

# A reference architecture for an enterprise knowledge infrastructure

Daniel Fitzpatrick, François Coallier, Sylvie Ratté

École de Technologie supérieure, Montréal, QC Canada  
Daniel.fitzpatrick@etsmtl.net

**Abstract.** With the emergence of social media, data available for market analytics has grown significantly, especially in the context of product lifecycle value analysis. Existing architecture frameworks do not support knowledge management, required to process massive amount of market data, or 'big data'. In order to perform product lifecycle value analysis, product managers need to access, in a seamless manner, data from several domains from systems within the organization and from external sources such as social media, government and industry sites, to name a few. The structured and integrated data must then be transformed into information, or contextualized data, and ultimately into actionable information or knowledge. To achieve this objective, this paper proposes an innovative approach, the Reference Architecture of an Enterprise Knowledge Infrastructure (RA-EKI) that provides a holistic approach to manage the complete knowledge lifecycle.

**Keywords.** Knowledge management, multi-domain ontology, data integration, product lifecycle management, enterprise architecture

## 1 Introduction

Product Lifecycle Management (PLM) reveals itself as a crucial business-enabling paradigm. It supports the business effort to ensure efficiency and consistency in the entire product lifecycle. Interoperability between PLM, Customer Relationship Management (CRM), Enterprise Resource Planning (ERP) and Manufacturing Execution Systems (MES) is also required to provide better process efficiency and customer satisfaction. These information systems need to be integrated in order to satisfy data quality requirements[1].

Within and between each of these systems, knowledge has become a strategic asset to the enterprise. It increases the innovative capability and core resource utilization, such as people and assembly lines[2]. It also favors knowledge reuse, helps improving workers efficiency and safety. Many obstacles still hinder the generalized adoption of cross-enterprise knowledge management in the PLM paradigm. Current PLM business process definition, product information model and information treatment approaches lack the required agnostic and holistic quality to support industry independent solution patterns[3]. The difficulty in extracting information stored in several types of formats such as image, video and audio constitutes an important hurdle to the enterprises' progress. Such a heterogeneous information environment conceals crucial

knowledge. Locating and retrieving such knowledge imposes costly processes and specialized human resources[4].

These many constraints may explain the great level of difficulty currently experienced by the organizations to adapt their PLM processes, and others, so that they can be more agile, flexible and adaptive to the rapidly changing market conditions[5, 6].

This research project intends to provide a Reference Architecture for an Enterprise Knowledge Infrastructure (RA-EKI), inspired from The Open Group Architecture Framework's TOGAF Reference Architecture - Information Integration Infrastructure (RA-III) to address these shortcomings. RA-EKI proposes a set of generic applications and a multi-domain ontology at its center. A description of the inductive research method used to elicit ontology patterns in this project can be found in [3]. RA-EKI's fundamental purpose is to deal with the issue of being data rich and knowledge poor, to be inundated with massive quantity of data but with limited capability to convert it into valuable knowledge.

The next section provides an insight on related projects that propose various ontology-based architecture approaches. Section 3 provides an overview of RA-EKI and a more detailed perspective on RA-EKI's generic applications. Section 4 covers a use case on the use of RA-EKI in support of product value analysis and section 5 concludes the paper.

## 2 Related work

As presented in [1], data quality constitutes one of the purposes of knowledge management models. Since enterprise knowledge is often unearthed from corporate data, data quality plays an important role in the presented models. Certain models extend their coverage of PLM and reach out to other process-centric paradigms such as Manufacturing Execution Systems (MES) [1, 7], Enterprise Resource Planning (ERP) [1] and Customer Relationship Management (CRM) [8], forming a synergy that would increase even more the potential of sustainable growth in revenue, profitability and market share.

Proposed methods use various semantic exchange mechanisms such as mediation web services[1, 9], Service-Oriented Architecture (SOA) semantic services[6, 9], intelligent agents[2, 10], Extraction, Transformation and Load (ETL)[4, 11] and Enterprise Application Integration (EAI) message broadcasting [5] to resolve the syntactic and semantic heterogeneity between the systems.

The reference models, such as SCOR [8] and VFDM [7] put forward knowledge management infrastructure functions that handle the transformation of raw data to refined knowledge. Table 1 synthesizes these knowledge management functions as described in the cited literature.

Table 1. Knowledge management functions

Functions	Description
Data Quality	A mediation approach to resolve syntactic and semantic heterogeneities between business processes [1].
Knowledge extraction	Applications can extract information from unstructured and semi-structured data, Information is grouped and ultimately forms unit of knowledge. The applica-

	tions annotate text files using W3C standards for XML, adding rich meta-information[2, 6, 9].
Knowledge structuring	Knowledge is structured after extraction from data into knowledge representation formats such as RDF triples[4, 6].
Knowledge storage	Archiving enterprise knowledge in a way to facilitate retrieval. Relational database and XML document technologies are used[2, 12].
Knowledge access	The treatment of queries to access stored knowledge. This can be accomplished through query processors or with the use of knowledge dashboards [4, 6, 12].
Intelligent agents	Applications dedicated to specific tasks such as supporting transactions between business entities[2, 10].
Process definition	An ontology-based design-time application that assists the enterprise to create or modify business processes with the use of generic process template and process modeling such as BPMN[5].
Data integration	Applications that map data from heterogeneous sources to a global schema and allowing to be restructured as contextualized data, or information, to be circulated and processed for knowledge extraction[3, 6, 13].
Natural Language Processing (NLP)	An problem solving method that uses stored knowledge to convert unstructured data in a form of text to structured data and information[6, 14].

The knowledge management models utilize various ontology approaches to provide a formal vocabulary to their semantic applications. Most models use widely known ontologies such as STEP, CPM, Onto-PDM and TOVE [1, 2, 7, 8]. With these ontologies and others built from within the projects, the models cover many concepts considered unrelated to PLM such as customer data. The pervasiveness of data subjects as explained in [15] highlights the changing nature of not only PLM but all of the other process-centric paradigms as well. The cited papers enunciated a significant set of concepts, such as: customer demographic data, orders, invoices, complaints, transactions, contracts, material orders, market information, new legislations affecting regulatory compliance, etc.

### 3 Description of RA-EKI

#### 3.1 Overview

RA-EKI, as illustrated in figure 1, comprises a set of processes, a collection of orthogonally linked ontologies and databases to process unstructured, semi-structured and structured data. Further processing converts data into information and ultimately

into knowledge. A more detail description of RA-EKI ontology structure can be found in [3]. Furthermore, since the inductive phase of the research is on-going, the final findings of the ontology structure including the multi-domain ontology will be the subject of future publications. Data can be extracted from within the organization, such as from structured databases, documents, emails, etc., and from external sources such as social networks, customer and government sites, etc. RA-EKI is described in greater depth in the next section. RA-EKI's applications may be implemented as agents or in other forms. Furthermore, although this paper covers RA-EKI with a warehouse style core database, this reference architecture may be implemented without a persistent central database, relying solely on mediated services or a hybrid configuration using both approaches.

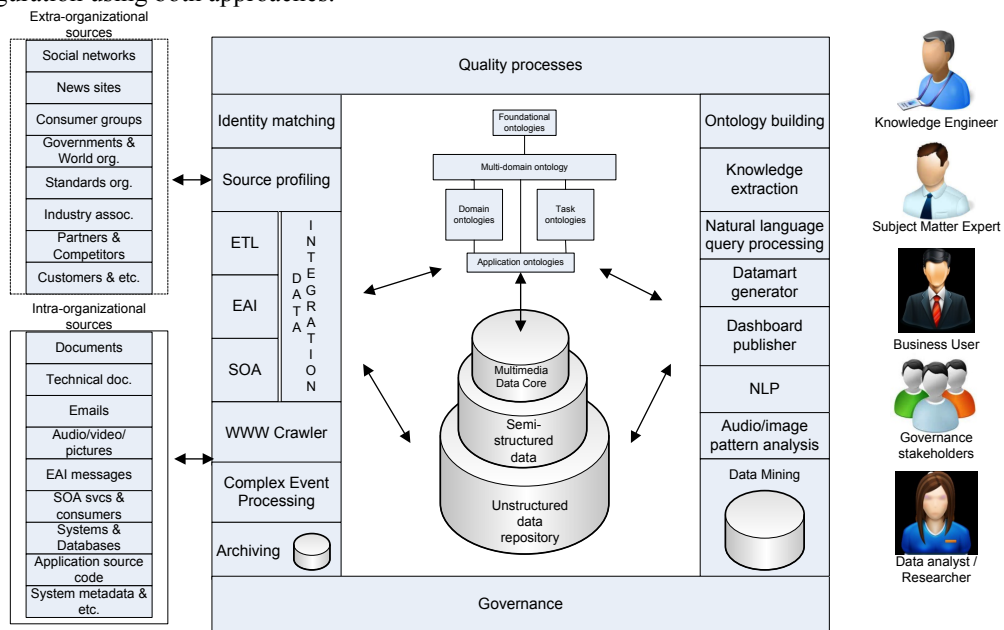


Fig. 1. Reference Architecture of an enterprise knowledge infrastructure (RA-EKI)

## 3.2 Description of RA-EKI's applications

### 3.2.1 Transformation of unstructured data and semi-structured to structured data

The data transformation applications, as an assembly line, progressively convert un-organized, massive amount of symbols into structured data, as illustrated in figure 2. There are three data structures used for storage: the unstructured data repository, the semi-structured database and the multimedia database core. The unstructured data repository contains raw text XML files pre-processed and summarily annotated by the crawler application. The semi-structured database contains refined XML text files processed by the NLP part 1 application with syntactic transformation. The core database, contains structured data and multimedia material. Along with the ontology structure, the core database is RA-EKI's central data structure. This database is struc-

tured semantically in line with the multi-domain ontology. Assertions are stored in RDF triples and incorporated in an object relational database such as the core database for performance reasons[16].

The crawler application, using a decay concept and genetic programming as proposed by [17], and following search goals stored in the core database, navigates through the web and detects internet material of interest. The crawler extracts HTML pages, per example, and pre-processes them by removing unnecessary items such promotional hyperlinks. The crawler annotates the raw XML file with meta-information such as author, location and time, thus providing a useful context for downstream applications[17].

The NLP part 1 application performs syntactic, morphological and lexical analysis on the raw data and provides a sentence structure to the text. Nouns, verbs and other sentence items are tagged and meta-information is added. The NLP part 2 application finalizes the text mining process by semantically converting the semi-structured text and by extracting concepts found in the ontology structure and by storing these concepts in the core database.[14, 18].

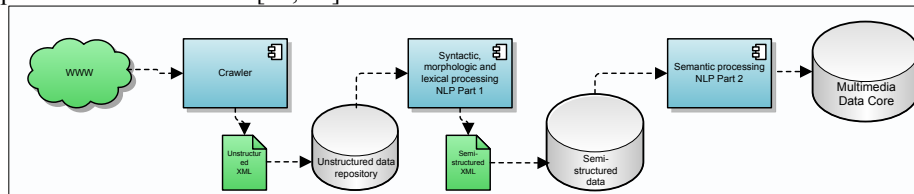


Fig. 2. Transformation of unstructured and semi-structured data into structured data

### 3.2.2 Source profiling, data integration and identity matching applications

The Source profiling application, as illustrated in figure 3, analyzes a new source database by generating volumetric and other statistics about databases. This information is then stored in the core database to be used especially for designing data integration mappings. It also provides an assessment on data quality issues.

The ontology-driven data integration applications convert in run time data from a heterogeneous source to a global schema that follows the same conceptualization as the multi-domain ontology and the core database. The three approaches used are EAI message broadcasting [5], SOA [6, 9, 10] and ETL [11, 19]. Although the underlying technologies are different, the fundamental mechanism remains the same. Mapping rules are also added to translate data from a heterogeneous source databases to the target global schema, represented in the ontology structure and core database. The mapping rules ensure that the source data are translated into the syntactic and semantic structures of RA-EKI's ontologies and core database. The same process is reversed when data is sent back to the source system either using SOA services or EAI message broadcasting. Current research issues can also be found notably in [13] about data integration in the context of knowledge management. Finally, the identity matching application is used in the attempt to associate objects, people and organizations per example, originating from different source databases but being potentially the same individual.

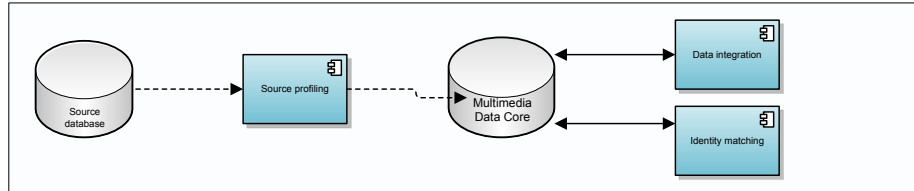


Fig. 3. Source profiling, data integration and identity matching applications

### 3.2.3 Transformation of structured data into information

As highlighted by [4] and illustrated in figure 4, one of the more efficient and popular mean to distribute information consists in pulling decontextualized data from RA-EKI's core database and then re-contextualizing the data in the form of a user-friendly dashboard. The dashboard publisher is ontology based in RA-EKI. The datamart generator uses corporate objectives, critical success factors, key performance indicators and other performance monitoring requirements that are stored in the ontology structure and the core database. A requirements ontology-driven datamart generator may reduce the time for delivering a datamart[20].

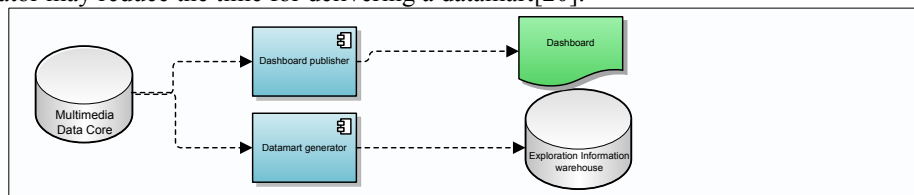


Fig. 4. Transformation of structured data into information

### 3.2.4 Transformation of information into knowledge

The data mining application, ironically, extracts knowledge not data. Through predictive and descriptive modeling, this design-time ontology-based application leverages the data scientist's skills and expertise to produce knowledge that can be encapsulated in a XML variant, a Predictive Model Markup Language (PMML). PMML is a standard adopted by most database technology manufacturers to facilitate the development and deployment of algorithms developed using data mining techniques. Research is currently conducted to perform ontology-driven knowledge extraction from PMML files[21]. RA-EKI knowledge extraction application stores extracted knowledge into the core database in transit to the downstream ontology building application. Various ontology-building methods are proposed in [4, 6, 16]. See figure 5 for the transformation of information into knowledge.

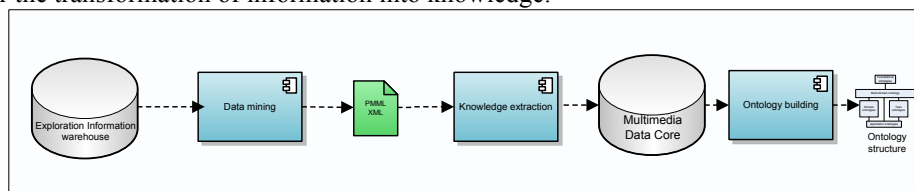


Fig. 5. Transformation of information into knowledge

### 3.2.5 Complex event processing

The Complex Event Processing (CEP) application, represented in figure 6, scans RA-EKI's databases on a continual basis. As these databases can be populated in right time, some data, such as those related to events, can be summararily analyzed by the CEP application. The ontology-driven CEP application attempts to detect any event of significance for the enterprise, guided by a set of prioritized subjects stored in the core database. This application may broadcast an alert through an EAI message, SOA service or using email services. The authors in [22] also describe an event ontology, an application and technology architectures dedicated to a CEP application. In RA-EKI, the CEP application shares with the other applications the same ontology structure and core database. The CEP application also stores the event in the core database.

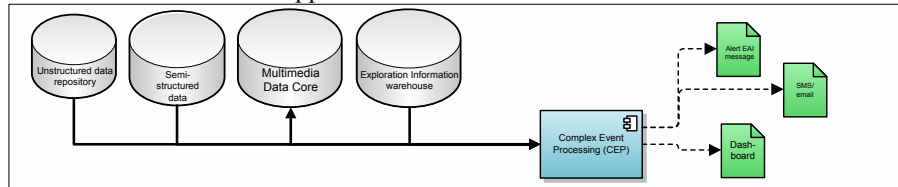


Fig. 6. Complex Event Processing (CEP)

### 3.2.6 Other applications

The other applications, although playing equally important roles, will be covered in a subsequent publication. The data quality application contributes significantly to the overall capacity of RA-EKI to meet its challenges. The Archiving application effectively stores historical versions of data, information and knowledge that can be used by other application. Challenging research issues associated with multimedia mining are being addressed notably by [23].

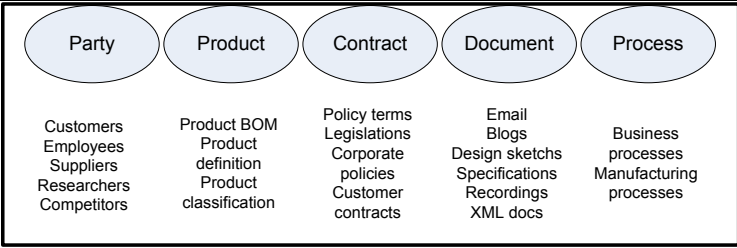
## 4 Uses case in product lifecycle value analysis

The following use case, outlined in table 2, illustrates the capacity of the RA-EKI model to provide assistance in performing product lifecycle value analysis by enhancing the process to transform massive amount of data into useful knowledge.[15] The use case formulates a competency question, describes the workflow used by ontology-driven components of the RA-EKI described, maps the concepts specific to the case against abstract concepts of RA-EKI's multi-domain ontology and core database, and identifies the sought benefits.

As indicated in[24], product design, that include product lifecycle value analysis, constitute one of the crucial stages in the entire product lifecycle. It often lacks procedural rigor and can progress through trial and error because of the complexity at hand. Furthermore, the massive amount of data affects product development, forcing the designers to spend a significant amount of time manipulating data in order to harvest crucial knowledge that may influence the product's commercial success.

Table 2. Use case in product lifecycle value analysis

Competency	"What are the factors that may influence the financial, customer and
------------	--

question:	environmental value of the new product currently under development?"										
RA-EKI workflow summary:	<ul style="list-style-type: none"> <li>• The crawler detects government, social media and competitor sites, notably, in all countries or regions considered for the market. The crawler will pre-process acquired text in removing unnecessary items, annotate the text and generating annotation tokens to be stored in the core database;</li> <li>• The NLP function further annotates unstructured text and extracts structured data from internal documents, competitor web sites, social media texts with relevant material for sentiment analysis, government regulatory documents, past similar product recall events, etc.;</li> <li>• Imaging and voice-recognition functions also annotate pictures (similar products...) and voice recording from customer service centers' logs[23];</li> <li>• The data mining support function allows the data analysts and researchers to develop targeting models to detect correlations and potential causalities that may influence the future product's market success;</li> <li>• The datamart generator, based on an goal oriented approach from [20], can produced in a semi-supervised mode, datamarts to provide a rich set of information to the product designers and to the product management governance stakeholders.</li> <li>• The knowledge extraction and integration functions pulls new semantic material from the core database and converts it into subsumed concepts, in a semi-supervised mode, in the ontology structure.[25]</li> </ul>										
Abstract concept mappings	 <table border="1" style="width: 100%; text-align: center;"> <thead> <tr> <th>Party</th> <th>Product</th> <th>Contract</th> <th>Document</th> <th>Process</th> </tr> </thead> <tbody> <tr> <td>Customers Employees Suppliers Researchers Competitors</td> <td>Product BOM Product definition Product classification</td> <td>Policy terms Legislations Corporate policies Customer contracts</td> <td>Email Blogs Design sketches Specifications Recordings XML docs</td> <td>Business processes Manufacturing processes</td> </tr> </tbody> </table>	Party	Product	Contract	Document	Process	Customers Employees Suppliers Researchers Competitors	Product BOM Product definition Product classification	Policy terms Legislations Corporate policies Customer contracts	Email Blogs Design sketches Specifications Recordings XML docs	Business processes Manufacturing processes
Party	Product	Contract	Document	Process							
Customers Employees Suppliers Researchers Competitors	Product BOM Product definition Product classification	Policy terms Legislations Corporate policies Customer contracts	Email Blogs Design sketches Specifications Recordings XML docs	Business processes Manufacturing processes							
Benefits sought:	<ol style="list-style-type: none"> <li>1. Provide the product designers and managers a 360-degree perspective on a new product to maximize its value and competitive position on the market.</li> <li>2. Mitigate the risks by leveraging on lessons learned.</li> </ol>										

## 5 Conclusion

A significant number of publications have addressed the challenge of developing integrative ontologies to assist in the various stages of the product lifecycle manage-



ment, and notably, in product lifecycle values analysis. This paper proposes a reference architecture, the RA-EKI, as the foundation of a comprehensive knowledge management that would allow a full cycle unstructured data to knowledge and know-how. The final results of this current project will constitute the foundation for future research involving the development of an implementation of the RA-EKI in the form of a prototype in laboratory settings and ultimately in a trial process to be performed the industry.

## References

1. Khedher, A., S. Henry, and A. Bouras. *Quality improvement of product data exchanged between engineering and production through the integration of dedicated information systems*. in *11th Biennial Conference On Engineering Systems Design And Analysis*. 2012. Nantes, France.
2. Marchetta, M., F. Mayer, and R. Forradellas, *A reference framework following a proactive approach for Product Lifecycle Management*. *Computers in Industry*, 2011. **62**(7): p. 672–683.
3. Fitzpatrick, D., F. Coallier, and S. Ratté, *A holistic approach for the architecture and design of an ontology-based data integration capability in product master data management*, in *9th International Conference on Product Lifecycle Management*, A.B. L.Rivest, B.Louhichi, Editor 2012, Springer: Montreal, QC, Canada. p. 559-568.
4. Mazumdar, S., et al. *A Knowledge Dashboard for Manufacturing Industries*. in *The Semantic Web: ESWC 2011 Workshops*. 2011. Heraklion, Greece.
5. Han, K. and J. Park, *Process-centered knowledge model and enterprise ontology for the development of knowledge management system*. *Expert Systems with Applications*, 2009. **36**(4): p. 7441–7447.
6. Raza, M. and R. Harrison, *INFORMATION MODELING AND KNOWLEDGE MANAGEMENT APPROACH TO RECONFIGURING MANUFACTURING ENTERPRISES*. *International Journal of Advanced Information Technology (IJAIT)*, 2011. **1**(3): p. 1-20.
7. TERKAJ, W., G. PEDRIELLI, and M. SACCO. *Virtual Factory Data Model*. in *Virtual and Mixed Reality - Systems and Applications*. 2011. Orlando, FL, USA.
8. Lu, Y., et al., *Ontology Alignment for Networked Enterprises Information Systems Interoperability in Supply Chain Environment*. *International Journal of Computer Integrated Manufacturing*, 2013. **26**(1-2): p. 140-151.
9. Yusuf, J., et al. *Extensive Overview of an Ontology-based Architecture for Accessing Multi-format Information for Disaster Management*. in *International Conference on Information Retrieval & Knowledge Management (CAMP), 2012*. 2012. Kuala Lumpur.
10. Wang, X., W. T., and G. Wang, *Service-oriented architecture for ontologies supporting multi-agent system negotiations in virtual enterprise*. *J Intell Manuf*, 2012. **23**: p. 1331–1349.
11. Thames, J., O. Eck, and D. Schaefer, *A Semantic Association Hardware Acceleration System for Integrated Product Data Management*. *Journal of*

- Computing and Information Science in Engineering, 2012. **12**(September 2012).
12. Liu, X. and G. Yang, *Research of Ontology-based coal Enterprise Knowledge Management Model system*. Applied Mechanics and Materials, 2011. **40-41**(625): p. 625-630.
  13. Calvanese, D., et al., *Conceptual modeling for data integration*. Conceptual Modeling: Foundations and Applications, 2009. **5600**: p. 173-197.
  14. Feldman, R. and J. Sanger, *The text mining handbook: Advanced approaches in analyzing unstructured data*2007: Cambridge University Press. 410.
  15. 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.
  16. Khouri, S. and L. Bellatreche. *DWOBS: Data Warehouse Design from Ontology-Based Sources*. in *16th International Conference on Database Systems for Advanced Applications 2011*. 2011. Hong Kong, China: Springer Berlin Heidelberg.
  17. Bazarganigilani, M., A. Syed, and S. Burki, *FOCUSED WEB CRAWLING USING DECAY CONCEPT AND GENETIC PROGRAMMING*. International Journal of Data Mining & Knowledge Management Process (IJDMP), 2011. **1**(1): p. 1-12.
  18. Wimalasuriya, D.C. and D. Dou, *Ontology-based information extraction: An introduction and a survey of current approaches*. Journal of Information Science, 2010. **36**(3): p. 306-323.
  19. Maynard, D., et al. *Natural language technology for information integration in business intelligence*. 2007. Springer.
  20. Ta'a, A. and M. Abdullah, *Goal-ontology approach for modeling and designing ETL processes*, in *World conference on information technology*2011, Elsevier. p. 942-948.
  21. Sottara, D., et al. *Enhancing a production rule engine with predictive models using PMML*. in *PMML '11 Proceedings of the 2011 workshop on Predictive markup language modeling*. 2011.
  22. Schaaf, M., et al., *Semantic Complex Event Processing*. Recent Researches in Applied Information Science, 2012: p. 38-43.
  23. Perner, P., *Learning an ontology for visual tasks*. MUSCLE 2011, 2012. **LNCS 7252**: p. 1-16.
  24. Chandrasegaran, S.K., et al., *The evolution, challenges, and future of knowledge representation in product design systems*. Computer-aided Design, 2012. **45**(2): p. 204-228.
  25. Bissay, A., et al., *Knowledge integration through a PLM approach*, in *15th international conference on new technologies and products in machine manufacturing technologies*2009: Sucuava, Romania.