USING ONTOLOGY META DATA FOR DATA WAREHOUSING

Alberto Salguero, Francisco Araque and Cecilia Delgado
Department of Software Engineering
E.T.S.I.I.T., University of Granada, Granada (Andalucía), Spain

Keywords: Data Warehouse, Ontology, Integration, GIS.

Abstract: One of the most complex issues of the integration and transformation interface is the case where there are multiple sources for a single data element in the enterprise Data Warehouse (DW). There are many facets due to the number of variables that are needed in the integration phase. However we are interested in the integration of temporal and spatial problem due to the nature of DWs. This paper presents our ontology based DW architecture for temporal integration on the basis of the temporal and spatial properties of the data and temporal characteristics of the data sources. The proposal shows the steps for the transformation of the native schemes of the data sources into the DW scheme and end user scheme and the use of an ontology model as the common data model.

1 INTRODUCTION

The ability to integrate data from a wide range of data sources is an important field of research in data engineering. Data integration is a prominent theme in many areas and enables widely distributed, heterogeneous, dynamic collections of information sources to be accessed and handled.

Many information sources have their own information delivery schedules, whereby the data arrival time is either predetermined or predictable. If we use the data arrival properties of such underlying information sources, the Data Warehouse Administrator (DWA) can derive more appropriate rules and check the consistency of user requirements more accurately. The problem now facing the user is not the fact that the information being sought is unavailable, but rather that it is difficult to extract exactly what is needed from what is available. It would be extremely useful to have an approach which determines whether it would be possible to integrate data from two data sources.

The use of DW and Data Integration has been proposed previously in many fields. In (Haller et al., 2000) the Integrating Heterogeneous Tourism Information data sources is addressed using three-tier architecture. In (Vassiliadis, 2001) a framework for quality-oriented DW management is exposed, where special attention is paid to the treatment of metadata. The problem of the little support for automatized tasks in Data Warehousing is considered in (Thalhammer, 2001), where the DW is used in combination with event/condition/action (ECA) rules to get an active DW. Nevertheless, none of the previous works encompass the aspects of the integration of the data according to the temporal and the spatial parameters of the data.

In this article a solution to this problem is proposed: a DW architecture for data integration on the basis of the temporal and the spatial properties of the data and temporal characteristics of the sources and their extraction methods. This architecture give as result the necessary data for the later refreshment of the DW. The use of a data model based on ontologies is proposed as a common data model to deal with the data sources schemes integration. Although it is not the first time the ontology model has been proposed for this purpose (Skotas and Simitsis, 2006), in this case the work has been focused on the integration of spatio-temporal data. Moreover, to our knowledge this is the first time the metadata storage capabilities of some ontology definition languages has been used in order to improve the DW data refreshment process design.

The remaining part of this paper is organized as follows. In Section 2, some basic concepts as well as our previously related works are revised; in section 3 our architecture is presented; Finally, Section 4 summarizes the conclusions of this paper.

Salguero A., Araque F. and Delgado C. (2008).
USING ONTOLOGY META DATA FOR DATA WAREHOUSING.
In *Proceedings of the Tenth International Conference on Enterprise Information Systems - DISI*, pages 497-500
DOI: 10.5220/0001710904970500

2 DATA WAREHOUSE

Inmon (Inmon, 2002) defined a Data Warehouse as "a subject-oriented, integrated, time-variant, non-volatile collection of data in support of management's decision-making process." A DW is a database that stores a copy of operational data with an optimized structure for query and analysis. A *federated database system* (FDBS) is formed by different *component database systems*; it provides integrated access to them: they co-operate (inter-operate) with each other to produce consolidated answers to the queries defined over the FDBS. Generally, the FDBS has no data of its own as the DW has.

We have extended the Sheth & Larson five-level FDBS architecture (Sheth & Larson, 1990), which is very general and encompasses most of the previously existing architectures. In this architecture three types of data models are used: first, each component database can have its own native model; second, a *canonical data model* (CDM) which is adopted in the FDBS; and third, external schema can be defined in different *user models*.

In order to carry out the integration process, it will be necessary to transfer the data of the data sources, probably specified in different data models, to a common data model, that will be the used as the model to design the scheme of the warehouse. In our case, we have decided to use an ontological model as canonical data model.

An ontology is a controlled vocabulary that describes objects and the relations between them in a formal way, and has a grammar for using the vocabulary terms to express something meaningful within a specified domain of interest. They allow the use of automatic reasoning methods. We have extended OWL with temporal and spatial elements. We call this OWL extension STOWL.

3 DATA SOURCES ANNOTATION

The ontology data models are a good option as a CDM but they are too generals. Extending any of the languages for defining ontologies seem the most suitable option. Among all of them, OWL language is selected as the base for defining the CDM.

Firstly an ontology in the OWL language will be built to define the generic temporal primitives found of interest. Then the spatial primitives will also be incorporated. Finally, the information that describes the characteristics of data to integrate, i.e. the metadata which will be useful to design the data extracting, loading and refreshing processes, will be incorporated to the data source scheme using the annotation properties of OWL. Annotation properties are a special kind of OWL properties which can be used to add information (metadata—data about data) to classes, individuals and object/datatype properties.

The result will be an ontology, expressed in OWL, defining the spatial and temporal primitives which will be used to build the schemes of the data sources and a set of properties which will allow the addition of information about the data sources characteristics. We call STOWL (Spatio-Temporal OWL) to this base ontology.

3.1 Temporal Annotation of Data

Due to the nature of the DW the annotation properties will usually refer to the temporal characteristics of data, so a set of annotation properties is defined to describe the sources according to some of the temporal concepts studied in (Araque et al., 2007a). All these properties can be associated directly to the ontology viewed as a resource or individually to each concept defined in the ontology.

```
<owl:DatatypeProperty rdf:ID="hasExtractionTime">
  <rdf:type rdf:resource="http://www.w3.org/2002/07/owl#AnnotationProperty"/>
  <rdfs:range rdf:resource="http://www.w3.org/2001/XMLSchema#int"/>
</owl:DatatypeProperty>
```

Figure 1: Extraction Time definition in STOWL.

STOWL defines, for instance, the Extraction Time of a change in a data source as shown in figure 1. The Extraction Time parameter can be defined as the time expended in extracting a data change from the source. Some examples of the temporal parameters (Araque et al., 2006) that we consider of interest for the integration process are: *Granularity, Availability*, *Extraction Time* , *Transaction time, Storage time*, *Temporal Reference System*…

3.2 Spatial Annotation of Data

After reviewing various spatial data models we have chosen the model proposed by the Open GIS Consortium. This standard, called Geography Markup Language (GML), is an XML grammar written in XML Schema for the modelling, transportation and storage of geographic information. GML is often used as a communication protocol between a large set of GIS applications, both commercial and open source. It has been

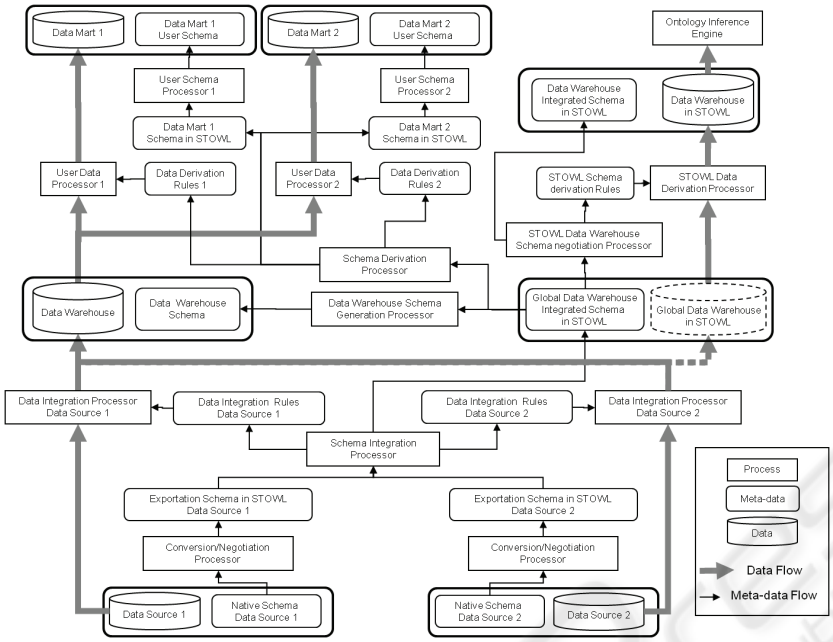

Figure 2: Functional architecture.

necessary the translation of GML, written in XML Schema, to OWL.

GML defines several kinds of entities (such as features, geometrical entities, topological entities…) in form of a object hierarchy. As well as the temporal properties they have been incorporated to STOWL in order to support the inclusion of spatial metadata in the data sources schemes.

We have also considered spatial characteristics like, for instance, the Coordinate Reference System (CRS) and the the *Spatial Granularity* (SGr). concepts. A CRS provides a method for assigning values to a location. All the data in a data source must share the same CRS. In this case, because we are dealing with spatiality, it is common to work with granules like meter, kilometer, country…

As with the temporal data source features, the annotation properties are used to describe the source spatial metadata.

4 DW ARCHITECTURE

Taking paper (Araque et al., 2006) as point of departure, we propose the reference architecture in figure 2. In this figure, the data flow as well as the metadata flow are illustrated. Metadata flow represents how all the data that refers to the data, i.e. the schemes of the data sources, the rules for integrating the data…, are populated through the architecture. Following are explained the involved component.

Native Schema. Initially we have the different data source schemes expressed in its native schemes. Each data source will have, a scheme, the data inherent to the source and the metadata of its scheme. In the metadata we will have huge temporal information about the source: temporal and spatial data on the scheme, metadata on availability of the source.

Preintegration. In the *Preintegration* phase, the semantic enrichment of the data source native schemes is made by the conversion processor. In addition, the data source temporal and spatial metadata are used to enrich the data source scheme with temporal and spatial properties. We obtain the component scheme (CS) expressed in the CDM, in our case using STOWL (OWL enriched with temporal and spatial elements). From the CS the negotiation processor generates the export schemes (ES) also expressed in STOWL. The ES represents the part of a component scheme which is available for the DW designer. It is expressed in the same CDM as the Component Scheme. For security or privacy reasons part of the CS can be hidden.

Integration. The DW scheme corresponds to the integration of multiple ES according to the DW designer needs. It is expressed in an enriched CDM (STOWL) so that temporal and spatial concepts could be expressed straightforwardly. This process is made by the Schema Integration Processor which suggests how to integrate the Export Schemes,

helping to solve semantic heterogeneities (out of the scope of this paper), and defining the Extracting, Transforming and Loading processes (ETL).

The integration processor consist of two modules which have been added to the reference architecture in order to carry out the integration of the temporal and spatial properties of data, considering the data source extraction method used:

The *Temporal and Spatial Integration Processor* uses the set of semantic relations and the conformed schemes obtained during the detection phase of similarities (Oliva and Saltor, 1996). This phase is part of the integration methodology of data schemes. As a result, we obtain data in form of rules about the integration possibilities existing between the originating data from the data sources (minimum resultant granularity…).

The *Metadata Refreshment Generator* determines the most suitable parameters to carry out the refreshment of data in the DW scheme. As result, the DW scheme is obtained along with the Refreshment Metadata necessary to update the DW according to the data extraction method and other spatio-temporal properties of a the data sources.

Data Warehouse Refreshment. After the schema integration and once the DW scheme is obtained, its maintenance and update will be necessary. Each Data Integration Processor is responsible of doing the incremental capture of its corresponding data source and transforming them to solve the semantic heterogeneities. Each Data Integration Processor accesses to its corresponding data source according to the temporal and spatial requirements obtained in the integration stage. A global Data Integrator Processor uses a parallel, fuzzy data integration algorithm to integrate the data (Araque et al., 2007b).

5 CONCLUSIONS

In this paper we have presented a DW architecture for the integration on the basis of the temporal and spatial properties of the data as well as the temporal and the spatial characteristics of the data sources.

We have described the modules introduced to the Sheth and Larson reference architecture. These modules are responsible of checking the temporal and the spatial parameters of data sources and determine the best refreshing parameters according to the requirements. This parameters will be used to design the DW refreshment process, made up by the extracting, transforming and loading data processes.

We used STOWL as the Canonical Data Model. All the data sources schemes will be translated to this one. STOWL is an OWL extension including spatial, temporal and metadata elements for the precise definition of the extracting, transforming and loading data processes.

ACKNOWLEDGEMENTS

This work has been supported by the Andalucía Research Program under project GR2007/07-2 and by the Spanish Research Program under projects EA-2007-0228 and TIN2005-09098-C05-03.

REFERENCES

Araque, F., Salguero, A., Delgado, C., 2007a. Monitoring web data sources using temporal properties as an external resources of a data warehouse. ICEIS. 28-35.

Araque, F., Carrasco, R. A., Salguero, A., Delgado, C., Vila, M. A., 2007b. Fuzzy Integration of a Web data sources for Data Warehousing. Lecture Notes in Computer Science (Vol 4739). Springer-Verlag.

Araque, F., Salguero, A.., Delgado, C., Samos, J., 2006. Algorithms for integrating temporal properties of data in DW. *8th Int. Conf. on Enterprise Information Systems* (ICEIS). Paphos, Cyprus. May.

Haller, M., Pröll, B. , Retschitzegger, W., Tjoa, A. M., Wagner, R. R., 2000. Integrating Heterogeneous Tourism Information in TIScover - The MIRO-Web Approach. Information and Communication Technologies in Tourism, ENTER. Barcelona (Spain)

Inmon W.H, 2002. Building the Data Warehouse. John Wiley.

Oliva, M., Saltor, F., 1996. A Negotiation Process Approach for Building Federated Databases. In Proceedings of 10th ERCIM Database Research Group Workshop on Heterogeneous Information Management, Prague. 43–49.

Sheth, A., Larson, J., 1990. Federated Database Systems for Managing Distributed, Heterogeneous and Autonomous Databases." ACM Computing Surveys, Vol. 22, No. 3.

Skotas, D., Simitsis, A., 2006. Ontology-Based Conceptual Design of ETL Processes for Both Structured and Semi-Structured Data. International Journal on Semantic Web and Information Systems, Vol. 3, Issue 4. pp. 1-24.

Thalhammer, T., Schrefl, M., Mohania, M., 2001. Active data warehouses: complementing OLAP with analysis rules, Data & Knowledge Engineering, 39 (3), 241-269.

Vassiliadis, P., Quix, C., Vassiliou, Y., Jarke, M., 2001. Data warehouse process management. Information System. v. 26. 205-236.