Available online at www.sciencedirect.com

ScienceDirect

Procedia Manufacturing 52 (2020) 307–312

Procedia MANUFACTURING

www.elsevier.com/locate/procedia

5th International Conference on System-Integrated Intelligence

Big Data Management Using Ontologies for CPQ Solutions

Alexander Binder[a], Eva-Maria Iwer[a], Werner Quint[a]*

[a]*Faculty of Design – Computer Science – Media, Rhein Main University of Applied Sciences, Unter den Eichen 5, 65195 Wiesbaden, Germany*

* Corresponding author. Tel.: +49 611 9495 2146; E-mail address: werner.quint@hs-rm.de

Abstract

In recent years, due to a progressive complexity of handling and processing business data, proper big data management has become a challenge, especially for SMEs that have limited resources for investing in the requested business transformation process.
As a solution, we suggest an ontology-based CPQ software approach, where we show how the implementation of semantic technologies and ontologies affects data integration processes. We also propose a method called "ontology-based data matching", which allows the semiautomatic generation of alignments used to formalize the coherence between ontologies.
The proposed method will ensure consistency during integration, significantly improving the productivity of enterprises.

Keywords: CPQ; Semantic Technologies; Ontologies; Ontology Matching; Data Quality

1. Introduction

The establishment of electronic business transactions has granted several improvements to enterprises, such as the automatic handling of order processes. Nowadays, product data and business processes are almost exclusively managed within the enterprises' information systems.

However, enterprises are challenged by the complexity increase required to handle more and more electronic data and processes. Frequently, this issue is being discussed in the context of big data management and chiefly product data.

The term "big data" describes large and complex data sets that traditional data applications are unable to process adequately. These data sets are often scattered across many different information systems. For example, Enterprise Resource Planning (ERP) systems are being set up to represent all business processes of an enterprise, in order to increase the overall cost effectiveness. In many cases, ERP systems do not include large data sets about products, e.g., product marketing descriptions, product pictures, or complementary technical product data. These data types are managed together with other types of information systems, e.g., Product Data (PDM), Customer Relationship (CRM), Content (CMS), and Product Information Management (PIM) systems. [1]

Other types of applications handling product data are Configure, Price and Quote (CPQ) solutions. They support companies that have to walk a fine line between profit margin and market share to adopt more data-aware systems by quoting complex and configurable products.

Many information systems are not fully capable of encompassing all data management needs of an enterprise. On the other hand, not all companies are able to implement proper information systems into their technology structure, as processes for data and business process integration are too costly and time-consuming for them. These issues can lead to data redundancy and a high number of inconsistencies within product information datasets of an enterprise.

In particular, Small and Medium-sized Enterprises (SMEs) face the problem that their product information is not well structured. Hence, high expenses can arise for maintenance, search, and presentation of product information. Also, the number of customer requests, wrong orders, and deliveries

10.1016/j.promfg.2020.11.051

may increase. Therefore, it is necessary to implement semantical restrictions or defaults to avoid redundant statements or wrong interpretations in complex data sets. [2]

Due to mass customization, the complexity of information about products increases with the products' complexity and with customers' demand for individuality. However, product information is created by specialized departments or its employees, that may have differentiating and individual perspectives on information, so it is not always consistent.

Furthermore, product information created outside of a company has to be managed, passed on, and taken over by other companies, e.g., suppliers, manufacturers, as well as by individuals like customers. However, companies might avoid data sharing and linking because of national-specific vocabulary and language for product data.

Additionally, if the product is highly configurable, the users may face combinatorial overload caused by the rapid complexity growth. Configuration engines needs to be employed to alleviate the problem.

For example, when composing a configurable product with a CPQ software, selecting a certain product feature may exclude other features. This information is essential not only for the employees, who are responsible for maintaining product data, but also for the customers, who want to configure a product or satisfy their needs for information.

Accordingly, it is necessary to capture, manage, and present the complex information in a clear and understandable way. Lastly, it has not yet been analyzed whether existing solutions for data and business process integration in big data management, consider different user perspectives, and, if they do, how satisfying they are for the users.

The necessary data integration processes and information systems can be improved by employing methods from the field of semantic technologies and ontologies, and by organizing data within one enterprise-wide centralized information system, that enables the company to handle big data and make use of it. Therefore, this paper is analyzing three application cases of complex and large product data integration and information system contexts.

We suggest an ontology-based CPQ software approach, where we show how the implementation of semantic technologies and ontologies affects data integration processes. We also propose an ontology-based data matching method, which allows the semiautomatic generation of alignments used to formalize the coherence between ontologies.

Section 2 sketches the approach for using ontologies and summarizes the related work regarding information systems, semantic technologies, and CPQ software. A detailed view on the examined application cases is presented in Section 3. A critical discussion of existing approaches as well as our approach is given in Section 4. The conclusion of the paper is then being presented in Section 5.

2. Background & Related Work

Due to the complexity of today's product data sets and their interconnections, new approaches will have to be applied to gain value from big data. When developing a method for handling data management, it will have to include data analysis, capture, querying, sharing, storage, visualization, or updating, as well as suitable algorithms for modelling and representation of data and individually customizable mechanisms for data linking in dynamic datasets.

2.1. Big Data

There are three acknowledged dimensions of big data, on which technology leaders have to focus: information volume, variety and velocity. While "volume" refers to the steady increase in the amount of data that has to be processed, "variety" focuses on the growing amount of different data types, like tabular data (databases), hierarchical data, documents, e-mails, metering data, videos, images, audios, stock tickers, or financial transactions. "Velocity" means how fast data is being produced, but also how fast it has to be processed in order to meet the demands of the stakeholders. Radio-Frequency IDentification (RFID) tags, sensors, and smart meters as well as Internet of Things (IoT) devices of all kinds are driving the need to deal with these torrents of data in near-real time. Another acknowledged dimension to measure the reliability of data is "veracity". Since data sets arrive from different sources connected to base data, they may refer to additional product information provided, e.g., by suppliers or different forms of information, like pictures, videos, or linked social media websites and analytic databases, and thus may not fully fulfil the required quality standards. [3]

2.2. Semantic Technologies and Ontologies

Semantic technologies help to interpret information by identifying the corresponding context when creating and representing complex information and interrelations, and thus make it easier to understand the meaning and purpose of data (e.g., symbols, words, etc.) and complex concepts, as well as to share knowledge between humans and machines. They can be based on metadata, which can be linked to other metadata in different data sources. The representation of metadata in a formal way requires standardized rules and is a prerequisite for metadata exchanging between information systems, applications, and workstations. Therefore, the Resource Description Framework (RDF) can be utilized, which also serves to describe resources in a web context.

For information systems, semantic technologies can be based on simple approaches, such as glossaries (lists of words and their definitions), taxonomies (hierarchies for terms), and thesauri (relations of similarity and synonyms) to avoid syntactical and semantical problems when creating and interpreting product data. Approaches with more semantic richness are topic maps and ontologies.

Ontologies are defined as an "explicit specification of a conceptualization" [4]. They allow definitions of concepts and relationships between them, while the specification representation provides a formal semantic. Of all common approach for knowledge representation, they offer the highest

degree of semantic richness. To provide the formal semantics of the specification, Ontologies use mathematical logic, which allows the inference of new knowledge. [5]

In the context of big data, semantic web technologies and linked data are used to turn data into knowledge. They have been adopted in scientific domains. In bioinformatics, several ontology-based systems like the gene ontology have helped researchers from different countries to communicate with each other and interlink their research data.

Besides describing product data, these semantic technologies can also be utilized to capture and represent their relations and connections to other products, product components, product features and further information. With ontologies, it is also possible to represent rules, which are associated with the aforementioned product relations.

2.3. CPQ software

As it is dealing with configuration, CPQ software faces complex challenges of combining components and parts into a more viable product. There are three main approaches used to alleviate the problem of combinatorial explosion.

Rule-based truth-maintenance systems, launched in the 1970s, were the first configuration engines. They were based on researches in artificial intelligence from the 1960s.

Constraint satisfaction engines, developed in the 1980s to 1990s, handle the full set of configuration rules for facing the problem of combinatorial overload. However, to reach the intended use, additional rules have to be written, making them sometimes complex and difficult to deploy.

Compile-based configurators build upon constraint-based engines and research in binary decision diagrams. All possible combinations are being compiled into a single distributable file and thus are independent of how rules are expressed by the author. In this way, companies can import rules from legacy systems and process more complex sets of rules and constraints associated with more customizable products.

Mostly, existing CPQ solutions still have to be customized with great effort, since their customers' structure is contrary to the standard functionality. This leads to manual preparation and intervention, e.g., manually decoding of databases and assigning of data fields and types. Besides that, they may have limitations, when presenting complex product information and relations to employees and customers, since they are usually based on coded databases only. Finally, licensing costs for CPQ software leaders may be too high for some SMEs.

3. Use Cases

As a solution of the challenges described in section 1, we propose the development of a new method based on semantic technologies, like data integration with ontology alignment, data quality measurement, and business process integration. When product information is modelled and represented using an ontology, several methods from the field of semantic technologies that have not yet been adopted in complex and big data management contexts, can be applied, achieving a higher level of business data and process integration. [6]

3.1. Data Integration

Existing products for complex product data integration require manual preparation and intervention, such as decoding databases and assigning data fields and types manually.

In contrast, ontology-based big data management based on semantic technologies enables enterprises to integrate and manage data with a high degree of automation. When the manual interference in the integration process is being minimized, less manual mistakes can occur.

Ontologies can help to integrate data from heterogeneous data sources, usually by creating a so-called ontology alignment. A highly simplified example is given in Fig. 1.

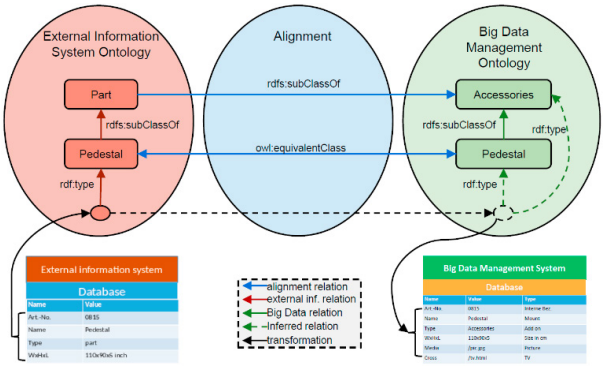

Fig. 1. Ontology Alignment.

This alignment contains the relationship between two data sources. It can be used to generate a transformation or to produce a mapping ontology that contains the relationship between them. Using this mapping ontology allows the interference of the types of entities from all aligned data sources in different layers, such as data, services and processes. Typical relationships between entities would be equivalence (owl: equivalentClass), subsumption (rdfs: subClassOf), and disjunction (owl: disjointWith). [7]

Alignments can be created manually or generated (semi-) automatically by using a process called Ontology Matching (OM). While for manual alignment, comprehensive manual input by a user is mandatory, for semiautomatic alignment, a user might also have to interfere, but most of the alignment is generated automatically by interpreting existing sources and deducing the relationship between them.

Usually, ontology-matching systems take two ontologies as input and provide the user with suggestions of correspondences between the data sources, which can be accepted or rejected. If the user has to make a decision during the integration process, the ontology will support him by limiting the possibilities he has to choose from. This leads to a decrease in the total number of manual decisions the user will have to make during the integration process. Enabled by a

pre-selection automatism based on ontology alignment, it will significantly reduce time and costs for the integration process.

3.2. Data Quality

The proposed approach incorporates the measurement of data integration quality and product information quality in a big data context. Hence, this can help enterprises discover insufficient data and improve the data quality of enterprises. Additionally, this will allow for a better assessment of the results of data integration than in previous big data management approaches. Also, this will support the evaluation of data integration results.

The Information Systems Success Model (ISSM) by DeLone and McLean incorporates the concept of information quality by identifying six dimensions and their relationships: information quality, system quality, service quality, use of system and intention to use, user satisfaction, and net system benefits. Information quality is regarded as a dimension used to measure the semantic success of an information system. [8] From the consumer perspective, Wang and Strong describe the concept of information quality as the "data, that fit for the use by data consumers" [9]. According to DeLone and McLean, information quality should be personalized, complete, relevant, easy to understand, and secure.

A theory to understand and predict technology acceptance is the Technology Acceptance Model (TAM) by Davis et al. It implies that the perceived usefulness and the perceived ease of use of a system directly impact the attitude toward usage, which then impacts the behavioral intent. While the ISSM focuses on the net benefits associated with information system use, the TAM focuses on expectations of net benefits from future information system use. Hence, the TAM and its application in a big data management context can be used for this research to measure the influence of data quality on the perceived usefulness of both the integration of data and the big data management system itself. [10]

For ontologies, several approaches that support data quality measurement exist in application cases such as sensor networks, data integration, and representation of data quality constraints in general.

3.3. Business Process Integration

While existing products only implement information systems, the proposed approach also supports the implementation of configurable product-related business processes and workflows. It will incorporate knowledge from other approaches, that model business processes in general as well as standards for modelling business processes, that have been transformed to ontologies, as far as they are necessary to operate big data successfully, e.g., management and extension of product taxonomies or management of product status.

As an example of such an approach, Garijo presented an ontology for representing Open Provenance Model (OPM) business processes. "Provenance" in informatics means the lineage of data. The conceptual model of causality and relation has been extended to include processes that act on data and agents, that are responsible for those processes. A design goal of OPM is establishing interoperability of systems through information exchange agreements. Its successor is a data model simply called "Provenance" (PROV).

Rospocher described an ontology for the Business Process Model and Notation (BPMN), a graphical representation for specifying business processes in a model. This would allow a unified representation of product data and business processes related to it. The goal of BPMN is to provide a standard notation for all business stakeholders: The business analysts, who create and refine the processes, the technical developers responsible for implementing them, and the business managers, who monitor and manage them. BPMN serves as a common language, bridging the communication gap that frequently occurs between business process design and implementation. BPMN has been complemented by two new standards for building case management and decision models, the Case Management Model and Notation (CMMN) and the Decision Model and Notation (DMN).

4. Method Comparison

4.1. Other Methods

The most easily employed method that can be used to cover all aspects presented in this paper is to perform the tasks of data integration and data quality measurement manually. While only few special technical skills are required to perform this task, its manual overhead is very high. Besides that, the process is very error-prone, usually unstructured, and not reproducible. This makes this solution insufficient for all but very simple application scenarios.

Another solution would be to develop custom applications that connect different data sources, implement business process integration, and compute data quality measures. It would allow a certain degree of automation of these processes, while the software development is feasible for programmers with basic programming skills.

This approach has been applied several times in the past. It leads to multiple problems, in particular the fact that the developed integration software contains a very limited reuse potential. Hence, any work, that is done will be repeated as soon as a new data source needs to be integrated. Also, data models, the semantics of the created integration, data quality and business processes are usually hard-coded within the software. Comprehension, maintenance, and changes of any of these aspects require a thorough study and/or change of the source code. In addition, it is difficult or virtually impossible to keep the existing technical solution when changing the technical platform. This inflexibility is very problematic for our target group of users, namely SMEs, since they usually have very limited resources to carry out the needed changes.

Using explicit model-driven processes, such as Model-Driven Software Development (MDSD), allows for an explicit meta model of the data integration and data quality models. The models could be reused in other applications, and

the model itself could be refined manually when changes occur. On the other hand, learning to work with MDSD-based tools requires time and effort on the part of the software developers, who do not already know them. Also, the semantics of these models is implicit, i.e., only represented in the program code that uses these models. Hence, reusing them is only possible if the source code or a very precise documentation of the source code is accessible. Automatically reusing the models over again is virtually impossible.

4.2. Challenges of our Approach

Our approach, which uses semantic technologies and ontologies and makes all major ingredients, e.g., data models, integration semantics, and data quality, explicit, will require more upfront effort from developers. Consequently, initial trainings as well as comprehensive documentation for developers, who want to adopt this method, will be needed.

The overall applicability of automated ontology matching product data integration might also be challenging. The state-of-the-art of it does not support the construction of complex alignments between ontologies to a satisfying degree.

Therefore, we need to evaluate, whether different product data integration scenarios require this type of alignment or whether simple alignments are expressive enough. The same applies to other ontology-based methods like data quality measures and modelling workflows: While these methods have proven to be working in general ontology-based settings, they are yet to prove their effectiveness for product data.

The complexity of the reasoning process in ontologies could be another challenge and might lead to performance issues. The standard language for ontologies, Web Ontology Language 2 (OWL2) is being divided into several parts, called profiles, each allowing different levels of expressivity. This separation is needed, because different levels of expressivity in the ontology description will lead to different execution times in the reasoning process when trying to derive new knowledge from the ontology. This possible issue can be addressed by carefully selecting an ontology language, that trades expressivity in for reduced reasoning complexity. For example, it can be evaluated, whether rule languages like Semantic Web Rule Language (SWRL) also could be used to represent these aspects. Depending on the origin of the increased reasoning complexity, it can also be helpful to customize or extend existing reasoners.

Representing product data as an ontology has been a topic in several publications. Their focus was mainly to represent the product data itself or to annotate websites for their processing through search engines. These approaches could be helpful in creating data representation, but have not been targeted to represent, exchange, and integrate product data inside an enterprise. Thus, new model features will be developed to support the new tasks presented before. This will lead to the technical question of how and to which degree the product information will be represented.

Scalability issues may affect our software as well. While having a unique set of limitations, nearly every system is struggling centralization and synchronous communication issues when trying to scale it, especially if there is a certain distance between single nodes of a system. Too much decentralization, however, may lead to product information data that is increasingly non-consistent, redundant, incomplete, or incorrect. Therefore, our approach will set decentralization spots and implement asynchronous communication only if needed within proposed tasks.

When designing a user-friendly domain-specific modelling and representation environment for product data based on ontologies, there is a trade-off between expressivity on the one hand, and usability on the other hand. While tools like WebProtegé promise usability improvements, they still focus on ontology engineers and not on end users. Thus, an evaluation of the required modelling complexity for them will be needed. Also, the required expressivity for ontology-based data management as well as new concepts for user-friendly modelling of complex topics have to be evaluated.

4.3. Benefits of our Approach

When our proposed approach is applied, the integration of new data sources into an existing system can be facilitated faster and it can make the results of the integration easier to assess. It might also improve the usability of software systems. At the same time, existing solutions can be used, where an application of existing approaches is desired.

Since the semantics of the metamodels are encoded in the model itself and not in the software, this allows a decoupling of metamodels and implementation and thus can be completely reused on other platforms and different software systems, even if the programming platform is changed. Through the included specified semantics, the meaning can be retained. This might also improve the maintainability of applications, while the model semantics do not need to be fully understood by the developers that maintain the software.

Through ontologies, complex product information, and relations between them, e.g. product specific attributes and formats, sector and nation specific terms/vocabulary, product taxonomies and classification standards and relations for up, down, and cross selling, can be administered easier. Product information can be represented in different forms, like visualized maps. They can enhance employees' and customers' understandings and increase the usage of complex product data and relations, making our approach user-friendly.

Ontologies can also represent different aspects of a domain, such as product catalogues or data quality definitions. Due to a high automation degree, their usage will minimize the manual interference in the integration process and thus enable SMEs to integrate and manage data with less errors. Besides that, preferences for different classes or properties of products can be communicated.

For the application of our approach, a cloud-based product information database will offer an integrative solution, while data and process integration quality can be measured using existing approaches and methods for integration quality assessment, as well as data quality evaluation, which can

incorporate multiple dimensions, e.g., completeness, consistency, and minimizing redundancy.

Our solution will incorporate the measurement of integration and data quality in the context of big data. This innovative feature will help SMEs to discover insufficient data and improve their data quality. Particularly, ontology-based search and navigation are promising solutions, able to improve the technical state-of-the-art significantly, e.g., in the form of full text search engines.

Also new in our proposed solution is the implementation of configurable data workflows which are necessary for the successful operation of product data as well as for the management and extension of product taxonomies and product status. A highly simplified example for a product configuration workflow of a CPQ software is given in Fig. 2.

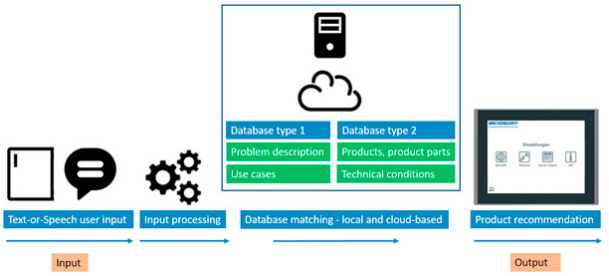

Fig. 2. Product configuration workflow.

While being expert-based in the beginning, in the following stages, more and more real user data will be added. Based on modern machine learning, deep learning, and AI concepts, multiple resources for product information can be analyzed and organized, and actual as well as upcoming user data for improvement and correction can be utilized.

Finally, we plan to link our CPQ software to standardized visualization systems, such as Supervisory Control And Data Acquisition (SCADA) and Human Machine Interface (HMI), so SMEs can create complex visualizations of company data by using a mobile app without any programming knowledge.

With ontology-based big data management, data integration services and the implementation of big data can be realized with less costs and less manual effort than existing solutions. Hence, more enterprises, that hitherto have been deterred by the high costs of data integration and implementation, can use a solution for exchanging, linking and reusing data in and across sectors at a lower price.

By using our approach, the willingness of enterprises to exchange, link, and reuse product data, and to implement a modern data management system will increase, and their productivity will improve, because expenses and time for managing and searching for product data will be reduced.

5. Conclusion

In this paper, we have presented three application cases for big data management, to be improved by the employment of semantic technologies and ontologies. The application cases are data integration from internal and external sources, data quality measurement, and business process integration. Existing research covering the specific problems of these application cases has been presented for each of them.

For data integration, a method called ontology matching has been proposed. This method allows the semiautomatic generation of alignments. Alignments are used to formalize the coherence between ontologies.

For process integration, modelling product-related business processes and workflows through ontologies based on case management and decision models, OPM/PROV as well as BPMN/CMMN/DMN have been proposed. They establish interoperability of systems through information exchange agreements by providing a standard notation for all business stakeholders like analysts, developers, and managers, who monitor and manage them.

For data quality, approaches for evaluating product information quality and data integration quality have been referenced. Also, examples of approaches for business process integration have been mentioned. Finally, it has been shown that enterprise solutions like CPQ software will benefit from our approach as well.

For future research, each of the approaches will have to be evaluated for their practical applicability through a prototype implementation and a user study to measure their effectiveness and usability. Since our proposed solution has never been implemented before, we intend to participate in a proper funding program, helping us to develop a software prototype to show the effectiveness of our method for different data sets. For this purpose, we invite industry partners, especially SMEs, to join our future R&D project consortium.

References

[1] Eine B, Jurisch M, Quint W. Semantic Technologies for Managing Complex Product Information in Enterprise Systems. In: ERP Future. Switzerland: Springer Cham; 2016. Vol. 245, p.111–118.
[2] Eine B, Quint W: User-oriented Product Information Management with Semantic Technologies. In: CENTRIC 2015, p. 86-87.
[3] Nalini C, Arunachalam, AR: A Study on Privacy Preserving Techniques in Big Data Analytics. In: International Journal of Pure and Applied Mathematics 2017. Vol. 116 (10), p.281-286.
[4] Guizzardi, G: Ontology-Based Evaluation and Design of Visual Conceptual Modeling Languages. In: Iris Reinhartz-Berger, I et al. (Eds.): Domain Engineering. Heidelberg: Springer 2013.
[5] Jurisch M.; Igler B: Knowledge-Based Self-Organization of Traffic Control Systems. In: GI-Jahrestagung 2016, p. 947-954.
[6] Eine B, Jurisch M, Quint W: Ontology-Based Big Data Management. In: Systems 2017, Vol. 5(45), p. 1-14.
[7] Jurisch M: Managing Ontology Mapping Change Based on Changing Inference Sets. European Knowledge Acquisition Workshop (EKAW). In: Satellite Events 2016. p. 256-263.
[8] Delone WH, McLean ER: Information Systems Success Measurement. In: Foundations and Trends in Information Systems 2016. Vol. 2(1), p. 1-116.
[9] Willcocks, LP; Sauer, C; Lacity MC: A Semiotic Information Quality Framework: Development and Comparative Analysis. In: Enacting Research Methods in Information Systems 2016, Vol.3, p. 219-250.
[10] Folkinshteyn D, Lennon M: Braving Bitcoin: A technology acceptance model (TAM) analysis. In: Journal of Information Technology Case and Application Research 2017, Vol. 18 (4), p. 220-24.