

# Expérience Colombienne dans la Mise en Œuvre du Processus Qualité du Répertoire Statistique d'Entreprises

Martha Poveda, Alexis Vladimir Maluendas, Oscar Mauricio Acosta et Gisela Castrillón  
Département Administratif National de la Statistique

[mpoveda@dane.gov.co](mailto:mpoveda@dane.gov.co) [avmaluendasp@dane.gov.co](mailto:avmaluendasp@dane.gov.co) [omacostao@dane.gov.co](mailto:omacostao@dane.gov.co) [gcastrillonm@dane.gov.co](mailto:gcastrillonm@dane.gov.co)

## Résumé

En Colombie depuis 2010 s'est développé le système d'information pour l'enregistrement des unités économiques. Le but est d'automatiser les processus et créer des règles et des contrôles pour améliorer la qualité des informations contenues dans la base de données. Cette dernière doit répondre aux exigences statistiques du système statistique national.

Cet article décrit les procédures mises en œuvre pour atteindre les objectifs de la SBR en termes de qualité (couverture, fiabilité, cohérence, opportunité, accessibilité et traçabilité). De la même façon nous verrons comment le système d'information du répertoire (SID) s'assure de la réussite du projet. Enfin seront présentés des indicateurs de mesure déterminés en fonction des paramètres de qualité du DANE.

Le SID (Système d'information du Répertoire Statistique) se compose de 9 modules qui sont: la gestion des fournisseurs, la préparation de l'information, le traitement de l'information, les opérations d'actualisation, l'analyse de la qualité, l'exploitation de l'information, l'administration et la configuration, les indicateurs de qualité et de gestion et l'historique.

Bien qu'il existe des processus de contrôle de qualité pour tous les modules, l'article se concentre sur deux modules, les modules de préparation et de traitement de l'information. Car ils sont les centres principaux où toutes ces règles sont appliquées pour que l'information des registres, des fournisseurs internes et externes remplissent les exigences nécessaires pour être incorporées à la base de données.

## 1. Présentation

Le registre statistique des entreprises - SBR en Colombie est l'instrument qui recense toutes les entreprises (morales ou naturelles) et les établissements qui développent l'activité économique dans le pays. Le registre ou SBR sert de base aux investigations économiques.

Cette base de données est une condition préalable au développement de toute recherche statistique dans le secteur économique, soit par des méthodologies de recensement (énumération complète des éléments d'ensemble) ou d'échantillonnage (énumération d'un échantillon ou de parties des unités statistiques sélectionné dans l'univers de l'étude et l'estimation respective des paramètres).

Le SBR est l'infrastructure d'un système d'information statistique et sa maintenance rigoureuse et systématique contribue à assurer la représentativité de l'information pour la prise de décisions, la formulation et le suivi des politiques publiques.

Nous présenterons une description du SBR en Colombie ainsi que le besoin de développer le système d'information afin d'améliorer et automatiser le processus de mise à jour et la maintenance qui étaient auparavant faits manuellement. Cela pour satisfaire l'objectif du registre et sa mission au sein de l'entité statistique

Le chapitre suivant présente les principales avancées et les résultats obtenus avec la mise en œuvre d'un ensemble de règles et de normes de validation ainsi que la description du processus de croisement des informations administratives avec celles du SBR. Enfin, nous proposerons cinq indicateurs qui visent à mesurer la qualité de l'information du SBR.

## **2. Description du registre statistique des entreprises**

La nécessité d'un instrument qui représente de manière organisée les structures opératives et juridiques des entreprises a pour but d'établir un cadre centralisé qui peut être utilisé par toutes les enquêtes économiques.

Dans ce sens, le DANE implémente depuis le début des années 90, ce qu'il appelle le Répertoire statistique des entreprises - DEST qui repose sur les résultats du Recensement économique multisectoriel et dont les objectifs sont compatibles avec les objectifs du registre statistique des entreprises, utilisé dans d'autres pays.

Cependant, parce qu'il tient compte des changements dans l'économie, de la disponibilité des nouveaux développements, des changements dans les renseignements techniques administratifs, et des besoins des utilisateurs, le cadre général a tendance à évoluer vers ce qui est connu comme le registre statistique des entreprises - SBR, pour son sigle en anglais.

Le DEST n'est pas étranger à ces nouvelles tendances et a subi une refonte qui se reflète dans le développement du système d'information pour la statistique d'entreprises – SID, lequel vise à répondre aux besoins d'information des enquêtes économiques et de chercheurs.

Dans ce cadre, le DEST à travers du SID vise à maintenir l'enregistrement des entreprises à jour en utilisant les informations des dossiers administratifs (téléphone, sites sur le terrain) et les opérations statistiques. Son but est de contribuer comme source aux différentes enquêtes menées par le DANE dans la construction d'échantillonnages adaptés à leurs besoins d'information.

Pour mener à bien le processus de mise à jour des informations est nécessaire d'être clair sur les éléments de base suivants qui découlent des besoins d'information des utilisateurs:

- Environnement: SBR doit inclure toutes les entreprises dans leurs unités légales et locales engagées dans une activité économique dans le pays.
- Couverture: l'Annuaire statistique doit garantir une couverture nationale de toutes les unités statistiques engagés dans des activités économiques dans tous les secteurs déterminés par la Classification Industrielle Internationale Uniforme, Révision 4 adaptée à la Colombie (CIIU Rev. 4 A. C.), en vigueur dans l'année de référence.
- Unités Statistiques: le modèle conceptuel de données DEST se réfère à trois unités de base: l'entreprise, l'unité homogène de production et l'établissement. Cependant, le modèle est adapté à un cadre plus général, qui prend en compte les unités légales, les entreprises, les unités locales, groupes d'entreprises plus innovantes.

- Entreprise : Entité économique ou combinaison d'unités économiques capables, de son propre chef, de posséder des actifs, des engagements, de s'engager dans des activités économiques et productives avec d'autres entités, pour le développement et la mise en œuvre des objectifs sociaux pour lesquels il a été créé.
- Unité homogène de production: Elle est caractérisée par une activité unique: par entrée de produits et des sorties de produits homogènes.
- Etablissement : entreprise ou partie d'une entreprise située dans un lieu topographiquement identifié. Dans ce lieu, ou à partir de lui, la société exerce une activité économique.

La mise à jour se concentre sur un ensemble de variables, en accord avec les besoins de l'information, lesquels sont classés en quatre catégories: les variables d'identification, les variables de localisation, les variables de stratification et les variables de gestion :

- Les variables d'identification: en plus des clés d'identification de chaque unité statistique on observe le numéro d'identification fiscale - NIT, nom, raison sociale, sigle, nom commercial, représentant légal et forme juridique.
- Variables géographiques: département, municipalité, adresse, site web, email, téléphone.
- Variables de stratification: Code CIIU, personnel, revenus (opératifs ou ventes).
- Les variables de gestion: date de constitution, état, date de début d'activité, date de cessation d'activité.

Le SID compte avec un ensemble entièrement automatisé de processus à travers desquels s'effectue le traitement de l'information provenant de sources administratives. L'objectif étant d'accroître l'intégration, l'harmonisation et la cohérence des registres. Pour réaliser cet objectif, le SID s'appuie sur l'actualisation de la plateforme technologique et l'amélioration de l'accès à l'information pour les usagers. Ces avancées visent à l'amélioration de la qualité de l'information présentée dans le SBR qui à son tour entraîne une amélioration de la qualité des enquêtes économiques.

### 3. Amélioration de la qualité de l'information administrative

Dans le cadre du registre statistique des entreprises en Colombie, la qualité vise à améliorer les aspects suivants : couverture, fiabilité, cohérence, rapidité, accessibilité et traçabilité. Pour répondre à ces idéaux, la mise à niveau et la maintenance du registre des entreprises sont basées sur les stratégies suivantes:

- Poursuite et croisement fichier d'enregistrements administratifs (information périodique).
- Retro alimentation entre le cadre statistique généré et des enquêtes économiques.
- Traitement de recensement des entités économiques réalisées au sein du DANE.
- Processus internes de vérification et validation des informations d'annuaire

Le traitement de l'information administrative est le plus dynamique dans le DEST. Ce traitement avec les paramètres de qualité préétablis au sein du le SID est mis en œuvre dans six grands modules: la gestion des fournisseurs, la préparation de l'information, traitement de l'information, les opérations d'actualisation, l'analyse de la qualité et l'exploitation de l'information.

Gestion de l'information: l'objectif est de faire le suivi des fournisseurs à partir du moment où l'information est demandée jusqu'à ce qu'elle soit reçue par l'entité. Par la suite, on contrôle que

les informations reçues répondent aux exigences et paramètres minimum pour permettre son utilisation dans le Registre. L'information s'accompagne nécessairement d'un ensemble de variables et dans un format spécifique. Actuellement on cherche à centraliser l'information de toutes les activités économiques qui contribuent au PIB du pays.

Un processus de révision générale de chacune des bases de données apportées par les fournisseurs est réalisé de même qu'une fiche statistique avec un diagnostic de l'information comme support des processus de de préparation et de chargement.

Préparation de l'information: ce module exécute automatiquement la normalisation, l'encodage, l'application de règles validation et la cohérence d'information consolidées des fournisseurs pour mettre à jour le registre. Dans ce traitement, nous distinguons les éléments de base suivants:

- Les règles de validation intégrées dans le système d'information.
- Génération de clés qui identifient de manière unique les unités statistiques.
- La normalisation des adresses pour chaque unité statistique.
- Détection et élimination des doublons.
- Création d'un catalogue pour la normalisation des mots (par exemple, le mot limitée: peut venir écrit comme Lta, Ltda, limite, etc, alors la procédure le normalise a Ltda.).

Les unités statistiques qui ne satisfont pas ces paramètres sont envoyées à une base de données pour examen. En outre, ce module croise les informations reçues de sources administratives, une fois le traitement ci-dessus réalisé avec l'information de l'annuaire.

Le système d'information détermine automatiquement s'il existe une unité statistique dans la base de données et identifie les hautes, les basses et les permanences. (Les unités présentes dans le registres et qui l'étaient déjà les années précédentes sont appelées «permanences». Les nouvelles unités sont appelé " hautes ", tandis que les unités qui étaient dans la base dans les années précédentes mais pas actuellement sont appelées " basses".)

Au moment de la mise à jour des informations des fournisseurs, on note la date de référence et l'origine de l'unité statistique pour chaque variable, ce qui permet de garder la traçabilité du processus.

Dans le cas où une variable est fournie par plus d'un fournisseur, les pondérations sont définies selon le niveau de confiance donné par la source. Par exemple, les données relatives à l'emploi total sont plus fiables dans le registre de la sécurité sociale.

Operations d'actualisation : Elles se réalisent par des moyens tels que centres d'appels, Internet et les formulaires électroniques à des unités économiques qui peuvent être élevés ou bas principalement. La plate-forme technologique de centre d'appels a été modernisée pour améliorer la qualité.

À cette fin, le formulaire web a un accès direct pour que les entreprises fournissent leurs informations.

L'analyse de la qualité : contient un ensemble d'outils qui facilitent l'analyse de tout type d'information qu'il provienne de registres administratifs ou contenus sur le DEST par des calculs «agrégats» déterminés par l'opérateur

L'exploitation des informations: en temps réel ce qui permet la génération de références pour les utilisateurs du système d'information.

#### 4. Résultats des processus de préparation et de traitement de l'information

Le module de préparation de l'information apporte deux produits principaux: une base consolidée et un bilan statistique. Le bilan statistique est structuré en quatre parties:

- Une liste des variables et leur description provenant du dossier présenté par le fournisseur (Dictionnaire).
- Une annexe au dictionnaire où se présentent les classements qui ont les variables.
- Un diagnostic de la base, qui inclut des doublons, des lacunes et incohérences. (Photo 1).
- Fréquences pour chaque variable catégorielle (département par exemple, de la municipalité, l'organisation juridique, etc.)

Photo 1. Diagnostic de la base de données pour actualisation

ENTIDAD QUE SUMINISTRA LA INFORMACION		FECHA EN LA QUE SE RECIBE LA INFORMACION		NOMBRE DEL ARCHIVO ORIGINAL		UBICACION ARCHIVO ORIGINAL		CANTIDAD DE VARIABLES		CANTIDAD DE REGISTROS	
DANE		31/12/2012		20121231163902_Aportes_2_2012_DANE_2.H4		c:\datos\DEE\ARCHIVOS_PLANOS_REGISTROS_ADMINISTRATIVOS\PILA		50		13444840	
DESCRIPCION DE LA INFORMACION											
CODI	NOMBRE ARCHIVO	ESQUEMA DIRECTORIO	PRESENCIA DE DATOS		VALIDACION DE DATOS		INFORMACION DUPLICADA POR VARIABLES				
			Registros validos	Registros vacios	Registros incorrectos	Registros correctos	Registros duplicados	Registros unicos	Registros faltantes		
1	tipo_identificacion	TIPO_DOCUMENTO	0	13444840	0	0	2	151464	2372614	0	
2	numero_identificacion	NET	0	13444840	0	0	32	0	12	0	
3	digito_verificacion	DIGITO_VERIFICACION	4329	13440519	0	0	0	0	0	0	
4	sexo_social	SEXO_SOCIAL	0	13444840	0	0	2302938	149567	2392505	0	
5	codigo_municipal	CODIGO_MUNICIPAL	11975723	2369125	0	0	2555	113	2669	0	
6	nombre_sucursal	NOMBRE_SUCURSAL	639068	12805040	0	0	10549	628	14247	0	
7	clase_aportante	CLASE_APORTANTE	639185	12805643	0	0	0	0	0	0	
8	sector_aportante	SECTOR_APORTANTE	639056	12805192	0	0	7	0	7	0	
9	tipo_persona	TIPO_PERSONA	639056	12805192	0	0	4	0	4	0	
10	direccion_correspondencia	DIRECCION_CORRESPONDENCIA	447618	12797230	0	0	193205	111268	204473	0	
11	codigo_ciudad	CODIGO_CIUDAD	639617	12805181	0	0	505	3	508	0	

Source: système d'information pour le registre statistique des entreprises.

Pour illustrer les processus de qualité concernant la consolidation de l'information, on présente les résultats obtenus avec la base du registre des contributeurs à la sécurité sociale.

Le formulaire complet de Liquidation et contributions - PILA - est un registre qui stocke les informations de toutes les personnes indépendantes ou non ou bien des sociétés qui versent des cotisations au Système de Protection Sociale et Paiements Parafiscaux.

La base fournie par le fournisseur compte initialement un total de 13'404.161 registres, où chacun correspond à un plan salariale, pour les différents mois de l'année, ainsi que le montre le tableau 1 pour chaque période de temps donnée.

PERIODO	No. Paillas	No. Aportantes Únicos
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
<b>TOTAL</b>	<b>13.404.161</b>	<b>11.993.411</b>

Dans ce cas, le nombre de cotisants uniques (en éliminant les doublons) est environ le nombre total d'entreprises qui sont dans la base de données. Après la consolidation d'un seul registre au niveau de l'unité économique (entreprises dans le présent cas) on obtient 2'137 .574 dossiers.

Cette base consolidée passe ensuite au module de la préparation de l'information, où les variables sont homologuées suivant les règles de validation et de normalisation.

Règles d'encodage: dans ce processus, l'information reçue de la base de données est chiffrées suivant les exigences factuelles (ex : la commune, le département, l'activité économique, le type de document et l'organisation juridique).

Par exemple, dans le cas du département et de la municipalité, il y a des cas où l'information ne vient pas codifiée sinon avec le nom du département ou de la commune où l'unité économique, dans ce cas, le processus transforme ces données conformément à la Division Politique Administratif Colombie– DIVIPOLA.-

Règles pour la normalisation:

Le système peut normaliser l'adresse (lieu), le nom de la société et raison sociale, téléphone, en autres, de l'unité statistique, les règles s'appliquent selon les catalogues de normalisation des mots.

Dans le cadre du processus de normalisation sont supprimés des caractères interdits, par exemple, des symboles tels que # ou - dans les adresses.

Règles de validation:

Cette étape comprend des règles visant à assurer l'exhaustivité et la cohérence de l'information.

Certaines des règles de validation sont:

- La longueur des numéros de téléphone doit être de 7 ou 10 chiffres, sans le code de ville. Le type de données doit être numérique.
- L'adresse doit être supérieure à 4 caractères.
- Le nom commercial doit être supérieur à 4 caractères
- Si l'unité statistique a été classé comme active et provient d'une enquête du DANE il doit y avoir des informations sur le revenu et les employés.
- Toutes les unités statistiques doivent avoir un numéro d'identification et il doit être d'une longueur spécifique en fonction du document associé.
- Les Emails doivent remplir des conditions comme le symbole @ et doivent appartenir à un domaine valide

Après application des règles ci-dessus, si certaines unités statistiques ne répondent pas à ces conditions, nous en informons le fournisseur et consultons différentes sources afin de les corriger et de les charger dans le système d'information. Après quoi sont identifiés les doublons sur toutes les variables pour être éliminés.

Mise à jour de la base de données répertoire statistique

Après avoir une base de données consolidée, purifiée et standardisée, le système effectue un croisement d'information avec la base de données DEST. Ce processus permet d'identifier les unités qui étaient déjà dans le registre («permanences»), celles qui sont nouvelles («hautes») et celles qui ont disparues («basses»).

Toutes les informations sont mises à jour, bien que soient générées des « hautes » et des « basses » lesquelles sont envoyés a des agents pour vérifier leur état.

Lorsque l'information d'une variable spécifique a déjà été mise à jour par un autre fournisseur avec la même date de référence, on tient compte du poids qui est attribué à chaque fournisseur.

À ce stade, l'information est transmise au module opérations des actualisations.

## **5. Vers un registre statistique des entreprises de qualité**

Dans le cadre d'une amélioration du DEST à travers la mise en œuvre du système d'information (SID), le DANE soulève un certain nombre de défis futurs, dont certains sont en cours d'exécution:

- Faire les ajustements nécessaires au SID pour consolider le processus d'intégration et de débogage des informations.
- Inclure les notions de continuité des unités statistiques. Une unité existante de la base est considérée comme une nouvelle unité lorsqu'au moins deux des trois critères suivants sont modifiés:
  - Changement d'identifiant de l'unité (Numéro d'identification fiscale - NIT et nom de l'entreprise),
  - Changement de la localisation spatiale de l'unité (adresse)
  - Changement de l'activité économique de l'entreprise.
- Générer la démographie des entreprises en fonction de critères de continuité des unités statistiques.
- Encourager l'utilisation généralisée du DEST a travers du SID par la diffusion dans les statistiques et leur intégration aux comptes nationaux.
- Gérez de nouvelles sources d'information pour renforcer le processus de mise à jour de la base du registre, notamment le registre des impôts.

Renforcer le cadre juridique accordant au DANE les pouvoirs nécessaires pour accéder aux dossiers administratifs requis par le DEST.

## **6. Proposition d'indicateurs de qualité**

Pour le suivi des mesures prises dans le contexte d'amélioration du registre des unités économiques grâce au système d'information, sont définis un ensemble d'indicateurs dont le but est d'évaluer l'instrument en fonction de son efficacité, l'efficience et l'effectivité.

### **Indicateur 1**

Nom: Niveau d'actualisation

Objectif: connaître le taux de mise à jour pour chacun des secteurs économiques

Classe d'indicateur: Efficacité

Variables utilisées dans les calculs:

$A_j$ : Total registres actualisés pour le secteur  $j$

$B_j$ : Total dossiers de mise à jour attendues  
dans le secteur  $j$

Méthode de calcul:

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

Périodicité: annuelle

Marges de tolérance:

Critique  $\leq 70$ ; 70 > Acceptable  $\leq 90$ ; Satisfaisant > 90.

## **Indicateur 2**

Nom: Suivi des hautes et basses

Objectif: détecter les signaux de capacité du registre à refléter la réalité économique

Classe d'indicateur: Efficacité

Variables utilisées dans les calculs:

$A_i$ : hautes pour l'année  $i$

$B_i$ : basses pour l'année  $i$

Méthode de calcul:

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

Périodicité : annuelle

Plages de tolérance: Il a établi en relation avec l'évolution historique des hauts et des bas. Il est attendu que la valeur ne s'éloigne pas de 1.

## **Indicateur 3**

Nom: La couverture

Objectif: déterminer la mesure dans laquelle le registre couvre la totalité des unités en termes de référence fiable.

Classe d'indicateur: Efficacité

Variables utilisées dans les calculs:

$t_i$ : Nombre total d'enregistrements contenus dans la base de données du système d'information pour l'année  $i$   
Sistema de Información en el año  $i$

$T_i$ : Total des enregistrements uniques identifiés dans la base du registre d'impôt national pour l'année  $i$

Méthode de calcul:

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

Périodicité: annuelle

Plages de tolérance:

Critique  $\leq 70$ ; 70 > Acceptable  $\leq 90$ ; Satisfaisant > 90

## **Indicateur 4**

Nom: Précision pour l'emploi



Objectif: déterminer si les informations contenues dans le système d'information sectorielle des bases de données se rapprochent de manière adéquate des statistiques officielles de l'emploi généré par le Département des statistiques.

Type d'indicateur: Effectivité

Variables utilisées dans les calculs:

$e_i$ : Nombre total d'employés selon l'information consolidé contenues dans le SID pour l'année  $i$

$PEA_i$ : Population économiquement active dans l'année  $i$ , selon les chiffres officiels

$L_i$ : Limite inférieure confidentiel établie pour estimer le chômage par l'Institut National Statistique

$L_s$ : Limite supérieure confidentiel établie pour estimer le chômage par l'Institut National Statistique

Méthode de calcul:

$$I_4 = \begin{cases} 1 & \text{si } L_i \leq 1 - \frac{e_i}{PEA_i} \leq L_s \\ 0 & \text{à d'autre cas} \end{cases}$$

Périodicité: annuelle

Plages de tolérance:

1 satisfaisant; 0 critique

## **Indicateur 5**

Nom: Précision pour les recettes

Objectif: déterminer si les informations contenues dans le système d'information se rapprochent suffisamment des statistiques officielles générées par l'Institut National Statistique.

Classe d'indicateur: Effectivité

Variables utilisées dans les calculs

$c_i$ : Total des revenus provenant de l'information consolidée contenue dans le SID pour l'année  $i$

$C_i$ : Total des revenus selon les chiffres officiels des comptes nationaux pour l'année  $i$

Méthode de calcul:

$$I_5 = \begin{cases} 1 & \text{si } 0.9 \leq \frac{c_i}{C_i} \leq 1.1 \\ 0 & \text{à autre cas} \end{cases}$$

Périodicité: annuelle

Plages de tolérance:

1 satisfaisant; 0 critique

## **Indicateur 6**

Nom: Opportunité

Objectif: établir la mesure dans laquelle le registre est disponible de manière opportune pour une utilisation par des utilisateurs.

Classe d'indicateur: l'efficience

Variables utilisées dans les calculs:

$u_j$ : Nombre d'utilisateurs qui ont qualifié la livraison en temps voulu du cadre  $j$  (timeline)

$U$ : Total de personnes interrogées

Méthode de calcul:

$$I_6 = \frac{u_j}{U} * 100$$

Périodicité: annuelle

Plages de tolérance:

Critique <= 70; 70> acceptable <= 90; Satisfaisant > 90

## **Les références**

Bérard, Hélène. Pursey, Stuart. and Rancourt, Eric. *Re-thinking Statistics Canada's Business Register*. Statistics Canada. 11<sup>th</sup>, 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