Available online at www.sciencedirect.com

ScienceDirect

Procedia CIRP 136 (2025) 73–78

www.elsevier.com/locate/procedia

35th CIRP Design 2025

Automated Information Extraction for Adaptation Design of Mechatronic Systems

Marc Behringer[a,]*, Stefan Goetz[a], Sandro Wartzack[a]

[a]*Engineering Design (KTmfk), Friedrich-Alexander-Universität Erlangen-Nürnberg (FAU), Martensstraße 9, Erlangen 91058, Germany*

* Corresponding author. Tel.: +49 9131 85-23659; fax: +49 9131 85-23223. *E-mail address:* behringer@mfk.fau.de

Abstract

Mechatronic product development is characterized by its interdisciplinary nature, requiring the linking of the various associated domain-specific information. Especially in adaptation design, the extensive reuse of existing information is reasonable. However, accessing and acquiring this information usually requires high manual effort. Thus, a novel approach for the automated extraction of information from various sources is presented. In order to achieve this, the establishment of a defined list of potential information elements provides the foundation. In light of the preceding analysis the sources of information pertaining to the aforementioned elements were identified and furnished with suitable extraction methods in accordance with the documentation form and type. This facilitates the automatized extraction and structuring of the information. Semantic processing provides the developer with the information in a comprehensible way, thus contributing to a reduced development effort and error avoidance. Finally, a conceptual graphical user interface is described to illustrate the application of the developed approach.

Keywords: Mechatronic Systems; Adaptation Design; Information Extraction

1. Introduction

The advent of digitalization has rendered data and information as an indispensable and potent commodity. It is particularly during the product development cycle, from the initial planning stages to the end of life, that information included in a considerable amount of product documentation are generated [1]. This documentation is created in various domains and divided among numerous individuals across a diverse range of contexts [2]. It is therefore necessary to comprehend the data in order to utilize the information it contains in a meaningful and targeted manner [2]. This offers significant potential for the adaptation of existing designs, optimizing processes and generating new insights. The analysis of the predecessor product and the acquisition of the requisite information represent a fundamental building block [3]. The value of this step can be considered as high and of crucial importance for the further course of the process. Nevertheless, it has been demonstrated that this potential is frequently unexploited [4]. The reasons for this are manifold. Firstly, the standardized manual processes that were previously sufficient for coping with the abundance of data in an industrial context are no longer viable in the context of rising costs and growing volumes of information [5]. Secondly, there is often a lack of background knowledge or explanations for a classification and linking. Consequently, there is an increased probability of erroneous outcomes, including misinterpretation.

2. State of the art

Mechatronic systems are pervasive in numerous fields and are characterized by a basic system to be controlled or regulated, an information processing system that employs sensors to record certain information from the basic system and the environment, and an actuator system that executes actions based on a defined behavioural logic [6]. Accordingly, a mechatronic system is defined by the integration of three distinct disciplines: mechanics, electronics, and information technology. The associated information is thereby stored in various domain-specific documents. Hira et al. [7] address the

2212-8271
10.1016/j.procir.2025.08.015

challenges of information extraction from published literature, including books, journal articles, and patents. In particular, they encounter a number of challenges in the context of machine learning, specifically with regard to the diverse formats and structures of the reporting style [7]. The process of extracting information represents a significant challenge in a multitude of fields. The research conducted by Jeblee et al. [8] investigates the automated extraction of pertinent data from medical records based on of machine learning techniques. Hong et al. [9] have investigated the extraction of pivotal information from documents comprising text in disparate formats, employing the pre-trained language model BROS (BERT Relying On Spatiality). In the field of technical product development, Kestel et al. [10] investigate methods for abstracting essential simulation knowledge from existing simulation models and text-based documents through the use of text and data mining techniques. Rahul et al. [11] address the automated extraction of information from pipework and instrumentation diagrams and the extraction of components and their relationships using deep learning techniques. In contrast, Wu et al. [12] investigate the issue of rule-based information extraction, with a particular focus on texts pertaining to the mechanical, electrical, and sanitary domain. In the context of improving information retrieval and the associated obtaining, semantics can represent a significant factor. Wei et al. [13] investigated the extraction of pertinent product features from customer reviews using semantics for refinement.

3. Research Need

As demonstrated in the preceding chapter, a number of approaches have already been developed for the purpose of obtaining and extracting domainspecific information. Nevertheless, these methodologies are situated within disparate domains, rendering direct transfer impractical. In particular, the initial analysis of a predecessor design, in conjunction with the necessary information gathering is an extensive and time-consuming process. However, the value of this analysis can be considered to be of significant merit in the adaptation design, forming an indispensable foundation for the identification of design-determining attributes [3]. This leads to the following research question:

How can the automated extraction and conversion of the relevant data from mechatronic products into usable semantic information be realized?

The following chapter presents a novel method for the acquisition and extraction of information from mechatronic systems. This approach addresses the identified challenges through automation and semantic endorsement, thereby providing sustainable support for the user. Subsequently, a demonstrator is provided, in order to illustrate the practical applicability of the proposed methodology. Finally, the paper closes with a conclusion and an outlook.

4. Method for Information Extraction

The following chapter presents a conceptual description of the proposed method, which is comprised of six steps (see Figure 1).

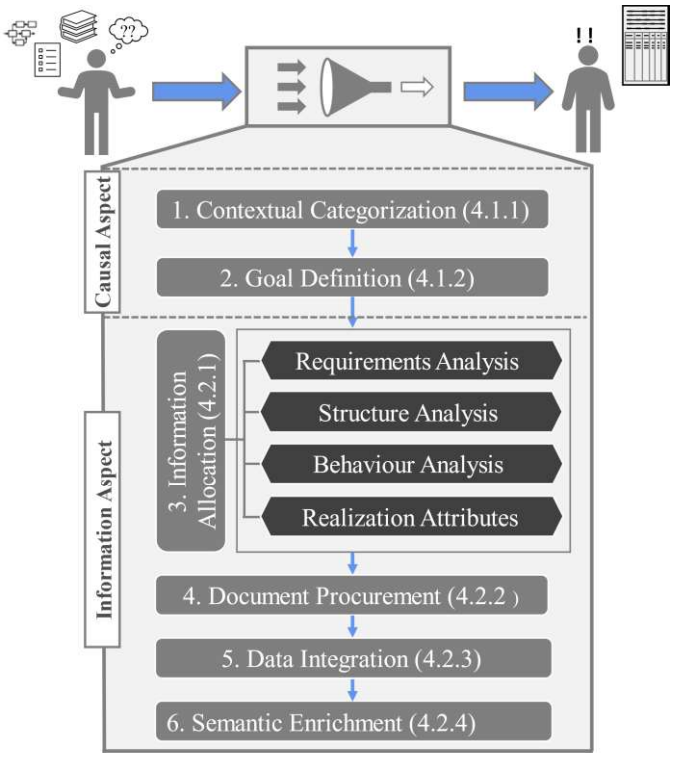

Fig. 1. Procedure of Information Extraction

To gain a comprehensive understanding of the overall context, it is necessary in the causal aspect to convey the system context and focus on the target state. Once the conceptual framework has been clarified, the subsequent chapter 4.2, will address the object of investigation in detail. Therefore, it is first essential to identify and collate the relevant information of mechatronics in the information allocation process. Subsequently, it is required to investigate and compile the relevant sources of information, given that this is often distributed across a wide range of locations. In order to ensure that the variety of forms is adequately addressed, appropriate extraction methods for the respective information source have to be defined. Ultimately, semantic enrichment is employed to enhance the quality of the results and facilitate user interaction.

4.1. Causal Aspect

The development of mechatronic systems is typically associated by a considerable investment of project effort. This is due to the high level of complexity, the integration of the system into one or more environmental systems, and their subsequent interaction. In order to emphasise the relevance of these aspects, 4.1.1 provides a categorization in the overall context, followed by an analysis of the objectives in 4.1.2.

4.1.1. Contextual Categorization

With the intention of exerting a positive influence on the subsequent design and its comprehensibility, the first step is the categorization of the situation or object in the overall context. It can be observed, in general, that a different background usually results in a different perception of the context, which can lead to errors. In order to reduce subjectivity and create a standardized basis, the widely used formal semantic dictionary ECLASS is employed as a general basis [14]. The data set is

integrated into the background in order to facilitate a description and classification of the object. The level of detail is augmented by a predefined class hierarchy with increasing depth. Given the novelty of this data set in the context of Industry 4.0, it is not yet complete. Furthermore, the definition of the application domain is beneficial in the subsequent steps, such as the definition of initial identifiers or the resolution of ambiguities in the context of formulation recognition.

4.1.2. Goal Definition

In examining the process adaptation design, it becomes evident that in most instances a particular element serves as the initial drivers and motivators. In the case of consumer goods, the design is the primary focus; in contrast, for capital goods, it is the technology that takes precedence [3]. It is therefore essential at the outset of the process to define the reasons and the desired outcome, given that further decisions will inevitably influence the final result. In order to ascertain whether a desired mathematical optimization is achievable, it is imperative to ascertain whether the target definition meets the criteria for verifiability. For this purpose, an analysis of the principal parameters of the target formulation is conducted, which is then classified according to the presence of common characteristics of quantity, such as assurance or values. The parameters included are also of particular significance with regard to the final evaluation and validation.

4.2. Information Aspect

In the initial phase of the adaptation, there is typically a confrontation with a substantial quantity of data, the appropriate utilization of which may not be immediately apparent. This can be attributed to the diversity of the areas involved and the lack of application relevance, which can give rise to considerable difficulties in the remainder of the project. The implementation of an automated data capture and extraction process is advantageous in this context. To circumvent this issue, an automated information collection and extraction process is necessary. To ensure this, four consecutive points need to be elaborated. The chapter 4.2.1 commences with the identification of all pertinent information. The second section of this chapter 4.2.2 is dedicated to the subject of procurement, with the aim of further elaborating on the principles previously outlined. Following, the process of data integration is demonstrated (4.2.3). Subsequently, chapter 4.2.4 addresses the topic of semantic enrichment.

4.2.1. Information Allocation

Considering mechatronic systems and significant amount of information, it is essential to gain a comprehensive understanding of the available information and encompassing its full diversity. Of particular interest are characteristics, as they can be defined by the product developer. At present, there is no comprehensive overview that can be readily applied. A number of steps were undertaken in order to create one. An initial guide is provided by VDI 2206 [6], which includes an integrated list of main characteristics. However, the level of abstraction is too high and the categories are too general. In order to focus these main characteristics on the actual tangible elements, detailed responses to the pivotal questions regarding the principal characteristics were derived. This preliminary list was subsequently refined in relation to the consistently present elements of mechatronic systems, including actuators and information processing. In order to encompass the full spectrum of mechatronic systems, the comprehensive list of additional attributes was compiled from the specialist literature and illustrative examples. The predefined list with the attributes of interest serves as a base for the following steps and is for a comprehensive overview divided into four areas (Figure 2).

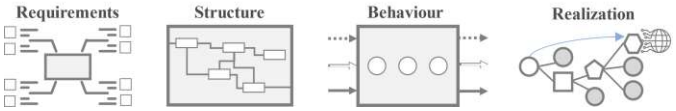

Fig. 2. Division of Information

Requirement's Analysis

Requirements constitute the foundation of subsequent considerations. In the context of adaptation design, this refers exclusively to quantitative changes to requirements, as the incorporation of additional ones would alter the fundamental solution [15]. The number of requirements inherent to complex projects can be considerable. In order to filter the relevant content and facilitate the automatic transfer of requirements and, it is essential to conduct a thorough analysis of both structure and semantics. In practice, requirements documents are often textual artefacts with an implicit structure. To ascertain the benefit or relevance with a view to subsequent optimization, a requirement classification (Figure 3) based on NLP and inspired by the criteria established by Horber et al. [16] is employed.

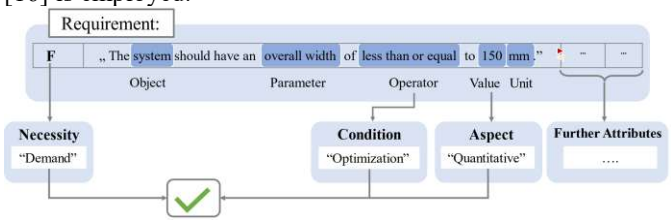

Fig. 3. Requirement classification

The classification is thereby based on the fundamental information from requirements data. In accordance with the specific available attributes, this may be expanded to encompass supplementary attributes, such as those pertaining to safety or security. Furthermore, relationships exert a considerable influence on subsequent advancement, as their accessibility can be leveraged to mitigate dependencies and the impact of alterations. When available, prospective change requests can be discerned from the relationships through the utilization of semantic data, thereby limiting the propagation of changes [17]. The requirements from the function model are operationalized on the hardware side by the structure model and on the software side by the behaviour model.

Structure Analysis

Mechatronic systems can considerably vary in their form and structure. Observations can be undertaken at a higher system level or the component level, in which only the pertinent details are addressed. The system is dismantled into its individual components and relations in accordance with the

guidelines set forth in VDI 2206 [6]. The fundamental structure is subdivided into the three principal domains of mechatronic systems. The basic system is examined within the context of mechanical engineering. The shape model is of particular significance in this context. The configuration of a component, an assembly, or a machine can be modified through the deliberate variation of primary functional elements, such as shape or position [3]. From a technical standpoint, the most crucial topics are primarily the design-determining main functional carriers, namely the modules and components that represent the core functionality of the product [3]. In the domain of electronics, the information elements as well as coherences are included for sensors, actuators as well as for power processing in the field of information technology.

Behavior Analysis

While the structure analysis focused on the composition, the behaviour analysis now concentrates on process flows and the interaction of the components. A correlation exists between the structural and behavioural aspects, which allocates variables to designated components as well as processes specific functional units. From one perspective, it is enlightening to examine the system in terms of its input variables, environmental exchange, and conversion to output variables. Conversely, it is also of interest to examine the behaviour and interaction of mechanical, electrical and software components in order to ensure the fulfilment of the principal function. In this instance, it is of particular importance to consider the functional and dynamic aspects. Certain attention is paid to time-dependency. In the context of actuators, the factor of movement as a function of time frequently represents a critical determining factor.

Realization Attributes

In addition to the product-classifying information, the framework conditions describe other aspects that exert an influence on the design. Among other things, it is significant to consider the external influences. Such occurrences may manifest in a variety of forms (predictability, temporal existence, etc.) and are typically evident during the realization phase. In this context the design rationale, as documented, provides relevant information. Design rationales are of particular interest to future generations of product developers, as they offer valuable insights into the rationale behind a product's design. This may result in a reduction in the effort required for future developments [15].

4.2.2. Documentation Procurement

A product typically progresses through a series of stages, from initial planning to the conclusion of its product life cycle. In the initial stages of the design process, engineers produce preliminary design documentation, which is typically expressed in abstract terms. As time progresses, there is an observable increase in the level of detail specified in product design phases, which is also reflected in the documents or models. An abstract representation of an object (type) is initially developed, from which a concrete realization of a type (instance) is ultimately derived. Information of relevance is stored thereby in a variety of information sources or formats. The range of document types is extensive, encompassing everything from one-dimensional text-based documents to three-dimensional volume models [18]. It is thus imperative to ascertain the potential sources of the previously identified information elements. The performed categorization is based on a literature analysis and existing development documents. The various boundary parameters are considered in the most optimal manner possible, with the intention of ensuring their continued realistic future utilization. The objective is to examine the ongoing applicability and utility of extant documents and models, including scaling them for different mechatronic systems. The evaluation criteria encompass a range of elements, including the degree of abstraction, the degree of formalization and the level of detail, the extent of interoperability between universally valid or proprietary file formats. It is preferable to have a low level of abstraction, a high level of formalization and a high level of detail. In light of the clearly delineated information sources, the methodologies for extraction are set forth below.

4.2.3. Data Integration

Cross-domain collaboration represents a fundamental element of mechatronic systems. The utilization of specialized tools and bespoke documentation is typical within a given domain, thereby enhancing a variety. The purpose is not to supplant these, but to establish a unified vision and interconnectivity within a system. A principal objective in this context is the formulation of a comprehensive information model, which serves as the foundation for subsequent modelling and further utilization. This requires data to be converted into usable information (Figure 4). The initial stage of the process is to **analyse the data set**, with particular attention paid to data format types and included structure. In this context, the primary focus is on images, natural language text, Extensible Markup Language (XML) and JavaScript Object Notation (JSON) files, tables, databases and boundary representations. In the majority of instances, the information elements are not readily accessible and is instead embedded within a variety of data formats, which contain a varying structure. The data may be presented in one of three forms: structured, unstructured, or semi-structured. In particular, unstructured data accounts for over 80% of all data [19]. The term "unstructured data" is used to describe information that does not adhere to a specific structure. In contrast to unstructured data, semi-structured data differs in particular with regard to the manner of its structuring and the formal semantics that it employs. The defining characteristic of semi-structured data is the clear delineation between form and structure through a level of abstraction, wherein text fragments are enclosed by declarative labels. In contrast, structured data is characterized by a fixed structure, comprising rows and columns. This has the particular advantage of facilitating analysis and enabling automated processing due to the evident data relationships.

The extraction methods are contingent upon the data formats with contained structures, which vary in their composition. The automatic extraction of key information necessitates the integration of multiple technical components from the domains of both computer vision and NLP [9]. A suitable extraction method was allocated to each previously defined element of the generated list, based on the assigned information source from the data set. Due to the different data formats with contained

structures, the application of the extraction methods differs. Subsequently, the important extraction methods employed are outlined. In particular, during the course of development, physical documentation is still prevalent. Subsequent processing necessitates the conversion of physical documents into a machine-processable format. This is achieved through the utilization of *Optical Character Recognition (OCR)* technology. To accomplish this, the text and content are extracted from an image, which needs to be captured, resulting in the generation of discrete text blocks in a machine-readable format. This approach allows for the creation of document text devoid of contained images, tables, or equations. Furthermore, a considerable amount of information is expressed in rich and ambiguous natural language [19]. In order to automate the processing and extraction of information using machine learning techniques, it is necessary to have a formal understanding of the underlying structures. Using *NLP techniques*, natural language texts are processed, enriched with formal linguistic knowledge [2]. The process commences with a lexical analysis, followed by an automated syntactic analysis and concluded by a semantic analysis. To then automatically extract and structure knowledge from these textual artefacts, computer linguistic and statistical methods from the field of *text mining* are adapted. Using text mining and defined templates, the relevant data consisting of content and metadata can be extracted. The transition from document-centred development to model-based development makes the usage of models increasingly prevalent. In this context, semi-structured data in form of *XML and JSON* files is of greater significance, from which the requisite information was then extracted into the desired format with the assistance of parsers bearing the same designation. As data is available in tables and SQL databases, a form of *structured data parsing* can now be used to extract the required information with the corresponding line information. The utilization of structured data thus enables the application of *data mining* techniques to ascertain and extract superordinate relationships and patterns, otherwise known as meta-models, from databases. This can be achieved through the deployment of machine learning, pattern recognition and statistical methods [10].

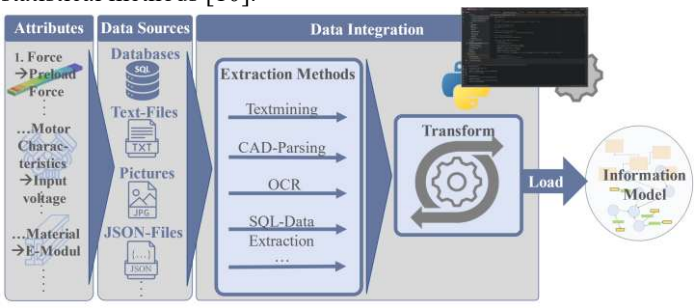

Fig. 4. Information Extraction method

Further models that are frequently utilized are for example CAD models and circuit diagrams. In particular, the CAD model is of great significance with regard to the adaptation design and the shape. The data of the model is for further utilization transferred to a subordinate dimension as part of the extraction process. The *CAD parser* reads the objects, their hierarchical relationships and existing documented design alterations and converts the object information into a text file. An attribute of a CAD object is created in accordance with the rules for text generation as outlined in [20].

Having delineated the extraction methods, we proceed with going into the **transforming of the extracted content**. The intention is to enhance the formal semantics through the defined structured arrangement. A template is provided within the demonstrator for the specification of the structure or process for all the information. The highest level of patterning occurs within the attribute's component, sub-model element, and value. The component is defined in accordance with the prescribed method catalogue and in relation to the extracted document. The sub-model element attribute serves to define the relationships or properties of the components. The data element property area is further classified into constants, parameters and variables in consideration of their temporal variability, while the relationship element contains the references. The 'value' attribute comprises the following elements: the attribute statement (assurance, measured value, etc.), the statement logic (less than, equal to, greater than), the numerical value, the type (integer, float, etc.) and the unit for categorization. Once the information has been structured, it can then be **transferred and utilized for further structuring and modelling** activities within the context of the information model.

4.2.4. Semantic enrichment

In the domain of data processing and extraction, an understanding of the subject matter is of fundamental importance, as the utility of the data is contingent upon its contextualization. The addition of supplementary attributes in the form of semantic enrichment provides data with its intrinsic meaning. This facilitates the exchange of understanding, increases knowledge between communication partners and reduces effort. In order to achieve this, the ECLASS database with its standards is integrated in the background in particular to enrich characteristics and attributes [14]. Semantic descriptions are employed to facilitate the semantic integration of disparate aspects (e.g. geometry, functions) of the technical design [15]. For example, an erroneous assessment of the unit can give rise to irritation and misinformation, which can have far-reaching consequences.

5. Prototypical Implementation

An assistance system was developed using Python to illustrate the practical application of the method for users and to facilitate their comprehension. The process is exemplified with the use of a 3D-printer. The object name must first be entered into the demonstrator, from which the hierarchical categorization and definition are determined (see Figure 5). Subsequently, the objective must be delineated. In this instance, the objective is to achieve optimal print accuracy in the z-direction, with a precision of 0.1 mm per layer. Based on the upload of the requirements source a first analysis containing the classification of the requirements from chapter 4.2.1 is pursued and in form of the structured core object points visualized to the user in form of a table. To initiate the process of filtration and extraction of fundamental attributes, an input of all pertinent documentation by the user is required.

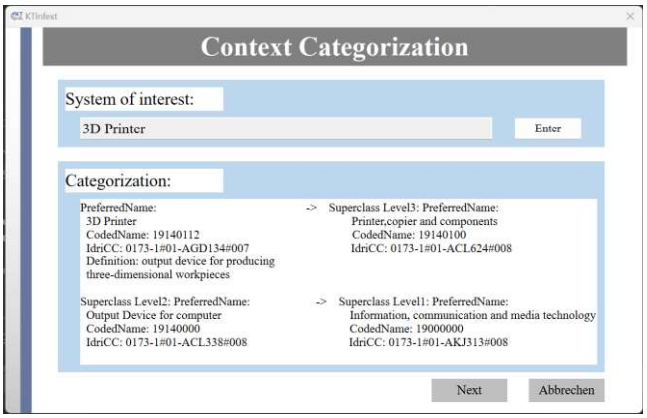

Fig. 5. Context Categorization step in the Assistance system

This will start a verification process in the background followed by an automatically extraction of all elements from the deposited list. The maximum print height represents a key factor in determining the target function. In order to facilitate this, the list defines an information storage location that is preferred as the technical specification. As the information for this is embedded in unstructured text written in natural language, the attributes are extracted using text mining. Afterwards the extracted data is converted into the pre-defined structure defined in 4.2.3 until it is semantically enriched with additional attributes from the E-Class Database. The information, can now be further utilized and employed as a foundation for structuring and modelling. It enabled the user to extract the relevant information from the data sets in a preferably automatized manner and thus achieving a time reduction while avoiding errors. The potential of this method is particularly evident when dealing with various different document types created by different persons.

6. Conclusion and future work

Considering the current limitations of data integration in the context of mechatronic systems, a novel approach was presented. The method addresses the primary challenges of information gathering and the considerable amount of manual time required. The high degree of automation is achieved through the utilization of automatized processes within the information procurement process. This considerably reduces the necessary exertion. Furthermore, the user is assisted by a variety of resources, including help documents like a collection of information elements, pre-existing extraction methods, and semantic enrichment. The procedure could be further elucidated by the assistance system, which serves the application for the user.

The method represents a further advance towards the provision of efficient and automated knowledge support throughout the development process. Nevertheless, further refinements and extensions are required. One area for further investigation is the modelling of extracted information in a meaningful way across domains. This should ultimately result in inter alia a knowledge representation, e.g., Ontologies, that is as automatized as possible. This will enable the exploration of the solution space using model-based, rule-based and use case-based reasoning techniques.

Acknowledgements

Funded by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) – 520437130 (WA 2913/58-1).

References

[1] VDI 2221. 2019. Design of technical products and systems Model of product design.
[2] Dengel A. (Hrsg.). Semantische Technologien: Grundlagen - Konzepte - Anwendungen. Heidelberg: Spektrum, Akad. Verl. 2012.
[3] Naefe Paul, Luderich Jörg. Konstruktionsmethodik für die Praxis: effiziente Produktentwicklung in Beispielen, Lehrbuch. Wiesbaden: Springer Vieweg. 2016.
[4] Abramovici M., Lindner A. Providing product use knowledge for the design of improved product generations. In: CIRP Annals Vol. 60 (2011). 1rd ed. p. 211–214.
[5] Fay Alexander, et al. Semantische Inhalte für Industrie 4.0: Modellierung technischer Systeme in kollaborativen Umgebungen. In: atp magazin Vol. 59 (2017). 07–08rd ed. p. 34–43.
[6] VDI/VDE 2206. 2021. Development of mechatronic and cyber-physical systems.
[7] Hira Kausik, et al. Reconstructing the materials tetrahedron: challenges in materials information extraction. In: Digital Discovery Vol. 3 (2024). 5rd ed. p. 1021–1037.
[8] Jeblee Serena, et al. Extracting relevant information from physician-patient dialogues for automated clinical note taking. In: Proceedings of the Tenth International Workshop on Health Text Mining and Information Analysis (LOUHI 2019). Hong Kong: Association for Computational Linguistics. 2019. p. 65–74.
[9] Hong Teakgyu, et al. BROS: A Pre-trained Language Model Focusing on Text and Layout for Better Key Information Extraction from Documents. In: Proceedings of the AAAI Conference on Artificial Intelligence Vol. 36 (2022). 10rd ed. p. 10767–10775.
[10] Kestel Philipp, et al. Ontology-based approach for the provision of simulation knowledge acquired by Data and Text Mining processes. In: Advanced Engineering Informatics Vol. 39 (2019). p. 292–305.
[11] Rahul Rohit, et al. Automatic Information Extraction from Piping and Instrumentation Diagrams, arXiv (2019).
[12] Wu Lang-Tao, et al. Rule-based information extraction for mechanical-electrical-plumbing-specific semantic web. In: Automation in Construction Vol. 135 (2022). p. 104108.
[13] Wei Chih-Ping, et al. Understanding what concerns consumers: a semantic approach to product feature extraction from consumer reviews. In: Information Systems and e-Business Management Vol. 8 (2010). 2rd ed. p. 149–167.
[14] Belyaev Alexander, et al. Modelling the Semantics of Data of an Asset Administration Shell with Elements of ECLASS. 2021.
[15] Dworschak Fabian, et al. Model and Knowledge Representation for the Reuse of Design Process Knowledge Supporting Design Automation in Mass Customization. In: Applied Sciences Vol. 11 (2021). 21rd ed. p. 9825.
[16] Horber D., Schleich B., Wartzack S. Conceptual model for (semi-) automated derivation of evaluation criteria in requirements modelling. In: Proceedings of the Design Society: DESIGN Conference Vol. 1 (2020). p. 937–946.
[17] Goknil Arda, et al. Change impact analysis for requirements: A metamodeling approach. In: Information and Software Technology Vol. 56 (2014). 8rd ed. p. 950–972.
[18] Pickel Jessica, et al. Integration of product development data for further ontological utilization. In: Proceedings of the Design Society Vol. 4 (2024). p. 463–472.
[19] University of Liechtenstein, et al. Text Mining for Information Systems Researchers: An Annotated Topic Modeling Tutorial. In: Communications of the Association for Information Systems Vol. 39 (2016). p. 110–135.
[20] Jeon Sang Min, et al. Automatic CAD model retrieval based on design documents using semantic processing and rule processing. In: Computers in Industry Vol. 77 (2016). p. 29–47.