

PAPER • OPEN ACCESS

Leveraging LLM Embeddings for Ontology Scope Optimisation and Reuse in Construction Transportation Data Integration to Enhance Construction Logistics

To cite this article: Jyrki Oraskari *et al* 2026 *IOP Conf. Ser.: Earth Environ. Sci.* **1672** 012025

View the [article online](#) for updates and enhancements.

You may also like

- [DeepSeq: high-throughput single-cell RNA sequencing data labeling via web search-augmented agentic generative AI foundation models](#)
Saleem A Al Dajani, Abel Sanchez and John R Williams
- [The PRISM semantic cohort builder: a novel tool to search and access clinical data in TCIA imaging collections](#)
Jonathan P Bona, Joseph Utecht, Sarah Bost et al.
- [Method for ontology representation from unstructured text corpus](#)
Sanaa Rashed Abdallah, Yasser F Hassan, Bothina Elsobky et al.

Leveraging LLM Embeddings for Ontology Scope Optimisation and Reuse in Construction Transportation Data Integration to Enhance Construction Logistics

Jyrki Oraskari¹, Jan Luca Fahrenndholz-Heiermann¹ and Sigrid Brell-Cokcan¹

¹Chair for Individualized Production, RWTH Aachen University, 52074 Aachen, Germany

E-mail: oraskari@ip.rwth-aachen.de

Abstract. Just-in-Time is a key lean construction principle aimed at reducing time, costs, and onsite emissions, but it increases reliance on precisely timed logistics. Integrating construction logistics with traffic and mobility data is currently hindered by fragmented open data sources and a lack of shared semantics. This paper investigates the use of large language model (LLM) embeddings to support ontology engineering for external mobility data that must be aligned with Just-in-Time construction logistics requirements. We propose a semi-automated workflow that uses embeddings to cluster domain terms, analyse ontology scope, identify relevant open data offers from a national mobility data platform, and suggest candidates for ontology reuse. We approach ontology scope optimization as a systematic refinement process driven by semantic embeddings. The scope of the ontology is to provide a structured representation of external datasets and their linkage points to internal concepts, rather than modeling the internal domain itself. A manually crafted baseline ontology serves as the reference for these mappings, and LLM-embedding-supported insights provide guidance. The results show that embedding-based methods can uncover missing concepts, improve concept alignment, and highlight reusable ontology fragments. At the same time, polysemy and unstable LLM outputs underline the need for systematic quality assurance.

1 Introduction

Just-in-Time (JIT) construction relies on precise logistics and tightly coordinated delivery processes. To transport prefabricated modules on time requires adherence to narrow time windows (Peiris et al. 2023) that are highly sensitive to traffic incidents, road closures, and adverse weather conditions. These disruptions reduce schedule reliability and increase planning complexity in time-critical supply chains.

In contrast to onsite fabrication, prefabricated construction methods reduce noise, dust, and emissions on-site by shifting labor-intensive and subtractive processes into controlled factory environments (Wang et al. 2024). This supports sustainability goals in urban settings, but also increases dependence on transport logistics. Large elements must be delivered just-in-time, synchronized with construction progress, and in sync with site operations, which requires reliable, interoperable access to real-time traffic, infrastructure, and environmental data.



Content from this work may be used under the terms of the [Creative Commons Attribution 4.0 licence](https://creativecommons.org/licenses/by/4.0/). Any further distribution of this work must maintain attribution to the author(s) and the title of the work, journal citation and DOI.

However, such data originates from many stakeholders, follows heterogeneous formats, and is often stored in isolated systems. Fragmentation limits interoperability and hinders informed decision-making. The rapid growth of open data offerings, such as real-time traffic conditions or road usage statistics, further complicates the situation, as inconsistent structures and missing semantics hinder their practical use.

In this context, semantic technologies offer a solution. Ontologies provide shared vocabularies, formalize relationships, and enable consistent interpretation of heterogeneous datasets. They serve as a key mechanism for implementing the principles for research data (Guizzardi 2020) and for linking traffic conditions, road works, weather information, delivery schedules, and project requirements into a unified conceptual model. FAIR means that data are Findable, Accessible, Interoperable, and Reusable.

Yet, ontology engineering is mainly manual. Experts interpret requirement documents, analyze data interfaces, review existing ontologies, derive concepts, formalize concepts, and maintain them for consistency. While this approach ensures precision, it is slow, labor-intensive, and difficult to scale. Reuse is limited by inconsistent documentation and the lack of systematic mechanisms for validating suitability. Recent advances in natural language processing (NLP) and large language models (LLMs) present new possibilities. Embedding-based similarity enables the automated extraction of domain terms, clustering related concepts, identification of structural gaps, and detection of reusable fragments in external ontologies. These techniques can reduce manual effort while improving the scalability and consistency of ontology development.

The main contribution of this work is the conceptualisation and operationalisation of ontology scope optimisation using LLM-derived embeddings to support semi-automated ontology engineering for construction-traffic data integration. Our research question is: To what extent can LLM-derived embeddings improve key ontology-engineering tasks—defining scope, selecting relevant open data sources, assessing LLM-generated ontologies, and identifying suitable external ontology fragments for reuse? A manually crafted baseline ontology is compared with insights from embeddings and LLM-generated variants. The results demonstrate the potential of modern NLP methods to enhance semantic modelling in construction logistics and to integrate transport data.

2 Related Work

LLM embeddings (Tao et al. 2024) represent text as high-dimensional numerical vectors that capture semantic meaning. LLM models such as BERT or all-MiniLM-L6-v2 (Reimers & Gurevych 2024) transform text into vector representations, aligning semantically similar content closely and enabling comparison based on meaning rather than exact wording.

Recent work has explored embedding-based clustering and LLM-supported ontology learning. Earlier approaches relied on Word2Vec and classical clustering methods (e.g., by Khan et al. (2020)), while more recent transformer-based models provide richer contextual representations (Tao et al. 2024). Several studies investigate the use of LLMs for ontology or knowledge graph generation (Kommineni et al. 2024, Babaei Giglou et al. 2023, Bakker et al. 2024), highlighting both the potential for automation and the challenges of structural consistency.

However, existing work focuses primarily on ontology generation rather than on systematic scope refinement and data-alignment analysis. Our contribution differs in conceptualising ontology scope optimisation as an embedding-supported diagnostic and reuse-oriented workflow for cross-domain data integration.

3 The Ontology Engineering Framework

3.1 *Ontology Engineering Workflow*

The proposed framework applies LLM-based embeddings to support semi-automated ontology engineering for external data sources that must align with internal requirements. Domain terms are curated from requirement documents and represent the information needs of internal construction and logistics systems. These terms are encoded into a shared embedding space that serves as a semantic reference frame. In parallel, textual descriptions of external data sources are processed and embedded using the same model, enabling consistent comparison between internal terminology and available datasets. By identifying semantically close pairs across both spaces, the method derives an ontology that describes the structure and linkage points of the external data. This ontology focuses on how external datasets should be organised and related so that their content can be mapped to internal concepts, rather than redefining the internal domain model itself. The approach, therefore, provides a general, reproducible pipeline for embedding-driven ontology construction well-suited for integration with existing lean construction models.

3.2 *Ontology Scope Optimisation*

In this work, Ontology Scope Optimisation is defined as the systematic alignment and refinement of an ontology's conceptual boundary, presented as competency questions (Noy & McGuinness 2001), to ensure consistency between internal requirement structures, external data availability, and reuse potential, while minimizing semantic redundancy and structural instability. The objective is not to expand the ontology arbitrarily, but to calibrate its conceptual scope to span the competency questions, to reflect both domain requirements and the structure of available external datasets. Operationally, scope optimisation can be conceptually assessed along four qualitative dimensions:

- **Requirement Coverage:** the extent to which extracted domain terms are represented in the ontology.
- **External Alignment:** the degree of semantic proximity between ontology concepts and external dataset descriptions.
- **Reuse Potential:** the extent to which ontology elements align with established external vocabularies.
- **Redundancy Control:** the reduction of semantically overlapping or weakly justified classes.

These dimensions provide an analytical framework for interpreting the embedding-based findings rather than constituting a fully quantitative optimisation procedure. Within this framework, embeddings function as a diagnostic instrument that supports scope refinement by revealing missing concepts, exposing structural overlaps, and identifying candidates for ontology reuse.

4 Methodology: Embedding-Based Ontology Engineering

4.1 *Manual Baseline Ontology*

A hand-crafted baseline ontology was first developed following the methodology of Uschold & King (1995). Its purpose was to represent open mobility information in a form usable for construction-related transport processes. Core concepts were derived from the use case of transporting prefabricated building elements and modelled using Chowlk (Chávez-Feria et al. 2022).

Relevant parts of the Internet of Construction Process Ontology (IOC) (Brell-Cokcan & Schmitt 2024) were reused. The resulting conceptual model is shown in Figure 1.

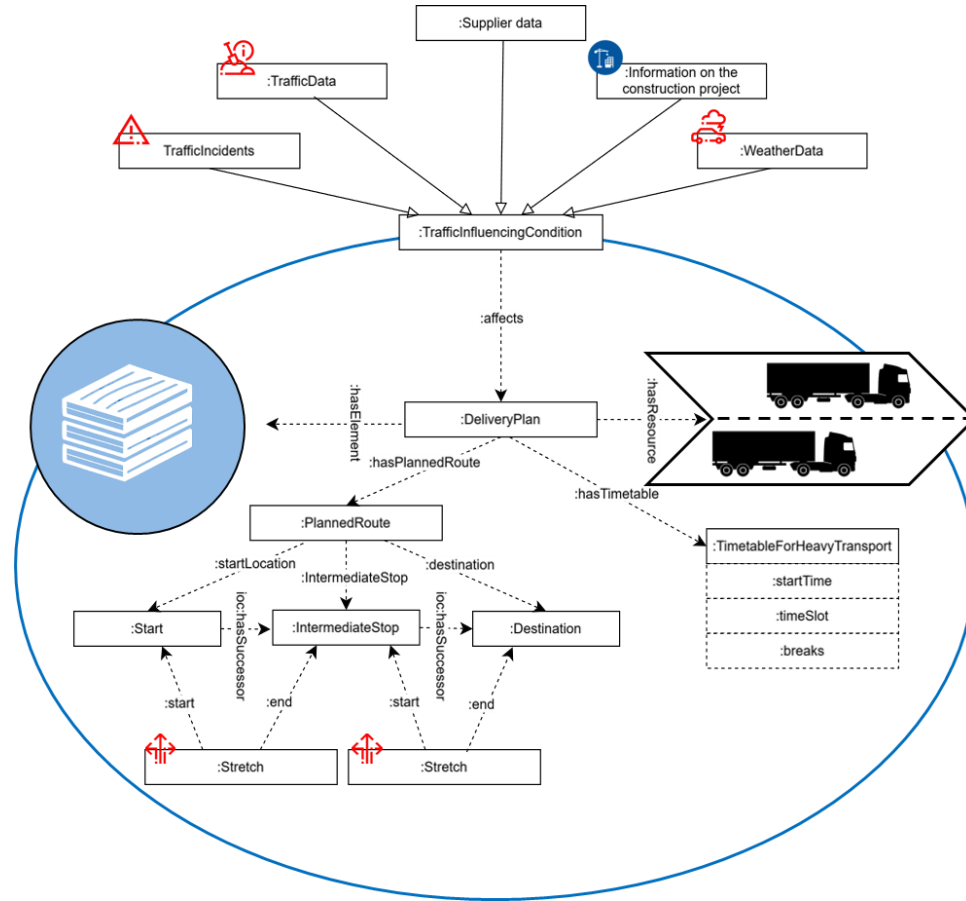


Figure 1: The first hand-crafted ontology for construction extra-logistics.

4.2 Domain Requirements and Term Extraction

A requirements analysis was conducted based on the expert-written data model document, following the recommendations of Simperl et al. (2009). Fifty-two relevant terms were identified and grouped into (i) internal concepts related to construction logistics and company systems, and (ii) external concepts expected from open data sources. These terms formed the basis for the embedding-based semantic analysis.

4.3 Embedding-Based Semantic Analysis

All terms were embedded using both the all-MiniLM-L6-v2 and the all-mpnet-base-v2 sentence-transformer model (Siino 2024) and form an Euclidean space. Both behaved similarly. Pairwise cosine distances, a standard method for semantic similarity comparison, were computed, and a similarity threshold was applied to generate a graph of the closest relationships. The all-mpnet-base-v2 model has provided the best quality in sentence embedding in the SBERT.net comparison ¹. Therefore, the semantic similarity visualizations are shown using that.

¹https://www.sbert.net/docs/sentence_transformer/pretrained_models.html

The similarity threshold determines which weak semantic connections are removed, i.e., those with a large similarity distance. That distance can be selected. When longer threshold values are used, the models tend to isolate into islets more likely and can reveal potential clusters.

The workflow is LLM model-agnostic and can be applied to alternative embedding models. Any of the current more than 10.000 sentence transformers referred in <https://sbert.net/>² can be used. All experiments were conducted using publicly available models and reproducible preprocessing steps. These included text normalisation, lowercasing, and consistent token preparation prior to embedding. The resulting visualisation, in which the layout optimizes preservation of defined distances between elements, is shown in Figure 2.

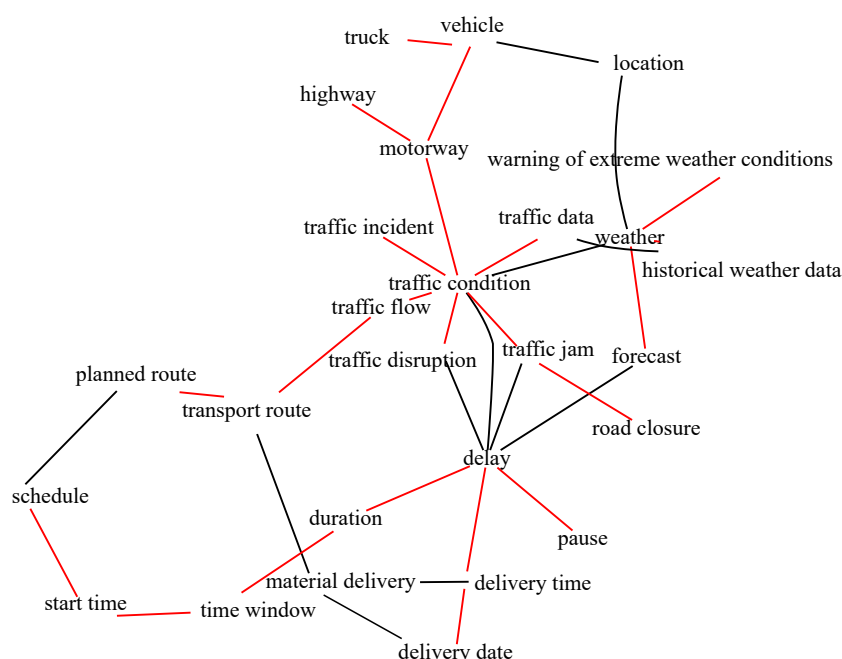


Figure 2: Semantic similarity visualisation of the open mobility data terms.

4.4 Internal-External Concept Alignment

Internal and external terms were embedded jointly to analyse cross-domain semantic alignment. Cosine distances between the two sets were computed, and the same similarity threshold selection strategy described in Section 4.3 was applied to ensure structural coherence while avoiding semantically weak connections. The resulting graph is shown in Figure 3.

²<https://sbert.net/>

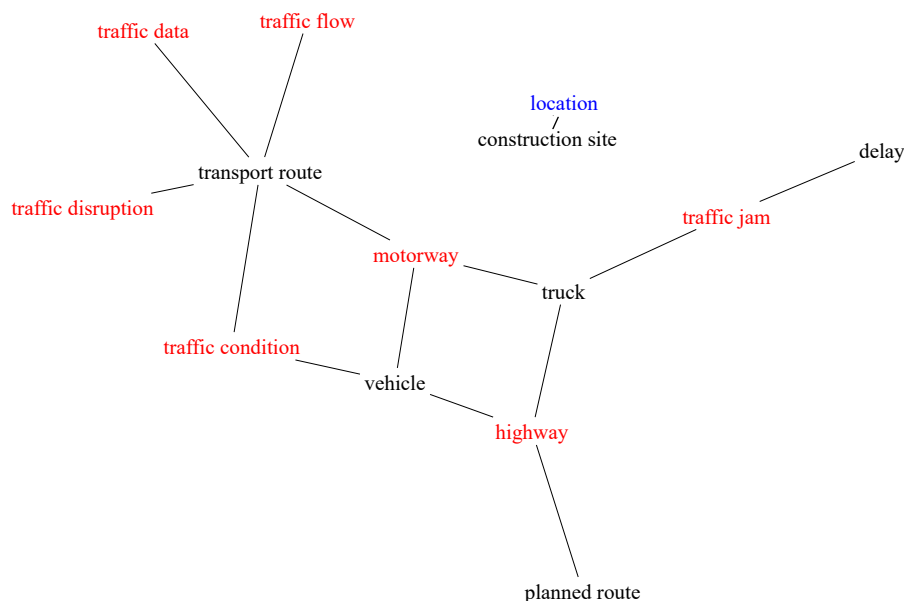


Figure 3: Semantic similarity between internal and external domain terms.

4.5 Identifying Relevant External Data Sources

Descriptions of external datasets were processed, normalised, and embedded using the same model to enable consistent comparison. Similarity scores between dataset embeddings and domain terms were computed following the pipeline illustrated in Figure 4.

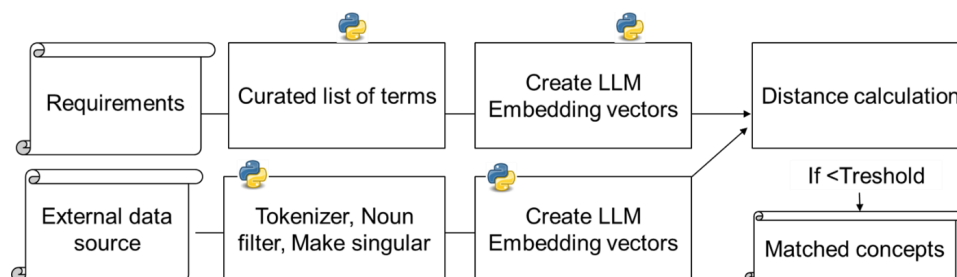


Figure 4: Pipeline used to identify relevant external datasets based on semantic similarity.

4.6 LLM-Based Ontology Generation Experiments

We evaluated whether large language models can generate ontology drafts from the requirement text. Microsoft Copilot, Meta’s 70-billion-parameter general-purpose model, and DeepSeek’s 8-billion-parameter reasoning-focused model were used to generate ontology candidates. They were qualitatively and manually evaluated for term coverage, structural coherence, and the detection of semantic outliers. The method is comparable to Retrieval-Augmented Generation (RAG), a framework that improves Large Language Model (LLM) outputs by fetching relevant, up-to-date information from external data sources.

4.7 Ontology Reuse Detection

The Resource Description Framework (RDF) is a framework for representing information as triples—subject, predicate, and object. It enables interoperability by providing a common,

graph-based model for sharing and connecting data across systems. `rdfs:label` and `rdfs:comment` are the standard Resource Description Framework Schema (RDFS) properties for giving human-readable names and descriptions to ontology classes. Additional annotation properties, such as SKOS or Dublin Core terms, and contextual data, can extend these descriptions when richer metadata is required.

RDF labels from the km4city ontology³ were embedded using the same sentence-transformer model as an experiment. The ontology was selected for its general topic relevance, and future work on complete set of building environment-related ontology extracts is planned. Nearest-neighbour matching was performed to identify ontology classes with high semantic similarity to the extracted domain terms. Table 1 summarises the closest matches identified through this procedure.

Table 1: Closest km4city ontology concepts for selected domain terms.

Domain Term	Ontology Concept	Cosine Distance
traffic flow	Traffic Flow Observation	0.13
highway	Highway stop	0.18
highway	Road	0.18
traffic data	Traffic observation	0.20
planned route	Route	0.22

4.8 Refined Ontology

A refined ontology for external mobility data was produced by consolidating insights from embedding-based analyses, dataset comparisons, and reuse identification. It specifies the structure of external datasets and how their elements relate to one another, with explicit links to the internal logistics concepts identified in the baseline ontology. The updated model is shown in Figure 5.

5 Findings

The embedding-based analysis yielded several insights that guided the refinement of the ontology and validated the usefulness of LLM-derived semantic similarity for integrating construction-traffic data. These findings emerged from term clustering, cross-domain alignment, external dataset matching, reuse analysis, and preliminary experiments with LLM-generated ontologies.

5.1 Semantic Structure of the Domain

The clustering results in Figure 2 revealed three interpretable semantic groups: (i) traffic-related concepts, (ii) temporal and scheduling concepts, and (iii) weather- and environment-related concepts. The major cluster structures remained stable under small variations of the similarity threshold during exploratory analysis. A small number of terms appeared as outliers, suggesting that they are either weakly related to the ontology scope or require more contextual refinement in the requirement documentation.

³<http://www.disit.org/km4city/schema#>

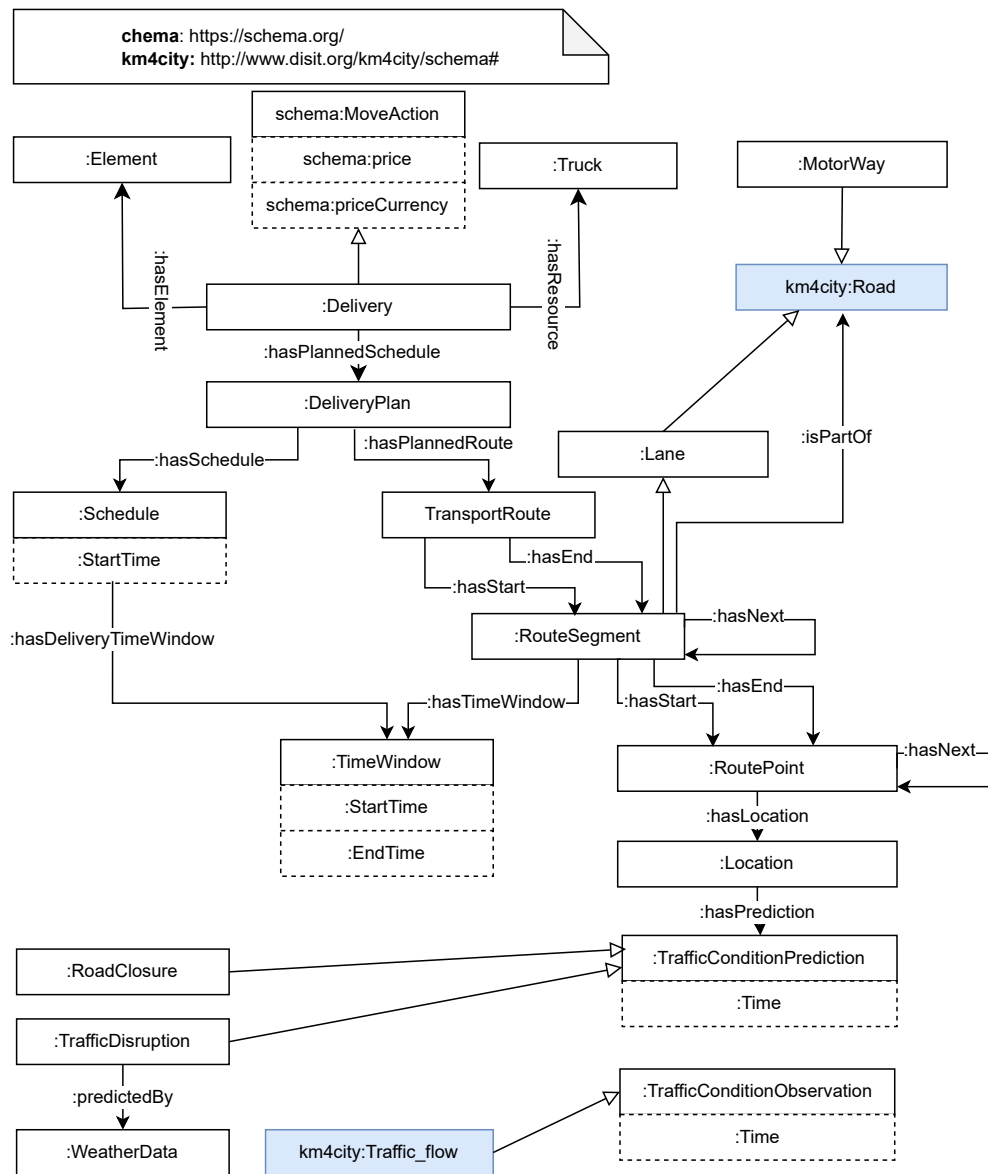


Figure 5: The refined ontology after embedding-based analysis and reuse integration.

5.2 Cross-Domain Alignment Insights

The joint embedding of internal and external terms, visualised in Figure 3, identified several key bridging concepts between construction logistics and open data sources. *transport route* and *planned route* aligned closely with external traffic concepts such as *traffic flow*, *traffic disruption*, and *road closure*. These connections confirm that embedding-based similarity can successfully expose conceptual junctions that a manual review may overlook. Polysemous terms, such as *forecast*, which are linked to both weather data and delivery scheduling, highlight the need for disambiguation mechanisms during ontology design.

5.3 Detection of Relevant External Data Sources

The embedding-based comparison of domain terms with dataset descriptions uncovered relevant external datasets that would not have been identified through keyword matching alone. Semantically similar datasets included lane-level traffic information, traffic volume observations, and geospatial road attributes. One concrete example is the identification of the concept *Lane*, which did not appear in the requirement text but is present in several mobility datasets. This insight directly motivated an ontology extension, demonstrating the value of embedding-driven discovery for aligning the ontology scope with real data availability.

5.4 Ontology Reuse Opportunities

The reuse analysis based on km4city embeddings (Table 1) identified several well-aligned concepts. The closest matches included *Traffic Flow Observation*, *Road*, and *Route*, which were subsequently incorporated into the refined ontology. This illustrates that embedding-based similarity can act as an effective filter to highlight semantically relevant elements from large existing ontologies, reducing lookup effort and supporting consistent reuse.

5.5 Evaluation of LLM-Generated Ontology Drafts

Experiments with models such as Microsoft Copilot, Llama3-70B, and DeepSeek-r1-8B showed that LLMs can propose coherent high-level ontology structures but are not yet reliable for fully automated ontology generation. Typical issues included out-of-scope classes, incorrect parent-child relationships, inconsistent property definitions, and occasional degradation into plain-text summaries. These observations confirm that LLMs require post-processing steps, validation metrics, and domain constraints to produce usable ontology artefacts. Nevertheless, the generated class descriptions (RDF:label) proved particularly useful for refining naming conventions.

To mitigate these issues, structural validation steps were applied to the generated ontology drafts. These included detection of orphan classes, inspection of subclass consistency, verification of mandatory object property links, and comparison of term coverage against the manually developed baseline ontology.

5.6 Impact on the Refined Ontology

The integration of these findings resulted in a refined ontology, shown in Figure 5, with several improvements: (i) the introduction of lane-level modelling and route segment structures; (ii) clearer temporal modelling through explicit time window and schedule concepts; (iii) strengthened links between traffic disruptions, weather effects, and delivery routes; (iv) alignment of start and end points of routes with location-based concepts; and (v) incorporation of km4city reuse candidates.

Cost-related properties (e.g., `schema:price` and `schema:priceCurrency`) were additionally integrated to represent transportation cost constraints in prefabricated construction logistics.

Compared to the initial baseline ontology (Figure 1), the refined model introduces explicit lane-level modelling, tighter integration of time-window semantics, formally linked reuse elements from km4city, and cost-related attributes reflecting the economic dimension of transport decisions. These modifications directly reflect embedding-driven insights from clustering, cross-domain alignment, and reuse detection.

Together, these enhancements produce a more coherent, data-aligned, and interoperable ontology that better reflects the structure of available mobility datasets and the operational and economic requirements of construction logistics.

6 Discussion and Conclusions

The findings of this study show that LLM-based embeddings can effectively support ontology engineering. The term clusters and cross-domain alignments indicate that the embedding space captures meaningful semantic structures of the construction–traffic domain. The comparison with external dataset descriptions enabled the identification of relevant data sources that keyword matching would miss, while the reuse analysis demonstrated that embedding similarity can surface appropriate ontology fragments from existing vocabularies.

In this context, optimisation refers to the systematic refinement of ontology scope through informed structural adjustments rather than to a numerical optimisation procedure.

These results are particularly relevant for JIT construction, where the transport of prefabricated elements depends on accurate, time-critical coordination. The alignment between logistics requirements and external mobility data suggests that the refined external-data ontology provides a coherent semantic layer for representing mobility information in a form that can be mapped to lean construction models. These structural refinements provide a semantically consistent foundation for improving the reliability of planning and execution in time-sensitive construction workflows.

The study also highlights several limitations. Polysemous terms, such as *forecast*, show that embeddings alone cannot resolve contextual meaning. Hallucinations also exist in embeddings, since they compress meaning into fixed-length vectors; they may not encode logical operators; they primarily capture surface-level semantic space, and their similarity metrics can't detect factual correctness.

Experiments with LLM-generated ontology drafts further indicate that large models can produce useful class descriptions but remain structurally inconsistent, often hallucinating relations or collapsing into plain-text output. These issues require explicit validation mechanisms and human oversight.

Methodological choices—such as model selection, similarity thresholds, and term-extraction criteria—constitute additional constraints. The reuse analysis considered only one external ontology, and the method primarily operates on isolated terms, thereby limiting semantic depth.

Future work should explore the use of richer textual context, domain-specific fine-tuning, and hybrid approaches that combine automated suggestions with expert review. Dynamic ontology maintenance and instance-level population supported by LLMs also represent promising directions, provided that robust quality-control mechanisms are available.

In conclusion, LLM-derived embeddings can meaningfully support semi-automated ontology development by identifying conceptual clusters, linking internal and external concepts, discovering relevant datasets, and suggesting reuse candidates. Although full automation is not yet feasible, the approach provides a structured methodology for semantic alignment in heterogeneous data-integration settings.

The embedding-based workflow itself remains domain-independent, as it relies on generic semantic similarity computation and graph-based thresholding. Domain specificity arises primarily from requirement extraction and dataset selection. This separation supports transferability of the methodological framework to other data-integration domains.

7 Data Availability

The datasets and source code will be published upon acceptance of the paper.

8 Acknowledgments

This work is part of the research project OnTimeOlogy, funded by the Federal Ministry of Transport (BMV) within the mFUND innovation initiative (funding number 01F1219A).

References

- Babaei Giglou, H., D'Souza, J. & Auer, S. (2023), Llms4ol: Large language models for ontology learning, *in* 'International Semantic Web Conference', Springer, pp. 408–427.
- Bakker, R. M., Di Scala, D. L. & de Boer, M. (2024), Ontology learning from text: an analysis on llm performance, *in* 'Proceedings of the 3rd NLP4KGC International Workshop on Natural Language Processing for Knowledge Graph Creation, colocated with Semantics', pp. 17–19.
- Brell-Cokcan, S. & Schmitt, R. H., eds (2024), *IoC – Internet of Construction: Informationsnetzwerke zur unternehmensübergreifenden Kollaboration in den Fertigungsketten des Bauwesens*, 1 edn, Springer Vieweg, Wiesbaden.
URL: <https://doi.org/10.1007/978-3-658-42544-9>
- Chávez-Feria, S., García-Castro, R. & Poveda-Villalón, M. (2022), Chowlk: from uml-based ontology conceptualizations to owl, *in* 'European Semantic Web Conference', Springer, pp. 338–352.
- Guizzardi, G. (2020), 'Ontology, ontologies and the "i" of fair', *Data Intelligence* **2**(1-2), 181–191.
- Khan, A., Shah, Q., Uddin, M. I., Ullah, F., Alharbi, A., Alyami, H. & Gul, M. A. (2020), 'Sentence embedding based semantic clustering approach for discussion thread summarization', *Complexity* **2020**(1), 4750871.
- Kommineni, V. K., König-Ries, B. & Samuel, S. (2024), 'From human experts to machines: An llm supported approach to ontology and knowledge graph construction', *arXiv preprint arXiv:2403.08345*.
- Noy, N. F. & McGuinness, D. L. (2001), 'Ontology development 101: A guide to creating your first ontology', *Stanford Knowledge Systems Laboratory Technical Report KSL-01-05 and Stanford Medical Informatics Technical Report SMI-2001-0880*. Available at: https://protege.stanford.edu/publications/ontology_development/ontology101.pdf (Accessed : 14April2026).
- Peiris, P. A. N., Hui, F. K. P., Ngo, T., Duffield, C. & Garcia, M. G. (2023), Challenges in transport logistics for modular construction: A case study, *in* R. Dissanayake, P. Mendis, K. Weerasekera, S. De Silva, S. Fernando & C. Konthesingha, eds, '12th International Conference on Structural Engineering and Construction Management', Springer Nature Singapore, Singapore, pp. 501–510.
- Reimers, N. & Gurevych, I. (2024), 'all-minilm-l6-v2: Sentence-transformers model'. Available at: <https://huggingface.co/sentence-transformers/all-MiniLM-L6-v2> (Accessed: 2025-11-26).
- Siino, M. (2024), All-mpnet at semeval-2024 task 1: Application of mpnet for evaluating semantic textual relatedness, *in* 'Proceedings of the 18th International Workshop on Semantic Evaluation (SemEval-2024)', pp. 379–384.
- Simperl, E., Mochol, M., Bürger, T. & Popov, I. O. (2009), Achieving maturity: The state of practice in ontology engineering in 2009, *in* R. Meersman, T. Dillon & P. Herrero, eds, 'On the Move to Meaningful Internet Systems: OTM 2009', Springer Berlin Heidelberg, Berlin, Heidelberg, pp. 983–991.

- Tao, C., Shen, T., Gao, S., Zhang, J., Li, Z., Hua, K., Hu, W., Tao, Z. & Ma, S. (2024), 'Llms are also effective embedding models: An in-depth overview', *arXiv preprint arXiv:2412.12591* . revised 2025.
- Uschold, M. & King, M. (1995), Towards a methodology for building ontologies. Available at: <https://api.semanticscholar.org/CorpusID:13963021> (Accessed: 2025-11-28).
- Wang, H., Yi, W., Zhen, L. & Chan, A. P. (2024), 'Cost-optimized transport planning for prefabricated modules with the implementation of the just-in-time strategy', *Journal of Construction Engineering and Management* **150**(8), 04024093.