Statistical Metadata Modeling and Transformations

Maria Vardaki
University of Athens, Greece

INTRODUCTION

The term metadata is frequently used in many different sciences. Statistical metadata generally used to denote "*every piece of information required by a data user to properly understand and use statistical data*." Modern statistical information systems (SIS) use metadata in relational or complex object-oriented metadata models, making an extensive and active usage of metadata. Early phases of many software development projects emphasize the design of a conceptual data/metadata model. Such a design can be detailed into a logical data/metadata model. In later stages, this model may be translated into physical data/metadata model.

Organisations aspects, user requirements and constraints created by existing data warehouse architecture lead to a conceptual architecture for metadata management, based on a common, semantically rich, object-oriented data/metadata model, integrating the main steps of data processing and covering all aspects of data warehousing (Pool et al, 2002).

In this paper we examine data/metadata modeling according to the techniques and paradigms used for metadata schemas development. However, only the integration of a model into a SIS is not sufficient for automatic manipulation of related datasets and quality assurance, if not accompanied by certain operators/ transformations. Two types of transformations can be considered: (i) the ones used to alleviate breaks in the time series and (ii) a set of model-integrated operators for automating data/metadata management and minimizing human errors. This latter category is extensively discussed.

Finally, we illustrate the applicability of our scientific framework in the area of Biomedical statistics.

BACKGROUND

Metadata and metainformation are two terms widely used interchangeably in various sciences and contexts. Until recently, metainformation was usually held as table footnotes. This was mainly due to the fact that the data producer and/or consumer had underestimated the importance of this kind of information.

When metadata integration in a pre-arranged format became evident, the use of metadata templates was proposed. This was the first attempt to capture metadata in a structured way. This approach was soon adopted since it reduced chances of ambiguous metadata as each field of the templates was well documented. However, they still had limited semantic power, as they cannot express the semantic links between the various pieces of metainformation.

To further increase the benefits of using metadata, attempts have been made to establish ways of automating the processing of statistical data. The main idea behind this task is to translate the meaning of data in a computer-understandable form. A way of achieving this goal is by using large, semantically rich, statistical data/metadata models like the ones developed in Papageorgiou et al (2001, 2002). However, in order to minimize compatibility problems between dispersed systems, the need that emerges is to build an integrated metadata model to manage data in all stages of information processing. The quantifiable benefits that have been proven through the integration of data mining with current information systems can be greatly increased, if such an integrated model is implemented. This is reinforced by the fact that both relational and OLAP technologies have tremendous capabilities for navigating massive data warehouses; nevertheless, brute force navigation of data is not enough. Such an integrated model was developed in Vardaki & Papageorgiou (2004), and it was demonstrated that an abstract, generally applied model, keeping information about storage and location of information as well as data processing steps, is essential for data mining requirements. Other related existing work focuses either mainly on OLAP databases (Pourabbas and Shoshani, 2006) or on semantically rich data models used mainly for data capturing purposes. In these cases, the authors focus their attentions on data

manipulations and maximization of the performance of data aggregations.

A number of modeling techniques and technologies have been discussed in literature and considered for the implementation of various research projects. Mainly in the case of medical statistics the Entity-Attribute-Value (EAV) (also known as "Object-Attribute-Value Model" and "Open Schema") database modeling technique has been extensively used (Dinu and Nadkarni, 2006). Also, Papageorgiou et al (2001) developed an Object Oriented metadata model for the series of processes followed in a Statistical Institute. Finally, the XML-based solution has been extensively used.

However, although a metadata model is an important step towards automation of data/metadata management and processing, the definition of a set of transformations/ operators is essential for harmonization purposes and for the automatic manipulations of statistical information. Two types of transformations can be considered: (i) the methodological transformations which are used when there are inconsistencies between practices of data collection, compilation and dissemination of statistics and (ii) operators that permit specific manipulations of the underlying data and their related metadata stored in the databases.

Regarding the methodological transformations, these are used to alleviate breaks in time series. When data collected in a specific time period are not fully comparable with the data of the following years we say that we have a break in time series. Breaks frequently occur in time series and involve changes in standards and methods that affect data comparability over time since they make data before and after the change not fully comparable. Information about breaks in time series is a quite important piece of statistical metadata because of the adverse effects they can have to statistical inference based on fragmented data. A number of transformations can be applied to minimize the effect of such incompatibility.

In case of simultaneous manipulation of both data and metadata, this is achieved by introducing a set of tools (transformations) to assist the data producer and user in manipulating both data and metadata simultaneously. Sets of such transformations have been discussed by Papageorgiou et al (2002) and Vardaki and Papageorgiou (2006).

MAIN THRUST

This chapter aims in discussing metadata modeling and related techniques and also introduce a set of underlying transformations. More specifically, topics that are covered include essential considerations in statistical metadata modeling development, modeling techniques and paradigms. Furthermore, a set of transformations is proposed for automation of data/metadata processing in a Statistical Information System.

The applicability of our scientific framework is further discussed in a case study in the area of medical statistics describing how the metadata model and the proposed transformations can allow for simultaneous handling of datasets collected during dispersed similar clinical trials performed by different medical centers.

Statistical Metadata Modeling

The design of a data/metadata model is a crucial task for further processing. If the model is undersized, it will be incapable of holding important metadata, thus, leading to problems due to missing metainformation. On the other hand, if it is oversized, it will keep information that is captured, rarely used and never updated, thus leading to severe waste of resources. Obviously, an oversized model cannot be easily implemented or used by the Institute's personnel. However, it is also difficult to predict the needs of data consumers, since the amount of required metainformation depends on the application under consideration. A step-by-step model development can serve the purpose.

We briefly consider three stages of data/metadata modeling:

- The conceptual metadata model (schema) consisting of entity classes (or objects), attributes and their relationships.
- The semantic data model describing the semantics.
- The operators/transformations abstract schema

Of course the database structure and coding as well as the data/metadata storage selection should be also represented.

Apart from choosing what metainformation is worth capturing, there is an additional difficulty in choosing the most appropriate modeling technique. Discussions