Distr. GENERAL
25 April 2013
WP.4
ENGLISH ONLY

UNITED NATIONS
ECONOMIC COMMISSION FOR EUROPE
CONFERENCE OF EUROPEAN STATISTICIANS

EUROPEAN COMMISSION
STATISTICAL OFFICE OF THE
EUROPEAN UNION (EUROSTAT)

ORGANISATION FOR ECONOMIC COOPERATION
AND DEVELOPMENT (OECD)
STATISTICS DIRECTORATE

Work Session on Statistical Metadata
(Geneva, Switzerland, 6-8 May 2013)
Topic (i): Metadata standards and models

A STRATEGY ON STRUCTURAL METADATA MANAGEMENT BASED ON SDMX AND THE GSIM MODELS

Working Paper

Prepared by Stefania Bergamasco, Alessio Cardacino, Francesco Rizzo, Mauro Scanu, Laura Vignola (ISTAT)

I. Introduction

1. Istat is developing a corporate metadata system ("Unified Metadata System", in short SUM) whose aim is: i) to have a unified vision of the metadata management inside Istat; ii) to facilitate the horizontal integration among statistical domains in order to reduce the stovepipes approach; iii) to foster a vertical integration with the European Statistical System, through the adoption of common standards and the harmonization of the content. The part of this system dealing with structural metadata will be organized in such a way that: i) it will be possible to trace the data production process, from the design to the dissemination phase; ii) it will consist of a set of harmonized concepts, where harmonization is sought through the different steps of a data production process and between different data production processes; iii) it will help a data producer to reuse already defined metadata.

2. In order to do this, different standards are available and should be taken into consideration. Two of them are particularly useful for the system purposes: SDMX and GSIM. SDMX is a pillar of data and metadata transmission between different institutions and relies on sound IT infrastructures, but it lacks of the statistical semantics that allows using it in a statistical metadata system with the purposes of the SUM. GSIM is helpful because it allows to structure metadata according to their role in the statistical process, introducing concepts that are not available in SDMX as: units/populations, i.e. the entities over which the statistical output is computed; statistical variables (with the corresponding classifications or measure units, according to their nature), i.e. the phenomena associated to the units of interest for the statistical output; statistical operators, i.e. the arithmetic/logical/statistical methods used to transform data from one phase to the other (e.g. from validated micro data to macro data).

3. This paper aims at illustrating the Istat standard and methodology to create and manage integrated and reusable classifications, highlighting its impact on technological standards. Furthermore, space is given to the necessary enhancements of the SDMX standard (or better of the features of its reference infrastructure) in the context of structural metadata management.

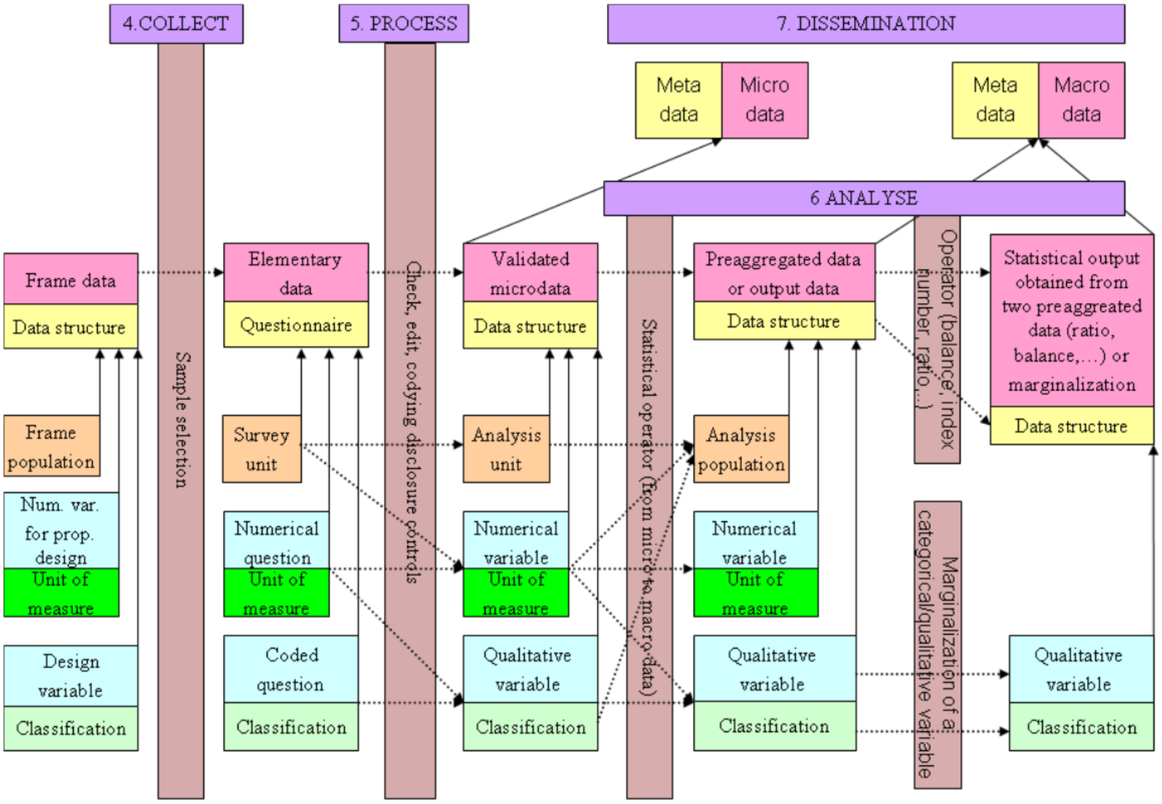
4. All these aspects will rely on the experiences gained since 2010 on the deployment of Istat new data warehouse (*I.Stat)* and on the dissemination of data contained in *I.Stat* through the Single Exit Point.

II. The Istat standard and methodology to create and manage integrated and reusable metadata

A. Structural metadata in the data production process

5. Structural metadata are commonly defined as those metadata that act as identifiers of the: a) structure of the data, e.g. names of columns of micro data tables or dimensions of statistical cubes; b) structure of associated metadata, e.g. units of measurement (Androvitsaneas et al., 2006, SDMX, 2009). Their nature as components of data definitions allow structural metadata to live on only some of the traditional GSBPM phases: mainly those named collect, process, analyse and disseminate, i.e. those characterized by the existence of data (respectively collected, processed, analysed and disseminated data). The following figure illustrates the main structural metadata to be considered in each of these phases, their relationship in each phase and their relationship along the data production process.

Figure 1: structural metadata and their relationships in a statistical process



6. From Figure 1 it comes out that the essential structural metadata are: the unit/population, the numerical/quantitative and categorical/qualitative variables (together with their unit of measures and classifications) and the statistical operator, i.e. the operation that transforms data in a previous phase into data of a subsequent phase. All together, these elements describe the “data structure” for that phase.

7. As far as the metadata relationships between different phases, these are essential for industrializing the data production process and for helping a user to understand the meaning of the data. Note that a population can give raise to another population (maybe a derived one, for instance selecting only those units that fulfil some rules, as in data conditioning) or a numerical variable (as in the case of two microdata sets on trips and households respectively, and then a dataset on household with the “number of trips per household” is considered); a categorical variable can only give raise to another

categorical variable (with a coarser classification) or can help in partitioning a population; a numerical variable can change drastically its role in a subsequent phase: it can remain a numerical variable, it can be categorised and be transformed in a categorical variable with a classification, and finally if a unit is associated with different countings related to the categories of one or more categorical variables, this numerical variables can give raise to a population (e.g. when the numerical variables “number of male” and “number of female” students are observed on the units “universities” in the process phase, and the table “number of students by gender” is then disseminated, where the students become the unit of interest, the numerical variable is the “counting” and there is the additional categorical variable “gender”).

B. How to model structural metadata for the data dissemination phase

8. Following both GSIM and SDMX, data structures will define a table content by using dimensions, measures and attributes. Focusing on dimensions, we will include the following characteristics: the macrodata output (most of the times termed “data type”), i.e. the actual content of the table; the categorical variables with their classifications used to cross-classify the macrodata output; the time and frequency dimensions; possible other operational dimensions as the adjustment for time series. The macrodata output is the core of the statistical content of the disseminated table specifying the meaning of what is measured in each table cell and can be one of two kinds:

(a) Simple macrodata output, i.e. aggregate data structures directly obtained from a microdata set through the application of a statistical technique (mean value, median, percentages,...);

(b) Composite macrodata output, i.e. aggregate data obtained by transforming two or more (simple or composite) macrodata outputs (ratios, index numbers, balances,...).

9. The two macrodata outputs are characterized by a different nature of the concept that describes what is measured in the table. In the case of a simple data structure, it is mandatory that the macrodata output specifies: the analysis unit, the numerical variable (if a counting, this is the numerical variable to use) and its associated unit of measure; the statistical operator used. An optional argument to include in the macrodata output is the “main” categorical variable, i.e. the categorical variable object of analysis. All the other categorical variables will form the so called conditioning variables (i.e. contribute to the definition of the reference subpopulations for each datum in the table) and should not be included in the data type. For complex data structures, the data type should include only the two macrodata that are the input of the statistical aggregate, and the operator (usually an arithmetic operator as a ratio or a subtraction) between them. For the arithmetic operator, additional pieces of information can be necessary, as the base year for index numbers. For both data types, additional arguments as the unit of measure and the unit multiplier can be given.

Table 1: description of a simple macrodata output

Macrodata output	Mandatory attributes			Non mandatory attributes		
	Analysis population	Numerical variable	Statistical operator	Unit of measure	Unit multiplier	Main categorical variable
Average monthly household income (in thousands of Euro)	households	monthly household income	average	Euro	thousands	---
Number of university graduates	graduates	counting	total	---	---	---
Persons aged 3 and over practising sports – Percentage	persons aged 3 and over	counting	percentage	---	---	practising sport: yes/no

Table 2: description of a composite macrodata output

Macrodata output	Mandatory attributes			Non mandatory attributes	
	Operator	Aggregate 1	Aggregate 2	Base period	Unit multiplier
Labour productivity - Value Added at basic prices, chain linked volume reference year 2005, per hour worked - Index 2005 =100	Index number (ratio)	Value Added at basic prices, chain linked volume reference year 2005 - Index 2005=100	Hours worked Index 2005=100	2005	---
Nuptiality rate (per 1000)	Rate (ratio)	Number of marriages	Mid-year resident population	---	Per 1000

10. A crucial aspect to consider in a metadata system is the management and governance rules related to each kind of metadata, from the variable names up to classifications. In the sequel we focus on classifications.

C. How to define integrated and reusable classifications

11. In order to create a unique Data Warehouse (DW for short), a National Statistics Institute has to face four major problems:

- integration among the data stored into the database;
- management, at the same time, of a great number of dimensions and data;
- the physical data base dimension and the performance and backup related issues
- dimension dynamism in the course of time

12. <u>From an integration point of view</u>: suppose to publish data in Table 3 and suppose that a user wants to analyse these data as well as to "work" with them.

Table 3: example of dissemination data

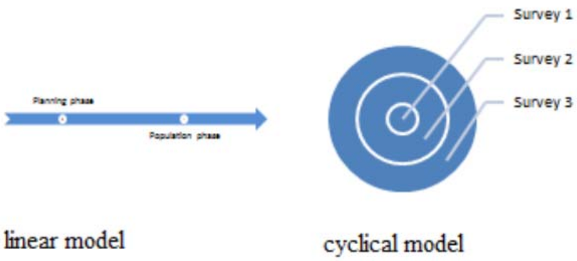
Data type	Males by marital status		Marriages by bridegroom marital status		Disabled people by marital status	
Variable name	Males marital status		Bridegroom marital status		Disabled marital status	
classifications	01	never married	01	never married	01	never married
	02	married	02	divorced	02	married
	03	divorced	03	widowed	03	divorced/widowed
	04	widowed				

13. There are some problems: different items with identical codes; same items with different codes. From an integration point of view there is the need to manage integrated classifications: each item has to have the same code in the whole database.

14. <u>From a dynamism point of view</u>: There are two different issues. On the one hand a survey naturally changes from time to time. Hence, it is possible that a survey publishes *"N° of people by gender, marital status and age class (5 age class)"* at year "x", while at year "x+k" the requested figure to be published corresponds to the *"N° of people by gender, marital status and age class (7 age class)"*.

15. Publication on excel files does not produce problems but in a data base, where there is the need to load data into the same table to allow users to surf among data along different time references, the consequence is that a classification can change in the different time references. On the other hand there are two different strategies to create a unique DW: *linear* and *cyclical models* respectively (Figure 2). Under the first model everything should be defined *ex ante*. In other words, all classifications and cubes should be collected and available before populating the DW. In a cyclical model, the DW is populated with just some surveys, with their classifications and data cubes, and the whole DW will be completed with other surveys step by step.

Figure 2: data warehouse development model

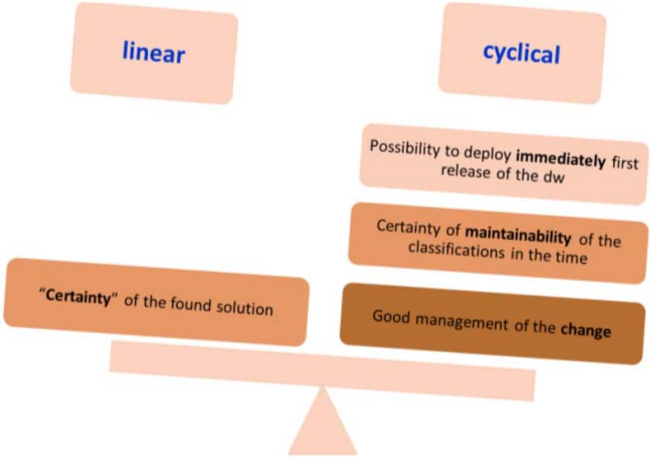


16. Each model has its own advantages but, in our opinion and according to our objectives, the cyclical model is better. Fig. 3 shows the main differences.

17. In the context of the cyclical model a classification can change from time to time. In a few words, classification management should follow two rules: integration and lack of stability. In order to fulfil these issues how should classifications be defined and used? The main ideas are in the following lines of this paragraph.

18. Let us consider again the example in Table 3 and, first of all, consider the integrated code list with all the necessary items (Table 4)[1].

Figure 3: data warehouse development model advantages


linear
cyclical
Possibility to deploy **immediately** first release of the dw
Certainty of **maintainability** of the classifications in the time
"**Certainty**" of the found solution
Good management of the **change**

[1] En example in the same direction in the code list *Age* of EUROSTAT

Table 4: example of integrated code list

Integrated code list	
Code list name: Marital status	
01	never married
02	married
03	divorced
04	widowed
05	divorced/widowed

19. The second thing to do is to distinguish among three different roles (Table 5): a <u>variable name</u> – which describes the statistical context of a dimension; a <u>code list</u> (classification) - which describes the possible categories that a variable can assume; a <u>group of items</u> – which describes the link between the first and the second for a particular publication (other examples are in Appendix 1).

Table 5: structural metadata and their relationships in a statistical process

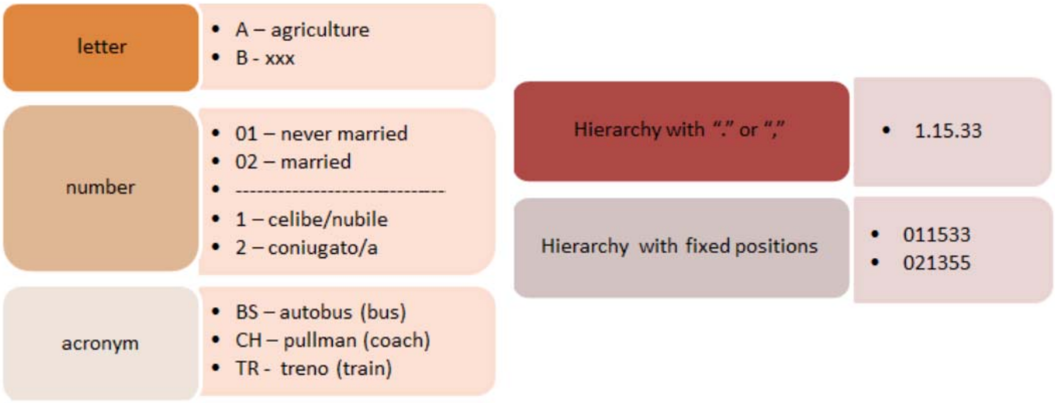
		Variable names			
Code list name: Marital status		Males marital status	Bridegroom marital status	Disabled marital status	
01	never married	x	x	x	Groups of items
02	married	x		x	
03	divorced	x	x		
04	widowed	x	x		
05	divorced/widowed			x	

20. The integration and lack of stability issues imply that a NSI has to manage also the classification items codification. Fig 4 shows the common statistical standards to code an item.

21. Suppose that in the first cycle of the DWH implementation the code list *size class* in Table 6.a (a) becomes available, while in the second cycle there is the need to introduce also the “0-1” and “2-9” items (Table 6.a (b)).

22. Which codes do we have to use? It is actually impossible changing the codification of the old items because of their link to already available the data. So we can just add new items. If we use *05* and *06* the classification will become unreadable: for example, the visualization order is jeopardized. The same holds also true if *letters are used instead of numbers* (Table 6.b).

Figure 4: common statistical standards of items



letter

- A – agriculture
- B - xxx

number

- 01 – never married
- 02 – married
- ------------------------------
- 1 – celibe/nubile
- 2 – coniugato/a

acronym

- BS – autobus (bus)
- CH – pullman (coach)
- TR - treno (train)

Hierarchy with "." or ","

- 1.15.33

Hierarchy with fixed positions

- 011533
- 021355

Table 6.a: the evolution of a code list during the DW development process

Step (**a**) of DW development process	
Code list: size class	
01	0-9
02	10-50
03	51-100
04	101 and over

Step (**b**) of DW development process	
	Code list: size class
01	0-9
	0-1
	2-9
02	10-50
03	51-100
04	101 and over

Table 6.b: the evolution of a code list during the dw development process

Code list: size class	
01	0-9
05	*0-1*
06	2-9
02	10-50
03	51-100
04	101 and over

	Code list: size class
a	0-9
e	*0-1*
f	2-9
b	10-50
c	51-100
d	101 and over

23. The hierarchical standard could be of help (Table 6.c). But can this hierarchical standard solve the problem if we need to add also the following items: "10-30", "31-60", "61-80", "81-100"?

Table 6.c: the evolution of a code list during the DW development process

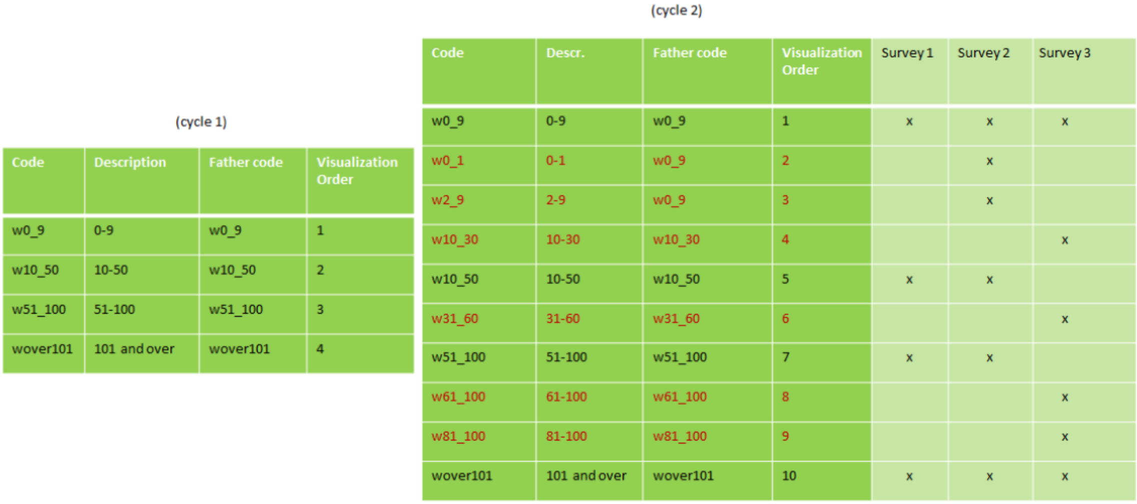
	Code list: size class
a	0-9
a.1	*0-1*
a.2	2-9
b	10-50
c	51-100
?	10-30
?	31-60
?	61-80
?	81-100

24. Hence, in order to manage the evolution of a classification our suggestion is:

- to distinguish the concepts of *code*, *visualization order*, *father code*
- to use an acronym.

25. Figure 5 shows the results of the previous procedure.

Figure 5: management of code list evolution



(cycle 1)

Code	Description	Father code	Visualization Order
w0_9	0-9	w0_9	1
w10_50	10-50	w10_50	2
w51_100	51-100	w51_100	3
wover101	101 and over	wover101	4

(cycle 2)

Code	Descr.	Father code	Visualization Order	Survey 1	Survey 2	Survey 3
w0_9	0-9	w0_9	1	x	x	x
w0_1	0-1	w0_9	2		x	
w2_9	2-9	w0_9	3		x	
w10_30	10-30	w10_30	4			x
w10_50	10-50	w10_50	5	x	x	
w31_60	31-60	w31_60	6			x
w51_100	51-100	w51_100	7	x	x	
w61_100	61-100	w61_100	8			x
w81_100	81-100	w81_100	9			x
wover101	101 and over	wover101	10	x	x	x

26. In this way, integration is obtained by:

- using a particular codification;
- showing the visualization order and father code in different fields (without using the item code).

27. As far as the hierarchical relationship between items in a classification, we stress that it is also important to consider that an item can have more than just one father.

III. The impact of this standard on technological standards

A. General technological standards

28. Each characteristic described before defines a specific standard for technological point of view. **In a few words a statistical software has to manage t**he c*lassification objects*:

- a classification has a name;

each item of a classification has the order of visualization;

- each item of a classification can have one or more fathers;

the codification of an item has to be a string (acronym);
as well as the *relationship* between a classification and a variable name:

- the relationship has specific name (variable name);
- once defined a relationship it is mandatory to select the subset of items of the corresponding code list.

29. We can also underline that if we consider the use of a classification at the same time for dissemination and survey phases probably we need to manage the visualization order in a different way. Hence, it would be better if the visualization order plays the role of an attribute of the code list/variable relationship instead of the classification.

B. Enhancements of the SDMX standard

30. SDMX is a statistical and technical standard worldwide used within the statistical community. Its initial aim was to facilitate data and metadata exchange between organizations but, nowadays, more and more organizations are exploiting SDMX in order to harmonize the dissemination process.

31. Within the European Statistical System, the Commission Communication 404/2009 "on the production method of EU statistics" and the draft EP/Council Regulation "on processes, standards and metadata" have leaded Eurostat in financing (under the form of grants) EU member states' projects in order to facilitate the horizontal and vertical integration. In this context SDMX is the reference standard for aggregated data. In general SDMX streamlines the process to define appropriated metadata for aggregated datasets through the definition of Data Structure Definitions (DSD) and Metadata Structure Definitions (MSD) and related artifacts, such as Concept Schemes and Code Lists. This process is simplified by using free tools such as the Data Structure Wizard that allows an easy data and metadata modeling.

32. The increasing use of SDMX by statistical organizations pushes a metadata system to be as much compatible as possible with SDMX not only to have a facilitation in creating the SDMX DSD and MSD for data and metadata exchange but also because SDMX defines how to describe the content of statistical data through its artefacts.

33. This standard allows also to maintain a synonymy between the codes of different code lists having the same statistical meaning. This is an important aspect of a metadata system because very often a code list defined internally in an organization can have a synonymy with other code lists managed by other organizations. In this case SDMX provides the way to define it through the "StructureSet", an artefact that allows the "mapping" between a set of metadata and a set of "target" metadata (code lists, codes, data structure definitions and so on).

34. However, pieces of information that are part of the SDMX-ML (artefacts) are not sufficient to implement a complete metadata system as described in the SUM.

35. Here, we focus on the following three aspects that the standard does not cover or covers only partially:

36. *Historicity of the codes*: In SDMX a Code List is defined by three identifiers: the id, the agency and the version. When a code in the Code List changes it is necessary to define a new Code List. For a metadata system this could represent a very heavy work, as in the geographical code list that is subject to frequent changes. SDMX allows the use of the Hierarchical Code List that however cannot be used in a DSD. In this case classifications with a large variability can be stored in the system as Hierarchical Code Lists and referred as Code Lists when the DSD must be produced.

37. *Definition of some peculiarities of concepts or "data types"*: Data can be described in SDMX using the following type of concepts: dimensions, attributes and measures (primary measure and measure dimension). This difference between concepts is linked to the role that they play in the data description but not the specification if a concept is a variable of the survey or if it is derived from an operation on data (for example the difference between "Activity" and the "Adjustment" in Short terms statistics). In order to take their statistical role into account, the "description" tag will be used. As far as "data types" are concerned, it should be mandatory for a data provider to specify all its attributes, in terms of population/variables/statistical operator for simple macrodata outputs and aggregate 1/aggregate 2/operator for composite macrodata outputs. Lack of some of these attributes harms the correct data interpretation. A massive use of annotations will be taken into consideration in order to allow this kind of modelization.

38. *Link between concepts and operations*: In order to know exactly the operation as well as the concepts used to calculate a new concept, it is necessary to keep trace of the transformations of data occurred during the statistical process. This kind of information is foreseen in the standard and it actually has the place in the "Transformations and Expressions" SDMX module. Anyway this part is not yet implemented, although it is currently under discussion in the SDMX Technical Working Group .

IV. References

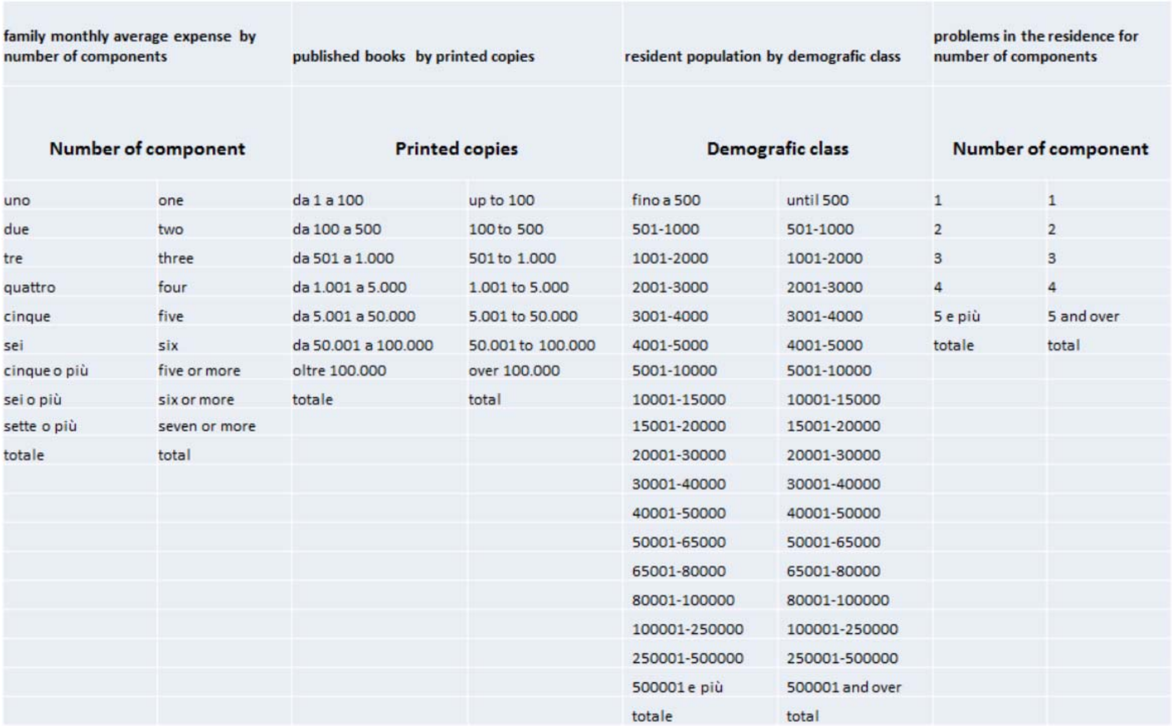
Androvitsaneas, C., Sundgren, B., Thygesen, L. (2006). Towards an SDMX User Guide: Exchange of statistical data and metadata between different systems, national and international, OECD Expert Group on Statistical Data and Metadata Exchange, Geneva, 6-7 April 2006.

SDMX (2009). Content oriented guidelines. Available at URL= http://www.sdmx.org/ (Accessed January 2013).

V. APPENDIX 1: Example for integrated classification

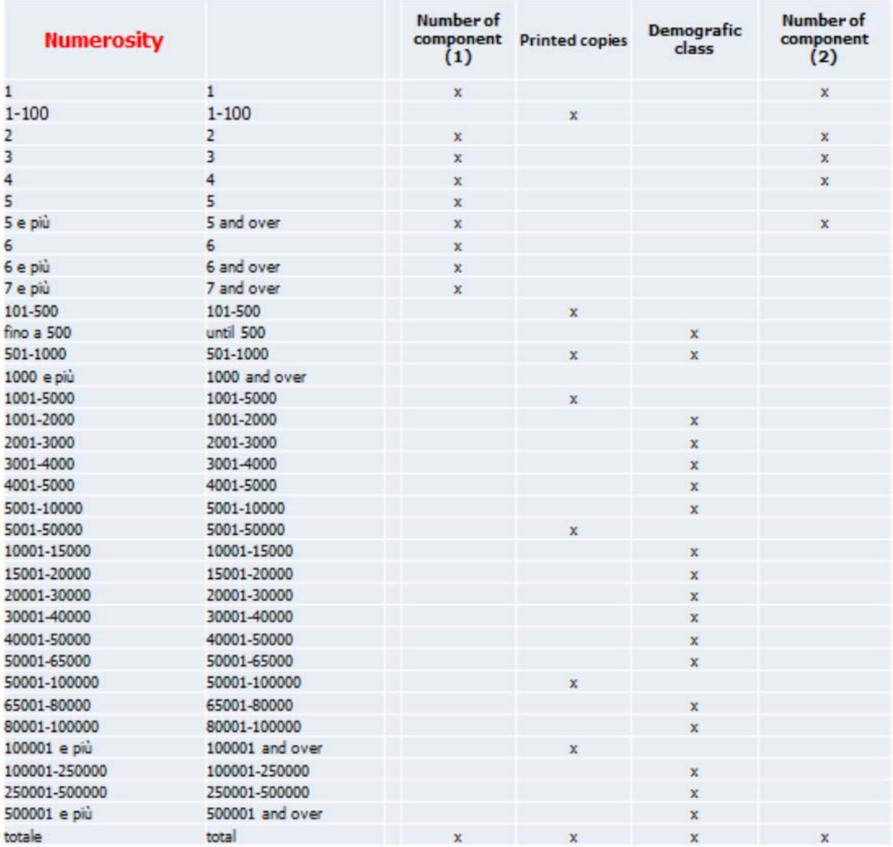
Example 1

Starting from:



family monthly average expense by number of components		published books by printed copies		resident population by demografic class		problems in the residence for number of components	
Number of component		Printed copies		Demografic class		Number of component	
uno	one	da 1 a 100	up to 100	fino a 500	until 500	1	1
due	two	da 100 a 500	100 to 500	501-1000	501-1000	2	2
tre	three	da 501 a 1.000	501 to 1.000	1001-2000	1001-2000	3	3
quattro	four	da 1.001 a 5.000	1.001 to 5.000	2001-3000	2001-3000	4	4
cinque	five	da 5.001 a 50.000	5.001 to 50.000	3001-4000	3001-4000	5 e più	5 and over
sei	six	da 50.001 a 100.000	50.001 to 100.000	4001-5000	4001-5000	totale	total
cinque o più	five or more	oltre 100.000	over 100.000	5001-10000	5001-10000		
sei o più	six or more	totale	total	10001-15000	10001-15000		
sette o più	seven or more			15001-20000	15001-20000		
totale	total			20001-30000	20001-30000		
				30001-40000	30001-40000		
				40001-50000	40001-50000		
				50001-65000	50001-65000		
				65001-80000	65001-80000		
				80001-100000	80001-100000		
				100001-250000	100001-250000		
				250001-500000	250001-500000		
				500001 e più	500001 and over		
				totale	total		

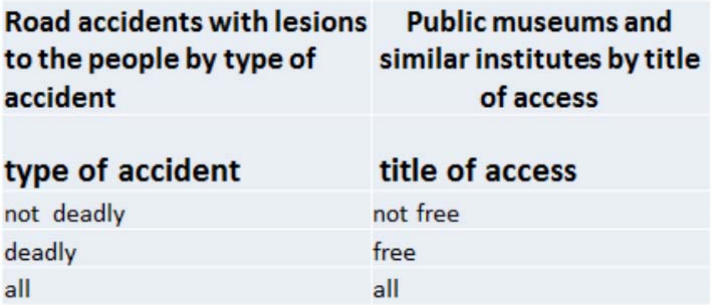
It is possible to obtain:



Numerosity			Number of component (1)	Printed copies	Demografic class	Number of component (2)
1	1		x			x
1-100	1-100			x		
2	2		x			x
3	3		x			x
4	4		x			x
5	5		x			
5 e più	5 and over		x			x
6	6		x			
6 e più	6 and over		x			
7 e più	7 and over		x			
101-500	101-500			x		
fino a 500	until 500				x	
501-1000	501-1000			x	x	
1000 e più	1000 and over					
1001-5000	1001-5000			x		
1001-2000	1001-2000				x	
2001-3000	2001-3000				x	
3001-4000	3001-4000				x	
4001-5000	4001-5000				x	
5001-10000	5001-10000				x	
5001-50000	5001-50000			x		
10001-15000	10001-15000				x	
15001-20000	15001-20000				x	
20001-30000	20001-30000				x	
30001-40000	30001-40000				x	
40001-50000	40001-50000				x	
50001-65000	50001-65000				x	
50001-100000	50001-100000			x		
65001-80000	65001-80000				x	
80001-100000	80001-100000				x	
100001 e più	100001 and over			x		
100001-250000	100001-250000				x	
250001-500000	250001-500000				x	
500001 e più	500001 and over				x	
totale	total		x	x	x	x

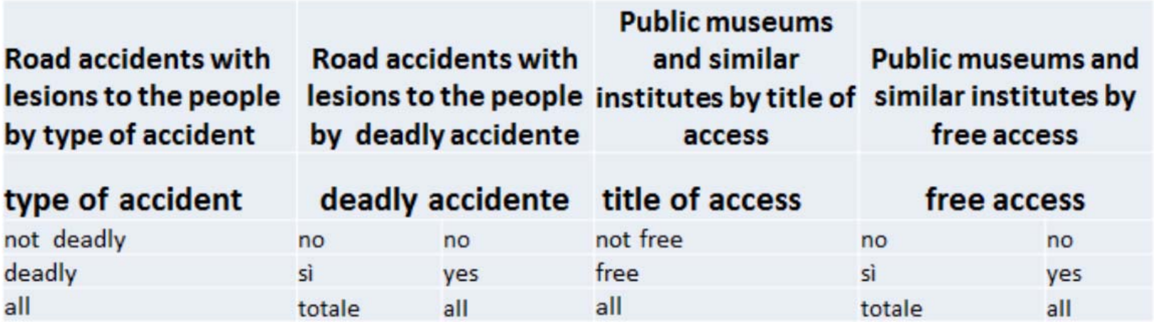
Example 2

Starting from:

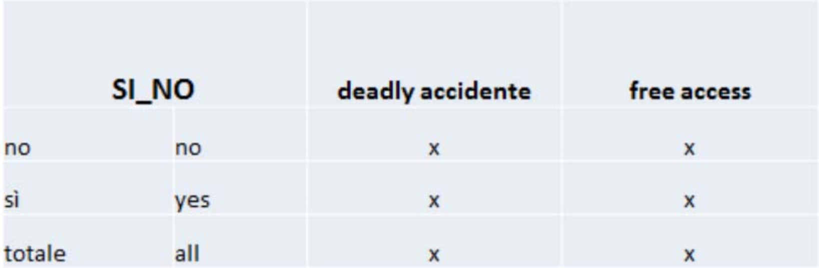


Road accidents with lesions to the people by type of accident	Public museums and similar institutes by title of access
type of accident	title of access
not deadly	not free
deadly	free
all	all

It is possible to obtain:



Road accidents with lesions to the people by type of accident	Road accidents with lesions to the people by deadly accidente		Public museums and similar institutes by title of access	Public museums and similar institutes by free access	
type of accident	deadly accidente		title of access	free access	
not deadly	no	no	not free	no	no
deadly	sì	yes	free	sì	yes
all	totale	all	all	totale	all



SI_NO		deadly accidente	free access
no	no	x	x
sì	yes	x	x
totale	all	x	x