A Semantic Approach for Automated Rule Compliance Checking in Construction Industry

DONGMING GUO [iD], ERLING ONSTEIN, AND ANGELA DANIELA LA ROSA [iD]
Department of Manufacturing and Civil Engineering, Norwegian University of Science and Technology, 2815 Gjøvik, Norway

Corresponding author: Dongming Guo (dongming.guo@ntnu.no)

ABSTRACT Automated Compliance Checking (ACC) of building/construction projects is one of the important applications in Architecture, Engineering and Construction (AEC) industry, because it provides the checking processes and results of whether a building design complies with relevant laws, policies and regulations. Currently, Automated Compliance Checking still involves lots of manual operations, and massive time and cost consumption. Additionally, some sub-tasks of ACC have been researched, while few studies can automatically implement the whole ACC process. To solve related issues, we proposed a semantic approach to implement the whole ACC process in an automated way. Natural Language Processing (NLP) is used to extract rule terms and logic relationships among these terms from text regulatory documents. Rule terms are mapped to keywords (concepts or properties) in BIM data through term matching and semantic similarity analysis. After that, according to the mapped keywords in BIM and logic relationships among keywords, a corresponding SPARQL query is automatically generated. The query results can be non-compliance or compliance with rules based on the generated SPARQL query and requirements of stakeholders. The cases study proves that the proposed approach can provide a flexible and effective rule checking for BIM data. In addition, based on the proposed approach, we also further develop a semantic framework to implement automated rule compliance checking in construction industry.

INDEX TERMS Automated compliance checking, data extraction, ifcOWL, natural language processing, SPARQL generation.

I. INTRODUCTION

Automated Compliance Checking (ACC) of building/construction projects is necessary in the Architecture, Engineering and Construction (AEC) industry to provide stakeholders with high-quality building design models that should comply with certain authorized rules (such as international, national or local authorities' laws, policies, and regulations). ACC has been researched near 70 years, and has been one of the popular research themes in construction related research in recent years [1]. Meanwhile, ACC is becoming more automated with the development of commercial software, semantics and intelligent technologies [2].

ACC can be generally classified into four stages: (1) building model preparation; (2) rule interpretation; (3) rule execution; and (4) the checking results reporting [3]. The proposal of the Industry Foundation Classes (IFC), which is a conceptual model schema and a neutral file format

to facilitate building related information interoperability in AEC industry, promoted some research for building rule compliance checking on this model schema [4]. IFC documents containing building models were generally viewed as inputs in building model preparation phase [4], [5]. Additionally, for some semantic ACC applications, IFC documents were generally converted into Resource Description Framework (RDF) standard model in the building model preparation phase [6], [7].

In rule interpretation stage, the representation of building rules in computer understandable or executable format is one of the important challenges [8]. In this processing, experts knowledge and interventions are usually required because rules/regulations are written and designed by people, and then are generally read and explained by people. The common practice is that building rule experts interpret the regulations and software engineers program the regulations into executable application codes. As a result, it is a time-consuming and error prone process. To solve these problems, some commercial software [9], [10] and Natural Language

The associate editor coordinating the review of this manuscript and approving it for publication was Saqib Saeed [iD].

Processing (NLP) [11]–[16] were used to analyze and explain the regulations.

In rule execution stage, the process relies on the executable codes of rules and implementation tools. Different rule presentation methods and building model formats generally require different rules implementation tools or software. For example, semantic building models and corresponding semantic representing forms of rules commonly adopt a reasoning engine for checking implementation, while building models of IFC schema can be checked in existing commercial software, such as Solibri Model Checker. Furthermore, rule check reporting includes "compliance results," "non-compliance results," "pass," "fail," or awarded points/scores, etc. For example, some evaluation scores can be gained as checking results in the Green Building Rating System (GBRSs), such as Leadership in Energy and Environmental Design (LEED).

ACC is a comprehensive and complicated engineering challenge and every stage (building model preparation, rule interpretation, rule execution, and checking reporting) may be an independent research hotspot. Although some automated approaches are developed, massive manual operations are still required. Additionally, subtasks of ACC received more attentions, while few studies have been conducted on the whole ACC process aiming. To reduce the workload of manual processing and automatically implement the whole ACC process, we develop a semantic approach to achieve the whole ACC process in an automated way in this paper. In this approach, multiple intelligent technologies are selected and combined to reduce manual operations or interventions, mainly including multiple information extraction technologies/methods. Natural Language Processing (NLP) is used to extract terms of rules from text regulatory documents. The term matching and semantic similarity analysis are used to map rule terms to concepts/properties/instances of BIM data that is enriched with BimSPARQL. An automatically generated SPARQL query is used to extract check results from BIM data. The query results can be non-compliance or compliance with rules based on the SPARQL query and requirements of stakeholders. Furthermore, based on the proposed approach, we further develop a semantic framework for ACC, which can be widely used in industry ACC applications. In the section 2, the main ACC approaches and related technologies are introduced. Our proposed approach is described in section 3, and the approach implementation and cases study are shown in section 4. Section 5 shows some discussions about our approach, meanwhile, we also further propose a semantic ACC framework based on our approach. The last section is a brief conclusion.

II. THE RELATED WORK
A. COMMERCIAL SOFTWARE/CHECKING SYSTEMS
The one of the popular rule checking approaches is to use existing commercial software or checking systems because of the convenience and ease of use. Some well-known systems or commercial software for ACC include Solibri Model Checker, SMARTreview, CORENET System, Jotne Express Data Manager (EDM) Model Checker, etc. The method of using existing software for ACC is generally called the hard-coded method, and some rules/regulations have been developed and contained in software or systems. Sacks *et al.* illustrated the basic processing flows for BIM model checking using commercial software [9]. Additionally, some rules can be manually interpreted and used as input in some rule checking software based on interfaces provided by software [10]. However, the correctness of the implementation is unverifiable in existing software because checking routines mainly are black boxes and not transparent for users [17].

B. OBJECT-BASED APPROACHES
Object-based (also called as object-oriented) approaches focus on building objects and rule objects. In the object-based approaches, machine learning is gradually being adopted. Some research showed machine learning had better performances than rule inferencing approaches in some specific applications [18]. Sacks *et al.* discussed the important positive effects of machine learning in ACC based on BIM models and pointed out some possible positions, where machine learning methods can be used during the whole checking processing, such as semantic enrichment, rule interpretation, and rule checking implementation [9]. Additionally, it is possible to use only a machine learning algorithm to implement the entire rule checking process [18]. However, the collection of training dataset for machine learning algorithms and the selection of data features among multitudinous geometric and non-geometric BIM information are tough tasks. For checking different rules, different data features of buildings may need to be defined or identified, because relevant and effective data features are an important guarantee of machine learning algorithms to obtain high accuracy results.

C. LOGICAL APPROACHES
In logical approaches, a general framework is based on two main parts: the conversion from textual building rules to computable format and a detailed BIM model [19]. The regulations with human language statements are interpreted and translated into a logical language, such as first-order predicate logic, while the semantic enrichment is generally used to enrich BIM model details. For interpretation of regulation documents, some research methods were proposed. For example, with the Requirement, Applicability, Select and Exceptions (RASE) marks, all elements of a rule were divided into these four basic components and marked within the regulations text [20]. Through this method, rules were analyzed and captured from regulatory sentences for model checking. Another approach, Building Environment Rule and Analysis (BERA) Language, was proposed to provide a descriptive categorization and an algorithm to implement rule checking [21]. Ilhan *et al.* proposed a framework to facilitate the green code generation for Building Research Establishment Environmental Assessment Methodology (BREEAM)

certification [22]. Through Structured Query Language (SQL), a BIM Rule Language (BIMRL) was developed to query and manage the rule data for supporting rule compliance checking [23]. Fan *et al.* designed a two-layered logic-based rule evaluation interface, in which users manipulated rules freely and utilized the defined rules for building model checking. The interface can identify topologically related design dependencies with BIM elements involved [24]. Additionally, mvdXML technology is a rule checking approach based on logical rules. Fahad *et al.* compared the pros and cons of mvdXML and Semantic Web Rule Language (SWRL) technologies for rule compliance checking [25].

D. SEMANTIC/ONTOLOGICAL APPROACHES

Ontological modeling can easily represent building basic concepts and their relationships through classes and properties of ontology. Semantic Web and ontology technologies have been widely applied in ACC [1]. Some sub-topics of ACC have been solved, such as ontology-based regulations modeling [26], [27], semantic information extraction [27]–[29], semantic mapping [30], [31], and compliance checking implementation [4], [31].

Zhong *et al.* explored an ontology-based semantic modeling approach for construction quality inspection and evaluation [26]. Similarly, Lu, *et al.* modeled the construction safety rule through a new ontology method. They also extracted information from construction safety regulations and manually represented constrain rules with SWRL [27]. For semantic information extraction, De Farias *et al.* proposed a flexible extraction method, called IfcView, to extract building views through the ifcOWL ontology and SWRL rules [28]. Lee *et al.* proposed an ontology-based framework to accurately recognize precast concrete domain knowledge and data exchange requirements [30].

In terms of the compliance checking implementation, a reasoning engine and a rule language operated on this engine are generally required. The common rule languages include SWRL, SPARQL Protocol and RDF Query Language (SPARQL), Semantic Query-Enhanced Web Rule Language (SQWRL), the N3Logic language, etc. For example, Beach *et al.* selected SWRL to describe rules [31]. Pauwels, *et al.* used N3-logic as the rule language to present related rules [4]. Nevertheless, these methods still need experts or software engineers to interpret rules and then to transform them into rule languages [4].

Semantic data supplement sometimes is necessary in ACC, when BIM data is not sufficient or not explicit for rule compliance checking in a semantic context. The implementation of supplementing extra data or inferring new knowledge from existing BIM data is called semantic enrichment, which generally requires experts' knowledge and interventions. BimSPARQL is an effective approach to enrich BIM data through extending functions and data properties [32]. For supplementing additional data in ACC applications, Xu *et al.* proposed a semantic approach to integrate heterogeneous data

for ACC of underground utilities [33]. Semantic web and ontology technologies are recently receiving more attentions in the ACC domain, however, executable codes for ACC (such as SWRL, SPARQL) are generally programmed manually based on rules requirements.

E. NATURAL LANGUAGE PROCESSING (NLP) IN ACC

To automatically recognize rules from regulatory documents and reduce the manual process, NLP techniques can be used to facilitate automated rules extraction from regulatory documents [13]–[16]. Zhang *et al.* proposed a semantic NLP-based approach to extract information from regulatory documents, including information representation, text pre-processing, part-of-speech tagging and semantics generation, target information analysis, information extraction rules generation, extraction execution and evaluation, in which information extraction and conflict resolution rules are the core and key technologies [13]. Based on this information extraction approach, authors also tried to transform the regulations from regulatory documents into logic clauses [15]. Additionally, because concepts in regulatory documents may be different than semantically equal concepts in IFC schema, Zhang *et al.* utilized NLP to extract concepts and used a machine learning approach to predict the relationships between the extracted regulatory concepts and the related IFC concepts [16]. Zhou *et al.* utilized NLP and multiple extraction techniques (ontology-based pattern-matching extraction, sequential dependency-based extraction, and cascaded extraction methods) to extract information from complex energy conservation codes that are long, hierarchically complex, and with exceptions [12].

III. A SEMANTIC APPROACH FOR ACC

According to current ACC research, manual processes in ACC aren't avoided and they are mainly connected to building model preparation, rule interpretation and rule checking execution. Additionally, few studies have been conducted on the whole ACC process. In order to reduce the workload of manual processing and automatically implement the whole ACC process, we integrate and improve multiple approaches to develop a new semantic approach for ACC. Our approach includes NLP technologies for semi-automated regulation analyses and information extraction from textual regulatory documents, semantic extended BIM, automated linking extracted regulatory information with BIM data and an automated SPARQL query generation. There are three main modules in our approach, which is shown in Figure 1. NLP is used to extract terms and logic relationships from regulatory documents. Through semantic similarity analyses and term matching, rule terms are mapped to concepts/properties in a BIM knowledge base that is enriched with functions extension and shortcuts of data properties. Utilizing these mapped terms and logic relationships, a corresponding SPARQL query will be automatically generated. According to the generated SPARQL query and stakeholders' requirements, the
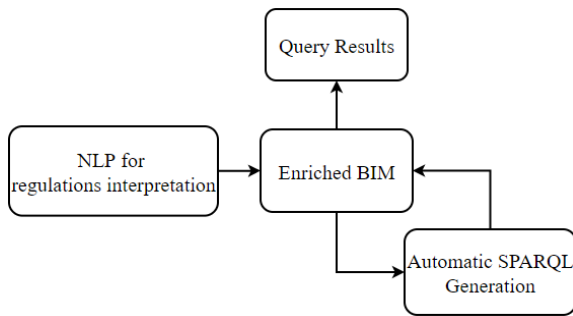
FIGURE 1. A semantic approach for automated rule compliance checking.

compliance or non-compliance query results will be extracted from the BIM knowledge base as results of ACC.

A. NLP FOR REGULATIONS INTERPRETATION

Multiple NLP approaches had been proposed for analyzing and extracting rule information from regulatory document [13], [15], [16]. We modify these approaches and partly utilize their research results for information extraction in our approach. The modified rule information extraction methodology is illustrated in Figure 2.
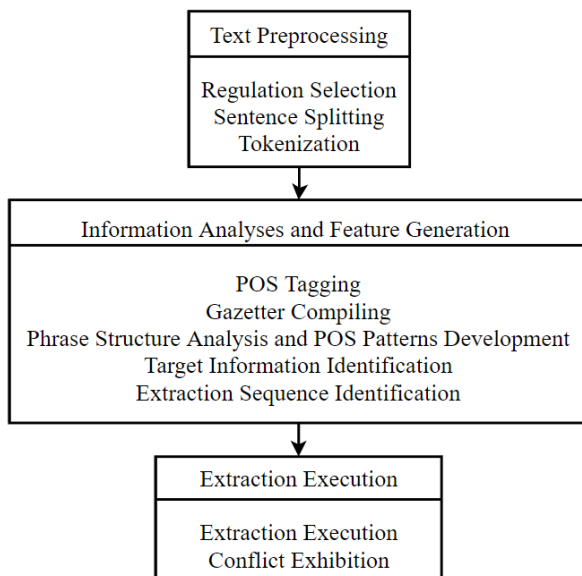
FIGURE 2. Modified rule information extraction methodology based on ref. [13], [15], [16].

1) TEXT PREPROCESSING

There are two main types of requirements in regulation documents: "quantitative requirement" and "existential requirement" [13]. The "quantitative requirement" mainly defines an attribute of a certain building element/part should comply with a requirement of a specific quantity value. For example, "Door openings within a dwelling unit or sleeping unit shall have a minimum clear opening height of 1981 mm." The "existential requirement" mainly requires the existence of a certain building element/part. For example, "The unit

(efficiency dwelling units) shall be provided with a separate bathroom containing a water closet, lavatory and bathtub or shower." Text classification for supporting ACC in construction was analyzed and implemented in ref [14], [34], which is beyond our research scope in this paper. We choose the quantitative requirements as research targets, because quantitative requirements are the most common and are generally more complex than existential requirements for ACC [13], [15]. Existential requirements may be converted into a specified kind of quantitative requirements.

Sentence splitting is to split the entire text into sentences, while tokenization is to tokenize the text as tokens, and a token can be a word, number, punctuation mark and a symbol (e.g., "#," "&") etc. The tokenization is a basic for the next further processing (e.g. POS tagging).

2) INFORMATION ANALYSES AND FEATURE GENERATION

This phase aims to implement information analyses and feature generation, and then create information patterns. Later, these patterns are used to extract information from text regulations.

a: PART-OF-SPEECH (POS) TAGGING

POS tagging is to tag each word in a sentence with POS of the word. POS tags are used to indicate lexical and functional categories of words. The POS of the word generally includes CC (Coordinating Conjunction), DT (Determiner), CD (Cardinal Digit), IN (Preposition), JJ (Adjective), JJR (Adjective, Comparative), NN (Noun, Singular), RB (Adverb), MD (Modal), and so on.

b: GAZETTEER COMPILING

Gazetteer compiling is to build sets of lists containing names of specific entities/concepts (e.g., cities, unit of measurement). In ACC, three Gazetteer lists are generally used: Comparative Relation List (e.g. less than, minimum, etc.), Negation List (e.g. no, not) and Unit List (e.g. square feet, meter) [13]. We can use these Gazetteer lists to identify parts of text content. There are some existing Gazetteer lists, such as NLTK NER tag, ANNIE (A Nearly-New information Extraction System) and Gazetteer of the GATE (General Architecture of Text Engineering). One can also create personal Gazetteer. We create own Gazetteer Unit List, Comparative Relation List and Negation List in our approach to implement the identification /conversion of units, comparative and negation relations.

c: PHRASE STRUCTURE ANALYSIS AND POS PATTERNS DEVELOPMENT

Based on POS tagging, some phrase structures can be identified, and then some POS patterns can be developed. Different sentences sometimes have similar phrase structures, such as NP (noun phrase, e.g. JJ+NN), VP (verb phrase, e.g. MD+VB), PP (prepositional phrase, e.g. IN+JJ+NN), etc. For obtaining phrase structures, an iterative, empirical process is required, including initial patterns construction,

NN	VBN NN	VBN NN NN
NNP	VBN NNS	VBN NN NNS
NNS	JJ JJ NN	JJ JJ JJ NN
VBG	JJ JJ NNS	JJ JJ NN NN
JJ NN	JJ NN NN	JJ NN NN NN
JJS NN	JJ NN NNS	NN NN NN NNS
JJ NNS	NN NN NN	NNP NN NN NN
NN NN	NN NN NNS	NNP NN NN NNS
NN NNS	NNP NN NN	NNP NNP NNP NNP
NNP NNP	NNP NNP NNP	RB VBN JJ NN
NNP NNS	NNP VBD NNS	RB VBN NN NN
VBG NN	NN VBG NN	RB VBN NNP NN
VBG NNS	RB JJ NNS	JJ VBN JJ JJ VBG NN

FIGURE 3. A developed NP set of POS patterns in ref. [16].

test and results analyses, and iterative phrase structure improvement [13]. Additionally, Zhang *et al.* developed a set of noun phrases in ACC, shown in Figure 3 [16]. We will use the NP set in our approach. Except NP phrase patterns, we will develop phrase patterns through a phrase generation approach in ref. [13], including quantitative and unit of measurement phrase, verb phrase, prepositional phrase, and so on. These developed POS patterns and Gazetteer lists compose phrase patterns for information extraction.

d: TARGET INFORMATION IDENTIFICATION

For representing the extracted information from text regulations, Zhang *et al.* proposed some extracted requirements of information identification: (1) the subject, comparative relation, quantity value, and quantity unit/reference respectively have one and only one instance of information elements; (2) the compliance checking attribute, deontic operator indicator, and quantitative relation respectively have at most one instance of information elements; and (3) the subject restriction and quantity restriction respectively have zero, one, or more instances of information elements [13]. We use the same extracted requirements and use an eight-tuple format in our approach. An example of the eight-tuple format is shown in Table 1.

e: EXTRACTION SEQUENCE IDENTIFICATION

The sequence of extracting information elements can affect the efficiency of extracting information, because the easier identifiable information/elements can be firstly extracted to reduce the obstruction of extracting the more difficultly identifiable information/elements. So, the extraction sequence of semantic information elements in our approach is "quantity value and unit of measure > subject > compliance checking attribute > comparative relation > relation operator indicator > subject restriction and quantity restriction," which is similar with ref [13].

3) EXTRACTION EXECUTION

After finishing feature generation and target identification, we use a bottom-up method to implement information extraction, which starts with split independent words, and then combines words into different phrase structures. However, some problems may be produced during the

TABLE 1. An information format in our approach.

A Regulatory Sentence: Door openings within a dwelling unit or sleeping unit shall have a minimum clear opening height of 1981 mm.		
Semantic Information Element	Extracts of this Sentence	Feasible POS Patterns
Subject	Door openings	NP
Subject restriction	a dwelling unit or sleeping unit	NP, PP
Compliance checking attribute	clear opening height	NP
Relation operation indicator	shall have	VP
Comparative relation	minimum/not less than	Comparative Relation or Negation List
Quantity value	1981	CD
Quantity unit/reference	mm	Unit List
Quantity restriction	N/A	NP, PP

extraction implementation. For example, the developed POS patterns and requirements of information identification are sometimes not suitable for a part of regulations. Regulations in regulatory documents can be described in various flexible ways by a natural language, while the developed POS patterns are limited and cannot cover all varieties of regulations descriptions. When extraction conflicting is identified, expert knowledge and interventions are necessary. In our approach, we choose to show all conflicting problems, instead of further processing them through setting conflict solution rules.

B. SEMANTIC ENRICHED BIM

BIM data that contains comprehensive building information is generally described by IFC schema to facilitate data exchange and sharing. However official IFC schema and BIM cases based on IFC schema only provide some basic properties and relationships of building data, while existing properties, relationships, functions in IFC schema are not sufficient for building ACC. Additionally, the concepts and properties in ifcOWL ontology, which was developed based on IFC schema, are also not sufficient for building ACC. Zhang *et al.* developed SPARQL extensions based on ifcOWL ontology, called BimSPARQL [32]. In BimSPARQL, there are six developed RDF prefixes, which define shortcuts of properties and relationships for IFC schema level semantics (in schm prefix file), instance level property and quantity sets (in pset and qto prefix files), and new properties and relationships based on geometry data (in pdt, spt, geom prefix files) [32]. We adopt BimSPARQL to enrich a BIM case through merging BimSPARQL six prefix RDF files with the BIM case.

C. AUTOMATED SPARQL GENERATION

Through information extraction from a regulatory document, we obtain some terms and logic relationships among

terms, an example shown in Table 1. However, concepts and properties in the enriched BIM RDF data can be different than the extracted terms from regulatory documents. In this paper, we map the extracted terms of regulations to enriched ifcOWL ontology through term matching and semantic similarity analysis. The term matching is based on string-based matching and semantic similarity analysis is to make a thorough analysis at a semantic level through an external linguistic database as background knowledge. The mapping is to identify which concepts/properties in the enriched BIM RDF data can correspond with the extracted terms in subject, subject restriction and compliance checking attribute of the eight-tuple format. Additionally, the terms of regulatory documents may have different forms, such as plural form, past tense, present participle and so on, so the morphological analysis will be used firstly before the terms matching and semantic similarity analysis. After mapping regulatory terms to the enriched BIM data, we can obtain some mapped keywords (concepts or properties) in BIM RDF data and use these keywords to automatically generate a SPARQL query based on ifcOWL ontology structure.

We introduced an automatic SPARQL generation approach in ref [29], in which a SPARQL query can be generated according to given query keywords. Utilized these query keywords, a shortest path connected all query keywords will be found in a BIM knowledge base, and then the structure of the shortest path will be used as a query sub-graph to generate a SPARQL query.

However, the SPARQL generation approach in ref [29] did not consider filter(s) of quantitative conditions. We extend our SPARQL generation approach to implement the information extraction through SPARQL query with a/some filter(s) of quantitative conditions. There are built-in filter functions in SPARQL, such as logical judgements (!, &&), math calculations ($+$, $-$) and comparison ($<$, $=$) etc., so the extracted information that is related with the filter of quantitative conditions (e.g. the logic relationships among compliance checking attribute, comparative relation and quantity value) are used to generate FILTER codes in a SPARQL query. The corresponding relationships between SPARQL sections and the eight-tuple format is shown in Figure 4. All extracted information elements in the eight-tuple format will form query keywords for SPARQL generation, except "relation operation indicator." The subject in the eight-tuple format will be used to create the SELECT section in SPARQL. When a subject term in the eight-tuple format is mapped to multiple keywords in BIM data, an extra FILTER sector will be generated to solve this problem (shown in a following case study). Meanwhile, stakeholders can add related variables that are listed in eight-tuple format in SELECT section to show required information in query results.

At last, query results can be obtained through executing the SPARQL query, which are generally compliance query results. When non-compliance query results will be required, stakeholders only invert the logic relationship in the FILTER sector, such as changing the ">" into "<=."
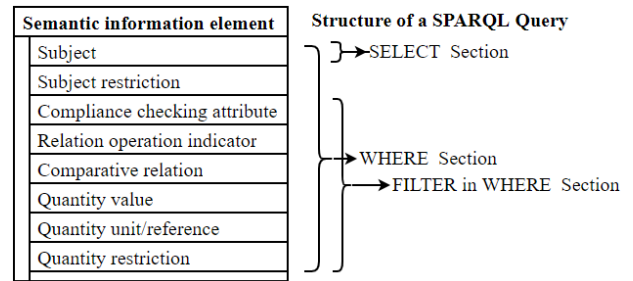

FIGURE 4. The corresponding relationships between the eight-tuple format and SPARQL structure.

IV. IMPLEMENTATION

To implement our approach and validate its effectiveness, we use some existing NLP libraries and some tools/software:

- IFCtoRDF tool (https://github.com/pipauwel/IFCto-RDF)
- BimSPARQL https://github.com/BenzclyZhang/Bim-SPARQL
- Protégé Desktop v.5.5.0.
- Stardog triple store and API version 7
- Natural Language Toolkit (NLTK) package
- PyCharm Edu 2020.1.1 x64

We use a public BIM case to test our approach, which is "Duplex_A_20110505," shown in Figure 5. The IFCtoRDF tool is used to convert BIM IFC data into RDF format, and Protégé Desktop v.5.5.0 is used to merge BimSPARQL RDF data into the BIM RDF data of "Duplex_A_20110505." After that, the enriched BIM data is stored in Stardog RDF database. Additionally, we develop python programs to implement our approach in PyCharm Edu 2020 and use some functions in NLTK package for extracting information from regulatory documents. At last, we utilize International Building Code (IBC) 2018 as a regulatory document to evaluate our approach.
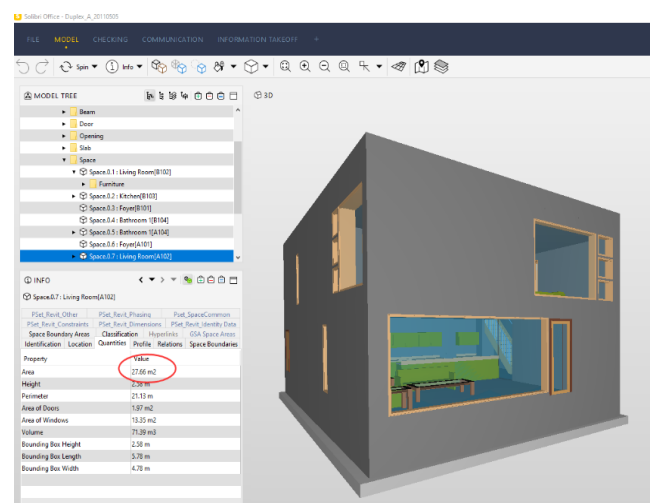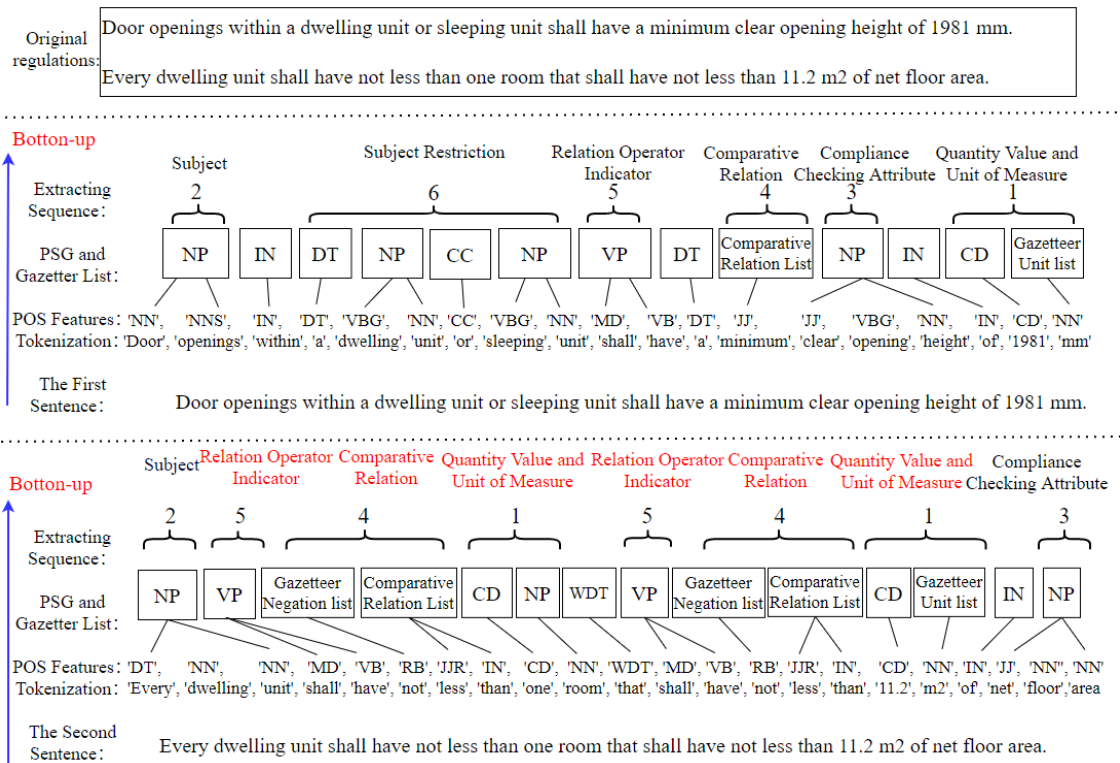

FIGURE 5. A duplex apartment BIM project.

FIGURE 6. Results of NLP execution in our approach.

A. NLP IMPLEMENTATION

For NLP, some existing functions in NLTK package are applied in our approach, such as "*sent_tokenize*" for sentences splitting, "*word_tokenize*" for words tokenization, "*nltk.pos_tag*" for tagging the words, "*stem.porter.PorterStemmer()*" and "*wordnet_lemmatizer.lemmatize()*" for morphological processing and lemmatization, etc. Because the quantity value and unit of measure in IBC 2018 have two units (metric unit and English unit) and the Duplex_A_20110505 BIM case uses the metric unit, we use the metric unit and quantity value in the following processing. After tagged all words with POS, POS patterns are used to identify the information elements and extract information.

Some feasible POS patterns are listed in the Table 1 to correspond to different information elements. The establishment of POS patterns can refer in ref. [13], [16]. We implement the NLP in python and the processing results are shown in Figure 6, in which we illustrate the processing of two regulatory sentences as cases. Because two comparative relations, quantity values, and quantity units are detected in information extraction of the second sentence, and they have conflicts with requirements of information identification, expert knowledge and interventions will be required to process conflicting problems. All conflicts are marked as red color in Figure 6 and the results of extracted information are listed in Table 2.

TABLE 2. The results of extracted information from rules.

Information element	The First Rule	The Second Rule
Subject	door openings	one room
Subject restriction	a dwelling unit or sleeping unit	every dwelling unit
Compliance checking attribute	clear opening height	net floor area
Relation operation indicator	shall have	shall have
Comparative relation	minimum	not less than
Quantity value	1.981	11.2
Quantity unit/reference	m	m^2

B. IMPLEMENTATION OF ENRICHED BIM

We use Protégé to implement the ontology merging of BimSPARQL and the BIM case. Because BimSPARQL was developed based on ifcOWL ontology and BIM RDF data is generally also based on ifcOWL ontology, merging them is easy. When opening six RDF files of BimSPARQL and the BIM RDF data in Protégé, we utilize the function of "Refactor -> Merge ontologies" in Protégé to merge RDF data into a new RDF file, and then save and load the new RDF file (enriched BIM data) in the Stardog database. The merged results of data properties are shown in Figure 7.
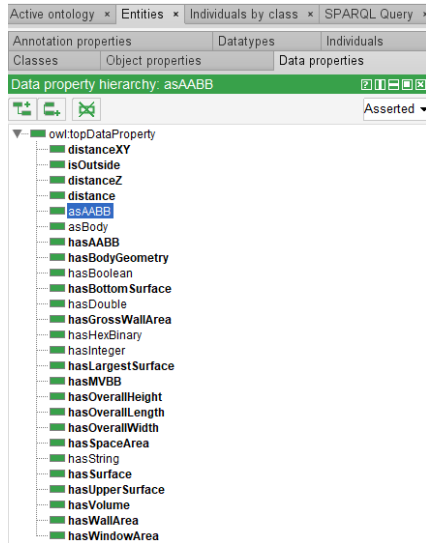
FIGURE 7. Data properties of the enriched BIM in Protégé.

FIGURE 8. The results of semantic similarity analyses among some keywords.

C. TERMS MATCHING

After enriching BIM data with BimSPARQL, we map the extracted terms of regulations to the enriched BIM data. For the term matching and semantic similarity analysis, we firstly establish a term dictionary from ABox (descripts the statements about conceptual entities and properties) and TBox (descripts the statements about concrete entities) in the enriched BIM RDF data that is stored in Stardog. The term dictionary consists of an index item and an index keyword item. For example, an ifcOWL data type "*IfcPositiveLengthMeasure*" will be separated into "Ifc Positive Length Measure" based on capital letters, and then "Ifc" is removed. The term "positive length measure" as an index and the "*IfcPositiveLengthMeasure*" as an index keywords are combined and stored in a term dictionary for term matching and semantic similarity analysis. When "positive length measure" is mapped through term matching, the corresponding term keyword "*IfcPositiveLengthMeasure*" is selected for the next processing. Because all extracted terms of regulations are processed through morphological processing and lemmatization, all index words in the term dictionary are lowercase and base form, while index keywords will retain their original forms in BIM RDF data, such as "live room" as index and "Living Room" as index keywords in term dictionary.

After establishing the term dictionary, every extracted regulatory term in the eight-tuple format is indexed in the term dictionary to find the matched index keywords based on string matching. Meanwhile, semantic similarity analysis will be adopted. WordNet in NTLK provides six semantic similarity analyses, and we use "wup_similarity" (Wu-Palmer Similarity) for semantic similarity analysis to denote the degree of similarity of two words senses. The function returns a score that is in the range 0 to 1. A score of 1 means two words semantics are identical and the higher score proves the more similar meanings between two words.

Figure 8 shows max similarity scores and corresponding synonyms between "room" and 8 different string labels of space instances in the BIM case. Because the semantic similarity scores among "room" and "kitchen," "bathroom," "bedroom," "living_room" or "foyer" are near 1, the keyword "room" is mapped to "kitchen," "bathroom," "bedroom," "living_room" or "foyer" in the BIM case. However, when the extracted terms cannot be mapped to an index in the term dictionary of BIM data, manual operations are necessary to explain this term/phrase and map this term to a keyword(s) in the term dictionary. Finally, the mapped keywords are listed in Table 3 in our approach. Because the BIM case used is a dwelling unit, the requirement of subject restriction is deemed to be satisfied and isn't required to be mapped to BIM data.

D. SPARQL QUERY GENERATION

After the extracted terms are mapped to the BIM term dictionary, these mapped keywords and logic relationships shown in Table 3 can be used to generate SPARQL query automatically. Here, we implement a SPARQL query generation with a/some filter(s) of quantitative conditions.

TABLE 3. The mapped keywords in our approach.

Information element	Mapped Keywords of the First Rule	Mapped Keywords of the Second Rule
Subject	door	Kitchen, Bathroom, Bedroom, Living Room, OR Foyer
Subject restriction		
Compliance checking attribute	OverallHeight_IfcDoor	hasSpaceArea
Relation operation indicator		
Comparative relation	>=	>=
Quantity value	1.981	11.2
Quantity unit/reference		
Quantity restriction		

To ensure to generate exact SPARQL query results, a generated SPARQL query with a quantitative filter about a property value of an instance should contain a class, an instance of this class, a property of this instance of this class and a

data value of this property. However, there are some different graph structures in the ifcOWL ontology about a class, an instance of this class, a property of this instance of this class and a data value of this property. We illustrate some different graph structures that contain these four items in the ifcOWL ontology in Figure 9, in which these four items are marked as bold font. These different graph structures bring some difficulties in automatic SPARQL query generation.
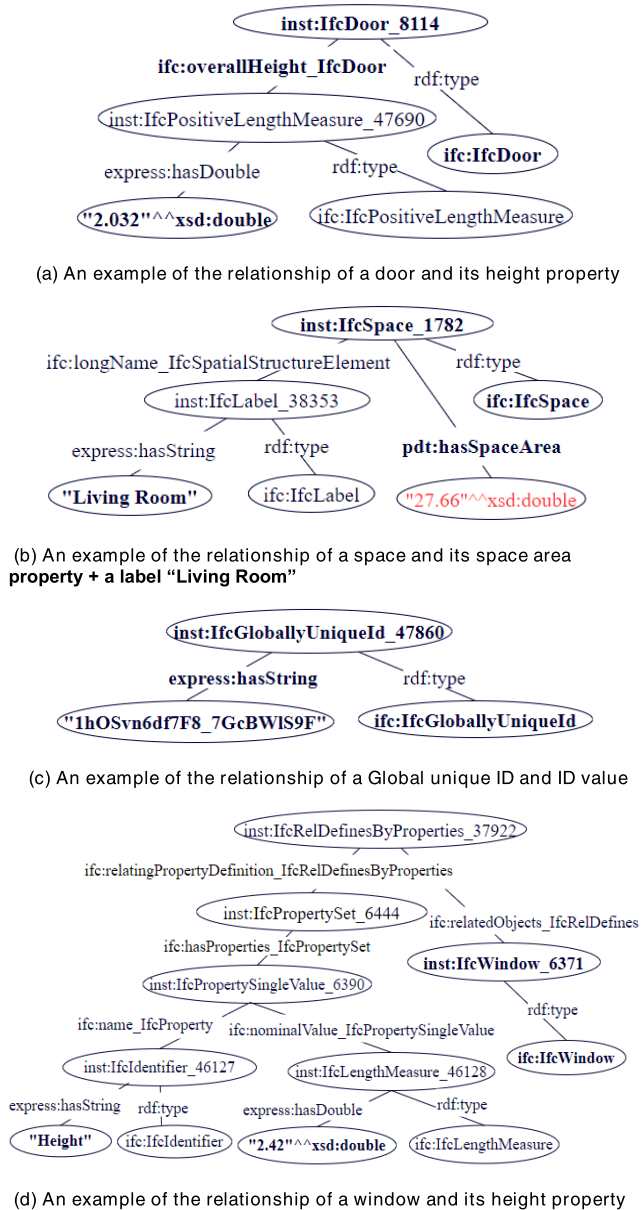

(a) An example of the relationship of a door and its height property

(b) An example of the relationship of a space and its space area property + a label "Living Room"

(c) An example of the relationship of a Global unique ID and ID value

(d) An example of the relationship of a window and its height property

FIGURE 9. Some sub-graph examples contained an instance, a class and a single value property of this instance.

The value data of space area (red color) in the graph structure of Figure 9(b) comes from a property defined in BimSPARQL. In all graph structures, two common characteristics can be found: (1) a quantity value of a property is always at the end point in a graph structure; (2) a property

concept/name has the shortest linked distance with its' quantity value, compared with distances among this property concept/name and all other quantity values. Based on these two characteristics, we can firstly explore the shortest linked structure between a specified property name and its data value, and then keep this shortest linked structure in the exploring the shortest path/structure that connects all mapped keywords in the BIM RDF data. We discussed an approach for generating the shortest path in ref [29]. The structure of shortest path can be used to create a SPARQL query. After that, SPARQL variables can replace instances and data value of the shortest path. The SPARQL variable of "compliance checking attribute" (e.g. ?c) and comparative relation (e.g. ">=") and quantity value (e.g. "1.981") in the eight-tuple format can be used to produce the *FILTER* codes in SPARQL. For example, Figure 9(a) shows a graph structure that connects all mapped keywords of the first regulation sentence, so the SPARQL query based on this structure can be generated and shown in Figure 10, in which the *FILTER* section is "*FILTER (?c >=1.981).*"

FIGURE 10. The generated SPARQL query based on the first regulatory sentence.

However, "space area" is not a property of a space instance in IFC schema, although an area property of a space is displayed in Solibri software, shown in Figure 5 (in a red circle). So, we obtain the data value of a space area through a BimSPARQL defined property that is 'pdt:hasSpaceArea'. Additionally, because multiple keywords are mapped to a subject term in the second regulation sentence, an extra *FILTER* code is required to describe the logic relationship among multiple subject keywords, which is "*FILTER (CONTAINS (?c, "Bathroom") ||CONTAINS(?c, "Kitchen") || CONTAINS(?c, "Living Room") || CONTAINS(?c, "Foyer") ||CONTAINS(?c, "Bedroom")).*" The other *FILTER* code explains a logic relationship between compliance checking attribute, data value and comparative relation that is "*FILTER (?d >=11.2),*" and the whole SPARQL query is show in Figure 11, based on the graph structure of Figure 9(b).

The compliance results based on these two cases can be generated through executing SPARQL queries shown in Figure 10 or Figure 11, while stakeholders can flexibly add variables of required information into "*SELECT*" section to obtain required information. For example, we add some query variables in the *SELECT* section to facilitate the understanding of query results, illustrated in Figure 12. Additionally, when converting the logic relationship in the FILTER code,

```
SELECT ?a
WHERE
{
?a a ifc:IfcSpace.
?a ifc:longName_IfcSpatialStructureElement ?b.
?b express:hasString ?c.
FILTER (CONTAINS(?c, "Bathroom") ||CONTAINS(?c, "Kitchen")
|| CONTAINS(?c, "Living Room") || CONTAINS(?c, "Foyer")
|| CONTAINS(?c, "Bedroom")).
?a pdt:hasSpaceArea ?d.
FILTER (?d >=11.2)
}
```

FIGURE 11. The generated SPARQL query based on the second rule.

FIGURE 12. The compliance results of the second rule checking.

such as " *FILTER* (?*d* >=*11.2*)" to " *FILTER* (?*d* <*11.2*)," the non-compliance results can be generated.

V. DISCUSSION

The proposed approach can implement the whole process of automated rule compliance checking in construction industry, including semi-automated information extraction from regulatory documents, automated term matching, automated SPARQL query generation, to reduce manual operations as much as possible.

Through case studies, our approach is proved to be feasible and effective. In the first case, we can fully automatically implement the whole process of rule checking. In the second case, expert knowledge and manual operations

are required to identify regulatory information and select matched terms from BIM data when one-to-many mapping happens, or mapping fails, or extracting information from regulatory document fails or conflicts. According to some NLP research [13]–[16], information extraction from documents inevitably brings some error/conflict problems, even after further processing by conflict solution methods. One of main reasons for this is that regulations/rules can be flexibly described by a natural language, while comprehensive and detailed information extraction patterns and identification rules are not easily implemented. For example, a pattern of "JJ VGB NN" is not included in an established NP set that was achieved in ref. [16], shown in Figure 3. Machine learning seems to be able to solve this kind of issues [35], [36]. However, when one uses machine learning technologies for ACC, a special training dataset of machine learning can be required for a special regulation checking, such as in ref [18]. It is not feasible to create different training dataset and machine learning algorithm for each construction regulation. Whether a training dataset or an algorithm of machine learning can be used for many regulations checking (and which rules) needs further research. In our approach, we choose to develop POS patterns and information identification rules and show all extraction conflicts/errors, and then manual operations solve these conflicts/errors.

Additionally, we use WordNet as data dictionary for similarity analysis, rather than a specific domain dictionary, so the results of similarity analyses are sometimes unsatisfactory. Furthermore, we only use BimSPARQL to enrich BIM data. Only BIM data or BIM extended functions are not sufficient for some regulations checking, such as underground utilities checking [33], so a specific domain dictionary and an effective integrated approach are sometimes necessary for integrating and mapping multi-source heterogeneous data for ACC applications.

The automated SPARQL query generation can convert extracted regulations into executable checking codes and compliance results can be obtained through executing the SPARQL query in a reasoning engine. The "existential requirement," one of two main types of requirements in regulation documents, can translated into the "quantitative requirement," so we only discuss the SPARQL generation with "quantitative requirement" in this paper. IfcOWL schema uses different sub-graph structures to describe connected relationships among a class, an instance of this class, a property of this instance of this class and a data value of this property. We find two common characteristics in these sub-graph structures and utilize these two common characteristics to explore the shortest path that connects all query keywords. Later, the structure of the shortest path will be used to generate SPARQL query. We have discussed an approach for exploring the shortest path in ifcOWL schema in ref [29], so we don't introduce it in this paper. The automatic generated SPARQL query can automatically implement the data extraction from BIM data and produce the compliance checking results.
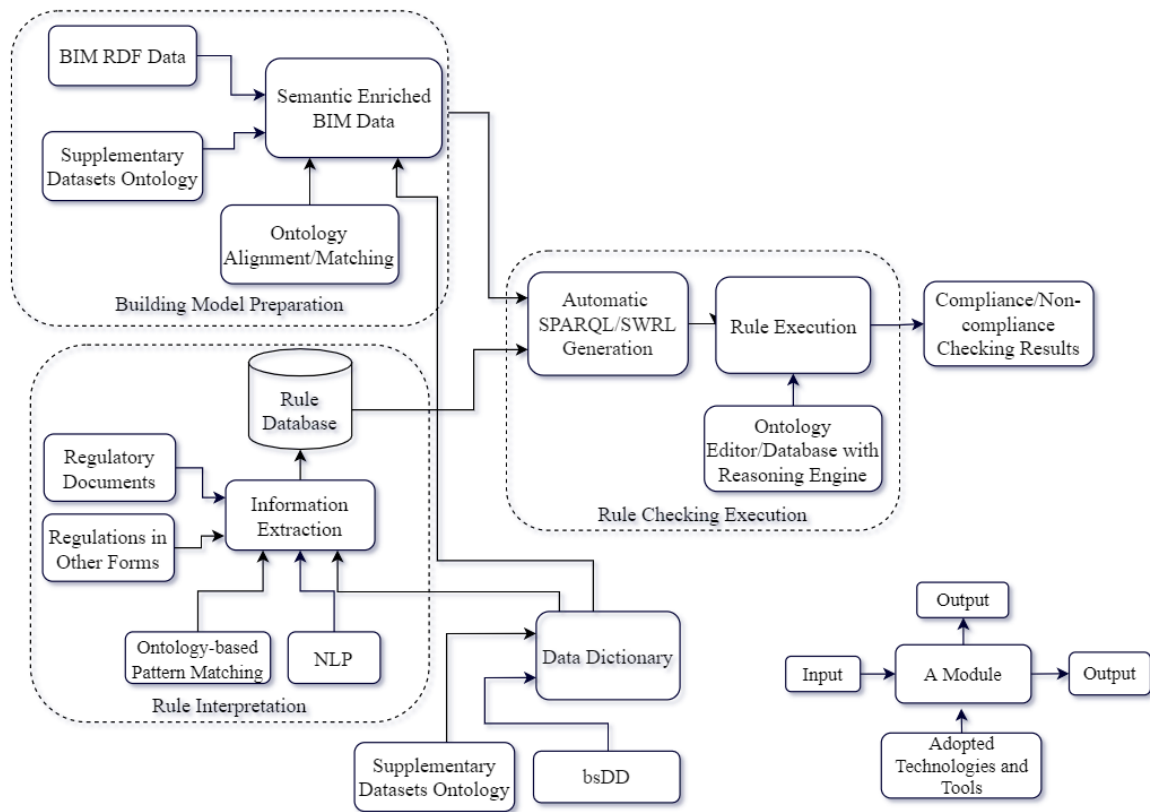
FIGURE 13. A new semantic framework for fully ACC in AEC industry.

We prove that the whole ACC process can be automatically implemented. However, we only achieve basic checking processing, and some problems remain to be solved on its way of further development, such as BIM semantic enrichment, conflict solving in NLP, etc. Based on our approach, we also further develop a semantic framework for ACC in AEC industry, shown in Figure 13. This framework can be used in wide ACC applications, which considers information supplementary, specific domain data dictionary, executable query codes generation and implement environments/tools, etc. This framework may better implement rule compliance checking, with the development of NLP, data mapping technologies and related intelligent technologies. Meanwhile, it can also solve some manual operations in our proposed approach and extend our approach in data supplementary and the form of executable query codes.

In this framework, main functions include:

(1) Data Dictionary, a specific domain dictionary, is to guarantee information consistency in regulatory documents, supplementary data and building data. It stores related terms in building construction domain such as classes, concepts, properties and instances, etc. The unified concepts/terms can ensure information consistency in extracting information from regulatory documents/other regulations sources and in supplementing extra data sources with BIM data to effectively implement to link extracted rule information with enriched BIM data.

(2) Semantic Enriched BIM Knowledge Base is to enrich BIM model with external supplementary data and functions. Supplementary data will sometimes be required and linked with BIM RDF data, because some rules may require not only building data, but also information about surrounding environment of buildings and building materials, etc. Establishing the enriched BIM knowledge base is to reduce repetitive manual operations because rule checking is an iterative processing. When elements of a building are changed, repeated manual linking supplementary data with building elements may occur. If the related supplementary data is stored and linked with a BIM knowledge base, the linking new building elements with supplementary data will be automatically executed and the workload of manual linking operations will be reduced.

Additionally, when some products or facilities knowledge bases already exist and are open in manufacturers' websites, utilizing the Web network protocols can retrieved the required information to reduce manual data input and acquirement. Simple Object Access Protocol (SOAP) can be used for this, just similar to ref [37] did.

(3) Rule interpretation is to explain, analyze and extract rules from regulatory documents or other regulation sources and can store extracted rules in a rule database with a logic

or certain format(s). Some rules extraction technologies can be NLP and other related intelligent technologies based on a domain data dictionary. It is also one of the important researches to realize ACC.

(4) Rule checking execution is to generate executable SPARQL or Semantic Web Rule Language (SWRL) query codes and implement executable SPARQL or SWRL in a certain reasoning engine to obtain query results. Some approaches for automatic SPARQL generation have been proposed based on ifcOWL ontology, such as ref [29] and an improved in this paper.

(5) Rule checking reporting mainly includes compliance or non-compliance checking results in this framework. Certainly, the compliance/non-compliance checking results can be easily converted into different formats, e.g. PASS, FAIL, or evaluation scores. For example, when non-compliance checking results can be found through SPARQL query, the checking report can be "FAIL."

Our proposed semantic approach can also prove the framework is feasible for ACC, because our approach is a part of this framework.

VI. CONCLUSION

In this paper, a semantic approach is proposed to implement the whole processing of automated rule compliance checking in construction industry. This approach contains extracting rule information from regulatory documents, semantic enrichment of BIM data, mapping rule keywords to BIM RDF data, automatically generating SPARQL query and compliance results. We try to reduce manual operations as much as possible in ACC and facilitate the automated rule compliance checking processing. The cases study shows our proposed approach is feasible and effective. Because conflict solving in NLP isn't achieved and only basic term mapping and semantic enrichment of BIM data are implemented in our approach, expert knowledge and manual operations are sometimes required. Based on our proposed approach and aiming to overcome shortages of our approach, a new semantic framework is also further developed, in which a specific domain data dictionary, the data supplementary and ontology mapping in rule checking processing are considered. The adopted technologies and tools for different modules are briefly introduced in this framework. This framework can better implement ACC in construction industry with the development of related technologies.

REFERENCES

[1] B. Zhong, H. Wu, H. Li, S. Sepasgozar, H. Luo, and L. He, "A scientometric analysis and critical review of construction related ontology research," *Autom. Construct.*, vol. 101, pp. 17–31, May 2019.

[2] A. S. Ismail, K. N. Ali, and N. A. Iahad, "A review on BIM-based automated code compliance checking system," in *Proc. Int. Conf. Res. Innov. Inf. Syst. (ICRIIS)*, Langkawi Island, Malaysia, Jul. 2017, pp. 1–6.

[3] C. Eastman, J.-M. Lee, Y.-S. Jeong, and J.-K. Lee, "Automatic rule-based checking of building designs," *Autom. Construct.*, vol. 18, no. 8, pp. 1011–1033, Dec. 2009.

[4] P. Pauwels, D. Van Deursen, R. Verstraeten, J. De Roo, R. De Meyer, R. Van de Walle, and J. Van Campenhout, "A semantic rule checking environment for building performance checking," *Autom. Construct.*, vol. 20, no. 5, pp. 506–518, Aug. 2011.

[5] S. Jiang, Z. Wu, B. Zhang, and H. Cha, "Combined MvdXML and semantic technologies for green construction code checking," *Appl. Sci.*, vol. 9, no. 7, p. 1463, Apr. 2019.

[6] M. Fahad and N. B. Bus Fies, "Semantic BIM reasoner for the verification of IFC models," in *Ework and Ebusiness in Architecture, Engineering and Construction*, J. Karlshoj and R. Scherer, Eds. Boca Raton, FL, USA: CRC Press, 2018, pp. 361–368.

[7] P. Pauwels and S. Zhang, "Semantic rule-checking for regulation compliance checking: An overview of strategies and approaches," in *Proc. 32rd Int. CIB W78 Conf.*, Eindhoven, The Netherlands, 2015, pp. 619–628.

[8] S. Macitlal and H. M. Günaydın, "Computer representation of building codes for automated compliance checking," *Autom. Construct.*, vol. 82, pp. 43–58, Oct. 2017.

[9] R. Sacks, "Automating design review with artificial intelligence and BIM: State of the art and research framework," in *Computing in Civil Engineering 2019: Visualization, Information Modeling, and Simulation*, Y. K. Cho, Ed. Reston, VA, USA: ASCE Press, 2019. 353-360.

[10] M. M. Hossain and S. Ahmed, "Developing an automated safety checking system using BIM: A case study in the Bangladeshi construction industry," *Int. J. Construct. Manage.*, vol. 4, pp. 1–19, Nov. 2019.

[11] J. Zhang and N. M. El-Gohary, "Integrating semantic NLP and logic reasoning into a unified system for fully-automated code checking," *Autom. Construct.*, vol. 73, pp. 45–57, Jan. 2017.

[12] P. Zhou and N. El-Gohary, "Ontology-based automated information extraction from building energy conservation codes," *Autom. Construct.*, vol. 74, pp. 103–117, Feb. 2017.

[13] J. Zhang and N. M. El-Gohary, "Semantic NLP-based information extraction from construction regulatory documents for automated compliance checking," *J. Comput. Civil Eng.*, vol. 30, no. 2, Mar. 2016, Art. no. 04015014.

[14] D. M. Salama and N. M. El-Gohary, "Semantic text classification for supporting automated compliance checking in construction," *J. Comput. Civil Eng.*, vol. 30, no. 1, Jan. 2016, Art. no. 04014106.

[15] J. Zhang and N. M. El-Gohary, "Automated information transformation for automated regulatory compliance checking in construction," *J. Comput. Civil Eng.*, vol. 29, no. 4, Jul. 2015, Art. no. B4015001.

[16] J. Zhang and N. M. El-Gohary, "Extending building information models semiautomatically using semantic natural language processing techniques," *J. Comput. Civil Eng.*, vol. 30, no. 5, Sep. 2016, Art. no. C4016004.

[17] C. Preidel and A. Borrmann, "Automated code compliance checking based on a visual language and building information modeling," in *Proc. 32nd ISARC*, Oulu, Finland, 2015, pp. 1–8.

[18] T. Bloch and R. Sacks, "Comparing machine learning and rule-based inferencing for semantic enrichment of BIM models," *Autom. Construct.*, vol. 91, pp. 256–272, Jul. 2018.

[19] N. O. Nawari, "Smartcodes and BIM," in *Proc. Conf. Struct. Congr.*, Pittsburgh, PA, USA, 2013, pp. 928–937.

[20] E. Hjelseth and N. N. Nisbet, "Capturing normative constraints by use of the semantic mark-up rase methodology," in *Proc. CIB*, Sophia Antipolis, France, 2011, pp. 1–10.

[21] J. K. Lee, "Building environment rule and analysis (BERA) language," Ph.D. dissertation, College Archit., Georgia Inst. Technol., Atlanta, GA, USA, 2011.

[22] B. Ilhan and H. Yaman, "Green building assessment tool (GBAT) for integrated BIM-based design decisions," *Autom. Construct.*, vol. 70, pp. 26–37, Oct. 2016.

[23] W. Solihin, "A simplified BIM data representation using a relational database schema for an efficient rule checking system and its associated rule checking language," Ph.D. dissertation, College Archit., Georgia Inst. Technol., Atlanta, GA, USA, 2016.

[24] S.-L. Fan, H.-L. Chi, and P.-Q. Pan, "Rule checking interface development between building information model and end user," *Autom. Construct.*, vol. 105, Sep. 2019, Art. no. 102842.

[25] M. Fahad and F. Andrieux, "Towards mapping certification rules over BIM," in *Proc. 33rd CIB W78 Conf.*, Brisbane, QLD, Australia, 2016, pp. 1–10.

[26] B. T. Zhong, L. Y. Ding, H. B. Luo, Y. Zhou, Y. Z. Hu, and H. M. Hu, "Ontology-based semantic modeling of regulation constraint for automated construction quality compliance checking," *Autom. Construct.*, vol. 28, pp. 58–70, Dec. 2012.

[27] Y. Lu, Q. Li, Z. Zhou, and Y. Deng, "Ontology-based knowledge modeling for automated construction safety checking," *Saf. Sci.*, vol. 79, pp. 11–18, Nov. 2015.

[28] T. M. D. Farias, A. Roxin, and C. Nicolle, "A rule-based methodology to extract building model views," *Autom. Construct.*, vol. 92, pp. 214–229, Aug. 2018.

[29] D. Guo, E. Onstein, and A. D. L. Rosa, "An approach of automatic SPARQL generation for BIM data extraction," *Appl. Sci.*, vol. 10, no. 24, p. 8794, Dec. 2020.

[30] Y.-C. Lee, C. M. Eastman, and W. Solihin, "An ontology-based approach for developing data exchange requirements and model views of building information modeling," *Adv. Eng. Informat.*, vol. 30, no. 3, pp. 354–367, Aug. 2016.

[31] T. H. Beach, Y. Rezgui, H. Li, and T. Kasim, "A rule-based semantic approach for automated regulatory compliance in the construction sector," *Expert Syst. Appl.*, vol. 42, no. 12, pp. 5219–5231, Jul. 2015.

[32] C. Zhang, J. Beetz, and B. de Vries, "BimSPARQL: Domain-specific functional SPARQL extensions for querying RDF building data," *Semantic Web*, vol. 9, no. 6, pp. 829–855, Sep. 2018.

[33] X. Xu and H. Cai, "Semantic approach to compliance checking of underground utilities," *Autom. Construct.*, vol. 109, Jan. 2020, Art. no. 103006.

[34] P. Zhou and N. El-Gohary, "Ontology-based multilabel text classification of construction regulatory documents," *J. Comput. Civil Eng.*, vol. 30, no. 4, Jul. 2016, Art. no. 04015058.

[35] Y. Kang, Z. Cai, C.-W. Tan, Q. Huang, and H. Liu, "Natural language processing (NLP) in management research: A literature review," *J. Manage. Anal.*, vol. 7, no. 2, pp. 139–172, Apr. 2020.

[36] S. L. Marie-Sainte, N. Alalyani, S. Alotaibi, S. Ghouzali, and I. Abunadi, "Arabic natural language processing and machine learning-based systems," *IEEE Access*, vol. 7, pp. 7011–7020, 2019.

[37] M. Niknam and F. S. Jalaei Karshenas, "Integrating BIM and manufacturer product data using the semantic web technologies," in *Proc. ITCoN*, vol. 24, 2019, pp. 424–439.

ERLING ONSTEIN received the M.S. degree in photogrammetry and surveying, and the Ph.D. degree in geographical information technology from Norwegian University of Life Sciences (NMBU), in 1983 and 2004, respectively. He has worked with map production in a private company, from 1983 to 1988; in Oslo municipality, from 1988 to 1989; and as a GIS Teacher, from 1990 to 1994. Since 1999, he has been with what is now Norwegian University of Science and Technology (NTNU), Gjøvik, where he has been an Associate Professor, since 2005, with a break from 2009 to 2013, working on national and international standardization at Norwegian Mapping Authority. His research interests include information modeling parts of geomatics and also BIM.

DONGMING GUO received the bachelor's and master's degrees in computer science from the Southwest Jiaotong University, China. He is currently pursuing the Ph.D. degree with Norwegian University of Science and Technology, Norway, with focus on semantic research for improving BIM interoperability. His research interests include semantic technologies, data integration in construction industry, BIM, and digital twin in construction industry.

ANGELA DANIELA LA ROSA received the Laurea degree in applied chemistry from the University of Catania, Italy, in 1994, and the Ph.D. degree in polymer science with most of the work done at the Strategic Technology Group of ICI, Imperial Chemical Industries, U.K., in 2001. She is currently an Associate Professor with the Department of Manufacturing and Civil Engineering, Norwegian University of Science and Technology (NTNU), Gjøvik. In 2006, she started to focus her research activity on the evaluation of life cycle assessment (LCA) of new formulations of polymer composites and bio-composites. Her research and teaching interests include LCA, sustainable development and circular economy, and the production and end-of-life processes of polymers and polymer composites.

· · ·