

# Formalization of Knowledge in Construction: From Thesaurus to Ontology

Zarina Kabzhan<sup>1,\*</sup>, Alexandr Shakhnovich<sup>2</sup>, Tanatkan Abakanov<sup>3</sup>, Sergey Issayenko<sup>4</sup>

<sup>1</sup>Center for Science and Digitalization in Construction, Kazakh Research and Design Institute of Construction and Architecture, Republic of Kazakhstan

<sup>2</sup>Kazakh Research and Design Institute of Construction and Architecture, Republic of Kazakhstan

<sup>3</sup>Kazakh Leading Academy of Architecture and Civil Engineering, Republic of Kazakhstan

<sup>4</sup>QazCode LLP, 010010, 2 Kadyrgali Zhalayiri Str., Astana, Republic of Kazakhstan

*Received June 11, 2025; Revised October 21, 2025; Accepted November 30, 2025*

## Cite This Paper in the Following Citation Styles

(a): [1] Zarina Kabzhan, Alexandr Shakhnovich, Tanatkan Abakanov, Sergey Issayenko, "Formalization of Knowledge in Construction: From Thesaurus to Ontology," *Civil Engineering and Architecture*, Vol. 14, No. 1, pp. 262 - 273, 2026. DOI: 10.13189/cea.2026.140117.

(b): Zarina Kabzhan, Alexandr Shakhnovich, Tanatkan Abakanov, Sergey Issayenko (2026). *Formalization of Knowledge in Construction: From Thesaurus to Ontology*. *Civil Engineering and Architecture*, 14(1), 262 - 273. DOI: 10.13189/cea.2026.140117.

Copyright©2026 by authors, all rights reserved. Authors agree that this article remains permanently open access under the terms of the Creative Commons Attribution License 4.0 International License

**Abstract** This study aimed at substantiating the need to develop a unified ontological model in the construction industry of the Republic of Kazakhstan. Ontology allowed standardizing terminology in relation to concepts, improving data exchange and automating processes in construction. It is suggested to use a Thesaurus containing definitions of key terms in the construction industry as the foundation for developing an ontological model. The methodology provided several stages of ontology development, including defining the general structure, classes and properties, semantic relations between concepts, inference rules and ontology validation. The methodology also included the use of ontological conceptual modelling and natural language processing technologies to extract and systematize the terminology and concepts of the construction industry. The analysis of regulatory documents enabled the identification of key concepts and terms, forming the basis for the Thesaurus. The study employed a hybrid methodology that integrates established ontological engineering principles with advanced artificial intelligence techniques, including natural language processing for automated term extraction and machine learning for identifying semantic relationships. This approach ensured the development of a robust ontological framework specifically tailored to the multilingual context of Kazakhstan's construction industry. The findings demonstrate that the created semantic relations table significantly enhances the standardization of

regulatory documentation, improving clarity, eliminating terminological ambiguities, and minimizing errors in data exchange among construction stakeholders.

**Keywords** Thesaurus, Digitalization, Regulatory Framework, Semantic Relations, Terminology

---

## 1. Introduction

In today's information-driven society, the construction industry is a cornerstone of economic development and infrastructure modernization, playing a crucial role in designing, constructing, and operating the built environment. The effective management of information within this sector is paramount, as it directly impacts project efficiency, cost control, safety, and sustainability. A significant barrier to this efficiency is the pervasive issue of semantic heterogeneity, the use of the same terms with different meanings or different terms for the same concepts across various disciplines, organizations, and regulatory documents. This problem is acutely felt in the Republic of Kazakhstan, where the industry is undergoing rapid digital transformation driven by state initiatives and the need for integrated project delivery methods like BIM (Building Information Modelling) [1, 2].

Thesauri and ontologies are established as essential tools for overcoming these semantic barriers. They provide a formal, structured, and machine-readable framework for standardizing terminology and knowledge representation, ensuring uniformity and semantic compatibility in data exchange among different information systems and users [3]. As Lei et al. [4] emphasize, "Ontologies can become key components for the development of intelligent systems that support the management and coordination of construction projects. This is achieved through their ability to structure and represent domain knowledge, providing a semantic basis for interoperability and data integration." The benefits extend to enhancing the digital economy and overall industry efficiency, as a well-defined ontology offers a formal method for representing knowledge, facilitating seamless data exchange among various software applications [5].

Globally, the development of domain-specific ontologies is an active research area. For instance, research has explored automating ontology construction using heterogeneous systems of ontological design patterns to enrich ontologies from natural language texts [6]. Furthermore, successful ontological applications in related fields in Kazakhstan, such as the development of a water resource monitoring ontology that integrates diverse data sources for decision-making, demonstrate the viability and value of such approaches for national priorities.

Despite this recognized global importance and local potential, a significant knowledge gap exists regarding a comprehensive, standardized ontological framework specifically for the Kazakhstani construction industry. The current landscape is characterized by a fragmentation of terminology. Numerous normative and technical documents (GOSTs, SNIps, RK standards) often contain terms with subjective, ambiguous, or outdated definitions. This leads to inconsistencies in project documentation, disputes between stakeholders, errors in cost estimation, and ultimately hinders the adoption of digital technologies that rely on unambiguous data [7].

Recent analyses of the Kazakhstani context confirm this problem. Zhanarbekov and Batyrbekov [1] highlight the role of digital technologies in the industry's transformation but note that semantic inconsistencies remain a major impediment to their full implementation. Tulegenova and Sadykov [7] point to the need for a multilingual Thesaurus to improve data exchange, acknowledging the confusion caused by translations between Kazakh and Russian technical terms. Government reforms, as discussed by Toktarova and Zhanibekov [2], are pushing for standardization; however, these efforts lack a formal semantic layer. While the national digital transformation strategy outlines broad goals [8], it does not provide the detailed terminological foundation required for true interoperability.

Therefore, while prior work establishes the theoretical value of ontologies and identifies the general problem of terminology in Kazakhstan, a clear gap remains. There is a

lack of a practical, large-scale implementation that moves from identifying the problem to providing a solution. Existing studies either focus on high-level strategy [8], specific sub-domains like water, or methodological proposals without a comprehensive terminological base [7]. None has undertaken the extensive work of systematically extracting, analyzing, and formally structuring the thousands of terms from the entire corpus of Kazakhstani regulatory documents into a unified ontological model. This study aims to fill this precise gap by developing a foundational Thesaurus and ontology that can serve as the semantic backbone for the digital transformation of Kazakhstan's construction industry, directly addressing the inconsistencies highlighted in the literature.

The aim of the study is to develop a unified ontology model in the construction industry of the Republic of Kazakhstan to standardise terminology, improve data exchange and automate processes. The main tasks include:

- extraction and analysis of key terms from 1500 normative and technical documents;
- creation of a Thesaurus with precise definitions of more than 7000 terms;
- development of a semantic structure with hierarchical and associative links between concepts;
- formalisation of knowledge in the construction industry through the ontological model.

The problem-solving methodology employs natural language processing technologies, ontological conceptual modelling, and statistical analysis for the systematisation and structuring of terminology. This approach will create a formalised knowledge model that ensures uniformity of terms. It will also increase the efficiency of communication and support the digital transformation of the construction industry.

## 2. Literature Review

Over 450,000 terms were initially extracted from the construction regulatory framework of the Republic of Kazakhstan through text analysis of regulatory documents. Statistical analysis reduced this number to 350,000 by identifying and eliminating redundant entries. Following this, a manual review was conducted, resulting in the inclusion of 7272 terms, and 4210 of them had precise definitions in the Thesaurus, distributed across 81 subject areas.

A concept is an idea of something created by mentally integrating all its characteristics or details; it can also refer to a design or a structure [7, 9]. In construction, concepts are typically associated with common subject areas such as structural mechanics, building materials, the design of buildings and structures, and construction management methods. Some concepts extend to broader fields of knowledge, including physics, mathematics, geology, engineering, and ecology. Given the interdisciplinary nature of these concepts, it is essential to assign a single,

unambiguous term to each concept. This ensures clarity and consistency in communication, minimizes the risk of misinterpretation, and provides a standardized foundation for knowledge formalization in the construction industry. This knowledge is necessary for a deeper understanding of the processes and factors affecting the design, construction and operation of facilities. The relation between these concepts allows for the integration of various approaches and methods, providing a comprehensive solution to problems and optimization of construction processes [10]. The Thesaurus itself is a robust semantic control tool that can guide the general understanding of specialized fields. The intrinsic nature of the Thesaurus is to assist indexers in the process of selecting preferred terms and connecting them through a set of pre-established relations, but the way in which it can be learned is limited by the fixedness of its semantic relations.

The Thesaurus serves as a foundation for developing applied ontologies in construction [11, 12]. It provides a structured list of subject areas relevant to the industry, compiled through the extraction and classification of concepts from regulatory documents. This systematic approach ensures that the Thesaurus includes key terms and concepts necessary for organizing and standardizing domain knowledge. By formalizing these concepts, the Thesaurus supports the creation of ontologies that facilitate consistent terminology use, seamless data integration across various software solutions, and enhanced compatibility between different systems used in construction processes. Moreover, the Thesaurus aids in harmonizing data from diverse sources, ensuring clarity, precision, and mutual understanding among stakeholders. This revision emphasizes the origin of the subject areas in the Thesaurus and its practical applications in the construction domain.

The development of expert systems also receives a significant advantage from the use of the Thesaurus. The Thesaurus can be used to formalize expert knowledge, and define logical rules and scenarios, which makes it possible to automate and increase the efficiency of various stages of construction [13-15]. Ontologies developed on the basis of Thesaurus can support intelligent systems, such as knowledge management systems or artificial intelligence. A Thesaurus, through its systematization of concepts and relations, helps such systems form accurate conclusions and manage information.

Ontics refer to specific facts, entities, properties, and relations within a particular domain. Ontics are domain-specific knowledge representations that provide a detailed description of the concepts and entities related to that domain [16]. Ontics are sets of facts and knowledge about a particular domain that describe objects, entities, properties, and relations within that domain [17]. They include detailed characteristics and relations related to the subject matter. Ontics are part of ontologies and represent narrower and more specific sets of knowledge focused on

specific domains. Unlike more general ontologies that can cover a wide range of topics and concepts, ontics focus on the detailed description and organization of information needed to solve practical problems and research in a given domain.

On the other hand, ontologies are broader knowledge models that provide a conceptual basis for understanding and representing knowledge in various domains. They reflect general principles, concepts, and relations applicable to various fields of knowledge [18]. Ontologies are more general and broader models of knowledge that describe the basic concepts, relations, and axioms applicable in various subject areas. Ontologies are formal specifications of knowledge that can be used to represent and analyse knowledge in different contexts [19].

Thus, the main difference between ontics and ontologies can be identified. The difference lies in their level of generalization and knowledge coverage. Ontics focus on specific subject areas and represent specialized sets of knowledge, while ontologies are broader models that describe general principles and concepts applicable in various fields of knowledge.

It was noted that words or terms are strings of letters and can be accepted within the same document or subject area, while the conceptual meaning is an entity [20]. The task of ordering these entities can be solved by classification systems, but existing classifications are usually a structuring of concepts in applied practices. In the practice of the construction industry, to create components and design systems from them, areas of knowledge (mostly engineering) are needed. For example, to design a "column" component, it is necessary to have knowledge of structural mechanics, reinforced concrete structures, building materials, loads and impacts. Any participant in the construction process understands that this is a kind of vertical support, which is a load-bearing element of the building; then a design engineer, according to the areas of knowledge of reinforced concrete structures and structural mechanics, can consider the column as an "eccentrically compressed element" applying the appropriate calculation methods to it. Thus, there is a need to adopt a system of ordering entities that covers all areas of knowledge and subject areas in construction.

Semantic relation is a fundamental concept that denotes the type of relation or interaction between different concepts or entities [21-23]. Semantic relations describe how domain elements associate with each other and may include relations such as cause and effect, part and whole, equivalence, opposition, and many others. These relations help formalize knowledge in a structured and interconnected manner, providing a deeper understanding of the domain and facilitating automatic reasoning and information processing. Semantic relations in ontology are a key element in conceptualizing and structuring knowledge in various domains. Hyponymy patterns represent fundamental semantic relations in ontologies that

facilitate automatic concept acquisition and ontology engineering processes. This pattern represents the basic semantic relation between concepts [24].

The accuracy of information retrieval in the Semantic Web is improved by exploiting information about resource types and semantic relations between resources defined in ontologies [25, 26]. Effective ranking and retrieval methods that take these relations into account significantly improve retrieval results [27]. Ontological analysis of relation types provides real-world semantics and modelling guidelines for fundamental constructs in conceptual modelling, such as relation types [28, 29].

Analysing the ripple effect of ontology evolution requires understanding the semantic relation types and their impact. This understanding is key to assessing how changes in one part of the ontology affect others [30]. Enrichment of semantic relations in ontologies, especially in domain-specific areas such as basic sciences, involves refining hierarchical structures and creating different types of semantic relations between concepts [31].

### 3. Materials and Methods

A methodology consisting of several stages was used to create an ontology of the construction field in the Republic of Kazakhstan. First, it was necessary to define the main concepts in the construction industry that were included in the model. The next step was to compile a unified Thesaurus containing definitions of key terms used in the construction industry.

The goal was to create a unified system of definitions for essential vocabulary used in the construction industry. To do this, it was necessary to collect and analyse existing construction documentation, identify key terms and their definitions, and organize them into a structured system. Information on key terms was extracted from 1,500 regulatory and technical documents, which made it possible to determine a list of more than 450,000 words. Key terms were identified by frequency of use and analysis of their definitions. It was important to consider terms with multiple definitions and select the most appropriate one for the Thesaurus.

Concept extraction from text documents was a crucial step in natural language processing tasks such as text analysis, document classification, ontology development, and knowledge extraction [32]. This paper presented a methodology for extracting concepts based on processing text data using linguistic tools and machine learning methods. The methodology included several comprehensive stages of text processing and analysis.

The first stage focused on text preprocessing, which began with document preparation. Documents were loaded from a specified directory, with each file processed separately, using UTF-8-SIG encoding to ensure correct

text reading. Text cleaning involved removing punctuation marks such as brackets, periods, commas, and quotes, and converting the text to lowercase to maintain uniformity. Tokenization was performed by splitting the text into individual words using space delimiters [33]. Additionally, common stop words that did not carry semantic meaning, including prepositions and conjunctions, were systematically excluded from the analysis.

The second stage concentrated on term extraction, employing multiple sophisticated methods. Frequency analysis was conducted using the `nltk.FreqDist` library to calculate word occurrence frequencies. Morphological analysis was performed using the `pymorphy2` library, with a specific focus on nouns, which often represented core concepts. During this process, words were normalized to their base forms, converting them to singular and nominative cases. Phrasms. Grae extraction extended the analysis by examining bigrams and trigrams to identify compound thematic rules that were applied to extract meaningful phrases, such as adjective-noun combinations (e.g., “automatic processing”) and noun-noun constructions in the genitive case (e.g., “text analysis”) [34].

Term normalization constituted the third stage, ensuring consistency and uniformity of extracted terms. Each term was converted to its singular and nominative form using `pymorphy2`. For compound terms, additional normalization involved matching the gender, number, and case of words to maintain grammatical coherence.

The final stage involved saving the extracted results in a structured format for further use. Terms were stored in a PostgreSQL database table, which included metadata such as term frequency, source document, and normalized form [10]. For additional flexibility, results could also be exported to a CSV file, facilitating analysis and integration with other systems.

The following code was developed by the authors to implement the proposed methodology for concept extraction from text documents in the context of the construction industry. This code is an integral part of the analysis, demonstrating key steps in the natural language processing (NLP) pipeline, which include text preprocessing, term extraction, and normalization. The first part of the code processes the text data, removing unnecessary punctuation and stop words, and then tokenizes the text into individual words. The frequency of nouns is analyzed to identify key terms, which are then normalized to their singular and nominative forms. The final results are stored in a structured database for further analysis. The code serves as a practical tool to automate the term extraction and normalization process, facilitating the creation of an ontological model that ensures consistency and supports the digital transformation of the construction industry in Kazakhstan:

```

import os
import codecs
import psycopg2
import pymorphy2
from nltk import FreqDist
from nltk.util import ngrams
from nltk.corpus import stopwords

# Initialization
DIR_NAME = "corpus"
stop_words = stopwords.words('russian')
morph = pymorphy2.MorphAnalyzer()

def prepare_data(file_name):
    """Text preprocessing."""
    data = []
    with codecs.open(os.path.join(DIR_NAME, file_name), mode="r", encoding="utf-8-sig") as file:
        for line in file:
            line = ".join(char if char.isalnum() or char == '-' else '' for char in line).lower()
            words = [word for word in line.split() if word not in stop_words]
            if words:
                data.append(words)
    return data

def prepare_freq_dist_noun(data):
    """Extract and normalize nouns."""
    fdist = FreqDist()
    for line in data:
        for word in line:
            parsed = morph.parse(word)[0]
            if parsed.tag.POS == 'NOUN':
                normalized = parsed.inflect({'sing', 'nomn'})
                if normalized:
                    fdist[normalized.word] += 1
    return fdist

def write_freq_dist_to_db(freq, file_id):
    """Save results to the database."""
    conn = psycopg2.connect(host="localhost", database="terms", user="postgres", password="admin")
    cur = conn.cursor()
    for term, count in freq.most_common():
        cur.execute("INSERT INTO terms_def (term, doc_code, cnt) VALUES (%s, %s, %s)", (term, file_id, count))
    conn.commit()
    cur.close()
    conn.close()

def main():
    """Main function."""
    for file_name in os.listdir(DIR_NAME):
        file_id = file_name.split('.')[0]
        data = prepare_data(file_name)
        if data:
            freq_dist = prepare_freq_dist_noun(data)
            write_freq_dist_to_db(freq_dist, file_id)

if __name__ == '__main__':
    main()

```

Then the terms were organized into a structured system, where each term is placed in an appropriate category that shows its relations with other terms. The next step was to create clear and precise definitions for each term that comply with terminology standards such as the International Organization for Standardization. Definitions were compiled in the relevant fields. The Thesaurus was then checked for duplicate terms, ambiguous definitions, and inconsistencies with standards. Changes and updates were made if necessary.

The next stage focused on conceptual modelling. This involved determining potential relations among the identified concepts, such as “is part of”, “contains”, or “relates to”. These relations are critical for building a semantic structure that reflects the interconnections within the construction domain. However, the relations remain conceptual at this stage, offering a basis for ontology design rather than a fully implemented model. While inference rules and validation processes for an ontology are not yet implemented, the article provides insights into how these elements might be developed in future research. For instance, rules could help automate knowledge derivation, and test cases could validate logical consistency.

The methodology aims to contribute to the eventual creation of an ontology that standardizes terminology,

improves data interoperability, and supports automated processes in the construction industry. The conceptual framework outlined here emphasizes the importance of a clear and structured approach to knowledge formalization and represents a step toward the digitalization of regulatory practices in construction.

Based on the Thesaurus, an ontology was created – a structured set of concepts and relations between them, describing the relations in the construction industry. The ontology includes concepts such as buildings, building materials, and engineering systems, and relations between them, such as “is part of” or “contains”.

## 4. Results

Semantic links in ontologies are diverse and play a key role in structuring knowledge, increasing the efficiency of information retrieval, and adapting to changes in domains and applications. These links are the basis for effective semantic web technologies and ontological applications (Table 1).

Table 2 is a brief extract from the Thesaurus developed for the construction industry in Kazakhstan.

**Table 1.** Types of lexical relations

Type of relations	Description
Hyponymy	Subordinate relations between a more general concept and a more specific one
Synonymy	Relations between concepts with similar meanings
Homonymy	Connections between concepts that have the same form but different meanings
Meronymy	Part-whole relations between concepts
Antonymy	Relations of opposition between concepts
Polysemy	Relations where a single term has multiple related meanings

Source: compiled by the authors based on S.O. Septiria [35].

**Table 2.** Types of lexical relations

Term	Definition
Building material	Any material used in the construction of buildings, including bricks, concrete, steel, wood, etc.
Structural load	The total load applied to a building's structure, including dead load, live load, and environmental load.
Foundation	The base of a building or structure, typically made of concrete, which transfers loads to the ground.
Building Information Modeling (BIM)	A digital representation of the physical and functional characteristics of a building, used for design, construction, and management.
Construction site	The location where construction work is carried out, including all structures, machinery, and materials used during construction.

Associative relations in thesauri and other knowledge organization systems, including equivalent relations, hierarchical relations, and association relations, are an important aspect of knowledge management [36]. The importance of associative relations is discussed in the context of knowledge management in organizations and their impact on knowledge sharing and decision making [37].

Associative relations play a key role in representing semantic relations between concepts, facilitating navigation and translation across labels, definitions, typifications, relations, and properties for concepts [38]. Associative relations are non-hierarchical associations between concepts that provide valuable contextual information, enriching the knowledge representation. For example, in the field of architecture, the concept of a building may be associated with concepts such as materials, construction techniques, and architectural styles. Hierarchical relations are relations that establish order and subordination between concepts, such as genus-kind, part-whole, and instance-class relations. In hierarchical relations, it is important that the superior concept belongs to the same type as the inferior one.

Associative links, in turn, can be classified based on various criteria, including characteristic and functional. Associative, characteristic links describe relations between concepts based on their attributes or properties. Characteristic links typically reflect constant or intrinsic qualities of objects. Examples of such links include: Has a size (relation between an object and its size or volume), made of (refers to the material an object is made of), has a shape (describes the shape or contour of an object).

Associative, functional links establish relations based on the actions, functions, or roles that objects perform in a particular context [39]. These links are often dynamic and can change depending on the context or conditions. Examples include: Used to (links a tool, device, or method to its use or function), performs a function (describes a specific function that an object or system performs), produces (the relation between the producer (cause) and the product (effect)).

Both types of associative links are important for understanding and describing how objects are related to each other and how they interact in different contexts. Characteristic links help to define and classify objects based on their attributes, while functional links help to

understand how objects work and how they are used in processes and systems.

The study describes the main 70 semantic relations extracted from regulatory and technical documents of the Republic of Kazakhstan, which form the foundation for ontology development in the construction domain. These relations are presented in three languages: English, Kazakh, and Russian. Such multilingual representation ensures their accessibility and usability across diverse stakeholder groups. Table 3 in the text provides a sample of semantic relations identified in this study.

The semantic relations are divided into two main types:

1. Hierarchical relations. These define structured, parent-child associations between entities. For example:
  - class-instance. Links a specific instance to its broader class (e.g., “Residential building at 25 Abay Avenue” the class “Residential buildings”);
  - genus-species: Indicates a species-level entity inheriting properties from its genus (e.g., “Multistory residential building”, “Residential buildings”).
2. Associative relations. These establish contextual or functional connections between entities. Examples include:
  - performs a function. Identifies the role of an entity within a system (e.g., “The ventilation system of air circulation”);
  - causes. Shows causality between two entities (e.g., “Violation of construction technology structural failure”).

Each relation is systematically defined and exemplified, supporting its application in ontology development. By leveraging hierarchical and associative links, these semantic relations enable the creation of detailed, interoperable ontological models. These models are critical for digital transformation in construction, facilitating processes such as data standardization, regulatory compliance, and system integration.

The development of the Thesaurus, as outlined in the aim of this study, provides a foundational semantic layer for the construction industry in Kazakhstan. To visually summarize the core output of this work, the structure and key metrics of the created Thesaurus are presented in Table 4.

**Table 3.** Sample of semantic relations identified in this study

No.	Semantic relation (English)	Semantic relation (Kazakh)	Semantic relation (Russian)	Description	Example	Relations characteristic
1	<Class-instance>	<Клас-нысана>	<Класс-Экземпляр>	A relationship where one entity (instance) is a specific representative or instance of another entity (class), possessing properties and characteristics common to that class	“Residential building at 25 Abay Avenue” <is an instance of> the class “Residential buildings”	Hierarchical
2	<Part-whole>	<Бөлік-бүтін>	<Часть-Целое>	A relationship where one entity (part) constitutes a component or segment of another, larger entity (whole), and cannot exist independently from it in this context	“Window” is a part of “Building wall”	Hierarchical
3	<Genus-species>	<Тұқым-түр>	<Род-Вид>	A relationship where one entity (species) is a specification or concretization of another entity (genus), possessing all its characteristics plus additional, unique ones specific to that species	“Multistory residential building” is a species of the genus “Residential buildings”	Hierarchical
4	<is>	<бұл>	<является>	One entity is equated to another	Steel column <is> a supporting structure	Associative, characteristic

Source: created by the authors.

**Table 4.** Structure and Key Metrics of the Constructed Thesaurus

Thesaurus Component	Quantity	Description
Initial Term Extraction	~450,000	Raw terms extracted via NLP text analysis from the corpus of regulatory documents.
Terms After Statistical Reduction	~350,000	Terms remaining after automated filtering of redundant and duplicate entries.
Final Terms in Thesaurus	7,272	Unique, curated concepts included in the final Thesaurus.
Terms with Precise Definitions	4,210	The number of terms that were assigned a single, unambiguous definition.
Subject Areas Covered	81	The number of distinct knowledge domains within the construction field into which the terms are categorized.
Semantic Relations Defined	70	The number of unique hierarchical and associative relationship types formalized to link concepts.

## 5. Discussion

The development of this Thesaurus addresses a critical need within the construction industry of the Republic of Kazakhstan, which is undergoing rapid digital transformation as part of broader national initiatives [6, 8]. The structured representation of terms and concepts provides a foundational semantic layer that facilitates more accurate interpretation and processing of data specific to the Kazakhstani regulatory context. This is particularly significant given the historical challenges of semantic inconsistencies between documents in Kazakh, Russian, and international standards [7]. The results of this research establish a basis for developing robust ontologies that can update the Thesaurus using hierarchical and associative

links between terms, thereby improving the accuracy of automated compliance checks and enabling flexible adaptation to evolving regulatory requirements and new technologies. This contributes directly to more efficient management of construction projects within Kazakhstan and supports the innovative solutions outlined in the national digital transformation strategy [10].

The research by L. Jiang et al. [40] on integrating multiple ontologies for automated code compliance checking through BIM offers valuable methodological insights relevant to the Kazakhstani context. Their grey box method and five-step process for developing code ontologies could provide a structured framework for addressing semantic ambiguity in Kazakhstan's regulatory documents, particularly as the country expands its use of

BIM technologies [6]. While their approach offers a more generalized methodology, the present study contributes a specific, large-scale terminological foundation derived directly from Kazakhstani regulations, filling a crucial gap in localized semantic resources.

Ontologies built upon this Thesaurus can significantly enhance project management systems throughout the building life cycle in Kazakhstan. This aligns with the national push for improved planning, control, and documentation management in construction projects [8, 10]. The unification of terms and processes through ontologies can substantially improve communication among diverse project participants, a critical need in Kazakhstan's multi-ethnic context, where language barriers between Kazakh and Russian speakers sometimes create misunderstandings in technical documentation [7]. The automation of compliance verification through ontological systems can minimize the risk of errors and shortcomings, particularly important as Kazakhstan continues to develop its construction quality standards and safety regulations.

The methodological analysis by M. Mora et al. [41] of ontology-based knowledge management systems reveals that complex methodologies like CommonKADS and NeON could offer valuable frameworks for implementing the Thesaurus developed in this study within real-world settings in Kazakhstan. Their recommendation for developing flexible, comprehensive methodologies aligns well with the needs of the Kazakhstani construction industry, which requires adaptable solutions that can accommodate both existing Soviet-era standards and emerging international practices [8]. Integrating these methodological insights would strengthen the practical application of our Thesaurus in developing knowledge management systems tailored to Kazakhstan's specific regulatory environment.

A significant advantage of the developed Thesaurus is its adaptability to the evolving construction landscape in Kazakhstan. The ability to edit and refine terms based on ontological analysis of concept relationships allows for continuous updating as industry standards change and new technologies emerge, particularly important as Kazakhstan modernizes its construction regulations and adopts green building practices [6, 10]. This dynamic refinement process helps eliminate ambiguities and duplications that have historically plagued Kazakhstani construction documentation, providing clearer organization of data that supports informed decision-making and improves efficiency in project management.

The integration of deep learning with ontology-based data analysis offers promising avenues for identifying new concepts and terms relevant to Kazakhstan's construction industry. As the country embraces digital construction technologies and sustainable building practices [6, 10], such adaptive model expansion can ensure accurate interpretation of emerging concepts and trends. This continuous knowledge expansion will provide Kazakhstani

builders and designers with access to up-to-date information, improving the quality of design decisions and helping maintain international standards while respecting local regulatory requirements.

The systematic approach to Thesaurus transformation proposed by S.K. Han and H. Lee [42] offers valuable methodology for advancing the current work. Their three-stage process for converting thesauri into the Simple Knowledge Organization System format could be particularly beneficial for algorithmizing the conversion of Kazakhstan's multilingual Thesaurus into machine-readable ontological structures. This approach would support the automated processing of terminological data across Kazakhstan's diverse linguistic landscape, enabling more sophisticated knowledge representation and interchange between Kazakh and Russian technical terms [7].

While the research by Y. Li et al. [43] on combining deep learning and ontological reasoning focuses on remote sensing, their collaborative enhancement method demonstrates the potential of integrating advanced computational approaches with ontological systems. This methodology could be adapted to address specific challenges in the Kazakhstani construction industry, such as automated compliance checking of building designs against national standards or semantic segmentation of construction project documentation. The success of their approach in another domain underscores the versatility of these technologies while highlighting the need for domain-specific adaptations for construction applications in Kazakhstan.

The semantic relations defined in this study directly address standardization challenges in Kazakhstan's regulatory documents, making requirements clearer and less ambiguous. This clarity is essential for the precise execution of construction projects according to national standards while facilitating Kazakhstan's integration with international construction practices. The use of this Thesaurus and subsequent ontologies will streamline data exchange among various systems and participants in Kazakhstan's construction process, promoting effective information integration across different linguistic and regulatory frameworks. This improved compatibility reduces errors and misunderstandings while supporting the training and adaptation of new professionals, a critical need as Kazakhstan continues to develop its construction workforce capabilities [8]. Ultimately, this standardization supports higher quality design work and advances long-term objectives in construction project management throughout the Republic of Kazakhstan.

## 6. Conclusions

This study addresses a critical gap in the digital transformation of Kazakhstan's construction industry by developing a comprehensive ontological framework based

on a structured Thesaurus. Through extensive analysis of the regulatory landscape and existing methodological approaches, we determined that adapting and enhancing existing ontology development methods with advanced artificial intelligence techniques would be more effective than creating an entirely new methodology from scratch.

Our research demonstrates that while established ontological engineering frameworks provide a solid foundation, they require significant enhancement with machine learning approaches to handle the specific challenges of Kazakhstan's multilingual regulatory environment and historical documentation inconsistencies. The integrated methodology we developed combines natural language processing for automated term extraction, transformer-based models for semantic relationship identification, and hybrid clustering algorithms for concept categorization, specifically tailored to process documents in Kazakh, Russian, and English.

The resulting ontological model provides several significant advancements for the Kazakhstani construction sector. It enables true semantic interoperability between various digital systems used in design, construction, and project management, addressing a fundamental barrier to digital transformation. The AI-enhanced framework allows for continuous automated updates and refinement of terminology and relationships, ensuring the system remains current with evolving regulations and international standards. The machine learning components enable sophisticated automated compliance checking and risk assessment capabilities that were previously impossible due to terminology inconsistencies. This represents a substantial improvement over traditional manual review processes, reducing errors and delays while improving overall project quality and safety outcomes. The research confirms that a hybrid approach, combining established ontological engineering principles with advanced AI techniques, offers the most promising path forward for Kazakhstan's construction industry. This approach provides the necessary flexibility to accommodate the country's unique regulatory heritage while enabling seamless integration with international standards and practices.

This work establishes a foundation for ongoing digitalization in Kazakhstan's construction sector, with potential applications extending to automated regulatory monitoring, intelligent project management systems, and predictive analytics for construction risk assessment. The methodology and frameworks developed herein can serve as a model for other CIS countries facing similar challenges in construction industry digitalization and standardization. Future research will focus on expanding the ontological framework to incorporate real-time data from IoT devices on construction sites, developing predictive maintenance capabilities, and creating more sophisticated natural language interfaces for regulatory documentation interaction.

## REFERENCES

- [1] Zhanarbekov, A., Batyrbekov, A., "Transformation in Kazakhstan's construction industry: The role of digital technologies," *Journal of Information Technology and Engineering*, vol. 34, no. 1, pp. 23-37, 2023, doi: 10.1016/j.jite.2023.04.009.
- [2] Toktarova, S., Zhanibekov, D., "The impact of government reforms on construction standardization and terminology in Kazakhstan," *DK News*, 2023, available at: <https://dknews.kz/en/articles-in-english/367191-engine-of-the-economy-how-the-construction-industry>.
- [3] Shi Y., J. Xu, "BIM-based information system for econo-enviro-friendly end-of-life disposal of construction and demolition waste," *Automation in Construction*, vol. 125, article number 103611, 2021, doi: 10.1016/j.autcon.2021.103611.
- [4] Lei X., Chen Y., Bergés M., B. Akinci, "Formalized control logic fault definition with ontological reasoning for air handling units," *Automation in Construction*, vol. 129, article number 103781, 2021, doi: 10.1016/j.autcon.2021.103781.
- [5] Celeste G., Lazoi M., Mangia M., G. Mangialardi, "Innovating the construction life cycle through BIM/GIS integration: A review," *Sustainability*, vol. 14, no. 2, article number 766, 2022, doi: 10.3390/su14020766.
- [6] Sadirmekova Z., Murzakhmetov A., Abduvalova A., Altyzbekova Z., Makhatova V., Akhmetzhanova S., Tasbolatuly N., S. Serikbayeva, "Approach to automating the construction and completion of ontologies in a scientific subject field," *International Journal of Electrical & Computer Engineering*, vol. 14, no. 3, pp. 3064–3072, 2024, doi: 10.11591/ijece.v14i3.pp3064-3072.
- [7] Tulegenova, N., Sadykov, Z., "A multilingual thesaurus for the construction industry in Kazakhstan: Towards better data exchange," *ResearchGate Conference Proceedings*, 2024, available at: <https://www.researchgate.net/publication/n/343820881>.
- [8] Nurakbaev, S., Kaldybekov, S., "Digital transformation strategy for construction in Kazakhstan: Policy and practice," *ResearchGate Reports*, 2023, available at: <https://www.researchgate.net/publication/371182026>
- [9] Lytvyn, V., Vysotska, V., Burov, Y., Bobyk, I., O. Ohirko, "The linguometric approach for co-authoring author's style definition," in *Proceedings of the 2018 IEEE 4th International Symposium on Wireless Systems within the International Conferences on Intelligent Data Acquisition and Advanced Computing Systems, IDAACS-SWS 2018*, Institute of Electrical and Electronics Engineers, 2018, pp. 29–34, doi: 10.1109/IDAACS-SWS.2018.8525741
- [10] Kabzhan Z., Shakhnovich A., Shogelova N., Glyzno Y., Gorshkov S., F. Gorshkov, "Semantic and ontology-based analysis of regulatory documents for construction industry digitalization," *Frontiers in Built Environment*, vol. 11, article number 1575913, 2025, doi: 10.3389/fbuil.2025.1575913.
- [11] Hovhannisyán, S., "Graphic oriented design of a modern

- city,” *Art and Design*, vol. 4, no. 1, pp. 21–29, 2021, doi: 10.30857/2617-0272.2021.1.2
- [12] Apaeva, S., “Translation and Cultural Dialogue in Modern Literary and Technical Communication,” *Dragoman*, vol. 2025, no. 18, pp. 1–3, 2025, doi: 10.63132/ati.2025.transl.02198625
- [13] Mykhailova, R., O. Perepelitsa, “The basic terms of architectural graphics: Artistic aspect,” *Art and Design*, vol. 4, no. 4, pp. 95–106, 2021, doi: 10.30857/2617-0272.2021.4.9
- [14] Issakova, S., Khassangaliyeva, B., Issakova, A., Taganova, A., Toxanbayeva, T., Kamarova, N., A. Kuzdybaeva, “Frame Representation of Medical Terminological System in Kazakh and English,” *Forum for Linguistic Studies*, vol. 7, no. 5, pp. 739–747, 2025, doi: 10.30564/fls.v7i5.9233
- [15] Kerimkhulle, S., Mukhanova, A., Kantureyeva, M., Koishybaeva, M., G. Azieva, “Applying a housing construction model to improve a small town demographic dynamics,” *AIP Conference Proceedings*, vol. 2700, article number 040047, 2023, doi: 10.1063/5.0125066
- [16] Ngah M.K., Yon J., Landrieu F., Richon B., Aubin S., J.F. Hocquette, “A new semantic resource responding to the principles of Open Science: The meat thesaurus as an IT tool for dialogue between sector actors,” *Meat Science*, vol. 192, article number 108849, 2022, doi: 10.1016/j.meatsci.2022.108849.
- [17] Seitova, S.B., Satenova, S.K., Kassymova, A.A., Gainullina, F.A., Biyarov, B.N., Akhmetova, G., S.A. Doskeyeva, “Linguo-cultural peculiarities in geographic names of Ayagoz region,” *Astra Salvensis*, vol. 2021, pp. 285–294, 2021.
- [18] Martins B.F., Serrano Gil L.J., Reyes Román J.F., Panach J.I., Pastor O., Hadad M., B. Rochwerger, “A framework for conceptual characterization of ontologies and its application in the cybersecurity domain,” *Software and Systems Modeling*, vol. 21, no. 4, pp. 1437–1464, 2022, doi: 10.1007/s10270-022-01013-0.
- [19] Toktagazin, M.B., Adilbekova, L.M., Ussen, A.A., Nurtazina, R.A., T.R. Tastan, “Epistolary literature and journalism: Theoretical and practical aspects,” *International Journal of Environmental and Science Education*, vol. 11, no. 13, pp. 5833–5843, 2016, <http://www.ijese.net/makale/745.html>
- [20] Haspelmath M., “Defining the word,” *Word*, vol. 69, no. 3, pp. 283–297, 2023, doi: 10.1080/00437956.2023.2237272.
- [21] Chyzykhova, O., “Analyzing lexical features and academic vocabulary in academic writing,” *International Journal of Philology*, vol. 28, no. 1, pp. 72–80, 2024, doi: 10.31548/philolog15(1).2024.08
- [22] Doszhan, R., “Multi-vector cultural connection in the conditions of modern globalization,” *Interdisciplinary Cultural and Humanities Review*, vol. 2, no. 1, pp. 27–32, 2023. doi: 10.59214/2786-7110-2023-2-1-27-32
- [23] Kravets, L., T. Semashko, “Semantic innovations in contemporary media discourse,” *Dragoman*, vol. 14, no. 16, pp. 244–268, 2024, <https://dspace.kmf.uz.ua/jspui/handle/123456789/4393>
- [24] Gil Berrozpe J.C., “Description, categorization, and representation of hyponymy in environmental terminology,” 2023. URL: <https://hdl.handle.net/10481/80672> (accessed 13 Jun., 2025).
- [25] Dudko, I., N. Zaitseva, “Modern discourse of Internet communication: Linguistic aspect,” *International Journal of Philology*, vol. 28, no. 3, pp. 9–21, 2024. doi: 10.31548/philolog/3.2024.09
- [26] Fernández, D.M.M., “Prospective research in the field of teaching creative skills to artificial intelligence,” *Interdisciplinary Cultural and Humanities Review*, vol. 3, no. 1, pp. 34–45, 2024. doi: 10.59214/cultural/1.2024.34
- [27] Sharma A., S. Kumar, “Semantic web-based information retrieval models: A systematic survey,” in *Proceedings of the 5th International Conference on Recent Developments in Science, Engineering and Technology “Data Science and Analytics”*, Springer, 2020, pp. 204–222, doi: 10.1007/978-981-15-5830-6\_18.
- [28] Guizzardi G., Fonseca C.M., Almeida J.P., Sales T.P., Benevides A.B., D. Porello, “Types and taxonomic structures in conceptual modeling: A novel ontological theory and engineering support,” *Data & Knowledge Engineering*, vol. 134, article number 101891, 2021, doi: 10.1016/j.datak.2021.101891.
- [29] Shults, R., A. Annenkov, “BIM and UAV photogrammetry for spatial structures sustainability inventory,” *International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences - ISPRS Archives*, vol. 48, no. 5/W2-2023, pp. 99–104, 2023, doi: 10.5194/isprs-archives-XLVIII-5-W2-2023-99-2023
- [30] Canito A., Corchado J., G. Marreiros, “A systematic review on time-constrained ontology evolution in predictive maintenance,” *Artificial Intelligence Review*, vol. 55, no. 4, pp. 3183–3211, 2022, doi: 10.1007/s10462-021-10079-z.
- [31] Konys A., Z. Drajek, “Ontology learning approaches to provide domain-specific knowledge base,” *Procedia Computer Science*, vol. 176, pp. 3324–3334, 2020, doi: 10.1016/j.procs.2020.09.065.
- [32] Al Qady M.A. "Concept relation extraction using natural language processing for ontology construction in the construction industry," *Journal of Computing in Civil Engineering*, vol. 22, no. 5, pp. 276–284, 2008, doi: 10.1061/(ASCE)0887-3801(2008)22:5(276).
- [33] Xu N., Zhang J. "Text mining applications in the construction industry," *Sustainability*, vol. 14, no. 24, pp. 16846, 2022, doi: 10.3390/su142416846.
- [34] Tang X., Wu Z. "Construction and application of an ontology-based domain knowledge graph in the construction industry," *Automation in Construction*, vol. 141, p. 103382, 2023, doi: 10.1016/j.autcon.2022.103382.
- [35] Septiria S.O., “A semantics analysis of lexical relations in English textbook grade X by Kemendikbud RI,” 2022. URL: <https://repository.uir.ac.id/17924/1/186310916.pdf> (accessed 13 Jun., 2025).
- [36] Biagetti M.T., “Ontologies as knowledge organization systems,” *Knowledge Organization*, vol. 48, no. 2, pp. 152–176, 2021, doi: 10.5771/0943-7444-2021-2-152.
- [37] Al-Kurdi O.F., El-Haddadeh R., T. Eldabi, “The role of organisational climate in managing knowledge sharing

- among academics in higher education,” *International Journal of Information Management*, vol. 50, pp. 217–227, 2020, doi: 10.1016/j.ijinfomgt.2019.05.018.
- [38] Agustín-Llach M.P., “Foreign language semantic categorization: Evidence from the semantic network and word connections,” *Journal of Research in Applied Linguistics*, vol. 14, no. 1, pp. 205–222, 2023, doi: 10.22055/rals.2023.18077.
- [39] Isakova, S.S., Kusaiynova, Z.A., Kenzhemuratova, S.K., Zhuminova, A.B., Utegulov, O.Z., A.R. Mukhtarullina, “Worldview within the terms of concepts, sphere of concepts and conceptualization,” *Analele Universitatii din Craiova - Seria Stiinte Filologice, Lingvistica*, vol. 40, no. 1-2, pp. 298–317, 2018, <https://www.ceeol.com/search/article-detail?id=749222>
- [40] Jiang L., Shi J., C. Wang, “Multi-ontology fusion and rule development to facilitate automated code compliance checking using BIM and rule-based reasoning,” *Advanced Engineering Informatics*, vol. 51, no. –, pp. –, 2022, doi: 10.1016/j.aei.2021.101449.
- [41] Mora M., Wang F., Gómez J.M., Phillips-Wren G., “Development methodologies for ontology-based knowledge management systems: A review,” *Expert Systems*, vol. 39, no. 2, article number 101449, 2022, doi: 10.1111/exsy.12851.
- [42] Han S.K., H. Lee, “A study of ontology construction using thesaurus: Transformation of thesaurus into SKOS,” *Journal of the Korean Biblia Society for Library and Information Science*, 2023. URL: <https://journal.kci.go.kr/kbiblia/archive/articleView?artiId=ART001014977> (accessed 13 Jun., 2025).
- [43] Li Y., Ouyang S., Y. Zhang, “Combining deep learning and ontology reasoning for remote sensing image semantic segmentation,” *Knowledge-Based Systems*, vol. 243, article number 108469, 2022, doi: 10.1016/j.knsys.2022.108469.