Distr. GENERAL
25 April 2013
WP.17
ENGLISH ONLY

UNITED NATIONS
ECONOMIC COMMISSION FOR EUROPE
CONFERENCE OF EUROPEAN STATISTICIANS

EUROPEAN COMMISSION
STATISTICAL OFFICE OF THE
EUROPEAN UNION (EUROSTAT)

ORGANISATION FOR ECONOMIC COOPERATION
AND DEVELOPMENT (OECD)
STATISTICS DIRECTORATE

Work Session on Statistical Metadata
(Geneva, Switzerland, 6-8 May 2013)
Topic (ii): Case studies and tools

DEVELOPMENT OF METADATA IN THE NATIONAL STATISTICAL INSTITUTE (INE) OF SPAIN

Working paper

Prepared by Ana Isabel Sánchez-Luengo Murcia, National Statistical Institute of Spain

I. Introduction

1. Statistical production should be regarded as a chain of processes that are carried out in order to obtain specific information. In international language this chain of processes is described through the GSBPM[1]. At INE we have considered this model to be adequate for describing any process[2] and its implementation is being studied. All the metadata tasks described here adhere to GSBPM.

2. Metadata are essential for the production of official statistics and for the exchange and interpretation of information both between institutions and users, at least in this sense has been understood by the majority of the National Statistical Institutes and other international organisations.

3. INE is working on the elaboration of metadata that embrace the whole statistical process, which provide cover for all the statistical domains and which are re-usable, in other words, which contribute towards greater efficiency and integration of the statistical process. This Project is the so-termed Integrated Metadata System (SIM), which sets out to achieve a gradual progressive transition from a separated metadata model, which had the prime aim of solving the specific problems of each one of the statistical operations, towards an integrated metadata system enabling the different statistical subprocesses to be carried out in a consistent, flexible and standardized way, while reducing the need for tailor-made developments.

4. This document will offer an up-to-date overview of metadata in the statistical process at INE. Firstly the paper will be given over to the general considerations that have to be taken into account for the development of the project. Secondly it will deal with the projects relating to reference metadata that describe the content, methodology and quality of the statistical operations, structural metadata

[1] http://www1.unece.org/stat/platform/display/metis/The+Generic+Statistical+Business+Process+Model
[2] On 8 March 2012 the INE Board of Directors took the decision to foment the use of the GSBPM model to describe the production model of any INE statistical operation and for communication between the different units.

describing the databases, and process metadata which describe each of the individual subprocesses used in the preparation of information.

II. Metadata Project, Integrated Metadata System

A. Aim of the SIM (Integrated Metadata System)

5. The production process of the National Statistical Institutes is undergoing a review due mainly to the need to meet ever-increasing demands for information with lower costs while at the same time, of course, maintaining or improving quality. The prime aim, therefore, of that review is to achieve a more efficient production process that will fulfil the good practice requirements called for in the international statistical system.

6. The system of statistical metadata stored in institutional metadata should be the mainspring of the statistical data elaboration process in the production system.

7. In this context, the main aims of the construction of the SIM are:

   (a) Propitiate **efficiency** and **re-use** in the statistical process, and facilitate integration of matching data.
   (b) Improve the **governance of INE**, i.e. the planning, design, implementation and assessment of the process.
   (c) Facilitate communication and **exchange of information** with both national and international bodies and end users.

8. The main aspects being taken into account for the implementation of the project are as follows:

   (a) Identification and design of the main databases to be taken into consideration at the institution.
   (b) Identification of the metadata that should form part of the SIM.
   (c) Compilation of the metadata identified in the previous point.
   (d) Implementation of the project, determination of teams, schedules, etc...

B. Description of the SIM

9. Since early in the last decade INE has conducted specific metadata projects in tune with the advances made in the international groups. We have compiled repositories of concepts, questions for questionnaires and lists associated with a large proportion of INE statistical operations, taking into account such operations as the Labour Force Survey (LFS) or the Consumer Price Index (CPI), and in the last year we have been working on the development of both reference metadata and process metadata.

10. The metadata project has been divided into different subprojects that will successfully assist in the elaboration of the overall project. Each one of these subprojects has been prepared with different teams, working plans, schedules and resources, but maintaining the interaction between them all the time.

11. These subprojects are considered in accordance with their purpose. Consideration is given to three key purposes:

    (a) For the description of a **statistical operation** reference metadata are used.
    (b) For the description and handling of the **databases** the structural metadata are used
    (c) For the description of the **process** and its subprocesses the GSBPM and process metadata are used.

C. Description of the statistical operations; Reference metadata

12. The function of reference metadata is to compile information on the whole set of statistical operations conducted at INE, on a systematic and uniform basis, considering the statistical process in terms of the GSBPM and using the ESMS[3] as the standard recommended by Eurostat, thus assuring harmonisation and comparability in the European Statistical System.

13. This project is being developed subject to two key premises: interactions with other national and international projects and re-use of the metadata already available at the institution.

14. The interaction with the assignments carried out with Eurostat relating to the ESMS is not complete due to national characteristics, since the reference metadata project takes into account any statistical operation, whether of a European or a national nature. Furthermore, on some occasions the standardised methodological report, which represents the reference metadata at national level, corresponds to more than one ESMS because in the national framework these reports are associated with statistical operations while on a European scale with statistical domains. For example, for a short-term survey as the Retail Trade Indices, on a domestic level a single methodological report is disseminated while at European level four ESMS are generated, one for each indicator, Turnover, Persons employed, Hours worked and Wages and salaries.

15. Figure 1 shows an example of re-use in point 3.4, which refers to the concepts used in a statistical operation and draws information from the SIM, specifically from the concept subsystem. We can observe the relation that exists between the concepts used in the reference metadata subproject and the different SIM systems, or else, for example, points 11-19, which will form part of the quality reports and which will provide information for the ESQRS (ESS Standard for Quality Reports Structure).

16. These metadata are already in the process of implementation on a national level and are available for any statistical operation published since February 2013 in Spanish. There are already 30 short-term statistical operations available, including the Labour Force Survey (LFS), the Consumer Price Index (CPI) or the National Accounts, and some 15 structural ones, such as the National Health Survey (NHS). The project is expected to be implemented completely by December 2013. Consideration is also being given to its availability in English and final completion of convergence with the international reference metadata based on the SDMX.

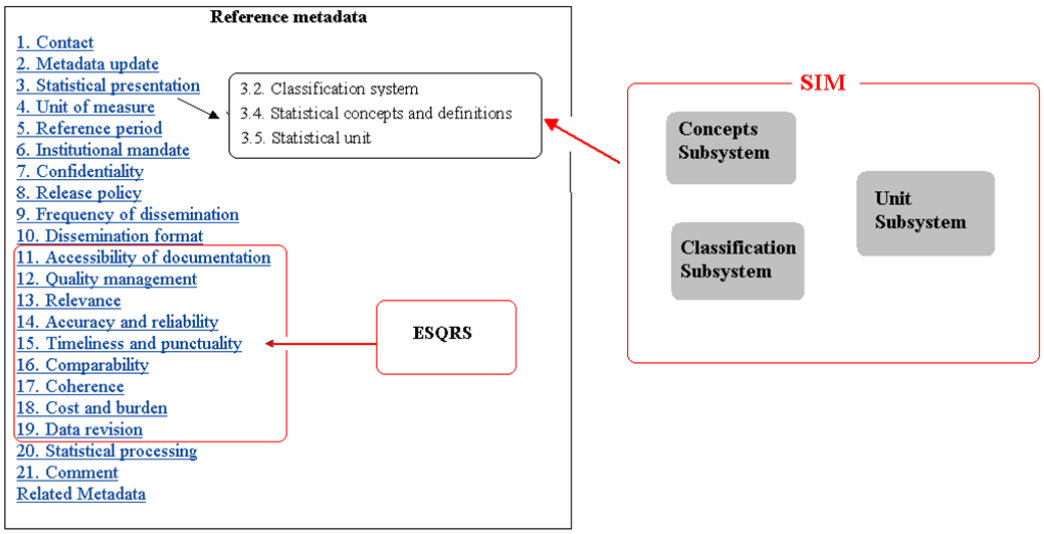

Fig. 1. Relation between reference metadata the SIM and the ESQRS

17. The most significant lessons learnt from this project are, on the one hand, the method followed in its implementation and management, the stage that we are at this time and which is being carried out on

[3] http://epp.eurostat.ec.europa.eu/portal/page/portal/statistics/metadata/metadata_structure

a joint basis by a multidisciplinary committee, with experts on methodology, quality, coordination and statistical dissemination, and the units responsible for the statistical operations, and on the other, the re-use of metadata that are already in the SIM, specifically of the concepts (point 3.4 of the standard methodological file) and in the future of the classifications too. It should not be forgotten either that for the first time at INE this project will provide users with structured, integrated and standardised information.

D. Description of the databases: structural metadata

18. In the GSBPM we may consider two highly important aspects: on the one hand, the activities or subprocesses that are being developed and, on the other, the products that are being obtained from those activities. The most valuable products for the institution are the databases, and hence the importance of the metadata associated with them.

19. In the course of the process, described always with the GSBPM, we go on obtaining different databases, which may in the end be considered as the backbone of the information. In figure 2 an example is described of how the databases stemming from a process described in the GSPM could be framed.

20. In the figure we may observe that stage 1 and 2 are especially important for the generation of information. A proper development of these two stages permits the re-use of processes in an efficient way, i.e. standardisation, for instance in point 1.5 checking the availability of data or in 2.2 design of the description of variables. The information is generated at stages 3 to 6, which is where the databases are therefore elaborated. For users the important stage is the dissemination stage 7.

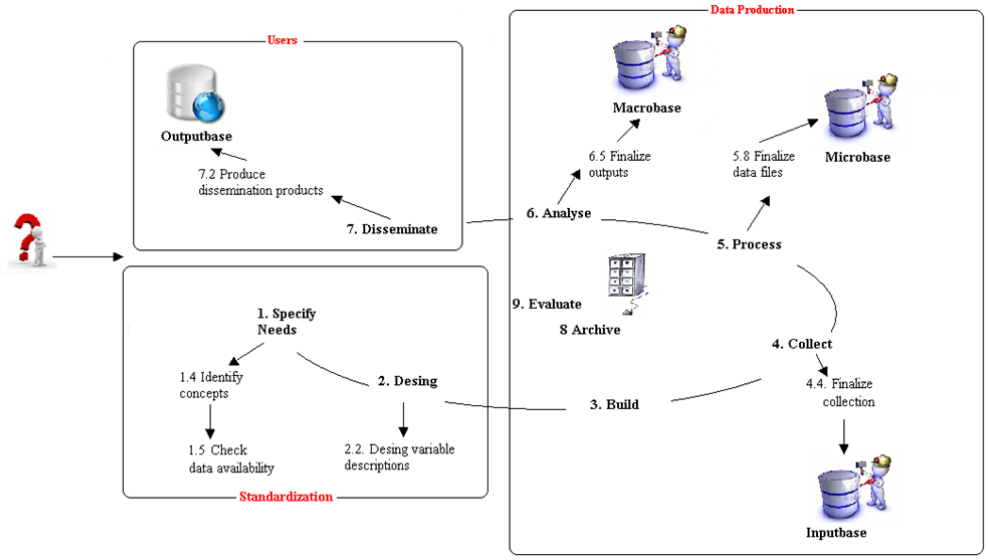

Fig.2 GSBPM databases

21. These corporate databases (inputbase, microbase, macrobase and outputbase) assemble all the information on each statistical item and the structural metadata represent a fundamental element of integration for describing any of these data. The use of structural metadata bases will make it possible, on the one hand, to replicate the production process, i.e. pass from one microdata to another easily, or what amounts to the same, from one base to another. On the other hand, it will facilitate re-use between different statistical domains both of the actual metadata and of the processes associated with them, and lastly it will permit interoperability between institutions.

22. The items that are going to be handled to describe the data will be: the variables, statistical units, their associated concepts and the classifications used. The definitions used for these items have the CMV[4] as their source.

23. Depending on stage of the GSBPM we are at or the database that we may be considering, we will have different variables and different classifications, while the associated definitions are maintained. The first two items will provide us with a smooth path in the database process, and the third one knowledge of the microdata and its comparability and coherence.

24. Figure 3 shows an example of how the structural metadata would work in a microdata file with its respective disseminated table. For example, the dissemination variable "Population of 16 years of age and over by marital status, sex and age group" is obtained from the microdata variables "Marital status", "Sex" and "Age". The classification used in the microdata file for example for the variable Sex; 1 Male, 6 Female is converted readily into the dissemination classification -Both Sexes, -Males, -Females. The same could be done for the variable “marital status” and its classification.

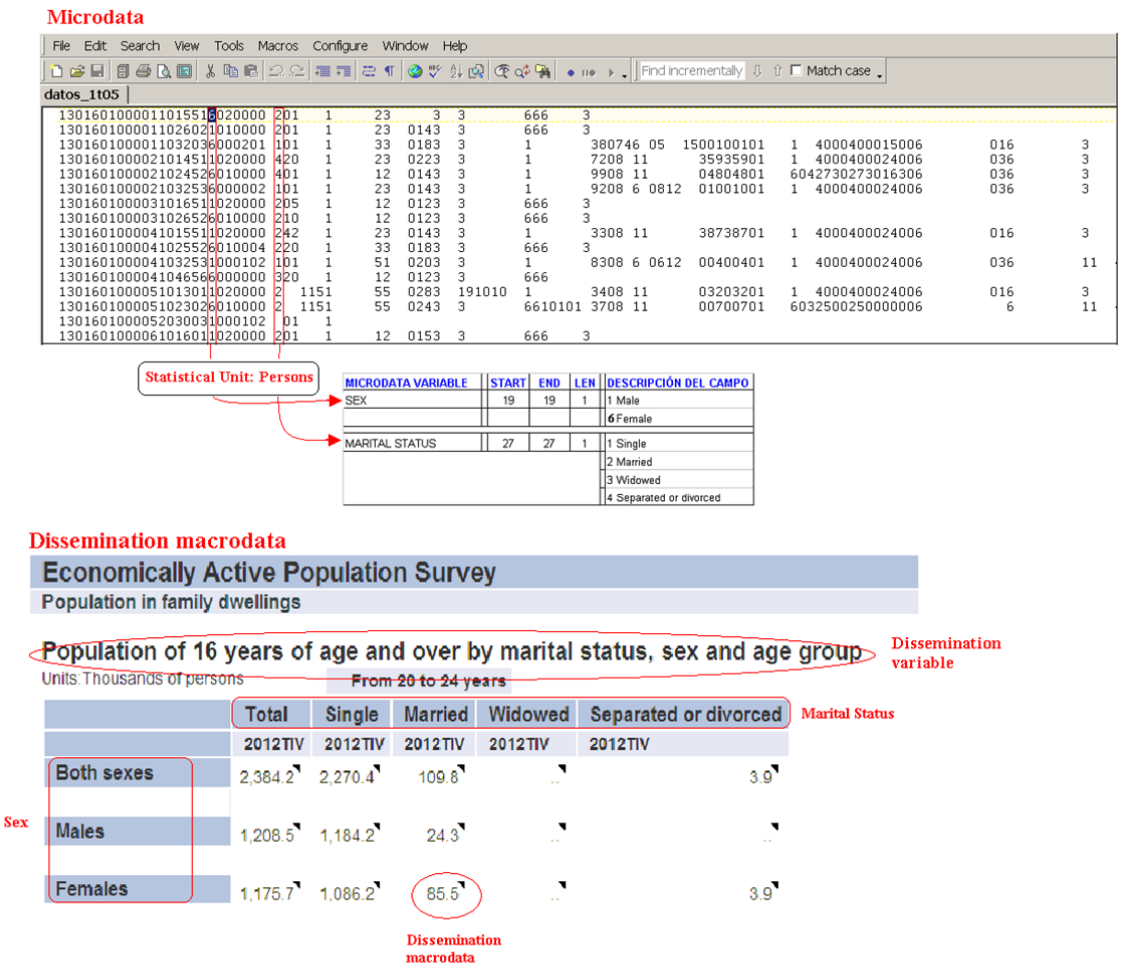

Fig. 3. Relation between microdata and macrodata

25. Having decided the use of the variable Sex and Marital Status in stage 1 of the GSBPM, in the design stage the process to be followed for obtaining microdata and macrodata would be analysed. For these

[4] http://sdmx.org/wp-content/uploads/2009/01/04_sdmx_cog_annex_4_mcv_2009.pdf

variables we could decide to use standard tools for collection, imputation, etc. we could take advantage of estimation, coding, adjustment, etc. subprocesses.

26. All this metainformation would be unmanageable and could not be re-used if it were not backed up by a good organization and search engines. The information is arranged on a topic basis and for each item there are statistical operations, variables, classifications, etc... topic classifications of their own. In addition, word search engines have been produced permitting the search for information without knowledge of the system.

27. This project has been in the process of elaboration for some years at INE, but it has not been implemented yet. There are currently repositories of classifications and questions for the majority of INE surveys in existence and work is being done on the elaboration of microdata variables.

28. The implementation stage has undergone problems mainly for two reasons: firstly because metadata had not been linked to data, but this has now been done since 2012; secondly because metadata were not present in the process. There has been a change of philosophy and implementation is now being attempted from the process.

E. Description of the GSBPM subprocesses; process metadata

29. Elaboration of process metadata is an extremely new project at INE on which work is now being started. As was pointed out at the beginning of the document, INE is anxious to foment use of the GSBPM and for this purpose a pilot study is being carried out on the use of GSBPM as a language for describing the process, which has the participation of two statistical operations, one short-term economic (Retail Trade Indices, RTI) and the other yearly, of a social nature (Survey on Equipment and Use of Information and Communication Technologies in the Households, ITC-H).

30. Short-term surveys are governed by a fixed schedule, usually monthly or quarterly. This means that the process is repeated periodically every month or three months.

31. There are some survey processes, however, as is going to be seen specifically for short-term surveys, that are not included in the normal monthly / quarterly production routine and which adhere to their own longer routine (yearly, for instance, as is the case with the calculation of the elevation factors that are applied to each company in the sample in the course of the year for the RTI). We may even pinpoint some non-periodic processes, for instance, a change of classification.

32. This leads us to differentiate three types of processes related to a set of short-term statistics, each of them organized in accordance with its own line:

    (a) Monthly processes: collection of monthly / quarterly data, includes GSBPM stages 4, 5, 6 and 7.
    (b) Occasional processes, for instance, changes in the questionnaire. Includes stages 1, 2, 3 and 4.
    (c) Yearly processes: For example, with the panel survey rotation: partial yearly change in the sample Let's assume that 25% of the sample is renewed every year. Before the new simple units start to take part in the survey like any other unit, there is a training period so as to make certain that the company has understood the questionnaire and thereby prevent unwanted distortions in the indices.

33. It is important to bear in mind not only that the process is not linear, but also that we have to consider certain subprocesses that are not separate and which converge on the main process at a given time.

34. Figure 5 shows the path of the RTI production process within the GSBPM for monthly tasks, which have been designed with Mi. The process metadata appearing in figure 6 may be derived from this production process.

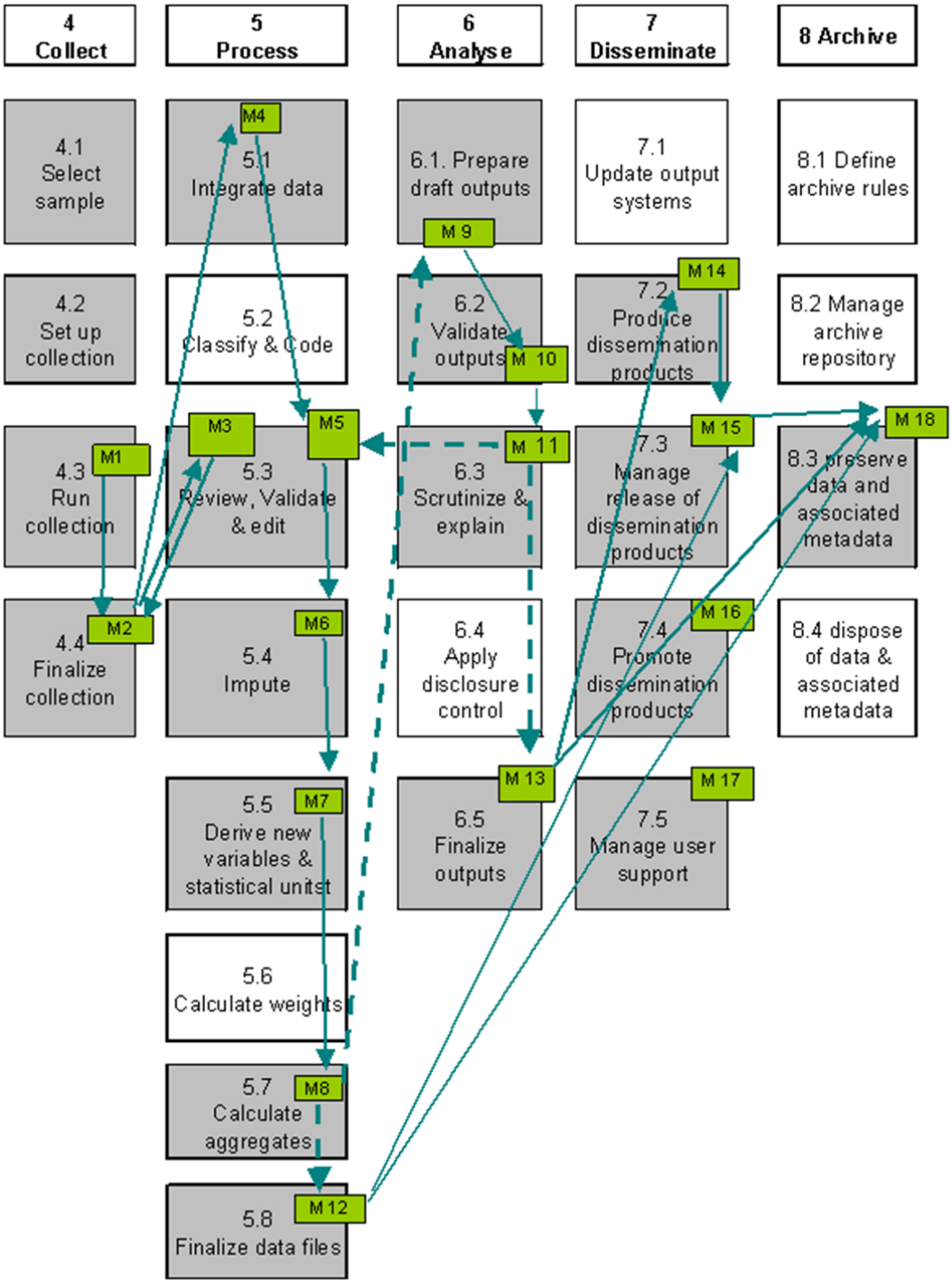

Fig. 5. Production process of monthly tasks in the RTI

Process metadata	**RTI**
Action	M7
Subprocess	5.5 Derive new variables & statistical units
Actions	Practically all the variables that are used for calculating the indices are obtained directly from the questionnaire, there are some, however, that are not obtained directly but which are calculated from the questionnaire information.
Action Periodicity (e.g. monthly, yearly,...)	Monthly
Starting date	t+19
Final date	t+25
Initial file	Validated and edited microdata file
Final file	Microdata file with the variables derived
Software (standard and/or tailor-made)	Internal or external development and type of language on which produced
Documentation, manual, handbook	Rules for validation, editing, sample selection ...
Unit in charge	Survey unit
Collaborating units	IT unit

Fig. 6. Process metadata associated with subprocess M7

35. The analysis carried out in figure 5 on the one hand allows the units to get to know their own processes and the latter to be properly documented. Besides each subprocess, we may obtain information on the associated process metadata, for example, execution schedule, units involved, which will permit appropriate planning, software used and manuals that will permit either re-use in other domains or improvement, or initial and final files which will enable us to identify the microdata or macrodata bases and manage them properly.

36. Furthermore, for a specific subprocess we may obtain additional information which could be linked to the reference metadata and the data. For example, currently INE is working in this direction and by way of a methodological group seasonal adjustment models have been produced for short-term data; these models will go on to become standard. After performing a statistical operation, we could obtain the following metadata associated with seasonal adjustment:

    - Series published
    - Method used
    - Software used
    - Aggregation
    - Reviews
    - Quality indicators

37. These metadata should be compiled and disseminated for the published series adjusted for seasonal effects and scheduling effects or at least the most significant aggregates.

III. Lessons learned

38. The assignments carried out over these years have progressed enormously thanks to the lessons learned. Projects of this kind should be considered as strategic in the institution as their outcome will lead to an improvement in the production processes and increase in quality. They are systems that permit re-use and enhance the efficiency of the institution, even though we have to take care with the cost both of development and of maintenance and strike the balance between cost and benefit. Metadata systems have to be linked to data within the statistical production system. In this way, we will be able to better identify the standards, as to promote and spread their use, raise the consistency of the data between statistical domains, and gain a better communication both with users and with other institutions.

IV. Future

39. INE will go on working on the elaboration of the SIM, using it as a tool for guiding the production process so as to describe not only the databases but also the actual process itself. As results of these works, the metadata will become the conducting wire that will get through the entire statistical process. The SIM will also be used as a tool for standardisation, thereby raising the efficiency of the institution and assuring and enhancing the quality of the information.

40. The use of administrative sources has a long tradition in our institution and we expect, in the near future, to address the development of metadata relating to the administrative registers used in the INE.

V. Bibliography

- "The Common Metadata Framework " UNECE, 2012
http://www1.unece.org/stat/platform/display/metis/The+Common+Metadata+Framework
- "Generic Statistical Business Process Model (GSBPM), Version 4.0", UNECE, April 2009
- "Generic Statistical Information Model (GSIM), Version 0.4" Alistair Hamilton, Steven Vale, 12 Jun, 2012
http://www1.unece.org/stat/platform/display/metis/GSIM+Version+0.4
- Commission recommendation of 23 June 2009 on reference metadata for the European Statistical System (2009/498/EC)
- "Harmonised structural metadata (code lists) progress report", Eurostat, 2012
http://ec.europa.eu/eurostat/ramon/miscellaneous/index.cfm?TargetUrl=DSP_GENINFO_SCL
- "Statistics Netherlands Architecture; Context of Change" Richard Bredero, Wim Dekker, René Huigen and Robbert Renssen, 2009
- "Variables Subsystem ", Portugal, Teodora Monica Isfan, 2009
- "Statistics Netherlands Architecture; Business and Information Model" René Huigen, Richard Bredero, Wim Dekker and Robbert Renssen, 2009
- "National Implementation of the GSBPM – The Swedish Experience" Mats Bergdahl, Klas Blomqvist, 2011
- "Implementation of the GSBPM in development of Integrated Business Survey Program" Statistics Canada, Tim Dunstan, Alice Born, 2011
- "Standardising & industrialising "end to end" flows of statistical metadata within the statistical production process.
- Initial practical steps at the ABS" Helen Toole, Jennifer Mitchell, Alistair Hamilton, 2011
- "METIS-wiki", UNECE, 2012
http://www1.unece.org/stat/platform/display/metis/METIS-wiki