THE COLOMBIAN EXPERIENCE ABOUT THE IMPLEMENTATION OF QUALITY PROCESSES ON THE STATISTICAL DIRECTORY OF ENTERPRISES (DEST)

Martha Poveda, Alexis Vladimir Maluendas, Oscar Mauricio Acosta & Gisela Castrillón
Statistic National Administrative Department
mpovedag@dane.gov.co avmaluendasp@dane.gov.co omacostao@dane.gov.co gcastrillonm@dane.gov.co

Abstract

The Information System for the Registration of Economical Units has been becoming developed in Colombia since 2010, in order to automatize processes and generate rules and controls allowing improving the quality of the information included into the database, which must fulfill the statistical requirements of the National Statistic System.

Those procedures implemented to achieve the aims in quality terms, namely, coverage, reliability, coherence, opportunity, accessibility and traceability, are exposed in this article. A description is presented and the how about its accomplishment is assured through the Directory Information System (SID); and measurement indicators are proposed according to the institutional parameter of quality.

The SID is composed of 9 modules which are: Supplier management, information preparation, information processing, actualization operations, analysis of quality, information exploitation, management and configuration, management and quality indicators, historic ones.
In spite of the quality control processes throughout the modules, the article is focused on the information preparation and processing ones, because these are the main where all the rules are implemented in order to the internal y external supplier registration information will accomplish the optimum characteristics so they can be included into the database.

1. Introduction

The Statistical Business Registration (SBR) is the instrument in which all the enterprises (legal or natural personalities), and establishments addressing economical activities in Colombia, are listed. This is the base of the economic investigations.

Having such a base is a fundamental condition to address any statistical research on the economic sector, either by means of the census (the total numbering of the universe elements) or sampling (numbering of a portion, sample, of statistical units, selected from the studied universe and its respective estimation of parameters) methodologies.

The SBR constitutes an infrastructure for a statistical information system and its maintenance in a rigorous and systematical way; so that it contributes to guarantee the representativeness of the information for making decisions, as well as the formulation and monitoring of public policies.

A description of the Colombian SBR and the need to develop an information system for improving and automatizing the actualization and maintenance process formerly made by hand are presented.

All this is focused to satisfy the objective that the Registration has and its mission within the statistical institution.

The main advances and results obtained by implementing a set of validation and normalization rules, as well as the description of the process to cross the administrative information with con the SBR information base are shown. Finally, five indicators looking for measure the quality of the SBR information are proposed.

2. Description of the Statistic Directory of Enterprises

The necessity to count with an instrument representing in an organized way the operative and legal structures of the enterprises rise from the effort to establish a centralized framework that can be used by every economic survey.

In this same way, since the early 90´s the DANE implements what it names the Statistic Directory of Enterprises – DEST, based on the results of the Multi-sectorial Economic Census, whose purposes keep coherent with the aims del Statistic Business Register utilized in other countries.

Moreover, by taking into account the transformations in the economy, the availability of new administrative information, the technological novelties and the changes in the user information requirements, the general framework tends to migrate to what is known as the Statistic Register System (or Statistic Business Register SBR).

The DEST is not out of these new trends, because of which it has become an object for re-designing. So, an Information System was developed for the Statistic Business Register (SID), looking for satisfying the information requirements of both the economic surveys and the researchers who make them, related to its own purpose as a statistic register.

Within this framework, the DEST, through the SID, aims to maintain an updated Business Register by means of the usage of information coming from administrative, updating operative (telephonic, field, Web) reports, and statistic operations, so its aim will be achieved as a input for the different investigations addressed by the DANE during the construction of frameworks with sampling purposes according to its information needs.

A clearness on the following basic components arising from the user information requirements is needed to work out the information updating process:

- Universe: the SBR must to contain all the enterprises (and its corresponding legal and local units) conducting economic activities in the country.
- Coverage: the Statistic Directory must guarantee a national coverage of all the statistic units addressing economic activities in every sector, this latter determined by means of the Uniform International Industrial Classification, Revision 4 adapted for Colombia (CIIU Rev. 4 A. C.)[1], extant in the year of reference.
- Statistic Units: the conceptual data model of the DEST refers to three basic units: the enterprise, the homogenous production unit and the establishment. Nonetheless, the model is being adequated into a more general framework in which the legal units, enterprises, local units and, in a more novel manner, the enterprise groups, are taken into account.
    - Enterprise: Economic entity or combination of the economic units capable, by their own right, of possessing actives, incur obligations and conduct economic and productive activities with other entities to develop and work out of the social objective to which it was created.
    - Homogenous production unit: It is characterized by an only activity: product inputs, production processes or homogeneous product outputs.
    - Establishment: An enterprise or an enterprise part located in a topographically delimited place in which, or from it, the economic activities are conducted.
    - The actualization is focused on an set of variables classified in four categories, according to the information needs: identification, location, stratification and management.

[1] That is a classification according to economic activity types, rather than a classification of goods and services. The activity made by an unit is the type of production to what it is devoted, and this will be the criterion by means of what those will be grouped with other units to form industries.

- Identification: the Tax Identification Number (NIT)[2], social reason, acronym, commercial (trade) name, legal representative and juridical form are observed, in addition to the identification keys for every statistic unit in the database.
- Location: department, municipality, address, web page, e-mail, telephone.
- Stratification: CIIU code, busy staff, incomes from operational or sales.
- Management: constitution date, status, activity initiation and cease dates.

The SID has completely automatized processes fundamentally directioned to the administrative information will be increasingly integrated, harmonized and coherent within the SBR. The SID is supported by the technological platform updating and the improved user access to the information to achieve this aim. In turn, the economic surveys will be improved too.

3. Quality improvement of the administrative information

The quality intends to the improvement of issues as the coverage, reliability, coherence, opportunity, accessibility and traceability in the frame of the Colombian SBR, by reason of which its updating and maintenance process is based on the following strategies:

- Processing and crossing of periodic administrative report files.
- Feed-back between the generated statistic frame and the economic surveys.
- Processing of economic units censuses made by the DANE.
- Own processes for verifying and validating the directory information.

The administrative information process is the most dynamic in the DEST, being implemented in six big modules by the SID, according to the established quality parameters: supplier information management, information preparation, information processing, updating operations, quality analysis and information exploitation.

Information management: The information management module is developed for monitoring the information since the moment when it is asked of the suppliers until it is received by the DANE. It is assessed whether the received information fulfill the parameters and minimum variable requirements so it can be used in the Register. The information is required with a variable dictionary and in a specific format. Nowadays, it is aimed to centralize the information about every economic activities contributing to the Colombia's GNP.

A process is included in which a general revision of every database given by the suppliers is worked out, and the statistic card is generated with the information diagnosis as a support for the preparation and loading processes.

Information preparation: the information preparation module consolidates the reports given by the suppliers at the level of the established economic units (Enterprises, Establishments and Homogeneous Production Units). A report consolidated with the different administrative sources is obtained as a result of this module. Evenly, the variables of the supplier consolidated file are semi-automatically homologized to those of the directory.

Information processing: the information processing module automatically normalizes, codifies, and applies rules of validation and consistency of the information consolidated from the suppliers to update the Report. The following basic elements are distinguished in this processing:

- Validation rules incorporated to the Information System.
- Key Generation to the statistic units will be uniquely identified.

---

[2] The Tax Identification Number (NIT) constitutes the identification number of those enrolled in the Tax Unified Register (RUT), allowing individualizing the contributors and users, for every effect on tax, custom and change matters, and especially for accomplishing such obligations.

- Address normalization for every statistic unit.
- Detection and elimination of identical duplicates.
- Creation of a catalog to normalize words. For example, the word “Limitada” (Limited) could be abbreviated as Lta, Ltda, Limit, etc., but the process normalizes it as “Ltda”.

Some statistic units are sent to a base for revision when they do not fulfill these parameters. Additionally, the information received from administrative sources is crossed in this module, once the anterior process has been worked out with the directory base information.

The Information System determines in automatic manner whether the statistic unit exists in the data base, and identifies the high ones, the permanencies and the low ones.

The reference date and the statistic unit origin are registered for every variable, as the supplier information is updated, so, the process traceability can be kept.

When a variable is provided by more than an only supplier, ponderations are defined according the level of confidence given by the source. For example, updating operations determine that the total amount of busy people can be more reliably registered as contributors to social security than the data gathered by the surveillant institutions.

Updating Operations: Information updating operations are conducted by means of call centers, internet and electronic forms for economic units that can be mainly either high or low. The call center technologic platform has been modernized to improve the quality. The web form allows the enterprises supplying its information to directly access.

Quality analysis: A set of tools making easy the analysis of any kind of information either coming from administrative reports or contained in the DEST to calculate "clusters" determined by the operator.

Information Exploitation: The real time consultation module is created to allow generating frames for the Information System users.

4. Results of the Information Preparation and Processing Processes

The information preparation module yields two main products: a consolidated base and a statistic card structured in 4 parts:

- A variable list, and their descriptions, coming from the file provided by the supplier (Dictionary).
- An attachment to the Dictionary in which the variable classifications are presented.
- A base diagnosis including duplicate, empty and inconsistent registers. (Image 1).
- Frequencies for every categorical variable, for example, department, municipality, legal organization, among others).

Image 1. Data base diagnosis for updating.

DEPARTAMENTO ADMINISTRATIVO NACIONAL DE ESTADISTICA
DIRECCION DE GEOESTADÍSTICA - FICHA TECNICA

CARACTERÍSTICAS DE LA INFORMACION

ENTIDAD QUE SUMINISTRA LA INFORMACION	PLANILLA INTEGRADA DE LIQUIDACIÓN DE APORTES
FECHA EN LA QUE SE RECIBE LA INFORMACION	31/12/2012
NOMBRE DEL ARCHIVO ORIGINAL	20121231163952_Aportes_2_2012_DANE_2.txt
UBICACIÓN ARCHIVO ORIGINAL	c:\datos\DEE\ARCHIVOS_PLANOS_REGISTROS_ADMINISTRATIVOS\PILA\
CANTIDAD DE VARIABLES	40
CANTIDAD DE REGISTROS	13444848

DESCRIPCIÓN DE LA INFORMACION

			PRESENCIA DE DATOS		VALIDACION DE DATOS		INFORMACION DUPLICADA POR VARIABL		
ORDEN	NOMBRE_ARCHIVO	NOMBRE_DIRECTORIO	Registros vacios	Registros no vacios	Registros inconsistentes	Registros consistentes	Registros duplicados	Registros unicos	Registros diferentes
1	tipo_identificacion	TIPO_DOCUMENTO	0	13444848	0	0	7	1	8
2	numero_identificacion	NIT	0	13444848	0	0	2221150	151464	2372614
3	digito_verificacion	DIGITO_VERIFICACION	4329	13440519	0	0	12	0	12
4	razon_social	RAZON_SOCIAL	0	13444848	0	0	2202938	149567	2352505
5	codigo_sucursal		11075723	2369125	0	0	2555	113	2668
6	nombre_sucursal		639808	12805040	0	0	15549	698	16247
7	clase_aportante		639185	12805663	0	0	8	0	8
8	sector_aportante		639656	12805192	0	0	7	0	7
9	tipo_persona	OJUR_ID_OJURIDICA	639656	12805192	0	0	4	0	4
10	direccion_correspondencia	DIRECCION	647518	12797330	0	0	1893205	111269	2004472
11	codigo_ciudad	MUNI_ID_MPIO	639667	12805181	0	0	595	1	596

Source: Information System for the Statistic Business Register.

The result obtained based on the social security contributor register is presented in order to exemplify the quality processes related to the information consolidation. The integrated form to liquidate contributions (PILA) is a register keeping the information of all either independent people or enterprises that pay contributions to the Social Protection System and Para-fiscal Payments[3]. The supplier provided base initially has 13,404,161 reports, each corresponding to a payment form for different month of the year. The total amount of forms is shown in the table 1 for each provided period.

PERIOD	No. Forms	No. Unique Contributors
01/01/2012	2,154,417	1,939,009
01/02/2012	2,152,264	1,925,639
01/03/2012	2,249,926	2,009,856
01/04/2012	2,320,143	2,068,211
01/05/2012	2,299,345	2,054,399
01/06/2012	2,228,066	1,996,297
TOTAL	**13,404,161**	**11,993,411**

In this case, the number of unique contributors (less the duplicates) is approximately the total number of enterprises that will be included in the data base. After the consolidation into an only register at the level of the economic unit, 2,137,574 registers were obtained.

This consolidated base passes to the information preparation module in which the variables are homologized, validation rules are applied, the addresses are normalized, and so on.

Codification Rules: The data base information received according to the reference tables is codified in this process. That is the case of the municipality, the department, the economic activity, the type of document and the legal organization. For example, there are cases in which the information does not come codified but named with the department or municipality where the economic unit locates. The process transforms these data into the official codes according to the Colombian Political-Administrative Division (DIVIPOLA).

[3] Resolution 1303 of 2005 by the Ministry of Health and Social Protection

The codifications joined to the variables are reviewed by taking into account the reference tables.

Normalization rules: The System allows normalizing the economic unit address (location), name, telephones, among other features, by applying defined updated rules, according to the word normalization catalogs.
Not accepted characters, for example, symbols like # or – in the addresses, are eliminated within the normalization process.

Validation rules: The information completeness and consistency ought to be guaranteed with these:

- The length of the telephonic numbers must be 7 or 10 digits, without the inclusion of the city indicative. The datum type must be numeric.
- The address must have more than 4 characters.
- The commercial (trade) name must have more than 4 characters.
- If the statistic unit was classified as active, and comes from a DANE's survey, it must have information about incomes and busy people.
- Every statistic unit must have an identification number, and this latter must be within a specific length range depending on the type of the associated document.
- The e-mails must have the symbol @ and belong to a valid dominion.

If some statistic units do not fulfill these conditions in spite the rules has been applied, the supplier is informed and the different sources are consulted in order to correct them and load them into the Information System. Then, the identical duplicates are identified in all their variables and eliminated.

Statistic directory data base updating: The System crosses a consolidated, depurated and normalized database with the DEST database. This process allows identifying units already found in that Register with the name of "Permanencies", other units new for the register are called "High", and other units found in bases during anterior years but not in the present ones named "Low". All the information is updated, but "High" and "Low" files are generated and obligedly sent to updating operations to verify their status.

When the specific variable information has been updated by another supplier with the same reference date, the weight assigned to each supplier is taken into account. In this point, the information passes to the Updating Operation module.

5. Toward a statistic business register with quality

The DANE intends a series of challenges to the future, some of which are in an implementation process through the Information System (SID) within the DEST improvement scheme:

- To adjust the SID as needed to consolidate all the information integration and depuration process.
- To include statistic unit continuity concepts. An unit extant in the base is considered a new one when at least two of the following three criteria change:
    - Unit identifier: either the Tax Identification Number (NIT) or the social reason (name)
    - Unit espacial location (Address)
    - Enterprise economic activity.
- To generate Enterprise demography based on statistic unit continuity criteria.
- To increase the DEST usage through the SID by means of the diffusion inside the statistic entity and the integration into the National Accounts.
- To manage new information sources to strength the register base updating processes, particularly, the tax register.

To enforce the legal frame by means of which the DANE will get the authority needed to access the administrative registers required by the DEST.

6. Proposed Quality Indicators

In this moment, Colombia is conducting the process of defining a set of indicators whose purpose is to evaluate the quality of SBR. The indicators proposed are:

Indicator 1
Name: Updating level
Objective: to know the updating rate for every economic sector in the frame.
Type of Indicator: Process quality
Variables used in the calculations are:
A_j: Total updated records for sector j
B_j: Total records expected for updating in sector j
The formula used for the calculation is:

$$I_{1_j} = \frac{A_j}{B_j} * 100$$

Calculation Frequency: Annually
Tolerance ranges:
Critical <= 70;
70> Fair <= 90;
Satisfactory> 90.

Indicator 2
Name: birth and death tracking
Objective: To evaluate the sector dynamics based on the economic unit demographics.
Type of Indicator: Process quality
Variables used in the calculations are:
A_i: birth in year i
B_i: death in year i
The formula used for the calculation is:

$$I_{2A} = \frac{A_i}{A_{i-1}}$$

$$I_{2B} = \frac{B_i}{B_{i-1}}$$

Calculation Frequency: Annually

Indicator 3
Name: Coverage
Objective: To establish the SBR coverage compared to the information from the Tax and Custom National Direction (DIAN) database.
Type of Indicator: Process quality.
Variables used in the calculations are:
t_i: Total records in the database in year i
T_i: Total unique records identified in the Tax database of the National Register in year i
The formula used for the calculation is:

$$I_3 = \frac{t_i}{T_i} * 100$$

Calculation Frequency: Annually
Tolerance ranges:
Critical <= 70;
70> Fair <= 90;
Satisfactory> 90.

Indicator 4

Name: Employment precision

Objective: To determine whether the information by sector kept in the SBR is approximated to the official employment statistics generated by any National Statistics Institute.

Type of Indicator: Quality of the process

Variables used in the calculations are:

e_i: Total employees according information contained in the SBR in year i.

PEA_i: Economically active population in the year i, according to official data.

Li: Lower limit established for estimating the unemployment by any Statistic National Institute

Ls: Upper limit established for estimating the unemployment by any Statistic National Institute

The formula used for the calculation is:

$$I_4 = \begin{cases} 1 & si\ L_i \leq 1 - \frac{e_i}{PEA_I} \leq L_s \\ 0 & otherwise \end{cases}$$

Calculation Frequency: Annually

Tolerance ranges:

1 Satisfactory;

0 critical

Indicator 5

Name: Income precision

Objective: To determine whether the information by sector kept in SBR is approximated to the official income statistics generated by any Institute of Statistics.

Type of Indicator: Quality of the process

Variables used in the calculations are:

c_i: Total income according with information keep in the SBR in year i

C_i: Total income according to official data of national accounts in year i

The formula used for the calculation is:

$$I_5 = \begin{cases} 1 & si\ 0.9 \leq \frac{c_i}{C_i} \leq 1.1 \\ 0 & en\ otro\ caso \end{cases}$$

Calculation Frequency: Annually

Tolerance ranges:

1 Satisfactory;

0 critical

Indicator 6

Name: Opportunity

Objective: To establish the level in which registers are available for being used by the SBR users.

Type of indicator: Output quality

Variables used in the calculations are:

u_j: Number of users qualifying the access as timely to SBR (schedule)

U: Total number of interviewees

The formula used for the calculation is:

$$\mathrm{I}_6 = \frac{\mathrm{u_j}}{\mathrm{U}} * 100$$

Calculation Frequency: Annually

Tolerance ranges:

Critical <= 70;

70> Fair <= 90;

Satisfactory> 90

7. References

Bérard, Hélène, Pursey, Stuart & Rancourt, Eric. *Re-thinking Statistics Canada's Business Register.* Statistics Canada. 11th, Floor, R.H. Coats Building,Tunney's Pasture, Ottawa, Ontario Canada K1A 0T.

DANE, *Metodología Directorio Estadístico de Empresas –DEST–,* Bogotá, D. C, Colombia. Versión 5. 2013.

Eurostat. *The French Business Register: From a quality approach to a statistical register*. Projects on the Improvement of Business Register. Session 5.

Eurostat. *Business Register Recommendations Manual: The use of Administrative Sources*. Chapter 20. December 2002.

Ritzen, Jean. *Statistical Business Register: Content, place and role in Economic Statistics*. Statistics Netherlands