INTERNATIONAL DESIGN CONFERENCE – DESIGN 2024
https://doi.org/10.1017/pds.2024.45

Automatic knowledge graph creation from engineering standards using the example of formulas

Janosch Luttmer ✉, Mostafa Kandel, Dominik Ehring and Arun Nagarajah

University of Duisburg-Essen, Germany

✉ janosch.luttmer@uni-due.de

Abstract

Engineering standards are an important source of knowledge in product development. Despite the increasing digitalisation, the provision and usage of standards is characterised by lots of manual steps.This research paper aims at applying automatic knowledge graph creation in the domain of engineering standards to enable machine-actionable standards. For this, a formula knowledge graph ontology as well as suitable information extraction techniques are developed. The concept is validated using the example of DIN ISO 281, showing the overall capability of automatic knowledge graph creation.

Keywords: *engineering standards, information extraction, knowledge representations, ontology, artificial intelligence (AI)*

1. Introduction

Product development is subject to constant change due to increasing product complexity, additional requirements towards sustainability, and increasing individualisation of products (Dumitrescu *et al.*, 2021; Isaksson and Eckert, 2020). For a company to improve quality, reliability and performance of its products while reducing development costs and time, the handling of knowledge is becoming increasingly important (Hicks *et al.*, 2002; Zhenyong *et al.*, 2020). Knowledge workers currently spend 16% of their time searching, understanding, and processing information from various sources, yet 44% of searches fail to find relevant information (Schubmehl and Vesset, 2014). Thus, solutions are required to enhance efficiency and effectiveness of knowledge work.

Engineering standards are a valuable source of knowledge and influence several decisions along the development process (Manoharan *et al.*, 2019). These documents are distributed by numerous organisations, such as the German Institute for Standardization (DIN), the International Organisation for Standardization (ISO) or the Association of German Engineers (VDI) (Hicks *et al.*, 2002), and enable companies to reduce development costs and time (Bender and Gericke, 2021). Despite the increasing digitalisation and automation of product development, standards are still distributed in the form of print or PDFs. As a result, engineers spend considerable time identifying useful standards, searching for information within them, manually extracting the information and transferring it to their target systems (Loibl et al., 2020; Ehring et al., 2023). As such, the current standards provision and workflow are not only inefficient, but also contradict efforts to digitise product development activities.

1.1. Trends towards machine-actionable and machine-interpretable standards

Due to the increasing demand from companies to obtain standards content according to their specific needs and to integrate it directly into target systems, standards development organisations have an interest in providing their content outside the traditional print and PDF formats. The German Initiative

for Digital Standards (IDiS) has derived 12 generic user stories from practical experience that describe the future creation, management, delivery, and use of standards. These range from finding specific standard content, such as requirements and formulas, to direct integration into target systems to support decision-making in product development (Czarny *et al.*, 2022). To meet future needs of industry, standards development organisations have already converted their documents into a standardised XML format, following the so-called NISO STS schema. DIN completed this transfer in 2018, transferring nearly 30,000 standards to a central XML database (Schuch and Wischhoefer, 2018).
Ongoing activities dealing with the provision of today's XML standards (machine-readable documents) and technologies beyond XML are summarised under the umbrella term SMART standards. The aim is to develop digital standards that are machine-actionable and machine-interpretable. As defined by Loibl et al. (2020), organisations are currently transitioning from machine-readable documents to machine-actionable content, which considers semantic information to automatically enforce operations in target systems. Machine-interpretability is achieved through the automatic interpretation of a target system based on the provided content, representing the highest form of automation. Although XML serves as the starting format for SMART Standards activities, it should be noted that the current version in NISO STS does not meet the requirements for machine-actionability (Loibl *et al.*, 2020), thus preventing the realisation of the generic user stories described above. One approach is to transform the XML content into a semantically enriched knowledge representation, while previous research has already shown that knowledge graphs (KG) can enable machine-actionability (Manoharan *et al.*, 2019; Ehring *et al.*, 2021). Hereby, a KG is a semantic network that interconnects data by representing entities and their relations whereas the knowledge structure is referred to as ontology (Huang et al., 2023). The KG creation from engineering standards was demonstrated using the example of mathematical equations that were automatically extracted from XML standards, transferred to graph patterns, and provided via flexible interfaces (Luttmer *et al.*, 2021). It should be noted, however, that only the equations and metadata, such as the standard number, were extracted, and no further contextual, descriptive elements for the correct interpretation of the equation were included. Addressing this limitation, Luttmer et al. (2022) introduced the so-called formula module incorporating descriptive elements based on a study of 400 standards documents. However, the authors also point out that these elements are currently added to the KG manually - an approach that is not feasible when scaling to almost 30,000 DIN standards. Solutions are therefore needed to automatically extract formulas and their contextual information from XML standards and merge them into KGs.

1.2. Information extraction and automatic KG creation

Information extraction aims at capturing and extracting useful information from unstructured or semi-structured data, such as text, images, audio, or video, whereas extraction techniques are divided into rule-based and learning-based approaches (Adnan and Akbar, 2019). Examples for extracting information from design relevant sources range from patents (Siddharth *et al.*, 2022) and engineering standards - including requirements (Luttmer *et al.*, 2023) as well as context and user specific information (Ehring *et al.*, 2023; Layer *et al.*, 2023) - to circular design guidelines (Gräßler and Hesse, 2023). While different research papers on extraction of valuable information from engineering standards exist, the automatic aggregation to KGs has not been investigated in detail.
Within the engineering design domain, Huang et al. (2023) present an approach for design knowledge acquisition in conceptual product design. Here, a KG is constructed using five layers, namely the data resources layer, the domain ontology layer, the entity extraction layer, the relation extraction layer, and the KG application layer. To extract entities and relations, a transformer-based method is used achieving an extraction accuracy of 94.2% for entity extraction and 68.5% for relation extraction (Huang et al., 2023). The final KG is stored in the graph database Neo4j. Huet et al. (2023) use a comparable strategy by creating KGs from design rules and contextual information, both integrated into a context-aware cognitive design assistant. For this, design rules are extracted from documents and transferred to a KG in Neo4j based on a defined domain ontology. Moreover, the solution interfaces with Computer-Aided-Design tools, actively supporting engineers in design tasks (Huet *et al.*, 2023). Zhao et al. (2020) present a research approach that focuses on the automatic extraction of information from technical documents and the construction of a domain KG. Leveraging learning-based extraction techniques, specifically

https://doi.org/10.1017/pds.2024.45 Published online by Cambridge University Press

TextCNN, the authors achieve a precision of up to 91% in experimental results. The KG resulting from the steps *data acquisition*, *definition of data models*, *information extraction*, *knowledge representation*, and *knowledge storage* is stored in Neo4j and serves as and knowledge platform (Zhao *et al.*, 2020).
In the domain of patents, which is also closely related to standards documents because of the technical language and standardisation of content, Siddharth et al. (2022) develop a comprehensive engineering KG by extracting "facts" from patent databases. Employing rule-based extraction with part-of-speech tagging and grammatical rules, the authors structure extracted rules in a dictionary and model them as RDF triples in a KG. This large and scalable graph aims to support inference, reasoning, and information recall in diverse engineering tasks (Siddharth *et al.*, 2022).
Collectively, the research works described above underscore the capability of automatic KG creation from diverse documents such as patents, design rules, and technical documents. Moreover, similarities in used methodologies are seen, typically involving the creation of a data model/ontology, extraction of relevant information, construction of the KG, and its storage in a graph database.

1.3. Research goals

To seamlessly integrate standards knowledge into development processes and enforce operations in engineering tools based on underlying information, a shift towards machine-actionable formats is evident. Existent content provision formats like PDF and even XML in NISO STS are not sufficiently considering semantic relations (Loibl *et al.*, 2020). Building upon the research work of Luttmer et al. (2021), knowledge graphs have demonstrated their efficacy in serving as semantically enriched knowledge representation of standards. However, the previous approach focused solely on mathematical equations whereas contextual elements such as symbol definitions and relations between equations were not included, limiting the correct interpretation and application of equations.
As existing approaches (see section 1.2) demonstrate the effectiveness of automated KG creation, this paper aims to transfer the approach to the domain of engineering standards. This research paper focuses on formulas as exemplary information object, mostly due to their high degree of formalisation which correlates to the suitability for machine-actionability (Loibl et al., 2020). In addition, formulas have a significant impact in standardisation, with approximately 75% of standards containing formulas, and are applied in almost every phase of the product development process - from design calculations to quality inspections (Luttmer et al., 2022). Hence, the automated formula KG creation will enable engineers to efficiently retrieve formulas from engineering standards and to directly apply them in company specific calculation tools. In detail, following research questions will be answered:

1. To what extend can formulas from engineering standards as well as their associated descriptive elements be automatically extracted and modelled in KGs?
2. Which conclusions can be drawn to optimize the future standards creation process?

The remainder of this paper is structured as follows: Section 2 provides on overview of the general concept for automatic creation of formula KG, while section 3 describes the detailed elaboration of this concept. The results are subsequently validated in section 4 and discussed in section 5. The paper concludes with a short summary and an outlook on future research in section 6.

2. Concept for the automatic creation of formula KGs

Based on the previously discussed literature, particularly the methodologies proposed by Huang et al. (2023) and Zhao et al. (2020), the automatic KG creation follows characteristics steps, namely the creation of an ontology, information extraction, the construction of the KG, and its subsequent storage. These steps are supported by the results of a systematic literature review on KG development processes, which comprises six key steps, following a top-down approach: (Tamašauskaitė and Groth, 2023)

1. Identify data,
2. Construct the KG ontology,
3. Extract knowledge: entity extraction, relation extraction, and attribute extraction,
4. Process knowledge: knowledge integration, mapping to ontology, and knowledge completion,
5. Construct the KG, including storage of KG, visualisation of KG, and enabling the usage,
6. Maintain the KG, including evaluation and updating of KG.

https://doi.org/10.1017/pds.2024.45 Published online by Cambridge University Press

These detailed steps provide the foundation for the concept developed in this section. While the final concept predominantly aligns with the process defined in Tamašauskaitė and Groth (2023), adjustments or tailoring of specific steps may be considered to optimally address the problem at hand. To further guide the concept development, several requirements are specified by domain experts (see Table 1).

Table 1. Requirements towards the automatic creation of formula KGs

ID	Requirements description
Req. 1	The KG shall use standards in XML format as input, considering NISO STS and MathML formats.
Req. 2	The KG shall incorporate formulas as well as their descriptive elements.
Req. 3	The KG shall merge formulas from different documents into one KG.
Req. 4	The KG shall be extendable by inserting additional documents.
Req. 5	The KG shall be created automatically.
Req. 6	The KG shall be of high transparency and understandability.
Req. 7	The KG shall automatically define relations based on logical decisions.
Req. 8	The KG shall store information in a single database.
Req. 9	The KG shall provide formulas via a standard interface.

While requirements *Req. 1* and *Req. 2* specify the input documents and the considered information objects within standards, requirements *Req. 3* to *Req. 5* focus on the knowledge extraction. Here, novel solutions need to be developed to automatically extract formulas and their descriptive elements as accurate as possible. Additionally, both the extraction and KG creation steps shall exhibit a high degree of transparency and understandability (*Req. 6*). Lastly, requirements *Req. 7* to *Req. 9* address postprocessing of knowledge as well as the storage of the KG. According to the solutions presented in section 1.3, the graph database Neo4j is chosen for storage and provision. Based on the literature as well as the defined requirements, the overall concept is developed (see Figure 1).

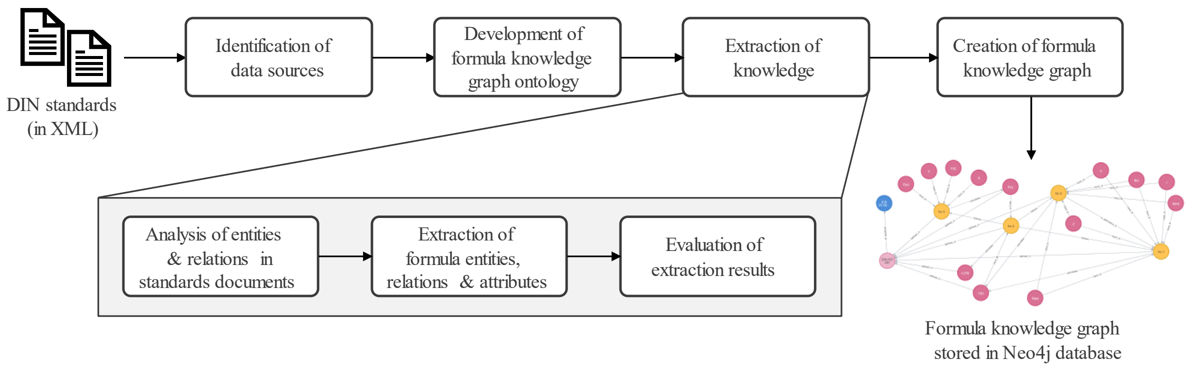

Figure 1. Process for automatic formula KG creation

3. Elaboration of concept

To realize the automatic creation of formula KGs, the previously described concept is elaborated in detail. Because of the importance of the knowledge graph ontology and the knowledge extraction step, this section will focus on these aspects while mentioning the key aspect of the remaining steps.

3.1. Identification of data sources

The process starts with identifying the domain of interest and data sources. While envisioning the integration of comprehensive standard knowledge in future, the focus of this research paper centres on formulas, their descriptive elements, and supplementary information. In terms of data source, only standards in XML format are considered. Because XML standards follow a standardized schema called NISO STS, defining the tags and attributes to be used, the automatic extraction of information is possible with a higher precision compared to PDF documents.

https://doi.org/10.1017/pds.2024.45 Published online by Cambridge University Press

3.2. Development of formula KG ontology

As this concept uses a top-down approach, the KG ontology is developed before the knowledge extraction. To guide the ontology development process, the Ontology Development 101 method (Noy and McGuiness, 2001) is used. The results are modelled in Protégé as RDF triples and stored in an OWL file. The classes and class hierarchy follow the work described in Luttmer et al. (2022). Here, two modules were introduced - the formula module and the variable module, both containing five descriptive elements. For this, 400 standards documents were analysed to identify relevant concepts and relations. The ontology´s completeness was then evaluated by expert reviews as well as the analysis of 50 additional documents, whereas no additional information was found. Based on this, the class diagram is developed. It contains the superordinate class *standards document* which is divided into the two subclasses *formula module* and *variable module*. These contain the descriptive elements defined in the corresponding module. Moreover, different types of relations are introduced. First, as indicated in Figure 2, symbols are related to the mathematical equation. Moreover, an equation can be related to another equation whereas two distinct types are defined: equation A (logically) follows equation B or equation A is alternative to equation B.

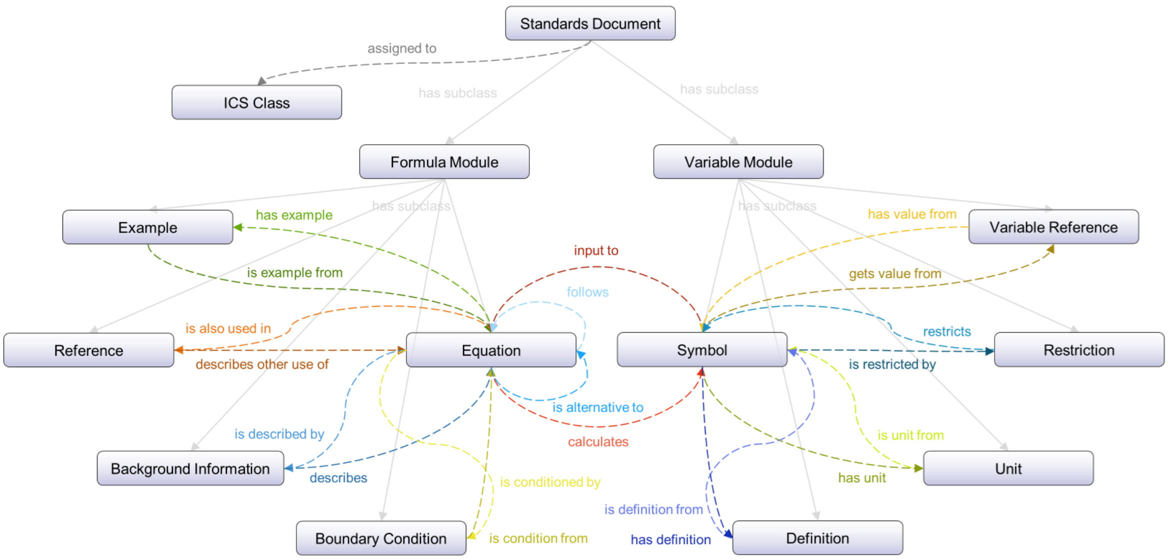

Figure 2. Formula KG ontology

3.3. Extraction of knowledge

The knowledge extraction is considered as main step of the automatic KG creation process as the overall quality depends on the accuracy of entity and relations extraction (Huang *et al.*, 2023). Previous analysis has shown that the mathematical equation itself, the used variables, their definitions as well as relations between equations are of high relevance to correctly interpret and apply formulas (Luttmer *et al.*, 2022). Thus, extraction solutions developed in this research will focus on these elements. As shown in Figure 1, the extraction of knowledge follows three steps that are described in the subsequent paragraphs.

Analysis of formula elements in standards

Before developing extraction solutions, an extensive analysis of standards documents with focus on formulas and their descriptive elements is performed. For this, a document corpus of 20 engineering standards - both in PDF and XML format - with a total of 1,400 formulas is manually analysed by two individual domain experts. The analysis aims at identifying syntactic and semantic characteristics of entities and their relations which includes not only the position in the text and specific keywords, but also the usage of XML tags as well as their nesting. The analysis results are displayed in Table 2.

Mathematical equations are represented as two main types in standards: displayed formulas and inline formulas. These two types are marked by specific XML tags, namely <disp-formula> and <inline-formula>. Because inline equations are mostly associated with boundary conditions and background

https://doi.org/10.1017/pds.2024.45 Published online by Cambridge University Press

information, this research work uses displayed equations as root elements of formula modules. In XML standards, displayed equations are moreover used in two ways: with or without a unique identifier (ID). The presence of an ID has implications for the development of later extraction solutions.
Variable modules - including the components *symbol* and *definition* - serve as integral components of equations. Their relation to the corresponding equation depends on the position relative to the equal sign, which is either left or right. Hence, based on the equal sign it can be determined whether a variable serves as input or output of the equation. Because the structure and notation of equations in standards document follow the NISO STS standard as well as the notation language *Presentation MathML 3.0*, variables and operators can be identified based on their XML tags. In addition to identifying variables, the assignment of definitions plays a significant role in ensuring the correct interpretation of variables. Definitions occur in two distinct ways. Firstly, symbols and their definitions are systematically listed in a dedicated chapter, following the rules defined by DIN 820 and the ISO Directives part 2. Furthermore, definitions are commonly found underneath equations - often presented as more detailed description.
According to section 3.2, two distinct general categories of relations between equations are defined. In this paper, these interrelated equations will be referred to as "root equation" and "related equation".

Table 2. Analysis results for considered entities and relations

Entity / Relation	PDF Analysis	XML Analysis
Mathematical equation	*Type 1:* displayed formula	*Option a:* Tag <disp-for> and ID *Option b:* Tag <disp-for> without ID
	Type 2: inline formula	Tag <inline-formula>
Variable / Symbol	Component of equation according to rules defined in DIN 1338 and ISO 80000	Tags according to NISO STS and MathML 3.0
Symbol Definition	*Type 1:* specific chapter or section for symbols and definitions	Tag <table>
	Type 2: symbols and definitions below equation	*Option a:* Tag <table> *Option b:* Tag <def-list>
Relation between equations	*Type 1:* usage of keywords to indicate relations	In total x different variants
	Type 2: relation out of context without usage of keywords	In total y different variants

Through an extensive analysis of PDF documents, five scenarios concerning relations have been identified, classifiable into keyword-based and non-keyword-based categories:

7. Related equation directly follows root equation with usage of keywords in text.
8. Related equation is mentioned with ID in text below root equation (within/outside of standard).
9. Related equation is mentioned with a reference to a specific chapter or section.
10. Root and related equation coexist in same chapter or section without explicit indication.
11. Root and related equation reside in different chapter or sections without explicit indication.

Development and evaluation of extraction approaches

Information extraction techniques can be differentiated into rule-based and learning-based methods. While learning-based approaches are predominant in literature, rule-based systems play a significant role in practical applications, mostly because of their high degree of understandability and transparency (Adnan and Akbar, 2019). As requirement *Req. 6* indicates that this feature is also required here and as the analysis results indicate the usage of specific patterns for entities and relations, a rule-based extraction system is developed. Based on the analysis, different patterns are developed by leveraging methods of natural language processing and keyword extraction. The results are iteratively evaluated using 20 standards documents. Figure 3 shows the final extraction process. Firstly, equations are identified and extracted from input documents. Afterwards, the equations are separated in variables and operators which also includes the rule-based interpretation of semantic features, such as whitespaces as substitutes for multiplication signs or the consideration of trigonometric functions. Moreover, multiple

https://doi.org/10.1017/pds.2024.45 Published online by Cambridge University Press

solutions are implemented dealing with inconsistencies and erroneous description of equations in XML standards. After separating variables and operators, the symbols as well as their definitions are extracted and assigned to the corresponding variable module. Lastly, the relations between equations are extracted which also includes references to other documents. Finally, two lists are generated as output of the knowledge extraction step: a list of equations including the variables and their symbol definitions and a list containing the relations between equations.

To evaluate the final extraction system, a set of 10 additional standards documents and the evaluation measure *Extraction Recall* are used. For this, a ground truth dataset is developed by manual annotation of two individual domain experts. Afterwards, to calculate the recall, the extracted elements are compared to the "true" relevant elements. Especially the extraction of equations and definitions show promising results with a recall of 100%. The extraction of variables poses some difficulties, mostly due to inconsistencies and errors in XML standards. Here, only 94% were extracted correctly. Moreover, the relation extraction between equations is limited, mostly due to the non-keyword-based scenarios. Additionally, the rule-based relation extraction relies on the identification of the exact same variables in possibly related equations. The described difficulties in sufficiently extracting variables lead to limitations in correctly extracting relations. However, a recall of 97% was achieved on the test data set.

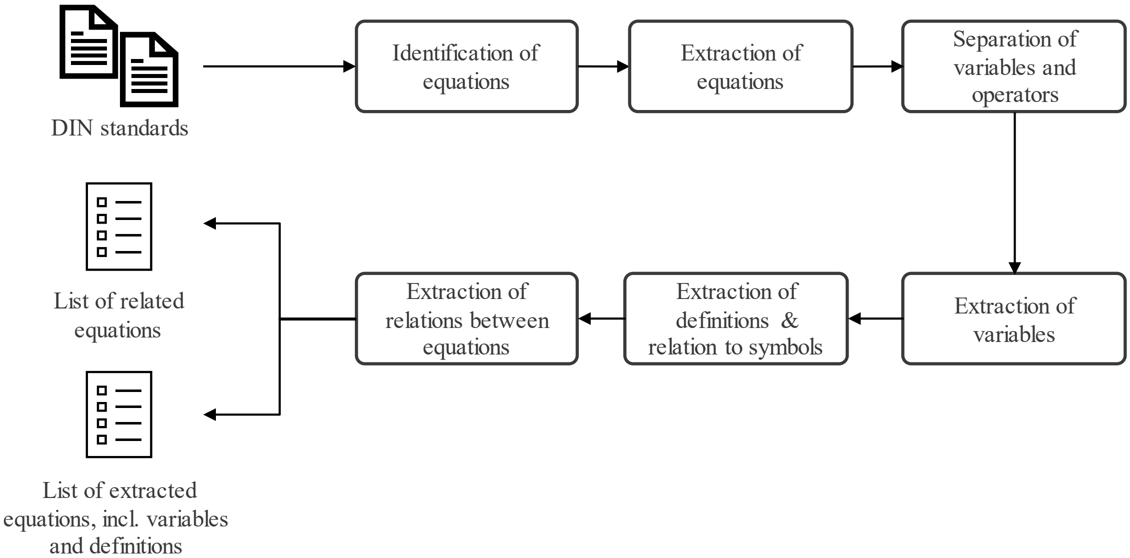

Figure 3. Automatic extraction of formula entities and relations

3.4. Creation of formula KG

Following the knowledge extraction, the formula KG is created. For this, the previously extracted lists are automatically transferred to graph patterns as RDF triples under consideration of the formula KG ontology defined in section 3.2. The KG is then stored in the graph database Neo4j using the query and modelling language Cypher. In addition to the ID, the mathematical equations nodes contain the MathML code which allows a direct integration of equations in target systems such as Maple or Mathematica. Moreover, the symbol definitions are modelled as properties of a variable node so that the problem of multiple symbol usage is mitigated. Additionally, the category of equation relation is assigned based on the provided input lists. The stored KG finally serves as a flexible interface which can be accessed by using standard APIs and the query language Cypher.

4. Validation

In order to validate the previously described concept, the SMART standards expert system described in (Luttmer *et al.*, 2021) is enhanced by using the automatic formula KG creation introducing advanced functionalities. As example, the standard DIN 281 is used dealing with the dynamic load ratings and rating life of rolling bearings.

https://doi.org/10.1017/pds.2024.45 Published online by Cambridge University Press

4.1. Creation of KG for DIN ISO 281

The standard DIN 281 contains 59 formulas in total. The relevant knowledge was extracted using the previously described extraction solutions and stored in the output lists. Then, the formula KG was created under consideration of the KG ontology and the modelling patterns. In summary, the KG contains 100 nodes: 59 equations nodes, 39 variable nodes including the definitions of symbols as well as one standard node and one ICS node, indicating the technical domain. In terms of relations between equations, 22 relations of type "follows" and 81 relations of type "is_variant_to" were extracted. The extraction of equations was executed with high quality. Only two of 59 equations were not extracted correctly (Recall = 97%) because of a misuse of summation signs in the XML standard. Symbols and symbol definitions were extracted without any errors. This indicates the high quality of DIN ISO 281 where symbols are either defined in a specific section or below the equation. Lastly, 90 out of 103 relations between equations were extracted correctly, leading to a recall of 87%. Here, issues were identified in the extraction of relations that are not described by using keywords and equations that are in different sections of the standards document. Figure 4 shows an exemplary extract of the KG, specifically for the network of equations that are connected to equation no. 4.

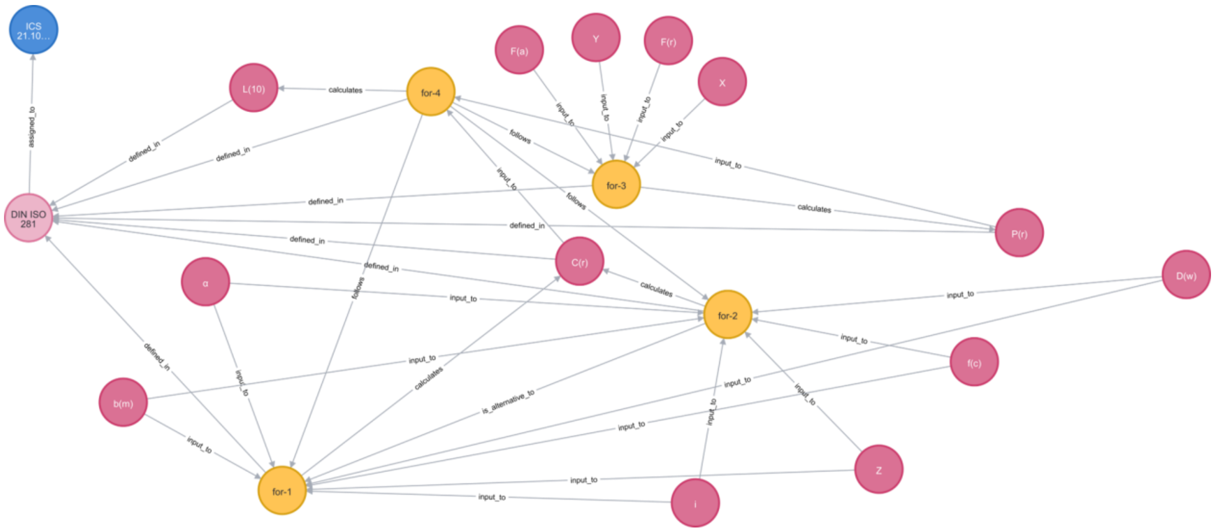

Figure 4. Exemplary representation of KG

4.2. Provision of formulas through user interfaces

The main benefit of using a graph database like Neo4j which is associated with a dedicated query language called Cypher, is the flexible interface to users. Previous work has introduced a chatbot application to support engineers in searching formulas and performing calculations (Luttmer *et al.*, 2021). Because of the limited information that was contained in the previous KG, only limited functionality was provided to the user. With the integration of variables and their definitions as well as relations between equations, users can run advanced queries, e.g., to extract networks of related equations including their logical flow (see Figure 4). Moreover, the chatbot´s user support in designing rolling bearings, which followed a rule-based interaction protocol, is enhanced due to the relations between equations. The chatbot can automatically extract the required equations from the KG and insert the variable values provided by the user. Lastly, formula modules can be automatically exported into mathematical tools, such as Mathematica and Maple, to perform calculations.

5. Discussion of results

Regarding the requirements defined in section 2 and the validation of the concept using the example of DIN ISO 281, the automatic formula KG creation can be proven sufficient. While XML standards are used as input documents - considering NISO STS and MathML standards (Req. 1) -, formulas as well as their descriptive elements are incorporated into a single KG (Req. 2, Req. 8) and can be accessed via standard interfaces (Req. 9). Currently, the application range of the extraction system is limited and will

https://doi.org/10.1017/pds.2024.45 Published online by Cambridge University Press

be further enhanced in future. Moreover, as the overall solution is validated using only one document, the merging of formulas from different documents cannot be evaluated. However, the KG and especially Neo4j offer merge functionalities which will be explored in future (Req. 3 - Req. 4). With respect to Req. 5, the automatic KG creation is proven to be sufficient. In particular, the extraction results using rule-based approaches show that the standards are largely structured according to the identified rules and follow the NISO STS and MathML standards. In addition, the rule-based approaches can be considered as so-called white-box solutions and therefore possess a high degree of understandability and transparency (see requirement Req. 6). For other elements of the formula and variable modules, however, it is necessary to examine the extent to which rule-based approaches lead to satisfactory results. Preliminary investigations show that rule-based approaches reach their limits when considering boundary conditions due to complex sentence constructions and semantically ambiguous descriptions. Further approaches need to be developed, especially for elements that are largely inferred from context. Concerning the automatic addition of relationships through logical inference in the KG (see requirement 7), no statement can be made based on the selected example. This becomes relevant when larger standard groups are considered. Regarding the first research question, it can be concluded that formula KGs can be automatically created from XML standards with a high level of quality.

In terms of extraction and the underlying XML standards, however, it should be noted that the correct decomposition of the extracted formulas into individual variables remains particularly challenging. Although several rules have been implemented - also to compensate for incorrect descriptions in standards - this step is prone to error. This is partly due to the usage of Presentation MathML, which leaves considerable creative freedom. Other formats, such as Content MathML or OpenMath, define semantic relationships in mathematical equations more clearly which would simplify extraction. Another problem is the relationship between equations where no keyword is used and relations result from context or expert knowledge. This is where rule-based approaches reach their limits, but it should be noted that even expert interpretation is not always unambiguous. This also indicates that the design or wording within the standard needs to be improved. These findings are used to answer the second research question, i.e., the implications for the future design of standards. From this research, four rules for future standards can be determined: 1) Create mathematical equations according to fixed patterns considering the by machines, i.e., no use of simplifications optimised for humans; 2) Clearly separate variables and operators in MathML; 3) Clearly indicate relationships between equations in the text above or below the equation; 4) Do not define symbols and definitions more than once in the document.

6. Conclusion and outlook

To provide machine-actionable and machine-interpretable standards content, standards need to be transferred to a semantically enriched format, such as KGs (Loibl *et al.*, 2020; Luttmer *et al.*, 2021). While previous work has demonstrated the overall effectiveness of KGs for standards provision, this paper has investigated the automatic creation of KGs, in particular for formulas and their descriptive elements. Therefore, a concept based on the general KG creation process (Tamašauskaitė and Groth, 2023) was developed and elaborated. This includes the development of a formula KG ontology as well as different rule-based extraction solutions for specific formula elements. The concept was validated using the example of DIN ISO 281 whereas a formula KG was created automatically and stored in the graph database Neo4j. The results confirm the automation potential for formula KG creation.

Future work will focus on the extension of the developed concept. Therefore, the remaining elements of the modules will be considered to develop appropriate extraction solutions. This includes learning-based approaches, especially for elements such as boundary conditions. Additionally, further object like tables and figures will be examined regarding their machine-actionability and possible incorporation into a KG. Ultimately, future work will focus on assessing the practical impact of the automatic KG creation in user studies. This will provide detailed insights into the differences between manual and automatic knowledge extraction from engineering standards as well as quantitative measure such as time and costs.

References

Adnan, K. and Akbar, R. (2019), "An analytical study of information extraction from unstructured and multidimensional big data", *Journal of Big Data*, Vol. 6 No. 1. https://doi.org/10.1186/s40537-019-0254-8

Bender, B. and Gericke, K. (Eds.) (2021), *Pahl/Beitz Konstruktionslehre: Methoden und Anwendung erfolgreicher Produktentwicklung*, 9th ed., Springer, Berlin, Heidelberg. https://doi.org/10.1007/978-3-662-57303-7

Czarny, D.A., Kriegsmann, S., Nagarajah, A., Schacht, M. and Sporer, R. (2022), *Use Cases for SMART Standards*, available at: https://www.din.de/resource/blob/868896/33ab0714368ab3cbb4ebe8614f2b065a/whitepaper-idis-2-en-data.pdf (accessed 15 November 2023).

Dumitrescu, R., Albers, A., Riedel, O., Stark, R. and Gausemeier, J. (2021), *Engineering in Deutschland - Status quo in Wirtschaft und Wissenschaft: Ein Beitrag zum Advanced Systems Engineering*, Paderborn.

Ehring, D., Ferraz-Doughty, P., Luttmer, J. and Nagarajah, A. (2023), "A first step towards automatic identification and provision of user-specific knowledge: A verification of the feasibility of automatic text classification using the example of standards", *Procedia CIRP*, Vol. 119. https://doi.org/10.1016/j.procir.2023.02.183

Ehring, D., Luttmer, J., Pluhnau, R. and Nagarajah, A. (2021), "SMART standards - concept for the automated transfer of standard contents into a machine-actionable form", *Procedia CIRP*, Vol. 100, pp. 163–168. https://doi.org/10.1016/j.procir.2021.05.025

Gräßler, I. and Hesse, P. (2023), "Considering Engineering Activities and Product Characteristics to achieve Material Circularity by Design", *Proceedings of the Design Society*, Vol. 3. https://doi.org/10.1017/pds.2023.108

Hicks, B.J., Culley, S.J., Allen, R.D. and Mullineux, G. (2002), "A framework for the requirements of capturing, storing and reusing information and knowledge in engineering design", *IJIM*, Vol. 22 No. 4, pp. 263–280. https://doi.org/10.1016/S0268-4012(02)00012-9

Huang, Y., Yu, S., Chu, J., Su, Z., Cong, Y., Wang, H. and Fan, H. (2023), "Combining Deep Learning with Knowledge Graph for Design Knowledge Acquisition in Conceptual Product Design", *Computer Modeling in Engineering & Sciences*, Vol. 138 No. 1, pp. 167–200. https://doi.org/10.32604/cmes.2023.028268

Huet, A., Pinquié, R., Segonds, F. and Véron, P. (2023), "A cognitive design assistant for context-aware computer-aided design", *Procedia CIRP*, Vol. 119, pp. 1029–1034. https://doi.org/10.1016/j.procir.2023.03.146

Isaksson, O. and Eckert, C. (2020), *Product Development 2040: Technologies are just as good as the Designer´s ability to integrate them*. https://doi.org/10.35199/report.pd2040

Layer, M., Neubert, S., Tiemann, L. and Stelzer, R. (2023), "Identification and Retrieval of relevant Information for instantiating Digital Twins during the Construction of Process Plants", *Proceedings of the Design Society*, Vol. 3, pp. 2175–2184. https://doi.org/10.1017/pds.2023.218

Loibl, A., Manoharan, T. and Nagarajah, A. (2020), "Procedure for the transfer of standards into machine-actionability", *JAMDSM*, Vol. 14 No. 2. https://doi.org/10.1299/jamdsm.2020jamdsm0022

Luttmer, J., Ehring, D., Pluhnau, R., Kocks, C. and Nagarajah, A. (2022), "SMART Standards: Modularization Approach for Engineering Standards", in *Proceedings of ASME 2022 IDETC-CIE* 2022, The American Society of Mechanical Engineers, New York, N.Y. https://doi.org/10.1115/DETC2022-88206

Luttmer, J., Ehring, D., Pluhnau, R. and Nagarajah, A. (2021), "Representation and Application of Digital Standards using Knowledge Graphs", *Proceedings of the Design Society: ICED21*, Vol. 1, pp. 2551–2560. https://doi.org/10.1017/pds.2021.516

Luttmer, J., Prihodko, V., Ehring, D. and Nagarajah, A. (2023), "Requirements extraction from engineering standards – systematic evaluation of extraction techniques", *Procedia CIRP*, Vol. 119, pp. 794–799. https://doi.org/10.1016/j.procir.2023.03.125

Manoharan, T., Loibl, A., Nagarajah, A. and Köhler, P. (2019), "Approach for a Machine-Interpretable Provision of Standard Contents Using Welded Constructions as an Example", *Proceedings of the Design Society: ICED 19*, Vol. 1 No. 1, pp. 2477–2486. https://doi.org/10.1017/dsi.2019.254

Noy, N.F. and McGuiness, D.L. (2001), *Ontology Development 101: A Guide to Creating Your First Ontology*.

Schubmehl, D. and Vesset, D. (2014), *The Knowledge Quotient: Unlocking the Hidden Value of Information Using Search and Content Analytics*.

Schuch, W. and Wischhoefer, C. (2018), "Abschluss des Projektes XML100", *DIN Mitteilungen*, 2018.

Siddharth, L., Blessing, L.T.M., Wood, K.L. and Luo, J. (2022), "Engineering Knowledge Graph From Patent Database", *J. Comp. Inf. Sci. Eng.*, Vol. 22 No. 2. https://doi.org/10.1115/1.4052293

Tamašauskaitė, G. and Groth, P. (2023), "Defining a Knowledge Graph Development Process Through a Systematic Review", *ACM Trans. Softw. Eng. Methodol.*, Vol. 32 No. 1, pp. 1–40. https://doi.org/10.1145/3522586

Zhao, H., Pan, Y. and Yang, F. (2020), "Research on Information Extraction of Technical Documents and Construction of Domain Knowledge Graph", *IEEE Access*. https://doi.org/10.1109/ACCESS.2020.3024070

Zhenyong, W., Xinguo, M., Lina, H. and Goh, M. (2020), "Product Development-Oriented Knowledge Service: Status Review, Framework, and Solutions", *IEEE Access*, Vol. 8, pp. 64442–64460. https://doi.org/10.1109/ACCESS.2020.2984631