2025 European Conference on Computing in Construction

CIB W78 Conference on IT in Construction

Porto, Portugal
July 14-17, 2025

ENHANCING CONSTRUCTION DATA INTEGRATION THROUGH DYNAMIC ONTOLOGY ALIGNMENT AND AUTOMATED ATTRIBUTE MAPPING

Jakob Martin[1], Cornelius Preidel[1], Simon Vilgertshofer[1], and André Borrmann[2]
[1]University of Applied Sciences, Munich, Germany
[2]Technical University of Munich, Munich, Germany

Abstract

The construction industry struggles with data integration due to semantic heterogeneity across stakeholders and lifecycle phases. Users must handle complex formats, inconsistent naming conventions, and diverse standards, detracting from core tasks. This work presents a hybrid matching system combining lexical similarity with domain-specific pattern recognition to align terms across terminologies. Automated enrichment from standardized sources and semantic fingerprinting enable property mapping between stakeholders, domains, and formats. A Human-in-the-Loop component ensures continuous refinement via expert feedback. Real-world evaluation in facility management scenarios achieved 86.6% initial accuracy in attribute matching and significant query processing time reductions over manual approaches.

Introduction

The construction industry faces a fundamental challenge in effectively leveraging digital processes: while digitization has introduced increasingly advanced data models, tools, and standards, their inconsistent and often user-unfriendly implementation often creates barriers for end users. These users – whether manufacturers, contractors, or operators – are frequently required to interact with complex, fragmented data structures that are not aligned with their practical needs or workflows. This disconnect is particularly evident throughout the building lifecycle, where different stakeholders often operate in isolated workflows with limited integration, making seamless collaboration and data exchange difficult.

The increasing digitalization has led to an overwhelming flood of data that practitioners must navigate daily. When faced with vast amounts of digital documents, models, and information systems, stakeholders are often overwhelmed, lacking efficient ways to find and interpret the specific information they need.

This information overload significantly impacts their ability to perform core tasks effectively.

For instance, door manufacturers need to efficiently match their product specifications against requirements from architectural designs to propose suitable products for construction projects. In today's digital, model-based processes, these manufacturers receive digital models – often in formats like Industry Foundation Classes (IFC) – from different architectural offices, each using distinct attribute naming conventions and structures. Before performing this essential matching process, manufacturers must interpret these diverse conventions and map them to their internal product ontology, creating unnecessary overhead in a straightforward product selection process. This process, however, remains largely manual and requires significant effort to match attributes and harmonize data, posing a major challenge to efficiency and scalability.

In practical terms, these data integration issues directly affect day-to-day operations. Studies indicate that facility managers can spend up to 25 % of their working time locating and interpreting building information scattered across various platforms and labeled inconsistently (Becerik-Gerber et al., 2012). This delays critical interventions – such as repairing malfunctioning fire safety systems or scheduling routine structural inspections –and increases administrative overhead. Recent industry analyses show that such inefficiencies can increase operational costs by up to 30% (Patacas et al., 2020) and significantly impact a building's overall performance by hindering proactive maintenance strategies and limiting the ability to fully capitalize on digital data models.

Stakeholders throughout the building lifecycle require seamless access to role-specific information – typically just a small subset of the project's complete knowledge base – without delving into the intricacies of data structures or standards. Users must interact with complex data formats like IFC and building management systems, often requiring technical expertise that diverts them from their core responsibilities.

A significant challenge in working with building information is that users often struggle to locate the data they need because it is stored in unexpected places or formats. For instance, a facility manager searching for ''lighting systems that need replacement in the next maintenance cycle'' might find them labeled as *luminaire* in an IFC model, *light fixture* in a maintenance tool, and *lamp* in component specifications. Such inconsistencies require users to develop technical expertise that distracts from core tasks. Our approach addresses this by scanning multiple sources, returning a concise list of rooms and item IDs so that maintenance orders can be generated directly. Once the manager

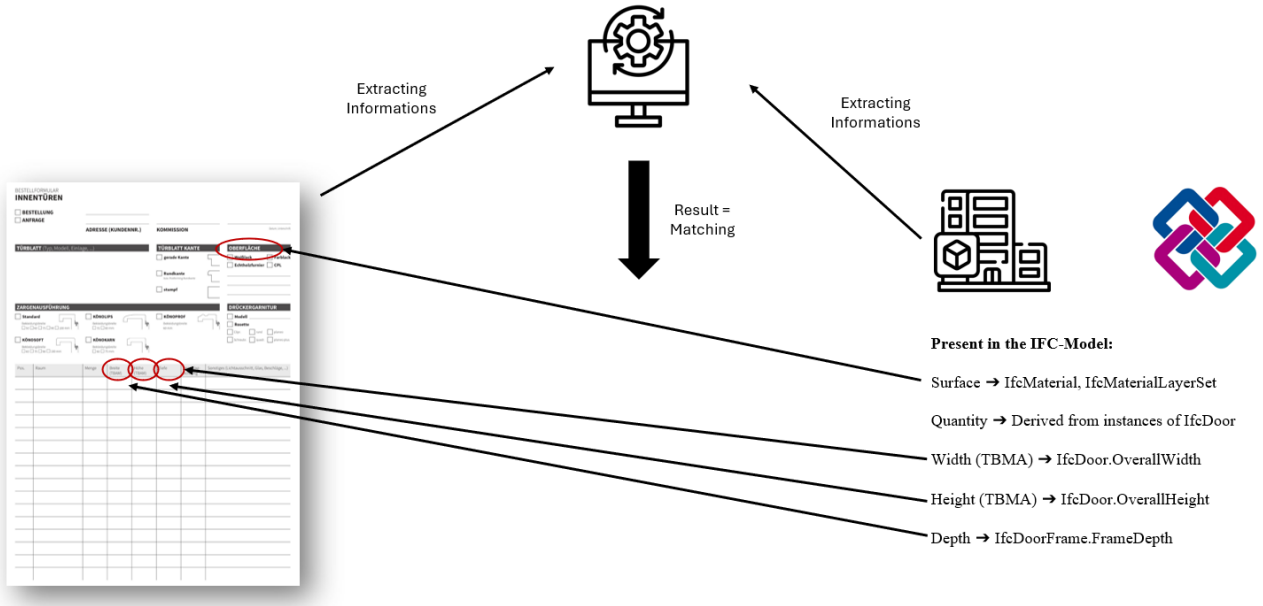

Figure 1: Overview of the data mapping process: Extracting information from IFC models and aligning it with target specifications like product order forms.

confirms which terms are correct synonyms and which are too general, the system learns and adapts (e.g., treating "lamp" and "light fixture" as equivalent), improving future queries.

Various approaches have been developed to address data integration challenges in the industry: data exchange standards (like IFC and COBie), knowledge representation frameworks (such as Resource Description Framework (RDF)/Web Ontology Language (OWL)), and, more recently, artificial intelligence solutions. These AI approaches include natural language processing for terminology matching, machine learning for pattern recognition in data structures, and deep learning models for semantic understanding of building information. While these technologies serve specific purposes in structuring and standardizing building information, they often focus more on technical implementation than practical usability for end users. They effectively structure building information, but were not designed to address end-users' challenges in accessing and interpreting this data. A different approach is needed to help users effectively work with these complex data structures.

The primary motivation for this work is to make building information seamlessly usable by all stakeholders, regardless of their technical expertise. By focusing on intuitive user interaction rather than technical complexity, we aim to streamline how building information is accessed and used throughout the lifecycle. With this goal in mind, we developed an approach that combines three complementary components to help users quickly find and work with building information:

- a matching system that bridges technical terminologies and industry-specific vocabularies with the long-term vision of supporting complex natural language queries
- Integration with multiple standardized sources, including classification systems, product catalogs, and domain dictionaries (such as buildingSMART Data Dictionary (bSDD)) to ensure consistent interpretation of building terms
- a HITL component that incorporates expert feedback to improve matching accuracy and adapt to organization-specific terminology continuously

The implemented system maps and aligns building information definitions through pattern recognition and semantic similarity analysis. It automatically identifies equivalent terms and characteristics across different data sources, from properties like *door_width*, *clear_opening*, and *passage_dimension* to broader concepts and domain-specific terminology, enabling standardized access without manual mapping. Our evaluation in facility management test scenarios demonstrates 86.6 % accuracy in matching with significant improvements in query processing time compared to manual selection and mapping approaches. While these initial results from our Proof-of-Concept implementation are promising, future work aims to extend the system toward natural language query processing. This extension would allow users to ask questions in their own words – for example, *show me all doors wider than 1 meter* – with the system handling the complex task of finding this information across different data sources and formats.

Use Case Scenarios and User Groups

The proposed approach addresses two primary user scenarios, reflecting the varying levels of formalization in their data workflows:

- **Stakeholders with well structured data or domain-specific ontologies:** such as manufacturers that maintain detailed product catalogs with standardized naming conventions or large organizations with advanced BIM workflows. Here, the system leverages existing ontologies to automate property mapping and reduces manual alignment effort.
- **Stakeholders with less formal or unstructured data:** for example, smaller facility management teams working with ad-hoc Excel files, PDF manuals, or text documents. The system's NLP and pattern-recognition components help to parse and align these informal terminologies with formal references without requiring extensive BIM or ontology expertise.

Although we describe these two extreme cases for clarity, real-world scenarios often fall somewhere in between. Some companies have partial data models or rely on both standardized properties and free-text annotations. Furthermore, it is useful to distinguish between data providers (e.g., manufacturers) and data consumers (e.g., facility managers) to clarify each group's primary interactions with the system. Nonetheless, the overarching goal remains the same: to provide seamless data integration and retrieval irrespective of the underlying data structure or the user's technical background.

Despite their differing contexts, both user groups share the common challenge of accessing information without directly interacting with the complexities of data models. The system bridges this gap by effectively utilizing formalized data structures for ontology users while interpreting unstructured queries from users without ontologies. By accommodating formalized and informal data structures, the proposed approach aims to shield users from underlying technical complexity while ensuring seamless, user-centric alignment of information. The system adapts to user-preferred terminologies dynamically, automates data integration across systems, and maintains consistency—effectively bridging the gap between diverse user needs and complex data systems while reducing technical barriers to entry.

Recent Technical Advances

Recent advancements in construction data integration have introduced a range of technological approaches, each addressing specific challenges within the Architecture, Engineering, and Construction (AEC) domain. These efforts can be categorized into semantic web technologies, query languages, and artificial intelligence-based methods, reflecting their underlying principles and historical development.

The Semantic Web aims to create a structured, machine-readable web of data, enabling seamless information exchange and interoperability across domains. Central to this vision are ontologies, which define formalized relationships between data entities and allow a deeper understanding of the data. For example, technologies such as RDF and OWL will enable the creation of detailed ontologies that capture complex interconnections within building information models (Pauwels et al., 2015). Querying such structured data is made possible by SPARQL, a query language for retrieving and manipulating linked data. Frameworks like the Building Topology Ontology (BOT) extend these capabilities specifically for the AEC sector, promoting interoperability between different stakeholders and tools (Rasmussen et al., 2017). However, adopting semantic web technologies often demands significant technical expertise, which limits their widespread acceptance, particularly among non-technical stakeholders such as facility managers (Curry et al., 2013). Additionally, these approaches are typically designed for structured data and may struggle to handle the unstructured or semi-structured data commonly encountered in real-world workflows.

Query languages provide powerful tools to streamline information retrieval from complex BIM models. Textual solutions, such as BIMQL (Mazairac and Beetz, 2013), simplify querying by allowing users to define model-specific information requirements without needing deep technical knowledge of IFC structures. Similarly, visual query languages, such as the Visual BIM Query Language (VBQL), offer graphical interfaces to specify filter queries, reducing the dependence on programming skills and making BIM data more accessible to non-BIM specialists, such as craftspeople (Wülfing et al., 2014). However, all query languages, whether textual or visual, share significant limitations: they require users to possess a minimum level of technical expertise, which many stakeholders lack, and they often fail to harmonize diverse property concepts or support semantic alignment across heterogeneous data sources. While these tools improve data accessibility, they cannot address the broader challenges of integrating unstructured data or enabling intuitive user interaction across domains.

Building on the concept of BIM querying, (Daum and Borrmann, 2014) propose a method for handling topological relationships in BIM. However, like other query languages, these methods demand programming knowledge and are not readily applicable to users without BIM programming expertise, such as facility managers. Additionally, they focus primarily on geometric aspects rather than addressing semantic heterogeneity or cross-domain data alignment.

Recent advancements leverage artificial intelligence (AI) and machine learning (ML) to automate and enhance data integration processes. These approaches focus on minimizing manual efforts by identifying patterns, aligning properties, and uncovering relationships in data. For instance, natural language processing (NLP) models have been developed to map terms based on semantic similarity, while clustering algorithms identify corresponding entities across heterogeneous sources (Zhang et al., 2018). Nevertheless, these methods often struggle with domain-specific nuances, such as abbreviations or implicit relationships, which can impact alignment accuracy (Shalabi

and Turkan, 2020). Furthermore, many AI-based methods, such as the deep learning approach proposed by (Yin et al., 2023), remain limited to batch-style ontology enrichment for IFC data and lack mechanisms for real-time adaptation or continuous feedback loops from domain experts.

While current solutions offer foundational capabilities, workflows across construction, facility management, and other domains require a more comprehensive solution. An effective framework must dynamically align ontology concepts to accommodate evolving terminologies and feedback, harmonize formal data models with unstructured naming conventions, and continuously adapt based on real-world usage scenarios. These capabilities are critical to bridge the gap between technical data systems and practical user needs, enabling stakeholders to interact with complex data structures intuitively and effectively.

Concept for an Integrated Framework

To address the challenge of semantic heterogeneity in construction data, we propose an integrated framework that combines three complementary approaches to create a robust and adaptable system for data alignment.

At its core, the framework employs multi-layered semantic integration to create a unified data environment where structured and unstructured information can seamlessly coexist and interact. This integration layer closes the gap between formal data models like IFC and informal documentation such as maintenance logs, enabling the system to recognize and align equivalent terms across different sources - for instance, automatically connecting concepts like *door width* with *clear opening*.

Building on this foundation, the framework implements dynamic alignment mechanisms that continuously adapt to evolving terminology and user-specific conventions. Rather than relying on static mappings, the system leverages HITL feedback to refine understanding iteratively. Domain experts provide validation and enrichment of proposed mappings, allowing the system to learn from real-world usage patterns and evolve its alignment rules accordingly. This adaptive approach ensures the framework remains relevant and accurate as industry terminology and user needs change.

The third key component employs advanced pattern recognition and contextual understanding to strengthen semantic alignment accuracy. The system analyzes individual terms, relationships, and usage contexts through domain-specific rules and similarity scoring. For example, when processing a query for *lamps*, the system considers related technical attributes like wattage and mounting specifications to ensure comprehensive and accurate recognition across heterogeneous data sources.

Semantic integration, dynamic alignment, and contextual understanding create a robust foundation to support future extensions, including natural language queries and advanced semantic processing while maintaining its core focus on practical user needs, as illustrated in Figure 2.

Workflow

The framework follows a systematic process to transform user queries into actionable results. Query Interpretation analyzes user inputs to identify intent and contextual elements. Automated translation supports multiple languages, enabling precise extraction of relevant information.

Ontology Mapping then aligns these extracted parameters with corresponding entities within existing ontologies. This stage employs linguistic and contextual analysis to accurately map informal data to formal representations.

Result Aggregation and Presentation organize the mapped entities into logical groupings enriched with contextual annotations. This ensures users receive information in a format that is intuitive and actionable.

Finally, Feedback Integration processes expert and user input through the HITL mechanism. This iterative loop continuously improves the system's adaptability, accommodating real-world requirements and domain-specific changes.

The overall workflow of the system is depicted in Figure 4, showing the key processing steps from initial user input to the final result presentation.

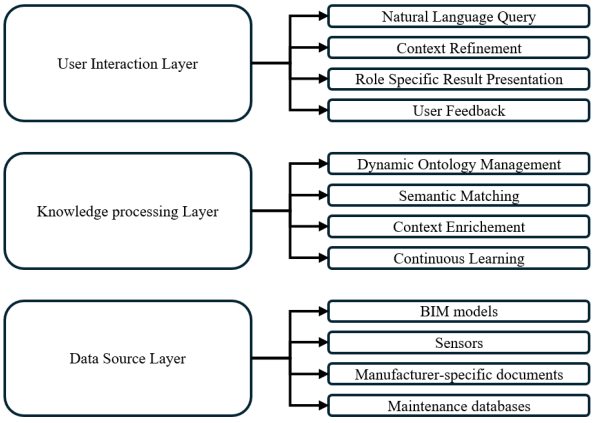

Figure 2: System architecture illustrating the interaction between the User Interaction Layer, Knowledge Processing Layer, and Data Source Layer.

HITL Feedback Mechanism

To improve accuracy over time, the proposed framework incorporates a Human-in-the-Loop component that enables continuous refinement through user validation. This iterative process enhances the reliability of future queries by learning from real-world use cases and domain expert input.

When a user submits a query, the system generates potential matches based on its current knowledge base. These suggested matches are then presented to the user for validation, where they can confirm correct matches, reject incorrect ones, and provide additional feedback or corrections. The validations are stored in the system's feedback repository, along with details such as the specific query, suggested matches, and the user's domain exper-

tise level. Periodically, the system processes this accumulated feedback to update its knowledge base. Matches consistently validated by multiple users, particularly those with high domain expertise, are assigned higher confidence scores, while rejected matches have their scores reduced. These updated confidence scores refine future query results, with higher-scoring matches being prioritized and lower-scoring matches being suppressed or presented with annotations.

In practice, feedback overhead remains manageable if the system is deployed in an organization with multiple users (e.g., 10–20 employees), each contributing occasional validations. To further reduce effort, an adaptive feedback loop only requests input on uncertain or new terms. Moreover, pilot deployments indicate that organizations are motivated to provide feedback, as improving data quality early on translates into measurable time and cost savings later.

Technical Architecture

The core system architecture implements a hybrid matching approach combining lexical similarity measures, pattern recognition, and domain-specific rule matching. (see Figure 3). The matching process utilizes complementary techniques: The Soft Pattern Matcher identifies technical terms and patterns, while The Enhanced Embedding handles semantic similarity through SciBERT embeddings. The Enhanced Preprocessor standardizes measurement units and handles compound terms through defined patterns like $\mathrm{W}/(\mathrm{m}^2 \cdot \mathrm{K})$ or kg/m^3.

The pattern matching engine, implemented in Soft Pattern Matcher, utilizes spaCy's *en_core_web_sm* model for linguistic analysis and extracts definition characteristics through rule-based matching. It identifies technical content via predefined patterns for measurements, ranges, equations, and technical symbols, preserving them during preprocessing through a placeholder mechanism. The system recognizes domain-specific terms across thermal, mechanical, geometric, material, and chemical domains, with specialized pattern sets for each category.

The property manager implements data persistence through JSON storage, using MD5 hashing for unique property identification. It maintains property mappings, relationships, and a caching system optimized for frequent property lookups. External data integration occurs through an asynchronous GraphQL client for buildingSMART Data Dictionary queries, implementing automatic retrying (max_retries=3) with exponential backoff for failed requests.

As an example, we consider a facility manager searching through a vast collection of digital documents, models, and building information systems for *luminaires*. Faced with this large amount of data, the system helps by initially suggesting matches such as *light fixtures*, *lamps*, and *lighting equipment*, based on linguistic similarity. The manager validates *light fixtures* and *lamps* as relevant but rejects *lighting equipment* as too broad. Additionally, they provide a new term commonly used in their organization:

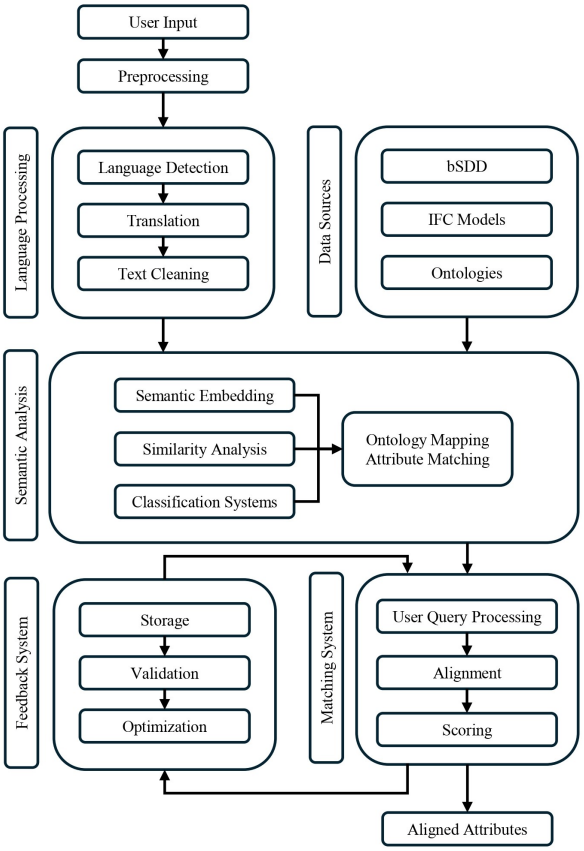

Figure 3: System architecture illustrating the interaction between Language Processing, Semantic Analysis, Matching System, and Feedback System.

lighting units. This feedback is stored, and as more users validate similar mappings, confidence scores for *light fixtures* and *lamps* increase, while the score for *lighting equipment* decreases. The new term *lighting units* is added to the knowledge base and appears in subsequent query suggestions. Through this iterative refinement, the system evolves to better reflect the terminology and needs of its users rather than relying solely on predefined standards or generic linguistic models. By incorporating domain expertise via the HITL approach, the framework bridges the gap between formal ontologies and practitioners' diverse, often informal, terminology. This adaptive, user-centric methodology represents a key strength of the proposed system, enabling continuous evolution and enhanced data integration capabilities aligned with real-world usage patterns and expert input.

Evaluation and Results

The system was tested using a dataset of 275 building properties, divided into four primary domains—thermal (75 properties), chemical (50 properties), mechanical (75 properties), and geometric (75 properties). A multi-component scoring methodology was used to evaluate classification accuracy and overall system performance, as summarized in Table 1.

The scoring mechanism integrates four components to

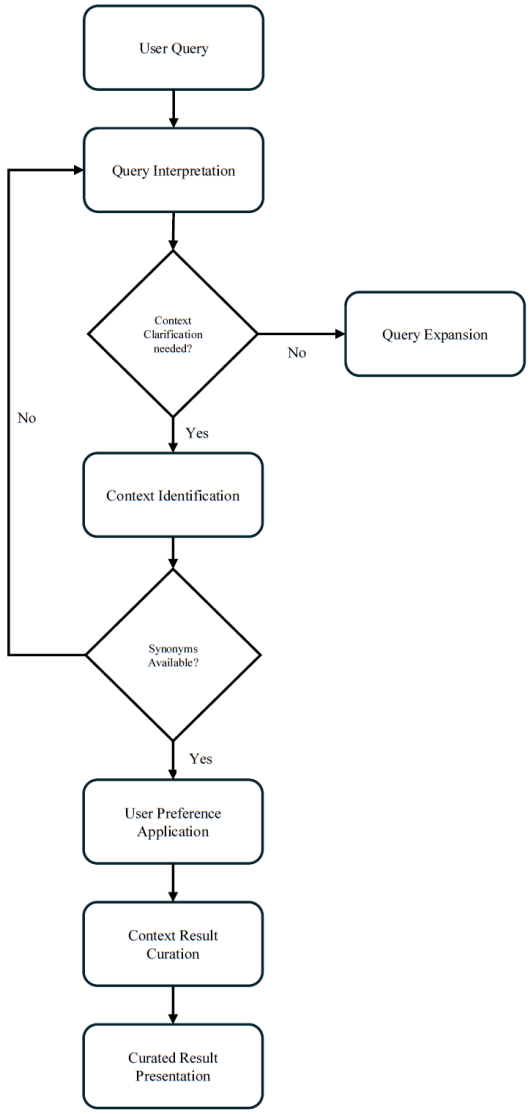

Figure 4: Workflow of query processing: From user input to curated result presentation, highlighting key steps such as query interpretation, context identification, and user preference application.

evaluate property classifications, each weighted equally at 25%. This equal weighting strategy was selected after evaluating different weighting configurations and was found to strike the best balance between robustness and simplicity. Specifically, the pattern-based and keyword-based modules are rule-driven, the ML component is data-driven, and the HITL mechanism reflects user feedback. By distributing weights evenly across these four components, the system avoids dominance by any single method and remains resilient, even when one module (e.g., the ML component) exhibits weaker performance due to limited training data.

Pattern Matching (25 % weight).
This component analyzes syntactic structures and domain-specific patterns. To prevent overconfidence, its score is capped at 0.9. We begin with a base of 0.3, then add 0.15 for each additional detected pattern (up to 0.3 total). If this interim sum exceeds 0.4, an extra 0.2 is applied, reflecting stronger confidence.

Keyword Matching (25% weight):
This component gauges domain-specific terminology according to a hierarchical scheme, capped at 0.9 to avoid overconfidence. High-priority terms receive a weight of 0.35, medium-priority terms 0.25, and low-priority terms 0.15, ensuring that key technical expressions are more strongly emphasized than generic or ancillary terms.

Machine Learning Confidence (25% weight):
Incorporates predictions from a TF-IDF/Random Forest classifier trained on domain-specific data, yielding a score between 0.0 and 1.0. This component ensures data-driven validation for ambiguous cases. If this score is weak, the rule-based components (Pattern/Keyword) and user feedback (HITL) can still maintain overall robustness.

Human-in-the-Loop Feedback (25%):
This module starts at a default score of 0.5 in the absence of user feedback and transitions to the ratio of correct vs. total validations once data is available. Such incremental refinement allows the system to incorporate expert knowledge over time, preventing classification drift. Finally, pattern and keyword scores are each capped at 0.9, and the total is limited to 0.95, ensuring no component creates an illusion of absolute certainty.

Boosting Factors
When both the pattern and keyword matching modules yield sufficiently high scores (e.g., above 0.7), we apply a 20% boost to the final classification confidence.
In more standardized technical domains (thermal, mechanical, geometric, chemical), the threshold is slightly reduced (e.g., 0.6), and a smaller 15% boost is used.
This mechanism leverages consistent domain-specific nomenclature to increase overall accuracy while preventing inflated confidence in uncertain scenarios.
Putting it all together, the complete scoring formula (with a final cap of 0.95) is:

$$\begin{aligned} S_{\text{final}} = \min\Big(\Big(&0.25 \times S_{\text{pattern}}[0{-}0.9] \\ &+ 0.25 \times S_{\text{keyword}}[0{-}0.9] \\ &+ 0.25 \times S_{\text{ml}}[0{-}1.0] \\ &+ 0.25 \times S_{\text{hitl}}[0{-}1.0]\Big) \times \text{boost\_factors}, 0.95\Big) \end{aligned} \tag{1}$$

Where boost_factors is the product of all applicable boosts (1.2 and/or 1.15, under the conditions noted above). The final limit of 0.95 leaves a small margin for uncertainty, even in very strong matches.

Table 1: Domain-Specific Results

Domain	Accuracy	Pattern Matching	Keyword Matching
Thermal	95.0%	83.3%	86.7%
Chemical	95.0%	90.0%	90.0%
Mechanical	82.3%	90.0%	58.3%
Geometric	74.1%	65.0%	45.0%

Discussion

Lessons Learned

This project highlighted several key lessons for designing robust and adaptable scoring systems for semantic alignment. First, distributing weights equally (25 % each) ensures no single component dominates, creating a balanced and reliable framework. Conservative score caps help avoid excessive confidence and reduce the risk of overfitting, particularly in early iterations. Domain-specific findings revealed that standardized terminologies, such as in thermal or chemical domains, simplify matching, while less uniform domains, like geometric properties, require deeper contextual analysis.

HITL feedback proved invaluable, with a neutral starting score of 0.5, effectively preventing initial bias and enabling gradual learning from user input. Finally, modular design, combined with context-aware weighting and a JSON-based feedback system, enhances maintainability and scalability, ensuring the framework can evolve to meet diverse and dynamic requirements.

Limitations

Integrating the system with existing software ecosystems is non-trivial, especially where proprietary formats or legacy infrastructure lack modern APIs. Large-scale deployments may encounter performance issues when handling thousands of concurrent property queries or scaling the HITL feedback loop. Moreover, reliance on external data sources (e.g., bSDD) introduces dependencies that could impact reliability if these resources become unavailable or outdated.

From a methodical perspective, ambiguous or fast-evolving terminology can exceed the capabilities of current pattern-recognition rules. While standardized domains yield robust performance, synonyms and rapidly changing terms require repeated expert validation. Without consistent feedback prioritization, enterprise-level implementations risk an overwhelming volume of validation requests. Additionally, organizational hurdles – such as inconsistent naming conventions, incomplete metadata, or inadequate stakeholder engagement – can limit the system's effectiveness. Extending beyond building-focused domains further complicates adaptation, requiring substantial modifications to rules, ontologies, and feedback procedures.

Another important aspect is the tension between standardized schemas (e.g., IFC, bSDD) and allowing for custom ontologies. While recognized standards ensure robust interoperability, they often cannot capture rapidly changing or highly specialized domain requirements. By contrast, fully customized approaches can adapt quickly but risk diverging from industry norms and complicating data exchange. Our solution adopts a middle path: we rely on standards as a stable reference baseline, yet enable iterative refinement and incorporation of new terms through the HITL mechanism. This dual strategy preserves broad compatibility while remaining flexible enough to accommodate dynamic operational demands.

Future Work

While this paper has established a proof-of-concept for dynamic ontology alignment and user-centric data access, several areas for further development remain, especially regarding technical, methodological, and organizational considerations. Future research could explore structured input masks or interactive interfaces that guide users toward more specific queries. By prompting for contextual details such as component type or targeted attributes, these systems could handle ambiguous or evolving terminology more effectively and reduce confusion from inconsistent data entry. To further improve accuracy, an optional pre-alignment step could roughly assign terms between the ontology and data sources before the actual matching process. This step would filter potentially corresponding entities using standard vocabularies, taxonomies, or simple lexicon mappings, allowing the matching process to start with a more focused candidate set. Another promising direction is to explore an adaptive weighting scheme in the scoring mechanism. Where abundant, high-quality training data are available, the ML component could be given more weight, while in early-stage or data-scarce scenarios, user feedback (HITL) and rule-based matching would be more important. Such an adaptive approach would allow the system to dynamically adjust to varying data qualities and operational contexts, improving its effectiveness across diverse use cases. A more advanced interface could allow users to preview potential data alignments before final confirmation. For example, when searching for *Glühbirne* (light bulb), the system could display candidate attributes and request validation of key aspects such as wattage. This would refine terminology handling and improve the scalability of the HITL mechanism by prioritizing more uncertain matches. Furthermore, future solutions could integrate with industry-standard software like Solibri Model Checker or other BIM tools to enhance ecosystem compatibility and enable immediate, in-context insights. Extending support to legacy systems with limited APIs would also address persistent integration barriers, especially for organizations operating in mixed environments. As more users from diverse backgrounds provide

feedback, the system could accumulate domain- and user-specific terminology, gradually becoming more robust and transferable to fields beyond building construction, such as manufacturing or facility operations. However, significant domain adaptation would be required to achieve this. Further research could enhance keyword matching and pattern recognition by incorporating advanced NLP techniques, such as context-aware embeddings or domain-specific ontologies. These enhancements would better handle ambiguous terminology and improve accuracy, especially when data quality varies or external ontologies like bSDD undergo frequent updates. A comprehensive real-world evaluation is planned to investigate user acceptance across various roles (e.g., facility managers, manufacturers, BIM coordinators). We aim to conduct both quantitative (time savings, error rates) and qualitative (usability tests, focus groups) assessments to validate the approach's practical benefits and gather detailed feedback for refining the user interface design. This larger study will be crucial for confirming the system's effectiveness under real project constraints and diverse organizational contexts.

Conclusion

This work represents a step towards addressing semantic heterogeneity in construction data through dynamic ontology alignment and user-centric data access. Integrating automated matching with HITL feedback, the proposed framework seeks to reduce barriers between technical data systems and practical user workflows. Features such as balanced scoring mechanisms, modular design, and domain-specific adaptability form a basis for further exploration and refinement.

Preliminary evaluations indicate that the approach can improve accuracy and query efficiency in standardized domains, while challenges remain in addressing less uniform contexts and ensuring scalability across diverse applications. Future efforts will focus on refining these aspects, expanding multilingual support, and exploring the framework's potential in other industries.

This methodology offers an incremental step toward more accessible and adaptable data systems, supporting stakeholders in navigating complex construction workflows more effectively.

Data availability statement

The data supporting this study's findings are available at *https://gitlab.lrz.de/jmartin/nlp-matching-tool*, ensuring that they are Findable, Accessible, Interoperable, and Reusable (FAIR).

Acknowledgments

This research was supported by the Nemetschek Innovation Foundation and benefited from collaboration with industry partners within the framework of the openDBL project. The openDBL project is funded by the European Union's Horizon Europe research and innovation program under grant agreement No. 101092161, providing valuable test cases during development.

References

Becerik-Gerber, B., Jazizadeh, F., Li, N., and Calis, G. (2012). Application areas and data requirements for BIM-enabled facilities management. Journal of Construction Engineering and Management, 138(3):431–442.

Curry, E., O'Donnell, J., Corry, E., Hassan, S., Costa, A., and Keane, M. (2013). Linked data in architecture and construction. International Journal of 3-D Information Modeling (IJ3DIM), 2(2):38–51.

Daum, S. and Borrmann, A. (2014). Processing of topological BIM queries using boundary representation based methods. Advanced Engineering Informatics, 28(4):272–286.

Mazairac, W. and Beetz, J. (2013). BIMQL – an open query language for building information models. Advanced Engineering Informatics, 27(4):444–456.

Patacas, J., Dawood, N., and Kassem, M. (2020). Bim for facilities management: A framework and a common data environment using open standards. Automation in Construction, 120:103366.

Pauwels, P., Zhang, S., and Lee, Y.-C. (2015). Semantic web technologies in AEC industry: A literature overview. In Ontology-Based Modeling in Physical Asset Integrity Management, pages 1–24. Springer, Cham.

Rasmussen, M. H., Lefrançois, M., Schneider, G. F., and Pauwels, P. (2017). BOT: The building topology ontology of the W3C linked building data group. In 5th Linked Data in Architecture and Construction Workshop (LDAC 2017), Dijon, France, volume 13.

Shalabi, F. and Turkan, Y. (2020). Challenges and opportunities for creating fully automated BIM-based cost estimation in the ksa construction industry. Construction Innovation.

Wülfing, A., Windisch, R., and Scherer, R. (2014). A visual BIM query language. In Proc. eWorkBau. Technische Universität Dresden. Accessed: 2025-01-13.

Yin, M., Tang, L., Webster, C., Yi, X., Ying, H., and Wen, Y. (2023). A deep natural language processing-based method for ontology learning of project-specific properties from building information models. Computer-Aided Civil and Infrastructure Engineering, 39(4):334–359.

Zhang, K., Xu, R., Chuang, W.-L., and Hsiang, T.-H. (2018). Knowledge representation in construction-oriented design decision support systems: A review. Automation in Construction, 95:64–76.