Joint UNECE/ EUROSTAT/ OECD Work Session on Statistical Metadata (METIS)
11-13 March, 2009

Variables Subsystem

Prepared by Teodora Monica Isfan[1]

I. Introduction

Science, technology and the information society are changing the way we live, learn and work. In this environment of rapid development of information and fast communication technologies, developing efficient strategies for the production and dissemination of statistics is a challenge. Due to these changes, integrated and transparent description of information inside and outside statistical organizations has become inevitable.

Under these conditions, the *variables subsystem* has been developed in an integrated way with methodological documentation of surveys, documentation of administrative sources and with the dissemination database. The main objective was to build a system that facilitates the integration and co-ordination of the whole statistical metadata.

II. Variables subsystem

The variables subsystem provides a database of variables standardised and harmonised with their respective concepts, classifications, explanatory notes and calculation formulae.

The main purposes of the variables subsystem are:

- To support the questionnaire and survey design;
- To improve statistical coordination;
- To support the dissemination of statistical data;
- To assist the definition of normalized and/or harmonized variables;
- To promote comparability of data by using normalized variables.

2.1. Subsystem architecture

Variables are the fundamental units of data an organization collects, process, and disseminates [6]. Metadata registries organize information about

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

variables [1], provide access to the information, facilitate standardization, identify duplicates, and facilitate data searching (figure 1).

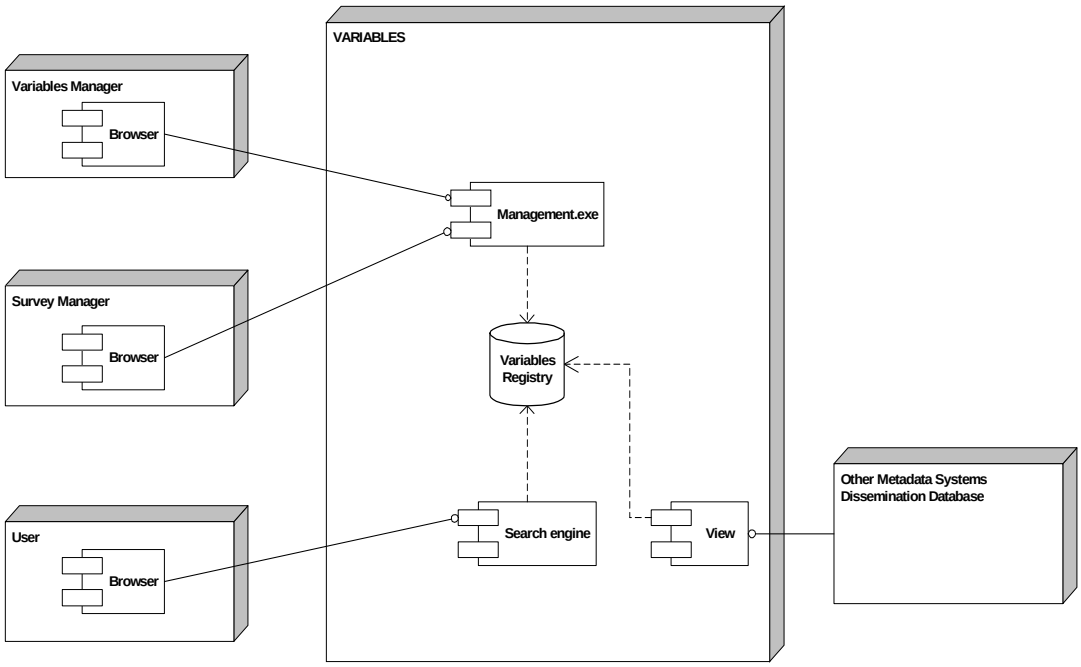

Figure 1. Components diagram

A variables registry is the core of the variables subsystem [10]. Each variable is the result of a process of development, involving several levels of abstraction (from the most general or "conceptual" to the most specific or "physical").

Different users access the subsystem via browser, and the interface with other systems is done through several views.

Another important feature of metadata registry is that variables are described by a concept and a representation of value domain (set of permissible values).

The advantages of this are the follows:

- Sets of similar variables are linked to a shared concepts, reducing search time;
- Every representation associated with a concept (i.e. each variables) can be shown together, increasing flexibility;
- All variables that are represented by a single (reusable) value domain can be located, assisting administration of a registry;
- Similar variables are located through similar concepts, again assisting searches and administration of a registry.

2.2. Conceptual model

The conceptual model (figure 2) is based on international standard ISO/ IEC 11179, "Information Technology – Specification and Standardization of Data Elements" and on Integrated Meta Database (IMDB) from Statistics Canada [7] (in particular the naming convention).

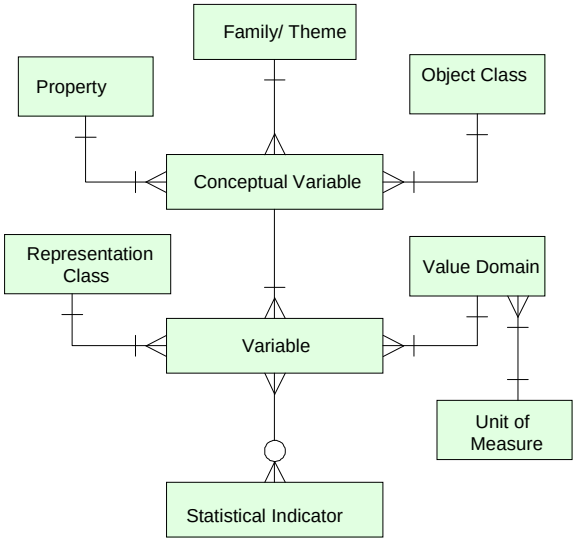

Figure 2. Conceptual model of the variables subsystem

Main entities:

Variables family – a classification for variables in general to facilitate the search for variables in the subsystem.

Property – characteristic or attribute common to all members of an object class; a property is a concept.

Objects class - a set of ideas, abstractions, or things in the real world that can be identified with explicit boundaries and meaning whose properties and behaviour follow the same rule.

Object classes in this subsystem are:

- Statistical units;
- Populations.

Conceptual variable – a property of an object class described independently from any particular representation.
Representation class – a component of the definition of the variable indicating the type of data it represents (code, ratio, quantity, etc).

Value domain - a set of permissible values and their associated meanings. The value domains may be:

- Categorical (or discrete);
- Continuous;
- Text.

Variable – the smallest identifiable unit of data in this subsystem for which a value domain, a unit of measure, versions, permissible values can be specified.

2.3. Naming convention

The variable subsystem was constructed to support statistical production and dissemination. These two facets have different requirements and sometime it is not easy to make them compatible. ISO 11179 has no specific naming convention rules, but we believe it is necessary to establish a convention in order to keep the database consistent and coherent.
Naming variables and their component entities (or statistical indicators) 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 ([4], [10]), we created:

- Formal name,
  Average income_Person_Value_[(0, ∞)/ Euro]
  Geographic localization_Establishment_Code_[Table of geography/ level = Municipality]

- External name,
  Average income (€) of person
  Geographic localization (Municipality) of establishment

- Abbreviate name,
  Average income (€)
  Geographic localization

without lost of information.

2.4. Variables subsystem management

Management was designed to be decentralised with central coordination. The *variables subsystem manager* ensures that the rules governing this subsystem are obeyed, so s/he checks proposed properties, object classes, representation classes and value domains to make sure that they are not duplicated. S/he also ensures that the names given to variables abide by the subsystem's rules. After conducting these checks, s/he approves or rejects the variables proposed by the *survey managers*[2]. S/he manages the subsystem's own decoding system and interacts with the IT technicians in the implementation and maintenance of the subsystem.

[2] *Survey manager* – This title is given to the statisticians (subject matter) in charge of each survey or to experts appointed by them. Their responsibilities in the system are as follows:
- At the end of the year and on an annual basis, the survey manager enters the survey plan for the following year into the planning system.
- S/he drafts the feasibility study.
- Propose the concepts, classifications and variables to be used in the survey.

2.5. Search and management applications

The applications supporting the variables subsystem are Web applications developed with the ".NET" platform. The subsystem has a bilingual consultation application and a management application.
The management application therefore implements two profiles: the *system manager* and the *survey manager.* There is a generic profile for consultation. The variables are accessible through the home page of the Official Statistics Portal.

III. Relationships with other systems

3.1. Concepts subsystem

The relation between these two subsystems is quite simple and direct, and is established among the conceptual variables and the concept itself. It is a bidirectional link, and it means that a new conceptual variable can induce a new concept in the database and a new concept can induce a new variable. Also a new version of concept brings a new version of variable (and vice versa).

3.2. Classification subsystem

The relation between variables subsystem and classification subsystem is more complex and is become fulfilled through a view. As in the case of the concepts, the relation is bidirectional and it is linking the value domain entity to level of classification. Also a new version of a classification brings a new version of variable (and conversely).

3.3. Methodological documents subsystem

The methodological documents [11] list the variables manipulated by the respective statistical operation. In the context of statistical operations, the variables are considered as observation, derivatives and statistical indicators. So, the survey manager can bring on, automatically, all variables that s/he needs to describe the survey and the information that disseminates.

3.4. Data collection instruments subsystem

In the component data collection instruments, both questionnaires and files have links to variables subsystem. The files may result from statistical operations, but also from administrative data deriving from external entities. For this relationship, the affectation of variables is executed only in one direction: from the variables subsystem.

3.5. Production system

Universes and samples management system

This system [9] is in its initial implementation phase and its purpose is the integrated management of an annual universe frame to support all the surveys based on the "enterprise" statistical unit. From the variables subsystem, we expect the use of statistical units and variables.

Statistical burden indicators

This system is in the planning stages and, when implemented, will be a tool for analysing statistical burden and the enterprise response rate [9]. From

the variables subsystem, we expect the use of the variables observed in the questionnaires.

Questionnaire design

The relation between questionnaires and variables, at this moment, is realised through an unidirectional link (see 3.4.), but the relation variables – questions/ question blocks/ questionnaires is not yet developed. Henceforth, through a dynamic link between variables - questions [4], we will be able to have a list of inquired variables and all associated metadata (code and designation, associated concept and classification, representation class, statistical unit, unit of measure, etc). We believe that the variable must be linked to questions because it expresses the concept that is being measured by the question. The variables value domain will be linked to response choices because each describes the valid values the data will take.

3.6. Dissemination database

The dissemination database was implemented to support the Official Statistics Portal [8] and represents the final repository of aggregated statistical information. The dissemination database 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 [12].
All the statistical indicators available in dissemination database have the associated metadata registered and approved previously in variables subsystem.

IV. Statistical Indicators

4.1. Statistical indicators definition

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 [14].
In practical terms, translating this definition for its applicability in variables subsystem and dissemination database, a statistical indicator is defined on the basis of variables [3] and results from the combination between aggregate variable and dimension variables (figure 3).
For the correct definition of the statistical indicator are indispensable two dimensions: the time dimension and the geographic dimension.

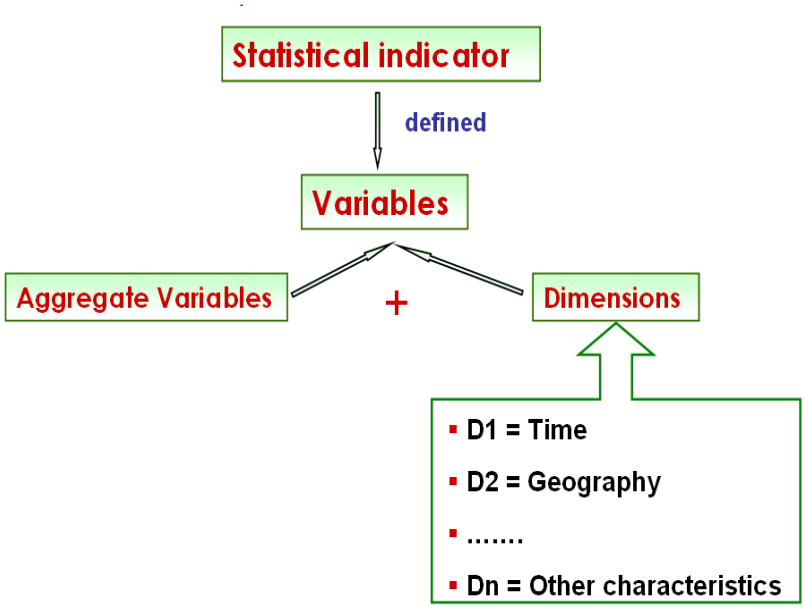

Figure 3. Statistical indicator – practical definition

Where:
Both aggregate variable and the different dimensions are considered variables and are registered previously in variables subsystem, in an independent way [4].

4.2. Statistical indicators naming convention

The naming convention [3] follows syntactic and semantic principles and the rules are basically identical for English and Portuguese.

Example:

- Resident population (No.) by place of residence and sex
- Enterprises (No.) by geographic localization, economic activity and legal form

4.3. Transmission and visualization

The definitions of approved statistical indicators are transmitted through a view (figure 4) and the unique identification is assured by the cross reference code, previously registered in variables subsystem [3].
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 databases, as we have specified.
The transmitted data for each statistical indicator must respect the structure and definition already registered in variables subsystem.
After data and metadata approval, the statistical indicator is published on the Official Statistics Portal.

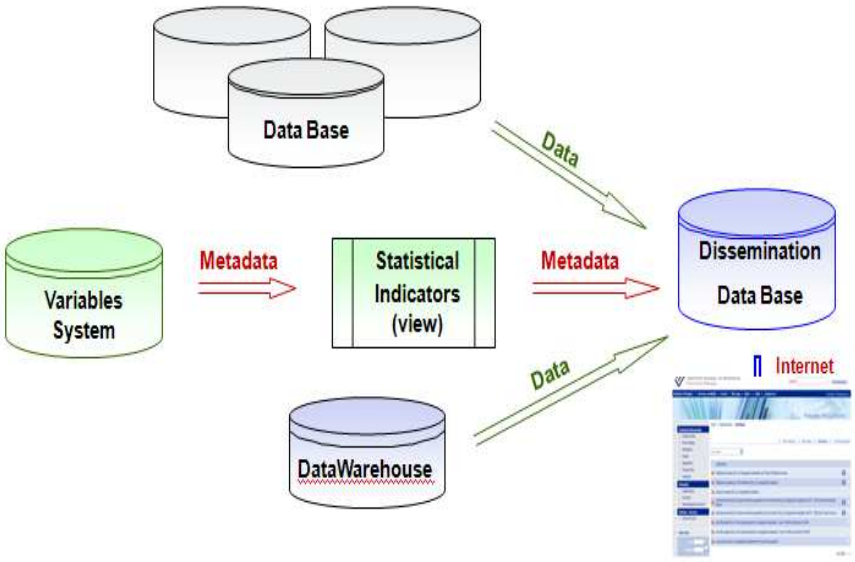

Figure 7. Statistical indicators - transmission and visualization

V. Normalization and harmonization

As a result of decentralized storage of data, we might find the same variable name defined in different ways in different parts of the organization, and also find the same variable definition named in different ways [1].
The main goal of normalization and harmonization is to introduce systematically a set of core harmonized variables in each survey or dissemination area [2]. This allows comparisons to be readily made over time, across geographic areas, industries and other domains. Data produced from different sources and at different times can be brought together to provide a richer dataset for analyses.
With these goals, come benefits in the form of reduced redundancies, fewer anomalies, improved efficiencies, increased comparability and decreased statistical burden.
The methodology adopted [5], consist basically on three major steps:

- Conceptual analysis

A major component of variable's conceptual analysis is identification and documentation of potential incompatibilities

- Normalization

The normalization represents a few steps, imposed by the rules of the Integrated Statistical Metadata System, specifically the variables subsystem (naming, definition, representation, etc.). All registered variables in the variables subsystem are normalized or normalized and harmonized.

- Harmonization.

Harmonization process is largely based on the first two phases and, in practical terms; the final result is a standard file (harmonization proposal) for the proposed variables. The harmonization proposal contain all the historical information, main concept, definition, coding process, associated classification, utilization, operational issues, international recommendation, good practices, etc.

The harmonized variable approved by the Consulting Group[3] is established as a variable preferred for use in data interchange and in new or updated applications. The harmonized variable may be unique within the registry, or it may be preferred variable among similar variables.

VI. Benefits

The implementation of *variables subsystem* and the application of normalization process increased the chances of sharing data and metadata with other statistical organization and improved the comparability of data and surveys.

Nowadays Statistics Portugal makes efforts to decrease respondent burden and production costs by increasing the use of administrative sources. With survey and administrative sources well documented and with a systematic registry of variables and their use, survey managers can even re-engineer statistical information system decreasing the number of surveys and increasing the use of administrative sources.

With a central reference of aggregated data (dissemination database) and metadata (variables subsystem), and through a standardized data and metadata transmission, we provided a common "look and feel" of all disseminated statistical information, improving quality and understandability.

VII. References

[1] Bargmeyer, E. B., Gillman, W. D., “Metadata Standards and Metadata Registries: An Overview”.

[2] INE/ DME/ SSM, (2002) “Sistema de Informação Estatística sobre as Famílias”.

[3] Isfan, M (2008), “Variables System – the bridge between metadata and dissemination”, paper presented at European Conference on Quality in Official Statistics, Rome, Italy.

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

[3] The Consulting Group is formed by delegates from Production Division, Dissemination Unit and Methodological Unit.

[5] Isfan, M (2007), “Variables – harmonization and normalization”, paper presented at 56th Session of the International Statistical Institute, Lisbon, Portugal.

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

[7] 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.

[8] 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.

[9] Morgado, I. and Isfan, M. (2008), “Case study – Statistics Portugal”, paper presented at METIS, Luxembourg.

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

[11] Morgado, I., (2004), “Metadata and Survey Documentation Portuguese NSI Experience”, paper presented at European Conference on Quality and Methodology in Official Statistics, Mainz, Germany.

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

[13] Sundgren, Bo, (2004), “Objects and their Classifications, Relations, and Life Histories – as Reflected by Official Statistics”.

[14] 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.