WP.11
ENGLISH ONLY

UNITED NATIONS STATISTICAL COMMISSION and ECONOMIC COMMISSION FOR EUROPE
CONFERENCE OF EUROPEAN STATISTICIANS

EUROPEAN COMMISSION
STATISTICAL OFFICE OF THE EUROPEAN COMMUNITIES (EUROSTAT)

ORGANISATION FOR ECONOMIC COOPERATION AND DEVELOPMENT (OECD)
STATISTICS DIRECTORATE

Joint UNECE/Eurostat/OECD work session on statistical metadata (METIS)
(Geneva, 3-5 April 2006)

Topic (ii): Metadata Concepts, Standards, Models and Registries

STATISTICAL METADATA MODEL DEVELOPED IN SPAIN: CURRENT AND FUTURE USE AND APPLICATIONS

Supporting Paper

Submitted by Instituto Nacional de Estadística (INE), Spain[1]

I. INTRODUCTION

1. This paper covers two main objectives. In a first place, the article describes the metadata model functional description and the current maintenance status. In addition to that, the core of the paper is devoted to explain the metadata model role in the Spanish National Statistics Institute.

2. The main objective of the metadata project is to build a tool in order to facilitate the integration and co-ordination of the whole information requested by INE to data providers. Generally speaking the project will allow INE to produce information more harmonised, rationalised and, specially, more comparable. Also data users will get ready a tool to get more information about every statistical operation performed by INE. Moreover, in the future this project will favour the co-ordination among the different actors of the Spanish Statistical System (INE, Ministerial Departments, Regional Statistical Offices, Spanish Central Bank…).

[1] Prepared by: Mar Blanco Frías (mblafri@ine.es) and Ana Isabel Sánchez-Luengo (anaisan@ine.es)

II. METADATA MODEL DESCRIPTION AND STATUS

3. Any model can be seen from quite different points of view. In order to give a complete view of the model, the following paragraphs explains it from the functional, management and user points of view.

A. Logical Architecture and management of the System

FIRST PHASE

4. The first phase of the project has as input the Statistical Operations performed in the Institute. To understand this phase the following definitions are needed:

a. *Statistical Operation*: set of activities leading to obtain statistical results about a specific sector or issue, from individually collected data using surveys or public sources.

b. *Variable*: information about the unit being studied. From the practical point of view, and in most cases these variables use to correspond to a unique question in the questionnaire.

c. *Classifications*: Set of classes or categories, defined using codes, which fix the values that a variable can get. When talking about an international regulation, we talk about standard classifications.

d. *Standard Classifications*: A standardized classification. There are international standard classifications, which derive from international recommendations and Regulations, and standard national classifications, which statistical use is recommended or mandatory due to a national regulation.

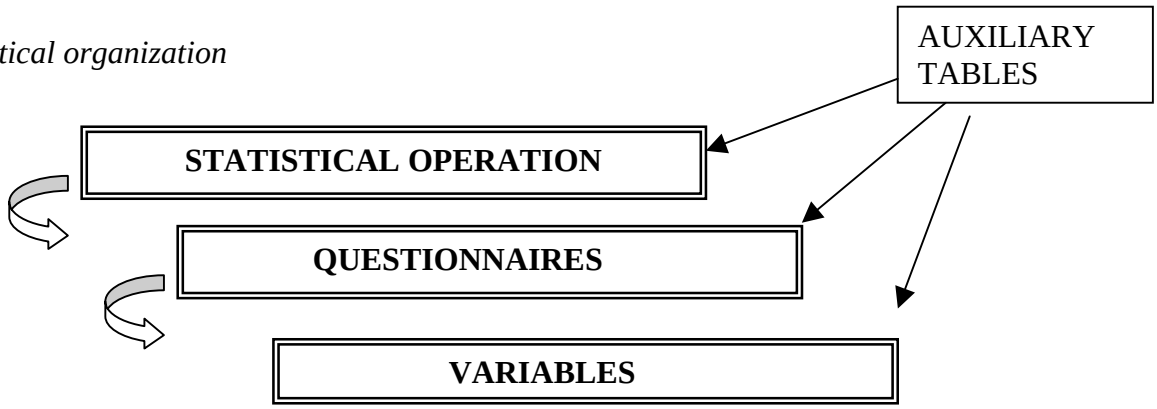
e. *Thematic classification of classifications and variables:* Internal classification at 3 levels that allows to organize information so that to make possible doing comparisons between different statistical domains.

5. The objective is to have a common repository in which all the surveys are included but not independently but establishing relations between variables and classification in different surveys by means of thematic classification, so that to be able to compare and therefore harmonise the information. The final objective is twofold:

a. In one hand, to achieve a high degree of both internal and international harmonization that allows to perform better comparisons.
b. On the other hand, to provide the users with a tool which allows them to achieve a better understanding of the information.

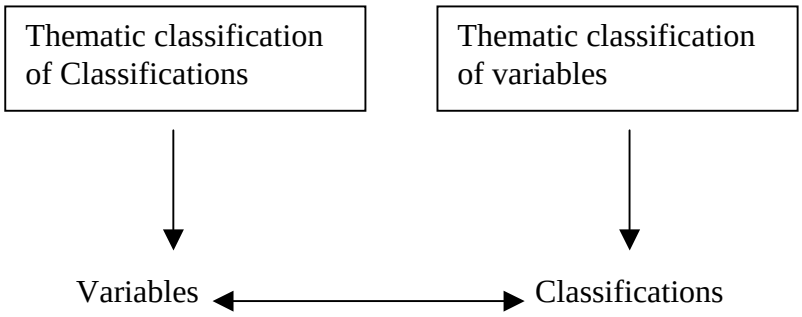
6. Functionally speaking the model has been built as a relational database supported by Oracle. This Data Base is composed by interrelated tables which follows a hierarchical structure. This hierarchical structure responds to a two types of organization of the information:

a. Vertical organization

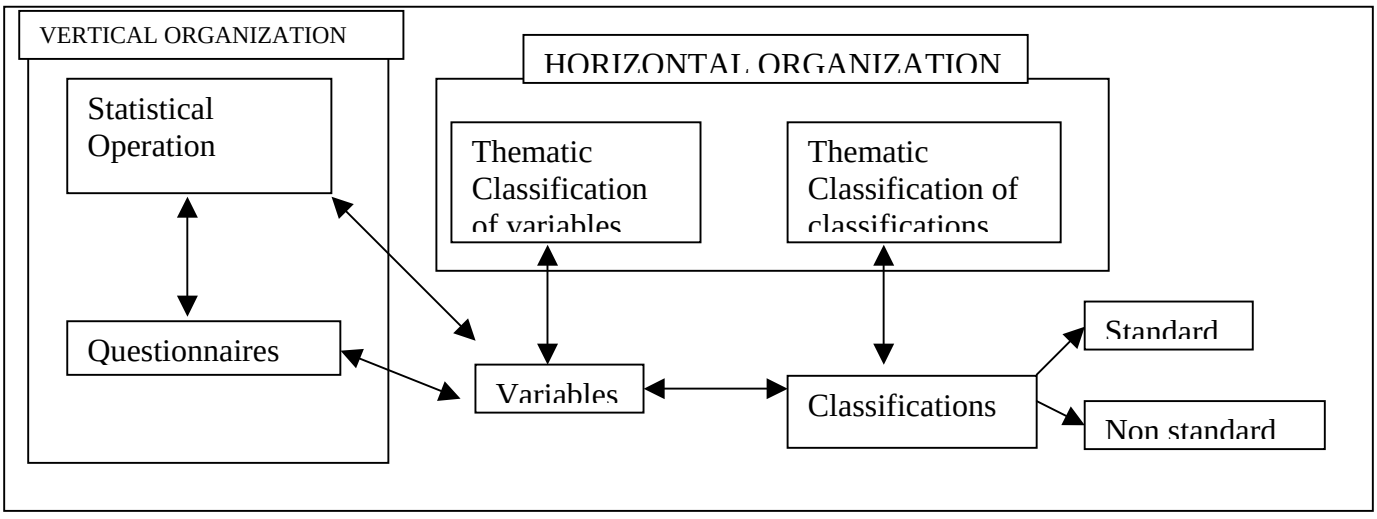


b. Horizontal Organization

7. Variables are linked to classifications. Both variables and classifications are classified thematically using 3 levels of detail. This will allow thematic searches used to compare the information in the different Statistical Operations, so that to be able to harmonize the way information is requested by the different Units. In addiction to thematic searches, there exists also the possibility to search information about variables, Statistical Operations and classification by literal and by global character chains, as we will see in the next paragraph.



8. The following chart summarises the main idea of this first phase:



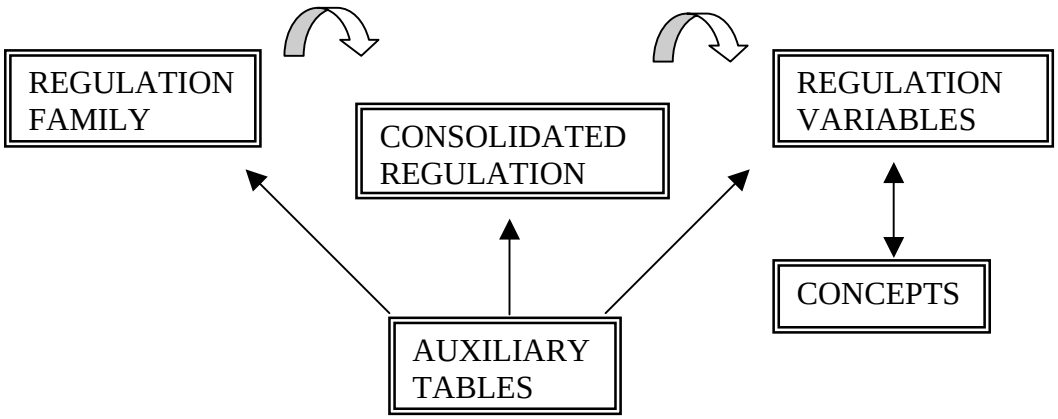
9. The maintenance of the system is carried out using a metadata management system instead of working directly on the database tables. This prevents the operator from making mistakes when introducing new information or when any modification is needed. See annex (FIGURE 2) for a picture of the main window of the management system (currently the development is available only in Spanish). The management system allows to introduce and modify information about all the elements in the system: Statistical operations, variables, classifications, and also to manage the thematic classification in the classification and variables, in a easy and friendly way.

SECOND PHASE

10. The second phase of the project is devoted to international Regulations and standardised concepts. The objective in this case is to bear a special attention to those variables included in a Regulation, among all variables. To fulfil this objective, a second relational database has been constructed, based also in Oracle, which contains in a similar hierarchical model the Regulations for each Statistical Operation and the variables contained on them. In addition to this, the concept for all the variables in the Regulations are also included in the data base, if exists. For the time being, the information is included in Spanish and sometimes in English.

11. This data base allows:
    a. To ensure that all variables requested in the Regulation are indeed included in the questionnaires, and in the contrary side, which variables included in the questionnaires are not requested in the Regulation.
    b. To have a repository of definitions which allows the Official Statistic both national and international harmonization in concepts.

12. The following chart summarised the idea:



13. The databases constructed in the first and second phases are closely interrelated by means of a transition table, which contains the Regulation variables that are also in the Statistic Operations.
The management of this database is performed directly in the tables instead of having a management system like in the first case, due to the fact that it is not expected to find frequent changes in the Regulations as they are quite consolidated. The same applies for the concepts.

B. Search Engine

14. A very important actor in the metadata project as a whole, especially from the user point of view, is the so-called "Enquire System". It is nor useful to have great amounts of information stored in tables if you are not able to efficiently exploit this information as much as possible; currently the amount of information contained in the Data Base is:

System element	Number
Statistical Operations	102
Survey Questionnaires	253
Variables	8.187
Non-standardised classifications	1.434

15. Concerning the thematic classification, the following number of themes are available at each level for variables and classifications:

	Thematic classification of Variables	Thematic Classification of Classifications
# First level	28	17
# Second level	181	133
# Third level	445	535

16. The Enquire System allows the user to easily find different types of information using different criteria. It is available in the Intranet.

17. The exploitation tools have been developed following the main necessities of the INE internal users. However, it is an alive system in the sense that new tools can be developed so as to respond to new necessities.

18. The tools that compose this System are broad and varied. It allows to:

a. Perform Thematic searches: The user is able to perform a thematic search on the classification and also on the variables. By selecting a theme (at 3 digit level), the tool shows all the classification (resp. variables) classified in that theme, coming from all the statistical operations included in the Data Base.

   The user is able to select the classification(resp. variable) he/she is interested in and retrieve all the information about it:
   - a "technical file" of the classification (resp. variable) containing the statistical operation and questionnaire in where is located the classification (resp. variable), if there exists a equivalent classification (resp. variable) in other statistical operations, the type of variable…
   - in case of variables, the tools allows to see the classification associated with the variable (if any) and also if there is more variables associated with the same classification.

   These thematic searches allow the comparison of how the different questionnaires request information about the same theme.

It is also possible to search classification or variables by character chain instead of theme.

b. Perform Vertical searches: due to the hierarchical structure of the Database, it is possible to search starting from Statistical operations →Questionnaire→Variable. The user sees all the variables and all the classifications in the selection.

   It is also possible to download the selection (the physical questionnaire as the responsible Unit spreads it).

c. Compare Statistical Operations: By selecting two statistical operations the tool allows to study, by means of an algorithm which analyse similarity between variables and classifications:
   - if they have common variables (both qualitative or quantitative)
   - if there are variables or classification classified with the same thematic classification
   - If there are common or similar classification
   - And shows all the variables in both Statistical Operations classified by themes (at 3 digit level), so that the user can see, even though there are no common variables, if there are related variables, in the sense that they are classified in the same way.

d. Perform searches in the Regulations: Many of the variables included in the Statistical Operations performed at INE are in accordance to requests from the European Union or an international organization. This application allows to consult the variables included in the Regulations and its attributes and the relationship with the variables included in the Statistical Operations. There are two types of searches:
   - By statistical Operation: it allows to see all the variables in the statistical operation and which of them are related with a variable in the associated Regulation
   - By Regulation: it allows to see all the variables in the Regulation and which of these variables are related with a variable in the corresponding Statistical Operations

e. Perform searches in the concepts: it is possible to search a concept by literal or by theme. The source of the concept is a Regulation and also CODED (the Eurostat concepts and definitions data base).

f. Questionnaire simulation: By selecting a specific questionnaire, it is possible to "reproduce" the questionnaire as it is stored in the data base: the user will not see the final questionnaire used by the responsible Unit, but a more technical version containing all the variables and classification and its internal attributes, internal codes, etc…

19. The following plot summarises the metadata model based on variables, concepts and Regulations:

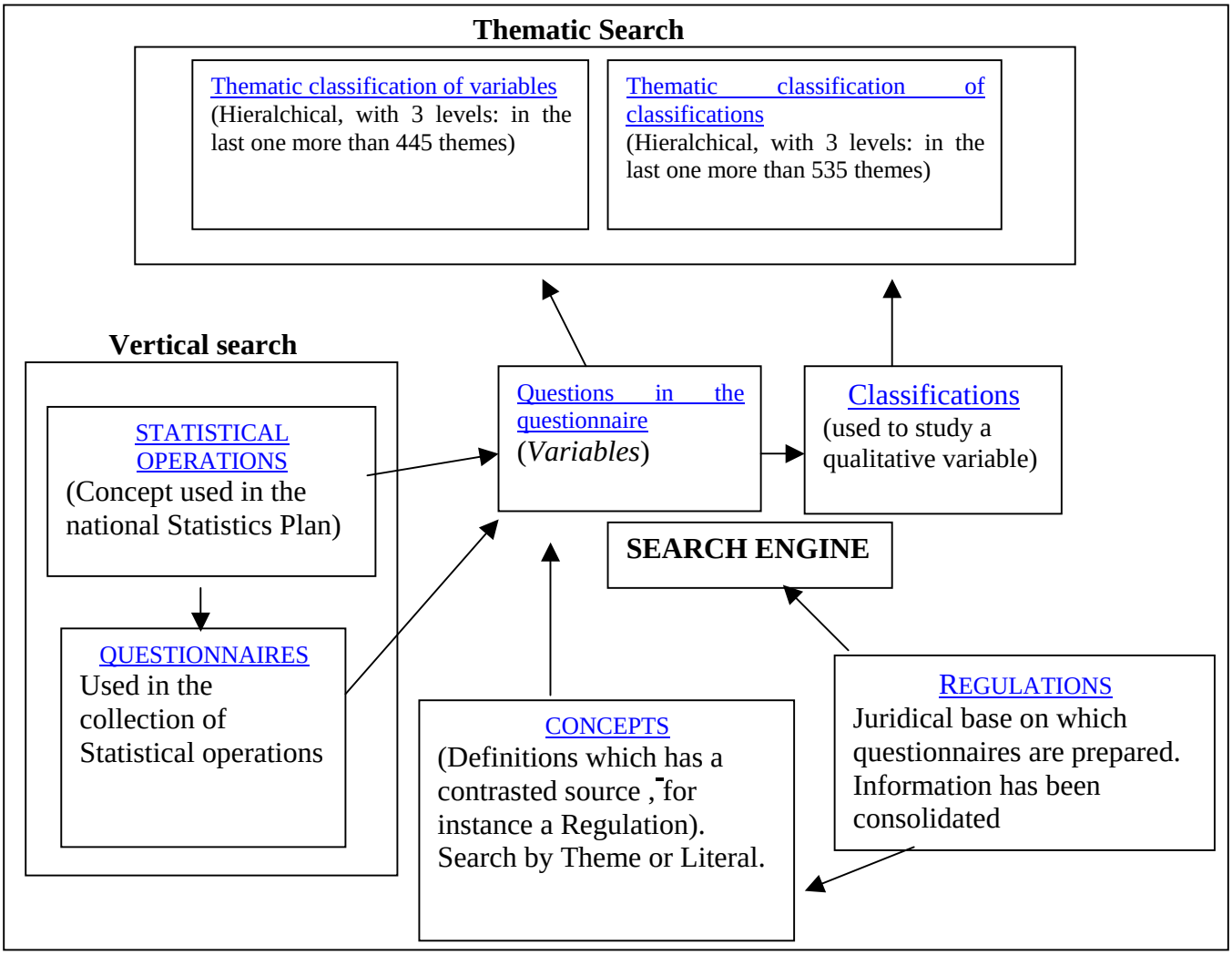


Figure 0: Metadata based on concepts, variables and Regulations

20. By means of the Enquire System, the information is organised so that its exploitation is quite complete and allows the user to have a complete picture of the whole data base and the relationships between variables, classification, Regulations…

III. ROLE OF THE METADATA INSIDE INE

21. Currently there is a growing interest in metadata inside INE, being a useful tool for all Units performing surveys and because it gives a general and complete view of the official statistics status. Nowadays, the Metadata Model is a strategy project inside INE and has much support from the Management Units.

22. Among all feasible uses we can emphasize:

a. Coordination and integration of the information
b. Rationalization of the information when it comes to elaborating questionnaires
c. Integration of data into metainformation
d. To provide a tool which allows a better analysis of the information by the users
e. To provide a coordination tool between all the statistical System (INE, Ministries, and Government Units)
f. Use in International projects

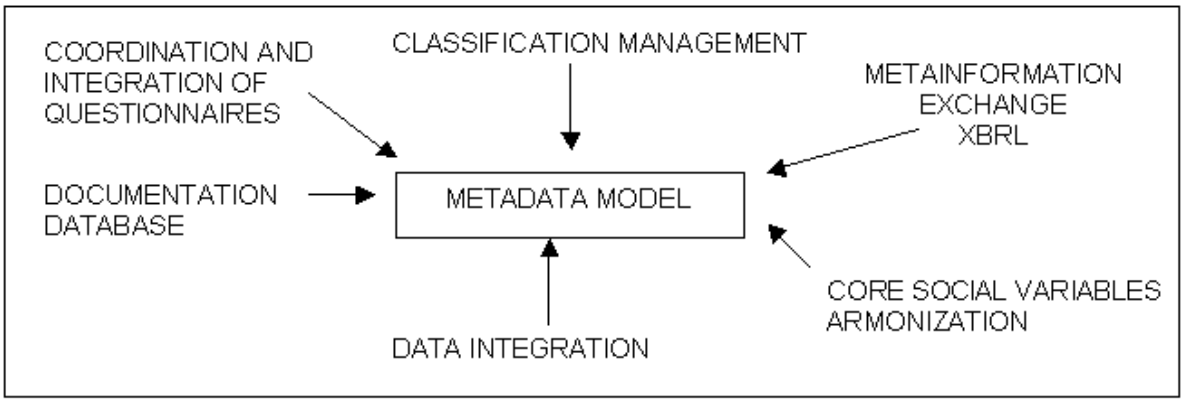


Figure 1: Metadata into INE

23. Our future works are aimed at two different goals:

(1) On one hand we want to include not also Statistical Operations under responsibility of INE, but also those carried out by other Ministries and Government units, as well as the public registers.

(2) On the other hand, our intention is to link the metadata to the data (micro and macro). We are facing a main technical problem: the great amount of data we need to manage. We are talking about annual statistical operations, but also of bi-annual and monthly ones. Subsequently, the problem is quantitative rather than qualitative. A possible solution is to keep the data under control of the responsible Unit, but connecting them with the variable in the metadata model using a unique identifier.

IV. ANNEX

24. Some of the interface developed for the management and exploitation of the information is included hereafter as an example.

- **Management System main window:**

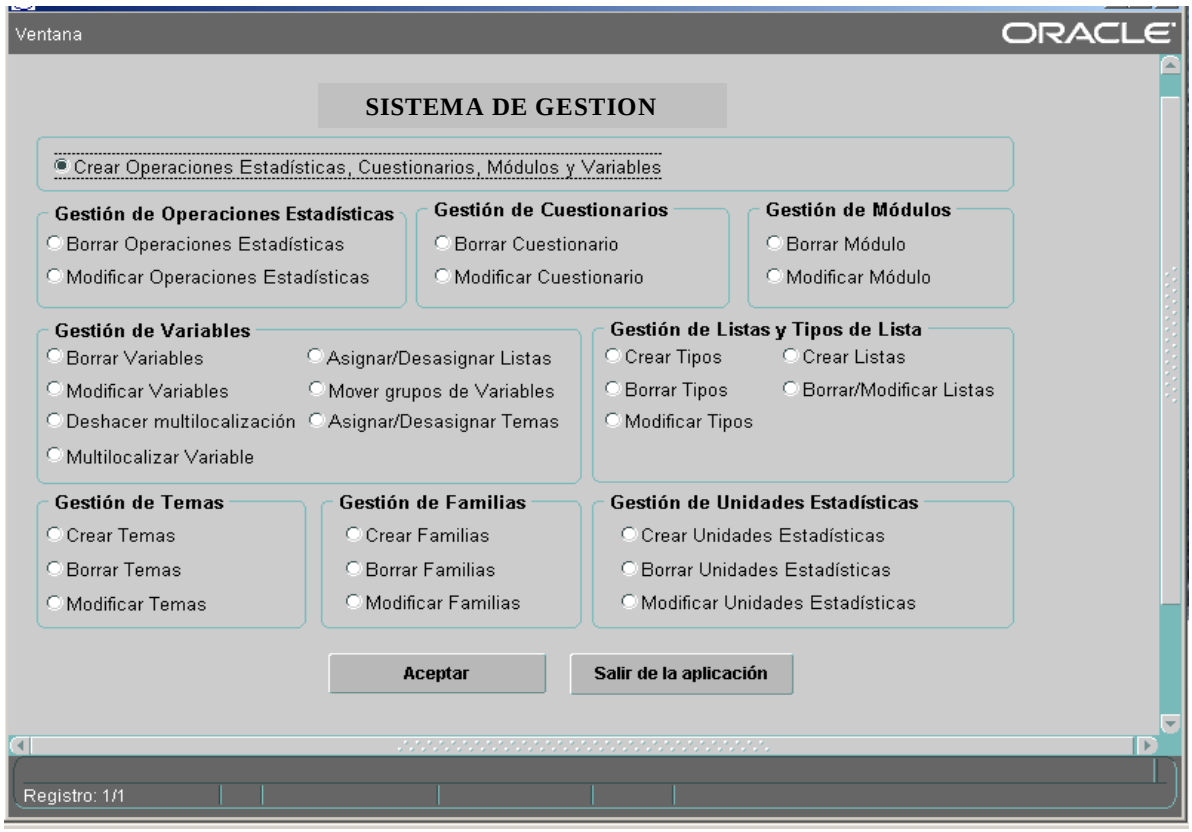


Figure 2: Management System Main window

- **Enquire System:**

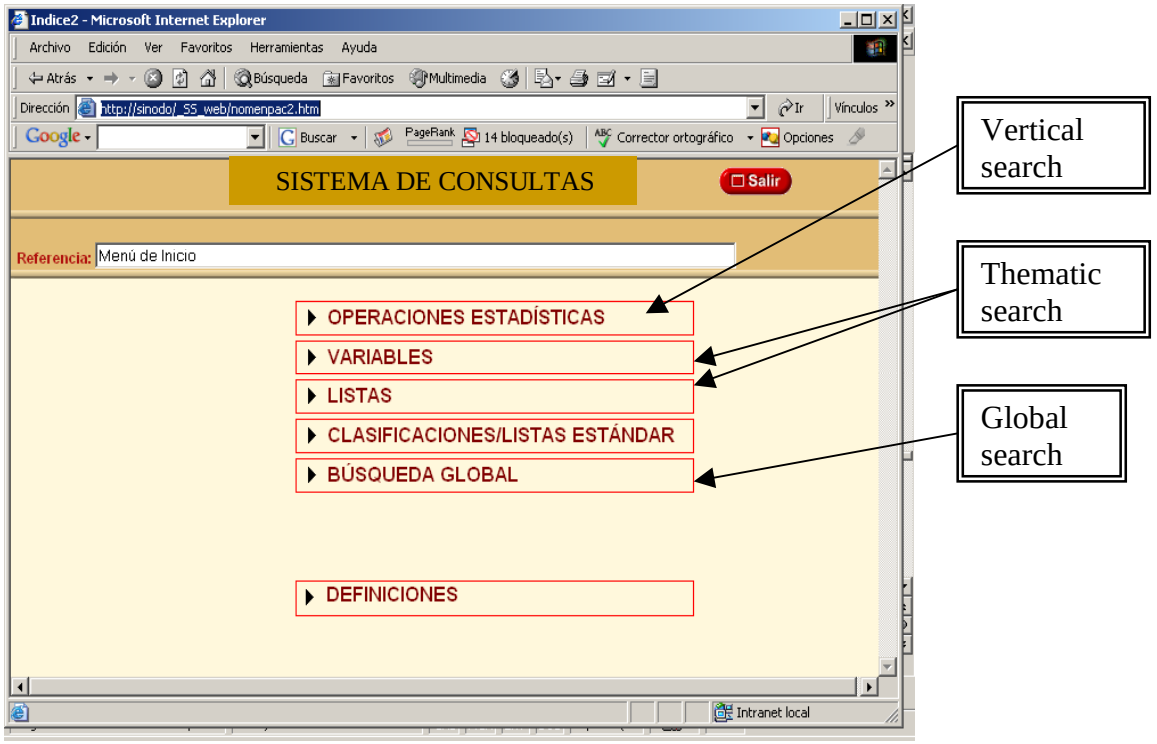


Figure 3: One of the Enquire System windows

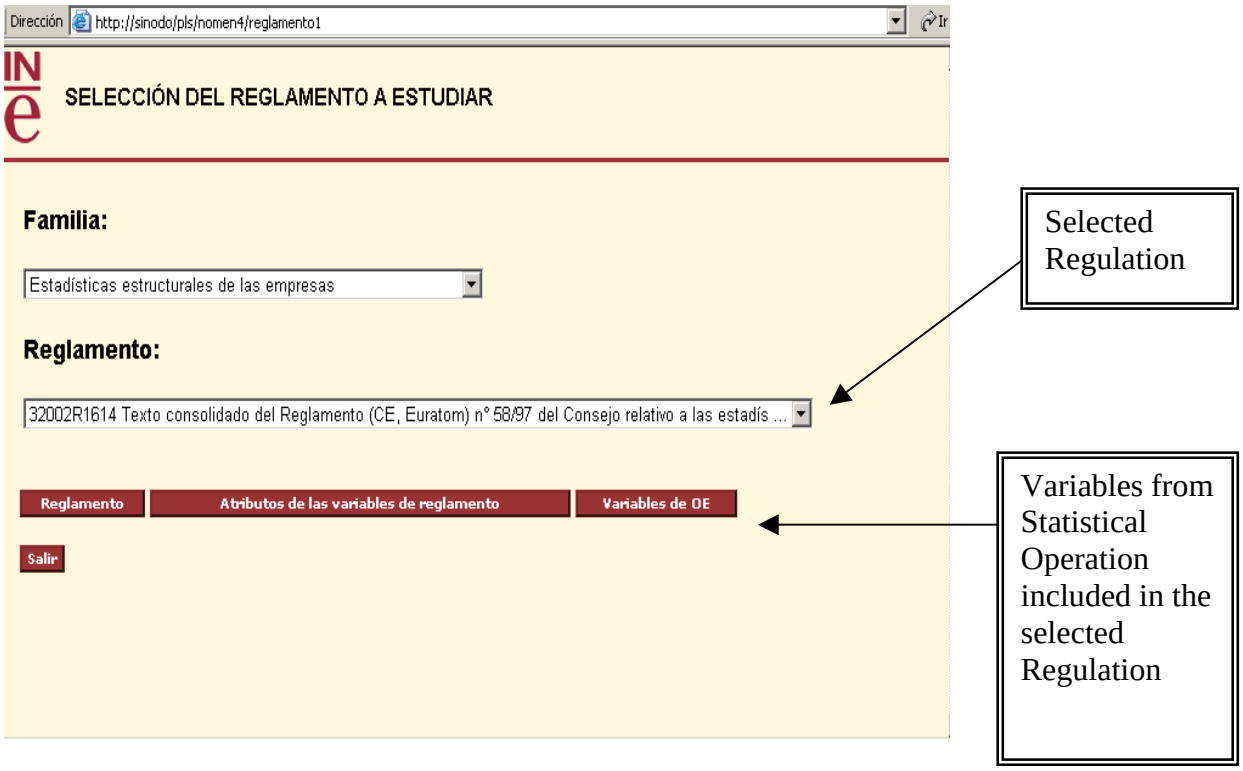


Figure 4: Performing searches in Regulations