2024 European Conference on Computing in Construction
Chania, Crete, Greece
July 15-17, 2024

NLP-BASED DATA-ENRICHMENT FOR BUILDING MANAGEMENT

Hamada Elshaboury[1], Fulvio Re Cecconi[2], Enrico De Angelis[3], Luciano Baresi[4], Vincenzo Scotti[5]
[1,2,3]ABC Department, Politecnico di Milano, Milan, Italy
[4,5]DEIB Department, Politecnico di Milano, Milan, Italy

Abstract

One of the main challenges in building management is dealing with the large amount of unstructured data produced throughout the asset's life cycle. At handover, Building Information Models often provide low-quality, incomplete data, necessitating extensive manual rework to elicit information from many sources. To minimize this manual rework, this study proposes an automated *Information Extraction* (IE) procedure, applied to the design and construction documents to extract information, enrich a model (COBie format), and maximize the transfer of structured data to the client. The proposed approach is based on *Natural Language Processing (NLP)* and adopts *Transformer-based Named Entity Recognition* (NER) and *Relation Extraction* (RE) methods for IE. It was evaluated and achieved good performance, with average F1 scores of 0.73 for NER and 0.91 for RE, representing a step toward a reliable tool for an enhanced data handover process.

Introduction

During the design and construction phase, a significant number of text documents are produced, not only by the design team but also by the contractor, suppliers, client, and consultants. Design documents and specifications are followed by building performance reports (structural safety, energy, acoustics...), contractor's method statements, risk assessments, health, safety, and maintenance plans, inspection and commissioning reports, product sheets and declarations, etc. However, all these documents contain poorly accessible, unstructured information that can only be retrieved through manual searches (Marzouk and Enaba, 2019).

Building Information Modelling (BIM) has emerged as the leading digital solution in the Architecture, Engineering, and Construction (AEC) industry to support the design, construction, and subsequent management phases, as facility managers can use the same physical and functional information defined in the early phases to plan and manage maintenance operations. However, the actual situation is not ideal. On the one hand, digital models are often unable to easily incorporate all the data produced (also because of the low digital readiness of stakeholders and their models). On the other hand, the information created is often incomplete, inaccurate, inconsistent, complex, large, and poorly standardized. As a result, BIM, as an enabler for providing reliable information for building management, faces numerous challenges (Tsay *et al.*, 2022; Ullah *et al.*, 2019).

In response to these issues, US agencies, with the help of the National Institute of Building Sciences, have developed a standard digital data schema known as Construction Operations Building Information Exchange (COBie) to streamline access and manipulation of BIM data. A COBie file can be exported as an Excel spreadsheet or IFC STEP file from any BIM model. However, a COBie file reflects the quality of the BIM model from which it was generated. In addition, some of the information in the BIM model dataset is not exported in the COBie file (Kumar and Teo, 2020), increasing the manual effort required to have complete information to support the management of maintenance and operations planning.

Recognizing these challenges, this study uses NLP to enhance the handover process, as shown in Figure 1, by developing a methodology in three steps: data preparation (1), information extraction (2), and integration into COBie (3).

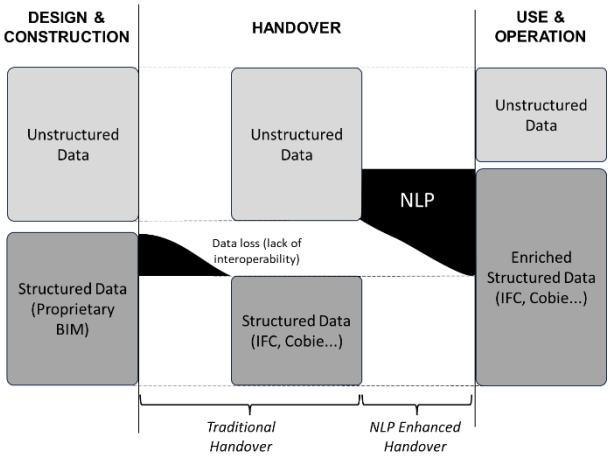


Figure 1: Handover problem statement and proposed solution

This paper is organized as follows: an introductory *State of the Art*, a *Methodology* section detailing the proposed approach, a third section for *Experiments*, and, eventually, a *Conclusion* and *References* section.

State of the Art

NLP and Data Management in the AEC Industry

NLP joins linguistics with computer science, empowering our lives with tools and techniques to perform machine

reading, comprehending, and analyzing human-written documents (Salama and El-Gohary, 2013). As a result of these capabilities, NLP has been widely applied in a highly documented sector as the AEC industry (Marzouk and Enaba, 2019) to improve the management of many processes such as safety (Zhang *et al.*, 2019), risk assessment (Zou *et al.*, 2017), quality control (Jeon *et al.*, 2021), document management (Qady and Kandil, 2014), and code compliance checking (Zhang and El-Gohary, 2013; Zhang and El-Gohary, 2015).

The cornerstone of improving construction management is built on the efficiency and effectiveness of data management through transforming unstructured texts into structured information, enabling organizing and analyzing information, and identifying patterns supporting the enhancement of planning and decision-making. This transformation from unstructured data into structured information, as shown in Figure 2, is executed through a suite of NLP tools, such as tokenization, part-of-speech (POS) tagging, phrase structure analysis (PSG), stemming and integrating with machine learning (ML) and deep learning (DL) techniques to perform advanced tasks such as Text Classification (TC), Document Clustering (DC), Information Extraction (IE), and Information Retrieval (IR) to support data management.

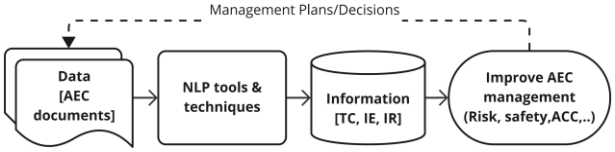


Figure 2: Methodology of NLP application in the AEC industry

Information Extraction (IE)

IE is a process that aims to automatically extract specific information from unstructured data and represent such information in a structured format (Salama and El-Gohary, 2013). IE includes sub-tasks such as *Named Entity Recognition* (NER), which recognizes and classifies entities into predefined categories, and *Relation Extraction* (RE), which extracts the semantic relations between these entities (Zhang and El-Gohary, 2013). IE methods include rule-based, machine learning (ML)-based, and deep learning (DL)-based approaches.

In rule-based approaches, the target information is extracted based on pattern-matching rules coded by experts. These rules utilize NLP tools such as POS and PSG for sentence analysis, term matching, and semantic analysis (Zhang and El-Gohary, 2015). For example, (Zhang and El-Gohary, 2013) proposed a semantic and rule-based approach to extract information from building codes to support compliance checking. Also, (Li *et al.*, 2016) developed a rule-based approach to extract spatial rules and check compliance of utility spatial specifications with requirements. (Liu and El-Gohary, 2021) proposed a dependency parsing model to extract dependency relations between the semantic information elements from bridge inspection reports and represent them in semantic information sets. Although rule-based approaches have low scalability and flexibility and require manual efforts for pattern formalization, they have high accuracy.

In ML-based approaches, models rely on learning from data (Russell and Norvig). For example, (Zhang and El-Gohary, 2016) developed an approach to classify relations between BIM and regulations concepts using ML classifiers (SMV, NB, DT, K-NN). ML-based approaches are more flexible for modifications, less costly, and require fewer manual efforts than rules-based approaches. However, manual work is still necessary for preprocessing and preparing datasets for training and testing.

DL-based models comprise multiple neural network layers from various algorithms, allowing them to deeply understand and represent complex and unstructured data for NLP tasks (Lecun *et al.*, 2015). Recently, sequence-to-sequence models such as Recurrent Neural Networks (RNNs), LSTM and RGU, along with Convolutional Neural Networks (CNNs), have been used extensively for IE. For example, (Zhong *et al.*, 2020) proposed an approach to support quality management checking using a Bi-LSTM-CRF-based NER model to identify and label entities in the clauses and an LSTM-MLP-based RE model to classify relations between entities into five predefined groups. (Zhang and El-Gohary, 2022) used a Bi-LSTM-MLP-based RE model to extract the semantic relations of building code sentences. (Wang and El-Gohary, 2023) proposed a CNN and RNN-based method to extract relations between entities that describe fall protection requirements from safety regulations and represent entities with their relations in query graphs.

Most recently, transformer-based models have gradually become a trend, utilizing multi-headed self-attention mechanisms to deeply understand text context and representation (Vaswani *et al.*, 2017) and enabling the development of pre-trained models (e.g., BERT) (Devlin *et al.*, 2018). For example, (Zhang and El-Gohary, 2023) used a BERT–based model to determine the semantic relation probability between regulatory and IFC concepts. DL-based approaches are more effective in handling various data types that exhibit high levels of complexity, achieving a high semantic analysis and word representation.

Data Handover for Building Management

There is a paramount interest in using project data for the life cycle with the evolution of Building Information Modelling, which promotes the incremental collection of data (Lindkvist and Whyte, 2013). Nevertheless, limited studies have focused on effective and automated building information sharing during the commissioning and close-out stage for efficient building handover, operations, and maintenance through its life cycle (Singh and Anumba, 2023). As a result of this lack of research in the field, the automated digital practices to perform the commissioning and handover process are still largely unknown.

Many studies (Cavka *et al.*, 2017; Tan *et al.*, 2018) collectively underscore the importance of knowledge trans-

fer, information flow, and the use of technology in enhancing information management during handover for better asset and facility management. COBie, a specification for the Construction Operations Building Information Exchange, aims to streamline data handover from design and construction to facility management (Schwabe *et al.*, 2018). It is a Model View Definition (MVD) of the Industry Foundation Classes (IFC) as defined in ISO (ISO 16739-1:2018) and was first introduced by the US Army Corps of Engineers to enable organizations to electronically capture and record important project data at the point of origin. Because the fundamental purpose of COBie is to efficiently obtain information (relevant to facility management) generated in the design and construction phases, an Excel-based spreadsheet that anyone can easily access and handle is often used (Shin *et al.*, 2022). A COBie datasheet consists of 20 workbooks in which data about the facility is stored systematically. The columns and their location in the workbooks are fixed, and changing the location of columns is restricted. Different colors are used to define data inside workbooks (Kumar and Teo, 2020). Although COBie has been widely used (Maltese *et al.*, 2017)(Kumar and Teo, 2020) (Asare *et al.*, 2023) in practice, creating COBie deliverables can be problematic due to misunderstandings among end users and insufficient software implementation. This can ultimately lower acceptance among practitioners.

COBie data capturing and compiling is a complex process; after each project phase ends, a COBie sheet deliverable is required, which needs to be verified with the project stage requirements known as COBie "data drops" (Love *et al.*, 2014). A typical data drop process requires, among other activities, extracting IFC models from native BIM models from different disciplines, merging and verifying for consistency, updating missing information through documents, and verifying that the dependency and links are maintained. This implies that data needs to be entered manually inside the COBie datasheet. The transfer of information from as-built documents is one of the biggest problems, as much of the information is not transferred in COBie format or is transferred with errors. Natural Language Processing (NLP) tools can improve the information management process by increasing the amount of structured data transferred to the client via COBie.

Methodology

In this section, we outline the proposed data enrichment methodology, which comprises three steps, as depicted in Figure 3: data preparation, information extraction, and integration into COBie. Recently, NLP has offered many DL-based tools to automate text analysis and process large batches of documents, so the proposed methodology exploits NLP techniques for language analysis to automate IE and integration.

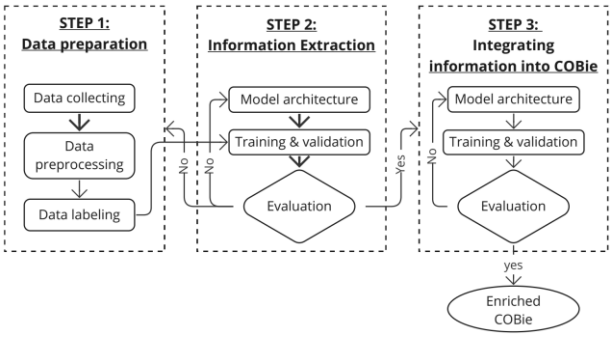


Figure 3: Methodology framework

Step 1: Data Preparation

Given the use of deep learning, a data-driven technique, ground truth data is required to train and evaluate the proposed models. This step, which includes data collection, preprocessing, and labeling, is crucial for the model's development.

For data collection, a corpus of unstructured documents for design and construction was collected from case studies provided by one of Italy's largest engineering companies in Milan and converted into plain text format for easier processing.

Table 1: Categories of the Information Entities

Entity Category	Definition
Object ID	Identify the ID code of an entity "Wall 01"
Object	Identify an entity "internal wall"
Description	Clarify a description for an entity, action, attribute
Section	Identify the geographical location of an object "north-facing windows"
Action	Identify an activity for an object "demolishing", "building", "painting"
Attribute	Identify a character that distinguishes an object's "thickness," "airtightness," and "or strength."
Value	Identify value "15", a class "U", or category "GBK(A)" for an attribute.
Unit	Identify a unit measure for attribute value "cm", or "REI" for fire resistance.
Constraint	Identify a specific condition for an attribute or value or action "equal to", "in a range", "partially, completely, "fully".
Reference	Identify a regulation, standard, or code explaining how the attribute must be evaluated or measured and the object or action must be realized.

After collecting the raw documents, data preprocessing was conducted to remove non-textual parts like figures, headers, and footers, and sentence segmentation was applied to isolate the individual sentences. Labeling involves annotating a specific word or more with the corresponding label or the relation between two entities. This

process begins with defining labels for the target information entities and relation types the proposed models intend to recognize and extract. In this research, the defined categories for the information entities and their relation types are presented in Tables 1 and 2 and illustrated in Figure 4.

Table 2: Relation types between entities

Relation Type	Definition
Identifies	Links the unique identifier code (*Object ID)* to the corresponding *Object*
PartOf	Links an *Object* to another *Object*, establishing a parent-child relationship and the hierarchical structure
DescribedBy	Links an *Object* or *Action* to a *Description* that provides more details
LocatedIn	Links an *Object* to a specific geographical location (*Section*) in a building
RequireAction	Links an *Object* to an *Action* that specifics the activities or tasks to be realized
HasAttribute	Links an *Object* to an *Attribute* that specifies its characteristics
HasValue	Links an *Attribute* with a specific *Value* for its characteristics
MeasuredIn	Links a *Value* to a *Unit*, specifying the unit of measurement for the value
HasConstraint	Links an *Action* or *Value* to a *Constraint*, specifying the conditions for this value or action.
Referenced	Links either an *Object*, *Action*, or *Value* to a *Reference* establish the connection between a regulation, standard, or code and its application to an object, action, or value.

An open-source data labeling platform, "Label Studio" was used to facilitate manual labeling, applying two distinct notations. For the NER task, the Beginning-Inside-Outside (BIO) labeling schema was used, where each word in a sentence was labeled manually to indicate whether it was the beginning or inside of a specific entity or not a part of any entity (outside). In the RE task, a single relation type was assigned between each input pair of entities using the tagging schema of the SemEval 2010-Task 8 datasets to mark the entity's positions with <e1> and </e2> XML-like tags, as shown in Figure 6.b.

Step 2: Information Extraction (IE)

In this study, we introduce a modular IE pipeline comprising two respective tasks: 1) *Named Entity Recognition* (NER) to process a plain text document or a specific segment and extract all entities that exist in the text according to Table 1; and 2) *Relation Extraction* (RE) to extract and classify the relation between each pair of entities following Table 2. By iteratively applying these two steps across the entire document corpus, we methodically construct a comprehensive and structured knowledge base that encapsulates atomic entities and delineates their interrelations, as shown in Figure 5. This step consists of three sub-steps: model development, training, and evaluation, as discussed below.

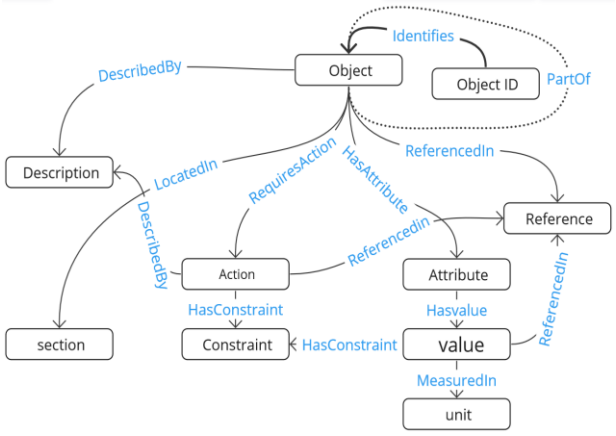


Figure 4: Entity categories and Relation types

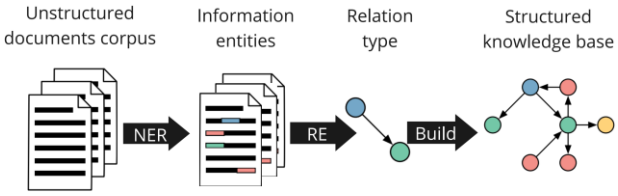


Figure 5: Information Extraction (IE) pipeline

Model development

Following current state-of-the-art NLP applications built using a transformer neural network, pre-trained language models were selected as the base for the NER and RE tasks composing the pipeline. As a classification task, a bi-encoder architecture is more suitable where the model architecture consists of an input, encoding, and output layer, as shown in Figure 6 and described below:

Input (embedding) Layer: converts each word in the input text to a vector representation by tokenizing sentences, padding, converting tokens to IDs, adding an attention mask and two special tokens [CLS] and [SEP] to determine the starting and ending of each sentence.

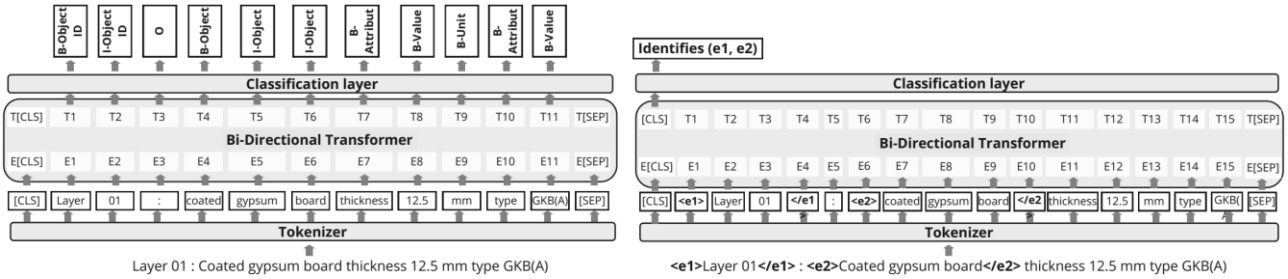
Encoding Layer: feeds the embedding vectors of tokens from the input layer into transformer encoder blocks passing with the multi-head attention followed by the feed-forward network, with add & normalize layers, and the output of each block is passed to the next one as input. As a result, the final output from the last block represents a highly contextual embedding vector (hidden state), considering both the left and right contexts, semantics, and synthetic relationships.

Output (classification) layer: takes the final hidden states from the last layer as input, applying a SoftMax function to calculate probabilities.

In the NER task, as depicted in Figure 6. a, given a piece of text passing through the input and encoding layers, the

language model computes a probability distribution over the class labels for each input token and selects the most probable class label associated with each token. The class label indicates whether the token corresponds to the beginning, inside, or outside of a specific entity. Once a text is tagged, named entities are extracted, searching for the beginning of each entity and taking that first token with all the inside tokens associated with that same entity.

In the RE task, as illustrated in Figure 6. b, the input piece of text has four additional tokens to mark the beginning and end of the first and second entities; passing through the input and encoding layers, the language model computes a probability distribution over the class labels and selects the most probable class label associated with the input text to tag the relation type between the two marked entities.



a. Transformer-based NER model *b. Transformer-based RE model*

Figure 6: Architecture of the language model for the NER and RE tasks

Model training

As pre-trained language models on a large general domain corpus, there is a need to fine-tune these models for specific downstream tasks (i.e., NER and RE). The objective is to minimize the negative log-likelihood of the target class(es) associated with the input text. Cross-entropy, serving as a cost function, calculates the discrepancy between the actual and predicted labels/relations, assessed after the model's forward pass; during the backpropagation, the optimizer updates the model weights based on the gradient of the loss and learning rate scheduler.

Evaluation

In the realm of IE, there are three common metrics for performance evaluation: Precision (P), Recall (R), and F1-score (Zhai and Massung, 2016), as delineated in Equations 1 to 3, with their values ranging from 0 to 1, where 0 is the worst score, and 1 is the best. For instance, R = 1 signifies that the model correctly identified all actual positives, P = 1 indicates that all positive predictions were correct, and F1 = 1 indicates a perfect harmonic balance between R and P.

$$Recall\ (R)\ =\ \frac{TP}{TP+FN} \qquad (1)$$

$$Precision\ (P)\ =\ \frac{TP}{TP+FP} \qquad (2)$$

$$F1 =\ 2\,x\,\frac{P\,x\,R}{P+R} \qquad (3)$$

Step 3: Information integration into COBie

In this step, we merge the extracted information with the COBie format, where the relations, as depicted in Table 2 , establish interactive connections and contextual associations between the entities, creating a semantic-structured knowledge base to support the integration process. To achieve this, two models will be developed: 1) a Transformer-based model to classify the extracted entities from Step 2 into COBie sheets, 2) a Rule-based model to match the entities with COBie sheets, and embed information into COBie data fields. Further details and experiments on this step will be pursued in future research directions.

An example to illustrate the application of the proposed methodology to enrich the COBie schema is shown in Figure 7.

Experiments

In this section, we describe the implementation of the IE pipeline, training settings, and evaluation results.

Data preparation

Wall specification documents from the collected corpus were used as a case study to prepare the dataset to train and test the models. These documents include detailed information regarding wall layers, materials, characteristics, and state-of-the-art construction procedures. Following the procedure outlined in Step 1, a dataset was developed encompassing approximately 2400 entities and 1000 relations representing all entity categories and relation types in Figure 4.

Implementation

The proposed IE pipeline was implemented using two prominent bidirectional language models, BERT (Devlin *et al.*, 2018) and RoBERTa (Conneau *et al.*, 2019), from the Transformers library in the Hugging Face model hub. Table 3 Provides the considered models' variants and reference names; some were pre-trained on multiple languages, while others were pre-trained on Italian. The implementation was conducted using PyTorch, built with Python 3, and executed on a T4 GPU provided by Google Colaboratory.

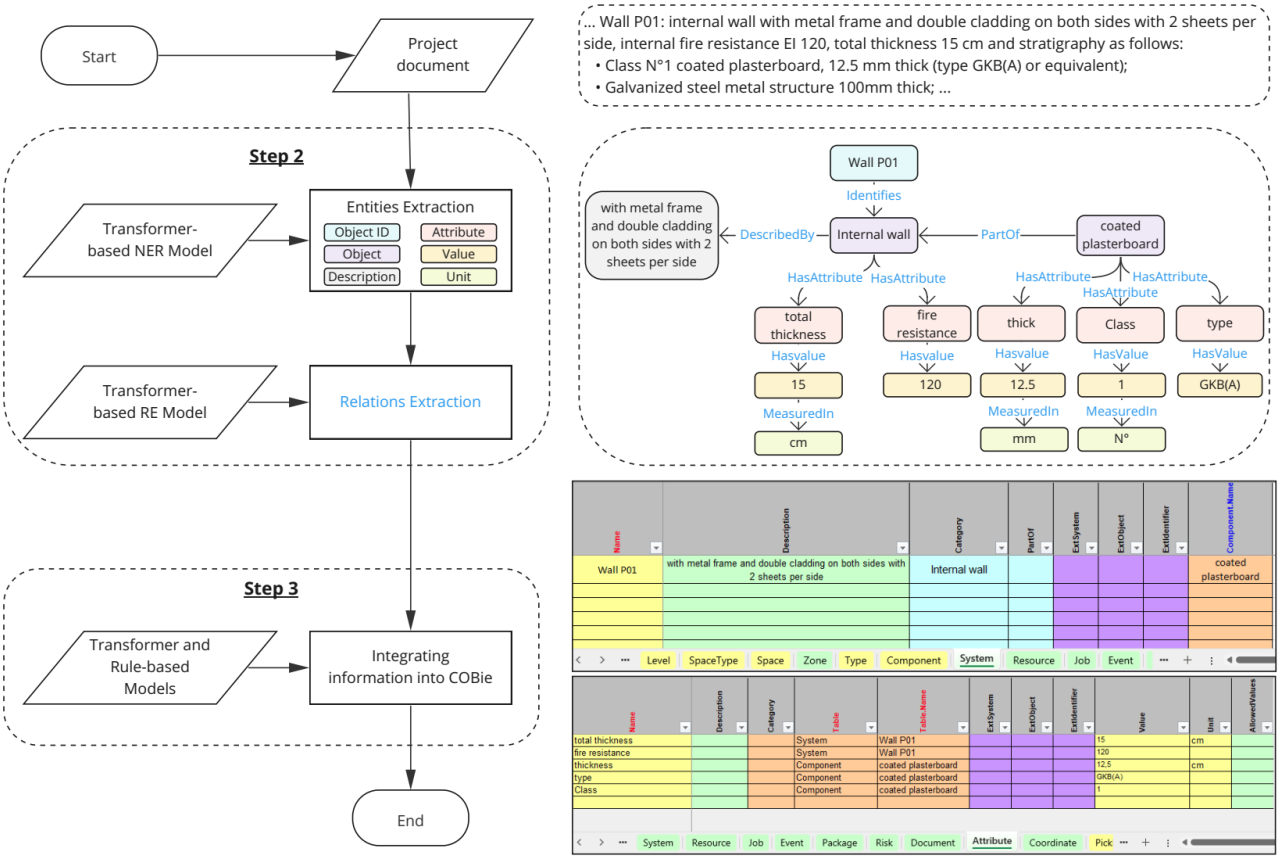


Figure 7: Enriche COBie schema using the proposed methodology

Table 3: Variants of BERT and RoBERTa models for the NER and RE tasks

Reference Name	No. of Layers	Parameters
Multilingual Bert (Cased - Uncased)	12 layers	110M
XLM-Roberta Large	24 layers	550M
Italian ALBERTO (Uncased) (Polignano *et al.*, 2019)	12 layers	110M
dbmdz-Italian Bert -XXL (Cased - Uncased)	12 layers	110M

Training

The language models were fine-tuned using the labeled dataset, which was split into training, validation, and testing sets with 80%,10%, and 10%, respectively, for the NER task and 64%, 12%, and 24%, respectively for the RE task. The configuration and hyperparameters stayed the same for the models, as provided in Table 4.

Evaluation and results

The evaluation was conducted using the testing dataset alongside selected transformer-based models. Table 5 Provides the results of the experiments where the "*xlm-roberta-large*" model emerges as the most effective in the IE pipeline, achieving the highest precision, recall, and F1 scores for both tasks. It achieved F1 scores of 0.73 for NER and 0.91 for RE, respectively, outperforming other models with an average increase of 8.8% for NER and 5.2% for RE tasks.

Table 4: Values of the hyperparameters of the language models

Hyperparameter	Value
Maximum length	128
Batch size of data loader	32
Adam learning rate	1e-5
Dropout Rate	0.1
Early stopping condition	5 epochs

Regarding the results of the RoBERTa model, the model indicated a good capability in identifying entities "Object", "Value", "Attribute", and "Unit", achieving relatively high F1-scores ranging from 0.75 to 0.90; conversely, it encountered challenges in recognizing entities "Description", "Action", and "Constraint" as reflected by lower F1-scores from 0.33 to 0.62. For the RE task, the model demonstrated a robust performance across various relation types achieving high F1-scores ranging from 0.88

to 0.97 for relations "HasValue","DescribedBy","HasAttribute","Partof","ReferencedIn", and "MeasuredIn" and moderate F1- scores from 0.67 to 0.78 for "LocatedIn", "RequiresAction", and "ConstrainedBy" ones. The limitations in the model's performance in identifying and recognizing these entities or relation types can be attributed to restricting their examples in the training dataset. Consequently, the model encountered interpretation complexities when distinguishing within the text. These results are satisfactory given the current volume of the dataset and underscore the potential for enhancing the performance once more data becomes available.

Table 5: Results of the proposed IE pipeline (NER and RE tasks) with the language models

Model	NER			RE		
	P	R	F1	P	R	F1
Bert-base-multilingual-cased	0.63	0.69	0.64	0.89	0.89	0.89
Bert-base-multilingual uncased	0.64	0.64	0.62	0.83	0.84	0.83
xlm-roberta-large	**0.74**	**0.74**	**0.73**	**0.91**	**0.91**	**0.91**
Bert_uncased_ italian_alb3rt0	0.69	0.68	0.67	0.86	0.86	0.86
dbmdz/bert-base-italian-xxl-cased	0.64	0.68	0.63	0.87	0.86	0.86
dbmdz/bert-base-italian-xxl-uncased	0.67	0.69	0.65	0.85	0.86	0.85

Conclusion

This study contributes to building management by enhancing information management during the handover process, enabling effective planning and management of maintenance operations. It proposed a methodology aiming to provide an enriched data schema (COBie) that integrates information from both the design and construction phases with less manual effort and time consumption, overcoming some of the prevalent challenges, such as data interoperability issues and the high level of documentation in the construction sector. The methodology comprised three main steps: data preparation, information extraction (IE), and integration into the COBie format. Entity categories and relation types were identified to prepare the labeled data, and two transformer-based models, BERT and RoBERTa, with their variants, were utilized to implement the proposed IE pipeline, automating *Named Entity Recognition* (NER) and *Relation Extraction* (RE) tasks. The proposed IE pipeline was tested using wall specification documents as a case study, and the "*xlm-roberta-large*" language model delivered the highest performance on both tasks, achieving an average F1-score of 0.74 in the NER task and 0.91 in the RE. These scores can be considered good results given the current data volume.

Future research directions include: 1) Implementing Step 3 to integrate the extracted entities and their relations into the COBie format; 2) Expanding the labeled dataset by considering a broader range of design and construction documents to improve the models' performance, scalability, and robustness; 3) Large Language Models (LLMs) will be used for the NER and RE tasks, and evaluated for their effectiveness in the given context, and 4) Performing a comparison to a COBie data extraction tool.

References

Asare, K.A.B., Liu, R. and Anumba, C.J. (2023). Building information modeling for airport facility management: the case of a US airport, Thomas Telford.

Cavka, H.B., Staub-French, S. and Poirier, E.A. (2017). Developing owner information requirements for BIM-enabled project delivery and asset management. *Automation in Construction*, 83:169–183.

Conneau, A., Khandelwal, K., Goyal, N., Chaudhary, V., Wenzek, G., Guzmán, F., Grave, E., Ott, M., Zettlemoyer, L., & Stoyanov, V. (2019). Unsupervised Cross-lingual Representation Learning at Scale. Proceedings of the Annual Meeting of the Association for Computational Linguistics, 8440–8451.

Devlin, J., Chang, M.-W., Lee, K. and Toutanova, K. (2018). BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. *NAACL HLT 2019*, 1:4171–4186.

ISO 16739-1:2018 - Industry Foundation Classes (IFC) for data sharing in the construction and facility management industries — Part 1: Data schema.

Jeon, J., Xu, X., Zhang, Y., Yang, L. and Cai, H. (2021). Extraction of Construction Quality Requirements from Textual Specifications via Natural Language Processing. SAGE Publications, 2675:222–237.

Kumar, V. and Teo, E.A.L. (2020). Perceived benefits and issues associated with COBie datasheet handling in the construction industry. *Facilities*, 39:321–349.

Lecun, Y., Bengio, Y. and Hinton, G. (2015). Deep learning. Nature Publishing, 521: 436–444.

Li, S., Cai, H., Asce, M. and Kamat, V.R. (2016). Integrating Natural Language Processing and Spatial Reasoning for Utility Compliance Checking. *Journal of Constr. Engineering and Management*, 142:04016074.

Lindkvist, C. and Whyte, J. (2013). Challenges and Opportunities Involving Facilities Management in Data Handover: London 2012 Case Study. *AEI 2013*, 670–679.

Liu, K. and 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.

Love, P.E.D., Matthews, J., Simpson, I., Hill, A. and Olatunji, O.A. (2014). A benefits realization management building information modeling framework for asset owners. *Automation in Construction*, 37:1–10.

Maltese, S., Moretti, N., Re Cecconi, F., Ciribini, A.L.C. and Kamara, J.M. (2017). A Lean Approach to Enable Sustainability in the Built Environment through BIM. *Journal of Technology for Architecture and Environment*, 13:278–286.

Marzouk, M. and Enaba, M. (2019). Text analytics to analyze and monitor construction project contract and correspondence. *Auto. in Construction*, 98:265–274.

Polignano, M., Basile, P., de Gemmis, M., Semeraro, G., & Basile, V. (2019). AlBERTo: Italian BERT language understanding model for NLP challenging tasks based on tweets. *CEUR Workshop Proceedings*, 2481.

Qady, M. and Kandil, A. (2014). Automatic clustering of construction project documents based on textual similarity. *Automation in Construction*, 42:36–49.

Russell, S. and Norvig, P. (n.d.). Artificial Intelligence-A Modern Approach (3rd Edition).

Salama, D.M. and El-Gohary, N.M. (2013). Semantic Text Classification for Supporting Automated Compliance Checking in Construction. *Journal of Computing in Civil Engineering*, 30: 04014106.

Schwabe, K., Dichtl, M., König, M. and Koch, C. (2018). COBie: A Specification for the Construction Operations Building Information Exchange. *Building Information Modeling*, 167–180.

Shin, S., Moon, H. and Shin, J. (2022). BIM-Based Maintenance Data Processing Mechanism through COBie Standard Development for Port Facility. *Applied Sciences*, 12:1304.

Singh, J. and Anumba, C.J. (2023). Building commissioning process and documentation: a literature review and directions for future research. *International Journal of Construction Management*.

Tan, A.Z.T., Zaman, A. and Sutrisna, M. (2018). Enabling an effective knowledge and information flow between the phases of building construction and facilities management. *Facilities*, 36:151–170.

Tsay, G.S., Staub-French, S. and Poirier, É. (2022). BIM for Facilities Management: An Investigation into the Asset Information Delivery Process and the Associated Challenges. *Applied Sciences*, 12:9542.

Ullah, K., Lill, I. and Witt, E. (2019). An overview of BIM adoption in the construction industry: Benefits and barriers. *Emerald Reach Proceedings Series*, 2:297–303.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., & Polosukhin, I. (2017). Attention Is All You Need. Advances in Neural Information Processing Systems, 5999–6009.

Wang, X. and El-Gohary, N. (2023). Deep learning-based relation extraction and knowledge graph-based representation of construction safety requirements. *Automation in Construction*, 47:104696.

Zhai, C. and Massung, S. (2016). *Text Data Management and Analysis: A Practical Introduction to Information Retrieval and Text Mining*.

Zhang, F., Fleyeh, H., Wang, X. and Lu, M. (2019). Construction site accident analysis using text mining and natural language processing techniques. *Automation in Construction*, 99:238–248.

Zhang, J and El-Gohary, N.M. (2015). Automated Information Transformation for Automated Regulatory Compliance Checking in Construction. *Journal of Computing in Civil Engineering*, 29(4).

Zhang, J. and El-Gohary, N.M. (2016). Extending Building Information Models Semiautomatically Using Semantic Natural Language Processing Techniques. *Journal of Computing in Civil Engineering*, 30(5).

Zhang, J. and El-Gohary, N.M. (2013). Semantic NLP-Based Information Extraction from Construction Regulatory Documents for Automated Compliance Checking. *Journal of Computing in Civil Engineering*, 30(2).

Zhang, R. and El-Gohary, N.M. (2022). Hierarchical Representation and Deep Learning–Based Method for Automatically Transforming Textual Building Codes into Semantic Computable Requirements. *Journal of Computing in Civil Engineering*, 36(5).

Zhang, R. and El-Gohary, N. (2023). Transformer-based approach for automated context-aware IFC-regulation semantic information alignment. *Automation in Construction*, 145:104540.

Zhong, B., Xing, X., Luo, H., Zhou, Q., Li, H., Rose, T. and Fang, W. (2020). Deep learning-based extraction of construction procedural constraints from construction regulations. *Advanced Eng. Informatics*, 43:101003.

Zou, Y., Kiviniemi, A. and Jones, S.W. (2017). Retrieving similar cases for construction project risk management using Natural Language Processing techniques. *Auto. in Construction*, 80:66–76.