Applying the Generic Statistical Business Process Model (GSBPM) to the Business Register; the Spanish experience

L.E. Barbado, National Statistical Institute, Madrid, Spain.

Abstract

The Generic Statistical Business Process Model (GSBPM) is a reference framework to describe the statistical processes in a coherent way, making them comparable within and between different Organizations.

The application of the GSBPM to the management of the Spanish Business Register was carried out by the NSI during 2015. This paper provides a first assessment of the work done, focusing on the selected approach for the description of the GSBPM phases and the criteria adopted for a proper assignation of the core parts of our business process. The main restrictions found and the potential value added of this exercise are also pointed out.

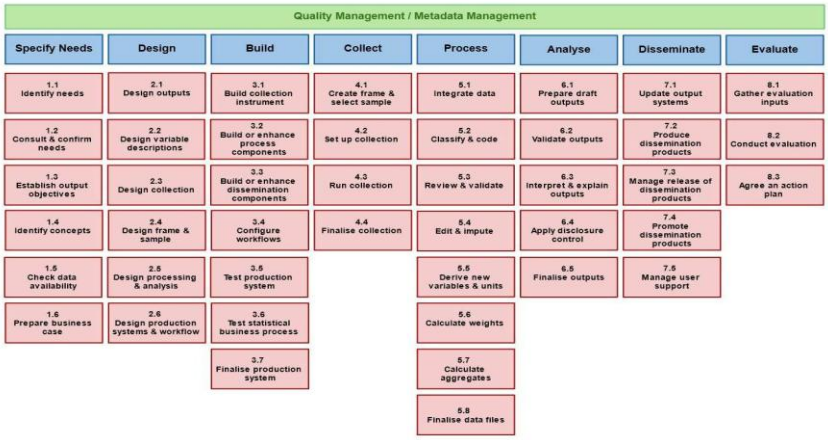
1. About the GSBPM

The Generic Statistical Business Process Model (GSBPM) is a reference framework developed by UNECE and the conference of European Statisticians Steering Group on Statistical Metadata. Its basic aim is to define and describe the statistical processes in a coherent way, making them comparable within and between different Organizations. This tool provides a standard framework and harmonised terminology to help statistical organizations to modernise their production processes as well as to share methods and components.

The GSBPM is closely connected to data quality management, providing a framework for its assessment. It comprises four levels: Level 0 (the statistical business process), Level 1 (the nine phases of the statistical business process), Level 2 (the sub-processes within each phase) and Level 3 (a description of those sub-processes). Levels 1 and 2 are illustrated in the Figure 1.

The National Statistical Institute (NSI) of Spain has adopted this standard as core element for the implementation of the Quality Assurance Framework of the European Statistical System.

Figure 1. Levels 1 and 2 of the GSBPM



2. GSBPM and Business Registers; general context

The management of National Business Registers (NBRs) for statistical purposes is a strategic action, usually incorporated within the official plans of the Statistical Offices. The key role of these infrastructural elements in data production, the increasing complexity of the related data architecture and the need for a continuous adaptation to international standards and methodologies are challenging issues undertaken by the daily work of the NBR teams.

Since 1992, the DIRCE (denomination of the Spanish Business Register) is the central reference as a sampling frame for official business surveys, which are carried out by the NSI and other Government Departments with statistical power. In the last year, more than 400,000 units were provided and investigated through more than 20 surveys.

The NSI of Spain is currently working under an explicit mandate of its Board of Directors, which is encouraging a progressive use of the GSBPM in all statistical domains identified in the national statistics plan.

Communication 404/2009, generally referred to as the Vision document, proposes several strategic principles for future statistics. Among them, the need for a re-engineering of the current production methods is particularly relevant, moving from a system based on parallel processes to a more integrated production model. In this line, Eurostat launched the 4-year initiative European System of Business Registers (ESBR, 2013-2017), with the aim to improve the relevance of these tools and reinforce their role as the backbone for the European Statistical System.

The Euro Groups Register (EGR) is the Statistical Register of the European Communities on multinational enterprise groups. The EGR is the authentic core of the ESBR system and includes information of the most influential multinationals operating in the EU and EFTA countries. It is built and maintained under a strict collaborative model involving all relevant stakeholders, mainly NSIs and Central Banks.

In the latest developments of the afore-mentioned initiative, a specific Business Architecture and its materialization through an Interoperability Frame will be available for NBRs and EGR. In this context, the application of the GSBPM to Business Registers becomes highly relevant because it will favour a mutual understanding of national procedures, the circulation of good practices and the identification of areas where efficiency can be gained. A wide application of this standard will also make future benchmarking tasks possible as well as the definition of a minimum set of interoperability requirements.

This descriptive process will also need to cover interactions with the EGR production, referring to the stage of the process where national data extractions and flows between the NBR and the EGR take part.

The experience in application of this standard to the Business Registers domain is quite new. In the scope of the ESBR project, Eurostat launched a grant with this purpose and the NSI of Spain participated in this action. A general assessment of this innovative experience is provided in the following paragraphs.

3. Evaluation of the work done

Preliminary activities focused on building capacity for using GSBPM. From 13 to 17 of October 2014, a training course was set up by national experts on standards and methodology. The DIRCE team participated in this cooperative action aiming to create basic knowledge. The relevant parts of our business process were identified and a proposal of allocation in the GSBPM structure was discussed.

Regarding the format adopted for this exercise, a combination of human and modelling language has been used. Among the possible options, the Business Process Modelling Notation version 2 was selected. The most important parts of the DIRCE business process have been graphically represented with this notation.

BPMN v2 offers possibilities to create several pools used for the representation of the donor Organizations and the actions carried out by the DIRCE Unit during the whole maintenance cycle. When an input source is received, it is subject to different kinds of processing and data quality programs. In order to provide a structured representation of all actions, three different internal areas have been considered:

- Source, where the input sources are received and evaluated
- Intermediate, where the sources are edited and transformed into statistical databases
- Final, where the integration and maintenance of the DIRCE is carried out

The Source and intermediate areas are closely related with the Collect phase. The Final area is highly relevant in the Process phase.

The main results acquired through this experience will be examined below in detail. The selected approach for the different GSBPM phases, the identification of the main register processes, their allocation to the standardized structure and the level of granularity adopted will be described. In addition, some lacks of relevance or restrictions found in this action will also be pointed out.

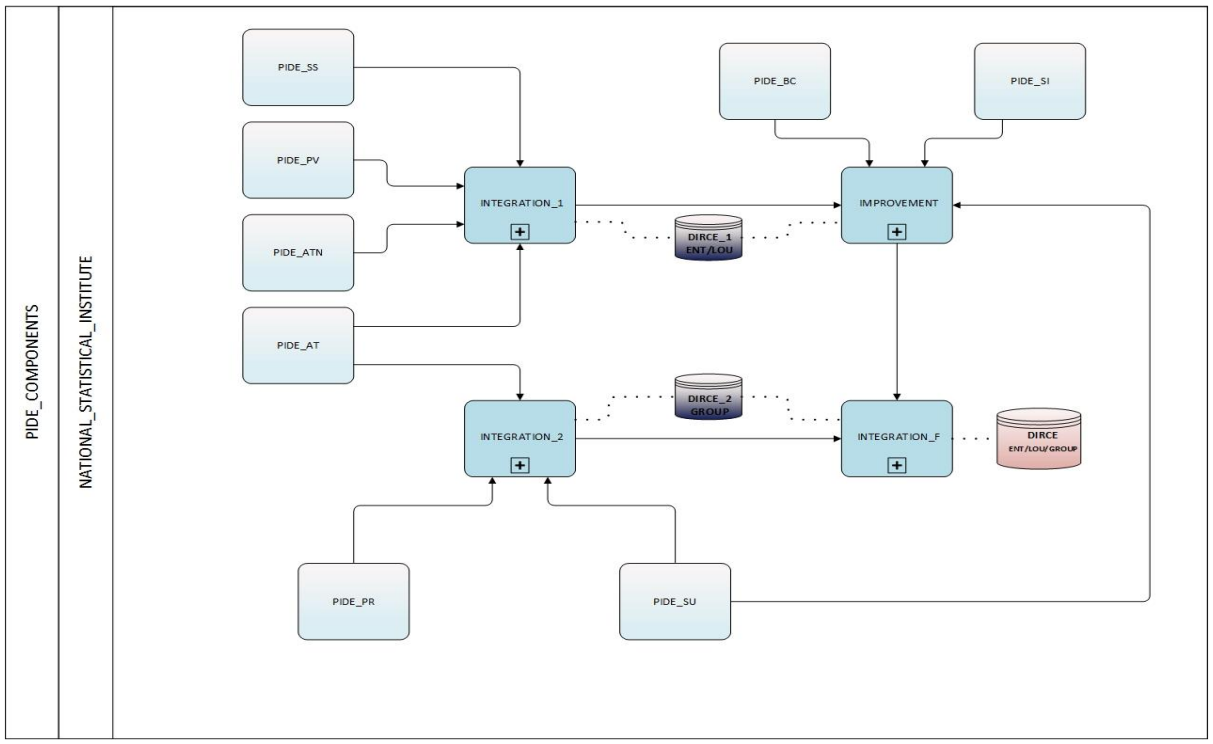
3.1 Specify needs, Design and Built phases

Management of NBRs has a long tradition in the majority of NSIs. User needs, output objectives and methods of production are continuously changing and being adapted to new emerging challenges. This context has a clear impact on the design and build of the respective data models and derived uses, which need to be continuously aligned with the new requirements.

For this reason, the approach followed for the description of these phases has been based on a historical dimension. This criteria will make a better understanding of the current state and the fundamentals of our national model easier. The methodological basis adopted for the

DIRCE management is known as the PIDE Project (*Proyecto de Integración de Directorios Económicos*, figure 2*).* This initiative started at the end of the 80's and was formulated under a modular approach, involving several components developed in successive steps and with different contributions to the maintenance of the DIRCE.

Figure 2. High level perspective of the PIDE project



The PIDE project is always considered *open*, because it is based on a continuous evaluation of the current and potential statistical needs, the development of specific actions to fulfill those needs and their definitive incorporation into the production model.

The documentation of Design and Built phases has been conceived as one unique part, focusing on the main milestones consolidated under a time-based perspective. From the beginning, both steps of the business process were jointly undertaken with a large degree of overlap. In addition, the description of the related sub-processes is not especially significant for our business case.

For the remaining phases, a static dimension has been adopted. All actions refer to the most recent cycles of maintenance for DIRCE and interactions with the EGR. More specifically, the description of the PIDE components is allocated in the Collect phase and all successive integration procedures are described in the Process phase.

3.2 Collect phase

This part is highly relevant for our business case and can be easily generalized to the big majority of Statistical Offices. It mainly involves a proper analysis of opportunity of data sources serving for the management of the NBR. It also includes all actions related to a successful and stable reception in the NSI.

The production of the DIRCE is cyclical with annual periodicity and is based on an intense use of data sources. Due to the diversity of the selected sources, the institutional and operative actions for their acquisition and processing were arranged in different modules, directly linked to the typology of sources (AT= Tax files, SS=Social Security files, PR=Private files,..) and making up the dynamic of the PIDE project.

When a source is received, a data quality program is applied according to their features and specific role in the business process. Some sources are core elements for the detection of new units or the elimination of previously existing ones. However, other sources are relevant for the maintenance of specific variables.

According to these parameters, the information provided in this phase basically refers to:

- The Inter Departmental context created allowing a stable access to input data (*4.2 set up collection*). Initially, institutional actions were addressed to the Tax and Social Security Authorities, formally established in both collaboration agreements. In other cases, specific service contracts are available for the acquisition of private databases
- The channels adopted for the reception of the data sources (*4.2 set up collection*). Diverse procedures are described, the most relevant of which being several IT tools allowing security requirements in data interchanges. In other cases, the direct download from official websites is applied
- The list of input sources used in each production cycle (*4.3/4.4 run and finalize collection*) classified by nature, in line with the components of the PIDE project

All input sources are numbered, this system being critical for the data integration and the DIRCE maintenance. A set of structured information is given for each database, including:

- Basic metadata: denomination, Managing Organization, reception date, timetable for data process, elementary observation unit and data structure

- Validation rules
- Editing and micro validation processes
- Transformation processes and adoption of statistical standards
- Production of statistical databases

Validation rules are designed with the purpose to make a decision about the source: acceptance or rejection. If rejected, the source is returned to the Managing Organization including a communication of errors to be corrected. If the source is accepted, a set of specific procedures is applied.

The lack of adaptation to statistical standards or the presence of low quality in particular variables can be pointed out as a classic restrictions of administrative data. These problems must be detected in the preliminary design of the project. Afterwards, the solution of these types of problems is normally the result of a close cooperation among statisticians and administrative managers, within the scope of the institutional context created. This is normally materialized by means of specific editing, micro validation or transformations processes.

The GSBPM offers possibilities for assigning this information in the next phase. However, in order to facilitate the understanding of the complete production chain, it has been decided to link all these register processes to the input sources in the collect phase.

3.3 Process phase

This phase is the core part of our business process. All statistical databases produced in the previous phase are used as input for the integration processes, the generation of the updated statistical units forming the data model and all the related characteristics linked to them. In summary, here is where the updating of the DIRCE takes part.

The integration procedures (*5.1 integrate data*) are mainly carried out by record linkage routines based on a universal presence of unique national IDs. During this action, several DIRCE frozen frames are generated with different levels of quality and uses. The features of the frames of reference t are described as a timeline over the year t+1 as the main result of the following iterative steps:

- INT_1_1 produces a preliminary updated version of enterprises, based on the new data sources

- INT_1_2 produces a second version of enterprises and local units fully consistent with the year t-1. The data quality is higher due to the incorporation of validated statistical information. This frame is used for sample selection in the STS domain and official dissemination of results

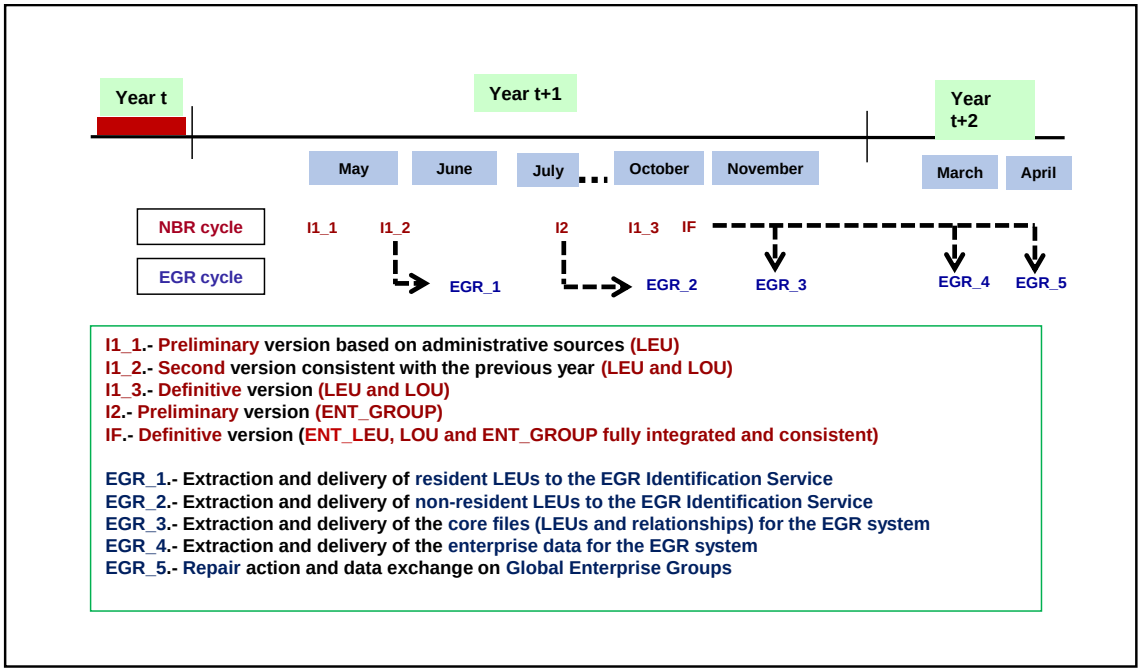
- INT_1_3 produces the definitive version of enterprises and local units incorporating the last updating of basic variables. In addition, specialized databases containing information on monetary characteristics are received during the last quarter of the year and this information is also incorporated

- INT_2 produces an updated version of enterprise groups based on private, tax and statistical sources

- INT_F produces the definitive updated system, integrating the results obtained in INT_1_3 and INT_2. Three levels of information formed by enterprises, local units and enterprise groups are available and fully consistent

In the last part of sub-process 5.1, the interactions with the EGR cycle are described. Due to the complexity of the EGR model and the need to give efficient answers from the NBRs, the uses of the frozen frames are described according to the different EGR interchange flows (extraction and delivery of resident/non- resident units to the EGR Identification Service, extraction and delivery of the *core* files on legal units and control relationships for the EGR system, extraction and delivery of the enterprise data to the EGR system, repair action and data exchange on Ultimate Controlling Units). Figure 3 shows these interactions as a timeline.

Figure 3. DIRCE frames and interactions with the EGR cycle



During the integration procedures, a definitive classification of all statistical units (*5.2 classify and code*) is also provided. For the core classification variables, a predefined set of decision rules is described, according to their presence in data sources, and their reliability.

New variables are also derived and systematically maintained based on information available or specific data sources (*5.5 derive new variables and units*). *Ad hoc* estimation procedures or deterministic rules are described for the delimitation of the number of persons employed, the institutional sector code or monetary variables like turnover, import and export.

The main restrictions were found in the documentation of review, validation, edit and imputation as separate sub-processes. As previously mentioned, all these practices are undertaken from the beginning of the cycle and they are allocated in the related sub-processes in order to facilitate the understanding of the whole production chain.

3.4 Analyse phase

The increasing demand for better and more detailed business statistics has put the focus on the NBRs and their key role in the statistical production chain. Originally, these tools were conceived as a vital component of statistical infrastructure, supporting data collection, monitoring the response burden and giving grossing up indicators for the production of

aggregates. All these tasks, closely related with the use of NBR as the survey frame, will be jointly considered in the application of the GSBPM to business surveys.

In recent decades, user demands have diversified and the role of the NBR as a source of data production has become more and more relevant. This aspect has been the approach adopted for the documentation of this phase and the following one. In the Spanish NSI, the DIRCE is the key data source for the statistical analyses of business activity from both a static and dynamic perspective. Two main references linked to the DIRCE macro-data are documented:

- Statistical Analysis of the DIRCE. A standard publication of results directly obtained from the updated frame
- Harmonized Business Demography. A product specifically elaborated to cover the national needs in this domain. Its production is fully consistent with the OCDE-Eurostat methodology

Both statistical operations incorporate the same metadata: type of operation, data source, periodicity, starting – ending date of processes, press release, presence in National Statistical Plan / Statistical Operations Inventory and methodological basis.

3.5 Dissemination phase

Dissemination of NBRs can be established at micro or macro-data levels. The first option is normally constrained by the confidentiality provisions applying to the national legal frame. This is the case of Spain, where access to the DIRCE micro data is restricted to national authorities in charge of official statistics. Dissemination of macro-data refers to the statistical operations previously mentioned. The main operational steps carried out up until their definitive publication are documented in this phase.

Joint meetings involving the DIRCE and Dissemination teams are held in the last part of each cycle. Information about the dynamic of the processes, the date foreseen for the generation of the aggregates and the innovations incorporated, form the basis for a proper adaptation of the output system (*7.1 update output systems*).

For the second stage, all components related to each operation are documented (*7.2 produce dissemination products*). They mainly refer to the list of data tables, metadata, standard methodological report, complementary reports, graphic annex and press release.

The external impact of these statistics is very relevant. A recent study of the number of web accesses to the INEBase, the generic brand for statistical information at NSI website consisting of 185 statistical operations, shows that DIRCE statistics are placed in the top 20 ranking.

Since the first year of publication, the DIRCE also provides a tailor-made service through direct use of register data. Requests from Public Administrations, Private Companies, Organizations, Professionals or Researchers are continuously increasing. The queries registered are very diverse in form and content and they are managed according to a specific protocol defined by the NSI dissemination policy (*7.5 manage user support*).

3.6 Evaluation phase

This phase is closely related to the quality policy implemented and the incorporation of successive improvements to our NBR. Two main orientations have been outlined:

- Internal evaluation, by developing a complete diagnosis of processes carried out during each annual cycle
- External evaluation, by using the feedback of business producers as a basic element for the improvement of the NBR

4. Final remarks

This has been a challenging and very positive experience for the DIRCE team. Although the GSBPM seems to be better adapted to a typical survey, this standard can also be applied to the Statistical Business Registers. However, the presence of national specificities in the management model makes the allocation to sub-processes sometimes difficult.

Different approaches can be adopted for the description of the phases, from a dynamic to a static perspective. Generally speaking, this decision should be made regarding the current level of implementation and the foreseen innovations to be included in the business process. In the case of management of BRs, which has a long-standing tradition in the Statistical Offices, the historical dimension could be more appropriate for the first phases and the static information linked to the most recent production cycle would be the appropriate approach for the remaining phases.

The GSBPM proposes a multi-focal description of the business process, allocating uniform parts in separate sub-processes. This exercise can be opportune when actions are addressed to the same dynamic database throughout the production chain. However, this philosophy can mean serious restrictions for projects involving a great amount and variety of data sources for which, specific actions must be designed. In the Spanish case, the longitudinal description for each input data source has been predominant in order to properly understand how our model actually works.

On an international scale, the results of these experiences will have to be jointly evaluated. As a starting point, the development of benchmarking activities will need to be undertaken. Expected results should lead to some agreements towards more coordinated and consistent production cycles. In addition, this context should facilitate the identification of specific tools for a common use or a preliminary identification of Data Quality Program for all BRs of the European System.

This progressive integration within an interoperable system will mean veritable added value for all statistical actors and an opportunity to modernise the production process of official statistics.

References

[1] Applying the Generic Statistical Business Process Model to business register maintenance. Economic Commission for Europe, Conference of European Statisticians. Paris, September 2011

[2]COM (2009) 404 final- Communication from the Commission to the European Parliament and the Council on the production method of EU statistics

http://eur-lex.europa.eu/LexUriServ/LexUriServ.do?uri=COM:2009:0404:FIN:EN:PDF

[3]Directive 2012/17/EU of the European Parliament and of the Council on interconnecting Business Registers

http://eur-lex.europa.eu/LexUriServ/LexUriServ.do?uri=OJ:L:2012:156:0001:0009:en:PDF

[4]http://www1.unece.org/stat/platform/display/metis/The+Generic+Statistical+Business+Process+Model

[5] http://www.ine.es/inebmenu/mnu_empresas.htm