"A modernization project in the INE of Spain: Microdata repository and structural metadata"

Ana I.Sanchez-Luengo, María del Mar Ballesteros, Aranzazu Baltanas
Dept. Methodology and Development of Statistical Production
Statistics Spain (INE)
Paseo de la Castellana, 183 28046 Madrid (Spain)
April, 2017

Abstract: Within the framework of standardisation of the statistical process, INE Spain is working on the development of a microdata repository. This project is based on three fundamental pillars; the development of structural metadata for improving data management, the development of sound methodologies for processing data in the different stages of the statistical process, and finally, the development of computer tools for accessing all the information easily and efficiently.
The model proposed in this paper uses structural metadata, i.e., variables, classifications, concepts and statistical units as an information system, key for the storage of microdata. These metadata are stored once in a single institutional repository and should be used and reused as much as possible. Designed and developed, the model is beginning to be implemented at INE-Spain.
The structure metadata model allows to build microdata bases and access them easily. The access can be made to data with different temporal references and in different subprocesses of the data collection phase.
As regards the quality of metadata, a strategy has been put in place where the metadata unit is responsible for building, maintaining and improving structural metadata, applying international standards as much as possible, always bearing in mind the institutional needs.

Introduction

The ESS Vision 2020 [1, 2] for the modernization of the European Statistical System defines a modular production system based on three pillars: sound methodologies, corporate metadata systems -both process and structural metadata-, and modular, reusable and interchangeable IT tools.

This kind of production systems must be built in an efficient way, what it clearly leads us to the need of cooperating between countries and applying to international standards. In this sense, several projects have been developed at international level, most of them supported by UNECE, such as GSBPM[1], GSIM[2] or CSPA[3], with the aim of paving the way towards the industrialization of statistical production and the sharing of methodologies and tools. At the same time, users are provided with standardized reports such as the Euro-SDMX Metadata Structure (ESMS) [3] or the ESS Standard for Quality Reports Structure (ESQRS) [4].

For some years now, INE has been working on the development of a structural metadata model based on international standards. The work carried out since 2000 by United Nations has been continued, adopting GSBPM v.5.0 as a standard language to describe the statistical production process, and using GSIM v.1.1 to describe the objects that will be used in that process.

Regarding the adoption of the GSBPM, INE has developed a national version that consists of the development of a third level, tasks, with two objectives: document and describe the processes so that they can be replicated. And identify the elements of production for the adoption of common tools and methodologies.

Since 2015 INE Spain has been working on the creation of a microdata repository [5], with the same data set structure based on a key-value scheme, for all statistical operations. This common data structure allows the design and development of standardized tools for statistical process. Since the end of 2016, this microdata is linked to a single institutional structural metadata system. The connection between the

[1] Generic Statistical Business Process Model (GSBPM): https://statswiki.unece.org/display/GSBPM/GSBPM+v5.0

[2] Generic Statistical Information Model (GSIM): https://statswiki.unece.org/display/metis/Generic+Statistical+Information+Model

[3] Common Statistical Production Architecture (CSPA):https://statswiki.unece.org/display/CSPA/CSPA+v1.5

repository and the structural metadata system is made through the key-value data structure, where the key is a composition of the structural metadata and the value of data itself.

1. Structural metadata for data management

The term "statistical data", originating from the Latin "datum", refers to information that provides access to precise and specific statistical knowledge. For statistical data to be useful, it must be organised and considered from a specific context. Data itself, studied as an isolated element, lacks interest. It is essential to link the metadata to the statistical microdata structures in order to organize the information for its production, process, dissemination and reuse.

Structural metadata make possible to identify and describe the data. They are necessary for the identification, use and processing of matrices and data cubes [Source: SDMX[4]]. We can differentiate two categories of structural metadata: metadata that let us describe the data conceptually (variables, concepts, classifications and statistical units) and metadata that allow us to identify the data, contextualizing it within its domain (for example, temporal reference, statistical operation, etc.)

2. Databases in the statistical production model according to GSBPM

Statistical institutions are committed to produce high quality, accurate and reliable statistics that are consistent and comparable in order to provide clear and accessible information. For this reason, the data sets are very important products that need to be adequately defined through the data structures.

In the GSBPM, in phases 5 "Process" and 6 "Analyse", the objects used as inputs are the databases. In the national version of the GSBPM [6], databases have been considered a production element, defined as "the organized collection of data and metadata, including microdata, estimated population aggregates (aggregates) or metadata. They can be structured as collections of files or have more complex designs (relational designs, NoSQL, etc.)".

The following tasks (level 3 of the GSBPM for national purposes) are related to the design and updating of the databases. Please note that the first number of the task code stands for the GSBPM phases, the second one for the sub-processes and the third one for the task.

- Task 2.6.4 "Design data and metadata databases": includes deciding which variables (microdata, estimates of population aggregates, structural metadata, etc.) are included in the database, its structure (relational model, pair structure: key-value, etc.) as well as any other detail that determines its final structure.
- Task 3.2.19 "Build IS (Information System) components for microdata and metadata databases" and 3.3.2 "Build IS components for data and metadata dissemination databases": include the computerized implementation of each database related to data processing.
- Task 5.8.1 "Update microdata and aggregate databases": includes the results (estimates and metadata) from the last execution cycle in the corresponding process databases.
- Task 7.1.2 "Update dissemination databases": includes the dissemination product from the last execution cycle in the corresponding databases.

Depending on the different spatial scale of data there are two types of databases: Microbases (individual data) and Macrobases (Aggregated data). Moreover, these bases can be defined according to the phase of the process in which they are. The repository of microdata has been designed and built for the microdata bases and depending on the task we can identify several microbases containing the following files:

- Task 4.3.3 "Execute data entry". Files with row data (FG files).
- Task 5.3.1 "Execute error detection and treatment (input)". Files with data edited during collection (FD files).

[4] SDMX: https://sdmx.org

- Task 5.8.1 "Update microdata and aggregate databases". Files with final data validated by the domain survey unit (FF files).
- Task 5.3-4.2 "Execute error detection and treatment (output)". Files with data for the construction of validation intervals in the longitudinal phase of the selective editing strategies (FL files).
- Task 6.2.1 "Execute error detection and treatment (macro)". Files with data for the selection of influential units in the transversal phase of the selective editing strategies (FT files).

3. Single Integrated Metadata System (SIMS)

INE Spain has been working on a Single Integrated structural Metadata System (SIsMS) for several years as a core of the SIMS, together with the reference and process metadata. A metadata system should be integrated into the statistical production process. Therefore, it must have two basis properties, completeness and updated. "Completeness" means that all statistical operations carried out by the institution must have their structural metadata input into the system. "Updated" means that it must be a "live" system that is continuously fed with new data.

3.1 Objects in the SIsMS according to GSIM

To define the objects in the SIsMS, the GSIM v.1.1 has been used, considering two of its four sections. "Structure group", where the objects in relation to Data Set, Referential Metadata Set and information about their structures can be found. And "Concepts group", which comprises the information on the objects that describe or define when the practical implementation of a phenomenon measured in a statistic is made. This last group includes objects such as variables, concepts, classifications and statistical units.

To clarify this, let´s analyse the following example: we wish to access the data on the "Survey on ICT and Electronic Commerce use in Companies", whose code in the national inventory of statistical operations is 30169. Specifically, we want to extract the final data validated (File FF) for the "Turnover" statistical variable for region 08 "Castilla-La Mancha" referred to January 2016. The name of the desired file is "E30169.FF_V1.MM012016.D_1". The variable instance is (E30169.FF_V1.MM012016.D_1, Turnover, 08 "Castilla-La Mancha"). The value domain is integer, so the represented variable is (Turnover, Integer). The population are all enterprises include into Business Register and the unit type is enterprise. The modelling according to the GSIM can be seen in Fig. 2:

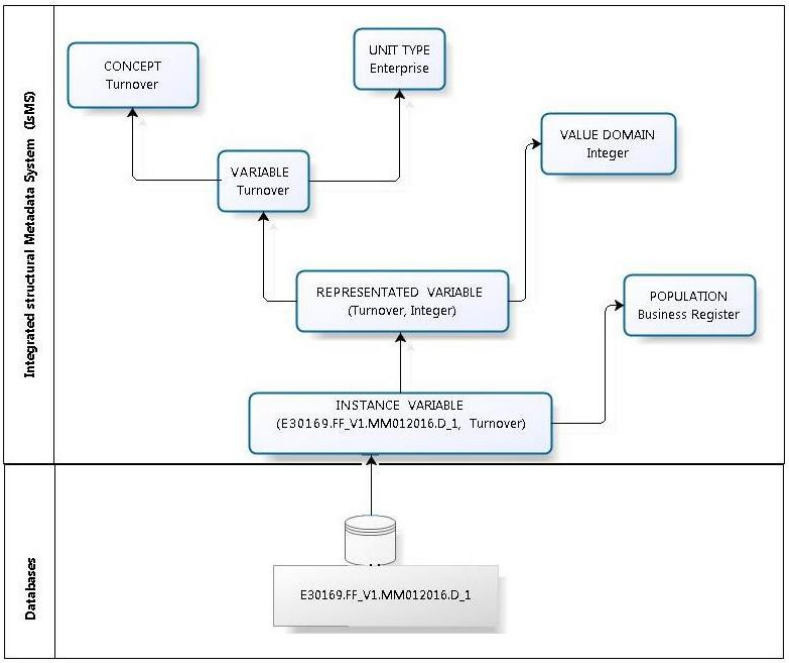

Fig.2 Metadata model in GSIM terms

3.2 SIsMS Information Systems

The model is based on four Information Systems (IS): concepts, variables, statistical units and classifications. These ISs are developed following international standards such as: Neuchâtel for classifications, ISO 11179[5] for variables and international definitions for concepts. In addition, as mentioned above, the GSIM has been followed to properly define the objects that will later be used in the statistical process described according to the national version of the GSBPM.

Concepts IS: Includes the definitions related to statistical variables, units and classifications used in the INE production system. The concept IS is available on Internet web page of INE[6].

The concepts are considered fundamental for the comprehension of the data, and it must be taken into account that a person's interpretation of the data depends on their frame of reference, which consists of the previously known information [Bo Sundgren States]. The concepts will help people, with different frames of reference, to understand the data in the same way and manage them properly.

Statistical Variables IS: Especially relevant in the statistical process in the need-identification phase and in the design phase. The identification of variables consists of highlighting the characteristics of the statistical units that are of interest. In the design phase, these variables will be made operational.

Classifications IS: the general idea of Classifications IS is to assign values to a statistical variable. This is the classic conception of a statistical classification, that is, a set of discrete, exhaustive and mutually exclusive observations, which can be assigned to one or more variables to be measured in the collection and/or presentation of data. The IS extends this idea. On the one hand, the IS also considers the classifications formed with the different values that a statistical variable can take. For example, the values of the statistical variable "Marital status" (translates into 1 data point). On the other hand, the classifications formed with the breakdown of a statistical variable are also included. For example, "Employed personnel according to compensation" (translates into several data points).

Statistical units IS: within the GSIM, the statistical unit is defined as a unit type for a particular statistical process, that is, the statistical unit is the object of study in a survey and the bearer of the statistical characteristics.

4. Project implementation

The project for the development of a microdata repository started with the current INE production model in order to develop a standardised modular statistical production process. At the meeting of December 18, 2015, the Board of Directors agreed to "the creation of a repository for final files collecting all the microdata files used for the products for the dissemination of statistical operations". At the end of 2016, work began on the "Survey on ICT and Electronic Commerce use in Companies". In the first quarter of 2017, it was agreed to implement this plan in some of the INE short-term statistical operation.

5. Conclusions

Database entries with a pair: key-value model has short-term benefits for the entire institution. This model allows to become more self-sufficient, reusing the standardized components developed for the extraction and inclusion of information. At the technological level, it will facilitate a more efficient way of working among different statistical domains when developing easily adaptable modular tools that comply with the requirements of the common statistical production architecture.

In addition to the information extraction and recording processes, other processes can be standardized, such as: the detection and treatment of input, output and macro errors or the updating of other databases, such as aggregates.

[5] ISO 11179: http://metadata-standards.org/11179/
[6] DEFine: http://www.ine.es/DEFIne/?L=0

6. References

[1] Eurostat (2014a). ESS Vision 2020

[2] Eurostat (2014b). Vision 2020 Implementation Portfolio

[3] COMMISSION RECOMMENDATION of 23 June 2009 on reference metadata for the European Statistical System

[4] ESS Standard for Quality Reports Structure (ESQRS) 2.0

[5] David Salgado, M. Novás, E. Esteban, S. Saldaña, L. Sanguiao (2017) "A lightweight key-value pair data model implementation for a standard production process"

[6] D. Salgado and A.-I. Sánchez Luengo (2016). Process metadata development and implementation under the GSBPM v5.0 at Statistics Spain. European Conference on Quality in Official Statistics (Q2016). Madrid, 31 May–3 June, 2016