Automated Extraction of Information Requirements from Building Regulations Using LLMs and Ontologies

Sven Zentgraf[1] and Leonie Zimmermann[1]
[1]Chair of Computing in Engineering, Ruhr-Universität Bochum, Bochum, Germany
E-mail(s): sven.zentgraf@rub.de, leonie.zimmermann@rub.de

Abstract: In the context of building requirements checking, the frequent process of information requirement extraction, even in digital Building Information Modeling (BIM) processes, involves a high amount of manual extraction of information requirements from non-machine-readable regulatory documents. This process is both time-consuming and prone to inconsistencies and errors. This poses significant challenges to efficient data processing and validation. Therefore, a method is needed that simplifies the process of extracting information requirements and deploying this information in a machine-readable format. To address these limitations, this paper proposes an automated approach that leverages a Large Language Model (LLM) and prompts engineering to extract relevant information requirements from building guidelines. The extracted information is then structured according to the pre-defined ISOProps Ontology as a corresponding A-Box ontology. The syntactic and semantic correctness of the generated A-box ontology is validated with the help of Shapes Constraint Language (SHACL) shapes generated from the associated T-box ontology. This ensures formal correctness and conformity with the ISOProps ontology. A human-in-the-loop approach qualitatively verifies the technical correctness. The feasibility of this approach was demonstrated on sample documents, highlighting its potential for streamlining the digital representation of building requirements.

Keywords: Building information modeling (BIM), Large Language Model (LLM), Prompt Engineering, Information Requirement Extraction, A-Box ontology

DOI: 10.18154/RWTH-CONV-254904. Published in the conference proceedings of the 36. Forum Bauinformatik 2025, Aachen, Germany,

1 Introduction

In the field of building requirements checking, the extraction of information requirements from regulatory documents remains a significant challenge, even within digital BIM processes. Typically, this task involves a considerable amount of manual effort, as the relevant information is often embedded in non-machine-readable documents. This manual extraction is not only time-consuming but also susceptible to inconsistencies and errors, which can hinder efficient data processing and validation [1].

To overcome these challenges, there is a clear need for a method that both simplifies the extraction of information requirements and enables their deployment in a machine-readable format [2], [3]. This paper introduces an automated approach designed to address these limitations by leveraging an LLM and prompt engineering techniques to extract relevant information from building guidelines. This involves not only identifying explicit requirements but also interpreting the implicit knowledge and contextual dependencies embedded in the text. The extracted requirements are then structured according to the ISOProps Ontology, as an A-Box ontology. The ISOProps Ontology is built upon the data model of ISO 23386, which provides standardized methods for describing, creating, and maintaining building-related information requirements as properties and property groups [4], [5].

To ensure both syntactic and semantic correctness, the generated A-Box ontology is validated using SHACL shapes derived from the associated T-Box ontology, guaranteeing formal compliance with the ISOProps Ontology. Additionally, the technical accuracy of the extracted information is qualitatively assessed through a human-in-the-loop approach. The feasibility and effectiveness of this automated method are demonstrated on sample regulatory documents, underscoring its potential to streamline the digital representation and validation of building requirements.

2 Background

2.1 Information Requirements checking for Regulatory Compliance

Regulatory compliance checking in the Architecture, Engineering, and Construction (AEC) industry is the process of verifying that a building design or BIM model adheres to applicable codes, standards, and regulations. Traditionally, this process has been labor-intensive as it is performed predominantly by human experts, who interpret written regulations and manually inspect drawings or BIM models for violations. In recent years, there has been a strong drive to automate regulatory checking using BIM, in order to reduce errors, save time, and ensure that complex regulations are consistently applied [1, 6]. Early implementations of Automated Code Compliance showed promise in improving design quality and shortening permitting times [7]. However, automating compliance checking is challenging because building regulations are often lengthy, complex, and sometimes ambiguous [8]. Over the last decade, multiple approaches have been researched and developed to represent regulatory information requirements and rules in a digital format and to implement rule-checking engines [6], [7], [8]. At a high level, these approaches can be grouped into a few categories [8]. These categories range from rule-based, proprietary BIM model checkers and standard-based schemas to model view definitions, as well as approaches based on the Semantic Web and linked data, or natural language processing and structured text. The approach employed in this study is based on Semantic Web and Linked Data technologies, relying on the IR-Ontology network [4], which is further introduced in Section 2.2. Pauwels et al. [9] presents a related work introducing a semantic rule-checking system where various technical requirements from regulations are encoded as SHACL rules and OWL axioms. Their results indicate that a large portion of regulatory requirements can be validated automatically using Semantic Web reasoning [9]. Another example is the work by Nuyts et al. [8] who used OWL and SWRL to encode parts of the Flemish building regulations and checked models with a reasoner. In addition, a comparative study was conducted that utilized the same set of regulatory requirements in eight

different ways, employing two proprietary/IFC-based tools, two generic data schema approaches, and four Semantic Web approaches.

2.2 Digital Standards and Ontologies for Information Requirements

The RE-ING are central technical guidelines issued by the German Federal Ministry for Digital and Transport (BMDV) that regulate the planning and design of engineering structures in federal infrastructure projects. In practice, they provide detailed specifications that go beyond the scope of general standards such as the Eurocodes and DIN norms, particularly concerning application-specific requirements for structures like bridges, retaining walls, and noise barriers [10]. As part of ongoing digitalization efforts, there are initiatives to convert RE-ING into machine-interpretable formats, such as XML compliant with the NISO Standard Tag Suite (NISO STS), a standardized XML format for encoding and exchanging technical and scientific documents. This is aimed at enabling automated processing and integration with BIM systems.

In parallel to these transformations, broader standardization initiatives are emerging that focus on formalizing information requirements. Notably, various methodologies have been developed to structure the exchange and validation of information throughout project lifecycles. Among these, the Information Delivery Manual (IDM)[1] and the Information Delivery Specification (IDS)[2] are promoted by buildingSMART International. However, despite the growing theoretical foundation, practical implementation often lags. Many projects lack clearly defined, machine-readable information requirements, which hinders the ability to automate the validation of delivered data.

To address these challenges, recent research has examined how semantic technologies such as Linked Data and ontologies can support a more rigorous and interoperable management of information requirements. A notable contribution to the digital management of information requirements in construction is presented by Filardo et al. [4]. An aligned ontology network that harmonizes the LOIN (Level of Information Need), Data Templates (DT), and ISOProps ontologies was introduced. It demonstrates how these ontologies can be coordinated to consistently define properties and construction objects, following international standards such as ISO 23386, ISO 23387, and ISO 7817. This modular and reusable ontology network supports the interoperability and scalability of digital workflows in construction, offering a flexible foundation for the automated validation and provision of project-specific information [4].

2.3 Large Language Models for Information Extraction

LLMs are advanced neural networks trained on extensive textual data to learn language patterns. After a pre-training phase, they can perform various language-based tasks such as translation, answering questions, and summarizing content [11].

In recent years, LLMs have rapidly gained attention, driven by significant advances in model architectures and training on bigger datasets comprising billions of parameters. Their performance has improved, enabling them to handle a broad spectrum of tasks, mainly depending on the diversity and quality of the data they are exposed to during training. State-of-the-art models, such as GPT-4o,

[1] https://technical.buildingsmart.org/standards/information-delivery-manual/
[2] https://technical.buildingsmart.org/projects/information-delivery-specification-ids/

exemplify this progress by extending beyond text generation to include capabilities like code synthesis, image generation, and the processing of audio and video inputs [12].

The Mistral LLM, developed by Mistral AI, is an open-source model that offers state-of-the-art language understanding and generation. Its public release provides researchers and developers with free access to a powerful tool for advanced NLP applications [13].

As these models become increasingly capable and accessible, the challenge shifts to effectively directing these models. This is where prompt engineering plays a central role. Prompt engineering is the practice of designing effective inputs to guide LLMs toward producing accurate, relevant, and domain-specific outputs. It has emerged as a crucial discipline for harnessing the full capabilities of LLMs across diverse tasks. Two foundational paradigms in this area are zero-shot and few-shot prompting. Zero-shot prompting in this regard refers to the process where an LLM is given a task description or instruction in natural language. Still, no explicit examples of the task are provided within the prompt. The model relies entirely on its pre-trained knowledge and generalization capabilities to interpret and perform the requested task. This approach is particularly advantageous in scenarios where labeled data is scarce or unavailable, as it enables the immediate application of LLMs to new domains without requiring additional training or annotation efforts. Few-shot prompting, in contrast, involves including a small number of input-output examples within the prompt, alongside the task instruction.

This technique is beneficial when some representative examples are available, but not enough to justify full-scale supervised fine-tuning [14].

Höltgen et al. [15] explore the capabilities of LLMs for semantic data integration, focusing on evaluating LLMs for generating R2RML (RDB to RDF Mapping Language) mappings [15]. Recent research also demonstrates the practical value of LLMs for information extraction in the context of automated compliance checking. For example, Iversen [16] from NTNU developed and evaluated an LLM-based artifact for validating building regulations against BIM models. The study focused on extracting and classifying regulatory requirements from natural language documents, identifying dependencies between rules, and mapping these requirements to BIM data for compliance assessment [16].

3 Methods

This study examines the application of LLMs for the automated conversion of regulatory documents into an ontology-based representation. Specifically, prompt engineering techniques are employed to guide an LLM in converting RE-ING text passages into A-box statements, grounded in the T-box defined by the ISOProps ontology network proposed by Filardo et al. [4].

The transformation process is implemented via a Python-based toolchain that interacts with the Mistral API. Within the prompts, a series of curated examples is included that demonstrate how textual guideline content should be interpreted and formalized as individual ontology instances. By iteratively adapting the prompt and refining examples, the accuracy and consistency of the generated A-box ontology are improved. To support this conversion task, we adopt a few-shot prompt engineering strategy. The prompt is carefully structured to guide the model in generating an ontology from

construction-related requirements by breaking the task into distinct, manageable components (cf. Table 1).

Table 1: Structured prompt for generating an ontology from construction-related texts.

Category	Description
Objective	Generate an ontology from construction-related requirements
Input	XML document: `{text}`
Task	- Extract explicit construction requirements - Map terms (only those found in the text) - Generate RDF triples
Type Definitions	- Group of Properties: `{group_of_properties_definition}` - Property: `{property_definition}`
Triple Generation	- Create for each property and property group - Format: Turtle (RDF) - Groups include category and definition - Properties reference their group
Additional Rules	- Do not invent new terms - Use `DefiningValuesList` for dependencies - Mark AI-generated definitions clearly
Example File	`{example_file}`
Expected Output	Turtle file with all generated RDF triples only

The prompt introduces the LLM as an expert in the analysis of construction-related documents. It provides clear instructions on how to extract entities of type *Group of Properties* and *Property*, using definitions aligned with the ISOProps ontology. It starts with a clear objective and input specification, ensuring the model understands the domain and expected data format. The task is divided into three steps: requirement extraction, term mapping, and generating RDF triples. Type definitions and output constraints are explicitly included to enforce semantic consistency and ensure compliance with formal Terse RDF Triple Language (Turtle) syntax. Additional rules prevent the model from inventing terms and ensure transparency in generated content. A detailed Turtle serialization schema is included, specifying how output triples should be structured, along with an annotated example derived from a real RE-ING section. This enables the LLM to learn both the structural and semantic patterns required for the ontology population.

```
tempo-usecase:propertyGroupNameAlsEinSubstantiv
       rdf:type isoprops:GroupOfProperties ;
       isoprops:categoryOfGroupOfProperties "Kategorisierung hier einfügen" ;
       tempo:definitionInLanguage "(KI-generiert) Wenn keine definition im Text steht, generiere eine
           und setze (KI-generiert) davor. Ansonsten entnehme die dem Text."@de ;
       tempo:nameInLanguage "Name der Gruppe von Eingeschaften"@de;
       isoprops:countryOfOrigin "DE" ;
       isoprops:countryOfUse "DE" ;
       isoprops:creatorsLanguage "de-DE" ;
.
```

Algorithm 1: Turtle template illustrating the structure of a GroupOfProperties instance.

The ontology itself comprises two main classes, namely *Groups of Properties*, which represents collections of related properties, and *Property*, which denotes individual, measurable, or descriptive

features of an Item. A template for defining a *Group of Properties* instance in Turtle syntax is provided in Algorithm 1. As illustrated, each GroupOfProperties is further classified using the object property `isoprops:categoryOfGroupOfProperties` into one of five predefined categories. These categories are *Domain*, *Alternative Use*, *Composed Property*, *Class*, or *Reference Document*, all of which are explicitly defined in the prompt. In addition, each group includes semantic metadata such as its name, the language of origin, country of use, and, where not explicitly provided in the source text, a marked AI-generated definition. For individual properties, the ontology captures metadata like measurement units, associated physical quantities, value boundaries, and precision. The ontology also supports enumerated values, boundary conditions, and measurement methods, where applicable. Each element is linked to its corresponding Group of Properties. Further key modeling elements are those that capture domain-specific constraints such as measurement precision, value ranges, and unit definitions. Rather than modeling these details separately, the methodology incorporates them directly into the prompt structure through clearly defined instructions and few-shot examples.

4 Evaluation

To evaluate the proposed approach, a set of exemplary regulatory texts was selected from publicly available building guidelines. These texts were carefully curated to cover a range of typical information requirements, including both explicit and implicit constraints. The size and complexity of the input documents were deliberately constrained to remain within the token limit of the Mistral API (approximately 32,000 tokens), ensuring reliable model performance and response consistency.

The evaluation addressed the two main aspects under consideration, namely the quality of the extracted requirements and the correctness of the generated ontology. Firstly, the LLM was prompted using tailored templates designed to elicit both specific requirement statements and contextual interpretations. A qualitative human-in-the-loop review assesses the technical accuracy and domain relevance of the results, with experts in building regulations and semantic modeling verifying that the generated requirements and ontology entries accurately reflected the intended regulatory constraints. For the second aspect, the extracted information was instantiated as an A-Box ontology conforming to the ISOProps Ontology. The generated ontology is validated using SHACL (Shapes Constraint Language) shapes derived from the corresponding T-Box ontology, ensuring both syntactic structure and semantic alignment with ISO 23386. Any violations or inconsistencies detected during the SHACL validation were analyzed to refine the prompt design or improve the mapping logic. After iterative refinement of the prompt design, the LLM ultimately produced ontology instances with correct syntax that complied with the defined SHACL shapes.

Overall, the evaluation demonstrates the feasibility of using LLMs to automate the extraction and structuring of building information requirements. While effective for small to moderately sized documents, performance scalability and accuracy across broader and more diverse regulatory corpora remain areas for future investigation.

5 Conclusion

This paper presented an automated method for extracting and structuring information requirements from regulatory building texts using an LLM combined with prompt engineering techniques. By

translating unstructured regulatory content into a machine-readable A-Box ontology aligned with the ISOProps Ontology and ISO 23386, the approach supports greater automation and consistency in digital building requirement validation workflows. The evaluation demonstrated that the method is capable of identifying both explicit and implicit requirements and encoding them in a semantically compliant format. Furthermore, the use of SHACL shapes for ontology validation ensures formal alignment with the defined T-Box structures[3].

However, the approach is not without limitations. One key constraint is the token limit imposed by the Mistral API, which restricts the size of input documents that can be processed in a single inference. This necessitates either careful document segmentation or future strategies for handling larger inputs in an iterative manner. While LLMs demonstrate significant capabilities in processing and interpreting natural language, they are not without limitations. Errors such as inaccurate information extraction, contextual misunderstandings, or improper mappings may arise, especially when dealing with ambiguous or complex regulatory language. Consequently, incorporating a human-in-the-loop is crucial to ensure the accuracy and relevance of the extracted information.

Despite these drawbacks, the proposed approach offers a promising direction for semi-automated digitization of building regulations. Future work will focus on extending to larger document sets, establishing a quantitative evaluation metric, refining prompt strategies to improve accuracy, and exploring hybrid techniques that combine rule-based and LLM-based extraction to enhance robustness.

References

[1] S. Fuchs, J. Dimyadi, M. Witbrock, and Amor Robert, “Improving the semantic parsing of building regulations through intermediate representations”, *The 30th EG-ICE: International Conference on Intelligent Computing in Engineering*, 2023.

[2] Z. Zhang, L. Ma, and T. Broyd, “Towards fully-automated code compliance checking of building regulations: Challenges for rule interpretation and representation”, *European Conference on Computing in Construction EC3*, 2022. DOI: 10.35490/EC3.2022.148

[3] Y.-C. Zhou, Z. Zheng, J.-R. Lin, and X.-Z. Lu, “Integrating nlp and context-free grammar for complex rule interpretation towards automated compliance checking”, *Computers in Industry*, vol. 142, p. 103 746, 2022. DOI: 10.1016/j.compind.2022.103746 [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0166361522001439

[4] M. M. Filardo, L. Liu, P. Hagedorn, S. Zentgraf, J. Melzner, and M. König, “A standard-based ontology network for information requirements in digital construction projects”, in *Proceedings of the 12th Linked Data in Architecture and Construction Workshop, Bochum, Germany, June 13-14, 2024*, P. Pauwels, M. Poveda-Villalón, and W. Terkaj, editors, ser. CEUR Workshop Proceedings, vol. 3824, CEUR-WS.org, 2024, pp. 77–90. [Online]. Available: https://ceur-ws.org/Vol-3824/paper6.pdf

[5] S. Zentgraf, P. Hagedorn, and M. König, “Multi-requirements ontology engineering for automated processing of document-based building codes to linked building data properties”, *IOP Conference Series: Earth and Environmental Science*, vol. 1101, no. 9, p. 092 007, 2022. DOI: 10.1088/1755-

[3] Relevant T-box ontologies available at: https://github.com/RUB-Informatik-im-Bauwesen/ir-ontologies

1315/1101/9/092007 [Online]. Available: https://iopscience.iop.org/article/10.1088/1755-1315/1101/9/092007

[6] W. Solihin, Z. Liu, Y. Lu, and L. Wei, "Bim-based automated rule-checking in the aeco industry: Learning from semiconductor manufacturing", *Automation in Construction*, vol. 162, p. 105 406, 2024. DOI: https://doi.org/10.1016/j.autcon.2024.105406 [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0926580524001420

[7] J. Zhang, B. Cui, Y. Gao, and D. Zhang, "Factors influencing the acceptance of bim-based automated code compliance checking in the aec industry in china", *Journal of Management in Engineering*, vol. 39, no. 6, p. 04 023 036, 2023. DOI: 10.1061/JMENEA.MEENG-5344

[8] E. Nuyts, M. Bonduel, and R. Verstraeten, "Comparative analysis of approaches for automated compliance checking of construction data", *Advanced Engineering Informatics*, vol. 60, p. 102 443, 2024. DOI: https://doi.org/10.1016/j.aei.2024.102443 [Online]. Available: https://www.sciencedirect.com/science/article/pii/S1474034624000910

[9] P. Pauwels, E. van den Bersselaar, and L. Verhelst, "Validation of technical requirements for a bim model using semantic web technologies", *Advanced Engineering Informatics*, vol. 60, p. 102 426, 2024. DOI: 10.1016/j.aei.2024.102426 [Online]. Available: https://www.sciencedirect.com/science/article/pii/S1474034624000740

[10] *Publikationen - Richtlinien für den Entwurf, die konstruktive Ausbildung und Ausstattung von Ingenieurbauten (RE-ING)*, 30.04.2025. [Online]. Available: https://www.bast.de/DE/Publikationen/Regelwerke/Ingenieurbau/Entwurf/RE-ING.html

[11] G. Sasson Lazovsky, T. Raz, and Y. N. Kenett, "The art of creative inquiry—from question asking to prompt engineering", *The Journal of Creative Behavior*, vol. 59, no. 1, 2025. DOI: 10.1002/jocb.671

[12] S. Shahriar et al., *Putting gpt-4o to the sword: A comprehensive evaluation of language, vision, speech, and multimodal proficiency*, 2024. arXiv: 2407.09519 [cs.AI]. [Online]. Available: https://arxiv.org/abs/2407.09519

[13] H.-C. Tsai, Y.-F. Huang, and C.-W. Kuo, *Comparative Analysis of Automatic Literature Review Using Mistral Large Language Model and Human Reviewers*. 2024. DOI: 10.21203/rs.3.rs-4022248/v1 [Online]. Available: https://assets-eu.researchsquare.com/files/rs-4022248/v1_covered_31b35741-397a-47a8-aa52-d71a4150153f.pdf

[14] S. Schulhoff et al., *The prompt report: A systematic survey of prompt engineering techniques*. [Online]. Available: http://arxiv.org/pdf/2406.06608v6

[15] L. Höltgen, S. Zentgraf, P. Hagedorn, and M. König, "Utilizing large language models for semantic enrichment of infrastructure condition data: A comparative study of gpt and llama models", *AI in Civil Engineering*, vol. 4, no. 1, 2025. DOI: 10.1007/s43503-025-00055-9

[16] O. Iversen, "Leveraging large language models for bim-based automated compliance checking of building regulations", Ph.D. dissertation, NTNU. [Online]. Available: https://ntnuopen.ntnu.no/ntnu-xmlui/handle/11250/3149999