A Survey of Ontology-based Data Integration Approaches to Support Smart Manufacturing

Saad Bin Abid

fortiss GmbH
Guerickestraße 25, 80805 München
abid@fortiss.org

Abstract. In todays world, modern manufacturing vicinities are data enrich environments that are continuously producing a huge amount of data. This data needs to be analysed, transmitted and integrated in order to be useful to improve the manufacturing intelligence. Manufacturing industry is now taking advantages of the concepts like that of Internet of Things (IoTs) and Cyber Physical Systems (CPSs) to have a competitive edge by making efficient and effective manufacturing processes. The introduction of intelligent physical devices on the factory floor to improve production are producing a massive amount of data that is in different source formats that is only understandable in a particular part of an operational environment. Data integration (DI) techniques allow different stakeholders (e.g., managers, manufacturing and operational engineers) to access data produced by intelligent devices across the board and helps them to understand the semantic meaning of the data. This understanding of the data also helps to make smart decisions that can improve the manufacturing and operational capabilities. In this context, we have conducted a survey on the current state-of-the-art (SOTA) on ontology-based DI approaches and technologies available. We summarise the current literature, compare the identified approaches, provide strengths and weakness and discuss the industrial use-cases and benefits of the ontology-based DI approaches to foster smart manufacturing (SM).

1 Introduction

Smart manufacturing (SM) is the application of advanced communication systems to conventional manufacturing processes, making them more flexible, efficient and responsive1. Smart manufacturing vicinity is a data-driven operational environment. Data-driven operational environment requires data produced by various sources in heterogeneous formats to be integrated, understood and analysed in order to provide operational intelligence. This intelligence can reduce the resource (e.g., energy, natural resources) usage, increase efficiency, productivity and minimize the time-to-market [32]. Various authors have made attempts to facilitate smart manufacturing, for instance ODonovan et al. [30-31] focused on equipment maintenance and provide industrial big data pipeline architecture. Data analytics applications based on the designed big data pipeline is suggested by the authors to achieve efficient maintenance of manufacturing equipment.

Data sources can be heterogeneous in syntax, schema, or semantics, thus making data integration and interoperability a difficult task [29]. Data heterogeneity can be classified as 1) Syntactic heterogeneity (i.e., use of different models or languages), 2) Schematic heterogeneity (i.e., structural differences) and 3) Semantic heterogeneity (i.e., different meanings or interpretations of data in various contexts) [29]. Ontology research is another discipline that deals with semantic heterogeneity in structured data and provides a conceptual representation of the data and of their relationships to eliminate possible heterogeneities [29]. The data sources to be integrated can be classified as 1) a real-time data sources (e.g., sensors) or 2) a pre-existing databases (i.e., legacy databases or log files). Without the Semantic DI, it is difficult to get the context and meaning of the data altogether. It can also become a difficulty for the stakeholders (e.g., operational manager, maintenance engineer, quality control experts) in the manufacturing environment to 1) perform data analytics in order to analyse, 2) comprehend the data and perform efficient decisions to increase time-to-market and 3) reduce the resource consumption (e.g., electricity from grid) to lower the equipment maintenance costs. In this context, we have performed a survey on ontology-based approaches that can support smart manufacturing by enabling semantic DI of the data produced by heterogeneous data sources (e.g., Sensors or databases). We have also discussed the strengths and weakness of the Ontology-based approaches and discussed the use cases where ontology-based approaches might not be feasible. The reminder of the paper is as follows, Section 2 discusses our literature collection methodology and research questions, Section 3 provides a brief overview on Ontologies, in Section 4 we summarise the identified approaches and provide comparison between them, Section 5 provides discussion of strengths and weakness of the ontology-based approaches and concludes the paper.

2 Literature Review Methodology

The aim of this Section is to familiarize the reader with the literature collection process and formulation of research questions.

2.1 Research Process

This sub-section discusses the investigation and analysis method inspired by the Review and Analysis method is utilized for literature collection in the context of smart manufacturing formulated and suggested by Kang.et.al.[33]. However, we have tailored that to our needs in order to provide a formal way of searching and analysing the DI techniques/approaches for smart manufacturing. The overall procedures are shown in Figure 1. The process starts by 1) searching and reviewing the DI approaches in SM domain, 2) selecting from the identified DI approaches targeting semantic DI, 3) categorizing the identified techniques under the main headings (e.g. approaches providing 1) reference model, 2) frameworks or 3) languages) and 4) discussion of the categorized and classified approaches. Discussion and analysis is performed by answering the research questions formalized in Section 2.2. We have not utilized a systematic literature review rather we have searched for ontology-based DI approaches and technologies. This limitation of selecting the research work has let us focus on only on the semantic DI approaches.

2.2 Research Focus

This sub-section provides a discussion on the research question formulated in order to initiate and facilitate the literature search and collection process. The main research question is formulated with the narrow scope which is intentional. The main reason for having the research question with a narrow scope is to serve the purpose of identification of literature on ontology-based DI approaches. The research question not only helped us to narrow the scope but also helped us to classify the identified approaches. Main Research Question: How are ontology-based data integration (DI) approaches utilized to support smart manufacturing? To answer the main research question, we have formulated two supportive/auxiliary questions (SQs). These supportive questions relates to various aspects of ontology-based DI approaches in smart manufacturing. The formulated smaller questions also address and answer different aspects of the main research question. SQ1: What ontology-based DI approaches/techniques suggested by the literature to support smart manufacturing? Rationale: The main idea of formulating this research question is to initiate exploring the current literature on data integration and focus on approaches that are focusing on ontology-based DI approaches that can support SM. SQ2: What are the weakness and strengths of the Identified ontology-based DI approaches/techniques? Rationale: The main idea of this supportive question to identify the strengths and weakness of ontology-based approaches and identify the use cases where semantic integration may or may not be applicable.

3 Ontology Overview

Definition: According to Gruber [34] ontology can be defined as follows, Ontology is a formal, explicit specification of a shared conceptualization. Ontology represents a schema that captures the domain concepts and relationships among the domain concepts. Ontology allows sharing common understanding of the structure of information among people or software agents. Ontology is a machine readable or formal in nature[29] Ontologies were initially developed by Artificial Intelligence (AI) community to enable knowledge sharing and reuse [29]. A common use of ontologies is data standardization and conceptualisation via a formal machine readable ontology language [5]. According to Noy [5], the following are the main steps to create ontologies,

- defining terms in the domain and relations among them,
- defining concepts in the domain (classes),
- arranging the concepts in a hierarchy (subclass-superclass hierarchy),
- defining which attributes and properties (slots) classes can have and constraints on their values
- defining individuals and filling in slot values (instances).

3.1 Languages for Ontology

As mentioned earlier, Ontologies are machine readable artefacts. There are a number of languages available to develop Ontologies, those are as follows, Ontology Web Language (OWL)1 is a semantic markup language that represents complex knowledge about things, group of things and relationships among them. OWL is utilized for publishing and sharing ontologies on the web. Extensible Markup Language (XML) Schema2 is considered to be semantic markup language developed for Web data. XML language has a simple, very flexible text format derived from SGML (ISO 8879). The database-compatible data types supported by XML Schema provide a way to specify a hierarchical model3. Resource Description Framework (RDF)4 developed by W3C is a metadata data model developed for describing Web resources. RDF extends the linking structure of Web to use URI (i.e., Universal Resource Identifier) to name the relationships between the things as well as two ends of the link (i.e., referred to as a triple). A triple in RDF forms a statement that looks like (Resource, Property, Value). Combining triples forms a directed, labelled graph. The directed graph can be visualized to to understand the relationships among the domain concepts. RDF also allows data merging even if the underlying schemas are non-similar. RDF Schema5 is a language describing the vocabularies of RDF data in terms of primitives such as rdfs:Class, rdf:Property, rdfs:domain, and rdfs:range. In other words, RDF Schema is utilized to define the semantic relationships between properties and resources. DAML+OIL (DARPA Agent Modelling Language)6 is also a semantic markup language for Web Resources that extends RDF and RDF Schema with richer modelling primitives. This ontology development language has XML-based syntax and layered architecture. Other ontology development languages are Unified Modelling Language (UML7), Ontology Exchange Language (XOL8) and Simple HTML Ontology Extensions (SHOE9). Apart of languages there are tools to build, edit and visualise ontologies. A few ones are Protg10, Ontosaurus11, OntoEdit12, WebODE13 and Apollo14. Open Services for Life Cycle Collaboration (OSLC15) is also a set of practical specifications (can be called Ontology language) built using XML alike schemas and RDF for data interoperability among software tools. Eclipse Modelling Framework (EMF16) also provides a UML like language called ECore17 language that also allows its user to perform ontology development process to build a domain-specific language (DSL).

3.2 Types of Ontologies for Data Integration

As explained earlier, ontologies are built with the purpose of capturing the domain knowledge and relationships among the domain concepts. In the context of DI, they are utilized as follows [35],

Single Ontology Approach A single ontology is considered as a global reference schema/model to all the sources in the system. All of the source schemas are directly related to the global schema which is shared among all source schemas.

Also the global ontology provides a uniform interface to the user. This approach requires all sources to have a common view of the domain concepts. This is the simplest approach that can be simulated by other approaches. A single ontology approach can utilized for data integration scenario where all the information or data sources to be integrated have almost a common view on the domain. For instance, consider a scenario where there are three heterogeneous data sources (Building Management System, energy monitoring system (MandT) and SCADA) wants to represent temperature sensors data and have a common granularity level. It would be easier to utilize single ontology approach in this situation. It would be really difficult to perform integration and find a common minimal ontology when one of the data source has a different view of the domain or if the data to be integrated are not at the same level of granularity. Depending on the nature of the changes in one information source it can imply changes in the global ontology and in the mappings to the other information sources. The limitations of single ontology approach led to the development of multiple ontology approaches. Figure 2 graphically represents single ontology approach.

Multiple Ontology Approach In multiple ontology approach, each of the data sources has its local schema defining it. For DI purpose there is a mapping between local ontologies (i.e., point-to-point DI). Example scenario could be when three HVAC systems (i.e., BMS, MandT and SCADA) require data integration and one wants to represent data sources (e.g., sensor temperature) according to the local domain. In such a scenario, each data source has its own local ontology (representing sensor temperature) and the integration is achieved via mappings between the local ontologies. Apparent advantages of multiple ontology approaches are 1) no need to manage a global ontology which puts limitations for each information source to have a common view of the whole domain, 2) modify the resources (i.e. add, remove and change) in local ontology without taking into other local ontologies and 3) resources can be represented at different levels of granularity. Although the multiple ontology approach tries to address single ontology limitations, however, there are challenges in applying multi ontology approach, for example, this kind of approach 1)requires additional mappings to be maintained per couple of local ontologies which is an overhead, 2) can be difficult to implement when same data has different semantic meaning and represented at different levels of granularity and 3) inter-ontology mappings are difficult to manage since they grow exponentially and 4) attempts to Integrate concepts that may resides at different level of granularity which becomes a really challenge when adding more data sources. Figure 3 represents a multi ontology approach.

Hybrid Ontology Approach In order to mitigate the issues and limitations of Single and Multiple Ontology approaches, there is another approach called Hybrid Ontology Approach in Literature. In hybrid approach, every data source has its local ontology providing semantic definition of sources and there is a global ontology representing a global view of the domain. The global ontology contains common concepts of the domain represents a common view of the sources of the domain. The interesting part is that how to develop the local ontology. Local ontology is made up of common concepts or vocabulary concepts and adding the local operators from the domain. These operators are local concepts that provide semantics to the global concepts and make them domain specific. A global ontology is utilized for comparing local ontologies with each other. For instance, a scenario could again be when one wants to integrate three data sources (BMS, SCADA and MandT) to integrate temperate sensors. In this situation temp will be a global concept residing in the global ontology and building temperature there is room1 underscore temp concept, for field sensors there is field1 underscore temp and for manufacturing machine temperature there is machine1 underscore temp. in this situation domain operators/attributes (i.e., room1, field1 and machine1) are attached to temp to make it more domain specific. Advantages of this approach are 1) it is easier for new data sources to be added without the need of additional mappings, 2) it reduces the number of mappings to be developed and managed between local and global concepts and 3) it supports evolution of ontologies. One of the biggest drawbacks is that one cannot reuse the existing local ontology leading towards development of local ontology from scratch. Our focus in this paper is on all of the aforementioned ontology-based DI approaches. In SM there exist data sources (e.g., Sensors, databases) that are heterogeneous in nature and produce data in multiple formats. These data sources can be integrated with single, multiple or hybrid ontology approaches. Cruz and Xiao [29] identified five uses of Ontologies in the context of DI, that are 1) Metadata representation, 2) Global conceptualisation, 3) Support for high-level queries, 4) Declarative mediation and 5) Mapping support.

4 Survey: Ontology-based Data Integration

This Section discusses the industrial use cases in manufacturing domain where ontology-based approaches are applied for DI and provides a comparison between them. A few of the comparison criteria factors are associated with guidelines provided by smart manufacturing leadership coalition (SMLC1). These comparison factors are, 1) Scalability, 2) Interoperability and 3) security. Rest of the comparison factors are perceived while reading the papers (i.e., 1) type of ontology utilized, 2) data source addition,3) domain in which utilized, 4) temporal data processing, 5) evaluation, 6) capacity to manage large data 7) types of data sources for DI and 8) data integrity loss). Our data collection process is inspired by Kang et al. [33]. Kang et al. [33] developed this framework to identify literature on smart manufacturing. We tailored work by Kang et al. [33] for our purpose to identify the literature on semantic integration of data sources in manufacturing domain that can support smart manufacturing paradigm (Section 2.1). Ontologies have been extensively used in DI systems because they provide an explicit and machine-understandable conceptualization of a domain [29]. This survey is considering ontology-based DI approaches that can support smart manufacturing. The following paragraphs will provide a brief overview of the approaches (frameworks, architecture and process models) that has taken ontologies as a primary way to integrate product information among data sources in enterprises. Work by Jian et al. [1], Chang and Terpenny[6],Bohlouli et al. [17], HEFKE et al. [3], Uzdanaviciute and Butleris [8], Fang and Wang [12] and Gangon [39] provide hybrid ontology-based DI approach, whereas, Alm et al. [38] provides a single ontology-based approach for DI. All of the approaches implementing hybrid ontology-based approach DI provide scalability because of the fact that it is relatively easier to add new data sources by adding a local ontology which represents the local/enterprise data source to be integrated. However, this gets manually effort demanding when it comes to single ontology (e.g., Alm et al. [38]). Most of the identified approaches attempt to solve the DI challenge in the manufacturing domain. Most of the approaches (e.g., Jian et al. [1], Chang and Terpenny[6],Bohlouli et al. [17], HEFKE et al. [3], Uzdanaviciute and Butleris [8], Fang and Wang [12] and Gangon [39]) provide DI between data sources that is product-related information residing in databases. The data sources can also be simple digital annotations and operational procedures (i.e., in Alm et al. [38]). Almost none of the identified approaches attempted to perform temporal data processing (i.e., temporal data is the data produced by sensors). Most of the identified approaches achieved interoperability by providing mappings between local and global ontologies. Only a few approaches have taken security into consideration and provided mechanism to deal with security related issues. For instance, Jiang et al. [1] provided security through access control layer of the proposed framework. Whereas Bohlouli et al. [17] provide cloud security layer and suggested a security mechanism (e.g., public key encryption, secure sockets layer (SSL) in their proposed cloud-based layered framework. Capacity to manage large data is also an important factor because as discussed earlier smart manufacturing is a data-driven environment. None of the approaches addressed the capacity to manage large data factor. Only work by Bohloui et al. [17] addressed this factor by providing a cloud based approach and implementing DI as a Service. Most of the identified approaches provide prototypes as a proof-of-concepts. Apart of a few approaches that utilize either manufacturing enterprises (e.g., Fang and Wang [12] and Chang and Terpenny [6]) or commercial tools (e.g., Plant@Hand in Alm et al. [38]). Almost all identified approaches for DI enables minimum data integrity loss because of the fact that there is a global schema and local ontologies present to define the concepts of the domain and provide semantics. Also, interoperability is achieved through mappings between local and global ontologies. Table 1 provides a comparison between the identified ontology-based DI approaches. Other noticeable works are by Noy [5], Wache [10], Cruz and Xiao [29] and Mansukhlal and Malathy [11] that provide surveys on semantic integration using ontologies.

5 Survey: Ontology-based Data Integration

5 Discussion and Conclusions This section discusses the advantages and disadvantages of ontology-based DI approaches and provides concluding remarks.

5.1 Discussion of Strengths and Weakness of Ontology-based DI Approaches

Section 4 provided discussion of uses cases of ontology-based approaches for DI in industry and literature. In this sub-section we are going to discuss advantages and disadvantages of ontology-based DI approaches. Ontology-based DI approaches using ontologies have some benefits that are as follows It is evident from the industrial use cases cited in Section 4 that the ontologies employed as the global schema are either an existing domain ontology or foundational ontology. The global schema provides enough coverage in the conceptualization of the domain to cater for heterogeneity from the potential data sources. The vocabulary provided by the global ontology serves as a stable conceptual interface to the databases and independent of the database schemas for DI Ontology-based DI approaches apply semantic meaning and expressiveness to the data. It is possible to perform semantic integration of the heterogeneous data. Ontology-based DI approaches also enable consistent management and recognition of inconsistent data. Ontology-based DI approaches advocate the utilization of reusable domain concepts. The language by the ontology is expressive enough to address the complexity of queries typical of decision support applications. Unlike approaches utilizing databases for DI, ontology-based DI approaches take advantage of ontologies that are designed with the purpose of improving interoperability. Based on completeness of a centralized or global schema/ontology (i.e., in the case of hybrid approach) it is possible to achieve data integrity (i.e., data completeness, accuracy and consistency). This leads towards intended DI. Security of accessing the data resources can be achieved by applying local security policies (e.g., local authentication and certificate exchanges). It is also possible to take advantage of proven manufacturing and production standards (e.g., ISO-10303 STEP, STEP-AP242), OPC-DA for data access and OPC-UA for Machine-to-Machine (M2M) communication protocol. Although, ontology-based DI approaches provide an efficient DI solution using common vocabulary for domain concepts. They still lack the following, To design and implement an ontology-based DI approach one has to have an adequate knowledge of semantics expressed in the ontology. Large and complex ontologies are more difficult to understand leading towards extra effort. Large and complex ontologies can be restrictive to be accepted by new parties involved for DI by asserting strong ontological commitments through complicated and tightly-coupled axiomatization. In such a situation ontologies might not be accepted by potential parties that joined later for DI purposes. This also leads to future extendibility issues. In the use cases where there is a no conceptual model/ontology available for the application. One has to develop a new ontology from scratch and this requires a large and complicated manual effort especially when one has to capture concepts and relationships between them in a large domain. This situation also contributes towards a large manual effort to design and develop DI approach. One of the problems with ontology-based approaches arises when there is not much of semantic understanding of the data sources required by the user. An example scenario can be when the user wants to perform manual analysis of the data integrated from various data sources (i.e., weather sensors) by putting it in simple csv files or excel sheets. In such a scenario, there is not much need of having semantics involved. In this situation the ontology-based approach might not be a viable option to take. It is a challenge to manage interfacing via explicit rules or writing adapters between local ontology and global ontology for DI. The interfacing should be consistent and vocabulary to be updated continuously in both the local and global ontology along with the mapping rules. Global ontology or schema grows as new enterprise ontologies are added representing heterogeneous data resources (locally from shop-floor or from geographically distributed enterprises). This also requires new mappings to be defined that can introduce more manual effort to develop and maintain new mappings. Data sharing is only possible upon the availability of global schema/ontology (in the case of hybrid approach) and also the mappings between local and global ontologies. Hence DI depends on availability of the global schema. Translation of real time sensory data to ontology requires additional cleansing of temporal data. Validity of global schema depends on the correctness of the local ontology

5.2 Concluding Remarks

Smart manufacturing vicinity is a data-driven operational environment. Data-driven operational environment requires data produced by various sources in heterogeneous formats to be integrated, understood and analysed in order to provide operational intelligence. In the context of DI, we have provided a survey focusing on ontology-based DI approaches that can support smart manufacturing by providing semantic integration. We summaries the identified literature provided comparison and discussed strengths and weakness of them. Ontology-based approaches are also well suited when, for instance, there is a requirement of semantic integration among heterogeneous data sources, semantic understanding of integrated data for data analytics and standardization of communication using proven standards like that of STEP, OPC is also a requirement. However, there are industrial use cases where ontology-based approaches are might be not feasible, for instance, 1) when there doesnt exist a conceptual model upfront of the domain leading towards additional effort of creating ontology from scratch, 2) when stakeholder doesnt require any semantic information of the integrated data and wants to perform manual analysis and 3) when data sources from heterogeneous manufacturing sites finds the global ontology to be very restrictive and rigid to map local domain concepts residing in local ontology. The takeaway message from this survey is that although ontology-based DI approaches proved an effective way for semantic integration in manufacturing domain. Ontologies can only be utilized based on context and availability of certain knowledge upfront and not in all industrial use cases. An example industrial use case could be when there is a requirement of manually performing data analytics by performing semantic DI that resides in plain CSV files. In this scenario, semantic integration will prove to be annotation or semantic information intensive. In order to improve the scalability and data availability issues, one can utilize the cloud computing paradigm and provide DI as a service (e.g., work by Bohlouli et al. [17]). For future work, we would like to add more DI approaches for semantic integration of data sources in manufacturing domain, provide empirical evidence to support our claim and apply systematic literature review to enhance our survey.

References