UDC: 004.8, 005.94

A METHODOLOGY FOR ONTOLOGICAL KNOWLEDGE CAPTURE FROM DATABASES

F. García-Sánchez, J.T. Fernández-Breis, R. Martínez-Bejar

Department of Computing and Systems, Faculty of Computer Science,
University of Murcia, Facultad de Informática, Campus de Espinardo,
30100 Espinardo (Murcia), Spain, +34 968 39 8107, frgarcia@um.es,
+34 968 36 4613, jfernand@um.es,
+34 968 36 4634, rodrigo@um.es

The successful emergence of the Information and Communication Technologies (ICT) has contributed to the efficiency improvement in a number of economic sectors. However, some strategic economic sectors, such as construction, have not been targeted enough yet. Construction-related ICT solutions lack mechanisms to permit the effective integration of the whole supply chain. Semantic Web can tackle these issues. This paper presents a methodology for acquiring knowledge from construction-related databases. A domain ontology has been developed that contains the relevant concepts regarding supply management in the construction domain. The methodology basically consists of mapping the database content onto the ontology and a further this one's population by applying a set of mapping rules.

Успешное появление информационно-коммуникационных технологий (ИКТ) внесло свой вклад в повышение эффективности многих секторов экономики. Однако, некоторые стратегические экономические сектора, такие, как строительство, не были все же достаточно исследованы. Связанные со строительством решения ИКТ испытывают недостаток в механизмах, позволяющих разрешать проблемы эффективной интеграции полной цепочки поставки. Семантическая Сеть может заняться этими проблемами. Эта статья представляет методологию, позволяющую извлекать знание из баз данных, связанных со строительством . Была разработана онтология домена, содержащая релевантные понятия, касающиеся управления поставками в домене строительства. Методология в основном состоит из отображения содержания базы данных на онтологию и дальнейшего ее заполнения , применяя набор правил отображения .

Introduction

The construction sector requires the effective management of large volume of data, information about building site processes, provider data, and so on. Most of the cost in this sector focuses on supplies and manpower, whose efficient management would produce both money and time savings. Besides, properties prices would decrease in the medium term. Each building is different and it should be treated as such. Customers should be able to get properties in accordance to their specific preferences. However, there are different elements in each building, contrasting ideas between customers and builders. In addition to this, the information regarding a particular building site can be so wide that a classification mechanism becomes necessary in order to take advantage of it. Furthermore, builders and customers must deal with supplies from different suppliers, and each supplier has its own information system and way of structuring the information concerning its supplies. This situation also happens with suppliers of the same type of product. Hence, mechanisms for harmonizing this heterogeneity should be pursued in order to facilitate the labour of both builders and customers.

Additionally, part of the knowledge and experience acquired from the development of a new building is currently kept by the personnel who have worked in that building site. Only in case the very same personnel works in another building site this knowledge and experience could be reused, otherwise it would be lost. Thus, if the knowledge acquired by the personnel is stored and matched against the potential customers' knowledge, a better supply schedule could be performed either automatically or semi-automatically (i.e. supervised). Furthermore, if that information is shareable and reusable by different members of the staff in the same company, the management and control of the supply material could be done more efficiently within the company.

Traditionally, adaptors and exchange formats have been applied to promote interoperability between information systems, without significant success yet. To face this problem, alternative approaches have been proposed that make use of semantic technologies to facilitate integration and interoperability [1]. An advantage of using semantic approaches is the fact that they do not require to replace current integration technologies, databases and applications. Moreover, they add a new layer that takes advantage of the already existing infrastructure [2]. Semantic Web technologies [3] are useful for our purpose. Amongst the core Semantic Web technologies, ontologies are basic to promote semantic interoperability between independent and heterogeneous systems such as the World Wide Web. Modelling the information by means of ontologies leads to an environment where builders can be aware of all the information regarding a building site at any time. Ontologies permit shareable, reusable, and machine-readable modeling of information, so most of the tasks regarding that information management information can be automated. Thereby, the organization increases its processes efficiency and has all the relevant elements needed to make an optimal control of the supplies integrated.

In practical settings, ontologies are more and more used in information management due to the advantages they have. On the one hand, ontologies are reusable, that is, a same ontology can be reused in different applications, either individually or in combination with other ontologies. On the other hand, ontologies are shareable, that is, their knowledge allows for being shared by a particular community.

The main goal of the approach presented here is to allow for the creation, integration and management of supply information in the construction domain. Such approach is based on Knowledge Management and Semantic Web technologies. Basically, the information that is obtained from databases and heterogeneous sources is modelled by means of knowledge management systems. After that, different tools can be ideated in order to allow an optimal access and management to the relevant supply information in the construction domain.

The rest of the paper is organized as follows. Section 2 offers an overview on the technologies applied in this approach. In Section 3, the methodology for knowledge acquisition from databases in construction is described and an example is depicted in Section 4 in order to show how the methodology works. Finally, in Section 5 some conclusions and further work plans are presented.

1. Methodological Foundations

The approach presented here aims at putting together different technologies related to the Semantic Web, such as ontologies (due to their adequacy in solving integration and interoperability problems) and Schema Integration. This section presents a brief overview of these technologies and how the proposed solution benefits from their use.

1.1 Ontologies. One of the most widespread definitions is ontology is Tom Gruber's [4]: "An ontology is an explicit specification of a conceptualization". An ontology represents a common, shareable and reusable view of a particular application domain. Moreover, ontologies are used to give meaning to information structures that are exchanged by information systems. An ontology is essentially a formal and structure information conceptual model. An ontology is here seen as a semantic model containing concepts, their properties, interconceptual relations, and axioms related to the previous elements. In this work, one of the objectives is to organize and model information about the construction domain into ontologies. For it, all taxonomies (e.g. there are different classes of bricks, tiles, slabs, etc), partonomies (e.g. a brick is part of a wall, the kitchen is part of a house), and chronologies (e.g. you have to paint after every wall and the roof have been pargeted) can be defined. In this work, the ontological content is expressed by using the Ontology Web Language (OWL), which is the W3C recommendation for exchange of ontologies on the Web (Web Ontology Working Group, 2004).

1.2 Ontologies for Integration and Interoperability. Nowadays, databases contain a huge amount of data. However, the integration of different databases in order to provide a uniform access to them has not been fully provided yet. Data integration requires real-time transformations of the information that flows between systems. The transformations must take into account the semantic differences between the applications. The most important factors that make it difficult to integrate and obtain interoperability between systems are the semantic and structural heterogeneity, as well as the different meaning assigned to information by different systems. In this context, ontologies facilitate the human understanding of the information besides the information-based access and the information integration from very different information systems. Ontologies allow for differentiating among resources, and this is especially useful when there are resources with redundant data. Thus, they help to fully understand the meaning and context of information. This is important for our objective of achieving semantic interoperability among different resources. Ontologies have been already used for the integration of databases in order to provide interoperability among different information systems in different domains such as biology and medicine. Examples can be found in [5], where ontologies were used to promote integration and interoperability between information systems for three medical communities by combining data with HL7 and terminologies such as UMLS, MEDCIN and SMOMED, or [6] where they are used to promote interoperability among electronic healthcare records information models.

1.3 Schema Integration. Another approach related to this work is that of Schema Integration. It is defined as "the process of generating one or more integrated schemas from existing schemas" [7]. The goal of the schema integration methods is to allow applications to transparently view and query data from multiple data sources as if they were one uniform data source. The idea in schema integration is to use mapping rules to handle the structural differences between the different data sources.

In [8] the authors present a four-phase integration process. The first phase is called 'Preintegration', moment in which database administrators and designers select schemas, decide the order of integration and set an integration policy or preference. During the second phase the schemas are analyzed and compared to detect possible schema and data conflicts. In the third phase, the requirements and conflicts for the merging are identified, requiring a close interaction between designers and users. Finally, the actual schema combination is performed. The blackboard architecture has been also use for schema integration [9]. Using the blackboard architecture, multiple knowledge agents were able to cooperate in spite of accessing different disparate knowledge sources.

2. A Methodology for the Semantic Management of Construction Supplies

The methodology presented in this work consists of four steps. The final aim of the methodology is to acquire knowledge (in the form of an instantiated ontology) from relational databases in the construction sector. It focuses on the construction domain but could be generalized to any other domain that shares some of the properties of the construction sector such as its stability (i.e. new elements that modify the domain model do not usually appear over time).

- Step 1: Build a general domain ontology scheme. During this step, the ontology scheme is developed. For this task, in-depth knowledge of the domain is required. The construction of this ontology is critical as it has an influence on the rest of the process. Thus, an expert is responsible for manually doing this step. Once the scheme is complete it can be extended according to changes in the domain.

Repeat for each database…

- Step 2: Get the map between the ontology and a relational database. In order to be able to instantiate the different concepts of the ontology resulting from carrying out Step 1 with the contents of the databases, a mapping between them is needed. The mapping process is manually done. Each element in the database scheme (i.e. all the database columns) is to be matched, on a one-by-one basis, against an element of the ontology (i.e., a concept, a property or an interconceptual relation). This step may also give rise to refinements in the ontology.
- Step 3: Populate the ontology. The third step of the methodology concerns the process of populating the ontology. Now, using the mapping rules previously obtained, the information contained in the database is mapped onto its correspondent element of the ontology. During this process, which is automatically performed, new instances of ontological elements are created along with the association between the attributes (i.e. concept properties) and their values.

End Repeat.

- Step 4: Ontology evolution. The general ontology schema should evolve according to the changes produced in the world (requirements, source databases, etc). Therefore, a continuous checking process needs to be performed to assure the consistency.

At this moment in time, most of the steps that comprise the methodology have to be performed manually. However, in the near future it is expected that research on different application fields will lead to a fully automated process.

3. Example

In this section an application of the methodology under question is illustrated by means of an example. This example use represents a typical problem in trying to integrate the access to two different databases, so that user queries are uniform. In particular, two different relational databases referring to the same real world elements but with different data schemes are integrated by means of a common general ontology model.

The first step, as indicated in the methodology, is to get a general domain ontology scheme. It has to be done once for each different application domain. As we are dealing with the construction industry, the ontology scheme to develop should show the concepts and specific features of this domain. Several different ontology models could be constructed depending on the modeller's point of view or the concrete properties of the problem to be solved. In Fig. 1, an extract of the domain ontology is depicted. The OWL file of the whole ontology can be found at http://klt.inf.um.es/ontologies/ConstructionSuppliesManagement.owl.

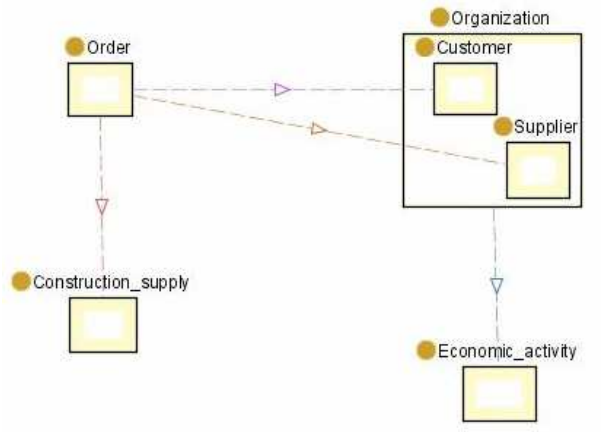

Fig. 1. Domain ontology

Four main concepts can be highlighted: the *economic activity* which a company is engaged in, the *organization* that needs or can provide a supply, the *order* a company makes to another related to a particular construction material, and the actual *construction supply* ordered. Apart from both customers and suppliers, two taxonomies have been designed, the economic activities taxonomy, and the building and construction taxonomy. In this figure, different interconceptual relations are depicted. For example, each organization performs its activity in a particular sector or economic activity, or the consumer organization (i.e. customer) makes an order of a supply to a provider organization (i.e. supplier). The economic activities taxonomy is based on the "International Standard Industrial Classification of all Economic Activities, Revision 3.1" that is maintained by the United Nation Statistics Division, Statistical Classification Section (http://unstats.un.org/unsd/cr/family2.asp?Cl=17).

The construction supplies taxonomy is based on a taxonomy previously developed by WAND Inc. (www.wandinc.com) and found through the Taxonomy Warehouse (www.taxonomywarehouse.com).

For the next step, according to the methodology described, the databases need to be identified. In this example, we are dealing with simple data schemes as shown in the following figures (Fig. 2 and Fig. 3). The schemes represent two different models related to construction supply transactions. Scheme A can be the one used by a supplier company while scheme B can be the one used by the builder company that needs supplies. Although they are different, they aim at modelling the same domain elements.

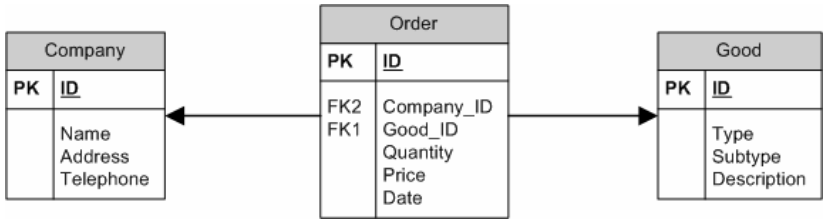

Fig. 2. Database Scheme A

Fig. 3. Database Scheme B

The entity-relation diagram in Fig. 2 presents three different entities (Company, Order, and Good) and two relations. The database itself would consist of three tables, each including a primary key (ID for every table) and table *Order* holding two foreign keys (Company_ID and Good_ID). The entity-relation diagram in Fig. 3 presents a similar scheme to that in Fig. 2 but with a number of differences. On the one hand, supplies are modelled by using different relational entities. Model A consists of three entities, one for companies, one for goods and one for orders, whereas Model B represents supplies orders with one single entity, *SupplyOrder*.

Besides, several attributes have changed in different ways. For example, *Quantity* in scheme A is termed *Amount* in scheme B. On the other hand, attribute Address in scheme A has been split up into four different attributes in scheme B, namely *Street*, *Number*, *City*, and *Country*. These differences would make it harder for companies to communicate with business partners. Once the databases have been identified, the manual mapping process starts. As it was explained above, this step is to be done manually because no satisfactory automatic solution to this problem has been obtained yet. For each database scheme, a separate mapping should be defined. Moreover, for each element in the database scheme a unique relation with an element of the ontology has to be identified. The mappings found between the ontology model and both database models are graphically represented in Fig. 4. It is worthy to explain some of the decisions taken when elaborating the mapping rules. For example, composed attributes in the databases, such as *Address* and *Date* in database A, are mapped onto ontology concepts. Therefore, a method for splitting up both attribute values and fulfilling the concepts properties has been applied. On the other hand, database foreign keys are mapped not to concept attributes but to interconceptual relationships. Thus, for example, attribute *Company_ID* in database A corresponds with the relation between concept *Order* and concept *Organization* in the ontology.

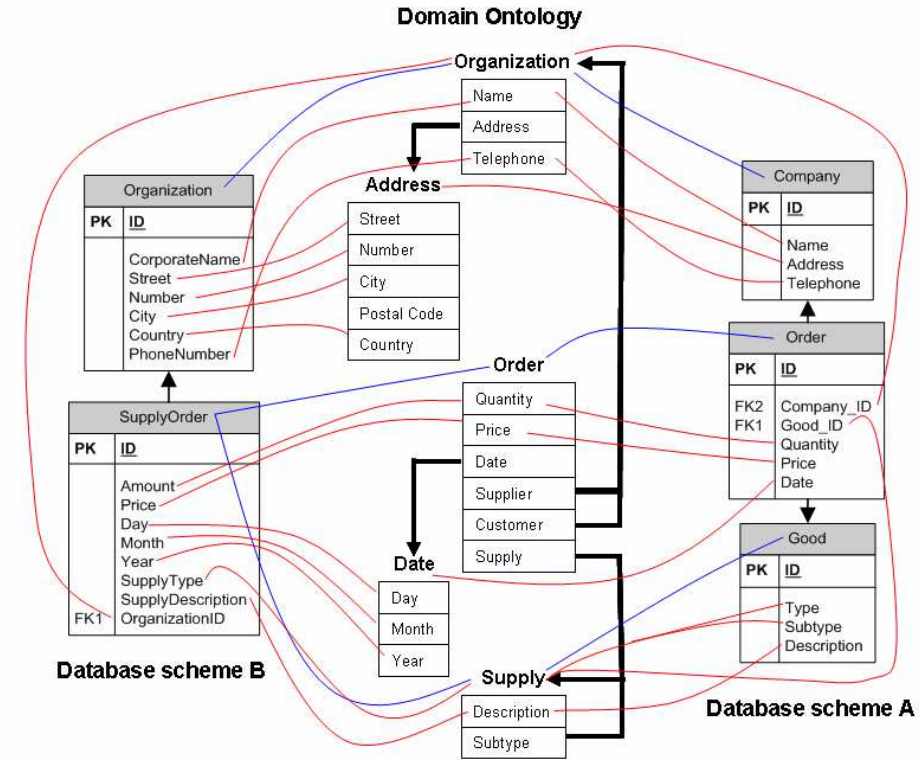

Fig. 4. Mapping rules

The final step in the methodology refers to the process of automatically populating the ontology. In order to do this, it is needed to take into account the mappings rules previously defined. From each attribute of each row in the database tables new instances in the ontology emerge and are fulfilled. In the following tables, a few of the rows in the tables are shown.

In Table 1, a few rows of table *Company* in database A are depicted. This table contains two different rows for two different companies and their data.

Table 1. Table 'Company'; Database A

ID	Name	Address	Telephone
1	Company A	Address A	+34 111 111111
2	Company B	Address B	+34 222 222222

In Table 2, two types of supplies are presented and a brief description of each is given. They belong to table *Good* in database A.

Table 2. Table 'Good'; Database A

ID	Type	Subtype	Description
1	Construction material	Brick	Clay bricks. The dimensions are 230 x 110 x 76 mm
2	Lighting	Bulb	E14 / E27 screw fittings, used in continental Europe. 100 W, 1700 lumens

In Table 3 the last table pertaining to database A is presented, the table *Order*. Two rows of this table are shown.

Table 3. Table 'Order'; Database A

ID	Company_ID	Good_ID	Quantity	Price	Date
1	1	2	500	1000	02/05/2005
2	2	1	1000	2000	10/12/2005

In Table 4, the entity *Organization* of database B is depicted. The data regarding two different rows are revealed.

Table 4. Table 'Organization'; Database B

ID	CorporateName	Street	Number	City	Country	Telephone
1	Company C	St C	45	Madrid	Spain	+34 111 111111
2	Company D	St D	21	Galway	Ireland	+34 222 222222

In Table 5, two supply orders are highlighted. They belong to table *SupplyOrder* in database B.
From these data stored in the different databases, a joint set of ontology instances emerges. In Fig. 5, some of the instances obtained from these heterogeneous data sources are depicted. They have been generated using Ontoviz plug-in for Protégé. Thus, the feasibility of offering a common view from heterogeneous data sources is proven. It should be noted that real names of the companies have been intentionally replaced.

Table 5. Table 'SupplyOrder'; Database B

ID	Amount	Price	Day	..	..	SupplyType	..	OrganizationID
1	20	500	25	..	..	Door	..	2
2	2000	4	10	..	..	Lighting	..	1

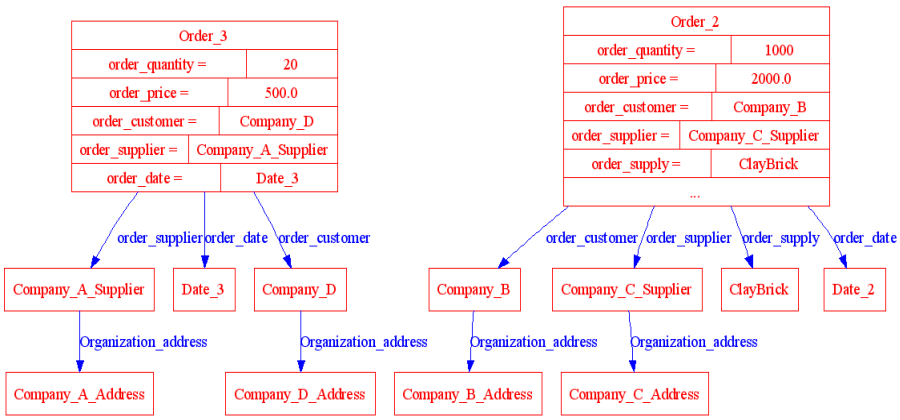

Fig. 5. Ontology instances

4. Conclusions and Future Work

The emergence of the Semantic Web has led to the establishment of ontologies as the de facto standard for knowledge representation. Ontologies permit shareable, reusable, and machine-readable modeling of information, so most of the task regarding the management of the information can be automated. The use of ontologies has then been widespread across many economic sectors. However, construction and building sector has not been targeted enough yet and there is no effective approach to allow for effective data exchange between business partners in this domain. Besides, a more sophisticated solution would permit a better understanding of company's own processes and an improved supply management.

In this paper, a methodology for knowledge acquisition from databases in the construction domain is presented. The methodology consists of a set of simple steps that leads to the elaboration and instantiation of a common and shared ontology. This ontology usefulness is twofold: it facilitates a proper control over data and a better understanding of supply statuses, and it allows for effective data exchange between builders and their suppliers. This methodology is based on previous research studies on knowledge discovery and knowledge discovery from databases, as well as on ontology learning and data integration. We claim that by applying ontology learning techniques and ontologies as knowledge representation, results in knowledge discovery from databases can be improved.

This constitutes a solution to a major issue in companies' relationships, intercommunication. In the building and construction industry, a common issue to solve appears when both the supplier and the builder do not share a common data model. By sharing a common ontology in an upper level of abstraction, instead of exchanging messages containing elements of the database, both the supplier and the builder use terms of the ontology to intercommunicate. The ultimate goal of the undergoing research presented here is to build a platform for supply information creation, integration, and

management in the building and construction domain based on knowledge management technologies and the Semantic Web. This integrated platform would make it possible for users to easily access to the knowledge acquired from heterogeneous information sources and databases. For this to be successfully accomplished, several milestones should be satisfied. The first step is the development of a Web application for cooperative and automatic building of construction supply ontologies. Then, a user-friendly interface need to be designed in order to enable final users (i.e. builders) to access in an intelligent manner to the supply information/knowledge stored. Finally, as knowledge acquisition is an incremental process, (semi)automatic mechanisms for knowledge refinement should be ideated. However, a number of challenges should be faced yet. Some of these challenges are, for example, the construction of large, useful ontologies that are shared by many, and the (semi)automatic creation of mappings.

1. *Semantic* Interoperability Community of Practice, 2005. White Paper Series Module 1: Introducing Semantic Technologies and the Vision of the Semantic Web.
2. *Missikoff, M. Harmonise*: An Ontology-based Approach for Semantic Interoperability. 2002. – ERCIM News 51.
3. *Berners-Lee T.* The Semantic Web. Scientific American, May 2001. – P. 34–43.
4. *Gruber T. R.* A translation approach to portable ontology specifications. Knowledge Acquisition. 1993. – Vol. 5. P. 199–220.
5. Ram, S., and Ramesh V. Schema Integration: Past, Current and Future, in Management of Heterogeneous and Autonomous Database Systems, A. Elmagarmid, M. Rusinkeiwicz, and A. Sheth (Eds.). San Francisco, Morgan Kaufmann, P. 119-155.
6. *Fernández-Breis J.T., Vivancos-Vicente P.J., Menárguez-Tortosa M., Valencia-García R., Miranda-Mena T.,Moner, D., Maldonado J.A, and Martínez Béjar R.* Towards the semantic interoperability of HER information systems. Pacific Rim Knowledge Acquisition Workshop, Guilin, China. 2006.
7. *Paterson G.I.* Semantic Interoperability for Decision Support Using Case Formalism and Controlled Vocabulary. Health'04, Challenges for Today for Success Tomorrow.
8. *Batini C.,* A Comparative Analysis of Methodologies for Database Schema Integration, ACM Computing Surveys, 1986. – Vol. 18. – No.4. P. 323–364.
9. *Ram, S., and Ramesh V.* A Blackboard-Based Cooperative System for Schema Integration. IEEE Expert, June 1995. – P. 56–62.