Proceedings of Q2008
European Conference on Quality in Official Statistics

Variables System – the bridge between metadata and dissemination

Teodora Isfan, Methodology and Information Systems Department, Statistics Portugal[1]

"Metadata is the key to ensure that information will survive and continue to be accessible into the future."

1. Introduction

"Two very similar paintings of circus performers by Picasso from 1904 are put on the auction block; one brings tens of millions of dollars, the other hundreds of thousands. What is the difference? In one case, the ownership of the painting can be traced through sales slips and auction house records back to the estate of Picasso's dealer. The other painting appeared suddenly on the art market. It looks almost identical, but lacking documentation, how can one be sure it's authentic?" (Wayne, 2006)

Just as a work of art can change many hands many times, so can data and the associated metadata. Once created, data and the associated metadata can travel almost instantaneously through Internet, can be used, analysed, transformed or retransmitted.

So, the need for metadata standards, particularly definitions that are consistent among the myriad of subject matters of a statistical agency is more urgent today than it has ever been. More recently, the availability of diverse types of data and information on the Internet means that it is possible for users to view and compare data from one subject matter area with that of another.

Organizations have attributed increasing importance to managing knowledge, as demonstrated by the growing implementation of metadata systems that systematise, standardise and formalise this knowledge so that it can be published within or outside the organization.

2. Variables System

The variables system (SVAR) was based on ISO 11179 international standard (regarding data elements, their constituents and relationships). Moreover, we introduced modifications on the ISO 11179 to maximize its practical utility, in representing and structuring the data holdings of our Agency.

Integrated Meta Database (IMDB) from Statistics Canada (Johanis, Brooks, Dunstan and Lévesque, 2003) also inspired our implementation of all the metadata system and in particular the work about variables, namely naming convention.

Variables are the fundamental units of data an organization collects, process, and disseminates. Metadata registries organize information about variables, provide access to the information, facilitate standardization, identify duplicates, and facilitate data searching.

[1] Teodora Monica Isfan, Systems and Metadata Unit, Methodology and Information Systems Department, Statistics Portugal, Av. António José de Almeida, 1000-xxx Lisbon, Portugal, monica.isfan@ine.pt.

The core entities that characterize define and designate a variable (Morgado and Isfan, 2006) are: property, object class (statistical unit or population), representation class and value domain.

SVAR (Isfan, 2007), like each subsystem in the integrated metadata system, has the following architecture: a database, two Web applications (one for consultation and the other for management) and a view that provides metadata to be reused by other systems.

Management was designed to be decentralised with central coordination. The management application therefore implements two profiles: the *variables manager* (VM) and the *survey manager* (SM). There is a generic profile for consultation.

3. Dissemination System

Electronic dissemination of statistical results is getting more and more important. Users want immediate access to the newest results, not only at national but also at regional level, in a language they can understand. Permanent availability of statistics, continual user support and free access to statistics are other user requirements (Knüppel and Kunzler, 2001), that must be considered.

The dissemination data base was implemented to support the Official Statistics Portal and represents the final repository of aggregated statistical information. The dissemination data base has aggregated statistical data (variables/ statistical indicators) as inputs, provided, directly or indirectly through the DataWarehouse, by the Production Departments. The outputs are, also, aggregated statistical data (variables/ statistical indicators) and all the associated metadata necessary to a correct interpretation of data (DMSI/ II, 2005).

4. Statistical Indicators

In accordance with “Terminology on Statistical Metadata”, a statistical indicator is a data element (variable) that represents statistical data for a specific time, place and other characteristics (UN/ ECE, 2000).

In practical terms, translating this definition for its applicability in SVAR and Dissemination Data Base, a statistical indicator is defined on the basis of variables (Isfan, 2007) and results from the combination between aggregate variable and dimension variables (figure 1).

For the correct definition of the statistical indicator are indispensable two dimensions: the time dimension and the geographic dimension.

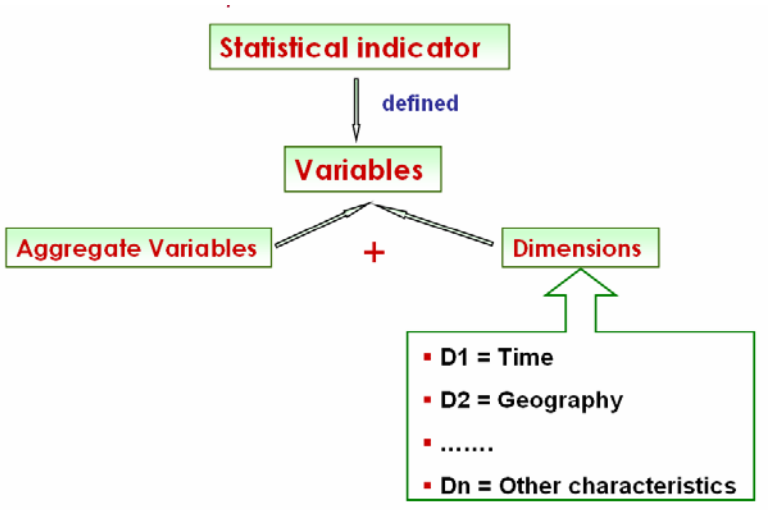

Figure 1. Statistical indicator – practical definition

Where:
The aggregate variable and the different dimensions are considered variables and are registered previously in SVAR, in an independent way (Isfan, 2007).

An aggregate variable (figure 2) is a variable defined by:

- Property,
- Object class = population (always),
- Representation class = Quantity, ratio, value, etc.
- Value domain = Non-enumerated + unit of measure (always),
- Definition or/ and concept,
- Formula,
- Use,
- Obligation.

Code:	259
Name:	**Resident population (No.)**
Short name:	Resident population (No.)
Validity:	**Start:** 05-09-2005 **End:** -
Status:	Valid
Representation Class:	Quantity
Value Domain:	
- Type	Non-enumerated
- Range	(0, ∞)
- Greatness	-
- Unity	Number (No.)
Definition:	➪RESIDENT POPULATION The persons who regardless of the fact that at the moment of observation ' 0:00 a.m. of the reference day ' are present or absent in a given housing unit, this unit being where they live during most of the year with their family, or where they have all or most of their belongings.
Formula:	Estimated value (Dem.)
Acronym:	-
Context:	-
Registration Authority:	INE - Instituto Nacional de Estatística
Submitting organization:	DME - Departamento de Metodologia Estatística
Use:	**Surveys** **Type** - (113) Annual estimates of resident population (version 1) Aggregate **Other sources** **Type** - INSA Aggregate
Obligation:	Obligatory

Figure 2. Variable's detail page – aggregate variable

A dimension (figure 3) is a variable defined by:

- Property,
- Object class = statistical unit (always),
- Representation class = code (always, except the time dimensions),
- Value domain = classification + level of classification or code list,
- Definition or/ and concept,
- Use,
- Obligation.

Code:	2710
Name:	**Place of residence (NUTS III - 2002) of person**
Short name:	Place of residence (NUTS - 2002)
Validity:	**Start:** 23-03-2007 **End:** -
Status:	Valid
Representation Class:	Code
Value Domain:	
- Type	Enumerated
- Classification (Version)	NUTS 2002 complete (PT, NUTS I, II, III, CC, FR) (v00320)
- Level	NUTS III
Definition:	-
Formula:	-
Acronym:	-
Context:	-
Registration Authority:	INE - Instituto Nacional de Estatística
Submitting organization:	DME - Departamento de Metodologia Estatística
Use:	**Surveys** / **Type** - (117) Demographic indicators (version 1) / Dimension - (5) Population and housing census (version 1) / Dimension
Obligation:	Obligatory

Figure 3. Variable's detail page – dimension

5. Naming convention

ISO 11179 has no specific naming convention rules, but we believe it is necessary to establish a convention in order to keep the data base consistent and coherent.

Naming statistical indicators, variables and their component entities is an integral part of the identification process.

The names are the primary means by which users of the data interact with variables and statistical indicators. So, we needed to generate and prepare "user friendly" names that must be brief, clear, and free of physical context. On these conditions, we created:

- Formal name,
- External name,

without lost of information.

The name convention we have developed (Morgado and Isfan, 2006) contains the rules that follow.

The formal name of statistical indicator is formed by the name of aggregate variable and the names of dimensions.

For the formal name the separators consists on a special character (underscore).

The registry of formal name is automatic and is important for management issues.

The rules of construction of the external name are identical for English and Portuguese.

The syntactic principles that we applied, specify the arrangement of components within

a name. This arrangement is considered as absolute, so specifies a fixed occurrence of the component, e.g., a rule requires that the aggregate variable is always the first component; the geography is the second component, the time dimension does not enter directly in the construction of the name, etc.
Semantics concerns the meaning of name components and the separators which delimit them.
The linking particle between the aggregate variable and the first dimension is: “by” The dimensions are separated by comma, except the “and” particle that separates the one before the last of the last dimension

Example:
Formal name:
Resident population (No.)_Reference period_Place of residence_Sex
External name:
Resident population (No.) by Place of residence and Sex

6. Register flow

As a start point we pick and choose the source that provide the selected statistical information and analyse the data and the associate metadata. Practically, begins the transformation of the "old" structure of statistical table into the "new" structure of statistical indicator.
At this point we must identify the definitional attributes of variables that will characterize the statistical indicator and design the combination aggregate variable and dimensions.
After the registry and approval of variables we proceed with the registry of statistical indicator(s). Following its definition, we start with the association of the aggregate variable to the source (figure 4).

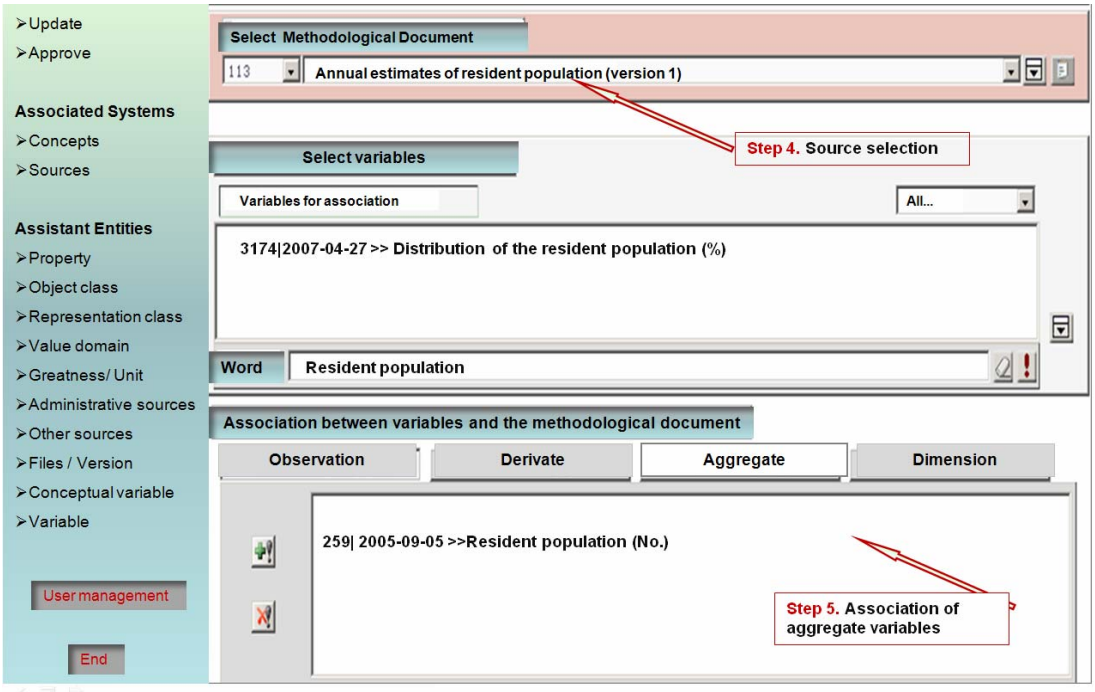

Figure 4. Association of aggregate variable to source

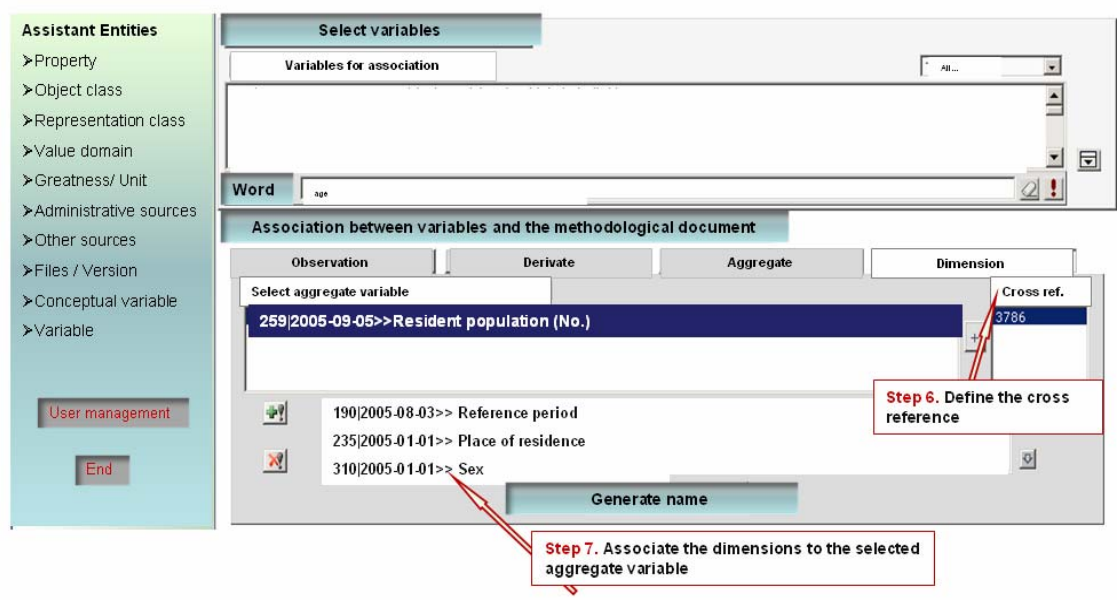

Figure 5. Cross reference definition

For the selected aggregate variable we define the cross reference. As a next step, we select and associate the dimensions. The definition of the cross reference is crucial, because in SVAR, the convention adopted, is that on the basis of the same aggregate variable we can define more than one statistical indicator (figure 5).

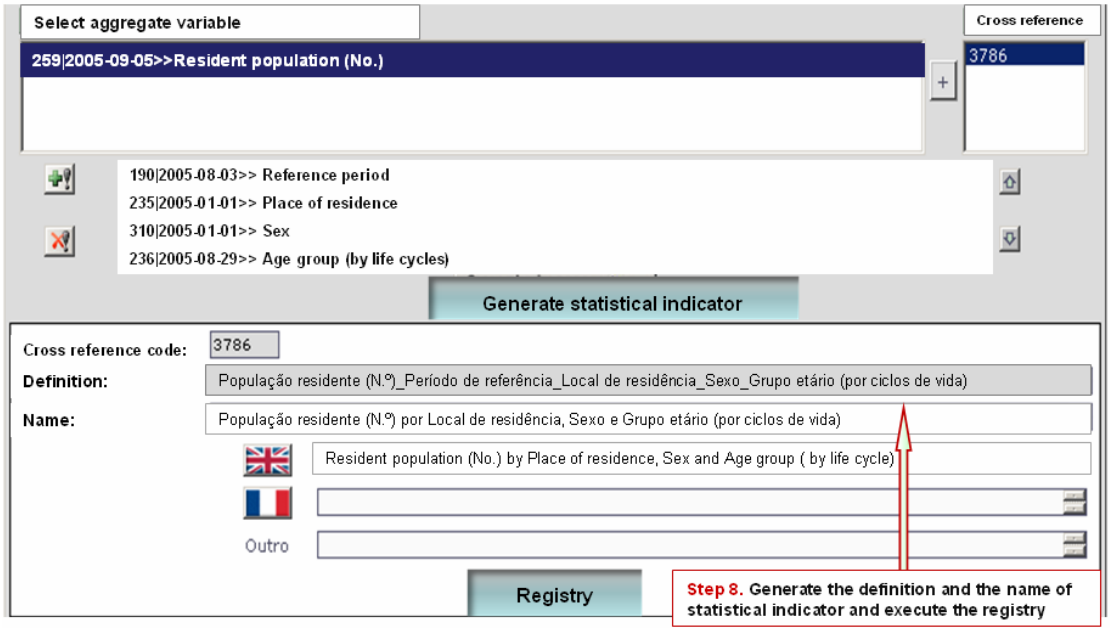

Figure 6. Registry of statistical indicator

Finally, we generate the definition and the name (formal and external) of the statistical indicator(s) and execute the registry (figure 6). After the registry, we perform the approval of the proposed definition and structure.

7. Transmission and visualization

One of the major rules established is, that all the statistical indicators available in Dissemination Data Base should have the associated metadata registered and

approved previously in SVAR. The definitions of approved statistical indicators are transmitted through a view (figure 7) and the unique identification is assured by the cross reference code.
The metadata attributes provided for each indicator are its name, frequency, source, unit of measure, associated concepts, definition, formula and other contextual information.
The data are transmitted directly from DataWarehouse, or indirectly (using XML) from other production data bases, as we already specified.
The transmitted data for each statistical indicator must respect the structure and definition already registered in SVAR.
After data and metadata approval, each statistical indicator receives a dissemination code and is published on the Official Statistics Portal.

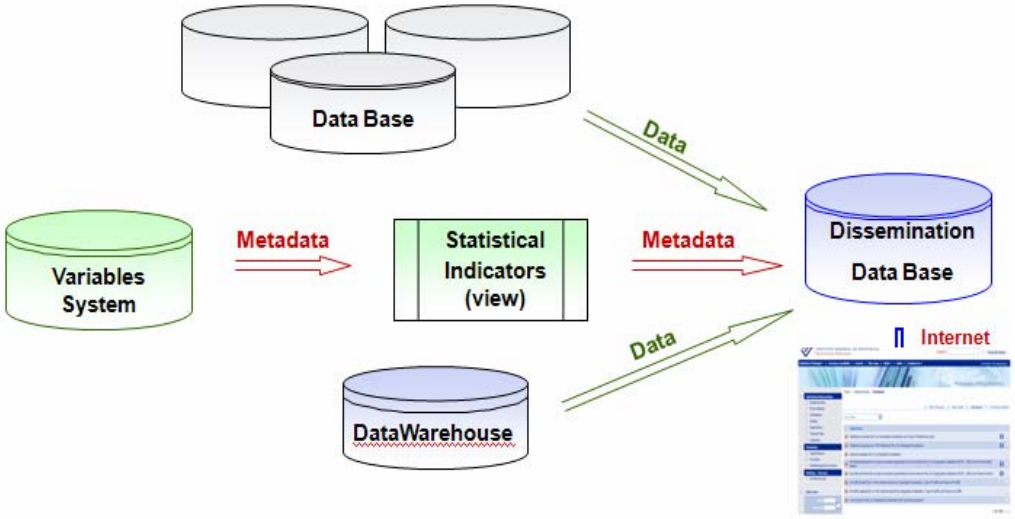

Figure 7. Statistical indicators - transmission and visualization

8. Benefits

The implementation of a central reference of aggregated data and metadata, the use of metadata standards, and the use of normalized data and metadata transmission, provided a common "look and feel" of all disseminated statistical information, improving quality and understandability and increasing the data and metadata sharing.

9. References

Isfan, M. (2007), “Sistema de variáveis – modelo conceptual”, unpublished report, INE, Portugal.

ISO/IEC 11179 (1999), “Information Technology – Specification and Standardization of Data Element”.

Johanis, P., Brooks, B., Dunstan, T., and Lévesque, J.P. (2003), “Statistics Canada’s Implementation of the Data Element Model”, Paper presented at Open Forum on Metadata Registries, Santa Fe, New Mexico, USA.

Knüppel, W. and Kunzler U. (2001), “Influence of the Internet on data collection and dissemination in the European Statistical System”, paper presented at IAOS Satellite Meeting on Statistics for the Information Society, Tokyo, Japan.

Morgado, I. and Isfan, M. (2006), “Documenting Variables”, paper presented at European Conference on Quality in Survey Statistics, Cardiff, UK.

Serviço de Infra-estrutura Informacional (DMSI/ II), (2005), “Gestão da Informação estatística a disponibilizar no Portal”, unpublished report, INE, Portugal.

United Nations Statistical Commission and Economic Commission for Europe (UN/ ECE), (2000), “Terminology on Statistical Metadata”, Conference of European Statisticians – Statistical Standards and Studies – Nº 53, Geneva, Switzerland.

Wayne, L. (2006), “The value of Metadata”, unpublished paper, Federal Geographic Data Committee, Reston, Virginia, USA.