WP. 10
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 MANAGEMENT AT ISTAT: CONCEPTUAL FOUNDATIONS AND TOOLS

Invited Paper

Submitted by [ISTAT, Italy][1]

I. INTRODUCTION

1. The present paper has two main sections: in the first one (Section II) we present some remarks that in our opinion justify the ISTAT approach to metadata management. In such a way we also intend to give our answers to some of the questions that have been specified for the theme of the session, namely Metadata models and terminology. In the second one (Section III) we outline the main features of the ISTAT strategy for metadata, with a particular focus on SDOSIS, the system for documenting the information content of ISTAT surveys.

II. CONCEPTUAL FOUNDATIONS FOR A METADATA MANAGEMENT STRATEGY

A. Role and the evolution of the notion of metadata: general remarks

2. The trend towards a massive exchange of pieces of information which involves various and heterogeneous organisations, information management systems, people is an outstanding evidence in latest years, a trend which concerns the producers and the users of statistical information too. Increasing bulks of exchanged and shared data require the availability of proper metadata. Generally speaking metadata are information that is required to retrieve, understand and properly use data. According to such a definition metadata are inherently context-dependent in principle. In different contexts different metadata are necessary: which pieces of information about data are required is a matter of the data exploiting activities and subjects that characterise the given contexts.

3. Nevertheless the present trend towards increasing data exchange makes the need arise for less context-dependent, more general-purpose metadata. Such a recent evolution resembles what happened in the early 1970 decade: as an integrated approach to data management within organisations became current, the conceptual specification of the content of any shared database according to a model (i.e. the Entity-Relationship model) turned into a necessity. An outcome of such an evolution is that the conceptual specification of the data

[1] Prepared by Giovanna D'Angiolini, dangioli@istat.it

.

semantics is now generally recognised as the minimal core of any metadata collection. The conceptual modelling approaches were aimed to support data sharing inside organisations; simple common sense based models such as the Entity-Relationship model or, more recently, the Fact-Dimension model are adequate for describing the real world of interest for any organisation as a whole, so as to provide all the subjects belonging to the organisation with a shared data semantics specification. Today a similar problem arises, at a higher level of generality. The travelling of data across the boundaries of different organisations as well as their exploitation by heterogeneous subjects in different contexts require more metadata as well as a more in-depth, general-purpose representation of data semantics. Research about ontologies is an example of a theoretical attempt to face the new situation. Moreover proper models and languages are still regarded as the main conceptual tools for attributing semantics to data.

B. The nature of statistical metadata

4. With respect to the outlined general scenario, the production and the use of statistical data have an interesting role: they support a particular activity, namely statistical analysis, and they are performed by several subjects that exchange data for this purpose. Therefore metadata for statistical data are both context-dependent and general-purpose. They are context-dependent if compared with generic metadata because they support a particular kind of activity. At the same time, given that the exchange of data between different subjects is an important component of statistical activity, statistical metadata are required not to depend on the particular characteristics of the subjects that produce and use data therefore they are general-purpose with respect to the data exploiting subjects. This is the reason why the need for general-purpose, exchangeable metadata has arisen much earlier for statistical data.

5. Statistical information is gathered and managed by people and organisations for the purpose of analysing real world phenomena. The statistical analysis of real world phenomena defines the reference context, which gives statistical data their proper semantics and determines which metadata are relevant for properly using them. This is the reason why the approaches and tools, which the market offered in the latest years, never matched the requirements of the statisticians in a satisfactory way: they concern information as a resource for an organisation, while the statisticians are interested in information as a resource for statistical analysis. The OLAP approach has encouraged the general misunderstanding about the nature of statistical data, because it spread the idea that the analytical usage of information is something internal to an organisation and that transactional data can be used for analyses simply by means of arranging them in a new format. On the contrary, any statistician knows that a) in principle statistical analyses exploit data from several sources and also, if needed, data produced by several organisations, even when they are performed for supporting decisions which concern a particular organisation b) the semantics of the statistical data should be specified in terms of concepts strictly related to statistical analysis, such as statistical unit, variable, classification c) in statistical analyses it is important to know the particular features of the observation process that generated the exploited data d) in order to be used for analytical purposes the transactional data must be evaluated for assessing their usability for statistical analyses, modelled by means of statistical concepts, equipped with proper metadata.

C. Singling out classes of statistical metadata

6. Due to the role and nature of statistical activity, an original approach is required to define which are the relevant classes of metadata for statistical data. What has been done for singling out classes of metadata for generic data amounted at answering a question such as: what is important to know for performing data processing activities inside organisations? Analogously, in order to single out all the relevant classes of statistical metadata we should ask: what is important to know for producing, exchanging and exploiting data for statistical analyses? Answering such a question implies analysing statistical activity, in the same way as an analysis of the generic data processing activity underlies generic metadata. As a first step we can classify the relevant bulks of knowledge about data by means of very general criteria so as to distinguish among coarsest classes of metadata. Note that such classification dimensions for metadata are the same for both generic and statistical data, because of their generality. A first dimension concerns the content of metadata, other dimensions are the level of abstraction, and the scope. As to the content, we discriminate between metadata concerning the data semantics, on one side, and metadata concerning the environment in which data are generated and exploited as well as other conditions which may influence the way in which data are managed and exploited, on the other side. Metadata concerning the data semantics are the minimal core of metadata. As

to the level of abstraction, we discriminate between conceptual and operational metadata, as to the scope we discriminate between local metadata, which concern a homogeneous amount of data, and global metadata, which concern heterogeneous data collections. According to such classification criteria, we have conceptual as well as operational metadata, which concern the data semantics, beside conceptual as well as operational metadata, which concern the other relevant aspects. All these metadata classes may be considered at a local level as well as at a global level. Inside these coarsest classes of metadata we single out more specific classes of metadata by means of in-depth analysis of the statistical activity. These singled out specific classes of statistical metadata differ from the analogous classes of generic metadata. At a conceptual layer the data semantics is specified by means of using statistical concepts such as statistical unit, variable, classification instead of common sense concepts such as entity, relationship, while the metadata about other relevant aspects describe the processes by which the statistical data have been generated instead of the owner organisation. The local metadata concern a single source of information instead of a single organisational entity. The analysis of statistical activity entails establishing boundaries between conceptual and operational metadata in a particular way: the conceptual statistical metadata are defined at a knowledge level, the operational statistical metadata include metadata for describing the organisations and subjects which are involved in statistical activities as well as operational metadata in the strict sense of the word.

7. It is worth noting that when we analyse the content and the characteristics of the statistical activity we are also producing a general ontology for such an activity. In fact an ontology is an explicit formal specification of how to represent the objects, concepts and other entities that characterise some area of interest and the relationships that hold among them. Therefore we obtain something more structured than a simple list of relevant metadata classes: we define a framework of logically related metadata classes, based on a proper general ontology.

D. The role of models and terminologies

8. The above remarks offer some directions for assessing the role of models and terminologies in specifying statistical metadata. Generally speaking a model is a set of related concepts, which is used for producing a structured specification of a metadata class. For specifying the data semantics a number of models has been proposed. The existence of several models is due to the different viewpoints from which the statistical activity can be regarded, as an example, the viewpoint of the producers of statistical data may differ from the viewpoint of the data analysts. Thus generally the existing models have their own rationale, even if some models may not respect the distinction between conceptual and operational aspects. As to the metadata classes concerning the other relevant aspects of statistical activity, it is worth noting that some model always underlies their specification, even when they are presented in a list of items format. How are the existing models related to the general ontology that is built by means of analysing the statistical activity? Models and ontologies are both conceptual tools for producing structured specifications of metadata classes. Both of them can be given a formal interpretation by means of logic. However the ontology for the statistical activity is something more fundamental and general than any particular model. It should be possible to establish a correspondence between each one of the existing metadata models and such an ontology: each component of any metadata model should be mapped to a component of the ontology.

9. We maintain that proper models for each class of metadata are the basic conceptual tools for metadata specification, in any concrete environment and for any subject, because they produce a structured representation, which may be justified by means of logic. Moreover the definition of a unique set of reference classes of metadata based on a proper ontology for the statistical activity is the main tool for ensuring the comparability of different models and ultimately the exchange of metadata. In fact by means of referring the components of different models to such a common ontology, a correspondence between different models can be established. In our opinion to define the ontology for the statistical activity is a difficult but affordable task for the statistical metadata community. The Metanet initiative can be judged as a first important step towards this goal.

10. What about terminologies? A terminology is a collection of terms, which have associated a definition involving other terms. In the following sections we use this notion to denote the network of statistical units, variables and classifications that are observed by a source of information (in particular a survey). This is consistent with the usage of the notion in computer science. Often the statistical metadata community use the

word terminology for denoting a collection of defined terms which describe relevant aspects of the statistical activity and therefore can be used for defining the content of metadata in an unambiguous way: such a kind of terminology will include for example the term “statistical unit”, therefore such terminologies are better defined as meta-terminologies. In our vision a terminology is something that is derived and justified by an underlying model or ontology. Therefore a common meta-terminology for statistical metadata can be obtained as a by-product of the definition of the ontology for statistical activity; moreover, comparing existing meta-terminologies implies comparing their underlying reference model. However the emphasis on meta-terminologies has the important role of attracting the attention on the importance of the definitions. In principle each concept in a model or an ontology, such as the concept “statistical unit”, should have attached a definition in terms of other defined or primitive concepts.

III. METADATA MANAGEMENT AT ISTAT: STRATEGY AND TOOLS

A. Foundations of the ISTAT strategy for metadata specification

11. The ISTAT strategy for metadata specification is based on the approach that is briefly outlined in the foregoing section. Given that the particular role of a subject in performing statistical activity has an influence on its adopted metadata framework and metadata models, our metadata framework results defined from the viewpoint of an organisation which produces and disseminates statistical data.

12. The produced and disseminated statistical data come from observing real world objects by means of specific techniques and procedures, for properly using data the analysts need to know what has been observed and the process by which it has been observed. Therefore our metadata specification is based on the following main concepts:

- a SOURCE of statistical information is any process, which is activated to observe real world phenomena so as to produce statistical information. A survey is a source, an administrative data collection is a source too. A source produces data collections by means of applying its own data production techniques and procedures;
- a STATISTICAL INFORMATION SYSTEM (SIS) is an integrated collection of pieces of statistical information, which concern related phenomena and are issued by different sources. SIS are built for satisfying various and/or unpredictable information requirements. They are often produced by the official statistical institutions for public usage.

These concepts define the environments in which statistical data are generated and organised. In order to document data semantics, proper metadata describe the information content of a source or a SIS. In order to document the other conditions, which may influence the interpretation and usage of the issued data, proper metadata describe the characteristics of each source as an observation process. These are the basic metadata classes. Further analysis of the two main metadata classes can be developed along two other dimensions: the level of abstraction and the scope. By means of considering sources and SIS at different levels of abstraction we single out three main metadata layers: the conceptual layer, the organisational layer and the operational layer.

13. In the conceptual layer we look at surveys (generally speaking, sources) and SIS as knowledge bases, from this viewpoint we describe the data semantics and the characteristics of the observation processes. It is worth noticing that a thorough specification of the data semantics requires that the meaning of each datum be ultimately expressed in terms of observed real world objects. For this purpose we single out several components in the specification of the information content of surveys and SIS. The first component is the specification of the observed real world objects (the content of the questionnaires, for surveys). Another component is the specification of those objects, which are derived from the observed ones by means of proper transformations. Both the observed and derived objects belong to categories, which correspond to statistical concepts, namely elementary concepts such as statistical unit, variable, classification as well as structured concepts such as statistical table. Each observed or derived object is described as a term with its own definition, which has a set of links with other object describing terms. The set of all observed and derived terms is the terminology of a survey or a SIS. On this basis we can produce the specification of the data semantics: the meaning of the data is represented by means of describing them as associations of terms of a terminology. We denote such associations by the name Information Frames. Each source or SIS has its own terminology, several

sources can share production procedures and terminology concepts, but they produce their own information frames.

14. In the conceptual layer the metadata concerning the characteristics of the observation processes specify for each source the conceptual features of the adopted data production techniques, namely the sample design, the estimation formulas and procedures, the performed operations for data capturing and processing.

15. At lower levels of abstraction we document the organisational and implementation characteristics of surveys and SIS, therefore we describe the activities of data interchange among data producers, data users and other organisations as well as the adopted procedures and systems for data production, transformation and dissemination.

16. Along the scope dimension we distinguish between local and global metadata. In the conceptual layer, local metadata are those metadata that describe the information content and the characteristics of each source of information. Global metadata are those metadata that are obtained by means of conceptually integrating or standardising local metadata. The specification of the information content of a SIS is an example of global metadata. Other examples are the standard terminologies such as the Eurostat standard terminology for surveys on enterprises. Conceptual integration is the activity, which is performed in order to produce global metadata, in particular for defining the information content of a SIS. It is a complex activity, which requires comparing and matching terms which belong to different source terminologies, by means of analysing their definitions; for this purpose, such definitions must be expressed in a structured format, as combinations of formally defined constructs. This is the reason why inside a large collection of sources, such as the set of all the surveys managed by a national statistical institute, we hardly find a complete conceptual integration of the terminologies and generally only delimited areas of integrated concepts are defined, corresponding to SIS or sets of partially integrated terms such as general standards or area standards. On the contrary it is easier to attain a conceptual layer standard description of the surveys' production processes, by means of exploiting thesauri of standardised descriptions of operations.

17. In the organisational and operational layer, global metadata should describe those practical interactions among organisations, processes, data and agents, which are involved in statistical data production and usage.

18. For each metadata class we define a model. It is worth noticing that our adopted models are defined from the viewpoint of an organisation that produces and disseminates statistical data. As an example, our model OSI, even if employs statistical concepts such as statistical unit, variable, classification, is common sense oriented at a certain extent, because it is conceived for specifying the content of very heterogeneous surveys, such as surveys on households and on enterprises, in a homogeneous way. Moreover, given that ISTAT mainly deals with the “basic stuff” for analysis, namely micro data and macro data, it does not offer constructs for describing the results of more sophisticated analytical activities.

B. The ISTAT strategy for metadata management

19. The core of the ISTAT strategy is the development of two centralised systems for metadata management, SIDI and SDOSIS, which manage metadata concerning the production processes of surveys and the information content of surveys and SIS, respectively. Both of them are based on proper metadata models. They will disseminate metadata to both data users and survey designers. Moreover they work as metadata servers for those data management systems and software tools that are exploited in the data production and dissemination activities.

20. SIDI is dedicated to manage metadata concerning the survey production processes. SIDI is based on a metadata model that allows for associating each survey with a set of OPERATIONS. An operation is a high-level description of survey procedures, such as Data capturing by means of CATI techniques. Each operation is associated with a set of CONTROL ACTIONS, namely particular operations that are performed for monitoring the production procedures. Operations and control actions are performed by AGENTS, moreover they produce and exploit DATA REPOSITORIES. SIDI warrants a standard specification of the survey production processes, which is ensured by means of a network of pre-defined thesauri. For each concept in our model we have built in the SIDI database a thesaurus of admissible descriptions. In particular, thesauri have been defined

for OPERATIONS and CONTROL ACTIONS and other auxiliary concepts. For describing a particular feature of the survey production process the survey manager may choose a description in the thesaurus or insert a new description; in the latter case, the new thesaurus item must be validated by a quality manager. This feature of SIDI ensures a meaningful concept-based inquiry. The end user chooses one or more operations, one or more control actions, one or more statistical units, and the system select those surveys whose production process specification matches such user-defined search criteria; then the end user can select a single survey in this list and view its metadata and quality indicators. At present SIDI is implemented and manages metadata describing the majority of the ISTAT surveys.

21. SDOSIS is aimed to document the information content of ISTAT surveys as well as the results of any integration activity. It is based on the OSI model, which is illustrated in the following section D. The first version of SDOSIS offers functionalities for specifying terminologies and information frames of surveys and SIS. Future versions of SDOSIS will directly support the integration activity, by means of offering functionalities for the analysis of terminologies. The first SDOSIS release, which will be delivered by May 2003, only manages the survey terminologies.

C. Main features of the ISTAT system for documenting the information content of surveys and SIS (SDOSIS)

22. The present version of SDOSIS encompasses a metadata specification environment, an inquiry environment and a classification repository.

23. In the metadata specification environment the survey manager specifies the survey's terminology according to the OSI model. Unlike the production process, the information content of the surveys cannot be specified in a homogeneous way by means of pre-defined thesauri. A standard specification would require the conceptual integration of the managed surveys, which would imply in-depth analysing their information contents. Therefore SDOSIS allows the survey managers to freely specify the name and definition for each term in the survey's terminology; in order to support such a task a function is available for text extracting from questionnaires and other survey documents. For the purpose of boosting standardisation however the survey managers may declare, for each term, a correspondence with a standard term. As an alternative choice, they may declare a correspondence with a local area standard term, which is shared by a set of similar surveys, or with a term in another survey's terminology. In such a way, SDOSIS documents all those situations in which there is a potential for standardising or integrating surveys. In order to allow the survey managers to establish such correspondences, SDOSIS manages standard terminologies together with survey terminologies. More precisely, the system documents those terminologies that are owned by official standards as well as by local area standards. As an example, the Eurostat nomenclature for long-term surveys on enterprises is documented in SDOSIS as an official standard with its own terminology; the ISTAT standard SIMIS, which collects the main concepts used in surveys on households, is documented in SDOSIS as a local area standard with its own terminology too. The standard terminologies are specified according to the OSI model by a particular system user, the manager of the standards, whose role is also to define integrated standard.

24. The SDOSIS database is structured as a network of term repositories. Each repository stores a collection of both survey and standard terms for a particular OSI concept; therefore we have repositories for statistical units, numerical variables, classification variables, classifications and the other relevant OSI concepts.

25. Because of the complex context that it documents, SDOSIS provides the end user with several inquiry functionalities. In particular, it offers term-based inquiry functionalities that enable the user to choose both standard and non-standard terms inside proper term repositories and looking for surveys or standards whose terminology includes the chosen term. In particular the user who has chosen a term describing a statistical unit has two other functions available. The first one allows the user to refine the initial choice by means of choosing another term among the subsets of the chosen statistical unit, the other one allows the user to choose other terms of interest among numerical variables, classification variables, classifications which are connected to the chosen statistical unit. On the basis of the user's choice the system produces a list of standards and surveys whose terminology includes terms that matches the user's specified terms. After having selected a single

standard or a single survey in such a list, the user can navigate across its terminology as well as, in the following SDOSIS releases, view the survey's information frames, and other characteristics. By means of its functional integration with a data dissemination system, in the future SDOSIS will allow the user to access the data issued by the selected survey too.

26. For the purpose of warranting meaningful inquiries, the inquiry environment of SDOSIS exploits a distinguished network of term repositories, which is updated by the system manager, by means of proper functionalities, starting from the SDOSIS database. The terms in such repositories are defined on the basis of the terms which set up the survey and standard terminologies by taking into account the declared correspondences between defined terms, on one side, the existence of synonyms among the names of terms, on the other side. In such a way the system manager builds a network of thesauri for supporting inquiries.

27. The SDOSIS classification repository stores the set of modalities of each documented classification together with correspondence tables for linking modalities of different classifications. The classification repository, which is an essential tool for survey design and management, is accessed directly as well as by the general management and inquiry functionalities.

28. SDOSIS is equipped with an environment for non-structured documentation in which the survey managers can describe the survey questionnaires and store their images, together with other documents. The future SDOSIS releases will manage some classes of operational metadata describing input and output data repositories such as print tables, files, database relations, data marts.

D. Main features of the OSI model

29. We have defined a conceptual model, called OSI (Objects-Information Frames), for specifying terminologies and information frames of surveys and SIS. From the viewpoint of statistical analysis each source is regarded as a distinguished way of observing the reality of interest. This is the reason why our model enforces a clear distinction between the specification of the observed part of the real world, which can be shared by different sources, and the description of the data issued by each source. Our modelling approach has two other main features. The first one is the explicit representation of the observed sets of individual objects as STATISTICAL UNITs. The other one is the distinguished specification for the observed qualitative properties of individual objects, on one side, and the set of admissible values for such properties, on the other side. They are represented as CLASSIFICATION VARIABLEs and CLASSIFICATIONs, respectively. In our opinion the lack of such a distinction is an important limit of the Fact/Dimension data model, which is used in the OLAP tools. In fact, the core of the statistical activity is to observe and measure homogeneous sets of objects: this implies singling out the basic sets of objects of interest (the STATISTICAL UNITs) and partitioning them according to pre-defined sets of values (CLASSIFICATIONs) for their observed qualitative properties (CLASSIFICATION VARIABLEs). Therefore the used classifications are basic concepts when the data are actually modelled for satisfying analytical purposes.

30. The terminology of a survey or SIS specifies those real world objects that are currently observed or are derived from the observed ones by means of proper transformations.

The OSI model assumes that these are the kinds of objects that it is important to specify: STATISTICAL UNITs, NUMERICAL VARIABLEs, CLASSIFICATION VARIABLEs, CLASSIFICATIONs, IDENTIFIERs, IDENTIFIER_SETs, ASSOCIATIONs.

A STATISTICAL UNIT is a set of observable individual objects. The notion of statistical unit describes observable populations such as Household, Business, as well as sets of observable events that involve instances of observable populations, such as Household-vacation, Person-hospitalisation.

A NUMERICAL VARIABLE is a quantitative property of observable individual objects, such as Family Income, Turnover, on which simple aggregation function (such as SUM, AVERAGE) can be applied. Any numerical variable has a numerical domain (such as INTEGER, REAL) and a unit of measure if its domain is Real. Weight is the special NUMERICAL VARIABLE that is used for counting the number of items of a statistical unit.

A CLASSIFICATION VARIABLE is a qualitative property of observable individual objects, such as Sex, Economic Activity, which can be used for classifying statistical units.

A CLASSIFICATION is a set of states, which can be observed for some qualitative property of observable individual objects. Each classification has an extension that is a list of identifiers and names of states, as an example, Sex-classification has associated the list {male, female}. These names of states are the classification's MODALITIES. A CLASSIFICATION VARIABLE is connected with one or more CLASSIFICATIONs, a CLASSIFICATION is connected with one or more CLASSIFICATION VARIABLEs.

It is well known that classifications can be organised in CLASSIFICATION SYSTEMs. A CLASSIFICATION SYSTEM is a set of aggregation relationships between classifications. An example of official classification system is the NACE Rev.1 classification for the economic activity.

An IDENTIFIER is a particular qualitative property of observable individual objects, which may be used to distinguish the single items of a statistical unit. An IDENTIFIER is connected with one or more IDENTIFIER_SETs, which are sets of identifiers of individual objects.

An ASSOCIATION is a one-to-many or a one-to-one relationship between statistical units. It is worth noting that for analytical purposes it is convenient to regard any many-to-many relationships between statistical units as a statistical unit too.

Any term in a terminology represents an object belonging to one of the above categories. Each term has associated the NAME of the represented object with its SYNONIMS and the OBJECT-DEFINITION, which is expressed in terms of other object describing terms.

The terminology of a survey or SIS is a network of connected terms. In a terminology STATISTICAL UNITs are connected to CLASSIFICATION VARIABLEs and NUMERICAL VARIABLEs; two STATISTICAL UNITs may be connected by an IS_A (subset) relationship; ASSOCIATIONs connect statistical units; two CLASSIFICATIONs may be connected by an AGGREGATION relationship, inside one or more CLASSIFICATION SYSTEMS.

An important feature of the OSI model is the availability of a set of TERM BUILDING CONSTRUCTS, which are used for specifying those transformations by means of which new terms are derived from the available ones. Obviously these constructs may be used to express the object definitions in a structured format too.

31. OSI encompasses two other categories of objects, namely CONSTRAINTs and STRUCTURED OBJECTs. The CONSTRAINTs are special relationships among objects, the STRUCTURED OBJECTs are objects that are built by associating elementary objects.

The most important example of a STRUCTURED OBJECT is the STATISTICAL TABLE.

A STATISTICAL TABLE describes the result of applying an aggregation function such as SUM, AVERAGE on the values of a numerical variable or a vector of numerical variables that have been observed for the instances of one or more sets. In a statistical context such an aggregation operation may concern a single statistical unit as well as the elements of a set of statistical units, which is obtained by means of partitioning a given statistical unit according to some qualitative characteristics. An example is the table Number and total Turnover of Businesses by Dimension and Economic Activity. In its definition it is implicit that we have observed a set of Businesses with their Dimension, Economic Activity, Turnover and Weight, moreover we have adopted two classifications for Business Dimension and Economic Activity, for instance Dimension Groups and NACE Groups. The table is defined by means of two transformations: a) partitioning Businesses according to the vector [BusinessDimension/DimensionGroups, EconomicActivity/NACEGroups], b) for each element of the attained set of statistical units applying an operator SUM on the values of the vector [Turnover, Weight] which are associated with its instances. The components of this statistical table are numbers, each number represents the total turnover or the total number of businesses for one of the subsets which we have obtained by partitioning Businesses by means of the qualitative characteristics: BusinessDimension/DimensionGroups, EconomicActivity/NACEGroups.

Generally statistical tables are used to specify the content of the issued data, but this is not the only context in which statistical tables occur. In fact, representing statistical tables in a terminology is mandatory because often a survey directly collects statistical tables, which are the result of a previous aggregation operation instead of, or in addition to, information related to individual real world objects. Note that an association with another statistical unit occurs in the definition of such observed statistical tables. An example is the table Total number of Students enrolled in Degree Courses by Sex, Age-class and Degree Course that may be defined by partitioning the statistical unit Students according to the

vector [Sex/Sex classification, Age-class/Five-years classes, Enrollment/Degree Course] and then applying the operator SUM on the values of the numerical variable Weight.
The OSI set of term building constructs includes proper constructs for deriving new tables from the available tables, or from collections of elementary terms.

32. Having defined a terminology for a SIS or a single survey, we can describe the meaning of its issued data. We need to model both observed data, which are the direct output of the data collecting and editing procedures, and data obtained by means of transformations. We specify the issued data as INFORMATION FRAMES. An information frame is a structured collection of terms of a terminology. Moreover an information frame refers to a TIMESET, which is a list of time references, representing a set of observation occasions. For data issued by surveys, TIMESET is a subset of the survey's replications. OSI distinguishes between two basic kinds of information frames, INDIVIDUAL DATA and SUMMARY DATA. The former models collections of individual items (so-called micro data), such as List of Students enrolled in Degree Courses with their Sex, Age-class, and Degree Course, the latter models aggregated data. More precisely, SUMMARY DATA are used for modelling totally aggregated data (so-called macro data), such as Total number of Students enrolled in Degree Courses by Sex and Age-class, as well as semi-aggregated data, such as Total number of Students enrolled in Degree Courses by Sex, Age-class and Degree Course. Note that the individual data may have associations with statistical units among their components, while the summary data have such associations among their components when they represent semi-aggregated data. A SUMMARY DATUM corresponds to a unique STATISTICAL TABLE, from which it inherits the component terms.

33. Both INDIVIDUAL DATA and SUMMARY DATA are specified according to a template in which a STATISTICAL UNIT is mandatory, together with a TIMESET and the special numerical variable Weight. The other components of an INFORMATION FRAME definition may be NUMERICAL VARIABLEs as well as CLASSIFICATION VARIABLE/CLASSIFICATION couples. An IDENTIFIER/IDENTIFIER_SET couple is a mandatory component in an INDIVIDUAL DATA definition. Special components of an INFORMATION FRAME may be necessary for representing the associations with statistical units. They have a role that is similar to the role of the external keys in the relational database tables, and are obtained by means of composing the couple ASSOCIATION/STATISTICAL UNIT with a couple IDENTIFIER/IDENTIFIER_SET connected with the associated statistical unit. For SUMMARY DATA, the same aggregation operation, such as SUM, AVERAGE, is associated with each NUMERICAL VARIABLE.

As an example, let us consider the datum List of Students enrolled in Degree Courses with their Sex, Age-class, and Degree Course and assume that it has been obtained by means of observing those students whose names have been listed in a list of Student identifiers in a number of years, listed in list of survey replications, and that we have recorded the age class according to a set of five-years classes. This is an individual datum, which is specified as:

[STUDENT,
STUDENT_IDENTIFIER/LIST_OF_STUDENT_IDs,
SEX/SEX_CLASSIFICATION,
AGE_CLASS/FIVE_YEARS_CLASSES,
ENROLLMENT/DEGREE_COURSE*COURSE_IDENTIFIER/LIST_OF_COURSE_IDs
WEIGHT,
LIST_OF_SURVEY_REPLICATIONS].

As another example, let us consider the datum Number and total Turnover of Businesses by Dimension, and Economic Activity. It corresponds to a statistical table, which is referred to the statistical unit Business and is obtained by means of partitioning Businesses on the basis of the vector [BusinessDimension/DimensionGroups, EconomicActivity/NACEGroups], and properly applying the SUM function on the vector [Turnover, Weight], for each subset of Business in the obtained partition. Such a summary datum is specified as

[BUSINESS,
BUSINESS_DIMENSION/DIMENSION_GROUPS,
ECONOMIC_ACTIVITY/NACE_GROUPS,
SUM(TURNOVER),
SUM(WEIGHT),
LIST_OF_SURVEY_REPLICATIONS].

IV. FINAL REMARKS

34. ISTAT is employing resources in developing metadata management systems, building a proper organisation for metadata capturing, launching metadata harvesting activities. Because of these priorities we could not undertake that systematic analysis effort which is required in order to define an ontology for statistical activity. Until now we focused our work on defining models for some important classes of metadata, namely conceptual metadata concerning data semantics on one side, the production processes on the other side. Nevertheless the ISTAT strategy for metadata strongly implies the definition of a general ontology for statistical activity. Indeed our opinion is that such an accomplishment can only result from the joined efforts of several subjects, experts as well as organisations. In fact, as ISTAT experience shows, any single subject has its own vision of the statistical activity and its own priorities. However in our opinion to define a general ontology for statistical activity is a difficult but affordable task. Perhaps only some lasting uncertainty, about both the goals to be reached and the conceptual tools to be exploited, prevented the statistical metadata community from attaining definitive results. The variety of conceptual approaches which are proposed by the computer science research, by the statisticians, as metadata users or metadata experts, by the market may have increased the confusion.

35. In principle to define an ontology for statistical activity requires in-depth analysis. According to a top-down approach performing such an analysis is the main way to work, however it is neither the only one nor the easiest one. It is difficult to perform a thorough analysis of such a wide field, by means of taking into account all the existing viewpoints. Bottom-up approaches which start from the comparison of existing models and terminologies, in the latter case after having specified their underlying reference models, are perhaps more feasible.

36. We maintain that the quality of metadata has several important dimensions: the conceptual soundness, the accuracy, the exchangeability. In our opinion the conceptual soundness is warranted by means of clearly distinguishing among classes of metadata and using a proper conceptual model for specifying metadata in each class. The accuracy is ensured by means of the metadata management organisation, which should warrant an efficient updating of the metadata collections. The functional integration of the metadata management systems with the tools for producing and exploiting statistical data is a way for pursuing such a goal. The main conceptual requisite for exchangeability, in particular semantic interoperability, is the definition of a common reference ontology.