SciBite
an ELSEVIER company

A CURATED COLLECTION

Enriching the information landscape in materials science

Developments in shifting data science workloads from data preparation to data analysis

A SPECIAL COLLECTION FOR YOU

Elsevier and SciBite couple extensive and deep content with cutting-edge semantic technology and services. Together we shift your burden of data cleaning, organizing and annotating to the benefit of improved scientific interpretation. By transforming full-text and open-source data and making it accessible and interoperable, we enrich the information used across the research landscape - ensuring solid data foundations for AI-based modelling or knowledge graph construction, irrespective of scientific domain. See how we can help in building domain specific vocabularies.

Featured Letter: Materials graph ontology, Materials Letters, Volume 295, 2021
— Sven P. Voigt a, Surya R. Kalidindi

A product lifecycle management methodology for supporting knowledge reuse in the consumer packaged goods domain, Computer-Aided Design, Volume 43, 2011
— Enrico Vezzetti, Sandro Moosa, Simona Kretli

A knowledge graph method for hazardous chemical management: Ontology design and entity identification, Neurocomputing, Volume 430, 2021
— Xue Zheng, Bing Wang, Yunmeng Zhao, Shuai Mao, Yang Tang

We hope you enjoy this collection! Contact us to discuss the application of these technologies at your company.

Julien Debeauvais
Global Head of Sales and Partnerships, SciBite

Contact us: https://www.scibite.com/contact-us

Materials Letters 295 (2021) 129836

Contents lists available at ScienceDirect

Materials Letters

journal homepage: www.elsevier.com/locate/mlblue

Featured Letter

Materials graph ontology

Sven P. Voigt [a], Surya R. Kalidindi [a,b,c,*]

[a] School of Materials Science and Engineering, Georgia Institute of Technology, Atlanta, GA 30332-0245, United States
[b] George W. Woodruff School of Mechanical Engineering, Georgia Institute of Technology, Atlanta, GA 30332-0405, United States
[c] School of Computational Science and Engineering, Georgia Institute of Technology, Atlanta, GA 30332, United States

ARTICLE INFO

Article history:
Received 18 January 2021
Received in revised form 17 March 2021
Accepted 4 April 2021
Available online 9 April 2021

Keyword:
Artificial intelligence

ABSTRACT

To maximize the use of the materials data being generated by various researchers and organizations, it is necessary to store the data such that it is findable, accessible, interoperable, and reusable (FAIR). Although current materials data repositories and databases partly address the FAIR principles, they do not adequately capture the critical metadata that represents the contextual information (e.g., relationship between materials data and terms typically used by materials scientists such as process, structure, and property). The collection and organization of this metadata along with the original data would allow advanced queries that implicitly improve FAIR characteristics. Recent work has attempted to define this necessary metadata through the development of materials ontologies. This paper introduces a new materials graph and develops the associated materials graph ontology needed to address shortcomings of the current materials ontologies. This novel ontology can be combined with existing ontologies to standardize the inter-relationships between materials data elements and related materials concepts. This paper demonstrates how the proposed materials graph ontology enables the conceptual description of a broad variety of materials data, improves the findability and usability of the different graph-connected material concepts and data, and formalizes a materials data ingest framework that is amenable for the extraction of process-structure–property relationships.

1. Introduction

An astonishingly increasing amount of data is being generated in the materials science field due in part to the advent of novel high throughput experimental assays, increased computational power for simulations, and national initiatives such as the Materials Genome Initiative (MGI) [1]. Leveraging this data in materials development efforts requires the design and deployment of a suitable data infrastructure [2,3] that allows the addition of the critical metadata needed to interpret the data correctly. For example, a microstructure image sitting on a remote server has very limited utility without the proper context. The metadata should include information describing the image, the type of image, relationships between this microstructure image and the material it describes, what/how additional data is derived from the microstructure image, what other information is available about the material structure at different length scales, or how the microstructure may change if the material is subjected to a new processing step. In this paper, all of the metadata about a materials data point and how it may be connected to other related data points will be collectively captured using a suitably defined materials ontology [4,5]. Finding connected materials data not only improves FAIR characteristics, but is also an essential step towards extracting process-structure–property (PSP) surrogate models needed to drive materials innovation [6,7]. Identifying related datasets or finding existing models that could be applied to a new dataset would dramatically improve the re-use of existing materials datasets, and is likely to produce significant cost and time savings in materials innovation efforts. Using a properly designed materials ontology can address this critical need.

Databases have been employed by the materials research community to realize some of the FAIR goals [8–11]. Some databases, such as the Materials Project [12], Automatic Flow for Materials Discovery [13], Materials Data Facility [9], and database technologies, such as Automated Interactive Infrastructure and Database for Computational Science [14], add indexable materials data directly to the database to allow users to quickly search by commonly used terms by materials specialists (e.g., attributes related to structure, property, and process). As a specific example, materialsproject.org [12] allows a researcher to search by elements, chemical formula, id, crystallographic information file (CIF), or an mpquery entry that lets users search over arbitrary database keys. All of the related

* Corresponding author at: School of Materials Science and Engineering, Georgia Institute of Technology, Atlanta, GA 30332-0245, United States.
E-mail address: surya.kalidindi@me.gatech.edu (S.R. Kalidindi).

https://doi.org/10.1016/j.matlet.2021.129836
0167-577X/

terms used in the database (i.e., database keys) have precise meanings, as established by the database schema. The only exception is mpquery, which allows any arbitrary database query string. The database schema, therefore, plays an important role in adding the contextual information to the data, improving its FAIR characteristics. Ontologies go beyond database schemas to enrich the metadata by including many features of language, such as subjects, predicates, objects, synonyms, etc., to describe the contextual information with desired precision. Ontologies are particularly known for their ability to describe complex heterogeneous information and integrate disparate data sources [15]. As such, ontologies are expected to play an important role in improving the FAIR characteristics of materials data, by connecting the typically dispersed data among the multiple databases and repositories.

Materials data is also stored and shared through data repositories. For example, the NIST materials data repository [9] allows searching for records by community collections, author, subject, title, date issued, and whether the record has associated files or not. However, materials specific information contained in repositories' files is inaccessible and cannot be searched. Further, these repositories do not enforce a schema; the files stored in repositories can include heterogeneous data in any format, limiting their interpretability and utility to anyone other than the original creator. Some data standards such as CIF [16], the chemical markup language [17], and Universal Spectroscopy and Imaging Data [18] aim to standardize file formats and address this problem. However, this standardized metadata has not yet been implemented as part of a schema or a materials ontology. Additionally, the materials repositories often save the materials data in zip or hdf5 format, which capture heterogeneous files in a hierarchical file structure. However, file structures can at best imply relationships, but these could easily be misinterpreted or could be sufficiently ambiguous hindering the re-usability of the data. A materials ontology using clearly defined terminology that can be accessed by a variety of software and software platforms would precisely capture the connections between materials data, and dramatically improves the FAIR characteristics of data in comparison to the typical schemas found in current materials databases and repositories. Furthermore, the emergent tools in AI reasoning and web ontologies can be leveraged in developing and deploying such a materials ontology.

2. Knowledge representation and ontology

Knowledge representation (KR) is a subfield of artificial intelligence concerned with the digital representation of human understanding. KR is a model of the real world and captures as much of reality as is needed for reasoning and inference [19]. Reasoning differentiates knowledge and data, where reasoning allows inferring new connections between data points based on rules [20], as depicted in Fig. 1. An ontology provides a vocabulary to define concepts and relationships [21], which can be used for KR. Ideally, an ontology should be designed to be capable of defining any concepts and rules needed to describe the real world; particular forms of ontology should be designed to represent the features most relevant or important to the specific field of implementation.

Ontologies are formally prescribed using ontology modeling languages, which take inspiration from *logics* (i.e., systems of logic) in which mathematical axioms are used to define rules and inference is computed within an established framework. First-order logics allow axioms to be declared with prepositions, which allow making statements about any number of variables. Second-order logics allow prepositions to make statements about other prepositions, but do not see use in ontology modeling languages. Instead, all logical statements must be mapped to first-order logic, if possible. Among the various ontology modeling languages, description logics (DL) [22] and the web ontology language (OWL) [21] restrict ontologies to decidable first-order logic statements, where decidable means that there exists a method to prove a new statement from existing statements (i.e., infer additional information). Decidable first-order logics constitute only a subset of all first-order logics. DL constructs ontologies using *concepts* (usually entities such as a Material or a Process or a Structure), *individuals* (specific instantiations such as Ti-6Al-4 V alloy), and *roles* (relationships) [22]. Analogous to these, OWL uses *classes*, *individuals*, and *properties*, respectively. Both of these ontological modeling techniques envision ontologies as graphs, where a graph is comprised of nodes and relationships [23]. The nodes reflect physical entities (i.e., nouns, objects, concepts, classes), while the relationships capture the contextual actions (i.e., verbs, predicates, roles, properties). A graph is easily visualized as shapes and arrows, where shapes indicate the nodes and arrows indicate the direction of a relationship. Additionally, data that follows an OWL ontology can be represented in the resource description framework (RDF) model, where all data is represented as a ⟨subject⟩⟨predicate⟩⟨object⟩ triples [24]. Each of these would be associated with defined classes, individuals, or properties in the OWL ontology. Further, the RDF is a directed, labeled graph. Here, the subject and objects are nodes and the predicate is the relationship that points from the subject to the object.

Despite being able to declare rules and reason about them, decidable first-order logics, and therefore the ontology modeling languages OWL and DL, have technical limitations that limit the type of rules that may be defined. In fact, many common first-order logic statements are in fact undecidable and may not be represented in the current ontological modeling languages [25]. However, there are many other graph analysis techniques, the latest of which being deep learning methods [26], which can address the problem of inferring new relationships in graphs. However, these statistical methods need to specify the probability of a relationship, which is not possible with the current RDF syntax, although RDF* [27] is under development and seeks to solve this issue.

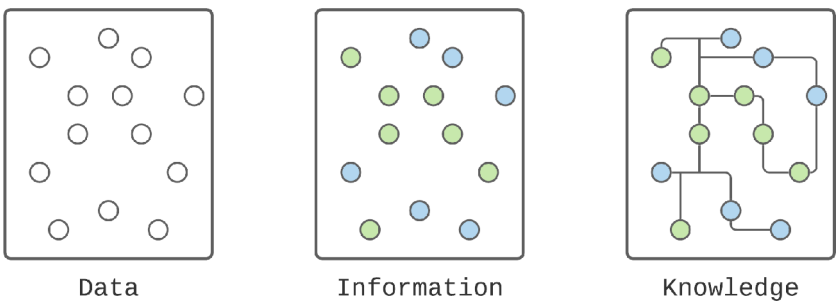


Fig. 1. KR deals with adding contextual information to data that allows us to reason about the data, and understand the patterns and connections present in the data.

Despite the drawbacks in defining rules mentioned above, OWL is still an extremely powerful ontological modeling language. OWL has many associated tools for ontology analysis and inherently allows the integration of any number of other ontologies, which are specified globally through the web using internationalized resource identifiers (IRI). The ability to extend existing ontologies allows new ontologies to draw on already defined classes and properties. This work plans to define a unifying ontology that can merge existing materials ontologies and make use of the extensive quantity of already defined classes and properties in the materials science domain.

3. Materials graph ontology

In materials science, ontologies have been developed for several specific applications such as general materials knowledge from wikidata [28], functionally graded materials [29], and additively manufactured materials [30]. Another ontology integrates two computational materials databases [15]. These ontologies are merged on the identical structure and spacegroup classes, which creates explicit links between the types of properties that can be found in each database. For example, one database schema has x-ray diffraction data associated with structures and another schema has associated prototype structures, which are now linked through their relationship to the structure class in the newly developed ontology. Another materials ontology has been developed for materials synthesis [31], which defines a process as a sequence of a precursor material node followed by a variable number of process operation nodes. This materials ontology is unique because it allows process operations to link to themselves, essentially creating a long chain of process steps that can define an arbitrarily complex processing history. However, despite these developments, there is very little consensus on what a material actually is, and "material" may not even be defined in the specific ontology. Callister [32] defines a material as having four components: Processing, Structure, Properties, and Performance. This implies that the characteristics of a material define it. In an ontology we can then define a material as anything having relationships to the four mentioned classes. Further, materials science is often concerned with how materials are related to each other. Therefore, a materials ontology should be able to answer how two materials are related, and if they are similar or dissimilar. This information can be captured by a materials graph that can link any two materials together in a connected network. The materials graph ontology could also be used to integrate other existing materials ontologies by merging them on the material class.

The proposed materials graph is designed to relate materials to other materials based on the concept of processing history, as discussed later. The ontology specifies the types of nodes (i.e., classes) and relationships (i.e., properties) that are permitted in constructing the materials graph to capture the desired contextual information (i.e., metadata). In other words, one can establish the desired graph by using any of the allowed nodes and relationships in any sequence, repeatedly as needed. Also, it is noted that we will be using the graph terminology node for OWL classes and relationship for OWL properties, as the word property has a different meaning in the materials domain.

Fig. 2 depicts the nodes and relationships of the proposed materials graph ontology. This materials graph ontology is also provided in the OWL syntax in the supplement to this paper, where WebProtégé [33] was used to generate the ontology. The four types of nodes identified in Fig. 2 as Material, Process, Data, and Tool and the six distinct relationships identified as *next_in_process*, *composed_of*, *describes*, *input_to*, *yields*, and *used_in* are proposed to be the minimal set needed to build a materials graph for any given

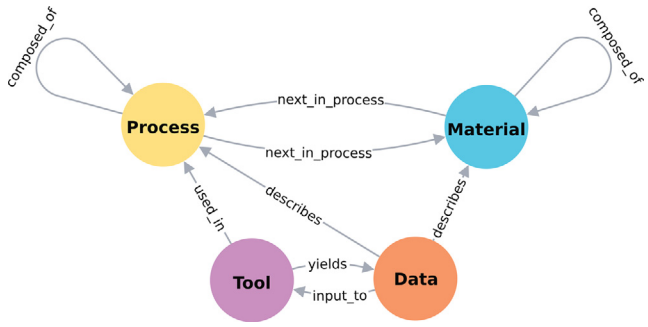


Fig. 2. Depiction of the foundational elements for the proposed materials graph ontology.

materials dataset or database. These can be used to broadly connect all materials concepts and data. Additional node types and relationships can also be added from other ontologies as needed.

The Material node represents a distinct material with associated properties, material structures at a hierarchy of length scales, process history (sequence of Material and Process nodes), and constituents (also represented as Material nodes). As many additional nodes as needed can be added to the Material node to produce the desired material graph. We also define the rule that any two Material nodes that are related to identical properties, material structures, and process history indeed represent identical materials. Furthermore, any two Material nodes that are related to some identical or similar properties, material structures, and process history, with others undefined, have a probability of being the same material. The graph model does not require that identical materials be merged into a single node; they may exist as different nodes in a graph database.

The Process node together with the *next_in_process* relationship is used to indicate the transformation of a material by a processing step. Additional information about the processing step can be captured using the Tool and Data nodes, along with the allowed relationships specified in Fig. 2. The relationship between a Material node and a Process node is defined exclusively by *next_in_process* which stipulates that a material can only change through a processing step and a process always acts on materials. A sequence of alternating and connected Material and Process nodes can then be used to capture a complex manufacturing process. The Material and Process nodes are allowed to relate to themselves, i.e., only these nodes are allowed to be connected to other nodes of the same kind. This is because one can visualize these entities as complex physical systems made of other entities of their type. A Material may have constituents defined at different length scales (e.g., phases, precipitates), which can be treated as distinct materials by themselves (i.e., they exhibit their own unique properties and structure). As another example, a heat treatment Process may have substep Processes such as ramp, soak, and quench, as shown in Fig. 3(a).

The Data node is used to capture all of the associated data and metadata that describe a material in terms of its structure, properties, or performance. For the Material and Process nodes, this is accomplished using the relationship *describes*. Data node can also store the input and output data from a Tool node using the relationships *input_to* and *yields*, respectively. Note that the relations *input_to* and *yields* are special as they can be linked. For example, if Tool acts as a function, it should map the exact input Data to the yielded Data. The Data node can be broadly used to store the results of experiments, simulations, or curated values from literature or domain experts.

The Tool node is designed to represent the different machines used by the materials experts. These may include a broad range

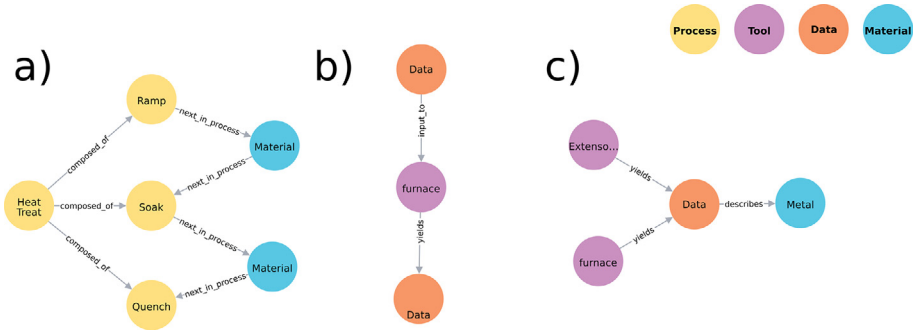


Fig. 3. Example Materials Graphs showing the (a) hierarchy of Processes, (b) conversion of Data by Tools, and (c) generation of Data from multiple Tool sources.

of equipment such as processing equipment, characterization equipment, and simulation/analysis software. Multiple Tools nodes may be connected to a single Process node to capture the desired metadata on how a specific materials processing step was achieved. For example, Fig. 3(b) shows a small materials graph that captures details of a heat treatment, where the temperature control information is *input_to* the furnace and the furnace *yields* temperature history. As another example, the use of multiple tools to measure a material property can be captured using multiple Tool nodes and a single Data node and a single Material node. This is illustrated in Fig. 3(c) for the measurement of the thermal expansion coefficient.

4. Case study

This case study examines an available dataset (containing both process history and material properties) taken directly from Ref. [34]. In this example, there are many discrete materials listed as rows in tables, without any notion of connection between those materials. One of the main benefits of the materials graph ontology is that these materials can be inter-related in a materials graph. Additionally, the implementation of the materials graph ontology would also allow us to store all of the original data in a connected manner together with the contextual information.

Fig. 4 describes the connected dataset generated by Ref. [34] as several disjoint materials graphs, using the ontology proposed in this work, where each material graph corresponds to a row in a table provided in Ref. [34]. The graphs start with a starting material and track their transformations through the different imposed process histories, while capturing the details on the tools employed and the data collected at different stages.

After identifying node equivalencies, the materials graphs in Fig. 4 are unified into the single graph shown in Fig. 5. The unified graph recognizes that there is a common starting material that was subsequently processed in overlapping process histories. For example, 780--00-000 is an ancestor common to both 780-05-170 and 780-10-170, and this contextual information is captured in an unambiguous manner in the proposed materials knowledge graph.

5. Discussion

Ontologies provide a powerful method for relating information, which would allow us to integrate heterogeneous datasets and make that data more useful by identifying data inter-relationships. However, there are still significant challenges to making the ontologies truly useful. For one, as observed in the case study presented earlier, a direct conversion from existing data to the graphical format produces the same disconnected data that we had before. Ontology rules are essential to generating truly connected data. Defining such rules is quite challenging, as it requires the conversion of higher level statements to first-order logic. For example, the rule that a material is the same as another material if the process history, properties, and constituent materials are the same requires multiple comparisons over a complex network of nodes and edges. Although this comparison is relatively straight-forward to define in graph query languages, defining such a rule in an ontology is quite complex. However, it is important that such definitions are represented in a formal ontology such that they can be automatically inferred by inference tools and defining them formally will be the subject of future work.

Secondly, an advantage of ontologies is that they can be combined by referencing the other ontologies via an Internationalized Resource Identifier (IRI). However, this requires that the other ontologies' IRIs be resolvable either locally or, preferably, through the internet. The current materials ontologies do not provide any IRI through which they could be referenced. There are online tools to help facilitate the sharing of ontologies, such as WebProtégé [33]. However, it is only possible to access ontologies in WebProtégé by making an account and by having the creator make the ontology public.

6. Conclusions

This work proposed a materials graph ontology that is capable of connecting disparate materials data with related materials concepts typically employed by domain experts. It describes the concept of a material in terms of its relationships to other concepts including process history, structure and property data, and other materials. This will enhance the ability to relate materials data to the actual concept of a material, improve the FAIR characteristics of the data, integrate heterogeneous datasets, and make it easier to define a class of materials in terms of related concepts. In materials science, large quantities of data are being produced by simulations, high throughput testing, and other data generation efforts. However, this data lacks the contextual information to make it reusable. Being able to add contextual information to the data using the materials graph ontology will not only improve the FAIR characteristics of the data, but give it the potential to be combined with other data, increasing its value. The materials graph ontology presented in this work allows integration with the other existing ontologies.

CRediT authorship contribution statement

Sven P. Voigt: Conceptualization, Methodology, Software, Writing - original draft. **Surya R. Kalidindi:** Conceptualization, Project administration, Funding acquisition, Writing - review & editing.

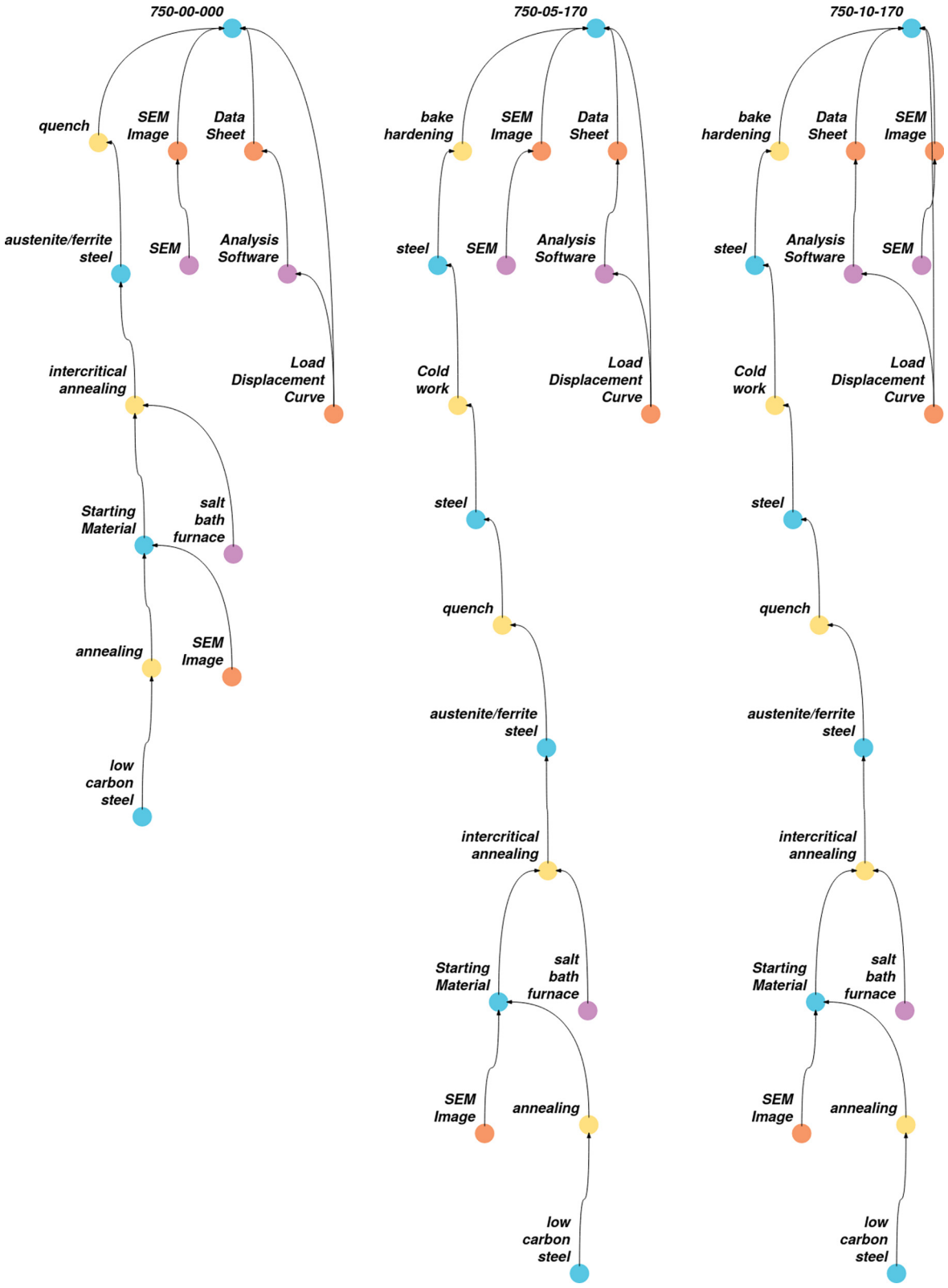


Fig. 4. The material graphs for samples 750-00-000, 750-05-170, and 750-10-170 from Ref. [34].

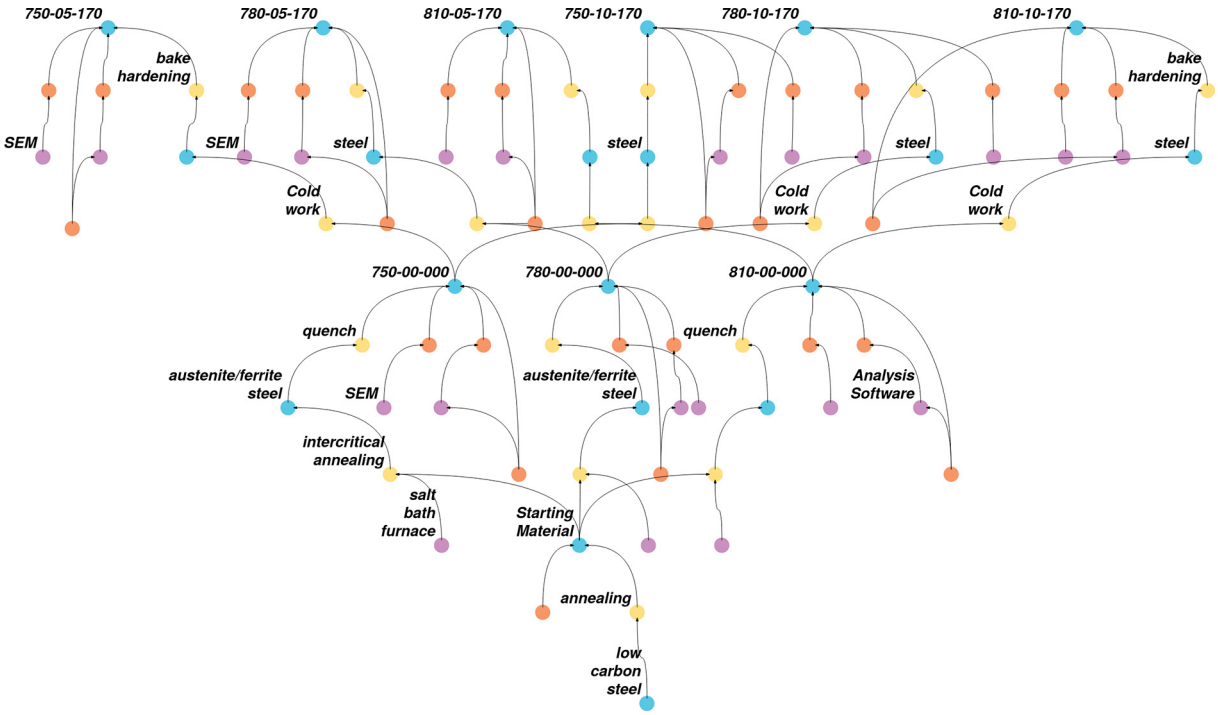


Fig. 5. Materials knowledge graph for all the samples processed in Ref. [34].

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgements

The authors acknowledge support for this work from NIST 70NANB18H039 (Program Manager: Dr. James Warren).

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.matlet.2021.129836.

References

[1] J.P. Holdren et al., Nat. Sci. Technol. Council (2011).
[2] D.L. McDowell et al., MRS Bull. 41 (2016) 326–337.
[3] S.R. Kalidindi et al., Integr. Mater. Manuf. Innov. 8 (2019) 441–454.
[4] B. Smith et al., Formal Ontol. Inf. Syst. (2001) 7.
[5] H. Li, et al., The Semantic Web – ISWC 2020 12507 (2020) 212–227
[6] S.R. Kalidindi, MRS Commun. 9 (2019) 518–531.
[7] S. Kalidindi, Butterworth-Heinemann (2015).
[8] K. Alberi et al., J. Phys. D Appl. Phys. 52 (2019) 013001.
[9] B. Blaiszik et al., JOM 68 (2016) 2045–2052.
[10] J. Hill et al., Comput. Mater. Syst. Design (2018) 193–225.
[11] S. Ramakrishna et al., J. Intell. Manuf. 30 (2019) 2307–2326.
[12] D. Gunter, et al., 2012 SC Companion: High Performance Computing, Networking Storage and Analysis (2012) 1244–1251
[13] S. Curtarolo et al., Comput. Mater. Sci. 58 (2012) 218–226.
[14] G. Pizzi et al., Comput. Mater. Sci. 111 (2016) 218–230.
[15] S. Zhao et al., AIP Adv. 7 (2017) 105325.
[16] S.R. Hall et al., Acta Crystallogr. A 47 (1991) 655–685.
[17] P. Murray-Rust et al., J. Chem. Inf. Comput. Sci. 39 (1999) 928–942.
[18] S. Somnath, et al., (2019) arXiv:1903.09515.
[19] R. Davis et al., AI Mag. 14 (1993) 17.
[20] L. Ehrlinger et al., SEMANTICS 48 (2016) 4.
[21] S. Bechhofer, et al., W3C (2004) www.w3.org/TR/owl-ref/.
[22] M. Krötzsch, et al., (2013) arXiv:1201.4089.
[23] M. Needham, et al., O'Reilly Media (2019).
[24] G. Klyne, et al. (Eds.), W3C (2014) www.w3.org/TR/rdf11-concepts/.
[25] M. Krötzsch, Description Logics (2017) 12.
[26] Q. Wang et al., IEEE Trans. Knowl. Data Eng. 29 (2017) 2724–2743.
[27] O. Hartig, (2014) arXiv:1409.3288.
[28] X. Zhang et al., Comput. Phys. Commun. 211 (2017) 98–112.
[29] M. Mohd Ali et al., Int. J. Prod. Res. (2020) 1–18.
[30] E.M. Sanfilippo et al., Comput. Ind. 109 (2019) 182–194.
[31] E. Kim et al., Matter 1 (2019) 8–12.
[32] W.D. Callister, et al., John Wiley & Sons (2010).
[33] T. Tudorache et al., Semantic Web 4 (2013) 89–99.
[34] A. Khosravani et al., Acta Mater. 123 (2017) 55–69.

Computer-Aided Design 43 (2011) 1902–1911

Contents lists available at SciVerse ScienceDirect

Computer-Aided Design

journal homepage: www.elsevier.com/locate/cad

A product lifecycle management methodology for supporting knowledge reuse in the consumer packaged goods domain

Enrico Vezzetti [a,*], Sandro Moos [a], Simona Kretli [b]

[a] Dipartimento di Sistemi di Produzione ed Economia dell'Azienda, Politecnico di Torino, Corso Duca degli Abruzzi 24, Torino, 10129, Italy

[b] Department of Science, Università degli Studi "G.D'Annunzio" Chieti-Pescara, Viale Pindaro 42, Pescara, 65127, Italy

ARTICLE INFO

Article history:
Received 5 November 2010
Accepted 30 June 2011

Keywords:
PLM
QFD
TRIZ
Knowledge sharing
Waste disposal

ABSTRACT

The present globalized market is forcing many companies to invest in new strategies and tools for supporting knowledge management. This aspect is becoming a key factor in the industrial competitiveness for the presence of extended enterprises that normally deal with huge data exchange and share processes. This scenario is due to the presence of partners geographically distributed over the entire globe, that participate in different steps of the product lifecycle (product development, maintenance and recycling). At present, Product Lifecycle Management (PLM) seems to be the appropriate solution to support enterprises in this complex scenario, even though a real standardized approach for the implementation of knowledge sharing and management tools does not exist today. For this reason, the aim of this paper is to develop a knowledge management operative methodology able to support the formalization and the reuse of the enterprise expertise acquired while working on previous products. By focusing on consumer packaged goods enterprises and on the concept development phase (which is one of the most knowledge intensive phases of the whole product lifecycle), this research work has developed a new systematic methodology to support knowledge codification and knowledge management operations. The new methodology integrates the Quality Function Deployment (QFD) and the Teoriya Resheniya Izobreatatelskikh Zadatch (TRIZ). Also, a case study on the problem of waste disposal has been conducted to validate the proposed methodology.

1. Introduction

The necessity to focus attention on the product has driven many companies to increase their efficiency by improving collaboration with partners, customers, and across all other company functions. This has been obtained by the adoption of Product Lifecycle Management (PLM). This business approach applies a consistent set of business solutions supporting the collaborative creation, management, dissemination, and use of enterprise data along the entire product lifecycle [1,2].

Although the data to be managed come in many different forms, they generally are classified into three main categories: product data, production data and operational support data.

Product data describe how the product is designed, manufactured, operated or used, serviced and then retired. Production data focus on all activities associated with the production and the distribution of the product. Operational support data deal with the enterprise's core resources, such as people, finances and other resources required for supporting the enterprise. This information is provided to all the organizational sectors in order to support an efficient integration of people, data, processes and business systems [3].

In the Consumer Packaged Goods (CPG) Industry, PLM plays an important role, particularly in the domain of packaging. To better understand this concept, it must be emphasized that the product conception should take into consideration not only the content but also the packaging. This definitely contributes to the overall product cost due to its own costs and transportation costs (which vary depending on weight). For example [4], the packaging cost of a jar containing beans is about 26% of the industrial selling price of the product. For a bottle of tomato sauce of 700 g, the bottle itself can reach up to 25% of the final selling price. For fruit juice in a box, the percentage is usually around 20%, while for milk in a plastic bottle, it is above 10%. Packaging also represents a significant portion of a product's selling price in other market sectors. For example [5], in the Cosmetics Industry the packaging cost may be as high as 40% of the product's selling price.

The environmental aspect is also an another factor to consider. Today the food sector is the one most responsible for the production of packaging waste. The 94/62/EC Directive [6] on packaging

* Corresponding author. Tel.: +39 011 564 7294.
E-mail addresses: enrico.vezzetti@polito.it (E. Vezzetti), kretli@sci.unich.it (S. Kretli).

0010-4485/$ – see front matter © 2011 Elsevier Ltd. All rights reserved.
doi:10.1016/j.cad.2011.06.025

and packaging waste presents clear specifications on the fundamental aspects of the packaging's environmental impact. The rules concern the prevention of packaging waste, packaging reuse, recycling and other forms of recovering for the volume reduction in the final disposal phase. To improve reusing and recycling, systems must be set up to guarantee and monitor the return of used packaging and packaging waste. Lifecycle assessments should be completed to define reusable, recyclable and recoverable packaging categories. Annex II specifies the requirements of packaging relative to manufacturing, composition, reusability and recoverability. Packaging designers should limit as much as possible its volume and weight, while maintaining the adequate level of safety and hygiene. Packaging ought to be conceived and commercialized so as to permit its reuse, recovery or "recyclability", and with limited noxious or hazardous substances. As for reusability, it is required that the packaging physical properties and characteristics allow a number of trips or rotations in normal use conditions. Packaging must be manufactured to allow the recycling of a certain percentage (by weight) of its component materials. The packaging waste destined for energy recovery processes shall have a minimum inferior calorific value. The biodegradable packaging waste needs to be able to undergo physical, chemical, thermal or biological decomposition. This would ultimately result in the decomposition of the packaging into carbon dioxide, biomass and water.

The main packaging function is to guarantee product safety and conservation throughout the logistic process. This includes shipping, storing, piling, loading and unloading, while being suitable for recycling. Moreover, packaging is the element that most influences customers' purchase process. A correct product identification, a beguiling aspect and a good market appeal are decisive factors which determine customers' choices [7,8].

The efficient exchange of product data made by PLM within the enterprise represented in the last years one of the key competitive factors in the global market. At present it does not seem to be enough. Currently products become obsolete quickly and customers are continuously asking for more added value in their products. In many companies, it is common to design the packaging ex novo or to outsource the design and the production to an external supplier. A new trend may be to exploit the knowledge and experiences related to existing packaging when developing new products inside of the product holding enterprise.

For this reason, data exchange should be replaced by knowledge management. This would allow product document and data to be organized and linked together to support the identification of similar engineering design process scenarios when new products are developed.

This knowledge is normally implicitly known by the designers and mainly based on their personal experience. However it should be explained and shared among designers and across the enterprise [9]. From inception to commercialization, a product generates documents that become the "know-how" of the enterprise. Losing this knowledge and not storing and organizing it, means losing the enterprise history, and losing time when in the future someone will need information regarding a past product.

2. Literature review

The PLM solutions now available on the market are characterized by efficient data sharing processes supporting collaboration inside extended enterprises. They lack the essential capabilities to manage methodologies that would facilitate the reuse of design processes knowledge. In other words, they are unable to identify similar scenarios [10].

To overcome this situation, it is necessary to employ mechanisms for easily codifying the design process to facilitate its retrieval and reuse.

The concept development phase [11] is one of the main knowledge intensive phases of the product lifecycle management. During this phase it is possible to verify that designers define a set of technical specifications that should be maximized or minimized in order to satisfy customers' requirements. As a consequence, a set of physical principles will be selected to implement the aforementioned technical specifications. This passage normally comes from personal expertise and from long testing and simulation work. The selection of the right technical specifications and consequently of the right physical principle represents a key point for a successful product development, and for the successive lifecycle phases (DFA, DFM, DFE,... DFX) [12].

The current working scenario describes a product development process where the product has been developed ex novo. As companies usually produce products that belong to similar domains, the "ex novo" scenario is not common in real situations. In the most usual scenario, in fact, companies implement modifications on an existing product to satisfy new requirements.

This means that very frequently companies work on similar projects (products) correlated with the knowledge and expertise developed during their working history. Unfortunately this historical expertise (product design scenarios) is currently "owned" by only a few people as an implicit competence. and is difficult to share with the whole enterprise.

In order to successfully identify completely or partially similar concept development scenarios, it is necessary to find a formalization method to synthesize the product technical specifications set. Technical specifications are usually expressed in different possible non-standardized and personalized formalizations. Synthesizing data is hence necessary to translate the technical specifications, into a discrete set of measurable standard attributes. This discrete set of attributes could be used for indexing the designer's intents of a previous product project into a common database. When developing a new product, the new designer will be able to use this attributes set as a search key to retrieve the previous solutions, that already were implemented in similar product design processes.

Regarding the reuse of engineering design knowledge, several existing strategies [13] begin by focusing on the most representative components of the product. Then, in order to capture the product knowledge for future projects reuse, an interview is implemented. The analysis of the interview results helps in reducing the product parameters into a small set synthesizing the product design. A different approach [14] regarding the most significant parameters that synthesize the technical specifications, works on similar geometrical features. From another point of view [15], knowledge reuse could be seen as a design repository system that includes descriptive product information such as functionality. Through the use of a representation schema built on logic functions, designers are able to gain insight on how a product functionally operates in a specific domain. Using the DSM matrix it is possible to describe the interaction between the tasks or system elements of a design project [16–19]. With the algorithm of partitioning and tearing, it is also possible to reorder the activities and to define a set of assumptions to reduce the design process iteration. Sharif [20] and Tang [21] used DSM to locate the phases of a new product development characterized by knowledge transfer between activities. DSM was there used for evaluating the enterprise performance and the reuse in redesign activities. To overcome its limitation, the DSM has been combined with Axiomatic Design [22] to obtain a better function requirement of a decomposition.

These examples highlight the necessity to focus on the key attributes. Some works are based on a refining process obtained by selecting only the most important attributes. Others employ a product representation based on product functionality. Even if these approaches could provide useful results, they seem to

employ a subjective point of view involving the personal designer expertise both in the refining process and in the vocabulary definition.

This paper will aim to solve this problem by trying to create a vocabulary set independent from the personal expertise of designers. This should allow the enterprise to synthesize the product technical specifications, which could be employed for indexing the enterprise experience in a common knowledge repository supporting its simple retrieve and reuse. Considering that many companies work dynamically with different partners, the vocabulary set must be composed of a discrete list of attributes sharable with all possible partners.

The ISO 10303 AP239 [23] which deals with product lifecycle support (PCLS) and covers the entire history of a product from conceptual design to disposal, could provide a standardized set of product attributes. As explained, standardized terminology represents a fundamental point for knowledge reuse. However, as the main focus of this paper is the "concept design" knowledge reuse, it is also necessary to identify a solution for organizing and storing the experimental approaches and methodologies employed to solve conflicts between technical specifications of past product projects. There are several different approaches based on ontology for supporting this process. Nonetheless, the use of the conflicts matrix provided by TRIZ gives an interesting challenge for supporting the development of a first operative tool for the reuse of the enterprise expertise. It also, does not require heavy investments, such as those necessary for creating a more elaborated knowledge database.

The Quality Function Design approach (QFD) [24] is necessary for codifying system requirements and correlating them with technical specifications. Its employment, together with the conflicts matrix proposed by TRIZ, render it possible to easily organize and store all the experimental approaches and methodologies (solutions) adopted for the first time in the solution of conflicts between two technical specifications. (this has been explained in the QFD roof). These experimental approaches and methodologies are stored together with the most fitting inventive principle. The inventive principles are located in the intersecting cell between the line and column where the two conflicting variables are located. By creating reduced conflicts matrices containing only the attributes of the specific product, it is possible to describe the product design scenario. Comparing these reduced matrices makes it possible to understand if the enterprise has already developed similar product design scenarios in the past.

QFD was chosen as the first operative tool, because it is able to represent the interactions between design objectives (as in DSM) and the correlation with the customer requirements (as in DMM). When combined with TRIZ, it is evident the added value of explicitly storing the conflicting design parameters and the inventive principle used to solve the conflict for subsequent research and reuse.

Regarding the user needs analysis and management, QFD could be integrated with the use Requirement Engineering (RE) [25,26] during the first methodology steps, where customer needs are analysed, so as to provide a rational background to the activity. This methodology investigates and describes the conditions in which a new or updated system is supposed to create the desired effects. It also designs and documents the behaviour of the system to be developed [27,28]. RE considers the needs of various stakeholders, which could be potentially conflicting, and selects those that would be implemented in the design phase. This selection is made to ensure the success of the product [29]. Several steps are fundamental, for instance the collection, analysis and negotiation of requirements allowing the selection of the technical specifications that will guide the system modelling. After that, the requirements are validated and managed [30,31].

Several problems usually arise in its application as the requirement's list is usually long and difficult to understand as a whole. Furthermore the list does not define requirement's priority, their interaction, or logical succession, hence causing difficulties in the knowledge capture tool implementation. That is why further evaluation is needed after this preliminary work.

This paper is organized as follows: the next session introduces the proposed methodology integrating the Quality Function Design (QFD) [24] method with the Teoriya Resheniya Izobreatatelskikh Zadatch (TRIZ) one [32] for organizing the design process knowledge. Section 3 illustrates through a case study how the proposed methodology works.

3. The proposed methodology

This paper proposes an operative methodology for supporting the product innovation design in the consumer packaged goods. This new operative methodology is created by combining the Quality Function Design approach (QFD) [24] with the Teoriya Resheniya Izobreatatelskikh Zadatch (TRIZ) methodology. The QFD is necessary for codifying system requirements and for explaining conflicts, whereas the TRIZ methodology is used to identify the discrete set of attributes to be employed for translating the product technical specifications into common vocabulary and for indexing them in a common knowledge repository. Thanks to this integration, it is possible to exploit the enterprise expertise by easily retrieving similar past product design scenarios.

The methodology starts (Fig. 1) from the evidence that the development of a product begins when the market stakeholders show new needs. As a consequence of this, it is necessary to collect as many needs as possible in order to have a clear idea of the market scenario. The goal is to satisfy all of them by presenting the right product at the right moment. Once collected and elaborated, the stakeholders' needs are converted into technical specifications to form the first description of the new product.

The customer needs codification in technical specification represents one of the key factors for the product development, and QFD represents one of the best tools for its implementation. Unfortunately, as the QFD approach is the only one employed, any engineer or technician involved in this codification employs his/her own glossary, based on his/her own personal experience. This represents a problem when this experience needs to be reused. When working only with the QFD, it is possible to work on a product with design problems similar to those previously faced and still misunderstand the problem because it has been described using a different vocabulary instead of a common one.

To minimize this issue, designers need to acquire experience and consequently the non-formal company knowledge, by actively working on the problem (by doing). The issue is clearly a critical one when a new member becomes part of the product development team and he/she ignores all previous experiences. For this reason, it is necessary to define a common language to be used for storing previous experiences and company knowledge. Using a common formalization language would make it easier to retrieve previous experiences. Furthermore, this formalization will allow an easy reuse of the company knowledge.

In order to fulfil this aim, the methodology proposed integrates the TRIZ methodology and uses it to introduce a common formalization language, which subsequently can be used to support company knowledge reuse. The TRIZ method provides 39 engineering parameters to describe the product (speed, weight, measurements accuracy, etc.). These parameters are divided into three categories: physical and geometric parameters, negative parameters independent by technique, and positive parameters independent by technique. All of these are used to translate the QFD technical specifications (Fig. 2).

Table 1
Proposed methodology scenarios.

Tasks	Scenario 1: product knowledge reuse	Scenario 2: physical principle reuse	Scenario 3: new domain problem
Control 1: At this step, the methodology requires to check if a similar product development scenario had already been analysed in the past	There is a similar conflicts matrix already stored	There is no similar conflicts matrix already stored	There is no conflicts matrix already stored
Control 2: At this step, the methodology requires to check if a similar variable conflict already occurred in the past. If so, the physical principle used would need to be identified.		There are some physical principles stored together with the inventive principles	There is no physical principle stored together with the inventive principles
Solution	*The conflict scenario was already analysed while working on a previous product. It is possible to reuse all the physical principles already employed*	*The stored physical principles are checked in order to verify if they fit the present conflict. If they do, then the principle is adopted; otherwise a new principle is developed and stored*	*A new physical principle is developed and, after its validation, stored in the company data base*

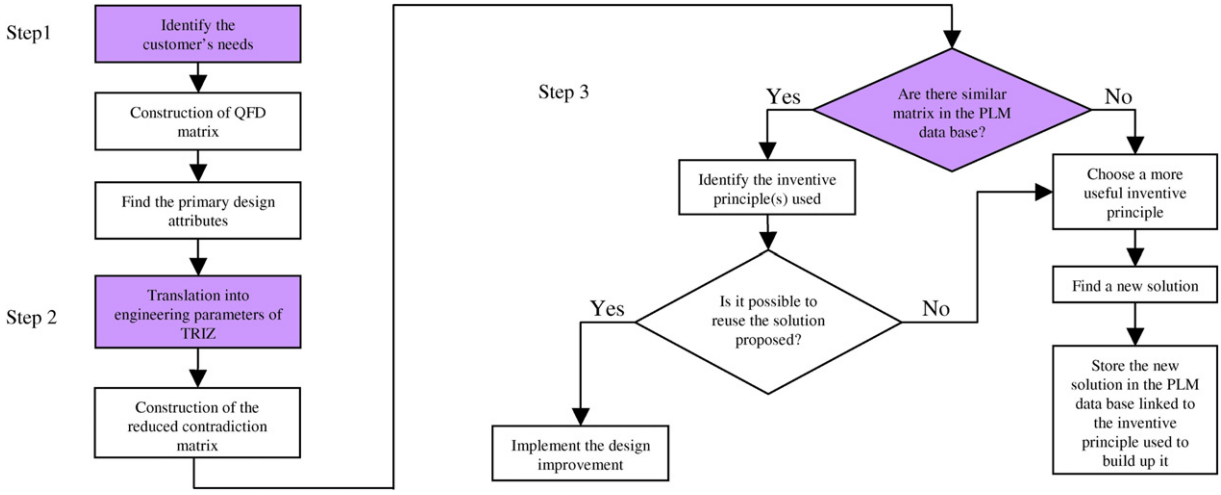


Fig. 1. Proposed methodology flowchart.

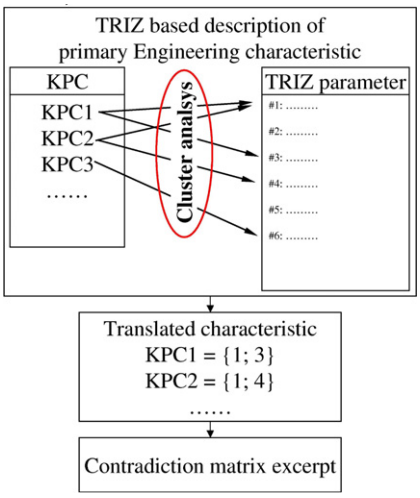


Fig. 2. QFD technical specifications translation in TRIZ variables.

Using the QFD correlation matrix and its roof, it is possible to understand which technical specifications are directly correlated, and which are inversely correlated. If two technical specifications are directly correlated, when one increases, the other will also increase. On the other hand, when two technical specifications are inversely correlated, when one increases the other will decrease. The QFD correlation matrix allows us to identify which technical specifications should be maximized and which, on the contrary, should be minimized.

As a consequence of this, the new method provides a set of TRIZ variables in conflict. Instead of using the complete conflict TRIZ matrix (39×39), a reduced one is adopted. This reduced version is composed only of those specifications involved in the semantic translation process. The reason behind this choice is that not all of the TRIZ variables are useful for the enterprise specific product tasks. As an example, TRIZ parameters such as speed (9) and power (21) are not compatible with the food and beverage packaging characteristics.

In the 39×39 matrix, the 39 engineering variables are located within the first column and the first row. Similarly, in the reduced matrix, the involved TRIZ variables are located in the first row and in the first column, while one or more of the 40 inventive principles are placed in the other cells. Those inventive principles have been identified by TRIZ developers by analysing the breakthroughs contained in international patents. Looking through these inventive principles, it will be possible to identify for each case one physical principle which suits the new product development process. The translation from QFD specifications to TRIZ variables is performed using a cluster analysis approach [33]. Hence, by observing which variables need opposite physical principles, the reduced conflicts matrix principles are analysed and a physical principle is studied. At this stage, three possible scenarios may result (Table 1).

In the first scenario, the proposed methodology has already been employed in previous projects. As a consequence, a certain number of reduced conflicts matrices have been developed and

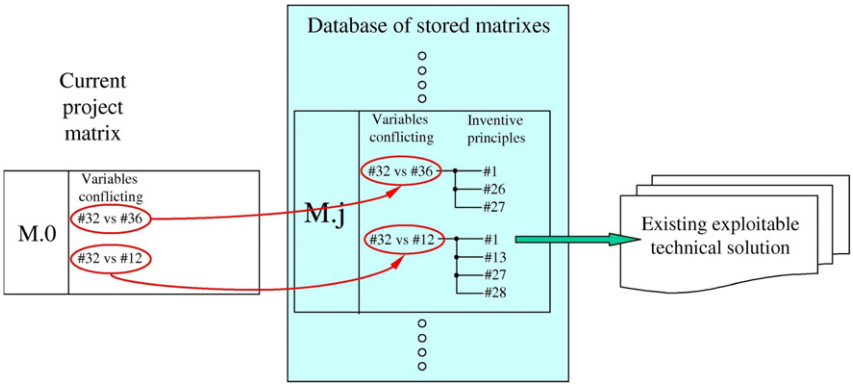


Fig. 3. Knowledge recovery from the enterprise PLM database when a similar matrix is found.

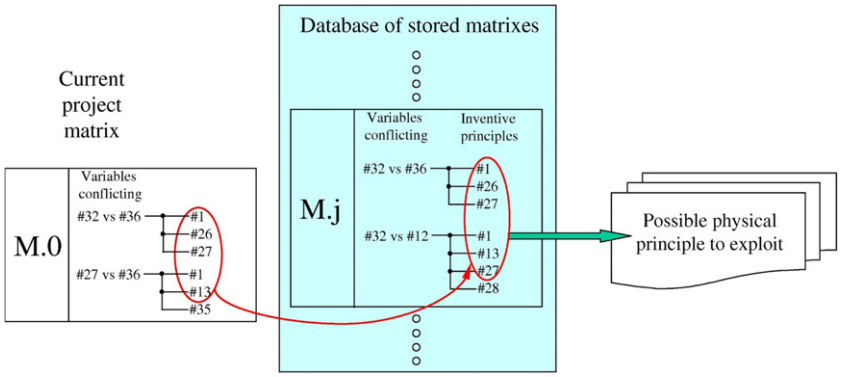


Fig. 4. Knowledge recovery from the enterprise PLM database, based on the inventive principles, when a full correspondence is not found.

stored and a certain number of physical principles have been integrated with the inventive principles (Fig. 3). In this situation, designers define the reduced conflicts matrix of the new development product and compare it with the others already stored in the company database. If the comparison shows that the matrices are equal, the previous product development scenario and the previous product experience can be reused to solve the new case.

In the second scenario similar conflict matrices are not found in the database. In this case, it is necessary to work at a lower level, because a similar product development scenario has not yet been analysed (Fig. 4). In such situations, the methodology proposes to verify if any of the stored physical principles could fit within the present variable conflict. Hence, the comparison must take place between the single inventive principles stored and the ones of the new reduced inventive matrix.

In the last scenario, no similar reduced conflict matrices have been found and no fitting physical principles have been identified while analysing single inventive principles (Fig. 5). This situation could occur when the company is working on a new domain or if the proposed methodology is employed for the first time. A new physical principle must hence be proposed. If the new principle solves the conflict and is accepted by the team, it will then be stored together with the inventive principle.

Thanks to this methodology, it is possible to store, organize and share the company's knowledge and avoid competence loss. The company knowledge will also be available to be searched and reused for future products, according to the PLM principle. The search key will be the reduced matrix arising from the TRIZ application, which provides a common vocabulary suitable for any of the product characteristics. Moreover, thanks to this TRIZ translation of the QFD Technical specifications (KPCs) (which are specific to the designer and the project) specifications are moved to a "higher" level. This means that they become more flexible and can be adapted to suit any product development typology.

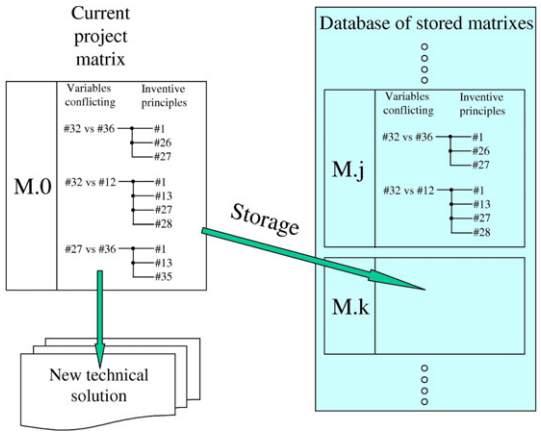


Fig. 5. Development of a new solution and storage of the matrix in the enterprise PLM database.

The use of the conflicts matrix as comparison element allows the specialization of the inventive principles used to solve the technical specification conflict. Every conflict is associated to one inventive principle and to one solution. As data are at first transformed into information and subsequently into knowledge (raw data), technical specifications, in this context, cease to be objective and independent from the possibility to be correlated with a specific meaning by an observer. Instead, they become contextualized thanks to the creation of the conflicts matrix and become information. The inventive principle and the proposed solution associated with the matrix allow to interpret this information, providing its translation into knowledge.

Table 2
The reduced conflict matrix. In the cells are listed the TRIZ inventive principles that can be applied to solve the conflicts. In bold are the conflicting parameters. In italic is the useful principle applied.

	14	16	23	32	**36**
14	–		35, 28, 31, 40	11, 3, 10, 32	2, 35, 22, 26
16		–	27, 16, 18, 38	35, 10	
23	*35*, 28, 31, 40	27, 16, 18, 38	–	15, 34, 33	35, 10, 28,24
32	11, 3, 10, 32	35, 16	15, 34, 33	–	27, 26, *1*
36	2, 13, 28		35, 10, 28, 29	27, 26, 1, 13	–

4. Case study

In order to understand the efficacy of the proposed methodology on consumer packaged goods, two packaging solution design scenarios are proposed below. In both scenarios, waste materials are easy to dispose of according to the design-for-environment principles. These principles aim to extend the enterprise design objectives to the end of the product's life.

This experimental validation is organized into two steps involving two different packaging design scenarios. In the first one it is assumed that this kind of problem was never dealt with before in the enterprise; hence, in this case, it will not be possible to find a former similar conflicts matrix. This new problem will be solved using the TRIZ methodology. The obtained results will be stored into the PLM database for future reuse. In the second scenario, the process begins by considering the fact that the problem may have been solved in the past. For this reason, the first step consists of comparing the reduced matrices and describing what happened when the packaging design came from reuse.

Food and beverage packaging is very often made of polylaminate material (i.e. packaging for milk, fruit juices, wine, soup and other liquid food), which is usually composed of aluminium alloy series 1000 (5%), polyethylene (20%) and paper (75%). The percentage of waste of polylaminate packaging is continuously increasing in municipal solid waste. In 2007 more than 137 billions of packaging were produced and used around the world [32].

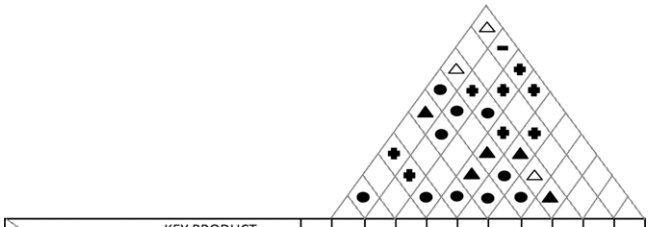
Customers were offered a questionnaire where they were asked to freely describe the product attributes that were more important to them. The answers were reworded considering similar statements of different users and then the reworded proposition were again subjected to the customers group which assigned a score on a scale from 1 to 5 to each of them. Fig. 6 shows the resulting customer requirement list, composed of 11 entries and coupled with the corresponding importance values.

Design experts of packaging enterprises were asked to identify the package design attributes corresponding to the given customer requirements list. They also had to define the weight of the relationship matrix.

In the case study, the last 2 rows of the QFD matrix show the importance ratings of the engineering characteristics. The main design attributes of packaging are: seal material, package material, geometry, seal shape, and seal structural elements. It is clear that the most important issues are the material of the container and the seal.

The TRIZ parameters describing the packaging (container) material are: loss of substance (23), complexity (36), strength (14), durability of non-moving object (16), ease to manufacture (32). Table 2 shows the reduced conflict matrix relative to the characteristics of the container material.

The engineering parameters in conflict for the container material are "#32, ease to manufacture" and "#36, device complexity". The feature to improve is the material eco-compatibility in row #32: "Ease to manufacture" and the undesirable secondary effect is polylaminate material in column #36, "Device complexity". The intersecting cell in the Table of Conflicts suggests the use of the inventive principles #27, #26, and #1 to solve the problem. Those



KEY PRODUCT CHARACTERISTICS / CUSTOMER REQUIREMENTS	IMPORTANCE	Complex geometry	Volume/Size	Material	Strength	Closure material	Shape of closure	Structural elements of closure	Type of process	Type of graphics	stability
Ergonomics	5	9	9	3	3	1	1	1			
Capacity	4	3	9	1	1						
Easy to pile	3	9	1		3	0	3				3
Fall-resistence	4	3		9	9	9	3	9			9
Compression-resistence	5	3		9	9	9	3	9			9
Hermetic	5			3		9	9	3	3		
Easy to close	3					9	9	9	3		
Cost	4		9	9	3	9	3	3	9	9	3
Design	5	9		1			3		1	9	3
Reciclability	2			9		9					
Respect of rules	4	9	3	9	1	9	3	1	3	3	1
Engineering chatacteristics importance rating — absolute		192	132	210	125	248	152	144	77	93	121
Engineering chatacteristics importance rating — relative		14	9,6	15	9,1	18	11	10	5,6	6,8	8,8

Fig. 6. The house of quality for the case study.

principles are to be analysed to see if and how they can be applied to the current problem. The useful inventive Principle #1 is "Segmentation":

a. Divide an object into independent parts.
b. Make a sectional object.
c. Increase the degree of an object's segmentation.

The solution could be obtained by dividing the container into two independent parts: an external paperboard box and an internal bag. The external paperboard guarantees the required strength for the package handling while the internal bag guarantees the correct preservation of the content.

The other conflict is between "#23, loss of substance" (hermeticity) and "#14, strength". To satisfy the strong requirement for material impermeability, not only to fluids but also to gas exchange, without worsening its strength, it is possible to use Principle #35 "Parameter changes":

a. Change an object's physical state (e.g. to a gas, liquid, or solid.)
b. Change the concentration or consistency.
c. Change the degree of flexibility.
d. Change the temperature.

The solution is to choose a flexible material with high density, such as aluminium, instead of plastic one, such as PET.

The stored matrix for the material attributes will contain the parameters description, the conflicting parameters and the useful inventive principles.

Table 3
The reduced conflict matrix.

	4	**6**	**12**	**13**	**32**
4	–	17, 7, 10, 40	13, 14, 15, 7	39, 37, 35	15, 17, 27
6	26, 7, 9, 39	–		2, 38	40, 16
12	13, 14, 10, 7		–	33, 1, 18, 4	1, 32, 17, 28
13	37	39	22, 1, 18, 4	–	35, 19
32	15,17, 27	16, 40	1, **28**, 13, 27	11, 13, **1**	-

Table 4
The reduced conflict matrix.

	4	**6**	**12**	**13**	**27**	**32**	**36**
4	–	17, 7, 10, 40	13, 14, 15, 7	39, 37, 35	10, 14, 29, 40	15, 17, 27	1, 26
6	26, 7, 9, 39	–		2, 38	29, 9	40, 16	1, 18, 36
12	13, 14, 10, 7		–	33, 1, 18, 4	10, 40, 16	1,32, 17, 28	16, 29, 1, 28
13	37	39	22, 1, 18, 4	–		35, 19	
27	15,29,28,11	32, 35, 4, 40	35, 1, 16, 11		–		13, 35, 1
32	15,17, 27	16, 40	1, 28, 13, 27	11, 13, 1		–	27,26,1
36	26	6, 36	15,37,1,8		13, 35, **1**	26, 27, 1, 13	–

It is important that the new packaging characteristics are the same as of those of polylaminate packaging: the internal bag must satisfy to the requirement "Ease to fill". The TRIZ engineering parameters that characterize this requirement are: length of stationary (4), area of stationary (6), stability of the object (13), shape (12), ease to manufacture (32). Table 3 shows the reduced conflict matrix relative to the features of the internal bag.

The engineering parameters in conflict in the case of the internal bag are #32 "Ease to manufacturing", #12 "Shape" and #13 "Stability of the object". The feature to improve in the first conflict is the simplicity of production (#32), and the undesirable secondary effect is a shape that is not simple to fill (#12). Looking at the Table of Conflicts, the principles 1, 28, 13 and 27 can be found in the intersecting cell. The useful inventive Principle is #28 "Mechanics substitution":

a. Replace the mechanical means with sensory (optical, acoustic, taste or smell) means.
b. Use electric, magnetic and electromagnetic fields to interact with the object.
c. Change from static to movable fields and from unstructured fields to those having a structure.
d. Use fields in conjunction with field-activated (e.g. ferromagnetic) particles.

The solution may be obtained creating a mobile canal for filling the internal bag. The feature to improve in the second conflict is the stability during the filling (#13 "Stability of the object") and the undesirable secondary effect is a shape that is not simple to fill (#12 "Shape"). Looking at the Table of Conflicts, it is possible to identify the principles 22, 1, 18 and 4. The useful inventive Principle is again #1, "Segmentation", and the solution may be dividing the internal bag into four pieces to make it more stable. An example is shown in Fig. 7. The four patches welded together provide more stiffness to the bag walls. The bottom is shaped like a plane so as to obtain a larger support surface. The square-like section of the bag will fit well into a parallelepiped cardboard box, which will also provide additional support for the bag walls. These will contribute to the stability of the object once filled.

A suitable seal must also be designed. The engineering parameters describing how this should be conceived are the following: length of stationary (4), area of stationary (6), shape (12), stability of the object (13), reliability (27), ease to manufacture (32), device complexity (36). Table 4 proposes the reduced conflict matrix concerning the seal.

In this matrix, the conflicts between the features to improve and the undesired secondary effects are: reliability against complexity of the closure; ease of manufacturing against device complexity; and stability and the possibility to reuse the closure against complexity in the production. The TRIZ matrix suggests the following principles to solve the conflict between improving #27, "Reliability", and avoiding the secondary effect of #36, "Device complexity". The suggested principles are:

Fig. 7. The possible shape of an internal bag.

#13 "The other way around":

a. Invert the action(s) used to solve the problem (e.g. instead of cooling an object, heat it).
b. Make movable parts (or in alternative, the external environment) fixed, and fixed parts movable.
c. Turn the object (or process) 'upside down'.

#35 "Parameter changes":

a. Change an object's physical state (e.g. to a gas, liquid, or solid.)
b. Change the concentration or consistency.
c. Change the degree of flexibility.
d. Change the temperature.

and #1, "Segmentation".

Based on these inventive principles, the conflicts could possibly be solved by segmenting the seal. In other words, this means

Fig. 8. A bag clip.

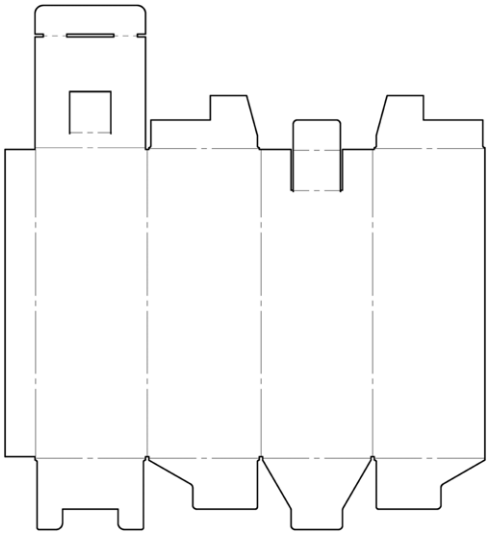


Fig. 9. The net of the paperboard container.

manufacturing the seal of the internal bag separately from the closure of the container (by using a different recyclable material, for example). A possible solution is to use a clip similar to the one shown in Fig. 8 for sealing the internal bag.

The last problem is the container closure, which must be functional and easy to use. When the container is open, the bag closure must be accessible and the liquid inside must be easy to pour. According to the TRIZ inventive principles, the solution proposed is to create a movable paperboard canal linked to the bag clip.

According to the results obtained by applying the methodology, a possible solution to the problem of waste disposal is avoiding to use polylaminate in packaging. Instead, the packaging should be made of two different parts: an external paperboard container and an internal aluminium bag. Fig. 9 shows the net of the paperboard container Fig. 10 shows the bag model, and Fig. 11 explains the new packaging functioning. This can be summarized as follow:

a. The beverage producer opens the paperboard container and
b. places the aluminium bag into the open container, ready for filling.
c. Once the bag is filled, it is sealed with the plastic clip, then
d. the paperboard container is closed and it is handled, stocked and transported to the shop.
e. The customer opens a smaller lid on the top surface of the container, extracts the bag neck, removes the plastic clip and pours the content. To seal again the bag, it is simply sufficient to re-clip it.

In addition, the parallelepiped shape of the container satisfies the requirements "ergonomics" and "ease to pile".

The model of internal bag shown in Fig. 10 has been analysed through a simple FEM simulation to verify the load resistance.

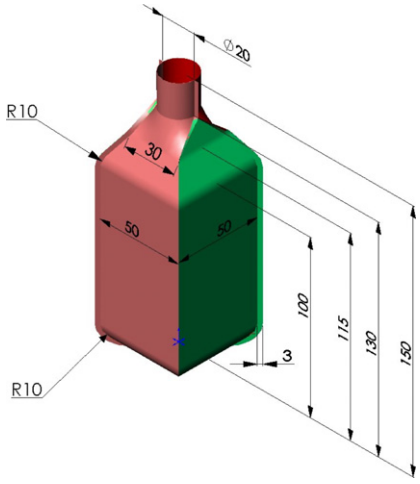


Fig. 10. The internal bag model.

Table 5
The reduced conflict matrix for the second case study.

	14	16	23	32	36
14	–		35, 28, 31, 40	11, 3, 10, 32	2, 35, 22, 26
16		–	27, 16, 18, 38	35, 10	
23	35, 28, 31, 40	27, 16, 18, 38	–	15, 34, 33	35, 10, 28,24
32	11, 3, 10, 32	35, 16	15, 34, 33	–	27, 26, **1**
36	2, 13, 28		35, 10, 28, 29	27, 26, 1, 13	–

On the internal walls of the bag a pressure corresponding to the maximum hydrostatic water pressure for the useful height of the bag (981 Pa) has been applied. In the initial condition the bag walls are separated by 1 mm gap from the external box (considered rigid), so that during the deformation caused by the applied pressure they can be supported by it. The results of numerical analysis are shown in Fig. 12. The maximum equivalent tension calculated is 70 MPa, which is lower than the yielding stress of the aluminium material considered for the calculation (aluminium alloy 1145-H19, yield tensile strength 145 MPa, 0.05 mm thickness, EN601, EN602). This pre-design shows that it is possible to obtain an adequate strength by the combination of the internal bag and the external paperboard container and that the sealing is assured by closing the bag neck with the plastic clip (Table 5).

For the next applications, the search of the conflicts matrices into PLM database will be done using the TRIZ parameters. Once the designers have chosen the TRIZ parameters that better describe the product, those can be used both to search for conflicts and to browse through the database to look for a TRIZ matrix containing a solution for the conflict. At this stage, the result of the technique application is a new package that has been implemented and assessed in the industrial practice. It also complies both with customers' requirements and with environmental regulations. Its conflicts matrices are stored into a database with the part models, the drawings and the other technical documentation of the developed package.

In a second case study, the enterprise has to develop a new package for a different food and different costumers. Once the customers' requirements are collected, the relative weight of all the requirements is defined (Fig. 13) by using the QFD.

Taking a look at the QFD matrix, it is possible to see that "cost" and "recyclability" have a higher importance for costumers than

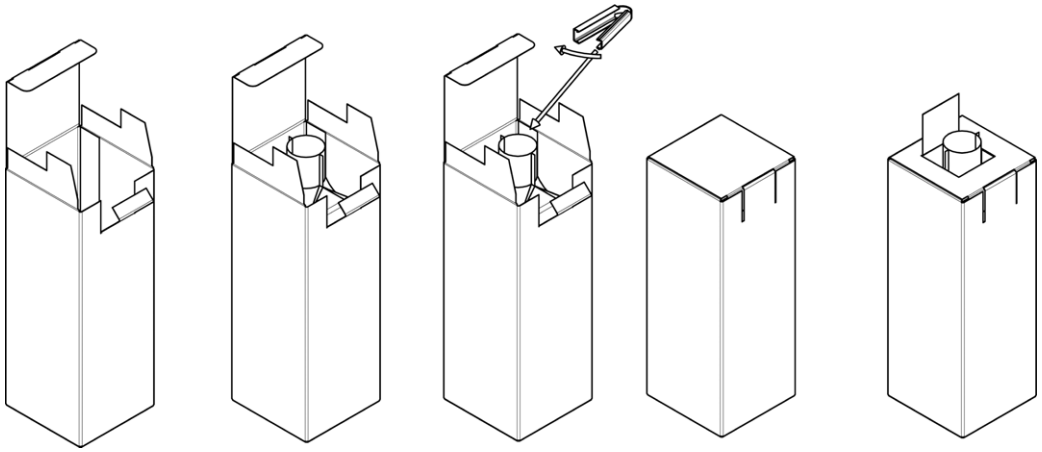


Fig. 11. The functional principle of the new packaging.

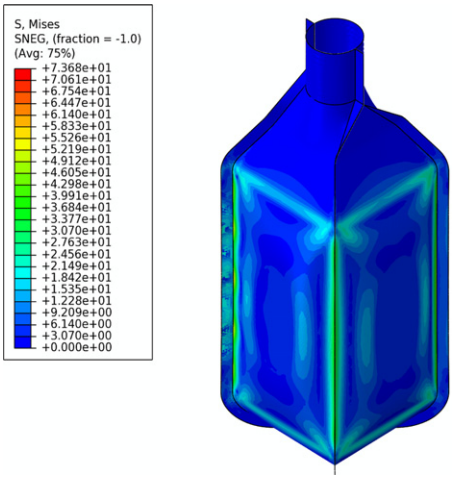
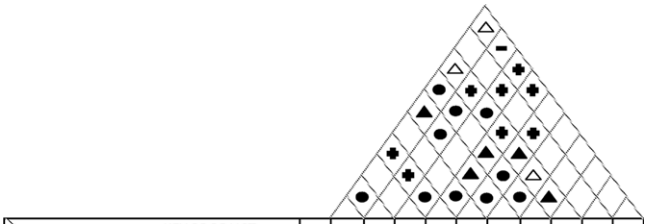


Fig. 12. The FEM analysis of the internal bag.

an "hermetic" seal. Moving the attention on the technical specification "Material", it seems clear that this is the most important one for customers. Focusing, for instance, only on this technical specification and following the TRIZ approach, the resulting conflicts matrix is composed of the following parameters: "Loss of substance (23)", "Complexity (36)", "Strength (14)", "Durability of non-moving object (16)" and "Easy to manufacture (32)". Going ahead with the method, it is possible to understand that the principal conflict is represented by #32 vs. #36. This conflict could be solved by employing the inventive principles 1, 26, 27. After a series of evaluations, principle 1 seems to be the most efficient.

At this stage, it is necessary to browse through the reduced conflicts matrices repository in order to look for a similar conflicts matrix, where the conflict #32 vs. #36 had already been solved by using principle 1. This matrix corresponds to the bag developed in the first case study, but includes one more conflict: #23 vs. #14. Comparing the previous application with the current one, the designer could conclude that principle #35 ("Parameter changes—change the concentration or consistency") is not that important and, as a consequence, choose a more economic and flexible material, such as PET, that can be easily separated for recycling purpose.

Moreover the designer will also be able to access all the previous project data (CAD parts, drawings, specifications, FEM analysis...) and reuse their contents (e.g. the same paperboard container). This will reduce some development costs.



KEY PRODUCT CHARACTERISTICS / CUSTOMER REQUIREMENTS		IMPORTANCE	Complex geometry	Volume/Size	Material	Strength	Closure material	Shape of closure	Structural elements of closure	Type of process	Type of graphics	stability
Ergonomics		5	9	9	3	3	1	1	1			
Capacity		4	3	9	1	1						
Easy to pile		3	9	1		3	0	3				3
Fall-resistence		4	3		9	9	9	3	9			9
Compression-resistence		5	3		9	9	9	3	9			9
Hermetic		3			3		9	9	3	3		
Easy to close		3					9	9	9	3		
Cost		5		9	9	3	9	3	3	9	9	3
Design		5	9		1			3		1	9	3
Reciclability		5			9		9					
Respect of rules		4	9	3	9	1	9	3	1	3	3	1
Engineering chatacteristics importance rating	absolute		192	141	240	128	266	137	141	80	102	124
	relative		14	10	17	9,3	19	10	10	5,8	7,4	9

Fig. 13. The house of quality for the second case study.

5. Conclusions

The proposed operative methodology offers great potential for food & beverage product developers who want to develop innovative packaging at a low cost. This can be achieved by reusing the historical knowledge of the enterprise, which is recorded and shared through the PLM database without heavy investments. Its implementation creates wide possibilities for the reuse of old designs and solutions adopted in other similar fields or products. This would reduce design costs, plant setup costs, and reduce the necessary time to reach the market.

The QFD matrix helps designers to identify the key product characteristics starting from customers' preferences. The KPCs are a useful tool to highlight the new packaging requirements that must be satisfied during the design phase. The TRIZ methodology successfully examines the conflict among the KPCs and proposes practical approaches to solve the problem. Moreover, the TRIZ description of the problem through conflicts matrices is useful to

index and research information into the PLM database, in order to facilitate the sharing and reusing of the enterprise knowledge.

The case study solves the conflict between key product characteristics and the designer's intent to consider packaging disposal and recycling (for example, in order to comply with the design-for-environment principles). It also demonstrates the strong ability of the combined methodology to develop a new packaging design with completely recyclable materials such as paperboard for the external container and aluminium for the internal bag.

References

[1] Saaksvuori Antti, Immonen Anselmi. Product lifecycle management. Berlin: Springer-Verlag; 2004.

[2] Stark John. Product lifecycle management paradigm for 21st century product realisation. London: Springer-Verlag; 2005.

[3] Günther Schuh, Henrique Rozenfeld, Dirk Assmus, Eduardo Zancul. Process oriented framework to support PLM implementation. Computers in industry 2008;59:210–8. doi:10.1016/j.compind.2007.06.015.

[4] http://www.packagingblog.it/2009/09/coldiretti-il-costo-delle-confezioni-supera-quello-degli-alimenti/.

[5] http://www.knowthis.com/principles-of-marketing-tutorials/product-decisions/factors-in-packaging-decision/.

[6] Direttive EC. 94/62/EC on packaging and packaging waste.

[7] Loschiavo dos Santos MariaCecilia, Franco Pereira Andrèa. Packaging: function, re-function, malfunction. from consumer society to the homeless material culture. In: EcoDesign '99. First international symposium on environmentally conscious design and inverse manufacturing. Proceedings. 1999. p. 492 –6. doi:10.1109/ECODIM.1999.747662.

[8] Meroni Anna. Active packaging as an opportunity to create package design that reflect the communicational, functional and logistical requirement of food product. Packaging Technology and Science. 2000;13(6):243–8. doi:10.1002/pts.524.

[9] Aziz H, Gao J, Maropoulos P, Cheung WM. Open standard. open source and peer-to-peer tools and methods for collaborative product development. Computers in Industry. 2005;56(3):260–71. ISSN:0166–3615.

[10] Mesihovic S, Malmqvist J, Pikosz P. Product data management system-based support for engineering project management. Journal of Engineering Design 2004;15(4):389–403.

[11] Ulrich K, Eppinger S. Product design and development. Mcgraw Hill; 2003.

[12] Huang GQ. Design for *X*-concurrent engineering imperatives. London: Chapman and Hall; 1996.

[13] Baxter David, Gao James, Roy Rajkumar. Design process knowledge reuse challenge and issues. Computer-Aided Design & Application 2008;5(6):942–52.

[14] Baxter David, Gao James. Development of a process based data driven engineering design knowledge reuse system. Computer-Aided Design &Application 2006;3(1–4):109–17.

[15] Bohm MR, Stone RB. Representing functionality to support reuse: conceptual and supporting functions, ASME 2004 design engineering technical conference, 2004.

[16] Steward D. The design structure system: a method for modelling the design of complex systems. IEEE Transsctions on Engineering Management 1981;3:71–4.

[17] S.T Pektas, M Pultar. Modelling detailed information flows in building design with the parameter-based design structure matrix. Design Studies 2006;27:99–122. doi:10.1016/j.destud.2005.07.004.

[18] Chen S, Huang E. A systematic approach for supply chain improvement using design structure matrix. Journal of Intelligent Manufacturing 2007;18:2:85–299. doi:10.1007/s10845-007-0022-z.

[19] Avnet MS, Weigel AL. An application of the design structure matrix to integrated concurrent engineering. Acta Astronautica 2010;66:937–49. doi:10.1016/j.actaastro.2009.09.004.

[20] Sharif SA, Kayis B. DSM as a knowledge capture tool in CODE environment. Journal of Intelligent Manufacturing 2007;18:497–504. doi:10.1007/s10845-007-0058-0.

[21] Tang D, Zhu R, Tang J, Xu R, He R. Product design knowledge management based on design structure matrix. Advanced Engineering Informatics 2010;24:159–66. doi:10.1016/j.aei.2009.08.005.

[22] Tang D, Zhang G, Dai S. Design as integration of axiomatic design and design structure matrix. Robotics and Computer-Integrated Manufacturing 2009;25:610–9. doi:10.1016/j.rcim.2008.04.005.

[23] ISO. 10303 AP 239 http://www.iso.org.

[24] Law Hang-wai, Hua Meng. Using quality function deployment in singulation process analysis. Engineering Letters 2007;14(1): EL_14_1_6.

[25] Bray IK. An introduction to requirements engineering. Reading: Addison-Wesley; 2002.

[26] Sommerville Ian. Integrated requirements engineering: a tutorial. IEEE Computer Society 2005.

[27] Dane B, Briand H, Barbier F. A use case driven requirements engineering process. In: Requirements Engineering, vol. 2. London Limited: Springer-Verlag; 1997. p. 79–91.

[28] Engelsman W, Jonkers H, Franken HM, Iacob ME. Architecture-driven requirements engineering. In: Proper E, Harmsen F, Dietz JLG, editors. PRET 2009. LNBIP, vol. 28. Berlin (Heidelberg): Springer-Verlag; 2009. p. 134–54.

[29] Tuunanen T, Peffers K, Hebler Simeon. A requirements engineering method designed for the blind. In: Winter R, Zhao JL, Aier S, editors. DESRIST 2010. LNCS, vol. 6105. Berlin Heidelberg: Springer-Verlag; 2010. p. 475–89.

[30] Perrouin Gilles, Brottier Erwan, Baudry Benoit, Traon YvesLe. Composing models for detecting inconsistencies: a requirements engineering perspective. In: Glinz M, Heymans P, editors. REFSQ 2009. LNCS, vol. 5512. Berlin (Heidelberg): Springer-Verlag; 2009. p. 89–103.

[31] Fabian B, Gurses S, Heisel M, Santen T, Schmidt H. A comparison of security requirements engineering methods. Requirements Engineering 2010;15:7–40. doi:10.1007/s00766-009-0092-x.

[32] Rau Hsin, Fang Yi-Tse. Conflict resolution of product package design for logistic using the TRIZ method. In: Proceedings of 8th international conference on machine learning and cybernetics. Baoding: July 2009. p. 12–5.

[33] Finch H. Comparison of distance measures in cluster analysis with dichotomous data. Journal of Data Science 2005;3:85–100.

Neurocomputing 430 (2021) 104–111

Contents lists available at ScienceDirect

Neurocomputing

journal homepage: www.elsevier.com/locate/neucom

A knowledge graph method for hazardous chemical management: Ontology design and entity identification

Xue Zheng, Bing Wang, Yunmeng Zhao, Shuai Mao, Yang Tang *

The Key Laboratory of Advanced Control and Optimization for Chemical Processes, Ministry of Education, East China University of Science and Technology, Shanghai 200237, PR China

ARTICLE INFO

Article history:
Received 27 May 2020
Revised 27 August 2020
Accepted 25 October 2020
Available online 10 November 2020
Communicated by Zidong Wang

Keywords:
Knowledge graph
Ontology
Hazardous chemicals management
Named entity recognition

ABSTRACT

Hazardous chemicals are widely used in the production activities of the chemical industry. The risk management of hazardous chemicals is critical to the safety of life and property. Hence, the effective risk management of hazardous chemicals has always been important to the chemical industry. Since a large quantity of knowledge and information of hazardous chemicals is stored in isolated databases, it is challenging to manage hazardous chemicals in an information-rich manner. Herein, we prompt a knowledge graph to overcome the information gap between decentralized databases, which would improve the hazardous chemical management. In the implementation of the knowledge graph, we design an ontology schema of hazardous chemicals management. To facilitate enterprises to master the knowledge in the full lifecycle of hazardous chemicals, including production, transportation, storage, etc., we jointly use data from companies and open data from the public domain of hazardous chemicals to construct the knowledge graph. The named entity recognition task is one of the key tasks in the implementation of the knowledge graph, which is of great significance for extracting entity information from unstructured data, namely the hazardous chemical accidents records. To extract useful information from multi-source data, we adopt the pre-trained BERT-CRF model to conduct named entity recognition for incidents records. The model achieves good results, exhibiting the effectiveness in the task of named entity recognition in the chemical industry.

1. Introduction

Hazardous chemicals are widely used in almost every corner of the chemical industries and many other manufacturing fields. They not only build up the basis of the economy but also are closely related to our daily life. Due to the flammability, explosion, high toxicity, and high corrosivity [1], hazardous chemicals have an impact on all phases of their lifecycles, such as production, transportation, and storage. Risk management of hazardous chemicals is an important topic worldwide. Hazardous chemical management includes many aspects, including hazard identification, consequence analysis, probability analysis, and risk assessment, etc.

One important reason for the frequent occurrence of hazardous chemical related incidents is the lack of relevant knowledge and poor training. Due to enormous differences in physical and chemical properties, the management of hazardous chemicals involves a vast amount of interconnected information, especially when conducted in the full life cycle. Though nontrivial data sets have been established to facilitate the hazardous chemical management, these isolated islands of information hinder the comprehensive understanding and utilization rate of the knowledge of hazardous chemicals [2]. Some information technologies have been used in the management of hazardous chemicals, like the expert system [3] in the process of safety analysis. Although the expert system is suitable for hazard identification and inference in a specific field, it cannot cover the whole life cycle of hazardous chemicals. In the field of process knowledge engineering, the integrated ontology, OntoCAPE [4] was developed. OntoCAPE defined the overall schema of the process industry including the meta layer, the upper layer, the conceptual layer, and the application layer, etc. OntoCAPE is an important comprehensive ontology that is not applied to process data.

Knowledge graph, as a new type of graph database content retrieval method proposed by Google in 2012, has developed vigorously and effectively so far [5]. As a semantic network, knowledge graph has powerful expressive ability and modeling flexibility, and

* Corresponding author.
E-mail addresses: tangtany@gmail.com, yangtang@ecust.edu.cn (Y. Tang).

https://doi.org/10.1016/j.neucom.2020.10.095
0925-2312/

it can model entities, concepts, attributes, and their relationships [6,7]. The knowledge graph is promising in knowledge retrieval [8], question-answering [9], knowledge recommendation [10], knowledge visualization [11], and other applications. The knowledge graph has been applied in many fields and have achieved good results [12–14]. For instance, some scholars [15] have tried to apply knowledge graphs in the field of traditional Chinese medicine (TCM) health care and have expanded the scale of the knowledge graph. The platform supported by knowledge graph can provide non-professionals with knowledge services such as knowledge retrieval of traditional Chinese medicine [15]. Besides, the knowledge graph has been applied in the field of geological hazards. By constructing the knowledge graph of geological hazard documents, scholars have improved the utilization rate of literature information and provided knowledge services and knowledge bases for preventing and responding to geological hazards [16]. However, the use of a knowledge graph in the field of hazardous chemical management is challenged by the complex interconnections between various risk influencing factors as well as the overwhelming amount of interconnected information.

There are unstructured documents in the field of hazardous chemicals management, and the knowledge graph should be established on the embedded entities and relations within these documents. The named entity recognition is one of the key tasks in the implementation of the knowledge graph. Different entities need to be identified following an overall schema. The mainstream methods of identifying entities are divided into three categories: rule-based methods, learning-based methods, and hybrid methods [17]. The most basic of rule-based methods are dictionary-based entity recognition. For the good performance of deep learning in natural language processing, most of the learning-based methods are implemented using deep learning.

Knowledge and data of hazardous chemicals are usually stored in separate tables. We try to establish a new knowledge graph to build connections among a substantial volume of chemical companies, chemicals, hazards, accidents, and other types of related knowledge. One major challenge in constructing a knowledge graph for hazardous chemicals management is how to build a appropriate ontology structure. A critical part is defining categories, relations, and attributes. Another major challenge is how to make better use of textual information in related documents, namely efficient entity identification. In this article, we propose an ontology framework for hazardous chemicals management, which provides the foundation and solution for improving process safety. We also carry out named entity recognition based on deep learning methods to identify entities in textual information so that the information can be better utilized in the chemical industry.

The main contributions of this paper are as follows:

- The contribution of introducing the knowledge graph to the chemical industry is linking the corresponding knowledge in the unstructured data source together and establishing a knowledge network full of connections. Besides, the application of the knowledge graph is beneficial to the identification of risk sourcing and propagation.
- Another contribution of this article is improving the utilization of text information in chemical documents. We apply the natural language processing technology to chemical data to solve the problem of identifying entities in chemical documents.

The paper is organized as follows. In Section 2, a framework of the proposed ontology for hazardous chemical management is given. Section 3 describes the establishment of the ontology-based knowledge graph. Section 4 presents the method that we adopt to recognize named entities in the chemical industry. Some concluding remarks are finally given in Section 5.

2. Ontology development in a top-down manner

The first step is the design of ontologies for the knowledge graph of hazardous chemical management from accumulated data resources and human knowledge. Data resources can be divided into structured data and unstructured data [18]. As shown in Fig. 1, the overall architecture is a combination of top-down and bottom-up methods: at first, the ontology is proposed based on human knowledge and experiences, then rules are applied to form the early entity maps by integrating exiting structural data.

The most important task of constructing a knowledge graph is to design ontology [19,20]. Before designing the ontology, we first determined the scope. The hazardous chemical management involves chemical production, storage, transportation, usage, treatment, etc. The class hierarchy should consider the concepts of company, equipment, hazardous chemical, human, incident, risk, etc. The subject concept was divided into the company, equipment, hazardous chemical, person, and incident. The typical relations between the listed classes included chemical reaction, production, and consumption, etc.

Based on the provided concepts, we first designed seven top-level classes. And subclasses were then added to the top-level classes. The structure of classes in ontology is shown in Fig. 2. Secondly, we defined the relation between classes, namely, the object attribute. For example, we defined the object attribute "relatedIncChe" (related incident chemical) between the class Incident and Chemical, the domain of the object attribute was Incident, and Chemical was the range of the object attribute. More relationships between classes are shown in Fig. 3. At last, we defined the properties of each class, also known as data properties. Fig. 4 shows examples of data properties. On the left, we defined the following data attributes for the chemical class: "hasCASNumber", "hasFormula", "Hasupperhlrelimit", etc. The right subgraph in Fig. 4 shows the data attributes of the company class: "hasTaxFileNumber", "hasEmail", "has Address", etc.

After defining the object properties and data properties, we used the properties to add constraints to the classes in ontology. As shown in Fig. 5, the description of a class includes an existential quantifier description (some) and a full quantifier description (only). For example, "relatedIncOrg some company" indicating that the class has a relationship associated with the company, and "incident and (relatedIncChe only (hazardous chemical))" indicating hazardous chemicals incidents are incidents only associated with hazardous chemicals.

Fig. 1(a) shows six types of structured data resources in the project. Among them, chemical attribute data are open data obtained from the chemical registration catalog. Related companies, major hazards evaluations, and incidents data are data provided by enterprises. All datasets are arranged in CSV files or tables. Tables 2 and 4 show the size and type of the datasets. The content of the chemical registration dataset is shown in Table 1, and the content of data such as companies and hazardous source incidents is shown in Table 3. Some of the datasets describe the attributes of the entity, such as "Hazardous chemicals catalog 2015". Some datasets also include related information between entities, such as "Dataset of chemicals and enterprise related to hazard sources".

3. Completion of ontology in a bottom-up way

We adopted a rule-based method to map the structured data resources with a knowledge graph and establish a mapping relationship between concepts in the database and ontology in the knowledge graph. We used different extraction rules to achieve semi-automatic extraction of database entities, attributes, and relations from different data structures. The data sets in the field of hazardous chemicals have the following characteristics: many

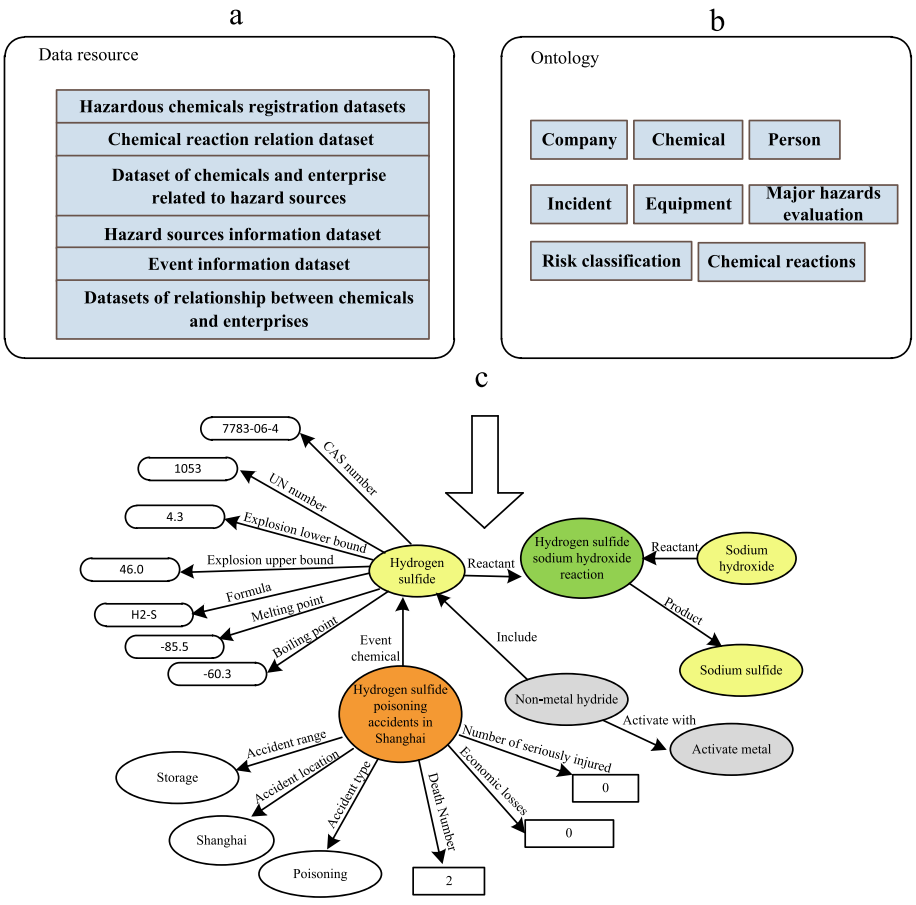


Fig. 1. Technical framework for building a knowledge graph of hazardous chemicals management.

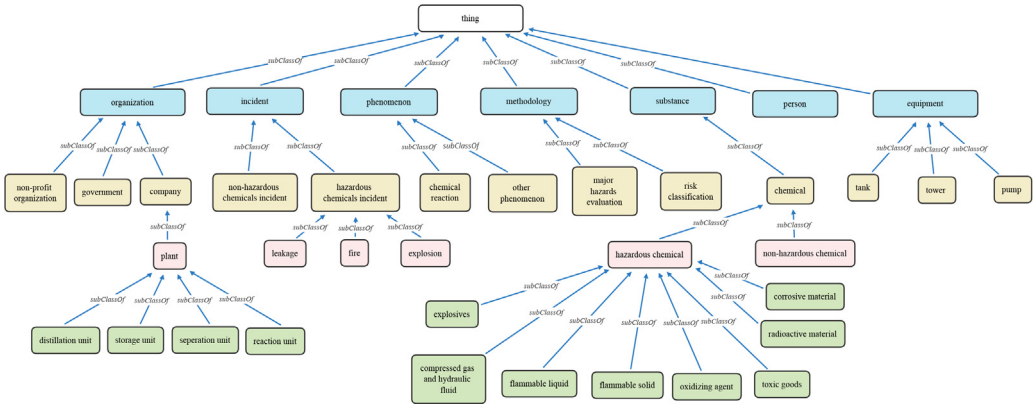


Fig. 2. The hierarchy of classes in the hazardous chemical ontology.

aliases and common names of chemicals, inconsistent names of chemicals in different data sets, and error and missing record in CAS number file. The above characteristics lead to the information island problem in the chemical industry. Compared to the non-exclusive names, the CAS number is an important basis for determining chemicals. Thus, to solve the above problems, we created and updated the CAS number file of chemicals, and we used the chemical name in the CAS number file as the standard chemical name. Also, we created and updated the chemical alias set.

With the aid of the integration of the structured data, we obtained a knowledge graph of hazardous chemicals management, which contained 124,593 attributes, 66,184 entities, and 223,640 relationships. However, the current graph construction had the following deficiencies:

- The amount of data was small, and the external chemical resources were not fully utilized. Various information such as chemical names, devices, processes, and accident types involved in the chemical process were included in the text, requiring more effort to identify safety-related entities.
- The graph did not completely describe the contents of the table, and the utilization rate of the table information could be further improved.
- The manual construction of the graph was inefficient.

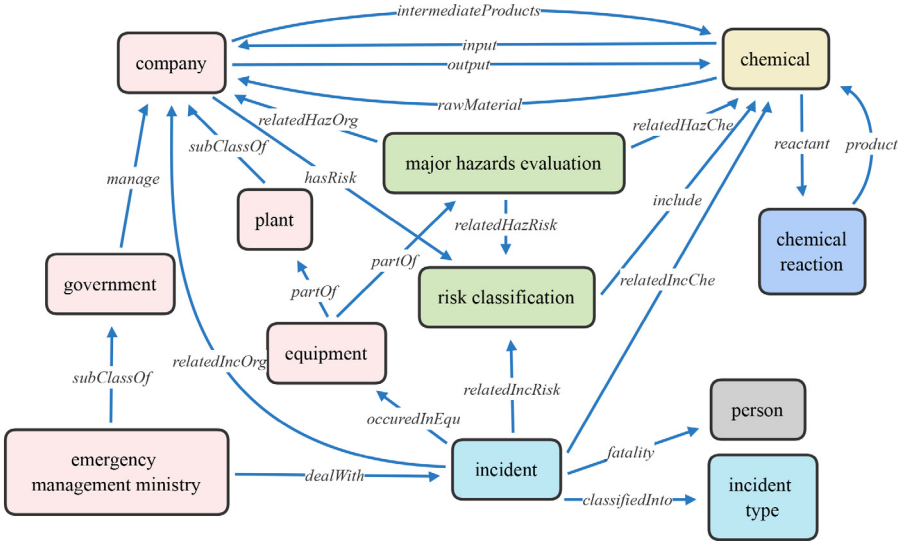


Fig. 3. The example of relations between classes in the hazardous chemical ontology.

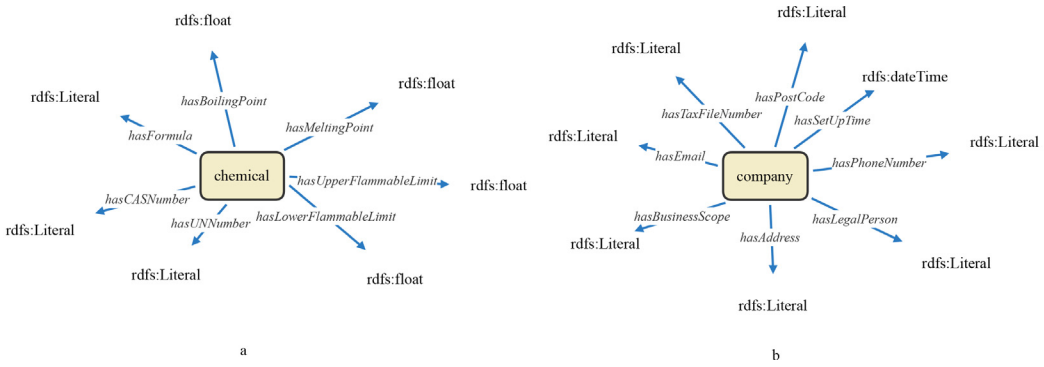


Fig. 4. Two examples of data properties of the class.

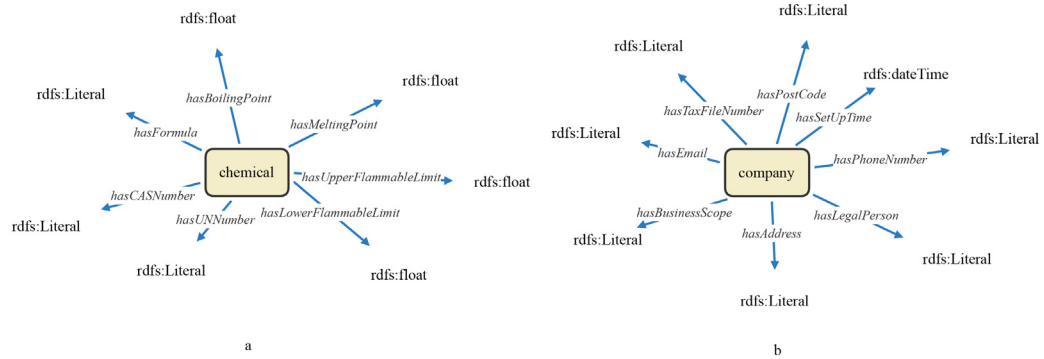


Fig. 5. Two examples of axioms in defining the class.

To overcome these deficiencies, the automation and efficiency of the construction of knowledge graphs need to be improved. Information extraction can automatically [21,22] or semi-automatically extract useful triples from the text, improving the speed and efficiency of constructing knowledge graphs. The information extraction task is divided into three steps: firstly, the named entities are correctly identified; secondly, the relationships between concepts are described; finally, the relationships are nicely classified. Therefore, the accuracy of the information extraction task depends on the correct named entity [23,24]. Named entity recognition is an important step in the information extraction task. Our goal in the next section is to train named entity recognizers to prepare for subsequent tasks such as relationship extraction and knowledge question answering. Specific identification methods will be developed in the next section.

4. Deep learning-powered named entity recognition

The representation of chemical knowledge includes structured, unstructured, semi-structured data. The relational database is the

Table 1
Hazardous chemicals registration datasets and content.

Hazardous chemicals registration dataset	Chemical attributes in related Database
Hazardous chemicals catalog 2015	CAS number, Chemical Chinese name
Dangerous goods list	UN dangerous goods number, Chemical English name, Chemical Chinese name, Category, Special dangers
Occupational disease classification and catalog	Chemical Chinese name, Occupational disease
List of hazardous chemicals under key supervision	CAS number, Chemical Chinese name
List of highly toxic substances 2003	CAS number, Chemical name, Chemical Chinese name, Chinese alias
Catalog of categories and varieties of precursor chemicals	Chinese name of chemical, Category
Catalogue of hazardous chemicals under key environmental management 2014	CAS number, Chinese alias,Chinese name of chemical
Explosive hazardous chemicals list 2011	name, CAS number, explosive danger category, un_num, chemId, classification

Table 2
The size and type of hazardous chemicals registration datasets and content.

Hazardous chemicals registration dataset	Size	Type
Hazardous chemicals catalog 2015	2997 items	csv
Dangerous goods list	2505 items	csv
Occupational disease classification and catalog	136 items	csv
List of hazardous chemicals under key supervision	76 items	csv
List of highly toxic substances 2003	54 items	csv
Catalog of categories and varieties of precursor chemicals	41 items	csv
Catalogue of hazardous chemicals under key environmental management 2014	84 items	csv
Explosive hazardous chemicals list 2011	99 items	csv

Table 3
Datasets and content provided by enterprises.

Datasets provided by the enterprise	Chemical attributes in related Database
Enterprise information	User enterprise ID, Unit name, Address, Province ID, City ID, County ID, Nature units, Post code, Business license number, Manager name, Phone number, Fax number, Mail, Unit code, Set-up time, Production range, Representative, etc.
Datasets of relationship between chemicals and enterprises	Unit name, Chemical code, Chemical property, Name, Alias name, English name, English alias name, CAS number, UN number, Formula
Dataset of chemicals and enterprise related to hazard sources	Hazard sources company, Danger source ID, Hazard source chemicals
Hazard sources information	Company ID, Hazard source name, Hazard source address, Hazard source level, Hazard source R value, Hazard source scale, Safety distance, People number, etc.
Event information Involved chemicals, Hazardous chemicals, Accident type, Accident level, etc.	Accident name, Domestic or international, Accident location,

Table 4
The size and type of datasets provided by enterprises.

Datasets provided by the enterprise	Size	Type
Enterprise information	47067 items	csv and unstructured data
Datasets of relationships between chemicals and enterprises	246562 items	csv
Dataset of chemicals and enterprise related to hazard sources	25541 items	csv
Hazard sources information	11295 items	csv
Event information	25522 items	csv and unstructured data

most common structured data format, and web documents are common semi-structured data types such as XML, JSON, etc. However, human-readable documents are one of the main data sources in the hazardous chemical industry. In Section 2 and Section 3, the ontology and knowledge graph have been constructed based on the structured data by combining the top-down and bottom-up approaches. For unstructured data in the chemical industry, such as position statement, operating procedures, and accident information, the identification of the entity is essential, and a named entity reorganization model is trained based on ontology definitions.

In this work, the original corpus contains five types of entities: chemicals, event type, enterprise organization, chemical equipment, and chemical operation system, all of which are derived from available documents such as job data, operating procedures, and accident information. For the available documents, we first performed sentence segmentation processing. We collected a total of 12,689 original corpus sentences. The training set contained a

Table 5
The number of entities in datasets.

Datasets	Chemicals	Accident type	Enterprise organization	Chemical equipment	Chemical operation system	Total number
Training dataset	7314	8263	4276	4171	1742	25,766
Validation dataset	1325	1326	731	672	302	4347
Testing dataset	1138	1369	722	604	309	4142

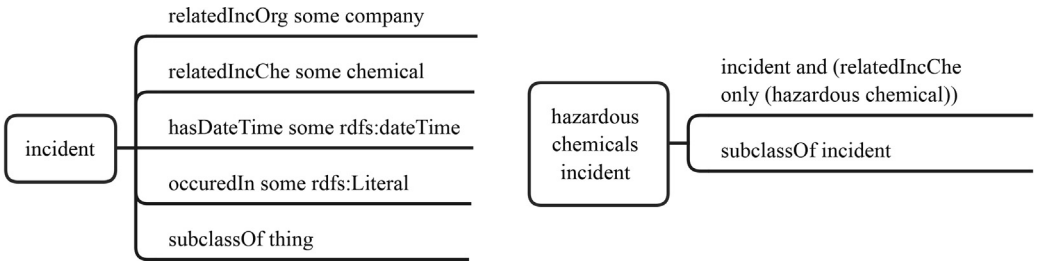


Fig. 6. Picture of the process of data preprocessing.

total of 9649 sentences, the verification set contained a total of 1577 sentences, and the test set contained a total of 1463 sentences. The dataset statistics are shown in Table 5.

4.1. Data preprocessing

The labeling strategies for named entity recognition include BIO mode, BIOE mode, and BIOS mode. We used the BIO labeling strategy, where B represents the beginning of the entity, I represents the non-starting part of the entity, and O represents the part that is not the entity in the sentence. When predicting the entity boundary, it is necessary to predict the entity type at the same time. So there are eleven types of labels to be predicted, namely "O", "B-CHE", "I-CHE", "B-ETP", "I-ETP", "B-ORG", "I-ORG ", "B-EQU", "I-EQU", B-SYS, I-SYS. CHE, ETP, ORG, EQU and SYS indicate the corresponding chemicals, event type, enterprise organization, chemical equipment and chemical operation system, respectively. The entire data processing diagram is shown in Fig. 6. The named entity identification in the hazardous chemicals industry was regarded as a sequence labeling problem. We organized the available documents related to hazardous chemicals, such as operating procedures and accident records. Then we preprocessed the documents into a training standard format through sentence segmentation and character splitting. The classic BIO labeling scheme was used to manually label industrial corpus sentences in the available documents. An example of labeling the original corpus data is shown in the right half of Fig. 6. The first part of each line is a word, and the second part is a word label. The word and the word label are separated by spaces. Sentences are separated by blank lines.

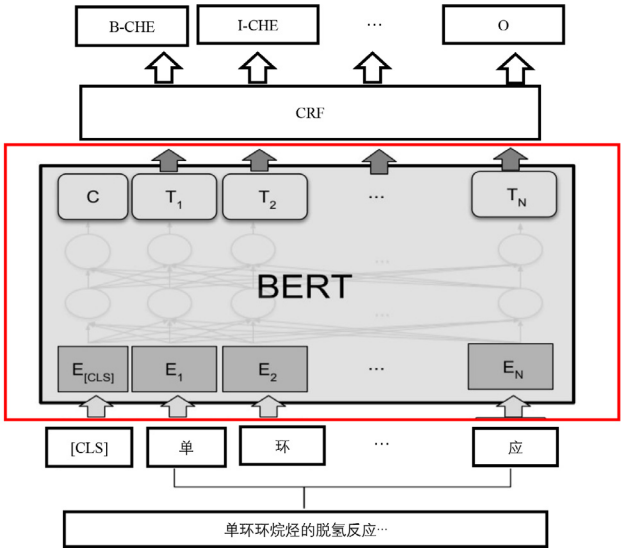


Fig. 7. This figure describes the structure of the BERT + CRF model, and the part in red box comes from the google [25].

4.2. Method

In the research of chemical named entity recognition, we took up a hybrid method to identify entities, and this model is based on bidirectional encoder representation from transformers (BERT) model and conditional random field (CRF) model. First, a deep neural network was used to obtain a deep representation of sentence semantics, and then through the constraint function of the CRF layer, the maximum probability sequence was output.

The overall structure is shown in Fig. 7. The entire model is divided into two parts. Among them, the BERT model [25] is a pre-trained language model that is trained from a large number of text corpora by using unsupervised training methods. The conditional random field [26] has been widely used in sequence labeling for a long time, and it is an undirected probability graph discriminant model. In this work, Chinese sentences were first split into single characters, each character was mapped to a character id by the dictionary that came with the BERT model, and then the character id was transformed into an embedding vector with complex semantics through the embedding layer. Then the word vector sequence was input to the CRF layer. The CRF layer [26] can learn the constraint conditions of the sentence and improve the accuracy of the prediction sequence.

4.3. Experimental parameters and results

Google provides two pre-trained language models: BERT-Base and BERT-Large. The network structures of these two models are the same, with only some parameters being different. It takes more graphics card memory to train BERT-Large. In contrast, BERT-Base model requires less memory and the accuracy of training has met our needs. Thus, we adopted the BERT-Base model. BERT-Base model had a total of 12 layers, and the hidden layer was 768 dimensions. We adopted a 12-head mode with a total of 110M

Table 6
The comparison of results on test datasets between different models. "Avg." means the average score.

Model	Evaluation	CHE	ETP	ORG	EQU	SYS	Avg.
Rule-based method	P	0.5382	0.3478	0.4212	0.6372	0.7480	0.4850
	R	0.4785	0.2655	0.8486	0.6527	0.9194	0.5309
	F	0.5066	0.3011	0.5630	0.6449	0.8249	0.4924
CRF [28]	P	**0.9643**	0.9039	0.9590	**0.9397**	**0.9605**	0.9395
	R	0.9500	0.9338	0.9500	0.9579	0.9419	**0.9452**
	F	**0.9571**	0.9186	0.9545	**0.9488**	**0.9511**	0.9423
BiLSTM + CRF [29]	P	0.9326	0.9171	0.8829	0.9052	0.8984	0.9123
	R	0.9500	**0.9396**	0.8865	0.9257	0.9129	0.9292
	F	0.9412	0.9282	0.8847	0.9153	0.9056	0.9206
BERT + CRF	P	0.9500	**0.9455**	**0.9623**	0.9190	0.8830	**0.9411**
	R	0.9508	0.9248	**0.9557**	**0.9586**	**0.9618**	0.9450
	F	**0.9504**	**0.9350**	**0.9590**	0.9384	0.9207	**0.9428**

parameters. The maximum sequence length was 230, the train batch size parameter was 32, the learning rate was 1e−5, the droop out rate parameter was 0.5.

We adopted precision (P), recall (R), and F indicators as our evaluation criteria, the three indicators can be calculated by true positives (TP), false positives (FP), and false negatives (FN). The specific calculation formula is as follows:

$$Precision = \frac{TP}{TP + FP} \tag{1}$$

$$Recall = \frac{TP}{TP + FN} \tag{2}$$

$$F = 2 \cdot \frac{Precision \cdot Recall}{Precision + Recall} \tag{3}$$

Simply, the precision is computed by our prediction results. It indicates the percentage of positive samples that are true positive samples. The recall rate indicates the percentage of positive examples in the sample that are predicted correctly. The F value combines the results of accuracy and recall [27].

The performance of these four methods on the test set is shown in Table 6. Among them, the rule-based method was based on the dictionary, which consisted of entities that appear in the training set. Besides, the external chemical dictionary was also used to identify chemical entities. The experiment result proved that the dictionary-based method was not effective in the test set of the chemical industry. The main reason is that the dictionary in the chemical industry is not complete. The other three models achieved good results. The test results of the best BERT model were as follows: the precision rate was 0.9411, the recall rate was 0.9450, and the F1 value was 0.9428. From the perspective of the model effect, the BERT model performed better than the BiLSTM model in the chemical dataset. In some entity types, the CRF model achieved superior results. For example, in terms of chemical entities, the CRF model was more accurate than the BERT model. However, in the training process, the CRF model needed to carefully select the feature template. In general, the BERT model did not require manual feature selection and performed well in named entity recognition tasks in the chemical industry.

In this section, we applied the methods in natural language processing to the chemical industry. Aiming at the large number and constant update of chemical names in the chemical industry, we used a combination of a pre-trained model and a probability graph model to identify entities in the text of the chemical industry. The experimental verification showed that the method had a significant effect on the test set.

5. Conclusion

Our work aims to extract chemical-related named entities from the considerable body of the unstructured document and build a hazardous chemical management knowledge graph. In this paper, an ontology for risk management of hazardous chemicals is designed. Besides, a deep learning method is adopted to identify named entities in the chemical industry, which greatly improves the effectiveness of named entity recognition. The method achieves the highest precision of 0.9411, recall rate of 0.9450, and an F1 score of 0.9428.

Compared with traditional data storage methods, the knowledge graph connects chemical industry-related datasets, laying a foundation for knowledge services in the chemical industry. The proposed hazardous chemicals ontology framework can provide basic support for information integration and inference. Named entity recognition lays the foundation for chemical corpus processing and chemical knowledge graph question and answer application, etc., leading to a significant improvement in the utilization of chemical-related document information.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgements

This work is supported in part by the National Key Research and Development Program of China under Grant 2018YFC0809302, the National Natural Science Foundation of China under Grants 61988101, 61751305, 61673176 and the Programme of Introducing Talents of Discipline to Universities (the 111 Project) under Grant B17017.

References

[1] B. Wang, C. Wu, G. Reniers, L. Huang, L. Kang, L. Zhang, The future of hazardous chemical safety in China: opportunities, problems, challenges and tasks, Sci. Total Environ. 643 (2018) 1–11.
[2] A. Menon, N.B. Krdzavac, M. Kraft, From database to knowledge graph—using data in chemistry, Curr. Opin. Chem. Eng. 26 (2019) 33–37.
[3] J. Zhao, L. Cui, L. Zhao, T. Qju, B. Chen, Learning hazop expert system by case-based reasoning and ontology, Comput. Chem. Eng. 33 (1) (2009) 371–378.
[4] J. Morbach, A. Yang, W. Marquardt, Ontocape—a large-scale ontology for chemical process engineering, Eng. Appl. Artif. Intell. 20 (2) (2007) 147–161.
[5] W. Liu, J. Liu, M. Wu, S. Abbas, W. Hu, B. Wei, Q. Zheng, Representation learning over multiple knowledge graphs for knowledge graphs alignment, Neurocomputing 320 (2018) 12–24.
[6] W. Li, R. Peng, Y. Wang, Z. Yan, Knowledge graph based natural language generation with adapted pointer-generator networks, Neurocomputing 382 (2020) 174–187.

[7] Q. Wang, Y. Hao, ALSTM: An attention-based long short-term memory framework for knowledge base reasoning, Neurocomputing (2020).
[8] X. Zhang, X. Hou, X. Chen, T. Zhuang, Ontology-based semantic retrieval for engineering domain knowledge, Neurocomputing 116 (116) (2013) 382–391.
[9] S. Zhu, X. Cheng, S. Su, Knowledge-based question answering by tree-to-sequence learning, Neurocomputing 372 (2020) 64–72.
[10] Z. Sun, J. Yang, J. Zhang, A. Bozzon, L.-K. Huang, C. Xu, Recurrent knowledge graph embedding for effective recommendation, in: Proceedings of the 12th ACM Conference on Recommender Systems, 2018, pp. 297–305.
[11] X. He, R. Zhang, R. Rizvi, J. Vasilakes, X. Yang, Y. Guo, Z. He, M. Prosperi, J. Bian, Prototyping an interactive visualization of dietary supplement knowledge graph, in: 2018 IEEE International Conference on Bioinformatics and Biomedicine (BIBM), IEEE, 2018, pp. 1649–1652.
[12] Y. Liang, F. Xu, S.-H. Zhang, Y.-K. Lai, T. Mu, Knowledge graph construction with structure and parameter learning for indoor scene design, Comput. Visual Media 4 (2) (2018) 123–137.
[13] H. Weng, Z. Liu, S. Yan, M. Fan, A. Ou, D. Chen, T. Hao, A framework for automated knowledge graph construction towards traditional Chinese medicine, in: International Conference on Health Information Science, Springer, 2017, pp. 170–181.
[14] P. Zhu, W. Zhong, X. Yao, Auto-construction of course knowledge graph based on course knowledge, Int. J. Perform. Eng. 15 (8) (2019).
[15] T. Yu, J. Li, Q. Yu, Y. Tian, X. Shun, L. Xu, L. Zhu, H. Gao, Knowledge graph for TCM health preservation: design, construction, and applications, Artif. Intell. Med. 77 (2017) 48–52.
[16] R. Fan, L. Wang, J. Yan, W. Song, Y. Zhu, X. Chen, Deep learning-based named entity recognition and knowledge graph construction for geological hazards, ISPRS Int. J. Geo-Inf. 9 (1) (2020) 15.
[17] M. Xiaofeng, W. Wei, X. Aiping, Incorporating token-level dictionary feature into neural model for named entity recognition, Neurocomputing 375 (2020) 43–50.
[18] Y. Jia, Y. Qi, H. Shang, R. Jiang, A. Li, A practical approach to constructing a knowledge graph for cybersecurity, Engineering 4 (1) (2018) 53–60.
[19] J. Dou, J. Qin, Z. Jin, Z. Li, Knowledge graph based on domain ontology and natural language processing technology for Chinese intangible cultural heritage, J. Visual Lang. Comput. 48 (2018) 19–28.
[20] M. Pietranik, N.T. Nguyen, A multi-attribute based framework for ontology aligning, Neurocomputing 146 (2014) 276–290.
[21] S. Zheng, Y. Hao, D. Lu, H. Bao, J. Xu, H. Hao, B. Xu, Joint entity and relation extraction based on a hybrid neural network, Neurocomputing 257 (2017) 59–66.
[22] J. Zhang, Y. Zhang, D. Ji, M. Liu, Multi-task and multi-view training for end-to-end relation extraction, Neurocomputing 364 (2019) 245–253.
[23] A. Goyal, V. Gupta, M. Kumar, Recent named entity recognition and classification techniques: a systematic review, Comput. Sci. Rev. 29 (2018) 21–43.
[24] M.W. Alnabki, E. Fidalgo, E. Alegre, L. Fernandezrobles, Improving named entity recognition in noisy user-generated text with local distance neighbor feature, Neurocomputing 382 (2020) 1–11.
[25] J. Devlin, M.-W. Chang, K. Lee, K. Toutanova, Bert: Pre-training of deep bidirectional transformers for language understanding, arXiv preprint arXiv:1810.04805 (2018).
[26] A. Chen, F. Peng, R. Shan, G. Sun, Chinese named entity recognition with conditional probabilistic models, in: Proceedings of the Fifth SIGHAN Workshop on Chinese Language Processing, 2006, pp. 173–176.
[27] W. Li, W. Song, X. Jia, J. Yang, Q. Wang, Y. Lei, K. Huang, J. Li, T. Yang, Drug specification named entity recognition base on BiLSTM-CRF model (2019) 429–433.
[28] N. Sobhana, M. Pabitra, S. Ghosh, Conditional random field based named entity recognition in geological text, Int. J. Comput. Appl. 1 (02 2010). doi:10.5120/72-166.
[29] G. Lample, M. Ballesteros, S. Subramanian, K. Kawakami, C. Dyer, Neural architectures for named entity recognition, in: Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Association for Computational Linguistics, San Diego, California, 2016, pp. 260–270. doi:10.18653/v1/N16-1030. https://www.aclweb.org/anthology/N16-1030.

Xue Zheng received the B.S. degree in the school of control science and engineering from Qingdao University, in 2018. She is currently pursuing an M.A. degree from East China University of Science and Technology. Her research interests include knowledge graph, natural language processing, and their applications in chemical engineering.

Bing Wang received the B.S. and Ph.D. degree in chemical engineering from Tsinghua University, Beijing China in 2011 and 2016, respectively. From 2014 to 2015, he was a Visiting Scholar with the Mary Kay O'Connor Process Safety Center in Texas A&M University, College Station, United States. After graduation, he has become an assistant professor in East China University of Science and Technology, Shanghai. Dr. Bing Wang's research interests includes atmospheric dispersion modeling, quantitative risk assessment to chemical spills, source term estimation for air pollution dispersion, knowledge base on chemical process safety.

Yunmeng Zhao received his B.S. in Chemistry from Nanjing University in 2014, and his PhD in Chemical Engineering from Monash University in 2019. He is currently an assistant professor at East China University of Science and Technology. His research interests are soft electronics, wearable sensors, and machine learning.

Shuai Mao received the B.S. degree in school of control science and engineering from East China University of Science and Technology, in 2017. He is currently pursuing the Ph.D. degree from East China University of Science and Technology. His research interests include multi-agent systems, distributed optimization and their applications in practical engineering.

Yang Tang received the B.S. and Ph.D. degrees in electrical engineering from Donghua University, Shanghai, China, in 2006 and 2010, respectively. From 2008 to 2010, he was a Research Associate with The Hong Kong Polytechnic University, Hong Kong. From 2011 to 2015, he was a Post-Doctoral Researcher with the Humboldt University of Berlin, Berlin, Germany, and with the Potsdam Institute for Climate Impact Research, Potsdam, Germany. Since 2015, he has been a Professor with the East China University of Science and Technology, Shanghai. His current research interests include distributed estimation/control/optimization, cyber-physical systems, hybrid dynamical systems, computer vision, reinforcement learning and their applications. Prof. Tang was a recipient of the Alexander von Humboldt Fellowship and the ISI Highly Cited Researchers Award by Clarivate Analytics from 2017. He is a Senior Board Member of Scientific Reports, an Associate Editor of IEEE Transactions on Neural Networks and Learning Systems, IEEE Transactions on Emerging Topics in Computational Intelligence and IEEE Systems Journal, etc.

Head Office:
SciBite Limited
BioData Innovation Centre
Wellcome Genome Campus
Hinxton, Cambridge CB10 1DR
United Kingdom

www.scibite.com
contact@scibite.com
Linkedin: SciBite
Twitter: @SciBite
+44 (0)1223 786 129