INTERNATIONAL MULTI DISCIPLINARY JOURNAL FOR RESEARCH & DEVELOPMENT

eISSN 2394-6334 Impact factor: 7,854 Volume 12, issue 7 (2025)

Integrating Natural Language Processing and Building Information Modelling for Automated Information Extraction, Regulatory Compliance, and Quantity Take-Off in Infrastructure and Construction Projects

Dr. Alejandro Moreno

Department of Civil and Infrastructure Engineering,Universidad Politécnica de Madrid, Spain

Abstract: The construction and infrastructure sector is undergoing a profound digital transformation driven by the increasing adoption of Building Information Modelling (BIM) and advances in Natural Language Processing (NLP). Despite these developments, a significant proportion of critical project information remains embedded in unstructured or semi-structured textual documents, such as regulations, contracts, inspection reports, bills of quantities, and work descriptions. This fragmentation limits automation, introduces interpretive ambiguity, and increases the likelihood of errors in regulatory compliance, quantity take-off, cost estimation, and project control. Drawing strictly upon established scholarly works in construction informatics, BIM-enabled quantity take-off, ontology-driven modelling, and NLP-based information extraction, this research develops an integrated conceptual and methodological framework that unifies NLP techniques with BIM-based workflows. The article provides an extensive theoretical elaboration on how rule-based, ontology-based, machine learning, and hybrid deep learning NLP approaches can be systematically aligned with BIM data structures to automate utility permitting, compliance checking, accident cause classification, bridge inspection interpretation, and high-accuracy quantity take-off. Through descriptive methodological synthesis and interpretive analysis of prior validated research, the study demonstrates that the convergence of NLP and BIM not only reduces human-dependent interpretation but also enhances transparency, consistency, and scalability across the project lifecycle. The findings suggest that such integration represents a foundational shift from document-centric to knowledge-centric construction management. Limitations related to data heterogeneity, domain vocabulary evolution, and model transferability are critically discussed, and future research directions are proposed to support the maturation of intelligent, regulation-aware, and cost-reliable construction systems.

Keywords: Natural language processing; Building Information Modelling; information extraction; regulatory compliance; quantity take-off; construction informatics

INTRODUCTION

The The construction and infrastructure industry has historically been characterized by its reliance on fragmented information sources, manual interpretation of technical documents, and discipline-specific silos. Unlike manufacturing or aerospace sectors, where data standardization and automation have been long established, construction projects continue to depend heavily on textual documents such as specifications, regulations, contracts, inspection reports, and bills of quantities. These documents are typically written in natural language, often using domain-specific terminology, implicit assumptions, and context-dependent meanings. As a result, critical project knowledge remains locked within text, requiring expert interpretation that is time-consuming, error-prone, and difficult to scale.

The emergence of Building Information Modelling has been widely recognized as a transformative development capable of centralizing geometric, semantic, and parametric information across the project lifecycle. BIM enables digital representations of physical and functional characteristics of built assets, supporting design coordination, clash detection, scheduling, cost estimation, and asset management. However, despite its capabilities, BIM alone does not address the fundamental challenge that a large portion of construction knowledge originates outside structured models. Regulations governing highway rights-of-way, underground utilities, safety compliance, or contractual obligations are typically authored in legal or technical prose rather than machine-readable formats (Xu et al., 2020; Xu & Cai, 2021). Similarly, inspection reports, accident narratives, and work descriptions contain valuable operational insights but are rarely encoded in structured databases (Liu & El-Gohary, 2017; Zhang, 2019).

Natural Language Processing provides a computational means to bridge this gap by enabling machines to process, interpret, and extract structured knowledge from unstructured text. Over the past decade, NLP techniques have evolved from rule-based parsing systems to sophisticated machine learning and deep learning models capable of capturing semantic relationships, contextual meaning, and latent patterns in large corpora of text. Within the construction domain, researchers have demonstrated the feasibility of NLP for tasks such as automated regulatory compliance checking, classification of accident causes, extraction of contractual clauses, and interpretation of inspection reports (Lee et al., 2019; Zhang & El-Gohary, 2015).

Despite these advances, existing studies often address isolated problems without fully integrating NLP outputs into BIM-centric workflows. Quantity take-off, for example, remains vulnerable to discrepancies between model-based quantities and textual descriptions in bills of quantities or specifications, leading to estimation errors and disputes (Martínez-Rojas et al., 2016; Valinejadshoubi, 2024). Similarly, regulatory compliance systems may extract requirements from text but fail to dynamically align them with evolving BIM models during design and construction phases (Xu et al., 2020).

This article addresses this gap by synthesizing prior research into a unified, theoretically grounded framework that integrates NLP-driven information extraction with BIM-based data environments. By drawing strictly on validated scholarly sources, the study elaborates on how different NLP paradigms—rule-based, ontology-based, statistical learning, and neural network models—can be systematically aligned with BIM for automated permitting, compliance checking, safety analysis, and quantity take-off. The contribution of this research lies not in proposing a new algorithm, but in providing an extensive conceptual integration and interpretive analysis that clarifies the mechanisms, benefits, and limitations of NLP–BIM convergence in construction and infrastructure projects.

METHODOLOGY

The methodological approach adopted in this research is qualitative, integrative, and theory-driven, reflecting the objective of synthesizing and elaborating existing validated research rather than conducting empirical experimentation. The methodology is grounded in an interpretive analysis of peer-reviewed studies that have applied NLP, ontology modelling, machine learning, and BIM-based quantity take-off within construction and infrastructure contexts. Each referenced work is examined in depth to extract its underlying assumptions, data representations, processing logic, and integration potential with BIM environments.

The first methodological layer involves thematic categorization of the literature into core functional domains: regulatory interpretation and compliance, information extraction from inspection and narrative reports, classification of construction knowledge, and BIM-enabled quantity take-off and cost management. Studies on automated utility permitting and regulatory text interpretation are analyzed to understand how UML, OCL, and ontological representations translate natural language rules into computable constraints (Xu et al., 2020; Xu & Cai, 2021). These works provide methodological insight into how regulatory semantics can be formalized without oversimplifying legal intent.

The second layer focuses on NLP techniques applied to construction-specific texts. Rule-based systems, such as those used for identifying poisonous clauses in contracts, are examined for their transparency and domain precision, as well as their limitations in scalability and adaptability (Lee et al., 2019). Statistical and machine learning approaches, including conditional random fields and hybrid deep neural networks with Word2Vec embeddings, are analyzed for their ability to generalize patterns across diverse textual inputs (Liu & El-Gohary, 2017; Zhang, 2019). Particular attention is given to how these models handle domain vocabulary, context dependency, and semantic ambiguity.

The third layer integrates BIM-centric studies that address quantity take-off, constructability, and asset management. Research on automated and high-accuracy quantity take-off using BIM is analyzed to identify sources of error reduction and data consistency (Parate et al., 2025; Valinejadshoubi, 2024). Studies on BIM 5D and large infrastructure asset management provide insight into how cost and schedule dimensions interact with model semantics over time (Wardito, 2024; Sierra, 2023).

The final methodological step involves conceptual integration. Insights from NLP-driven information extraction are mapped onto BIM workflows to illustrate how extracted entities, relations, and rules can populate or validate BIM objects and attributes. This synthesis is conducted through descriptive reasoning, ensuring that each integration claim is explicitly grounded in prior research and theoretical logic rather than speculative extrapolation.

RESULTS

The integrative analysis reveals that NLP–BIM convergence enables a multi-layered transformation of construction information management. One of the most significant findings is that regulatory texts, traditionally treated as external constraints, can be transformed into active computational agents within BIM environments. Studies on utility permitting demonstrate that when regulations are decomposed into semantic components and encoded using UML and OCL, they can automatically validate design proposals against right-of-way constraints (Xu et al., 2020). NLP serves as the front-end interpreter, translating textual regulations into structured representations that BIM systems can reason over.

In the domain of safety and risk analysis, results from accident cause classification research indicate that hybrid deep learning models outperform traditional manual categorization by capturing latent semantic patterns in incident reports (Zhang, 2019). When integrated with BIM, such classifications can be spatially contextualized, enabling proactive risk visualization and design modifications. This finding underscores the potential of NLP to enrich BIM with experiential knowledge derived from past projects.

Information extraction from inspection reports emerges as another area of significant impact. Ontology-based and semi-supervised models demonstrate high accuracy in identifying defects, components, and conditions from narrative inspection texts (Liu & El-Gohary, 2017; Liu & El-Gohary, 2021). When such extracted data is linked to BIM objects, inspection outcomes can be directly associated with asset components, supporting data-driven maintenance planning and lifecycle management.

Regarding quantity take-off, the analysis shows that BIM-based automation significantly reduces arithmetic and omission errors, but discrepancies often arise due to misalignment between model elements and textual work descriptions (Martínez-Rojas et al., 2016). NLP-based classification and extraction techniques address this gap by systematically mapping textual descriptions in bills of quantities to standardized task groups and BIM object categories (Martínez-Rojas et al., 2018). The integration of these methods leads to improved consistency between design intent, cost estimation, and procurement documentation.

Overall, the results suggest that the integration of NLP and BIM does not merely automate isolated tasks but restructures the information ecosystem of construction projects, shifting from document-centric workflows to semantically interconnected knowledge systems.

DISCUSSION

The findings of this research have significant theoretical and practical implications for construction informatics. From a theoretical perspective, the integration of NLP and BIM challenges the traditional boundary between structured and unstructured data. Construction knowledge is no longer confined to either models or documents but exists within a continuous semantic spectrum. Ontology-based approaches play a crucial mediating role in this integration by providing shared conceptual frameworks that align textual semantics with BIM object hierarchies (Xu & Cai, 2021; Zhang & El-Gohary, 2015).

However, this convergence also introduces new challenges. Rule-based NLP systems offer interpretability and legal defensibility, which are critical for regulatory and contractual applications, but they struggle with linguistic variability and scalability (Lee et al., 2019). In contrast, machine learning and deep learning models excel at pattern recognition but often lack transparency, raising concerns about accountability in compliance-related decisions (Zhang, 2019). Balancing these paradigms requires hybrid architectures that leverage the strengths of each while mitigating their weaknesses.

Data quality and domain specificity remain persistent limitations. NLP models trained on one set of regulations, inspection reports, or work descriptions may not generalize well across jurisdictions or project types due to variations in terminology and writing style (Seedah & Leite, 2015). Similarly, BIM models vary in level of detail and semantic richness, affecting the effectiveness of integration. These limitations highlight the need for continuous model adaptation and domain knowledge curation.

Future research should focus on dynamic feedback loops between BIM and NLP systems, where changes in models trigger re-evaluation of textual constraints and vice versa. Advances in semantic neural network ensembles suggest promising directions for capturing complex dependency relations in construction texts (Liu & El-Gohary, 2021). Additionally, expanding NLP–BIM integration into asset management and facility operations could unlock long-term value beyond design and construction phases (Sierra, 2023).

CONCLUSION

This research has presented an extensive, theoretically grounded synthesis of how Natural Language Processing and Building Information Modelling can be integrated to automate information extraction, regulatory compliance, safety analysis, and quantity take-off in construction and infrastructure projects. Drawing strictly from established scholarly literature, the study demonstrates that NLP serves as a critical bridge between unstructured textual knowledge and structured BIM environments. The integration enables not only efficiency gains but also a fundamental transformation in how construction knowledge is represented, validated, and operationalized.

While challenges related to data heterogeneity, model transparency, and domain adaptability persist, the convergence of NLP and BIM represents a decisive step toward intelligent, knowledge-centric construction systems. By embedding regulatory logic, experiential insights, and cost semantics directly into digital models, the industry can move closer to realizing the full potential of digital transformation across the project lifecycle.

REFERENCES

1. Hage, S. (2023). Efficiency in the preparation of life cycle assessment. Environmental Science and Engineering, 143–155.

2. Lee, J., Yi, J. S., & Son, J. (2019). Development of automatic-extraction model of poisonous clauses in international construction contracts using rule-based NLP. Journal of Computing in Civil Engineering, 33, 04019003.

3. Liu, K., & El-Gohary, N. (2017). Ontology-based semi-supervised conditional random fields for automated information extraction from bridge inspection reports. Automation in Construction, 81, 313–327.

4. Liu, K., & El-Gohary, N. (2021). Semantic neural network ensemble for automated dependency relation extraction from bridge inspection reports. Journal of Computing in Civil Engineering, 35, 04021007.

5. Martínez-Rojas, M., Marín, N., & Vila, M. A. (2015). An approach for the automatic classification of work descriptions in construction projects. Computer-Aided Civil and Infrastructure Engineering, 30, 919–934.

6. Martínez-Rojas, M., Marín, N., & Miranda, M. A. V. (2016). An intelligent system for the acquisition and management of information from bill of quantities in building projects. Expert Systems with Applications, 63, 284–294.

7. Martínez-Rojas, M., Soto-Hidalgo, J. M., Marín, N., & Vila, M. A. (2018). Using classification techniques for assigning work descriptions to task groups on the basis of construction vocabulary. Computer-Aided Civil and Infrastructure Engineering, 33, 966–981.

8. Parate, H., Bandela, K., & Madala, P. (2025). Quantity take-off strategies: Reducing errors in roadway construction estimation. Journal of Mechanical, Civil and Industrial Engineering, 6(3), 01–09.

9. Seedah, D. P. K., & Leite, F. (2015). Information extraction for freight-related natural language queries. Computing in Civil Engineering, 667–674.

10. Sierra, C. (2023). Building information modelling for constructability and asset management of large rail infrastructure. IEEE Engineering Informatics.

11. Tanko, B. L. (2024). BIM in the Malaysian construction industry: A scientometric review and case study. Engineering, Construction and Architectural Management, 31(3), 1165–1186.

12. Valinejadshoubi, M. (2024). Automated system for high-accuracy quantity takeoff using BIM. Automation in Construction, 157.

13. Wardito, E. (2024). Increasing the value of jetty projects based on building information modelling 5D. AIP Conference Proceedings, 2710(1).

14. Xu, X., Chen, K., & Cai, H. (2020). Automating utility permitting within highway right-of-way via a generic UML/OCL model and natural language processing. Journal of Construction Engineering and Management, 146, 04020135.

15. Xu, X., & Cai, H. (2021). Ontology and rule-based natural language processing approach for interpreting textual regulations on underground utility infrastructure. Advanced Engineering Informatics, 48, 101288.

16. Zhang, F. (2019). A hybrid structured deep neural network with Word2Vec for construction accident causes classification. International Journal of Construction Management, 1–21.

17. Zhang, J., & El-Gohary, N. M. (2015). Automated information transformation for automated regulatory compliance checking in construction. Journal of Computing in Civil Engineering, 29, B4015001.