A Hybrid BERT–LLM Approach for Regulation Graph Generation & Visualization from Fire Safety Documents

Tomás Jorge[1][0009-0003-5458-7150], Diogo Ribeiro[1][0000-0001-6199-3120], Jéssica Reis[2], Rui Gavina[2][0009-0000-9180-639X], Alex Donkers[3][0000-0002-8809-3277], Ekaterina Petrova[3][0000-0002-8651-0671], Alda Canito[4][0000-0002-4338-7239], Cedric Driesen[5][0000-0002-0294-4756]

[1] iBuilt, School of Engineering, Polytechnic of Porto, Porto, Portugal
[2] VN2R-Engineering Innovation Consulting LDA, Porto, Portugal
[3] Information Systems in the Built Environment, Eindhoven University of Technology, De Groene Loper 6, Eindhoven, The Netherlands
[4] GECAD/LASI, ISEP, Polytechnic of Porto, Porto, Portugal
[5] Buildwise, Kleine Kloosterstraat 23, 1932 Zaventem, Belgium

tosjo@isep.ipp.pt

Abstract. The efficient application of fire safety regulations in the early stages of a project is often hindered by their unstructured language and dispersion across multiple legal documents. This fragmentation poses challenges for automatic interpretation and hinders the digitalization of compliance verification processes. In this context, the paper introduces an innovative hybrid methodology that integrates pre-trained neural models (BERT) with large language models (LLMs), specifically Gemini Flash 2.5. The main objective is to automatically transform natural language regulations into a machine-readable format 'representing regulatory requirements as knowledge graphs, using a pre-defined regulation ontology. The proposed methodology was validated using Portuguese fire safety regulations and demonstrated strong adaptability across different regulatory documents. The results reveal a significant reduction in manual effort and human error traditionally associated with compliance tasks, while incorporating a Human-in-the-Loop strategy to ensure expert validation at critical stages of the process. Overall, this work makes a substantial contribution to intelligent regulatory automation and provides a solid foundation for automated compliance checking.

Keywords: Fire Safety, Natural Language Processing, Large Language Models, BERT, Knowledge Graphs, Human-in-the-Loop

1 Introduction

Fire safety is a critical issue in the construction sector, directly impacting the protection of human lives, built heritage, and the environment. However, despite its widely recognized importance, fire safety standards and strategies are often considered only late in the building design process. This predominantly reactive approach not only reduces the technical and economic efficiency of projects but also significantly compromises overall safety, frequently resulting in costly design changes at advanced stages or, in extreme cases, critical failures during emergencies [1]. One of the main barriers to the

early integration of fire safety lies in the complexity and fragmentation of existing regulations. These documents, typically written in natural language, are dispersed across multiple legal texts and technical standards, often under different jurisdictions. This dispersion creates terminological ambiguities and semantic inconsistencies, posing major challenges to their automatic interpretation [2]. Furthermore, the widespread lack of structured digital formats greatly limits the potential for technologies that could facilitate the systematic application of regulatory requirements. In this context, the advancement of digital technologies has driven the development of artificial intelligence (AI) and natural language processing (NLP) techniques aimed at converting natural language into machine-interpretable formats.

Various researchers aimed to convert building codes into machine-interpretable formats. Some rely on manual approaches, such as Nisbet and Ma [3] who applied the RASE methodology for semantic mark-up of regulatory documents. Various authors applied some form of AI, specifically NLP or Large Language Models (LLMs) to generate machine-interpretable versions of building codes. Zhang & El-Gohary [4] used NLP techniques and ontologies to map semantic concepts to words in regulations and linked those terms to IFC models [5], while Donkers and Petrova [2] combined NLP techniques and ontologies to generate SHACL shapes from regulatory text. Costa et al. [6] suggest the organization of regulations in a machine-interpretable format. Al-Turki et al. [7] aimed to combine LLM methods with human-in-the-loop learning to format regulations into structured YAML.

However, these approaches present certain limitations. Many current processes are either overly automated, frequently lacking rigorous validation by domain experts, or heavily reliant on manual procedures, which significantly limits their scalability. Therefore, it becomes essential to adopt a Human-in-the-Loop approach, where automation is complemented by targeted expert intervention, ensuring both operational efficiency and semantic accuracy throughout the process [8].

This article proposes an innovative hybrid methodology aimed at the automatic generation of Regulation Graphs for fire safety regulations, specifically applied to the Portuguese context, with the goal of supporting subsequent processes of automated compliance checking. The methodology combines (1) the semantic capabilities of neural models (BERT), (2) the generative power of large language models (LLMs), specifically Gemini Flash 2.5, and (3) a human-in-the-loop component to guide and validate the process. Accordingly, this study seeks to contribute to the current state of the art, with particular emphasis on the following aspects:

- Automated conversion of regulatory PDF documents into a graph-based format, enabling semantic representation of fire safety requirements in a machine-readable format and supporting downstream automated compliance checking;
- Combination of LLM-based processing with Human-in-the-loop validation to ensure accuracy and regulatory alignment;
- A flexible approach that can be adapted to several regulations and countries with different structural and semantic characteristics.

2 Proposed Methodology

The proposed methodology, implemented in Python[1], aims to convert regulatory documents into a structured regulation graph to support analysis, integration, and automated compliance checking. As shown in **Fig. 1**, it is composed of three main stages: conversion to machine readable text, enrichment & convert to RDF graph, and visualization.

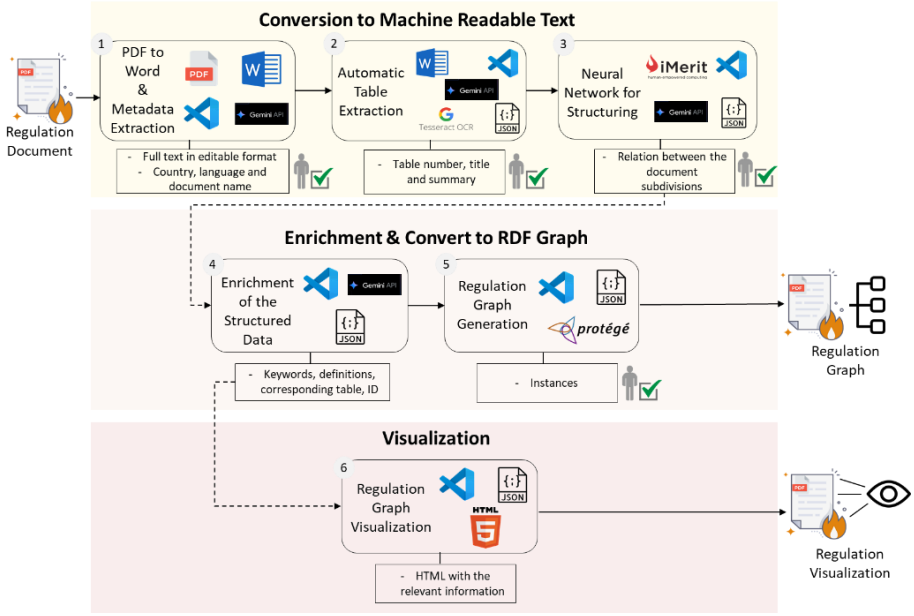


Fig. 1. Flowchart with the three main stages of the proposed methodology

In Step 1, documents in PDF format are converted to Word format to facilitate structured content extraction while preserving the original layout as much as possible. This editable format is necessary to enable accurate parsing and segmentation of the text, which supports the automatic extraction of key metadata, such as country, language, and the title of the legislation, using custom Python functions built on top of Gemini Flash 2.5 model. This metadata is essential for the subsequent stages of the process.

In Step 2, the information contained in all tables within the document is extracted. This is achieved through Optical Character Recognition (OCR) using Tesseract[2]. The output is a JSON file containing the table number, its title, and its content. Additionally, Gemini Flash 2.5 is used to generate an interpretative textual summary of each table, based on the analysis of its constituent elements. This summary helps capture the semantic content of complex tabular data and is later used to support the construction of the Regulation Graph.

In Step 3, a BERT-type neural network is applied. To this end, a set of real regulatory documents was manually labeled using iMerit[3], assigning the labels to each line (details provided in a later section).

[1] https://www.python.org/

[2] https://tesseract-ocr.github.io/

[3] https://imerit.ango.ai/login

Based on this annotated dataset, a BERT model was trained and fine-tuned to recognize the typical structural elements of such documents. When applied to new texts, the model can automatically infer the label of each line. A dedicated interface is used to validate and correct the model's predictions before proceeding. Based on this classification, a JSON file is generated representing the document's structure, including the assigned labels and their hierarchical relationships.

Once the conversion stage is completed, the semantic enrichment stage begins. This part aims to assign semantic meaning to the extracted data, again supported by Gemini Flash 2.5 and using the previously collected metadata. In Step 4, keywords are identified, their definitions are provided, and links are established with previously identified tables as well as their unique identifiers (ID). The enriched information, structured in JSON format, is then used in Step 5 to generate regulation graphs in RDF (Resource Description Framework), using the rdflib[4] library. The construction of these graphs is based on a pre-defined ontology that specifies relevant classes, subclasses, and properties. From this structure, individuals are automatically created to form a structured knowledge base that can be queried, and integrated with other ontologies. The RDF graphs are visualized in Protégé[5] for human-in-the-loop validation.

The final stage of the methodology focuses on data visualization. In Step 6, the information contained in the RDF graph is transformed into an HTML-based visualization layer, enabling user to explore the data in a clear and structured way. This interface is designed to support the interpretation of the extracted content, while the information it contains may serve, in the future, as a foundation for automated compliance checking. Moreover, the resulting HTML can be published as an interactive webpage, enabling the creation of a repository of linked, country-specific interactive pages.

To ensure the quality and accuracy of the extracted and structured information, human-in-the-loop validation is incorporated at Step 1, 2, 3, and 5 of the methodology.

3 Results

3.1 PDF to Word & Metadata Extraction

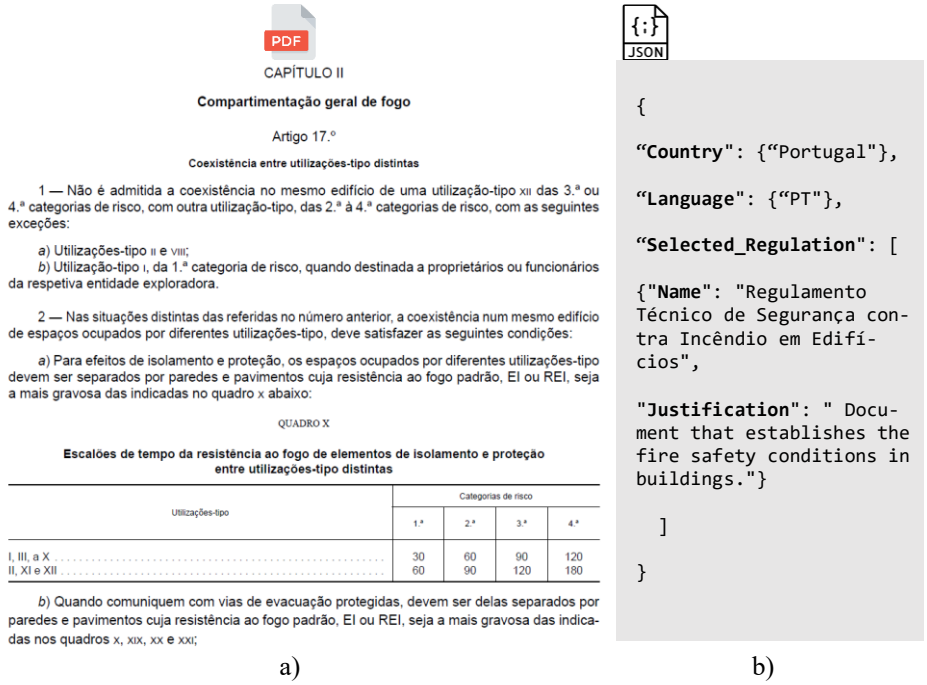
A significant number of technical regulations in Portugal, particularly those related to fire safety, are published in PDF format. Although this format is effective in preserving the visual layout of documents, it represents significant challenges for the automatic extraction of information. To address these limitations, the first step of the proposed methodology consists of converting the original PDF file into a Microsoft Word document. This initial conversion is performed using an online PDF to Word converter, followed by a format refinement step using Visual Basic for Applications (VBA). This ensures a more consistent and structured document, suitable for further processing.

The decision to use Word format, as opposed to alternatives such as Markdown or TXT, is due to its superior ability to preserve the visual and logical structure of the original document. Simpler formats tend to result in significant structural losses, especially in the case of more complex elements such as tables. After conversion, the Word document is analyzed by an LLM-based Python functions, specifically Gemini Flash 2.5,

[4] https://rdflib.readthedocs.io/en/stable/

[5] https://protege.stanford.edu/

which processes text excerpts to identify important metadata, such as the language of the document, the country of origin and a list of potentially relevant regulations. At this stage, the user is asked to confirm the correct regulation from the suggested options. This metadata is essential for correctly contextualizing the LLM-based Python functions, resulting in a much more accurate extraction of information. **Fig. 2**a) and **Fig. 2**b) show the metadata obtained from the analysis carried out by the LLM-based Python functions.



CAPÍTULO II

Compartimentação geral de fogo

Artigo 17.º

Coexistência entre utilizações-tipo distintas

1 — Não é admitida a coexistência no mesmo edifício de uma utilização-tipo XII das 3.ª ou 4.ª categorias de risco, com outra utilização-tipo, das 2.ª à 4.ª categorias de risco, com as seguintes exceções:

a) Utilizações-tipo II e VIII;
b) Utilização-tipo I, da 1.ª categoria de risco, quando destinada a proprietários ou funcionários da respetiva entidade exploradora.

2 — Nas situações distintas das referidas no número anterior, a coexistência num mesmo edifício de espaços ocupados por diferentes utilizações-tipo, deve satisfazer as seguintes condições:

a) Para efeitos de isolamento e proteção, os espaços ocupados por diferentes utilizações-tipo devem ser separados por paredes e pavimentos cuja resistência ao fogo padrão, EI ou REI, seja a mais gravosa das indicadas no quadro X abaixo:

QUADRO X

Escalões de tempo da resistência ao fogo de elementos de isolamento e proteção entre utilizações-tipo distintas

Utilizações-tipo	Categorias de risco			
	1.ª	2.ª	3.ª	4.ª
I, III, a X	30	60	90	120
II, XI e XII	60	90	120	180

b) Quando comuniquem com vias de evacuação protegidas, devem ser delas separados por paredes e pavimentos cuja resistência ao fogo padrão, EI ou REI, seja a mais gravosa das indicadas nos quadros X, XIX, XX e XXI;

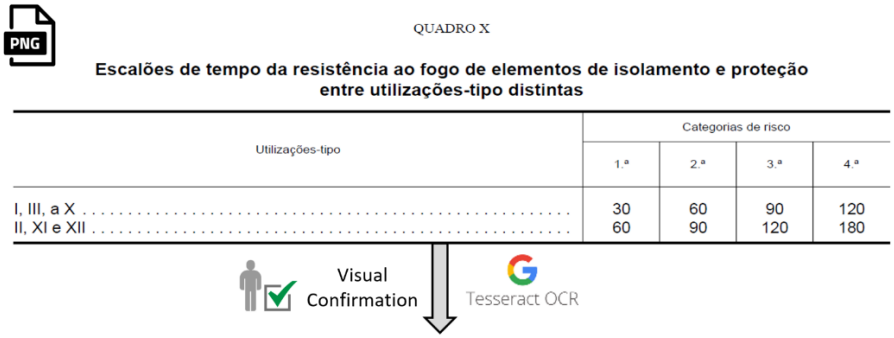
a) b)

Fig. 2. Example of the document processing in the first step: a) original regulation in PDF format; b) metadata extracted using the LLM-based Python functions

3.2 Automatic Table Extraction

Automatic table extraction proved to be one of the most critical steps of the entire process, due to the diversity and structural complexity of Portuguese regulations. These documents frequently include tables with multiple columns, asymmetric headers, merged cells, and sometimes continuation across several pages, which complicates their automatic detection and interpretation. To address these challenges, techniques combining OCR and text processing were employed. Given the consistent pattern of the tables, starting with the table number, followed by the title and respective content, it was possible to implement an automated extraction process without resorting to more complex methods. The procedure begins by converting the PDF into high-resolution images, where morphological operators are applied to identify tabular structures based on horizontal and vertical lines. The detected table regions are then cropped and processed using Tesseract for text recognition. In parallel, the Word file is analyzed to

identify the table titles, which are correctly associated with the extracted regions. The outcome is a collection of tables stored in a JSON file containing the numbering, title, content, and an indication of whether manual review is necessary, important in cases of gaps or structural ambiguities. Additionally, the LLM-based Python functions automatically generate a descriptive summary for each table based on its content. Thus, tabular information is extracted efficiently and in a structured manner, allowing its subsequent association with the logical hierarchy of the regulatory document. **Fig. 3** shows how the tabular information from **Fig. 2** is extracted and converted to a JSON file.




```
   [   {
        "table_number": "QUADRO X",
        "table_title": "Escalões de tempo da resistência ao fogo de
elementos de isolamento e proteção entre utilizações-tipo distintas",
        "table_content": [
            ["Utilizações-tipo", "Categorias de risco", "Categorias de
risco", "Categorias de risco", "Categorias de risco"],
            ["Utilizações-tipo", "1.ª", "2.ª", "3.ª","4.ª"],
            ["I, III, a X", "30", "60","90","120"],
            ["II, XI e XII", "60", "90", "120", "180"]
        ],
        "needs_review": false,
        "table_summary_pt": " QUADRO X apresenta os “Níveis de tempo de resis-
tência ao fogo para os elementos de isolamento e de proteção entre diferentes
utilizações-tipo”, categorizados por “Categorias de Risco”. Para a utiliza-
ção-tipo I, III a X, a resistência ao fogo exigida é de 30 minutos para a 1ª
categoria de risco e de 60 minutos para a 2ª. Por fim, para a utilização-tipo
II, XI e XII, a resistência exigida aumenta para 60 minutos e 90 minutos,
respetivamente.
    }  ]
```

Fig. 3. Example of the automatic table extraction process

3.3 Neural Network for Structuring

To automatically infer the hierarchical structure of regulatory documents, a neural model based on the BERT (Bidirectional Encoder Representations from Transformers) architecture was employed, widely recognized in the literature for its performance in text classification tasks. This model, named bert-base-multilingual-cased[6], is a transformer-based language model developed by Google to capture the bidirectional context of words. In this case, the model was trained on a dataset derived from real fire safety documents, in which each line was manually annotated using the iMerit platform, following a single-label per line approach. The annotated categories included labels such

[6] https://huggingface.co/google-bert/bert-base-multilingual-cased

as "Article Number", "Article Title", "Section Number", and "Subitem", covering a wide range of structural elements, including less frequent ones like "Annex Title", thus enabling the model to effectively learn to distinguish between different normative components. In total, 3223 instances were obtained, with the most represented label being "Item" with 1324 instances.

To mitigate class imbalance, a data augmentation strategy based on direct translation of underrepresented examples into multiple languages was applied (from 255 to 3000). This approach ensured a more balanced training dataset, with at least 300 examples per class.

The training process also incorporated a weighted loss function and early stopping mechanisms to prevent overfitting and promote generalization. Key hyperparameters included a learning rate of 3e-5, a batch size of 8, and a maximum of 100 training epochs. The dataset was split into training (70%), validation (20%), and test (10%). Additionally, to reduce ambiguities in classification, a sliding window approach was adopted, in which sequential blocks of text are processed together. This technique allows the model to observe the typical ordering of regulatory elements, such as "Chapter" followed by "Section" and "Article", improving classification accuracy.

To ensure the reliability of the results, a Human-in-the-Loop strategy was implemented. After automatic classification, the results are presented in an interactive graphical interface that enables experts to manually validate the assigned labels. Although this interface does not allow direct retraining of the model, it ensures that only validated content proceeds to the subsequent phases. Upon completion of this validation, a JSON file representing the document's logical structure is automatically generated, serving as the input for the next stage of the methodology. **Fig. 4**, shows the interface used by experts to review and validate the classifications automatically assigned.

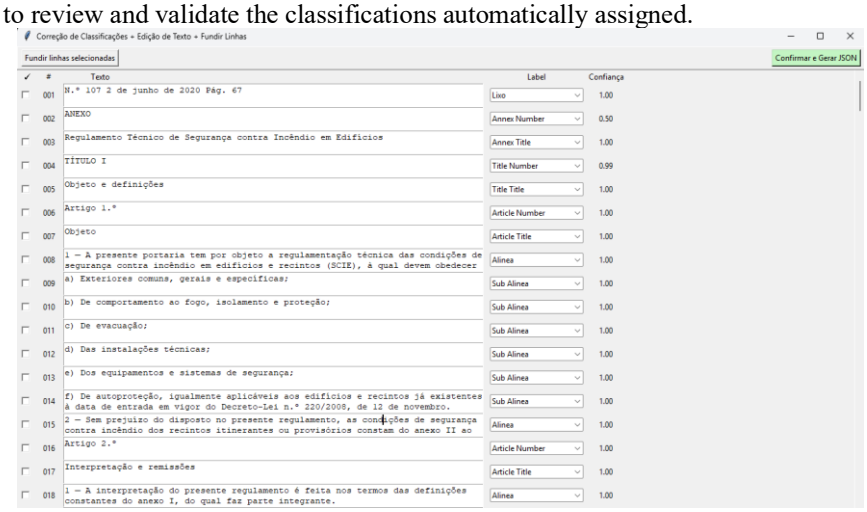


Fig. 4. Interface for manual validation and adjustment of automatic classifications

3.4 Enrichment of the Structured Data

After the structural classification and hierarchical organization of the document, the semantic enrichment phase begins. This step aims to complement the content with

additional relevant information. The process is carried out using the LLM-based Python functions, which analyse each element within its specific context.

At this stage, enrichment focuses exclusively on the lower-level elements of the document, such as articles, paragraphs, or items. For these parts, keywords are automatically extracted in both the native language of the document and in English, representing the main concepts addressed. Additionally, the model generates concise and contextual definitions for each keyword, based on the metadata collected during the initial step (e.g., country, language, type of regulation). Each element also receives an ID, for example, "PT-Ane0_CapII_Art17_Ite2_Suba", which clearly indicates its exact location within the document structure (i.e., SubItem 'a' of Item 2 of Article 17, located in Chapter II of Annex 0 of a Portuguese document). In addition, elements that have tables also receive a specific indicator (ID_T), which indicates the presence of a table in this element. As a final output, shown in **Fig. 5**, an enriched JSON file is produced, in which each textual element is associated with its corresponding keywords, definitions, unique identifiers, and links to previously identified tables. This file serves as a structured foundation for the generation of regulation graph.


```
{ "type":"annex",
  "number":"Annex 0",
  "title": "Regulamento Técnico de Segurança contra Incêndio em Edifícios",
  "titles":[
    {"type":"chapter",
      "number":"Capítulo II",
      "title":"Compartimentação geral contra incêndios",
      "sections":[],
      "articles":[
        {"type":"article",
          "number":"Artigo 17º",
          "title": "Coexistência de diferentes tipos de utilização",
          "alineas": [
            {"type": "item",
              "text": " 2 – Nas situações distintas das referidas no número ante-
rior, (…) diferentes utilizações-tipo, deve satisfazer as seguintes condições:",
                "type": "sub_item",
                "text": " a) Para efeitos de isolamento e proteção, os espaços
(..), seja a mais gravosa das indicadas no quadro X abaixo:",
                  "type": "sub_sub_items": [],
                "table_refs": ["QUADRO X", "Escalões de tempo da resistência ao
fogo de elementos de isolamento e proteção entre utilizações-tipo distintas",
                "table_summary": "QUADRO X apresenta os (…). Por fim, para os ti-
pos de utilizações-tipo 'II, XI e XII', os valores correspondentes às categorias
de risco são '60' na '1.ª', '90' na '2.ª', '120' na '3.ª' e '180' na '4.ª'".
                "table_summary_en":" QUADRO X presents the (…). Finally, for the
occupancy types 'II, XI e XII', the corresponding values for the risk categories
are '60' in the '1.ª', '90' in the '2.ª', '120' in the '3.ª', and '180' in the
'4.ª'",
                "keywords": {
                   "Isolamento": {
                     "Portuguese": "Isolamento é a capacidade de um elemento con-
strutivo ou sistema de limitar a passagem de chamas...",
                      "English": "Insulation is the ability of a building element
or system to limit the passage of flames..."},
                              (…)
               "ID":"PT-Ane0_TitIII_CapII_Art17_Ite2_Suba","ID_T":"(…)_Suba_TabX"]
}}}]}]}]}
```


Fig. 5. Example of the enriched JSON file

3.5 Regulation Graph Generation

After the enrichment step, the structured content is converted into an RDF graph, making it possible to clearly represent the structure, internal relationships, and other relevant information of the regulatory documents. This representation facilitates advanced queries, promotes interoperability with other knowledge bases, and makes it possible to use it in later stages, such as data visualization. To accomplish this step, Donkers et al. [2] developed a specific ontology, called the FireBIM Regulation Ontology (FRO). This lightweight, modular ontology is based on the recommendations of the W3C Linked Building Data Community group[7], also incorporating concepts already established by the AEC3PO ontology[8].

The FRO was specifically created to structure regulatory documents in a machine-readable format, allowing subsequent automated compliance checking. At an early stage, this ontology does not directly address specific technical elements (e.g. fire compartments or fire doors), but rather focuses on the logical and structural organization of regulatory texts. The FRO ontology is made up of three main classes that interact with each other in a complementary way: fro:Authority, which identifies the entity responsible for publishing and maintaining the regulation; fro:DocumentSubdivision, which represents the different internal parts of documents (such as chapters, articles or annexes); and fro:Reference, which allows internal or external references to be established between documents. The fro:DocumentSubdivision class is detailed through a hierarchy of specific subclasses, such as Annex, Article, Item, and Block, the latter being subdivided into more concrete elements, such as Table, Equation, Appendix, Figure, Chapter, and others. The structural relationships between these subclasses are established by specific properties, such as fro:hasAnnex, fro:hasTable or fro:hasParagraph. Based on this ontology, the enriched JSON file obtained in earlier steps is automatically analyzed to identify and extract its underlying structure. This process automatically generates the individuals associated with the concrete elements of the document, thus creating the final RDF graph (**Fig. 6**). During this conversion, the ID´s, multilingual information, and original textual content are used, guaranteeing a trustworthy and complete representation of the extracted data. The result of this procedure is exported in Turtle format (.ttl), allowing advanced queries and visualizations using specialized tools. Thus, the resulting RDF graph constitutes a robust semantic infrastructure suitable for complex semantic queries.

[7] https://github.com/w3c-lbd-cg/

[8] https://github.com/Accord-Project/aec3po

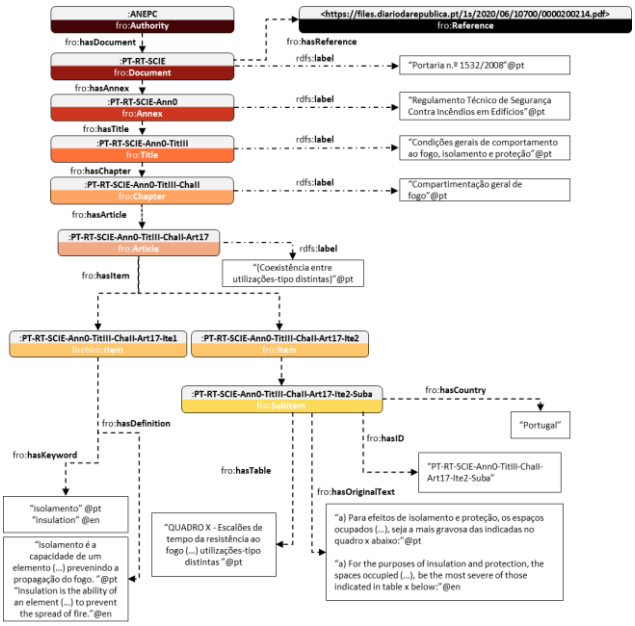


Fig. 6. Part of the RDF graph generated from the Portuguese regulations

3.6 Regulation Graph Visualization

The final phase of the methodology consists of the interactive visualization of the generated knowledge graphs. Based on the RDF graph extracted from the regulation under analysis, this step enables a clear, hierarchical, and semantically enriched exploration of the information. Designed as a support tool for final visualization, the interface (consolidates the most relevant elements of the processed document within a single, coherent environment.

The HTML interface is organized into distinct functional areas. In the top-left corner, a search panel allows users to quickly locate terms, articles, or specific expressions. Directly below is the keywords and definitions panel, which displays the main terms identified in the regulation, along with their definitions in both the original language of the document and in English. This information, sourced from the RDF, is linked to the content in the main display area, allowing users to access full definitions by clicking on the highlighted terms. The table of contents of the document, located in the bottom-left corner, is automatically generated based on the hierarchical structure of the RDF graph and displays only the top-level elements, namely Title, Chapter, Section, and Article. Presented in a compact format, it provides direct navigation to the main sections of the regulation, facilitating access to its structural content. At the center of the interface is the main visualization area, which displays the textual content of the regulation enriched with semantic links. This section also includes tables automatically extracted in Step 3.2, preserved in their original layout as informative and complementary elements. In the top-right corner, a visualization menu allows users to adjust parameters such as dark mode, article numbering, and level of detail. While offering simple functionalities,

this menu enhances the reading experience and adapts the content display to user preferences. In summary, the main function of this interface (**Fig. 7**) is to serve as a consolidated query layer and a final validation support tool. Although the data have already been verified in earlier stages, this integrated view allows potential inconsistencies or omissions to be more easily detected. If such issues are identified, the corresponding processing step must be repeated or the JSON file generated in Step 3.4 must be manually corrected. Beyond validation, publishing the interface as interactive webpages for different regulations would also make it possible to compare definitions of similar terms across countries, adding practical value. It is important to note that this interface was not designed for direct integration with other tools.

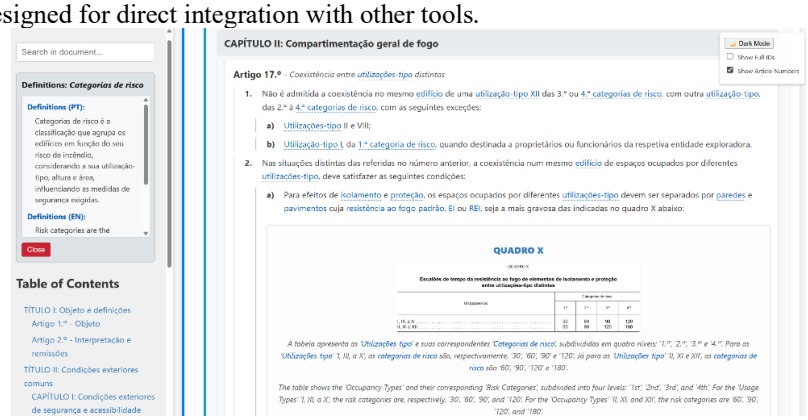


Fig. 7. Interactive HTML interface for visualization and validation of the processed document

4 Conclusions

The effectiveness automated compliance checking in fire safety regulations depends on the ability to convert these documents into structured, machine-readable formats. While manual conversion is accurate, it's extremely time-consuming and prone to inconsistency. On the other hand, fully automated solutions still struggle to deal with the complexity, variability, and structure of real regulatory texts. This contrast between manual and automated approaches highlights a key challenge: how to make the process more efficient without losing reliability. To address this, the study proposes an innovative hybrid methodology that combines the fast and automated information extraction capabilities of BERT and LLMs, with efficient, continuous, and rigorous expert validation. The methodology was evaluated through its practical application to Portuguese fire safety regulations, involving documents with more than 150 pages and characterized by complex and detailed normative structures. The results demonstrated a high level of adaptability and effectiveness in capturing, structuring, and semantically enriching regulatory content. Each document was processed in approximately 1.5 hours (using less than €1 for the processing), representing a substantial improvement in time and resource efficiency when compared to traditional manual methods, which could take several days and are more prone to human error and inconsistencies. This study highlights the critical importance of document digitalization as a foundational step in

developing robust automated compliance checking systems. Moreover, it reinforces the value of balancing automation with expert human validation to ensure accurate and reliable outcomes.

Despite promising first results, certain limitations were identified, such as the dependency on the quality of the underlying ontology and the occasional need for manual adjustments, due to the integration of AI models. Future work includes the exploration of another different techniques, like LayoutLMv2/v3, which explicitly capture the sequential logic and structural context of text, enabling more accurate classification of regulatory elements. In addition, the integration of multi-shot learning strategies into LLM-based Python functions will also be considered. Despite requiring large training datasets, these strategies can produce more generalizable and robust models, reducing manual intervention. These improvements are particularly relevant for the semantic classification stage (Step 3.3), which is a central role in the quality of the final output. Moreover, the methodology will be extended across regulatory documents from different countries, whose structural and semantic characteristics may change considerably.

Acknowledgments. This work has been funded by the NORTE 2030 Programme through the European Regional Development Fund (ERDF), under project FireBIM, reference NORTE2030-FEDER-00358400. Also, this work was supported by: UID/04708/2025 and https://doi.org/10.54499/UID/04708/2025, of the CONSTRUCT - Instituto de I&D em Estruturas e Construções - funded by Fundação para a Ciência e a Tecnologia, I.P./ MECI, through the national funds.

References

[1] R. K. Chaudhary, A. Lucherini, T. Gernay, and R. Van Coile, "Evaluation of anticipated post-fire repair cost for resilient design of composite slab panels," *Journal of Building Engineering*, vol. 52, p. 104460, 2022, doi: https://doi.org/10.1016/j.jobe.2022.104460.

[2] A. Donkers and E. Petrova, *Converting Fire Safety Regulations to SHACL Shapes Using Natural Language Processing*. 2024.

[3] N. Nisbet and L. Ma, "Using RASE semantic mark-up for Normative, Definitive and Descriptive knowledge," *Taylor & Francis*, pp. 1–7, 2023, Accessed: Jul. 15, 2025. [Online]. Available: https://discovery.ucl.ac.uk/id/eprint/10178378/1/ECPPM-2022-NNa.pdf

[4] J. Zhang and N. El-Gohary, "Semantic NLP-Based Information Extraction from Construction Regulatory Documents for Automated Compliance Checking," *Journal of Computing in Civil Engineering*, vol. 30, p. 141013064441000, Jul. 2013, doi: 10.1061/(ASCE)CP.1943-5487.0000346.

[5] J. Zhang and N. M. El-Gohary, "Extending Building Information Models Semiautomatically Using Semantic Natural Language Processing Techniques," *Journal of Computing in Civil Engineering*, vol. 30, no. 5, Sep. 2016, doi: 10.1061/(asce)cp.1943-5487.0000536.

[6] Costa, G., Vakaj, E., Beach, T., Lavikka, R., Lefrançois, M., Zimmermann, A., ... & Keberle, N. Formalization of building codes and regulations in knowledge graphs, in Digital Building Permit conference 2024. Barcelona, Spain, Apr. 2024.

[7] Al-Turki, D., Hettiarachchi, H., Gaber, M. M., Abdelsamea, M. M., Basurra, S., Iranmanesh, S., Vakaj, E. Human-in-the-Loop learning with LLMs for efficient RASE tagging in building compliance regulations, IEEE Access, vol. 12, pp. 185291-185306, 2024, doi: 10.1109/ACCESS.2024.3512434.

[8] N. L. Schroeder, C. Davis Jaldi, and S. Zhang, "Large Language Models with Human-In-The-Loop Validation for Systematic Review Data Extraction," Jan. 2025. doi: https://doi.org/10.48550/arXiv.2501.11840Focustolearnmore.