CONCEPT FOR ENRICHING NISO-STS STANDARDS WITH MACHINE-READABLE REQUIREMENTS AND VALIDATION RULES

Sven Zentgraf, Sherief Ali & Markus König
Chair of Computing in Engineering, Ruhr-Universität Bochum, Germany

ABSTRACT: *During building project planning, various standards, such as material specifications, value ranges, and construction regulations, must be considered. When analyzing a regulation for its BIM-based use, it must be identified which information can be checked directly or indirectly using a BIM model. The basis for the directly checkable information requirements is the explicit description of object classes, object types, properties, and values. Additionally, complex validation rules can be derived from the standards. These information extractions are mostly performed manually and laboriously on text-based regulatory documents. To provide a better data format, the NISO proposed the Standard Tag Suite (NISO-STS), which is an XML format for publishing and exchanging full-text content and metadata of standards. This paper proposes a concept to enrich standards in NISO-STS format with information requirements and validation rules to provide a machine-interpretable semantic knowledge base for BIM processes. To achieve this, the concept utilizes natural language processing (NLP) methods to extract semantic information from the standards. Furthermore, the paper introduces a workflow to transfer the gathered knowledge into the XML-based standard. This allows the acquired semantic knowledge to be used BIM-based and directly updated in future versions of the standards. To show the applicability of the concept an approach is presented in which the obtained information is stored and used as a queryable knowledge base. The resulting database is used by a querying assistant, in which a user can enter keywords and questions that are translated into SPARQL queries to provide answers for the given input.*

KEYWORDS: *Natural Language Processing (NLP), NISO-STS (Standard Tag Suite), Smart Standards, Rule-based model checking, Semantic knowledge*

1. INTRODUCTION

In civil engineering, managing and exchanging information is a non-trivial task due to the complex nature of construction projects and the involvement of numerous stakeholders (Alani et al., 2021; Tomczak et al., 2022). These stakeholders in the Architecture, Engineering, Construction, and Operations (AECO) industry put their priority on ensuring compliance with standards, regulatory documents, and other requirements (Beach et al., 2015). However, formalizing the Information Requirements (IR) and intricate validation rules poses a significant challenge and has hindered the widespread adoption of Building Information Modeling (BIM) practices (Tomczak et al., 2022). The current manual compliance checking process is prone to errors, time-consuming, and costly. Due to the high effort required for the testing processes, in practice testing is only carried out on a random sample and not in its entirety (Z. Zhang et al., 2022, Fauth, 2021). These drawbacks have motivated extensive research into Automated Compliance Checking (ACC).

Acquiring precise IRs and validation rules from guidelines and standards remains a significant obstacle, particularly considering that many of these documents exist in non-machine-readable formats (Schönfelder & König, 2021). Thus, there is a need for a machine-readable and interchangeable format of representation for regulatory documents and standards to extract the desired information for ACC checking. To tackle the challenge of not machine-readable standards, the German Institute for Standardization (DIN) has started the Initiative Smart Standards (Czarny et al., 2021). Their objective is to convert their published standards into machine-readable documents. To fulfill this, they have adopted the NISO Standard Tag Suite (NISO-STS), an XML-based extendable data format introduced by the National Information Standards Organization (NISO). This standardized format is designed to present and preserve the content and metadata of standards and regulatory documents (NISO Standards Tag Suite Working Group, 2017).

This paper presents a concept for enhancing standards and regulatory documents in the NISO-STS format with IR and validation rules. The aim is to establish a machine-interpretable semantic knowledge base that can be seamlessly integrated into BIM processes. To achieve this, natural language processing (NLP) methods are facilitated to extract relevant semantic information from standards in NISO-STS representation. Furthermore, the concept introduces a workflow to transfer the gathered knowledge back into the XML-based standard to link the extracted IRs and rules with the original text. This allows the extracted semantic knowledge to be used on a BIM basis and updated directly when new versions of the standard are created. To show the applicability within BIM

Referee List (DOI: 10.36253/fup_referee_list)
FUP Best Practice in Scholarly Publishing (DOI 10.36253/fup_best_practice)

workflows, the paper concludes with a demonstrator that can be used to query the semantic knowledge to obtain relevant validation rules and IRs. Overall, this paper aims to address the challenges faced in ACC, information management, and integrating digital information into the BIM workflow.

2. BACKGROUND

This section presents the key terms, concepts, and relevant research for the concept presented in this paper. Initially, the section provides an overview of the Standard Tag Suite developed by the NISO, followed by an introduction to the Initiative Smart Standards proposed by the DIN, which employs this tag suite. Following, an introduction to NLP-based information extraction and knowledge representation is provided. The chapter concludes with a presentation of the current state of research on the topic of code compliance.

2.1 Digital standards

As denoted in Section 1 there is a need for machine-readable, interchangeable, and maintainable regulatory documents and standards for ACC (Schönfelder & König, 2021). For this purpose the NISO in particular the NISO Standards Tag Suite Working Group published the NISO Standards Tag Suite (NISO-STS) in 2017 (NISO Standards Tag Suite Working Group, 2017). The NISO-STS defines a set of XML elements and attributes that describe the complete content and metadata of standards. This includes co-produced standards and standards bodies' adoptions of existing standards, to establish a universal format for publishing and exchanging standards content in all shapes. The primary objective of the NISO-STS is to preserve the content of standards, irrespective of how they were created and delivered. It enables the acquisition of structural and semantic components without being bound to a specific order or textual arrangement. The standard consists of two implementations, referred to as the Interchange Tag Set and the Extended Tag Set. These Tag Sets are constructed from the elements and attributes defined in the NISO-STS and are designed to function as models for publishing and enhanced interoperability of standards and regulatory Documents (NISO Standards Tag Suite Working Group, 2017).

Within Initiative SMART Standards (IDiS) of the DIN, the concepts and implementations of the NISO-STS are used to advance the digitalization of German regulations and standards (Czarny et al., 2021). The IDiS facilitates the establishment of digital standards, which offer information necessary for standardization tasks in a suitable format and scope. A whitepaper has been developed to foster a common understanding and clear action scenarios for implementing digital standards. The document provides a comprehensive understanding of various scenarios concerning standards, encompassing aspects such as maturity, readability, feasibility, interpretability, and even the potential for machine-driven creation. It also addresses the different levels of autonomy in the creation and application of standards and regulatory documents. These levels span from level 0, representing the traditional paper-based format, to a potential level 5, depicting a future scenario where standards are directly influenced and optimized by artificial intelligence (AI). Currently, the DIN is actively engaged in converting all its rules and regulations into a machine-readable XML serialization using the aforementioned NISO-STS model. To effectively implement rules for the ACC of building information models, digital standards at level 3 or higher are necessary. This indicates that the standards must reach a level of autonomy where they can be interpreted and applied by humans and machines (Czarny et al., 2021).

In this contribution, Autonomy Levels 2 and 4 are utilized. A document in Level 2 is a machine-readable XML document and allows the extraction of its textual content and other structural elements for further processing. Within the document’s chapters, sentences, graphics, and tables are distinguishable which simplifies a separate examination of the individual components. A Level 4 document contains, in comparison to a Level 2 document, not only machine-readable but also machine-interpretable content that enables a close linkage with execution and application information. These features allow seamless integration of the contained information into other information systems and software tools (Czarny et al., 2021).

2.2 NLP-based information extraction

In the construction industry, the checking of specifications from building codes, regulations, and standards plays a crucial role. Stakeholders involved in a construction project must adhere to precise guidelines in both design and realization, and adherence to these guidelines needs to be consistently proven. One way to ensure compliance is through ACC of building designs. However, for ACC to work effectively, it requires converting the natural language specifications found in regulatory documents into machine-readable constraints (Fuchs & Amor, 2021). To achieve this, NLP methods can be facilitated. NLP is a subfield of AI and computer linguistics that focuses on the interaction between computer or formula languages and human language. It involves the development of

algorithms and language models that enable machines to understand, interpret, generate, and manipulate human language effectively (Chowdhary, 2020). NLP utilizes a diverse set of techniques to gain a comprehensive understanding of natural language. These techniques have been applied in various studies to facilitate the full automation of the code compliance process during information extraction from regulatory documents.

Schönfelder and König (2021) proposed a Named Entity Recognition (NER) based model that trained German building code documents on the pre-trained German corpus BERT (Bidirectional Encoder Representations from Transformers). BERT is a language representation model, which is pre-trained on the deep-directional representation of unlabeled text. BERT has shown promising results in eleven natural language processing tasks such as question answering and language inference (Devlin et al., 2019). Schönfelder and König (2021) used the NER technique to label text in the building code as they aimed to train the network based on supervised learning. The study demonstrated results of average performance values of 95.7 % precision and 95.2 % recall. The authors discussed limitations in the study as the proposed concept could provide good results only to the German corpus used.

Some studies employ another technique besides NER as Zhou et al. (2022), and R. Zhang and El-Gohary (2020). Zhou et al. (2022) proposed an approach that deploys NER and Context-Free Grammar (CFG) to formulate a generalized rule interpretation framework. The study focused on analyzing regulatory text to create a syntax tree representing roles and concepts and developing a deep learning network using transfer learning to label the semantic elements in the text. Zhou et al. (2022) presented outcomes with an accuracy of 99.6 % and 91.0 % for parsing single- and multi-requirement sentences. The study stated that it focused on quantitative sentences in the regulatory documents and that more types of sentences will be addressed.

In their work, R. Zhang and El-Gohary(2020)introduced a new machine learning-based approach to automatically match building-code concepts and relations with their equivalent concepts and relations in the Industry Foundation Classes (IFC). The approach was implemented and tested on chapters from the 2009 International Building Code (IBC) and the Champaign 2015 IBC Amendments. The preliminary results achieved a semantic matching performance of 77 % accuracy for matching building-code concepts to IFC elements and 78 % accuracy for matching building-code relations to IFC relations.

2.3 Knowledge representation

There are two fundamental concepts for data and knowledge representation, which are ontologies and knowledge graphs. Ontologies are used widely in various domains to represent the semantics of a specific domain and provide standardized data representation (Ehrlinger & Wöß, 2016). The Interconnected Data Dictionary Ontology (IDDO), developed by Zentgraf et al. (2022), was designed to digitize knowledge from building regulations and construction guidelines. It offers a data schema to describe and manage properties in accordance with ISO 23386 (ISO 23386, 2020). The ontology organizes the digitized knowledge into a hierarchically structured tree of property groups and properties, extracted from natural language texts. Its main purpose is to provide an architecture for transforming building codes into a structured, knowledge-represented format. Encoded in Web Ontology Language (OWL) (Motik et al., 2012), it enables seamless integration and utilization of digitized knowledge in various applications.

Other studies focus on rule formulation from regulatory documents that help the development of ACC frameworks. Wessel et al. (2013) proposed an approach with two parts. The first part involves building an ontology to capture and represent the standards and information found in regulatory documents. In the second part, the enriched ontology is utilized to extract rules. These rules are derived from the information contained within the ontology, facilitating the automated extraction and formalization of regulatory guidelines and requirements. The author states that there are further improvements they are targeting in the future, which aim to enrich the knowledge base and improve automatic reasoning and extraction of rules (Wessel et al., 2013).

2.4 Code Compliance

Building codes are a part of the construction work, which aims to ensure the integrity and compliance of the planned structure. As the review of building codes is a time-consuming and error-prone process, there are a lot of efforts to digitize the process. Accordingly, many research studies use different methodologies to reach the goal of rule extraction. Eastman et al. (2009) stated that the process of rule checking is composed of four steps, which are (1) rule interpretation and logic structuring ; (2) building model preparation; (3) rule checking, and (4) reporting the checking results.

Schwabe et al. (2019) proposed an approach that aims to create a model-based rule checking for the planning of construction site layouts. The study used the open-source rule engine Drools and Industry Foundation Classes (IFC) to extract information from building models and apply rules to the information extracted. Drools is a Business Roles Management System (BRMS) based on Java language, which allows users to create decision models that are based on rules formulated as *when-then* statements, to take an action when a condition is true (Browne, 2009). Schwabe et al. (2019) stated that they are planning to extend the rule sets and experiments with other rule languages in future research.

In their work, Beach et al. (2015)introduced a method that utilizes the semantic web to achieve a comprehensive understanding of regulatory documents. The authors divided the semantic web knowledge into three main concepts, where each concept targets a specific knowledge in the regulation document. Additionally, the authors have used RASE tags to annotate the regulatory document, which is a markup technique applied to text and it is composed of four operators; Requirement, Applicability, Selection, and Exceptions (Hjelseth & Nisbet, 2011). Accordingly, Beach et al. (2015) converted the tags into Semantic Web Rule Language (SWRL), which can conceivably detect if the regulation is in the scope or not.

3. CONCEPT

This paper proposes a concept for the enrichment of digital standards in Level 2 with machine-readable IRs and validation rules to create digital standards at autonomy Level 4. The objective is to establish a machine-interpretable semantic knowledge base that can be integrated into the BIM methodology to support planning processes, ACC, and other BIM practices. To achieve this, an XML-Crawler first processes a digital standard at autonomy Level 2 to extract all contained textual information. This textual information is then forwarded into an NLP pipeline that extracts all relevant IRs and validation rules from the natural language texts. After the extraction the requirements and rules are further processed into suitable data representations and stored in respective databases. The stored data is then used to enrich the analyzed standard to make it compliant with autonomy Level 4 and additional areas of application for the stored data. Additionally, the entire concept is designed to be realized using open data formats and interfaces, aligning with the principles of the openBIM concept.

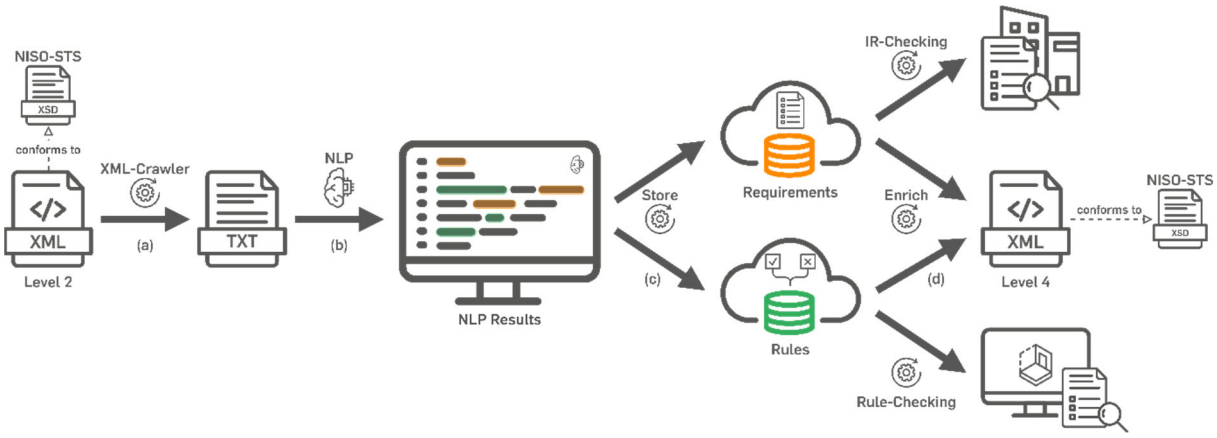

Fig. 1: Schematic representation of the concept

The input for this framework is as mentioned above an XML-serialized digital standard at autonomy Level 2. It complies with the XML Schema Definitions (XSD) specified by the NISO-STS. At Level 2, the document is machine-readable, enabling automated extraction of its structural elements. Granular contents such as chapters, sentences, graphics, and tables can be distinguished and extracted from the document (cf. Section 2.1). Moreover, the separation between representation and content allows a more streamlined and efficient processing of the information (Czarny et al., 2021).

In the initial step of the developed workflow, the uniform structure provided by the autonomy level is utilized. The structure allows the creation of an XML crawler that can efficiently extract all relevant textual information from the standard. A document or in this case, XML crawler is a software or script designed to automatically search and navigate through documents to extract information from these documents and store it for further processing, analysis, or database indexing. Document and XML crawlers are commonly used in search engines and knowledge

management systems. The extracted textual data taken from the NISO-STS compliant standard is stored in a plain text format for further processing (cf. Fig. 1 (a)).

Subsequently, algorithms from the field of NLP can be trained using these text-based input data. The aim of the conceptualization, implementation, and training of NLP algorithms within the proposed concept is to find two or optimally one NLP pipeline to find and extract IRs and validation rules which can be further processed (cf. Fig. 1 (b)). To enable the training of such NLP pipelines, further preprocessing of the provided textual information may be necessary. This preprocessing could involve tasks such as sentence tokenization, lemmatization to revert words to their base forms, or conversion of input texts into vectorized representations.

In the third step, the output generated by the NLP pipeline, which includes the identified IRs and validation rules, is further processed (cf. Fig. 1 (c)). The goal is to find suitable data representations for both the IRs and the validation rules. This step involves transforming the extracted information into structured and machine-interpretable formats, making it more accessible for downstream tasks. To achieve this, databases are created to store and manage the extracted IRs and validation rules. These databases are designed to support the creation, storage, organization, exchange, and utilization of data in structured and machine-readable formats. Special attention is given to utilizing open interfaces to ensure interoperability and flexibility. By leveraging open REST (Representational State Transfer) APIs, the databases make the stored information available for seamless integration with other systems and applications.

The last step of the concept is divided into two parts. In the enrichment, IRs and validation rules stored are accessed through REST APIs. Leveraging this extracted data, the analyzed standard of Level 2 is enriched with the extracted elements to a smart standard in autonomy Level 4. As denoted in section 2.1 a Level 4 standard encompasses machine-interpretable content, strongly linked with execution and application information. This capability enables direct executability and seamless integration with other relevant information sources (Czarny et al., 2021). By incorporating machine-readable features, the standard becomes easily interpretable and executable by machines, minimizing the need for manual intervention. This automation enhances efficiency and precision in processes reliant on the standard. Furthermore, the integration of execution and application information allows the automation of specified actions and enables interactions between interconnected systems and data sources. With this enriched autonomy, the standard gains agility in handling dynamic tasks and adapting to changing scenarios. ACC and other complex procedures can be executed with greater effectiveness and consistency, supporting workflows, and enhancing data interoperability.

The second part of the final step focuses on other areas of application of the information extracted from the standards. Fig. 1 (d) illustrates an exemplary area of application of the automatic validation of IRs and validation rules directly at a BIM model, both formally and technically. This type of validation could be integrated into the lifecycle of a construction project during a digital building permit review process. There are several other potential applications for the provided information. For instance, it could be utilized to define the Level of Information Need (LOIN) during the tendering process of buildings or to formulate general modeling guidelines for construction projects. Moreover, it can also support the creation and versioning of new or existing standards, facilitating a more streamlined and efficient standardization process.

By offering different possibilities for leveraging the extracted information requirements and validation rules, this shows the relevance and impact of the NLP-based analysis of standards and regulatory documents in the AECO domains. It enables stakeholders in the construction industry to implement automated validation processes, more streamlined tendering procedures, and maintain consistent modeling practices, leading to improved efficiency and enhanced collaboration throughout the entire construction lifecycle.

4. USE CASE

We aim to develop an approach to convert the knowledge in regulatory documents into a machine-interpretable representation. As the regulatory documents are mostly not computer processable, the purpose is to represent the knowledge in the regulatory document in a knowledge base, which is computer interpretable. Accordingly, the next step is that we apply the rules based on the retrieved data as illustrated in Fig. 1 on a regulatory document.

The regulatory documents used in this demonstrator are from the Research Society for Roads and Traffic (FGSV) (FGSV Verlag GmbH, 2023), which creates the technical regulations for the entire road and traffic system in Germany. The regulation's language is German, but the concept has broad applicability and can be extended to any other language. The FGSV has multiple regulations in this area, while we focused on FGSV 499 – RStO12

(Forschungsgesellschaft für Straßen- und Verkehrswesen, 2012). The proposed use case places particular emphasis on one chapter of the mentioned regulatory document. The chapter encompasses introductory text, definitions, tables, and interrelated constraints. The approach of knowledge extraction required extensive and comprehensive reading and understanding of the document, to extract the correlations and interrelationships in the text. The knowledge acquisition is performed manually by highlighting the logical sentences, descriptive texts, and the relationships between the tabulated data and the plain texts.

4.1 Data Preparation

The subsequent action entails gathering the knowledge extracted in a machine-readable format, to be able to formulate rules upon the acquired information. The representation of knowledge is based on the semantic web by employing the OWL and the Resource Description Framework (RDF). The ontology hierarchy and relationships between classes is a complex stage that requires a significant amount of attention to achieve the accurate formulation of knowledge extracted. The software used to create the ontology is Protegé (Musen, 2015). The acquired knowledge from the regulatory document is centered around the construction of roadways. The regulatory document encompasses different classes of soils and the requirements for the subsoil or substructure according to frost sensitivity or other constraints. The soil classes have a minimum thickness delegated to each class and this thickness could be increased or decreased whenever exposed to local conditions, for instance depending on the zone, underground water conditions, or drainage of the roadway. Each local condition possesses a value, which has a positive or negative sign, to raise or lower the thickness of the soil class.

Table 1: Increased or reduced thicknesses due to local conditions (translated (Forschungsgesellschaft für Straßen- und Verkehrswesen, 2012))

Local conditions		**A**	**B**	**C**	**D**	**E**
Frost effect	Zone I	±0 cm				
	Zone II	+5 cm				
	Zone III	+15 cm				
small-scale Climate differences	unfavorable climatic influences, e.g., due to North-facing slopes or in ridge locations of mountains		+5 cm			
	no special climatic influences		±0 cm			
	favorable climatic influences with closed lateral development along the street		-5 cm			
Water conditions in the subsoil	No groundwater and stratum water down to a depth of 1.5 m below the subgrade level			±0 cm		
	Groundwater or stratum water permanently or temporarily higher than 1.5 m below ground level			+5 cm		
Location of the gradient	Incision, gating				+5 cm	
	Terrain height up to 2.0 m				±0 cm	
	Dam > 2.0 m				-5 cm	
Drainage of the roadway/execution of the edge width	Drainage of the roadway via swales, ditches, or the embankments					±0 cm
	Drainage of the roadway and peripheral areas via gutters or drains and pipelines					-5 cm

Furthermore, the regulatory document comprises a section on asphalt base courses, where tabulated data shows different types of base layers without binders and each type has a thickness depending on the load-bearing capacity (cf. Table 1).The document encompasses abundant data about how to construct the roadways and the constraints encountered, which affect the substructure or superstructure. The stated knowledge is represented as an ontology through different classes and properties. To ensure the ontology's coherence and compatibility, it is structured based on the IDDO (Zentgraf et al., 2022), which provides an architecture for transferring building codes into a structured format based on the data schema of ISO 23386 (ISO 23386, 2020). As a result, the knowledge is well structured through classes and properties in a standardized way.

4.2 Information retrieval

As a prerequisite, it is assumed that the conversion of the regulatory document into a machine-readable format was achieved successfully. The prevailing stage is to retrieve the information stored in the knowledge base in order to be able to apply rules to the retrieved data. The chosen retrieval method is the SPARQL Protocol and RDF Query Language. SPARQL is selected due to its ease of use and applicability to manipulate RDF data, which supports the proposed framework. The ontology is established and structured using Protegé. To enhance the concept's capabilities and prepare for the subsequent step of formulating rules, the queries are interconnected with the RDF data. The querying stage starts with trials to ensure the functionality of the ontology, whether the data is retrieved correctly or not. The objective is to retrieve the classes, properties, and annotations from the ontology. We used to identify the vocabulary from the ontology precisely as SPARQL queries are sensitive and comply with the class naming pattern in the ontology. As a case in point, to begin with, we utilized the first part of the regulatory document, which aims to find the minimum thickness for soil classes. The soil classes are *F1*, *F2*, and *F3*, where each one has a relationship to a local condition. The main class in the ontology, which stores the soil classes is named *Frostempfindlichkeitsklasse*, thus we call the main class in the queries, to get the subclasses, annotations, and relationships assigned to it. The objective is to retrieve the thickness of the soil class, as well as the thickness assigned to the local condition. As a result, the total thickness of the soil class could be calculated, and the final thickness is the sum of the soil thickness and the local condition thickness variable. The final thickness calculation will not be executed through SPARQL queries. This step will be calculated during the formulation of the validation rule instead.

```
WHERE
{{
        ?frostklasse rdfs:subClassOf ont:{class_name} .
        ?klasse rdf:type/rdfs:subClassOf* ?frostklasse .
        ?klasse ont:hasThickness ?Dicke .
        ?klasse ont:BoundaryValue ?Constraint .
        ?Constraint ont:hasThickness ?ConstraintDicke .
        ?GroupOfProperties ont:DateOfCreation ?DateOfCreation .
        FILTER (regex(str(?klasse), "{klass}", "i") && regex(str(?Constraint), "{condition}", "i"))
 }}
GROUP BY ?frostklasse ?Dicke ?klasse ?Constraint ?ConstraintDicke ?DateOfCreation
```

Fig. 2: Excerpt of an example query for detecting the Soil class, the Thickness, the Local condition, and the Local condition

As an example, Fig. 2 shows an excerpt of a query formulated for Rule 2, which detects the Frostklasse (Soil class), Dicke (Thickness), Constraint (Local condition), and Constraint Dicke (Local condition thickness). The same pattern of query formulation is employed for other knowledge data stored in the ontology, taking into account the distinctions in parameters.

4.3 Rule Checking

Our purpose is to build a framework that consists of three components, which are ontology, SPARQL queries, and the Python programming language for conducting additional analysis on the retrieved data and rule development. The rule's structure consists of a condition part *if statement* and the outcome part *Then and Else statements*. Thus, the rules are formulated to execute the results based on specific conditions.

For instance, a rule logic is *if the user selects a class named X*, and *if the user selects a local condition named Y,* then the data for the assigned class based on the local condition selected by the user will be retrieved. Accordingly, some actions will be executed and applied to the retrieved data. For example, the aim is to calculate the final

thickness of the subgrade, where the thickness depends on the minimum thickness. This is specified depending on the soil class and local conditions, to increase or decrease the total thickness of the substructure. As a result, the rules calculate the final thickness of the substructure based on the class selected by the user and the local condition selected as well. One of the significant aspects taken into account is that not all the data can be retrieved from the knowledge base with only one SPARQL query, thus every rule formulated has its query. Three rules have been derived from the knowledge extracted from the regulation document.

4.4 User Interface

In the upcoming step, a user interface is implemented, which uses the created validation rules and the converted excerpt from the regulatory document. The aim is to allow the user to interact with the regulatory document through input provision and rule selection. We developed a user interface, which prompts the user for input, such as keywords, filters, or specific entities directly into the interface. Accordingly, the system uses the input to construct a SPARQL query to retrieve the relevant data in the ontology. Furthermore, the interface provides users with a list of pre-defined descriptive rules, each designed to process data based on a certain logic. Additionally, once the user selects a rule, the interface provides an instruction statement, which guides the user on what the system needs to process the data as illustrated in Fig. 3.

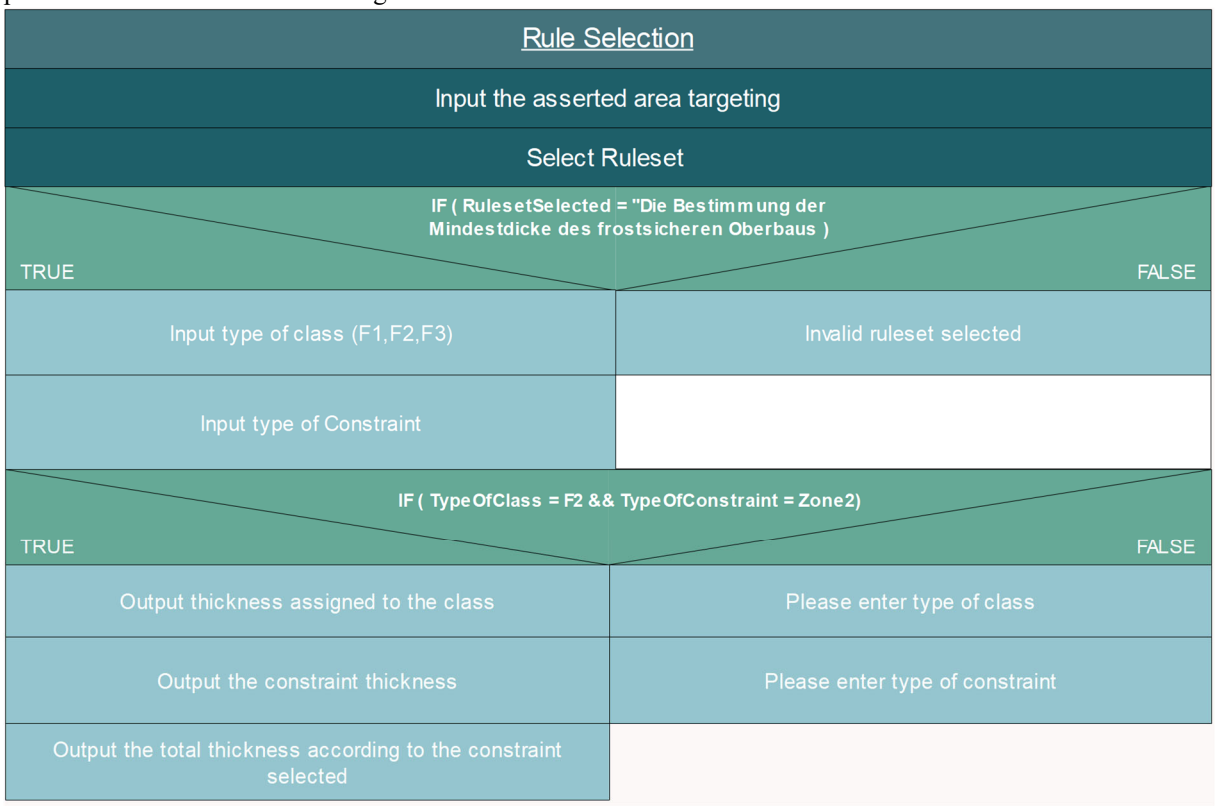

Fig. 3: Nassi-Schneidermann diagram of the program logic of the user interface

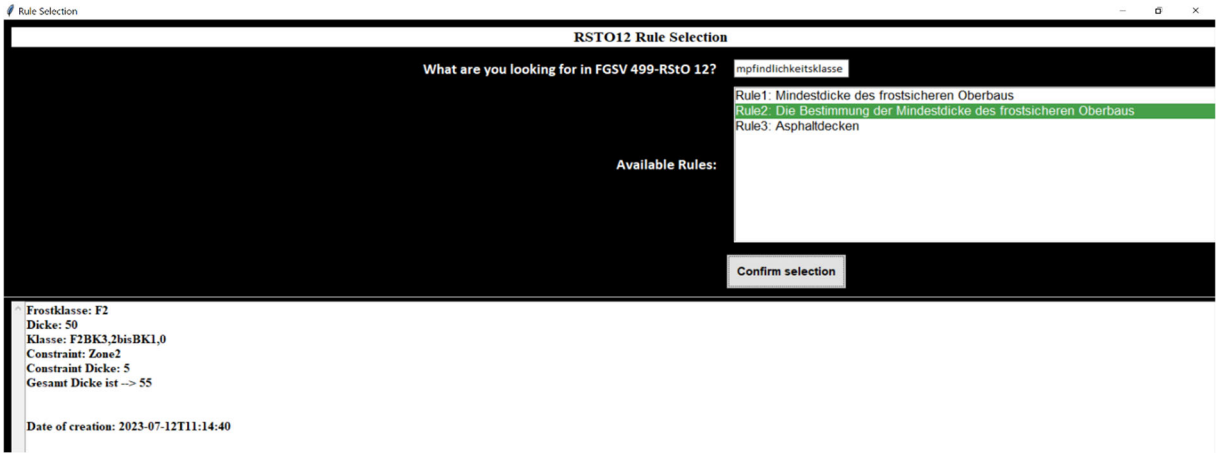

Fig. 4: Results based on the class name specified and the selected rule

The integration of SPARQL and Python rules delivers a comprehensive solution to interact with the knowledge base and perform rule-checking. This process enables the users to interact with the ontology. Fig. 4 shows the prototype of the implemented user interface.

5. CONCLUSION

This contribution provides a concept to enhance digital standards in Level 2 by incorporating machine-readable IRs and validation rules to advance these digital standards into autonomy Level 4. The concept uses an XML document crawler and NLP algorithms to analyze the textual information of an examined standard. The extracted IRs and validation rules are converted and stored in semantic knowledgebases and are made available via open RESTful APIs. With these open interfaces, the rules and requirements can be used to transform the considered standard into a machine-interpretable standard of autonomy Level 4. Furthermore, the gathered rules and requirements can be used in other areas of application within the BIM methodology (cf. Section 3).

One of these application areas is presented in detail in the use case (cf. Section 4). An approach is presented in which the obtained information can be used as a queryable knowledge base. For this purpose, it is assumed that a standard was processed by the NLP algorithm in advance and that the results are available for further processing. In the next step, the obtained information is structured according to the IDDO ontology and stored in a graph database. The resulting database is used as a semantic knowledge base for a querying assistant in the following. Within the assistant, the user can input a fixed set of inputs to query the knowledge base. The entered keywords and questions are translated into SPARQL queries which search the knowledge base to provide the user with an answer to the given input. The presented use case shows the feasibility of the presented concept with a restricted set of possible filters and questions. In future research, it can be considered to extend pre-trained networks, like ChatGPT, by incorporating extracted information from standards and regulatory documents.

Current parts of the concept have already been implemented while others need to be addressed in future work. In their work, Kandt and Zentgraf (2023) presented the implementation of an XML-Crawler for NISO-STS-compliant standards. With the outcome of this contribution, the process step shown in Fig. 1 (a) can be realized. The aforementioned IDDO ontology published by (Zentgraf et al., 2022) can be used to create a graph database for information requirements (cf. Fig. 1 (c)). Apart from outlining the data schema, the paper also introduces a method demonstrating how IDDO can ensure the accuracy of information requirements through the application of Shapes Constraint Language (SHACL) shapes.

In future research, several areas need to be addressed to realize the whole of the presented concept. Firstly, the identification of suitable NLP algorithms for extracting information requirements and validation rules is important. In the following step, it is necessary to define and establish a dedicated database to store, manage, and maintain the extracted validation rules. Building upon this, methods must be developed to automate the enrichment of a Level 2 standard using the extracted information requirements and validation rules, in order to obtain a Level 4 standard. Additionally, potential application areas within the building lifecycle need to be explored. Possible areas where the extracted information could be utilized effectively include Facility Management, refurbishments in combination with sustainability assessments, and other areas.

REFERENCES

Alani, Y., Dawood, N., Patacas, J., Rodriguez, S., & Dawood, H. (2021). A semantic common model for product data in the water industry. *Journal of Information Technology in Construction*, *26*, 566–590. https://doi.org/10.36680/j.itcon.2021.030

Beach, T. H., Rezgui, Y., Li, H., & Kasim, T. (2015). A rule-based semantic approach for automated regulatory compliance in the construction sector. *Expert Systems with Applications*, *42*(12), 5219–5231. https://doi.org/10.1016/j.eswa.2015.02.029

Browne, P. (2009). *JBoss Drools business rules*. *From technologies to solutions*. Packt Publ.

Chowdhary, K. R. (2020). Natural Language Processing. In *Fundamentals of Artificial Intelligence* (pp. 603–649). Springer, New Delhi. https://doi.org/10.1007/978-81-322-3972-7_19

Czarny, D. A., Diemer, J., Schacht, M., Bülow, G., Noll, M., Lochner, D., Gayko, J., Pervin, J. N., Rauh, P., &

Diedrich, C. (06/2021). *Scenarios for digitizing standardization and standards*. 10787 Berlin. https://www.din.de/resource/blob/801106/0251eb1280a9a97e53285d42d3bf1fea/whitepaper-idis-en-data.pdf

Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (2019). BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding, 4171–4186. https://doi.org/10.48550/arXiv.1810.04805

Eastman, C., Lee, J, Jeong, Y., & Lee, J (2009). Automatic rule-based checking of building designs. *Automation in Construction*, *18*(8), 1011–1033. https://doi.org/10.1016/j.autcon.2009.07.002

Ehrlinger, L., & Wöß, W. (2016). Towards a Definition of Knowledge Graphs. *Joint Proceedings of the Posters and Demos Track of 12th International Conference on Semantic Systems*. https://api.semanticscholar.org/CorpusID:8536105

Fauth, J. (2021). *Ein handlungsorientiertes Entscheidungsmodell zur Feststellung der Genehmigungsfähigkeit von Bauvorhaben* , Bauhaus-Universitätsverlag.

FGSV Verlag GmbH. (2023, August 10). *Forschungsgesellschaft für Straßen- und Verkehrswesen*. https://www.fgsv-verlag.de/

Forschungsgesellschaft für Straßen- und Verkehrswesen. (2012). *Richtlinien für die Standardisierung des Oberbaus von Verkehrsflächen: RStO 12* (Ausg. 2012). *FGSV: 499: R1*. FGSV-Verl.

Fuchs, S., & Amor, R. (2021). Natural Language Processing for Building Code Interpretation: A Systematic Literature Review. *Proceedings of the 38th International Conference of CIB W78*.

Hjelseth, E., & Nisbet, N. (2011). Capturing normative constraints by use of the semantic mark-up RASE methodology. In *Proceedings of CIB W78-W102 Conference*.

ISO 23386 (2020-03). *Building information modeling and other digital processes used in construction* (ISO 23386). Vernier, Geneva. ISO copyright office.

Kandt, K., & Zentgraf, S (2023). Development of a Python-based NISO-STS document crawler for the creation of NLP pipeline input data. *Proceedings of the 34th Forum Bauinformatik, Bochum, Deutschland*.

Motik, B., Patel-Schneider, P. F., & Parsia, B. (2012, December 11). *OWL 2 Web Ontology Language: Structural Specification and Functional-Style Syntax (Second Edition)*. https://www.w3.org/TR/2012/REC-owl2-syntax-20121211/

Musen, M. A. (2015). The Protégé Project: A Look Back and a Look Forward. *AI Matters*, *1*(4), 4–12. https://doi.org/10.1145/2757001.2757003

NISO Standards Tag Suite Working Group (2017-06). *ANSI/NISO Z39.102-2017, STS: Standards Tag Suite*. Baltimore, MD. National Information Standards Organization.

Schönfelder, P., & König, M (2021). Deep Learning-Based Entity Recognition in Construction Regulatory Documents. *Proceedings 38th International Symposium on Automation and Robotics in Construction (ISARC 2021).* Advance online publication. https://doi.org/10.22260/ISARC2021/0054

Schwabe, K., Teizer, J., & König, M (2019). Applying rule-based model-checking to construction site layout planning tasks. *Automation in Construction*, *97*, 205–219. https://doi.org/10.1016/j.autcon.2018.10.012

Tomczak, A., Berlo, L. v., Krijnen, T., Borrmann, A., & Bolpagni, M. (2022). A review of methods to specify information requirements in digital construction projects. *IOP Conference Series: Earth and Environmental Science*, *1101*(9), 92024. https://doi.org/10.1088/1755-1315/1101/9/092024

Wessel, C., Humberg, T., Poggenpohl, D., Wenzel, S., Ruhroth, T., & Jürjens, J. (2013). Ontology-based Analysis of Compliance and Regulatory Requirements of Business Processes. In F. Desprez (Ed.), *Proceedings of the 3rd International Conference on Cloud Computing and Services Science: Aachen, Germany, 8-10, May 2013.* SciTePress.

Zentgraf, S, Hagedorn, P., & König, M (2022). Multi-requirements ontology engineering for automated processing of document-based building codes to linked building data properties. *IOP Conference Series: Earth and*

Environmental Science, *1101*(9), 92007. https://doi.org/10.1088/1755-1315/1101/9/092007

Zhang, R., & El-Gohary, N. (2020). A Machine-Learning Approach for Semantic Matching of Building Codes and Building Information Models (BIMs) for Supporting Automated Code Checking. In H. Rodrigues, G. Morcous, & M. Shehata (Eds.), *Sustainable Civil Infrastructures. Recent Research in Sustainable Structures* (pp. 64–73). Springer International Publishing. https://doi.org/10.1007/978-3-030-34216-6_5

Zhang, Z., Ma, L., & Broyd, T. (2022). Towards fully-automated code compliance checking of building regulations: challenges for rule interpretation and representation. *European Conference on Computing in Construction EC3.* Advance online publication. https://doi.org/10.35490/EC3.2022.148

Zhou, Y.-C., Zheng, Z., Lin, J.-R., & Lu, X.-Z. (2022). Integrating NLP and context-free grammar for complex rule interpretation towards automated compliance checking. *Computers in Industry*, *142*, 103746. https://doi.org/10.1016/j.compind.2022.103746