7 Metadata and Data Warehouse Quality

In the traditional view, data warehouses provide large-scale caches of historic data. They sit between information sources gained externally or through online transaction processing systems (OLTP), and decision support or data mining queries following the vision of online analytic processing (OLAP). Three main arguments have been put forward in favor of this caching approach:

1. **Performance and safety considerations**: The concurrency control methods of most DBMS do not react well to a mix of short update transactions (as in OLTP) and OLAP queries that typically search a large portion of the database. Moreover, the OLTP systems are often critical for the operation of the organization and must not be in danger of corruption by other applications.

2. **Logical interpretability problems**: Inspired by the success of spreadsheet techniques, OLAP users tend to think in terms of highly structured multidimensional data models, whereas information sources offer at best relational, often just semi-structured data models.

3. **Temporal and granularity mismatch**: OLTP systems focus on current operational support in great detail, whereas OLAP often considers historical developments at a somewhat less detailed granularity.

Thus, quality considerations have accompanied data warehouse research from the beginning. As shown in the previous chapters of this book, a large body of practical experience and research literature has evolved over the past few years in addressing the problems introduced by the DW approach, such as the trade-off between freshness of DW data and disturbance of OLTP work during data extraction; the minimization of data transfer through incremental view maintenance; and a theory of computation with multidimensional data models.

However, the heavy use of highly qualified consultants in data warehouse applications indicates that we are far from a systematic understanding and usage of the interplay between quality factors and design options in data warehousing. The goal of the European DWQ project [JaVa97] is to address these issues by developing, prototyping and evaluating comprehensive Foundations for Data Warehouse Quality, delivered through enriched metadata management facilities in which specific analysis and optimization techniques are embedded.

After giving a short overview of the state of the practice in handling data warehouse quality, this chapter further develops the DWQ architecture and quality management framework introduced in chapter 2.7, and links it to other work on data and software quality.
7.1 Metadata Management in Data Warehouse Practice

Metadata play an important role in data warehousing. Before a data warehouse can be accessed efficiently, it is necessary to understand what data is available in the warehouse and where is the data located. In addition to locating the data that the end users require, the metadata may contain [AdCo97, MStr95, Micr96]:

- data dictionary: contains definitions of the databases being maintained and the relationships between data elements
- data flow: direction and frequency of data feed
- data transformation: transformations required when data is moved
- version control: changes to metadata are stored
- data usage statistics: a profile of data in the warehouse
- alias information: alias names for a field
- security: who is allowed to access the data

Metadata is stored in a repository, where it can be accessed from every component of the data warehouse. Because metadata is used and provided by all components of the warehouse, a standard interchange format for metadata is necessary. The Metadata Coalition has proposed a Metadata Interchange Specification [MeCo96]; additional emphasis has been placed on this area through the recent efforts of Microsoft to introduce a repository product in their Office tool suite, including some simple Information Models for data warehousing [BBC*99]. In addition, a number of meta database systems developed in research have been successfully used in industry. In the following subsections, three such approaches are described.

![MDIS metamodel](MeCo96)