Working Papers

05/2012

Implementing a corporate-wide metadata driven production process at INE Spain[1]

Pedro Revilla, José Luis Maldonado, José Luis Bercebal, Francisco Hernández

[1] This document has been published at the Congress European Conference on Quality 2012

Implementing a corporate-wide metadata driven production process at INE Spain

Abstract

As other national statistical institutes, INE has started the transition from the numerous stovepipe-like chains of production to more integrated production processes. The Generic Statistical Business Process Model (GSBPM) provides a framework for the development of this goal. This paper describes INE experiences developing this new model, based on a single standardized production line for all surveys, supported by metadata systems, generic and standardized tools and corporative databases.

Keywords

Process reengineering, Enterprise architecture, European Statistical System

Authors and Affiliations

Pedro Revilla

D. G. of Methodology, Quality and Information and Communications Technology, INE

José Luis Maldonado

S. G. for Information and Communications Technology, INE

José Manuel Bercebal

S. G. for Information and Communications Technology, INE

Francisco Hernández

S. G. for Standards and Training, INE

Implementing a corporate-wide metadata driven production process at INE Spain

Pedro Revilla, José Luis Maldonado, Francisco Hernández and José Manuel Bercebal
National Statistical Institute, Spain

Abstract

As other national statistical institutes, INE has started the transition from the numerous stovepipe-like chains of production to more integrated production processes. The Generic Statistical Business Process Model (GSBPM) provides a framework for the development of this goal. This paper describes INE experiences developing this new model, based on a single standardized production line for all surveys, supported by metadata systems, generic and standardized tools and corporative databases.

1. Introduction

The National Statistical Institute of Spain (INE) has started the transition from the numerous stovepipe-like chains of production to more integrated production processes in line with the principles of the Commission Communication "Re-engineering the Production Systems of European Statistics: a vision for the next decade" *(COM 404/2009).* The ideas of the High Level Group for Strategic Developments in Business Architecture in Statistics *(HLG-BAS)* are also taken into account.

The new production model is based on a single standardized production line for all surveys and is supported by metadata systems and generic and standardized tools. Each of the properties and parameters that define the production phases, will be reflected in a metadata corporate-wide system, standardized and integrated for all statistical operations, allowing its reusability whenever necessary to perform a new operation.

The Generic Statistical Business Process Model (GSBPM) and the Generic Statistical Information Model (GSIM) provide a framework for the construction of the production processes.

This new production model is difficult to perform in the short term. The main difficulty to address is the great diversity of surveys carried out by INE. Another difficulty is the conflict between the modernization and the continuous production compromises. Hence, a step-by-step approach is used in a way the *stovepipe model* would be gradually abandoned in favor of a more integrated one.

Inside the general framework of a new enterprise architecture design some IT tools are implemented before than others. The systems for the design, construction, implementation, and management of surveys using *CAWI* channel are already in production since January 2012. The other data collection channels *(CAPI, PAPI, CATI)* will be implemented progressively between January 2012 and September 2013. The other phases of the production process

following the GSBPM framework are being designing and will be implementing in the coming years.

The first step in this project is to implement a corporate-wide data collection system. The system is flexible enough to accommodate a variety of surveys. In 2011 we started the development of a parameterized tool (IRIA) that will allow the data collection of all INE surveys, whether they are households or businesses, short-term or structural surveys, and by all the established collection channels. This tool has been designed to be executed in all of its phases by the end users. Users can define, design, build and exploit data collection systems.

IRIA supports the GSBPM primarily in phases 2 (sub-processes 2, 3 and 4), 3 (sub-processes 1, 2, 4, 5 and 6) and 4 (sub-processes 2, 3 and 4). Sub-process 5.3 is partially supported since other validations and editing are performed after data collection. With regard to the possibility of producing complete prototypes for new surveys or making improvements to the same, IRIA is a good support tool in the phase 1. Lastly, IRIA has a powerful management and security module, which supports Level 0 GSBPM in Phases 2 to 4.

In the following section, the main principles and general characteristics of the new INE Production Process Architecture are introduced. In section 3, the INE corporate-wide data collection system project is presented. In section 4 the adaptation to GSBPM is discussed. The paper ends with some final remarks.

2. New INE Production Process Architecture

Traditionally in INE, the production of statistics has been based on a *stovepipe model*, where statistics of different domains have been developed independently from each other. Changes in circumstances (increasing needs of data-users, excessive respondent burden, budget cuts), put pressure to redesign the way its statistics are produced in order to improve the efficiency of statistical production processes. In particular, the stovepipe model presents two main drawbacks: the difficulty to reuse procedures that are similar from survey to survey and the difficultly of integrating data from different surveys.

In recent years, INE is making great efforts to transform the existing production model based on stovepipe processes to a more industrialized and integrated model based on internationally recognized standards (in particular the GSBPM and the GSIM).

The new architecture enables configurable, rule-based and modular ways of producing statistics, thus minimizing human intervention in the production process. In particular, it will allow optimizing time and tasks upstream in the production process, which can be used downstream in phases such as analysis or evaluation.

The integration and reuse of components is only possible if you get a parameterized tool, which allows different behaviors in different surveys and collection channels.

The parameterization will allow the IT tools to be actually an "application of applications".

The different units in charge of statistical projects will be able to decide the properties that they wish to apply to each process. Another of the basic aspects of this architecture will be the reusability of the information. It will allow to design and implement statistics in a simple way when its components are common with others, by reusing the information stored within the same corporate database.

Each of the properties and parameters that define the production phases will be reflected in a metadata corporate-wide system, standardized and integrated for all statistical operations, and allowing its reusability whenever it is necessary to perform a new operation. Such metadata corporate-wide system will include structured metadata (variables, concepts, categories, etc.), reference metadata (associated with the survey methodology), process metadata (on different process phases) and quality metadata (quantitative and qualitative descriptions of the quality of each statistical operation).

The data that this system will accumulate for each statistical unit at microdata level will come from surveys and administrative data. This repository will allow the use of the information already obtained, either for its confirmation or correction request, as support for the collection or validation of other surveys. The system feeds back to itself allowing the reusability of existing information, either to define the properties of a new survey collection, either as support for its own data collection or its validation. The new collection tool is designed to store historical information about the sampling units collaboration, as well as to update its identification information and the hierarchical relationship among units.

This production system is based on the following elements:

1. A corporate-wide data collection system (IRIA) for generating surveys and collecting data

2. A system for making use of administrative records, whose purpose is to obtain a corporate micro-database relating INE master databases to those of administrative sources

3. An information exploitation system for standardized information processing in order to support Phases 5 and 6 of the GSBPM, resulting in the generation of corporate micro and macro-databases that serve as input to data dissemination phases

4. A metadata system that drives and centralizes the entire production process

The metadata system is a set of applications to built, maintain, query and reuse the metadata. The main applications include the following.

(1) Reference metadata editor
Through this editor, each unit before publishing the survey will create the methodological sheet and from this information will be published the standardized methodology on the web. This can reuse everything that has been considered for this statistical operation in other realizations of the same, updating only the part of the information deemed necessary.

(2) Structural metadata manager
Through this manager may be recording and updating information on concepts, survey questions, variables, lists and classifications used in any phase of a statistical operation. In addition, you will have the possibility to relate this information to the structure of the microdata file of the survey, taking into account the component variables. Through this manager can be defined from a conceptual point of view a questionnaire, ie the questions collected and the associated lists. In developing this questionnaire will be offered the best practices or standards that are approved at the institution, facilitating their practical implementation. In fact, when you want to choose a variable that has an associated standard or best practices, priority will be offered to that option, having to justify the choice of which is not part of good practice or standard.

(3) Metadata database search engine
Search algorithms are developed that quickly and reliably allow access to metadata repositories. These include variables, questions, classifications, concepts, standards and best practices. It also can perform a number of predefined queries that will allow us to compare data between different surveys, or see the information associated with a particular topic.

(4) Process and quality metadata editor
This editor includes the main process characteristics and quality indicators of the process. GSIM and GSBPM standard languages are used.

When a new survey is projected a cycle similar to this one will be followed.

1. The Meta-Data Corporate-wide Data Base is the initiator of the generation of surveys. It will store metadata standard questions about the different surveys and their relationship to variables and concepts, and process metadata, allowing for reusing that information in other surveys.

2. With this information and by means of the Surveys Generator, we will assign all the necessary properties for the collection of information (collection channels, each channel properties, etc.). The questionnaires will be generated based on the information of the Meta-Data Corporate-wide Data Base, but it will be required the incorporation of the logic functionality (flows, complex validations, etc.). The generated metadata will be incorporated into the Metadata Data Base as metadata about the process.

3. The data collection will be carried out on a corporate database where original data will remain unchanged. Each micro-data of each statistical operation will be associated to the question used during the information

collection and therefore the variables and concepts used (i.e. each microdata will be associated with its metadata).

4. Once the data collection is ended, from the corporate database with original data, an editing process and the following business processes will be carried out.

The first system tackled by the INE was the corporate-wide data collection *System* (IRIA), whose first phase (system administration, survey management for the *CAWI* mechanism, questionnaire design and the performance of interviews) is in production since January 2012.

3. Corporate-wide data collection system

The aim of IRIA system is to serve as a tool allowing end users to define, design, build and exploit data collection systems for both structural and short-term surveys of households and businesses and for different collection channels.

The first phase of development is now complete and in production. It allows for administration of the system, survey management, questionnaire design and execution of interviews through *CAWI* channel.

The remaining implementation phases scheduled by the INE are:

1. Development of general modules for all collection mechanisms. This includes all common developments for collection management, and the actions of interviewers, supervisors, persons responsible for data collection, etc. Scheduled implementation date: October 2012.

2. *CATI* channel. Covers all of the specific characteristics of this collection mechanism, CTI integration, modification of the questionnaire design for actions during telephone interviews, etc. Scheduled implementation date: January 2013.

3. *CAPI* channel. Includes characteristics specific to the this collection mechanism, control and monitoring of portable devices, control and monitoring of interviewers, communications, etc. Scheduled implementation date: May 2013.

4. *PAPI* channel, self- or staff-administered. Covers management specific to this collection mechanism and integration with mass scanning. Scheduled implementation date: September 2013.

IRIA is an end-user tool with simple, user-friendly interfaces and on-line support that allows computer users without an advanced knowledge of computing to manage their survey, design questionnaires and make these available to users through the desired collection channels.

It allows the performance of all checks to ensure that information is collected with assurance.

IRIA currently consists of four applications, which will be expanded with future developments for the other collection channels.

MANAGER management of surveys.

DESIGNER design and testing of questionnaires.

ENGINE executes the interviews designed in *DESIGNER*.

PORTAL on-line collection of information.

Questions are divided into four basic types: simple, multiple, quantity (string, numeric, date...), and grid or table type. The questions can be obtained from the corporate metadata system. Questions are contained in superior hierarchical structures with specific properties. These structures are the screen, the block and the form. One of the most interesting features of *DESIGNER* is the possibility of displaying these items in the navigation menu, allowing the user to return directly to previous screens.

The screens also have a Preview feature, so the user designing the questionnaire can view the on-screen layout, application of the styles applied, etc. Questionnaires can be run in demo mode as they are designed. This mode includes an editing module for users to check the values taken by the different variables, change values, verify the correct behavior of features, flows, validation, etc.

The required steps for designing a questionnaire are:

- Add the required global items.
- Add the questions, which are obtained either from the corporate metadata system or as new questions.
- Add the logic needed to obtain the desired interview behavior: flows, validations, etc.
- Add appearance-related elements: templates, style sheets, etc.

ENGINE is responsible for executing the interview and interpreting the elements and properties assigned in *DESIGNER*. It controls the interview from start to finish, from the moment the user enters and navigates to its completion or cancellation.

ENGINE stores the information entered by the user but also saves other information of interest, often for subsequent studies of quality:

- The time spent on each screen.
- User navigation.
- The data initially and ultimately entered by the user.
- Validation errors detected for the first time and the final errors included in the questionnaire.

PORTAL has the following features:

- Access is afforded by a key pair that is unique to each interview. Users may also be registered for access to display information from one or more selected sampling units, or certificates may be used to identify them

- For enterprise surveys, where a sampling unit is selected in several surveys, registered users or those with a digital certificate can obtain information on the units, the surveys associated to each one, besides information on completed or pending interviews. Registered users can also receive e-mail alerts when the term for completing an interview begins. *PORTAL* has been designed to provide general information on the situation of sampling units in each survey, regardless of the collection mechanism used to conduct the interview.

Reuse is a key feature of the IRIA system. It covers the reuse of both components and information. Given that the IRIA system covers a range of surveys and questionnaires, reuse is essential for significant reductions in time to start gathering information.

The *MANAGER* application has an interesting feature for copying diverse items from one survey to another, including the design of the questionnaire itself. It is quick and easy to obtain a survey with all of its items configured when it is similar to an existing one. Moreover, libraries of different items are available for use, including question libraries.

Since IRIA is a unique system for all surveys, the reuse of information comes naturally to it. All surveys may use data from earlier periods of the same survey or data from other surveys or, indeed, data residing in other information repositories. This is accessed directly through a database query, web services, or by uploading additional information to the IRIA system.

Integration is another key feature of IRIA. Integration is not only possible with the other applications of the production system; it is also established as a key element: integration of the data collected by the various collection mechanisms for each survey and integration of all survey data into a corporate micro-database.

This corporate collection micro-database is a key element for integration with other subsystems. Through this micro-database, the global data from all surveys is made available to the users who require it, whether independently or through the record exploitation system. Lastly, the micro-database is the starting point for the performance of all general explotation tasks *(Phases 5 and 6 of the GSBPM),* closing the cycle of statistical production.

Regarding the technological environment IRIA tool is a full web application developed in Java 6 using the following technologies:

4. Adaptation to GSBPM

The difficulty of designing and implementing a new standardized and industrialized production model has advised the use of a stepwise strategy. The implementation follows a modular system, where different stages of production

correspond to different applications, designed and implemented at different times. For this reason, and to implement this “great puzzle” safely and efficiently, it is essential to have a standardized production model, valid for all surveys. In this way there will be no inconsistencies or gaps between different applications.

The GSBPM is a model internationally recognized and is being used successfully in the *INE*. It has recently been adopted as a standard.

As an example is described below the adaptation of IRIA to GSBPM. The adjoining figure shows the GSBPM model, indicating the sub-processes fully supported by IRIA in green, partially supported sub-processes in orange and those for which IRIA may be useful in yellow.

In addition to adaptation to the phases and sub-processes indicated below, *IRIA's "MANAGER"* module has specific tools to meet all GSBPM Level *0* administrative requirements.

PHASE 1. SUB-PROCESSES 1.1 AND 1.2.

With regard to the possibility of producing complete prototypes for new surveys or making improvements to the same, IRIA is a good support tool in the "Specify needs" phase.

PHASE 2. SUB-PROCESSES 2.2., 2.3. AND 2.4.

IRIA can provide information on national or international standards and can be integrated with other information systems or retrieve data or classifications via web services.

IRIA allows the creation of repositories of questions or groups of the latter that are common to several surveys (e.g. household composition and questions related to each household member). This ensures quick and easy access to tools for cognitive testing or for the performance of pilot collection tests.

IRIA also maintains a history of the units that have been selected at some point in a survey, allowing for the reuse of this information; this, combined with the possibility of using information from other sources, should allow the optimal design of questions and even support for decisions on sample selection.

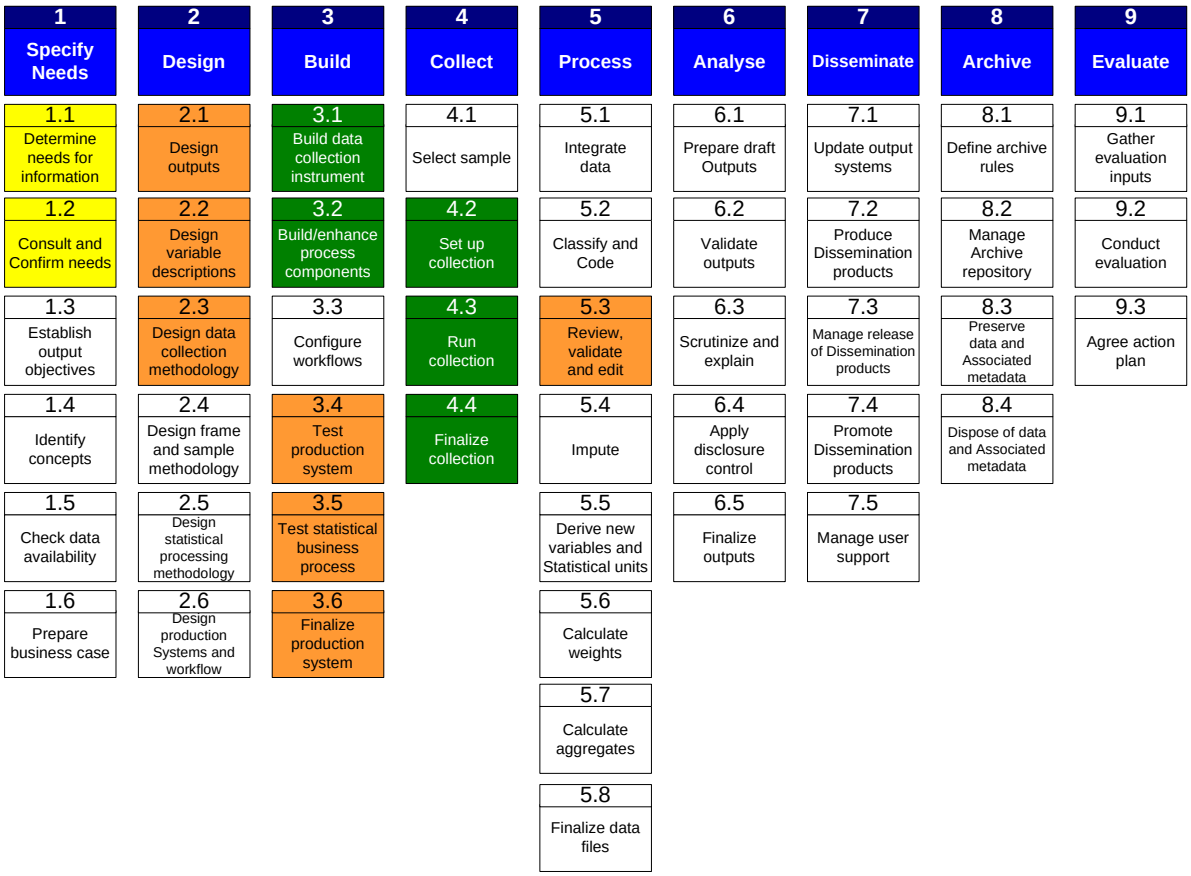
PHASE 3. SUB-PROCESSES 3.1., 3.2., 3.4., 3.5. AND 3.6

IRIA offers a straightforward and user-friendly method of constructing the collection tool.

It also allows new components to be created in order to design questions to suit the collection mechanism or features of different surveys.

Surveys are available for the performance of pilot tests to check both the tool and the structure or formulation of the questions designed.

As a unique tool, all necessary user manuals will be made available, which will be reused in the different surveys. User training will be fast and stable over time, since the same work tool is always used.



PHASE 4. SUB-PROCESSES 4.2., 4.3. AND 4.4.

Obviously, in this phase, the framing of the IRIA system is more advanced, which is why its adaptation to GSBPM is considered to be virtually complete.

With regard to sub-process 4.3, much of the analysis of the development to be undertaken over the coming months has already been completed, including the management of interviewers with reporting units and, hence, the monitoring, control and exploitation of the field work.

PHASE 5. SUB-PROCESS 5.3

In all cases, the surveys developed with IRIA include the pre-validation of responses and give the respondent the opportunity to edit these, which means that the data collected by the applications developed with this system have been validated previously.

5. Final remarks

Nowadays, public statistical offices are under continuous pressure from society, which demands more and more data, to be produced at a lower cost and with a lower respondent burden. In this context, many official statistical offices and international organizations find the current stovepipe production method unsustainable.

Changes must occur in the acquisition of data and in the production of statistical information to succeed. New IT tools and statistical methodologies offer the opportunity for re-engineering statistical production processes in a

way the stovepipe model would be abandoned in favor of a more integrated and industrialized one. Since this is a very difficult goal, and requires significant research and development, international cooperation may be essential.

6. References

[1] Commission of the European Communities (2009), Communication from the Commission to the European Parliament and the Council on the Production method of EU Statistics: A Vision for the Next Decade.

[2] Pedro Revilla, Jose Luis Maldonado and Jose Manuel Bercebal (2011), Towards a Corporate-Wide Electronic Data Collection System at the National Statistical Institute of Spain.