the records of 277 contracts obtained from the three reference cities presented in the thesis. These results, depicted in Figure 4.1 on Page 127 to Figure 4.7 on Page 130, provide an in-depth understanding of the various facets of the contracts and tenders. A detailed breakdown of the data reveals that of the total contracts, 221 tenders belong to City A, 52 belong to City B, and 10 tenders belong to City C. The distribution reflects the magnitude and scale of operations in these cities and offers an insight into the spatial dynamics of the projects.

The approach adopted for calculating the unit cost index for each type of project closely follows the methodology previously employed by Rehan et al. [Rehan et al., 2016]. This methodological alignment ensures continuity with past research and supports a robust comparative analysis. The sections below elaborate on the plots' specifics to represent the analyzed data.

In the respective plots, each data point is denoted with distinct markers and colours to represent the cities: star and green for City A, circle and red for City B, and square and blue for City C. These markers illustrate the geometric mean value of all individually scaled contract items for each project type within a given scaled contract to unit project. Accompanying each data point is a vertical line matching the marker's colour, indicating the range of item values. The upper bound of this line, marked with a small horizontal line, signifies the maximum value of the scaled item found for that specific contract. In contrast, the lower bound represents the minimum value.

This graphical representation offers a comprehensive view of the data, allowing for immediate visual differentiation between cities and a clear understanding of the range and central tendency of contract values over time. Figures 4.1, 4.2, and 4.3 focus on sanitary sewerage projects, including manholes, pipes, and overall unit projects, respectively. Similarly, Figures 4.4, 4.5, 4.6, and 4.7 pertain to watermain projects, encompassing hydrants, pipes, valves, and overall unit projects. These visual representations enable a nuanced understanding of the variability and trends in contract values across different cities and project types, thereby enhancing the analysis of spatial and temporal patterns in urban infrastructure development.

### 4.2.1 Analysis of Contract Value Variability

Regarding differences in the price variations presented in the figures and reasons behind the varying sizes of error bars, our analysis reveals several key factors influencing these disparities. The range of error bars, representing the upper and lower bounds of scaled item values for each contract, varies notably across different cities and years. This variation is primarily attributed to the impact of inflation, which is evident in the plots where the leftmost projects, representing earlier years, show a smaller range of items compared to the rightmost items, indicative of more recent years.

Additionally, the number of items in a contract significantly influences the range of prices. Contracts with a larger number of items tend to exhibit a broader range of scaled-item values, resulting in longer vertical range lines in the plots. Furthermore, the pricing strategy employed for different items within a contract can lead to increased variability. For instance, contractors may allocate higher costs to material procurement and lower costs to service items to receive higher payments in advance. This discrepancy in pricing contributes to greater variability in the vertical range lines of the corresponding contracts.

In contrast to City A's procurement approach, where the municipal authority outlines contract specifications and invites competitive bidding based on a predefined template, City B adopts a contractor-driven bid proposal model. In this model, contractors assume the role of primary engineers in formulating the bid, encompassing the detailed proposal of contractual elements, material specifications, service deliverables, and provisional items. Although there are differences in procedure, our analytical assessment, grounded in unit price calculation methodology and item categorization aided by the AI model, reveals a remarkable parity in unit project costs between the two cities. This observation shows no significant markups or disparities in unit project cost calculations. It highlights the effectiveness and methodological robustness of our AI-driven classification and scaling algorithm in standardizing unit cost computations across heterogeneous tendering frameworks.

In the context of City C, a municipality characterized by dense urban development, our analysis shows a potential escalation in unit costs. The rise in unit cost values is attributable to the densely built environment and the proximity of complex infrastructural networks. These factors inherently amplify the logistical and material expenditure components, reflected in the escalated service and material costs associated with urban infrastructure projects.

However, it is essential to note that comparing different contracts solely based on the range they exhibit is not straightforward and may yield inaccurate insights. The vertical lines in the plots offer a more qualitative approach to understanding pricing variability in each contract item. A consistent range of geometric mean item prices, and to a lesser extent, their extreme points, are expected. Contracts showing inconsistent pricing should be scrutinized for potential errors or underlying issues causing this inconsistency.

The figures indicate that City A exhibits higher variation in sanitary sewer items compared to City B. However, this variation is less pronounced in watermain contracts. In the case of City C, due to the limited number of contracts and their recent nature, it is challenging to comment on their spread or make direct comparisons with the other cities. The inclusion of City C in the figures demonstrates the feasibility of calculating scaled contract values for cities with different characteristics and layouts. It also indicates that, despite the inability to draw definitive conclusions, the pattern of contract item ranges in City C appears to follow a similar trend to Cities A and B, especially when considering inflation and price increases over time.

These insights, drawn from the analysis of the figures, underscore the complexities involved in interpreting contract data across different urban settings. They highlight the importance of considering various factors, such as inflation, contract size, and pricing strategies, in understanding the variability and trends observed in urban infrastructure development contracts.

### 4.2.2 Comparison with Shapton's Findings

The results obtained in this study, particularly visible through Figures 4.1 to 4.7, can be contextualized within the framework of Shapton's work in [Shapton, 2017]. Shapton's analysis of City A provides an in-depth examination of the impact of government policy changes on contract prices within the infrastructure sector. Shapton scrutinizes the effects of the Infrastructure Stimulus Fund (ISF), which significantly increased project applications and approvals, especially in the water and wastewater infrastructure sector. This policy led to a disproportionate funding allocation towards these sectors, with Ontario receiving substantial federal and provincial support [Shapton, 2017].

Shapton further illustrates this impact by analyzing specific projects in City A, highlighting the reconstructions and infrastructure revitalization projects. Though categorized differently, these projects included substantial components of watermain, sanitary sewer, storm sewer, and road construction, aligning with the broader infrastructure development trends under the ISF [Shapton, 2017]. These changes in the

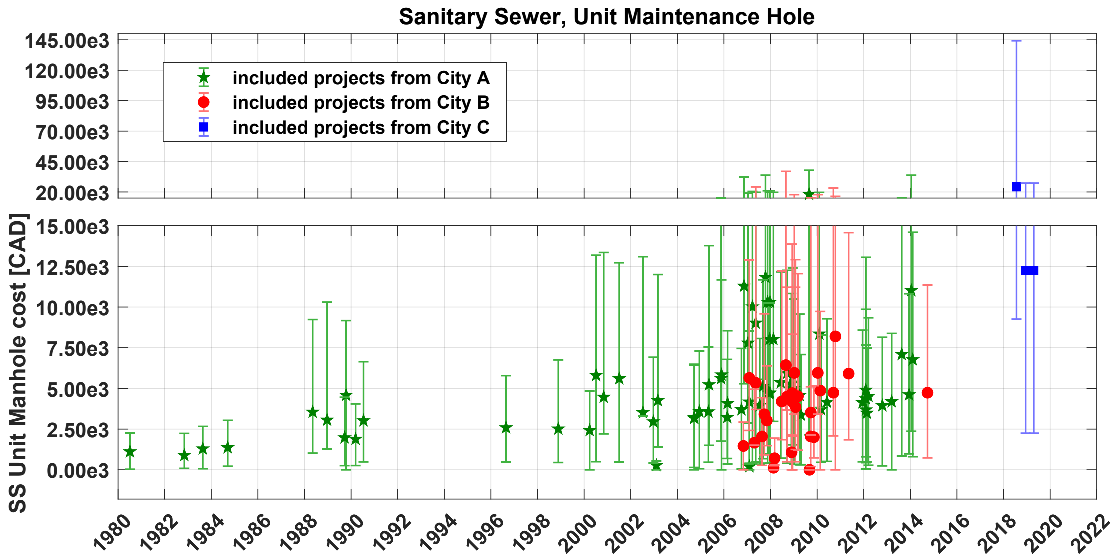



Figure 4.1: Plot of the unit cost index values for projects in the three cities, encompassing maintenance hole items. Each entry in the figure delineates the minimum, geometric mean, and maximum value, offering a comprehensive understanding of the cost dynamics.

number and unit costs of the projects can be seen in the current results in Figures 4.1, 4.2, 4.3, 4.4, and 4.7.

An important observation from Shapton's thesis is the fluctuation in the number of capital works projects and the corresponding unit prices post-ISF. Shapton notes a heightened number of projects during the ISF years (2009-2010) and a subsequent decrease in later years (2011, 2013, 2014). This trend mirrors the findings in this thesis, where although a direct trend analysis is not conducted, the subtleties of these changes are apparent in the provided figures as mentioned above. Shapton's work clearly indicates how ISF influenced project prioritization and funding in City A, leading to a temporary spike in infrastructure projects, followed by a reduction in subsequent years [Shapton, 2017].

Moreover, Shapton's analysis extends to the tender prices of these projects. They observe that the total tender prices of water and wastewater projects were higher during the ISF years compared to the post-ISF years. It is consistent with the observations in this thesis, where 2009 and 2010 show higher total tender prices for these projects, albeit less pronounced due to the need for explicit trend analysis in our figures. Thus, while this thesis provides a quantitative overview of contract prices in City A, City B, and partially City C, Shapton's detailed examination offers a more nuanced understanding of the specific impacts of governmental policy changes on these prices [Shapton, 2017].

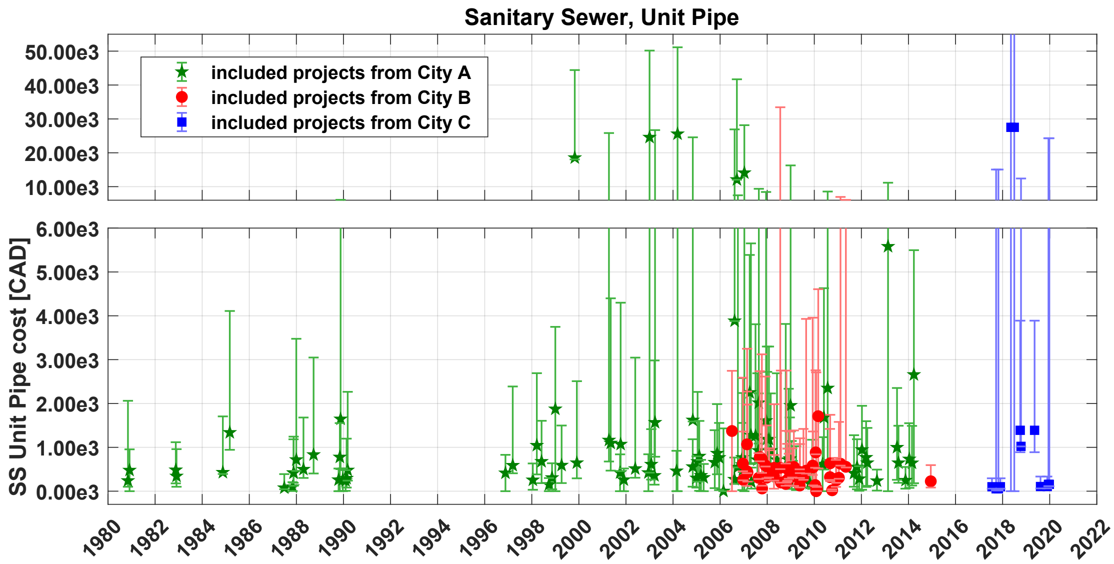



Figure 4.2: Plot of the unit cost index values for projects in the three cities, encompassing sanitary sewer pipe items. Each entry in the figure delineates the minimum, geometric mean, and maximum value, offering a comprehensive understanding of the cost dynamics.

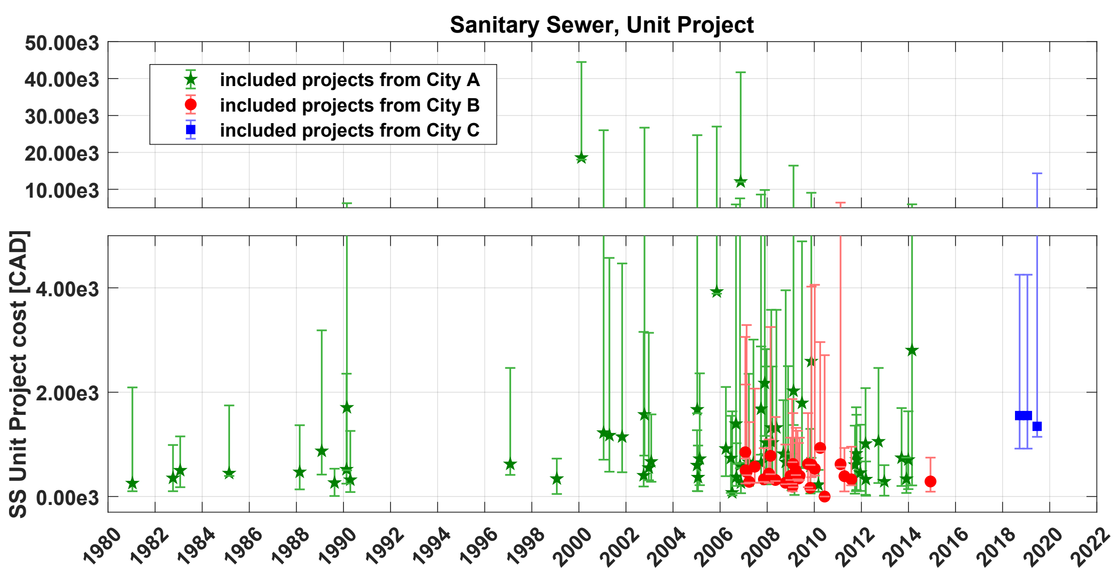



Figure 4.3: Plot of the sanitary sewer unit project values for tenders in the three cities, covering all standard sub-parts of the sanitary sewer standard part. Each entry signifies the minimum, geometric mean, and maximum values, encapsulating the unit project values variations.

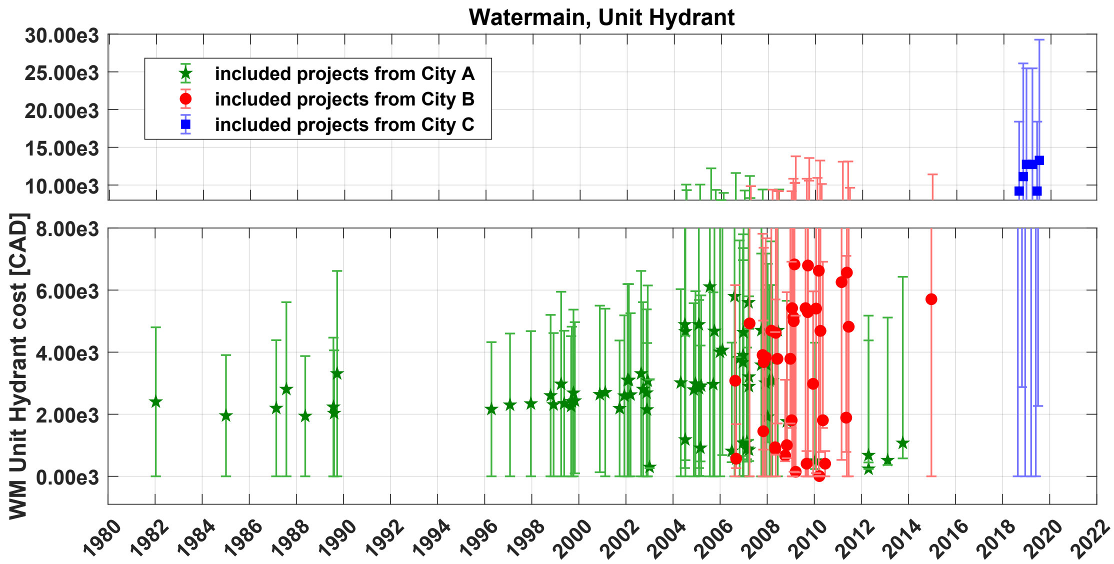



Figure 4.4: Plot of the unit cost index values for projects in the three cities, encompassing hydrant items. Each entry in the figure delineates the minimum, geometric mean, and maximum value, offering a comprehensive understanding of the cost dynamics.

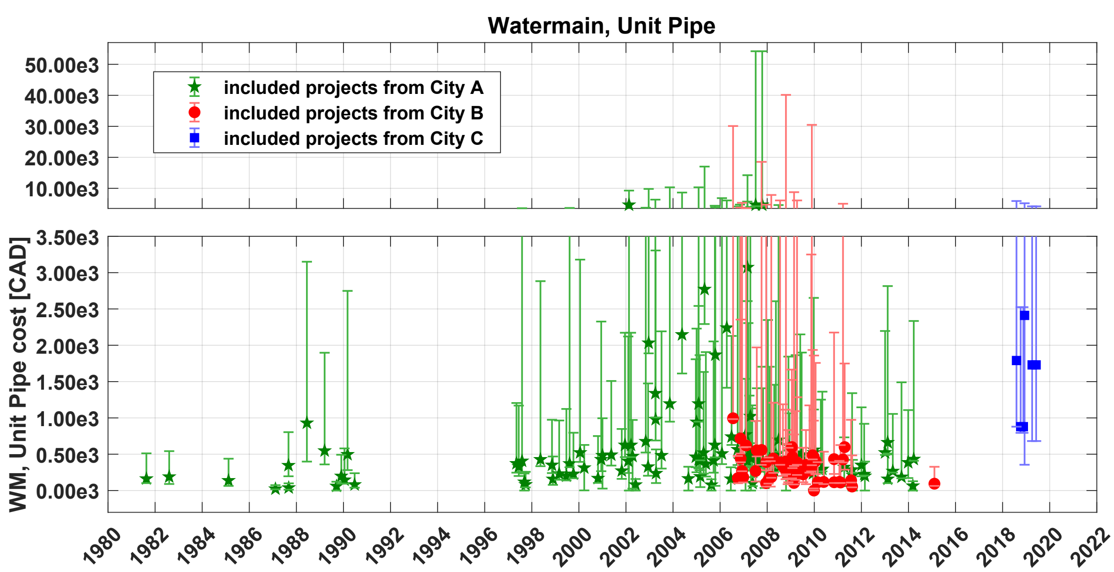



Figure 4.5: Plot of the unit cost index values for projects in the three cities, encompassing watermain pipe items. Each entry in the figure delineates the minimum, geometric mean, and maximum value, offering a comprehensive understanding of the cost dynamics.

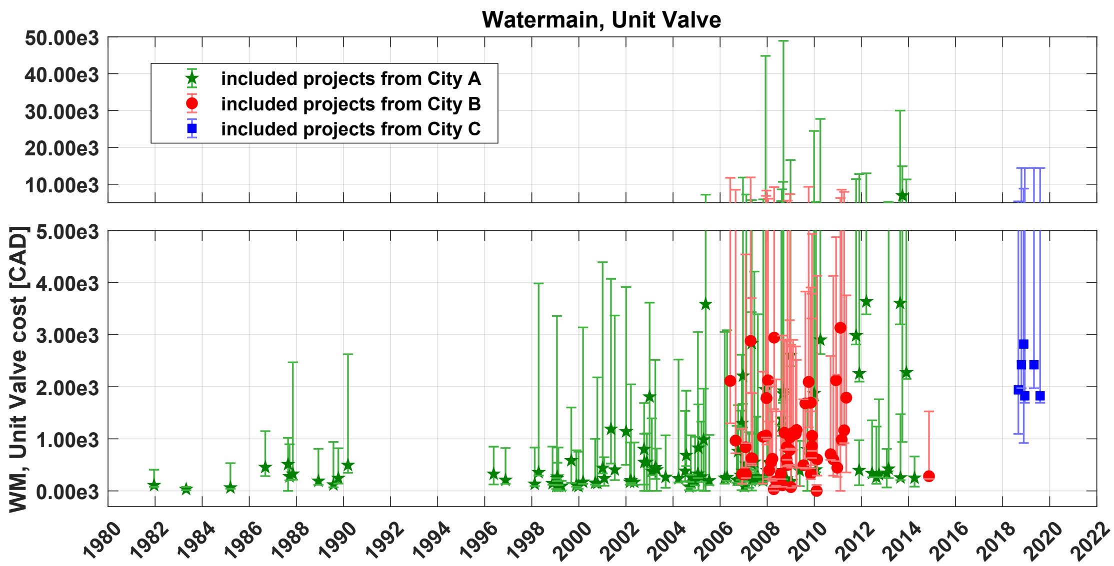



Figure 4.6: Plot of the unit cost index values for projects in the three cities, encompassing watermain valve items. Each entry in the figure delineates the minimum, geometric mean, and maximum value, offering a comprehensive understanding of the cost dynamics.

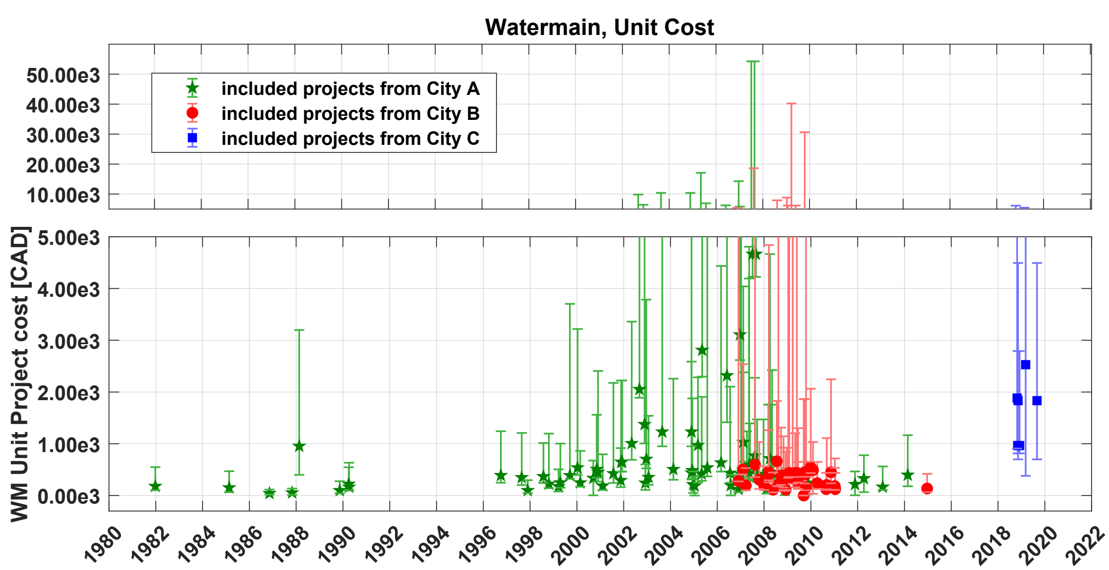



Figure 4.7: Plot of the watermain unit project values for tenders in the three cities, covering all standard sub-parts of the watermain standard part. Each entry signifies the minimum, geometric mean, and maximum values, encapsulating the unit project values variations.

## 4.3 Data Analysis Toolbox

The development of the Data Analysis Toolbox represents a significant step towards realizing the project's goals, demonstrating an innovative contribution to academia and industry. This section elucidates a web server interface's design, objectives, and implementation, encapsulating the methodologies and databases delineated in the preceding chapters. It offers an efficient Decision Support System that potentially meets the needs of industry stakeholders and municipalities.

### 4.3.1 Overview of the Interface

The prior version of this project, developed by Shapton et al., employed an offline application in Microsoft Access. Although functional, it was confined to limited and simplified capabilities. In contrast, the current project has evolved into an online, web-based application driven by the necessities detailed in the objectives section below. This transformation caters to a broader spectrum of real-time interactions, bridging the gap between scientific research and practical application.

### 4.3.2 Objectives of the Web Server Interface

The web server interface is designed to fulfill specific objectives catering to user requirements. These objectives are detailed as follows:

(A) **Real-Time Interaction:** The interface should be online to facilitate real-time interaction with tools and databases.

- Immediate access for operators to server information without additional processing.
- Reasonable processing time for filtered data or processed information proportionate to the computational load.

(B) **Data Import and Alignment:** The interface must import new contracts, ensuring alignment with predefined standards. Required order: (a) item number, (b) specification number, (c) item description, (d) unit, (e) unit price, (f) quantity, (g) total price. Extraneous fields are ignored.

(C) **Export Functionality:** Users can export the standardized tender and download it as a CSV file.

(D) **Analysis Capability:** Facilitates the analysis of unit cost index and inflation for specific standard-parts with acceptable delay parameters.

### 4.3.3 Implemented Features and Outcomes

The web server's functionality is multifaceted, emphasizing user interaction and data privacy. Key features and outcomes include:

- **User Responsiveness:** The server's architecture minimizes response time and caches repeated queries for immediate future access.
- **Contract Acceptance and Standardization:** The server accepts new contracts, processes them through a filtration and mapping approach, and stores them for future analysis. Downloadable standardized tenders are made available to users.
- **Analytical Toolbox:** The current revision focuses on basic unit cost index and inflation rate calculations. This implementation serves as a proof of concept, demonstrating that the platform can potentially extend to diverse financial analyses within the confines of standardized methodologies. It is important to note that the system is not designed to diagnose specific industry trends, such as past cost inflation or collusion among bidders, and any such claims would require further validation and domain expertise.

In summary, the Data Analysis Toolbox embodies a critical component of the project, transforming theoretical methodologies into practical tools. The interface's online nature enhances accessibility, ensuring the data's privacy is safeguarded and municipalities retain exclusive access to their respective information. It sets the stage for further enhancements, aligning with the scientific community's aspirations and the industry's pragmatic needs.

### 4.3.4 Methods

This section elucidates the various toolboxes implemented within the webserver, integral to the functioning and practical application of the system. Each toolbox is meticulously designed to execute specific tasks, contributing to the overall efficiency and responsiveness of the server. A brief description of each toolbox and illustrative examples are provided to explain the corresponding functionality and its integration within the larger webserver framework.

Figure 4.8 provides a comprehensive illustration of the client and server-side implementation of the software, demarcating the functional relationships and architectural design. This visual representation aids in understanding the collaborative dynamics between different components and how they interact to deliver the desired outcomes.

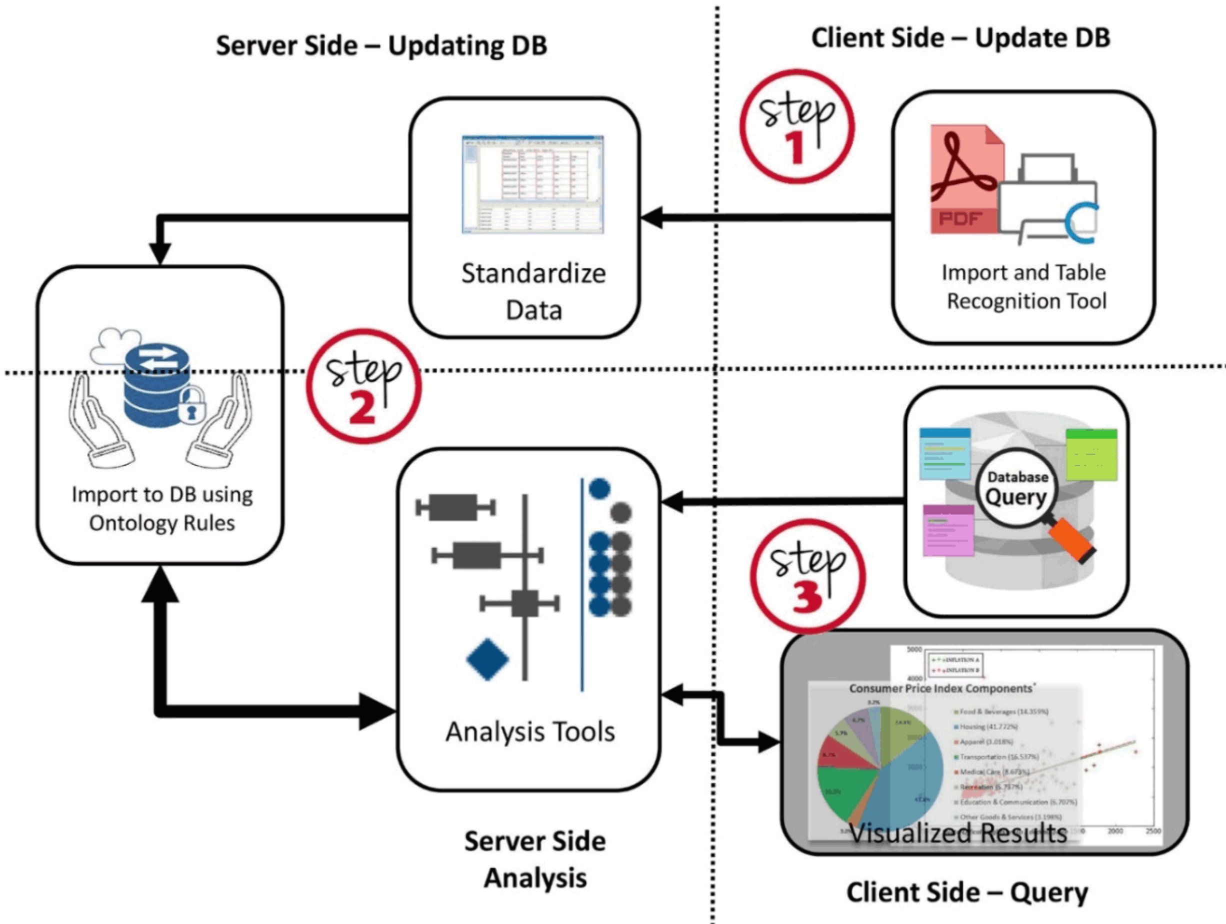



Figure 4.8: Illustration of the client and server-side implementation of the software.

The ensuing sections will delve into the details of the individual toolboxes, elucidating their design, operational principles, and contributions to the overall system's capabilities.

### Tender Summary Index

The Tender Summary Index is an essential component within the WaterIAM server, functioning as the primary interface for operators seeking to harness the platform's capabilities. Acting as the gateway to the system's various features, this toolbox allows users to manage contracts seamlessly, ensuring efficient integration with the core database.

As depicted in Figure 4.9, one of the principal functionalities of this toolbox is the option to import a new contract. Post-importation, operators gain the ability to download this specific contract, as well as other existing tenders, through the associated "Tender Query" toolbox. This feature allows for greater accessibility and control over the stored data, enhancing the user's interaction with the system.

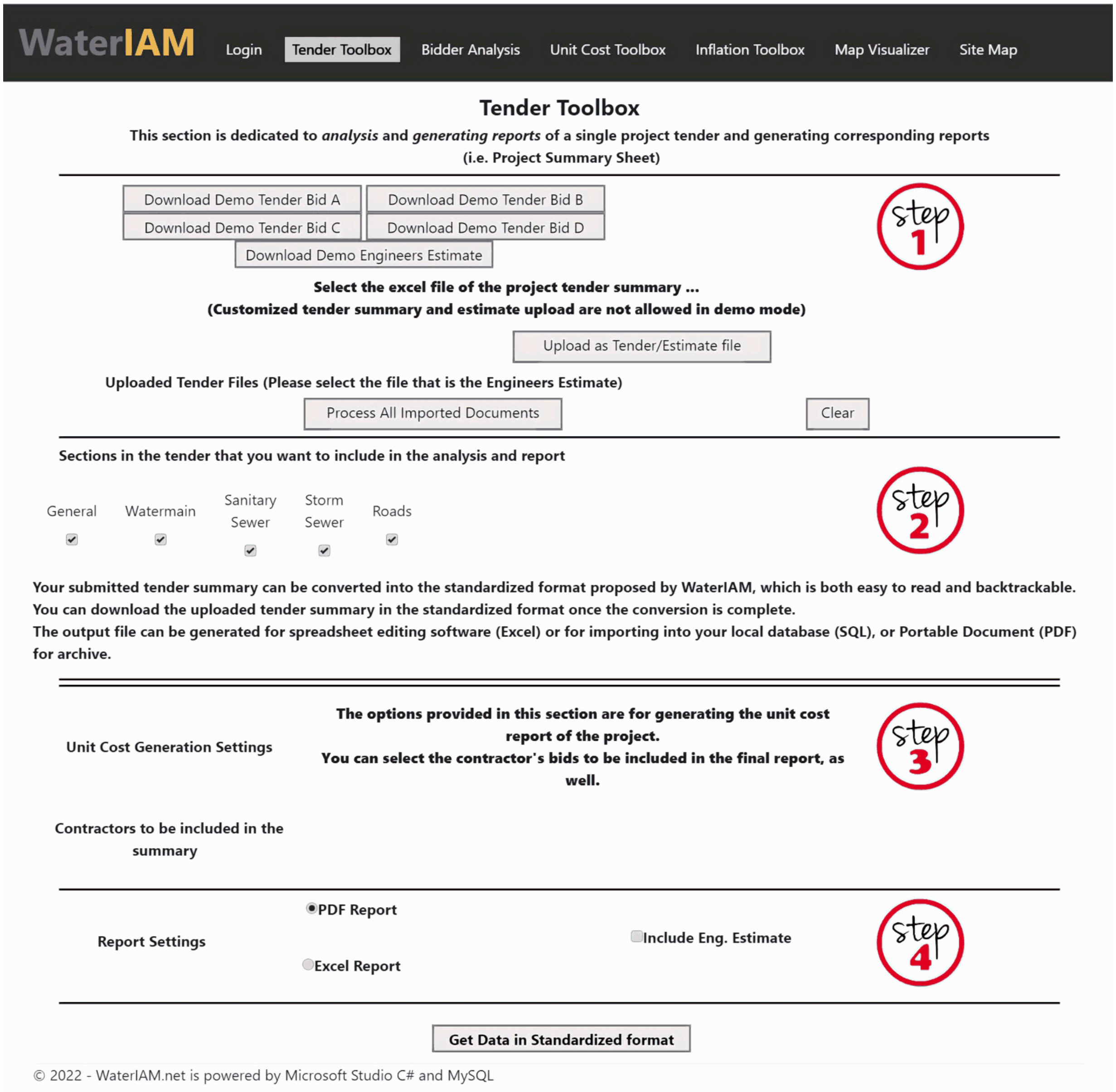



Figure 4.9: Sample webserver output while using the Tender Analysis toolbox.

The Tender Summary Index toolbox underscores the flexibility and user-centred design of the WaterIAM server, catering to the varied needs of the operators. Its intuitive layout and comprehensive functionality establish it as a pivotal aspect of the system's architecture, playing a vital role in navigating and exploiting the diverse features embedded within the platform.

### Bidder Analysis Toolbox

The Bidder Analysis Toolbox represents an advanced feature within the server's architecture, enabling operators to gain insight into the statistical information associated with bidders across various contracts. This aspect of the system focuses on providing a comprehensive view of the bidders' annual activities, detailing the total number of bids made and the subsequent contracts awarded, as illustrated in Figure 4.10.

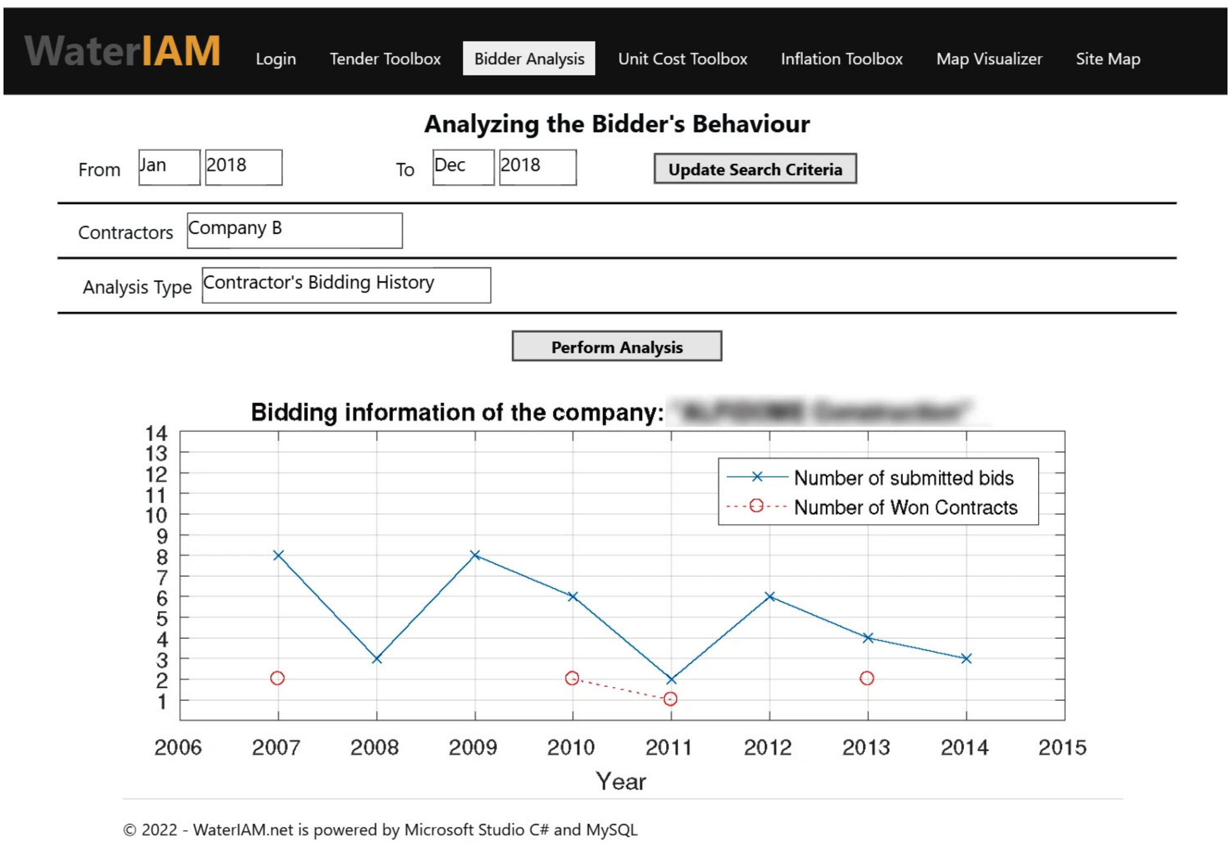



Figure 4.10: Sample webserver output while using the Bidder Analysis toolbox.

It is pertinent to note that the bidder information is not directly manageable through the web interface. Instead, the server relies on previously uploaded statistical profiles specific

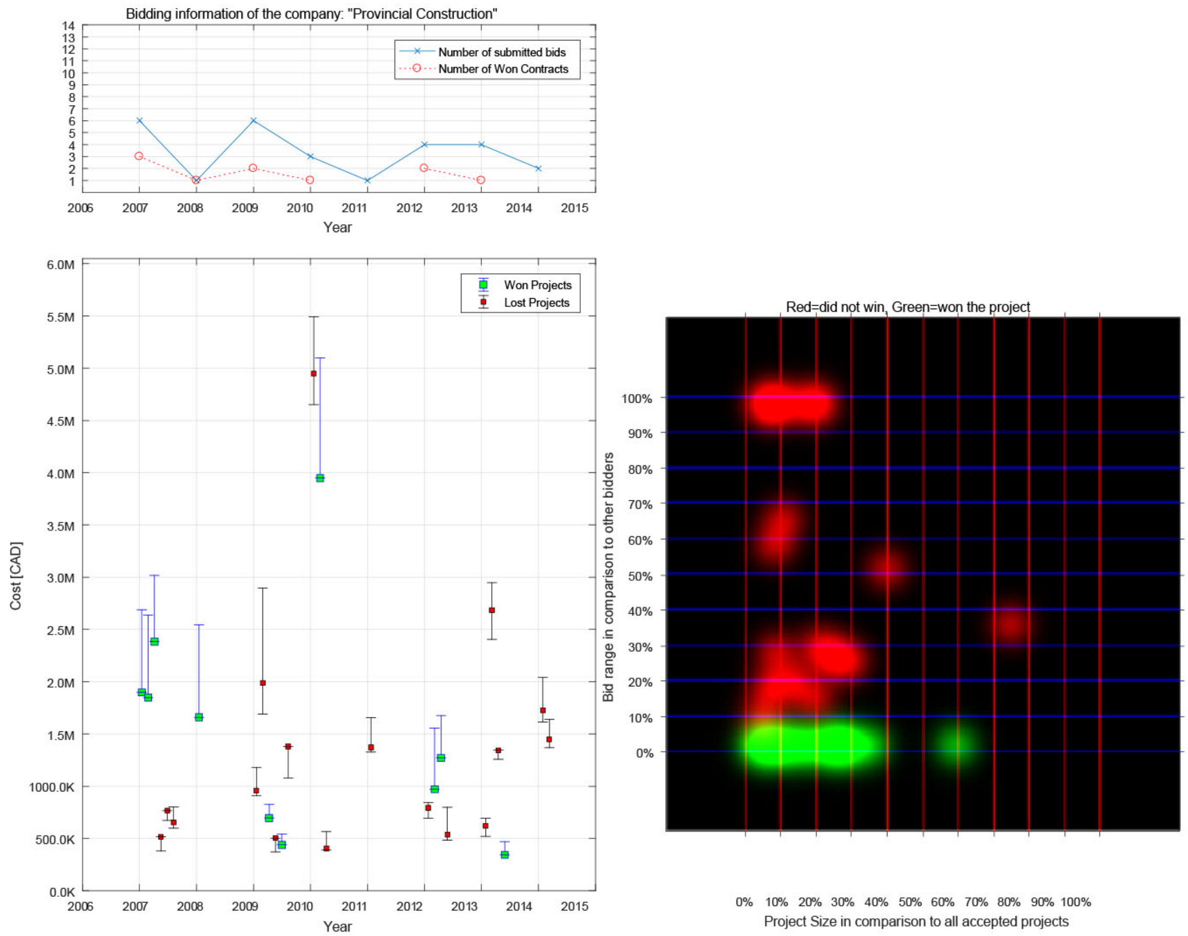



Figure 4.11: An example of contractor analysis plots that compare the historical bidding of a contractor.

to contractors or bidders. As such, the toolbox primarily serves as a demonstrative feature, exemplifying the potential benefits and delivery of such information to the operator.

Though not currently implemented, an extended representation of bidder statistics is proposed in Figure 4.11 on Page 136. This additional layer of analysis would delineate the annual bidding profile and bid range for individual contractors, adding depth to the system's analytical capabilities.

While the Bidder Analysis Toolbox within the WaterIAM server constitutes a robust component for understanding bidder activities, it is pivotal to recognize its limitations in the current implementation stage. Specifically, this toolbox does not possess the functionalities required to diagnose or predict collusion among bidders or to scrutinize past project costs.

These aspects represent complex areas that are beyond the toolbox's current capabilities.

Despite these constraints, the Bidder Analysis Toolbox holds significant potential for future enhancements. Its current design facilitates a nuanced comprehension of bidder activities, offering municipalities and operators valuable insights into the bidding landscape. Such features contribute substantially to the system's efficacy, even as they leave room for further exploration and development in subsequent research efforts.

### Unit Cost Analysis Toolbox

The Unit Cost Analysis Toolbox serves as a critical component within the server's interface, encapsulating the main functionality of the proposed algorithm. This toolbox facilitates the operator's ability to execute a refined version of the unit cost calculation method, as delineated in previous studies [Shapton, 2017], [Rehan et al., 2016].

Designed with versatility in mind, the toolbox can be applied to various city datasets and tailored to specific periods. Moreover, it supports distinct standard parts, such as Watermain" or Sanitary Sewer," providing further customization in the context of standard-parts. Each standard part's unit cost calculation parameters can be adjusted through the user-friendly interface, offering the operator control over the analysis parameters.

Figure 4.12 presents a sample webserver output when utilizing the Unit Cost Analysis Toolbox. As depicted, the webserver not only calculates unit costs for reference projects but also allows for selecting default values for specific project units. It includes parameters like unit pipe size, valve size, and the number of valves and hydrants, thereby providing a more comprehensive understanding of the unit cost dynamics.

The Unit Cost Analysis Toolbox is a specialized component within the system, explicitly designed to facilitate detailed cost assessments. Rooted in methodologies previously outlined in the literature, this toolbox has been developed with a clear and specific purpose in mind. It should be noted, however, that the toolbox's capabilities are confined to the execution of cost evaluations and do not extend to the analysis or prediction of past cost project costs.

It is paramount to cautiously approach any assertions regarding the toolbox's predictive capabilities, as its current design and functional parameters do not support these. Such limitations must be acknowledged to ensure a clear understanding of the system's capabilities and intended applications.

In summary, the Unit Cost Analysis Toolbox symbolizes a significant achievement in translating theoretical concepts into practical applications. Its dynamic design and


Figure 4.12: Sample webserver output while utilizing the Unit Cost Analysis Toolbox.

adaptability make it valuable for in-depth cost assessment operators. The toolbox's integration within the more extensive system enhances the overall robustness and illustrates a practical realization of academic methodologies, thereby contributing substantively to the field.

### Geographical filtering of the Projects and visualization toolbox

The "Map Visualizer" toolbox accurately represents how geographic information corresponding to various projects can be harnessed and utilized. Serving as a demonstrative example, this tool seeks to enrich financial analysis by overlaying historical financial data with their corresponding spatial information, thereby enhancing the depth and context of the interpretation.

On an interactive map, the toolbox visually presents the location of three distinct types of standard-parts: watermains, sanitary sewers, and roads. This visual representation promotes a more intuitive understanding of the spatial distribution of projects, enabling analysts to discern patterns and correlations that might otherwise remain obscured.

Each project's location is meticulously extracted from the contract or tender information provided by collaborating municipalities. This information not only adds to the authenticity of the data but also fosters a collaborative approach to information sharing between different governmental bodies.

To ensure compliance with privacy standards and maintain the data's anonymity, the locations shown in Figure 4.13 are randomly selected and do not correspond to actual projects. This precaution reflects the ethical considerations inherent in handling sensitive information and demonstrates a commitment to responsible data management.

## Conclusion

This chapter has methodically presented the implementation and outcomes of an AI model, detailing its multifaceted application in urban infrastructure projects across various cities. The AI methodology, designed to address specific challenges in data processing and analysis, successfully integrates natural language processing, deep learning classification, and unit cost optimization. These components form the foundation of a comprehensive system capable of standardizing and analyzing vast datasets with precision and efficiency.

The implementation of the AI model is structured into several progressive layers, each serving a distinct yet integral role in the data processing pipeline. Starting with Layer A, which focuses on importing and validating raw data tables, the model demonstrates

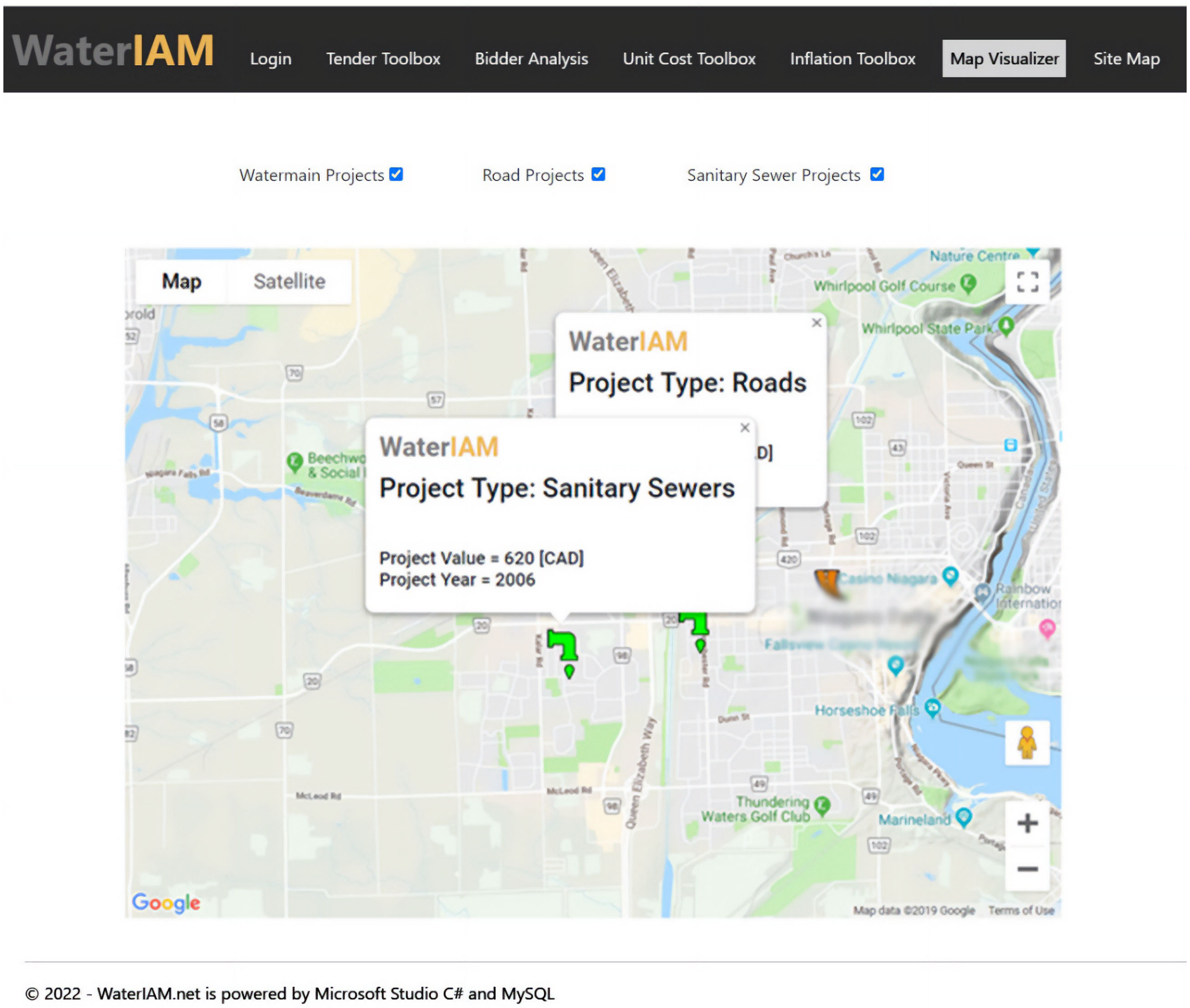



Figure 4.13: sample output of the webserver while using the map filtering toolbox.

meticulous attention to detail in handling data. The subsequent layers, B through H, extend this approach, ensuring that each step, from data cleaning to visualization, adheres to rigorous standards of accuracy and relevancy. The structured approach not only enhances the data's usability but also aligns with object-oriented programming principles, paving the way for potential migration to advanced programming platforms.

A key highlight of this chapter is the analysis of 277 contracts from three different cities, offering insights into the variability of contract values and the influence of various factors such as inflation, contract size, and pricing strategies. This analysis, underpinned by the unit cost index methodology, brings to light the complexities of urban infrastructure development and the need to consider contextual factors in interpreting data. The comparison with Shapton's findings further enriches this analysis, linking governmental policy changes to fluctuations in project numbers and unit costs.

The development of the Data Analysis Toolbox marks a significant leap in bridging the gap between theoretical research and practical application. The transition from an offline Microsoft Access application to an online, web-based platform underscores the project's evolution, catering to the dynamic needs of industry stakeholders and municipalities. The web server interface, with its focus on real-time interaction, data import and alignment, export functionality, and analysis capability, epitomizes the practical utility of the AI model. The various toolboxes, from the Tender Summary Index to the Unit Cost Analysis Toolbox, each contribute uniquely to the system's robustness and adaptability.

Furthermore, the introduction of the "Map Visualizer" toolbox exemplifies the potential of integrating geographical information with financial data, offering a more nuanced perspective on project distribution and trends. While ensuring privacy and ethical considerations, this tool enhances the depth and context of financial analyses, enriching the interpretation with spatial dynamics.

In conclusion, this chapter encapsulates the successful application of an AI model in analyzing and visualizing urban infrastructure data. It demonstrates the model's capacity to handle complex datasets, adhere to high standards of data processing, and provide insightful analyses that are crucial for decision-making in urban development. The integration of advanced AI techniques with a user-friendly interface and practical toolboxes underscores the project's commitment to making sophisticated methodologies accessible and relevant to industry and municipal stakeholders. The AI model, with its layered approach and comprehensive analysis, stands as a testament to the potential of AI in transforming data into actionable insights, thereby fostering scientific inquiry and operational efficiency in urban infrastructure management.

# Chapter 5

# Conclusions, Contributions, and Future Research

## 5.1 General Conclusions

This thesis has systematically unravelled various facets of civil engineering data management, analysis, and application. The four chapters are designed to work together, each building on the other to create a cohesive and innovative framework not currently seen in existing academic literature or industry solutions. A general summary is presented below:

**Chapter One** established the imperative for a systematic and automated approach towards importing, standardizing, classifying, and analyzing data in civil engineering. This chapter sets the stage for the rest of the thesis by identifying the fundamental problem that the subsequent chapters address.

**Chapter Two** provided the innovative ontology tool to structure data, contributing to enhancing the quality and sustainability of data handling in civil engineering. Introducing a lexicon specific to infrastructure capital works and the concept of data provenance are vital features that facilitate error correction and data refinement.

**Chapter Three** presented a significant advancement in classification methodologies through the use of LSTM, addressing the specific challenges of language constructs in tender-bid document records. This chapter's contribution to automating unit cost computations with high accuracy is a notable achievement that serves municipalities with standardized, consistent categorizations.

**Chapter Four** demonstrated the feasibility of the entire approach through the implemented web server. It distilled the theoretical insights into a practical application, showcasing the adaptability and diversity of the methodology. This chapter's contribution lies in simplifying complex tasks like unit cost calculation, an essential enhancement in efficiency and precision within civil engineering.

In essence, this thesis makes three major contributions:

1. Introduction of ontology as a tool for structuring data in civil engineering.
2. A novel approach to classification through LSTM, leading to automation in unit cost computation.
3. Implementation of a web server that translates theoretical concepts into real-world applications.

These contributions are woven together to create a pathway toward advanced data management and analysis tools in civil engineering. They significantly depart from existing methodologies, emphasizing data as a dynamic resource that fuels informed decision-making and elevates operational standards. This coherent and innovative framework heralds a new era of industry transformation and academic exploration.

## 5.2 Contributions

This research brings together the findings and methodologies from all chapters to enrich the existing body of knowledge by making the following impactful and original contributions that are not commonly found in current academic literature or industry solutions:

1. Introduction of a unique methodology that integrates data preprocessing, deep learning, and professional engineering insights to automate contract interpretation for watermain and sanitary sewer capital works. This methodology bridges the gap between traditional practices and modern AI-driven processes, significantly improving efficiency and accuracy.
2. Advancement of machine learning techniques tailored to civil engineering, including developing an AI model that emulates the unit cost computation from tender-bid documents. This replaces human-guided mapping, reduces overhead, and enhances classification accuracy.

3. Creation of a repeatable and adaptable approach, such as a design that allows model refinement with incoming data. This ensures continuous improvement and relevance, making the methodology more robust and versatile across different municipalities or regions.
4. Broadening the applicability of the methodology, extending potential applications beyond water systems to other civil engineering domains. This opens new avenues for innovation, setting a benchmark for data-driven decision-making in the industry.

Additionally, the research offers these significant features:

- A comprehensive lexicon and ontology specific to the industry, enhancing data contextualization and error detection.
- Compilation of a curated inventory of materials and services, paired with data provenance records, for future compatibility.
- Application of natural language processing for efficient classification and standardization across various municipalities.

The contributions and features detailed in this thesis collectively form a framework that addresses the current needs of civil engineering professionals while offering a basis for further refinements and advancements. This framework exhibits a significant shift from existing practices, indicating a step towards modernized approaches within the industry.

Future work could extend and refine this framework to better meet the evolving demands of the civil engineering domain. For instance, exploring the integration of real-time data processing capabilities, further customizing the methodology for diverse civil engineering subdomains, or enhancing the user interface of the implemented web server for a more intuitive user experience are potential avenues for future exploration. Additionally, collaboration with industry stakeholders to test and validate the framework in real-world settings could provide valuable insights and drive further innovations in practical applications.

## 5.3 Future Research Directions

Building upon the comprehensive framework established in this thesis for automating capital work planning and enhancing the assessment of project costs and tender contracts, several promising areas for future research emerge. These areas, while extending the current work, also open new avenues for exploration and innovation in civil engineering data management and application:

1. **Expansion to Other Civil Engineering Sub-Domains:** While the current framework is tailored to watermain and sanitary sewer capital works, extending this methodology to other areas such as transportation, urban planning, and environmental engineering could significantly broaden its impact.

2. **Enhancement of User Interface and Interaction:** Improving the user interface of the developed web server for a more intuitive and user-friendly experience would facilitate broader adoption and ease of use, especially for professionals less familiar with advanced data analysis tools.

3. **Multilingual and Cross-Cultural Adaptation:** Adapting the framework for use in different languages and cultural contexts would be beneficial, especially considering the global nature of civil engineering projects. It would involve not just language translation but also the customization of ontologies to reflect different construction norms and regulations.

4. **Advanced Machine Learning Models for Predictive Analytics:** Investigating the use of more advanced machine learning models, such as deep reinforcement learning or generative adversarial networks, could enhance predictive capabilities in cost estimations and risk assessments.

5. **Collaborative Validation and Refinement:** Working in collaboration with industry professionals to apply the framework in real-world settings would provide valuable feedback for refinement. It could include case studies or pilot projects in different municipalities or regions.

These potential research directions not only build upon the existing contributions of this thesis but also align with the ongoing evolution of civil engineering practices. By pursuing these avenues, future research can continue to advance the field, offering innovative solutions to complex challenges and further transforming industry practices.

# References

[Abanda et al., 2013a] Abanda, F., Tah, J., and Keivani, R. (2013a). Trends in built environment semantic web applications: where are we today? *Expert Systems with Applications*, 40(14):5563--5577.

[Abanda et al., 2013b] Abanda, F., Zhou, W., Tah, J., and Cheung, F. (2013b). Exploring the relationship between linked open data and building information modelling. In *Proceedings of the Sustainable Building Conference 2013*, pages 176--185.

[Abdalla et al., 2015] Abdalla, G., Damasceno, C. D. N., Guessi, M., Oquendo, F., and Nakagawa, E. Y. (2015). A systematic literature review on knowledge representation approaches for systems-of-systems. In *2015 IX Brazilian Symposium on Components, Architectures and Reuse Software*, pages 70--79. IEEE.

[Abdallah and Rosenberg, 2019] Abdallah, A. and Rosenberg, D. (2019). A data model to manage data for water resources systems modeling. *Environmental Modelling & Software*, 115:113--127.

[Abdul-Ghafour et al., 2007] Abdul-Ghafour, S., Ghodous, P., Shariat, B., and Perna, E. (2007). A common design-features ontology for product data semantics interoperability. In *IEEE/WIC/ACM International Conference on Web Intelligence (WI'07)*, pages 443--446.

[Adamson, 2010] Adamson, C. (2010). *Star Schema*. McGraw-Hill.

[Aggarwal and Karl, 2006] Aggarwal, N. and Karl, W. (2006). Line detection in images through regularized hough transform. *IEEE Transactions on Image Processing*, 15(3):582--591.

[Agostinho et al., 2007] Agostinho, C., Dutra, M., Jardim-Goncalves, R., Ghodous, P., and Steiger-Garcao, A. (2007). Express to owl morphism: making possible to enrich iso10303 modules. In Loureiro, G. and Curran, R., editors, *Complex Systems Concurrent Engineering*, pages 391--402, London, UK. Springer.

[Altman, 1992] Altman, N. S. (1992). An introduction to kernel and nearest-neighbor nonparametric regression. *The American Statistician*, 46(3):175--185.

[Anumba et al., 2008] Anumba, C., Issa, R., Pan, J., and Mutis, I. (2008). Ontology-based information and knowledge management in construction. *Constr. Innov.*, 8(3):218--239.

[Auer et al., 2015] Auer, S., Ermilov, I., Lehmann, J., and Martin, M. (2015). Lodstats -- 9960 datasets. (Last Accessed July 6, 2023).

[Baader and Nutt, 2003] Baader, F. and Nutt, W. (2003). Basic description logics. In Baader, F., Calvanese, D., McGuinness, D., Nardi, D., and Patel-Schneider, P., editors, *Description Logic Handbook: Theory, Implementation, and Applications*, pages 47--100. Cambridge University Press, Cambridge, MA, USA.

[Barbau et al., 2012] Barbau, R., Krima, S., Rachuri, S., Narayanan, A., Fiorentini, X., Foufou, S., and Sriram, R. (2012). Ontostep: Enriching product model data using ontologies. *Computers in Industry*, 44(6):575--590.

[Batini et al., 2021] Batini, C., Bellandi, V., Ceravolo, P., Moiraghi, F., Palmonari, M., and Siccardi, S. (2021). Semantic data integration for investigations: Lessons learned and open challenges. In *2021 IEEE International Conference on Smart Data Services (SMDS)*, pages 173--183.

[Batini et al., 2009] Batini, C., Cappiello, C., Francalanci, C., and Maurino, A. (2009). Methodologies for data quality assessment and improvement. *ACM Computing Surveys*, 41(3):16:1--16:52.

[Baumgartel et al., 2014] Baumgartel, K., Kadolsky, M., and Scherer, R. (2014). An ontology framework for improving building energy performance by utilizing energy saving regulations. In *Proceedings of the 10th European Conference on Product and Process Modelling (ECPPM)*, pages 519--526.

[Beetz et al., 2010] Beetz, J., van Berlo, L., de Laat, R., and van den Helm, P. (2010). BIMserver.org–An open source IFC model server. In *Proceedings of the CIP W78 Conference*.

[Beetz et al., 2005] Beetz, J., van Leeuwen, J., and de Vries, B. (2005). An ontology web language notation of the industry foundation classes. In *Proceedings of the 22nd CIB W78 Conference on Information Technology in Construction*, pages 193--198.

[Beetz et al., 2009] Beetz, J., van Leeuwen, J., and de Vries, B. (2009). Ifcowl: a case of transforming express schemas into ontologies. *Artificial Intelligence for Engineering Design, Analysis and Manufacturing*, 23(1):89--101.

[Bengio et al., 1994] Bengio, Y., Simard, P., and Frasconi, P. (1994). Learning long-term dependencies with gradient descent is difficult. *IEEE transactions on neural networks*, 5(2):157--166.

[Berners-Lee, 2003] Berners-Lee, T. (2003). WWW Past & Future. (Last Accessed 30 June 2023).

[Berners-Lee, 2006] Berners-Lee, T. (2006). Linked Data -- Design Issues. (Last Accessed July 6, 2023).

[Berners-Lee et al., 2001] Berners-Lee, T., Hendler, J., and Lassila, O. (2001). The semantic web. *Scientific American*, 284(5):35--43.

[Biau and Scornet, 2016] Biau, G. and Scornet, E. (2016). A random forest guided tour. *Invited Paper*, 25:197--227.

[Bilgin et al., 2018] Bilgin, G., Dikmen, I., and Birgonul, M. T. (2018). An ontology-based approach for delay analysis in construction. *KSCE Journal of Civil Engineering*, 22(2):384--398.

[Borgo et al., 2015] Borgo, S., Sanfilippo, E., Sojic, A., and Terkaj, W. (2015). Ontological analysis and engineering standards: an initial study of IFC. In Ebrahimipour, V. and Yacout, S., editors, *Ontology Modeling in Physical Asset Integrity Management*, pages 17--43. Springer.

[Bosch et al., 2005] Bosch, V., Gonzalez, E., and Tamayo, F. (2005). Tqm and qfd: exploiting a customer complaint management system. *International Journal of Quality & Reliability Management*, 22(1):30--37.

[Breiman, 2001] Breiman, L. (2001). Random forests. *Machine learning*, 45:5--32.

[Brickley and Guha, 2014] Brickley, D. and Guha, R. V. (2014). RDF Schema 1.1 -- W3C Recommendation 25 February 2014. (Last Accessed 30 June 2023).

[Buneman et al., 2001] Buneman, P., Khanna, S., and Wang-Chiew, T. (2001). Why and where: A characterization of data provenance. *Database Theory - ICDT 2001*, pages 316--330.

[Chen and Han, 2016] Chen, Y. and Han, D. (2016). Big data and hydroinformatics. *Journal of Hydroinformatics*, 18(4):599--614.

[Choat et al., 2022] Choat, B., Pulido, A., Bhaskar, A., Hale, R., Zhang, H., Meixner, T., McPhillips, L., Hopkins, K., Cherrier, J., and Cheng, C. (2022). A call to record stormwater control functions and to share network data. *Journal of Sustainable Water in the Built Environment*, 8(2):02521005.

[Choi et al., 2021] Choi, S.-W., Lee, E.-B., and Kim, J.-H. (2021). The engineering machine-learning automation platform (emap): A big-data-driven ai tool for contractors' sustainable management solutions for plant projects. *Sustainability*, 13(18).

[Christopher Pereira, 2020] Christopher Pereira, P. (2020). Text-mining maintenance records to automate the identification and grouping of failure modes. In *Offshore Technology Conference*, page D041S055R002. OTC.

[Codd, 1989] Codd, E. F. (1989). Relational database: A practical foundation for productivity. *Readings in Artificial Intelligence & Databases*, pages 60--68.

[Connolly and Beg, 2005] Connolly, T. and Beg, C. (2005). *Database Systems: A Practical Approach to Design, Implementation, and Management*. Pearson, Boston, 4th edition.

[Costin et al., 2017] Costin, A. M., Eastman, C., and Issa, R. R. A. (2017). The need for taxonomies in the ontological approach for interoperability of heterogeneous information models. In *Computing in Civil Engineering 2017*, pages 9--17. American Society of Civil Engineers.

[Coussement and Van den Poel, 2008] Coussement, K. and Van den Poel, D. (2008). Improving customer complaint management by automatic email classification using linguistic style features as predictors. *Decision Support Systems*, 44(4):870--882.

[Cruz et al., 2013] Cruz, I. F., Palmonari, M., Caimi, F., and Stroe, C. (2013). Building linked ontologies with high precision using subclass mapping discovery. *Artificial Intelligence Review*, 40(2):127--145.

[Dai et al., 2008] Dai, C., Lin, D., Bertino, E., and Kantarcioglu, M. (2008). An approach to evaluate data trustworthiness based on data provenance. In Jonker, W. and Petkovic, M., editors, *Secure Data Management*, Lecture Notes in Computer Science, pages 82--98. Springer.

[Dang et al., 2020] Dang, N. C., Moreno-García, M. N., and De la Prieta, F. (2020). Sentiment analysis based on deep learning: A comparative study. *Electronics*, 9(3).

[Daraio et al., 2016] Daraio, C., Lenzerini, M., Leporelli, C., Naggar, P., Bonaccorsi, A., and Bartolucci, A. (2016). The advantages of an ontology-based data management approach: openness, interoperability and data quality. *Scientometrics*, 108(1):441--455.

[de Farias et al., 2015] de Farias, T., Roxin, A., and Nicolle, C. (2015). IfcWoD: Semantically adapting IFC model relations into OWL properties. In *Proceedings of the 32rd International CIB W78 Conference*, pages 175--185, Eindhoven, NL.

[Devlin and Cote, 1996] Devlin, B. and Cote, L. D. (1996). *Data warehouse: from architecture to implementation*. Addison-Wesley Longman Publishing Co., Inc.

[Devlin et al., 2018] Devlin, J., Chang, M.-W., Lee, K., and Toutanova, K. (2018). Bert: Pre-training of deep bidirectional transformers for language understanding.

[Dietterich, 2000] Dietterich, T. (2000). An experimental comparison of three methods for constructing ensembles of decision trees: Bagging, boosting, and randomization. *Machine learning*, 40:139--157.

[EIU, 2020] EIU, E. I. U. (2020). Benchmarking infrastructure costs: A case of road and water basket of locally-obtained commodities (bloc). *Asian Infrastructure Investment Bank*.

[El-Diraby, 2013a] El-Diraby, T. (2013a). Domain ontology for construction knowledge. *Journal of Construction Engineering and Management*, 139(7):768--784.

[El-Diraby, 2013b] El-Diraby, T. (2013b). Domain ontology for construction knowledge. *Journal of Construction Engineering and Management*, 139(7):768--784.

[El-Diraby and Osman, 2011] El-Diraby, T. and Osman, H. (2011). A domain ontology for construction concepts in urban infrastructure products. *Automation in Construction*, 20(8):1120--1132.

[El-Diraby and Zhang, 2006] El-Diraby, T. and Zhang, J. (2006). A semantic framework to support corporate memory management in building construction. *Automation in Construction*, 15(4):504--521.

[El-Diraby et al., 2005] El-Diraby, T. A., Lima, C., and Feis, B. (2005). Domain taxonomy for construction concepts: Toward a formal ontology for construction knowledge. *Journal of Computing in Civil Engineering*, 19(4):394--406.

[El-Diraby, 2013c] El-Diraby, T. E. (2013c). Domain ontology for construction knowledge. *Journal of Construction Engineering and Management*, 139(7):768--784.

[El-Gohary and El-Diraby, 2010] El-Gohary, N. and El-Diraby, T. (2010). Domain ontology for processes in infrastructure and construction. *J. Constr. Eng. Manag.*, 136(7):730--744.

[Fisher and Kingma, 2001] Fisher, C. W. and Kingma, B. R. (2001). Criticality of data quality as exemplified in two disasters. *Information & Management*, 39(2):109--116.

[Gao et al., 2018] Gao, T., Du, J., Dai, L.-R., and Lee, C.-H. (2018). Densely connected progressive learning for LSTM-based speech enhancement. In *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 5054--5058. IEEE.

[Gao and Pishdad-Bozorgi, 2020] Gao, X. and Pishdad-Bozorgi, P. (2020). A framework of developing machine learning models for facility life-cycle cost analysis. *Building Research & Information*, 48(5):501--525.

[Garijo et al., 2014] Garijo, D., Alper, P., Belhajjame, K., Corcho, O., Gil, Y., and Goble, C. (2014). Common motifs in scientific workflows: An empirical analysis. *Future generation computer systems*, 36:338--351.

[Ghang et al., 2014] Ghang, L., Jiyong, J., Jongsung, W., Chiyon, C., Seok-joon, Y., Sungil, H., and Hoonsig, K. (2014). Query performance of the IFC model server using an object-relational database approach and a traditional relational database approach. *J. Comput. Civ. Eng.*, 28(2):210--222.

[Goldberg et al., 1993] Goldberg, D. E., Deb, K., Kargupta, H., and Harik, G. (1993). Rapid, accurate optimization of difficult problems using messy genetic algorithms. In *Proceedings of the Fifth International Conference on Genetic Algorithms,(Urbana, USA)*, pages 59--64. Proceedings of the Fifth International Conference on Genetic Algorithms ....

[Goldberg and Levy, 2014] Goldberg, Y. and Levy, O. (2014). word2vec explained: deriving mikolov et al.'s negative-sampling word-embedding method. *arXiv preprint arXiv:1402.3722*.

[Grosof and Poon, 2003] Grosof, B. N. and Poon, T. C. (2003). SweetDeal: Representing agent contracts with exceptions using XML rules, ontologies, and process descriptions. In *Int. Conference on World Wide Web*, page 10. Cite 174.

[Han et al., 2011] Han, J., Pei, J., and Kamber, M. (2011). *Data Mining: Concepts and Techniques*. Elsevier.

[Hartmann et al., 2019] Hartmann, J., Huppertz, J., Schamp, C., and Heitmann, M. (2019). Comparing automated text classification methods. *International Journal of Research in Marketing*, 36(1):20--38.

[Hausenblas and Kim, 2012] Hausenblas, M. and Kim, J. (2012). 5-star open data. (Last Accessed July 6, 2023).

[Hitzler et al., 2012] Hitzler, P., Krotzsch, M., Parsia, B., Patel-Schneider, P., and Rudolph, S. (2012). *OWL 2 Web Ontology Language Primer (Second Edition) -- W3C Recommendation 11 December 2012*. (Last Accessed 30 June 2023).

[Hoaglin, 2013] Hoaglin, D. C. (2013). Volume 16: How to detect and handle outliers. In *https://api.semanticscholar.org/CorpusID:208231456*.

[Hong et al., 2022] Hong, S., Kim, J., and Yang, E. (2022). Automated text classification of maintenance data of higher education buildings using text mining and machine learning techniques. *Journal of Architectural Engineering*, 28(1):04021045.

[Horrocks et al., 2005] Horrocks, I., Parsia, B., Patel-Schneider, P., and Hendler, J. (2005). Semantic web architecture: Stack or two towers? In Fages, F. and Soliman, S., editors, *Principles and Practice of Semantic Web Reasoning*, volume 3703 of *Lecture Notes in Computer Science (LNCS)*, pages 37--41. Springer, Berlin Heidelberg.

[Horsburgh et al., 2016] Horsburgh, J., Aufdenkampe, A., Mayorga, E., Lehnert, K., Hsu, L., Song, L., Spackman, Jones, A., Damiano, S., Tarboton, D., Valentine, D., Zaslavsky, I., and Whitenack, T. (2016). Observations data model 2: A community information model for spatially discrete earth observations. *Environmental Modelling & Software*, 79:55--74.

[Horsburgh et al., 2008] Horsburgh, J., Tarboton, D., Maidment, D., and Zaslavsky, I. (2008). A relational model for environmental and water resources data. *Water Resources Research*, 44(5).

[International Organization for Standardization, 1994] International Organization for Standardization (1994). ISO 10303: STEP Overview -- Product Data Representation and Exchange.

[Isozaki, 2001] Isozaki, H. (2001). Japanese named entity recognition based on a simple rule generator and decision tree learning. In *Proceedings of the 39th Annual Meeting on Association for Computational Linguistics*, ACL '01, pages 314--321, USA. Association for Computational Linguistics.

[Issa and Mutis, 2015] Issa, R. and Mutis, I. (2015). *Ontology in the AEC industry: a decade of research and development in architecture, engineering, and construction*. American Society of Civil Engineers.

[Jasper and Uschold, 1999] Jasper, R. and Uschold, M. (1999). A framework for understanding and classifying ontology applications. In *IJCAI*, page 20.

[Jeong et al., 2010] Jeong, J., Lee, G., and Kang, H. (2010). Preliminary performance evaluation of an ORDB-based IFC server and an RDB-based IFC server by using the BUCKY benchmark method. In *Proceedings of CIB World Congress*.

[Jiang et al., 2015] Jiang, Y., Yu, N., Ming, J., Lee, S., DeGraw, J., Yen, J., Messner, J., and Wu, D. (2015). Automatic building information model query generation. *ITcon*, 20:518--535.

[Jotne Co., 2014] Jotne Co., l. (2014). EDM Model Server (IFC). http://www.epmtech.jotne.com/. Last access: July 6, 2023.

[Kadolsky et al., 2014] Kadolsky, M., Baumgärtel, K., and Scherer, R. (2014). An ontology framework for rule-based inspection of eebim-systems. *Procedia Engineering*, 85:293--301.

[Karimi et al., 2020] Karimi, D., Dou, H., Warfield, S. K., and Gholipour, A. (2020). Deep learning with noisy labels: Exploring techniques and remedies in medical image analysis. *Medical Image Analysis*, 65:101759.

[Khaki, 2021] Khaki, M. (2021). Progressive cleaning and mining of uncertain smart water meter data. In *2nd Global Conference on Artificial Intelligence and Applications*, pages 229--240. CRC Press. Best Paper Award.

[Khalili and Chua, 2015] Khalili, A. and Chua, D. (2015). IFC-based graph data model for topological queries on building elements. *J. Comput. Civ. Eng.*, 29(3):401--4046.

[Kim and Grobler, 2009] Kim, H. and Grobler, F. (2009). Design coordination in building information modeling (bim) using ontological consistency checking. *Journal of Computing in Civil Engineering*, pages 410--420.

[Kim et al., 2005] Kim, H., Howland, P., Park, H., and Christianini, N. (2005). Dimension reduction in text classification with support vector machines. *Journal of machine learning research*, 6(1).

[Kim et al., 2003] Kim, W., Choi, B.-J., Hong, E.-K., Kim, S.-K., and Lee, D. (2003). A taxonomy of dirty data. *Data Mining & Knowledge Discovery*, 7(1):81--99.

[Kimball and Ross, 2011] Kimball, R. and Ross, M. (2011). *The Data Warehouse Toolkit: The Complete Guide to Dimensional Modeling*. John Wiley & Sons.

[Koleck et al., 2019] Koleck, T. A., Dreisbach, C., Bourne, P. E., and Bakken, S. (2019). Natural language processing of symptoms documented in free-text narratives of electronic health records: a systematic review. *Journal of the American Medical Informatics Association*, 26(4):364--379.

[Krima et al., 2009] Krima, S., Barbau, R., Fiorentini, X., Sudarsan, R., and Sriram, R. (2009). Ontostep: Owl-dl ontology for step. In *Proceedings of the 2009 International Conference on Product Lifecycle Management*, pages 770--780.

[Lakshmanan et al., 1993] Lakshmanan, L. V. S., Sadri, F., and Subramanian, I. N. (1993). On the logical foundations of schema integration and evolution in heterogeneous database systems. In Ceri, S., Tanaka, K., and Tsur, S., editors, *Deductive and Object-Oriented Databases*, volume 760, pages 81--100, Berlin, Heidelberg.

[Le and Jeong, 2016] Le, T. and Jeong, H. (2016). Interlinking life-cycle data spaces to support decision making in highway asset management. *Automation in Construction*, 64:54--64.

[Lee et al., 2016] Lee, Y.-C., Eastman, C. M., Solihin, W., and See, R. (2016). Modularized rule-based validation of a bim model pertaining to model views. *Automation in Construction*, 63(March):1--11.

[Li et al., 2016] Li, H., Liu, H., Liu, Y., and Wang, Y. (2016). An object-relational ifc storage model based on oracle database. *Int. Arch. Photogramm. Remote. Sens. Spat. Inf. Sci.*, 41.

[Li et al., 2015] Li, S., Sun, Y., and Soergel, D. (2015). A new method for automatically constructing domain-oriented term taxonomy based on weighted word co-occurrence analysis. *Scientometrics*, 103(3):1023--1042.

[Lima et al., 2003] Lima, C., El-Diraby, T., Fies, B., Zarli, A., and Ferneley, E. (2003). The E-Cognos project: current status and future directions of an ontology-enabled IT solution infrastructure supporting knowledge management in construction. In *Proceedings of the Construction Research Congress*, pages 1--8.

[Lima et al., 2005] Lima, C., El-Diraby, T., and Stephens, J. (2005). Ontology-based optimization of knowledge management in construction. *Journal of Information Technology in Construction*, 10:305--327.

[Lima et al., 2002] Lima, C., Fies, B., Zarli, A., Bourdeau, M., Wetherill, M., and Rezgui, Y. (2002). Towards an IFC-enabled ontology for the building and construction industry: the e-cognos approach. In *Proceedings of the eSM@RT 2002 Conference*, pages 254--264.

[Lindén et al., 2018] Lindén, J., Forsström, S., and Zhang, T. (2018). Evaluating combinations of classification algorithms and paragraph vectors for news article classification. In *2018 Federated Conference on Computer Science and Information Systems (FedCSIS)*, pages 489--495.

[Liu et al., 2016] Liu, H., Lu, M., and Al-Hussein, M. (2016). Ontology-based semantic approach for construction-oriented quantity take-off from BIM models in the light-frame building industry. *Adv. Eng. Inform.*, 30(2):190--207.

[Liu and Ma, 2015] Liu, Z. and Ma, Z. (2015). Establishing formalized representation of standards for construction cost estimation by using ontology learning. *Procedia Engineering*, 123:291--299.

[Loffredo, 1998] Loffredo, D. T. (1998). *Efficient database implementation of EXPRESS information models*. Rensselaer Polytechnic Institute.

[Magdy and Elsayed, 2016] Magdy, W. and Elsayed, T. (2016). Unsupervised adaptive microblog filtering for broad dynamic topics. *Information Processing & Management*, 52(4):513--528.

[Maidment, 2008] Maidment, D. (2008). Bringing water data together. *Journal of Water Resources Planning and Management*, 134(2):95--96.

[Mazairac and Beetz, 2013] Mazairac, W. and Beetz, J. (2013). BIMQL–an open query language for building information models. *Adv. Eng. Inform.*, 27(4):444--456.

[Metral et al., 2010] Metral, C., Billen, C., Cutting-Decelle, A., and van Ruymbeke, M. (2010). Ontology-based approaches for improving the interoperability between 3d urban models. *Journal of Information Technology in Construction*, 15:169--184.

[Metral et al., 2009] Metral, C., Falquet, G., and Cutting-Decelle, A. (2009). Towards semantically enriched 3d city models: an ontology-based approach. In *Proceedings of the GeoWeb 2009 Academic Track --- Cityscapes --- International Archives of Photogrammetry, Remote Sensing and Spatial Information Sciences*, pages 40--45.

[Mika et al., 1999] Mika, S., Ratsch, G., Weston, J., Scholkopf, B., and Mullers, K.-R. (1999). Fisher discriminant analysis with kernels. In *Neural networks for signal processing IX: Proceedings of the 1999 IEEE signal processing society workshop (cat. no. 98th8468)*, pages 41--48. Ieee.

[Missier et al., 2013] Missier, P., Belhajjame, K., and Cheney, J. (2013). The w3c prov family of specifications for modelling provenance metadata. In *Proceedings of the 16th International Conference on Extending Database Technology*, pages 773--776.

[Mohanta and Das, 2016] Mohanta, A. and Das, S. (2016). Ict-based facilities management tools for buildings. In *Proceedings of International Conference on ICT for Sustainable Development: ICT4SD 2015 Volume 1*, pages 125--133. Springer.

[Moreau et al., 2013] Moreau, L., Missier, P., Cheney, J., and Soiland-Reyes, S. (2013). Prov-n: The provenance notation. *W3C Recommendation*.

[Mubarak et al., 2020] Mubarak, S. A. et al. (2020). *How to estimate with RSMeans data: basic skills for building construction*. John Wiley & Sons.

[Musen, 1998] Musen, M. A. (1998). Modern architectures for intelligent systems: Reusable ontologies and problem-solving methods. *AMIA Symposium*, page 7.

[Musen, 2015] Musen, M. A. (2015). The protégé project: a look back and a look forward. *AI matters*, 1(4):4--12.

[Muñoz-Soro et al., 2016] Muñoz-Soro, J. F., Esteban, G., Corcho, O., and Serón, F. (2016). PPROC, an ontology for transparency in public procurement. *Semantic Web*, 7(3):295--309.

[Navigli, 2009] Navigli, R. (2009). Word sense disambiguation: A survey. *ACM Comput. Surv.*, 41(2).

[Ngo et al., 2020] Ngo, J., Hwang, B.-G., and Zhang, C. (2020). Big data and predictive analytics in the construction industry: Applications, status quo, and potential in singapore 2019;s construction industry. *Construction Research Congress 2020*, pages 715--724.

[Niknam and Karshenas, 2013] Niknam, M. and Karshenas, S. (2013). A semantic web service approach to construction cost estimating. In *Computing in Civil Engineering*, pages 484--491. American Society of Civil Engineers.

[Paliouras et al., 2000] Paliouras, G., Karkaletsis, V., Petasis, G., and Spyropoulos, C. D. (2000). Learning decision trees for named-entity recognition and classification. In *ECAI Workshop on Machine Learning for Information Extraction*.

[Pauwels et al., 2011] Pauwels, P., De Meyer, R., and Van Campenhout, J. (2011). Interoperability for the design and construction industry through semantic web technology. In Declerck, T., Granitzer, M., Grzegorzek, M., Romanelli, M., Ruger, S., and Sintek, M., editors, *Semantic Multimedia*, pages 143--158, Berlin, Heidelberg. Springer Berlin Heidelberg.

[Pauwels et al., 2015] Pauwels, P., Terkaj, W., Krijnen, T., and Beetz, J. (2015). Coping with lists in the ifcowl ontology. In *Proceedings of the 22nd EG-ICE International Workshop*, pages 113--122.

[Pauwels et al., 2017] Pauwels, P., Zhang, S., and Lee, Y. (2017). Semantic web technologies in aec industry: A literature overview. *Automation in Construction*, 73:145--165.

[Perez-Urbina et al., 2012] Perez-Urbina, H., Sirin, E., and Clark, K. (2012). Validating rdf with owl integrity constraints. (Last Accessed 24 August 2016).

[Pileggi and Amor, 2013] Pileggi, S. and Amor, R. (2013). Addressing semantic geographic information systems. *Future Internet*, 5(4):585--590.

[Pipino et al., 2002] Pipino, L. L., Lee, Y. W., and Wang, R. Y. (2002). Data quality assessment. *Communications of the ACM*, 45(4):211--218.

[Pustejovsky and Boguraev, 1993] Pustejovsky, J. and Boguraev, B. (1993). Lexical knowledge representation and natural language processing. *Artificial Intelligence*, 63(1-2):193--223.

[Pyon et al., 2011] Pyon, C., Woo, J., and Park, S. (2011). Service improvement by business process management using customer complaints in financial service industry. *Expert Systems with Applications*, 38(4):3267--3279.

[Qi and Davison, 2009] Qi, X. and Davison, B. D. (2009). Web page classification: Features and algorithms. *ACM Comput. Surv.*, 41(2).

[Quinlan, 1986] Quinlan, J. R. (1986). Induction of decision trees. *Machine learning*, 1:81--106.

[Raganato et al., 2017] Raganato, A., Camacho-Collados, J., and Navigli, R. (2017). Word sense disambiguation: A unified evaluation framework and empirical comparison. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 1, Long Papers*, pages 99--110.

[Rahm and Do, 2000] Rahm, E. and Do, H. H. (2000). Data cleaning: Problems and current approaches. *IEEE Data Eng. Bull.*, 23(4):3--13.

[Ratinov and Roth, 2009] Ratinov, L. and Roth, D. (2009). Design challenges and misconceptions in named entity recognition. In *Proceedings of the thirteenth conference on computational natural language learning (CoNLL-2009)*, pages 147--155.

[Reeder and David, 2016] Reeder, B. and David, A. (2016). Health at hand: A systematic review of smart watch uses for health and wellness. *Journal of Biomedical Informatics*, 63:269--276.

[Rehan et al., 2016] Rehan, R., Younis, R., Unger, A. J. A., Shapton, B., Budimir, F., and Knight, M. A. (2016). Development of unit cost indices and database for water and wastewater pipelines capital works. *J. of Cost Analysis & Parametrics*, 9(2):127--160.

[Rezgui et al., 2011] Rezgui, Y., Boddy, S., Wetherill, M., and Cooper, G. (2011). Past, present and future of information and knowledge sharing in the construction industry: towards semantic service-based e-construction? *Computer-Aided Design*, 43(5):502--515.

[Ricquebourg et al., 2007] Ricquebourg, V., Durand, D., Menga, D., Marhic, B., Delahoche, L., Loge, C., and Jolly-Desodt, A. (2007). Context inferring in the smart home: an swrl approach. In *Proceedings of the 21st International Conference on Advanced Information Networking and Applications Workshops*, pages 290--295.

[Rischmoller et al., 2000a] Rischmoller, L., Fox, R., Williams, M., and Alarcon, L. (2000a). Automation and visualization tools to improve support for process integration in the construction industry. In *International Conference in Construction Information Technology*, pages 17--18.

[Rischmoller et al., 2000b] Rischmoller, L., Fox, R., Williams, M., and Alarcon, L. (2000b). Automation and visualization tools to improve support for process integration in the construction industry. In *International Conference in Construction Information Technology*, pages 17--18.

[Ruckert and Sjogren, 2022] Ruckert, L. and Sjogren, H. (2022). *Exploring State-of-the-Art Natural Language Processing Models with Regards to Matching Job Adverts and Resumes*. PhD thesis, Uppsala University, Division of Systems and Control.

[Ruikar et al., 2007] Ruikar, D., Anumba, C., Duke, A., Carrillo, P., and Bouchlaghem, N. (2007). Using the semantic web for project information management. *Facilities*, 25:507--524.

[Sadiq et al., 2011] Sadiq, S. W., Indulska, M., and Jayawardene, V. (2011). Research and industry synergies in data quality management. In *ICIQ*.

[Sawhney et al., 2004] Sawhney, A., Walsh, K. D., and Brown IV, A. (2004). International comparison of cost for the construction sector: towards a conceptual model. *Civil Engineering and Environmental Systems*, 21(3):151--167.

[Schevers and Drogemuller, 2006] Schevers, H. and Drogemuller, R. (2006). Converting the industry foundation classes to the web ontology language. In *Proceedings of the First International Conference on Semantics, Knowledge, and Grid*, pages 73--75.

[Schmachtenberg et al., 2014] Schmachtenberg, M., Bizer, C., and Paulheim, H. (2014). State of the lod cloud 2014. (Last Accessed July 6, 2023).

[Schreiber and Raimond, 2014] Schreiber, G. and Raimond, Y. (2014). RDF 1.1 Primer -- W3C Working Group Note 24 June 2014. (Last Accessed 30 June 2030).

[Sebastiani, 2002] Sebastiani, F. (2002). Machine learning in automated text categorization. *ACM Comput. Surv.*, 34(1):1–47.

[Shah et al., 2011] Shah, N., Chao, K., Zlamaniec, T., and Matei, A. (2011). Ontology for home energy management domain. In *Digital Information and Communication Technology and Its Applications*, volume 167 of *Communications in Computer and Information Science*, pages 337--347.

[Shapton, 2017] Shapton, B. (2017). Development of unit price indices and estimating inflation for potable water and wastewater pipeline capital works construction. phdthesis, University of Waterloo.

[Shvaiko and Euzenat, 2005] Shvaiko, P. and Euzenat, J. (2005). A survey of schema-based matching approaches. *Journal on Data Semantics IV*, 3730:146--171.

[Siami-Namini et al., 2019] Siami-Namini, S., Tavakoli, N., and Namin, A. S. (2019). The performance of LSTM and BiLSTM in forecasting time series. In *2019 IEEE International Conference on Big Data (Big Data)*, pages 3285--3292. IEEE.

[Simmhan et al., 2005] Simmhan, Y. L., Plale, B., and Gannon, D. (2005). A survey of data provenance in e-science. *SIGMOD Rec.*, 34(3):31--36.

[Single et al., 2020] Single, J. I., Schmidt, J., and Denecke, J. (2020). Knowledge acquisition from chemical accident databases using an ontology-based method and natural language processing. *Safety Science*, 129:104747.

[Smith et al., 2023] Smith, V., McGauley, M., Newman, M., Garzio-Hadzick, A., Kurzweil, A., Wadzuk, B., and Traver, R. (2023). A relational data model for advancing stormwater infrastructure management. *Journal of Sustainable Water in the Built Environment*, 9(1):04022023.

[Solihin et al., 2017] Solihin, W., Eastman, C., Lee, Y., and Yang, D. (2017). A simplified relational database schema for transformation of bim data into a query-efficient and spatially enabled database. *Automation in Construction*, 84:367--383.

[Stuckenschmidt, 2009] Stuckenschmidt, H. (2009). A semantic similarity measure for ontology-based information. In *Flexible Query Answering Systems: 8th International Conference, FQAS 2009, Roskilde, Denmark, October 26-28, 2009. Proceedings 8*, pages 406--417. Springer.

[Tao et al., 2010] Tao, J., Sirin, E., Bao, J., and McGuinness, D. (2010). Extending owl with integrity constraints. In *International Workshop on Description Logics (DL2010)*, volume 573 of *CEUR Workshop Proceedings*, pages 137--148.

[Terkaj and Sojic, 2015] Terkaj, W. and Sojic, A. (2015). Ontology-based representation of ifc express rules: an enhancement of the ifcowl ontology. *Automation in Construction*, 57(September):188--201.

[The MathWorks, 2022] The MathWorks, I. (2022). Matlab deep learning toolbox. [Online; accessed Jul 19, 2022], Available at https://www.mathworks.com/help/deeplearning/ug/long-short-term-memory-networks.html.

[Traver and Ebrahimian, 2017] Traver, R. and Ebrahimian, A. (2017). Dynamic design of green stormwater infrastructure. *Frontiers of Environmental Science & Engineering*, 11:1--6.

[Turk, 2001] Turk, Z. (2001). Phenomenological foundations of conceptual product modelling in architecture, engineering and construction. *Artificial Intelligence in Engineering*, 15(2):83 -- 92.

[Turk, 2006] Turk, Z. (2006). Construction informatics: Definition and ontology. *Advanced Engineering Informatics*, 20(2):187--199.

[Van Houdt et al., 2020] Van Houdt, G., Mosquera, C., and Nápoles, G. (2020). A review on the long short-term memory model. *Artificial Intelligence Review*, 53(8):5929--5955.

[Vaswani et al., 2017] Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., and Polosukhin, I. (2017). Attention is all you need. *Advances in neural information processing systems*, 30.

[Veltman, 2001] Veltman, K. (2001). Syntactic and semantic interoperability: new approaches to knowledge and the semantic web. *The New Review of Information Networking*, 7:159--184.

[Venugopal et al., 2015] Venugopal, M., Eastman, C., and Teizer, J. (2015). An ontology-based analysis of the industry foundation class schema for building information model exchanges. *Advanced Engineering Informatics*, 29(4):940--957. Collective Intelligence Modeling, Analysis, and Synthesis for Innovative Engineering Decision Making Special Issue of the 1st International Conference on Civil and Building Engineering Informatics.

[W3C OWL Working Group, 2012] W3C OWL Working Group (2012). *OWL2 Web Ontology Language Document Overview (Second Edition) -- W3C Recommendation 11 December 2012*. (Last Accessed 30 June 2023).

[W3C Report, 2014] W3C Report (2014). W3C Linked Building Data Community Group. (Last Accessed 24 August 2016).

[Wang et al., 2019] Wang, C., Li, M., and Smola, A. J. (2019). Language models with transformers.

[Wang et al., 2020] Wang, M., Wang, C. C., Sepasgozar, S., and Zlatanova, S. (2020). A systematic review of digital technology adoption in off-site construction: Current status and future direction towards industry 4.0. *Buildings*, 10(11):204.

[Wang et al., 1995] Wang, R., Storey, V., and Firth, C. (1995). A framework for analysis of data quality research. *IEEE Transactions on Knowledge and Data Engineering*, 7(4):623--640.

[Wang and El-Gohary, 2021] Wang, X. and El-Gohary, N. (2021). *Deep Learning-Based Named Entity Recognition from Construction Safety Regulations for Automated Field Compliance Checking*, pages 164--171. ASCE Library.

[Whyte and Donaldson, 2015] Whyte, A. and Donaldson, J. (2015). Digital model data distribution in civil engineering contracts. *Built Environment Project and Asset Management*, 5(3):248--260.

[Wicaksono et al., 2010] Wicaksono, H., Sven, R., and Kusnady, E. (2010). Knowledge-based intelligent energy management using building automation system. In *Proceedings of the 2010 IPEC Conference*, pages 1140--1145.

[Wu et al., 2022] Wu, C., Li, X., Guo, Y., Wang, J., Ren, Z., Wang, M., and Yang, Z. (2022). Natural language processing for smart construction: Current status and future directions. *Automation in Construction*, 134:104059.

[Wu et al., 2006] Wu, Y.-C., Fan, T.-K., Lee, Y.-S., and Yen, S.-J. (2006). Extracting named entities using support vector machines. In *International workshop on knowledge discovery in life science literature*, pages 91--103. Springer.

[Yang and Bayapu, 2020] Yang, E. and Bayapu, I. (2020). Big data analytics and facilities management: a case study. *Facilities*, 38(3/4):268--281.

[Yang and Zhang, 2006] Yang, Q. and Zhang, Y. (2006). Semantic interoperability in building design: Methods and tools. *Computer-Aided Design*, 38(10):1099--1112.

[Yin et al., 2012] Yin, H., Costa, J. A., and Barreto, G. (2012). *Intelligent Data Engineering and Automated Learning–IDEAL 2012: 13th International Conference, Natal, Brazil, August 29-31, 2012, Proceedings*, volume 7435. Springer.

[You et al., 2004] You, S.-J., Yang, D., and Eastman, C. (2004). Relational DB implementation of STEP based product model. In *CIB World Building Congress 2004*.

[Younis et al., 2016] Younis, R., Rehan, R., Unger, A. J. A., Yu, S., and Knight, M. A. (2016). Forecasting the unit price of water and wastewater pipelines capital works and estimating contractors' markup. *Journal of Cost Analysis and Parametrics*, 9(1):46--68.

[Zanzi, 2013] Zanzi, A. (2013). *Data quality evaluation through data quality rules and data provenance.* Doctoral thesis, Università degli Studi dell'Insubria.

[Zhang, 2004] Zhang, H. (2004). The optimality of naive bayes. *Aa*, 1(2):3.

[Zhang and El-Gohary, 2016] Zhang, J. and El-Gohary, N. M. (2016). Semantic nlp-based information extraction from construction regulatory documents for automated compliance checking. *Journal of Computing in Civil Engineering*, 30(2):04015014.

[Zhao and Liu, 2008] Zhao, W. and Liu, J. (2008). Owl/swrl representation methodology for express-driven product information model: part i. implementation methodology. *Computers in Industry*, 59:580--589.

[Zhou and El-Gohary, 2016] Zhou, P. and El-Gohary, N. (2016). Ontology-based multilabel text classification of construction regulatory documents. *Journal of Computing in Civil Engineering*, 30(4):04015058.

[Zhou et al., 2016] Zhou, Z., Goh, Y. M., and Shen, L. (2016). Overview and analysis of ontology studies supporting development of the construction industry. *Journal of Computing in Civil Engineering*, 30(6):04016026.

# Appendix A

# APPENDICES

## A.1 OCR and related issues

The flowchart depicted in Figure A.1 outlines the procedure for preparing scanned images of tables for Optical Character Recognition (OCR) detection and conversion. The process can be summarized as follows:

1. **Receive a Contract:** The process begins with receiving a contract other than in table format.
2. **Check Image/Table Format:** The system checks if the PDF or image is of a table and not in a scanned format.
3. **Evaluate Brightness and Contrast:** If the image is a scanned table, its brightness and contrast are assessed. If acceptable, it moves to the OCR processing phase.
4. **Quality Assessment:** If the image is not scanned, it is considered that quality degradation has not occurred, and the system proceeds to process the table using OCR software.
5. **Brightness and Contrast Adjustment:** If the image requires adjustments, the corresponding routines for brightness and contrast are called.
6. **Correct Skewness:** If the table image is skewed, the corresponding routines to correct it are invoked.
7. **Final Processing:** Once all the above steps are performed, the image is ready for the ABBYY software and can be exported to a table with minimal errors.

This process ensures that the scanned images of tables are in the appropriate format and quality for further OCR detection and conversion, contributing to the efficiency and accuracy of the WaterIAM system.

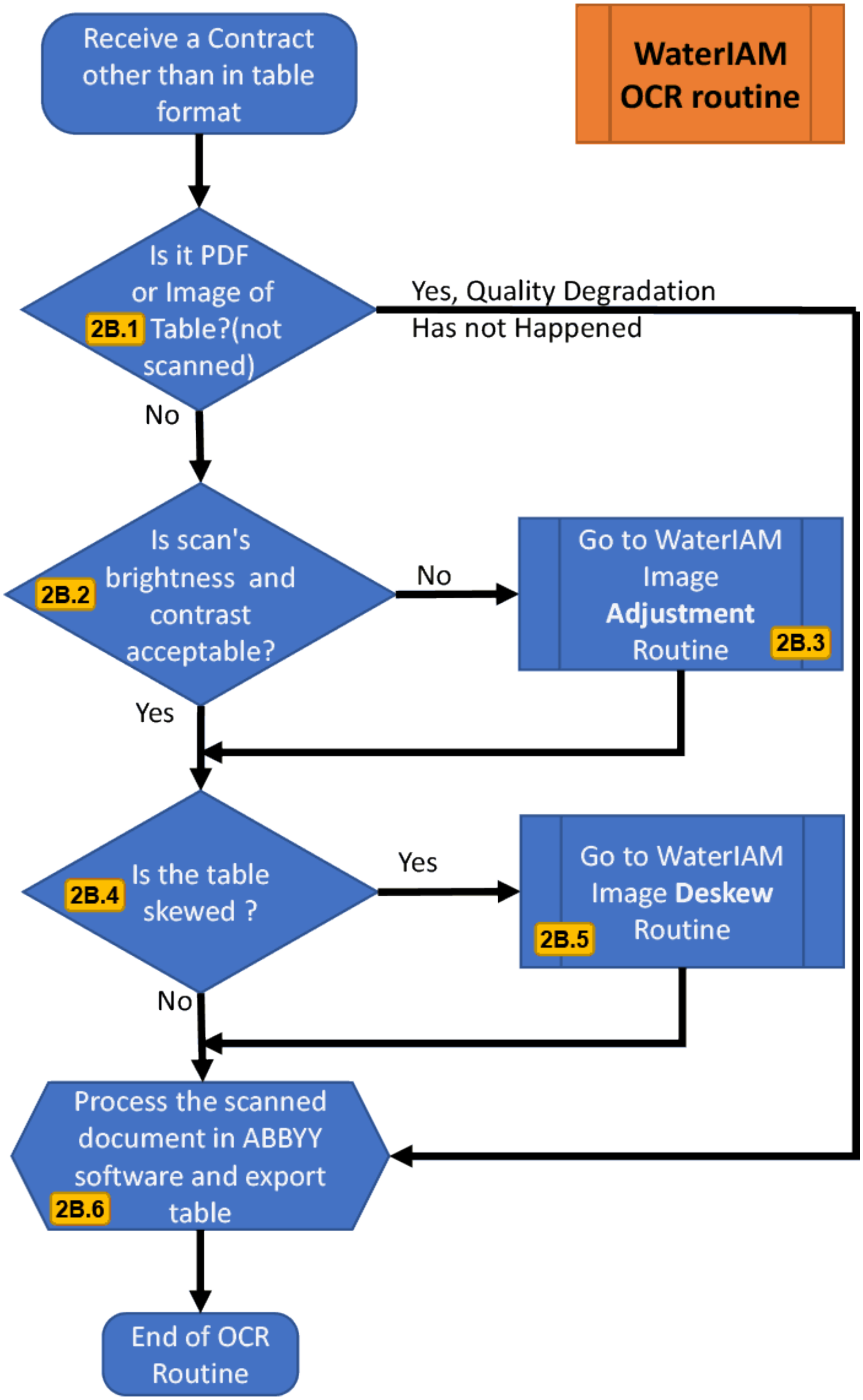



Figure A.1: Flowchart of using a hard copy contract for the WaterIAM system and passing through the OCR check routine.

## A.2 Main word-frequency table

The main word-frequency table is a comprehensive lexicon representing key terms in the field of watermain and sanitary sewer systems capital works. Due to the extensive nature of this table, spanning four pages, it is included in its entirety in this Appendix. A description and a concise one-page sample are provided in Section 2.2.2 on Page 52 of Chapter Two. The following tables represent the full details, compiled from approximately three hundred tender documents.

| Column1 | Frq1 | Column2 | Frq2 | Column3 | Frq3 | Column4 | Frq4 | Column5 | Frq5 | Column6 | Frq6 | Column7 | Frq7 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| mm | 2822 | sanitary | 869 | 0 | 734 | number | 731 | manhole | 675 | remove | 644 | road | 622 |
| and | 548 | curb | 546 | driveway | 541 | supply | 503 | + | 483 | install | 480 | asphalt | 477 |
| concrete | 471 | 600 | 461 | section | 456 | construct | 454 | exist | 440 | sewer | 435 | sidewalk | 433 |
| 100 | 416 | include | 410 | 50 | 403 | connect | 400 | excavate | 397 | base | 396 | catchbasin | 391 |
| watermain | 391 | pipe | 381 | 150 | 377 | depth | 355 | material | 350 | grade | 349 | storm | 348 |
| valve | 348 | place | 342 | st | 342 | repair | 335 | box | 334 | granular | 334 | cover | 330 |
| provision | 321 | diameter | 317 | 300 | 313 | type | 310 | stone | 308 | or | 304 | water | 301 |
| service | 298 | boulevard | 294 | to | 294 | test | 293 | backfill | 291 | hole | 290 | traffic | 290 |
| 310 | 289 | all | 289 | 200 | 286 | restore | 281 | dispose | 280 | for | 280 | site | 278 |
| adjust | 277 | cut | 277 | maintain | 270 | bed | 267 | lead | 259 | of | 259 | control | 256 |
| line | 255 | hydrant | 253 | as | 251 | in | 250 | new | 248 | trench | 248 | stop | 247 |
| tee | 246 | topsoil | 244 | total | 244 | street | 239 | gutter | 237 | 250 | 233 | sod | 233 |
| opsd | 231 | at | 230 | infrastructure | 230 | sign | 227 | width | 225 | precast | 219 | bond | 216 |
| 10 | 215 | on | 212 | application | 211 | replace | 210 | with | 203 | pvc | 202 | tree | 201 |
| main | 200 | clean | 198 | interlock | 196 | direct | 193 | require | 193 | up | 193 | by | 191 |
| 20 | 189 | commercial | 189 | complete | 189 | item | 186 | contingency | 184 | roadway | 184 | 450 | 181 |
| calcium | 181 | the | 181 | 12 | 180 | 15 | 179 | fill | 177 | hl8 | 177 | chloride | 175 |
| frame | 175 | class | 174 | engineer | 173 | from | 170 | mix | 169 | appurtenance | 168 | clear | 167 |
| mill | 167 | nfs | 167 | single | 167 | thickness | 167 | hl3 | 165 | 705 | 164 | 75 | 163 |
| protect | 163 | grate | 162 | temporary | 161 | 40 | 160 | per | 160 | location | 159 | any | 157 |
| private | 156 | general | 154 | pavement | 154 | sawcut | 150 | 1200 | 149 | approve | 149 | dwg | 149 |
| abandon | 145 | hot | 144 | 25 | 143 | 701 | 143 | an | 143 | pave | 142 | additional | 140 |
| plan | 140 | ramp | 140 | 35 | 138 | apply | 137 | reconnect | 137 | run | 136 | subdrain | 135 |
| subexcavate | 135 | draw | 134 | placement | 134 | ave | 133 | copper | 133 | verify | 133 | wall | 133 |
| fence | 132 | thick | 132 | lateral | 131 | salvage | 131 | residence | 130 | wire | 130 | size | 129 |
| area | 128 | cross | 127 | structure | 127 | mpa | 126 | crusher | 125 | pole | 125 | 19 | 124 |
| pre | 124 | filter | 123 | property | 123 | layout | 122 | barrier | 118 | be | 118 | coat | 118 |
| hs | 118 | open | 118 | allowance | 117 | anode | 117 | fit | 117 | ordinary | 117 | mark | 116 |
| not | 115 | management | 114 | out | 114 | deep | 113 | surface | 113 | cloth | 110 | joint | 110 |
| survey | 110 | use | 110 | sta | 109 | cap | 107 | leave | 107 | mesh | 107 | perforate | 107 |
| purpose | 106 | tangent | 106 | white | 106 | rock | 105 | flush | 104 | sdr | 104 | inspect | 103 |
| standard | 103 | 11 | 102 | iron | 102 | dr18 | 101 | grind | 101 | wide | 101 | 375 | 100 |
| restraint | 100 | unshrinkable | 100 | chamber | 99 | bend | 98 | double | 98 | signal | 98 | 400 | 97 |
| basin | 97 | pressure | 96 | show | 95 | sleeve | 95 | tack | 94 | top | 94 | high | 92 |
| large | 92 | 110 | 91 | condition | 91 | disinfect | 90 | set | 90 | straight | 90 | curve | 89 |
| extension | 89 | grub | 89 | hoe | 89 | hydro | 89 | ram | 89 | work | 89 | light | 88 |
| cable | 87 | reducer | 87 | shore | 86 | earth | 84 | price | 84 | solid | 84 | temp | 84 |
| brace | 83 | compact | 83 | length | 82 | pdc | 82 | plastic | 82 | culvert | 81 | bear | 80 |
| cast | 80 | mulch | 80 | taper | 80 | seed | 79 | anchor | 78 | cathodic | 78 | soft | 77 |
| 130 | 76 | duct | 76 | face | 76 | final | 76 | into | 76 | lift | 76 | off | 76 |
| old | 76 | unit | 76 | 125 | 75 | side | 75 | specification | 75 | steel | 75 | hl3f | 74 |
| record | 74 | rigid | 74 | tap | 74 | 1500 | 73 | 550 | 73 | bar | 73 | ductile | 73 |
| fine | 73 | rebuild | 73 | 24 | 72 | dzp | 72 | inlet | 72 | office | 72 | paint | 71 |
| 18 | 70 | break | 70 | insulation | 70 | thrust | 70 | build | 69 | down | 69 | piece | 69 |
| 1104 | 68 | catch | 68 | grout | 68 | rod | 68 | south | 68 | typical | 68 | zinc | 68 |
| 525 | 67 | ditch | 67 | limit | 67 | shut | 66 | dust | 65 | dr | 64 | sub | 64 |
| vacuum | 64 | 1105 | 63 | 38 | 63 | detail | 63 | leak | 63 | local | 63 | method | 63 |
| outline | 63 | package | 63 | polyethylene | 63 | retain | 63 | under | 63 | yellow | 63 | contract | 62 |
| end | 62 | delay | 61 | foot | 61 | pedestrian | 61 | information | 60 | progress | 60 | brick | 59 |
| compaction | 59 | permanent | 59 | video | 59 | 750 | 58 | butt | 58 | exercise | 58 | photography | 58 |
| plate | 58 | silt | 58 | trenchless | 58 | 30 | 57 | density | 57 | dr28 | 57 | gate | 57 |
| hand | 56 | minimum | 56 | north | 56 | plug | 56 | 1800 | 55 | 900 | 55 | assembly | 55 |
| import | 55 | reinforce | 55 | shrub | 55 | b01 | 54 | fix | 54 | 1110 | 53 | city | 53 |
| electrical | 53 | full | 53 | head | 53 | labour | 53 | lane | 53 | performance | 53 | riser | 53 |
| stub | 53 | than | 53 | circular | 52 | drain | 52 | fire | 52 | over | 52 | adjacent | 51 |
| walkway | 51 | ± | 51 | cm | 50 | continue | 50 | directional | 50 | offset | 50 | park | 50 |
| each | 49 | less | 49 | luminaire | 49 | necessary | 49 | note | 49 | only | 49 | swab | 49 |
| system | 49 | where | 49 | 675 | 48 | cold | 48 | east | 48 | finish | 48 | power | 48 |
| reuse | 48 | shoulder | 48 | via | 48 | exclude | 47 | field | 47 | long | 47 | nursery | 47 |
| see | 47 | black | 46 | project | 46 | setback | 46 | swale | 46 | wood | 46 | 65 | 45 |
| loop | 45 | median | 45 | rail | 45 | west | 45 | after | 44 | combine | 44 | conduit | 44 |
| pay | 44 | post | 44 | vegetation | 44 | cost | 43 | sweep | 43 | tv | 43 | way | 43 |
| 60 | 42 | 80 | 42 | elevate | 42 | hydraulic | 42 | mount | 42 | two | 42 | administrator | 41 |
| csa | 41 | landscape | 41 | 14 | 40 | bench | 40 | drop | 40 | equipment | 40 | have | 40 |
| low | 40 | treat | 40 | vi | 40 | brush | 39 | collector | 39 | dump | 39 | etc | 39 |

Table A.1: Word frequency table generated from all contracts (Refer to Section 2.2.2 on Page 52), Part 1 of 4.

| Column1 | Frq1 | Column2 | Frq2 | Column3 | Frq3 | Column4 | Frq4 | Column5 | Frq5 | Column6 | Frq6 | Column7 | Frq7 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| miscellaneous | 39 | rd | 39 | relocate | 39 | stump | 39 | subgrade | 39 | utility | 39 | 16 | 38 |
| arm | 38 | chlorinate | 38 | continuous | 38 | edge | 38 | step | 38 | waste | 38 | 50% | 37 |
| course | 36 | geotextile | 36 | junction | 36 | locate | 36 | regrade | 36 | normal | 35 | 500 | 34 |
| case | 34 | hst | 34 | replacement | 34 | stm | 34 | symbol | 34 | mechanical | 33 | municipality | 33 |
| pit | 33 | rc | 33 | reinstate | 33 | 90 | 32 | controller | 32 | device | 32 | divide | 32 |
| hdpe | 32 | queen | 32 | region | 32 | blow | 31 | foundation | 31 | native | 31 | plant | 31 |
| seal | 31 | back | 30 | couple | 30 | mobilize | 30 | pad | 30 | propose | 30 | recycle | 30 |
| safety | 30 | storz | 30 | support | 30 | table | 30 | awg | 29 | bury | 29 | duty | 29 |
| flake | 29 | great | 29 | kg | 29 | relocation | 29 | walk | 29 | 1050 | 28 | 28 | 28 |
| detector | 28 | drive | 28 | intersection | 28 | non | 28 | part | 28 | reinstall | 28 | shallow | 28 |
| tender | 28 | widen | 28 | 219 | 27 | around | 27 | bell | 27 | board | 27 | crosswalk | 27 |
| glass | 27 | inch | 27 | match | 27 | spot | 27 | treatment | 27 | turn | 27 | 160 | 26 |
| 225 | 26 | 840 | 26 | cl | 26 | expose | 26 | extra | 26 | guide | 26 | industrial | 26 |
| payment | 26 | riprap | 26 | 2010 | 25 | 825 | 25 | bead | 25 | below | 25 | cabinet | 25 |
| cone | 25 | crew | 25 | entrance | 25 | equal | 25 | gap | 25 | if | 25 | list | 25 |
| make | 25 | organic | 25 | outlet | 25 | push | 25 | reflectorize | 25 | solvent | 25 | vertical | 25 |
| 2000 | 24 | 22 | 24 | 299 | 24 | 599 | 24 | arterial | 24 | behind | 24 | bracket | 24 |
| button | 24 | cc | 24 | core | 24 | fuse | 24 | handhole | 24 | rap | 24 | various | 24 |
| 610 | 23 | arrow | 23 | cash | 23 | drill | 23 | green | 23 | gst | 23 | hl | 23 |
| horizontal | 23 | hour | 23 | lie | 23 | other | 23 | patch | 23 | prior | 23 | prune | 23 |
| raise | 23 | red | 23 | roadside | 23 | rope | 23 | rwu | 23 | stockpile | 23 | voltage | 23 |
| 2009 | 22 | access | 22 | cement | 22 | crescent | 22 | demobilize | 22 | invert | 22 | 305 | 21 |
| aggregate | 21 | basis | 21 | bollard | 21 | court | 21 | crush | 21 | dewater | 21 | do | 21 |
| emulsion | 21 | expansion | 21 | fish | 21 | flexible | 21 | level | 21 | mainline | 21 | previously | 21 |
| saddle | 21 | sand | 21 | station | 21 | strip | 21 | timber | 21 | 101 | 20 | approximate | 20 |
| bag | 20 | barricade | 20 | chain | 20 | durable | 20 | fabric | 20 | link | 20 | rebench | 20 |
| rip | 20 | sc | 20 | sw | 20 | bulkhead | 19 | crack | 19 | hl8hs | 19 | hp | 19 |
| independent | 19 | maple | 19 | monument | 19 | pour | 19 | pump | 19 | short | 19 | signboard | 19 |
| weave | 19 | yard | 19 | 100% | 18 | 180 | 18 | 2011 | 18 | 8501 | 18 | adaptor | 18 |
| block | 18 | connector | 18 | contaminate | 18 | csp | 18 | deflection | 18 | dowel | 18 | increase | 18 |
| lawn | 18 | lot | 18 | paver | 18 | qpr | 18 | rib | 18 | sample | 18 | store | 18 |
| Styrofoam | 18 | subbase | 18 | warn | 18 | 2012 | 17 | 68 | 17 | b21 | 17 | bike | 17 |
| bridge | 17 | dead | 17 | degree | 17 | extend | 17 | insurance | 17 | kit | 17 | lid | 17 |
| London | 17 | pedestal | 17 | remobilize | 17 | report | 17 | sac | 17 | series | 17 | st4 | 17 |
| 1000 | 16 | 102 | 16 | 21 | 16 | camera | 16 | change | 16 | chip | 16 | hedge | 16 |
| land | 16 | order | 16 | parge | 16 | profile | 16 | reinstatement | 16 | root | 16 | select | 16 |
| signage | 16 | slab | 16 | this | 16 | tube | 16 | viii | 16 | within | 16 | 105 | 15 |
| 26 | 15 | 350 | 15 | 912 | 15 | 975 | 15 | ac | 15 | aluminum | 15 | approach | 15 |
| bank | 15 | bare | 15 | bolt | 15 | bus | 15 | corrugate | 15 | design | 15 | directly | 15 |
| during | 15 | facility | 15 | forcemain | 15 | limestone | 15 | machine | 15 | meter | 15 | platform | 15 |
| point | 15 | provide | 15 | reprocess | 15 | special | 15 | square | 15 | unsuitable | 15 | add | 14 |
| common | 14 | cul | 14 | dress | 14 | flow | 14 | gran | 14 | hanger | 14 | hardware | 14 |
| island | 14 | manual | 14 | mat | 14 | photocell | 14 | separate | 14 | super | 14 | tactile | 14 |
| truck | 14 | watt | 14 | wortley | 14 | wrap | 14 | against | 13 | amount | 13 | apron | 13 |
| bay | 13 | centre | 13 | channel | 13 | corrugation | 13 | dash | 13 | dip | 13 | drainage | 13 |
| flag | 13 | gas | 13 | hl2 | 13 | improve | 13 | interconnect | 13 | kor | 13 | manufacture | 13 |
| Niagara | 13 | perimeter | 13 | poly | 13 | scratch | 13 | spring | 13 | sr | 13 | termination | 13 |
| tie | 13 | underground | 13 | 103 | 12 | 1350 | 12 | audible | 12 | beam | 12 | bypass | 12 |
| Clarke | 12 | co | 12 | completion | 12 | contractor | 12 | damage | 12 | delivery | 12 | determine | 12 |
| electric | 12 | guard | 12 | heavy | 12 | hf | 12 | mast | 12 | minor | 12 | modify | 12 |
| need | 12 | operation | 12 | overlie | 12 | partial | 12 | Philip | 12 | rack | 12 | requirement | 12 |
| ring | 12 | roadview | 12 | roadwork | 12 | rt | 12 | sdr35 | 12 | streetlight | 12 | subtotal | 12 |
| synertech | 12 | wellington | 12 | which | 12 | William | 12 | 104 | 11 | accidental | 11 | both | 11 |
| caliper | 11 | capital | 11 | debry | 11 | decommission | 11 | disconnect | 11 | dr35 | 11 | encase | 11 |
| ferry | 11 | highway | 11 | interceptor | 11 | member | 11 | nozzle | 11 | outside | 11 | overflow | 11 |
| police | 11 | polypropylene | 11 | prefabricate | 11 | reposition | 11 | rodent | 11 | specify | 11 | Stanley | 11 |
| steamer | 11 | time | 11 | trim | 11 | weld | 11 | wheelchair | 11 | armourstone | 10 | avg | 10 |
| awwa | 10 | backboard | 10 | backflow | 10 | basket | 10 | blowoff | 10 | caution | 10 | cctv | 10 |
| certificate | 10 | conductor | 10 | date | 10 | deliver | 10 | distribution | 10 | dogwood | 10 | except | 10 |
| fluorescent | 10 | follow | 10 | form | 10 | front | 10 | gabion | 10 | handle | 10 | height | 10 |
| illumination | 10 | insulate | 10 | medium | 10 | modification | 10 | mountable | 10 | pathway | 10 | pile | 10 |
| polara | 10 | polymer | 10 | pond | 10 | pool | 10 | preventer | 10 | roll | 10 | stopbar | 10 |
| sump | 10 | suppressant | 10 | tall | 10 | TERRAFIX | 10 | trailer | 10 | trap | 10 | winter | 10 |
| 235 | 9 | armour | 9 | asbestos | 9 | Astrobrac | 9 | band | 9 | before | 9 | between | 9 |
| binder | 9 | book | 9 | central | 9 | chainlink | 9 | check | 9 | close | 9 | Cogeco | 9 |

Table A.1, continued: Word Frequency table generated from all contracts (Refer to Section 2.2.2 on Page 52) Part 2 of 4.

| Column1 | Frq1 | Column2 | Frq2 | Column3 | Frq3 | Column4 | Frq4 | Column5 | Frq5 | Column6 | Frq6 | Column7 | Frq7 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| colour | 9 | con | 9 | credit | 9 | dechlorinate | 9 | deck | 9 | decorative | 9 | detect | 9 |
| emergency | 9 | every | 9 | fabricate | 9 | factory | 9 | fibre | 9 | flower | 9 | garden | 9 |
| grass | 9 | guy | 9 | July | 9 | lieu | 9 | Marley | 9 | npei | 9 | Ontario | 9 |
| Opticom | 9 | PEX | 9 | pick | 9 | playground | 9 | reconstruct | 9 | schedule | 9 | sdr28 | 9 |
| sealant | 9 | waterproof | 9 | wooden | 9 | $ | 8 | above | 8 | across | 8 | advance | 8 |
| aerial | 8 | appropriate | 8 | average | 8 | barrel | 8 | beyond | 8 | cl150 | 8 | conflict | 8 |
| creek | 8 | dig | 8 | drip | 8 | epoxy | 8 | estimate | 8 | external | 8 | fixture | 8 |
| good | 8 | guild | 8 | hl3hs | 8 | indicate | 8 | jack | 8 | liner | 8 | load | 8 |
| messenger | 8 | mini | 8 | one | 8 | pc | 8 | photo | 8 | pushbutton | 8 | quantity | 8 |
| railway | 8 | receptacle | 8 | return | 8 | rout | 8 | sanitarysewer | 8 | shall | 8 | space | 8 |
| strength | 8 | tape | 8 | tax | 8 | that | 8 | thread | 8 | tight | 8 | tracer | 8 |
| urban | 8 | weekend | 8 | also | 7 | auger | 7 | avoid | 7 | berm | 7 | bicycle | 7 |
| blend | 7 | brown | 7 | Bruce | 7 | bush | 7 | cedar | 7 | chemical | 7 | dam | 7 |
| day | 7 | disturb | 7 | due | 7 | embed | 7 | erection | 7 | erosion | 7 | excess | 7 |
| feature | 7 | fee | 7 | flex | 7 | fourth | 7 | gasmain | 7 | globe | 7 | handrail | 7 |
| house | 7 | implement | 7 | inc | 7 | interval | 7 | key | 7 | late | 7 | manager | 7 |
| metre | 7 | name | 7 | near | 7 | notice | 7 | officer | 7 | opposite | 7 | orange | 7 |
| overhead | 7 | picnic | 7 | play | 7 | position | 7 | revise | 7 | revision | 7 | sack | 7 |
| same | 7 | sectional | 7 | sheet | 7 | shrink | 7 | slope | 7 | small | 7 | soil | 7 |
| stage | 7 | stem | 7 | subsurface | 7 | sufficient | 7 | tennis | 7 | track | 7 | twin | 7 |
| upon | 7 | visit | 7 | zone | 7 | acer | 6 | air | 6 | associate | 6 | banner | 6 |
| bituminous | 6 | blast | 6 | bottom | 6 | bucket | 6 | car | 6 | castiron | 6 | chimney | 6 |
| cippsr | 6 | cl65d | 6 | contain | 6 | detour | 6 | duplex | 6 | dye | 6 | early | 6 |
| fernco | 6 | find | 6 | flat | 6 | ft | 6 | galvanize | 6 | gravel | 6 | haul | 6 |
| identification | 6 | insert | 6 | investigate | 6 | ladder | 6 | latch | 6 | Lawrence | 6 | live | 6 |
| major | 6 | marker | 6 | maximum | 6 | may | 6 | oak | 6 | operate | 6 | parapet | 6 |
| pass | 6 | portage | 6 | portion | 6 | preparation | 6 | program | 6 | provincial | 6 | put | 6 |
| reference | 6 | river | 6 | rogers | 6 | sealer | 6 | semi | 6 | shademaster | 6 | specie | 6 |
| stair | 6 | stamp | 6 | sugar | 6 | surround | 6 | synertec | 6 | temperance | 6 | through | 6 |
| tulip | 6 | unknown | 6 | upper | 6 | vehicle | 6 | analysis | 5 | applicable | 5 | assume | 5 |
| august | 5 | away | 5 | basketball | 5 | big | 5 | blanket | 5 | category | 5 | celtis | 5 |
| charge | 5 | clair | 5 | clay | 5 | coarse | 5 | crysler | 5 | description | 5 | elliptical | 5 |
| emulsify | 5 | grey | 5 | hatch | 5 | hf150s | 5 | HVAC | 5 | hydrostatic | 5 | incentive | 5 |
| it | 5 | jam | 5 | lamacoid | 5 | layer | 5 | lirodendron | 5 | magnesium | 5 | max | 5 |
| metal | 5 | moisture | 5 | monitor | 5 | november | 5 | occupancy | 5 | panel | 5 | path | 5 |
| patio | 5 | percentage | 5 | phase | 5 | plus | 5 | pound | 5 | premium | 5 | prepare | 5 |
| prestress | 5 | prop | 5 | protrude | 5 | rebar | 5 | result | 5 | round | 5 | rural | 5 |
| second | 5 | september | 5 | shape | 5 | snow | 5 | spillway | 5 | split | 5 | straw | 5 |
| stripe | 5 | superpave | 5 | tool | 5 | transition | 5 | tulipifera | 5 | vehicular | 5 | wash | 5 |
| well | 5 | % | 4 | abrasive | 4 | along | 4 | arch | 4 | attach | 4 | bale | 4 |
| bid | 4 | blade | 4 | boot | 4 | but | 4 | cock | 4 | coir | 4 | collection | 4 |
| company | 4 | compliance | 4 | compound | 4 | countdown | 4 | crane | 4 | dixon | 4 | dr25 | 4 |
| drummond | 4 | dry | 4 | durostar | 4 | easement | 4 | eastwood | 4 | eccentric | 4 | echo | 4 |
| enbridge | 4 | ent | 4 | excavator | 4 | exploratory | 4 | extract | 4 | fall | 4 | filler | 4 |
| flagstone | 4 | future | 4 | gasket | 4 | generator | 4 | go | 4 | grand | 4 | grosvenor | 4 |
| hackberry | 4 | hard | 4 | holder | 4 | hump | 4 | hydroseed | 4 | index | 4 | investigation | 4 |
| invoice | 4 | kinsman | 4 | landfill | 4 | liquid | 4 | loader | 4 | log | 4 | message | 4 |
| mortar | 4 | multi | 4 | navigator | 4 | net | 4 | norway | 4 | obliterate | 4 | october | 4 |
| oil | 4 | operator | 4 | orifice | 4 | orlando | 4 | overland | 4 | permit | 4 | pi | 4 |
| pine | 4 | plane | 4 | pleasant | 4 | plunge | 4 | polyester | 4 | portable | 4 | potable | 4 |
| preserve | 4 | princess | 4 | priority | 4 | proctor | 4 | public | 4 | pull | 4 | railroad | 4 |
| receive | 4 | refer | 4 | regent | 4 | reserve | 4 | retap | 4 | rubrum | 4 | saint | 4 |
| scaffold | 4 | school | 4 | screen | 4 | setup | 4 | shear | 4 | silver | 4 | smooth | 4 |
| spare | 4 | speed | 4 | spruce | 4 | stake | 4 | standby | 4 | stormwater | 4 | streetline | 4 |
| stuart | 4 | suitable | 4 | supervision | 4 | surplus | 4 | switch | 4 | third | 4 | three | 4 |
| tonnes | 4 | trail | 4 | transport | 4 | trash | 4 | trunk | 4 | update | 4 | upgrade | 4 |
| valley | 4 | variable | 4 | vary | 4 | warranty | 4 | waterline | 4 | waterway | 4 | when | 4 |
| while | 4 | active | 3 | adhesive | 3 | adjuster | 3 | ahead | 3 | alba | 3 | allow | 3 |
| alloy | 3 | apart | 3 | approval | 3 | april | 3 | areg | 3 | ash | 3 | aspen | 3 |
| authority | 3 | autocad | 3 | autumn | 3 | basement | 3 | beech | 3 | blacktop | 3 | boss | 3 |
| branch | 3 | brute | 3 | btuc | 3 | can | 3 | chevron | 3 | cl100d | 3 | cl51 | 3 |
| clamp | 3 | class150 | 3 | closure | 3 | coa | 3 | collect | 3 | confirm | 3 | conn | 3 |
| coordination | 3 | corporation | 3 | curbstop | 3 | curt | 3 | demolition | 3 | difference | 3 | discolor | 3 |
| disincentive | 3 | disk | 3 | documentation | 3 | doghouse | 3 | durastar | 3 | eastern | 3 | egg | 3 |
| elmwood | 3 | enclosure | 3 | energy | 3 | entry | 3 | equivalent | 3 | erect | 3 | est | 3 |

Table A.1, continued: Word Frequency table generated from all contracts (Refer to Section 2.2.2 on Page 52) Part 3 of 4.

| Column1 | Frq1 | Column2 | Frq2 | Column3 | Frq3 | Column4 | Frq4 | Column5 | Frq5 | Column6 | Frq6 | Column7 | Frq7 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| execute | 3 | family | 3 | fertilize | 3 | first | 3 | flagman | 3 | forsythe | 3 | furnish | 3 |
| gale | 3 | gallon | 3 | garner | 3 | gauge | 3 | gladstone | 3 | glauca | 3 | glue | 3 |
| gravity | 3 | guardrail | 3 | hammer | 3 | handicap | 3 | harmonize | 3 | hit | 3 | holdback | 3 |
| hyd | 3 | incidental | 3 | irrigation | 3 | keller | 3 | labourer | 3 | larch | 3 | larix | 3 |
| leader | 3 | liability | 3 | library | 3 | locust | 3 | lump | 3 | mailbox | 3 | mall | 3 |
| masonry | 3 | miller | 3 | mixture | 3 | mobile | 3 | mod | 3 | monolithic | 3 | mto | 3 |
| narrow | 3 | night | 3 | obtain | 3 | octagonal | 3 | operational | 3 | opss | 3 | original | 3 |
| percent | 3 | picea | 3 | pillar | 3 | pinus | 3 | planter | 3 | plywood | 3 | polymeric | 3 |
| populous | 3 | positive | 3 | powder | 3 | preliminary | 3 | previous | 3 | proceed | 3 | proper | 3 |
| react | 3 | rehabilitate | 3 | release | 3 | reset | 3 | right | 3 | saccharum | 3 | sale | 3 |
| salix | 3 | scada | 3 | sceptre | 3 | sediment | 3 | segmental | 3 | settle | 3 | sewage | 3 |
| shelter | 3 | shoe | 3 | shop | 3 | should | 3 | sieve | 3 | speer | 3 | spin | 3 |
| stall | 3 | std | 3 | stella | 3 | stormceptor | 3 | strobus | 3 | subdivision | 3 | subrain | 3 |
| subsection | 3 | substructure | 3 | summary | 3 | sydenham | 3 | tandem | 3 | teck | 3 | terra | 3 |
| then | 3 | tile | 3 | trace | 3 | transportation | 3 | tremble | 3 | trial | 3 | tunnel | 3 |
| unlocated | 3 | upstream | 3 | upto | 3 | volume | 3 | weatherproof | 3 | western | 3 | will | 3 |
| willow | 3 | year | 3 | zebra | 3 | accommodate | 2 | acess | 2 | acoustic | 2 | actual | 2 |
| actuator | 2 | addition | 2 | adhere | 2 | advisory | 2 | alder | 2 | alnus | 2 | amelanchier | 2 |
| american | 2 | amp | 2 | authorize | 2 | available | 2 | axle | 2 | backhoe | 2 | beak | 2 |
| bebbiana | 2 | become | 2 | bevel | 2 | biamonte | 2 | bicolor | 2 | bill | 2 | birdge | 2 |
| blaze | 2 | boil | 2 | boy | 2 | bring | 2 | bulk | 2 | burn | 2 | burr | 2 |
| burst | 2 | business | 2 | butterfly | 2 | calm | 2 | capacity | 2 | carry | 2 | cat | 2 |
| catalpa | 2 | cathcart | 2 | cell | 2 | cemetery | 2 | chad | 2 | chair | 2 | changeable | 2 |
| chase | 2 | cherry | 2 | chicane | 2 | choke | 2 | circuit | 2 | cl52 | 2 | clearly | 2 |
| clenray | 2 | cleveland | 2 | cobble | 2 | comb | 2 | combar | 2 | come | 2 | comer | 2 |
| comission | 2 | comm | 2 | commission | 2 | composite | 2 | compressive | 2 | compressor | 2 | containment | 2 |
| coordinate | 2 | corner | 2 | coupler | 2 | cumulus | 2 | cushion | 2 | cylinder | 2 | daylily | 2 |
| dear | 2 | delete | 2 | deposition | 2 | dept | 2 | detectable | 2 | detention | 2 | deteriorate | 2 |
| diesel | 2 | dimension | 2 | director | 2 | dissipate | 2 | distribute | 2 | division | 2 | dla | 2 |
| door | 2 | downtime | 2 | dr14 | 2 | dr21 | 2 | drum | 2 | dsa | 2 | dura | 2 |
| dynamic | 2 | eagle | 2 | edition | 2 | elder | 2 | electrode | 2 | encounter | 2 | english | 2 |
| entire | 2 | envelope | 2 | environmental | 2 | escalation | 2 | excessive | 2 | extruder | 2 | fagus | 2 |
| FALSE | 2 | flap | 2 | flood | 2 | force | 2 | foreman | 2 | formliners | 2 | formwork | 2 |
| forward | 2 | fraxinus | 2 | french | 2 | frontier | 2 | fuel | 2 | furniture | 2 | geogrid | 2 |
| geotechnical | 2 | girl | 2 | goal | 2 | gobain | 2 | golden | 2 | goldflame | 2 | gracefield | 2 |
| gradation | 2 | graffito | 2 | grandifolia | 2 | grassy | 2 | gray | 2 | grit | 2 | grubbind | 2 |
| hale | 2 | handwell | 2 | handwork | 2 | health | 2 | hewitt | 2 | hickory | 2 | hold | 2 |
| holophane | 2 | honey | 2 | hose | 2 | hydrocarbon | 2 | hydrovac | 2 | impact | 2 | individual | 2 |
| ingleside | 2 | injury | 2 | inn | 2 | insertion | 2 | inside | 2 | inspector | 2 | instrumentation | 2 |
| jackhammer | 2 | job | 2 | just | 2 | keep | 2 | label | 2 | laevis | 2 | larcinia | 2 |
| largo | 2 | lat | 2 | lean | 2 | lemon | 2 | lentago | 2 | lexington | 2 | lightweight | 2 |
| lily | 2 | linden | 2 | loose | 2 | lucida | 2 | macrocarpa | 2 | manufacturer | 2 | marshall | 2 |
| meadow | 2 | measure | 2 | measurement | 2 | mech | 2 | mechanic | 2 | mechnical | 2 | membrane | 2 |
| mesic | 2 | mewburn | 2 | midblock | 2 | mismark | 2 | model | 2 | moduloc | 2 | modulock | 2 |
| month | 2 | montrose | 2 | more | 2 | motor | 2 | mud | 2 | nameplate | 2 | nannyberry | 2 |
| northern | 2 | nuclear | 2 | onsite | 2 | optic | 2 | option | 2 | ornamental | 2 | osier | 2 |
| otherwise | 2 | oval | 2 | ovata | 2 | paddock | 2 | page | 2 | pair | 2 | parkway | 2 |
| pattern | 2 | pear | 2 | penalty | 2 | perform | 2 | person | 2 | photograph | 2 | pickup | 2 |
| piezometer | 2 | plain | 2 | plaque | 2 | plumb | 2 | possible | 2 | pot | 2 | preconstruction | 2 |
| probe | 2 | procedure | 2 | production | 2 | proofroll | 2 | protective | 2 | prunus | 2 | publication | 2 |
| quality | 2 | racemosa | 2 | radius | 2 | raspberry | 2 | rate | 2 | read | 2 | rear | 2 |
| reassemble | 2 | recap | 2 | receipt | 2 | reditch | 2 | regional | 2 | reimburse | 2 | relay | 2 |
| reline | 2 | remain | 2 | rental | 2 | request | 2 | reverse | 2 | rhus | 2 | riffle | 2 |
| rot | 2 | route | 2 | royal | 2 | rubble | 2 | rubra | 2 | rubus | 2 | rugosa | 2 |
| sambucus | 2 | scenic | 2 | seat | 2 | secondary | 2 | selectra | 2 | self | 2 | separator | 2 |
| serviceberry | 2 | settlement | 2 | shagbark | 2 | sheer | 2 | shine | 2 | shingle | 2 | sidetap | 2 |
| silane | 2 | sluice | 2 | sock | 2 | sound | 2 | southgate | 2 | specialty | 2 | spirea | 2 |
| spoil | 2 | stainless | 2 | stolonifera | 2 | strand | 2 | streetprint | 2 | strom | 2 | submersible | 2 |
| submit | 2 | substantial | 2 | substrate | 2 | sum | 2 | sumac | 2 | sunburst | 2 | supplemental | 2 |
| surcharge | 2 | suspend | 2 | swamp | 2 | sweet | 2 | tab | 2 | take | 2 | tapestry | 2 |
| tecumseh | 2 | terminus | 2 | thermal | 2 | toe | 2 | torque | 2 | train | 2 | transformer | 2 |
| transit | 2 | truack | 2 | turfstone | 2 | twist | 2 | typhina | 2 | ultimate | 2 | unacceptable | 2 |
| underneath | 2 | union | 2 | unmark | 2 | unused | 2 | upland | 2 | valour | 2 | value | 2 |
| vent | 2 | vibration | 2 | vibrunum | 2 | viginiana | 2 | vocomp | 2 | warren | 2 | washer | 2 |
| watercourse | 2 | watertight | 2 | weatherhead | 2 | wetland | 2 | wheel | 2 | wier | 2 | without | 2 |

Table A.1, continued: Word Frequency table generated from all contracts (Refer to Section 2.2.2 on Page 52) Part 4 of 4.

## A.3 Ontology definition and implementation

### A.3.1 Data Preprocessing and Word Tokenization

The process of preparing the data for subsequent analyses comprises several essential stages, including data type conversion, cleaning, handling special fields, and word tokenization. The following sections delineate the processes and mechanisms employed for data preprocessing and tokenization.

**Data Type Conversion and Checking** Each field of the raw table data is inspected and converted to its required data type if necessary. This process ensures the accuracy and consistency of the data types throughout the dataset, enabling correct and efficient analysis.

**Description Cleaning and Splitting** For the Description field of each entry in StructTbl, the code performs text cleaning and word splitting operations. This operation ensures that the description field is in a suitable format for subsequent analyses, aiding in feature extraction and improving the quality of data-driven insights.

**Multiple Field Handling** There is an optional field Multiple, which if present, triggers the creation of additional copies of the current record. The 'OrgSection' field is randomly updated in these duplicated entries, allowing for the generation of varied and comprehensive training data for the model.

**Word Exclusion and Substitution** The ontology is designed to standardize data and preclude errors to enhance the subsequent analyses. Specific words, predominantly comprising conjunctions, prepositions, and numerals, are entirely removed. These are outlined in the Ont_RemWords list. Similarly, certain words are replaced with alternatives to standardize synonymous terms and correct common misspellings or abbreviations.

**Character Exclusion and Substitution** Certain characters for removal from words are specified, predominantly pertaining to punctuation. Certain characters are also replaced with alternatives to standardize the use of particular punctuation marks and symbols.

**Root Extraction** Words are processed through the WTM_NLP_RootFinder___v4p0 function, which reduces words to their base or root forms. This function employs common Natural Language Processing techniques to either lemmatize or stem words.

**Unit Separation** Strings that contain unit need this essential rule for cleanup. This rule plays a vital role in the data preprocessing stage. It takes an input string and a list of units, sorts the units based on their length in descending order, and separates the input string into multiple sub-strings based on predefined separators such as ' ',';','-',':','/'. The rule checks for numbers and units in the sub-strings and if found, separates and arranges them properly. Each sub-string is further separated if it contains one of the units. The rule specifically handles several cases:

- When a unit is found at the beginning of the sub-string and is followed by a number (i.e. "mm275 pipe diameter" is replaced by "mm 275 pipe diameter").
- When a unit is found at the end of the sub-string and is preceded by a number (i.e. "43sqft" is replaced by "43 sqft").
- When a unit is found in the middle of a sub-string and is both preceded and followed by a number (i.e. "15PVC50mm pipe" is replaced by " 15 PVC 50 mm pipe").

The output of this rule is a list of processed strings with numbers and units separated.

**Project Program Availability** The program developed for this project, including all the tools and functions described in this section, is available online for download. Interested parties can access the repository at the following Git address: `https://github.com/mld-khaki/WaterIAM-Prj-MiladKhaki`.

**Special Case Handling and Dynamic Ontology Update** Conditions are in place to handle unique cases, such as removing a period at the end of a string with a significant number of numeric characters or single quotation marks at the beginning or end of a word. Additionally, the system incorporates a dynamic ontology update functionality through the **WTM_ONT_UpdateByWTMTable___v1p0** rule.

The **WTM_ONT_UpdateByWTMTable___v1p0** rule takes a mapping table, referred to as the *Input Table*, and a list of words, referred to as *Input Words*, as inputs. The purpose of this rule is to update the words in *Input Words* based on the mapping table. The specific rules followed by this rule are as follows:

1. The rule accepts two inputs: *Input Table*, a table or structured data that includes mapped unknown words, and *Input Words*, a list of words or strings.

2. If any of the elements in *Input Words* are not already in string format, they are converted to strings.

3. The rule iterates through each word in *Input Words*. For each word, it checks against all rows in the *MapUnknownTable* within the *Input Table*. The *MapUnknownTable* contains the list of the words that need to be updated.

4. If a word from *Input Words* is found as an entry in the *MapUnknownTable*, it is updated to the corresponding mapped word(s) found in the same row of *MapUnknownTable*. If multiple mapped words are found, they are concatenated and separated by a space to form a single string.

5. This process continues until all words in *InpWords* have been checked and possibly updated based on the mapping table.

6. The output, referred to as *OutWords*, is the list of updated words.

7. The following are a few example rows and rules of the MapUnknownTable:

   - Input Word: "chlonde", Corresponding Row in MapUnknownTable: "chlonde" is replaced by "chloride" (typo and OCR error correction).
   - Input Word: "chlorination", Corresponding Row in MapUnknownTable: "chlorination" is replaced by "chlorinate" (word consistency)
   - Input Word: "ci", Corresponding Row in MapUnknownTable: "ci" is replaced by "castiron" (abbreviation removal for consistency)
   - Input Word: "cicbmh", Corresponding Row in MapUnknownTable: "cicbmh" is replaced by "castiron catchbasin manhole", (abbreviation removal for consistency)
   - Input Word: "clcbs", Corresponding Row in MapUnknownTable: "clcbs" is replaced by "castiron catchbasin" (abbreviation removal for consistency) Input Word: "cliftonvale", Corresponding Row in MapUnknownTable: "cliftonvale" is replaced by "" (word specific to a place or street, no additional value, removal).
   - Input Word: "colborne", Corresponding Row in MapUnknownTable: "colborne" is replaced by "" (word specific to a place or street, no additional value, removal).
   - Input Word: "conc", Corresponding Row in MapUnknownTable: "conc" is replaced by "conrete" (abbreviation removal for consistency)

If multiple mapped words are associated with a single unknown word, they are joined together, separated by a space. The **WTM_ONT_UpdateByWTMTable___v1p0** rule enables the ontology to dynamically update based on a mapping table, thereby enhancing its adaptability.

**Separation of Numbers from Strings** Numerical values are separated from string data, reducing noise and enabling the model to focus on both numbers and text data.

**Removal of Subitem Numbers** The WTM_NLP_RemoveSubitemNumber___v1p0 function removes subitem numbers from a given word. It entails the exclusion of any numeric identifiers linked to an item or subitem in a list.

**Handling Quotation Marks and Leading Zeroes** Special care is taken to remove single quotation marks at the beginning or end of a word and to check if a word begins with 0. If so, and the word length is two or more characters, the leading 0 is removed.

**Period at the End of String** For words with more than four characters and containing more than three alphabetic characters that conclude with a period ('.'), the period is eliminated. The minimum four-letter length condition ensures that numbers remain untouched and only the punctuational character "." is removed from the strings.

**Tokenization Rules** Tokenizing text in natural language processing tasks is essential. It allows a program to understand different inflections of the same word as having the same root meaning. The general structure of each word is taken as input. A set of rules transform this word into its most simple form. These rules include the transformation of plural forms to singular, explicit rules for certain English words that do not follow regular spelling conventions, and specific handling for words ending in 'ly', 'ing', 'ment', 'ion', etc (i.e. "paving" and "pavement" are replaced by "pave").

**Record Sanity Check Rule** This rule verifies the presence and type of a specified field ('ItemStr') within a given record.

1. The rule accepts three inputs: 'Line', which represents the structure to be checked; 'ItemStr', indicating the field name to be examined within the structure; and 'DigStr', specifying the expected field type.

2. If the ‘ItemStr’ field does not exist in the ‘Line’ structure, the function returns ‘-1’ and throws an error with the message "Unacceptable."

3. When the ‘ItemStr’ field exists in the ‘Line’ structure, the function proceeds to validate if the field is non-empty and not a NaN value.

4. If the value of ‘DigStr’ is "digit" and the ‘ItemStr’ field in ‘Line’ is numeric, the function returns ‘true’.

5. Similarly, if the value of ‘DigStr’ is "string" and the ‘ItemStr’ field in ‘Line’ is of character or string type, the function returns ‘true’.

6. If none of the above conditions are met, indicating a mismatch in the field type, the function returns ‘false’.

7. Example usage of the rule: assert(WTM_UCI_CheckLineItems___v2p0(CurItemInfo, "FinalPrice," "digit") == true). In this example, the ‘WTM_UCI_CheckLineItems___v2p0’ function is called to check if the field named "FinalPrice" within the ‘CurItemInfo’ structure is numeric. The assertion confirms that the condition ‘WTM_UCI_CheckLineItems___v2p0( CurItemInfo, "FinalPrice," "digit")’ evaluates to ‘true’, ensuring that the "FinalPrice" field is indeed a numeric value in the ‘CurItemInfo’ structure.

### A.3.2 Standardizing Item Categories

The process of standardizing item categories covers a wide range of classifications, such as general items, roadwork, and various types of pipes and sewers, among others. String operations are used extensively to handle the various input forms, even managing unusual cases such as empty inputs, not string data types, or fall under a category data type. If an item's input does not contain a category, it remains unclassified and is labelled as "UNKNOWN" for future classification.

#### Function: WTM_ONT_ItemUpdater___v2p0

The function 'WTM_ONT_ItemUpdater___v2p0' plays a key role in this standardization process. It takes an input mapping table and an original item. The function then iterates through the table to find an interval where the original item fits. Specifically, the original item should be less than the upper limit and greater than or equal to the lower limit of an interval in the table. The function asserts that the original item fits into only one table interval. If the original item fits into multiple intervals, it raises an error. The item index is then set as the index of the interval that the original item fits into. Finally, the function updates the original item to be the upper limit of the interval from the input mapping table that it fits into.

#### Classification and Numerical Equivalents

Each specific string input corresponds to a standardized category type and is assigned a numerical equivalent. The classifications are:

- The category "General" covers various string inputs, such as "GNRL" and "GENERAL". It is assigned a numerical equivalent of 1.
- The "Road" category represents sections like "ROAD", "ROADWORKS", "ROADWORK", "ROADS", "REMOVAL", "REMOVALS", and more. Its numerical equivalent is 2. Special road categories like "RD_General", "RD_ConcSidewalk", and "RD_Manhole" are also considered preserved for future project expansions.
- The "Miscellaneous" category represents "MISC" and "MISCELLANEOUS" sections. As no further rule is defined for miscellaneous items, it has no numerical value and is set to NaN.

- The "Watermain" category covers sections like "WTMN", "WATERMAIN", "WATERMAINS", and corresponds to the number 4. Further, there are subcategories within "Watermain", each with specific string representations and numerical equivalents.

- The "SanitarySewer" category represents sections like "SNSW", "SANITARYSEWER", "SANITARY SEWER" , "SANITARY SEWERS" , and more. It corresponds to the number 5, and also has its own subcategories.

- The category "ProvisionalItem" includes "ChangeWorkOrder" and is represented by numerical equivalents of 3 and 10, respectively.

- The "StormSewer" category and its subcategories are currently not expanded further to keep the focus on watermain and sanitary sewers; however, this can be a consideration for future work.

### A.3.3 Watermain Item Surcharge Calculation

The function, 'WTM_UCI_ItemSurcharge___v5p0', calculates and adds a surcharge to each item's unit price to compute the item's unit cost. The function requires three parameters: 'ItemInfo', 'Costs', and 'Prms'.

**Total Price Calculation** , the total price for an item is computed by multiplying the unit price by the quantity ('ItemInfo.UnitPrice * ItemInfo.Quantity').

**Cost Summation for Specific Parts** : the summation of all costs is calculated which contains the standard sub-types: "WM_Pipe", "WM_Hydrant", "WM_Valve", "WM_Service", "StormSewer", and "Road" are summed up to compute "BCDEM_Cost". This cost is used in the following calculations.

- Surcharge Calculation Part 1: The first part of the surcharge is computed as a proportion of the item's total price relative to its contribution to the "BCDEM_Cost". This is done separately for General costs ("General") and Provisional Item costs ("ProvisionalItem"), and these two surcharge amounts are then added together. If "BCDEM_Cost"is zero, the surcharge is set to zero.
- Surcharge Calculation Part 2: The second part of the surcharge calculation is conditional based on the standard sub-part of the item. If it is "WM_Pipe', the surcharge is computed as the proportion of the item's total price to the cost of the Watermain Pipe "WM_Pipe", multiplied by the total cost of the Watermain Service "WM_Service", inflated by a factor dependent on the General and Provisional costs relative to "BCDEM_Cost" . On the other hand, if standard sub-part is "SS_Pipe", the surcharge is computed as the proportion of the item's total price to the cost of the StormSewer Pipe ("SS_Pipe"), multiplied by the total cost of the StormSewer Lateral ("SS_Lateral"), inflated by a factor dependent on the General and Provisional costs relative to "BCDEM_Cost". If standard sub-part is anything else, the second part of the surcharge is set to zero.

The total surcharge at this point is the sum of the two parts of the surcharge (Surcharge1 and Surcharge2), divided by the quantity of the item. If the item quantity is zero, the total surcharge is set to zero.

### A.3.4 Normalizing Attributes of Sanitary Sewer Items

This section details the normalization process of different attributes related to sanitary sewer items such as manholes and pipes.

#### Sanitary Sewer Manholes

The Diameter and Depth of ’SS Manhole’ items is determined based on their descriptions in the provided ’Item’. The pertinent rules extracted are as follows:

- Initialization, Set initial values for ‘OutDiameter’ and ‘OutDepth’. Initialize the ‘SizeValues’ array with ‘[1200, 1500, 1800]’, which are possible diameters.
- Misleader Removal, There are some known misleading terms that are found in the description of the items but are not relevant for determining the diameter and depth. These are removed from the ‘Item.Description’ using ‘regexprep’ and ‘strrep’ functions. These are usually the item number in the description of an item ( "a) sanitary manhole 1200 mm"to "sanitary manhole 1200 mm")
- Diameter Determination, if the string representation of the size value is contained in the item’s description it will be assigned to the corresponding field: ‘OutDiameter’. After setting the size value and the size string, the string equivalents will be removed from the item’s description.
- Depth Determination, using a regular expression the potential depths value of an item is captured in the item’s description. If no depth value is found, the field should be marked by ‘Could not find Item Depth!!’ and sets ‘OutDepth’ to ‘-1’.

#### Sanitary Sewer Pipes

Each "SS Pipe"item requires the Diameter and Type. Determination is based on their descriptions in the provided item.

- Pipe Type Determination, the item’s description should contain either "PVC" or "concrete". If one of these is found, the type is determined. If not, a flag is raised for the operator to check.

- Diameter Determination, it then initializes arrays 'SizeValues' and 'TypeValues' for potential diameters and corresponding types, respectively. Then, for each size value, it checks if the string representation of the size value is contained in the item's description. This process is performed multiple times until the known misleading strings are removed, and it is possible to ascertain whether the item description has the pipe dimensions.

### Observed Material Prices

The scaling process for observed material prices related to sanitary sewer items, such as manholes and pipes, is managed through the rule 'WTM_UCI_SS_Scale_Observed_MaterialPrice___v2p0'. This rule manipulates the observed material price based on restrictions, parameters, and data extracted from the sewer cost table. It accepts three arguments: 'ObservedMaterial', 'Item', and 'Prms'.

The application of the rule involves scaling the observed material price by extracting information from the sewer cost table, which is filtered and selected based on specific criteria. The 'ItemNum' and 'MaterialType' fields from 'Item', along with the 'PipeSize' and 'PipeType' fields from 'Prms', play a crucial role in this process. Additionally, the function utilizes a fixed 'Size' value of 375 during the filtering processes. The adjustment formula consists of a sequence of multiplication and division operations involving the 'MaterialCost' field from 'Input Item', 'RefTable', 'ItemAdj', and 'SelectionAdj'. In essence, this rule scales the observed material price according to the item size, material type, and sewer analysis parameters. As a result, this rule encompasses complex transformation regulations for the price of pipe and maintenance holes in sanitary sewer items.

# Glossary

**Black flag** This flag indicates that the record is a raw item with no change to the original information. The flag is not permanent, but it is advisable to keep this copy of the record untouched for a provenance check. An item is not readily usable for analysis. 41

**Brown flag** This flag indicates that the record is from a hard-copied source. The flag is permanent, and an item with this flag could be used for analysis (but this flag does not guarantee safety for analysis). 65

**DLC** Deep Learning Classification module 115, 116

**Green flag** This flag indicates that the record has standard-part and standard-sub-part that were predicted using the deep learning classifier. Therefore, this flag comes after the resolution of the issue in an item with a pink flag. It is a removable flag if the operator deems the classification wrong. An item with this flag is suitable for analysis. 55

**Meta-Data** while storing information in a database, the information that is the main content is called the data. In the case of WaterIAM, Data is the contracts and items that are stored in the main database. In contrast, any information not directly usable by the user and supporting the primary Data is called meta-data. In the WaterIAM database, the contract's additional information (consultant, dates, personnel) is considered meta-data. Additionally, the provenance records that indicate what error corrections are performed are also considered meta-data. 41

**OCR** Optical Character Recognition xi, xiv, 38, 166

**Pink flag** This flag indicates that the record does NOT have pre-determined standard-part and standard-sub-part. It is a removable flag (after using the deep learning classifier and determining the standard-part and standard-sub-part). An item with this flag is not suitable for analysis before the issue resolution. 55

**Red flag** A flag assigned to record with errors. This flag indicates that manual handling is required. It is removable after the error is removed, and the item cannot be analyzed before resolution. 55

**standard-part** The standardized Part of an item that is compatible with the classification performed via the DLC. These parts are: "General", "ProvisionalItem", "Miscellaneous", "Road", "SanitarySewer", "StormSewer", and "Watermain" xii, 35, 37--39, 43, 55, 56, 81, 83, 103, 115, 116, 132, 137, 139

**standard-PSP** The standardized part and standardized sub-part of an item together that are compatible with the classification performed via the DLC. The acceptable standard-psps are:"General_NoSubPart", "ProvisionalItem_NoSubPart", "Miscellaneous_NoSubPart", "Road_NoSubPart", "StormSewer_NoSubPart", "SanitarySewer_SS_Pipe", "SanitarySewer_SS_Manhole", "SanitarySewer_SS_Lateral", "Watermain_WM_Pipe", "Watermain_WM_Valve","Watermain_WM_Service", and "Watermain_WM_Hydrant" 115

**standard-sub-part** The standard subpart of an item that is compatible with the classification performed via the DLC. Within the contex of the current thesis and project, standard-sub-parts are defined only for "Watermain", and "Sanitary Sewer" Parts. The default standard-sub-part use for other standard-parts is "NoSubPart". The standard-sub-parts for the "Watermain" standard-part are: "WM_Pipe", "WM_Valve", "WM_Service", and "WM_Hydrant". Also the standard-sub-parts for the "Sanitary Sewer" standard-part are: "SS_Pipe", "SS_Manhole", and "SS_Lateral". 35, 38, 43, 55, 56, 81, 103, 116

**Violet flag** A flag assigned to records that the errors in them are removed and have provenance information. The item with a violet flag can be analyzed. 41, 65

**WaterIAM-Khaki** It is part of the WaterIAM project done by Milad Khaki and is in Milad's Ph.D. thesis scope. The proof of concept of WaterIAM-Khaki is a website that is developed by Milad Khaki and is based on C Sharp, Java, and MySQL implementation. 34, 37

**Yellow flag** A flag assigned to records with minor error(s). This flag indicates that the record should be used with caution. It is removable after the minor error(s) are removed. The usage of the record in analysis depends on the nature of the minor error. 39