Ontology-based Foundations for Data Integration

Virginija Uzdanaviciute
Department of Information Systems
Kaunas University of Technology
Studentu street 50-313a, Kaunas
virginija.uzdanaviciute@stud.ktu.lt

Rimantas Butleris
Department of Information Systems
Kaunas University of Technology
Studentu street 50-313a, Kaunas
rimantas.butleris@ktu.lt

Abstract—**The integration of data is one of the most complicated tasks which need to be addressed by IT researchers. Despite the critical importance, the current approaches to semantic interoperability of heterogeneous databases have not been sufficiently effective. We apply an idea of ontology as the foundation for data integration. In the paper the most common data integration methods and their essential features are discussed; the major problems of integration tasks are highlighted. Requirements for implementing data integration tasks are defined for the model, the architecture, the content, and the representation. In order to specify data source semantics, a meta-model is created, which is used to describe the concepts of source and relations. The following architectural ontology-based models are analyzed by selected criteria: a global, a multiple and a hybrid. The model of data integration process is then built upon the hybrid architecture model.**

Keywords-system interoperability; data integration methods; ontology-based data model; integration approaches; semantic integration

I. INTRODUCTION

One of the biggest current research challenges in the human–computer interaction and information retrieval is to provide users with intuitive access to the growing amount of data present in different database management systems (DBMS). Databases (DB) are designed and filled over time by different people, but they represent the same or related areas. These data express real world facts, attributes and interrelationships. The origin of data is their history which covers source collection, processing technologies, executors, etc. All businesses collect data using diverse information systems (IS). Typical ISs are designed to enable users to perform business operations and not to exchange data with other ISs. The interoperability of systems is, unfortunately, not relevant from the rate of interest standpoint. This has been a critical issue within the database community for the past two decades [27].

The interoperability of a system is seen as a consequence of technical, semantic, organizational, legal and political tools. It empowers transfer and usage of data in other information resources by following types:

- Organizational. It specifies the regulation of resource interaction.
- Technical. It describes the compatibility of IT tools, establishment and usage of open interfaces, standards and protocols in order to ensure effective data exchange.
- Semantic. This characteristic ensures that data from one IS are understood and interpreted in the same way in other systems.

Systems must be able to exchange data. Data exchange between ISs is determined by reciprocal agreements which are different in each case: web-based services, open standards, specifications [30]. Direct data integration is impossible if data is processed by applied IS logic [2]. This process can be performed in real time in source system changes occur, fixed time intervals automatically or manually, using popular methods: Extract, Transform and Load (ETL), data replication, federation, event-based integration, web-based technologies and open standards [18]. The aforementioned methods have essential disadvantages in the context of heterogeneous DBMS [15][16]: the problems of automatic update are neither considered nor solved, the same data is stored in several sources. Besides, there is no possibility to get data or information messages on databases using direct access interaction. The researchers of distributed heterogeneous databases have applied ontologies to support semantic interoperability: to integrate data sources developed using different vocabularies and to see data from a different perspective [22].

The process model presented in this paper describes the foundations for ontology-based data integration system. The described data integration task automatically performs data extraction and integration from both structured and semi-structured data sources. In addition, semantic IS interaction type is analyzed; searching for solution of ensuring unified understanding of the same data which is in heterogeneous data source systems, clearly describing semantics of commonly used data. The proposed process model integrates and reuses data using ontologies by relevant criteria.

The structure of the paper is as follows. Section 2 introduces the foundations of data integration. Section 3 relates the different concepts of the task. ER and ontology-based data models are compared; essential features of the models are highlighted. In addition, we also derive the requirements for an ontology modeling language. An ontology-based data source (OBDS) model is proposed for the development of systems. The evaluation of different aspects of the architectural models based on ontologies

ISBN: 978-1-61208-160-1

(global, multiple and hybrid) is also presented. Furthermore, an overview of the most popular data manipulation methods is in order. Section 4 describes the process of data integration based on ontology. Section 5 presents the conclusions of our research and thoughts on future work.

II. Data Integration

Normally, the organizational data resides in multiple data sources. For typical business intelligence (BI) data integration projects [18], the design and development of data integration processes involve collecting facts for the integration, analyzing data structures and their descriptions [4]. However, it is inappropriate to focus on the management of data requirements [2] only: it is very important to discern that integration is more than data. It also covers:

- Data sources: what data from where has to be integrated?
- Business rules (BR): which BRs have to be evaluated for data processing and keeping in data sources?
- Transformations: which transformations have to be done in order to avoid structural and semantic conflicts?

The integration of data – data management in the way that they would be unambiguously identified in IS and it is possible to transfer, transform, load and use them in other IS or source without changing program code [18]. Currently prevailing IS infrastructures are characterized as complex, distributed and heterogenic environments [1]. For this reason, data and integration of various programs are unceasing challenges for system developers, as well as providing accurate information which is necessary in today‘s competitive markets [18]. Ontology-guided data integration makes the process more efficient – reducing the cost, maintenance and risk of the project [18].

In order to consolidate and integrate data, we have to know which data is required, where it is and which method will ensure this process. First of all, we have to identify data, determine suitable ISs and DBMSs. Moving on, we have to specify the relationships, place, and accessibility of the data. Then we can create logical schemes, i.e. perform reverse engineering and use data dictionaries (if they exist). The next (very important) step is to evaluate DB integrity (triggers, relationships), data structure (types and lengths of fields), the time and frequency of their updates. A further step is to describe meta-data. The third step is to create the meta-data model, which describes not only data structure, but also their reciprocity. The model can be formally defined by Entity-Relationship (ER) diagram [10], which allows visual specification of data structure and relationships [17]. Alternatively we can use Data Flow Diagrams (DFD), which identify the manipulated data and visualize the data flows among processes, repositories and external environment entities [17]. Unified Modeling Language (UML) class diagrams, which describe data structures using interrelated classes, are also an option. Consider even the flexible RDF/OWL data model [15][16][23][24]. After carrying out these steps, data integration can be performed.

III. The Task Of Data Integration

In this section, we describe the requirements for the task of data integration. As sketched in Figure 1, the integrated data takes into consideration the four following aspects designed in the composition model: a model, architecture, the content and the representation. Each aspect is described below:

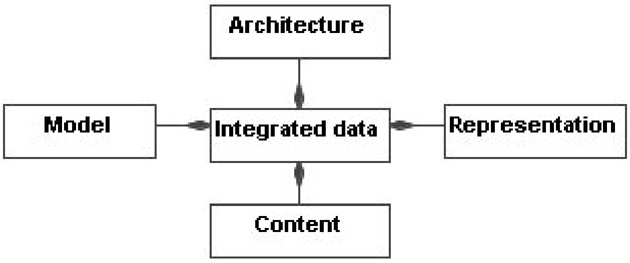

Figure 1. Integrated data.

The data model of a particular data source definition. The model must allow extension with new data, retrieval of important, highest-quality, semantically meaningful information, and the re-use of data.

The architecture, which is the core of integration and has to perform the high level autonomy to data sources. The architecture must provide semantic interoperability for the systems. It must allow the system architect to manage the development of data collections when data sources have different formats (text files, XML schemas, relational models) and to re-aggregate the application.

A neutral representation abstracts specific syntax; therefore, all the structured and semi-structured data sources first need to be expressed in a neutral format. A set of content data elements must be able to receive high-quality, semantically meaningful information. The content is heavily affected by the semantic conflict types to be resolved.

The content, i.e., the meaning of the information that is interchanged, must be understood. Data and relations have to be visualized and represented in the best, most appropriate way. It follows that each representation must bind a single expression to a single meaning using the Resource Description Framework (RDF) language.

A. *Requirements for Modeling Language*

The criteria for ontologies as modeling language of a domain modeling and formal semantic specification of data sources are chosen [5][11][22]. Minimalism is of paramount importance – only the necessary concepts must be represented in ontology. Expression – all the required concepts of a domain must be represented. Clarity – ontology must be unambiguous, easy to learn and remember; the meaning of diagrams or text expressions must be intuitively obvious, the language concepts and notations should be understandable by non-technical domain experts. Semantic stability – possibility to remain in changes of a domain. Semantic suitability – only conceptually suitable entities of a domain are modeled. Verifiability – domain experts must be able to verify if model corresponds to domain. Abstract mechanisms – possibility to hide unwanted details. Formality – model is unambiguous and executable.

B. Ontology-based Data Model

The term “ontology” refers to a machine-readable representation of knowledge, particularly for automated inference. Ontology is a data model which consists of these parts: classes, properties and relationships between them [19]. The power of ontologies lies in the ability to represent relationships between the classes. The main benefit of using the ontology-based model is its runtime interpretation [21]. One of the major advantages of the ontology model is an assumption of open-world [12]. The reason for the popularity – clearly interpreted dissemination of knowledge between people and applications [7]. Moreover, ontology supports the integration task as it describes the exact content and semantics of these data sources more explicitly [1].

Gardner [18] proposes to focus explicitly on the representation of knowledge rather than just its management. He ensures that if a highly descriptive semantic representation of the available knowledge could be built, it could be reused to power a variety of business applications without the need for repeated integration exercises. Furthermore, the new knowledge gathered from different sources can build upon the current knowledge because all of it exists in a semantically consistent system. Thus we conclude that knowledge is the foundation of all successful decisions.

Semantic technologies in data integration solutions allow representing relationships of data definition area, relating data using data sets and identifying relationships for new associations. The reuse of legacy data provides the following opportunities: store and represent any types of data, easily modify the model, expand it with new data, evaluate changes.

An overview of the requirements which will be automatically satisfied by an ontology-based process is given in Figure 2.

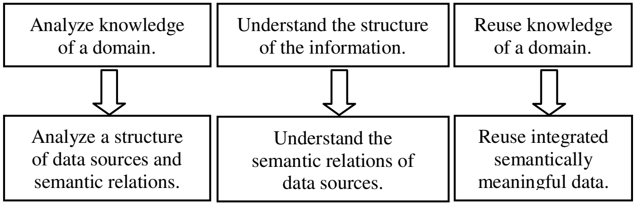
Analyze knowledge of a domain.
Understand the structure of the information.
Reuse knowledge of a domain.
Analyze a structure of data sources and semantic relations.
Understand the semantic relations of data sources.
Reuse integrated semantically meaningful data.

Figure 2. Transformations of ontologies features to data integration systems requirements.

Besides, ontology-based model is used to solve semantic and syntax conflicts of the heterogeneous data sources [9][22]:

- Schematic: when data is stored in heterogeneous DBMS. Such conflicts are caused by different designers, different area interpretation and usage of different data models.
- Semantic: when different class / attribute names (issues of synonyms and homonyms), different output formats (coding, data formats), different meanings are used. Conflicts arise in attributes when semantically equivalent (having the same meaning) attributes have different domains in several schemes.

C. Ontology-based data source integration architectures

In this section, we describe the three main ontology-based architectures: global, multiple, and hybrid in Figure 3 [4][7][14][20].

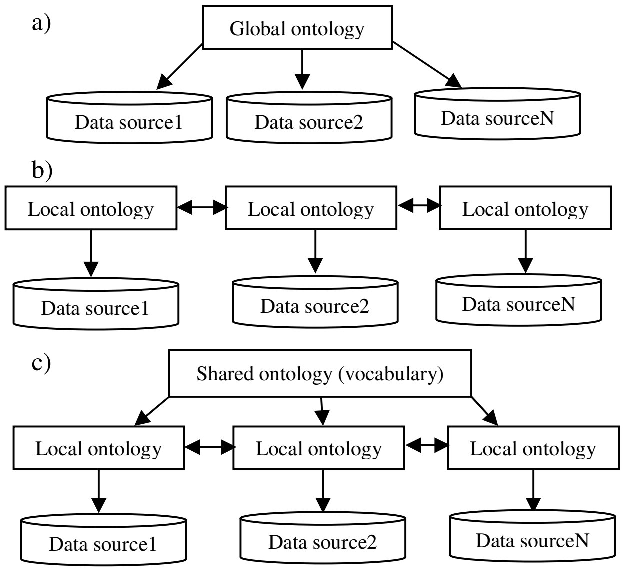

Figure 3. The methods: a) global, b) multiple, c) hybrid.

The method of data integration generally depends on which integration architecture the developer is most familiar with, what is known about heterogeneity of the data sources, etc. In models [14] shown in Figure 3 ontologies and data sources are represented as classes and semantic relationships between data source content and ontology as links, and links between local ontologies as mappings.

The global ontology method in a) of Figure 3 uses a single ontology. The method has a single main stage: building global ontology by domain expert, who knows the semantics for all data sources. The global ontology can also be a combination of several specialized ontologies. The global ontology describes data from heterogeneous data sources; query is executed via the main ontology. All data sources are related to the global ontology. This method can be applied to integration solutions where all data sources to be integrated provide the same view on a domain.

The multiple ontology method in b) of Figure 3 uses local ontologies and mapping rules between them. Each data source is described by its own ontology. The mapping rules can be modified according to the dynamic change of data source. The method has two main stages: building local ontologies and defining mappings. The local ontologies describe data from heterogeneous data sources; integrated query is executed via the local ontologies. The essential feature of this method is that the ontologies for individual data sources could be developed or changed without respect to other semantic relations, data sources or their ontologies.

The hybrid ontology method in c) of Figure 3 uses a vocabulary of a domain to represent a shared ontology, a local ontology and mapping rules between them. The specification of a vocabulary includes definitions of classes, relations, functions, and other objects [8]. The mapping rules can be modified according to the dynamic change in data source. The method has three main stages: building the shared vocabulary, building local ontologies and defining the mappings. Similarly to the multiple ontology method, the semantics of each source is described by its own ontology. However, in order to make the source ontologies comparable to each other, they are built upon one global shared vocabulary.

The advantage of the hybrid method is that new data sources can easily be added without the need to modify the mappings or the shared vocabulary. Therefore, the hybrid ontology architecture gives more autonomy to data sources [28]. The use of shared vocabulary makes the source ontologies comparable and avoids the disadvantages of single or multiple ontology methods. Table 1 presents the different ontology architecture methods resulting from this analysis.

TABLE I. BENEFITS AND DRAWBACKS OF THE ONTOLOGY-BASED INTEGRATION METHODS

Criterions	Ontology-based architectures		
	Global	Multiple	Hybrid
Evaluation of semantic heterogeneity	Useful for systems which have the same view on a domain.	Useful for systems, which have the same view on a domain.	Useful for systems, which have the different view on a domain.
Appending new data sources	Some modification is necessary in the global ontology.	Supports an opportunity to append the new data source with some adaption in other ontologies.	New data sources can easily be added without the need of modification.
Elimination of data sources	Some modification is necessary in the global ontology.	Supports an opportunity to remove the data source with some adaption in other ontologies (need to remove relation between ontologies).	Data sources can easily be removed without the need of modification.
Comparison of multiple ontologies	Impossible.	Difficult, because of lack of a common vocabulary.	Simple, because ontologies use a global shared vocabulary.

D. *An Ontology of Data Source*

Ontology-based data integration is an effective method to cope with the heterogeneous data. This solution is based on the idea of decoupling information semantics from the data sources. Moreover, ontologies support dynamic domains better. For this reason, it is necessary to analyze data source elements: data, schema, schema elements and content, values, entities and attributes, query result classes. It is known that ontology-based search system gives the user more meaningful query results than the normal search system [13], which queries data with syntactic parameters. The query result is based on data retrieval methods [23][24][25][26]. Figure 4 gives an overview of data source meta-model.

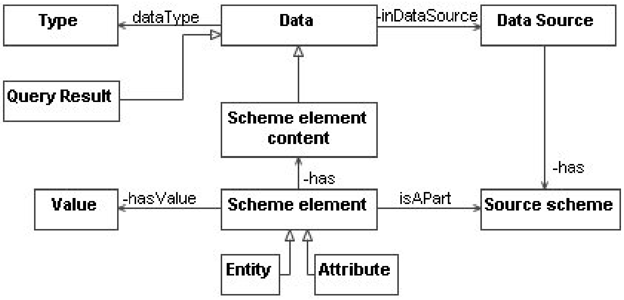

Figure 4. Meta-model of data source.

IV. CONCEPTUAL ONTOLOGY-BASED DATA INTEGRATION PROCESS

In this section, we describe the process for constructing a suitable system to semantically integrate the data from heterogeneous data sources and ensure the interoperability of it. The process model is based on hybrid method of architecture which has a shared ontology – vocabulary. Each step of the process model has its own ontology. A shared ontology is established using local ontology for each data source and a method of ontology alignment to match them. The usage of this method allow us to enhanced usability, the possibility to model mechanisms that are closer to the way we understand the real world. According to Ram *et al.* [22], ensuring interoperability of systems and knowledge-based information sharing is one of the key aspects of successful implementation.

We propose to evaluate business rules (BRs) and constraints, which ensure data integrity and correctness in the processes of data update, processing and integration. It is known that a BR is a logical statement of what to do (what actions to take) in different situations [6]. BRs can be classified by the actions in the ISs as shown in [6]. All characteristics of the data including BRs, and constraints we extract automatically from DBMS, or describe manually in a vocabulary.

Moreover, we propose to detect and solve conflicts at both data and schema levels. Han *et al.* [29] affirms that different treatment of the data structure and semantics plays a major role in IS. In this context, the scientist tries to achieve semantic coherence by eliminating semantic conflicts with a common ontology. Semantic Conflict Resolution Ontology (SCROL) provides a dynamic mechanism of comparing and manipulating contextual knowledge of each data source, which is useful in achieving semantic interoperability among heterogeneous data sources. A more detailed description of the conflict classification and the method for automatic detection and resolution of various semantic conflicts in heterogeneous databases using SCROL are found in [22].

Our process model in Figure 5 is similar to the approach proposed by Skoutas *et al.* [3], which consists of five aspects. One of the chief drawbacks of that approach is inability to resolve all the semantic conflicts detections and solving processes. In addition, it does not evaluate BRs and constraints, which play the main part in the integration of data. Compared with Chang *et al.* [30], the proposed process of ontology-based data integration and analysis solves only data format conflicts and duplication.

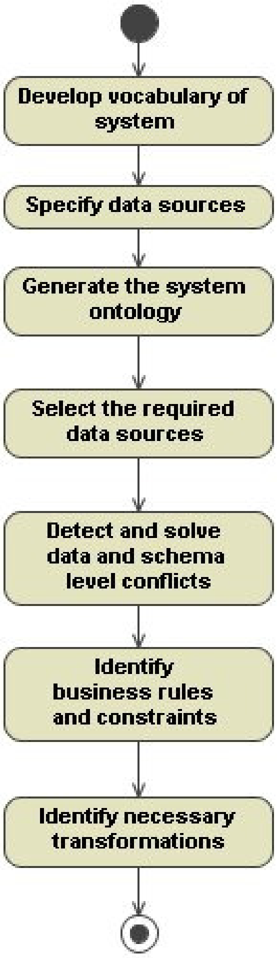
Develop vocabulary of system
Specify data sources
Generate the system ontology
Select the required data sources
Detect and solve data and schema level conflicts
Identify business rules and constraints
Identify necessary transformations

Figure 5. Process model.

The activities of our process model are specified as follows:

- *Develop vocabulary of system*. The vocabulary consists of the concepts of the domain, the attributes characterizing each concept, the different representation formats, and values for each attribute (feature values). The concepts of the domain are represented by classes, while the relationships between concepts, as well as the attributes of the concepts are represented by properties. The values of features consist of concept features and representation formats. Also, data sources and system requirements are described in the vocabulary of the system.
- *Annotate the data sources*. This one involves the mapping of attributes. Attribute mappings relate attributes to features. The types of mappings are: 'one-one', 'one-', 'many-many', and 'none'. The attribute specification consists of representation format, range of values (min, max), cardinality, referenced relations, function and attributes used for aggregation. Representation formats belong to the concept features.
- *Generate the system ontology*. System ontology is used to model the domain and to formally specify the semantics of data in the data sources. The system ontology consists of: a set of classes corresponding to the specified domain concepts, a set of properties corresponding to specified features of the concepts of the domain, and a set of classes representing the data sources.
- *Select the required data sources*. In this step we need to identify required data sources for integration.
- *Detect and solve data and schema level conflicts*. This stage is useful to determine the semantic match of data sources. It is necessary to decouple meaningful data and its semantics from the data sources with conflicting constraints.
- *Identify BRs and constraints*. It is the processing of complex BRs and constraints, including complex data transformation logic for output from integration of heterogeneous sources. Conceptual BRs and constraints provide the rationale for the correct data values in the data sources, warns of errors in the updating, processing and integrating of data. BRs ensure that the integrated data records have the same semantics. Besides, they prevent data integration between data sources with conflicting constraints and guard data correctness and integrity.
- *Identify required data conceptual transformations*. Transformation rules describe how the required data is extracted from the sources, combined and re-used in other ISs according to predefined BSs and constraints. They ensure the consolidation of data quality and detail level requirements.

V. CONCLUSION AND FUTURE WORK

We applied the idea of process based on ontology for data integration. The problem of data integration is data exchange, defined as the problem of transforming data structured under one schema into data structured under another schema.

The proposed hybrid data integration process is based on the use of ontology that explicitly captures knowledge about different types of data sources. While database schemas are generally regarded as static, the ontology schemas are typically assumed to be highly dynamic and evolving objects. The key feature of the data integration process model is: it evaluates BRs, constraints and transformations for identification of semantic conflict and solving processes at both data and schema levels. The advantage of our method is that it is based on hybrid architecture method. It relies on the following elements: system vocabulary and local ontology per each heterogeneous data source. In order to integrate data from heterogeneous data sources using the hybrid method, the relations between central and local ontologies, and the relations between local ontology and the corresponding data sources should be built up.

The prospective work is to describe ontology-based data integration methodology using ontologies of the data source and resolution of semantic conflicts, BRs, and transformations.

REFERENCES

[1] M. Gagnon, “Ontology-Based Integration of Data Sources”, Information Fusion, 10th International Conference, ISBN: 978-0-662-45804-3, pp. 1-8, 2007.

[2] A. Taa and A. S. Abdullah, “Norwawi M. Rameps: A Goal-Ontology Approach to Analyse the Requirements for Data Warehouse Systems”, Wseas Transactions On Information Science And Applications, ISSN: 1790-0832, iss. 2, vol. 7, pp. 295-309, 2010.

[3] D. Skoutas and A. Simitsis, “Designing ETL Processes Using Semantic Web Technologies”, In DOLAP’06: Proceedings of the 9th ACM international workshop on Data warehousing and OLAP, ACM Press, pp. 67–74, 2006.

[4] Skoutas D. and A. Simitsis, “Ontology-Based Conceptual Design of ETL Processes for Both Structured and Semi-Structured Data”, International Journal on Semantic Web and Information Systems, Special Issue on Semantic Web and Data Warehousing, vol. 3, no. 4, pp. 1-24, 2007.

[5] T. R. Gruber, “Toward Principles for the Design of Ontologies Used for Knowledge Sharing”, Technical Report KSL 93-04, Knowledge Systems Laboratory, Stanford Univ., 1993.

[6] L. Tutkute and R. Butleris, “Template based business rules modelling from UML to executive code”, Proceedings of BIR’2009 : the 8th International Conference on Perspectives in Business Informatics Research, Kristianstad University College, 01-02 October 2009 / J. Aidemark, S. Carlsson, B. Cronquist (Eds.), Kristianstad : Kristianstad Academic Press, 2009, ISBN 9789163355097, pp. 7-14, 2009.

[7] D. Fensel, “Ontologies: Silver Bullet for Knowledge Management and Electronic Commerce”, Springer Verlag, pp. 138, 2001.

[8] T. R. Gruber, “A Translation Approach to Portable Ontology Specifications”, Knowledge Acquisition, vol. 5, no. 2, pp. 199-220, 1993.

[9] S. Spaccapietra and C. Parent, “View Integration: A step forward in solving structural conflicts”, IEEE Transactions on Data and Knowledge Engineering, vol. 6, no. 2, pp. 258-274, 1994.

[10] P. McBrien and A. Poulovassilis, “A Formal Framework for ER Schema Transformation”, International Conference on Conceptual Modeling, the Entity Relationship Approach, Springer, pp. 408-421, 1997.

[11] E. Vysniauskas and L. Nemuraite, “Transforming Ontology Representation from OWL to Relational Database”, Information technology and control, vol. 35, no. 3A, pp. 333-343, 2006.

[12] J. Bock, S. Grimm, J. Henß and J. Kleb, “A Database Backend for OWL”, In: Proceedings of the 5th International Workshop on OWL: Experiences and Directions OWLED 2009, vol. 529, pp. 1–8, 2009.

[13] Oracle Database Semantic Technologies Overview, White Paper, 2007.

[14] L. Zhang, Y. Ma and G. Wang, “An Extended Hybrid Ontology Approach to Data Integration”, 2nd International Conference on Biomedical Engineering and Informatics, pp.1-4, 2009.

[15] L. Dong and H. Linpeng, “A Framework For Ontology-based Data Integration”, International Conference On Internet Computing in Science and Engineering ICICSE 2008, pp. 207-214, 2008.

[16] T. Berners-Lee, “The Semantic Web”, Sci. Am. 284, pp. 34-43, 2001.

[17] P. P. Chen, “The entity-relationship model – toward a unified view of data”, ACM Transactions on Database Systems (TODS), vol. 1, no. 1, pp. 9-36, 1976.

[18] S. P. Gardner, “Ontologies and semantic data integration”, DDT, vol. 10, no. 14, pp. 1001-1007, 2005.

[19] N. F. Noy and D. L. McGuinness, “Ontology Development 101: A Guide to Creating Your First Ontology”, Technical Report, SMI-2001-0880, Stanford Medical Informatics, Stanford, 2001.

[20] H. Wache, T. Vogele, U. Visser, H. Stuckenschmidt, G. Schuster, H. Neumann and S. Hubner, “Ontology-Based Integration of Information - A Survey of Existing Approaches”, In Proceedings of the IJCAI-01 Workshop on Ontologies and Information Sharing, pp. 108-117, 2001.

[21] D. Calvanese and G. Giacomo, “Ontology-Based Data Integration”, Tutorial at the Semantic Days 2009 Conference Stavanger, Norway, 2009.

[22] S. Ram and J. Park, “Semantic Conflict Resolution Ontology (SCROL): An Ontology for Detecting and Resolving Data and Schema-Level Semantic Conflict", IEEE Transactions on Knowledge and Data Engineering, vol., 16, no. 2, pp. 189-202, 2004.

[23] OWL Web Ontology Language Reference. Available at: http://www.w3.org/TR/owl-ref.

[24] Resource Description Framework (RDF). Available at: http://www.w3.org/TR/rdf-concepts.

[25] SPARQL Query Language for RDF. Available at: http://www.w3.org/TR/rdf-sparql-query.

[26] XML Query. Available at: http://www.w3.org/XML/Query.

[27] A. P. Sheth, “Semantic Issues in Multidatabase Systems”, SIGMOD Record, vol. 40, no. 4, pp. 5-9, 1991.

[28] L. Bellatreche, G. Pierra, “OntoAPI: An Ontology-based Data Integration Approach by an a Priori Articulation of Ontologies”, 18th International Workshop on Database and Expert Systems Applications, IEEE Computer Society, pp. 799- 803, 2007.

[29] L. Han and L. Qing-zhong, “Ontology Based Resolution of Semantic Conflicts in Information Integration”, Wuhan University Journal of Natural Sciences, vol. 9, no. 5, pp. 606-610, 2004.

[30] X Chang. and J. Terpenny, “Ontology-based data integration and decision support for product-Design”, Robotics and Computer-Integrated Manufacturing 25, pp. 863–870, 2009.

[31] J. Heflin and J. Hendler, “Semantic Interoperability on the Web”, In Proceedings of Extreme Markup Languages 2000, Graphic Communications Association, pp. 111-120, 2000.