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

The Portuguese classifications database (SINE)

Isabel Valente
Methodology and Information Systems Department
Statistics Portugal[1]

1. The conceptual model

The Integrated Statistical Nomenclatures System (SINE) is an integrant part of one vaster system implemented on Statistics Portugal, the Integrated Statistical Metadata System (Fig. 1), started in 2002. Of the global Statistical Metadata System other subsystems are integrant parts nominated: concepts, classifications, statistical sources (includes data collection instruments and surveys) and variables.

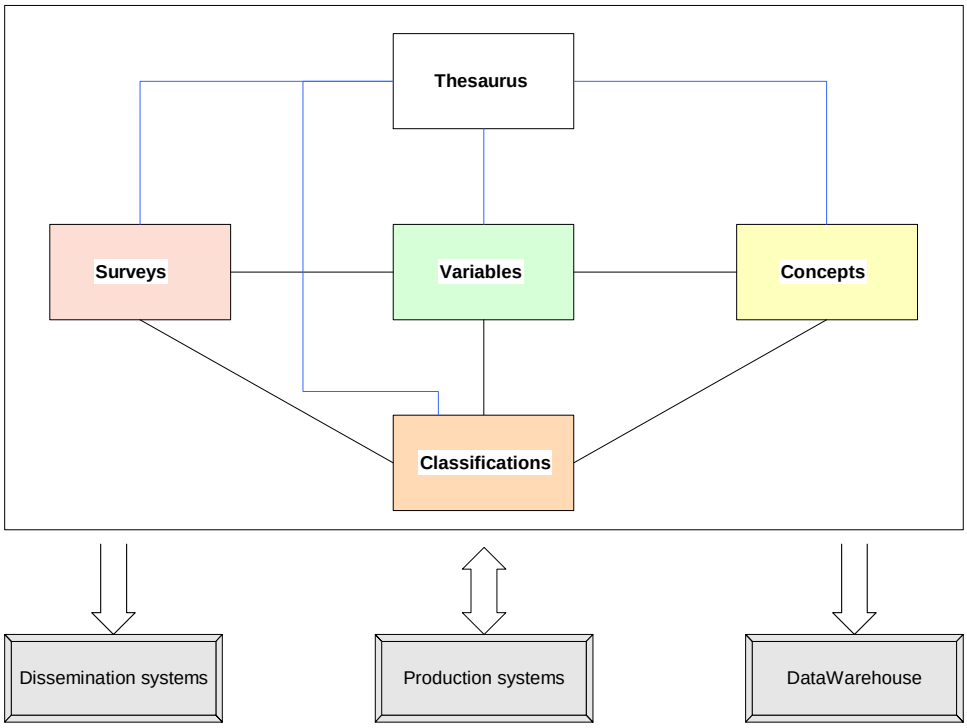

Fig.1 *Macro Architecture of the Statistical Metadata System* [2]

[1]Isabel Valente, Metadata Unit, Methodology and Information Systems Department, Statistics Portugal, Av. António José de Almeida, n.º2, 1000-043 Lisbon, Portugal, isabel.valente@ine.pt.

The study of SINE has begun in 2002 and was the first subsystem to be developed in accordance with this new perspective of integration of statistical metadata. The conceptual model developed by the Neuchâtel[3] group of which result a common terminology for use in object types of databases of classifications and its attributes served as base for SINE. The SINE database is constituted around a set of key concepts and related attributes considered as relevant. The conceptual model of SINE is presented in figure 2.

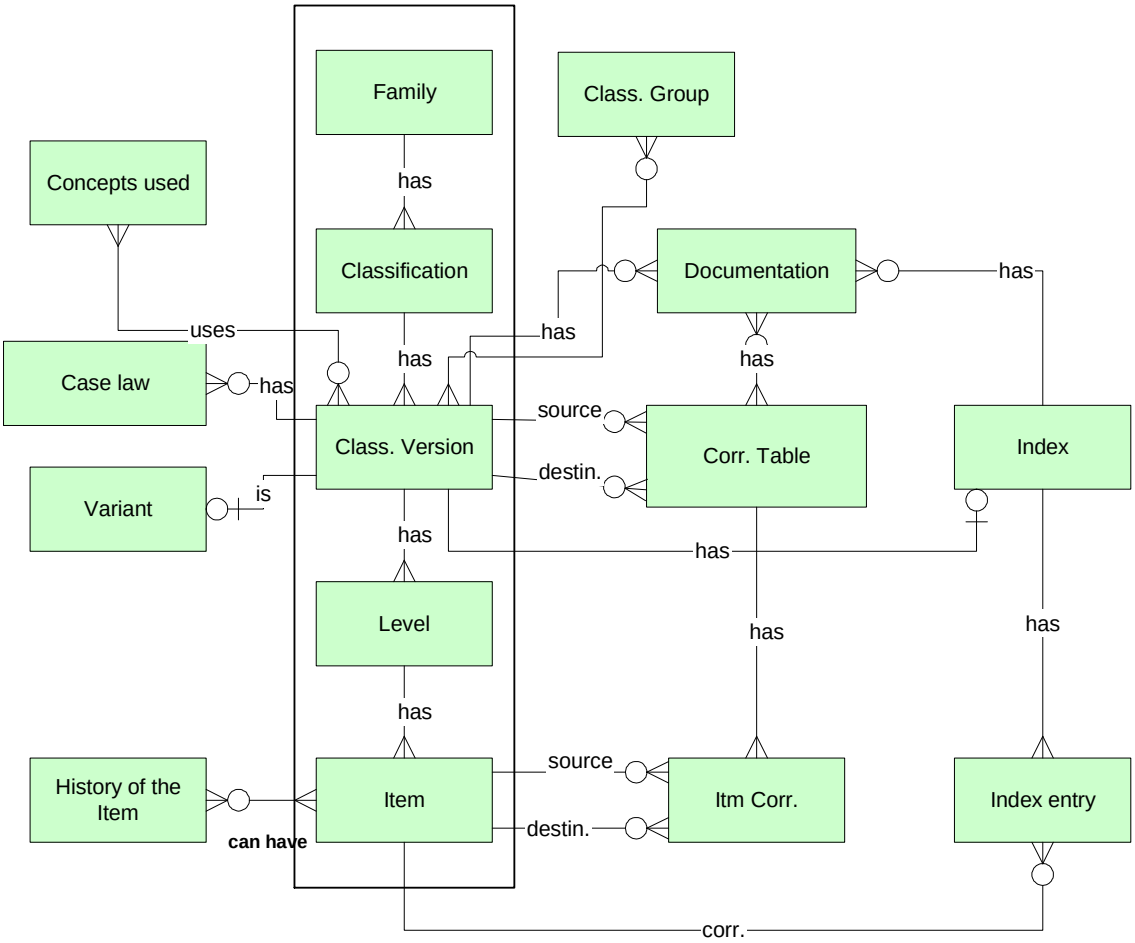

Fig.2 Conceptual model of the Integrated System of Statistical Nomenclatures (SINE)

[2] In, Morgado, Isabel, "Metadata and survey documentation portuguese NSI experience", European Conference on Quality and Methodology in Official Statistics (Q2004), 24-26, May, 2004, Mainz-Germany.

[3] Statistics Denmark, Statistics Norway, Statistics Sweden, Swiss Federal Statistical office, Run Software-Werkstatt; Neuchâtel Terminology: Classification Database Object Types and Their Attibutes", 5th September 2002.

2. SINE and SINEG components

SINE is a reference system on classifications of national, communitarian and international scope, used for statistical purposes. When we speak in classifications we mean not only standard classifications but also code lists that support variables. This system is accessible inside and outside INE (through intranet and through the home page of the Official Statistics Portal) however, the code lists are, saved some exceptions, of internal use.

The classifications system is, at present, composed by two components one of consultation (SINE) and other of management (SINEG). The applications who support SINE/SINEG are Web applications developed with the ".NET" platform. The consultation component is multilingual supporting, at present, three different languages namely Portuguese, English and French while management component exists in Portuguese language only. In the case of translations they are only allowed if the information already exists in Portuguese language. Both components are structured around a hierarchic structure arrogated by the entities referred in the central axle of figure 2. It is also around these entities that the navigation system of SINE and SINEG was established.

Classifications are arranged by families, being a family understood as a set of classification with technical and functional affinities (ex.: products, economical activities, etc.). They had been established 18 families not being foreseeable, at short or medium term, to increase or to diminish the number of families fixed.

A classification is understood as a structured list of discrete, exhaustive and mutually exclusive categories, identified by codes and names, that aims to assign all the elements of a certain population to categories representing the values of a certain variable. A version is an instance of a classification valid for a given period of time. So, in SINE, a classification has always associated one or several versions. (ex.: the Portuguese classification of economic activities has 4 different versions, CAE Rev.1, CAE Rev.2, CAE Rev.2.1 and CAE Rev.3.)

Classifications are distinguished towards time variability being this of two types: floating (when their items have different validities ex.: Code of the administrative division) non-floating (when they are valid for a specific period in time ex.: CAE Rev.2.1). They are also distinguished in terms of their grade of formalization. That is, we consider classifications those who possess a certain formal statute being considered the other code lists. Those distinctions, established at the management level, are not reflected to SINE. In terms of SINE, at present, is only visible the distinction between classifications floating and non-floating given that the first ones present a calendar for opposition to the second ones.

Is associated to the version that appears the items of a classification and is also associated to the version entity that are available the biggest number of functionalities namely: characterization, levels, variants, correspondences, indexes, publications, concepts, downloads, search of code and word. The number of versions in SINE is variable, existing at present (January 2009) 1513 versions, 284 of which available for the exterior. The significant difference between what it is made available outside INE and what exists internally is arrested with the fact of great part of these versions being code lists. For the opposite, downloads of versions and correspondence tables work in the same way inside and outside the institution being available in both cases in csv format.

SINE also makes available a set of reading functionalities on classifications versions grouped by the title “Other searches”. This block allows the consultation of the information through specific fields as they are owner, obligatoriness among others. Inside this block we point out the options Group and Word for its specificity. A group is understood as a set of classifications which integrate methodological systems of reference (ex.: System of National Accounts, 1995). By selecting a specific group all the versions that integrate this group appear. A version can integrate one or several groups. This functionality is also used to collect versions used by surveys. The search by word allows search to the database by designation or acronym of classifications, versions or items entities or in alternative to the three entities simultaneously.

The development of consultation (SINE) and management (SINEG) applications was not given in simultaneous. The gap occurred did not figure a good solution therefore,

the availability of the management module showed to be a basic step to gain autonomy in the management of the classifications and, over all to get better the consultation and management model.

2.1 The management of SINEG

The management of SINE is, at present, a centralized management being the Metadata Unit (SM) the responsible of the management of all classifications. The decentralization of SINE was foreseen however, for several changes did not figure itself, until the moment, as a viable solution; however, inside SM, some operational managers of classifications have the possibility to manage the classifications for which they are responsible.

In SINEG the insertion screens are variable according to the entity where we are. The insertion of information on those screens is fulfilled in four different ways: by direct insertion, by selection through combo boxes, by association from a list and by importation.

The information inserted by direct entry respects information that cannot be standardized. The information inserted by combo boxes is information that has a repetitive character and that is therefore susceptible to normalization. That kind of information constitutes what we call the reference tables of the system. Information like the “Owner” or “Contacts” works by accessing to a total list. Being in that list is only necessary to choose the entity wanted and associate it to the respective version. In the case of concepts SINEG has access to a view of all the concepts existing in the concepts database. By introducing a specific number of a concept the application search that number and shows the different concepts that exist under that number. Chosen the wanted concept this is associated to the specific version where we are.

The last form of insertion used is importation. It is mainly used for the insertion of items and correspondence items, where the set of information is usually very extent. It works through mdb files where tables have been standardized according with the fields need in SINEG application. Those mdb files also have a set of associated queries which convert the information in those tables to a txt format readable by SINEG application.

In SINE exists at present a wide set of information, because of that, in elapsing of 2008, we feel the necessity to establish rules for the normalization of the writing of classifications, versions and items names viewing to simplify and harmonize this same writing. We count in the future to deepen these rules as well as finding a form to distinguish in a formal way classifications from code lists.

3. Relationships of SINE with other systems

As we relate previously, SINE, is available in intranet and in Portal of INE, what it showed to be a privileged vehicle in the accessibility of this information and, consequently in the biggest visibility of the classifications. From the moment where this database started to exist, with some volume of information, other systems had been able to take off part of its existence. For example, the system of variables it goes to search classifications registered in the system, as a value domain of observation variables as of diffusion. Also the system of methodological documentation uses SINE because is uses the register of the SINE version for the description of classifications used in some statistical operation and, on the other side, helps to feed SINE by registering new classifications that might be able to be integrated in this system. Also connexions between the classification database and the concepts database have been established providing that concepts have an association to the classifications where they are used.

SINE constitutes now a central repository of classifications, because of that the use of SINE is not limited to the relationships with the other metadata subsystems. Different projects and applications make use of SINE in Statistics Portugal. This use is made essentially by two ways: through downloads; by accessing the views of existing tables. Once gotten the intended information this is reverted for the system or project that uses it according with its own specific purposes (ex.: international trade or Data Warehouse).

4. Conclusions

The systematization, centralization, normalization and a relative easiness in the use of SINE, had allowed a bigger generalization in the use of classifications. At the same time the process of register, systematization and normalization of classifications allowed at internal level, to prevent duplications, to increase the coherence of the information and to improve the knowledge on existing classifications.

However the use of SINE and the tests of usability also disclosed that the consultation format, available at present, does not answer to the needs of the different types of users. So, despite the conceptual scheme remains in the future the same, it will be looked for a visualization format that simplifies the use of SINE for inexperienced users but that, in simultaneous, allows to demanding users the access to all available information. The diversification of the search options, the joining of some screens, considered superfluous, and the change in the access to some functionalities constitutes some of the improvements that we intend to implement.

Also the experience in last years with SINEG proof to be very reach. Because of that some changes will be foreseen as predictable and desirable towards its simplification. We also intent to proceed at improvements in the structure of the underlying database in order to decrease response times in consultation and management of SINE with sight to a better performance.

To the similarity of the applications we also intend to effect improvements in the congregated information. To keep updated the existing information, to complete it, improve its coherence, relationships and harmonization are also purposes to reach.

In conclusion, in the next years, we will look for to improve, deepen and strengthen the system having in account that this end will be only reached when we will be able to give reply to the needs of the different users.

References

Anne Gro Hustoft, Statistics Norway, AMRADS training workshop on Metadata, 13th March 2003.

Eivind Hoffmann, Bureaus of Statistics, International Labour Office and **Mary Chamie**, United Nations Statistics Division, Standard Statistical Classifications: Basic Principles, Statistical Commission, Thirtieth session, New York, 1-5 March 1999.

Morgado, I. and Isfan, M., "Case study – Statistics Portugal", paper presented at METIS, Luxembourg, April 2008.

Statistics Denmark, Statistics Norway, Statistics Sweden, Swiss Federal Statistical office, Run Software-Werkstatt; Neuchâtel Terminology: Classification Database Object Types and Their Attibutes", version 2.0, 5th September 2002.

Sundgren, Bo, Objects and their Classifications, Relations, and Life Histories – as Reflected by Official Statistics", 2004.

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

United Nations, Economic and Social Council, ECE/CES/2008/3, of 2 April 2008 prepared by the **Economic Commission for Europe and Statistical Commission, for the** "Statistical Metadata", Conference of European Statisticians, fifty-sixth plenary session, Paris 10-12 June 2008.

UN, Glossary of Classification Terms, UN, 2003