*European Conference on Quality and Methodology in Official Statistics (Q2008), 8-11, July, 2008, Rome - Italy*

# Metadata Life Cycle – Statistics Portugal

*Isabel Morgado*
*Methodology and Information Systems Department*, *Statistics Portugal*
*e-mail: isabel.morgado@ine.pt*

## 1. Introduction

A major coordination tool to Statistics Portugal is the integrated statistical metadata system. It supports both statistical production and dissemination of statistical data, and is built upon four main subsystems: Concepts, Classifications, Variables and Statistical Sources, with tightly coupled interrelationships.
The Metadata Unit, responsible for the central coordination of this system, has conceived and implemented it in accordance with survey managers, and deals with harmonization issues.
Statistical Council, which assures the coordination of the National Statistical System, approve concepts, classifications and other technical coordination tools. This paper aims to describe the workflow established for these approvals and the interrelationship between the survey life cycle and the metadata life cycle.

## 2. National Statistical System

The National Statistical System (NSS) consists of:

- Statistical Council (SC);
- Statistics Portugal - National Statistical Institute (SP);
- Central Bank – Bank of Portugal (BP);
- Regional Statistical Offices (Madeira, Azores);
- Entities producers of statistics by delegation of Statistics Portugal;

The Statistical Council (SC) is the state body that guides and coordinates the National Statistical System. Its mission includes, according to the Law 22/2008 of 13th May (Statistical Act):

- *"Define, every year, official surveys at a national level and those of regional interest, according to the proposals of the statistical authorities;*
- *Approve technical instruments of statistical coordination, of mandatory use in the production of official statistics, promote their dissemination and use and propose to the Government their use in Public Administration;*
- *Approve and regulate standard procedures for the registration of data collection instruments submitted by statistical authorities and other sources that can be used for statistical purposes;*
- *Formulate recommendations in the definition of methodologies, concepts and statistical nomenclatures, to be used on administrative acts, to the production of official statistics and ensure their application;"*

All the above is carried out by the “Planning, Coordination and Dissemination” Standing Section (PCDSS).
The job of Statistics Portugal, beyond the production of official statistics, is to supervise and make technical and scientific coordination of the NSS, taking into account the general guidelines laid down by the Statistical Council. It may also delegate the production of official statistics to other public departments, called delegated bodies. SP has the responsibility to conceive and manage the statistical metadata system of the NSS, having as presumption that the concepts, classifications and other technical instruments of statistical coordination have to be approved by the SC. The metadata unit coordinates all the work related to the statistical metadata system.

### 2.1 Approval of concepts, classifications and methodological documentation

In these processes exists a strong interaction between SP and the Statistical Council. SP gathers all the information and prepares the documentation that is submitted to the SC for approval. The SP centralises the statistical concepts used in its own and the delegated bodies’ statistical surveys in a database. These concepts are classified by subject areas and are loaded into the database with the status of “proposed concept”, when they are used for the first time. Groups of new concepts or changes to approved concepts are sent to the SC periodically for analysis and new approval. The SC has working groups by subject area to analyse them and recommend their approval to the PCDSS. After the approval, their status in the database is changed to “SC-approved concept” and, is of mandatory use whenever applicable.
The classifications used in all statistical activity, such as the Portuguese Classification of Economic Activities, National Classification of Occupations, National Classification of Goods and Services, Administrative Division Code and List of Countries are also approved by the SC for mandatory use in the NSS.
In 2005, the SP submitted to appreciation to the PCDSS a standard format for the methodological document for the NSS’s statistical surveys because it was considered to be a coordination instrument. The format was approved and adopted as mandatory in the NSS.
By December 2007, 75% of the surveys in the NSS were documented accordingly to this format.

## 2. Statistical Metadata System

The Statistical Metadata System is an integrated system composed by several subsystems: Concepts, Statistical Classifications, Surveys (including the components: Methodological Documents, Data Collection Instruments, and in future Administrative Sources and Questions) and Variables.

The main purposes of this system are:

- To support the whole *life cycle* of surveys;
- To act as a *central repository* for statistical metadata serving as a source for other databases that support: design, production, dissemination of statistics and management;

- To establish *terminology* for statistical metadata;
- To be an *instrument for statistical harmonisation and coordination* of the NSS, standardising the documentation of surveys, among other elements;
- To implement a *homogeneous environment* for its technological infrastructure.

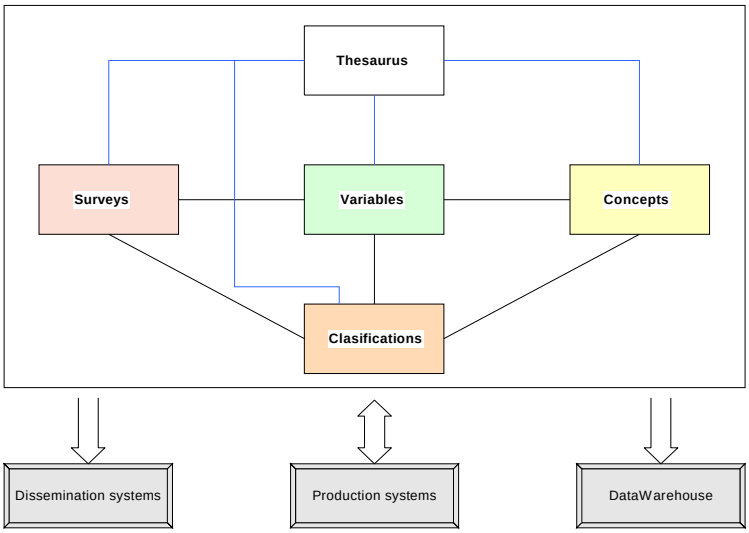


Fig. 1. Macro architecture of the Integrated Metadata System

## 2.1 Concepts subsystem

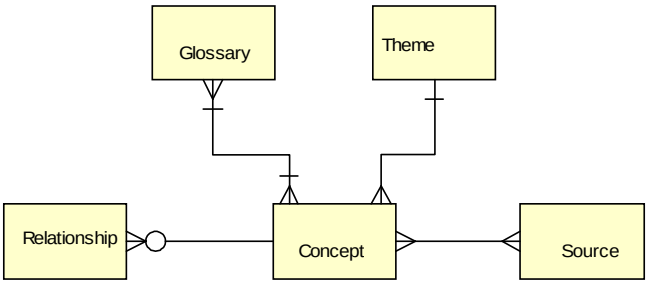


Fig. 2. Conceptual model of the Concepts Subsystem

A concept is a unit of knowledge created by a unique combination of characteristics (ISO 1087-1:2000, *Terminology work -- Vocabulary -- Part 1: Theory and application*).
The concepts and definitions recorded in the database are classified by subject area and organised in glossaries. Each glossary corresponds to a theme in the Official Statistics Portal. The main attributes of the concepts are: code, name, definition, notes on the definition and source. Other attributes are required for the management of the system, such as status (proposed, in use, SC-approved), dates on which it was proposed, came into use and was approved by the SC. It is possible to establish a relationship between two concepts, from which synonymy and homonymy have already been implemented.
There is a generic glossary of concepts used throughout statistical activity entitled “Metadata Terminology” and a list of abbreviations and acronyms used in the documentation of surveys.

There is a plan to enlarge the scope of the system so that other types of relationships can be implemented enabling to view the concepts of a particular area as a conceptual system. Due to the integration of the different subsystems, the detail page of each concept shows its use in methodological documents, classifications and variables.

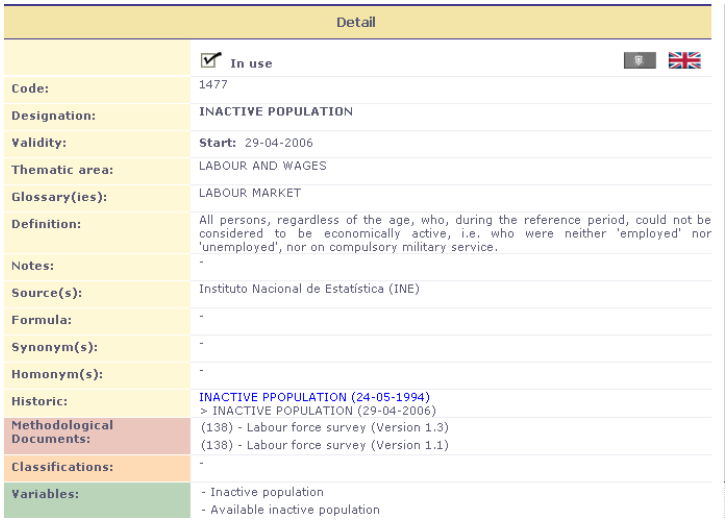


Fig. 3. Detail of a concept

The concepts are available on the Official Statistics Portal, with access from the home page, and are searchable by alphabetical order in each glossary.

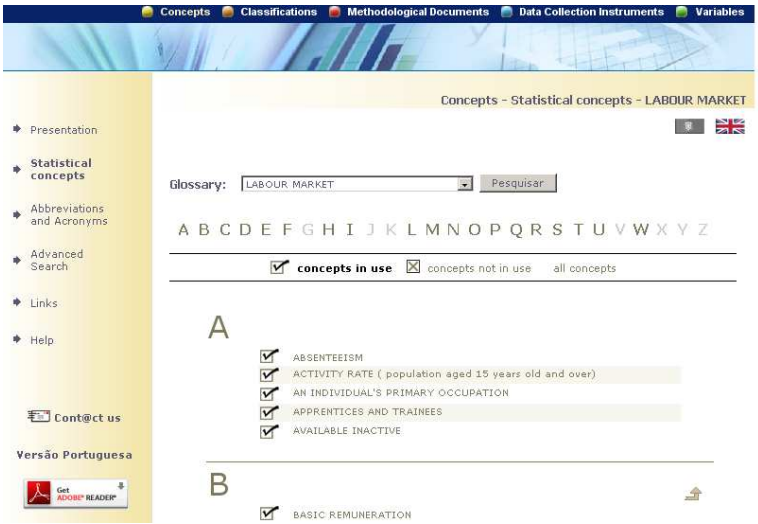


Fig. 4. List of the concepts of a glossary

An advanced search was implemented with the possibility of the combination of more than one search criteria.
It is in course the translation to English, of the concepts registered in the database. 49% of the concepts are, at the present, available in English.

## 2.2 Classifications subsystem

The conceptual model of the classifications subsystem was developed on the basis of the Neuchâtel model, a simplified version of which is shown in Figure x.

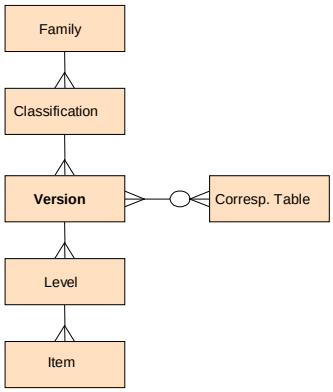


Fig. 5. Conceptual model of the Classifications Subsystem

Essentially, it provides access to three different types of information:

- National and international classifications and their description;
- Code lists that are value domains of variables;
- Correspondence tables.

Related classifications are grouped in families, and a classification can have more than one version.


Fig. 6. List of Families of classifications

According to the Neuchâtel terminology a classification version is a structured list of discrete, exhaustive and mutually exclusive categories defined by codes and designations intended to typify all units of a certain population in relation to a defined

property. A classification version has a certain normative status and is valid for a given period of time.
This subsystem allows:

- To consult and export classification versions, respective correspondence tables and indexes, when they exist;


Fig. 7. The hierarchical structure of a classification

- To consult a set of normalised attributes that characterise a specific version of each classification;
- To consult other specific and relevant attributes in specific classification versions;
- To consult documentation related with each classification version;
- To consult variants of a classification version;
- To consult, by date, “floating” classification versions.


Fig. 8. Characterization of a classification version

The classifications are accessible through the home page of the Official Statistics Portal.

## 2.3 Variables subsystem

The conceptual model is based on international standard ISO/IEC 11179, “Information Technology – Specification and Standardization of Data Elements”.

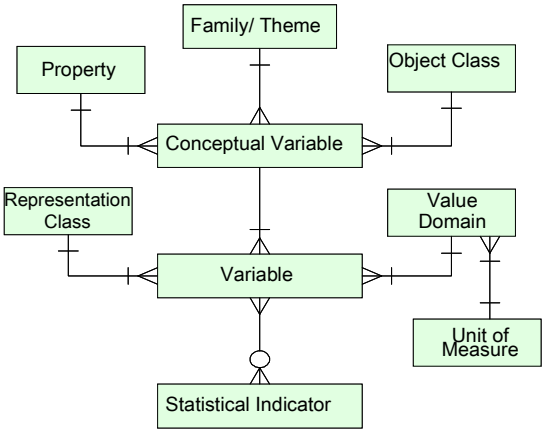


Fig. 9. Conceptual model of the Variables Subsystem

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 support the dissemination of statistical data;
- To assist the variables harmonization work;
- To improve statistical coordination;

Variables are classified in themes. Choosing a theme, a list of alphabetic ordered conceptual variables is provided.


Fig. 10. List of conceptual variables of a theme

Clicking on a conceptual variable, we access to it's details and to the list of the variables that depend on it and clicking again on the name of the variable we access to it's details.

| Conceptual variable | In use |
|---|---|
| Code: | 2915 |
| Name: | **Average stay (No.) in hotel establishments** |
| Short name: | Average stay (No.) |
| Validity: | **Start:** 11-04-2007 **End:** - |
| Status: | Valid |
| Representation Class: | Ratio |
| Value Domain: | |
| - Type | Non-enumerated |
| - Range | (0, ∞) |
| - Greatness | - |
| - Unity | Number (No.) |
| Definition: | AVERAGE STAY IN THE ESTABLISHMENT<br>Ratio of the number of nights spent to the number of guests that gave rise to these nights spent. |
| Formula: | Number of nights spent/ Number of guests that originated those nights |
| Acronym: | - |
| Context: | - |
| Registration Authority: | INE - Instituto Nacional de Estatística |
| Submitting organization: | DME - Departamento de Metodologia Estatística |
| Use: | **Surveys** **Type**<br>- (305) Guests stays and aother data on hotel activity survey (version Aggregate 2) |
| Obligation: | Obligatory |

Fig. 11. Detail of a variable

A statistical indicator is a data element that represents statistical data for a specified time, place, and other characteristics. It is composed by several variables with different roles: a variable measure and several dimensions. Time and geography are mandatory dimensions.
At present, all the statistical indicators disseminated on the Official Statistics Portal, are registered in this subsystem, with complete metadata in Portuguese and English.

## 2.4 Data collection instruments subsystem

Data collection instruments are the means of transporting information from source to destination. The data collection instruments subsystem stores and publishes in user interface, all the questionnaires (files still in preparation) that represent an instrument of reference on data used in NSS surveys. Images of questionnaires are available too, as well as some of its characteristics such as frequency and the observation variables.
There are two main types of statistical data collection instruments:

- Questionnaires;
- Files.

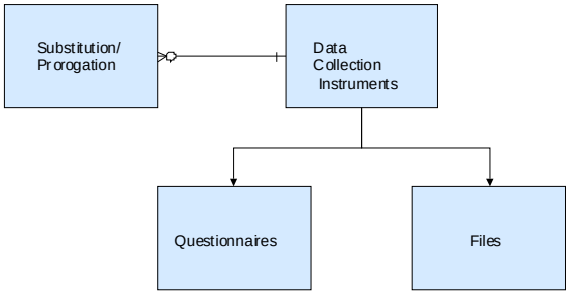


Fig. 12. Conceptual model of the Data Collection Instruments Subsystem

This subsystem makes it possible to:

- Consult and manage questionnaires and files;
- Consult and manage the history of different collection instruments;
- View their images and layouts;
- Find out how they are used in methodological documents;
- Find out what variables they observe.

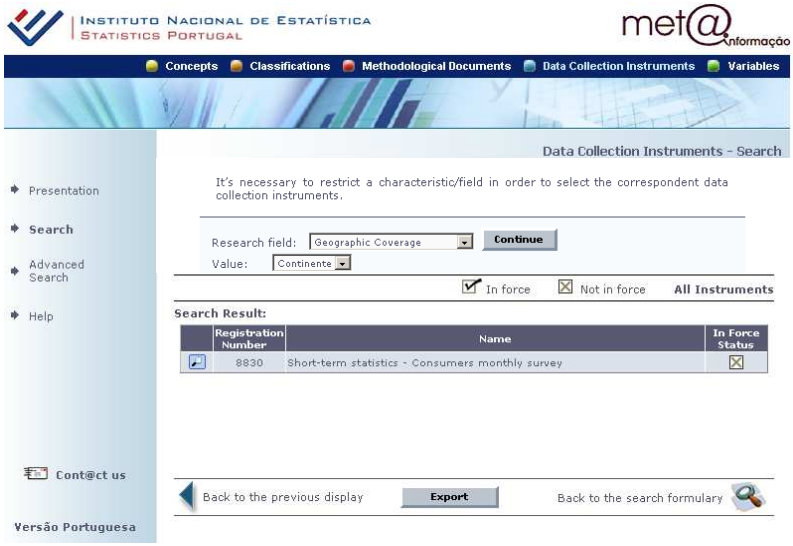


Fig. 13. Filter to search Data Collection instruments

Registration of a data collection instrument is the final step of the technical certification process of a survey and guarantees the overall quality of the survey's object. Data collection instruments are given a *registration number* and *period of validity*, whenever:

- It is a collection instrument in a new survey;
- There have been changes to the *content* of a collection instrument in a routinized survey resulting from:
    - Inclusion or exclusion of variables;
    - Changes to questions;
    - Change on the name of survey.

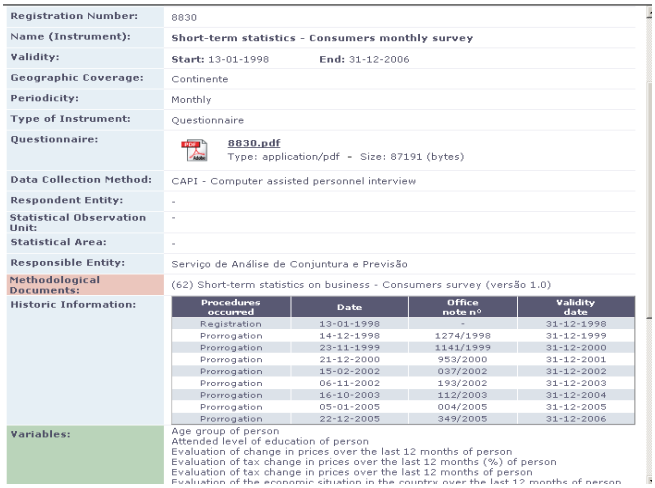


Fig. 14. Detail of a Data Collection instrument

## 2.5 Methodological document subsystem

This is the core subsystem in statistical production and the one that interacts most directly with the life cycle of surveys: in the *design phase*, survey managers define the methodologies, concepts and classifications to be used, questionnaires and their connection to the list of observation variables and definition of data for dissemination. The methodological documents of surveys have a standard format in order to facilitate and increase their usability.

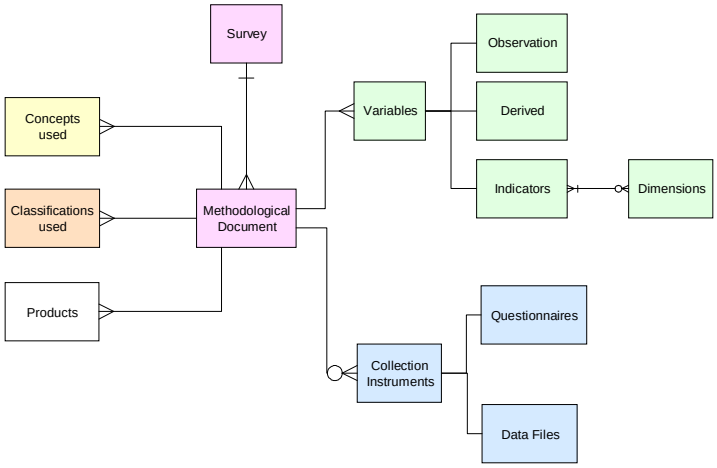


Fig. 15. Conceptual model of the Methodological Document Subsystem

This standard format was approved by the Statistical Council to support all the documentation of all the surveys in the NSS and it is composed of 8 chapters:

*I – General characterization*
*II – Methodological characterization*
*III – Concepts*
*IV – Classifications*
*V – Variables*
*VI – Data collection instruments*
*VII – Abbreviations and acronyms*
*VIII – Bibliography*

In this context a s*urvey* is a statistical activity belonging to a predefined statistical method and involving the collection, processing, refinement, analysis, study and dissemination of data on the characteristics of a population. Four basic types of surveys are considered: sample survey, census, analytical study and statistical study.


Fig. 16.Filter to search methodological documents

The search is done applying a filter using theme, code, version or a string in the name of the survey. Clicking on the name, a PDF version of the document is shown. This documentation is available in Portuguese only.

## 3. Business Process Model

The life cycle of primary statistical operations is the subject of the Statistical Production Procedures Handbook:

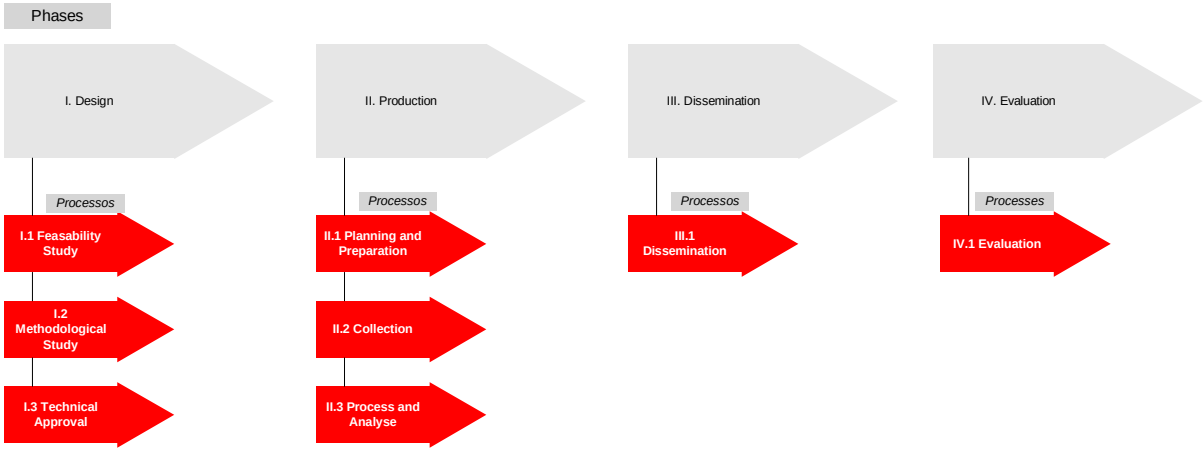


Fig. 17. Phases and processes of the life cycle of a survey

The four main phases are decomposed into processes, sub-processes and tasks. Each task has as a responsible entity.
The *Design* phase is the first phase of a survey, including a feasibility study, defines its conceptual, methodological and technical characteristics, as well as the pertaining documentation. This phase ends with a technical approval of the survey. In this phase we emphasize the Technical Approval procedure where the methodological document, containing all the metadata about the survey, is approved by the Board.
The *Production* phase is the operational phase composed by statistical processes of data collection, treatment and data analysis.
The *Dissemination* phase is where statistical products are produced and made available.
The *Evaluation* phase identifies strong and weak points of the survey and consequently, improvement actions to be taken.

### 2.6 Technical approval of surveys

The process for technical approval of surveys, which it is implemented at the SP's level without intervention of the SC, until now, is closely linked to their life cycle and consists of the following stages:

- The preliminary methodological document and the questionnaire(s) produced in the methodological study are sent to the units directly involved or users of the results of surveys, the Planning Unit, the Data Collection Department and the Methodology and Information Systems Department to get their judgement.
- At this point in the workflow, the Metadata Unit analyses the appropriate use of concepts and classifications approved by the SC, ensures correct application of

the standard format in the methodological document also approved by the SC, analyses questionnaires, introduces new concepts into the concept base and issues an opinion on the basis of its analysis.

- The department responsible for the survey updates the methodological document and/or the questionnaire(s) with the proposed changes or justifies its rejection to the unit that proposed them, submitting then the new version of the methodological document and questionnaire(s) for approval by the Board.
- The Metadata Unit prepares a memo, on the basis of all the judgments of the different units and respective answers, to send to the Board, proposing their approval or rejection.
- The Board then approves or rejects the survey: if it approves, the methodological document and the questionnaire(s) become final; if it rejects them, the process starts again.
- In the approved situation, the Metadata Unit records the questionnaire(s) in the data collection instruments database, giving them a registration number and publishing the methodological documentation in the Intranet and in the Official Statistics Portal.

In the new Statistical Act is envisaged that SC approves standard procedures to the technical approval of surveys applied to all the surveys in the NSS.

## 4. Metadata Life Cycle

The interaction of the Statistical Metadata System with the Business Process Model (BPM) generates a workflow of metadata called Metadata Life Cycle.
Majority of the metadata is born on the Design Phase of a survey and remain in the repository over time even though may be no longer valid. It becomes historic metadata.

### 4.1 Design phase

| Level of attachement | Design | | |
|---|---|---|---|
| | Feasibility Study | Methodological Study | Technical approval |
| Survey | Insert | Insert, Update, Retrieve | |
| Variable | | Insert, Update, Retrieve | |
| Concept | | Insert, Update, Retrieve | |
| Classification | | Insert, Update, Retrieve | |
| Statistical indicator | | Insert, Update, Retrieve | |
| Data collection instrument | | Retrieve | Insert |

Fig. 18. Interaction between metadata entities and processes in the Design phase

*Feasibility study:* The result of this process is a document describing a preliminary analysis of the survey, and contains:

- Methodological document – Chapter I - General characterization;
- Generic characterization of the methods for the survey;
- Planned Schedule;

- Human and material resources; estimates of income and costs; sources of funding;

The Board is responsible for the approval of the feasibility study. This approval is indispensable for the development of the following processes.
*Methodological study:* is the process where the methodology of the survey is defined. The result of this process is the methodological document of the survey whose content was described earlier in this document.
*Technical approval:* is a very important process where the dialogue between the technical coordinator of the survey and the other units is formalized.
The Board approves the survey, through the methodological document and the questionnaire;
At the end of this process the methodological document and the questionnaire enter into production and are made available in the Intranet and in the Web, if it is the case.

## 4.2 Production phase

In the operation phase, the data collection process is the one that gather and use more metadata.
*WebInq* is an online service available on the Official Statistics website for electronic data collection. It allows respondents to answer SP surveys in different ways: Filling in an electronic form online; Filling in XLS (Excel) files and sending them by email; Uploading XML files. For each survey whose data can be collected in this system, we have a description of some characteristics included in its methodological document and an image of the questionnaire is shown in the data collection instruments system. The information from the methodological document visible on WebInq comprises: description, objectives, legal framework, type of survey, geographical scope, date reference period, data collection period, concepts and classifications used. The surveys are identified in the system by the survey code used in the metadata system.
Other systems, in implementation yet, create and update other metadata entities, beyond those already stored in the repositories. Examples of these entities are: survey instance, sample frame, sample and stratum. Surveys are identified by the code used in the metadata system.

| Level of attachement | Production | | | |
|---|---|---|---|---|
| | Planning | Collection | | |
| | | WebInq | Frames and Samples | Data collection control |
| Survey | Insert | Retrieve | Retrieve | Retrieve |
| Variable | | | | |
| Concept | | Retrieve | | |
| Classification | | Retrieve | | |
| Statistical indicator | | | | |
| Data collection instrument | | Retrieve | | Retrieve |
| Survey instance | Insert | | | Retrieve |
| Population frame | | | Insert, Retrieve | |
| Sampling frame | | | Insert, Update; Retrieve | |

Fig. 19. Interaction between metadata entities and processes in the Production phase

### 4.3 Dissemination phase

In this phase survey managers define statistical indicators and prepare data to be available in the Dissemination Data Base (DDB); although theoretically these indicators should have been defined in the design phase, in practice is in this phase that this task is executed.
Also in this phase, survey managers prepare reports to monitor data exchanged with European and international organizations.
Quality reports about surveys must be systematized in this phase, because most of the quality components and concepts are needed to document data.

| Level of attachement | Dissemination |
|---|---|
| Survey | Retrieve |
| Variable | Insert, Retrieve |
| Concept | Insert, Retrieve |
| Classification | Insert, Retrieve |
| Statistical indicator | Insert, Retrieve |
| Data collection instrument | Retrieve |

Fig. 20. Workflow of metadata in the Dissemination phase

Once data is disseminated, metadata remains in the repositories with documentation purposes. As reality changes over time, metadata changes accordingly through new versions of the different metadata entities. So, there is a history that will be preserved. Technical constraints will determine the time frame history will be available.