systems

MDPI

Article

Ontology-Based Big Data Management

Bastian Eine [1,*], Matthias Jurisch [2] and Werner Quint [1]

[1] Department of Media Management, CAEBUS Center for Advanced E-Business Studies, RheinMain University of Applied Sciences, Unter den Eichen 5, 65195 Wiesbaden, Germany; werner.quint@hs-rm.de

[2] Department of Applied Computer Science, RheinMain University of Applied Sciences, Unter den Eichen 5, 65195 Wiesbaden, Germany; matthias.jurisch@hs-rm.de

* Correspondence: bastian.eine@hs-rm.de; Tel.: +49-611-9495-2278

Received: 15 May 2017; Accepted: 4 July 2017; Published: 6 July 2017

Abstract: Big data management is no longer an issue for large enterprises only; it has also become a challenge for small and middle-sized enterprises, too. Today, enterprises have to handle business data and processes of increasing complexity that are almost entirely electronic in nature, regardless of enterprises' size. Enterprises' information systems need functions based on specific technologies to be able to reduce and interpret the complexity of business data and processes. This paper pursues the question: how can state-of-the-art information systems be improved by the use of semantic technologies, and particularly ontologies? For this purpose, three use cases of information systems that could be improved are described, and approaches based on semantic technologies and ontologies are proposed. The selected use cases relate to data integration, data quality, and business process integration.

Keywords: ontologies; big data management; complexity management; systems tools

1. Introduction

The establishment of electronic data management and business processes brought a number of improvements for enterprises, such as the automatic handling of purchasing and selling products. Therefore, information about products and business processes are almost exclusively managed as data in enterprises' information systems [1]. However, enterprises are challenged by the increase of complexity required to handle more and more electronic data and processes [2]. Frequently, this issue is discussed in context of big data. The term "big data" describes large and complex data sets that traditional data applications are unable to process adequately. Likewise, not all enterprises' information systems are capable of fully encompassing the big data management needs of an enterprise. For example, enterprise resource planning systems are set up to represent all business processes of an enterprise, in order to increase the overall cost effectiveness [3]. Often, enterprise resource planning systems do not include large data set about products, e.g., product-marketing descriptions, product pictures or complementary technical product data. Generally, these types of data are managed with the help of a number of other types of information systems, e.g., product data management systems [4], customer relationship management systems [5], and content management systems [6]. For example, product information management systems, which had been the focus of the approach that is proposed in this paper [7], provide centralized and media-neutral data storage, data management of product information in enterprises [8]. Additionally, these systems offer a number of functions, which enable companies to manage and use product data consistently in many areas inside and outside of the enterprise. For instance, product information management systems can assist companies and employees with product classification, translation management, media asset management and data output to different media (e.g., print catalogue, online shop, technical data sheets) [9].

However, companies have to handle more and more complex information about products as product complexity increases and customers' demand for customizable products grows [10].

Also, companies face the challenge of integrating large amount of data that is scattered across many different information systems into one enterprise-wide centralized information system that enables the company to handle and make use of complex and large data (big data). Today, many companies are not able to implement such an information system into their information technology structure, as processes for data and business process integration are still too costly and time-consuming for them. This can lead to data redundancy and a high number of inconsistencies within the product information of an enterprise. Hence, high expenses for maintaining, searching for and presenting product information can arise [11]. Also, the number of customer requests, wrong orders and wrong deliveries may increase. Therefore, it is necessary to implement syntactical, as well as semantical restrictions or defaults to avoid redundant statements or wrong interpretations in complex and/or big data sets.

Another challenge is to capture and represent complex and big data about products, e.g., relations between products or product features. For example, when composing a configurable product, it has to be taken into account that selecting a certain product feature may exclude other product features. This information is essential not only for employees, who are responsible for maintaining product data, but also for customers, who want to configure a product or satisfy their information needs. Accordingly, it is necessary to capture, manage and present the complex information in a clear and easy to understand manner. Lastly, it has not yet been analyzed whether existing solutions for data and business process integration in big data management context take different user perspectives into account, and if they do, how satisfying are they to the user.

We propose that the necessary data integration processes and information systems can be improved by employing methods from the field of semantic technologies. This paper will analyze three use cases of complex and large product data integration and information system contexts that can be improved. Section 2 sketches the approach for using ontologies and summarizes the related work regarding information systems and semantic technologies. A detailed view on the use cases is presented in Section 3, while Section 3.1 covers data integration, Section 3.2 describes data quality measures and Section 3.3 gives an overview of business process integration. A critical discussion of the approach is given in Section 4. The conclusion is presented in Section 5.

2. Background & Related Work

Regarding the complexity of today's product information data sets and their interconnections, new approaches will have to be developed [12] to gain value from this big data. When developing a method or software for big data management, handling big data includes analysis, capture, querying, sharing, storage, visualization or updating [13]. In particular, suitable algorithms for modeling and representation of different data as well as individually customizable mechanisms for data linking in dynamic datasets are relevant for such an approach.

2.1. Big Data

In 2011, Gartner identified three dimensions of big data, on which business and information technology leaders have to focus: information volume, variety and velocity [14]. 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-mail, metering data, video, images, audio, stock ticker data, financial transactions. "Velocity" means how fast data is being produced, but also how fast it must be processed to meet the demands of the stakeholders. Besides volume, variety and velocity, the International Business Machines Cooperation (IBM) also suggests "veracity" as another dimension to measure the reliability of data since data sets arrive from different sources and thus may not fully fulfill the required quality standards [15]. When developing an approach, these "3 + 1Vs" should be taken into account.

2.2. Semantic Technologies

The use of semantic technologies when creating and representing complex information and relations between concepts can help to interpret information by identifying the corresponding context. Semantic technologies can make it easier to understand the meaning and purpose of data (e.g., symbols, words etc.) and complex concepts, as well as share knowledge for humans and machines [16]. Semantic technologies can be based on metadata that contain more information about other data and therefore help to more efficiently find information and documents in general. In addition, metadata can be linked to other metadata in different data sources. However, this requires standardized rules that make it possible to represent metadata in a formal way. These rules are prerequisites in order for metadata to be exchanged between information systems, applications and workstations [17]. Therefore, the Resource Description Framework (RDF) can be utilized, which also serves for describing resources (meta-information) 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 semantic problems when creating and interpreting product data. Approaches with more semantic richness are topic maps [18] and ontologies [19].

2.2.1. Ontologies

Ontologies are usually defined as an "explicit specification of a conceptualization" [20]. This means that ontology allows for the definition of concepts and relationships between these concepts, and that the specification representation provides a formal semantic of the specification. Ontologies are the approach with the highest degree of semantic richness of all common models for knowledge representation [16]. They are based on models that will be explained in order of increasing degree of semantic richness, while none of these models achieves the degree of semantic richness that ontologies provide. A glossary is a model with the lowest degree of semantic richness out of all of the other models mentioned here. A glossary is a list of words in alphabetical order with definitions of words, but without any explanation of relations between these words. On the next level of semantic richness, taxonomy is a model for the hierarchical classification of words. It describes relations between words using super- and sub-relations. These relations provide an ordering of generality between these terms. A thesaurus is an extension of taxonomy. A thesaurus describes any type of relations between words. A topic map is a model that is most similar to an ontology. A topic map is an abstract model and data format for the formulation of knowledge structures. Based on a thesaurus, associations describe the relationships between different topics. Additionally, occurrences can be used to embed external documents into topic maps [16].

In the case of full ontologies, usually, some kind of mathematical logic is used to provide the formal semantics of the specification, which allows the inference of new knowledge from the ontologies. Building on the ideas of ontologies, Berners-Lee et al. proposed the semantic web [21]. This idea is a technical approach applying ontology-based technologies to the web. The basis for the semantic web are the core technologies XML, which is a format for interchanging data, and URI/IRI, which is a schema for addressing resources [22]. Based on these, there are technologies and norms that were developed specifically for the semantic web. The underlying framework is RDF for describing resources and the relations among these resources. With RDF, content can be expressed semantically by using so-called triples of subject-predicate-object for describing data [17].

The next level in the layered architecture of the semantic web is the ontology layer, which is based on RDF. The ontology layer includes further language formats, including ontologies and vocabularies for modeling semantic knowledge. Overlying layers like Rules, Query and Logic control the access to semantic data sets (SPARQL). The final layers, Proof and Trust, should establish structures of security and trust in the semantic web. RDF can be used to describe resources, but within a certain domain, additional vocabulary is needed to describe data. A vocabulary is a collection of concepts and

properties of a domain for describing resources and their relationships to each other [23]. In context of semantic modeling, vocabulary can also be interpreted as ontologies. For this purpose, vocabulary- and ontology-languages are needed. RDF Schema (RDFS) and Web Ontology Language (OWL) are examples of such ontology languages and are verbalized in RDF. The relationship between these two languages is established by inheritance. The main difference between these two languages is their power of expression. More expressive ontologies can be developed with OWL than RDFS. Additionally, OWL's higher degree of expression compared to RDFS makes it possible to infer complex previously implicit knowledge in the ontology by using Reasoners [18].

2.2.2. Semantic Web

In context of big data, scientists proposed to make use of semantic web technologies and linked data to turn data into knowledge [24]. The semantic web has been adopted in various scientific domains. In bioinformatics, several ontology-based systems like the gene ontology [25] have helped researchers from different countries communicate with each other and interlink their research data. A semantic web-based approach for integrating e-commerce systems has been proposed by Hepp [26] and since been adopted by search engines such as Google and Yahoo.

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 mentioned product relations.

2.2.3. Semantic Desktop

The Semantic Desktop is an example of the application of semantic technologies for supporting employees with information-intense tasks. The Semantic Desktop provides the support that is unavailable in today's information systems for individual knowledge management, and enables the integration of existing (not individual) knowledge management systems [27]. Often, companies use distributed and proprietary content management systems (CMS) to manage their data. Thereby, many connections between information systems arise that are not efficient and get even more complex because of Big Data. Another disadvantage is that users of CMS have to use several different CMS when searching for information. The solution is a so-called semantic middle layer, which would significantly reduce the number of links and represent a consistent interface. Thus, all data sources can be accessed easily and combined. By the concepts presented before, e.g., RDF and ontologies, this middle layer can enable a consistent and centralized access to content and knowledge networks [28]. Hence, the Semantic Desktop provides the integration of metadata through semantic technologies, and ontologies and RDF are the primary technologies that operate in the background. RDF is utilized as a data structure for the serialization of metadata, thus for the process of saving an object (metadata). Ontologies provide the formal description of metadata so that it can be interpreted, exchanged and reused by machines (applications). By developing these knowledge networks, future task-related requests can be made easier. In this case, the semantic functions are an integral part of the system; they are not an "add-on" that has to be installed separately [29].

2.3. Semantic Technologies and Information Systems

The approach proposed in this paper will enable the application of ontology-based concepts to the field of big data management, which are new to this field.

One solution of the problems described in Section 1 would be the development of a new method based on semantic technologies. In this method, product information is modeled and represented using an ontology. Through the ontology, it would be possible to apply several methods from the field of semantic technologies that have not yet been adopted in complex and large product data management contexts. The applied methods from semantic technologies are data integration through ontology alignment, data quality measurement, and business process integration.

Previous research in these areas has shown that semantic technologies are applicable to the field of information systems, especially for product data. In 2005, Hepp enabled the representation of eClass catalogues as an ontology by proposing a special ontology vocabulary named eClassOWL [30]. This work focused on representing the information from an eClass catalogue in an ontology. Brunner gave an overview of the possible improvements of information systems for product data management by semantic web technologies in 2007 [31]. Brunner also presented a product information system built on semantic technologies; however, he concluded that the existing semantic web technologies at that time did not allow for the implementation of an efficient and scalable product information system. Thus, proprietary extensions had to be applied. In 2008, Hepp presented the ontology GoodRelations as a semantic web-based representation for products and services [26]. Since then, it has been adopted as a widespread method to annotate existing websites with product data that can be parsed by search engines such as Google and Yahoo. The Aletheia architecture is a work that focuses on the integration of structured and unstructured data sources based on semantic technologies [32]. This approach is based on a service hub that allows data exchange and data transformation, and separates between a certain and uncertain knowledge repository. Stolz et al. presented another integration-focused approach [33]. This work consisted of a transformation from the standard catalogue format BMEcat to the already mentioned GoodRelations ontology. Fitzgerald et al. presented a holistic approach for master data management [34]. The focus of this approach is the usage of the same data structures in different phases of the product life cycle. A research approach and preliminary results in the form of a reference architecture for a semantic master data management system are presented.

To integrate data from different sources, several tools have already been implemented that demonstrate the capabilities of automated ontology alignment computation in general. Shvaiko and Euzenat [35], and more recently Rodriguez-Martinez and Gomez-Rodriguez [36], present a survey of the most important tools in this area. The tools combine several approaches to find correspondences between schema entities, e.g., using terminological similarities (finding similarities in descriptions and labels for schema entities) and structural similarities (finding graph similarities in the hierarchies of the schemas). The implementation of these type of tools might enable automated support in big data management and the integration of different data sources into big data management systems, as it would allow data integration experts to perform the integration more easily, with less errors and faster than when integrating data manually. This will lower the resources required to integrate the data compared to executing the process using a less automated approach, which is not reasonable for the requirements of big data management. However, the technique of ontology alignment has not yet been applied to information systems for complex and large product data, and its applicability in this area has not yet been proven. It is expected that the application of ontology matching to complex product information structures will lead to a higher degree of data automation and business process integration. Also, this will contribute to the research about big data management as well as data and business process integration approaches.

3. Use Cases

3.1. Data Integration

Existing products for complex and large product data integration and the implementation of big data management require a lot of manual preparation and intervention, such as decoding databases and assigning data fields and types manually. In contrast, ontology-based big data management will be based on semantic technologies, which will enable enterprises to integrate and manage data with a high degree of automation. Also, less manual mistakes can occur because this approach will minimize the manual interference in the integration process. The advantage of ontology-based big data management is that if the user/enterprise has to make a decision during the integration process, the ontology will support the user by limiting the possibilities the user has to choose from. Hence, there will also be

a decrease in the total number of manual decisions the user will have to make during the integration process. This will be enabled by a pre-selection automatism based on ontology alignment, and will reduce time and cost for integration process significantly.

Ontologies can be used to integrate data from heterogeneous data sources. Usually, this is done by creating a so-called ontology alignment (a highly simplified example is given in Figure 1).

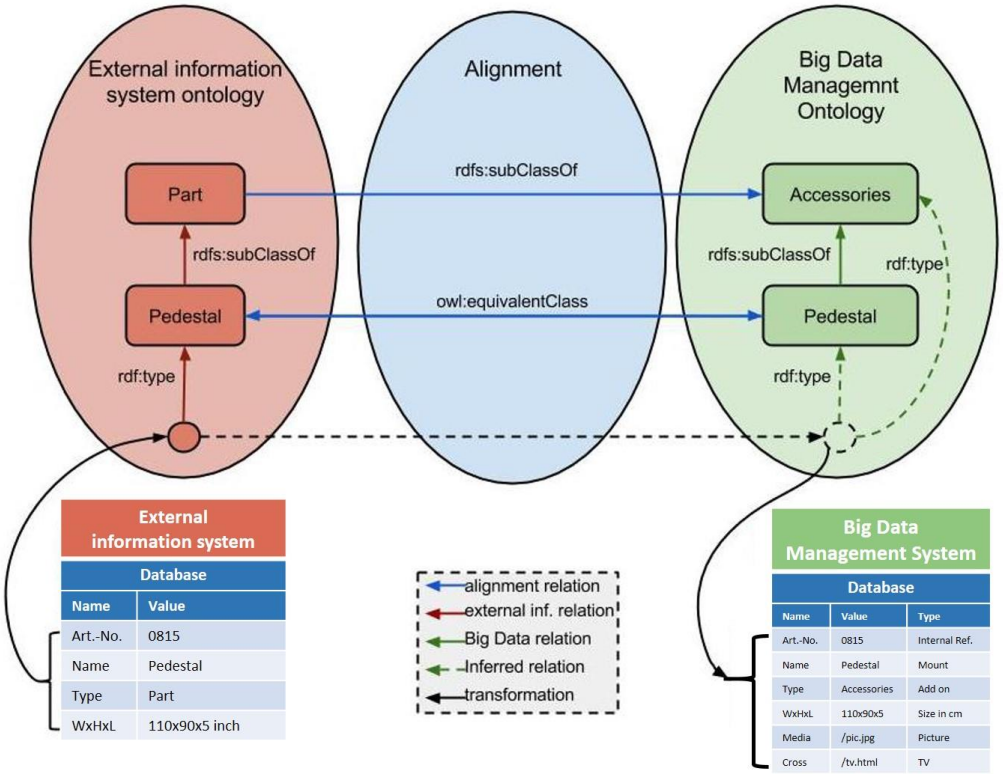

Figure 1. Ontology alignment.

This alignment contains the relationship between two data sources. Typical relationships between entities in the different data sources are equivalence (owl: equivalentClass), subsumption (rdfs: subClassOf) and disjunction (owl: disjointWith). For example, this alignment can be used to generate a transformation between the data sources or to produce a mapping ontology that contains the relationship between them. Using this mapping ontology allows to infer the types of entities from all aligned data sources, much like how it is depicted for one record in the external information system in Figure 1.

Alignments can be created manually, but it is also possible to generate them (semi-) automatically by a process called ontology matching. 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. Ontology matching is a research topic that deals with integrating data from different sources by automatically (or semi-automatically) deducing the relationship between them [37]. Usually, ontology-matching systems take two ontologies as input and provide the user with suggestions of correspondences between the data sources, which the user can accept or reject.

A technical challenge is the applicability of automated ontology matching to big data management. The current state of the art in ontology matching does not support the construction of complex alignments between ontologies to a satisfying degree [38]. Therefore, it has to be evaluated whether big data management requires this type of alignments, or simple alignments are expressive enough for big data management scenarios.

3.2. Data Quality

The proposed approach incorporates the measurement of integration quality and data 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.

In information system research, the information systems success model by DeLone and McLean is a very frequently cited model that incorporates the concept of information quality [39]. The information systems success model helps understand information system success by identifying six dimensions and their relationships: information quality, system quality, service quality, use of system and intention to use, user satisfaction, and net benefits of system. Information quality is regarded as a dimension used to measure the semantic success of an information system [40]. From the consumer perspective, Wang and Strong describe the concept of information quality as the "data that fit for the use by data consumers" (p. 6) [40]. According to DeLone and McLean, information quality should be personalized, complete, relevant, easy to understand, and secure [41]. A theory to understand and predict technology acceptance is the technology acceptance model by Davis et al. [42]. The technology acceptance model 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 information systems success model focuses on the net benefits associated with information system use, the technology acceptance model focuses on expectations of net benefits from future information system use [43]. Hence, the technology acceptance model and its application in a big data management context can be taken into account 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. For ontologies, several approaches that support data quality measurement exist in use cases such as sensor networks [44], data integration [45] and representation of data quality constraints in general [46].

3.3. Business Process Integration

Ontology-based big data management can also support the implementation of big data management business processes. While existing products only implement information systems for big data management, the proposed approach can also support the configurable business processes and workflows that are necessary to operate big data successfully, e.g., management and extension of product taxonomies, or management of product status. Therefore, modeling product-related business processes and workflows through ontologies can be incorporated into an ontology-based big data management. Other approaches that model business processes in general already exist. Standards for modeling business processes have also been transformed to ontologies as well. Garijo [47] presents an ontology for representing Open Provenance Model (OPM) business processes and Rospocher [48] describes an ontology for the Business Process Model and Notation (BPMN). This will allow a unified representation of product data and business processes related to it.

4. Method Comparison

The most easily employed method that can be used to cover all aspects presented here 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 this, 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

had been applied several times in the past. It leads to several problems. In particular, 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, the data models, 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 therefore a thorough study and/or change of the source code. In addition, it is usually difficult or virtually impossible to keep the existing technical solution when changing the technical platform.

Using explicit model-driven processes such as model-driven software development (MDSD) allow 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 by software developers who do not already know them. Also, the semantic 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 again is virtually impossible.

The approach proposed in this paper, which uses semantic technologies and makes all major ingredients, e.g., data models, integration semantics and data quality explicit, will require more upfront effort from developers, but also has several significant advantages. The approach will allow using several advanced techniques employed in ontology integration, such as ontology matching. This will make integration with different software systems easier (see Figure 1). In addition, the models including their semantics will be completely reusable on other platforms and different software systems. This will allow for a complete reuse of the models, even if the programming platform is changed. The usage of ontologies will also enable the reuse of established ontologies for representing different aspects of the domain, such as product catalogues or data quality definitions.

Because of high demand on big data management software, the market and competitors both grow very fast. The variety of solutions and variable volumes of big data management systems lead to an enormous range of competitive products, most of which are sales-oriented. Increasingly, end-user organizations are exploiting master data management as the heart of digital business transformation. However, in most cases, these solutions have to be customized with great effort for customers as soon as the customers' structure is contrary to the standard functionality. In Figure 2, we created an overview of products in this market segment, which sketches the business process and data integration support of these solutions. The overview is based on similar studies conducted by Gartner [49] and Forrester [50]. It is expected that the proposed approach for ontology-based big data management (DaMonto), which is described in this paper, can enhance the market for business process and data integration solutions in both competencies. In the paragraphs following Figure 2 in this section, we will explain how the position of our approach and the positions of existing solutions were estimated in the matrix in Figure 2. In addition, a more detailed, exemplified comparison of an existing solution of software company SAP with our approach is given below. For the actual application of our proposed approach, data and process integration quality can be measured using existing approaches and methods for integration quality assessment, as proposed by Nigel et al. [51] and Batista and Salgado [52], and using models for data quality evaluation in general (see Section 3.2). These methods can incorporate multiple dimensions, e.g., completeness, consistency, and minimizing redundancy. We expect to achieve better results in regards to these quality assessment dimensions than existing solutions, because our ontology-based approach should enable the implementation of centralized big data management systems that aim to maximize these criteria (completeness, consistency) in and across enterprise business processes.

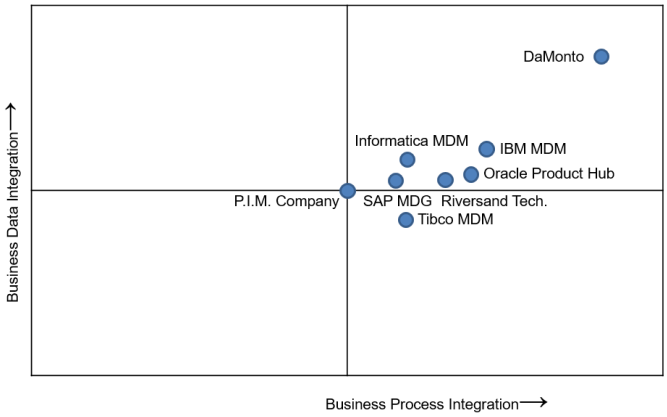

Figure 2. Illustration of competitive products.

For classifying and evaluating competitive products, qualified vendors were considered and compared regarding their data integration (y-axis) and business process integration (x-axis) in the year 2017. At the same time, the application of semantic technologies was analyzed. In general, the term "data integration" refers to the integration of data from different internal areas of operation to make these data useful for various tasks [53]. "Process integration" is important for operational functions that are scattered across various applications on different platforms and can be joined in terms of data and business process integration. "Process integration" is the extent to which these functions allow a consistent process management along the value-added chain [54]. By the interconnection of both types of integration, data redundancy will be avoided and data integrity will be ensured. Also, the complexity will decrease because of less interfaces, and therefore time and effort for administration and service will decrease too. However, to access big information pools efficiently, data have to be structured and categorized. A technological approach that has proven of value are ontologies, but not many companies use this approach. Particularly, ontology-based search and navigation are promising solutions that are able to improve the technical state of the art significantly, e.g., in the form of full text search engines. The additional use of the W3C standard OWL promotes the integration of heterogeneous sources of information. An OWL-based ontology establishes a common vocabulary in different systems [55].

SAP is such an enterprise that uses ontologies. For example, in the context of logistics, the enterprise developed an ontology-based distribution platform for managing the logistic processes of single or multiple enterprise locations. As a result, OWL was used to describe transportation services. Each service is represented as a class that contains distribution properties, e.g., the location of origin and ship-to location. Further classes include information about articles to be shipped, e.g., package or documents. By using ontologies, preferences can be communicated, too, such as specifying which transport service is practical for the order [56]. In relation to enterprise resource planning (ERP) systems, the ERP business suite technology by SAP is divided into different modules for different divisions of an enterprise, e.g., logistics, controlling, production etc. These modules, e.g., material management in the logistics division, contain in turn many thousands of attributes and millions of attribute values. The SAP product Master Data Governance (MDG) for master data management offers only slightly above-average data integration. Only approximately half of the data that are generated in individual process steps are collected into a centralized big data management system. This can be exemplified with the process "outbound logistics". This business process contains 20 steps. The dataset "client base handling" is not part of all of the steps for which it would be relevant; hence, the data integration is incomplete. Thus, the incomplete data integration leads to incomplete process management along the value-added chain.

As a result, SAP is in the middle (average) of the matrix of the competitive products in Figure 2. The matrix shows that other competitive products are also in the middle. They have slightly

above-average data integration, but no product has complete process integration. Hence, the aim of the proposed approach should be to ensure complete data integration by using ontologies to collect all generated product data from internal and external information systems into a centralized big data management system. At the same time, continuous process management will be created by linking applications and reusing and processing data that is generated in other applications.

There are also various other reports that show an overview of the market big data management solutions from different perspectives.

In the area of proprietary software, the latest Gartner's Magic Quadrant for Business Intelligence and Analytics Platforms 2016 [49] points out Qlik is a leader, which scores significantly higher on complexity of analysis than its competitors Tableau and Microsoft. Gartner recognizes a modern business intelligence architecture, a rapid implementation approach and a strong partner network.

The latest Forrester Wave product information management report [50] names Hybris as a leader in product information management. Their software is "unique in having a combined and tightly integrated product information management/ecommerce offering that is built on the same flexible Java architecture and data model". The enterprise Informatica is also considered a leader. The report cites its strong business process capabilities, which help businesses to create and manage product data, as well as their usability to support an omni-channel environment. Compared to Gartner, it identifies mostly different competitors, but companies like SAP and IBM appear on both.

In the field of open source software, vendors are especially focused on sales-oriented software, for example Pimcore or PIMagento. Another ambitious and aspiring enterprise is Akeneo, which is a finalist in the 2015 eCommerce Awards. All of the competitors have different strengths, which might be exceeded by using ontologies, such as Pimcore, which has a 'connect anything' architecture, tight integration of e-commerce, web-to-print and web experience management, and easy integration via APIs. However, Akeneo is based on a powerful and modular platform leveraging the Symfony2 framework. The Oro Platform allows Akeneo to easily interact with other Oro applications such as OroCRM. The current products offered by the P.I.M. company are also still structured as sales-oriented. In contrast, an ontology-based big data management approach might provide an enterprise-wide database in order to offer an integrative solution. Thus, a unique position in the field of open source providers and direct competition to proprietary vendors could be reached, because of the sometimes high license and maintenance costs of the proprietary systems.

5. Discussion

A potential problem that could arise from the application of ontology-based big data management is caused by the complexity of the reasoning process in ontologies. This might lead to performance issues. These issues can be addressed by carefully selecting an ontology language that trades expressivity for reduced reasoning complexity.

Another potential shortcoming is the required initial training for developers who want to adopt this method. Also, the migration to an ontology-based implementation might induce high costs for the users. A cost-benefit analysis of the conversion to the ontology-based big data management method is planned as a part of future research to further investigate this question.

The method also offers several benefits. When the method is applied, the integration of new data sources into an existing system can be facilitated faster, easier and make the results of the integration easier to assess, too. The application of semantic-web-based methods might also improve the usability of software systems. Integration of semantics into the model itself also allows a decoupling of metamodels and implementation. Since the semantics of the metamodels of the data are part of the model, it can still be reused in other applications. When reusing the data, the meaning can be retained in the new application through the included specified semantics. This might also improve the maintainability of applications, because the semantics of the model are not encoded in the software, but in the model itself. Hence, the model semantics do not need to be fully understood by the developers that maintain the software.

Additionally, several examples have been shown for using the proposed ontology-based big data management on product data. The approach can help to manage complex product information and relations between these, 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. Existing big data management solutions might have limitations when it comes to presenting complex product data and relations to employees and customers, because they are usually based on coded databases only. In contrast, the proposed ontology-based big data management approach can manage product data with semantic technologies, which can represent information in different forms, e.g., visualized maps. These can enhance employees' and customers' understandings, and thus increase the usage of complex product data and relations. Hence, this approach is more user-friendly and easier to administer than other solutions. This can lead to a higher willingness to implement a big data management system with this approach.

It is expected that 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 had 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. It can be expected that the willingness to use big data integration and implementation services will increase.

Furthermore, the willingness of enterprises to exchange, link and reuse product data will increase long term because the proposed approach incorporates data quality measures and ensures consistency when integrating product data. Thus, the productivity of enterprises can improve, because expenses and time for managing and searching for product data can be reduced.

Representing product data as an ontology was part of several publications [16]. The focus of these was mainly to represent the product data itself or annotate websites, so that search engines can process it. These approaches will probably be helpful in creating data representation, but have not been targeted to represent, exchange and integrate product data inside an enterprise. Therefore, new model features will be developed to support the new tasks presented before. This will lead to the technical question of how the product information will be represented, and to which degree existing solutions can be used where an application of existing approaches is desired.

The applicability of automated ontology matching to product data integration is still a technical challenge. The current state of the art in ontology matching does not support the construction of complex alignments between ontologies to a satisfying degree [57]. Therefore, it has to be evaluated whether this type of alignment is required or if simple alignments are expressive enough.

The same applies for other ontology-based methods such as modeling business processes and data quality measures. While these methods have been proven to work in general ontology-based settings, it is not clear if they will work for product data and big data in general. Therefore, the applicability of these methods has to be evaluated.

The reasoning complexity of the modeled ontology is also a challenge. The standard language for ontologies, OWL2, is divided into several parts, called profiles, that each allow 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. For ontology-based big data management, the required expressivity has to be evaluated.

When designing a user-friendly domain-specific modeling 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. This will require an evaluation of the required modeling complexity for end users. Also, new concepts for user-friendly modeling of complex topics have to be evaluated.

6. Conclusions

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

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 data quality, approaches for evaluating product information quality and data integration quality were referenced. Also, two examples of approaches for business process integration have been mentioned.

These ideas have not yet been evaluated for their practical applicability. For future research, each of the approaches will have to be evaluated through a prototype implementation and a user study to measure their effectiveness and usability. This research is part of the proposed research project DaMonto, which is currently in the process of being evaluated by the Horizon2020 funding agency. If the project proposal is granted, the research will start in the midst of 2017.

Author Contributions: Bastian Eine contributed all contents about requirements of complex and large data integration and data quality. Matthias Jurisch contributed all contents about semantic technologies and ontology alignment. Werner Quint contributed all contents about business process management and integration.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. O'Brien, J.A.; Marakas, G. *Management Information Systems*; McGraw-Hill Irwin: New York, NY, USA, 2011.
2. Manyika, J.; Chui, M.; Brown, B.; Bughin, J.; Dobbs, R.; Roxburgh, C.; Byers, A.H. *Big Data: The Next Frontier for Innovation, Competition, and Productivity*; The McKinsey Global Institute: New York, NY, USA, 2011. Available online: http://www.mckinsey.com/business-functions/digital-mckinsey/our-insights/big-data-the-next-frontier-for-innovation (accessed on 10 May 2017).
3. Sumner, M. *Enterprise Resource Planning*; Prentince Hall: Upper Saddle River, NJ, USA, 2005.
4. Philpotts, M. An introduction to the concepts, benefits and terminology of product data management. *Ind. Manag. Data Syst.* **1996**, *96*, 11–17. [CrossRef]
5. Ngai, E.W.T. Customer relationship management research (1992–2002): An academic literature review and classification. *Market. Intell. Plan.* **2005**, *23*, 582–605. [CrossRef]
6. Boiko, B. *Content Management Bible*; John Wiley & Sons, Inc.: New York, NY, USA, 2001.
7. Eine, B.; Jurisch, M.; Quint, W. Semantic Technologies for Managing Complex Product Information in Enterprise Systems. In *Innovations in Enterprise Information Systems Management and Engineering (ERP Future 2015), Lecture Notes in Business Information Processing*; Springer: Cham, Switzerland, 2016; Volume 245, pp. 111–118.
8. Gartner Magic Quadrant for Product Information Management 2007. Available online: ftp://public.dhe.ibm.com/software/emea/de/db2/Gartner_MDM_Magic_Quadrant_PIM.pdf (accessed on 10 May 2017).
9. Lucas-Nülle, T. *Product Information Management in Deutschland—Marktstudie*; Pro Literatur Verlag: Mammendorf, Germany, 2005.
10. Sheldon, P.; Goetz, M. The Forrester Wave: Product Information Management, Q2 2014. Available online: https://www.informatica.com/resources.asset.84dbaa93b4463fbae96a596b2068676d.pdf (accessed on 10 August 2016).
11. Gartner Magic Quadrant for Product Information Management (2005). Available online: ftp://public.dhe.ibm.com/software/data/mdm/pdf/ibmPIMMQ.pdf (accessed on 10 May 2017).
12. Dumbill, E. What Is Big Data? Available online: https://www.oreilly.com/ideas/what-is-big-data (accessed on 10 May 2017).
13. Labrinidis, A.; Jagadish, H.V. Challenges and Opportunities with Big Data. *Proc. VLDB Endow.* **2012**, *5*, 2032–2033. [CrossRef]

14. Beyer, M. Gartner Says Solving "Big Data" Challenge Involves More Than Just Managing Volumes of Data. 2011. Available online: http://www.gartner.com/newsroom/id/1731916 (accessed on 10 May 2017).
15. Zikopoulos, P.C.; deRoos, D.; Parasuraman, K.; Deutsch, T.; Corrigan, D.; Giles, J.; Melnyk, R.B. Harness the Power of Big Data—The IBM Big Data Platform. 2011. Available online: http://www-01.ibm.com/software/data/bigdata (accessed on 10 August 2016).
16. Blumauer, A.; Pellegrini, T. Semantic Web und semantische Technologien. Zentrale Begriffe und Unterscheidungen. In *Semantic Web. Wege Zur Vernetzten Wissensgesellschaft*; Pellegrini, T., Blumauer, A., Eds.; Springer: Heidelberg, Germany, 2006; pp. 9–25.
17. Dengel, A. *Semantische Technologien: Grundlagen-Konzepte-Anwendungen*; Springer: Berlin/Heidelberg Germany, 2012.
18. Pepper, S. The TAO of Topic Map—Finding the Way in the Age of Infoglut. Available online: http://www.ontopia.net/topicmaps/materials/tao.html (accessed on 10 May 2017).
19. Gruber, T. Ontology. In *Encyclopedia of Database Systems*; Ling, L., Tamer Öszu, M., Eds.; Springer Science + Business Media: New York, NY, USA, 2009.
20. Gruber, T. Toward Principles for the Design of Ontologies Used for Knowledge Sharing. *Int. J. Hum. Comp. Stud.* **1995**, *43*, 907–928. [CrossRef]
21. Berners-Lee, T.; Hendler, J.; Lassila, O. The semantic web. *Sci. Am.* **2001**, *284*, 28–37. [CrossRef]
22. Halpin, H. Sense and references on the web. *Minds Mach.* **2009**, *21*, 153–178. [CrossRef]
23. Bizer, C.; Cyganiak, R.; Heath, T. How to Publish Linked Data on the Web. 2007. Available online: http://www4.wiwiss.fu-berlin.de/bizer/pub/LinkedDataTutorial/ (accessed on 10 May 2017).
24. Knowledge Processing with Big Data and Semantic Web Technologies. Available online: https://www.insight-centre.org/sites/default/files/publications/kcap_2015_copy_.pdf (accessed on 11 May 2017).
25. Ashburner, M.; Ball, C.A.; Blake, J.A.; Botstein, D.; Butler, H.; Cherry, J.M.; Davis, A.P.; Dolinski, K.; Dwight, S.S.; Eppig, J.T.; et al. Gene Ontology: Tool for the unification of biology. *Nat. Gen.* **2000**, *25*, 25–29. [CrossRef] [PubMed]
26. Hepp, M. GoodRelations: An ontology For Describing Products and Services Offers on the Web. In *Knowledge Engineering: Practice and Patterns. Proceedings of the 16th International Conference (EKAW), Acitrezza, Italy, 29 September–2 October 2008*; Springer: Berlin/Heidelberg, Germany; New York, NY, USA, 2008; pp. 329–346.
27. Staab, S.; Schnurr, H.-P.; Franz, T.; Hansch, D. Semantische Technologien und Auswirkungen auf Informations und Wissensmanagementsysteme. Available online: http://userpages.unikoblenz.de/~staab/Research/Publications/2008/SemantischeSysteme_Artikel_I_M-preprint.pdf (accessed on 17 June 2016).
28. Gams, E. Semantische Content Management Systeme. In *Social Semantic Web*; Blumauer, A., Ed.; Springer: Berlin, Germany, 2009; pp. 207–226.
29. Mantsch, M.-T. Anwendungsintegration für den Social Semantic Desktop mittels Publish/Subscribe. Ph.D. Thesis, Universität Koblenz-Landau, Koblenz, German, 2009.
30. Hepp, M. eClassOWL: A Fully-Fledged Products and Services Ontology in OWL. In Proceedings of the 4th International Semantic Web Conference (ISWC), Galway, Ireland, 6–10 November 2005.
31. Brunner, J.S.; Ma, L.; Wang, C.; Zhang, L.; Wolfson, D.C.; Pan, Y.; Srinivas, K. Explorations in the use of semantic web technologies for product information management. In Proceedings of the 16th International Conference on World Wide Web, Banff, AB, Canada, 8–12 May 2007; pp. 747–756.
32. Wauer, M.; Schuster, D.; Meinecke, J. Aletheia: An architecture for semantic federation of product information from structured and unstructured sources. In Proceedings of the 12th International Conference on Information Integration and Web-based Applications & Services, Paris, France, 8–10 November 2010; pp. 325–332.
33. Stolz, A.; Rodriguez-Castro, B.; Hepp, M. Using BMEcat catalogs as a lever for product master data on the semantic web. In Proceedings of the Extended Semantic Web Conference, Montpellier, France, 26–30 May 2013; pp. 623–638.
34. Fitzpatrick, D.; Coallier, F.; Ratté, S. A holistic approach for the architecture and design of an ontology-based data integration capability in product master data management. In Proceedings of the Product Lifecycle Management: Towards Knowledge-Rich Enterprises—IFIP WG 5.1 International Conference, Montreal, QC, Canada, 9–11 July 2012; Volume 388, pp. 559–568.
35. Shvaiko, P.; Euzenat, J. Ontology Matching: State of the Art and Future Challenges. *IEEE Trans. Knowl. Data Eng.* **2011**, *25*, 158–176. [CrossRef]

36. Otero-Cerdeira, F.J. Rodriguez-Martinez, A. Gomez-Rodriguez: Ontology matching: A literature review. *Expert Syst. Appl.* **2015**, *42*, 949–971. [CrossRef]
37. Euzenat, J.; Shvaiko, P. *Ontology Matching*; Springer: Heidelberg Germany, 2013.
38. Otero-Cerdeira, L.; Rodriguez-Martinez, F.J.; Gomez-Rodriguez, A. Ontology matching: A literature review. *Expert Syst. Appl.* **2015**, *42*, 946–971. [CrossRef]
39. Delone, W.H.; McLean, E.R. The DeLone and McLean Model of Information Systems Success: A Ten-Year Update. *J. Manag. Inf. Syst.* **2003**, *19*, 9–30.
40. Wang, R.Y.; Strong, D.M. Beyond Accuracy: What Data Quality Means to Data Consumers. *J. Manag. Inf. Syst.* **1996**, *12*, 5–33. [CrossRef]
41. Delone, W.H.; McLean, E.R. Measuring e-Commerce Success: Applying the DeLone & McLean Information Systems Success Model. *Int. J. Electron. Commer.* **2004**, *9*, 31–47.
42. Davis, F.D.; Bagozzi, R.P.; Warshaw, P.R. User acceptance of computer technology: A comparison of two theoretical models. *Manag. Sci.* **1989**, *35*, 982–1003. [CrossRef]
43. Wang, Y. Assessing E-Commerce Systems Success: A Respecification and Validation of the DeLone and McLean Model of IS Success. *Inf. Syst. J.* **2008**, *18*, 529–557. [CrossRef]
44. Esswein, S.; Goasguen, S.; Post, C.; Hallstrom, J.; White, D.; Eidson, G. Towards ontology-based data quality inference in large-scale sensor networks. In Proceedings of the 2012 12th IEEE/ACM International Symposium on Cluster, Cloud and Grid Computing (ccgrid), Ottawa, ON, Canada, 13–16 May 2012; pp. 898–903.
45. Martin, N. A Methodology and Architecture Embedding Quality Assessment in Data Integration. *J. Data Inf. Qual.* **2014**, *4*, 1–40. [CrossRef]
46. Fürber, C.; Hepp, M. Towards a Vocabulary for Data Quality Management in Semantic Web Architectures. In Proceedings of the 1st International Workshop on Linked Web Data Management (LWDM), Uppsala, Sweden, 21–24 March 2011; pp. 1–8.
47. Garijo, D.; Rey, M. A New Approach for Publishing Workflows: Abstractions, Standards, and Linked Data. In Proceedings of the 6th Workshop on Workflows in Support of Large-Scale Science, Seattle, WA, USA, 12–18 November 2011; pp. 47–56.
48. Rospocher, M.; Ghidini, C.; Serafini, L. An ontology for the Business Process Modelling Notation. *Front. Artif. Intell. Appl.* **2014**, *267*, 133–146.
49. Gartner Magic Quadrant for Business Intelligence and Analytics Platforms (2016). Available online: https://www.gartner.com/doc/reprints?id=1-2XXET8P&ct=160204 (accessed on 10 August 2016).
50. Sheldon, P.; Goetz, M. The Forrester Wave: Product Information Management (PIM), Q2 2014. Available online: http://www.hybris.com/medias/sys_master/formsCollaterals/formsCollaterals/hec/h36/8808347926558/Hybris-Forrester-PIM-2014.pdf (accessed on 10 May 2017).
51. Nigel, M.; Poulovassilis, A.; Wang, J. A methodology and architecture embedding quality assessment in data integration. *J. Data Inf. Qual.* **2014**, *4*, 17.
52. Batista, M.D.C.M.; Salgado, A.C. Information Quality Measurement in Data Integration Schemas. In Proceedings of the 5th International Workshop on Quality in Databases at VLDB, Vienna, Austria, 23 September 2007; pp. 61–72.
53. Gabler Wirtschaftslexikon Datenintegration. Available online: http://wirtschaftslexikon.gabler.de/Archiv/74965/datenintegration-v8.html (accessed on 10 May 2017).
54. Conrad, S.; Hasselbring, W.; Koschel, A.; Tritsch, R. *Enterprise Application Integration–Grundlagen. Konzepte, Entwurfsmuster, Praxisbeispiele*; Spektrum: München, Germany, 2005.
55. Heinrich, M.; Boehm-Peters, A.; Knechtel, M. A platform to automatically generate and incorporate documents into an ontology-based content repository. In Proceedings of the 9th ACM Symposium on Document Engineering, Munich, Germany, 16–18 September 2009; pp. 43–46.
56. Oberle, D. How ontologies benefit enterprise applications. *Semant. Web* **2014**, *5*, 473–491.
57. Otero-Cerdeira, L.; Rodríguez-Martínez, F.J.; Gómez-Rodríguez, A. Ontology matching: A literature review. *Expert Syst. Appl.* **2015**, *42*, 949–971. [CrossRef]