Ontology-Based Data Integration

Agustina Buccella  
*Universidad Nacional del Comahue, Argentina*

Alejandra Cechich  
*Universidad Nacional del Comahue, Argentina*

Nieves Rodríguez Brisaboa  
*Universidade da Computación, Spain*

**INTRODUCTION**

Nowadays, different areas of large modern enterprises use different database management systems to store and search their critical data. Competition, evolving technology, geographical distribution, and the inevitable growing decentralization contribute to this diversity. All of these databases are very important to an enterprise, but their different interfaces make their administration difficult. Therefore, recovering information through a common interface becomes crucial to realize, for instance, the full value of data contained in the databases (Hass & Lin, 2002).

In the 1990s, the term *federated database* emerged to characterize techniques for proving an integrating data access, giving a set of distributed, heterogeneous, and autonomous databases (Busse, Kutsche, Leser & Weber, 1999; Litwin, Mark, & Roussopoulos, 1990; Sheth & Larson, 1990). Even these concepts are rather well-known today, a brief introduction would clarify their meanings in the context of our paper. They are as follows:

- **Autonomy:** Users and applications can access data through a federated system or by their own local system. Autonomy can be classified into three types: *design autonomy*, *communication autonomy* and *execution autonomy* (Busse et al., 1999; Oszu & Valduriez, 1999). The first type refers to the data model independence, the second type involves the different ways of communication among the systems, and the third type refers to the independence of the execution of the local operations.

- **Distribution:** In the last years and with the arrival of the Internet it is very common to see computers connected by some type of network. Generally speaking, data may be distributed 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.

- **Heterogeneity:** Different meanings that may be inferred from data stored in databases. In Cui and O’Brien (2000), heterogeneity is classified into four categories: *structural*, *syntactical*, *system*, and *semantic*. Structural heterogeneity deals with inconsistencies produced by different data models; syntactical heterogeneity deals with consequences of using different languages and data representations; system heterogeneity deals with having different supporting hardware and operating systems; and semantic heterogeneity is further classified as (a) dealing with semantically equivalent concepts; some models use different terms to refer to the same concept (e.g., synonyms, or properties, are modeled differently by different systems); (b) dealing with semantically unrelated concepts; the same term may be used by different systems to denote completely different concepts; and (c) dealing with semantically related concepts by using generalization/specification, different classifications, and so forth. Additionally, a similar classification of heterogeneity can be found in Goh (1996).

In this paper, we will focus on the use of ontologies because of their advantages when using them 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; and an ontology may support consistency management and recognition of inconsistent data. 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 into federated information systems (FIS) because of the diversity of new information sources involved in the federation, such as HTML pages, databases, and filing, 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). Figure 1 shows this classification where the addition of an additional dimension—autonomy—makes the term Federated Information Systems appears.

Besides, the work reported in Busse et al., 1999 defines the classical architecture of federated systems (based on Sheth & Larson [1990]), which is widely referred by many researches. Figure 2 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 the semantic heterogeneity, as was previously introduced. So far, different approaches have been used to model this layer. They are as diverse as complementary in some cases, and can involve different perspectives such as the use of ontologies (Ambite et al., 1997; Buccella, Cecich, & Brisaboa, 2003; Goh, Bressan, Siegel, & Madnick, 1999; Gray et al., 1997), the use of metadata (Busse et al., 1999; Nam & Wang, 2002; Seligman & Rosenthal, 1996).

ONTOLOGY-BASED DATA INTEGRATION

The term ontology was introduced by Gruber (1993) as an “explicit specification of a conceptualization.” A conceptualization, in this definition, refers to an abstract model of how people commonly think about a real thing in the world; and explicit specification means that concepts and relationships of the abstract model receive explicit names and definitions.

An ontology gives the name and description of the domain-specific entities by using predicates that represent relationships between these entities. The ontology provides a vocabulary to represent and communicate domain knowledge along with a set of relationships containing the vocabulary’s terms at a conceptual level. Therefore, because of its potential to describe the semantic of information sources and to solve the heterogeneity problems, the ontologies are being used for data integration tasks.

Some surveys on ontology-based systems for data integration can be found in literature. For example, Wache et al. (2001) focused on some aspects of the use of ontologies: the language representation, mappings, and tools. This work also classified the use of ontologies into three approaches: single ontology approach, multiple ontology approach, and hybrid ontology approach (see Figure 3).

Another different survey comparing the expressiveness of the languages can be found in Corcho and Gomez-Perez (2000), but only languages to represent ontologies are compared in this case.

This section will focus on how ontology-based systems address the semantic heterogeneity problems. We have investigated many systems, which follow some of the approaches of Figure 3, considering the relevant aspects shown in the hierarchy of Figure 4. This hierar-
Related Content

[www.igi-global.com/article/rich-base-schema/3247?camid=4v1a](www.igi-global.com/article/rich-base-schema/3247?camid=4v1a)

A Multiple-Bits Watermark for Relational Data
[www.igi-global.com/article/multiple-bits-watermark-relational-data/3388?camid=4v1a](www.igi-global.com/article/multiple-bits-watermark-relational-data/3388?camid=4v1a)

Clustering Vertices in Weighted Graphs
[www.igi-global.com/chapter/clustering-vertices-weighted-graphs/58615?camid=4v1a](www.igi-global.com/chapter/clustering-vertices-weighted-graphs/58615?camid=4v1a)

Adaptive XML-to-Relational Storage Strategies
[www.igi-global.com/chapter/adaptive-xml-relational-storage-strategies/20771?camid=4v1a](www.igi-global.com/chapter/adaptive-xml-relational-storage-strategies/20771?camid=4v1a)