UNITED NATIONS ECONOMIC COMMISSION FOR EUROPE
CONFERENCE OF EUROPEAN STATISTICIANS

Working paper
27 July 2016

Workshop on Implementing Standards for Statistical Modernisation,
21 – 23 September 2016

The Generic Statistical Information Model (GSIM) and the Sistema Unitario dei Metadati: state of application of the standard

Mauro Scanu, Cecilia Casagrande***

[*] ISTAT, Department for data collection and the development of methods and technologies for the production and dissemination of the statistical information, Rome, ITALY – email: scanu@istat.it

[**] ISTAT, Department for data collection and the development of methods and technologies for the production and dissemination of the statistical information, Rome, ITALY – email: casagran@istat.it

1. Introduction

The Generic Statistical Information Model GSIM (UNECE, 2013b) is a reference framework of internationally agreed definitions, attributes and relationships that describe the pieces of information that are used in the production of official statistics. It aims at modelling "information objects", i.e. data, metadata, as well as the rules and parameters needed for production processes to run (e.g. data editing rules). In the current version GSIM identifies around 110 information objects, which are grouped into four broad categories: Business, Exchange, Concepts and Structure. The information objects are then used in order to describe the inputs and outputs of the statistical process phases, as defined in the Generic Statistical Business Process Model (GSBPM, UNECE 2013a). Hence, GSIM contains many concepts useful for describing the input and output of different phases from the assessment of user needs to data dissemination.
These internationally agreed concepts have been used for the construction of the Sistema Unitario dei Metadati, for the part devoted to structural metadata (SUM-MS, Signore et al, 2015, Scanu, 2015). This system contains metadata that define data in their data structures (i.e. micro data sets or macro data tables and hypercubes). The SUM-MS aims at ensuring data retrieval and usability (by associating proper meaning to data), allowing metadata reuse (in order to harmonize concepts), documenting traceability (with the objective of statistical process transparency and process automation) and performing integration (with the aim of making the different sectors within a statistical institute speak with one voice and support standardization). The following sections will show how GSIM has been used, and the enhancements that we put in practice in order to fulfil the SUM-MS goals.

As shown in the figure below (see Figure 1), SUM-MS contains metadata describing data produced only in some of the GSBPM phases, i.e. from collection process step to dissemination process step. At the moment, in terms of input and output specifications, SUM-MS has a total coverage of the output produced in the dissemination phase and partial coverage with respect to the input and the output of the validation's phase.

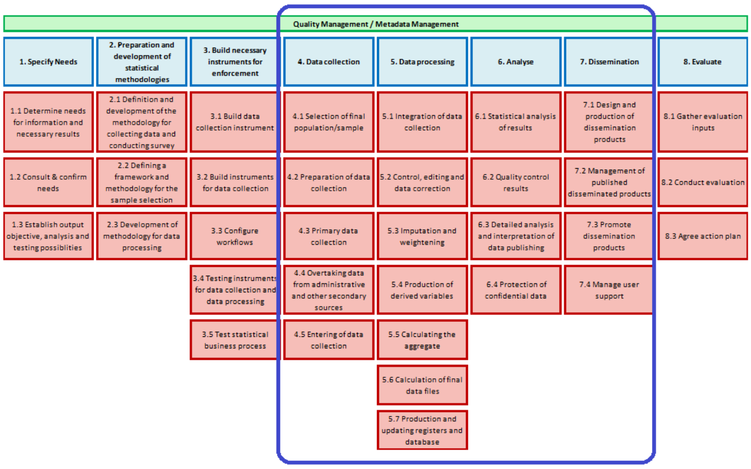

Quality Management / Metadata Management
1. Specify Needs
1.1 Determine needs for information and necessary results
1.2 Consult & confirm needs
1.3 Establish output objective, analysis and testing possiblities
2. Preparation and development of statistical methodologies
2.1 Definition and development of the methodology for collecting data and conducting survey
2.2 Defining a framework and methodology for the sample selection
2.3 Development of methodology for data processing
3. Build necessary instruments for enforcement
3.1 Build data collection instrument
3.2 Build instruments for data collection
3.3 Configure workflows
3.4 Testing instruments for data collection and data processing
3.5 Test statistical business process
4. Data collection
4.1 Selection of final population/sample
4.2 Preparation of data collection
4.3 Primary data collection
4.4 Overtaking data from administrative and other secondary sources
4.5 Entering of data collection
5. Data processing
5.1 Integration of data collection
5.2 Control, editing and data correction
5.3 Imputation and weightening
5.4 Production of derived variables
5.5 Calculating the aggregate
5.6 Calculation of final data files
5.7 Production and updating registers and database
6. Analyse
6.1 Statistical analysis of results
6.2 Quality control results
6.3 Detailed analysis and interpretation of data publishing
6.4 Protection of confidential data
7. Dissemination
7.1 Design and production of dissemination products
7.2 Management of published disseminated products
7.3 Promote dissemination products
7.4 Manage user support
8. Evaluate
8.1 Gather evaluation inputs
8.2 Conduct evaluation
8.3 Agree action plan

Fig. 1.1 Metadata collected in SUM-MS relative to some phases of GSBPM

2. Concepts useful for identifying a data set

Lines 47-50 in the GSIM Specification document (UNECE, 2013b) declares that: "Each data is a result of a *Process step* through the application of a *Process method* on the necessary *Inputs*". This is not something new. That sentence essentially describes in harmonized concepts the essential statistical task of transforming a data input into an output. Tracing back to Fisher (1925), "Statistics may be regarded i) as the study of populations, ii) as the study of variation, iii) as the study of methods of the reduction of data" meaning that the statistician focuses her/his attention on a population on which at least a variable is observed (input), studies the variation of the variable(s) of interest by summarizing this variation into something easier to manage (the process method: e.g. a cumulative distribution function, or some numbers that describe the cumulative distribution function characteristics, as the mean, median, variance, Gini index, …). Fisher was focusing only on the production of statistics, essentially data as available in the dissemination phase. GSIM has the merit of extending the same approach to the whole set of actions that characterize the statistical production. That sentence has also another merit: it declares what are the essential concepts to take into consideration in order to insert the data in an appropriate place in space (if the space considered is that of all the statistical programs as available in an NSI) and time. Taking in mind that *Process Step* and *Process method* are concepts to be used in a *Statistical Program*, the concepts that characterize the position of data in space and time are:

a. *Statistical Program* and *Statistical Program Cycle* (e.g. the Labour Force Survey, first quarter 2014);
b. *Process Step* (i.e. phase, e.g. collection or dissemination);
c. *Process Method* (e.g. web questionnaire filled in by a sample selected according to a survey design strategy in the collection phase, or an average in the dissemination case);
d. *Input*(s) (e.g. a questionnaire and a sample frame in the collection phase, or a numerical variable and the (sub-)population of interest on a validated data set in the dissemination phase).

If the a. and b. items are easy to include in this framework, the c. and d. items are less trivial. As a matter of fact, a data producer in a *Process Step* of a specific *Statistical Program cycle* can produce data sets with the same *Data Structure* (see Section 4) that differ for the *Process Method* or the *Input*. For instance, in passing collected to validated data, a data producer can produce data according to different check, edit and imputation procedures. The outputs are part of the same *Statistical Program*, *Statistical Program Cycle* and *Process Step*, make use of the same *Input*, consist of data structures of the same form (same population, same observed variables) but obtained through different methods, and this last element is the one that characterizes the different outputs as different data. The same concepts described here will be the ones used for describing the *Data Content* of macro data, see Section 3.4.

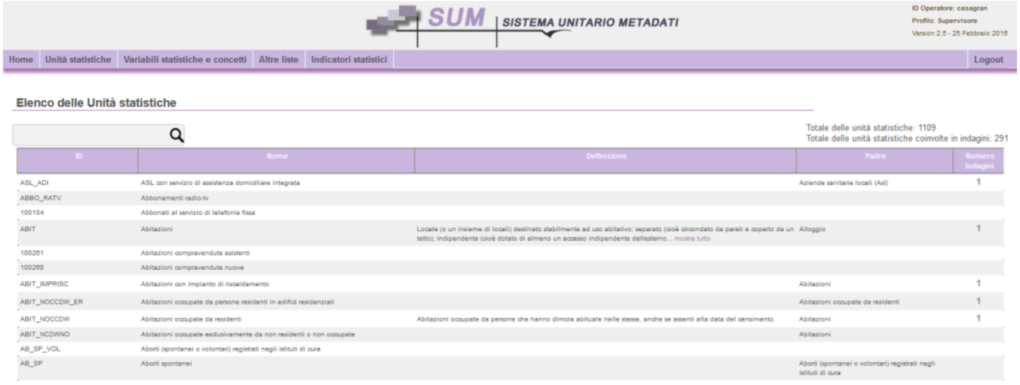
3. Concepts useful for describing statistical data

If the previous concepts place data in the correct space and time point, its content should be explained by other concepts. These are essentially in the *Concepts Group* in GSIM. Anyway there is an additional concept that Istat has implemented for describing macro data: the *Data Content*. This is the overview of the concepts used in SUM.

3.1 Population

The concept of Population coincides with the collective of reference of the datum. As planned in GSIM, the populations have a hierarchical relationship starting from the unit type concept. The population is defined on the basis of a certain criteria, such as all units that have a particular mode of a variable. For example the population of "Enterprise with more than 50 employees" is a subset of the more general unit type "enterprises".

The unit type is a more generic concept than that of the population, but extremely useful because it allows to identify the different types of populations which refer to a specific unit type. For example it is possible to investigate all types of populations related to women.



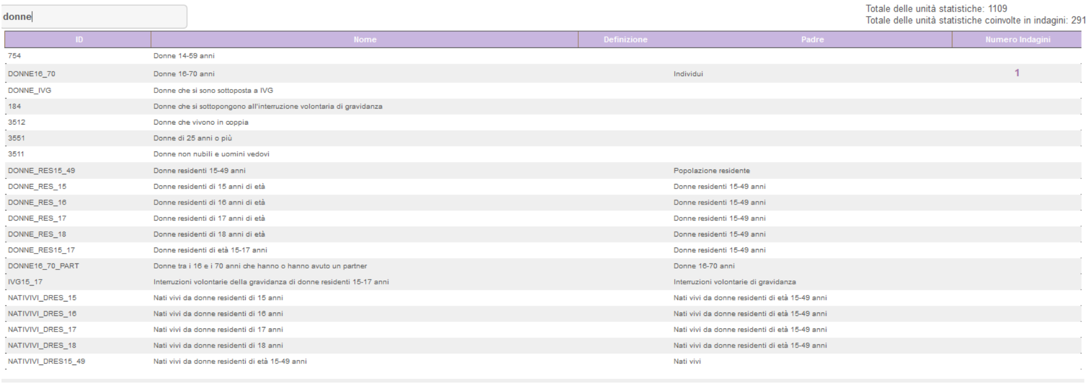
ID Operatore: casagran
Profilo: Supervisore
Version 2.5 - 25 Febbraio 2016

Elenco delle Unità statistiche

Totale delle unità statistiche: 1109
Totale delle unità statistiche coinvolte in indagini: 291

ID	Nome	Definizione	Padre	Numero Indagini
ASL_ADI	ASL con servizio di assistenza domiciliare integrata		Aziende sanitarie locali (Asl)	1
ABBO_RATV	Abbonamenti radio-tv			
100104	Abbonati al servizio di telefonia fissa			
ABIT	Abitazioni	Locale (o un insieme di locali) destinato stabilmente ad uso abitativo; separato (cioè circondato da pareti e coperto da un tetto); indipendente (cioè dotato di almeno un accesso indipendente dallesterno... mostra tutto	Alloggio	1
100261	Abitazioni compravendute esistenti			
100260	Abitazioni compravendute nuove			
ABIT_IMPRISC	Abitazioni con impianto di riscaldamento		Abitazioni	1
ABIT_NOCCDW_ER	Abitazioni occupate da persone residenti in edifici residenziali		Abitazioni occupate da residenti	1
ABIT_NOCCDW	Abitazioni occupate da residenti	Abitazioni occupate da persone che hanno dimora abituale nelle stesse, anche se assenti alla data del censimento.	Abitazioni	1
ABIT_NCDWNO	Abitazioni occupate esclusivamente da non residenti o non occupate		Abitazioni	
AB_SP_VOL	Aborti (spontanei o volontari) registrati negli istituti di cura			
AB_SP	Aborti spontanei		Aborti (spontanei o volontari) registrati negli istituti di cura	

Fig. 1.2 Unit type section in SUM-MS


donne

Totale delle unità statistiche: 1109
Totale delle unità statistiche coinvolte in indagini: 291

ID	Nome	Definizione	Padre	Numero Indagini
754	Donne 14-59 anni			
DONNE16_70	Donne 16-70 anni		Individui	1
DONNE_IVG	Donne che si sono sottoposta a IVG			
184	Donne che si sottopongono all'interruzione volontaria di gravidanza			
3512	Donne che vivono in coppia			
3551	Donne di 25 anni o più			
3511	Donne non nubili e uomini vedovi			
DONNE_RES15_49	Donne residenti 15-49 anni		Popolazione residente	
DONNE_RES_15	Donne residenti di 15 anni di età		Donne residenti 15-49 anni	
DONNE_RES_16	Donne residenti di 16 anni di età		Donne residenti 15-49 anni	
DONNE_RES_17	Donne residenti di 17 anni di età		Donne residenti 15-49 anni	
DONNE_RES_18	Donne residenti di 18 anni di età		Donne residenti 15-49 anni	
DONNE_RES15_17	Donne residenti di età 15-17 anni		Donne residenti 15-49 anni	
DONNE16_70_PART	Donne tra i 16 e i 70 anni che hanno o hanno avuto un partner		Donne 16-70 anni	
IVG15_17	Interruzioni volontarie della gravidanza di donne residenti 15-17 anni		Interruzioni volontarie di gravidanza	
NATIVIVI_DRES_15	Nati vivi da donne residenti di 15 anni		Nati vivi da donne residenti di età 15-49 anni	
NATIVIVI_DRES_16	Nati vivi da donne residenti di 16 anni		Nati vivi da donne residenti di età 15-49 anni	
NATIVIVI_DRES_17	Nati vivi da donne residenti di 17 anni		Nati vivi da donne residenti di età 15-49 anni	
NATIVIVI_DRES_18	Nati vivi da donne residenti di 18 anni		Nati vivi da donne residenti di età 15-49 anni	
NATIVIVI_DRES15_49	Nati vivi da donne residenti di età 15-49 anni		Nati vivi	

Fig. 1.3 Unit type section in SUM-MS

Up to now within the SUM-MS, about 549 reference populations have been identified. The system is arranged in this way: each population has an ID code of identification, a definition that describes its properties (this feature is partially developed) and the possible hierarchical relationship with another population. The hierarchical relationship is defined by the fact that there are more populations related to the same unit type, and these populations, as illustrated above, are identified based on the specific values of the variables associated with the unit type. Moreover, for each population the system shows the (micro or macro) data sets that refer to them as well as the surveys that produce data sets on them.

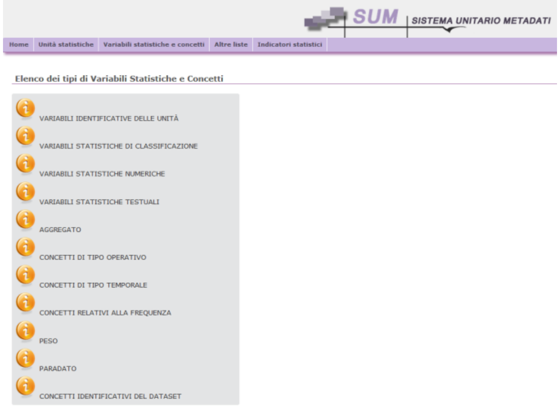
3.2 Variable

One of the main concepts on which GSIM has focused his attention and, consequently, also the SUM-MS has given importance, is the concept of variable. In a statistical process each unit of a population is associated with one or more characteristics. The association between the units of the population and the concept is called variable. GSIM distinguishes between the conceptual and representational levels of variables. At the conceptual level belongs the “Conceptual Domain” of the variables and at the representational level belongs the “Value Domain” of the variables. Among the variables as defined in GSIM, the SUM-MS takes note of different elements:

- Statistical variable: it is a characteristic of interest associated to each unit of a population for which a statistician is interested in representing its statistical distribution on the population, or a peculiar term of a distribution (frequency, mean, etc), i.e. an aggregate value (i.e. macro data, indicator, …). SUM-MS further specifies those statistical variables that are numeric, those that are textual (as the address, or those variables where it is possible to specify what “other” means) from those that are categorical. The last ones are associated to classifications. Note that only for categorical statistical variables it is possible to define appropriate levels of a classification, as specified in GSIM in the classification chapter.
- Identification variables: this variables are typically available in micro data sets and describe the association of each record to a unit in a population. It can be just one variable (e.g. a personal identification number) or a set of identifying variables (as an identifier for the household and a numeric value for each component). It is the primary source for linking records from different files.
- Time variable: the reference time of a survey is not a statistical variable, it is not the object of any statistical analysis. It is a specific element that allows to position data in time, as described in the beginning of Section 3: for instance it is useful to appropriately show in tables or figures aggregate data/indicators in time series.
- Operative variables: some macro data sets show explicitly how macro data have been obtained, usually for comparison. This is the case of the Adjustment used in time series, where data can be used as they were detected, or seasonally adjusted according to different criteria. Other examples are the Edition for the national accounts, the kind of currency used (current or in a specific year), the base year for index numbers.
- Indicator: again for macro data, this concept has usually a specific role as a dimension of a table. This is true for those macro data tables that show different indicators for comparison purposes. The name of that dimension will be clearly assessed in the Section on Data Content.

Currently in SUM-MS we started collecting Represented Variables: therefore, for instance, SUM-MS contains the terms "age of father" and "age of the mother" with respect to data on births, postponing in a second time the establishment of the corresponding conceptual variable, Age.

SUM-MS dedicates to the statistical variables and concepts a part of the system, in conformity to what describe before. As indicated with the figure below (see Figure 4) the concept “variable” included in GSIM has been split in statistical numerical variables, statistical categorical variables, statistical textual variables and other general concepts. The distinction between the categorical variables and other general concepts is governed by the possibility to associate the variable with a statistical classification with one or more levels (for the categorical ones) or with a code list.



Elenco dei tipi di Variabili Statistiche e Concetti

- VARIABILI IDENTIFICATIVE DELLE UNITÀ
- VARIABILI STATISTICHE DI CLASSIFICAZIONE
- VARIABILI STATISTICHE NUMERICHE
- VARIABILI STATISTICHE TESTUALI
- AGGREGATO
- CONCETTI DI TIPO OPERATIVO
- CONCETTI DI TIPO TEMPORALE
- CONCETTI RELATIVI ALLA FREQUENZA
- PESO
- PARADATO
- CONCETTI IDENTIFICATIVI DEL DATASET

Fig. 1.4 Types of variables section in SUM-MS

As arranged in GSIM, the “Represented Variable” has been associated with a “Value Domain” that can be enumerable or descriptive.

- Enumerable value domain: the variable, called categorical variable, can assume a finite number of different states on the population. These states correspond to the item of a classification (male and female for the "sex" variable detected on individuals)
- Descriptive value domain: the variable can take values according to a rule (for example not negative integers related to the variable "number of animals" observed on surveys of farms). These variables are typically numerical.

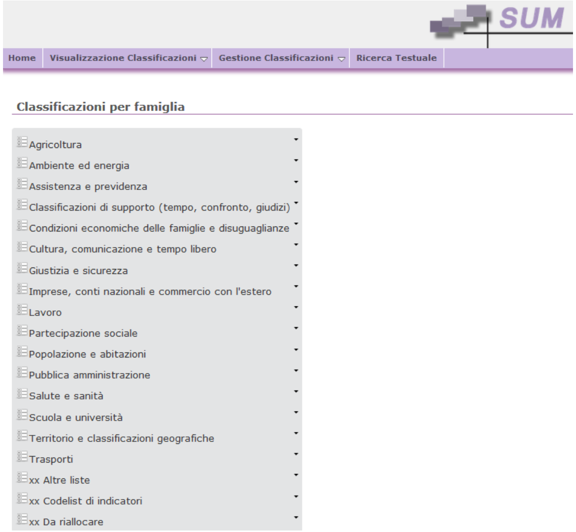
For the statistical variables with enumerable domain it is important to the management of classifications (see Section 3.3).

Among the represented variables, a prominent part is occupied by "derived variable": this variable is created in a step of the business process by applying one or more methods of transformation to other variables belonging to a previous phase. For instance, the variable "Age"- traceable in the dissemination phase - is constructed starting from the date of birth variable recorded in the collection phase.

The features, not yet completely developed, will provide, as well as for statistical units, the textual research of variable and concepts names, their definition, the possibility of study the number and which are the sources of information involved and the export in different formats.

3.3 Classification

Classifications are a key component of a metadata system. They have mainly a twofold role: on the first hand they are used in order to categorize in groups (or in other words “classify”) the unit of a reference population; on the other hand, they can be used in order to detail the outputs (macro data) of a statistical process. The SUM-MS makes available classifications according to the international standards, more precisely the Generic Statistical Information Model (hereinafter referred GSIM, UNECE, 2012) and the Neuchâtel Model (Netterstrøm, 2004).



Classificazioni per famiglia

- Agricoltura
- Ambiente ed energia
- Assistenza e previdenza
- Classificazioni di supporto (tempo, confronto, giudizi)
- Condizioni economiche delle famiglie e disuguaglianze
- Cultura, comunicazione e tempo libero
- Giustizia e sicurezza
- Imprese, conti nazionali e commercio con l'estero
- Lavoro
- Partecipazione sociale
- Popolazione e abitazioni
- Pubblica amministrazione
- Salute e sanità
- Scuola e università
- Territorio e classificazioni geografiche
- Trasporti
- xx Altre liste
- xx Codelist di indicatori
- xx Da riallocare

Fig. 1.5 Classification family section in SUM-MS

Each classification is embedded in a Classification Family, that groups classifications related to the same general theme. Up to now, SUM-MS contains 16 different themes (from agriculture to transportation, one of them is devoted to the "supporting classifications", i.e. those related to time, judgments, comparisons, that can be generally reused in different themes). For example, the Italian version of the NACE (ATECO) is contained in the enterprise and national accounts theme.

In the SUM-MS, classifications are composed of a set of schemes/versions which contain mutually exclusive and collectively exhaustive categories, which produce a partition of populations (Statistical Classification in GSIM).

The presence of several versions of the same classification is due to various reasons. First of all some classifications are related to a new version due to availability of different years related to each version (for example the NACE exists as NACE2002 version and as NACE2007 version, each one belonging to a specific reference year). It also considers the case in which more than one system of data production and/or dissemination adopts different versions of a classification. It also includes the case in which some changes occur in the items of the version (for a substantial modification or for an addition of new items) and it becomes necessary an upgrade of the version that has added to the previous one.

As far as versions are concerned, each one is characterized by two attributes: balanced (i.e. categories are mutually exclusive), and universal (i.e. each unit of a population can be associated to one category).

A version may have a linear structure or may be hierarchically structured, such that all categories at lower levels are sub-categories of categories at the next level up.

Each version is related to one or several code lists. The code lists are created to meet particular needs, that may be of statistical type or related to dissemination. Through the code lists, it is possible to split or to regroup each category in order to provide additions or alternatives to the reference standard version. Two types of code lists have been identified in SUM-MS, System code lists and data Structure code lists:

a. System code lists, containing all the codes of a data system actually in use on that system (e.g. all the NACE codes used by the Istat dissemination system, including groups of codes that do not form a level); this code list is the one actually used to derive code lists for the SDMX translation of a data structure.

b. Data structure code lists, containing only the items actually in use in a data structure (so that it is possible to know the cross-cutting level detail of a Data Content in terms of a classification variable).

SUM-MS introduces a new component in the classification system: simultaneous levels. These new components meet Institute needs to describe specific output.

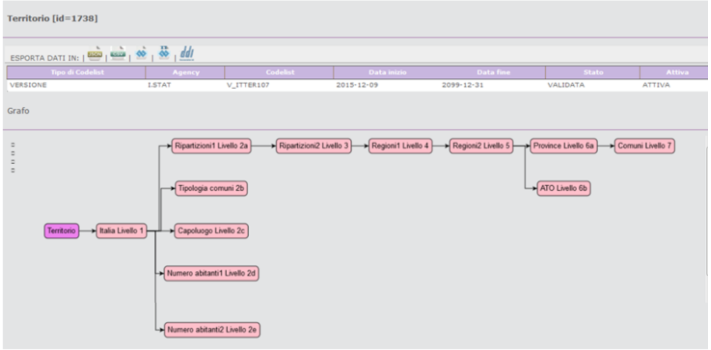
They provide one more tool than the standard GSIM. They are contained within the same classification and share the same basic atomic level (atomic item). They are constituted by different groupings of the atomic level which are not in a hierarchical relationship between them and so can exist in parallel.

Simultaneous levels can describe the relationship between levels just in case the classification categories are not necessarily organized in a hierarchical way.

As follows we are going to show a table containing the definitions used in the international standards also adopted by SUM-MS. This table aims to make a comparison and, at the same time, to underline both the differences in the nomenclature used by each system both the simultaneous levels component integrated in the SUM-MS.

GSIM	**SUM-MS**	**EXAMPLE**
Classification family	Classification family	Classification on Enterprises
Classification series	Classification	Economic activity (NACE)
Statistical classification /Version	Classification scheme/Version	Nace 2007 Nace 2002
Level	Level	Level composed by the first digit of the NACE 2007 codes (A – Agriculture)
	Simultaneous levels	Level 1a, Level 1b, … Atomic item in a territorial classification: Municipalities Simultaneous levels in a territorial classification: NUTS 1 and number of inhabitants (atomic level: municipalities)

Currently, as provided by GSIM, SUM-MS is developing the object types that has still left out by the system, i.e. classification index and classification index entry.



Territorio [id=1738]

ESPORTA DATI IN:

Tipo di Codelist	Agency	Codelist	Data inizio	Data fine	Stato	Attiva
VERSIONE	I.STAT	V_ITTER107	2015-12-09	2099-12-31	VALIDATA	ATTIVA

Grafo

Fig. 1.6 Structure of a classification with simultaneous level

3.4 An enhancement for macro data: the data content

The data content concept corresponds to the minimal but complete set of information on macro data. It corresponds to the title of tabular data when data was printed on sheet, and tables could not be too complex. Nowadays data hypercubes are very common, and they include the most diverse data in content and meaning. In order to identify and retrieve easily data, Istat has organized this minimal and complete set of information on the data as in Section 2 (with the exception of the *Statistical Program Cycle*, usually described in a specific time dimension in *Data Structures*, see Section 4). Hence, the *Data Content* is a structured concept and each *Data Content* item's main components are:

a. *Statistical Program*,
b. *Process Step* (although the phase is generally the dissemination one),
c. *Process Method*,
d. *Inputs*.

The data content is a code list, each item of this code list contains the description of the data content. In order to be complete, this description should contain all the a.-d. items described before.

Example - For instance "Households average monthly income, in thousands of Euro (source: SILC)" is a complete description of an output macro data: it contains the *Statistical Program* (SILC) and the description of the *Input* in terms of the reference population (households), the numerical variable (monthly income), the transformation method (average), the unit of measure (Euro), the unit multiplier (in thousands).

Each code of the *Data Content* code list is also linked to the corresponding concepts that define the a.-d. items. These links are the core for search tools and for data traceability. Any difference in one of the a.-d. items define a new data content. As a rule, data contents that are identical on the a.-d. items are the same, and should be fused in just one item.

Apart the a. and b. items that put the data in the correct point (in time and along a process) of a Statistical Program, the c. and d. issues are the actual statistical content of the data. There are essentially two types of data content.

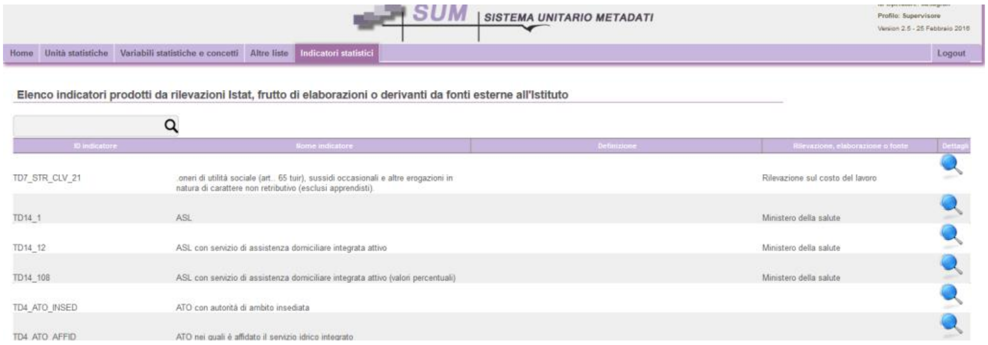
1. A data content obtained as a direct analysis of a micro data set. In this case, the c. and d. items are explained by:
   a. The transformation method (mean, media, total, percentage, Gini index, variance,…)
   b. The input data set, with details on which part of the micro data set has been used:
      i. Reference population of the macrodata;
      ii. Numerical variable on which it has been computed (up to now, Istat analyses for output macrodata just one numerical variable – no correlations etc. In case of only categorical variables, the numerical variable is the counting variable);

        iii. Conditional categorical variables (only for transformation method=conditional percentages, e.g. for tables containing the percentage of smokers given age class and region of residence, the conditional categorical variable is “Smoker” and the only category of the classification that is used in the tables is “Yes”);
        iv. Unit of measure (when necessary);
        v. Unit multiplier (when necessary).

2. A data content obtained from other macro data (usually for comparison purposes), where the c. and d. items are explained by:
    a. The transformation method (ratio, balance, percentage variation, index number,…);
    b. The input macro data, as the macro data corresponding to the numerator and denominator in a ratio. Each macro data component is again modelled either as an a. or b. macro data.
        i. The description of this kind of macro data is completed by the unit of measure, unit multiplier, base year concepts, when necessary.

The benefits of the use of the *Data Content* modelled as illustrated before are different.

- Macrodata description is highly standardized with what used in the different data production process phases. The components of the *Data Content* should be all declared clearly. For this reason, in a joint action with the Istat Corporate Data Warehouse (I.Stat: http://dati.istat.it), data producers are asked
    - to fill in a form that describes the data content, filling it in with the already available codes on the statistical program, reference population, numerical variable, data transformation method, macro data, …, or including new codes if necessary,
    - to compare it with what described in words in the *Data Content* item and add or modify whatever is necessary in order to make the *Data Content* complete (nothing should be given for granted).
- The *Data Content* facilitates the coherent modelling of *Data Structures*, so that two different people or organizations can model the same data cube in the same way.
- Macro data are described by linking their data content with concepts that are essentially already in use for micro data descriptions (e.g. reference population of the data, numerical variable,…) or that trace back to micro data (just unrolling the data content whose inputs are other macro data).
- Traceability is easily re-usable.
- Search procedures are enriched by all the connections that the data content ensures.
    - For instance, it is easy to ask for micro- and macro data that refer to the same reference population and look on their metadata (variables, classifications) in order to compare them and propose harmonization tasks.
- Data users find the *Data Content*, i.e. the meaning of data, in a unique place. This is especially useful in data cubes, containing huge amounts of data with different meanings. If a structured model as the *Data Content* is not used there is the risk that different data producers put the *Data Content* components in different places (e.g. one in the hypercube title, another in a query title, another in one or more dimensions). In this way it is not easy to count and understand how many different outputs the data cube is containing (not only for data users, also for our managers and colleagues). The *Data Content* is a complete list, each item of this list is structured along the lines described before.


SUM
SISTEMA UNITARIO METADATI
Profilo: Supervisore
Version 2.5 - 25 Febbraio 2016
Home
Unità statistiche
Variabili statistiche e concetti
Altre liste
Indicatori statistici
Logout

Elenco indicatori prodotti da rilevazioni Istat, frutto di elaborazioni o derivanti da fonti esterne all'Istituto

ID indicatore	Nome indicatore	Definizione	Rilevazione, elaborazione o fonte	Dettagli
TD7_STR_CLV_21	.oneri di utilità sociale (art.. 65 tuir), sussidi occasionali e altre erogazioni in natura di carattere non retributivo (esclusi apprendisti).		Rilevazione sul costo del lavoro	🔍
TD14_1	ASL		Ministero della salute	🔍
TD14_12	ASL con servizio di assistenza domiciliare integrata attivo		Ministero della salute	🔍
TD14_108	ASL con servizio di assistenza domiciliare integrata attivo (valori percentuali)		Ministero della salute	🔍
TD4_ATO_INSED	ATO con autorità di ambito insediata			🔍
TD4_ATO_AFFID	ATO nei quali è affidato il servizio idrico integrato			🔍

Fig. 1.7 Macrodata section in SUM-MS

4. Data structures

As anticipated in section 2, data sets are the output of a process step, that can become the input of the next process step. GSIM states that “all Data Sets must have a structure associated with them”. For better understand the meaning of a dataset and in order to promote the harmonization and the comparability, each data sets is described from a Data Structure by means of Data Structure Components (Identifier Components, Measure Components and Attribute Components). A data structure has defined by the set of these three components.

It is possible to distinguish two main different type of data structure:

- Dimensional Data Structure (DSD);
- Unit Data Structure.

Up to now the repository SUM-MS allows to investigate the data structure distinctly for Dimensional Data Structure type and for Unit Data Structure type (see figure 8).

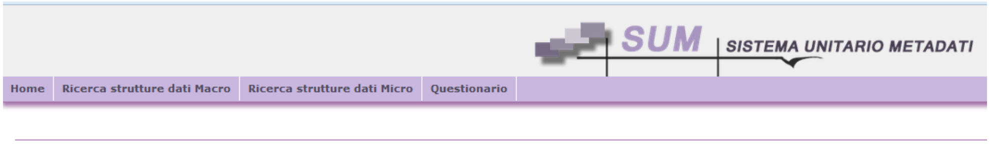

SUM
SISTEMA UNITARIO METADATI
Home
Ricerca strutture dati Macro
Ricerca strutture dati Micro
Questionario

Fig. 1.8 Macro and micro data structure section in SUM-MS

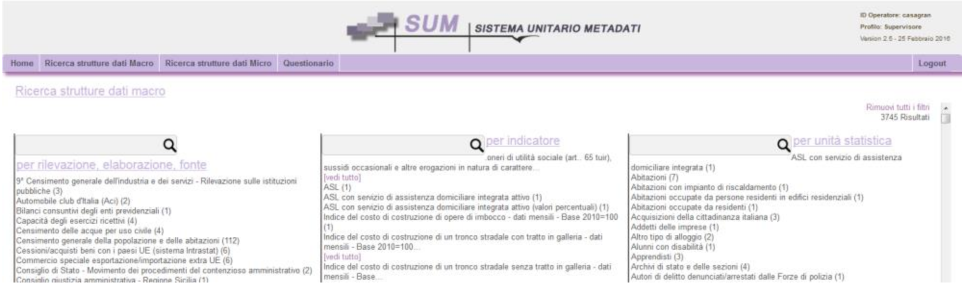
The first component of a Data Structure are the “Dimensions”. As usual, these concepts are those that identify the figures in a table. In a Unit Data Structure it is comparable with the identifying/key variable of each single record.

The second data structure component “Measure” describes the content of a statistics table. For macro data, SUM_MS uses the data content (see section 3.4) as a measure. In the case of Unit Data Structures this component corresponds to the value that the statistical variables assume for each unit.

The last component, “Attribute”, both for Dimensional and Unit Data Structure, refers to concepts which help to specify the meaning of a data set without being identifying variables. The main attributes are unit measure, scale factor and others attributes that specify the nature of the data (observation status, confidentiality of data, number of decimals, ...).

With regard to the data structures, it is necessary to underline the key role of the other “Dimensions”, compared to measure, of the Hypercube. The dimensions include all the concepts that appropriately define each number in the table: at least a categorical variable, a time dimension, any other element useful to distinguish different types of data (such as the different types of correction for the seasonal adjustment of time series).

The Dimensional Data Structure, as showed in the figure below (see Figure 9), allows to analyze each DSD actually included in the system starting from three different ways: surveys/source, indicator and population/unit type.


SUM
SISTEMA UNITARIO METADATI
ID Operatore: casagran
Profilo: Supervisore
Version 2.5 - 25 Febbraio 2016
Home
Ricerca strutture dati Macro
Ricerca strutture dati Micro
Questionario
Logout
Ricerca strutture dati macro
Rimuovi tutti i filtri
3745 Risultati
per rilevazione, elaborazione, fonte
per indicatore
per unità statistica

9° Censimento generale dell'industria e dei servizi - Rilevazione sulle istituzioni pubbliche (3)
Automobile club d'Italia (Aci) (2)
Bilanci consuntivi degli enti previdenziali (1)
Capacità degli esercizi ricettivi (4)
Censimento delle acque per uso civile (4)
Censimento generale della popolazione e delle abitazioni (112)
Cessioni/acquisti beni con i paesi UE (sistema Intrastat) (6)
Commercio speciale esportazione/importazione extra UE (6)
Consiglio di Stato - Movimento dei procedimenti del contenzioso amministrativo (2)
Consiglio giustizia amministrativa - Regione Sicilia (1)

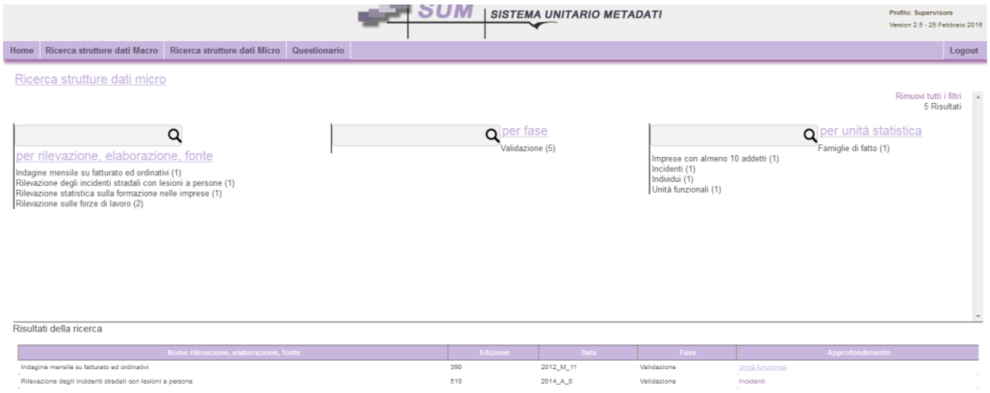
oneri di utilità sociale (art.. 65 tuir), sussidi occasionali e altre erogazioni in natura di carattere...
[vedi tutto]
ASL (1)
ASL con servizio di assistenza domiciliare integrata attivo (1)
ASL con servizio di assistenza domiciliare integrata attivo (valori percentuali) (1)
Indice del costo di costruzione di opere di imbocco - dati mensili - Base 2010=100 (1)
Indice del costo di costruzione di un tronco stradale con tratto in galleria - dati mensili - Base 2010=100...
[vedi tutto]
Indice del costo di costruzione di un tronco stradale senza tratto in galleria - dati mensili - Base...

ASL con servizio di assistenza domiciliare integrata (1)
Abitazioni (7)
Abitazioni con impianto di riscaldamento (1)
Abitazioni occupate da persone residenti in edifici residenziali (1)
Abitazioni occupate da residenti (1)
Acquisizioni della cittadinanza italiana (3)
Addetti delle imprese (1)
Altro tipo di alloggio (2)
Alunni con disabilità (1)
Apprendisti (3)
Archivi di stato e delle sezioni (4)
Autori di delitto denunciati/arrestati dalle Forze di polizia (1)

Fig. 1.9 Types of research available for macro data structure in SUM-MS

For each data structure some useful information is included concerning, in a first instance, the single indicators measured and the name of the dataset disseminated from Istat. It is possible to have further information about the others dimensions, the type of variables and the classifications or code lists associated with each variable.

In the case of a Unit Data Structure search tools allow to make searches through the surveys/source, process step and unit type (see Figure 10). In a first instance it is possible to know information about the name of the surveys included up to now, the edition, the process step and the unit type.


SUM
SISTEMA UNITARIO METADATI
Profilo: Supervisore
Version 2.5 - 25 Febbraio 2016
Home
Ricerca strutture dati Macro
Ricerca strutture dati Micro
Questionario
Logout
Ricerca strutture dati micro
Rimuovi tutti i filtri
5 Risultati
per rilevazione, elaborazione, fonte
per fase
per unità statistica

Indagine mensile su fatturato ed ordinativi (1)
Rilevazione degli incidenti stradali con lesioni a persone (1)
Rilevazione statistica sulla formazione nelle imprese (1)
Rilevazione sulle forze di lavoro (2)

Validazione (5)

Famiglie di fatto (1)
Imprese con almeno 10 addetti (1)
Incidenti (1)
Individui (1)
Unità funzionali (1)

Risultati della ricerca

Nome rilevazione, elaborazione, fonte	Edizione	Data	Fase	Approfondimento
Indagine mensile su fatturato ed ordinativi	390	2012_M_11	Validazione	Unità funzionali
Rilevazione degli incidenti stradali con lesioni a persone	510	2014_A_0	Validazione	Incidenti

Fig. 1.10 Types of research available for micro data structure in SUM-MS

References

Fisher, R.A. 1925. *Statistical Methods for Research Workers*. London: Oliver and Boyd.

Scanu M. 2015. “GSIM implementation in the Istat Metadata System: focus on structural metadata”. Workshop on International Collaboration for Standards-Based Modernization, Geneva 5-7 May.

Signore M., Scanu M., Brancato G. 2015. Statistical Metadata: A Unified Approach to Management and Dissemination. *Journal of Official Statistics*, vol. 31, 1-23.

United Nations Economic Commission for Europe (UNECE). 2013a. GSBPM v5.0

United Nations Economic Commission for Europe (UNECE). 2013b. Generic Statistical Information Model (GSIM): Specification (Version 1.1, December 2013)