Chapter LI
An Overview of Ontology-Driven Data Integration[1]

Agustina Buccella
Universidad Nacional del Comahue, Argentina

Alejandra Cechich
Universidad Nacional del Comahue, Argentina

INTRODUCTION

New software requirements have emerged because of innovation in technology, specially involving network aspects. The possibility enterprises, institutions and even common users can improve their connectivity allowing them to work as they are at the same time, generates an explosion in this area. Besides, nowadays it is very common to hear that large enterprises fuse with others. Therefore, requirements as interoperability and integrability are part of any type of organization around the world. In general, large modern enterprises use different database management systems to store and search their critical data. All of these databases are very important for an enterprise but the different interfaces they possibly have make difficult their administration. Therefore, recovering information through a common interface becomes crucial in order to realize, for instance, the full value of data contained in the databases (Hass & Lin, 2002).

Thus, in the '90s the term *Federated Database* emerged to characterize techniques for proving an integrating data access, resulting in a set of distributed, heterogeneous and autonomous databases (Busse, Kutsche, Leser & Weber, 1999; Litwin, Mark & Roussoupoulos, 1990; Sheth & Larson, 1990). Here is where the concept of *Data Integration* appears. This concept refers to the process of unifying data sharing some common semantics but originated from unrelated sources. Several aspects must be taken into account when working with Federated Systems because the main characteristics of these systems make more difficult the integration tasks. For example, the *autonomy* of the information sources, their *geographical distribution* and the *heterogeneity* among them, are some of the main problems we must face to perform the integration. *Autonomy* means that users and applications can access data through a federated system or by their own local system. *Distribution* (Ozsu & Valduriez, 1999) refers to data (or computers) spread among multiple sources and stored in a single computer system or in multiple computer systems. These computer systems may be geographically distributed but interconnected by a communication network. Finally, *heterogeneity* relates to different meanings that may be inferred from data stored in databases. In (Cui & O'Brien, 2000), heterogeneity is classified into four categories: *structural*, *syntactical*, *system*, and *semantic*. Structural heterogeneity deals with inconsistencies produced by different data models whereas syntactical heterogeneity deals with consequences of using different languages and data representations. On the other hand, system heterogeneity deals with having different supporting hardware and operating systems. Finally, semantic heterogeneity (Cui & O'Brien, 2000) is one of the most complex problems faced by data integration tasks. Each information source included in the integration has its own interpretation and assumptions about the concepts involved in the domain. Therefore, it is very difficult to determine when two concepts belonging to different sources are related. Some relations among concepts that semantic heterogeneity involves are: synonymous, when the sources use different terms to refer to the same concept; homonymous, when the sources use the same term to denote completely different concepts; hyponym, when one source contains a term less general than another in another source; and hypernym, when one source contains a term more general than another in another source; etc.

In this paper we will focus on the use of ontologies because of their advantages when using for data integration. For example, an ontology may provide a rich, predefined vocabulary that serves as a stable conceptual interface to the databases and is independent of the database schemas; knowledge represented by the ontology may be sufficiently comprehensive to support translation of all relevant information sources; an ontology may support consistency management and recognition of inconsistent data; etc. Then, the next section will analyze several systems using ontologies as a tool to solve data integration problems.

BACKGROUND

Recently, the term *Federated Databases* has evolved to *Federated Information Systems* because of the diversity of new information sources involved in the federation, such as HTML pages, databases, filing, etc., either static or dynamic. A useful classification of information systems based on the dimensions of distribution and heterogeneity can be found in (Busse et al., 1999). Besides, this work defines the classical architecture of federated systems (based on Sheth & Larson (1990)) which is widely referred by many researches. Figure 1 shows this architecture.

In the figure, the *wrapper layer* involves a number of modules belonging to a specific data organization. These modules know how to retrieve data from the underlying sources hiding their data organizations. As the federated system is autonomous, local users may access local databases through their local applications independently from users of other systems. Otherwise, to access the federated system, they need to use the *user interface layer*.

The *federated layer* is one of the main components currently under analysis and study. Its importance comes from its responsibility to solve the problems related to semantic heterogeneity, as we previously introduced. So far, different approaches have been used to model this layer. In general they use ontologies as tools to solve these semantic problems among different sources