DATA WHAREHOUSES: AN ONTOLOGY APPROACH

Alexandra Pomares Q.
Systems Engineer Department, Javeriana University,Cra 7 No 40-62, Bogotá D.C, Colombia

José Abásolo P.
Systems Engineer Departments, Los Andes University, Carrera 1 N° 18A 10, Bogotá D.C, Colombia

Keywords: Data warehouse, ontologies, dimensional model, data integration, data warehouse design.

Abstract: Although dimensional design for data warehouses has been used in a considerable amount of projects, it does have limitations of expressiveness, particularly with respect to what can be said about relations and attributes properties and restrictions. We present a new way to design data warehouses, based on ontologies, that overcomes many of these limitations. In the proposed architecture descriptive ontologies are used to build the data warehouse and taxonomic ontologies are used during data preparation phase. We discuss the expressive power of Ontology approach showing a semantic comparison with dimensional model both applied to a case study.

1 INTRODUCTION

The complexity of data warehouse models based on the e-r model was one of the biggest driving forces behind dimensional modeling, which was created so that the designed models where easily understood by a business expert and easily analyzed by the final user. Nevertheless, the evolution of the dimensional paradigm has showed that the representation of the business world is so complex that it is necessary to introduce new concepts to the models like bridge tables, heterogeneous dimensions, factless fact table, etc. (according to Kimball & Ross (2002)) to allow a greater level of representation. As a result, the designed model lacks the desired simplicity and does not yet guarantee the representation of all the semantics of the domain.

This article explores an alternative to the design of data warehouses that allows the creation of a model that reflects in a greater proportion the semantic of the business world and that can be exploited by the final user through different analysis tools. The alternative, based on ontologies, is shown through a comparison with dimensional model with regards to the level of semantic representation, exploring all the limitations and ease of use derived from the standard language for ontologies known as OWL (Web Ontology Language).

The objective is to make a comparison between the dimensional and the ontology design, stressing out the semantic richness of each of the approaches. In order to do so, the article will explore briefly, in Section 2, the applied ontologies in data integration; then, in Section 3, it will show the proposed architecture that will be applied in a real case study in Section 4; and finally it will make a comparative analysis of both approaches in Section 5.

2 ONTOLOGIES AND DATA INTEGRATION

2.1 Ontologies General Concepts

In 1993 Tom Gruber defined an ontology is "a formal and explicit specification of a conceptualization" (cited in Antoniou & van Harmelen 2004). Its objective, according to Heflin (2004), is "to be used by persons, data bases and software applications that need to share the information of a domain" and produce knowledge from it.

One of the biggest advances in the area of ontologies was the creation of design language, known as the Web Ontology Language (OWL) by the World Wide Web Consortium (W3C). The elements that OWL uses to represent a domain creates a powerful semantic that allows representing a knowledge domain more accurately than other

Pomares Q. A. and Abásolo P. J. (2006).
DATA WHAREHOUSES: AN ONTOLOGY APPROACH.
In *Proceedings of the Eighth International Conference on Enterprise Information Systems - DISI*, pages 187-192
DOI: 10.5220/0002460601870192

languages created to model ontologies like the RDF Schema, DAML or OIL.

2.2 Data Integration

Data integration is concerned with unifying data that shares common semantics but originates from heterogeneous sources. The level of unification depends on the type of heterogeneity: structural, when the source data models are different; syntactic, when the source data models use different languages; or semantic, when there are different concepts with similar meaning or similar concepts with different meanings.

Most of the problems related to syntactic and semantic heterogeneity have been solved with ontologies that are used for mapping concepts between different data models. In these cases, the ontologies allow the translation between different sources so that they arrive unified to the destined data model. An example of this type of integration is showed in Kedad, & Métais (2002) where a domain ontology was defined to unify data from sources with different syntactic terminology but semantically related. In this type of problems the use of the ontology is not to conceptualize the entire domain, but only those zones that have syntactic or semantic problems.

For the problem of structural heterogeneity the ontology is used not as a translator but as a reference data model in which all sources must stay within. One of the areas that has used a lot this type of ontologies to integrate knowledge is bio-informatics in which the semantic and structural heterogeneity is solved as shown in Clusters & Smith Fielding, (2004) through a case study.

In this context, data warehouses been task independent and defining a reference model that allows to integrate multiple sources can be seen like an ontology that solves the problem of structural integrity of organizational databases. The compatibility between data warehouses and ontologies is so close, that the concept of data warehouse can be materialized through an ontology.

3 PROPOSED ARCHITECTURE: ONTOLOGY - BASED DATA WAREHOUSES

The principle of the architecture is that the design of a data warehouse must be done looking to reflect the domain of the world most close to reality, independently of the complexity of the resultant model, because for presentation purposes this can be reduced to the level of simplicity required by the final user.

In the proposed architecture (shown in figure 1) the data warehouse is filled with data from operating systems and data obtained from external ontologies that are treated in an intermediate preparation layer. The objective of this layer is to transform and generate the correct structures so that they can be loaded to the warehouse. It is in this layer that the taxonomic ontologies, that allows the integration of semantic and syntactic heterogeneity, are located.

The data warehouse is built upon ontologies that allow representing the world through structures of great semantic power, obtaining as a result a model much closer to reality than the dimensional model. The warehouse is accessed through a mediator which generates the correct views (virtual or materialized) based on the level of comprehension and detail required by each type of user. Depending on the type of tool that each of them uses, the data warehouse will be accessed directly or using the mediator.

The data warehouse is constituted by a descriptive ontology (a kind of ontology that according to Kedad & Métais 2002 contains instances of their classes that are stored in a database or other semi-structured store media) that represents the world domain. This ontology is administered by an Ontology Management System (OMS) which according to Cullot & Parent et al (2003) offers four functionalities: allow data modeling, provide efficient store services and instance management, provide tools of reasoning, and allow queries over the model and its instances. The OMS provides inference engines that enrich the model even more, because from facts originated in the sources they can infer additional facts called derived facts Lee & Goodwin et al (2003).

The warehouse can be built incrementally adding more classes, properties and restrictions to the ontology in accordance to the business process that is been modeled. The data integration of the different business processes is guaranteed by the preparation layer and the equivalence properties provided by OWL-like equivalentClass, equivalentProperty and class consructors like unionOf and intersectionOf, among others.

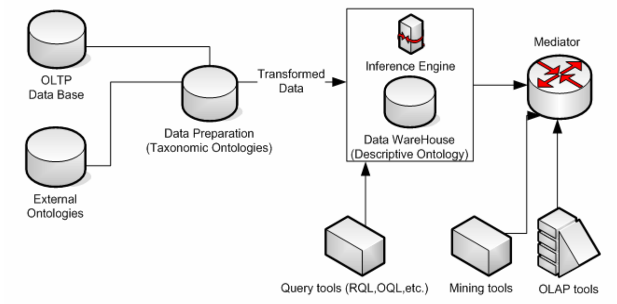
OLTP Data Base
External Ontologies
Data Preparation (Taxonomic Ontologies)
Transformed Data
Inference Engine
Data WareHouse (Descriptive Ontology)
Mediator
Query tools (RQL,OQL,etc.)
Mining tools
OLAP tools

Figure 1: Proposed Architecture.

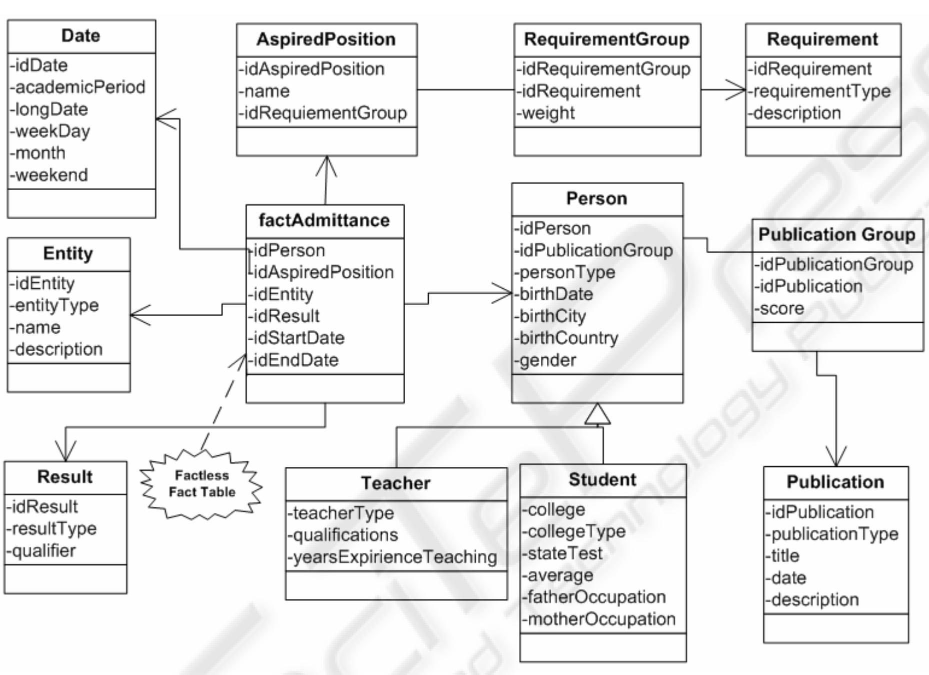

Figure 2: Dimensional model – Admittance.

4 CASE STUDY

To test the ontology approach and compare it at the same time with the dimensional approach, both of them were applied to a University domain, to analyze their semantic representation power.

One of the requirements of the data warehouse is to represent the admittance and number of applicants of an academic program for a period, including administrative, teaching and student positions. In the section 4.1 and 4.2 we show this requirement using both models.

4.1 Dimensional Model

Figure 2 shows the dimensional diagram of admittance. In the formalism used in the figure, the arrow symbol was used to represent 1 to n relations, pointing to the 1 side of the relation; in heterogeneous dimensions, inheritance symbol is used. During the design, there were various difficulties that, even though resolved, made the final data model very complex. Some of them were:,

- Each position has a set of requirements with a defined weight. These requirements could not be modeled inside the dimension *AspiredPosition* because the grain of the dimensions would be violated. Neither could they be related directly with the table of facts, because they where related to a position. The only alternative was to create a bridge table to relate Requirement with *AspiredPosition*.
- For the model to be flexible, the table of facts of admittance must support the record

of admittance for every type of person; nevertheless, depending on the type, the attributes will be different. For this reason, it was necessary to use heterogeneous dimensions, in which a table is added for each type of person to extend the dimension depending on the case.

- Another complex issue was that the table *Person* has attributes with multiple values (e.g. *Publications*) so another bridge table had to be included to take account of the values.

The semantic limitations were identified when attempting to represent the following restrictions of the domain:

- Depending on the aspired position, it is necessary to restrict the possible related records of the entity dimension. For example, if the aspired position is Undergraduate Student, it is only possible to relate it to the Entity dimension where the field *entityType* is equal to *undergraduateStudent*.
- All types of persons can record publications, nevertheless, only those that come from Person type Teacher will have a score.
- At the model level, it is not possible to limit the record of publications in accordance with the type of person. For example, publications can be assigned to the type of person *Administrative Worker*.
- It is not possible to represent that every position should have at least one requirement of Academic type.
- It is not possible to make distinctions between the type of students or teachers in accordance of their characteristics.

4.2 Ontology Model

The first issue raised when using the ontology approach to model data warehouses was how to join time to object type properties and data type properties. The following alternatives can be used:

1. 3-nary relation: When a property exists between two classes, an intermediate class is created to join both classes with *Date*.
2. Date as a subclass: Through the creation of a class named *Date* that is a subclass of all classes (equal to the *Nothing* class). This approach seeks to include the date between the range of any property and then define for all the properties the following two restrictions:
    - The range of the property must have some value from the **Date** class; and
    - All ranges of the property must have a minimum cardinality of 2: one of the elements is the direct range of the relation and the other is the relation with the *Date* class.
3. Range modifications: The range of all properties is defined as the *join* of the date with the class, that was originally the only range, and the cardinality of 1 is established as the minimum related to the Date class.

    The form of time representation in the model is a choice of each designer, but is subject to the chosen "flavor" of OWL. If the designer chooses OWL Lite, for example, the only choice to use is the 3-nary relation.

The designed ontology to support the need of the case study is partially shown in Figure 3. It was created following OWL recommendation (Dean & Schreiber 2004).

In this ontology, restriction of range and cardinality was defined to describe the business world more accurately. Some of the defined restrictions that allow representing the limitation of the dimensional model are:

- The property *inSelection* of the *Person* class has a cardinality of 1.
- In the *Position* class over the property *hasRequirement* a restriction was defined to have some values from the *AcademicRequirement* class, which is defined as the union between Requirement class and the condition hasType equals to Academic.
- Different types of students where defined through new classes (like undergraduateStudent) form the union of Students and the property OrganizationRelated that has some valued from Undergraduate class.
- The class academicApplicant was created from the union of Position class and the property positionType equals to student. For this new class, the values for the property isPartOfOrganization must be in Program class.

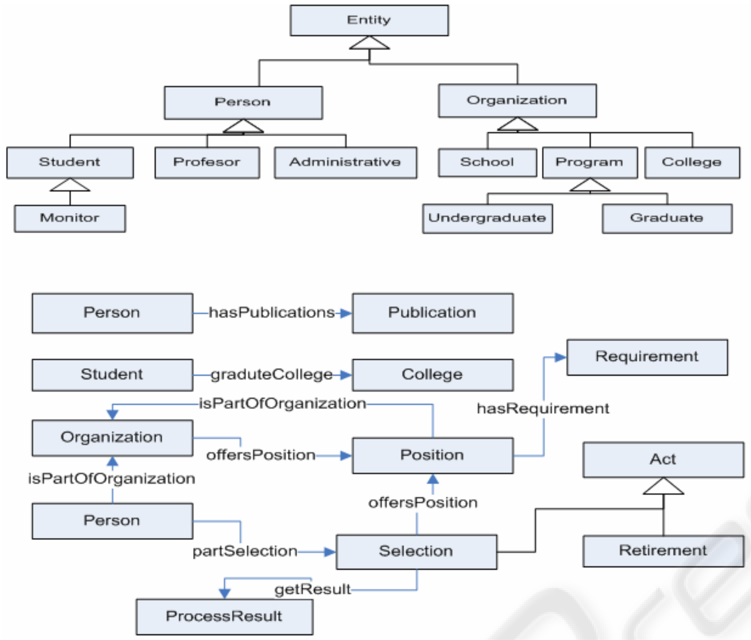

Figure 3: Ontology model – Admittance.

Table 1: Approaches Comparison.

Characteristic	Dimensional Approach	Ontology Approach
World elements representation	Fact tables, dimensions, attributes	Classes, properties, restrictions, axioms.
Representation of simple relations between elements	Relations 1 to N.	- Binary relations. - Domain restrictions, range and cardinality for properties. - Relation between properties.
Representation of complex relations between elements	- The heterogeneous dimensions use the generalization – specification concept.	- Relations of union, difference and complement between the elements of the world and the restrictions. - Inheritance relations between classes and between properties.
Representation of internal restrictions of the elements	The integrity restrictions are at preparation level, not in the model itself.	- Restriction in the data preparation layer and the model itself.
Representation of restriction within the relation between elements	The referential integrity is a product of the foreign keys.	- Restriction of the possible values or range of value within a property.at any level of specialization. - Restriction to establish the number of individuals related to a property. - It's not possible to define a property as the union or intersection of others
Time inclusion	Time dimension is included and the management of the changes to each attribute is defined.	There is liberty to establish the inclusion of Time in the model.
Data integration	The integration is defined by the methodology used in the project.	Uses concepts like equals to, different from and disjunction of classes and properties.
Complexity of final design	The representation of a real domain is more complex than a star diagram.	The model is complex to the final user
Conditionals	Inside the model, it's not possible to define conditions to establish relations between elements. For example, it's not possible to define that an Admittance of one type of Person should only have one kind of Aspired Position.	Each class can have conditions that defined characteristics of the individuals that contains. Conditions can be established as necessary or necessary and sufficient. It is not possible to define conditions like: If element hasValue X then property Y is applicable.
Knowledge generation	The inference of facts is a responsibility of the final users.	Derived facts can be inferred of original facts automatically.
Paradigm evolution	The dimensional model is not a standard.	OWL is a recommendation of w3c which encourages its upgrade and evolution.

5 COMPARISON OF THE APPROACHES

To look more clearly the semantic differences of both approaches for data warehouse design, a comparison of the core characteristics of each one is presented in Table 1.The semantic analysis was made using the OWL specification (Dean & Schreiber 2004.).

6 CONCLUSIONS

The domain representation through ontologies provides more flexible mechanisms to represent the complexity, relations and restrictions of the business World than those offered by the dimensional model. Nevertheless, the approach has nowadays limitations related to the creation of properties and restriction qualifications.

Although the dimensional model offers additional mechanisms, different from the dimensions and the facts, to represent most of the elements of the world, they are not enough to model the complexity, relations and restrictions of the business world.

The proposed architecture for the construction of data warehouses, based on ontologies generates more semantically rich models which are easier to integrate them than the traditional architecture.

REFERENCES

Antoniou G. & van Harmelen F. (2004) A Semantic Web Primer. First Edition. MIT Press

Clusters, W. & Smith, B. Fielding, J. (2004). On the Application of Formal Principles to Life Science Data: A Case Study in the Gene Ontology. [Electronic Version]. DILS 2004.

Cullot N. & Parent C & Spaccapietra s. & Vangenot C. (2003) Ontologies: A contribution to the DL/DB debate. Proceedings of Semantic Web and Databases 2003: Berlin, Germany . http://lbdsun.epfl.ch/e/publications_new/articles.pdf/Cullot_SW_DB2003_CR.pdf (Published Septiembre 2003; accessed 8 July 2005)

Dean M. & Schreiber G. (2004). OWL Web Ontology Language Reference W3C Recommendation. http://www.w3.org/TR/owl-ref/ (Published 10 February 2004; accessed 11 March 2005)

Heflin J (2004). OWL Web Ontology Language Use Cases and Requirements. W3C Recommendation. http://www.w3.org/TR/webont-req/ (Published 10 February 2004; accessed 22 April 2005)

Kedad, Z. & Métais E. (2002). Ontology – Based Data Cleaning. [Electronic version]. NLDB 2002 Record.

Kimball R. & Ross M. (2002) The data warehouse toolkit : the complete guide to dimensional modeling. Second Edition. New York : Wiley, c2002

Lee J. & Goodwin R. et al (2003). Towards Enterprise-Scale Ontology Management. IBM T. J. Watson Research Center, 2004.