Developing a Metadata Infrastructure for Official data: the ISTAT experience

Giovanna D'Angiolini

e-mail dangioli@istat.it
Istituto Nazionale di Statistica (ISTAT), via C. Balbo 16
00186 Roma, ITALY,

Abstract. Mining official data implies retrieving data from different surveys or administrative sources and properly interpreting them as measures of observed phenomena. In order to support such an activity, ISTAT is developing a metadata infrastructure which is based on two centralized metadata management system, and implies the development of tools for exploiting metadata in the data manipulation activities. Such a strategy is supported by the definition of a conceptual metadata framework together with proper documentation models for each class of metadata.

1 The ISTAT Strategy for Metadata Management: choices and involved problems

Mining official data implies retrieving data from different surveys or administrative sources and properly interpreting them as measures of observed phenomena. Such an activity requires the availability of several classes of metadata concerning the characteristics and the information content of each exploitable source of information. To ensure the dissemination of such metadata to the data users is a primary task for National Statistical Institutes (NSI), nevertheless to introduce metadata management practices in the official data production is often a challenge. Most NSIs consider the development of a metadata infrastructure a long-term goal, which requires a carefully devised strategy. Moreover the increasing need for integrating data from several sources obliges the NSIs to pursue a policy of centralised metadata management. By means of homogeneously documenting data from different sources in a unique environment, a centralised metadata system provides the rough material for data integration.

Therefore 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. They will disseminate such metadata to both data users and survey designers. Moreover they are conceived as metadata servers for those data management systems and software tools which are exploited in the data production and dissemination activities. Another important experience which we are carrying on

is the development of ESPLORIS, a system which exploits metadata for assisting the user in extracting data from data collections produced by several sources. As analysts of real world phenomena, the end users of official data have a basic requirement: to search the large data collections which are produced by NSIs for retrieving those data which are better suited to the analysis of a given class of phenomena, and properly transform them. This is an exploratory activity, which requires the availability of proper metadata. Many OLAP/DW tools exist which are well suited to perform data exploration inside an organisation. Inside a single organisation the role of the different classes of users constrains their information requirements. On the contrary the official data users have unpredictable information requirements, therefore a richer amount of metadata is required for steering them in retrieving and transforming data. This is the reason why NSIs should invest in developing specialised tools for supporting the data users' exploratory activities. Moreover tools for metadata-based data retrieving are an important component of a general metadata infrastructure because a good metadata quality is ensured only if metadata are actually used in data manipulation activities.

Last but not least, a NSI's metadata strategy requires a sound conceptual foundation. This implies answering such questions: which are the relevant classes of metadata for properly retrieving, transforming, analysing official data? How should we model such metadata?

In sections 2 and 3 we outline our general conceptual framework for metadata specification and present our model OSI, which we use for modelling the information content of surveys and statistical information systems. In section 4 we discuss the main characteristics of the systems which compose our metadata infrastructure, namely SIDI, SDOSIS and ESPLORIS.

2. Classes of Metadata for Documenting Official Data

There is a general agreement on the need for a general conceptual framework which specifies relevant classes of metadata, starting from an in depth understanding of the role of metadata in the data production and analysis activities (see for example [13], [14], [18]) In our opinion, a general conceptual framework for metadata specification should specify classes of metadata and relationships among them according to several dimensions, and provide the metadata managers with documentation models for each singled out class of metadata.
Moreover, the different contexts in which metadata are used and the different ways of communicating and using metadata should be analysed.

We propose to single out relevant metadata classes on the basis of such dimensions: the content of metadata, their level of abstraction, their scope, that is "what metadata describe", "how we look at metadata", "which extents of data are described", respectively .

Our specification of the content of metadata, that is, "what metadata describe", 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 proper 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 typically produced by NSIs for public usage.

These concepts define the contexts which are described by metadata: the documented 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 way it has been observed.

Therefore two main classes of metadata are singled out, concerning the information content of a source or a SIS and the characteristics of each source as an observation process, respectively. These are the basic metadata classes. Other relevant metadata concern the quality and the characteristics of the issued data when regarded as the result of a particular repetition of a production process. Such metadata have a different explanation and specification, therefore we separately analyse them. An outcome of our work on such other classes of metadata is our implemented system SIDI, which manages quality indicators together with metadata about the survey production process, in an integrated way (see [8]). Further analysis of the two main metadata classes can be developed along two other dimensions: the level of abstraction and the scope.

A level of abstraction defines a particular “way of looking at" the objects which we want to represent. By means of considering sources and SIS at different levels of abstraction we single out three main metadata layer: the conceptual layer, the organisational layer and the operational layer.

Sources and SIS as knowledge bases are the main objects which are defined in the conceptual layer. Proper classes of conceptual metadata describe the production process which is associated with each source as well as the information content of sources and SIS.

We single out two components in the specification of the information content of sources and SIS. The first component encompasses the specification of the observed part of the real world, which is expressed in terms of elementary statistical concepts such as statistical unit, variable, as well as the specification of structured objects which are derived from such elementary objects. Each concept has a definition and a set of links with the other concepts. We call such a specification a terminology. The second component is the specification of the issued data: their meaning is documented by means of representing 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, however sources can share production procedures and terminology concepts, but they produce their own information frames.

At lower levels of abstraction we describe the distribution of SIS among organisations and the implementation of SIS and sources as data production, transformation and dissemination systems.

In the organisational layer we distinguish among data producers, data users and other organisations, in the operational layer we document the physical data repositories and the data manipulation procedures, that is, those data management and

manipulation systems which are built for supporting the production and the usage of statistical data.

Along the dimension "scope" we distinguish between local and global metadata (see [14]). In the conceptual layer, local metadata are those metadata which describe the information content and the characteristics of each source of information. Global metadata are those metadata which 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. Conceptual integration (or, equivalently, harmonisation) is the activity which is performed in order to produce global metadata concerning the information content. It is a complex activity, which requires comparing and matching the conceptual objects which belong to different source terminologies, by means of analysing their definitions; for this purpose, the object definitions must be expressed in a structured shape, as combinations of formally defined constructs. This is the reason why inside an NSI's collection of sources there is never complete conceptual integration of the terminologies and generally only delimited areas of integrated concepts are defined, corresponding to SIS or other sets of partially integrated surveys 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 shared descriptions of operations (see [8]). In the organisational and operational layer, global metadata should describe those interactions among organisations, processes, data and agents which are involved in statistical data production and usage.

In our metadata framework each metadata class plays the role of a homogeneous viewpoint with which we can associate a documentation model. A documentation model is the specification of a set of meta-objects which have their own meta-properties and are linked by meta-relationships. It allows for specifying in a standard way those metadata which belong to a given metadata class.

3 Modelling metadata

As a consequence of our activity we have especially studied how to model the production process from a documentation viewpoint, for SIDI, and how to model the information content of surveys and other sources, for SDOSIS and ESPLORIS.

SIDI is based on a conceptual layer model which allows for associating each survey with a set of OPERATIONS (see [8]). 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 which are performed for monitoring the production procedures. Operations and control actions are performed by AGENTS, moreover they produce and exploit DATA REPOSITORIES.

We have defined a conceptual layer model, called OSI (Objects-Information Frames), for specifying terminologies and information frames of surveys and SIS (see [7] and [3]). OSI aims to support analysing concepts for integration as well as performing complex data manipulation activities. Therefore it strongly differs from the OLAP data model (see [6], [11]), based on the notions of cube, fact, dimension,

which is only aimed to let the analyst perform simple operations on pre-defined data marts. Unfortunately, to support non planned data manipulation is not an objective of the OLAP/DW research, even if many researchers in this area (see for example [1], [4], [10]) stress the need for richer conceptual specification languages. OSI borrows some concepts from those conceptual models which were proposed in the statistical database research area, whose main goal, however, was to support an operational layer activity, namely statistical database design activity (see for example [15], [16], [17]).

The main particular features of our model depend on its underlying analysis-oriented conceptualisation (for similar approaches see for example [2], [5], [9], [13]). From this viewpoint each source is regarded as a different way of observing the reality of interest, data from distinct sources are evaluated and used differently. 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 important 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 most of the existing OLAP tools. As an example, let us consider two data such as Total number of University Students by Sex and University Location and Total number of Persons by Age-Class, Sex and Person Residence. Let us simply define Sex and University Location as dimensions for the fact table Total number of University Students by Sex and University Location, and Sex and Person Residence as dimensions for the fact table Total number of Persons by Age-Class, Sex and Person Residence. In this case, the name of the variable is the name of the dimension. Let us suppose that we want to calculate the derived indicator Number of University Students/ Number of 19-25Aged Persons, by Sex and Region. Given the above definition of dimensions, we cannot state if this indicator may be obtained, because we do not know if the two fact table share the classification Region. A different choice is possible, that is, to define Region as a shared dimension for the two fact tables instead of University Location and Person Residence respectively: in such a way however, the exact meaning of the data is not conveyed to the user. 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). The available classifications are established on the basis of the analysts' goals. Therefore, the latter ones are basic meta-objects when the data are actually modelled for satisfying analytical purposes. It is worth noting that in our vision the ultimate goal of the metadata representation is to enable the data user to directly build new data by means of performing meaningful data manipulations. From this viewpoint the problem of the poor definition of the Fact/Dimension model cannot be solved by means of simply adopting proper naming conventions for the modelled

objects because the data user may need to manipulate variables and classification as distinct objects, as in the above example.

The terminology of a survey or SIS is a collection of concepts which describe the information content of the survey or SIS. A set of basic meta-objects is provided for describing the observed part of the real world, namely STATISTICAL UNIT, CLASSIFICATION VARIABLE, NUMERICAL VARIABLE, CLASSIFICATION, IDENTIFIER_SET, ASSOCIATION. The observed part of the real world is described by specifying a network of elementary concepts, each one belonging to one of these meta-objects.

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

A CLASSIFICATION VARIABLE is a qualitative property of observable real world individual objects, such as Sex, Economic Activity, which can be used for classifying statistical units. *Identifier* is a special CLASSIFICATION VARIABLE, which is used to name the single items of a statistical unit.

A NUMERICAL VARIABLE is a quantitative property of observable real world individual objects, such as Family Income, Turnover, on which simple summarising 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 which is used for counting the number of items of a statistical unit.

A CLASSIFICATION is a set of states which can be observed for some qualitative property of observable real world individual objects. Each classification is associated with an extension which is a list of names of states; as an example, Sex-classification is associated with the set {male, female}. It is well-known that classifications can be organised in CLASSIFICATION SYSTEMS, which establish a set of hierarchical relationships between classifications. An example of official classification system is the NACE classification for the economic activity.

An IDENTIFIER_SET is a name for a set of names of observable real world individual objects. Each IDENTIFIER_SET is associated with an extension which is a list of names of individual objects.

A CLASSIFICATION VARIABLE may be associated with one or more CLASSIFICATIONs, and in particular cases with IDENTIFIER_SETs; a CLASSIFICATION may be associated with one or more CLASSIFICATION VARIABLEs. *Identifier* may only be associated with one or more IDENTIFIER_SETs.

An ASSOCIATION is a one-to-many or a one-to-one relationship between statistical units. It is worth noting that for the analytical purposes it is convenient to regard any many-many relationship between statistical units as a statistical unit too. PART_OF, GROUP are names for special associations.

A SIS or survey terminology specifies relationships among real world objects belonging to one of the above meta-concepts (see at the end of the paper for a diagram which presents the main relationships among OSI meta-objects). STATISTICAL UNITs are associated with CLASSIFICATION 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. The properties of the documented real world objects encompasses the NAME of the represented object with its SYNONIMS, the OBJECT-DEFINITION, which is expressed in terms of other objects, one or more relationships with other objects of the above described types. An important feature of the OSI model is the availability of a set of CONCEPT DEFINITION CONSTRUCTS, which are used for specifying the definitions of terminology objects in a formal way, so as to support their analysis and conceptual integration. These constructs are also used for deriving new elementary objects from the available ones.

OSI encompasses two other classes of meta-concepts, which are used to describe CONSTRAINTS on the terminology objects and their relationships, and for defining STRUCTURED OBJECTS which are built by associating terminology objects, respectively.

Example of CONSTRAINTS concerning terminology objects are STATISTICAL UNIT PARTITION RELATIONSHIP, NUMERICAL VARIABLE SUM RELATIONSHIP.

A STATISTICAL UNIT PARTITION RELATIONSHIP may be established between a statistical unit and a vector of statistical units which are connected by IS_A (subset) relationships with the given statistical unit. Such a constraint states that the vector of statistical units is a partition of the given statistical unit. A vector of couples CLASSIFICATION VARIABLE/CLASSIFICATION may have been used for defining the partition. A simple example is the partition of Person into Male with Age_class≥14, Male with Age_class<14, Female with Age_class≥14, Female with Age_class<14, which is based on the vector [Sex/SexClassification, Age_class/Age_classGroups].

A NUMERICAL VARIABLE SUM RELATIONSHIP is established between a numerical variable and a vector of numerical variables, when the given numerical variable corresponds to the sum of the vector components.

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

The meta-object STATISTICAL TABLE describes a common outcome of a basic data manipulation operation, namely the result of applying a summarising function such as SUM, AVERAGE on the values of a numerical variable or a vector of numerical variables which are associated with the instances of one or more sets. In a statistical context the sets of interest for a summarising operation are statistical units or components of a vector of statistical units, in the latter case, the vector of statistical units is defined by means of establishing a partition relationship with a given statistical unit. As a consequence, generally a statistical table is a collection of elementary components, where each component is the result of applying the given summarising function on one of the given numerical variable for one of the defined subsets of a partitioned statistical unit. An example is the object Number and total Turnover of Businesses by Dimension and Economic Activity. In its definition it is implicit that we have a set of Businesses with their Dimension, Economic Activity, Turnover and Weight, and two classifications for Business Dimension and Economic Activity, for instance Dimension Groups and NACE Groups. We make two

operations: a) partitioning Businesses according to the vector [BusinessDimension/DimensionGroups, EconomicActivity/NACEGroups], b) for each component of the attained vector of statistical units, which is a partition of Businesses, 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 which represents the total turnover or the number of businesses for one of the subsets which we have obtained by partitioning Businesses in the described way.

Obviously statistical tables may be used to specify the content of the issued data, but issued data are not the only context in which statistical tables occur. In fact, representing statistical table in a terminology is mandatory because often a statistical source directly collects statistical tables instead of, or in addition to, information related to individual real world objects; moreover, components of statistical tables are often involved in the definition of elementary objects such as numerical variables. Another example of a STRUCTURED OBJECT is the RATIO meta-object, which is used for describing those statistical indicators, such as Number of Students for each Professor, which are obtained as ratios between two couples of the kind statistical unit/numerical variable (in the example, the ratio is between Student/Weight and Professor/Weight).

Having defined a terminology for a SIS or a single survey, we can describe the meaning of its issued data. Among such data we consider those data collections which are the direct output of the data capturing and editing procedures as well as those data collections which are the output of further transformation procedures. At a conceptual layer, we specify the content of the issued data as a set of interrelated INFORMATION FRAMES. An information frame is a conceptual object which is specified as a tuple of terms of a terminology. Moreover an information frame refers to a TIMESET, which is a list of temporal references, representing the set of its observation occasions. In most cases TIMESET is the same for all those data which have been issued by a survey which has been repeated several times in a given period. We single out two basic kinds of information frames, SET_OF_INDIVIDUALS and SUMMARY. The former models collections of individual items (so-called microdata), such as List of Students with Sex and Age-class, for each Degree Course, the latter models data which have been obtained by means of summarising pre-existing collections of individual items. More precisely, SUMMARYs are used for modelling so called macrodata, such as Total number of Students by Sex and Age-class in Italy, as well as pre-aggregated data, such as Total number of Students by Sex and Age-class, for each Degree Course, which have links with microdata, in the example List of Degree Courses. It is worth noting that the conceptual relationship linking two information frames is the same which links the statistical units to which they refer (Enrolled in the example). It is worth noting that another kind of information frame, which we do not analyse in this paper, is used for modelling those data which are obtained by means of combining macrodata, such as indicator tables.

Both SET_OF_INDIVIDUALS and SUMMARY information frames 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 VARIABLES/CLASSIFICATION couples. An

Identifier/IDENTIFIER_SET couple is a mandatory component in a SET_OF_INDIVIDUALS definition. For SUMMARYs, an operation is associated with each NUMERICAL VARIABLE, such as COUNT, SUM, AVERAGE.

A SET_OF_INDIVIDUALS denotes a set of tuples whose components are single instances of the specified component variables, that is, modalities of classifications or values in the domain of a numerical variable. Each tuple contains the instances of the specified concepts which have been collected for an observed individual instance of the specified statistical unit. As an example, let us consider the datum List of Businesses with their Dimension, Economic Activity, Turnover and assume that it has been obtained by means of observing those businesses whose names are listed in List of Businesses Identifiers in those occasions which are listed in List of observation occasions, and that we have recorded the business dimension according to a Business Dimension Classification and the business economic activity according to NACE Groups. Such a datum is specified as

[BUSINESS,
IDENTIFIER*LIST_OF_IDENTIFIER,
BUSINESS_DIMENSION*BUSINESS_DIMENSION_CLASSIFICATION,
ECONOMIC_ACTIVITY*NACE_GROUPS,
TURNOVER, WEIGHT,
LIST_OF_OBSERVATION OCCASIONS].

A SUMMARY denotes a set of tuples which arranges the components of a statistical table. As an example, let us consider the datum Number and total Turnover of Businesses by Dimension, and Economic Activity. It corresponds to the statistical table which is defined for the statistical unit Business in the above described way, 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 datum is specified as

[BUSINESS,
BUSINESS_DIMENSION*DIMENSION_GROUPS,
ECONOMIC_ACTIVITY*NACE_GROUPS,
SUM(TURNOVER),
SUM(WEIGHT),
LIST_OF_OBSERVATION OCCASIONS].

This SUMMARY denotes a set of tuple, in which each tuple is referred to a component of the partitioning vector of the specified statistical unit, Business. Therefore each tuple is composed by that particular combination of modalities of the specified couples BusinessDimension/DimensionGroups and EconomicActivity/NACEGroups which uniquely identifies the referred component of the partitioning vector, together with the values of SUM(TURNOVER) and SUM(WEIGHT) which have been calculated for such a component.

It is worth noting that, despite their different meaning, the two kinds of information frames can be processed in a very similar way. This is a feature that OSI shares with the Fact/Dimension model which is generally used in OLAP applications.

In order to describe the whole set of data issued by SIS or by single sources of statistical information we need to specify relationships among information frames.

Two SET_OF_INDIVIDUALS may be connected by an IS_A (subset) relationship, which is the same which connects their statistical units.

Two SET_OF_INDIVIDUALS may be connected by an ASSOCIATION, which is the same which connects their statistical units. A SUMMARY may be connected with a SET_OF_INDIVIDUALS by an ASSOCIATION, in the role of son. In these cases, special CLASSIFICATION VARIABLE/IDENTIFIER_SET couples occur in the information frame which has the role of son, where IDENTIFIER_SET is referred to the father information frame. In particular, this is the case when a SUMMARY is used for modelling pre-aggregated data.

The data which are observed and released by each source are thoroughly described as associations of terms, when their meaning is what matters.

However, such data are also characterised by the operations which produced them as well as by the operations which can modify them. OSI specifies all the admissible TRANSFORMATIONS which can be applied on information frames. Generally speaking an information frame is obtained by another information frame by means of applying transformations which belong to one of these classes:

a) simply ruling out components: CLASSIFICATION VARIABLE/CLASSIFICATION couples and NUMERICAL VARIABLES for SET-OF-INDIVIDUALS, only NUMERICAL VARIABLES for AGGREGATES;

b) summarising by means of selecting a subset of CLASSIFICATION VARIABLE/CLASSIFICATION couples and then applying a summarising function;

c) summarising by means of choosing, for any CLASSIFICATION VARIABLE/CLASSIFICATION couple, another existing classification which has the role of father in an AGGREGATION relationship with the given classification and then applying a summarising function;

d) selecting a subset of observation occasions;

e) applying elementary transformations to the information frame components which implies deriving new elementary objects: new CLASSIFICATIONs, new CLASSIFICATION VARIABLEs, new NUMERICAL VARIABLEs, or a new reference STATISTICAL UNIT. In all these transformations the OSI concept definition constructs are involved. In particular, a new statistical unit is built by means of specifying selection criteria which involve the given CLASSIFICATION VARIABLEs, or derived CLASSIFICATION VARIABLEs.

OSI provides the data user with a set of information frame transformations which include a rich set of CONCEPT DEFINITION CONSTRUCTS for deriving new objects from the observed ones and therefore is richer than the set of OLAP operations.

4. The ISTAT metadata infrastructure

SIDI (see [8]) is the component of the ISTAT metadata infrastructure which is dedicated to the specification and maintenance of metadata concerning the survey production processes. Indeed SIDI has been designed as a tool for monitoring the

quality of the ISTAT surveys from both a qualitative and a qualitative viewpoint. Therefore it is not only a metadata management system, it also allows for calculating and disseminating standard quality indicators for each ISTAT survey. As a metadata management system, SIDI warrants a standard specification of the survey production processes, which is ensured by means of a network of thesaura. For each meta-concept in our model we have built in the SIDI database a thesaurus of admissible descriptions. In particular, thesaura have been defined for OPERATIONS and CONTROL ACTIONS and other auxiliary concepts. Thesaura have been defined for STATISTICAL UNITS and for the observed phenomena, too, which will be shared with SDOSIS. The most complex thesaura have been given a hierarchical structure, so as to steer the user in choosing the most suitable description, starting from general descriptions and navigating towards more precise concepts. The conceptual links among thesaura which are established by our model are represented in the database. 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 particular system user, the quality manager, whose role is to keep a good level of standardisation by means of properly managing the content of the thesaura network. This features 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 or phenomena, 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 navigate across its metadata and quality indicators, or select several surveys and compare their quality indicators.

At present SIDI is implemented and manages metadata describing the majority of the ISTAT surveys, we are now designing the first version of SDOSIS.

SDOSIS is aimed to document the information content of ISTAT surveys as well as the results of any integration activity. Future versions of SDOSIS will directly support the integration activity, by means of offering functionalities for the analysis of terminologies. The present version of SDOSIS is equipped with functionalities for specifying terminologies and information frames of surveys and SIS. In order to warrant a good quality of metadata SDOSIS manages some classes of operational metadata concerning input and output data repositories. In particular, the survey managers can describe the survey questionnaires and store their image, moreover they can specify the physical characteristics of those administrative archives which they exploit as data sources and document the way by which the survey data are disseminated: print tables, files, database relations, data marts. Moreover, the SDOSIS database encompasses a classification repository, in which the set of modalities of each documented classification is stored, together with correspondence tables for linking modalities of different classifications. Unlike the production process, the information content of surveys and SIS cannot be specified in a homogeneous way by means of pre-defined thesauri. A standard specification would require conceptual integration, which is only obtained by means of in-depth analysing the information content of the involved sources. SDOSIS manages a standard terminology, based on official standards, which is represented by means of a network of thesaura which store standard terms for statistical units, variables and classifications. However, the survey managers are not obliged to adopt such a terminology nor to define compatible terms.

They define the survey's own terminology and may declare, for each term, a correspondence with a standard term. As an alternative choice, they may declare a correspondence with a term in another survey's terminology, or with an area standard term, which is shared by a set of similar surveys. In such a way, SDOSIS documents all those situations in which a partial integration of surveys have been established. Because of the more complex context which it documents SDOSIS provides the end user with more inquiry functionalities. In particular, it offers two distinguished concept-based inquiry functionalities which exploit standard and non-standard terms, respectively. Both of them enable the user to choose terms in a network of term repositories concerning statistical units, numerical variables, classification variables, classifications and search for surveys whose description matches the specified criteria. However the first functionality allows the user to choose terms in the standard terminology thesaura, while the other one allows the user to choose terms in repositories of non-standard terms. For the purpose of warranting meaningful inquiries, such repositories include those terms which are shared by several surveys as well as those survey terms which have no correspondence with other terms. After having selected a single survey of interest, the user can navigate across its terminology as well as view its information frames, and view the characteristics of the input and output data repositories. Proper inquiry functionalities are provided for the other system entities: SIS, standard terminologies, local standard terminologies.

As we discussed in the foregoing sections, we have decided to implement ESPLORIS after having observed that most OLAP/Data warehousing tools are not suitable for the requirements of the SIS users (see [3] and [13] for a similar approach). ESPLORIS is a specific tool for implementing multi-source SIS, which employs metadata for steering the users in selecting sources of information and extracting new data from data collections produced by several sources, through navigation and manipulation. ESPLORIS is built around a knowledge base (KB in the following) in which the information content of the implemented SIS is described in terms of the OSI model. The interaction with the data users is based on the conceptual metadata specification stored in the KB. The user interface represents the relevant classes of metadata and their relationships by means of graphs. Operational metadata which describe logical and physical structures and their correspondence with conceptual metadata are represented, too, for the use of several system components.

In the ESPLORIS knowledge base, each source of information is associated with its own information frames, but all information frames are defined by employing shared sets of statistical units, numerical variables, classification variable/classification couples. These sets of elementary concepts describe the real world which is observed by the SIS as a whole. Moreover, a unique network of classification systems is implemented and represented inside the knowledge base.

The system allows the data user to explore the whole set of conceptual metadata in the ESPLORIS KB. Such an activity produces a query on the Data Base component of the system, which stores the data issued from the information sources. To build such a query the user employs a graphical interface, which assists him/her in a step-by-step fashion. Use cases have been employed to model all phases of the user's activity. An interface panel corresponds to each phase.

Exploration of the conceptual domain and definition of the statistical unit and variables of interest: this is the user activity in the first interface panel, called *Navigation Panel*. The *Navigation Panel* presents the network of Statistical Units which are connected by means of IS-A relationships, together with the corresponding network of Information Frames. When the user selects a statistical unit, the Information Frames which have the selected Statistical Units as a component are enlightened. Once the end user has chosen a reference Information Frame, the system presents a star-shaped graph, in which the Association links between the selected Information Frame and other Information Frames are showed. In such a way the user can choose variables of interest among the Classification and Numerical variables of the selected Information Frame as well as build new variables by means of navigating across the linked Information Frames, therefore he/she is enabled to carry out a non-planned data manipulation. As an example, let us consider the variable Type of the Degree Course in which he/she is Enrolled, which is referred to the statistical unit Student. Such a kind of a variable is obtained by means of navigating along the association Enrolled between Student and Degree Course. The existing OLAP tools require to have it previously defined by a data warehouse administrator, in order to include it in a data mart. Thanks to its richer metadata specification and user-friendly interface, ESPLORIS allows the end user to on-line build such a variable when needed. All the components which have been selected in the *Navigation Panel* define the *Conceptual Query*, the first structured object built by the user.

Query building: this is the user activity in the second interface panel, called *Query Panel*. Here the user, starting from a *Conceptual Query*, defines the logical structure of target data, by means of performing classical OLAP operations, such as defining a new Statistical Unit on the basis of a selection condition, choosing more aggregated Classifications, summarising and ruling out variables. In such a way the second structured object is built, called *Logical Query*. On user request, the *Logical Query* is transformed in a set of SQL commands which are required for data retrieval, called *Physical Query*. Data retrieval: this is the user activity in the third interface panel, called *Presentation Panel*. The *Physical Query* is executed and the *Presentation Panel* returns the data set, which is finally presented to the user. In the current version, the available options are data visualisation and export in standard formats (Excel, text files).

Future versions of ESPLORIS will provide the end user with a richer set of transformation operations, in the *Information Frame Transformation* panel. Moreover it will possible to store the new created Information Frames, thus enabling the end user to dynamically build his/her own data marts.

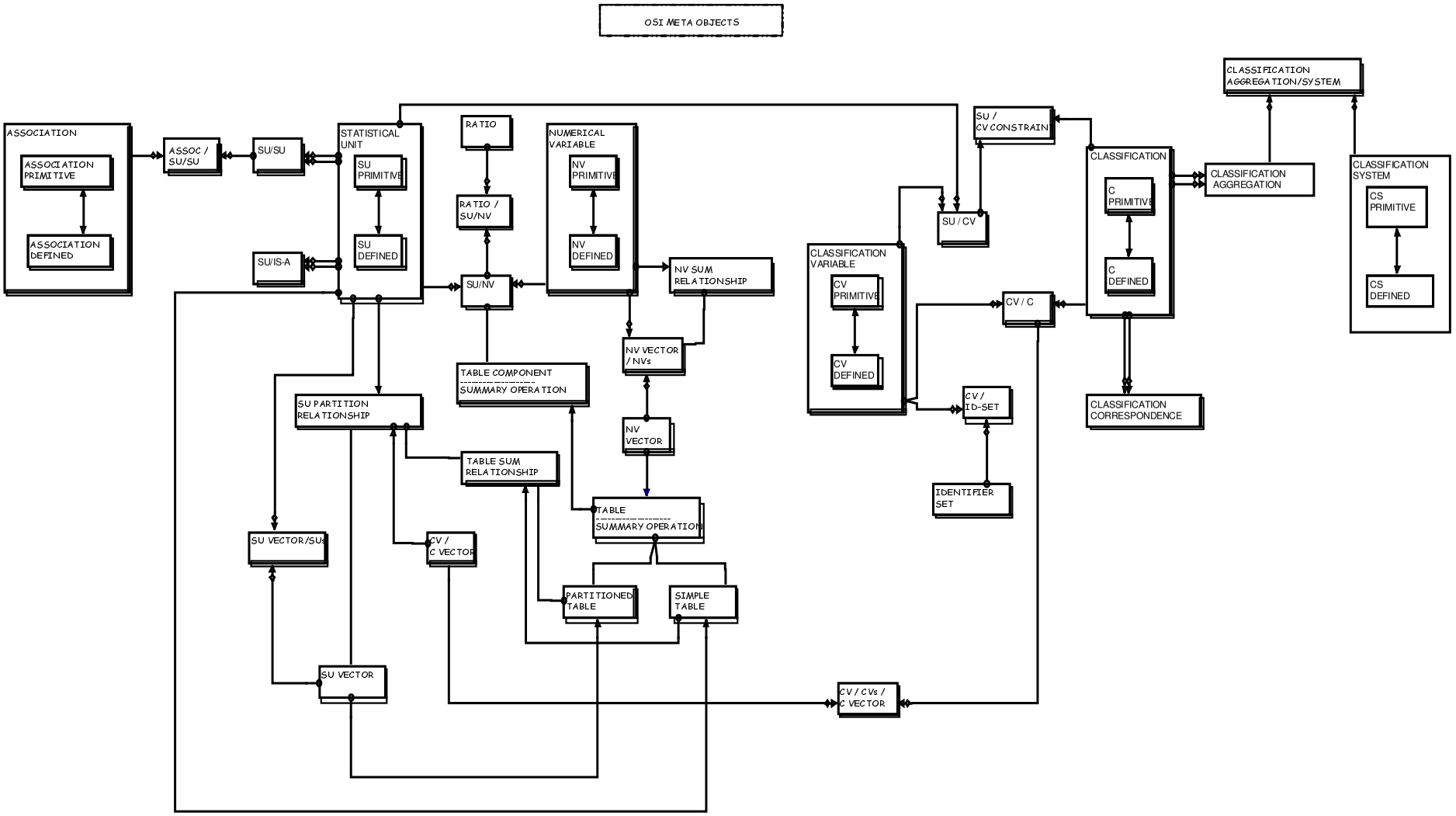
5. Future work

At present, to implement our devised metadata infrastructure by means of developing a metadata management infrastructure is our main activity. A related theoretical work aims to define a complete conceptual framework and a well established methodology for metadata specification, in particular for the information content metadata integration. This also implies documenting the relationships among

metadata at different levels of abstraction, in particular how to make the conceptual specification of data and processes correspond to the description of the same objects in operational terms, as sets of concrete procedures which have input and output datasets. There is another requirement for a metadata specification methodology: to define which views of the conceptual layer specification of data and processes should be communicated and used in different concrete contexts. Finally, what is most important is to promote the extensive usage of metadata in practical data manipulation activities: this is the only way to warrant a good quality of the defined conceptual metadata.

References

[1] R. Agrawal, A. Gupta, S. Sarawagi, “Modeling multidimensional databases,” IBM Research Report, 1995
[2] G. Appel, “Metadata Driven Statistical Information Systems”, Statistical Metainformation Workshop, Luxembourg 1993
[3] P. Barboni, G. D'Angiolini, L. Fanfoni, M. Paolucci, G. Sulsenti "ESPLORIS. Exploring Multi_source Statistical Information Systems through Metadata", NTTS-ETK Conference Crete 2001
[4] L. Cabibbo, R. Torlone, "A logical approach to multidimensional databases", EDBT98
[5] T. Catarci, G. D’Angiolini, M. Lenzerini, “Concept Language for Statistical Data Modeling”, Data and Knowledge Engineering, 1995
[6] E. F. Codd, “Providing OLAP (on-line analytical processing) to user-analysts: an IT mandate”, Technical Report, Codd and Associates, 1993.
[7] G. D’Angiolini, L. Fanfoni, M. Paolucci, "Modelling and Managing Metadata: the ISTAT experience", METANET Project Conference, Voorburg (Netherlands) 2001
[8] G. D’Angiolini, M. Fortini, M. Signore, (1996) “Metainformation Management Systems in the Survey production Process: A System for Survey Quality Control”, Proc. 2nd ASC
[9] G. De Giacomo, P. Naggar, “Conceptual Data Model with Structured Objects for Statistical Databases”, SSDBM 1996
[10] M. Jarke, M. Lenzerini, Y. Vassilou, P: Vassiliadis "Fundamentals of Data Warehouses" Springer- Verlag 1999
[11] R. Kimball, L. Reeves, M. Ross, W. Thornthwaite, "The Data Warehouse Lifecycle Toolkit", Wiley & Sons, 1998
[12] SJ. P. Kent, M. Schuerhoff (1997) "Some Thoughts about a Metadata Management Systems", SSDBM 1997
[13] S. I. Mc Clean, W. Grossman, K. A. Froeschl, (1998) "Towards Metadata-Guided Distributed Statistical Data Processing" NTTS '98
[14] ONU-ECE - "Guidelines for the modelling of Statistical Data and Metadata", 1995
[15] M. Rafanelli, A. Shoshani, “STORM: A Statistical Object Representation Model”, SSDBM 1990
[16] A. Shoshani, "Statistical databases and OLAP: similarities and differences", International Conference on Information ad Knowledge Management, 1996
[17] S.Y.W. Su “SAM*: A semantic association model for corporate and scientific/statistical databases”, Information Sciences, vol.29, No 2-3, May-June 1983
[18] B. Sundgren, “Some properties of statistical information: Pragmatic, Semantic and Syntactis”, Statistics Sweden 1991


OSI META OBJECTS
ASSOCIATION
ASSOCIATION PRIMITIVE
ASSOCIATION DEFINED
ASSOC / SU/SU
SU/SU
SU/IS-A
STATISTICAL UNIT
SU PRIMITIVE
SU DEFINED
RATIO
RATIO / SU/NV
SU/NV
NUMERICAL VARIABLE
NV PRIMITIVE
NV DEFINED
NV SUM RELATIONSHIP
SU / CV CONSTRAIN
SU / CV
CLASSIFICATION VARIABLE
CV PRIMITIVE
CV DEFINED
CV / C
CV / ID-SET
IDENTIFIER SET
CLASSIFICATION
C PRIMITIVE
C DEFINED
CLASSIFICATION CORRESPONDENCE
CLASSIFICATION AGGREGATION/SYSTEM
CLASSIFICATION AGGREGATION
CLASSIFICATION SYSTEM
CS PRIMITIVE
CS DEFINED
NV VECTOR / NVs
NV VECTOR
TABLE COMPONENT
SUMMARY OPERATION
SU PARTITION RELATIONSHIP
TABLE SUM RELATIONSHIP
TABLE
SUMMARY OPERATION
SU VECTOR/SUs
CV / C VECTOR
PARTITIONED TABLE
SIMPLE TABLE
SU VECTOR
CV / CVs / C VECTOR