# Statistical Metadata in Data Processing and Interchange

**Maria Vardaki**
*University of Athens, Greece*

## INTRODUCTION

The term *metadata* is frequently considered in many different sciences. Statistical metadata is a term generally used to denote data about data. Modern statistical information systems (SIS) use metadata templates or complex object-oriented metadata models, making an extensive and active usage of metadata.

Complex metadata structures cannot be stored efficiently using metadata templates. Furthermore, templates do not provide the necessary infrastructure to support metadata reuse. On the other hand, the benefits of metadata management depend also on software infrastructure for extracting, integrating, storing, and delivering metadata.

Organizations 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).

## BACKGROUND

Metadata and metainformation are two terms widely used interchangeably in many different sciences and contexts. In all those cases, these terms are defined as data about data; that is, *metadata are every piece of information needed for someone to understand the meaning of data.*

Until recently, metainformation usually was held as table footnotes. This was due to the fact that the data producer and/or consumer had underestimated the importance of this kind of information.

When metadata consideration in a prearranged format became evident, the use of metadata templates was proposed. This was the first true attempt to capture metadata in a structured way. The advantage of this approach was reduced chances of having ambiguous metadata, as each field of the templates was well documented. Templates succeed in capturing metadata in a structured way. However, they have limited semantic power, as they cannot natively express the semantic links between the various pieces of metainformation.

To capture the semantics of metainformation, a metadata model must be used. In this case, metainformation is modeled as a set of entities, each having a set of attributes. The real advantage comes from the fact that these entities are interrelated. This enables the user to follow a navigation-style browsing in addition to the traditionally used, label-based search.

Froeschl (1997) created an object-oriented model for storing and manipulating metadata. A number of European projects deals with metadata models development and their subsequent integration into statistical information systems. Currently, automated statistical information systems allow for complex data aggregations, yet they provide no assistance in metadata manipulation.

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., (2001a, 2001b, 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 usage in all stages of information processing. The quantifiable benefits that have been proven through the integration of data mining with current information systems will 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, but brute force navigation of data is not enough.

Such an integrated model was developed in Vardaki & Papageorgiou (2004), and it was demonstrated that such a generally applied model, keeping information about storage and location of information as well as data processing steps, was essential for data mining requirements.

Other related existing work focuses either mainly on data operations (Denk et al., 2002) and OLAP databases (Scotney et al., 2002; Shoshani, 2003) or on semantically rich data models used mainly for data capturing purposes. In these cases, the authors focus their attention on data

manipulations and maximization of the performance of data aggregations.

## MAIN THRUST

This paper aims to summarize some of the latest results of research in the area of metadata. Topics that are covered include a possible categorization of statistical metadata, the benefits of using structured metainformation, standardization, metadata databases, modeling of metainformation, and integration of metadata in statistical information systems.

### Types of Metadata

In the literature, a number of categories has been proposed to classify metainformation according to different criteria. The following division in four overlapping categories (Papageorgiou et al., 2000) is proposed, since the partitioning criterion is the role that metainformation plays during the life cycle of a survey.

- **Semantic Metadata:** These are the metadata that give the meaning of the data. Examples of semantic metadata are the sampling population used, the variables measured, the nomenclatures used, and so forth.
- **Documentation Metadata:** This is mainly text-based metainformation (e.g., labels), which are used in the presentation of the data. Documentation metadata are useful for creating user-friendly interfaces, since semantic metadata are usually too complex to be presented to the user. Usually, an overlap between the semantic and documentation metadata occurs.
- **Logistic Metadata:** These are miscellaneous metadata used for manipulating the data sets. Examples of logistic metadata are the data's URL, the type of RDBMS used, the format and version of the used files, and so forth. Mismatches in logistic metadata are easily discovered, since the used information tools immediately produce error messages. However, many times, logistic metadata can be corrected only by specialized personnel.
- **ProcessMetadata:** Process metadata are the metadata used by information systems to support metadata-guided statistical processing. These metadata are transparent to the data consumer and are used in data and metadata transformations.

### Benefits of Using Metadata

Even though competition requires timely and sophisticated analysis on an integrated view of the data, there is a growing gap between more powerful storage and retrieval systems and the users' ability to effectively analyze and act on the information they contain.

The benefits of using metadata are several. Some of the most important can be summarized as follows: By capturing metadata in a structured way and providing a transformations framework, computers are enabled to process metadata and data at the same time. Thus, the possibility of human errors is minimized, since user intervention is generally not necessary. Furthermore, the possibility of errors is reduced by the fact that metadata can be used by computers for asserting data manipulations. For example, a metadata-enabled statistical software can warn the user of a possible error when adding two columns that use different measure units. Finally, errors due to misunderstanding of footnotes are eliminated, since structured metadata are unambiguously defined (Foeschl, 1997). Hence, it is easy to show that metadata are important for assuring high levels of data quality at a low cost. However, it should be noted that the benefits of using metadata are subject to the quality of the metadata.

### Metadata Standards Affecting Quality of Results

During the design of a survey, the statistician implicitly produces metainformation. Usually, for small non-periodic surveys, the statistician might choose to use an ad hoc solution. However, for large periodic surveys, the statistician definitely will follow a standard. Depending on the authority describing a standard, we can identify three types of metadata standards:

- **The Ad Hoc (Internal) Standards:** These are defined internally by each statistical office. Due to the versatility of a small statistical office, these standards are highly adaptive to the latest needs of the data consumers. However, the compatibility of an internal standard with respect to an internal standard of a different office is not guaranteed.
- **National Standards:** These are defined by the National Statistical Institutes of each country. Although they may not be as current as their respective internal statistical standards, they offer statistical data compatibility at country level, which is the level that interests mostly the data consumers.
- **International Standards:** These might be nomenclatures or classifications that are defined by supranational organizations such as OECD and Eurostat. The usage of international standards provides the maximum intercountry compatibility for the captured data. However, the approval of an international standard is a time-consuming process. In any