REACHING SEMANTIC INTEROPERABILITY

Ivan Mutis[1], Raja R.A. Issa[2], Randy Chow[3], Su-Shing Chen[4], and Chu Huang[5]

ABSTRACT

There is a need in specific domains to develop strong interaction between agents using emerging technologies to facilitate communication and collaborative use of information. A successful implementation of these technologies enables interoperability based on real time and transparent information transfers to make decisions or to work on concurrent basis.

New approaches using standards, which allow stakeholders' computer systems to exchange information between actors, have been used for the implementation of this interoperation. However, the main obstacle in enabling interoperability of the standards developed is the matching process involved in relating data representations. This process is complex but it is a 'glue' of the interoperability process and is very critical in making the exchange of information possible.

This research is intended to develop new approaches using data representations structured in Ontologies. The research uses an example from the construction industry domain. The approach intends to facilitate the mapping process between ontologies coping with the semantic heterogeneity problem for the construction industry domain. The research will determine the main characteristics of the data representations needed to perform matching operations, based on the assumptions of the autonomy of the information sources, in order to formulate alternative solutions and expedite the information exchange process. The proposed approach uses plug-in matching prototypes from the **protégé** ontology editor. The result of this approach is to produce a framework that contains an analysis of the typical data representation of the actors in the construction domain.

KEY WORDS

Mapping, Interoperation, Ontology, Information Integration, Schema Matching, Protégé

INTRODUCTION

Agents have to comprehend what to exchange, and how and when the action has to occur. 'What' refers to the content of information that generally one actor queries from another, 'how' refers to the process that is going to be used to exchange the information, and 'when' refers to the instants when a transaction is suitable to be executed.

[1] Ph.D. student, Rinker School of Building Construction, University of Florida, Gainesville, FL 32611; Ph: (352) 273 1179; imutis@ufl.edu.

[2] Rinker Professor, Rinker School of Building Construction, University of Florida, Gainesville, FL 32611; Ph: (352) 273 1152; raymond-issa@ufl.edu.

[3] Professor. Computer and Information Science and Engineering (CISE), University of Florida, Gainesville, FL 32611; Ph: (352) 392 1487; chow@cise.ufl.edu.

[4] Professor. Computer and Information Science and Engineering (CISE), University of Florida, Gainesville, FL 32611; Ph: (352) 392 1220; suchen@cise.ufl.edu.

[5] Ph.D. Student, Computer and Information Science and Engineering (CISE), University of Florida, Gainesville, FL 32611; Ph: (352) 392 1220; shuang@cise.ufl.edu.

One approach that tackles the aforementioned constituents of the exchange process is that of executing agreements, in partnering or alliance models, between actors. Recent explorations in this area of work have been aimed at the development of ontology products (Jeusfeld and Moor 2001). In this approach, essentially the actors have to enable mechanisms to understand that the elements of the information in their data representations that have the same meaning for them. In addition, they also have to implement in their system additional modules to support the transactions and type or format of information. Hence, the actors have to know what operations they are going to execute in the exchange process. However, this type of specific implementation is extremely expensive and is not generic, i.e.; that is, it will be valid only for a particular group of stakeholders that achieve the agreement between each other. Additionally, firms lack the level of technology sophistication to implement these types of solutions (Veeramani et al. 2002).

Other approaches are the integration and merging of information. Both approaches have drawbacks. Integration forces the creation of a single source of information leading to the loss of information source autonomy, a property that stakeholders generally prefer to maintain. The merging of information is not a scalable process and is costly. Furthermore, both approaches are difficult to implement considering disparate, multiple source and large amount of information.

In general terms, the basis of the problem of interoperability that any approach needs to tackle, without any rigorous conceptual analysis, are:

- The different methods to represent information (Partridge 2002)[6];
- The different levels of specification of data representations;
- The various levels of systematization or sophistication of the actors' systems.

In order to obtain an effective exchange of information process, this research will explore the mapping approach. A framework for interoperation using a Plug in for Protégé will be also developed to obtain better specifications on the contents of the information sources, which in this case are the ontology sources.

COMPLEXITY OF MATCHING FROM HETEROGENEUS SOURCES

The technique used to determine correspondences between data representation is the matching operation. These operations tackle integration from different sources – applications, data warehouses, web-oriented data integration, e-commerce, etc. As an operation, matching works independently from the implementations of system architecture. Matching is a schema manipulation operation that takes two schemas as input and returns a mapping that identifies corresponding elements in the two schemas (Madhavan et al. 2001).

Matching works on an operation called *mapping*, which transforms requests and results between levels in schemas, which represent metadata from autonomous sources. In simple terms, *semantic* matching is the process of semantically relating elements from different schemas that have the same meaning.

[6] A good discussion about this disagreement problem is found in Partridge, C. (2002). "The role of ontology in integrating semantically heterogeneous databases." *05/02*, National Research Council, Institute of Systems Theory and Biomedical Engineering (LADSEB-CNR), Padova - Italy.

Manually constructed, these mappings are both labor-intensive and error prone, and have proven to be a major problem in deploying data integration systems (Doan et al. 2000). Figure 1 shows that despite the identification of data representation, their mapping is extremely cumbersome: schemas may have structure and naming differences; may have similar but not identical content; may dissimilarly express data representation; may use similar syntax but have different semantics, and so on. Thus, human intervention is required to execute the mapping or matching between elements of the data structures. Hence, data representation matching with a high degree of certainty cannot be executed automatically under this scenario.

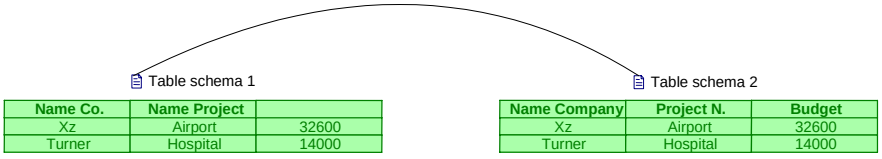

Figure 1: Mapping Complexity: Different Syntax, Same Semantics

The most important issues to consider in this complexity are those related to finding the meaning or semantics involved between the elements that are being mapped. One factor is that the elements of representations are frequently subjective. For instance, one source could call one of its elements "automobile" but the other source could call the correspondent element of its source "car". Then, what is the semantic difference between the labels of this transportation means? Another important factor is that *data representations in any structure are the products of designers or creators of that data*. These designs in the structures are more representative to the authors than other users that intend to interpret the source of information. Conspicuously, mapping is a complex and pervasive process that has to be critically and independently analyzed.

Although there are good tools and approaches to help find correspondences with the syntax of elements, the most challenging is to find semantic relationships or semantic mappings. Formally, the difficulty encountered in trying to integrate information is due to its heterogeneous nature. Thus, for two data representations, the following differences may be found (Garcia-Molina et al. 2002):

- **Data type differences** (syntax), the data might be represented in each source by strings or other different values;
- **Value differences** (syntax), the same constant might be represented by different constants at different sources -
- **Semantic differences**, terms may be given distinct interpretations at different sources, and,
- **Missing values**, a source might not record information of a type that all or most of the other sources provide.

SEMANTIC AND SYNTACTIC SIMILARITIES

The starting point about how the matching process works with data representations is the components of information such as syntax, semantics, and methods of construction. Labels or syntax terms contain a meaning we implicitly give to them. Inclusively, authors

discriminate the meaning of labels as lexical (linguistic) and the relations in a structure as structural knowledge (Bouquet et al. 2003; Giunchiglia and Shvaiko 2003a).

Consider the conceptual graph structures shown in Figure 2. The results of having a simple conceptual graph structure, mapping two syntax terms, and performing a syntactic matching indicates that the meaning to the correspondent labels depends only on their labels; However, if we perform a semantic matching, the meaning on the labels depends not only on the current meaning but also on the relations with other concepts in conceptual graph structure. Observe that *relations* that are components of the structure among the labels; they are represented by the links shown in the conceptual graph.

As shown in Figure 2, the structure of each conceptual graph is the same, but the meaning is different. The syntax of some components is the same; the labels are the same and the meaning may be the same. The syntax on the label of the component "Frame" indicates that the meaning is the same. However, taking into account the *relations* among the components in the structure, the intended meaning of the "Frame" may be different in graphs. For instance, using natural language and linking the components with their *relations*, the first conceptual graph indicates a "sliding frame window" and the second a "fixed frame window".

The example in Figure 2 shows the syntax and semantic mapping complexity in a simple conceptual graph data representation. This problem exists in all types of schemas. As a result from this example, it is clear that in any case *the syntax terms or labels conduct the semantic mapping process*.

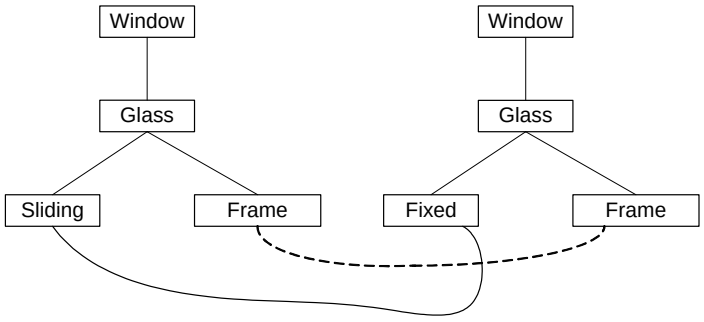

Figure 2. Structure Represented in Graphs

THE MATCHING PROBLEMS

The most common cases in the mapping operation processes are:

- **The One to One Element Relationship Problem:** Two different actors may agree that the meaning of one element or concept, which exists in each one of the data representations, is the same. Thus, they agree that one element of an actor's data representation corresponds to only one element of another actor's data representation. However, most of the cases with one element in a data representation do not fully correspond to only one element of the other data representation, because they could match to a different element. Therefore, the problem is to find the ones most semantically similar to each other. The word correspondence in this context means the association of two elements, the mapping or the semantic matching of these elements. Only partial solutions to find correspondences among elements, which have more than

one semantic similarity, are possible. They consist of using available or additional information like data instances, integrity constraints or user knowledge (Rahm 2001). Consequently, obtaining matches of two schemas or data representations using this additional information is likely. This type of matching is called 1:1 mapping, otherwise it is called complex mapping or complex matching.

- **The Complex Matching Problem:** Recall that the purpose of mapping is to find the best correspondences between two elements of data representations, utilizing available information such as schema information, user knowledge, and data instances. In the mapping process, there are not only 1:1 mappings from two data representations; a significant portion of a semantic mapping that involves two elements or more of a data representation also exists. This semantic mapping is called *complex mapping*. This mapping involves a set of rules or operations to find the "right" or "best" correspondence. Complex mappings involve two or more elements from one or more data representations, and applying axiomatic operations to match them to the elements of another data representation. Again the problem is how the user finds the semantic mapping that best matches elements of data representations using auxiliary information. The problem is enhanced due to the need to find axiomatic rules constructed over the elements of one representation. Complex matching subsumes a 1:1 mapping problem. The availability of multiple options requires that the most semantically "correct" association, among the paths to execute a superior map, should be evaluated.

- **Granularity Problem:** Two actors in an information exchange process may relatively understand the concept of an object or activity. One actor may understand that the object or activity, or concept, holds certain characteristic while the other actor comprehends the same concept at a different level of specifications. This dilemma is labeled as a granularity problem. In other words, how can an actor semantically map two data representations when they hold different concepts? Granularity means the relative size, scale, and level of specification that characterizes concepts, such as objects or activities (Giunchiglia and Shvaiko 2003a; Rahm and Bernstein 2001). Then the question becomes about how granularity affects the matching process of two data representations. There are two levels to evaluate granularity: *the element level* and *the structural level*. The *element level* performs mappings between individual concepts, and the *structural level* maps the semantics of two concepts, where each one holds other concepts

MATCHING APPROACHES

Due to the complexity of the operations, mapping the process manually is slow and expensive. Although there are some solutions or prototypes in the Computer Science field that help diminish human intervention in matching operations, manual procedures are unavoidable to get accurate results. The high cost of manual operations has spurred numerous solutions, which can roughly be classified into two groups:

- ***Development Standards***. All representations must conform to a common vocabulary or set of rules. However, these developments cannot be a general solution to reconciling representations. The reasons often cited are that a domain generates multiple competing standards, which defeats the purpose of having a standard in the

first place. Furthermore, organizations need to expand standards because extensions from different organization are generally incompatible with each other. Finally, developing standards demands engaging in a time-consuming consensus building process.

- ***Automatic approaches***. Although automatic matching approaches are being developed, user interventions are needed to perform matches using the existing approaches. For this reason, these approaches are still under the label of semi-automatic prototypes. If these approaches are not domain specific, then, any solution can be completed with additional tools or applications, e.g. an application for incomplete data representations or application interfaces (Doan 2002). The reason is that the approaches employ single matching criterion and are developed for certain types of information. As a consequence, the prototypes have less accuracy and have limited applicability.

Consequently, this research is focused on the ***semantic mapping*** process between two data representations. This direction complements the purpose of the standards. Hence, an ontological analysis is proposed, due to its flexible nature, to represent relationships, axiomatic rules, and different type of data representations. The work in this context is to define semantics, to aid in the analysis of the structure of poorly specified ontologies.

ONTOLOGY ANALYSIS FOR REACHING SEMANTIC MAPPING

Ontology is a set of concepts describing a specific domain following a subclass hierarchy, assigning properties, and defining relations between concepts. They have a set of identifiable classes and relationships. Ontology resembles taxonomy since it describes concepts, or knowledge representation language. Thus, ontology has a taxonomy-based sub-class hierarchy. Data representations are subject to be represented in ontological representations in a specific domain, such as the construction industry domain. If a specific domain uses taxonomies to represent its data, then its ontology could resemble that taxonomy. In summary, ontology resembles taxonomy as it describes concepts, or knowledge of representation language and ontology has a taxonomy-based sub-class hierarchy.

Assume a user wants to perform a query of information from two-construction businesses. The data representations and the ontologies are shown in Figures 3 and 4. Remember that the nodes represent concepts that have levels of specializations derived from their parent's node. Additionally, as shown in Figures 3 and 4, there are different types of relations between the nodes, for example *Isa(CPVC, Plastic Pipe Fittings).*

Suppose the user queries information about availability and costs of specific items from the sources, say pipes for internal water distribution for buildings. Therefore, he needs to map a road between the two sources and the query. Specifically, he needs to perform a semantic mapping in the most similar nodes of the two ontologies that contain information from the construction businesses.

Again, the problem lies in how the user can semantically map a node or concept of one data representation to another node of the other data representation. Moreover, as one can intuitively notice in Figures 3 and 4, how can the user semantically map more complex matching when one node is semantically similar to a concatenation of two or more nodes?

In addition, with a careful reading of the previous problems, we notice that they resemble one to one, complex and granularity problems. For instance, consider one to one

semantic match in one of the levels. With the aid of auxiliary information such an "expert knowledge", node 7 from Figure 3 *(Steel Pipe, Black Weld, Screwed)* with node 8 from Figure 4 *(Metal, Pipes & Fittings)* match. Observe, however, they semantically match, although they fully syntactically mismatch.

Remember the user queries specific information from the two sources; he/she needs to map specific instances from one source to the other source. Then, as an illustration of the problems of the matching process, suppose the user semantically maps node 4 *(Plumbing Piping)* in Figure 3 to node 6 *(Pipes and Tubes)* in Figure 4. But this mapping does not resolve the query. The user should follow the relations of specialization of the concepts. It is akin to following down a hierarchy of taxonomy. A taxonomy is a central component of an ontology (Noy and McGuinness 2001). Assume the expert's information ushers that a *Polymer Pipe* between *1"* and *1"1/2 diameter* is suitable for internal water distribution, like *PVC (Polyvinyl Chloride)* pipe. Hence, the user matches nodes 9 *(PVC 1¼")* from Figure 3 to nodes 10 *(PVC)* and node 13 *(1¼" Diameter)* from Figure 4. Observe that node 10 is concatenated with node 13 to perform the match. Node 10 is a more general concept than node 13 or, vice versa, node 13 is more specific than node 10. This is a complex type of match that includes a joint of two nodes and a specialization of one node.

Following the hierarchy to map the nodes in Figures 3 and 4, we notice the relations between nodes of two representations could match in a more general or specific level, or they could overlap or could mismatch. These types of intuitively semantic relations have what is called a *level of similarity* (Doan 2002; Giunchiglia and Shvaiko 2003b). Thus we could say that the concept of node 4 *(Plumbing Piping)* of Figure 3 is similar to node 4 *(Building Service Piping)* of Figure 4.

Additionally, it is important to remark that the relation between nodes such as *Isa(HVAC, Mechanical)* are also possible maps to other nodes of the ontology. These mappings between different types of elements make the process more complex.

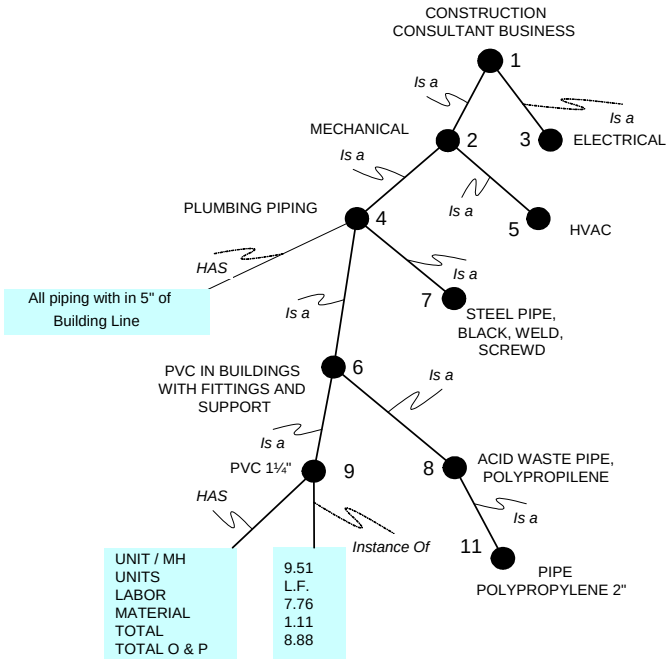

Figure 3. Construction Company Ontology

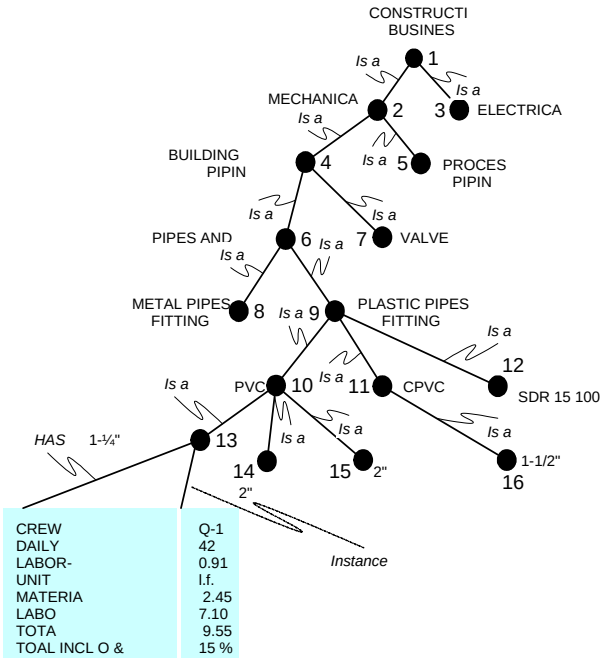

Figure 4. Construction Business Ontology

CONTEXT TO FIND ONTOLOGY ARTICULATIONS

Ontology representations facilitate the interoperation of concepts. For this purpose a context to 'improve' the semantics of the ontology is proposed, thus enlarging its relations. This will result in a sufficient representational adequacy of the ontology to establish a better definition of the concepts.

In order to be able to find relationships with the terms, concepts, and differences with real world objects of concepts the so-called Ogden's triangle (Ogden and Richards 1989); as shown in Figure 5 will be used. Eckholm (1996) extended this schema by establishing differences between *reference* and *representation*. This extended schema will be used to find fundamentals of the concepts of the domain and to achieve a unified high-level representation. As shown in Figure 5, the concepts are referenced to some common class models as synthetic references using standard product data models.

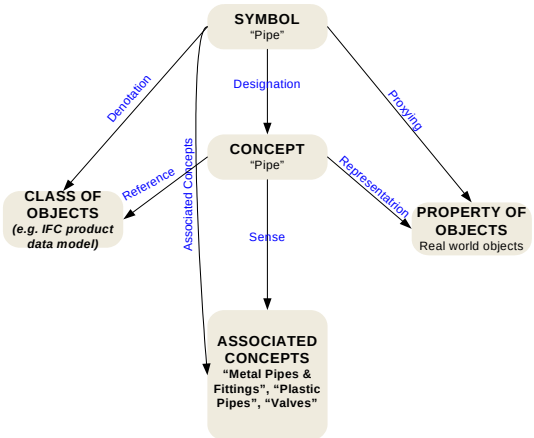

Figure 5. Fundamental Semantics.

The proposed system architecture as shown in Figure 6 has the PROTEGE ontology definition module, one translator, a plug in PROTÉGÉ mapping tool, and application for customizing the results. Each actor has a defined ontology as data representation, according to its view of the concepts. The data representation uses a translator that will transform the ontology file into a valid input for the semi-automatic tool. The knowledge base is a tool that supports operations of the semi-automatic engine and it works as a library that contains additional information or information from experts. The result then is a mapping file in a text-based format. This is used to customize construction applications where the mapping process between two actors is needed.

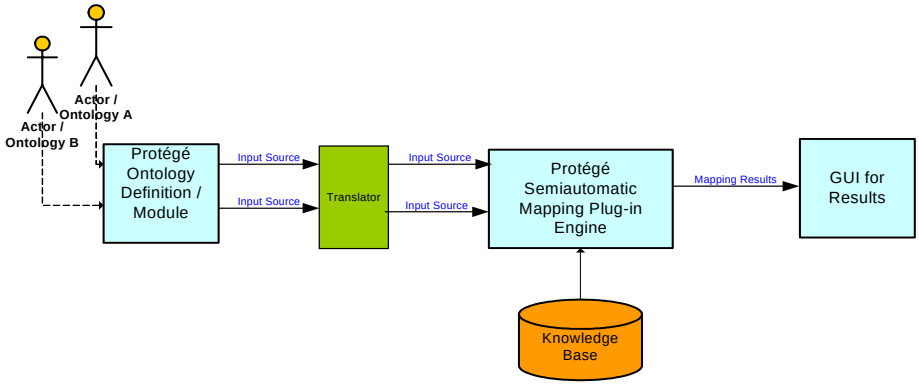

Figure 6. Prototype Architecture.

CONCLUSIONS

The nature of information creates a barrier in its flow, making the exchange and sharing of information intricate and complex even with the use of common standard developed for *interoperability* operations. This study proposes a more comprehensive definition of the information components, or of the concepts of the data representations through an analysis of ontologies. In these ontological representations, the focus is on redefining the semantics of the concepts. Based on these semantic definitions a better-structured data representation will be created thus improving the specifications and the comprehensiveness of the relations. Additionally, the use of semi-automatic tools based on PROTEGE to perform the mappings will open up the possibility of finding ontology articulations. Although this type of research contributes to having a better picture of the interoperability problem using mapping approaches, additional efforts to perform efficient mapping in the industry have to be undertaken through the use ontology representation paradigms.

LIST OF REFERENCES

Bouquet, P., Serafini, L., Zanobini, S., and Benerecetti, M. "An algorithm for semantic coordination." *Semantic Integration*, Sanibel Island, Florida, USA.

Doan, A. (2002). "Learning to Map between Structured Representations of Data," Doctor of Philosophy, University of Washington, Seattle, Washington.

Doan, A., Domingos, P., and Levy, A. (2000). "Learning source descriptions for data integration." Department of Computer Science. University of Washington. Seattle, WA 98195, 1-6.

Garcia-Molina, H., Ullman, J. D., and Widom, J. (2002). *Database systems: the complete book*, Prentice-Hall, New Jersey.

Giunchiglia, F., and Shvaiko, P. "Semantic Matching." *Semantic Integration*, Sanibel Island, Florida, USA.

Giunchiglia, F., and Shvaiko, P. (2003b). "Semantic Matching." *DIT-03-013*, University Of Trento Department Of Information And Communication Technology.

Jeusfeld, M. A., and Moor, A. d. "Concept Integration Precedes Enterprise Integration." *34th Annual Hawaii International Conference on System Sciences (HICSS-34)*, Island of Maui, Hawaii, 10.

Madhavan, J., Berntein, P. A., and Rahm, E. (2001). "Generic Schema Matching with Cupid." Roma, Italy, 10.

Noy, N. F., and McGuinness, D. L. (2001). "Ontology Development 101: A Guide to Creating Your First Ontology'." Stanford Knowledge Systems Laboratory Technical

Stanford Medical Informatics Technical, Stanford, Ca.

Ogden, C. K., and Richards, I. A. (1989). *The Meaning of Meaning: A Study of the Influence of Language upon Thought and of the Science of Symbolism*, H. B. Javanovich, translator, Harvest Books, Orlando, Florida.

Partridge, C. (2002). "The role of ontology in integrating semantically heterogeneous databases." *05/02*, National Research Council, Institute of Systems Theory and Biomedical Engineering(LADSEB-CNR), Padova - Italy.

Rahm, E., and Bernstein, P. A. (2001). "A survey of approaches to automatic schema matching." *The VLDB Journal*, 10, 334-350.

Veeramani, R., Russell, J. S., Chan, C., Cusick, N., Mahle, M. M., and Roo, B. V. (2002). "State of Practice of E-Commerce Application in the Construction Industry." *180-11*, Construction Industry Institute, The University of Texas at Austin, Austin, TX.