Describing Data Sources Semantically for Facilitating Efficient Creation of OLAP Cubes

Santtu Toivonen[1] **and Tapio Niemi**[2]
santtu.toivonen@vtt.fi, tapio.niemi@cern.ch

[1] VTT Information Technology, P.O.Box 1203, FIN-02044 VTT, FINLAND
[2] Helsinki Institute of Physics, CERN Offices, CH-1211 Geneva, SWITZERLAND

Abstract

Traditionally data used in on-line analytical processing (OLAP) has limited to the content of one data warehouse of the company. However, analysis needs are often more advantages and data is needed from different sources. We study how the semantics of data sources can be described to allow combining data from several sources into an OLAP cube. We apply Semantic Web technologies for defining contents of the data sources. RDF descriptions conforming to an ontology can be used to find suitable data from different sources and to semantically integrate the needed heterogeneous data into one OLAP cube.

1 Introduction

OLAP has gained popularity as a method to support decision making in situations where raw data on measure, such as sales or profit, needs to be analysed at different levels of statistical aggregation [Pendse, 2004]. In OLAP, queries are made against a multidimensional database, called an OLAP cube, in which the dimension attributes (i.e. coordinates) determine the measure values. For example, if the dimensions are *time*, *location*, *product*, the measures can be *sales* and *profit*. A dimension can have a hierarchical structure to enable that the analysis can be performed, e.g., at daily or monthly level. In the latter case, the monthly data is aggregated from the daily data.

The contents of OLAP databases are typically collected from other data repositories, such as databases in operational use. The traditional approach is to implement a large multidimensional database and upload the data frequently from the operational databases. However, it should be possible to use different kinds of databases and other data sources on the intranet of a company or even on the Internet in the analysis. This traditional design and implementation method does not work very well since it is obvious that all potential data cannot be collected a priori into one analysis database. Our suggestion is that the user beforehand defines what kind of data will be needed in the analysis and the data is collected after that [Niemi *et al.*, 2003].

To be able to collect semantically appropriate data we utilise Semantic Web technologies. More specifically, we use Resource Description Framework (RDF) for describing the instances, and Web Ontology Language (OWL) for creating the ontology that restricts the semantics of the RDF descriptions. Using this approach it is possible to implement a software to assist the user in designing a suitable OLAP schema and performing the data extraction, transformation and loading (ETL) processes.

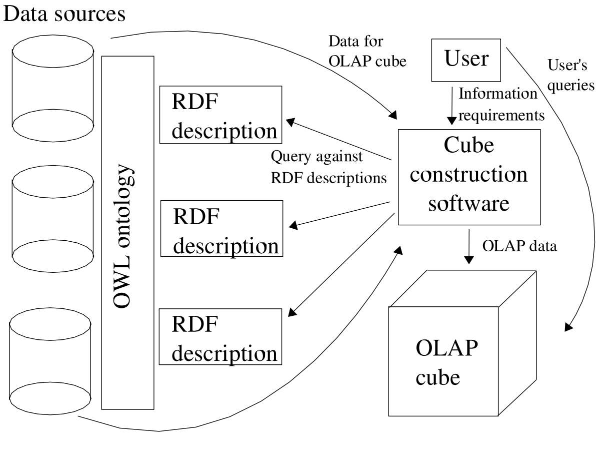

Figure 1: System architecture

Our framework is illustrated in Figure 1. We assume a set of data sources with OWL ontologies and RDF descriptions. Typically, the description is given by the administrator of the data source, or alternatively the user of the framework. Descriptions can locate differently from the data sources. The process works as follows:

1. The user starts the analysis by defining the data that needs to be collected to the OLAP cube.
2. The system locates the data sources based on RDF descriptions conforming to an ontology of the application area, constructs a logical schema for the OLAP cube, and suggests the design to the user.
3. The user can accept or modify the suggested design.
4. The system suggests how data should be manipulated before storing to the OLAP cube. This includes, for example, modifications because of different measure units or currencies.
5. The OLAP cube is constructed to an OLAP server by the system and the user can start the analysis.

2 Using Semantic Web Technologies for Working with OLAP Data

XML is intended for describing the structure of Web documents, whereas RDF and the ontology languages are intended for describing the meaning of any resources—whether on the Web or not. With Semantic Web technologies it is more straightforward to define for example cardinalities, class-relationships, data types, and logical relationships (such as disjoint, inverse, union, etc.). There is a role for XML, however, and that is the common serialisation syntax for Semantic Web descriptions [Decker *et al.*, 2000].

To illustrate the semantic composition of OLAP cubes, we apply world trade data. The data contains pairwise import/export figures of countries classified according to product groups. Figure 2 depicts a simple OWL-ontology for defining an OLAP cube of trade data. The concept of *OLAP Cube* has *FactTables*, which consist of *FactRows*. One *FactRow* specifies one trade instance. In such instance exactly one *ImportCountry*, one *ExportCountry*, one *Year*, one *Product*, and one *Value* are given. A product belongs in exactly one *SubGroup*, which in turn belongs in exactly one *MainGroup*.

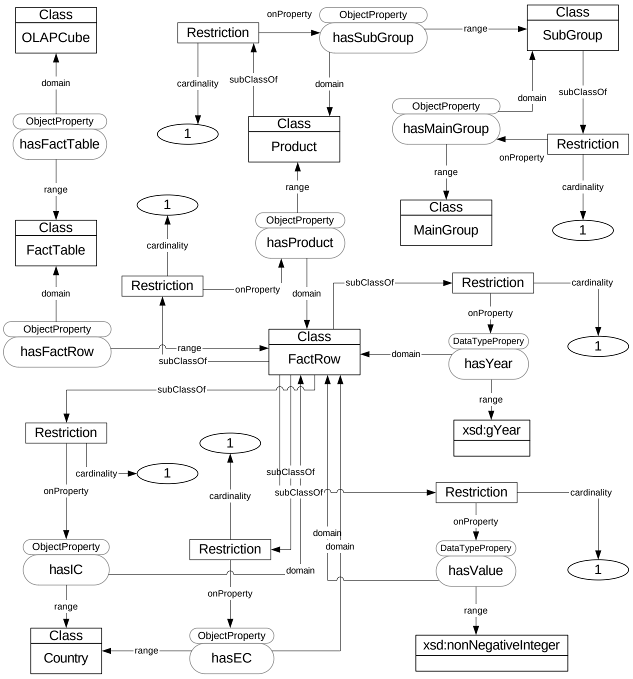

Figure 2: OWL model of world trade data

Our RDF approach builds on earlier XML models for OLAP data, such as the star schema described in [Niemi et al. 2003]. However, in order to achieve richer semantics, we now consider Semantic Web technologies instead of plain XML. An example RDF description of the data is as follows:

```
<rdf:Description rdf:about="Trade">
     <rdf:type rdf:resource="http://olapcubes.org#OlapCube" />
     <trade:hasFactTable rdf:resource="TradeFactTable" />
</rdf:Description>

<rdf:Description rdf:about="TradeFactTable">
     <rdf:type rdf:resource="http://olapcubes.org#FactTable" />
     <trade:hasFactRow rdf:resource="TradeFactRow1" />
</rdf:Description>

<!--  Row 1 -->
<rdf:Description rdf:about="TradeFactRow1">
     <rdf:type rdf:resource="http://olapcubes.org#FactRow" />
     <trade:hasIC rdf:resource="http://countries.org#UK" />
     <trade:hasEC rdf:resource="http://countries.org#Finland" />
     <trade:hasYear>1988</trade:hasYear>
     <trade:hasValue>200</trade:hasValue>
     <trade:hasProduct rdf:resource="FinePaper" />
</rdf:Description>

<rdf:Description rdf:about="FinePaper">
     <rdf:type rdf:resource="http://olapcubes.org#Product" />
     <trade:hasSubGroup rdf:resource="Paper" />
</rdf:Description>

<rdf:Description rdf:about="Paper">
     <rdf:type rdf:resource="http://products.org#SubGroup" />
     <trade:hasMainGroup rdf:resource="Forest" />
</rdf:Description>

<rdf:Description rdf:about="Forest">
     <rdf:type rdf:resource="http://products.org#MainGroup" />
</rdf:Description>

<!--  Row 2 -->
<rdf:Description rdf:about="TradeFactRow2">
     :
</rdf:Description>
```

3 Conclusions and Future Work

Our RDF representation for OLAP determines the basic structure of the OLAP cube, i.e. the set of attributes forming the dimension hierarchies, functional dependencies inside dimensions, and measure attributes. The definition could easily contain some addition information, e.g. the aggregation type of measure attributes (flow, stock, etc.), the type of the dimension (time, geographical, etc.), and the type of the dimension hierarchy. Moreover, the model is capable to indicate relationships, e.g. functional dependencies, among the attributes in different hierarchies or even in different databases. Also, the 'meaning' of attributes (sales values, temperature, etc.) could be described.

Utilising Semantic Web makes OLAP systems more flexible and efficient, since Semantic Web enables the user to integrate data from different sources. Searching the required information, integrating data, and data transformation and uploading can partially be automated by applying semantic descriptions of data sources. Moreover, we believe that the quality of the data in an analysis database will be improved due to explicit semantics of data sources.

Acknowledgments

We wish to thank Marko Niinimäki and Jyrki Nummenmaa for their comments and ideas. The work was partly support by the Academy of Finland, project 203554.

References

[Decker *et al.*, 2000] Stefan Decker, Sergey Melnik, Frank Van Harmelen, Dieter Fensel, Michel Klein, Jeen Broekstra, Michael Erdmann, and Ian Horrocks. The Semantic Web: The roles of XML and RDF. *IEEE Internet Computing*, 15(3):63–74, October 2000.

[Niemi *et al.*, 2003] Tapio Niemi, Marko Niinimaki, Jyrki Nummenmaa, and Peter Thanisch. Applying Grid technologies to XML based OLAP cube constraction. In *Design and Management of Data Warehouses DMDW'03*, 2003.

[Pendse, 2004] Nigel Pendse. What is OLAP?, 2004. Available on: http://www.olapreport.com/fasmi.htm.