WP. 11
ENGLISH ONLY

**UNITED NATIONS STATISTICAL COMMISSION and ECONOMIC COMMISSION FOR EUROPE
CONFERENCE OF EUROPEAN STATISTICIANS**

**EUROPEAN COMMISSION
STATISTICAL OFFICE OF THE EUROPEAN COMMUNITIES (EUROSTAT)**

**ORGANISATION FOR ECONOMIC COOPERATION AND DEVELOPMENT (OECD)
STATISTICS DIRECTORATE**

Joint UNECE/Eurostat/OECD work session on statistical metadata (METIS)
(Geneva, 9-11 February 2004)

Topic (iii): Metadata models and terminology

METADATA MODELS IN SURVEY COMPUTING

Invited Paper

Submitted by University Vienna, Austria[1]

I. INTRODUCTION

1. In the last years there was growing interest in metadata inside computer science, mainly motivated by the widespread use of data warehouses, by the development of geographical information systems and by the intensified usage of resources from the Internet, which implied also new challenges for the traditional field of library sciences and archivists. Also inside statistics, where metadata have been studied for more than three decades, the interest in metadata has intensified for a number of reasons. Accelerated growth rate of statistical data production, new types of data production and a higher level of labour division in data processing are the most important reasons. In response to this increased demand on metadata in statistics EUROSTAT established the METANET project as part of the IST research program of the European Union. METANET was established as a network of excellence, coordinated by Joanne Lamb from the University of Edinburgh and offered the opportunity to bring together practitioners as well as researchers in the area of statistical metadata from Europe, USA and Canada. The activities of the network were organized in four work groups responsible for the following topics: Work Group 1, with Statistics Netherlands as main responsible partner, investigated technical aids for implementing metadata systems and exchanging metadata descriptions; The main responsible partner for Work Group 2 which considered harmonisation of metadata based on current practice in classification and definitions of metadata was University Vienna; Work Group 3 was lead by Statistics Sweden and contributed to best practices for migration in statistical offices; Work Group 4 was coordinated by Statistics Norway and considered adoption issues. An additional work group 5 on terminology was constituted during the activities of the network.

2. Inside Work Group 2 the problem of harmonisation was pursued from two different perspectives. The first one was the so called Terminology Model (cf. Working paper 12 of this session) and the second one was the so called UMAS model trying to establish a **U**nified **M**etadata **A**rchitecture for **S**tatistics. This presentation gives a short summary of the UMAS model and is organized as follows: Section 2 presents some examples of metadata usage in survey processing as motivation for the UMAS model, section III gives a short outline of the model and section IV considers some implications for representation of terminology.

[1] Prepared by Wilfried Grossmann wilfried.grossmann@univie.ac.at

II. MOTIVATION

3. Survey computing encompasses usually a number of different activities like sampling, data editing (data validation), weighting, or tabulation. All these activities have specific requirements on the (meta)information accompanying the data and produce different pieces of new metainformation. In order to make this statement more precise the following paragraphs discuss briefly some aspects of the above mentioned activities in survey computing.

4. Sampling is one of the standard procedures for obtaining raw datasets. It presupposes the availability of an existing sampling frame, which represents the underlying universe of interest. Such a sampling frame is not always at hand, hence it must be generated by processing activities. One method could be merging a number of existing registers representing subpopulations, another method may be selection of some units with specific properties from a sampling frame in order to avoid overcoverage. As a result of these processing activities we obtain a new object of the type population together with some structural properties like the selection probabilities for the units. From computer science point of view such sampling frames may be interpreted as datasets, but such an interpretation does not take into account a number of specific statistically important properties, for example the above mentioned issue of coverage.

5. Data cleaning and editing is another typical activity in survey computing. Basically one can formulate most types of data validation and data editing as processing activities for the value sets (or value domains) of variables of interest. These editing activities have been analysed in some detail by the INSPECTOR project within the EUROSTAT research program. In the most simple case validation deals with checking of the value sets (or value domains) of one variable, or with checking of logical conditions for the combined value sets for a number of variables. However, there exist much more complicated types of validation procedures using summary information of the already existing dataset or summaries of other datasets. Take as an example a validation rule for a survey about persons, which states that within a household there is only one person which is allowed to be head of the household. Checking such a condition depends essentially on the underlying data model. If we use a flat file for the person data it would imply that we have to summarise the data with respect to the partition defined by the households and check the condition that for each household the count of the variable 'head of household' is exactly one. The results of such editing procedures are twofold: On the one hand they define constraints on combined value domains within the conceptual domain to which these value sets belong, on the other hand they define quality measures for datasets within this conceptual domain. Note that such constraints for the value sets can be defined without usage of existing dataset. Moreover the set of constraints will be in most cases incomplete because definition of the constraints is done always relative to a predefined environment.

6. In case of weighting information about the structure of the underlying population and about the method of data production is required. The information about the underlying population may be given in different forms of datasets and limits to some extent possible weighting procedures. For example, in the classical case of adaptation of the sample to strata one needs data about the population available in one or more summary tables of sample sizes for these strata. In some cases of model assisted or model based weighting procedures the data about the population have to be available in detailed form as auxiliary variables for the statistical units of the register. Note that such data are only loosely coupled to the original dataset and may be used for many different samples. Also information about the sampling procedure may be available in different forms, for example as additional variables in the dataset, as selection probabilities for each unit or as sampling fractions for a number of subpopulations. The output of a weighting procedure may be an additional dataset containing different weighting schemes and should accompany the original dataset in all further processing activities.

7. Computation of statistical tables ranges from simple projections, based only on a limited amount of information about the underlying data, up to sophisticated procedures taking into account not only statistical information but also subject matter information about the dataset and the objects constituting the dataset. Take as an example the task of summarizing data from different sources based on more or less similar statistical units which show differences only with respect to some aspects. In order to make such data comparable one defines many times so called analytical units, i.e. we construct new units which allow better comparison of the data. The definition of such analytical units is rather independent from the existence of concrete datasets (although

the definition may be motivated by these datasets). In fact it can be interpreted as a processing activity for statistical units based on subject matter information which is captured in some auxiliary variables.

8. The examples above show that survey computing involves computational procedures not only for datasets but also for other objects of the statistical information system which have a designated meaning within the statistical system. Reuse of such procedures could be facilitated if we consider these objects not only in connection with the dataset, but as rather independent objects of interest. All these activities are based on a data representation of the objects and on appropriately structured information about the objects. Specification of the structure needs besides a description as standardized data element also a description at a more conceptual statistical level taking into account the designation of the objects inside a statistical information system. In particular, the examples point to a number of description elements which may be summarized as follows:

(i) We need description elements containing the concepts represented by the data;
(ii) We need a description how the different objects are related according to statistical methodology;
(iii) Concrete computational procedures have to take into account the specific format of realisation for the objects as physical datasets;
(iv) The description has to give also information about responsibility, access rights and manipulation rights for the objects;
(v) In order to facilitate the required flexibility of usage in survey computing we need a mechanism of loose coupling of the different objects;
(vi) Last but not least we have to take into account the information requirements of people outside the system.

9. The UMAS model developed within the activities of METANET Work Group 2 made an attempt to set up a descriptive metadata model meeting these requirements. It is built upon the synthetic classification approaches employed by librarians and archivists, notably recognising the flexibility and expressive power of the facet classification. Five facets called *structure*, *view*, *form*, *stage* and *function* were found to support the three semantic dimensions (operation, function, representation) of statistical metadata employed upon the network's inception: In particular, the stage and function facet correspond closely to the operational and functional semantics dimensions respectively, with the representational semantics dimension refined three-fold into structure, view and form facets.

III. SKETCH OF THE UMAS MODEL

A. The Structure Facet: Statistical Categories

10. The structure facet defines the carriers of information in statistics, so called *categories*, resembling in a natural way the basic setup for statistical analysis defined by event space and variables generated in agreement with the event structure. The main categories are *statistical unit*, *statistical population* defined by the statistical units, *statistical variables* representing the process of measurement and *statistical values* defining the range (or co-domain) of the statistical variables. According to a survey carried out by Work group 4 of METANET there is rather large agreement among statisticians on the following definitions of these terms:

(i) Statistical units are the entities for which information is sought and for which statistics are ultimately compiled. Statistical units may be real world objects (e.g. persons, enterprise) or abstract objects (e.g. accidents, transactions).
(ii) Statistical variables are defined characteristics for the statistical units that are used in the measurement process, e.g. income, sex, age or production volume.
(iii) Statistical values are the concrete result of the measurement process for each statistical variable and statistical unit, e.g. 112 000 EURO, female, 37 years, 3.5 million tons. Values can be determined on different measurement levels or modalities and by using a classification procedure.

11. Based on these basic terms one can define *statistical datasets* as the organized collection of the outcome of measurements for a number of statistical variables. Statistical datasets may occur in different guise, for example as case level data (the well-known case by variates matrix), as summary level data (multidimensional tables), or vectors of observations (for example time series).

12. Besides these basic categories we need a number of additional categories describing the structure of statistical value sets: *grouping levels* defined by aggregation of statistical values, *classifications* describing hierarchical schemes of grouping levels, *scales* capturing information about the measurement process and *measurement units* representing the meaning of quantitative variables. A further category called *statistical domain* is needed as organization principle for a statistical information system. Basically a statistical domain binds together the different categories occurring in connection with one investigation or a number of investigations. The statistical domain plays a central role in statistical survey processing in that sense that it supports the loose coupling of different information entities. Roughly speaking one can think of a statistical domain as a catalogue system for the different categories which are put together according to some subject matter consideration.

13. From data modeling point of view these categories can be interpreted as abstract classes. A concrete realization of a category occurring in connection with practical applications has to be represented in a twofold way: As category instance model (CI-model) and as category instance data (CI-data). The CI-data correspond to data for the different objects occurring in the context of survey computing and the CI-model describes these data. Obviously CI-data for the category statistical dataset are the most important type of data, but also other types of CI-data are well known: an administrative register may be seen as CI-data representing statistical units, a census file used in a sampling procedure may be seen as CI-data of a statistical population or a hierarchical tree structure of value sets may be interpreted as CI-data of a classification.

14. The CI-model is usually represented only partially in extensional format as a ‘file description’ or a ‘codebook’. Other parts of the description are many times only present in the mind of the statistician. Such an implicit consideration has often drawbacks with respect to documentation of results and implies many times a lot of additional and tedious processing activities, in particular in case of complex procedures consisting of many tasks and involving different people. It is one of the main goals of the UMAS model to overcome these obstacles by giving an explicit representation of all CI-data occurring in context of survey computing, together with the description of the corresponding CI-models. In ideal case such close connection between CI-data and CI-models allows formulation of preconditions for processing in a more formalized way and supports in that sense the work of statisticians. Prerequisite for such an active use of CI-models in the survey computing is representation of the CI-models in extensional format, i.e. as data. Such a representation of CI-models as data may be denoted by the term *statistical metadata*.

15. Formulation of CI-models in extensional format as data is based on a unified description principle captured in the so called view facet. It consists of four different views together with a structural description of the interconnections between the categories implied by four views.

B. View Facet 1: The Conceptual Category View

16. The *conceptual category view* represents the subject matter definition of the category instance and builds in that sense the bridge to reality. Usually this view is represented by a verbal definition, in the most simple case an appropriate name or label for the object of interest. Besides this description we need in any way a temporal and geospatial specification stating time and location of validity for the definition.

17. Relationships between different category instances at the conceptual view resemble in some sense the traditional data modeling, for example ER-diagrams. Sundgren (1975) denoted modeling of such relationships by the term infological approach.

C. View Facet 2: The Statistical Category View

18. The statistical methodological category view describes the statistical properties of the objects using a number of formal parameters. The most important parameters are *type parameters* characterizing the object of interest within the class. For example in case of statistical datasets one type parameter specifies whether the dataset is based on case level data (often called microdata) or summary data (so-called macrodata). In the former case information is given for each statistical unit in the underlying population whereas in the latter case information is given for classes of statistical units. Another type parameter for datasets is usually needed for determination of the temporal structure of the dataset, distinguishing between cross sectional and time series data. In case of statistical variables the type parameter distinguishes between the different scale levels, for example qualitative variables and numeric variables. This distinction determines the additional categories needed for the description: in case of qualitative variables we need a discrete set of statistical values characterized by conceptual definition of the values and in case of numeric variables we need a measurement unit for designation of the meaning. Note that in general the values of such parameters may be combined in arbitrary way, for example we can consider time series data at the case level as well as at the summary level. In that sense the parameter concept is more flexible than the traditional inheritance structure in object oriented modeling.

19. Besides type specification we need also *role parameters* describing the actual role of the category in the context of a specific application. Contrary to the type parameters, which are usually fixed for a category instance within one statistical domain, role parameters may change within one statistical domain depending on specific processing activities. Well known examples for role parameters occur in connection with statistical variables: A variable may have in context of a statistical dataset the role of an identifier for statistical units (cases), the role of a cross classification variable identifying a class of cases, the role of a filter variable, the role of an explanatory variable in a model and so on.

20. Statistical relationships between the different categories are of utmost importance for establishing the statistical data model. For example, in case of statistical datasets the relationship to other categories may be described at a top level as shown in Figure 1. The relationships to statistical population and statistical unit are quite obvious and the structural relationships defined by the variables correspond to restrictions given usually by the role parameters of the involved statistical variables. Besides the relationships to the categories Figure 1 contains also some additional information about production. This objects as well as the numeric information are usually obtained and augmented in connection with processing activities captured in the stage facet.

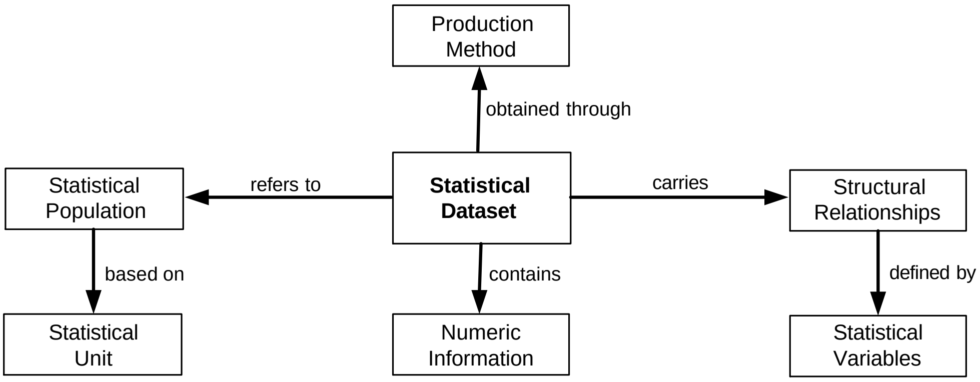

Figure 1: Sketch of statistical relationships between a statistical dataset and other objects

D. View Facet 3: The Data Management Category View

21. The *data management category view* is geared towards machine-supported manipulation, storage and retrieval of data. Main task of this view is management of CI-data in terms of files through properties often called *logistic metadata*. In general, the data management view concerns issues of how to represent, or encode, and manipulate entities and processes symbolically, especially in regard of storage and exchange, referring to data models and information structures as developed by computer science and mathematics.

22. Depending on the specific category under consideration different elementary data structures are necessary:

- For statistical units a useful data structure would be a list with operations like insertion and deletion.
- For statistical populations and statistical value sets a set structure together with the standard set operations would be appropriate.
- In case of statistical datasets a number of different structures may occur: matrices in case of cross sectional case level data (the well-known case by variates matrix), multidimensional tables (cubes) in case of summary level data, or vectors in case of time series data.
- A special case of data structure is necessary for statistical domains as the basic organization principle of statistical systems. CI-data of a domain may be thought in the simplest case as catalogues of the different CI-models for the different CI-data used in the domain, for example a catalogue of variables, a catalogue of registers, a catalogue of statistical value sets, or a catalogue of statistical datasets.

23. Due to the fact that statistical practice makes often explicit use of such elementary data structures, one has to consider these structures additionally to the usual data models. This does by no means deprive the importance and usefulness of traditional data modelling, which is of utmost importance for documentation of the different categories within a domain, in particular for capturing the conceptual category view.

24. In case of data archives such a data management category view is well known. The Data Documentation Initiative (DDI) (2001, see also http://www.icpsr.umich.edu/DDI) defines in its codebook a rather detailed description of such logistic attributes using a XML specification for the elements. Although the intention of the DDI group is mainly documentation of statistical datasets it can be applied to all other categories with minor modifications.

E. View Facet 4: The Data Administration Category View

25. The administration category view addresses administrative management and bookkeeping of all the structures. It is necessary for documentation of all kinds of activities in connection with definition of structures and schemas, insertion, update, and deletion of structures, as well as for search and retrieval activities. It has to take into account that production and storage of statistical data is often managed, or hosted, by public (often national) agencies or supranational (international) organizations with different subject matter orientation (e.g. economic data, social science data, biometric data, …). This implies administrative structures exceeding by far the conventional horizons and functionalities of – more or less local – data (base) administration.

26. Frequently, these structures also reflect *administrative* processes often implying that responsibility for data production and maintenance is spread among various administrative bodies, or agencies. Apparently, effective *statistical* usage of any of these data sources presupposes a fairly detailed knowledge of the administrative systems providing these data holdings. Moreover, legal aspects such as data privacy and data linkage prohibition rules have to be obeyed. While recent proposals for arranging data combination processes in the domain of data warehousing – as part of the so-called ETL-(extract-transform-load)-process – provide a range of technical solutions, administrative structures of data sources typically receive little attention.

27. In order to obtain appropriate documentation standards for the administration category view it seems reasonable to start from already existing proposals. An acceptable balance between level of detail and documentation effort seems to be the widely accepted Dublin Core standard for documentation of resources (see http://dublincore.org/), which is also used in the above mentioned DDI model for data archives. In this connection it should be mentioned that the METIS terminology offers a number of additional descriptive elements for administrative description.

F. The Stage Facet

28. The stage facet is responsible for support and documentation of all kinds of processing of categories within a statistical information system. There exist a number of documentation templates for statistical datasets, which cover usually the entire processing chain. Two important approaches are the proposal of Rosen and Sundgren (1991) in the area of official statistics and the already mentioned DDI scheme for social science data archives. In both cases the documentation is more a passive metadata repository. In order to make active use of metadata in survey processing one has to decompose these rather elaborated documentation templates into more operational building blocks. At the top level such decomposition can be defined by the four main stages: *definition and design*, *production*, *processing* and *dissemination and exchange*. A similar proposal was made by Statistics Netherlands using the terms 'input', 'output' and 'throughput'. A schematic representation of this decomposition is shown in Figure 2 and described in the following paragraphs.

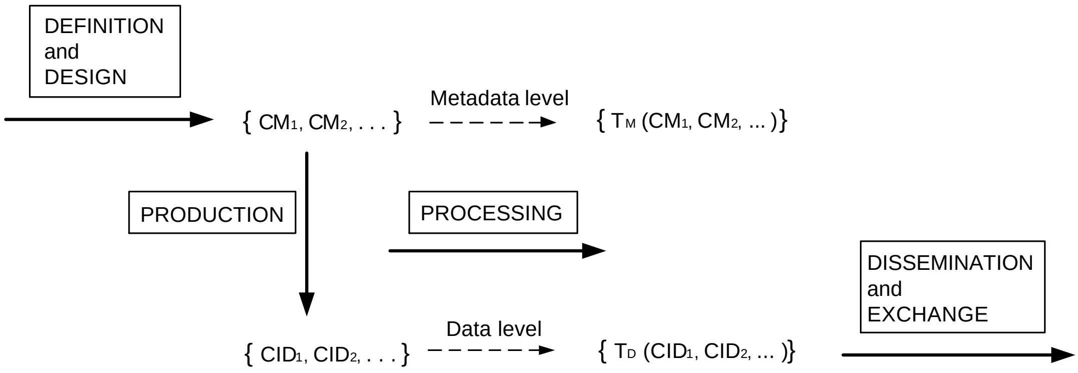

Figure 2: Schematic description of survey processing

29. The definition stage defines the work plan for setting up CI-data in advance to the production of the data itself. It is based on the analysis of the object system and the intended mapping of the object system into statistical categories. Main result of the definition part is a so called *CI-blueprint* of the CI-models for the envisaged category instances, denoted by CM_i in Figure 2. These CI-models describe the intended CI-data according to the different category views described in subsections B. – E., together with specification of the relationships to other category instances. In case of statistical datasets the file description for data used by statistical analysis systems can be interpreted as sketch of such a CI-blueprint referring mainly to the statistical and data management view.

30. Main result of the design part is an operational plan for the activities necessary for obtaining CI-data. Documentation of the planned activities is kept in a so called *CI-production-blueprint* describing the activities according to the different category views. The methodological background for the design phase are sampling theory and planning of experiments, often considered only as a side branch in statistical data analysis. Consequently, there is often a gap in documentation of the CI-production blueprint. Notable exemptions are specific tools for data capture and the already mentioned DDI standard, which keeps a quite complete documentation about the production steps.

31. The production stage establishes the CI-data for the category instances, denoted by CID_i in Figure 2, according to the CI-production-blueprint and completes also the blueprints of the CI-models with respect to production dependent parameters. In case of statistical datasets such parameters may refer to sample sizes or non-responses and are shown in Figure 1 as Numeric Information. For objects of the categories statistical dataset and statistical population theory of survey collection offers a number of instruments for computer assisted data production. Such data capture tools offer also a number of options for metadata management in the definition and design stage and use these metadata sometimes for active control of data capture. Furthermore, let us mention that a number of preprocessing steps, for example editing, are considered many times as part of the production stage. Another example for documentation of the production step occurs in context of classifications. International organizations put quite a lot of effort into documentation of the development of international standard classifications, usually in verbal form from the conceptual point of view.

32. The processing stage generates new category instances out of already existing ones by a sequence of transformations. As shown in Figure 2 each transformation operates at the data as well as at the metadata level with strong interconnection between the two levels. Roughly speaking one can say that processing of the CI-models refer to preconditions and admissibility checks for processing at the data level, which corresponds more to traditional processing activities. For each processing activity a *planning phase* and an *execution phase* may be distinguished. The planning phase of a transformation defines an activity plan for the envisaged processing and as result of the planning phase one obtains one or more *CI-blueprints* and a *CI-processing-blueprint*.

33. The CI-blueprints, denoted by $T_M(CM_1, CM_2,\ldots)$ in Figure 2, describe the CI-models for the output category instances according to the description of the statistical categories and the relationships between statistical categories as outlined in subsection B. – E.

34. The CI-processing-blueprint describes the envisaged processing activities for the CI-data – denoted by $T_D(CID_1, CID_2,\ldots)$ in Figure 2 – with respect to the *conceptual, statistical, data management*, and *administrative processing view:*

- The *conceptual processing view* describes the processing method from a general point of view, in particular the intention of the method and the connection to subject-matter issues.
- The *statistical processing view* describes the transformation from a statistical methodological point of view. In particular, the following details have to be specified:
  - (i) The *input CIs* used in the transformation;
  - (ii) The *output CIs* produced by the transformation;
  - (iii) The *operators (statistical methods)* applied to the input CI-data;
  - (iv) The *operator parameters* necessary for detailed specification of the algorithm.
- The *data management processing view* is responsible for keeping additional results occurring besides the CIs in connection with processing, for example process parameters.
- The *administrative processing view* informs about administrative details of processing.

35. Dissemination and exchange are the main operations for obtaining information about category instances and processing activities of the category instances. All information required for dissemination is based on a specific view towards the category instances, the CI-production-blueprints and the CI-processing-blueprints. Contrary to dissemination of CI-data, which is usually well defined, recommendations for the dissemination of CI-models is often rather vague.

G. The Function Facet

36. The function facet has to be seen in close connection with dissemination and exchange and considers communication and usage aspects of statistical meta-information by humans. In order to fulfill the communication needs one has to specify the *person* involved, so-called consumers, the content requirements for information, so-called *transfer standards*, and the encoding schemes for transmitting the content, so-called *transmission standards*.

37. There are numerous potential consumers of statistical information, which may be classified with respect to different criteria. On such criterion identified within METANET is the level of background knowledge. Information end consumers with limited background information and interested mainly in serious summary information, institutional consumers with advanced background knowledge, and scientific consumers with high level background knowledge may be distinguished. Another criterion considered within METANET is the role of the person seeking information from a statistical system. Such roles range from citizens rather far from the statistical system over press officers or teachers interested in dissemination of results up to persons who work in close connection with the system like IT specialists or data administrators.

38. In the area of transfer and transmission standards a number of proposals are available. Well known examples for transfer standards are the SDDS standard for dissemination of economic data, the EUROSTAT quality standard or the rather general standard for disseminating statistical data on the Internet of the UN-ECE. A major drawback of all these standards is that the content is well defined by a number of information items but the methods for obtaining such standardized information is not well defined. With respect to the transmission format different specifications are well known in the area of official statistics, for example

GESMES (see http://www.gesmes.org/) for dissemination of statistical tables together with appropriate descriptive elements. However, note that this is not only a transmission standard but contains also some normative ideas for the content. Nowadays practical all transmission standards are based on XML.

IV. IMPLICATIONS FOR TERMINOLOGY

39. The investigations of METANET showed that there is abundance of terminology in statistical information systems which may be traced back to different roots. On the one hand there is a lot of statistical terminology, in particular in connection with survey processing, on the other hand terminology from other scientific disciplines like computer science, mathematics or library sciences is used. Moreover we have the subject matter fields for which statistics are compiled as additional sources for terminology, in particular economics and social sciences. All these fields have their own history and developed terminology for describing their data according to their view onto reality. Due to the fact that statistics is only one area which contributes to terminology it seems unrealistic to aim for standardization of all these developments under the umbrella of statistics. Moreover, we must be aware that even inside statistics different terminologies exist within specialized sub-communities. More realistic seems an approach which tries to bring some order into the wilderness of terminology by defining a limited number of generally accepted classification criteria for all the terms.

41. Based on the UMAS model a scheme for classification of terminology was developed, which is represented as XML schema in Figure 3.

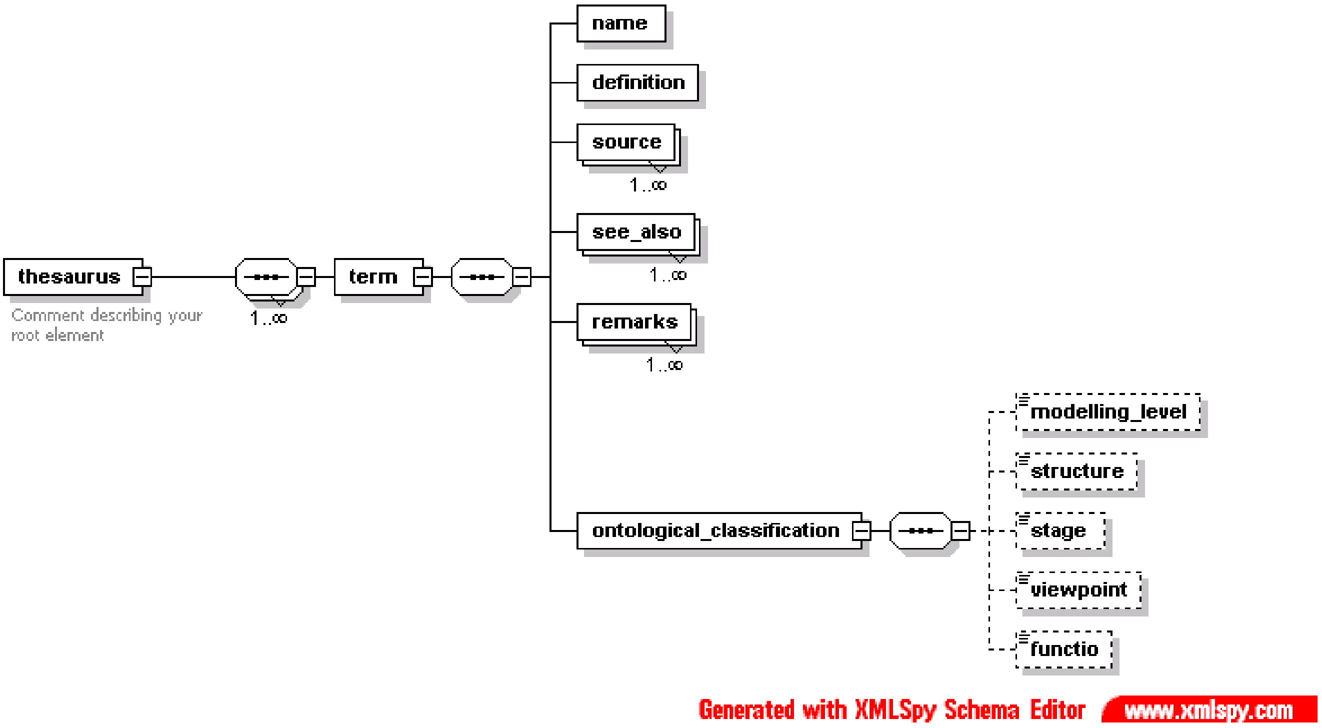

Figure 3: XML Schema for Terminology

The ‘name’, ‘definition’, ‘source’, and ‘remarks’ elements correspond to the usual descriptors for terminology and contain the name of the term, the definition, the source of the definition and some additional remarks for explanation. With respect to relation to other terms usage of a ‘see_also’ element seems more appropriate than defining relations like broader and narrower term which are in many cases ambiguous. The elements for the ontological classification reflect the outline of the UMAS model described in the previous section. In addition to the facet elements an element called ‘modeling level’ is included. It is used to distinguish between the different sources of terminology. The following values for this element are possible:

(i) *Tools and methods* is designated to terms which are borrowed from other methodological discipline like computer science or mathematics. Such terms can be used within a statistical information system but have no additional semantics in the sense of statistical data modeling.

(ii) *Subject matter* is used for terms which stem from subject matter disciplines like economics or social sciences. Similar to the tools and methods such terms should not have any additional semantics in the sense of statistical data modeling.

(iii) *Statistical* is used for terms which are used inside the statistical data modeling paradigm. For these terms the other elements of the ontological classification apply according to the specifications of the different facets.

40. Usage of a classification for ordering terminology was recently also proposed by the SDMX initiative, a consortium which encompasses a number of international statistical organizations. The SDMX approach distinguishes between the following headings for the different items: 1. Administration; 2. Concepts, Definitions Standards; 3. Data Collection, manipulation/accounting convention, etc.; 4. Quality and Performance Metadata. These headings refer to some of the facets and category views presented in the setup of the UMAS model. The main difference is probably item 4. Terminology considered under the heading ‘Quality and performance metadata’ would fall within the UMAS model into the dissemination view and the function facet. The reason for this is that the UMAS model puts more emphasis on survey processing and views quality as the result of the processing activities. Hence the items under heading 4 of the SDMX proposal correspond more to a specific view onto the UMAS proposal and an evaluation of the description for the objects and for the processing activities inside the statistical system.

V. REFERENCES

Data Documentation Initiative (DDI) (2001), “Codebook Document Type Definition (DTD)”, http://www.icpsr.umich.edu/DDI/CODEBOOK/.

Froeschl, K. A., Grossmann, W., Del Vecchio, V. (2003), “The Concept of Statistical Metadata”. Deliverable 5 METANET (EPROS Project IST- 1999-29093), University of Edinburgh.

Rosen, B. Sundgren, Bo (1991), “Documentation for reuse of microdata from surveys carried out by Statistics Sweden (SCBDOK)”, Statistics Sweden.

Sundgren, Bo (1975), “Theory of Databases”, Mason/Charter, New York.