A holistic approach for the architecture and design of an ontology-based data integration capability in product master data management

Daniel Fitzpatrick, François Coallier, and Sylvie Ratté

École de Technologie supérieure, Montréal, QC Canada
{Daniel.fitzpatrick,francois.coallier,Sylvie.ratte}@etsmtl.net

Abstract. In the context of a broadened product lifecycle management environment, a traditional product information management, also referred to as product master data management (P-MDM) needs to be complemented by other MDM domains. Such MDM domains may include Customers, Financials, Suppliers, Human Resources, Events and other domains. To satisfy such a transversal set of requirements requires a true cross-enterprise semantic integration capability. This capability cannot be met by current off-the-shelf technologies. This paper proposes a research approach that would elicit the definition of a reference architecture and a multi-domain ontology, from research and development work performed notably in ontology engineering, in both academic and industry domains.

Keywords. Product lifecycle management, product master data management, ontology-based data integration, data architecture, qualitative research

1 Introduction

Product lifecycle management (PLM) is one of the keystone paradigms that bring value to the stakeholders, notably shareholders and customers. In the aftermath of what is currently called the great recession, PLM processes are focused to sustain growth, to improve products and processes on a continuous basis and eliminate wasteful activities and constraints.

PLM is defined by [1] as a «product-centric – lifecycle-oriented business model, supported by ICT (Information and Communication Technologies), in which product data are shared among actors, processes and organizations in the different phases of the product lifecycle for achieving desired performance and sustainability for the product and related services.».

While not contradicting this definition, this paper, and the underlying research, considers master data as data that constitute the foundation of all business transactions, as pervasive, cross-enterprise assets that contributes not only to PLM, but to other keystone paradigms such as customer-centric (CRM), supply-chain/vendor-

centric (ERP) and employee-centric (HRM). Taking an epistemological stance, this research distinguishes between factual data, information and knowledge. Here, information is defined as contextual data, and knowledge being actionable information [2, 3]. Information and knowledge are produced contextually by the processes of each business paradigms. Therefore, a holistic architectural approach in implementing master data management (MDM) is needed to ensure that not only PLM's data requirements are met, but those of a collaborative environment in which the other process paradigms relate with PLM.

To illustrate the importance of the distinction between data and information, one can compare the customer lifetime value (CLV) concept, a key metric used in the CRM processes, and the product margin concept, equally important for PLM. While drawing from the same transactional data, both elements of information are produced using different analysis axes in different contexts (PLM vs. CRM).

One of the key MDM function is data integration. It is considered as a daunting scientific and industry research subject. While important advances have been made, especially with the use of artificial intelligence technologies and of the ontology concept, more research is needed to make data integration a stable background function.[4]

Considering that Information technology (IT) is relatively deprived of a theoretical foundation, [5] a theory building approach vs. a theory testing approach is prescribed. For the purpose of building a proper theoretical framework, a qualitative research design such as the phenomenological research method can be used to induce new knowledge and know-how from the industry, in this research case related to data integration. As [5] posits: «… knowledge can be developed (in drawing from) from academic research and (also) from practice. ». The proposed research intends, more specifically, to determine data architecture patterns that can be used as generic solutions to address data integration issues that affect PLM.[5] The research project intends to accomplish this by inducing knowledge and know-how from experienced practitioners, which include one of the authors, with an actual active involvement in the design of multi-domain data integration capabilities in PLM.

2 Research objective

The research project aims in the formulation a reference architecture and a formal ontology from data architecture patterns induced from the field research phase of the project. This multi-domain ontology would intend to provide the basis for a reusable semantic structure of a data integration capability.

Such multi-domain ontology would help the enterprise architects and designers to put in place the data services to integrate data within an enterprise data warehouse or a SOA infrastructure for the specific purpose to supply master data to PLM and other business paradigms. This method should also assist developers using various tech-

nologies such as Extract-Transformation-Load (ETL) and semantic reasoning tools, to resolve mappings from heterogeneous sources to the data integration ontology.

3 Literature review

3.1 Product Lifecycle Management

This business paradigm comprises human, material and informational resources, along with processes to guide and operate the various activities involved for each product from the early stages of R&D and design, or beginning-of-life (BOL), thru the commercial stage of the product life, or middle-of-life (MOL), and terminating at its decommissioning, or end-of-life (EOL). [1]

PLM has become a more complex set of processes, involved in creating value for shareholders and customers alike. It involves using information, knowledge and know-how to continuously improve on product efficiency, performance and quality. Some of its processes involve the capacity to trace manufacturing errors and other quality and performance issues, to track product through logistics store and transport, material recycling and energy saving, to name a few. Finally, PLM also involves optimal decision-making through product lifecycle stages, from BOL to EOL. A data integration capacity ensures that proper timely information and knowledge is made available for PLM processes and also for collaborative activities with other business paradigms, such as the customer-centric CRM. Table 1 illustrates various types of data needed for the PLM product life stages. [1, 6] This is only a minimal list of types of data. This research is likely to unearth a much greater list.

Table 1. Types of data needed at the PLM product lifecycle stages

PLM Product life stages	Types of data
Beginning-of-life	Product, equipment, material, plant, employees, tools, techniques, methodologies, document, suppliers,
Middle-of-life	Product, customer, employees, services, service providers, events, geography, financials, document
End-of-life	Product, customer, service, service providers

3.2 Master data management (MDM)

Master data can be defined as the most important data that would assist the organization in reaching its objectives. Master data is used to produce valuable contextualized information and knowledge to support PLM. [7]

There exist several data taxonomies. These taxonomies reflect the context in which data is considered, i.e. managed, stored and exploited. The proposed data classifica-

tion scheme by [8] comprises the following key kinds: Metadata, reference data, master data, transaction data and historical data. The authors in [8] recognize a relative state of confusion in respect to this taxonomy. Any of the types of data indicated in the aforementioned taxonomy could be considered as crucial, therefore being master data, for a given enterprise in its business dealings. This paper considers any data as potentially master data in the context of a specific enterprise's PLM environment. Ultimately, data domains, such as parties, products and others constitute a more reliable method to classify data. A more complete list of data domains and associated ontology axioms is found in the preliminary results section of this paper.

[8] and [9] propose notably the Coexistence implementation style with trickle feed (see figure 1) that integrates data from heterogeneous sources in a batch mode in the context of an enterprise data warehouse environment. It returns integrated master data to its sources, but usually also in a batch mode. Although it produces a golden record that can be used to alter master data located in source system, it does not constitute a system of record since change is not instantaneous. Great care must be taken in the correcting master data in operational systems using the MDM's golden record. It uses a physical database instance and uses a read-only mode approach. In some cases, a direct Enterprise Application Integration (EAI) feed is added to allow some near-real time or even real-time events or other data to be loaded for intraday event processing. The coexistence implementation style is often used for the design of an enterprise data warehouse for the PLM paradigm.

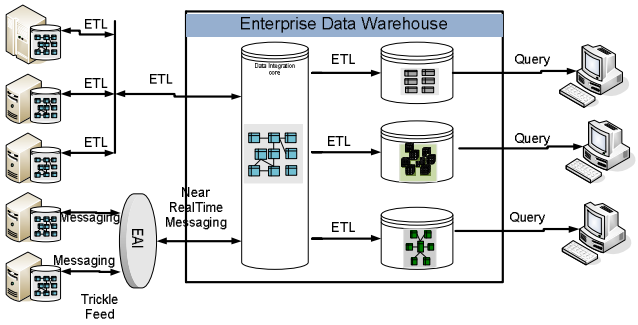

Fig. 1. Coexistence implementation style with trickle feed

3.3 Ontology

An ontology is defined as an «explicit representation of a shared conceptualization». [10] The basic purpose of the ontology is to produce a shareable and reusable set of information elements to be used by people and computer systems. Also, the ontology must distinguish between domain knowledge that may be extra organizational versus localized application level knowledge. The criterion of orthogonality is defined as the requirement of basing a newly created ontology on one or more existing ontologies. This practice, if generalized, would help reduce the silo effect in the development of

ontologies. It would therefore favor the trend toward a greater universal interoperability across all industries and government domains [11]. The preliminary results outlined in this paper illustrate how the criterion of orthogonality is applied.

The ontology is encoded in a computer treatable format to insure interoperability over a network [12]. [13] Distinguishes ontology engineering from conventional data modeling. Ontology engineering covers a domain wide semantics and relationships within. The emphasis in defining the ontological approach is put on the shareability aspect. Data modeling establishes, according to the authors, a semantics structure specifically suited for an individual system with no consideration with other systems or applications. This dichotomy between ontology and data modeling is disputed by [14] who promotes a continuum approach to ontology engineering and data modeling. Such dispute on the theoretical foundation of ontology engineering may hinder its wide adoption in the industry.

3.4 Data integration

Taken holistically, data integration represents the computerized capability to address the problem of providing data thru a single perspective from heterogeneous sources located within an organization [15]. Along with data quality, data profiling and other MDM functions, data integration attempts to service the organizations and the community at large with the widest perspective possible. Data is usually located in specialized systems. These silos are difficult to link together to provide transversal views of the data. There is a growing need to deliver cross-domain data, a usually highly difficult task considering that there are rarely any common semantic convention that may allow interoperability amongst systems [16].

[16] proposes a common data integration architecture composed of wrappers and mediators. In this architecture, source databases or systems are wrapped by specialized software components that convert the source's local semantics into a global set of shared concepts. The wrappers allow the source to which it is attached to interact with the rest of the world. Mediators are components that issue queries or sub-queries to wrappers or other mediators to gather data. Mediators are views that are designed to satisfy queries issued by humans and systems. Persistent forms of mediators are also designed in the form, notably, of data warehouses.

Within the framework of master data management, mediators are data services. Some mediators may be implemented using a registry architecture style, in a pure virtual view approach. Others would be implemented using a persisted data structure such as in the coexistence or the transactional hub implementation styles. An ontology-based approach for data integration in the context of PLM would be the best suited for the PLM's demanding requirements. [6]

3.5 Interpretation of the literature review

One question still remains about what exactly is being actually integrated: data, information or knowledge. This fundamental question may influence the efforts to ultimately build a comprehensive theoretical framework dedicated to semantic integration. The current understanding of what is master data and data integration is continuously being questioned.

The adoption of the ontology discipline has been hailed as a significant step in the right direction for true intra and extra organizational interoperability. However, it remains unclear how to design an ontology dedicated to semantic integration that would address the issues of dealing with siloed legacy systems and semantic reusability. Furthermore, it also remains unclear how the creation of such ontology would fit as a step in a data integration capability architectural approach. This lack of clarity originates from what the researchers perceive as divergences amongst the authors. This paper considers the concept of a multi-domain ontology as the keystone of a cross-enterprise data integration capability, such as presented in [17].

4 Preliminary results

Along mainly with data collected from some of the participating practitioners, material consulted from [18-20] in semantic data warehousing and in semi-structured and unstructured data treatment specifically on corporate documents from [21], were used to provide an up to date position of the research project. Inspired by the MDM coexistence implementation style, discussed in section 3.2, a reference architecture of a semantic enterprise data warehouse, as illustrated in figure 2, is proposed to provide a multi-domain data integration capability to support contemporary PLM.

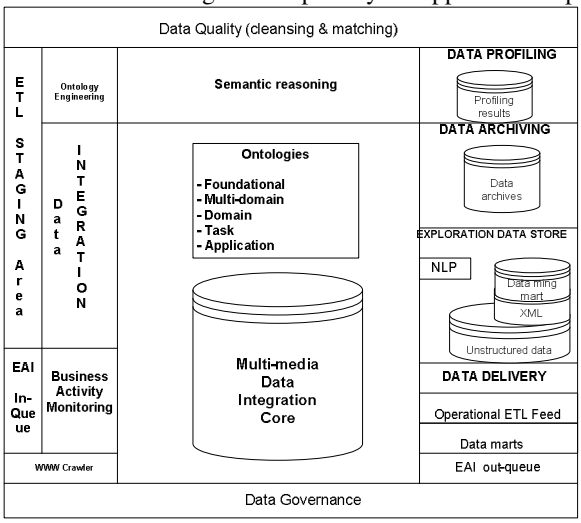

Fig. 2. Reference architecture of a semantic enterprise data warehouse

The proposed reference architecture of the semantic enterprise data warehouse could be used to design a multi-domain data integration capability, notably, to support PLM processes as defined by [1]. It would also include other MDM functions such as data quality, data profiling and data archiving, which are essential in insuring effective cross-enterprise data integration for operational and business intelligence applications. Semi-structured and unstructured data can also be extracted internally in the enterprise and externally on the web, and, be annotated with tokens allowing linking with structured data. In light of the criterion of orthogonality, figure 3 subsumes the proposed multi-domain data integration ontology in respect with the foundational ontologies such as SUMO and others. Domain specific ontologies such as Onto-PDM proposed by [7] which incorporates product technical data standards STEP and IEC62264 are subsumed to the multi-domain ontology proposed in this paper. Then, the ontology structure comprises generic task ontologies, such as for natural language processing (NLP), for dealing with semi-structured and unstructured data, and for mapping heterogeneous sources to the Data Integration Core. Finally, the structure is completed with application ontologies to support domain specific tasks such as processing unstructured text from social media regarding PLM.

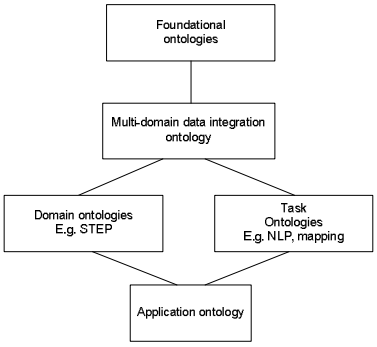

Fig. 3. Reference architecture ontology structure

Figure 4 identifies data domains that would compose the multi-domain data integration ontology. In its final formal form, each of these data domains, and others, would include one or more axioms that would serve as the core concepts allowing cross-enterprise interoperability to fully support PLM. Some of these data domains are already well known in the data modelling community. The Party concept was first published by [22] and successfully used in several enterprises and industry data models to represent customers, vendors, employees, partners, organizational structures and more.

Through the remaining part of the research project, these artefacts will be detailed while validated by a committee of experts from the scientific and industry realms. The completion of these artefacts will be done through knowledge extraction performed using the research method described in the following section.

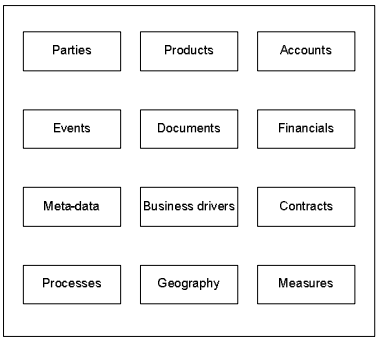

Fig. 4. Data domains for the multi-domain data integration ontology

5 Research method

As noted earlier, the current IT theoretical frameworks do not adequately support the industry in terms of knowledge and know-how in respect to ontology-based data integration. No existing methodology would allow, without research, to elaborate an ontology-based architecture approach of a cross enterprise data integration capability to support PLM. A qualitative research project to achieve the research objective is therefore warranted. For this purpose, a theory building qualitative research approach is considered here to tackle this research project problem and question.

Several research designs were investigated and analyzed for this specific project, such as ethnographic, content analysis, case study, grounded theory, analytical induction and others. The researchers find that a research approach based on the phenomenological method, as pioneered by Clark Moustakas [23] would be the most appropriate and effective to fulfill this research's objective. The phenomenology-inspired research protocol in this project involves a series of semi structured interviews and focus group sessions to collect architecture patterns related to the implementation of a data integration capability, complementing the analysis of the available technical documentation.[24]

In addition to allow the extraction richer pattern-like information throughout the field research part of the project, the phenomenological approach provides two other important benefits: it assists the researchers to better select the interviewees («first-persons») and allows one of the researchers to submit himself to a very rigorous and effective preparation to better conduct interviews. [25] The data collection processes are executed in the context of the field research phase of the project in which a minimum of 15 participants are interviewed individually. A semi-structured interview approach is used. The interview questionnaire is designed to elicit rich information and knowledge from industry experts and seasoned practitioners that have actually contributed to the design of a multi-domain data integration capability. Through the analysis processes, conceptual data modeling patterns would be identified along with

valuable methodological heuristics such as how to ensure the reusability and robustness of the underlying conceptualization, used for the specific purpose of data integration. These findings will be used to formulate the intended reference architecture and multi-domain ontology. The final results of this project will be subjected to a validation process with the contribution of a 20-member committee composed of subject matter experts from the scientific and industry realms.

6 Conclusion

As the PLM paradigm evolves and its data requirements become more complex, there is a need for a holistic architecture approach to design and implement a master data management environment. A new trend in data integration academic research that uses a formal ontology processed by a semantic reasoner, provides a promising direction to resolve system interoperability issues. In addition to the academic advances, this research project will leverage on the industry's efforts to implement cross-enterprise multi-domain conceptualization, through an adapted qualitative research method. Although the multi-domain conceptualization expertise developed in the industry, and sought in this research, was meant for designing databases, its contribution can be invaluable in solidifying a badly needed theoretical framework for ontology engineering for the design of data integration capabilities not only for PLM, but for other process-centric paradigms as well.

References

1. Terzi, S., et al., *Product lifecycle management - from its history to its new role.* International Journal of Product Lifecycle Management, 2010. **4**(4): p. 360-89.
2. Liew, A., *Understanding data, information, knowledge and their inter-relationships.* Journal of Knowledge Management Practice, 2007. **8**(2).
3. Bouthillier, F. and K. Shearer, *Understanding knowledge management and information management: the need for an empirical perspective.* Information research, 2002. **8**(1): p. 8-1.
4. Halevy, A., A. Rajaraman, and J. Ordille. *Data integration: The teenage years*. 2006: VLDB Endowment.
5. Gregor, S. *Building theory in the sciences of the artificial*. in *4th International Conference on Design Science Research in Information Systems and Technology, DESRIST '09, May 7, 2009 - May 8, 2009*. 2009. Philadelphia, CA, United states: Association for Computing Machinery.
6. Matsokis, A. and D. Kiritsis, *An ontology-based approach for Product Lifecycle Management.* Computers in Industry, 2010. **61**(8): p. 787-797.
7. Panetto, H., M. Dassisti, and A. Tursi, *ONTO-PDM: Product-driven ONTOlogy for Product Data Management interoperability within manufacturing process environment.* Advanced Engineering Informatics, 2012. **26**(2): p. 334-348.

8. Dreibelbis, A., et al., *Enterprise Master Data Management: An SOA Approach to Managing Core Information.* 2008: IBM Press. 617.
9. Dyché, J. and E. Levy, *Customer data integration: reaching a single version of the truth*. 2006: Wiley. 294.
10. Gruber, T.R., *A translation approach to portable ontology specifications.* Knowledge Acquisition, 1993. **5**(Copyright 1993, IEE): p. 199-220.
11. Smith, B., *Ontology (science).* Nature Precedings, 2008.
12. Noy, N.F. and D.L. McGuinness (2001) *Ontology development 101: A guide to creating your first ontology*.
13. Spyns, P., R. Meersman, and M. Jarrar, *Data modelling versus ontology engineering.* ACM SIGMOD Record, 2002. **31**(4): p. 12-17.
14. Andersen, B., *Ontologies, data models, and ontology*, in *Interdisciplinary Ontology Conference (InterOntology08)*. 2008: Tokyo. Japan.
15. Lenzerini, M. *Data integration: A theoretical perspective*. 2002: ACM.
16. Ullman, J., *Information integration using logical views.* Database Theory—ICDT'97, 1997: p. 19-40.
17. Jinxin, S., et al. *An environment for multi-domain ontology development and knowledge acquisition.* in *Engineering and Deployment of Cooperative Information Systems. First International Conference, EDCIS 2002. Proceedings, 17-20 Sept. 2002*. 2002. Berlin, Germany: Springer-Verlag.
18. Nazri, M.N.M., S.A. Noah, and Z. Hamid. *Using lexical ontology for semi-automatic logical data warehouse design*. in *5th International Conference on Rough Set and Knowledge Technology, RSKT 2010, October 15, 2010 - October 17, 2010*. 2010. Beijing, China: Springer Verlag.
19. Villanueva Chavez, J. and X. Li. *Ontology based ETL process for creation of ontological data warehouse*. in *2011 8th International Conference on Electrical Engineering, Computing Science and Automatic Control, CCE 2011, October 26, 2011 - October 28, 2011*. 2011. Merida, Yucatan, Mexico: IEEE Computer Society.
20. Jiang, L., H. Cai, and B. Xu. *A domain ontology approach in the ETL process of data warehousing*. in *IEEE International Conference on E-Business Engineering, ICEBE 2010, November 10, 2010 - November 12, 2010*. 2010. Shanghai, China: IEEE Computer Society.
21. Ratté, S., W. Njomgue, and P.A. Ménard. *Highlighting document's structure*. in *World Academy of Science, Engineering and Technology* 2007.
22. Hay, D.C., *Data model patterns: conventions of thought*. 1996: Dorset House Pub.
23. Moustakas, C.E., *Phenomenological research methods*. 1994: Sage Publications, Inc.
24. Patton, M.Q., *Qualitative research and evaluation methods*. 2002: Sage.
25. Tesch, R., *Qualitative research: Analysis types and software tools*. 1990: Routledge.