

Distr.
GENERAL

CES/SEM.47/24
9 January 2002

ENGLISH ONLY

**STATISTICAL COMMISSION and
ECONOMIC COMMISSION FOR EUROPE**

**COMMISSION OF THE
EUROPEAN COMMUNITIES**

CONFERENCE OF EUROPEAN STATISTICIANS

EUROSTAT

**Joint UNECE/Eurostat Seminar on Integrated Statistical
Information Systems and Related Matters (ISIS 2002)**

(17-19 April 2002, Geneva, Switzerland)

Topic IV: Ways of making statistical information systems more responsive to users

**THE USE OF STATISTICAL METADATA MODELLING AND RELATED TRANSFORMATIONS
TO ASSESS THE QUALITY OF STATISTICAL REPORTS**

Invited paper

Submitted by the University of Athens¹

I. INTRODUCTION

1. The amount of information processed and disseminated by National Statistical Institutes (NSIs) in the various statistical reports is constantly growing, as demands for well-timed data, of high quality are increasing. Currently, NSIs are facing a dual task: i) to embody harmonisation and transformation procedures in their workflow processing, since methodological peculiarities in the way of collecting, storing and disseminating information may lead to inconsistencies of their statistical results; ii) to upgrade their infrastructure with new metainformation systems, aiming in increasing the quality of the provided services.

2. Concerning the harmonisation problem, 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. Breaks can appear in space as well as in time. It is becoming ever more common to compare data and indicators between countries by forming tables with a space dimension. This is also becoming more important for shaping policy in the face of growing international co-operation in the era of globalisation.

¹ Prepared by H. Papageorgiou and M. Vardaki, University of Athens, Department of Mathematics, e-mail: {hpapageo, mvardaki}@cc.uoa.gr; E. Theodorou and F. Pentaris, University of Athens, Department of Informatics, e-mail: {i.theodorou, frank}@di.uoa.gr.

3. In the second case, NSIs are forced to look for different ways of extending the automation of their internal procedures. This attempt is influenced by the use of *Internet*, as well as by the construction of new *metadata-enabled* Statistical Information Systems (SIS). Unfortunately, most of these systems treat metadata as plain collections of text that are only used for documentation reasons and for locating specific data. This *passive* use of metadata ignores the existence of process metadata [4], thus reducing the advantages of using meta-information. Advanced systems use process metadata for automating the statistical processing of information so that users need only to describe the statistical table they are interested in and not how to construct it. The important feature of these systems is that the resulting table is also *automatically* accompanied by the appropriate documentation (metadata), which in turn, greatly improves the quality of the produced tables and statistical reports by reducing the dangers of data and metadata mismatches [9].

4. A prerequisite for building such metadata-enabled systems is the modeling of data and metadata [2],[10], [11], [13], using a semantically rich data model and subsequently, the definition of a set of operators, that are used for manipulating the stored (meta)information.

5. In this paper we primarily describe in detail a statistical data and metadata model, based on the results of the IPIS project [3]. The presented model is general enough to support capturing ad hoc statistics ranging from micro data to official statistics' indicators, thus making it perfectly suitable for metadata harmonization procedures. We discuss the semantics of the model and subsequently examine a set of seven operators that can be used for the *simultaneous manipulation* of both data and statistical metadata. Then, we illustrate how breaks in time series and in space can be treated by discussing *mapping* [11] and *methodology-correcting transformations* [12] to alleviate the breaks in time series under consideration. Finally, we demonstrate the importance of the model and its operators by presenting, as a case study, of a metadata-enabled website.

II. THE DATA AND METADATA MODEL

6. The design of a data and metadata model is the most important step in the creation of a SIS. If the model is undersized, it will be incapable of holding important metadata, thus, leading to problems due to missing meta-information. 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 human resources. Obviously, an oversized model is difficult to be implemented or used by the NSI's personnel. However, it is difficult to predict the needs of data consumers, as the amount of required meta-information depends on the application under consideration [16]. In any case, a metadata model should at least capture a minimum set of *semantic*, *documentation*, *logistic* and *process* metadata [13]. Some relevant metadata categories have been described in [4].

7. Apart from choosing what meta-information is worth capturing, there is an additional difficulty in deciding on the most appropriate modeling technique. Knowledge gained from previous projects directed us to design our model using an Object-Oriented (OO) approach (UML diagrams [6]) since it allows for class inheritance and class operations, thus making the model more flexible, extensible and expressive. Figure 1 (page 11) gives an overview of the IPIS model (see also [10], [11]). It is based on a standard metadata list published by OECD [15] and contains more than thirty different classes, which together are used to describe a single statistical table (macrodata) produced from a typical survey.

8. In our model, the central class is **SURVEY (TABLE)**. It is related with a **STATISTICAL POPULATION** that holds the abstract definition of the set of units that must be examined. This set of units is either described using a **MASTER LIST** or it is a subset of a larger **STATISTICAL POPULATION**. In the second case, the two relevant **STATISTICAL POPULATIONS** are related with each other through a **CONDITION**. Note, that usually it is not desirable or feasible to examine every unit included in the **STATISTICAL**

POPULATION, so a **SAMPLING METHOD** is applied to derive a smaller, but representative enough, set of units called **SAMPLE**.

9. The statistical service responsible for the **SURVEY** gathers the elements of the **SAMPLE** (*microdata*) and stores them in electronic form. Thus, **SURVEY** is related with the classes **COLLECTION INFO** and **DATA STORAGE**. The first one keeps information about the questionnaire, the non-response rate as well as the reporting date and method. The second one is divided into the classes **RDBMS STORAGE** and **FILE STORAGE**. In the case of **RDBMS STORAGE**, we need to know about the **TABLES**, the **COLUMNS**, and the kind and location of the **RDBMS** used, whereas, in the case of **FILE STORAGE**, we need additional information about the name, type, format, location and URL of the files. A **SURVEY** has a set of **VARIABLES** and each of them takes values from a **GROUPING LEVEL**, consisting of **MEASURE UNITS**. A **MEASURE UNIT** is either an **ATOM VALUE** (either nominal or ordinal), or a **REAL NUMBER**. It is related with other **MEASURE UNITS** through the class **EQUALS**, which specifies the kind of this relation, e.g. contains, equivalent with, etc. An ordered set of **GROUPING LEVELS** is a **CLASSIFICATION**. One or more **VARIABLES** are used to calculate an **INDICATOR**, which is made to reflect the value of the whole population through a **GROSSING-UP** method. Finally, the used **CLASSIFICATIONS** and **INDICATORS** may comply with certain **STANDARDS**.

10. Moreover, a **SURVEY** is related to one or more **SOURCE AGENCIES** that may provide **MULTIPLE SOURCE ITEMS**, corresponding to either an **ADMINISTRATIVE SOURCE ITEMS** or a second **SURVEY**. There are two kinds of **SOURCE AGENCIES**: a source agency for compilation of statistics and a source agency for data collection. Additionally, a **SURVEY** may have **ADJUSTMENTS**, e.g. seasonal, when the results of a survey depend on the specific time period when the survey takes place.

11. Finally, the model holds information about the **DATA QUALITY AND TIMELINESS** of the data, as well as about the **ERRORS** that might have been detected and corrected (**CORRECTIONS**).

III. OPERATORS THE MODEL

12. The presented model keeps information about the series of processes that have been applied on the data of a survey through the class **OPERATIONS** and its subclasses **SELECT**, **PROJECT**, **CONCATENATE**, **GROUP BY**, **RECLASSIFY**, **JOIN**, and **ALGEBRAIC**. Each of these subclasses (Figure 2) implements one of the *operators* described later on. The existence of these classes is extremely important as they hold the processing history of the table, which is very useful to end-users for proper evaluation of data quality.

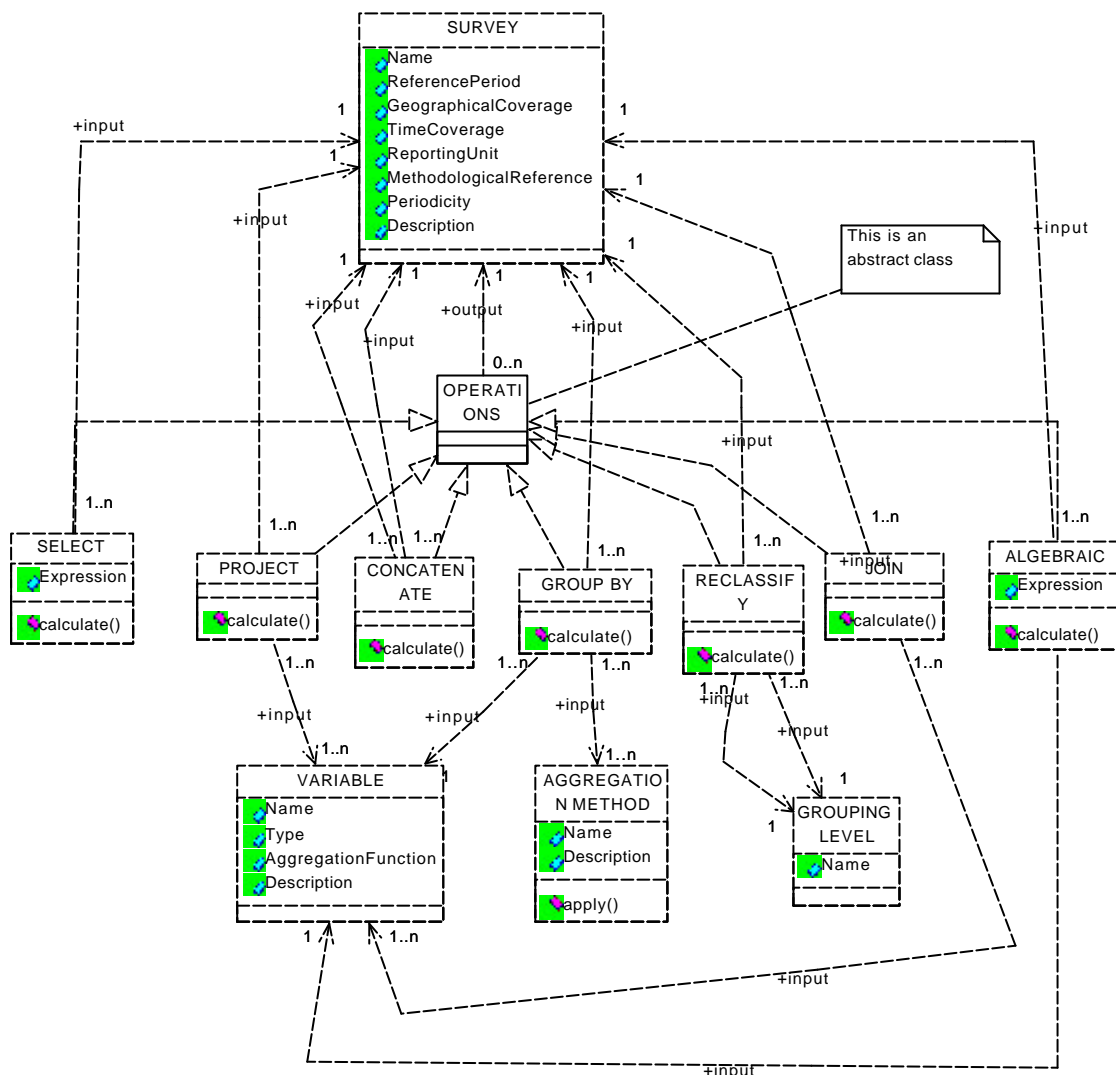


Figure 2: The model operators

13. All the operations of our data/metadata model have the closure property, meaning that the application of an operator on a TABLE always produces a new TABLE. The definition of each operator contains three parts that have been defined and illustrated in [10], [11].

- ? **Pre-conditions.** Each operator must satisfy a set of rules (preconditions) to be valid. The preconditions are used by an automated system to decide whether an operator can be applied or not. In this way, many errors can be avoided.
- ? **Processing instructions.** These are a description of the physical processing that must be carried out for evaluating the transformation.
- ? **Post-conditions.** These are a description of the properties of the operators' results. Someone can use these post-conditions to decide whether the operator produces the desirable results.

14. The importance of operators is that they allow for simultaneous manipulation of both data² and metadata³. This means that the result of an operation is a new, automatically documented TABLE. Consequently, they severely reduce the possibility of data and metadata mismatches. In addition, operations are necessary for building SISs that support metadata-guided statistical processing. In this case, the user simply describes the statistical Table, which interests him and the system automatically finds a series of operators that can be applied onto existing tables in order to produce the requested one.

15. A brief description of each operator is as follows:

Selection: The selection operator closely resembles the one from the relational algebra. The result of applying this operator is a new table holding only a subset of the initial data satisfying the selection criterion. This operator has no pre-conditions.

Projection: The projection operator like the selection, closely resembles the one from the relational algebra. The result of applying this operator is a new table holding only a subset of VARIABLES and INDICATORS of the initial data. Projection also has no pre-conditions.

Concatenation: The result of applying the concatenation operator on two tables is a new table combining the data of the initial tables. This operator has the following pre-conditions in order to be applicable:

- Both tables must have equivalent sets of variables measured with the same MEASUREUNITS, otherwise, the resulting table will have many null/unknown values.
- Both tables must refer to the same kind of units.
- The intersection of the two STATISTICALPOPULATIONS must be empty (void). This last pre-condition is needed to ensure that the resulting table does not incorrectly contain duplicates that may lead to biased results.

Group by: The GroupBy operator is used for creating new INDICATORS from existing data. The new INDICATOR is calculated after grouping (partitioning) the values of a Table according to the distinct values of one or more existing VARIABLES denoted as grouping variables. The derived indicator is calculated by applying an aggregation function onto the values of a subset of the Tables' VARIABLES. The only prerequisite of applying a GroupBy operator is that if the Table already has an INDICATOR, then the grouping variables of the operator must be the same as the grouping variables of the existing INDICATORS. The post-condition of this operator is that the resulting Table contains the *Grouping Variables* the previously existing INDICATORS and, obviously, the INDICATOR that was calculated with the GroupBy operator.

Reclassification: The reclassification operator converts the values of a VARIABLE or INDICATOR from one GROUPINGLEVEL to a different one. The reclassification operator poses no constraints on the target GROUPINGLEVEL. Therefore, the results of a reclassification may lead to tables with missing values, or with inaccurate values, especially when the existing GROUPINGLEVEL is not fully convertible into the new one (for example, converting from NABS [1] classification into the OECD [1] classification).

Join: The join operator resembles the join operator of the relational algebra. It is applied on two tables having one or more common variables (joined VARIABLES). The result is a new table having all the VARIABLES and INDICATORS of both tables (obviously, the common VARIABLES and INDICATORS are included once.). Since the result of the operator is a new Table, the INDICATORS of this table must all have the same grouping variables. Consequently, the join operator has at least the following prerequisites:

² This happens during the physical processing of the transformations

³ This happens during the evaluation of pre- and post-conditions

- The INDICATORS of the joined Tables must have the same grouping VARIABLES measured with the same GROUPINGLEVELS.
- If grouping VARIABLES' set is not empty (i.e the joined tables have INDICATORS), then the joined variables must be equal to grouping VARIABLES.

Algebraic Operators: This is a general operator used to denote simple mathematical operations (e.g. additions, multiplications, etc.) that are frequently applied into a TABLE. The algebraic operator is the only one, which requires additional documentation and semantic metadata to be specified by the user, during its application. The reason is that there is no automated way for SISs to understand the semantic meaning of such an operator.

IV. BREAKS IN TIME SERIES DATA

16. In the time-period for which a specific statistical data series is available there can be several instances where methodologies have been revised. Although this tends to result in more accurate or cost-effective data, it generates breaks in times series. Consequently, data before and after the revision are not fully comparable. In addition, the existence of a variety of International and National classifications and nomenclatures as well as their revisions affects the possibility of comparability of data collected and compared between countries and through the years. *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* [7],[14]. These breaks can be **spatial**, when there is a difficulty in comparing data from different countries that use various nomenclatures and **inter-temporal** when revisions of nomenclatures are introduced in specific years at the same country. Therefore, comparisons need to become even more frequent to increase international cooperation and to harmonise practices or, failing that, to be able to ascertain the level of fragmentation of a statistical table synthesized from data supplied from different NSIs.

17. Using the structured approach of OECD's standard list of metadata items and the possibility of applying the afore-mentioned operators, we have proceeded in grouping possible alleviations of breaks in time series due to methodological peculiarities by applying i) *mapping transformations* that are related to different classifications adopted by various countries or to revisions of the same classification and ii) to specific *methodology-correcting* ones developing a two level taxonomy of breaks according to the aspect of the statistical methodology that creates it.

A. Mapping Transformations

18. These transformations include all transformations that are required in order to convert data collected under previous classifications into the ones used nowadays. The definition of such functions demonstrates that it is possible via mapping transformations to convert existing data into a different format. When applying *mapping transformations* at least the following cases have to be considered:

- *the introduction of new international nomenclature in the same area of application*
- *a revision of the existing one*
- *the introduction of an international nomenclature in a different area of application*
- *the introduction of national nomenclatures that may be difficult to discern if they are related or derived to a reference one* [14].

19. It is essential to define a classification that will serve as a target one for comparisons and conversions of other classifications into it (our Quality Frame, QF). For the specificities of the IPIS project we have selected NACE Rev1 (General Industrial Classification of Economic Activities within the European Communities) [1] as our QF. In this case, the breaks in time series data that can be alleviated via *mapping transformations* are classified into four main groups according to the structure of the nomenclature that must be converted:

Group 1: it includes conversions of nomenclatures produced by adopting NACE Rev 1 classification's structure and categories, and then possibly providing additional detail (like NAF [1]) or through rearrangement or aggregation of items from one or more reference classifications.

Group 2: includes conversions of nomenclatures that provide a set of organised categories for the same variables as NACE Rev1, but for which the categories may only partially refer to those defined in NACE Rev1, or that may only be associated to it at specific levels of structure (i.e. NAICS [1]).

Group 3: includes transformations of revisions of already existing nomenclatures. This group is related to the previous ones for their second step of harmonisation.

Group 4: concerns mapping transformations of nomenclatures where statistical units are different than those of NACE Rev1 (i.e. CPC [1]).

20. There are conditions that have to be considered, as for example the possibility when the `GROUPINGLEVEL` under conversion is not fully convertible with the target one. Take for example NACE Rev1 as QF and CPA as the classification under conversion. Then the corresponding Variables for a related table classifying amounts allocated under CPA according to Country and Year are: Year, Country and Product/ Service Type.

21. Consequently, in order to enable mapping procedures, the *Classification* object has to be related to specific characteristics (variables) that will permit transformations between them and therefore, corresponding metadata items should be captured [5], [14]. In addition, it should also be considered that each statistical unit of a classification is a specific entity defined in such a way that it can be recognised, identified and not confused with any other unit. Classification of statistical units has been described in the Council Regulation on Statistical Units⁴, in order to allow for comparability between classifications. Of course, in order to compare a source classification with the pre-defined QF, a number of preconditions need to be fulfilled (see for more information [14]).

B. Methodology-Correcting Transformations

22. The *methodology correcting transformations* include all the possible transformations that are used to enhance the data homogeneity by correcting errors that were introduced due to methodological inconsistencies and restore their effects in breaks in time series. In order to evaluate fragmentation of data in both time and space, several coefficients of fragmentation in both space and time can be developed (see for more information [12]).

23. Evidently, methodological differences might not affect statistical data substantially. The first step in the identification of the problem is the verification whether these peculiarities do indeed cause a problem. A function g is therefore introduced for a methodological difference.

$$g(v) = \begin{cases} 1 & \text{substantial break} \\ 0 & \text{no substantial break} \end{cases}$$

24. For any number of differences in a data series either in time or in space a first descriptor will naturally be:

⁴ Council Regulation (EEC) No 696/93 of 15 March 1993 (OJ No L 76, 30.3.1993, p.1).

$$p ? \frac{\text{Number of substantial breaks}}{\text{total Number of breaks}}$$

25. Furthermore, an evaluation of the severity of the break has to be performed. This can range from small variations in the error of measurement to disrupting changes in variable definitions. We therefore weight the breaks using three ordered levels, low medium and high severity [12].

26. Last but not least, it is essential to characterize breaks according to kind. We assume that the data provider has accurately described changes in statistical methodologies. This can be performed in two major directions, firstly the reason of the break, that is, the part of methodology that creates it (i.e. data collection method or measurement unit used) and secondly the severity of the problem (i.e. if the difference creates a problem at all and how significant it is). For the specificities of our metadata model, we adopted a two-digit classification of breaks in order to classify breaks in time series reported to the OECD [12].

V. CASE STUDY

27. Our metadata model designates what metadata should be kept, how they are related with each other, and defines what kind of manipulations can be executed on them. Therefore, the metadata model affects both storage and execution levels –the two lower tiers in a 3-tier architecture⁵ (see Figure 3). In order to prove the efficiency of this model, a case study is presented below.

28. Assume that a user would like to access a table that has aggregated data about unemployment in the EU-15 during 2000 *grouped by age, sex, occupation and geographical area*. Also, assume that the NUTS classification is used for geographical area and the ISCO '88 [1] for occupations. Then, the user can perform through an appropriate interface certain manipulations on this table with the help of *operators*:

- a) **Selection:** the user can pose a criterion on the variable 'geographical area'. For example, from the countries in EU-15, only data concerning Greece, Italy, Spain and Portugal will be displayed.
- b) **Projection:** the user may want to see data independently from the variable sex, thus he chooses to project all variables except for sex.
- c) **Concatenation:** another useful operation is to append to this table the respective one concerning the same countries during 1999, if available, in order to easily compare the relevant surveys.
- d) **Grouping by:** in our example, data about unemployment are already grouped by the variables age, sex, occupation and geographical area.

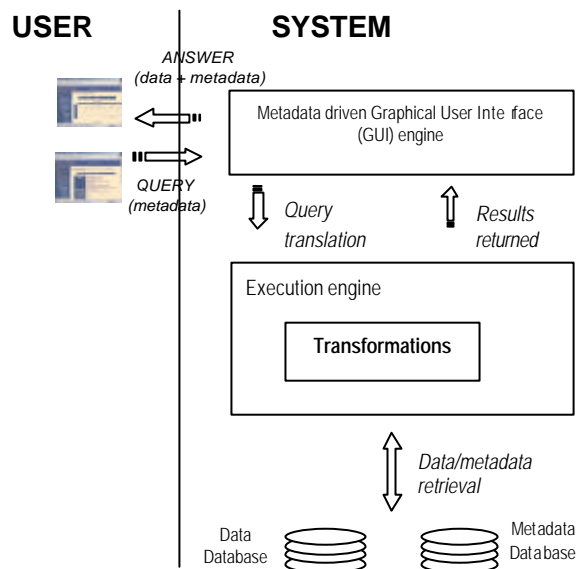


Figure 3 The architecture of a metadata-enabled web site.

⁵ The corresponding tiers are: data tier, execution tier and presentation tier

- e) Reclassification: the user may want to see the values of occupation according to the ISCO '68 classification instead of ISCO '88 for several reasons. The result will be calculated according to the correspondence tables available in the system.

Remark: This example of Reclassification operator can also be considered as a case of mapping transformation between ISCO-88 (source classification) and ISCO-68 (target one).

- f) Join: the table can be joined, for example, with the one containing data about job vacancies for the same countries. Thus, a user can visualize the relation of the two indicators (unemployment and job vacancies) for each country.
- g) Algebraic: this operator is used to calculate simple mathematical operations that can be applied on data. For example, if the system does not “know” how to convert a unit of measurement into another, the user can give the specific algebraic formula so that the conversion can be calculated.

29. As it can be concluded, the user is more “active” as he declares somehow the operations or the combinations of operations he wants by himself and gets back a quick answer. Thus, the metadata model with an appropriate user-friendly interface can make a system tool more responsive, while it offers more user-oriented services.

VI. CONCLUSIONS

30. We have described in detail a semantically rich metadata model and discussed the properties of seven important operators. The proposed framework is essential for designing general statistical information systems supporting automated, metadata-guided processing, which is vital for the next generation of statistical web sites as well as for asserting data quality.

31. Future plans include research on data quality issues arising from the use of operators, as well as relating the proposed model with relevant data warehousing and OLAP technologies. Additionally, it appears that the quality of TABLES that can be created using more than one operators' plan may vary, due to the initial missing values and the possibly biased reclassifications that may be included in the plan. It seems that up to now, this problem has not been properly investigated. Therefore, the automation of statistical processing calls for the derivation of some quality metrics that will subsequently be used during the plan optimization and selection stage

References

- [1] **EUROSTAT (1999)**, “*Inventory of International Statistical Classifications*”, Statistical Office of the European Communities, Luxembourg, ISBN 92-828-8204-7.
- [2] **Grossmann, W. (1999)**. “*Metadata*”, Encyclopedia of Statistical Sciences, update Vol. 3, pp. 811-815, 1999, S. Kotz, Editor-in-Chief, John Wiley and Sons, New York.
- [3] **IPIS IST project (2000)**. <http://www.instore.gr/ipis>
- [4] **Kent, J-P. & Schuerhoff, M. (1997)**. “*Some Thoughts About a Metadata Management System*”, in Proc. of the Ninth International Conference on Scientific and Statistical Database Management, Olympia, Washington, USA, pp. 174-185, IEEE Computer Society.
- [5] **Neuchatel Group (2000)**, “*The Neuchatel Terminology: Classification database object types and their attribute*”, UN/ECE Work Session on Statistical Metadata, Working Paper no 10, Geneva, Switzerland.
- [6] **OMG (1999)**. OMG Unified Language Specification, Object Management Group (OMG) Inc., available at <http://www.omg.org>

- [7] **Papageorgiou H., Artikis G. & Vardaki M. (1999).** “*On updating highly aggregated economic time series*”, *Wiadomosci Statystyczne*, 46 (4), pp. 4-11.
- [8] **Papageorgiou, H., Vardaki, M. & Pentaris, F. (2000a).** “Recent advances on metadata”, *Computational Statistics*, Vol.15-1, pp. 89-97.
- [9] **Papageorgiou, H. Vardaki, M. & Pentaris, F. (2000b).** “*Quality of Statistical Metadata*”, *Research in Official Statistics*, Vol. 2 No1, pp. 45-57.
- [10] **Papageorgiou, H., Pentaris, F., Theodorou E., Vardaki M. and Petrakos M. (2001a),** “*Modelling Statistical Metadata*”, *Proceedings of the Thirteenth International Conference on Scientific and Statistical Database Management (SSDBM)*, Virginia, USA, pp.25-35, IEEE Computer Society.
- [11] **Papageorgiou, H., Pentaris, F., Theodorou E., Vardaki M. and Petrakos M. (2001b),** “*A statistical metadata model for simultaneous manipulation of data and metadata*”, *Journal of Intelligent Information Systems (JIIS)*, Vol 17, No 2/3, pp. 169-192.
- [12] **Papageorgiou H., Petrakos M., Vardaki M., Theodorou E. and Pentaris F. (2001c),** “*Metadata based Assessment of the level of fragmentation of Data Series and Multisource Statistical Tables*”, presented to the NTTS-ETK 2001 Conference, June 18-22, 2001, Crete, Greece Pre-Proceedings, (1), pp.263-272.
- [13] **Papageorgiou, H. Vardaki, M. & Pentaris, F. (2001d),** “*Data and Metadata Transformations*”, *Research in Official Statistics*, Vol.3, No2, pp. 27-43.
- [14] **Papageorgiou H., Vardaki M., Petrakos M., Theodorou E. and Pentaris F. (2001e),** “*Harmonisation of Economic Classifications and related Transformations*” presented NTTS-ETK 2001 Conference, June 18-22, 2001, Crete, Greece, Pre-Proceedings, (1), pp.345-354.
- [15] **Petit G., Beziz P., and van Eck R., OECD Directorate (1996).** “*List of Metadata Items for OECD’s Main Economic Indicators*”, *Statistical Commission and Economic Commission for Europe, Conference of European Statisticians.*
- [16] **Sundgren B. (1996).** “*Making Statistical Data More Available*”, *International Statistical Review*, Vol. 64, pp. 23-38.

