Chapter 11
Construction and Maintenance of Heterogeneous Data Warehouses

M. Badri
Crip5 Université Paris Descartes, France & Lipn Université Paris Nord, France

F. Boufarès
Lipn Université Paris Nord, France

S. Hamdoun
Lipn Université Paris Nord, France

V. Heiwy
Crip5 Université Paris Descartes, France

K. Lellahi
Lipn Université Paris Nord, France

ABSTRACT

The data necessary to decisional ends are increasingly complex. They have heterogeneous formats and come from distributed sources. They can be classified in three categories: the structured data, the semi-structured data and unstructured data. In this work, we are interested in the field of data integration with the aim of construction and maintenance of warehouses whose sources are completely heterogeneous and belonging to the various categories. We propose a formal framework based on the definition of an integration environment. A set of “integration relationships” between the components of the sources is thus defined: an equivalence relation and a strict order relation. These relationships are independent of any data sources modelling. These last can be then heterogeneous and having different models and/or categories. Two different physical architectures, to create and maintain the warehouses and the materialized views, are given.

DOI: 10.4018/978-1-60566-756-0.ch011
INTRODUCTION

Nowadays, the current informational environment is characterized by strongly distributed heterogeneous data. Complex applications such as knowledge extraction, data mining, learning and web applications use heterogeneous and distributed data sources (Boussaïd, Darmont, Bentayeb & Loudcher-Rabaseda, 2008). Thus the need of integrating and manipulating of large amount of data is more and more increasing. In the absence, first, of tools for the heterogeneous data integration, and second, of formalisms for modelling the integration of these data, we propose in this chapter early attempts to formalise the integration of heterogeneous data and their maintenance. Indeed, the data can be classified in three categories: structured (relational and object data), semi-structured (HTML, XML, graphs) and unstructured (text, images, sounds) (see figure 1).

Our contribution is twofold: the first part of our work concerns the beginning of the Data Warehouse (DW) life cycle: the building of DW from heterogeneous sources and the second part is related to the maintenance phase.

In the literature (Da Silva, Filha, Laender & Embley, 2002; Saccol & Heuser, 2002; Kim & Park, 2003; Beneventano, Bergamaschi, Castano, Antonellis, Ferrara, Guerra, Mandreoli, Ornetti & Vincini, 2002), there is no consensus on the significance of heterogeneity. According to the domain and the type of considered application, the treatment and the interpretation of heterogeneity were made in several ways. Considering this ambiguous interpretation of heterogeneity, we adopt, in our integration work, the definitions below which will enable us to treat all categories together.

Data sources are known as heterogeneous if they check one of the two following properties:

1) They belong to the same category of data but they have different modellings;
2) They belong to different data categories.

Thus, the integration of a relational database (DB) and of an object-relational one is an example of handling heterogeneous data sources. It is the same case for a relational and an XML DBs.

A DW results from data sources integration. It is a subject-oriented, integrated, time-variant,