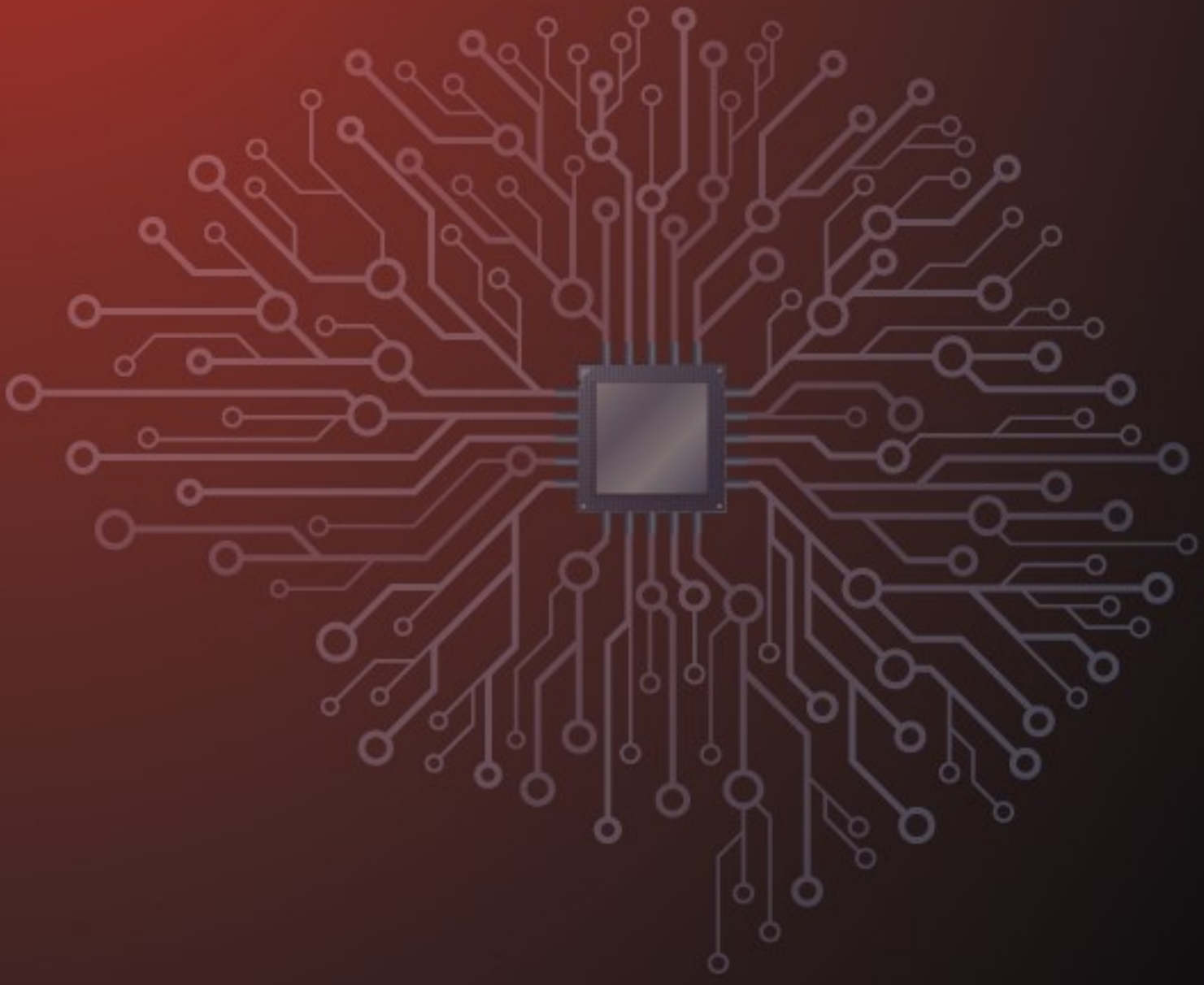
Marco conceptual y metodológico del Sistema Integrado de Registros Estadísticos y Encuestas (SIREE)

Diseño conceptual del Data Warehouse Estadístico del INE



Junio 2021
Montevideo - Uruguay

URUGUAY
INE
Instituto Nacional de Estadística

Marco conceptual y metodológico del Sistema Integrado de Registros Estadísticos y Encuestas – SIREE

Diseño conceptual del Data Warehouse Estadístico del INE

Junio 2021

Montevideo – Uruguay

Director Técnico

Diego Aboal

Sub-Director General

Federico Segui

Documento elaborado por

Federico Segui

Contenido

Introducción

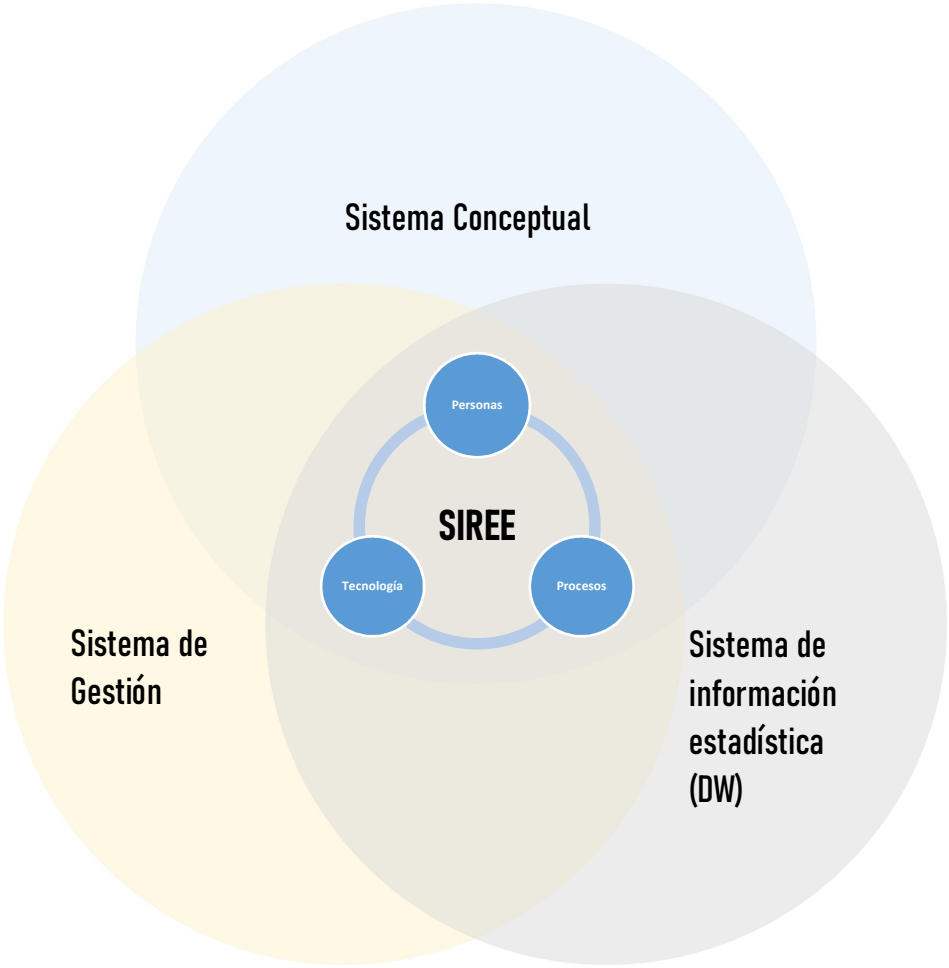
Este documento está compuesto por dos capítulos principales, el primero orientado al marco conceptual y metodológico que sustenta el Sistema Integrado de Registros Estadísticos y Encuestas (SIREE), y el segundo dirigido a los aspectos de diseño conceptual de la solución tecnológica que soporta el SIREE, a través de un Data Warehouse Estadístico. Se incluye un tercer capítulo enfocado en el rol del INE como Data Steward dentro del Sistema Estadístico Nacional.

¿Qué es el SIREE?

El SIREE es un sistema, en la acepción más amplia del término, formado por un conjunto de elementos relacionados entre sí que funciona como un todo. Es, por un lado, un sistema conceptual (conceptos, definiciones, metadatos, metodología), un sistema de gestión (procesos, administración) y un sistema de información estadística (Data Warehouse Estadístico), por otro.

El SIREE se basa en tres pilares fundamentales: personas, procesos y tecnología, alineados con la estrategia del INE.

Figura 1. Sistemas que forman parte del SIREE.



1. Marco conceptual y metodológico del SIREE

En este capítulo se desarrolla el marco metodológico y conceptual que sustenta el diseño e implementación del sistema integrado de registros estadísticos de personas, empresas e inmuebles y su integración con encuestas por muestreo.

El marco conceptual y metodológico está basado, por un lado, en la metodología de los países nórdicos que establece un sistema de registros conformado por cuatro registros base (personas, empresas, actividades e inmuebles), pero además, se ha adaptado del marco conceptual y metodológico del SIREPI de la Comunidad Andina[1].

Cabe aclarar que este documento metodológico es revisado en forma continua para detectar oportunidades de mejora.

1.1. Antecedentes

En los últimos años, ha habido una demanda creciente de información oportuna con mayor desagregación territorial, cobertura temática y periodicidad por parte de los sistemas sub-nacionales de estadística y los gobiernos locales.

Las nuevas demandas de información para la generación de indicadores de los Objetivos de Desarrollo Sostenible en muchos casos son muy difíciles de satisfacer a través de censos y encuestas por sus elevados costos de producción de los datos.

Esto ha obligado a los INE a buscar nuevas fuentes de información que permitan generar estadísticas a menores costos y a un mayor nivel de desagregación geográfica, como son los registros administrativos. Es decir, información administrativa que ya ha sido capturada por las instituciones del estado en el marco de sus obligaciones y puede ser utilizada además con fines estadísticos (previo procesamiento y transformación) a bajo costo.

El uso de registros administrativos con fines estadísticos presenta una serie de ventajas, oportunidades, desventajas y desafíos que podemos resumir en el siguiente cuadro:

[1] Segui Stagno, Federico (2016b). *Marco conceptual y metodológico que sustenta el diseño, desarrollo e implementación de un sistema integrado de registros estadísticos de población e inmuebles*. "Proyecto Estadística de Población e Inmuebles a partir del uso de registros administrativos oficiales en la Comunidad Andina". Cooperación Técnica No Reembolsable No. ATN/OC-14340-RG – Banco Interamericano de Desarrollo.

Cuadro 1. Ventajas y desventajas del uso de registros administrativos con fines estadísticos.

Ventajas	Desventajas
• Menor costo. • Menor carga para los informantes. • No hay errores de muestreo. • Mayor cobertura. • Evita duplicación de esfuerzos entre entidades públicas. • En general tienen mayor tasa de respuesta. • Es posible generar estadísticas con mayor desagregación. Estadísticas locales y subpoblaciones. • Permite fortalecer los sistemas de información nacional, a todo nivel. • Mejora la calidad de la investigación, pues permite combinar varias fuentes y realizar análisis multidimensional. • Facilita análisis longitudinales. • Facilita la construcción de series temporales. • Fuente de información de los indicadores de los ODS (Objetivos de Desarrollo Sostenible).	• Los registros administrativos (en general) no son diseñados con fines estadísticos. • Diferentes unidades de análisis entre unidades administrativas y las estadísticas. • Diferencias en definiciones y conceptos. Definiciones de variables. • Falta de documentación, rigor metodológico. • Difícil implementar cambios en los instrumentos y procesos de captura de datos de las fuentes administrativas. • Cambios políticos e institucionales pueden afectar la continuidad de los registros administrativos. • Falta de identificador común estandarizado para la integración de datos de diferentes fuentes. • Dependencia de la cooperación interinstitucional y marco legal adecuados. • Cambio cultural en los INE, resistencia al cambio, se “pierde el control” sobre el proceso de captura de datos.

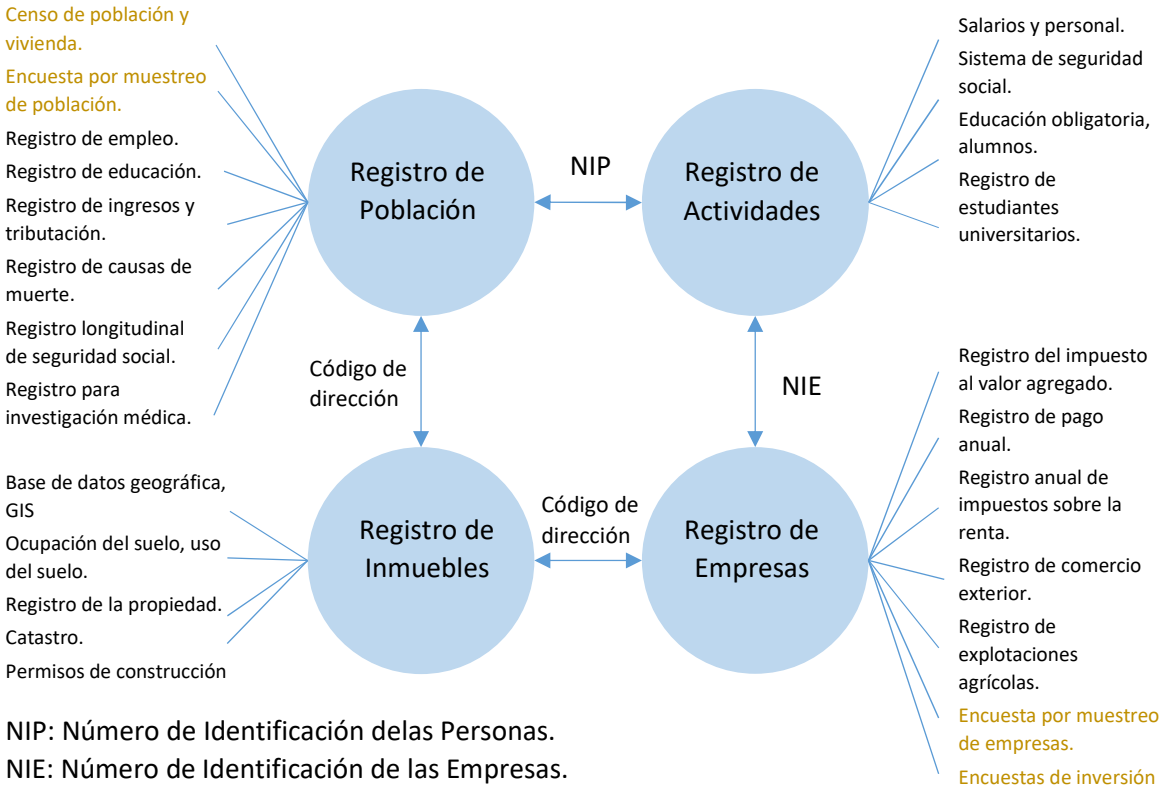
1.2. Referencias internacionales

La amplia experiencia y de larga data de los países nórdicos en el aprovechamiento estadístico de los registros administrativos, siendo los pioneros en este campo, está siendo replicada o adaptada por varios países alrededor del mundo.

En particular, la metodología de la Oficina de Estadística de Suecia sobre el uso de registros administrativos con fines estadísticos ha sido ampliamente difundida en nuestra región a través del trabajo de A. Wallgren y B. Wallgren (2012). *Estadísticas basadas en registros. Aprovechamiento estadístico de los registros administrativos*. INEGI; y es la base conceptual del presente marco conceptual y metodológico del SIREE.

El modelo de los autores Wallgren se enfoca en un sistema integrado de registros estadísticos conformado por cuatro registros base: población, inmuebles, empresas y actividades, como lo ilustra el siguiente esquema.

Figura 2. Sistema de registros estadísticos por tipo de objeto y campo de estudio.



Fuente: Anders Wallgren, Britt Wallgren. (2012). *Estadísticas basadas en registros. Aprovechamiento estadístico de los registros administrativos.* INEGI.

Entre los países que han adoptado un modelo integrado de registros estadísticos se encuentran algunas diferencias en cuanto a la cantidad de registros base que contienen. Actualmente, la oficina de estadística de Suecia cuenta con un sistema de registros conformado por tres registros base (población, empresa e inmuebles)[2] al igual que el resto de los países nórdicos, en cambio Holanda ha definido un sistema con más de diez registros base.

Lo relevante en este punto es contar con un sistema integrado de registros estadísticos compuesto por ciertos registros base que cuenten con buena cobertura y calidad.

En Uruguay, a diferencia de otros países donde existe un registro central de población (países nórdicos, Holanda, España, entre otros) y es obligatorio

[2] Klas Blomqvist and others (2011). *A strategy to improve the register system to store, share and access data and its connections to a generic statistical information model (GSIM)*. Invited paper. Work Session on Statistical Data Editing of the Conference of European Statisticians – UNECE. Ljubljana, Slovenia, 9-11 May 2011.

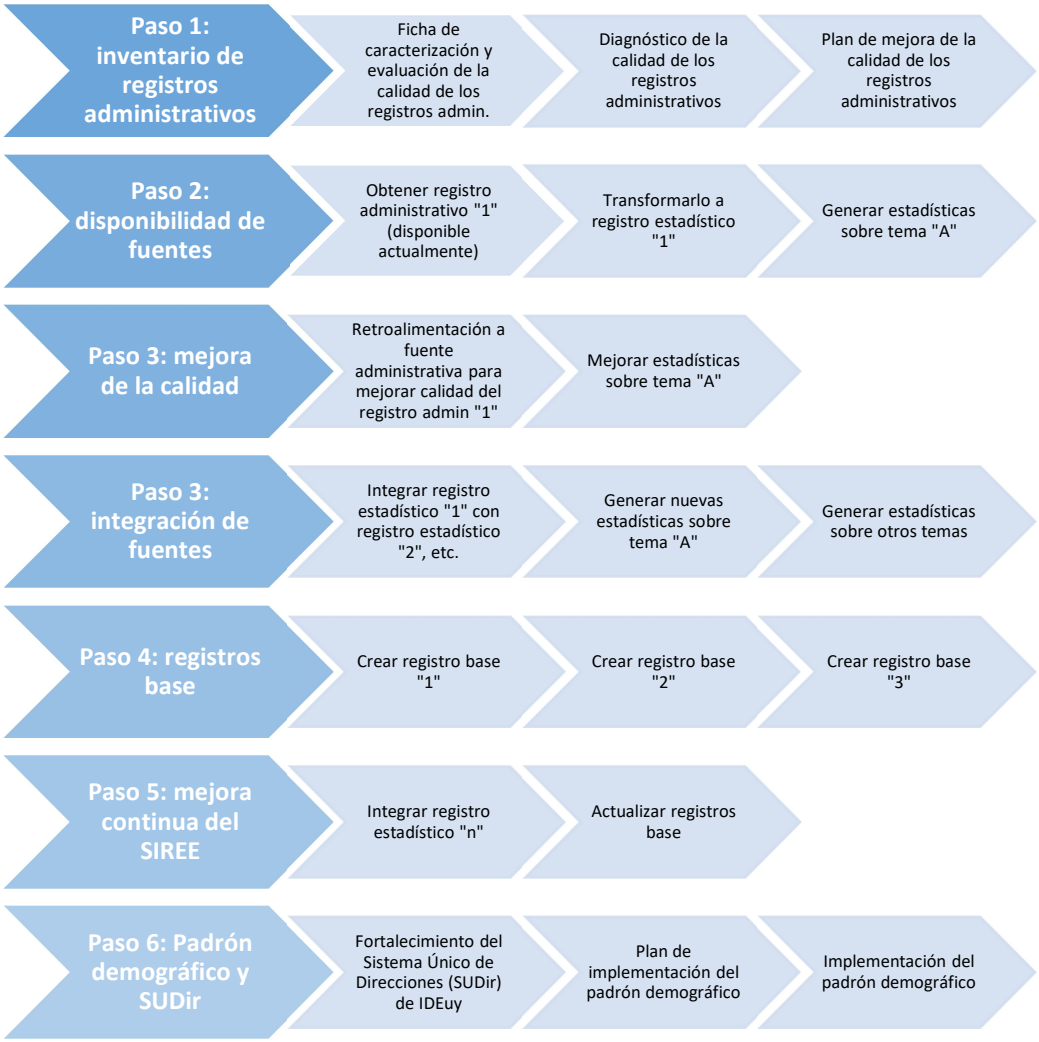
registrar cualquier cambio de domicilio, debemos partir por el criterio de unir múltiples fuentes de datos administrativos existentes para construir el registro base de población; y en paralelo avanzar hacia la construcción de un padrón demográfico nacional.

Oficinas Nacionales de Estadística como la de Nueva Zelanda y el Reino Unido están analizando la factibilidad de realizar un censo basado en registros administrativos sin contar con un registro nacional (central) de población.

1.3. Plan de acción

El enfoque es avanzar paso a paso, empezando por lo que se tiene acceso y de a poco en paralelo ir conformando los registros base, como así también lo recomienda UNECE[3].

Figura 3. Plan de acción para el aprovechamiento estadístico de los registros administrativos y la conformación del Sistema Integrado de Registros Estadísticos y Encuestas - SIREE.

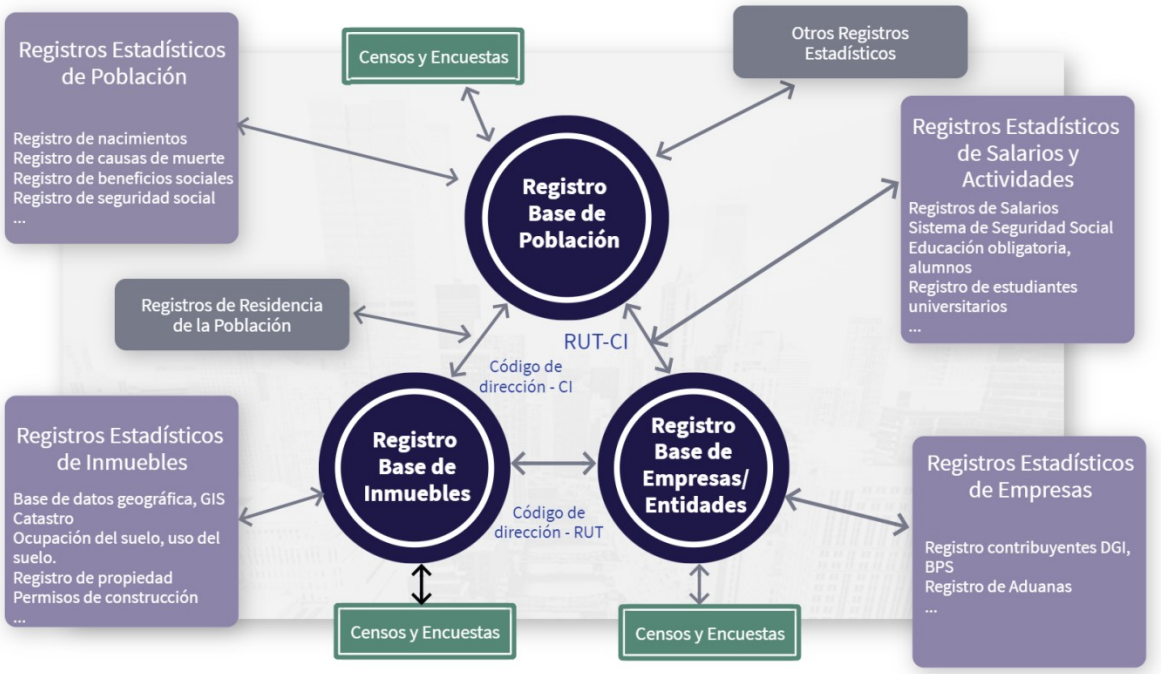


[3] UNECE (2007). *Register-based statistics in the Nordic countries. Review of best practices with focus on population and social statistics*. Naciones Unidas. Nueva York y Ginebra 2007.

1.1. Componentes del sistema

Este apartado contiene una descripción de los elementos que forman parte del sistema integrado de registros estadísticos y encuestas.

Figura 4. Elementos del Sistema Integrado de Registros Estadísticos y Encuestas - SIREE.



Objetos, ***unidades*** o ***elementos*** del sistema de registros: objetos, elementos, entidades o individuos del mundo real, ya sean personas, hogares, viviendas, empresas u organizaciones, vehículos, u otros, son los elementos que forman parte de una población.

A modo de ejemplo, una población de individuos se basa en un tipo de objeto *persona*, entonces cada individuo de la población es una instancia de ese tipo de objeto (representado por filas en un sistema de archivos o base de datos).

Población: conjunto de objetos o unidades que la componen (incluido lugar y tiempo). Se debe definir claramente del tipo de objeto (qué significa hogar, empresa, inmueble, etc.) para tener una definición clara de la población.

Población de interés o población objetivo: *"son todos los casos o unidades que forman parte del Registro Estadístico que cumplen con un conjunto de características particulares, es decir, ciertas variables tienen determinados valores en común. La población de interés puede o no coincidir con la totalidad de unidades que componen el Registro Estadístico. Es posible, entonces, definir a la población de interés como un subconjunto del Registro Estadístico, esto va a depender del uso con fines estadísticos que se quiera hacer del mismo"*.

Registro estadístico: *"registro consolidado de datos estandarizados y procesados provenientes de uno o más Registros Administrativos, que originalmente no (necesariamente) fueron captados con fines estadísticos, pertenecientes a una o más fuentes de datos administrativos, pero que han sido adaptados para su uso estadístico"*.

Los ***casos*** o ***filas*** de los registros (administrativos o estadísticos) se refieren a cada uno de los objetos o unidades individuales (instancias de objetos), contienen información individualizada de cada elemento que compone el registro.

Las ***variables*** son una serie de **atributos medibles** que forman parte de los **objetos o unidades estadísticas**. A nivel de los casos o filas correspondientes a las instancias de los objetos o unidades las variables son representadas mediante **campos**. El archivo o dataset del registro representa una matriz de datos cuyas columnas son estos campos.

*"Una **variable estadística** está definida por el tipo de objeto que presenta la característica (por ejemplo, ingreso para personas e ingreso para hogares son dos variables distintas), por el método de medición y la escala aplicados, y por el momento o período a los que refiere la medición".*

1.2. Tipologías de Registros Estadísticos

El sistema integrado de registros estadísticos está conformado por una serie de registros estadísticos: *registros base, registros primarios* y *registros integrados*.

Wallgren y Wallgren[4] señalan que los ***registros base*** tienen como función definir los objetos y poblaciones del sistema de registros. Son la columna vertebral del sistema pues contienen los tipos de objetos y los vínculos más relevantes. La calidad del sistema está determinada por las definiciones de los objetos y la cobertura de los registros base.

Un registro base tiene las siguientes características (Wallgren y Wallgren):

- "*Define tipos de objetos importantes.*
- *Define conjuntos de objetos o poblaciones estandarizadas importantes.*
- *Contiene vínculos con objetos de otros registros base.*
- *Contiene vínculos con otros registros relacionados con el mismo tipo de objeto.*
- *Es importante para el sistema en su conjunto, por lo que resulta esencial que sea de alta calidad y esté bien documentado.*

[4] Wallgren, A. Wallgren, B. (2012). *Estadísticas basadas en registros. Aprovechamiento estadístico de los registros administrativos. INEGI.*

- *Es importante para el marco muestral.*

- *Se puede usar para estadísticas demográficas relacionadas con personas, actividades, inmuebles o empresas.*

  Las fechas de nacimiento y defunción deben estar presentes en el registro base para producir estadísticas demográficas."

***Registro base de población**: es el registro de todas las personas nacidas o que residen permanentemente o temporalmente en el país (adaptado de CAN[5]).*

***Registro base de inmuebles**: es el registro de los predios urbanos y rurales del país, así como de las construcciones o edificaciones y viviendas construidas dentro de ellos.*

***Registro base de empresas**: es el registro de todas las entidades comerciales y no comerciales, públicas y privadas, constituidas en el país.*

Registros primarios: se basan directamente en al menos una fuente administrativa.

Registros integrados: combinan exclusivamente información ya existente en los registros estadísticos del sistema. Los *registros longitudinales* son un caso particular de registros integrados, los cuales consolidan información de varios registros estadísticos anuales para hacer seguimiento a los mismos objetos o unidades a lo largo del tiempo.

Registros satélites o ***asociados:*** son registros que están disponibles para el INE (utilizados en determinadas ocasiones), contienen información acerca de unidades y variables de interés, y tienen las siguientes características:

- No son parte integral del sistema de registros, pero pueden ser vinculados al mismo.
- Son más limitados en alcance, pero pueden tener mayor cobertura de unidades y/o variables.
- Contienen una o más variables que no se encuentran en los registros estadísticos del sistema. Tales variables son generalmente usadas con fines de estratificación.
- Son herramientas para incorporar datos administrativos que sólo son relevantes para un subconjunto de unidades en un registro estadístico.
- Pueden contener unidades adicionales, o variables, o ambas cosas.
- Pueden construirse utilizando información de fuentes administrativas, encuestas por muestreo o una combinación de ambas.

[5] CAN (2013). Gaceta Oficial del acuerdo de Cartagena. CAN. Pág. 5.16. http://intranet.comunidadandina.org/Documentos/Gacetas/Gace2163.pdf

- En algunos casos pueden agregar, combinar o transformar variables, aunque en otros casos pueden ser más o menos idénticos a una fuente particular.
- Para garantizar que los registros satélite sean suficientemente coherentes con los registros estadísticos, puede ser útil considerar criterios adicionales, como identificadores comunes de unidades, definiciones y clasificaciones comunes. Cuanto mayor sea la coherencia, más útil será el registro satélite.
- No son registros satélites las bases de datos o archivos de datos en los cuales son normalmente almacenados los resultados de las encuestas por muestreo.

1.3. Proceso de transformación de registros administrativos en registros estadísticos

El registro administrativo, como en general no ha sido concebido para fines estadísticos, debe transformarse en un registro estadístico. Esto implica controles de consistencia y calidad, depuración de datos, estandarización de variables, creación de variables agregadas/derivadas, creación de objetos o unidades derivadas, unión de registros, documentación de metadatos y el aseguramiento de la calidad durante todo el proceso.

Figura 5. Proceso de transformación de registros administrativos a registros estadísticos.

1.3.1. Controles de consistencia de los datos

En la mayoría de los casos, en los registros administrativos no se aplican las buenas prácticas, recomendaciones o estándares metodológicos requeridos para producir estadísticas, pues no han sido diseñados con este fin.

Las fuentes administrativas son responsables de ejecutar los procesos de captura de datos de los registros administrativos y, en general, el INE no tiene control

sobre estos procesos, su conocimiento acerca de los mismos es escaso, y la documentación es exigua.

Además, los procesos de ingreso de datos, validación y control de consistencia de la información que aplican las fuentes administrativas están más orientados y se aplican con más rigurosidad sobre las variables administrativas de interés de cada institución, que las variables de interés estadístico para el INE, pues podrían revestir menor importancia para las fuentes administrativas y éstas no siempre aplican los mismos criterios que utilizaría el INE en esos procesos de validación de datos.

Asimismo, en cualquier proceso de captación de datos, ya sean registros administrativos, censos o encuestas, se generan errores tanto en los datos que reporta el informante, como quien registra (encuestador o administrativo, si corresponde), en el ingreso de datos (registro en el formulario, digitación, escaneo, etc.), o en la depuración de datos.

Todos estos elementos provocan errores en los datos capturados en los registros administrativos, por lo que es necesario implementar controles de consistencia y calidad para detectarlos y proceder con el proceso de depuración de datos para minimizarlos.

Los errores pueden ocurrir tanto en las variables como en las unidades estadísticas, por lo tanto los controles de calidad y consistencia de los datos deberán aplicarse a ambos niveles.

A partir de las definiciones (metadatos) de las variables estandarizadas (ver siguientes apartados) se deben establecer en un documento los controles de consistencia y calidad (reglas de validación o consistencia) y definir el plan de depuración de los datos. Los resultados de la validación y depuración también deben ser documentados.

1.3.2. Variables del sistema

1.3.2.1. Categorías de variables

Es fundamental comprender la utilidad e importancia dentro del sistema de registros de las diferentes categorías de variables que conforman los registros estadísticos.

Categoría de variable	Descripción
Variables **clave de identificación** o variables identificadoras:	Como su nombre lo indica, se usan para identificar objetos o unidades. Estas variables se usan para hacer la unión entre registros (método determinístico) con el mismo tipo de objetos o unidades. En general se trata de variables numéricas que representan un código de identificación, pero también pueden ser claves alfanuméricas. En el apartado 1.3.3 se presentan diferentes métodos de unión de registros (métodos probabilísticos) cuando no se dispone de una clave identificadora estandarizada o es de mala calidad y se debe recurrir a otras variables como nombres y direcciones.
Variables **de contacto** (identificadores explícitos):	El nombre, dirección, teléfono, correo electrónico se usan cuando el INE necesita contactarse con el objeto o unidad al que corresponde el caso (en general cuando se utilizan cuestionarios). Pero también pueden utilizarse como llaves para la unión de registros por métodos probabilísticos, como se ha mencionado en el párrafo anterior.
Variables de **ubicación geográfica**:	Son utilizadas para asociar los objetos a ubicaciones físicas en el territorio. Entran en esta categoría las coordenadas geográficas determinadas por GPS y las variables codificadas correspondientes a nomenclátores de divisiones político-administrativas-geográficas del territorio nacional.
Variables de **unión o fusión** (claves foráneas):	Son utilizadas para hacer la unión con otras tablas de una base de datos, describen relaciones entre diferentes tipos de objetos. A diferencia de las claves o llaves de identificación, que tienen una correspondencia uno a uno cuando se hace la unión entre registros (salvo que existan duplicados en alguna de las tablas de los registros), las claves foráneas mantienen una relación uno a muchos. Por ejemplo, supongamos que la tabla correspondiente al registro de población tiene una variable de fusión o clave foránea para determinar en qué vivienda reside cada persona, esta variable está vinculada con la variable llave de identificación del registro de viviendas. Entonces la tabla

	del registro de viviendas tiene una relación uno a muchos con la tabla del registro de personas, a través de la variable código de vivienda, pues en una vivienda residen muchas personas[6].
Variables **estadísticas**:	Son las variables provenientes del registro administrativo de interés estadístico, que se utilizan para generar estadísticas o hacer análisis estadístico. Se clasifican en variables cuantitativas, continuas (escalares), como por ejemplo edad, ingresos, superficie, ventas, etc.; o variables cualitativas, categóricas (nominales u ordinales), como estado civil, sexo, tipo de vivienda, etc. También entran en esta categoría las variables de estratificación y de ubicación geográfica, que pueden ser a su vez clasificadas como variables estadísticas cuantitativas o cualitativas, como por ejemplo el código compuesto de ubicación geográfica de los inmuebles, código de rama de actividad, ingresos de las personas, volumen de ventas de las empresas o el personal ocupado.
Variables de **uso administrativo** interno:	Se utilizan para indicar resultados de procesos internos como por ejemplo códigos de error de las variables estadísticas, valores editados, resultados de la depuración de variables, código de la fuente administrativa de las variables, código de la fuente o versión del código de clasificación de las variables codificadas (CIUO-08, CIIU-Rev4, etc.), fecha de actualización de las variables, estado de los objetos o unidades (personas: Activo, Inactivo por emigración, Inactivo por muerte, Eliminado por proceso administrativo, Inactivo por duplicado; viviendas: Activa, Inactiva por demolición, Eliminada por proceso administrativo, Inactiva por duplicado), entre otros.
Variables de **referencia temporal**:	Indican el momento en que ocurre un evento relativo a los objetos o unidades del registro. Por ejemplo, la fecha de creación de una empresa (inscripción en el registro de la oficina de impuestos, por ejemplo), fecha de construcción/habilitación del inmueble, período al que corresponden los ingresos de una persona (ingresos del último año o del mes anterior), fecha de nacimiento, fecha de cambio de residencia de las personas. Se utilizan para generar diferentes versiones del registro estadístico de acuerdo a los períodos establecidos por esta variable, (por ejemplo, stock de viviendas en determinado período) y para describir la evolución de determinados fenómenos a lo largo

[6] Una misma persona podría ser propietario o inquilino de más de una vivienda, pero para el ejemplo se ha considerado la residencia habitual (en el caso de personas que residen en varias viviendas, se considera una sola de ellas).

	del tiempo.
Variables **derivadas o agregadas**:	Son variables que no están presentes en ningún registro administrativo pero son necesarias para facilitar la producción de estadísticas. El apartado 1.3.2.3 describe el proceso de creación de variables derivadas y las diferentes formas de hacerlo.

1.3.2.2. Estandarización de variables

Los conjuntos de datos de los registros administrativos están conformados por variables que, en la mayoría de los casos, no se corresponden con los conceptos y definiciones estadísticas, pues no han sido concebidas con estos fines.

Los instrumentos de captura utilizados por las fuentes administrativas responsables de los registros administrativos, en general, no siguen los estándares, recomendaciones o buenas prácticas de los instrumentos diseñados con fines estadísticos. Por ejemplo, las categorías de respuesta de las variables no se ajustan a los clasificadores/codificadores utilizados habitualmente por el INE.

Las definiciones de las variables de registros administrativos se ajustan a las necesidades administrativas y no siempre se corresponden con las definiciones estadísticas de acuerdo a un uso estadístico específico.

Es así que resulta imprescindible estandarizar las variables del registro administrativo para incorporarlas al sistema de registros estadísticos y poder utilizarlas en diferentes proyectos estadísticos.

La estandarización de variables del registro corresponde al INE, quien es responsable de los metadatos de las variables (denominación, definición, categorías, calidad y documentación).

El nombre o identificación de la variable debe ser único dentro del sistema de registros estadísticos.

Las variables estandarizadas no deberán ser modificadas por otras áreas del INE, puesto que cualquier cambio en la denominación o definición de las variables estandarizadas podría tener un impacto en otros proyectos estadísticos que estén utilizando dichas variables, provocando además que dejen de ser variables estandarizadas.

El hecho de contar con variables estandarizadas minimiza errores, evita duplicidad de esfuerzos de documentación (quienes utilizan las variables estandarizadas no necesitan volver a documentarlas) y facilita su utilización.

Además, la estandarización de variables a través de sus metadatos, facilita la comparación con variables de nuevos registros administrativos que se deban incorporar al sistema. Mediante esta comparación entre definiciones, categorías y

documentación es posible especificar el procedimiento necesario para la transformación de las nuevas variables en variables estadísticas estandarizadas o la creación de variables agregadas, que también estarán estandarizadas.

En el INE se ha adoptado el estándar de metadatos ISO/IEC 11.179 Metadata Registry[7] para la documentación de las variables estandarizadas.

Es importante mantener copias de los formatos o cuestionarios de captura de datos, manuales y otro tipo de documentos utilizados por la fuente administrativa para capturar la información del registro administrativo. Asimismo, se deben transcribir las instrucciones de llenado de los instrumentos de captura de datos.

La fuente de datos también debe registrarse en la ficha de metadatos de las variables estandarizadas, indicando el nombre de la variable o variables administrativas originales utilizadas para generar la variable estandarizada, nombre del registro al cual pertenece(n) y la fuente administrativa.

En el caso que la variable estandarizada requiera de algún proceso de transformación, recodificación o creación (variables agregadas) a partir de una o más variables del registro administrativo, deberá indicarse en el apartado de "Recodificación y derivación".

La estandarización de variables implica además, la transformación de formatos de datos y en algunos casos los datos en sí mismos. Por ejemplo, la variable fecha de nacimiento de un registro administrativo almacena los datos en el formato dd/mm/aaaa, pero la variable estandarizada fecha de nacimiento del registro estadístico de población se ha definido con el formato aaaa-mm-dd, lo cual implica que los formatos de datos originales de la variable del registro administrativo deben transformarse (siguiendo un algoritmo) para convertirlos al formato estandarizado de la variable del registro estadístico.

Otro caso es el de las variables categóricas donde los códigos de categoría y sus descripciones podrían cambiar al transformarlos a las correspondientes variables estandarizadas del registro estadístico.

Supongamos que una variable categórica del registro administrativo como tipo de material de los pisos de la vivienda tiene las siguientes categorías de respuesta: 1-Tierra; 2-Arena; 3-Cemento; 4-Madera o tabla; 5-Otro material. Pero su correspondiente variable estandarizada en el registro estadístico tiene las siguientes categorías estandarizadas: 1-Tierra o arena; 2-Cemento; 3-Madera o tabla; 4-Otro material. El proceso de transformación o estandarización implica que los códigos de la variable sean convertidos de la siguiente manera: 1 y 2 ->1; 3->2; 4->3; 5->4. Los códigos originales de la variable administrativa son 1 y 2 que se han unido en un solo código (1) en la variable estandarizada y eso provoca que también el resto de los códigos deban ser transformados.

Todos estos procesos de transformación o estandarización deben ser documentados en la ficha de metadatos de las variables estandarizadas.

[7] International Standard ISO/IEC 11179 – Information Technology – Metadata Registries (Parts 1 – 6)

MasterData como herramienta de Gestión de Datos Maestros y Metadatos:

El INE ha incorporado una herramienta de gestión de metadatos denominada *MasterData* basada en el estándar de metadatos ISO 11.179, que permite documentar esta asociación entre los catálogos o categorías de las variables y además es un sistema completo de documentación de metadatos del sistema de registros estadísticos.

1.3.2.3. Variables derivadas o agregadas

En la generación de estadísticas basadas en registros administrativos no se tiene la posibilidad de diseñar los instrumentos de captura de datos, no es posible definir las preguntas del cuestionario. En estos casos se debe apelar a la creación de variables agregadas utilizando las variables disponibles de los registros administrativos.

Las variables derivadas o agregadas se crean de seis modos diferentes:

1) Por medio de **cálculos aritméticos** o **procedimientos lógicos**.

   En el caso de las variables cuantitativas se pueden hacer operaciones aritméticas o lógicas para generar variables derivadas. Las variables cualitativas sólo permiten hacer operaciones lógicas. Por ejemplo, el área total construida de un inmueble se calcula como la suma de la superficie de cada una de las plantas.

2) Mediante **agrupamiento** de valores.

   Se crean agrupando valores en determinados intervalos. Por ejemplo, la ocupación de una persona se podría agrupar en una nueva variable por clases de ocupación.

3) Mediante **codificación**.

   En el caso de variables cuyos datos están almacenados en forma de texto no estructurado pues corresponden a preguntas abiertas, como por ejemplo ocupación, rama de actividad, estudios cursados, etc., es necesario convertirlas a datos estructurados mediante un procedimiento de codificación.

4) Por **asociación** de otras variables.

   Se calcula la variable derivada usando variables de otro registro. Por ejemplo, se pueden vincular o asociar variables de la vivienda a las personas como ser el tipo de vivienda donde reside cada persona, uniendo ambos registros (personas y viviendas) a través de variables (clave) comunes.

5) Mediante **agregación** de otras variables.

   La variable derivada se genera en un registro usando variables de otro(s) registro(s). Se utilizan operaciones aritméticas (suma, promedio, etc.) para agregar los valores de la variable del segundo registro (vinculado) y el resultado se almacena en la variable derivada para cada objeto vinculado. Por ejemplo, una variable derivada que tenga el promedio de horas trabajadas por las personas en cada empresa; o el ingreso total de los hogares o viviendas a partir de los ingresos de las personas que las integran.

6) Mediante **modelos estadísticos**.

   Se diseña un modelo estadístico mediante el cual se analiza la relación entre la variable derivada que se desea crear y las variables administrativas, y permite crear la variable en cuestión.

Las variables derivadas se pueden generar combinando estos métodos, incluso aplicando cálculos más complejos con procedimientos basados en ciertas reglas.

Estas variables, al igual que el resto, también deben ser documentadas de acuerdo al estándar de documentación definido para tales fines. Se debe hacer especial hincapié en describir detalladamente la fórmula o regla de cálculo utilizada para crear la variable.

1.3.2.4. Mapeo de variables

Mantener la trazabilidad de cada variable que conforma el registro estadístico es fundamental para facilitar el mantenimiento y actualizaciones del SIREE, dar transparencia al proceso y para fines de auditoría.

Se debe asegurar el mapeo de las variables desde su origen en el registro administrativo, donde fue creada por la respectiva fuente administrativa, pasando por el proceso de validación, depuración y estandarización y justificación de su selección e inclusión en el registro estadístico.

El sistema de gestión de metadatos *MasterData* asegura la trazabilidad de todas las variables del sistema de registros estadísticos con sus respectivas fuentes administrativas. Además, existe una vinculación entre el inventario de registros administrativos y los metadatos del SIREE.

El nombre o identificación de la variable (único dentro del sistema de registros estadísticos, como se ha mencionado anteriormente) es la clave de vinculación con las variables administrativas.

Las variables derivadas del registro estadístico se crean combinando variables, por medio de fórmulas, cuyas fuentes pueden ser otras variables administrativas o estadísticas, o la combinación de ambas, y deben ser especificadas en la ficha de metadatos

A continuación se presentan algunos ejemplos:

Cuadro 2. Ejemplo de ficha de metadatos.

FICHA DE METADATOS		
Nuevas Variables del Registro Estadístico	**Origen de los datos**	
	Variables del Inventario de Registros Administrativos	**Variables del Registro Estadístico**
Nivel_Educativo	MEC-RA1-Nivel_edu	
Genero	MSP-RA1-Genero	
Ingresos_Persona	SegSocial-APORTES-Salario; SegSocial-APORTES-HrsExtra; SegSocial-PENSIONES-Pension	Ingresos_Salario Ingresos_HrsExtra Ingresos_Pension

En la ficha de metadatos se debe indicar el criterio de selección de la(s) fuente(s) para crear cada variable del registro estadístico, como se establece en el siguiente apartado.

1.3.2.5. Selección de variables y fuentes del Registro Estadístico

Las variables que conformarán el registro estadístico son seleccionadas de acuerdo a las necesidades de información de los usuarios y los requerimientos técnicos-metodológicos (uso de variables clave para la unión de registros, variables de identificación, validación de datos, para mejorar el tiempo de respuesta de las consultas a la base de datos, entre otros).

Cada una de las variables seleccionadas para el registro estadístico son creadas a partir de una o más variables (combinándolas mediante fórmulas u otros cálculos) provenientes de diversos registros administrativos o de una única fuente.

Estas fuentes de datos, es decir las variables de los registros administrativos que generan las variables del registro estadístico, deben ser seleccionadas de acuerdo a criterios y priorizaciones preestablecidos.

Estos criterios deben estar asociados a la calidad de los datos (tasa de respuesta de la variable, reglas de integridad y consistencia de los datos, cobertura del registro y de la variable) y disponibilidad total o parcial del registro administrativo (si la fuente administrativa responsable del registro no lo entrega al INE o no habilita a éste para acceder a los microdatos del registro, o alguna de las variables del registro no está disponible).

Cualquier cambio en los criterios de selección debido a excepciones o condiciones particulares deberá ser justificado y documentado.

A los efectos de documentación, transparencia y organización del trabajo se debe utilizar una matriz de selección de variables por registro administrativo similar al ejemplo que se presenta a continuación.

Cuadro 3. Matriz de selección de variables por registro administrativo según prioridad (1 – alta, 2 – media 3 – baja).

	VARIABLES DEL REGISTRO ESTADÍSTICO											
REGISTRO ADMINISTRATIVO	Tipo de docum.	Número de documento	Nombres y apellidos	Fecha de nacim.	Sexo	Tel.	Edad	Nacio-nalidad	Estado civil	Domi-cilio	Fec. Fallecim.	Fec. Emi-gró
Registro de cédulas de identidad – DNIC.	1	1	1	1	1			1	1	3		
SIIAS – MIDES.	3	3	3	3	3	1	1	3	3	2	2	
Registro Único de Cobertura de Asistencia Formal (RUCAF) – MSP.	2	2	2	2	2	2				1		
Registro de nacimientos. Certificado electrónico de nacido vivo – MSP.	3	3	3	3	3	2	2		2			
Registro de defunciones – MSP.											1	
Registro de entrada y salida de personas del país. Migraciones.												1

Según el ejemplo anterior, la variable *Fecha de Nacimiento* del registro estadístico se creará a partir del dato de la misma variable proveniente del registro administrativo *Registro de cédulas de identidad* de la Dirección Nacional de Identificación Civil (DNIC). Es decir, se le dará más alta prioridad a dicha fuente (pues según el ejemplo la variable proveniente de esa fuente tiene prioridad de selección = 1), pero en el caso que alguna fila/caso de ese registro administrativo no disponga del dato se deberá tomar la variable proveniente del registro RUCAF (pues tiene prioridad = 2), y así sucesivamente de acuerdo al orden de prioridad de selección establecido previamente.

El software *MasterData* permite registrar y visualizar la información contenida en la matriz anterior de una forma amigable para el usuario.

1.3.3. Unión de registros

Según Fellegi y Sunter *record linkage* o unión de registros "es una solución al problema de identificar aquellos registros (filas o casos) en dos archivos que representan personas, objetos o eventos idénticos"[8].

[8] Fellegi I.P., Sunter A.B. (1969) *A theory for record linkage*. Journal of the American Statistical Association 64, 1183-1210. USA.

La unión de registros es la tarea de identificar de forma rápida y precisa los registros correspondientes a la misma entidad/objeto/individuo de una o más fuentes (archivos) de datos[9].

En Uruguay tenemos registros administrativos de buena calidad que, en general, cuentan claves de identificación de los casos. Además, estas claves están estandarizadas y son utilizadas por todos los registros administrativos referidos al mismo tipo de objetos o elementos, como en el caso de los registros de población que utilizan la cédula de identidad. De esta forma se podrían unir filas de diferentes registros que refieren al mismo objeto o elemento, en los casos que coincida exactamente la clave de identificación de ambos registros (método determinístico).

Sin embargo, existen casos de registros administrativos que no utilizan una clave única estandarizada común para identificar los casos, sobre todo en los registros de inmuebles.

En las situaciones donde sí se utiliza una clave de identificación común, ésta podría presentar ciertos problemas, como por ejemplo: contener duplicados, o la estandarización no es de alcance nacional entonces cada institución tiene su propio código y procedimiento para crearlo.

Si se aplica el método determinístico de unión de registros sin evaluar la calidad de la variable clave utilizada, se corre el riesgo de unir filas o casos de ambos registros que no se tiene la certeza que correspondan realmente al mismo objeto o elemento. Asimismo, si se tienen duplicados no se sabrá cuál de las filas es la que corresponde efectivamente al elemento u objeto en cuestión. Además, están los casos cuya clave de identificación está en blanco o contienen datos inválidos y no será posible unirlos con otros registros.

Por estas razones se deben plantear métodos alternativos para la unión de registros. Los métodos probabilísticos de unión de registros utilizan algoritmos específicos para determinar con cierta certeza (probabilidad) que dos filas o casos de diferentes registros corresponden al mismo elemento u objeto. Estos métodos utilizan otras variables del registro (aparte de las variables clave de identificación), combinándolas para lograr una pseudo-clave. Las variables de población más comúnmente utilizadas para estos propósitos son nombre, apellido, fecha de nacimiento o edad y otras dependiendo de la disponibilidad en el registro. En el caso de inmuebles se utiliza la variable de dirección del inmueble.

Las variables alfanuméricas que podrían ser utilizadas para hacer la unión de registros presentan una serie de problemas:

- Errores, variaciones y datos en blanco.

[9] Baxter, R. Gu, L. Vickers, D. Rainsford, C. *Record Linkage: Current Practice and Future Directions*. CSIRO Mathematical and Information Sciences. CMIS Technical Report No. 03/83. Canberra, Australia.

- Diferencias en definiciones, períodos, formatos de los datos capturados por los diferentes registros administrativos.
- Cambios en los datos a lo largo del tiempo, por ejemplo cambio de domicilio de las personas.
- Errores de digitación, letras o palabras ingresadas en diferente orden.
- Palabras fusionadas o divididas, palabras incompletas, letras faltantes o excedentes.
- Puntuación o acentuación incorrecta.
- Abreviaciones.

Por lo tanto es imprescindible realizar la depuración de datos y estandarización de variables antes de iniciar el proceso de unión de registros.

Los métodos de unión probabilística de registros implican el cálculo de pesos de unión estimados con base en todos los casos coincidentes y no coincidentes observados de los valores de la(s) variable(s) utilizada(s) para la unión.

Los métodos probabilísticos permiten obtener una mejor unión de registros que un simple método de unión determinístico. Además, pueden ser utilizados para detectar casos/filas duplicados en un archivo de datos del registro administrativo (cuando dos o más casos/filas tienen diferentes valores en la clave de identificación, pero en realidad corresponden a la mismo objeto/elemento/individuo).

Los métodos probabilísticos de unión de registros se describen en detalle en el *Anexo I - Métodos probabilísticos de unión de registros*.

Caso particular: el problema de la falta de identificador común entre personas y viviendas, e incluso entre los registros de inmuebles donde tampoco se cuenta encuentra una clave de identificación estandarizada. En su lugar se deben utilizar las variables de domicilios/direcciones.

1.5. La gestión por procesos en la producción de estadísticas a partir de registros administrativos

El siguiente cuadro presenta las principales diferencias entre la gestión tradicional centrada en la organización funcional y la gestión orientada a los procesos del INE (enfoque moderno).

Cuadro 4: cuadro comparativo entre la Gestión centrada en la organización funcional del INE y la Gestión centrada en los procesos del INE.

Gestión centrada en la organización funcional del INE	Gestión centrada en los procesos del INE
Las relaciones interpersonales constituyen la fuente de los problemas de la organización	Los procesos inadecuados son el problema de gestión del INE
Evaluar el desempeño individual	Evaluar el rendimiento de los procesos
Cambiar estructura, personas, funciones, objetivos, etc.	Cambiar los procesos
Orientación a productos y costos	Orientación a resultados, creación de valor y satisfacción de los usuarios internos y externos del SEN
Siempre se puede hallar un funcionario mejor	Siempre se puede mejorar los procesos, “reducir su variabilidad” (W.E. Deming)
Controlar empleados	Desarrollo de las personas, gestión participativa. Auto-control; auto-medida y auto-supervisión
Responsabilidad fragmentada de las tareas de la organización	Responsabilidad colegiada de los resultados finales de la actividad central del INE
Optimizar funciones con atención a la eficiencia	Optimizar procesos con atención a la creación de valor
Relación jerárquica entre superior y subordinado	Relación proveedor-cliente
Enfrentar y corregir errores	Reducir la variación de los procesos
Consigna a los funcionarios a realizar su trabajo	Comprensión del lugar que ocupa el trabajo de cada funcionario en el proceso y colaboración en el desarrollo de los procesos en que participa

Fuente: Medina, Alejandro y Segui, Federico (2013). *¿Cómo mejorar el desempeño y crear valor público en las oficinas y sistemas estadísticos nacionales en América Latina y el Caribe?* USA: innovacionestadistica.com

El cuadro anterior presenta una serie de ventajas en cuanto a la implementación de un modelo de gestión por procesos en la producción estadística. El modelo de gestión basada en procesos está enfocado al cambio operacional de la organización, dejando de lado la operación y organización funcional para pasar a

una gestión orientada a los procesos, siendo éste un aspecto fundamental para la implementación de un sistema integrado de registros estadísticos y encuestas.

El INE ha adoptado la metodología *BPM* (por su sigla en inglés de *Business Process Management* o Gestión de Procesos de Negocio/Misionales) para la gestión de procesos misionales relativos a la producción estadística a partir de registros administrativos.

La metodología *BPM* tiene por objetivo mejorar el desempeño de la organización por medio de la optimización de sus procesos de negocio o procesos misionales. Provee un modelo de gestión flexible que permite a las organizaciones adaptarse rápidamente y con el menor impacto posible a los cambios que cada vez son más frecuentes en la actualidad.

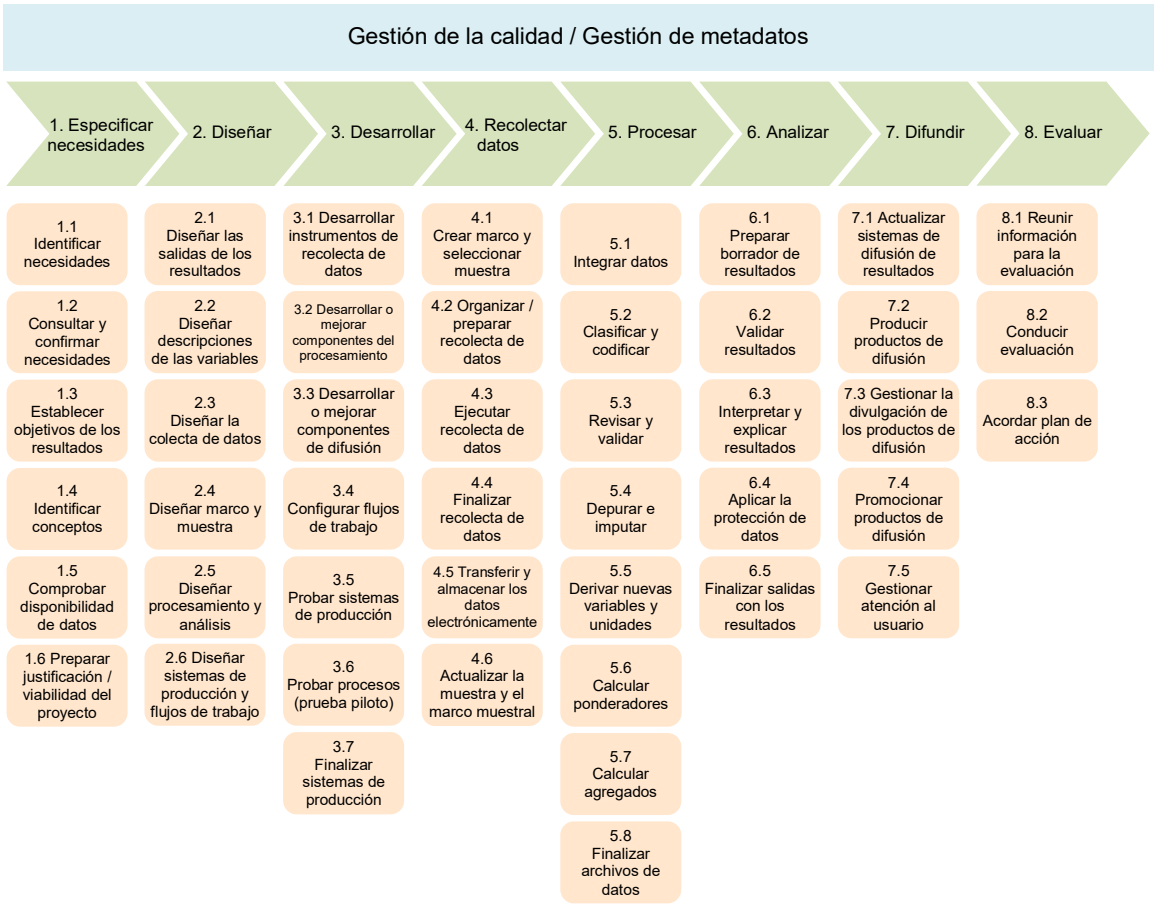
1.4.1. GSRBPM – Modelo Genérico de Procesos de Producción de Registros Estadísticos (adaptado de GSBPM - UNECE)

UNECE desarrolló el modelo *GSBPM*[10] (*Generic Statistical Business Process Model*) que establece, a través de un marco de referencia estándar y terminología armonizada, las fases y procesos misionales necesarios para elaborar estadísticas oficiales, ayudando a los INE a modernizar sus procesos de producción de estadísticas.

El *GSBPM* establece una secuencia de pasos dentro del proceso de producción de estadísticas, pero como no es un modelo rígido, los pasos no tienen por qué seguir el estricto orden preestablecido. Sino que se pueden realizar en diferente secuencia dependiendo del contexto y repetir los mismos pasos en caso de ser necesario.

[10] UNECE (2013). *Generic Statistical Business Process Model – GSBPM v.5.0*

Figura 6. Modelo Genérico de Procesos Misionales de Producción Estadística – GSBPM v.5.0.



Fuente: UNECE (2013). *Generic Statistical Business Process Model – GSBPM v.5.0*.

Según se indica en el "*Marco conceptual y metodológico que sustenta el diseño, desarrollo e implementación de un sistema integrado de registros estadísticos de población e inmuebles. Proyecto Estadística de Población e Inmuebles a partir del uso de registros administrativos oficiales en la Comunidad Andina"* (Segui Stagno, Federico [2016b]), el modelo *GSBPM* de UNECE está orientado a los procesos de generación de estadísticas a partir en censos, encuestas y registros estadísticos pero tienen un inicio y un fin (incluso cuando la investigación se realiza en forma continua, cada ciclo inicia y termina en un período determinado). En cambio, los procesos de producción (creación y actualización) de registros estadísticos que forman parte de un sistema integrado de registros son procesos continuos, independientemente de los usos estadísticos en diferentes proyectos o estudios (obviamente están vinculados a los procesos de los proyectos estadísticos).

Los expertos Anders y Britt Wallgren también hacen una crítica del modelo GSBPM de UNECE en cuanto a su aplicación en un modelo de producción de estadísticas basado en un sistema integrado de registros.

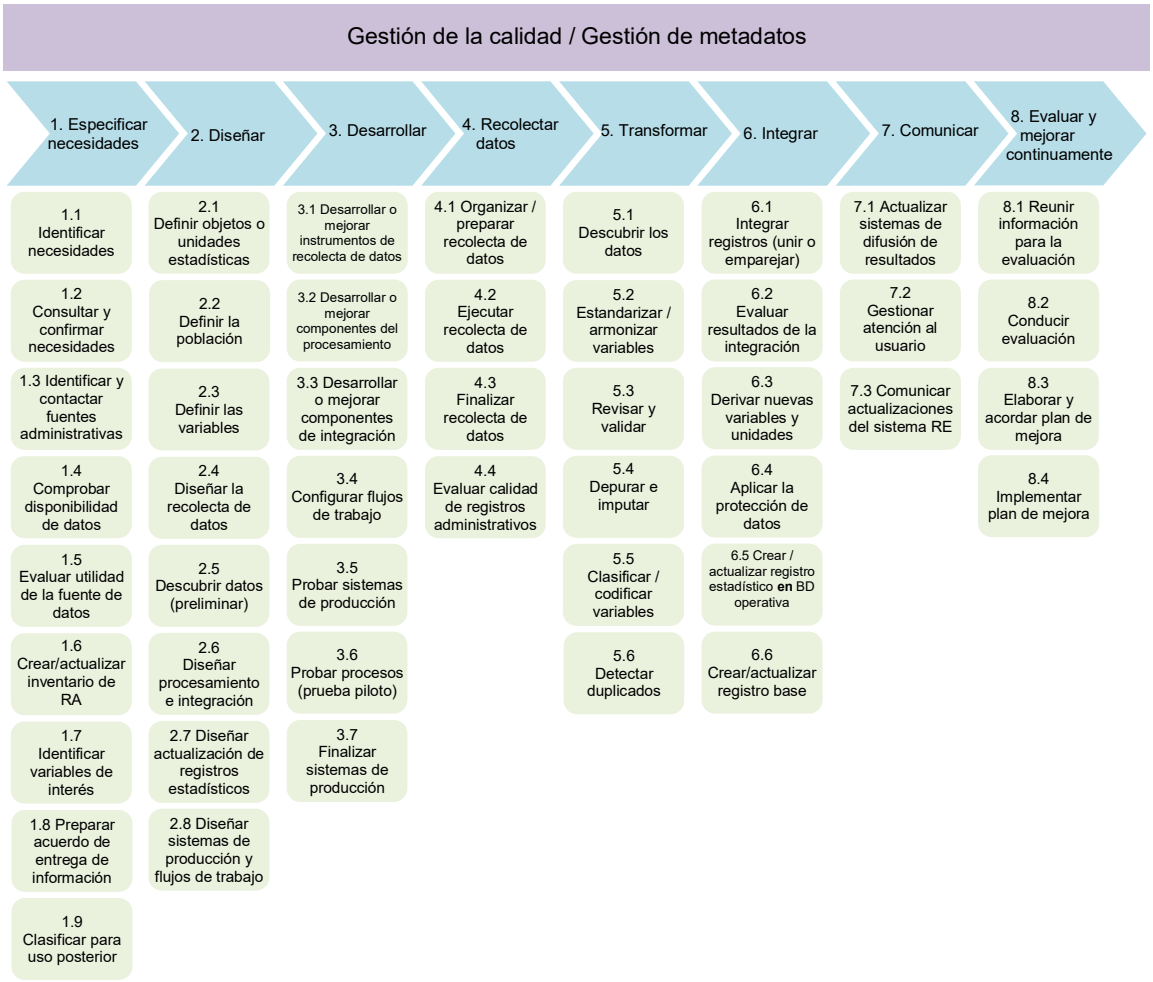
"*El modelo del proceso es aplicable a encuestas por muestreo y censos, pero no resulta de ayuda para el enfoque sistémico que es esencial en las encuestas por registro. El importante trabajo de crear, mantener y usar el sistema de registros no se incluye en el modelo del proceso. El modelo ilustra que el enfoque*

sistémico no se necesita en un sistema de producción tradicional sin registros, ya que es suficiente pensar en una encuesta por muestreo a la vez"[11].

Es por esta razón que en el mencionado marco conceptual y metodológico se ha hecho una **adaptación del modelo GSBPM** de UNECE y se ha creado el *Modelo Genérico de Procesos de Producción de Registros Estadísticos* (**GSRBPM** – Generic Statistical Registers Business Process Model).

La siguiente figura representa el nuevo **modelo GSRBPM** desde el punto de vista de la creación/actualización de registros estadísticos que forman parte del sistema integrado de registros estadísticos.

Figura 7. Modelo Genérico de Procesos de Producción de Registros Estadísticos (GSRBPM).

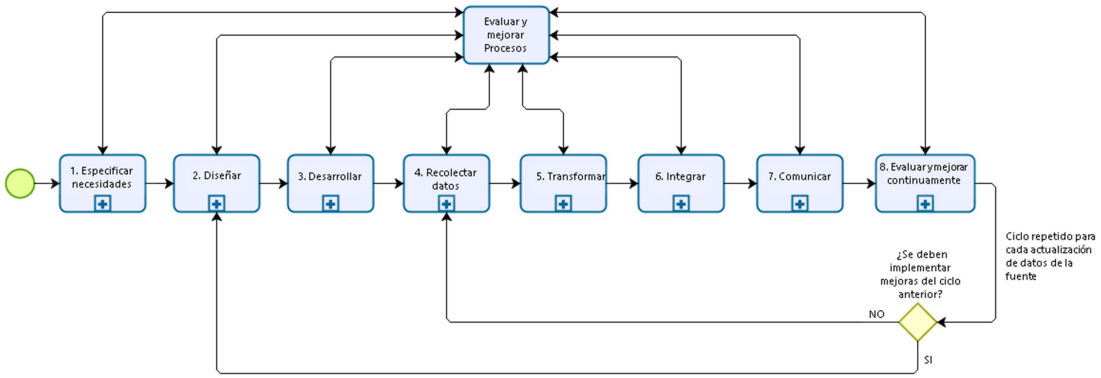


Fuente: Segui Stagno, Federico (2016b). *Marco conceptual y metodológico que sustenta el diseño, desarrollo e implementación de un sistema integrado de registros estadísticos de población e inmuebles. Proyecto Estadística de Población e Inmuebles a partir del uso de registros administrativos oficiales en la Comunidad Andina.*

[11] Wallgren, A., Wallgren, B. (2021). Hacia un sistema estadístico integrado y basado en registros. Banco Interamericano de Desarrollo BID. https://publications.iadb.org/publications/spanish/document/Hacia-un-sistema-estadistico-integrado-y-basado-en-registros.pdf

Las fases y procesos que integran el modelo GS**R**BPM no tienen por qué seguir el orden estricto de ejecución preestablecido. Sino que se pueden realizar en diferente secuencia dependiendo del contexto y ejecutar los mismos pasos formando bucles repetitivos en caso de ser necesario.

1.4.2. Descripción de los procesos y sub-procesos que forman parte del modelo GSRBPM:



A continuación se describen y especifican los procesos y subprocesos de cada fase, según se establece en el *"Marco conceptual y metodológico que sustenta el diseño, desarrollo e implementación de un sistema integrado de registros estadísticos de población e inmuebles. Proyecto Estadística de Población e Inmuebles a partir del uso de registros administrativos oficiales en la Comunidad Andina"* (Segui Stagno, Federico [2016b]):

1. Especificar necesidades

1.1 Identificar necesidades

1.2 Consultar y confirmar necesidades

1.3 Identificar y contactar fuentes administrativas

1.4 Comprobar disponibilidad de datos

1.5 Evaluar utilidad de la fuente de datos

1.6 Crear/actualizar inventario de RA

1.7 Identificar variables de interés

1.8 Preparar acuerdo de entrega de información

1.9 Clasificar para uso posterior

1. Especificar necesidades

Esta fase se ejecuta cada vez que se identifica la necesidad de nuevas estadísticas o se requiere una revisión de las estadísticas producidas actualmente. Incluye todas las actividades asociadas con la participación de los usuarios para identificar sus necesidades estadísticas, identificar las posibles fuentes de datos administrativos, crear o actualizar el inventario de registros administrativos y preparar los acuerdos de entrega de información de las fuentes administrativas.

1.1. Identificar necesidades

Este subproceso incluye la investigación inicial y la identificación de las estadísticas que se necesitan y de lo que se necesita de las estadísticas. Puede ser provocado por una nueva solicitud de información, un cambio en el contexto, como una reducción presupuestal o por la necesidad de mejorar los registros base. Los planes de acción a partir de evaluaciones de iteraciones anteriores del proceso, o de otros procesos, podrían proporcionar una entrada a este subproceso.

1.2. Consultar y confirmar necesidades

Este subproceso se centra en la consulta con las partes interesadas y la confirmación en detalle de las necesidades estadísticas. Para que el INE sepa qué se espera producir, cómo y por qué. Para las subsiguientes iteraciones de esta fase, el enfoque principal será determinar si las necesidades previamente identificadas han cambiado. Esta comprensión detallada de las necesidades de los usuarios es la parte crítica de este subproceso.

1.3. Identificar y contactar fuentes administrativas

El objetivo de este subproceso es identificar las posibles fuentes administrativas y los registros administrativos que podrían proveer de los datos necesarios para cubrir las necesidades de los usuarios identificadas y confirmadas en los subprocesos anteriores. Una vez identificadas las fuentes de datos se establece el contacto con la fuente para iniciar el intercambio de información y sobre todo establecer o mejorar el

relacionamiento colaborativo mutuo. Se confirma con la fuente de manera primaria y general si dispone en sus registros administrativos de la información requerida para satisfacer las necesidades de los usuarios. Se obtiene información y documentación sobre los registros administrativos de interés.

1.4. Comprobar disponibilidad de datos

Luego de establecer el contacto con la fuente administrativa se comprueba la disponibilidad de los datos necesarios para satisfacer las necesidades de los usuarios y las condiciones en que estarían disponibles, incluidas las restricciones sobre su uso. Se verifica con la fuente si los datos están disponibles en medios electrónicos o deben digitarse, escanearse, etc. Se constata si la fuente está en condiciones de entregar los microdatos identificados del registro administrativo. Cuando se han evaluado las fuentes existentes, se prepara una estrategia para llenar los vacíos restantes en los requerimientos de datos. Este subproceso también incluye una evaluación más general del marco jurídico en el que se recopilarán y utilizarán los datos y, por lo tanto, podrá identificar propuestas de modificación de la legislación vigente o la introducción de un nuevo marco jurídico.

1.5. Evaluar utilidad de la fuente de datos

La utilidad de los datos administrativos es evaluada en términos de cobertura, relevancia, oportunidad y calidad. En el caso que no se considere útil para los propósitos establecidos en el subproceso 1.1, se clasifica la fuente de datos para su uso posterior.

1.6. Crear/actualizar inventario de Registros Administrativos

En este subproceso se crea o actualiza el inventario de registros administrativos y las fuentes responsables. Se registra información básica sobre identificación y denominación de la fuente administrativa y los registros administrativos que gestiona, sus principales variables, cobertura, alcance temático, etc.

1.7. Identificar variables de interés

Se identifican las variables del registro administrativo necesarias para cumplir con los requerimientos estadísticos de los usuarios. Se registran no sólo las variables estadísticas, sino también las variables clave de identificación, variables de referencia temporal, variables de contacto, ubicación geográfica y variables de unión.

1.8. Preparar acuerdo de entrega de información

En este subproceso se elabora el acuerdo de entrega de información con la fuente administrativa. Allí se establece la forma de entrega, formato de los archivos y medio de entrega, periodicidad de las actualizaciones, variables y documentación necesaria.

1.9. Clasificar para uso posterior

En los casos que el registro administrativo no cumpla con los requisitos para ser utilizado con los fines estadísticos previstos o no sea de utilidad para cubrir las necesidades de información de los usuarios para el uso previsto actual, será evaluado y clasificado su potencial uso en el futuro.

Flujo del proceso de la Fase 1. Especificar necesidades:

Elementos de Entrada		Elementos de Salida
Requerimientos de información de los usuarios. Evaluaciones anteriores del proceso. Interfaz de captura de requerimientos.	1.1 Identificar necesidades	Formulario con especificaciones de requerimientos.
Formulario con especificaciones de requerimientos.	1.2 Consultar y confirmar necesidades	Requerimientos confirmados por los usuarios.
	¿Necesidades confirmadas? (NO / SI)	
Formulario con especificaciones de requerimientos confirmados.	1.3 Identificar y contactar fuentes administrativas	Lista de posibles fuentes administrativas para cubrir requerimientos. Documentación sobre los RA.
Lista de fuentes de información.	1.4 Comprobar disponibilidad de datos	Confirmación de disponibilidad de información y medio de entrega de los datos.
Documentación sobre los RA. Requerimientos de información de los usuarios.	1.5 Evaluar utilidad de la fuente	Informe de evaluación
Interfaz de actualización del inventario de RA. Documentación del RA. Datos de la fuente.	1.6 Crear/actualizar inventario de RA	Inventario RA
	¿Se puede utilizar? (NO / SI)	
Requerimientos de los usuarios. Documentación del RA.	1.7 Identificar variables de interés	Lista de variables a utilizar del RA. Los RA deben tener un mínimo de variables relacionadas a su identificación y ubicación
Modelo de acuerdo de entrega de información. Requerimientos de los usuarios. Lista de variables a utilizar del RA.	1.8 Preparar acuerdo de entrega de información	Acuerdo con la fuente administrativa
Documentación del RA.	1.9 Clasificar para uso posterior	Actualización del inventario de RA (clasificación para uso posterior)

2. Diseñar

Esta fase está compuesta por las actividades de diseño y desarrollo y cualquier trabajo de investigación práctica necesario para definir los productos, conceptos, metodologías, instrumentos de recolecta y procesos operativos. Incluye todos los elementos de diseño necesarios para definir o refinar los registros estadísticos. Esta fase especifica todos los metadatos relevantes, listos para su uso más adelante en el proceso GSRBPM, así como los procedimientos de aseguramiento de calidad. Esta fase suele ocurrir durante la primera iteración y siempre que se identifican acciones de mejora en la fase Evaluar de un ciclo anterior.

2.1. Definir objetos o unidades estadísticas

Los denominados objetos o unidades estadísticas corresponden a entidades, objetos, elementos o individuos del mundo real, ya sean personas, empresas u organizaciones, hogares, viviendas, inmuebles, vehículos, etc. Son los elementos que forman parte de la población.

Existen diferentes tipos de objetos dentro del sistema de registros estadísticos. Todos los tipos de objetos deben tener una definición clara y precisa, y se deben documentar tanto las definiciones administrativas como las estadísticas.

Se deben documentar las definiciones de las variables clave de identificación de los objetos. Serán utilizadas para detectar duplicados y unir o emparejar registros.

Se debe mantener un registro de referencias cruzadas con todos los cambios de identificación de los objetos a lo largo del tiempo (cambios de número de identificación personal de la población, o cambio del número de predio del inmueble).

2.2. Definir la población

Cuando se genera un nuevo registro estadístico para una investigación específica, se debe definir la población del nuevo registro. Cada registro fuente tiene su propio conjunto de objetos (población), que se incluirá total o parcialmente en el nuevo registro.

Según Wallgren y Wallgren: *la definición de una población debe mostrar claramente qué objetos están incluidos en esa población. El tipo de objeto*

también se especificará con claridad. Asimismo, siempre se incluirán una referencia temporal y una delimitación geográfica. Esta última deberá indicar la relación que existe entre los objetos o unidades estadísticas y el área geográfica.

Se deben usar los registros base para definir los conjuntos de objetos o población del RE por dos razones: los conjuntos de objetos del registro base son los mejores, en teoría, y porque las estadísticas basadas en registros deben ser consistentes.

Requisitos que deben cumplir los registros base para ser utilizados en la definición de poblaciones de los RE:

- Contener referencias temporales (fechas de ocurrencia de los eventos que afectan a los objetos).
- Tener una buena cobertura.
- Tener variables de unión de buena calidad.
- Tener variables de estratificación actualizadas y de buena calidad.

2.3. Definir las variables

Este subproceso define las variables que se utilizarán de los registros administrativos, así como cualquier otra variable que se derivará de ellas en el subproceso 6.3 (derivar nuevas variables y unidades) y cualquier clasificación estadística que se utilizará. Los metadatos de las variables del RA y las derivadas deben documentarse antes de iniciar las fases subsiguientes. Se debe documentar también el criterio de selección de las variables de los diferentes registros administrativos que conformarán el registro estadístico.

2.4. Diseñar la recolecta de datos

Este subproceso determina el método o métodos de recolecta de datos más apropiados o disponibles con la fuente administrativa. Las actividades reales en este subproceso variarán según el tipo de instrumentos de recolección utilizados, que pueden incluir cuestionarios en papel digitados en el INE, transferencia de archivos de datos, webservices para acceso a los datos administrativos, integración de bases de datos, etc. Este subproceso incluye el diseño de instrumentos de recolecta. También incluye los ajustes de cualquier acuerdo formal relacionado con el suministro de datos, como memorandos de entendimiento y confirmación de la base legal para la entrega de datos, ambos elementos considerados en el subproceso 1.8 *Preparar acuerdo de entrega de información*, pero que deben ser ratificados o ajustados en esta etapa.

2.5. Descubrir datos (preliminar)

Descubrir los datos de los registros administrativos a partir de entregas parciales del archivo completo del RA. Este subproceso podrá llevarse a cabo junto con los procesos de la etapa “5. Transformar” si no se acuerda una entrega parcial de datos del RA (archivos de prueba), en cuyo caso deberá iniciarse esa etapa con el descubrimiento de los datos para ajustar los algoritmos y funciones desarrollados en esta etapa de Diseño. El objetivo es conocer los datos de antemano para elaborar un perfil de los mismos que facilite el diseño de los procedimientos de estandarización y armonización de variables, revisión y validación de datos, procesamiento e integración de los registros administrativos. Detectar posibles problemas, conocer el estado de los datos, contenido de los archivos o tablas de los registros administrativos. Elaborar un perfil de los datos (data profiling) para facilitar el diseño o ajuste de los procedimientos y algoritmos de validación, depuración y transformación del registro administrativo.

2.6. Diseñar procesamiento e integración

Este subproceso diseña la metodología de procesamiento estadístico que se aplicará durante las fases "Transformar" e "Integrar". Esto puede incluir la especificación de rutinas para estandarizar variables, codificar, depurar, procesar, integrar, derivar variables, validar y finalizar datasets.

2.7. Diseñar la actualización de registros estadísticos

Este subproceso establece cómo se realiza la actualización de los archivos o tablas de la base de datos de producción estadística correspondientes a los registros estadísticos que se están creando/actualizando (primarios o integrados y base, si corresponde) para su uso por parte de las áreas temáticas o usuarios externos.

2.8. Diseñar sistemas de producción y flujos de trabajo

Este subproceso establece el flujo de trabajo desde la recolecta de datos hasta la comunicación, tomando una visión general de todos los procesos requeridos en todo el proceso de creación y actualización del sistema de registros estadísticos, asegurando que se acoplan eficientemente sin brechas o redundancias. Se necesitan varios sistemas y bases de datos durante todo el proceso. Se debe hacer hincapié en reutilizar procesos y tecnología. Este subproceso también considera cómo el personal interactuará con los sistemas y quién será responsable de qué y cuándo.

Flujo del proceso de la Fase 2. Diseñar:

Elementos de Entrada		Elementos de Salida
Requerimientos de los usuarios.	2.1 Definir objetos o unidades estadísticas	Metadatos sobre definición de objetos y variables clave de identificación.
Requerimientos de los usuarios.	2.2 Definir la población	Metadatos actualizados con definición de la población.
Requerimientos de los usuarios.	2.3 Definir las variables	Metadatos actualizados con definición de variables.
Acuerdo de entrega de información.	2.4 Diseñar la recolecta de datos	Documento de diseño de ETL de la recolecta de datos.
Documentos de diseño.	Diseño de rutinas para extracción de datos (lectura de archivos, integración de BD, webservices, etc) / 2.8 Diseñar sistemas de producción y flujos de trabajo	Documento de diseño de ETLs y flujos del BPMS.
Documentación del RA. Datos (preliminares o de prueba) del RA.	2.5 Descubrir datos (preliminar)	Documento perfil (preliminar) de los datos.
Documentación del RA.	Análisis de reglas de validación, estandarización de variables, etc. / 2.6 Diseñar Procesamiento e integración	Documento de especificación de diseño de ETL y reglas de consistencia.
Documentación del RA.	2.7 Diseñar la actualización de RE	Documento de diseño de ETL de actualización del RE.

3. Desarrollar

En esta fase se construye y prueba la solución hasta el punto en que está lista para su uso en el entorno de producción. Los resultados de la fase "Diseñar" orientan la selección de procesos, instrumentos, información y servicios reutilizables que se montan y configuran en esta fase para crear el entorno operacional completo para ejecutar el proceso. Los nuevos servicios se crean por excepción, y desarrollados en respuesta a las brechas existentes en el catálogo de servicios (sistemas) de la organización y externamente. Estos nuevos servicios serán construidos para ser ampliamente reutilizables dentro de la arquitectura de producción de registros estadísticos.

3.1. Desarrollar o mejorar instrumentos de recolecta de datos

Este subproceso abarca las actividades para desarrollar los instrumentos de recolecta que se utilizarán durante la fase "Recolectar datos". El instrumento de recolecta se genera o construye sobre la base de las especificaciones de diseño creadas durante la fase "Diseñar". Este subproceso también incluye preparar y probar el funcionamiento de ese instrumento. Se recomienda asociar los instrumentos de recolecta al sistema de metadatos estadísticos, de modo que los metadatos puedan capturarse más fácilmente en la fase “Recolectar datos”. La conexión de metadatos y datos en el momento de la recolecta puede ahorrar trabajo en fases posteriores.

3.2. Desarrollar o mejorar componentes del procesamiento

En este subproceso llevan a cabo las actividades para desarrollar nuevos y mejorar componentes y servicios existentes necesarios para la fase "Transformar", tal como se diseñó en la fase "Diseñar". Esto incluye el desarrollo de rutinas para estandarizar variables, detectar duplicados, codificar, depurar, procesar y validar datasets.

3.3. Desarrollar o mejorar componentes de integración

Este subproceso incluye las actividades para desarrollar nuevos y mejorar componentes y servicios existentes necesarios para la fase "Integrar", tal como se diseñó en la fase "Diseñar". Esto incluye el desarrollo de rutinas para integrar RA y RE, derivar nuevas variables y unidades, des-identificar registros (proteger datos), actualizar y finalizar datasets.

3.4. Configurar flujos de trabajo

Este subproceso configura el flujo de trabajo, los sistemas y las transformaciones utilizados en los procesos de creación y actualización del

sistema de registros estadísticos, desde la recolecta de datos hasta la comunicación. Asegura que el flujo de trabajo especificado en el subproceso 2.6 *Diseñar sistemas de producción y flujo de trabajo* funcione en la práctica.

3.5. Probar sistemas de producción

Este subproceso se refiere a la prueba de los sistemas implementados y configurados y sus respectivos flujos de trabajo. Incluye las pruebas técnicas y la aprobación de nuevos programas y rutinas, así como la confirmación de que las rutinas existentes de otros procesos de negocio estadísticos son adecuadas para su uso en este caso. Mover los componentes del proceso al entorno de producción y asegurar que funcionen como se espera en ese entorno.

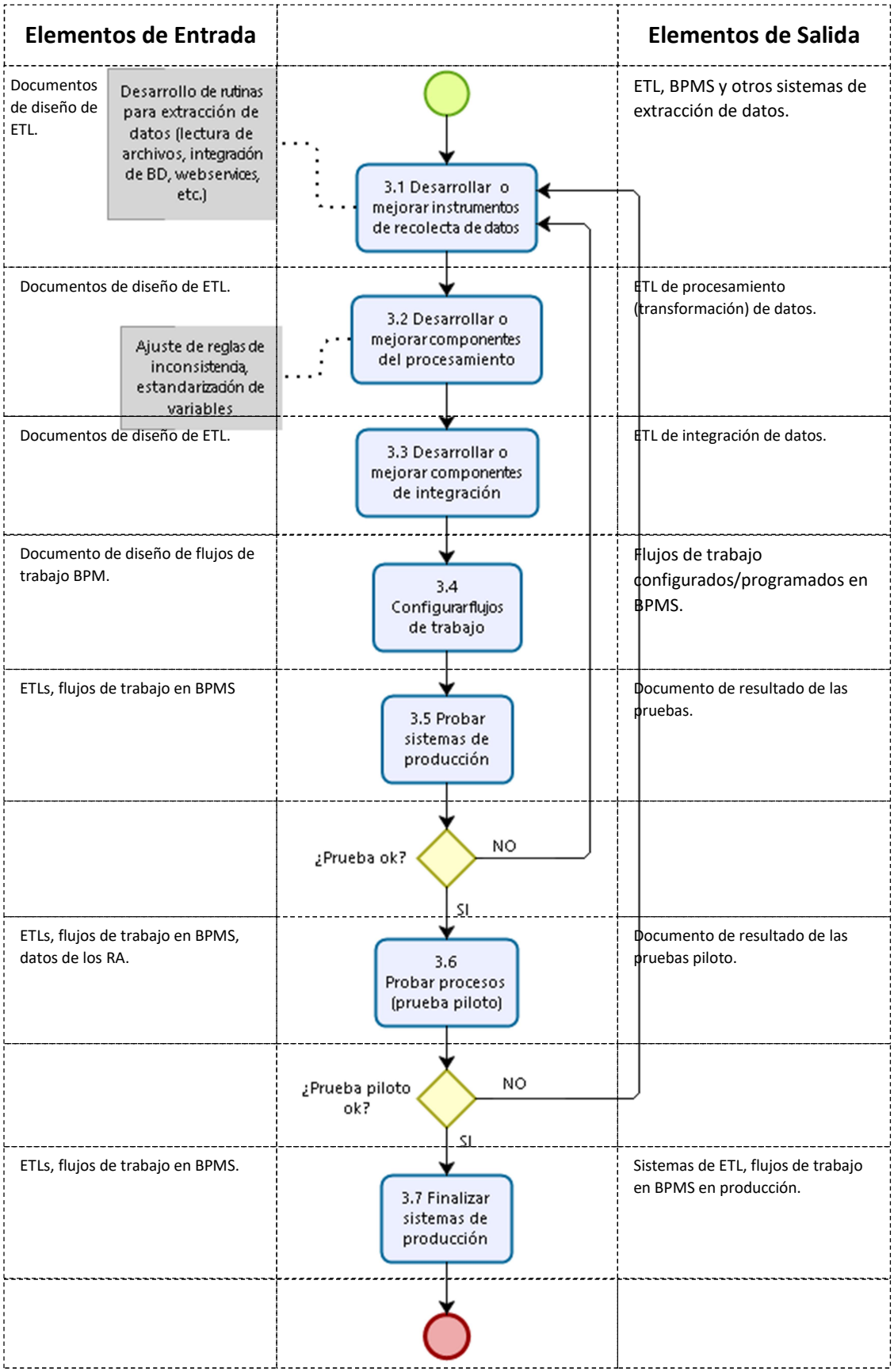
3.6. Probar procesos (pruebas piloto)

Este subproceso abarca las actividades para gestionar una prueba piloto del proceso de actualización y mantenimiento del sistema de registros estadísticos. Normalmente incluye una recolecta de datos a pequeña escala, para probar los instrumentos de recolecta, seguido por el procesamiento e integración de los datos, para asegurar que el proceso se ejecute como se esperaba. Después del piloto, puede ser necesario volver a pasos anteriores y hacer ajustes a instrumentos, sistemas o componentes.

3.7. Finalizar sistemas de producción

Este subproceso incluye las actividades para que los procesos y sistemas implementados y configurados, incluyendo los servicios modificados y recién creados en producción, queden listos para el uso por las áreas temáticas. Las actividades incluyen: producir documentación sobre los componentes del proceso, incluida la documentación técnica y los manuales de usuario; y capacitar a los usuarios sobre cómo operar el proceso.

Flujo del proceso de la Fase 3. Desarrollar:



4. Recolectar datos

Esta fase recoge o reúne toda la información necesaria (datos y metadatos), utilizando diferentes modos de recolección (digitación de formatos papel, extracción de datos de archivos, integración de bases de datos, webservices, etc.) y los carga en el entorno adecuado para su posterior procesamiento. Aunque puede incluir la validación de formatos del archivo de datos (dataset), no incluye ninguna transformación de los datos en sí, ya que esto se realiza en la fase "Transformar".

4.1. Organizar/preparar recolecta de datos

Este subproceso garantiza que las personas, los procesos y la tecnología están preparados para recopilar datos y metadatos. Este subproceso incluye:

- Garantizar la disponibilidad de recursos de recolecta, por ejemplo: servidores, medios de transferencia de archivos, software de extracción de datos, sistemas de integración de bases de datos.
- Configurar sistemas de recolecta para solicitar y recibir los datos.
- Asegurar la confidencialidad de los datos a recolectar.
- Preparar instrumentos de recolecta.

4.2. Ejecutar recolecta de datos

Este subproceso implica la recolecta de datos en sí misma. Se ponen en marcha las actividades de recolecta de datos utilizando los instrumentos especialmente diseñados y desarrollados para recolectar los datos provistos por las fuentes administrativas. Se realiza la extracción de datos (lectura de archivos, integración de BD, webservices, etc.).

4.3. Finalizar recolecta de datos

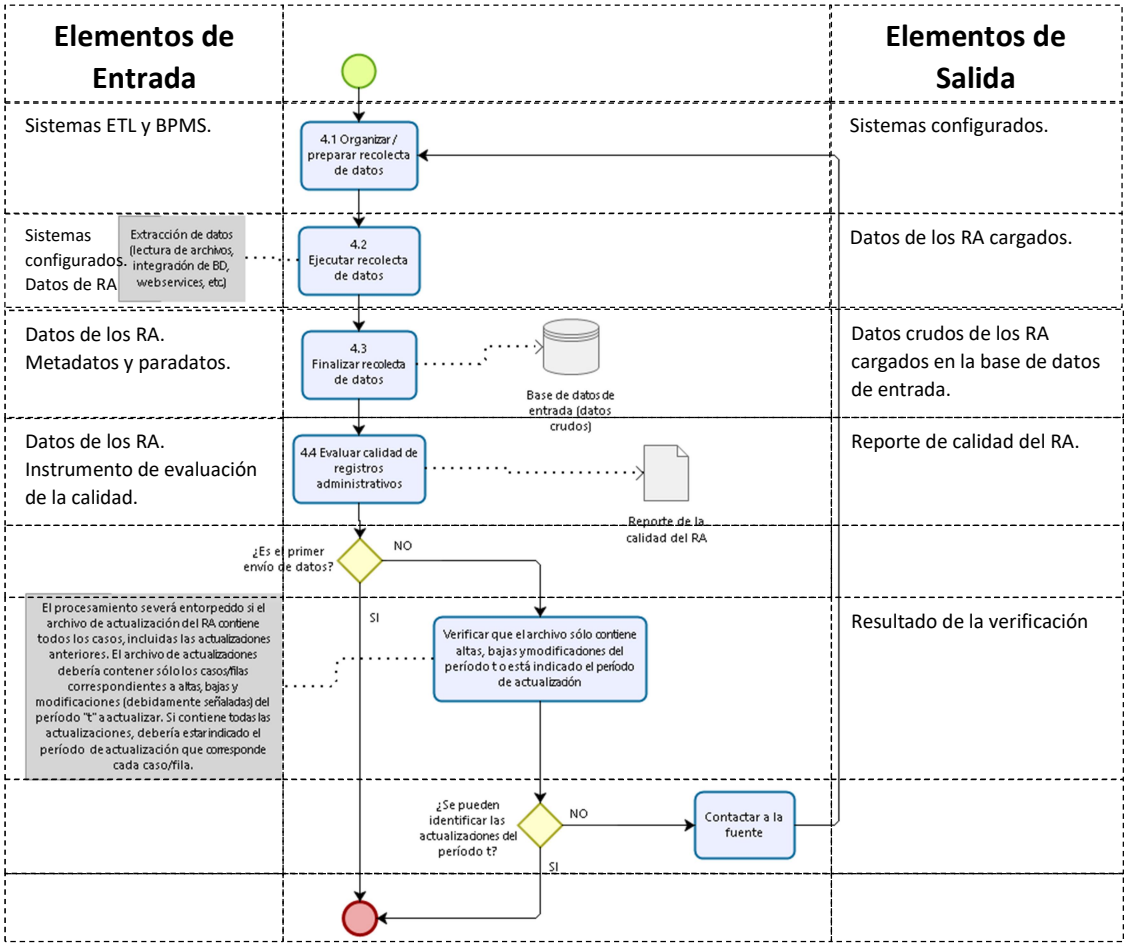
Este subproceso incluye la carga de los datos y metadatos recopilados en un entorno electrónico adecuado (base de datos de entrada, con datos crudos) para su posterior procesamiento. Puede incluir la toma de datos manuales o automáticos, por ejemplo, utilizando personal administrativo u herramientas ópticas de reconocimiento de caracteres para extraer información de cuestionarios en papel o convertir los formatos de archivos recibidos de las fuentes administrativas. También puede incluir análisis de los metadatos del proceso (paradata) asociados con la recolecta para asegurar que las actividades de recolección cumplan con los requisitos.

4.4. Evaluar calidad de registros administrativos

Se debe evaluar la calidad de los datos de los registros administrativos una vez que son recolectados. Se debe seguir la metodología de evaluación de la calidad de los datos del registro y se debe generar el

correspondiente reporte de la calidad. Este subproceso será clave para determinar efectivamente la utilidad de los datos y sus restricciones en cuanto al uso estadístico.

Flujo del proceso de la Fase 4. Recolectar datos:



5. Transformar

Esta fase implica la transformación del registro administrativo en registro estadístico y su preparación para la integración con otros registros. Se compone de subprocesos que estandarizan variables, validan, depuran y transforman los datos de entrada, para que puedan ser utilizados con fines estadísticos. Puede repetirse varias veces si es necesario. Para las actualizaciones de datos recolectadas regularmente, esta fase se produce en cada iteración.

5.1. Descubrir los datos

En el subproceso "2.5 Descubrir datos (preliminar)" se realiza el descubrimiento de datos (data profiling) para tener un perfil de los datos para facilitar el diseño y desarrollo de los algoritmos de procesamiento y transformación del registro administrativo. En el caso que no se haya podido obtener datos preliminares o parciales del archivo del registro administrativo para hacer este descubrimiento de datos, se debe ejecutar en esta etapa y realizar los ajustes necesarios a los algoritmos y funciones para el procesamiento, transformación e integración de registros administrativos. Si por el contrario, se logró obtener los datos del registro en la etapa de diseño, aquí se debería verificar que los datos finales del archivo completo del registro administrativo presentan las mismas características, es decir el mismo perfil que los datos utilizados para el diseño; y en el caso que existan diferencias se deberá hacer los ajustes necesarios en los algoritmos y funciones para adaptarlos al nuevo perfil de los datos.

5.2. Estandarizar/armonizar variables

Las definiciones de las variables del RA se ajustan a las necesidades administrativas y no siempre se corresponden con las definiciones estadísticas de acuerdo a un uso estadístico específico.

La estandarización o armonización de variables implica la adecuación de las definiciones de las variables del RA a las definiciones estandarizadas del RE. Para lo cual podría ser necesario transformar las categorías de respuesta de variables, los clasificadores/codificadores utilizados, los formatos de datos o los datos en sí mismos. Además, implica la normalización de textos y creación o actualización de fichas de metadatos.

5.3. Revisar y validar

Este subproceso examina los datos para tratar de identificar problemas potenciales, errores y discrepancias tales como valores atípicos, falta de

respuesta en las variables y errores de codificación. También puede denominarse validación de datos de entrada. Puede ejecutarse iterativamente, validando datos contra reglas de inconsistencias predefinidas, generalmente en un orden establecido. Puede marcar los datos para la revisión o edición automática o manual. La revisión y validación puede aplicarse a los datos de cualquier tipo de fuente, antes y después de la integración de datos. Si bien la validación se trata como parte de la fase "Transformar", en la práctica, algunos elementos de validación pueden ocurrir junto con las actividades de recolecta y luego de la fase “Integrar”, cuando se tienen los registros unidos y se pueden utilizar variables adicionales para las validaciones. Si bien este subproceso se refiere a la detección de errores reales o potenciales, cualquier actividad de corrección que realmente cambie los datos se realiza en el subproceso 5.4.

5.4. Depurar e imputar

Cuando los datos se consideran incorrectos, faltan o no son fiables, este subproceso puede incorporar nuevos valores en su lugar. Los términos depuración/edición cubren una variedad de métodos para hacer esto, a menudo usando un enfoque basado en reglas. Los pasos específicos típicamente incluyen:

• Determinación de agregar o cambiar datos.
• Selección del método a utilizar.
• Agregar / cambiar valores de datos.
• Escribir los nuevos valores de los datos en el conjunto de datos y marcarlos como cambiados.
• Documentación de metadatos sobre el proceso de depuración/edición.

5.5. Detectar duplicados

Los archivos del RA pueden contener duplicados (casos o filas con los mismos datos, claves de identificación duplicada, claves diferentes pero los casos corresponden al mismo objeto o unidad). Se utilizan métodos determinísticos y probabilísticos para detectar duplicados.

Se detectan duplicados a través de las claves de identificación, que luego serán eliminados (borrado lógico, se marca la fila con el estado *inactivo por duplicado*) si a través de los métodos probabilísticos se detecta que hay una alta probabilidad de que se trate del mismo objeto o unidad en la realidad.

Las actividades previas a la aplicación de los métodos probabilísticos son: normalización de textos, transformación de variables numéricas o fecha a texto, concatenación de variables (también se utilizan las variables clave de identificación).

Luego, se aplican las técnicas probabilísticas de unión de registros (el archivo del RA se empareja con sí mismo para detectar duplicados), donde a cada caso/fila del archivo se le asigna una probabilidad de coincidencia (posible duplicado) con un caso/fila del mismo archivo. Se aplican técnicas de blocking, indexing o filtering, para trabajar con bloques del archivo de menor tamaño y mejorar la velocidad de procesamiento.

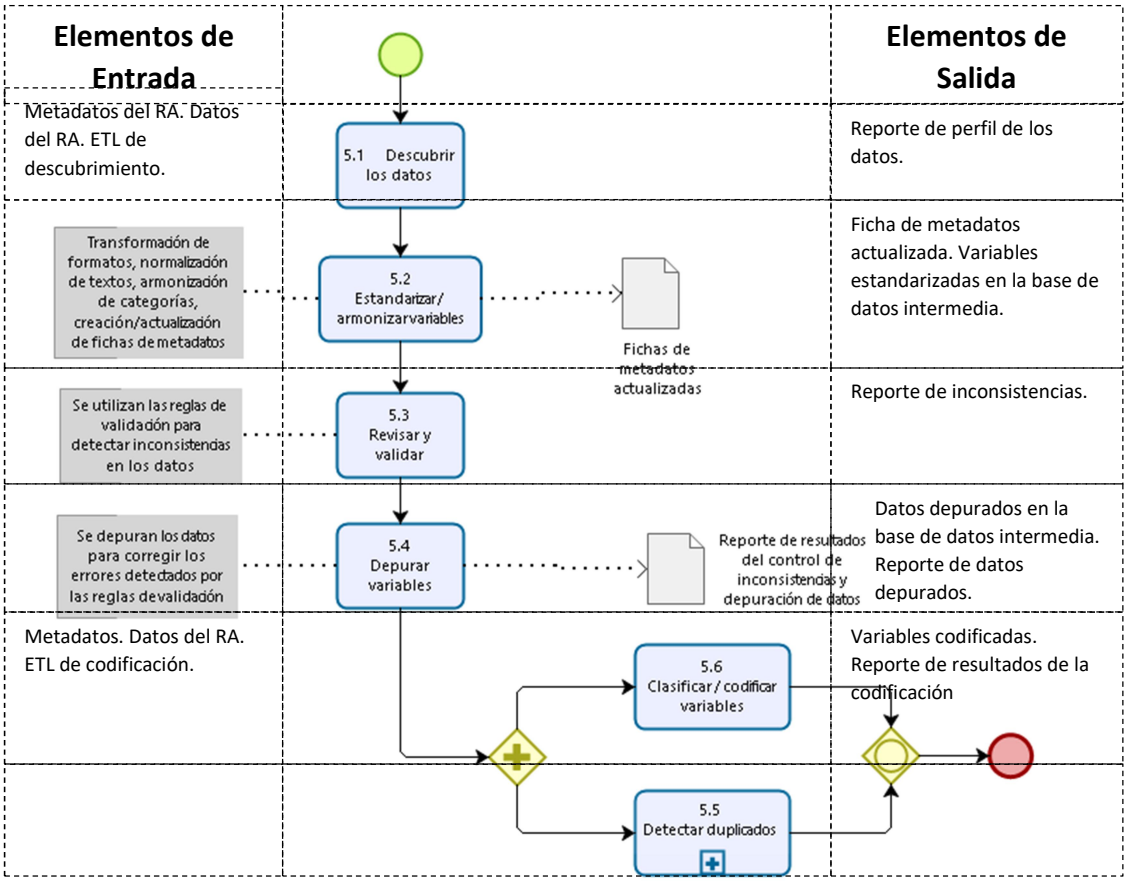
Las pares de casos cuya probabilidad de coincidencia supera cierto umbral son considerados como duplicados y serán marcados como eliminados (borrado lógico, se marca la fila con el estado *inactivo por duplicado*) del archivo de datos.

Finalmente, la calidad de los resultados es evaluada.

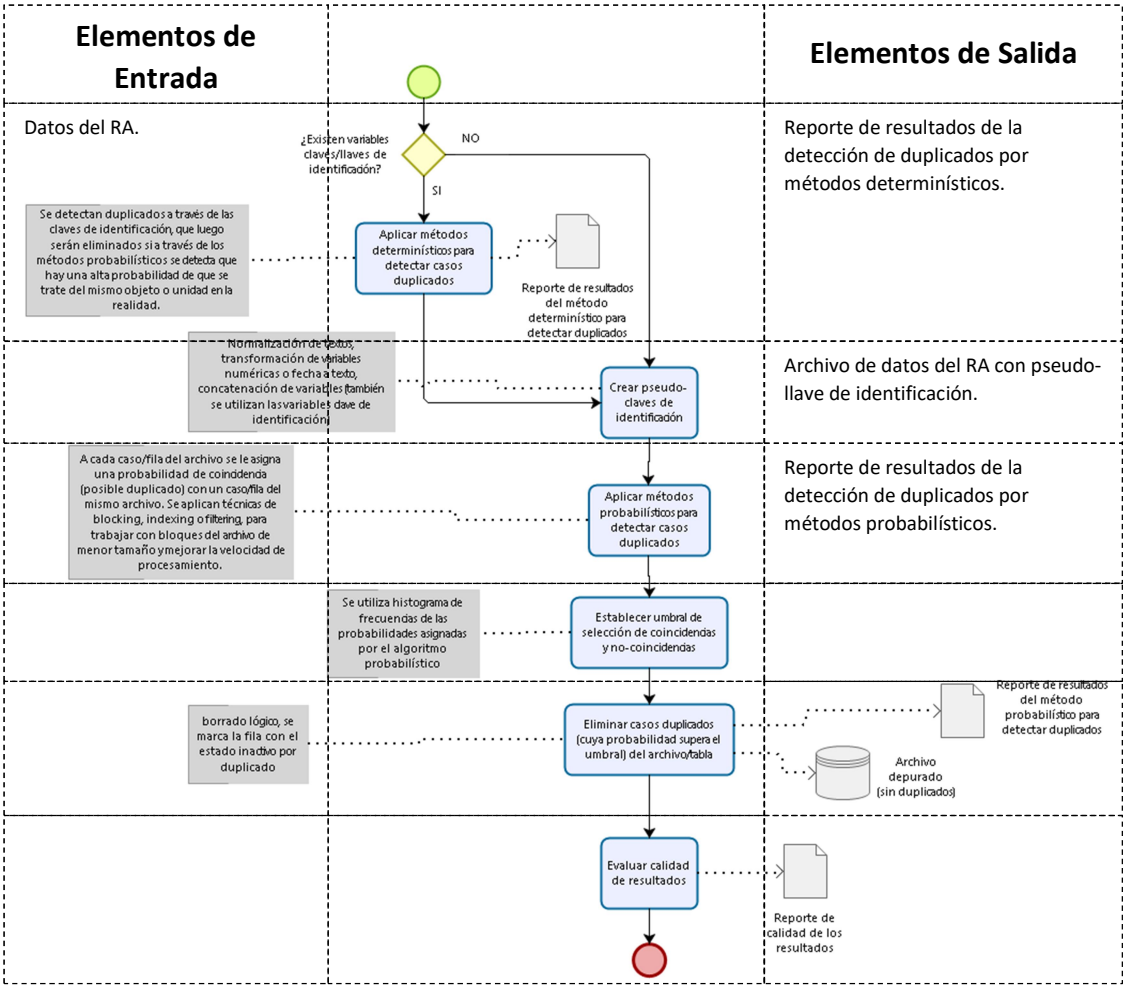
5.6. Clasificar/codificar variables

Este subproceso clasifica y codifica los datos de entrada. Por ejemplo, las rutinas de codificación automáticas (o manuales) asignan códigos numéricos a las respuestas de preguntas de texto abiertas de acuerdo con un esquema de clasificación predeterminado.

Flujo del proceso de la Fase 5. Transformar:



Flujo del subproceso 5.5 Detectar duplicados:



6. Integrar

En esta fase se realiza la integración de registros a través de métodos de unión o emparejamiento de archivos. Se crean nuevas variables y unidades luego de la unión de registros. Se des-identifican los registros para asegurar la confidencialidad de la información previamente a la actualización en la base de datos de producción estadística.

6.1. Integrar registros (unir o emparejar)

Se aplican métodos determinísticos y probabilísticos para realizar la unión de registros.

La unión de registros consta de tres etapas: pre-unión, unión y post-unión. La pre-unión corresponde a la normalización de variables de texto y estandarización otras variables (realizada en la fase 5 “Transformar”). La unión de registros se realiza aplicando algoritmos estadísticos estandarizados y ampliamente utilizados en diversas disciplinas. Finalmente, la post-unión o proceso de evaluación de los resultados (siguiente subproceso).

6.2. Evaluar resultados de la integración

La evaluación de resultados de la integración permite identificar oportunidades de mejora de los procesos y métodos utilizados, para lograr mejores tasas de coincidencias, minimizando los “falsos positivos” incluidos en las coincidencias y los “falsos negativos” excluidos del emparejamiento.

6.3. Derivar nuevas variables y unidades

Este subproceso deriva los datos de variables y unidades que no se proporcionan explícitamente en el archivo del registro administrativo, pero son necesarios para cumplir con los requerimientos de los usuarios. Se crean nueva variables a partir de otras variables del RA o de los RE con los que se ha integrado (se crean por varios métodos diferentes). Esta actividad podría ser iterativa, ya que algunas variables derivadas pueden basarse, a su vez, en otras variables derivadas. Por lo tanto, es importante asegurarse de que las variables se derivan en el orden correcto. Las nuevas unidades pueden derivarse agregando o dividiendo datos de las unidades del RA, o por otros métodos de estimación. Los ejemplos

incluyen los hogares derivados de las unidades del RA de personas o las empresas donde las unidades del RA son unidades legales.

6.4. Aplicar la protección de datos

En este subproceso se hace la desidentificación del registro estadístico. Es decir, se sustituyen las variables clave de identificación por claves aleatorias y se eliminan las variables de contacto o identificadores explícitos (nombres, teléfono, dirección, email, coordenadas geográficas) del RE que estará disponible para los usuarios internos y externos.

ATENCIÓN: este proceso de protección de datos simplemente hace la des-identificación de registros. No se trata de un proceso de anonimización completo, el cual debería implementarse antes de liberar una base de microdatos a usuarios externos.

6.5. Crear/actualizar registro estadístico en BD de producción estadística

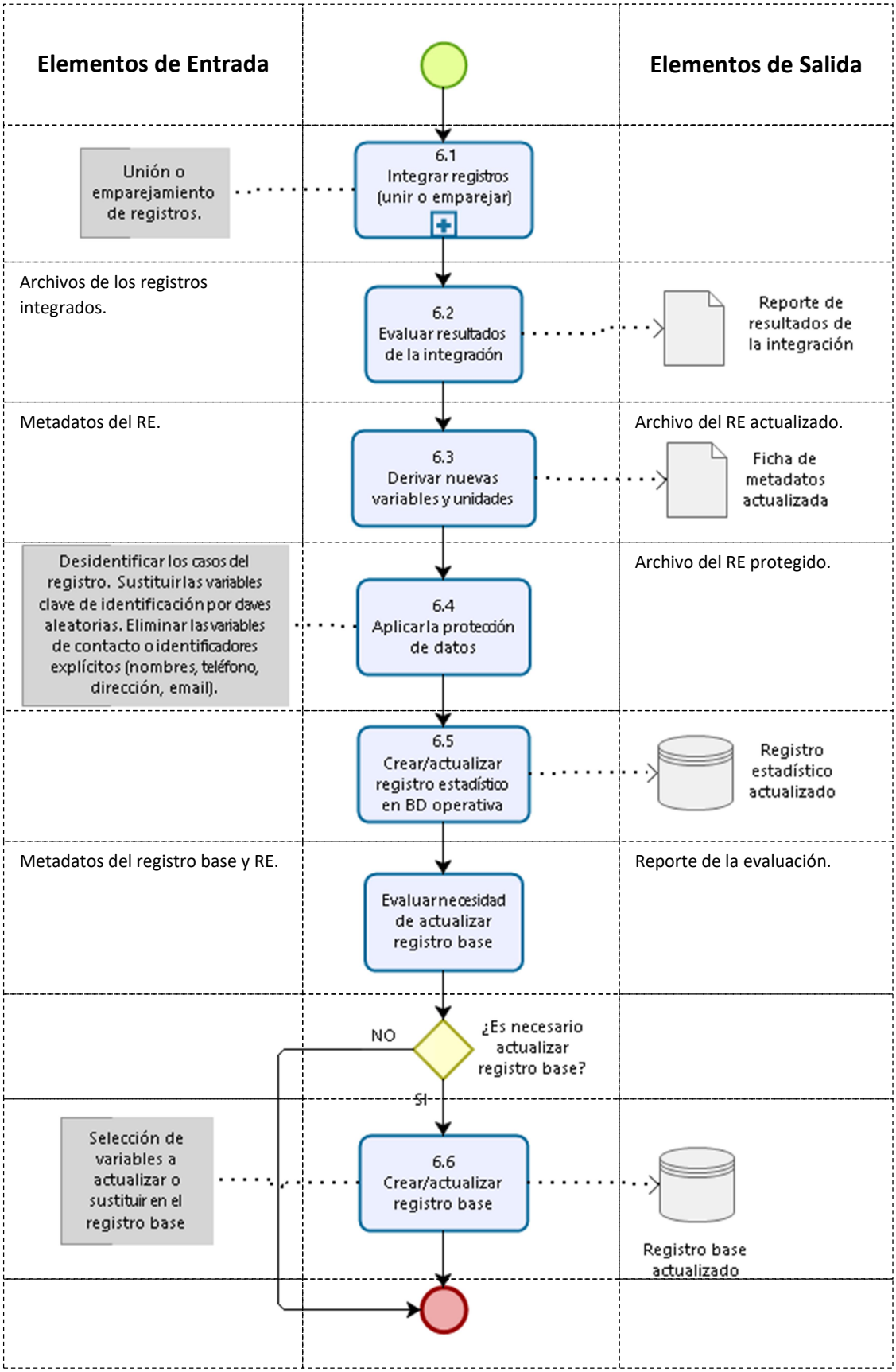
Este subproceso realiza la actualización de los archivos o tablas de la base de datos de producción estadística correspondientes a los registros estadísticos que se están creando/actualizando (primarios o integrados y base, si corresponde), para su uso por parte de las áreas temáticas o usuarios externos. Se actualiza el estado de los casos/filas del registro (Activo, Inactivo por emigración, Inactivo por muerte, Eliminado por proceso administrativo, Inactivo por duplicado, etc.).

6.6. Crear/actualizar registro base

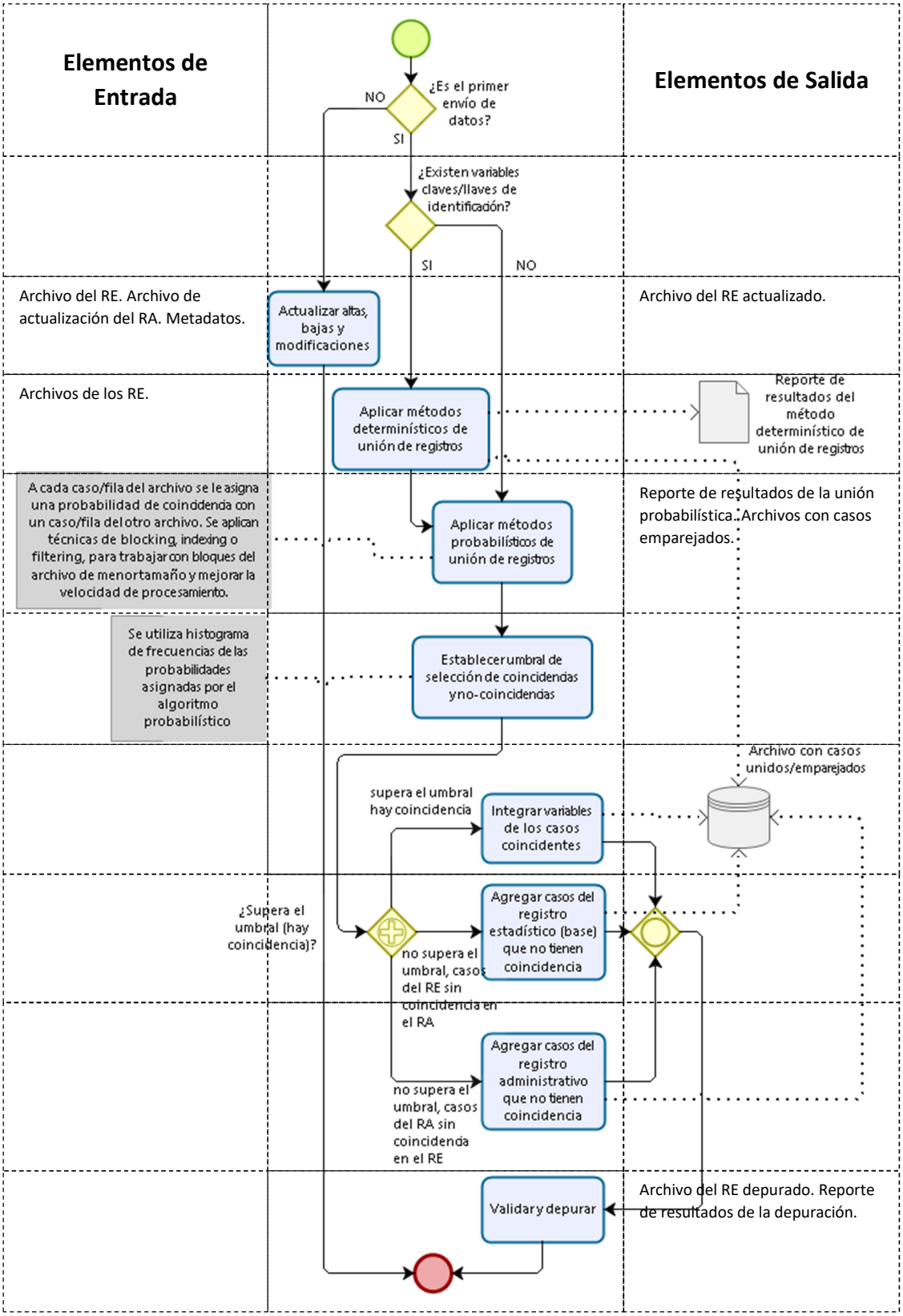
Se evalúa la necesidad de actualizar o mejorar el registro base. Se seleccionan las variables y unidades a actualizar o sustituir en el registro base para mejorar su cobertura, calidad y actualidad de los datos. Se actualiza la ficha de metadatos del registro base, y se verifica que se actualiza la información correspondiente a la trazabilidad o mapeo de variables.

Se actualiza el estado de los casos/filas del registro base (Activo, Inactivo por emigración, Inactivo por muerte, Eliminado por proceso administrativo, Inactivo por duplicado, etc.).

Flujo del proceso de la Fase 6. Integrar:


Elementos de Entrada
Elementos de Salida
Unión o emparejamiento de registros.
6.1 Integrar registros (unir o emparejar)
Archivos de los registros integrados.
6.2 Evaluar resultados de la integración
Reporte de resultados de la integración
Metadatos del RE.
6.3 Derivar nuevas variables y unidades
Archivo del RE actualizado.
Ficha de metadatos actualizada
Desidentificar los casos del registro. Sustituir las variables clave de identificación por claves aleatorias. Eliminar las variables de contacto o identificadores explícitos (nombres, teléfono, dirección, email).
6.4 Aplicar la protección de datos
Archivo del RE protegido.
6.5 Crear/actualizar registro estadístico en BD operativa
Registro estadístico actualizado
Metadatos del registro base y RE.
Evaluar necesidad de actualizar registro base
Reporte de la evaluación.
NO
¿Es necesario actualizar registro base?
SI
Selección de variables a actualizar o sustituir en el registro base
6.6 Crear/actualizar registro base
Registro base actualizado

Flujo del subproceso 6.1 Integrar registros (unir o emparejar):



7. Comunicar

Es importante comunicar los cambios y actualizaciones de datos realizadas en el sistema de RE y en los sistemas de difusión y exploración de datos, de modo tal que los usuarios internos y externos conozcan el nuevo alcance y posibilidades del sistema en cuanto a la producción de estadísticas.

7.1. Actualizar sistemas de difusión de datos y herramientas de análisis

Este subproceso gestiona la actualización de los sistemas en los que los datos y los metadatos se almacenan listos para su difusión o explotación, entre ellos:

- Dar formato a los datos y metadatos para que estén listos para ser cargados en bases de datos de salida;
- Cargar datos y metadatos en las bases de datos de salida;
- Asegurar que los datos estén vinculados a los metadatos relevantes.
- Actualizar el Data Warehouse (si corresponde).

El formateo, la carga y la vinculación de los metadatos deberían realizarse principalmente en fases anteriores, pero este subproceso incluye una verificación final de que todos los metadatos necesarios están listos para su difusión o explotación.

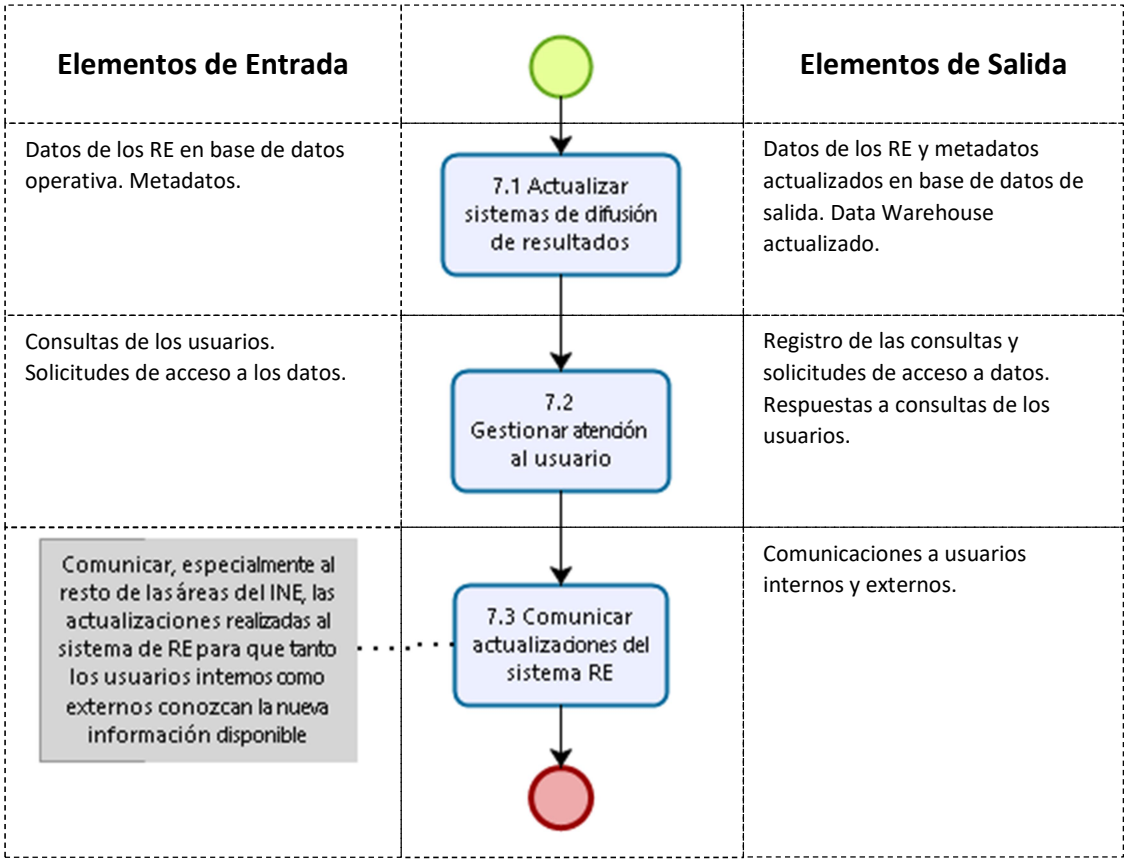
7.2. Gestionar atención al usuario

Este subproceso garantiza que se registran las consultas de los usuarios y las peticiones de servicios como el acceso a los microdatos, y que las respuestas se proporcionen dentro de los plazos acordados. Estas consultas y solicitudes deben revisarse periódicamente para proporcionar una entrada al proceso de gestión de calidad general, ya que pueden indicar cambios o nuevas necesidades de los usuarios.

7.3. Comunicar actualizaciones del sistema de registros estadísticos

Se deben comunicar adecuadamente los cambios y actualizaciones realizadas en los datos del sistema de RE, de modo tal que los usuarios internos y externos conozcan el nuevo alcance y posibilidades del sistema en cuanto a la producción de estadísticas.

Flujo del proceso de la Fase 7. Comunicar:



8. Evaluar y mejorar continuamente

Esta fase gestiona la evaluación de un caso o ciclo específico del proceso de gestión del sistema de RE, en contraposición con el proceso general de gestión de la calidad que abarca a todos los procesos del INE. Lógicamente tiene lugar al final de la instancia del proceso, pero se basa en los insumos reunidos a lo largo de las diferentes fases. Incluye la evaluación del desempeño de una instancia específica del proceso, considerando una serie de insumos cuantitativos y cualitativos, y la identificación y priorización de mejoras potenciales.

Para los procesos de actualización regular de los datos de los mismos RA, donde se ejecutan ciclos iterativos, la evaluación debería (al menos en teoría) ocurrir para cada iteración y si se deberían implementar mejoras en los siguientes ciclos.

8.1. Reunir información para la evaluación

El material para la evaluación se puede producir en cualquier otra fase o subproceso. Puede tomar muchas formas, incluyendo retroalimentación de los usuarios, metadatos del proceso (paradata), métricas del sistema y sugerencias del personal. Los informes de avances con respecto a un plan de acción acordado durante una iteración anterior también pueden ser un aporte a las evaluaciones de las iteraciones posteriores. Este subproceso reúne todos estos insumos y los pone a disposición de la persona o equipo que realiza la evaluación.

8.2. Conducir evaluación

Este subproceso analiza los insumos de la evaluación y los sintetiza en un informe de evaluación. El informe resultante debe incluir cualquier problema de calidad específico de esta iteración del proceso y debe hacer recomendaciones de mejora, si corresponde. Estas recomendaciones podrían considerar cambios en cualquier fase o subproceso para futuras iteraciones del proceso, o podrían sugerir que el proceso no se repita.

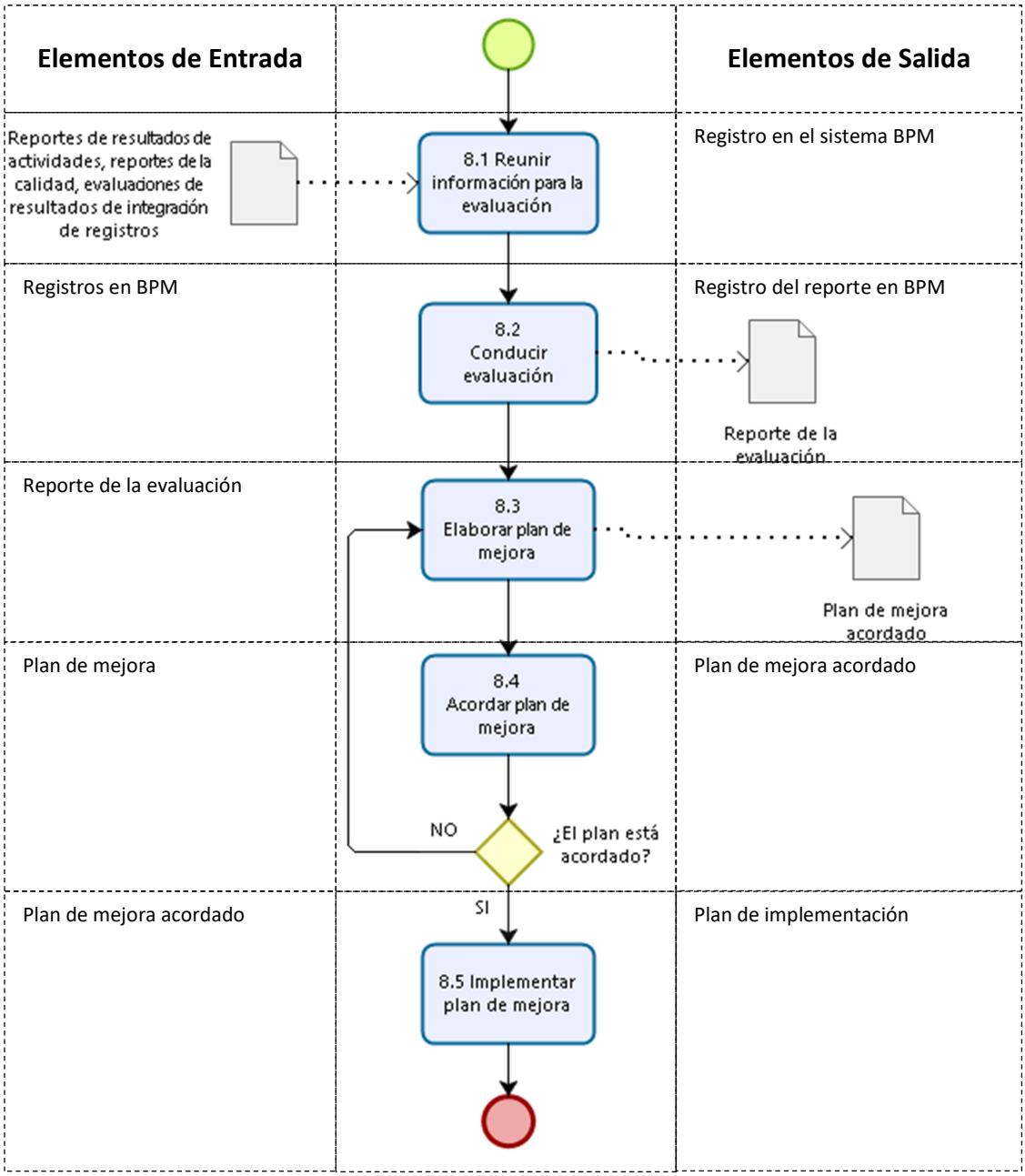
8.3. Elaborar y acordar plan de mejora

Estos subprocesos reúnen a quienes tienen el poder de decisión necesario para elaborar y acordar un plan de acción basado en el informe de evaluación. También debería incluir la consideración de un mecanismo para monitorear el impacto de esas acciones, las cuales, a su vez, pueden aportar una contribución a las evaluaciones de futuras iteraciones del proceso.

8.4. Implementar plan de mejora

Las acciones acordadas en el plan de mejora se implementan de acuerdo a las prioridades y cronograma establecidos. Los responsables designados en el plan coordinan la correcta ejecución de las acciones de mejora.

Flujo del proceso de la Fase 8. Evaluar y mejorar continuamente:

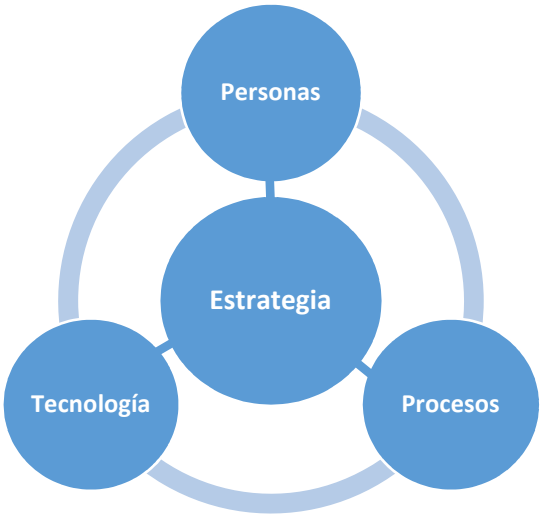


1.5 Resumen de las características generales del SIREE

El SIREE es un sistema, en la acepción más amplia del término, formado por un conjunto de elementos relacionados entre sí que funciona como un todo. Es, por un lado, un sistema conceptual (conceptos, definiciones, metadatos, metodología), un sistema de gestión (procesos, administración) y un sistema de información estadística (Data Warehouse), por otro.

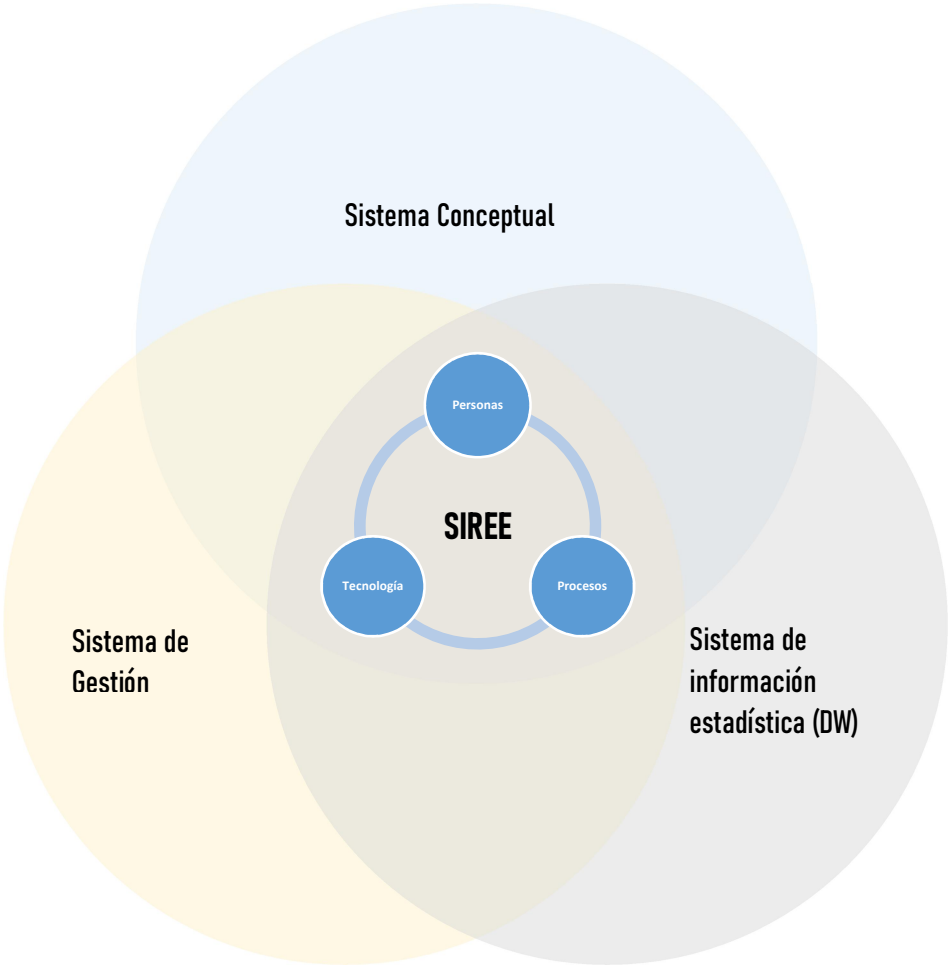
El SIREE ha sido diseñado tomando en cuenta un modelo de gestión basado en tres pilares fundamentales: personas, procesos y tecnología, es decir, no está enfocado en la tecnología exclusivamente, sino que además se sustenta en los procesos y las personas que los ejecutan, y estos tres elementos se alinean, a su vez, con la estrategia del INE.

Figura 8. Modelo de gestión base del SIREE.



La adopción del modelo de producción estadística basada en el SIREE exige que los datos sobre diversos temas se produzcan como partes integradas de un sistema de información integral, en lugar de hacerlo como "compartimientos estancos" de forma independiente entre sí. Esto implica un cambio de paradigma en la producción estadística que se hacía habitualmente en el INE.

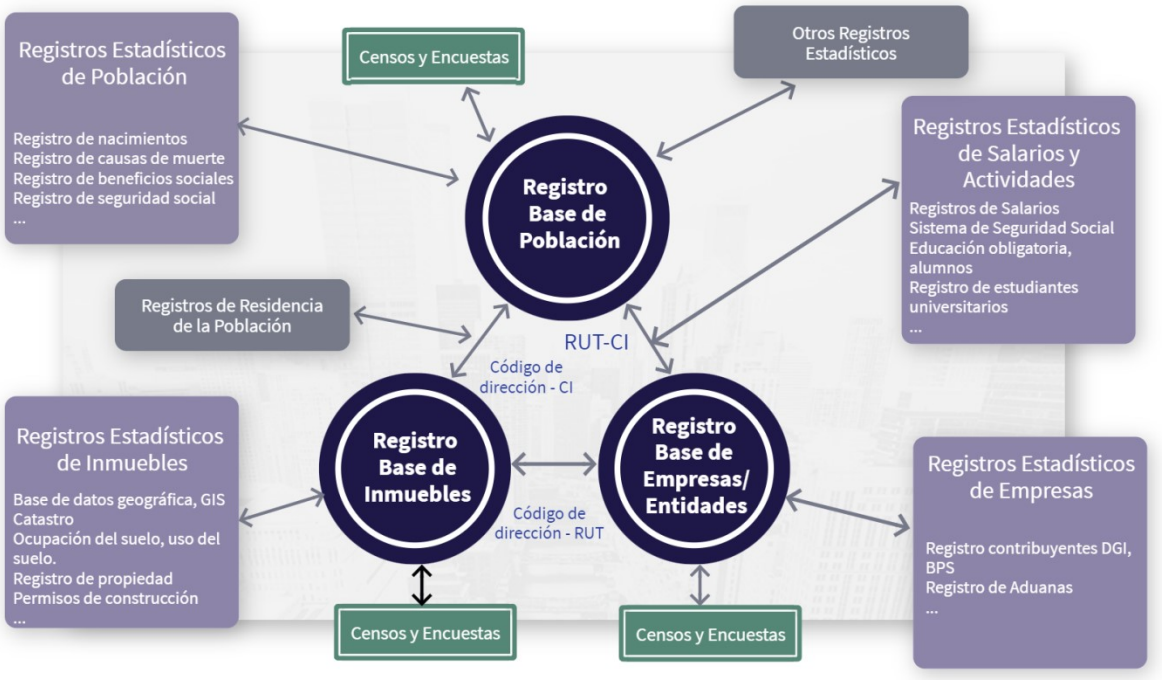
Figura 9. Sistemas que forman parte del SIREE.



El SIREE se sustenta en el Data Warehouse Estadístico (DWE), como sistema de información y de base de datos, pero abarca también aspectos conceptuales y de gestión de procesos.

Este Data Warehouse Estadístico del SIREE cuenta con una base de datos central de producción estadística conformada por las tablas correspondientes a los registros estadísticos que forman parte del sistema y las tablas auxiliares necesarias para el funcionamiento del mismo.

Figura 10. Elementos del Sistema Integrado de Registros Estadísticos y Encuestas - SIREE.



A partir de los datos operativos de la base de datos central (microdatos) se crean diferentes conjuntos de datos agregados para un área de investigación o temática en particular (Data Marts). Estos almacenes de datos permiten explotar al máximo el sistema de registros estadísticos realizando análisis y visualizaciones de datos a través de herramientas analíticas de datos.

Los procesos misionales del SIREE que corresponden a la extracción, transformación, integración y carga de datos los podemos resumir en los siguientes:

1. El primer paso es la extracción de los datos provenientes de las fuentes administrativas y almacenarlos en un repositorio de datos de entrada sin ningún procesamiento o con alguna transformación mínima (datos crudos). Se des-identifican los conjuntos de datos originales y se sustituyen las claves de identificación (cédula de identidad o RUT) por un ID-Estadístico.

2. Luego, los datos del registro administrativo almacenados en el repositorio de datos de entrada son transformados a un registro estadístico, como se ha descripto en los apartados anteriores, y son almacenados en un repositorio de datos intermedio mientras se realiza la transformación (procesamiento de datos e integración de registros y encuestas).

3. Posteriormente, los datos del registro estadístico en el repositorio de datos (intermedio) de transformación se des-identifican para preservar la confidencialidad de la información. En este proceso se eliminan las variables clave de identificación y los identificadores explícitos (nombres, dirección,

teléfono, email, coordenadas geográficas, etc.) y se sustituyen por variables clave subrogadas.

4. Finalmente, se cargan los microdatos des-identificados en la base de datos del Data Warehouse y los datos agregados en los Data Marts, si corresponde.

El sistema debe asegurar la trazabilidad de las operaciones realizadas por los usuarios con los datos, para lo cual cuenta con un log de auditoría (qué, quién y cuándo se hizo cada operación en el sistema), que permite además, conservar la historia de las transformaciones/actualizaciones de datos y permitir volver atrás o recuperar estados anteriores.

El SIREE promueve una mayor estandarización de procesos y métodos en la producción estadística, y el uso completo de toda la información (encuestas y registros estadísticos) que tenemos sobre toda la población objetivo.

Las ventajas que provee un sistema de registros estadísticos debido a la inexistencia de errores de muestreo por un lado y la integración y estandarización de procesos por el otro, superan las desventajas debidas al aumento de errores ajenos al muestreo y la pérdida parcial de control sobre los procesos de captura y validación de los datos administrativos.

Para su implementación se necesitan cambios técnicos y metodológicos importantes, un enfoque gradual y colaboración intensiva tanto interna como externa (con los proveedores de datos).

Gestión de Datos Maestros y Metadatos, sistema *Masterdata*

Los procesos de gestión de la calidad y gestión de metadatos funcionan en paralelo a los procesos misionales de extracción, transformación y carga del DW-SIREE, de forma tal que en cada etapa o actividad se recogen los metadatos correspondientes y se evalúa la calidad a lo largo de todo el proceso. Para lo cual, el SIREE cuenta con herramientas de Gestión de Datos Maestros, metadatos y calidad. El módulo de gestión de metadatos permite manejar las fichas de metadatos para la estandarización de variables.

Masterdata está basado en el estándar de metadatos ISO 11.179.

La Gestión de Datos Maestros representa un elemento clave en la orquestación del DW en particular y del SIREE en general. Garantiza que el INE, o en un sentido más amplio el SEN, trabaja siempre con una única versión de los datos “verdaderos” y actualizados.

La gestión de datos maestros consiste en definir procesos y reglas de negocio para la gestión de datos comunes en sistemas dispares de la institución o datos provenientes de diferentes fuentes externas a la organización.

Como se ha mencionado anteriormente, los metadatos son el núcleo o el ADN del sistema. Proveen trazabilidad de las fuentes y de las transformaciones de los datos, brindan transparencia al proceso, facilitan la integración de datos, es un único punto de referencia de los datos.

Los metadatos documentados a través de la herramienta *MasterData* brindan información sobre los datos pero también sobre los procesos. Contienen definiciones de reglas de consistencia y transformación de variables, homologación de variables (conceptos, vincular categorías), descripción de datasets fuente y destino, establecen reportes de la calidad de los datos. Así como también es posible documentar metadatos de indicadores (elementos de entrada, salida, fórmulas de cálculo).

Figura 11. Módulo de consulta de metadatos del software *MasterData*.

Figura 12. Módulo de mapeo de variables del registro estadístico hasta su fuente de datos, software *MasterData*.

Figura 13. Módulo de armonización de variables, software *MasterData*.

2. Diseño conceptual del Data Warehouse Estadístico

El objetivo principal de un Data Warehouse (DW) en el ámbito empresarial es integrar y almacenar los datos generados como resultado de las actividades de una organización.

En los INE, el DW se ha usado generalmente como un sistema de salida, recopilando datos agregados finales; pues los procesos de producción estadística de diferentes temas se realizan de forma independiente, como silos.

El enfoque moderno de producción estadística basada en un Data Warehouse exige que los datos agregados sobre diversos temas se produzcan como partes integradas de un sistema de información completo, en lugar de hacerlo de forma independiente entre sí. Los datos en un dominio estadístico común se almacenan una vez para múltiples propósitos.

El Data Warehouse Estadístico, según ESSnet[12], es un "*almacén central de datos estadísticos para gestionar todos los datos de interés disponibles, que permite al INE (re)utilizarlos para crear nuevos datos/resultados, producir la información necesaria y realizar informes y análisis, independientemente de la fuente de los datos*".

En resumen: "*almacén central o concentrador de datos estadísticos, independientemente de la fuente*".

El DW está diseñado para proveer una visión integral y en múltiples dimensiones de los datos provenientes de diferentes fuentes, rompiendo así con los silos de información compartimentada.

Además, el DW se convierte en la parte central de toda la infraestructura de tecnología de la información que soporta la producción estadística del INE. Es capaz de gestionar tanto microdatos como datos agregados y sus metadatos de diferentes fases del proceso de producción estadística.

Data Warehouse Estadístico-Espacial o Data Warehouse Geo-Estadístico

El DW Estadístico del INE contiene, además, la geografía del dato, añadiendo así la dimensión espacial que lo convierte en un DW Geo-Estadístico.

El DW Geo-Estadístico es una colección de datos orientados a temas, integrada, variante en el tiempo, no volátil, que incorpora la dimensión espacial, donde las variables geográfica y tiempo son imprescindibles.

[12] Goossens, H. ESSnet (2012). *The statistical data warehouse: a central data hub, integrating new data sources and statistical output*. UNECE.
https://unece.org/fileadmin/DAM/stats/documents/ece/ces/ge.44/2012/mtg2/WP18.pdf

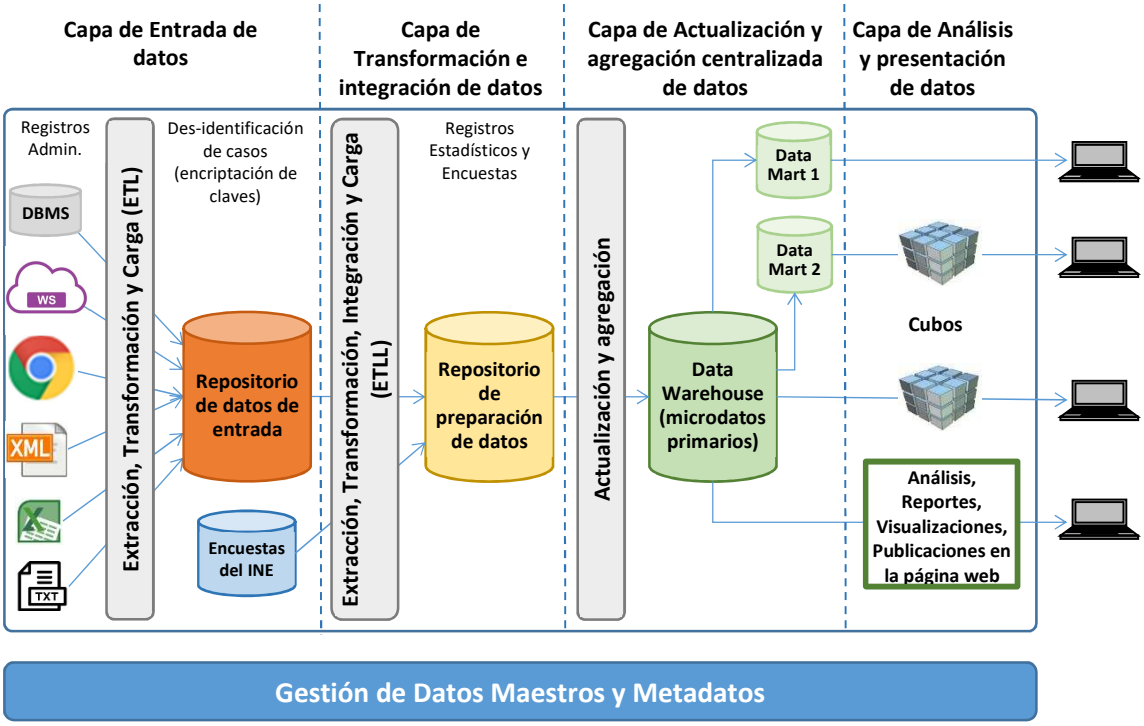
Un DW Geo-Estadístico bien estructurado debería reportar los siguientes beneficios:

- Los datos históricos se mantienen y son accesibles. Es diseñado para proveer almacenamiento de datos históricos a largo plazo.
- Los datos existentes se pueden combinar y reutilizar fácilmente para crear nuevos resultados.
- Diferentes usuarios pueden trabajar en un único entorno común.
- Los datos están estructurados para generar consultas analíticas y presentar la información de manera coherente.
- Proporciona trazabilidad (respecto a los datos originales).
- Flexible ante cambios en sistemas transaccionales de los registros administrativos originales.
- Útil para auditoría, velocidad de carga, y resiliencia a los cambios. Se registran los cambios a nivel de cada variable (fecha-hora carga, fecha-hora vigencia, usuario, fuente del dato).

2.1. Arquitectura del Data Warehouse Geo-Estadístico

La arquitectura del DW está basada en cuatro capas que estructuran los componentes en cuanto a las tecnologías de la información, pero también desde el punto de vista de la gestión de procesos.

Figura 14. Arquitectura del Data Warehouse Geo-Estadístico del INE.



Las cuatro capas del DW:

1. Capa de Entrada de datos

La capa de entrada de datos es el punto de acopio de todos los datos que serán almacenados en el DW, tanto de fuentes internas (encuestas) como externas (registros administrativos).

Cuando se reciben datos externos (registros administrativos) se verifica la completitud y coherencia de los metadatos.

Los datos crudos de los registros administrativos (externos) se almacenan sin modificar, tal cual se reciben, en el repositorio de datos de entrada con la misma estructura con la que llegan.

Se crea o actualiza la ficha de metadatos del dataset de entrada y todas sus variables. Se catalogan todas las variables del dataset.

Esta capa, o más bien el repositorio de datos de entrada, es de acceso restringido al equipo de gestión del DW, pues contiene información personal y confidencial.

2. Capa de Transformación e integración de datos

A partir del repositorio de datos de la capa de entrada de datos, los datos extraídos, transformados y cargados en el repositorio de preparación de datos, mediante procesos ETL diseñados y desarrollados por el equipo de desarrolladores especializados de ETL.

Los procesos de transformación e integración con otras fuentes, tal como se ha descripto en capítulos anteriores, consta de las siguientes actividades:

- Controles de consistencia y calidad.
- Detección y depuración de duplicados.
- Depuración de datos.
- Estandarización de variables.
- Recodificación de variables.
- Creación de variables agregadas/derivadas.
- Creación de objetos/unidades derivadas.
- Unión de registros.

El proceso de transformación de variables es documentado en el sistema de gestión de metadatos *Masterdata*. También se realiza el mapeo de las variables originales del registro administrativo a las correspondientes variables del registro estadístico.

Es obligatorio implementar mecanismos de protección de datos en las tablas de producción estadística, mediante la desidentificación de los registros estadísticos del sistema, donde se sustituyen las variables clave de identificación (cédula de identidad, RUT, etc.) por claves subrogadas. También se encriptan las claves de identificación naturales. Se actualizan las tablas de correspondencia de las claves del negocio (naturales) y las claves subrogadas.

Se actualizan las tablas de "contactos" de personas o empresas, según corresponda. Además de las claves de identificación naturales, se eliminan los identificadores directos (nombres, teléfonos, dirección). Estas tablas de contactos son de acceso restringido, y se utilizan a los efectos de generar marcos de muestro de encuestas. Estos datos están disociados de cualquier información estadística o administrativa, confidencial o sensible.

Esta capa, o más bien el repositorio de preparación de datos, es de acceso restringido al equipo de gestión del DW, pues contiene información personal y confidencial.

3. Capa de Actualización y agregación centralizada de datos

Durante los diversos procesos ETL, es probable que aparezca una variable en varias versiones. Cada vez que un valor se corrige o cambia por alguna razón, el valor anterior no debe borrarse, sino que debe almacenarse una nueva versión de

esa variable. Ese es un mecanismo que se utiliza para garantizar que todos los elementos de la base de datos puedan seguirse a lo largo del tiempo.

Finalmente, los datos se cargan en las tablas de la base de datos del DW.

Esta capa contiene todos los datos recopilados procesados y estructurados para ser optimizados para el análisis. Está especialmente diseñada por expertos en estadística para admitir pruebas de hipótesis, minería de datos y el diseño de cubos multidimensionales.

Su modelo de datos subyacente no es específico para un informe o requisito analítico en particular. En lugar de centrarse en un diseño orientado a procesos, el diseño del repositorio se modela en función de las interrelaciones de datos que son fundamentales para la organización en todos los procesos.

La capa de Actualización y agregación contiene microdatos, hechos observados elementales, agregaciones y valores calculados. Contiene todos los datos al nivel granular más fino para poder cubrir todas las consultas y combinaciones posibles.

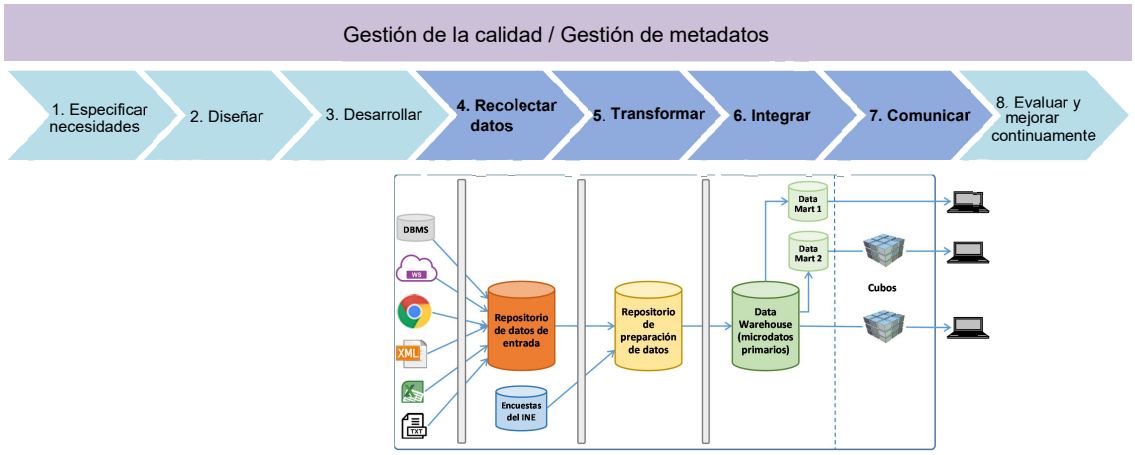
4. Capa de Análisis y presentación de datos

Esta es la capa para la presentación final, difusión y entrega de información. Los datos están optimizados para ser compilados y presentados de forma efectiva. Se pueden presentar en cubos multidimensionales y en diferentes formatos para soportar diferentes herramientas informáticas para el análisis y visualización de datos como R y Tableau.

Esta capa no tiene restricciones de acceso a los datos por parte de los analistas de datos del INE, salvo que no tienen acceso a los datos de contacto ni a las claves de identificación naturales. Las consultas a la base de datos mediante joins de SQL se realizan mediante las claves subrogadas (Id_Estadistico).

Las 4 capas, que se pueden ver también como fases de creación o actualización y de explotación del DW, corresponden a las fases el modelo GSRBPM: 4. Recolectar datos, 5. Transformar, 6. Integrar y 7. Comunicar.

Figura 15. Fases del modelo GSRBPM que se corresponden con las 4 capas del DW.



2.2. Diseño conceptual del DW

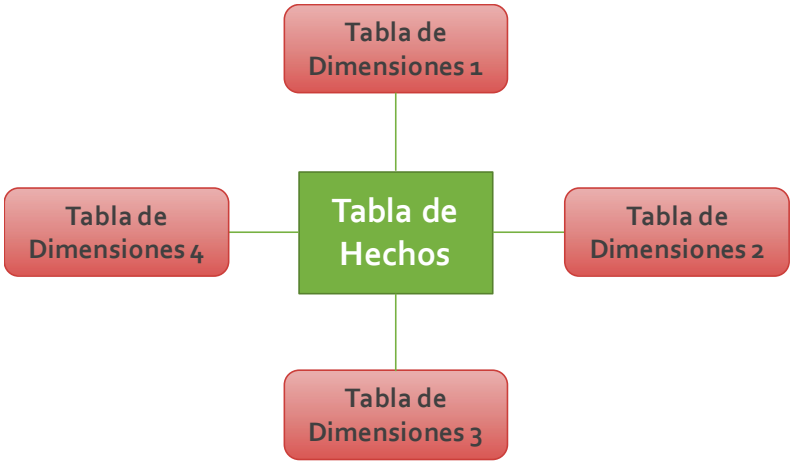
Modelo Estrella (*Kimball*[13]):

El modelo "estrella" propuesto por Kimball coincide con las percepciones de los usuarios, tiene una estructura predecible, estándar que facilita el desarrollo de consultas y análisis. Presenta las siguientes características:

- Puede ser modificado fácilmente.
- Usa perspectivas de modelización comunes.
- Simplifica la agregación.
- De-normalizado (generalmente).
- Tiene caminos de unión bien diseñados.
- Paraleliza la visión de los datos por el usuario.
- Simplifica la comprensión y navegación por los metadatos.
- Amplía las opciones de herramientas de usuario final.

[13] Kimball, R (2008). The Data Warehouse Lifecycle Toolkit (Second Edition), Wiley, USA.

Figura 16. Esquema del modelo Estrella de Kimball.

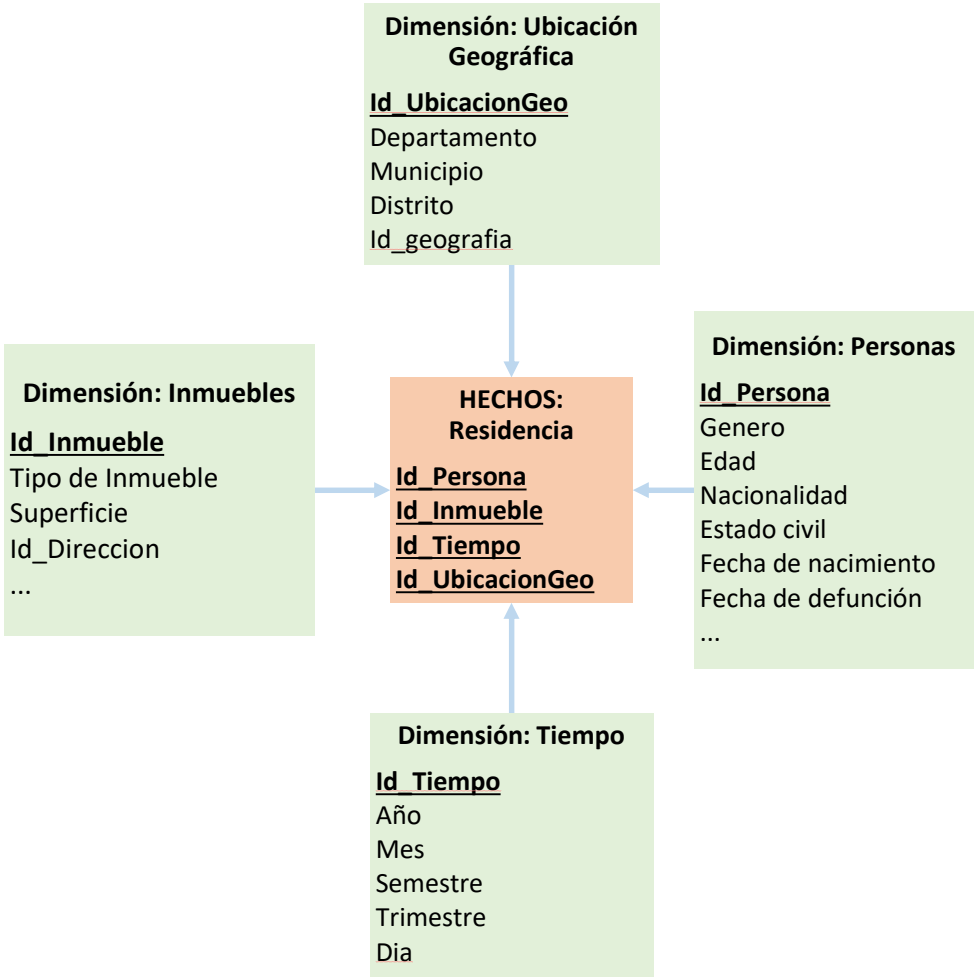


Los Esquemas en Estrella están conformados por dos tipos de tablas: tablas de hechos y tablas de dimensiones.

- **Tablas de Hechos**: contienen datos *cuantitativos* (medidas) sobre el fenómeno de estudio.
    - La clave primaria es una concatenación de claves de dimensión, incluyendo el tiempo.
    - Cada elemento de la clave primaria compuesta es una clave de integridad referencial hacia una tabla de dimensión.
    - Contienen menos atributos, pero muchos más registros.
- **Tablas de Dimensiones:** gestionan datos descriptivos que reflejan las diversas dimensiones del fenómeno de estudio.
    - Contienen muchos atributos pero menos (pocos) registros.
    - La clave primaria 'ayuda' a componer las claves primarias de las tablas de hechos.
    - Sirven para facilitar las consultas de los usuarios (filtros, agregaciones y orden).

A continuación se presenta, a modo de ejemplo, una versión simplificada del diseño lógico del modelo Estrella:

Figura 17. Esquema (simplificado) del modelo estrella de la residencia de las personas.



En el *Anexo II – Implementación del Data Warehouse Geo-Estadístico* se detallan las tablas de hechos y dimensiones diseñadas para cumplir con los requerimientos del SIREE.

Campos (variables) utilizados para la gestión de los datos del DW

Los siguientes campos, que se incorporan en todas las tablas del DW, se utilizan para la gestión de los datos dentro la base de datos del DW.

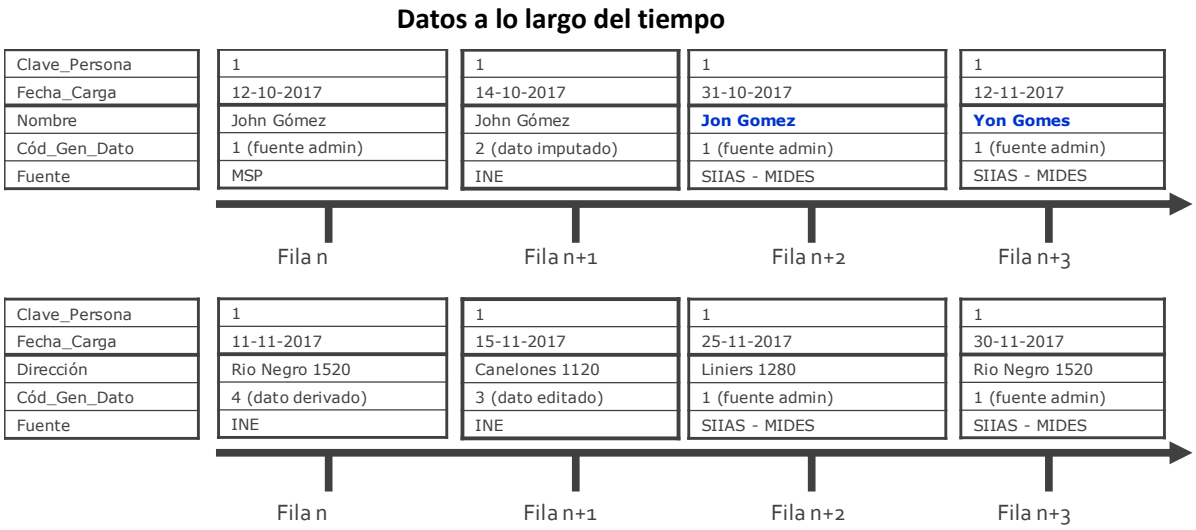
Cuadro 5. Campos para la gestión de los datos del DW.

Nombre del campo	Descripción
Id-Estadistico	Clave (principal) subrogada de las tablas de dimensiones de los registros base. Se utiliza para reemplazar a la clave de negocio (CI, RUT, según corresponda).
Fecha-Vigencia-Desde	Fecha y hora (timestamp) de entrada en vigencia del dato.
Fecha-Vigencia-Hasta	Fecha y hora (timestamp) de fin de la vigencia del dato.

Nombre del campo	Descripción
Fecha-Carga	Fecha y hora (timestamp) en que se carga el dato a la tabla.
Codigo-Generacion-Dato	Código que establece la forma en que se generó el dato (fuente administrativa, cuestionario de encuesta, dato derivado, dato editado, dato imputado)
Fuente	Código de la fuente que provee el dato (fuente administrativa, encuesta).
Version-actual	Número de versión actual del dato.
Usuario	Usuario del sistema que ejecutó el proceso ETL de actualización del dato.

El modelo de DW propuesto provee el valor agregado del tiempo, forma de generación del dato y la fuente.

Figura 17. Esquema que representa los cambios en las variables del DW a lo largo del tiempo y cómo se almacena la información en la base de datos.



2.3. Implementación del Data Warehouse Geo-Estadístico

En el *Anexo II – Implementación del Data Warehouse Geo-Estadístico*, se describe el modelo lógico y físico del DW, así como también los procesos de ETL desarrollados con la herramienta *Pentaho PDI* para la extracción, transformación, integración y carga de datos en la base de datos Oracle del DW.

2.4. Herramientas de ETL

El INE ha adoptado el software libre *Pentaho PDI* como herramienta de ETL para simplificar, sistematizar y (en la mayoría de los casos) automatizar los procesos de extracción, transformación, integración y carga de datos en la base de datos del DW.

En lugar de programar los procesos de extracción, transformación, integración y carga de datos manualmente en SQL, Java o R, las herramientas de ETL como *Pentaho PDI* simplifican el proceso y proveen ciertas ventajas en lugar de programar "a mano":

- Representación visual de flujos de datos en lugar de programación lineal.

  Con las herramientas ETL se diseñan flujos de datos con bloques que representan procesos o actividades de transformación específicas. No es necesario escribir ni una sola línea de código, por lo que no es necesario tener conocimientos de programación en ningún lenguaje. Los desarrolladores ETL se enfocan en las reglas del negocio, en lugar de los detalles del lenguaje de programación.

- Menos errores de programación, mayor eficiencia y performance.

  Los bloques ETL que proveen estas herramientas han sido desarrollados para adaptarse fácilmente a diferentes escenarios y el código que utilizan ha sido debidamente probado. Son más eficientes y logran una mayor performance que el código de programación desarrollado a mano.

- Reutilización de bloques ETL.

  La gestión, soporte y reutilización de bloques ETL es más simple que con la programación manual. Las herramientas de ETL crean una representación visual de un flujo de datos que es mucho más fácil de comprender. Que un solo desarrollador aprenda el código escrito a mano de otro desarrollador puede resultar muy difícil, ni que hablar de reutilizarlo.

- Menores costos de mantenimiento de los bloques ETL.

  Programando manualmente, muchas veces los desarrolladores reescriben el código de otros, porque es más sencillo reescribir que aprender lo que ha hecho otra persona. Esta es la razón por la que los costos de mantenimiento suelen ser el doble que si se programa a mano.

- Más funcionalidades con las herramientas ETL.

  Con las herramientas de integración de datos como Pentaho PDI se obtendrán funcionalidades avanzadas automáticamente, como paralelización, seguimiento y tolerancia a fallos, todo integrado. Si se quisiera obtener las

mismas funcionalidades programando manualmente, necesitaríamos programadores muy calificados que aprendieran todas estas técnicas.

- Escalamiento e innovación.

  Las herramientas ETL facilitan el escalamiento de los procesos inicialmente diseñados para una aplicación puntual y de alcance limitado. Permiten el desarrollo de procesos paso a paso y en forma incremental.

2.5. Metadatos del Data Warehouse Geo-Estadístico

"Los metadatos son el ADN del Data Warehouse, que define sus elementos y cómo funcionan juntos.
[...] Los metadatos juegan un papel tan crítico en la arquitectura del DW que tiene sentido describirla como impulsada por metadatos"[14].

Los metadatos proporcionan acceso a los datos y deben dar una descripción clara e inequívoca de los datos y sus elementos.

El INE utiliza el sistema *MasterData* de Gestión de Datos Maestros y Metadatos, como se ha indicado en apartados anteriores.

2.6. Evaluación de la Calidad del SIREE – DW

El modelo de evaluación de la calidad del SIREE está orientado en las cuatro capas del DW, pues abarca todo el ciclo de vida de los datos, desde el punto de entrada (calidad de entrada) pasando por el proceso de transformación e integración (calidad del proceso) hasta la generación de los productos estadísticos (calidad de salida).

Los elementos de la calidad de entrada a ser evaluados son los atributos propuestos por la herramienta HECRA[15] del Banco Mundial, para las dimensiones de Fuente administrativa de datos, Metadatos y Datos. A los que se le agregarán los atributos de la calidad asociados al proceso de transformación e integración de registros administrativos en estadísticos y los elementos de calidad correspondientes al producto estadístico final.

- *Calidad de entrada*

[14] Kimball (2008), The Data Warehouse Lifecycle Toolkit (Second Edition), Wiley, p. 117
[15] Segui Stagno, Federico (2012). *Guía de la herramienta para la evaluación de la calidad de Registros Administrativos (HECRA) a ser usados con fines estadísticos*. Banco Mundial.

Puesto que los elementos de entrada provienen de un proceso de recolección de datos que está fuera del control del INE, es de suma importancia comprobar la calidad de la fuente de datos administrativos adquirida, teniendo en cuenta el uso previsto de los datos. Se debe aplicar la herramienta HECRA para evaluar la calidad de las dimensiones Fuente de datos administrativa, Metadatos y Datos.

- *Calidad del proceso de transformación e integración*

Las fases y sus correspondientes subprocesos y los posibles errores en unidades y variables se muestran en la siguiente matriz.

Cuadro 6. Fases del modelo GSRBPM y los potenciales errores en unidades y variables.

Fases y subprocesos del modelo GSRBPM	Errores potenciales (unidades)	Errores potenciales (variables)
1. Especificar necesidades 1.1 Identificar necesidades 1.4 Comprobar disponibilidad de datos 2. Diseño	No hay correspondencia en conceptos (sub-cobertura, sobre-cobertura)	Errores de especificación No hay estabilidad conceptual
4. Recolectar datos 4.1 Organizar/preparar recolecta de datos 4.2 Ejecutar recolecta de datos 4.3 Finalizar recolecta de datos	Errores de selección (falta de respuesta por unidad o missings, duplicados, demoras)	Errores de medición
5. Transformar - 6. Integrar 5.3 Depurar y editar 6.1 Integrar registros (unir o emparejar) 6.2 Derivar nuevas variables y unidades	Errores de unión de registros Errores de derivación de unidades	Errores de procesamiento (errores de clasificación, datos faltantes, errores de consistencia intra e inter fuente, errores implícitos y explícitos de suposición del modelo)

Fuente: adaptado de Brancato, G. y otros (2016).Guidelines for the quality of statistical processes that use administrative data. Istat – Italia.

El resultado de la integración de datos por cualquier método (determinístico, probabilístico, mixto) debe evaluarse tomando en cuenta dos indicadores principales: coincidencias falsas (falsos positivos) y falsas no coincidencias (falsos negativos), para lo cual se debería seleccionar una muestra de casos a evaluar.

- *Calidad de salida o del producto estadístico final*

Se debe aplicar la herramienta HECRA para evaluar la calidad de la dimensión Producto estadístico final.

2.7. Data Marts

Un Data Mart es un conjunto de datos para un área temática en particular que se crea con la arquitectura del data warehouse.

Los datos contenidos en el DW están a un nivel muy detallado, desagregado, mientras que los datos que contiene un data mart podrían estar a un nivel más resumido o agregado.

En general, los data marts se crean para resolver consultas particulares sobre determinado tema y representan una solución más simple. También se desarrollan a los efectos de mejorar la performance de las consultas a la base de datos del DW. Los data marts son también denominados bases de datos de reportes, OLAP o multidimensionales.

Los data marts tienen las siguientes características, según Poe, Klauer y Brobst (1998):

- Implementación rápida, simple y de bajo costo.
- Cubre requerimientos de información de un área temática.
- Proporciona protección a la información sensible.
- Mayor performance, debido a que maneja menor volumen de información.

2.7.1. Modelo del negocio del data mart

Metodología:

Los requerimientos del área temática deben establecerse con base en entrevistas con los usuarios del SIREE.

A modo de ejemplo, a continuación se presentan algunos indicadores básicos definidos por los usuarios a ser generados por el sistema SIREE, relativos a estadísticas demográficas (muy) básicas de población e inmuebles:

- Número de personas por sexo y grupos de edad, según nacimiento por áreas geográficas (departamento y localidad).
- Porcentaje de personas por sexo, según nacimiento por áreas geográficas.
- Número de personas por sexo y grupos de edad, según residencia por áreas geográficas (departamento y localidad).
- Porcentaje de personas por sexo, según residencia por áreas geográficas.
- Número de inmuebles por tipo, uso y destino, según construcción en áreas geográficas (departamento y localidad).
- Porcentaje de inmuebles por tipo, según construcción en áreas geográficas.

El siguiente paso es identificar las medidas, dimensiones, granularidad y definiciones y reglas del negocio que permiten generar los indicadores planteados:

- **Medidas** (el atributo varía continuamente):
  - Cantidad de personas nacidas.
  - Cantidad de personas residentes.
  - Cantidad de inmuebles construidos.
- **Dimensiones** (el atributo se percibe como una constante o discreto):
  - Grupo de Edad.
  - Género.
  - Tiempo.
  - Ubicación.
  - Inmueble.
  - Tipo de inmueble.
- **Granularidad** (nivel de detalle al que se desea almacenar información sobre la actividad a modelar. Define el nivel atómico de los datos en el data mart. Determina el significado de las filas de la tabla de hechos. Determina las dimensiones básicas del esquema):

  Dimensiones que caracterizan la actividad al nivel de detalle o granularidad que se ha elegido:
  - Dimensión temporal (cuándo se produce la actividad): Anual, semestral, trimestral, mensual.
  - Dimensión cuál es el objeto de la actividad: Personas, Inmuebles.
  - Dimensión geográfica (dónde se produce la actividad): Departamento o Localidad.
- **Definiciones y reglas del negocio:**
  - Grupos de edad: grupos quinquenales de edad (0 – 4, 5 – 9, 10 – 14, etc.).
  - Ubicación geográfica: Departamento -> Localidad

2.7.2. Modelo dimensional del data mart

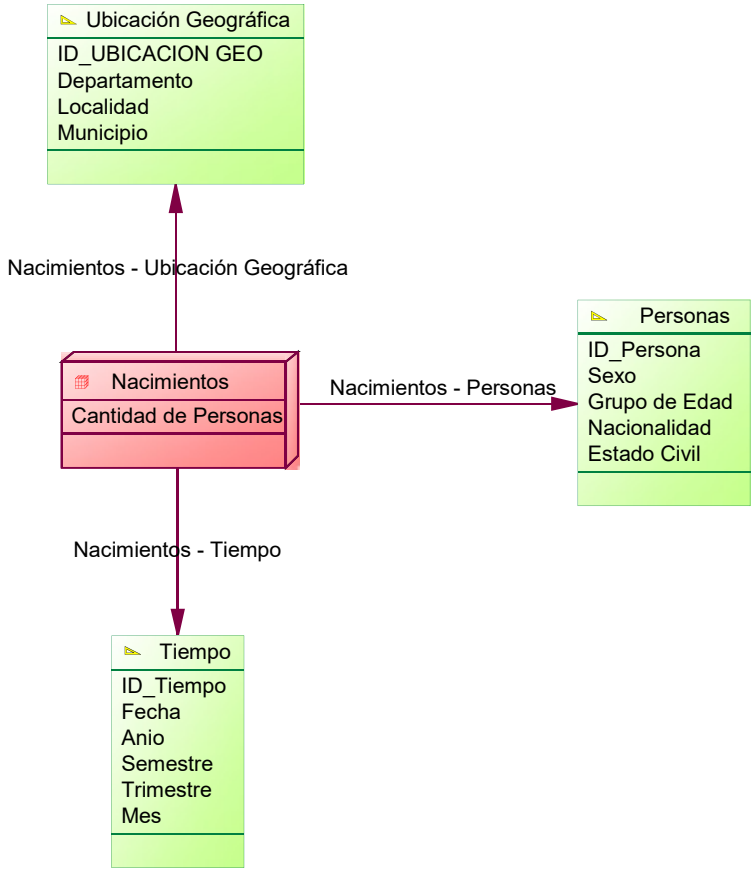
- **Tablas de hechos** (se traducen las medidas de negocio en tablas de hechos):
  - Nacimientos.
  - Residencia.
  - Construcciones.
- **Tablas de dimensiones** (contienen información textual que representa los atributos del negocio o área temática a analizar, almacenan datos

relativamente estáticos y están vinculadas a las tablas de hechos a través de referencia de clave foránea).

- Persona.
- Tiempo.
- Ubicación.
- Inmueble.

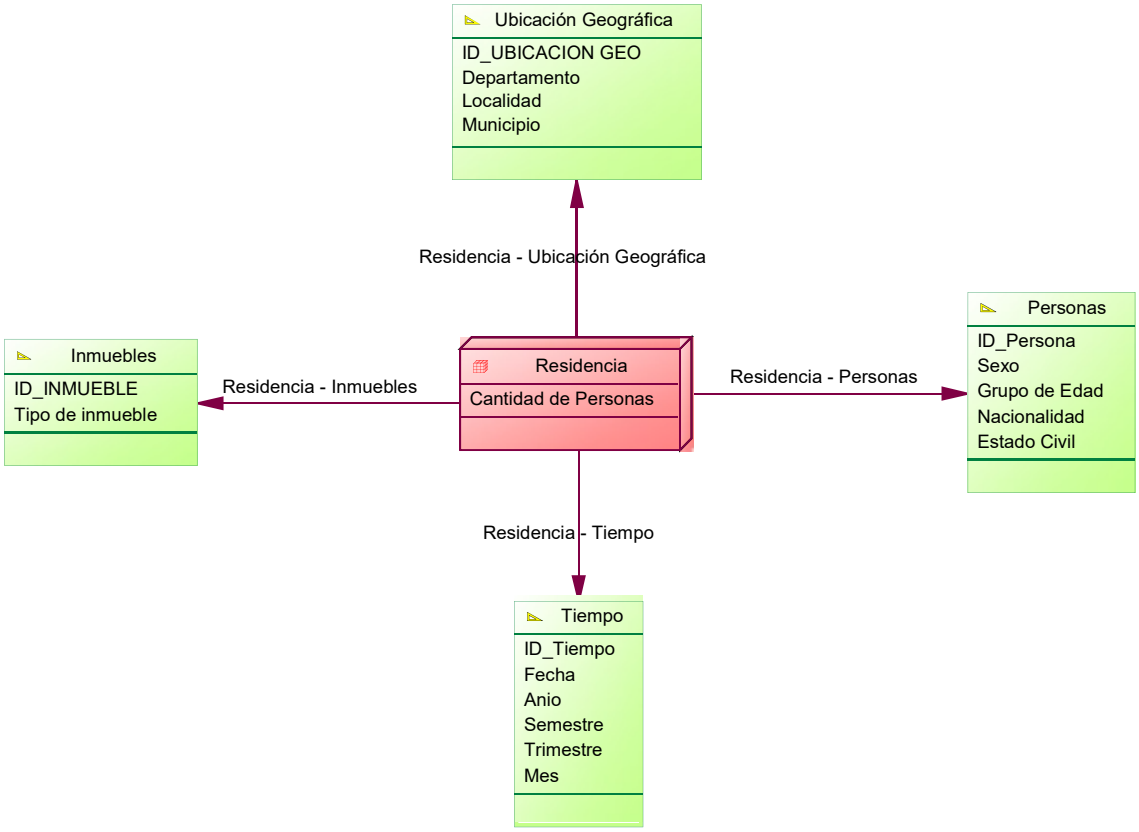
Ejemplo de modelo dimensional de nacimientos (simplificado):

Figura 18. Modelo dimensional de nacimientos (simplificado.



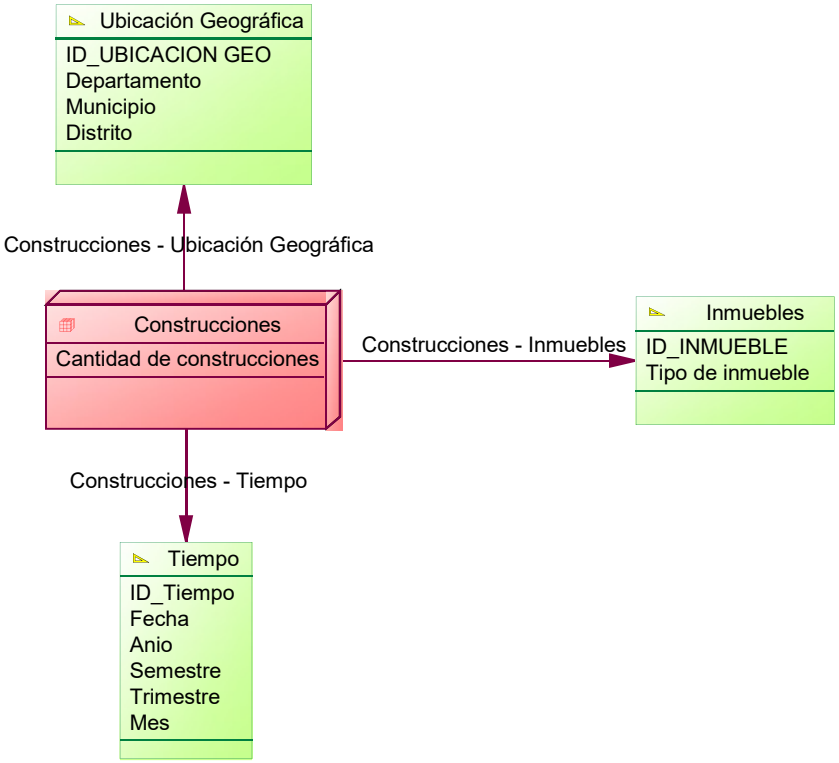
Ejemplo de modelo dimensional de residencia (simplificado):

Figura 19. Modelo dimensional de residencia (simplificado).



Ejemplo de modelo dimensional de construcciones (simplificado):

Figura 20. Modelo dimensional de construcciones (simplificado).



3. El INE como administrador de datos (Data Steward) del Sistema Estadístico Nacional

Definición de "Stewardship"[16]*:*

1: el cargo, deberes y obligaciones de un administrador.

2: la realización, supervisión o gestión de algo, especialmente la gestión cuidadosa y responsable de algo confiado al cuidado de uno.

Definición de "Data Stewardship"[14]*:*

1: los roles, funciones y responsabilidades de un administrador de datos.

2: la gestión sistemática, sostenible y responsable de datos para beneficio público.

¿Quiénes son los Data Stewards?[14] Los administradores de datos son líderes o equipos organizacionales capacitados para crear valor público mediante la reutilización de los datos de su organización (y el conocimiento de los datos); identificar oportunidades para la colaboración productiva intersectorial y responder de forma proactiva a las solicitudes externas de acceso funcional a los datos, conocimientos o experiencia. Actúan tanto en el sector público como en el privado, promoviendo la confianza dentro y fuera de su organización.

El término "Data Steward" (administrador o custodio de datos) se ha utilizado hasta ahora, principalmente en un contexto limitado y estrecho de gobernanza y gestión de datos internos, con una fuerte connotación técnica. En realidad abarca un conjunto más amplio de funciones y responsabilidades de administración dirigidas a aprovechar los activos de datos para abordar los desafíos sociales y mejorar la vida de las personas. Dado el creciente llamado de la sociedad a la responsabilidad sobre los datos, incluida la responsabilidad de proporcionar un impacto social, existe la necesidad de redefinir las funciones y responsabilidades de un administrador de datos más allá de la gestión de datos técnicos. Se debe establecer una función ampliada para iniciar de manera responsable los proyectos colaborativos de datos y alinear a todas las partes interesadas en torno a los objetivos de la colaboración en datos de una manera rápida y ágil.

Los administradores de datos (Data Stewards) tienen **tres responsabilidades**:

1. **COLABORAN**, trabajando con otras áreas (internas y externas) para hacer aparecer el valor inherente de los datos cuando existe un caso de uso claro.
2. **PROTEGEN** a los usuarios, intereses institucionales y al público en general de los daños que pudieran derivarse de compartir o usar los datos; es decir, gestionan los datos de forma responsable.

[16] GovLab (2020). Wanted Data Stewards. GovLab. https://thegovlab.org/static/files/publications/wanted-data-stewards.pdf

3. **ACTÚAN**, asegurándose de que las partes relevantes pongan en práctica los conocimientos generados a través del acceso funcional a los datos.

Los administradores de datos, para cumplir con esas tres responsabilidades, deben ocupar **cinco roles principales**:

1. Asociación y compromiso comunitario;

   Encontrar socios potenciales y, al mismo tiempo, informar a los beneficiarios de los conocimientos generados a partir de los esfuerzos.

2. Coordinación interna y participación del personal;

   Coordinar a los actores internamente y obtener su aprobación.

3. Auditoría de datos, ética y evaluación de valor y riesgo;

   Monitorear y evaluar el valor, el potencial y el riesgo de todos los datos almacenados dentro de la organización.

4. Difusión y comunicación de resultados;

   Actuar como la "cara visible" de los proyectos de datos de la institución y comunicar los resultados compartidos a los actores externos.

5. Fomentar proyectos colaborativos de datos sostenibles.

   Trabajar con las partes interesadas para reunir los recursos y el apoyo necesarios para lograr un impacto amplio a largo plazo.

El INE ha adoptado la metodología GovLab[17] sobre proyectos colaborativos de intercambio de datos que se basa en 7 fases:

1. *Comprender la Demanda*. Establecer una iniciativa colaborativa de datos eficaz requiere una comprensión profunda del problema que se debe abordar y la oportunidad que brinda el intercambio de datos entre instituciones.
2. *Explorar la Oferta*. El intercambio de datos entre instituciones requiere no solo la entrega de diversos conjuntos de datos, sino también la colaboración entre personas y organizaciones con diferentes habilidades y normas institucionales. Una vez que la demanda se comprende y articula bien, se debe analizar la oferta de datos y el capital humano para determinar cómo la oferta puede cumplir la demanda.
3. *Diseñar*. ¿Cómo se compartirán y usarán los datos?, evaluación de los riesgos más destacados y los daños probables, creación de una estrategia específica para mitigar esos riesgos y determinar el marco de gobernanza continuo para la iniciativa colaborativa de datos.

[17] The GovLab Data Collaborative Methodology. https://datacollaboratives.org/canvas.html

4. *Establecer Asociaciones.* Definir la base legal y operativa para la colaboración.
5. *Gestionar el Conocimiento.* Realizar análisis de datos responsable.
6. *Decisiones basadas en conocimientos*. Utilizar conocimientos para informar la acción y la toma de decisiones.
7. *Sistematizar*. Iterar e institucionalizar lo que funciona.

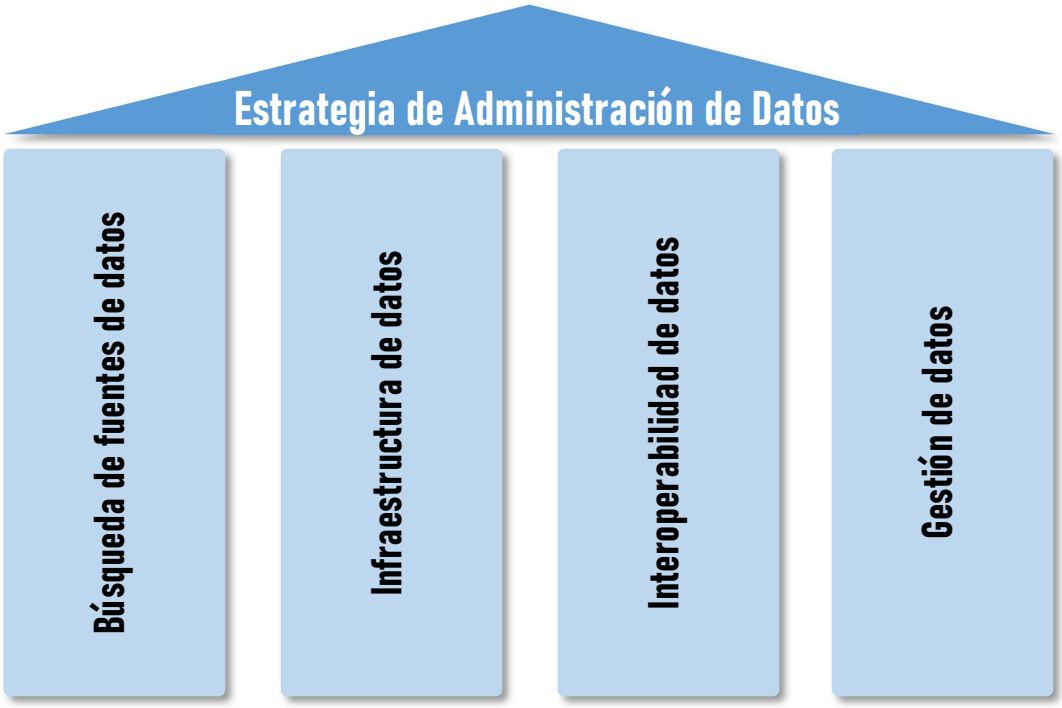
Figura 21. Esquema de la metodología de GovLab sobre proyectos colaborativos de intercambio de datos.

La Estrategia de Administración de Datos del INE

El INE considera los datos como un activo, por lo que la estrategia de administración de datos se centra en asegurar que sean de alta calidad, de fácil acceso y se utilicen de manera eficiente y adecuada.

La Estrategia de Administración de Datos del INE se basa en 4 pilares: búsqueda de fuentes de datos, infraestructura de datos, interoperabilidad de datos y gestión de datos.

Figura 22. Estrategia de Administración de Datos (Data stewardship) del INE.



1. Búsqueda de fuentes de datos

- El INE está trabajando intensivamente en la reutilización de datos administrativos públicos y privados con fines estadísticos.
- Mantener un inventario de registros administrativos (activos de información). Identificar, caracterizar y evaluar la calidad de las fuentes de datos.
- Priorizar el uso de datos administrativos o datos ya recopilados, en primer lugar, en lugar de realizar encuestas.
- Establecer asociaciones estratégicas y relaciones mutuamente beneficiosas con los proveedores de datos.
- Alinear oferta y demanda. Encontrar fuentes de datos para construir indicadores de los ODS.
- Establecer una cultura de reutilización e intercambio de datos para garantizar la eficiencia, especialmente con los activos de información (registros administrativos) del sector público.
- Actualizar el marco legal para facilitar el acceso a más fuentes de datos.
- Establecer procesos para apoyar la adquisición de datos de diferentes fuentes, formatos y múltiples puntos de acceso.

2. Infraestructura de datos

- Existe una necesidad creciente de herramientas para recopilar, almacenar, analizar, administrar, compartir y visualizar datos. Uso extendido de herramientas ETL y BI (Pentaho, Tableau, R).
- Fortalecimiento del SIREE alimentado por fuentes administrativas y estadísticas. Mantener y mejorar el Data Warehouse Geo-Estadístico.
- Fortalecimiento de los recursos humanos, contratación de científicos de datos, analistas de datos, expertos en Data Warehouse y business intelligence y profesionales de ciberseguridad. Crear oportunidades de aprendizaje mutuamente beneficiosas. Crear experiencias laborales prácticas para estudiantes universitarios como fuerza laboral futura y obtener acceso a talentos nuevos y en desarrollo.
- Organizar concursos, como hackatones de datos, de forma regular.
- Implementar la norma ISO 27001 (Sistema de gestión de seguridad de la información).

3. Interoperabilidad de datos

- Promover el uso de la Plataforma de Interoperabilidad del Gobierno para el intercambio de datos entre INE y sus proveedores de datos. Establecer acuerdo con AGESIC.
- Implementar el estándar SDMX para compartir datos y metadatos. Promover su uso en el Sistema Estadístico Nacional.
- Promover la investigación y mejora continua hacia la interoperabilidad total.

4. Gestión de datos

- Organizar y mantener los procesos sobre los datos de acuerdo con las necesidades del ciclo de vida de los datos. Incluye: inventario de datos de registros administrativos, gestión de datos maestros y gestión de la calidad.
- El ciclo de vida de los datos está alineado con el modelo GSRBPM.
- Ampliación del uso del software MasterData para la gestión de datos maestros, la gestión de la calidad de los datos y la gestión de metadatos. MasterData cumple con la norma ISO 11.179. Tecnología de la información: registros de metadatos.

Siguientes pasos:

De aquí en adelante vemos al INE ocupando el rol de Data Steward del Sistema Estadístico Nacional, coordinando proyectos colaborativos de intercambio de datos.

Además, el INE podría convertirse en el concentrador de datos (Data Hub) del SEN. Por ejemplo, recibe de una institución del SEN una solicitud para "pegar" uno de sus datasets con datos del DW, el INE los une dentro de su DW y luego los entrega pseudo-anonimizados o da acceso al DW para consulta de los datos unidos a otros registros estadísticos del SIREE.

También, en el marco de la estrategia de Data Stewardship surge la siguiente pregunta *¿cómo gestionar nuevas fuentes de información?* Como por ejemplo, registros administrativos del sector privado, redes sociales, imágenes, sensores, internet de las cosas (IoT), etc.

Son temas que se están abordando en estos momentos en el INE y seguramente serán incorporados en una nueva versión de este documento.

4. Glosario

Aprovechamiento o explotación de Registros Administrativos (con fines estadísticos). “método para generar datos estadísticos, mediante el uso de los sistemas de registro de hechos o sucesos individuales que realizan las dependencias y organismos públicos como parte de su función”. *Fuente: Instituto Nacional de Estadística y Geografía de México (INEGI). DGE. DGAIN. DN. Glosario sobre la Generación de Estadística Básica. Marzo de 2006.*

Archivo de microdatos. Es un arreglo matricial de datos individuales (personas, viviendas, establecimientos, u otras unidades de observación) en medios computacionales, donde las filas representan a cada unidad o individuo y las columnas son las variables que contienen información (datos) sobre las características de la población.

Calidad. Existen varias definiciones de calidad universalmente aceptadas. Algunas de ellas son:
“Propiedad o conjunto de propiedades inherentes a una cosa que permiten apreciarla como igual, mejor o peor que las restantes de su especie”. *Fuente: Real Academia de la lengua española*.
“Grado en el que un conjunto de características inherentes cumple con los requisitos”. *Fuente: Norma ISO 9000*.
“Calidad es adecuación al uso del cliente”. *Fuente: Joseph Juran.*

Captación. Serie de actividades para obtener los datos de cada unidad objeto de registro, siguiendo las estrategias determinadas en los programas y procedimientos de trabajo. *Fuente: Instituto Nacional de Estadística y Geografía de México (INEGI). Captación en Registros Administrativos. Serie: Lineamientos para la generación de estadística básica. 2010.*

Captura de datos. Procedimiento para transformar la información del cuestionario en un archivo electrónico de datos. *Fuente: ONU. (2001); Departamento de Asuntos Económicos y Sociales. División de Estadística. Manual sobre Gestión de Censos de Población y Habitación. Estudios de Métodos. Serie F. No. 83. Naciones Unidas, Nueva York. p: 157*

Codificación. Procedimiento para asignar identificadores numéricos o alfanuméricos a conceptos en un orden establecido. *Fuente: Instituto Nacional de Estadística y Geografía de México (INEGI). Glosario sobre la Generación de Estadística Básica. Marzo de 2006.*

Criterios de validación. Conjunto de reglas de naturaleza conceptual y estadística, que sirven de base para la identificación y solución de los problemas que se presentan en los datos estadísticos. *Fuente: Instituto Nacional de Estadística y Geografía de México (INEGI). Diseño conceptual. Serie: Documentos técnicos para la generación de estadística. 2006.*

Dato administrativo. Valor de una variable correspondiente a una acción, hecho

o evento que forma parte de un Registro Administrativo.

Dato estadístico. Valor cuantitativo de un conjunto específico respecto a una variable, con referencia de tiempo y de espacio. *Fuente: Instituto Nacional de Estadística y Geografía de México (INEGI). Glosario sobre la Generación de Estadística Básica. Marzo de 2006.*

Edición de datos: procedimiento para detectar y corregir datos que están en blanco o no cumplen con ciertas reglas de consistencia de la información.
Según la definición de la Organización para la Cooperación y el Desarrollo Económico (OCDE), la edición de datos es la actividad que tiene por objetivo detectar y corregir errores (inconsistencias lógicas) en los datos. Las técnicas de edición se refieren a ciertos procedimientos y procesos utilizados para detectar y manejar los errores en los datos.

Estadísticas basadas en registros administrativos: elaboración de productos estadísticos (ver definición) a partir del procesamiento y la transformación de registros administrativos en registros estadísticos.

Fuente de datos administrativa o fuente administrativa: es definida por la OCDE como: "la unidad de organización responsable de implementar una regulación administrativa (o grupo de regulaciones), cuyo registro correspondiente de unidades y transacciones se ven como fuente de datos estadísticos". Dependencia u organismo público, privado o mixto responsable del Registro Administrativo y sus oficinas donde se llevan a cabo los procesos de captación y mantenimiento de los datos del registro. El Registro Administrativo es el resultado de la acción de registrar los datos de determinado evento administrativo; mientras que la fuente administrativa se refiere a la organización responsable del RA, los procesos y entorno institucional que involucran al mismo. Ver definición de Registro Administrativo.

Fuente informante: entidad o persona que informa sobre la acción, evento o hecho objeto de registro.

HECRA o Herramienta. Herramienta para la Evaluación de la Calidad de Registros Administrativos a ser usados con fines estadísticos, elaborada por el consultor internacional Federico Segui para el Banco Mundial.

Imputación de datos: procedimiento estadístico para asignar valores a un dato faltante o en sustitución de valores de respuesta no válidos o inconsistentes. El objetivo es reemplazar los valores faltantes, erróneos o inconsistentes utilizando variables auxiliares, por medio de procedimientos estadísticos estandarizados.

Indicador. Es un parámetro o un valor derivado de parámetros, que proporciona información y/o describe el estado de determinado fenómeno, y tiene un significado que se extiende más allá de aquel directamente asociado a cualquier valor paramétrico dado.
"Herramientas para clarificar y definir, de forma más precisa, objetivos e impactos (...) son medidas verificables de cambio o resultado (...) diseñadas para contar con un estándar contra el cual evaluar, estimar o demostrar el progreso (...) con

respecto a metas establecidas, facilitan el reparto de insumos, produciendo (...) productos y alcanzando objetivos". *Fuente: Organización de las Naciones Unidas (ONU). Integrated and coordinated implementation and follow-up of major United Nations conferences and summits. Nueva York, Estados Unidos de América, 10 y 11 de mayo de 1999, p. 18. Consultado en internet en la página www.un.org/documents/ecosoc/docs/1999/e1999-11. 17 de julio de 2011.*
Una de las definiciones más utilizadas por diferentes organismos y autores es la que Bauer dio en 1966: "Los indicadores sociales (...) son estadísticas, serie estadística o cualquier forma de indicación que nos facilita estudiar dónde estamos y hacia dónde nos dirigimos con respecto a determinados objetivos y metas, así como evaluar programas específicos y determinar su impacto". *Fuente: Horn, Robert V. Statistical indicators for the economic and social sciences. Cambridge, University Press, Hong Kong, 1993, p. 147.*

Indicador estadístico. Es un elemento de datos que representa datos estadísticos para determinado período, lugar y otras características. *Fuente: Economic Commission for Europe of the United Nations (UNECE), "Terminology on Statistical Metadata", Conference of European Statisticians Statistical Standards and Studies, No. 53, Geneva, 2000.*

Instrumento de captación. Formato que se utiliza para el registro de los datos a nivel de cada unidad objeto de registro. *Fuente: Instituto Nacional de Estadística y Geografía de México (INEGI). Glosario sobre la generación de estadística básica. Marzo de 2006.*

Marco legal. Conjunto de leyes, reglamentos, políticas y normas que fundamentan jurídicamente los registros que las dependencias y organismos de la administración pública realizan como parte de su función. *Fuente: Instituto Nacional de Estadística y Geografía de México (INEGI). Glosario sobre la generación de estadística básica. Marzo de 2006.*

Medio del formato en que se hace el registro. Tipo de instrumento físico utilizado para contener los formatos. Puede ser de dos tipos: impreso y electrónico. *Fuente: Instituto Nacional de Estadística y Geografía de México (INEGI). Captación en Registros Administrativos. Serie: Documentos técnicos para la generación de estadística básica. 2010.*

Metadatos estadísticos. Información acerca de los datos estadísticos. Los metadatos comprenden datos y otra documentación que describe objetos de una manera formalizada. *Fuente: UNECE. "Terminology on Statistical Metadata", Conference of European Statisticians Statistical Standards and Studies, No. 53, Geneva, 2000*. Los metadatos brindan información sobre los datos estadístico y los procesos de producción y uso de los mismos. *Fuente: UNECE, "Guidelines for the Modelling of Statistical Data and Metadata", 1995.*

Microdato. Cada uno de los datos referentes a cada una de las unidades de observación obtenidos en un proyecto de estadística. *Fuente: Instituto Nacional de Estadística y Geografía de México (INEGI). Procesamiento de la información. Serie: Lineamientos para la generación de estadística básica. 2006.*

Operación estadística: conjunto de procedimientos y actividades llevados a cabo para elaborar uno o varios productos estadísticos.

Población de interés o estudio: son todos los casos o unidades que forman parte del Registro Estadístico que cumplen con un conjunto de características particulares, es decir, ciertas variables tienen determinados valores en común. La población de interés o estudio puede o no coincidir con la totalidad de unidades que componen el Registro Estadístico. Es posible, entonces, definir a la población de interés como un subconjunto del Registro Estadístico, esto va a depender del uso con fines estadísticos que se quiera hacer del mismo.

Población objetivo: ver Población de interés o estudio.

Procesamiento de datos. Serie de actividades para preparar los archivos de datos, asegurándose que sean congruentes y ordenados para su aprovechamiento. *Fuente: Instituto Nacional de Estadística y Geografía de México (INEGI). Glosario sobre la Generación de Estadística Básica. Marzo de 2006.*

Procedimiento documentado. Documento que describe la secuencia de actividades para ejecutar efectivamente un proceso. Además, significa que el procedimiento se ha establecido, documentado, implementado y mantenido.

Proceso: conjunto de actividades mutuamente relacionadas o que interactúan, las cuales transforman elementos de entrada en resultados. *Fuente: ISO 9000 - Sistemas de gestión de la calidad - Conceptos y vocabulario. 2000.*

Producto estadístico: uno o varios indicadores, cuadros estadísticos y/o archivos de microdatos preparados para ser aprovechados con fines estadísticos, ya sea para la toma de decisiones, definición de políticas públicas o análisis de series históricas.

Registro Administrativo (RA): "todo registro resultante de necesidades fiscales, tributarias u otras, creado con la finalidad de viabilizar la administración de los programas de gobierno o para fiscalizar el cumplimento de obligaciones legales de la sociedad". *Fuente: CEPAL, II CEA, 2003: 10.*
"Serie de datos sobre un tipo de sujeto, acción, hecho o evento, obtenidos mediante un proceso de captación, con base en un formato específico ya sea impreso en papel o en medios computacionales, y que realiza una institución pública, bajo un marco de funciones y facultades formalmente establecidas en instrumentos jurídicos, reglamentarios o programáticos". *Fuente: Instituto Nacional de Estadística y Geografía de México (INEGI). Proceso estándar para el aprovechamiento de Registros Administrativos. Serie: Documentos técnicos para la generación de estadística básica. 2010.* Esta última definición será la utilizada en el presente documento.

Registro con fines estadísticos o Registro Estadístico (RE): es el registro consolidado de datos estandarizados y procesados provenientes de uno o más Registros Administrativos, que originalmente no (necesariamente) fueron captados con fines estadísticos, pertenecientes a una o más fuentes de datos administrativos, pero que han sido adaptados para su uso estadístico.

Regulación administrativa. Son los trámites y las formalidades administrativas con que los gobiernos recogen información e intervienen en decisiones económicas individuales. *Fuente: OCDE. Glosario de Términos Estadísticos.*

Responsable del Registro Administrativo. Unidad o cargo responsable del mantenimiento de los datos del Registro Administrativo. Ver también: Fuente de datos administrativos.

Seguridad de la información - Confidencialidad, Integridad y Disponibilidad. La seguridad de la información tiene por objetivo la preservación de la confidencialidad, integridad y disponibilidad de la información. La confidencialidad se refiere a la información que no se revela ni se encuentra a disposición de individuos, organizaciones o procesos no autorizados. La integridad es la propiedad de salvaguardar la exactitud y completitud de la información. La disponibilidad está referida a la información que puede ser accesible y utilizable a pedido de un agente autorizado. *Fuente: ISO 27001:2005 – Sistemas de Gestión de Seguridad de la Información. Requerimientos.*

Sistema integrado de registros estadísticos: Conjunto ordenado de principios y procedimientos que establecen la interrelación e interacción de los registros estadísticos que forman parte del sistema.

Transferencia de datos administrativos: es el envío de los datos del Registro Administrativo al usuario primario, por parte de la fuente administrativa.

Unidad de la población: caso o unidad que forma parte de la población de interés del Registro Administrativo. Las unidades de la población cumplen con un conjunto de características particulares, es decir, ciertas variables tienen determinados valores en común.

Usuario final: es quien recibe finalmente el producto estadístico. Se trata de quien consume la información estadística contenida en la publicación final sobre el fenómeno de estudio basado en el Registro Administrativo. En algunos casos podrá tratarse de un usuario de la misma institución responsable del Registro Administrativo, y en otros casos podrán ser usuarios externos a ésta.

Usuario primario del Registro Administrativo o **productor estadístico**: es el usuario directo -en ocasiones interno- de los datos administrativos. En algunos casos puede tratarse del estadístico que trabaja en la Oficina Nacional de Estadística o Departamento de estadística de la institución responsable del Registro Administrativo, cuyo objetivo es producir indicadores sobre un fenómeno en particular, es decir, quien explota el registro con fines estadísticos.

Validación de datos. Actividades de control realizadas en los campos de un registro en particular de un archivo de datos. Esto incluye la comprobación de cada campo de cada registro para determinar si contiene un dato válido (criterios de validación) y la comprobación de que los datos de todos los campos son coherentes entre sí.

Variable. Una variable es una característica de una unidad observada que puede asumir más de uno de un conjunto de valores a los que se puede asignar una medida numérica o una categoría de una clasificación (por ejemplo, ingreso, edad, peso, etc., y "ocupación", "Industria", "enfermedad", etc.). *Fuente: OCDE. Glosario de Términos Estadísticos.*

Variable estadística. Una variable es un atributo medible de un objeto o unidad estadística. Una variable estadística está definida por el tipo de objeto que presenta la característica (por ejemplo, ingreso para personas e ingreso para hogares son dos variables distintas), por el método de medición y la escala aplicados, y por el momento o período a los que refiere la medición.

5. Bibliografía

Baxter, R. Gu, L. Vickers, D. Rainsford, C (2016). *Record Linkage: Current Practice and Future Directions*. CSIRO Mathematical and Information Sciences. CMIS Technical Report No. 03/83. Canberra, Australia.

Berglund, B., Laureti, A. (2013). *Functional Architecture of the Statistical Data Warehousing*. ESS - NET on micro data linking and data warehousing in production of business statistics.

Berning, M. y otros (2013). *Data Quality Assessment Tool for Administrative Data*. U.S. Census Bureau. USA.

Biemer P., Lyberg L. (2003). *Introduction to survey quality*. Wiley, New York.

Blomqvist, Klas and others (2011). *A strategy to improve the register system to store, share and access data and its connections to a generic statistical information model (GSIM)*. Invited paper. Work Session on Statistical Data Editing of the Conference of European Statisticians – UNECE. Ljubljana, Slovenia, 9-11 May 2011.

Booch, Grady (2006). *The Accidental Architecture*. IEEE Software. Mayo-Junio 2006.
https://pdfs.semanticscholar.org/7fdc/6cde5c79f046e4a33e4fb36210abb041b0c6.pdf

Bowler, C., Lindelauf, M., Dressen, J. (2013). *Recommendations on the Impact of Metadata Quality in the Statistical Data Warehouse*. ESS - NET on micro data linking and data warehousing in production of business statistics.

Brancato, G. y otros (2016).*Guidelines for the quality of statistical processes that use administrative data*. Istat. Italia.

Bycroft, C (2010). *A register-based census: what is the potential for New Zealand?* Wellington. Statistics New Zealand.

Elfeky, M. y otros (2003). *Record Linkage: A Machine Learning Approach, A toolbox, and a Digital Government Web Service*. Computer Science Technical Reports. Paper 1573. Purdue University. USA. http://docs.lib.purdue.edu/cstech/1573

Fellegi I.P., Sunter A.B. (1969) *A theory for record linkage*. Journal of the American Statistical Association 64, 1183-1210. USA.

Fingar, P. Smith, H (2003). *Business Process Management. The third wave*. Meghan-Kiffer Press. USA.

GovLab (2020). *Wanted Data Stewards*. GovLab.

https://thegovlab.org/static/files/publications/wanted-data-stewards.pdf

Goossens, H. ESSnet (2012). *The statistical data warehouse: a central data hub, integrating new data sources and statistical output*. UNECE, Conference of European Statisticians. Seminar on New Frontiers for Statistical Data Collection (Geneva, Switzerland, 31 October-2 November 2012). https://unece.org/fileadmin/DAM/stats/documents/ece/ces/ge.44/2012/mtg2/WP18.pdf

Kimball, R. Caserta, J (2004). *The data warehouse ETL toolkit. Practical techniques for extracting, cleaning, conforming, and delivering data*. Wiley. USA.

Kimball, R (2008), The Data Warehouse Lifecycle Toolkit (Second Edition), Wiley, USA.

Laitila, T., Wallgren, A. Wallgren B. (2011). *Quality Assessment of Administrative Data*. Research and Development – Methodology reports from Statistics Sweden 2011:2. Suecia.

Lundell, Lars-Göran (2013). *Metadata Framework for Statistical Data Warehousing*. ESS - NET on micro data linking and data warehousing in production of business statistics.

Medina, Alejandro. Segui, Federico (2013). *¿Cómo mejorar el desempeño y crear valor público en las oficinas y sistemas estadísticos nacionales en América Latina y el Caribe?* CreateSpace. USA. innovacionestadistica.com

Segui Stagno, Federico (2009). *Implementing ISO 20252:2006 "Market, opinion and social research" in a statistical office*. Special Topic Paper. Actas del 57º Congreso Mundial de Estadística, ISI 2009, Durban - Sudáfrica).

Segui Stagno, Federico (2011). *Certifying the Quality & Information Security Management Systems of the NSO according to the international standards ISO 9001, ISO 20252 and ISO 27001*. Special Topic Paper. Actas del 59º Congreso Mundial de Estadística, ISI 2011, Dublín-Irlanda.

Segui Stagno, Federico (2012). *Guía de la herramienta para la evaluación de la calidad de Registros Administrativos (HECRA) a ser usados con fines estadísticos*. Banco Mundial.

Segui Stagno, Federico y otros (2012). *Quality improvement of administrative registers statistically exploited to generate the indicator-based decision-making system in the State of Yucatan, Mexico*. Actas de la Conferencia sobre Estadísticas Oficiales IAOS 2012. Kiev, Ucrania.

Segui Stagno, Federico (2016a). *Guía para la gestión de inventarios de registros administrativos de entidades del SEN*. innovacionestadistica.com

Segui Stagno, Federico (2016b). *Marco conceptual y metodológico que sustenta el diseño, desarrollo e implementación de un sistema integrado de registros*

estadísticos de población e inmuebles. "Proyecto Estadística de Población e Inmuebles a partir del uso de registros administrativos oficiales en la Comunidad Andina". Cooperación Técnica No Reembolsable No. ATN/OC-14340-RG – Banco Interamericano de Desarrollo.

Segui Stagno, Federico (2017). *Diseño de un sistema integrado de registros estadísticos de población e inmuebles – SIREPI*. "Proyecto Estadística de Población e Inmuebles a partir del uso de registros administrativos oficiales en la Comunidad Andina". Cooperación Técnica No Reembolsable No. ATN/OC-14340-RG – Banco Interamericano de Desarrollo.

UNECE (2007). *Register-based statistics in the Nordic countries. Review of best practices with focus on population and social statistics.* Naciones Unidas. Nueva York y Ginebra, 2007.

UNECE (2011). *Using administrative and secondary sources for official statistics: A handbook of principles and practices*. Naciones Unidas. Nueva York y Ginebra, 2011.

UNECE (2013a). *The Generic Statistical Business Process Model GSBPM v5.0* http://www1.unece.org/stat/platform/display/metis/The+Generic+Statistical+Business+Process+Model

UNECE (2013b). *Generic Statistical Information Model GSIM v1.1*. http://www1.unece.org/stat/platform/display/metis/Generic+Statistical+Information+Model

UNECE (2015). *Common Statistical Production Architecture*. UNECE. http://www1.unece.org/stat/platform/display/CSPA/Common+Statistical+Production+Architecture

Wallgren, A. Wallgren, B. (2012). *Estadísticas basadas en registros. Aprovechamiento estadístico de los registros administrativos*. INEGI.

Wallgren, A. Wallgren, B. (2016). *Hacia un sistema estadístico integrado basado en registros*.

Wallgren, A., Wallgren, B. (2021). Hacia un sistema estadístico integrado y basado en registros. Banco Interamericano de Desarrollo BID. https://publications.iadb.org/publications/spanish/document/Hacia-un-sistema-estadistico-integrado-y-basado-en-registros.pdf

Zhang L.C. (2012). *Topics of statistical theory for register-based statistics and data integration*. Statistica Neerlandica (2012) Vol 66, no.1, pp. 41-63.

Anexo I – Métodos probabilísticos de unión de registros

El proceso de unión probabilística de registros consta de tres fases:

1) **Pre-unión**. Esta etapa implica la depuración de datos y estandarización de variables, como se ha explicado en apartados anteriores.

   Además, en el caso que se utilicen variables alfanuméricas para realizar la unión de registros, éstas presentan una gran cantidad de errores tipográficos o problemas de conversión de caracteres especiales, por lo cual deben normalizarse, es decir, los textos deben ser procesados para homogeneizarlos de forma tal que los métodos de unión de registros utilizados logren una mayor tasa de coincidencias.

   Para esto, se deberían aplicar los siguientes criterios de normalización de textos:

   - Reemplazar las ocurrencias del símbolo ¥ (o el símbolo correspondiente al sistema de codificación utilizado en las bases de datos o archivos) por el carácter Ñ, y el símbolo ◆ por la ñ [18].
   - Sustituir las vocales con acento o diéresis por las respectivas vocales.
   - Se convierten las minúsculas a mayúsculas.
   - Eliminar partículas. (D, DE, DEL, DA, DI, DO, L, LA, LAS, EL, LOS, Y)
   - Reemplazar los ceros por la letra O. Excepto en la variable de direcciones.
   - Eliminar los números. (1, 2, 3, 4, 5, 6, 7, 8, 9). Excepto en la variable de direcciones.
   - Se eliminan los caracteres diferentes a las 27 letras mayúsculas y minúsculas del alfabeto. Se eliminan los guiones, puntos, comas y otros caracteres.

   Las variables de domicilios son un caso particular de variables alfanuméricas y se deben normalizar siguiendo los criterios de estandarización de cada país. En el caso que los domicilios se almacenen en más de una variable (calle, número, complemento, piso, apartamento, etc.) éstas se deberían concatenar en una sola variable siguiendo los mismos criterios de estandarización.

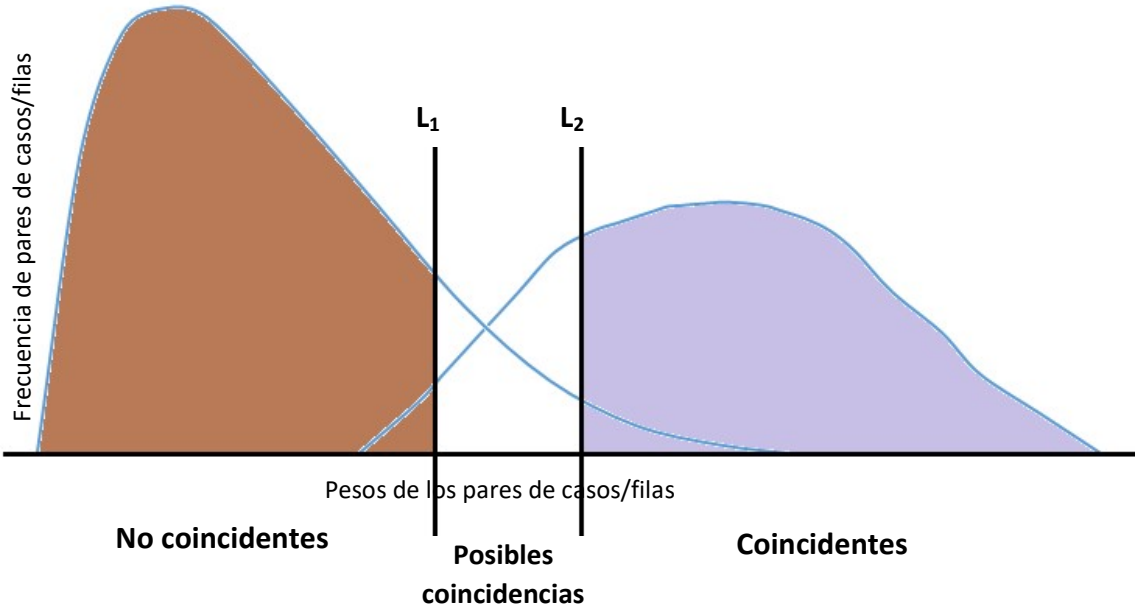
2) **Unión**. Se aplica la unión de registros para decidir cuando dos casos o filas de diferentes registros coinciden (match), o sea pertenecen al mismo objeto o elemento, o no coinciden (no-match) es decir, se trata de diferentes objetos o

[18] Los símbolos ¥ y ◆ podrán variar dependiendo del sistema de codificación o set de caracteres utilizado (UTF-8, ISO-8859, etc.).

elementos. También puede utilizarse para encontrar duplicados en el mismo archivo del registro administrativo. Se utilizan herramientas informáticas que proveen varios algoritmos para realizar la unión probabilística. Estas aplicaciones primero calculan los pesos para cada uno de los posibles casos de unión y luego se determinan los umbrales de los casos unidos y los no unidos. A modo de resumen, uno de los métodos más conocidos y comúnmente aplicado es el método Fellegi-Sunter que sigue estos pasos:

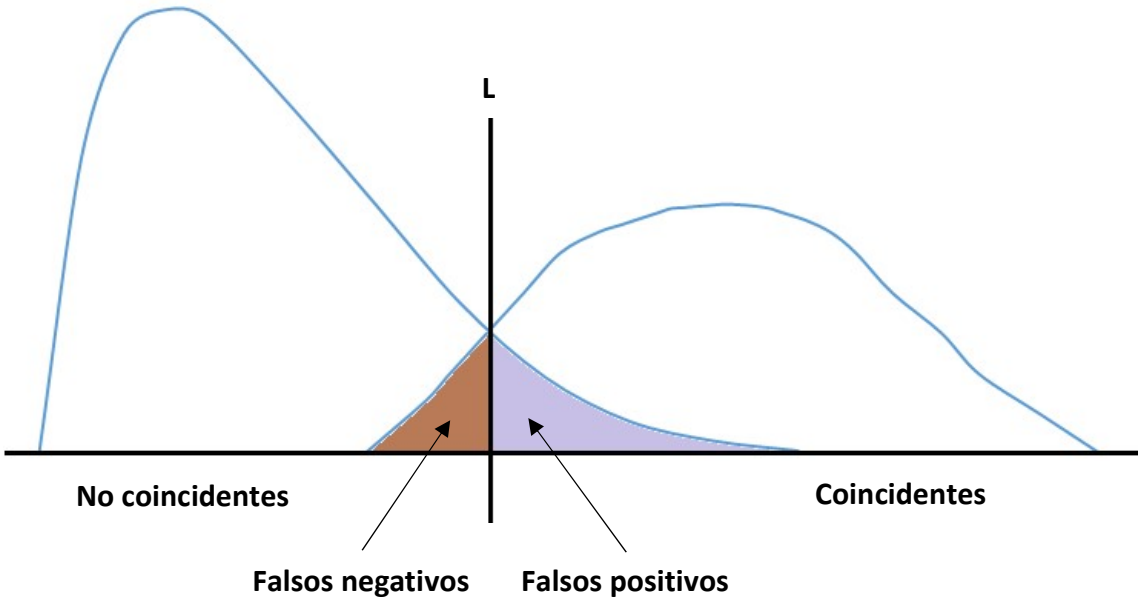
- Cada variable utilizada en el proceso de unión es comparada y se le asigna un puntaje (peso) basado en qué tan bien coincide.
- Se calcula un puntaje para cada campo (variable) que indica, para cada par de casos/filas, cuan probable es que ambos correspondan al mismo objeto o elemento. Este puntaje es basado en la probabilidad de que la coincidencia entre los campos corresponda a una verdadera coincidencia de los objetos o elementos.
  El método
- Se suman los puntajes de todos los campos. El puntaje total de la unión entre dos casos/filas de dos registros es la suma de los puntajes generados de la unión de los campos individuales.
- Se ordenan los pares de casos/filas de acuerdo a sus puntajes (pesos).
- Se establecen ciertos valores de corte de los puntajes para distinguir entre coincidencias (matches) y no coincidencias (no-matches).
- Por encima de cierto umbral de puntaje (peso) se considera como coincidencia (match).
- Por debajo de cierto umbral se considera como no coincidencia (no-match).
- Todo lo que se encuentre entre ambos umbrales se considerará como “posible coincidencia” que necesita ser revisada manualmente (si es que se elige esta alternativa). La siguiente figura ilustra cómo los dos límites o umbrales dividen los pares de casos/filas en tres zonas. La zona de la izquierda del límite L_1 corresponde a los pares de casos/filas clasificados como no coincidentes, en la zona derecha del límite L_2 se encuentran los pares de casos/filas catalogados como coincidentes y en la zona del medio entre los límites L_1 y L_2 se hallan los pares de casos/filas denominados “posibles coincidencias” y los cuales deberán pasar por un proceso de revisión manual para decidir en qué grupo clasificarlos.

Figura 23. Umbrales que delimitan las zonas correspondientes a los tres grupos de pares de casos/filas: no coincidentes, posibles coincidencias y coincidentes.



- Otra opción es seleccionar un único umbral de corte (L) y todos los puntajes que estén por encima serán considerados como coincidencias y los que estén por debajo de ese mismo umbral serán calificados como no coincidentes, sin dejar lugar a "posibles coincidencias" (ver siguiente figura), es decir, todo el proceso será automático.

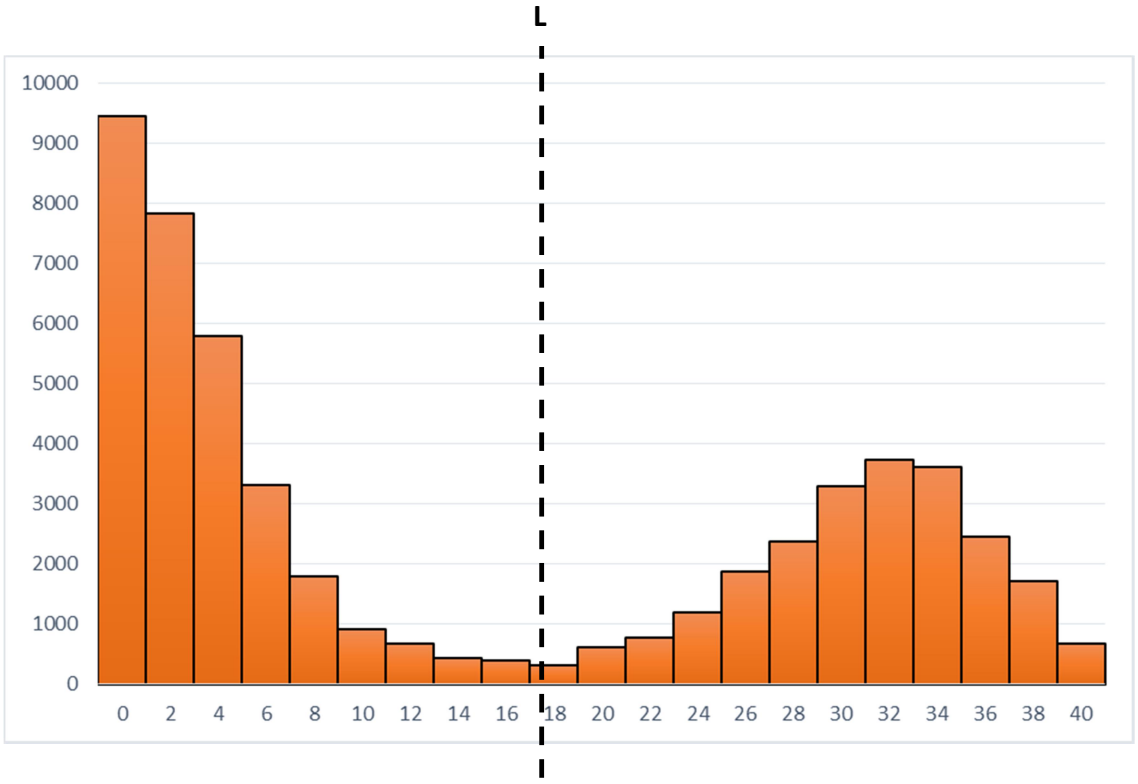
Figura 24. Fijación de un único umbral de corte para dividir en dos grupos de pares de casos/filas: no coincidentes y coincidentes.



- La forma más simple de determinar adecuadamente este límite (L) es mediante la observación del histograma de frecuencias de todos los pesos

totales calculados para todos los pares de casos/filas. La forma que presenta el histograma ayudará a tomar la decisión acerca de dónde ubicar el umbral (L). El siguiente gráfico representa un histograma de frecuencias de todos los pesos totales asignados a los casos/filas.

Gráfico 1. Histograma de frecuencias de los pesos totales de los pares de casos/filas, utilizado para la fijación de un único umbral de corte para dividir los pares en dos grupos: no coincidentes y coincidentes.



Los pares de casos/filas correspondientes al mismo objeto o elemento tienen, en general, un mayor peso, mientras que los pares que no corresponden al mismo objeto, elemento o individuo deberían tener un menor peso.

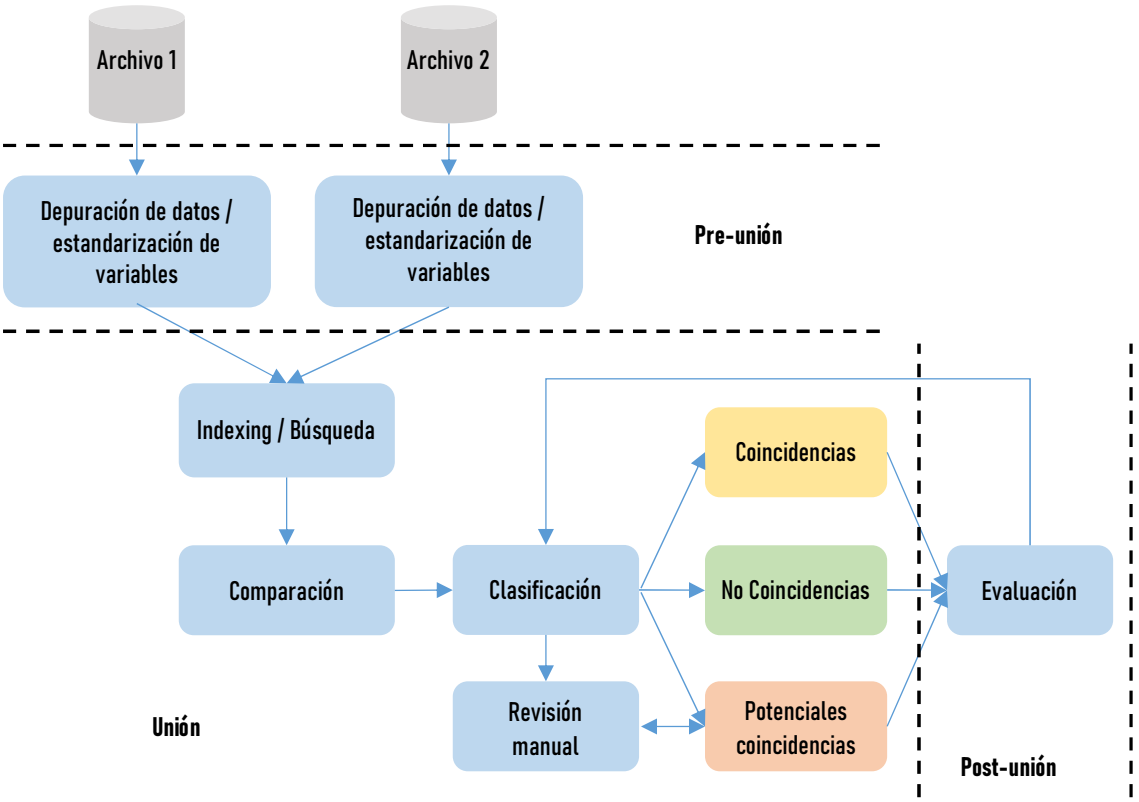
- Es así que mediante la observación del histograma de frecuencias se puede establecer el valor del umbral (L) y todos aquellos pares de casos/filas que tienen un peso superior a este límite se catalogan como coincidentes y aquellos con pesos por debajo del límite serán considerados como no coincidentes.

Otros métodos probabilísticos de unión de registros son Damerau-Levenshtein, Needleman Wunsch, Jaro-Winkler, Pair letters Similarity, Metaphone, SoundEx.

3) **Post-unión**. Revisión manual de los casos/filas no vinculados (si se opta por esta estrategia). Evaluación de los resultados.

El diagrama de procesos de la siguiente figura representa esquemáticamente el proceso de unión de registros.

Figura 25. Diagrama resumen del proceso de unión de registros.



La unión probabilística de registros presenta los siguientes desafíos:

- No hay identificadores únicos.
- Los datos de los registros administrativos contienen errores.
- Archivos de datos demasiado grandes.
- No hay datos de entrenamiento (o testing para hacer ajustes) en muchas aplicaciones de unión de registros.
- Privacidad y confidencialidad.

Para solucionar el problema de manejar archivos de datos demasiado grandes se pueden utilizar técnicas de *blocking*, *indexing* o *filtering*, para trabajar con bloques del archivo de menor tamaño.

La técnica de *indexing* permite fácilmente remover pares de casos/filas que obviamente no coinciden.

El *blocking* solo compara pares de casos/filas que tienen el mismo valor en la variable elegida como variable de blocking. Se recomienda utilizar más de una variable de blocking para minimizar los errores debidos a datos erróneos en dichas variables.

Recientemente se han estado aplicando técnicas de *aprendizaje automático* o *machine learning* para resolver los problemas que presenta la unión de registros.

Los métodos de aprendizaje automático como arboles de decisión, redes neuronales, aprendizaje basado en ejemplos o *instance-based learning*, agrupamiento o *clustering*, entre otros, se utilizan ampliamente para la clasificación de patrones. Un algoritmo de aprendizaje automático construye un modelo a partir de información suministrada en forma de ejemplos, para generalizar comportamientos y reconocer patrones.

Los métodos de aprendizaje automático se clasifican en dos grupos: aprendizaje supervisado (se cuenta con información que especifica qué conjuntos de datos son satisfactorios para el objetivo del aprendizaje) y aprendizaje sin supervisión (encontrar patrones que permitan separar y clasificar los datos en diferentes grupos, en función de sus atributos).

Anexo II – Implementación del Data Warehouse Geo-Estadístico

En este anexo se ha incluida la documentación técnica referida a la implementación del Data Warehouse Geo-Estadístico del INE en la base de datos corporativa Oracle, a través de la herramienta de ETL Pentaho PDI.

Esta documentación es el producto final de la "Consultoría para la implementación de un data warehouse como soporte del sistema integrado de registros estadísticos del INE" financiada por UNFPA, realizada por la consultora María Eugenia Pastor.

Este anexo se estructura en tres módulos que corresponden a cada uno de los tres registros estadísticos base que forman parte del Data Warehouse.

Registros de Población

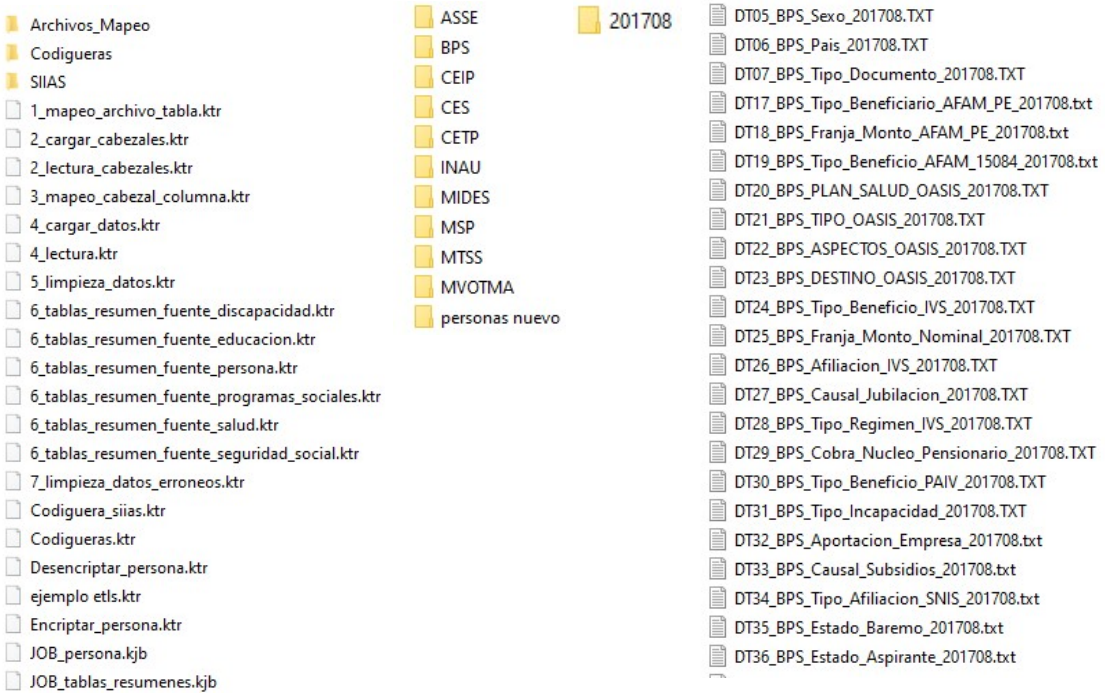
1. Convención de nombres y buenas prácticas

1.1. Estructura de carpetas

Se propone una estructura de carpetas siguiendo la siguiente jerarquía:

1. **ETL Persona:** repositorio donde se almacenarán los ETL. Dentro del mismo se crearán 3 carpetas:
    a. **Archivos_Mapeo**: repositorio donde se almacenarán todos los archivos que se utilizarán para realizar los mapeos del módulo.
    b. **Codigueras**: repositorio donde se almacenarán todos los archivos que se utilizarán como codigueras.
    c. **SIIAS:** repositorio donde se almacenarán los archivos a cargar en el DW. Dentro del mismo estarán las carpetas de los organismos que integran el SIIAS (Sistema de Información Integrada del Área Social) y dentro de ellas habrá una subcarpeta con la fecha. Dentro de las fechas se encontrarán los archivos para cargar, todos con extensión .txt

A continuación, se detalla un ejemplo:



1.2. Nombre y formato de los archivos

Para la ejecución de los ETL se requieren varios archivos de entrada. Se deberá revisar el archivo “Personas_direcciones.txt” y eliminar todas las comillas que presente. También se deberá agregar en el nombre del archivo el año y mes de los datos en el formato AAAAMM.

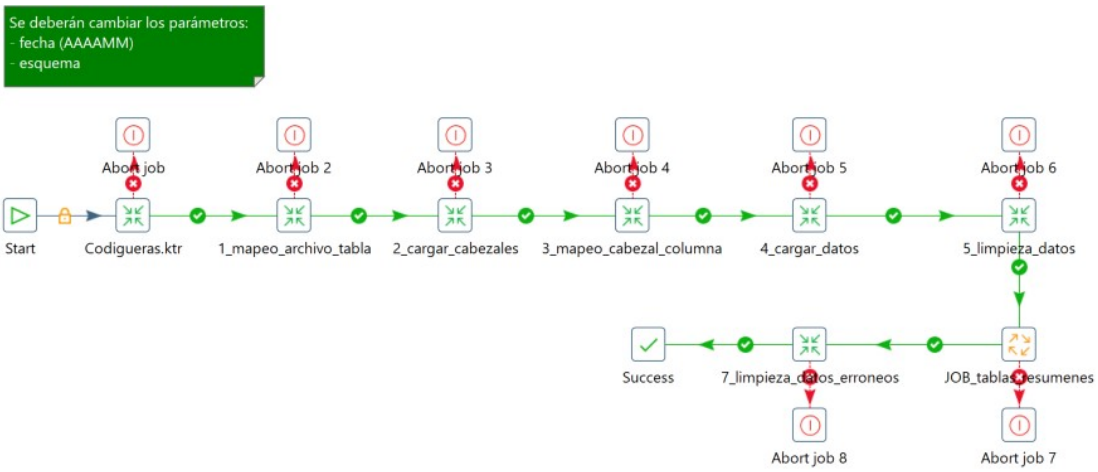
2. Flujo de trabajo

2.1. Comentarios generales

Para realizar el módulo de trabajo, se pensó una metodología lo más genérica posible. La idea era que fuera capaz de leer diferentes tipos de archivos, con diferentes formatos, campos, nombres de columnas, etc. Qué posibilitara la inclusión de nuevos archivos sin la necesidad de modificar los ETL.

2.2. ETL en Pentaho PDI

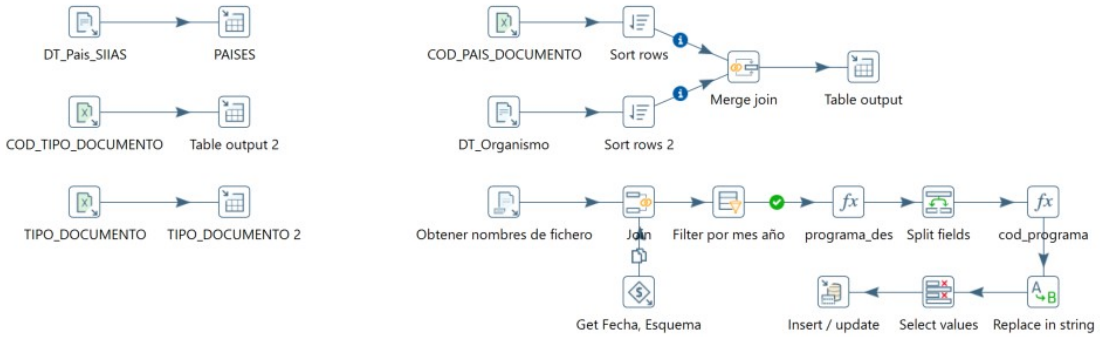
JOB_persona



El JOB de persona tienen definido todos los ETL que se describirán a continuación. Se ejecutan uno a continuación del otro. De presentar algún error se detalla el mensaje de error.

Antes de comenzar a ejecutar los ETL se deberá cambiar el parámetro de fecha por el AAAAMM correspondiente a la carga. Se deben definir o revisar los parámetros de Esquema, Directorio de codiguera, de archivos de mapeo y de SIIAS.

Codigueras



La finalidad de este ETL es cargar las tablas de CODIGUERA_PAISES, CODIGUERA_TIPO_DOCUMENTO, MAPEO_TIPO_DOCUMENTO, MAPEO_PAIS_DOCUMENTO, CODIGUERA_PROGRAMAS.

- La tabla de codiguera_paises se nutre del archivo DT_Pais_SIIAS, donde detalla el código y descripción que se utilizan en el SIIAS para países.
- La tabla codiguera_tipo_documento, es creada para definir los tipos de documentos que se utilizarán en el DW de personas del INE.
- La tabla mapeo_tipo_documento se utiliza para mapear los distintos tipos de documentos que vienen en las fuentes del SIIAS y normalizarlos.
- La tabla mapeo_pais_documento se utiliza para mapear los distintos códigos de países que vienen en las fuentes del SIIAS y normalizarlos.
- La tabla codiguera_programas se nutre de leer todos los archivos existentes en el repositorio del SIIAS y definir los códigos de programa que se utilizarán en el DW de personas del INE.

Codiguera_SIIAS

La finalidad de este ETL es cargar las tablas de codigueras del SIIAS. Fue creado de forma genérica por lo que lee las diferentes hojas del Excel fuente "Codigueras.xlsx" y las carga en las tablas con los nombres correspondientes. Se cargaron aquellas codigueras que se utilizan en las tablas resúmenes.

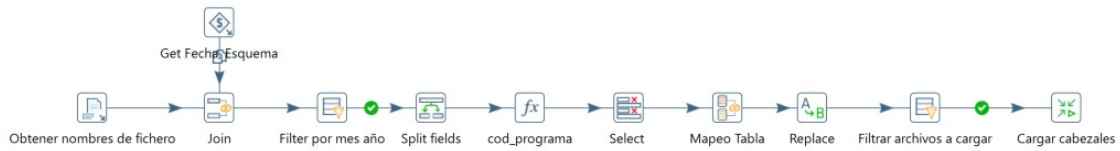
2.2.1. Cargar datos fuentes

1_Mapeo_archivo_tabla

Se definió un archivo Excel "mapeo_archivo_tabla.xlsx" el cual tiene el código del programa (generado con el nombre del archivo) asociado a la tabla del DW en la cual se van a cargar los datos. Dicho mapeo queda registrado en la tabla del DW "ODS_MAPEO_ARCHIVO_TABLA" ejecutando el ETL "1_mapeo_archivo_tabla.ktr".

Cuando se incorpora un archivo nuevo se debe asociar el código del programa a la tabla.

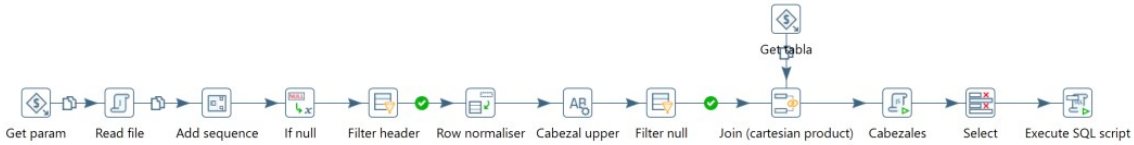
2_Cargar_cabezales



La finalidad de este ETL es obtener las url de todos los archivos cuya terminación es ".txt" dentro de la carpeta SIIAS para luego ejecutar la transformación 2_Lectura_cabezales. Filtra los archivos cuya fecha corresponde a la fecha de carga y fueron mapeados a una tabla del DW en el paso anterior.

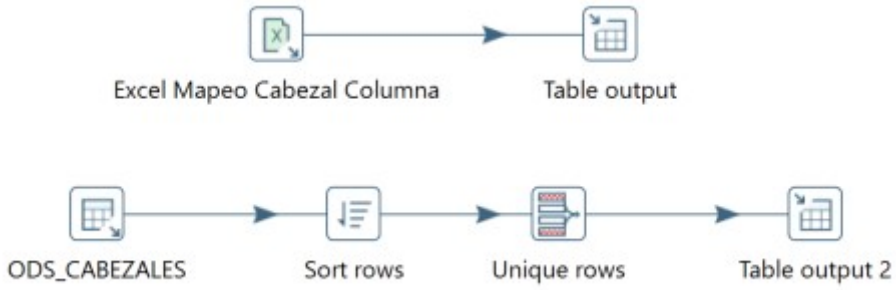
Como parámetros del JOB recibe el nombre del esquema, el año y el directorio donde se encuentran los archivos del SIIAS.

2_Lectura_cabezales



La finalidad de este ETL es leer todos los cabezales de los archivos que fueron previamente mapeados en el ETL anterior. Como resultado genera un registro en la tabla del DW “ODS_CABEZALES” detallando el nombre del cabezal y la tabla a la cual pertenece.

3_Mapeo_cabezal_columna

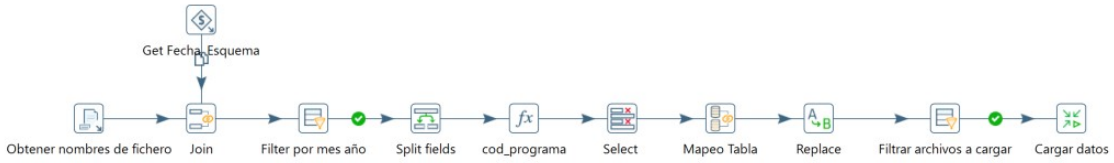


El paso anterior sirve de input para mapear a través de un archivo Excel “mapeo_cabezal_columna.xlsx” los cabezales y las columnas de las tablas fuentes donde se realizará el volcado de información. Dicho mapeo permite eliminar redundancia de datos y homogenizar los nombres de las columnas de las tablas de fuentes que serán utilizadas para la generación de las tablas resúmenes.

Crear nueva tabla con las columnas mapeadas

Con los datos obtenidos de la tabla “ODS_CABEZALES” se crean las nuevas columnas de las tablas de fuentes de las diferentes áreas a trabajar: “Programas sociales”, “Salud”, “Discapacidad”, “Seguridad Social”, “Persona” y “Educación”.

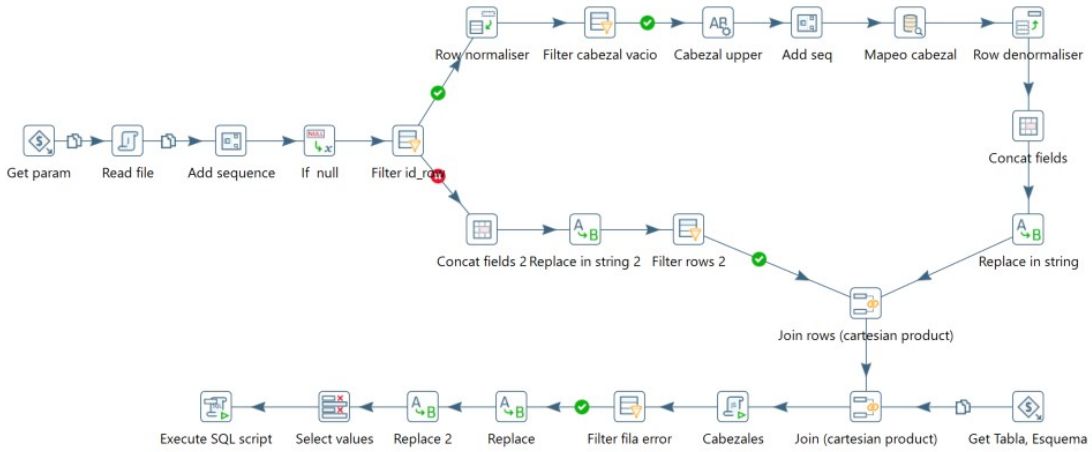
4_Cargar_datos



Una vez finalizada la etapa de preparación de las tablas del DW, se prosigue a realizar el volcado de información utilizando el ETL para leer y cargar todos los archivos fuente.

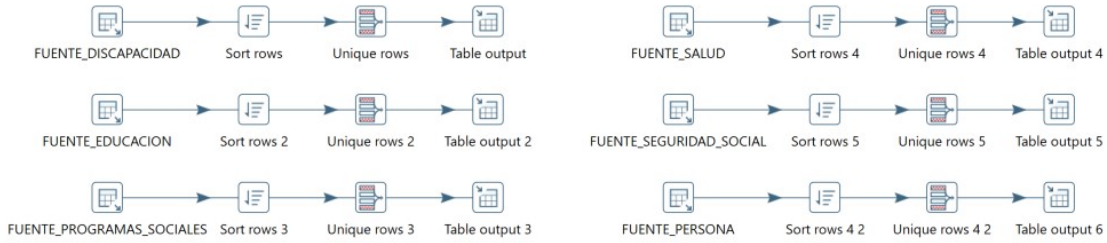
La finalidad de este ETL es obtener las url de todos los archivos cuya terminación es “.txt” dentro de la carpeta SIIAS para luego ejecutar la transformación 4_Lectura. Filtra los archivos cuya fecha corresponde a la fecha de carga y fueron mapeados a una tabla del DW.

4_Lectura



La finalidad de este ETL es leer todos los archivos que fueron previamente mapeados en el ETL anterior. Se cargarán todos los archivos nuevos, que hayan sido mapeados, en sus respectivas tablas fuentes dependiendo el área de interés.

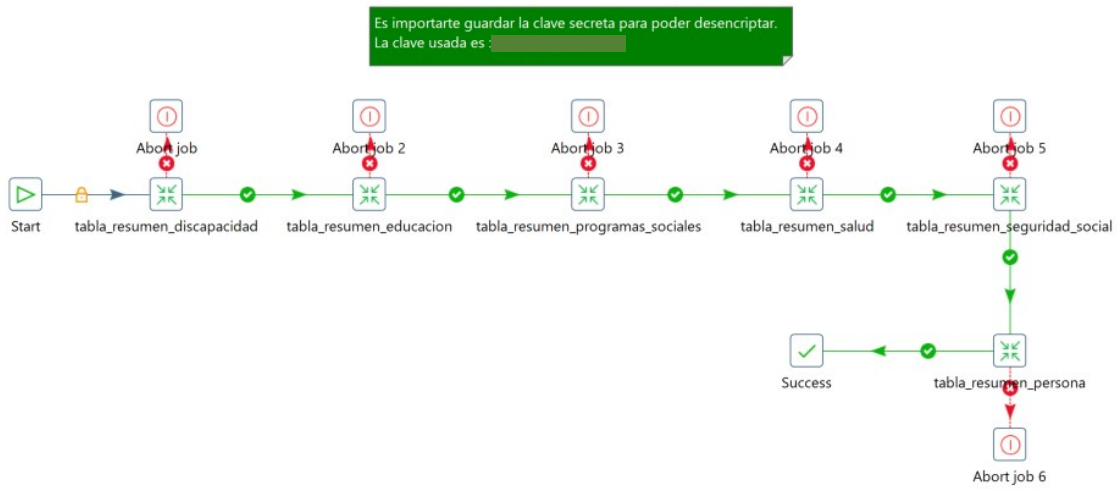
5_Limpieza_datos



La finalidad de este ETL es eliminar repetidos, redundancia, etc de las tablas fuentes. Este paso se realiza por si hubiera algún problema de carga y los datos quedasen repetidos.

2.2.2. Cargar tablas resúmenes

JOB_tabla_resumenes



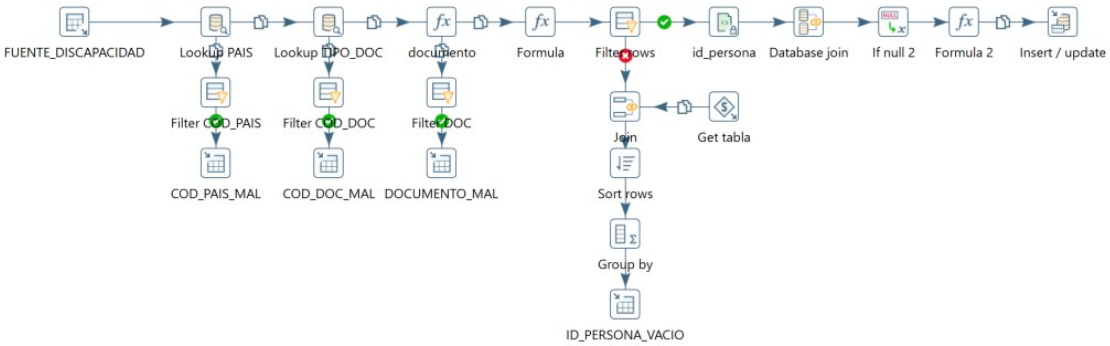
Una vez finalizada la etapa de carga de la información, se prosigue a realizar las tablas resúmenes. El JOB de tablas resúmenes tienen definido todos los ETL de tablas resúmenes que se describirán a continuación de acuerdo con el área de interés. Se ejecutan uno a continuación del otro. De presentar algún error se detalla el mensaje de error.

Todos los ETL de tablas resúmenes tienen mapeado y normalizado los códigos de documento y código de país de documento de la persona, ya que diferentes fuentes generan diferentes códigos para la misma referencia, siguiendo los mapeos provistos por el Ministerio de Desarrollo Social (MIDES).

En los casos donde no se logre mapear ni el código de documento, ni el código de país de documento o que el documento no presente el formato correspondiente, se filtran dichos datos en las tablas COS_PAIS_MAL, COD_DOC_MAL y DOCUMENTO_MAL respectivamente. También se filtra, cuentan y guardan los casos donde el ID de persona quedase vacío en la tabla ID_PERSONA_VACIO. Se indica la tabla de donde proviene el dato. Estas tablas tienen como objetivo la depuración y monitoreo de los datos provenientes de los archivos fuentes.

Los ID de persona fueron creados concatenando las tres variables descriptas anteriormente. Posteriormente se realiza un encriptado de dicha información utilizando el step de Pentaho “Symmetric cryptography”. Es importante recordar que la key utilizada para este paso es la misma que se utilizará para realizar el desencriptado si fuese necesario. Se adjuntan un ejemplo de un ETL para encriptar y uno para desencriptar, detallados más adelante.

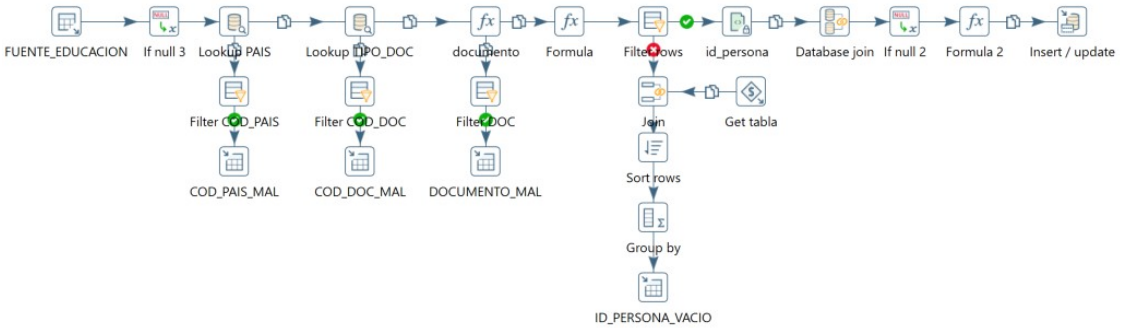
6_tablas_resumen_fuente_discapacidad



La finalidad de este ETL es crear la tabla RESUMEN_DISCAPACIDAD. Las variables incluida en dicha tabla son:

- FECHA_DESDE
- FECHA_HASTA
- PROGRAMA_DISCAPACIDAD
- VERSION
- ID_PERSONA

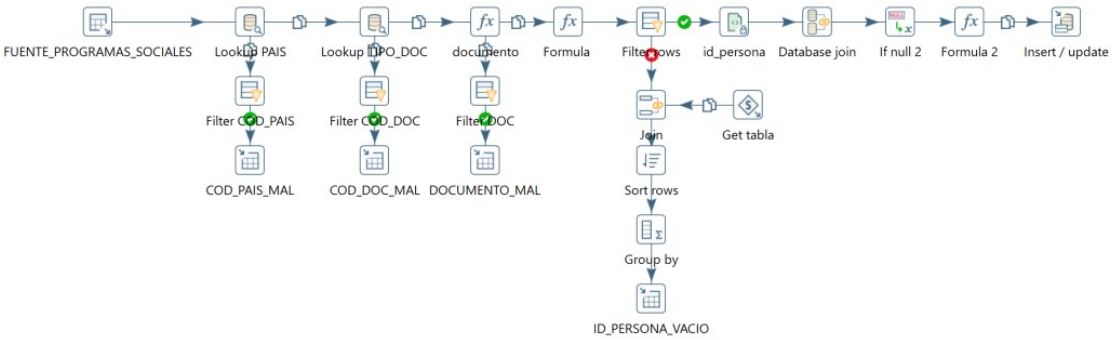
6_tablas_resumen_fuente_educacion



La finalidad de este ETL es crear la tabla RESUMEN_EDUCACION. Las variables incluida en dicha tabla son:

- FECHA_DESDE
- FECHA_HASTA
- COD_PROGRAMA
- MAX_NIVEL
- VERSION
- ID_PERSONA

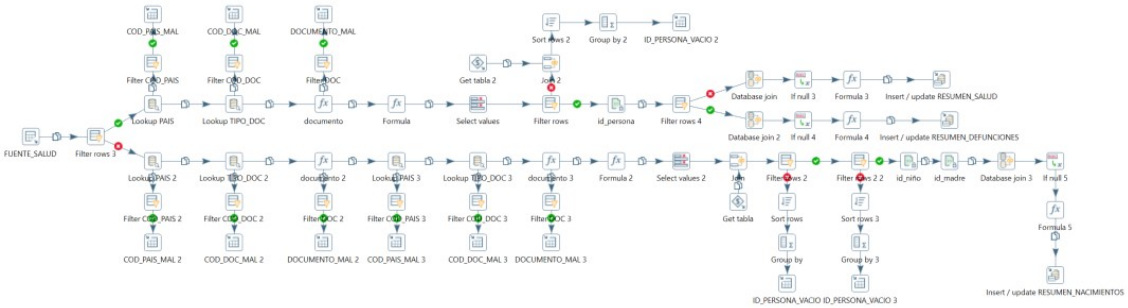
6_tablas_resumen_fuente_programas_sociales



La finalidad de este ETL es crear la tabla RESUMEN_PROGRAMAS_SOCIALES. Las variables incluida en dicha tabla son:

- FECHA_DESDE
- FECHA_HASTA
- PROGRAMA_SOCIAL
- VERSION
- ID_PERSONA
- PRESTACION_INDRA
- ES_TITULAR
- TUS_DOBLE
- MONTO_MENSUAL_HOGAR

6_tablas_resumen_fuente_salud



La finalidad de este ETL es crear 3 tablas: la tabla RESUMEN_SALUD, la tabla RESUMEN_DEFUNCIONES y la tabla RESUMEN_ NACIMIENTOS.

Las variables incluida en la tabla RESUMEN_SALUD son:

- FECHA_DESDE
- FECHA_HASTA
- VERSION
- ID_PERSONA
- COD_PROGRAMA
- TIPO_CARNE_ASISTENCIA
- TIPO_COBERTURA_SANITARIA
- INSTITUCION
- FECHA_REGISTRO

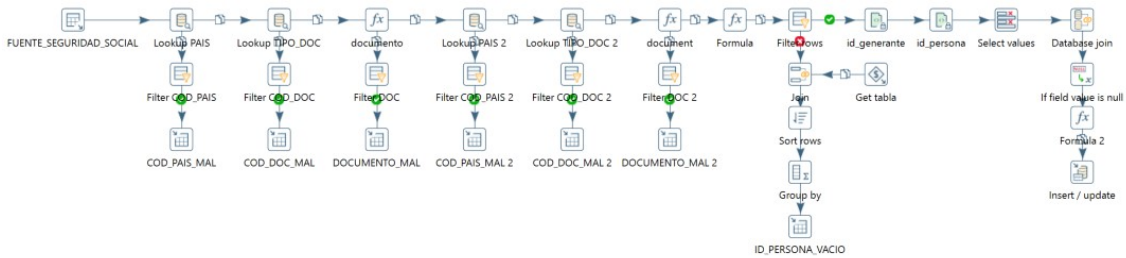
Las variables incluida en la tabla RESUMEN_ DEFUNCIONES son:

- FECHA_DESDE
- FECHA_HASTA
- VERSION
- ID_PERSONA
- COD_PROGRAMA
- FECHA_DEFUNCION
- DEPARTAMENTO_OCURRENCIA
- SECCION_JUDICIAL_OCURRENCIA

Las variables incluida en la tabla RESUMEN_ NACIMIENTOS son:

- FECHA_DESDE
- FECHA_HASTA
- VERSION
- COD_PROGRAMA
- ID_PERSONA_NIÑO
- ID_PERSONA_MADRE
- NIVEL_INSTRUCCION_MADRE
- CANTIDAD_CONSULTAS_PRENATALES
- SEMANA_PRIMER_CONSULTA
- PESO_AL_NACER
- PARTO_MULTIPLE
- ESTABLECIMIENTO
- FECHA
- TIPO_EMBARAZO
- CONVIVENCIA_CON_PADRE
- TIPO_UNION_CON_PADRE
- LUGAR_PARTO
- DIA_HORA_PARTO
- TIPO_PARTO
- SEMANA_GESTACION

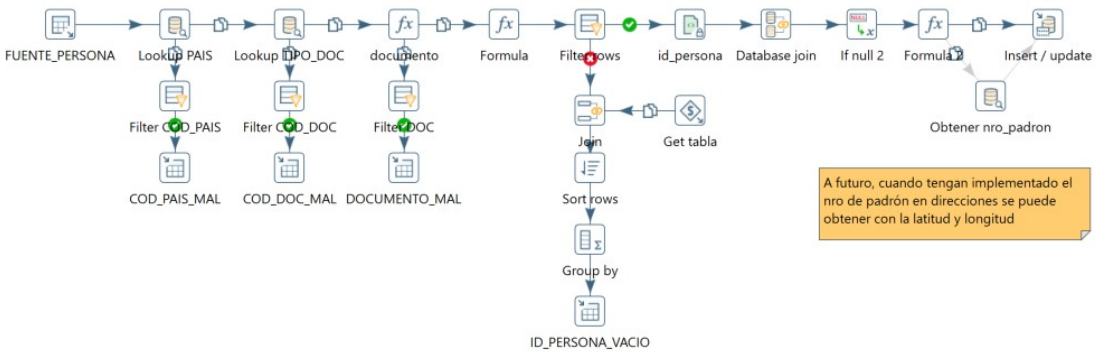
6_tablas_resumen_fuente_seguridad_social



La finalidad de este ETL es crear la tabla RESUMEN_SEGURIDAD_SOCIAL. Las variables incluida en dicha tabla son:

- FECHA_DESDE
- FECHA_HASTA
- COD_PROGRAMA
- VERSION
- ID_PERSONA
- ID_PERSONA_GENERANTE
- TIPO_BENEFICIARIO
- TIPO_BENEFICIO
- FRANJA_MONTO_GENERANTE_MENSUAL

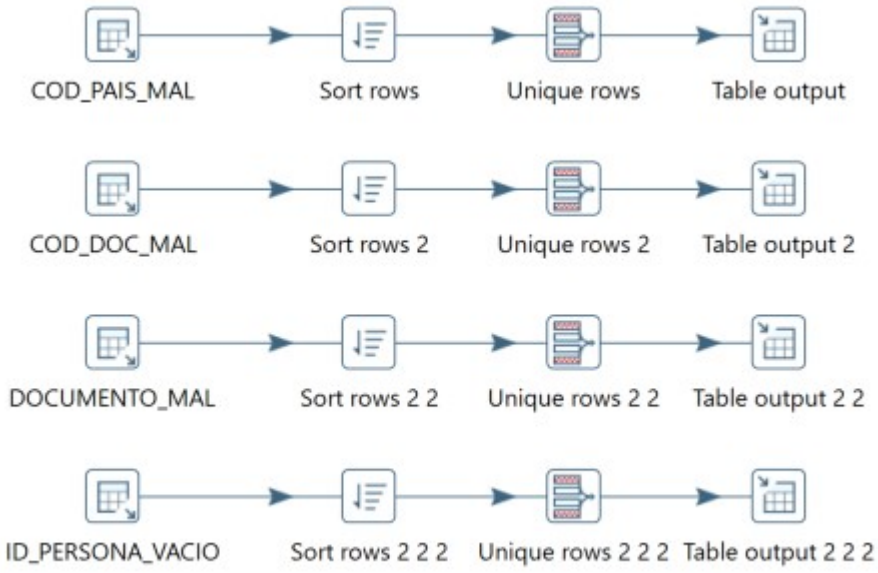
6_tablas_resumen_fuente_persona



La finalidad de este ETL es crear la tabla RESUMEN_PERSONA. Las variables incluida en dicha tabla son:

- FECHA_DESDE
- FECHA_HASTA
- VERSION
- ID_PERSONA
- FECHA_NACIMIENTO
- FECHA_FALLECIMIENTO
- COD_PROGRAMA
- PAIS_NACIMIENTO
- SEXO
- PRIMER_NOMBRE
- SEGUNDO_NOMBRE
- PRIMER_APELLIDO
- SEGUNDO_APELLIDO
- TELEFONO
- LATITUD
- LONGITUD
- ESTADO_CIVIL: No disponible en los archivos fuentes.
- NRO_PADRON: A futuro, cuando INE tenga implementado la tabla de direcciones normalizadas, se podrá mapear la latitud y longitud de la tabla fuente de persona con la de direcciones normalizadas y obtener el número de padrón.

7_limpieza_datos_erroneos



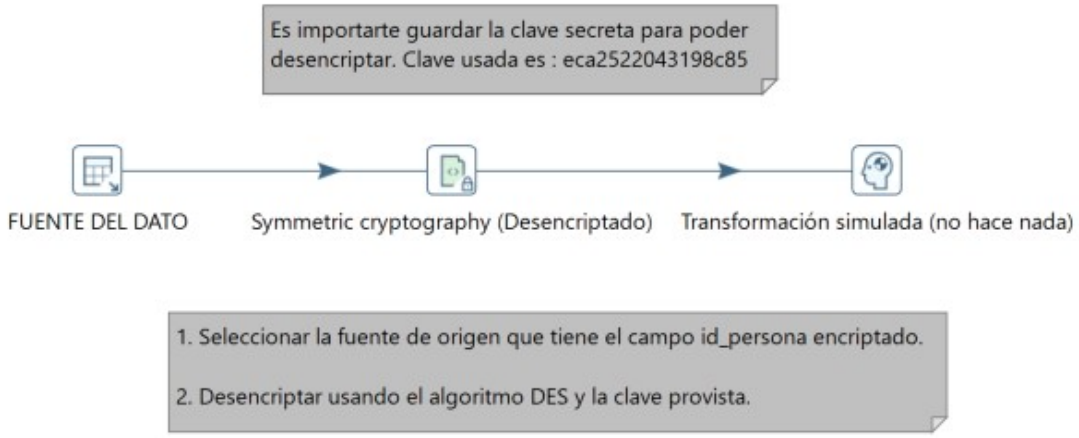
La finalidad de este ETL es eliminar repetidos, redundancia, etc. en las tablas COD_PAIS_MAL, COD_DOC_MAL, DOCUMENTO_MAL y ID_PERSONA_VACIO. Este paso se realiza por si hubiera algún problema de carga y los datos quedasen repetidos.

2.2.3. ETL extras

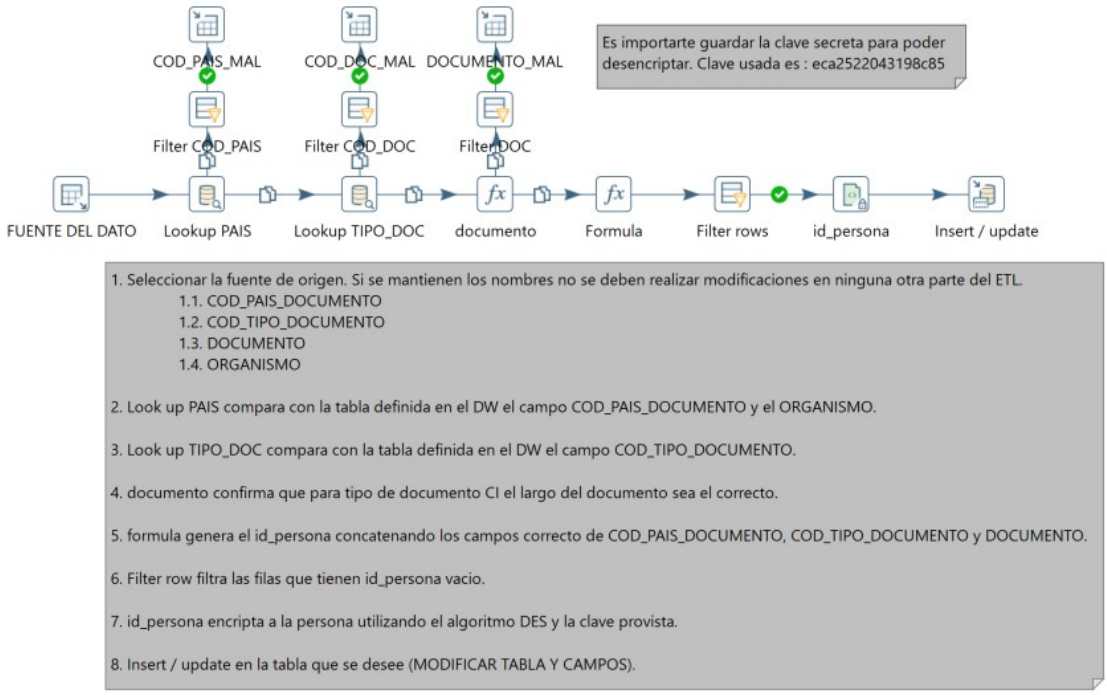
Como se mencionó anteriormente, todos los ETL de tablas resúmenes tienen mapeado y normalizado los códigos de documento y código de país de documento de la persona siguiendo los mapeos provistos por el MIDES.

Los ID de persona fueron creados concatenando las variables: código de documento, código del país de documento y documento. Posteriormente se realiza un encriptado de dicha información utilizando el step de Pentaho “Symmetric cryptography”. La key utilizada para este paso es la misma que se utilizará para realizar el desencriptado si fuese necesario. A continuación, se muestran dos ETL de ejemplo, uno para realiza el encriptado y el otro para desencriptar.

Desencriptar_persona



Encriptar_persona



3. Problemas encontrados

Se puede procesar los archivos que les falten columnas por completar, sin embargo, aquellos que no sigan con el formato de los archivos serán descartados. Por ejemplo, el archivo “FT03_INAU_Club_De_Ninos_201708.txt” no tiene salto de línea entre la fila 3 y 4 y por lo tanto no se puede procesar.

4. Mejoras sugeridas

A continuación, se detallan algunas mejoras sugeridas para realizar una vez finalizado el proyecto o como versión 2 del mismo.

- Utilizar una herramienta centralizada de información.
- Armar un esquema de visualización.
- Armar módulo de carga: carga de archivos a través del módulo y no de forma manual.
- Mapeo de nombre de cabezales a través de un programa y no de forma manual.
- Realizar analítica avanzada.

Dimensiones Generales

DIMENSIONES GENERALES

Tabla de dimensiones: Fecha				
#	Dato	Comentario	Tabla de origen	Observaciones
1	sk_fecha	Fecha formato integer	Autogenereado ETL de fecha	
2	fecha	Fecha formato date	Autogenereado ETL de fecha	
3	anio	Año	Autogenereado ETL de fecha	
4	anio_mes_cod	Código año - mes	Autogenereado ETL de fecha	
5	anio_mes_des	Descripción año - mes	Autogenereado ETL de fecha	
6	semestre_cod	Código semestre	Autogenereado ETL de fecha	
7	semestre_nom	Descripción semetres	Autogenereado ETL de fecha	
8	trimestre_cod	Código trimeste	Autogenereado ETL de fecha	
9	trimestre_nom	Descripción trimeste	Autogenereado ETL de fecha	
10	mes_cod	Código mes	Autogenereado ETL de fecha	
11	mes_nom	Descripción mes	Autogenereado ETL de fecha	
12	diasem_cod	Código día de la semana	Autogenereado ETL de fecha	
13	diasem_nom	Descripción día de la semana	Autogenereado ETL de fecha	

Tabla de dimensiones: ID persona estadístico				
#	Dato	Comentario	Tabla de origen	Observaciones
1	id_persona	Se deberá realizar un mapeo	Autogenereado ETL de Id_persona	
2	Documento & Cod_Tipo_Documento & Cod_Pais_Documento		Autogenereado ETL de Id_persona	

Tabla de dimensiones: Departamento				
#	Dato	Comentario	Tabla de origen	Observaciones
1	Fecha desde	Fecha desde la que el dato es valido		Fecha_Dato
2	Fecha hasta	Fecha hasta la que el dato es valido		Fecha_Dato anterior
3	Cod_Departamento	Departamento	DT08_MSP_Departamento.txt	
4	Código de Departamento	Departamento	DT08_MSP_Departamento.txt	

Tabla de dimensiones: Localidad				
#	Dato	Comentario	Tabla de origen	Observaciones
1	Fecha desde	Fecha desde la que el dato es valido		Fecha_Dato
2	Fecha hasta	Fecha hasta la que el dato es valido		Fecha_Dato anterior
3	Cod_Localidad	Código de Localidad	DT09_MSP_Localidad.txt	
4	Localidad	Localidad	DT09_MSP_Localidad.txt	
5	Cod_Departamento	Departamento	DT09_MSP_Localidad.txt	

Tabla de dimensiones: Sexo				
#	Dato	Comentario	Tabla de origen	Observaciones
1	Fecha desde	Fecha desde la que el dato es valido	DT05_ASSE_Sexo.txt	Fecha_Dato
2	Fecha hasta	Fecha hasta la que el dato es valido	DT05_ASSE_Sexo.txt	Fecha_Dato anterior
3	Cod_Sexo	Código de Sexo	DT05_ASSE_Sexo.txt	
4	Sexo	Sexo	DT05_ASSE_Sexo.txt	

Tabla de dimensiones: Típo de vínculo				
#	Dato	Comentario	Tabla de origen	Observaciones
1	Fecha desde	Fecha desde la que el dato es valido	DT03_ASSE_Tipo_Vinculo.txt	Fecha_Dato
2	Fecha hasta	Fecha hasta la que el dato es valido	DT03_ASSE_Tipo_Vinculo.txt	Fecha_Dato anterior
3	Cod_Tipo_Vinculo	Código de Tipo de Vínculo	DT03_ASSE_Tipo_Vinculo.txt	

4	Tipo_Vinculo	Tipo de Vínculo	DT03_ASSE_Tipo_Vinculo.txt	

Tabla de dimensiones: País				
#	**Dato**	**Comentario**	**Tabla de origen**	**Observaciones**
1	Fecha desde	Fecha desde la que el dato es valido	DT06_ASSE_Pais.txt	Fecha_Dato
2	Fecha hasta	Fecha hasta la que el dato es valido	DT06_ASSE_Pais.txt	Fecha_Dato anterior
3	Cod_Pais_Documento	Código de País del Documento	DT06_ASSE_Pais.txt	
4	Pais	País del Documento	DT06_ASSE_Pais.txt	

Tabla de dimensiones:				
#	**Dato**	**Comentario**	**Tabla de origen**	**Observaciones**
1	Fecha desde	Fecha desde la que el dato es valido		Fecha_Dato
2	Fecha hasta	Fecha hasta la que el dato es valido		Fecha_Dato anterior
3				
4				

Salud

HECHOS

Tabla de hechos: Salud				
#	Dato	Comentario	Tabla de origen	Observaciones
1	Fecha desde	Fecha desde la que el dato es valido	Fecha_Dato	
2	Fecha hasta	Fecha hasta la que el dato es valido	Fecha_Dato anterior	
3	Version	version de actualización		Autonumeral incremental
4	Id_persona	Id estadísitico de la persona	FT01_ASSE_Padron_Beneficiarios_ASSE.txt FT02_ASSE_Aduana.txt FT03_MSP_RUCAF.txt	Documento & Cod_Tipo_Documento & Cod_Pais_Documento Se hace una correspondencia y se crea el Id de persona
5	Asistencia_ASSE	codigo de asistecia (1 =Se asiste, 0 = No se asiste)	FT01_ASSE_Padron_Beneficiarios_ASSE.txt	Variable nueva si se encuentra o no en la tabla FT01_ASSE_Padron_Beneficiarios_ASSE.txt
6	Tipo_Carne_Asistencia	Tipo de Carné de Asistencia	FT01_ASSE_Padron_Beneficiarios_ASSE.txt	Hay codiguera?
7	Plan_Aduana	Código de participación (1 = Asiste, 0 = No asiste)	FT02_ASSE_Aduana.txt	Variable nueva si se encuentra o no en la tabla FT02_ASSE_Aduana.txt
8	Tipo_Cobertura_Sanitaria	Código de tipo de cobertura sanitaria	FT03_MSP_RUCAF.txt	
9	Institucion	Código de institución de cobertura sanitaria	FT03_MSP_RUCAF.txt	
10	Fecha_registro	Fecha de registro de la afiliación en la institución	FT03_MSP_RUCAF.txt	Convertir en integer para unión con la dimensión fecha

Tabla de hechos: Defunciones				
#	Dato	Comentario	Tabla de origen	Observaciones
1	Fecha desde	Fecha desde la que el dato es valido	Fecha_Dato	
2	Fecha hasta	Fecha hasta la que el dato es valido	Fecha_Dato anterior	
3	Version	version de actualización		Autonumeral incremental

4	Id_persona	Id estadísitico de la persona		Documento & Cod_Tipo_Documento & Cod_Pais_Documento Se hace una correspondencia y se crea el Id de persona
5	Fecha_Defuncion	Fecha de Defunción	FT05_MSP_CDEF.txt	Convertir en integer para unión con la dimensión fecha
6	Departamento_Ocurrencia	Departamento de ocurrencia	FT05_MSP_CDEF.txt	
7	Seccion_Judicial_Ocurrencia	Sección judicial de ocurrencia	FT05_MSP_CDEF.txt	Hay codiguera?
8	Zona_Geografica_Ocurrencia	Zona geográfica de ocurrencia	FT05_MSP_CDEF.txt	
9	Cod_Causa_Muerte_cie10	Código de causa de muerte según cie10	FT05_MSP_CDEF.txt	A confirmar si están, y si es codificada
10	Cod_Causa_Tres	Código de causa tres por la cual fallece	FT05_MSP_CDEF.txt	A confirmar si están, y si es codificada

Tabla de hechos: Nacimientos				
#	**Dato**	**Comentario**	**Tabla de origen**	**Observaciones**
1	Fecha desde	Fecha desde la que el dato es valido	Fecha_Dato	
2	Fecha hasta	Fecha hasta la que el dato es valido	Fecha_Dato anterior	
3	Version	version de actualización		
4	Id_persona_niño	Id estadísitico de la persona		Documento & Cod_Tipo_Documento & Cod_Pais_Documento Se hace una correspondencia y se crea el Id de persona
		Deberíamos hacer un id de persona para la madre y alguna tabla de madre - hijo y poner acá el vínculo no el id de la madre. Podría dar problemas		Documento & Cod_Tipo_Documento & Cod_Pais_Documento Se hace una correspondencia y se crea el Id de persona
5	Documento_Madre	Documento de identidad Madre	FT01_MSP_CNV.txt	
6	Cod_Tipo_Documento_Madre	Código de tipo de documento de identidad Madre	FT01_MSP_CNV.txt	
7	Cod_Pais_Documento_Madre	Código de pais del documento de identidad Madre	FT01_MSP_CNV.txt	
8	Nivel_Instruccion_Madre	Nivel de instrucción de la Madre	FT01_MSP_CNV.txt	

9	Cantidad_Consultas_Prenatales	Cantidad de consultas prenatales	FT01_MSP_CNV.txt	
10	Semana_Primer_Consulta	Semana de la primer consulta	FT01_MSP_CNV.txt	
11	Peso_Al_Nacer	Peso al nacer	FT01_MSP_CNV.txt	
12	Parto_Multiple	Parto múltiple	FT01_MSP_CNV.txt	
13	Establecimiento	Establecimiento	FT01_MSP_CNV.txt	
14	Fecha	Fecha	FT01_MSP_CNV.txt	Convertir en integer para unión con la dimensión fecha
15	Tipo_Embarazo	Tipo de embarazo	FT01_MSP_CNV.txt	
16	Convivencia_con_Padre	Convivencia actual con el padre del recién nacido	FT01_MSP_CNV.txt	Hay codiguera?
17	Tipo_Union_con_Padre	Tipo de unión con el padre del recién nacido	FT01_MSP_CNV.txt	
18	Lugar_Parto	Lugar donde ocurrió el parto	FT01_MSP_CNV.txt	
19	Dia_Hora_Parto	Día y Hora del parto	FT01_MSP_CNV.txt	
20	Tipo_Parto	Tipo de parto (Vaginal, Cesarea)	FT01_MSP_CNV.txt	
21	Semana_Gestacion	Semana de gestación	FT01_MSP_CNV.txt	

DIMENSIONES

Tabla de dimensiones: Tipo de carne asistencial				
#	Dato	Comentario	Tabla de origen	Observaciones
1	Fecha desde	Fecha desde la que el dato es valido		Fecha_Dato
2	Fecha hasta	Fecha hasta la que el dato es valido		Fecha_Dato anterior
3				
4				

Tabla de dimensiones: Tipo de cobertura sanitaria				
#	Dato	Comentario	Tabla de origen	Observaciones
1	Fecha desde	Fecha desde la que el dato es valido	DT15_MSP_Tipo_Cobertura_Sanitaria.txt	Fecha_Dato
2	Fecha hasta	Fecha hasta la que el dato es valido	DT15_MSP_Tipo_Cobertura_Sanitaria.txt	Fecha_Dato anterior
3	Cod_Tipo_Cobertura_Sanitaria	Código Tipo Cobertura Sanitaria	DT15_MSP_Tipo_Cobertura_Sanitaria.txt	
4	Desc_Tipo_Cobertura_Sanitaria	Tipo de Cobertura Sanitaria	DT15_MSP_Tipo_Cobertura_Sanitaria.txt	

Tabla de dimensiones: Institución de cobertura sanitaria				
#	Dato	Comentario	Tabla de origen	Observaciones
1	Fecha desde	Fecha desde la que el dato es valido	DT16_MSP_Institucion_Cob_Sanitaria.txt	Fecha_Dato
2	Fecha hasta	Fecha hasta la que el dato es valido	DT16_MSP_Institucion_Cob_Sanitaria.txt	Fecha_Dato anterior
3	Cod_Institucion	Código Institución Cobertura Sanitaria RUCAF	DT16_MSP_Institucion_Cob_Sanitaria.txt	
4	Desc_Institucion	Institución Cobertura Sanitaria RUCAF	DT16_MSP_Institucion_Cob_Sanitaria.txt	
5	Cod_Dpto_Casa_Central_Inst	Código del departamento de la casa central de la institución	DT16_MSP_Institucion_Cob_Sanitaria.txt	

Tabla de dimensiones: Sección Judicial				
#	Dato	Comentario	Tabla de origen	Observaciones
1	Fecha desde	Fecha desde la que el dato es valido		Fecha_Dato
2	Fecha hasta	Fecha hasta la que el dato es valido		Fecha_Dato anterior
3				
4				

Tabla de dimensiones: Zona Geográfica

#	Dato	Comentario	Tabla de origen	Observaciones
1	Fecha desde	Fecha desde la que el dato es valido	DT34_MSP_Zona_Geografica_Fallecim.txt	Fecha_Dato
2	Fecha hasta	Fecha hasta la que el dato es valido	DT34_MSP_Zona_Geografica_Fallecim.txt	Fecha_Dato anterior
3	Cod_Zona_Geografica_Fallecim	Código Zona Geográfica de Fallecimiento	DT34_MSP_Zona_Geografica_Fallecim.txt	
4	Desc_Zona_Geografica_Fallecim	Zona Geográfica de Fallecimiento	DT34_MSP_Zona_Geografica_Fallecim.txt	

Tabla de dimensiones: Nivel de instrucción				
#	Dato	Comentario	Tabla de origen	Observaciones
1	Fecha desde	Fecha desde la que el dato es valido	DT24_MSP_Nivel_Instruccion.txt	Fecha_Dato
2	Fecha hasta	Fecha hasta la que el dato es valido	DT24_MSP_Nivel_Instruccion.txt	Fecha_Dato anterior
3	Cod_Nivel_Instruccion	Código Nivel de Instruccion	DT24_MSP_Nivel_Instruccion.txt	
4	Desc_Nivel_Instruccion	Nivel de Instruccion	DT24_MSP_Nivel_Instruccion.txt	

Tabla de dimensiones: Establecimiento				
#	Dato	Comentario	Tabla de origen	Observaciones
1	Fecha desde	Fecha desde la que el dato es valido	DT31_MSP_Establecimiento_CNV.txt	Fecha_Dato
2	Fecha hasta	Fecha hasta la que el dato es valido	DT31_MSP_Establecimiento_CNV.txt	Fecha_Dato anterior
3	Cod_Establecimiento_CNV	Código Establecimiento CNV	DT31_MSP_Establecimiento_CNV.txt	
4	Desc_Establecimiento_CNV	Establecimiento CNV	DT31_MSP_Establecimiento_CNV.txt	

Tabla de dimensiones: Tipo embarazo				
#	Dato	Comentario	Tabla de origen	Observaciones
1	Fecha desde	Fecha desde la que el dato es valido	DT30_MSP_Tipo_Embarazo.txt	Fecha_Dato
2	Fecha hasta	Fecha hasta la que el dato es valido	DT30_MSP_Tipo_Embarazo.txt	Fecha_Dato anterior
3	Cod_Tipo_Embarazo	Código Tipo de Embarazo	DT30_MSP_Tipo_Embarazo.txt	

4	Desc_Tipo_Embarazo	Tipo de Embarazo	DT30_MSP_Tipo_Embarazo.txt	

Tabla de dimensiones: Tipo de unión				
#	**Dato**	**Comentario**	**Tabla de origen**	**Observaciones**
1	Fecha desde	Fecha desde la que el dato es valido	DT29_MSP_Tipo_Union.txt	Fecha_Dato
2	Fecha hasta	Fecha hasta la que el dato es valido	DT29_MSP_Tipo_Union.txt	Fecha_Dato anterior
3	Cod_Tipo_Union	Código Tipo de Unión	DT29_MSP_Tipo_Union.txt	
4	Desc_Tipo_Union	Tipo de Unión	DT29_MSP_Tipo_Union.txt	

Tabla de dimensiones: Lugar parto				
#	**Dato**	**Comentario**	**Tabla de origen**	**Observaciones**
1	Fecha desde	Fecha desde la que el dato es valido	DT32_MSP_Lugar_Parto.txt	Fecha_Dato
2	Fecha hasta	Fecha hasta la que el dato es valido	DT32_MSP_Lugar_Parto.txt	Fecha_Dato anterior
3	Cod_Lugar_Parto	Código Lugar del Parto	DT32_MSP_Lugar_Parto.txt	
4	Desc_Lugar_Parto	Lugar del Parto	DT32_MSP_Lugar_Parto.txt	

Tabla de dimensiones: Tipo parto				
#	**Dato**	**Comentario**	**Tabla de origen**	**Observaciones**
1	Fecha desde	Fecha desde la que el dato es valido	DT33_MSP_Tipo_Parto.txt	Fecha_Dato
2	Fecha hasta	Fecha hasta la que el dato es valido	DT33_MSP_Tipo_Parto.txt	Fecha_Dato anterior
3	Cod_Tipo_Parto	Código Tipo del Parto	DT33_MSP_Tipo_Parto.txt	
4	Desc_Tipo_Parto	Tipo del Parto	DT33_MSP_Tipo_Parto.txt	

Seguridad Social

HECHOS

Tabla de hechos: Seguridad Social				
#	Dato	Comentario	Tabla de origen	Observaciones
1	Fecha desde	Fecha desde la que el dato es valido	Fecha_Dato	
2	Fecha hasta	Fecha hasta la que el dato es valido		
3	Version	version de actualización		Autonumeral incremental
4	Id_persona	Id estadísitico de la persona beneficiaria		Documento & Cod_Tipo_Documento & Cod_Pais_Documento Se hace una correspondencia y se crea el Id de persona
5	Beneficiario_seguridad_social	Variable que indica el programa de seguridad social del cual la persona es beneficiaria	FT01_BPS_AFAM_PE.txt FT02_BPS_AFAM_15084.txt FT03_BPS_OASIS.txt FT12_BPS_SH_Adulto_Mayor.txt FT13_BPS_AI_Calle.txt FT14_BPS_Beneficios_AP.txt	Codiguera creada
6	Id_persona_generante	Id estadísitico de la persona Generante	FT01_BPS_AFAM_PE.txt	Documento_Generante & Cod_Tipo_Documento_Generante & Cod_Pais_Documento_Generante Deberíamos hacer un id de persona para el generante y hacer alguna tabla de generante - beneficiario y poner acá el vínculo
9	Tipo_Beneficiario	Tipo de Beneficiario de AFAM PE	FT01_BPS_AFAM_PE.txt	
10	Franja_Monto_Generante_Mensual	Franja del monto mensual total que recibe el generante (por todos sus beneficiarios)	FT01_BPS_AFAM_PE.txt	
10	Tipo_Orden_Asistencial	Tipo de orden asistencial (OASIS)	FT03_BPS_OASIS.txt	
12	Fecha_Desde_Vigencia_Orden	Fecha desde vigencia de la orden	FT03_BPS_OASIS.txt	
13	Fecha_Hasta_Vigencia_Orden	Fecha hasta vigencia de la orden	FT03_BPS_OASIS.txt	
14	Jubilado	La persona es jubilada	FT04_BPS_Jubilacion.txt	Variable nueva si se encuentra en la tabla FT04_BPS_Jubilacion.txt

15	Tipo_Beneficio	Tipo de Beneficio	FT04_BPS_Jubilacion.txt	Codiguera de tipo de beneficiario PAIV?
16	Franja_Monto_Nominal	Franja monto nominal	FT04_BPS_Jubilacion.txt	
17	Fecha_Desde_Beneficio	Fecha desde beneficio (fecha alta)	FT04_BPS_Jubilacion.txt	
18	Fecha_Hasta_Beneficio	Fecha hasta beneficio	FT04_BPS_Jubilacion.txt	
19	Afiliacion	Afiliación	FT04_BPS_Jubilacion.txt	Codiguera de tipo de afiliación
20	Tipo_Causal	Tipo_Causal	FT04_BPS_Jubilacion.txt	
21	Tipo_Regimen	Tipo de Régimen	FT04_BPS_Jubilacion.txt	
22	Pension	La persona recibe alguna pensión (1-Fallecimiento, 2-Invalidez, 3-vejez)		Variable nueva si se encuentra en la tabla FT05_BPS_Pension_Fallecimiento.txt o FT06_BPS_Pension_Invalidez.txt o FT07_BPS_Pension_vejez.txt
23	Tipo_Beneficio	Tipo de Beneficio	FT05_BPS_Pension_Fallecimiento.txt FT06_BPS_Pension_invalidez.txt FT07_BPS_Pension_vejez.txt	Codiguera de tipo de beneficiario IVS
24	Franja_Monto_Nominal	Franja monto nominal		
25	Fecha_Desde_Beneficio	Fecha desde beneficio (fecha alta)		
26	Fecha_Hasta_Beneficio	Fecha hasta beneficio		
27	Tipo_Regimen	Tipo de Régimen	FT05_BPS_Pension_Fallecimiento.txt	Se puede unir con la fila 23
28	Vinculo_Parentesco_Fallecido	Vínculo de parentesco con fallecido	FT05_BPS_Pension_Fallecimiento.txt	
33	Tipo_Incapacidad	Tipo de Incapacidad	FT06_BPS_Pension_invalidez.txt	
38	Subsidio	La persona recibe algun subsidio (1-Desempleo, 2-Enfermedad, 3-Maternidad)		Variable nueva si se encuentra en la tabla FT08_BPS_Subsidio_Desempleo.txt o FT09_BPS_Subsidio_Enfermedad.txt o FT10_BPS_Subsidio_Maternidad.txt
39	Franja_Monto_Nominal	Franja monto nominal	FT08_BPS_Subsidio_Desempleo.txt FT09_BPS_Subsidio_Enfermeda	Se puede unir con las filas 26, 27, 28

40	Fecha_Desde_Beneficio	Fecha desde beneficio	d.txt FT10_BPS_Subsidio_Maternidad .txt	
41	Fecha_Hasta_Beneficio	Fecha hasta beneficio		
42	Aportacion_Empresa_Amparante	Aportación de la empresa amparante		Codiguera?
43	Causal	Causal		
54	Id_persona_generante	Documento de Identidad Generante	FT11_BPS_SNIS.txt	Se hace una correspondencia y se crea el Id de persona
55	Cod_Tipo_Doc_Generante_Derecho	Tipo de Documento de Identidad Generante	FT11_BPS_SNIS.txt	
56	Cod_Pais_Doc_Generante_Derecho	Pais del Documento de Identidad Generante	FT11_BPS_SNIS.txt	
57	Tipo_Afiliacion	Tipo de Afiliación	FT11_BPS_SNIS.txt	Se puede unir con la fila 19?
58	Fecha_Desde_Afiliacion	Fecha desde la afiliación	FT11_BPS_SNIS.txt	
59	Fecha_Hasta_Afiliacion	Fecha hasta la afiliación	FT11_BPS_SNIS.txt	
60	Cod_Mutualista_SNIS	Código de Mutualista a la que está afiliada el Beneficiario	FT11_BPS_SNIS.txt	

DIMENSIONES

Tabla de dimensiones: Beneficiario de seguridad social				
#	**Dato**	**Comentario**	**Tabla de origen**	**Observaciones**
1	Fecha desde	Fecha desde la que el dato es valido	Fecha de carga	Fecha_Dato
2	Fecha hasta	Fecha hasta la que el dato es valido	Fecha de carga anterior	Fecha_Dato anterior
3	Cod_beneficiario_seguridad_social	Código del beneficiario de seguridad social	Autogenerado	Detalle en la tabla
4	Des_beneficiario_seguridad_social	Descripción del beneficiario de seguridad social	Autogenerado	

Tabla de dimensiones: Tipo de beneficiario AFAM PE				
#	Dato	Comentario	Tabla de origen	Observaciones
1	Fecha desde	Fecha desde la que el dato es valido	DT17_BPS_Tipo_Beneficiario_AFAM_PE	Fecha_Dato
2	Fecha hasta	Fecha hasta la que el dato es valido	DT17_BPS_Tipo_Beneficiario_AFAM_PE	Fecha_Dato anterior
3	Cod_Tipo_Beneficiario_AFAM_PE	Código Tipo de Beneficiario AFAM PE	DT17_BPS_Tipo_Beneficiario_AFAM_PE	
4	Desc_Tipo_Beneficiario_AFAM_PE	Tipo de Beneficiario AFAM PE	DT17_BPS_Tipo_Beneficiario_AFAM_PE	

Tabla de dimensiones: Tipo orden asistencial y plan de salud (OASIS)				
#	Dato	Comentario	Tabla de origen	Observaciones
1	Fecha desde	Fecha desde la que el dato es valido	DT20_BPS_Plan_Salud_OASIS DT21_BPS_Tipo_OASIS	Fecha_Dato
2	Fecha hasta	Fecha hasta la que el dato es valido	DT20_BPS_Plan_Salud_OASIS DT21_BPS_Tipo_OASIS	Fecha_Dato anterior
3	Cod_Tipo_OASIS	Código Tipo OASIS	DT21_BPS_Tipo_OASIS	
4	Desc_Tipo_OASIS	Tipo OASIS	DT21_BPS_Tipo_OASIS	
5	Cod_Plan_Salud	Código Plan de Salud OASIS	DT20_BPS_Plan_Salud_OASIS DT21_BPS_Tipo_OASIS	
6	Desc_Plan_Salud	Plan de Salud OASIS	DT20_BPS_Plan_Salud_OASIS	

Tabla de dimensiones: Tipo de beneficio PAIV				
#	Dato	Comentario	Tabla de origen	Observaciones
1	Fecha desde	Fecha desde la que el dato es valido	DT30_BPS_Tipo_Beneficio_PAIV	Fecha_Dato
2	Fecha hasta	Fecha hasta la que el dato es valido	DT30_BPS_Tipo_Beneficio_PAIV	Fecha_Dato anterior
3	Cod_Tipo_Beneficio_PAIV	Código Tipo de Beneficio PAIV	DT30_BPS_Tipo_Beneficio_PAIV	
4	Desc_Tipo_Beneficio_PAIV	Tipo de Beneficio PAIV	DT30_BPS_Tipo_Beneficio_PAIV	

Tabla de dimensiones: Afiliación				
#	Dato	Comentario	Tabla de origen	Observaciones

1	Fecha desde	Fecha desde la que el dato es valido	DT34_BPS_Tipo_Afiliacion_SNIS	Fecha_Dato
2	Fecha hasta	Fecha hasta la que el dato es valido	DT34_BPS_Tipo_Afiliacion_SNIS	Fecha_Dato anterior
3	Cod_Tipo_Afiliacion_SNIS	Código Tipo Afiliación SNIS	DT34_BPS_Tipo_Afiliacion_SNIS	
4	Desc_Tipo_Afiliacion_SNIS	Tipo Afiliación SNIS	DT34_BPS_Tipo_Afiliacion_SNIS	

Tabla de dimensiones: Causal de jubilación				
#	**Dato**	**Comentario**	**Tabla de origen**	**Observaciones**
1	Fecha desde	Fecha desde la que el dato es valido	DT27_BPS_Causal_Jubilacion	Fecha_Dato
2	Fecha hasta	Fecha hasta la que el dato es valido	DT27_BPS_Causal_Jubilacion	Fecha_Dato anterior
3	Cod_Causal_Jubilacion	Código de Causal de Jubilación	DT27_BPS_Causal_Jubilacion	
4	Desc_Causal_Jubilacion	Causal de Jubilación	DT27_BPS_Causal_Jubilacion	

Tabla de dimensiones: Tipo de régimen				
#	**Dato**	**Comentario**	**Tabla de origen**	**Observaciones**
1	Fecha desde	Fecha desde la que el dato es valido	DT28_BPS_Tipo_Regimen_IVS	Fecha_Dato
2	Fecha hasta	Fecha hasta la que el dato es valido	DT28_BPS_Tipo_Regimen_IVS	Fecha_Dato anterior
3	Cod_Tipo_Regimen_IVS	Código Tipo de Régimen de IVS	DT28_BPS_Tipo_Regimen_IVS	
4	Desc_Tipo_Regimen_IVS	Tipo de Régimen de IVS	DT28_BPS_Tipo_Regimen_IVS	

Tabla de dimensiones: Pensión				
#	**Dato**	**Comentario**	**Tabla de origen**	**Observaciones**
1	Fecha desde	Fecha desde la que el dato es valido	Fecha de carga	Fecha_Dato
2	Fecha hasta	Fecha hasta la que el dato es valido	Fecha de carga anterior	Fecha_Dato anterior
3	Cod_pension	Código pensión	Autogenerado	Detalle en la tabla
4	Des_pension	Descripción pensión	Autogenerado	

Tabla de dimensiones: Tipo de regimen IVS

#	Dato	Comentario	Tabla de origen	Observaciones
1	Fecha desde	Fecha desde la que el dato es valido	DT28_BPS_Tipo_Regimen_IVS	Fecha_Dato
2	Fecha hasta	Fecha hasta la que el dato es valido	DT28_BPS_Tipo_Regimen_IVS	Fecha_Dato anterior
3	Cod_Tipo_Regimen_IVS	Código Tipo de Régimen de IVS	DT28_BPS_Tipo_Regimen_IVS	
4	Desc_Tipo_Regimen_IVS	Tipo de Régimen de IVS	DT28_BPS_Tipo_Regimen_IVS	

Tabla de dimensiones: Franja monto nominal				
#	Dato	Comentario	Tabla de origen	Observaciones
1	Fecha desde	Fecha desde la que el dato es valido	DT25_BPS_Franja_Monto_Nominal	Fecha_Dato
2	Fecha hasta	Fecha hasta la que el dato es valido	DT25_BPS_Franja_Monto_Nominal	Fecha_Dato anterior
3	Cod_Franja_Monto_Nominal	Código de Franja Monto Nominal	DT25_BPS_Franja_Monto_Nominal	
4	Desc_Franja_Monto_Nominal	Franja Monto Nominal	DT25_BPS_Franja_Monto_Nominal	

Tabla de dimensiones: Vínculo con el parentesco fallecido				
#	Dato	Comentario	Tabla de origen	Observaciones
1	Fecha desde	Fecha desde la que el dato es valido	DT43_BPS_Vinculo_Parentesco_Fallecido	Fecha_Dato
2	Fecha hasta	Fecha hasta la que el dato es valido	DT43_BPS_Vinculo_Parentesco_Fallecido	Fecha_Dato anterior
3	Cod_Vinculo_Fallecido	Código de Vínculo de Parentesco con el Fallecido	DT43_BPS_Vinculo_Parentesco_Fallecido	
4	Desc_Vinculo_Fallecido	Descripción de Vínculo de Parentesco con el Fallecido	DT43_BPS_Vinculo_Parentesco_Fallecido	

Tabla de dimensiones: Subsidio				
#	Dato	Comentario	Tabla de origen	Observaciones
1	Fecha desde	Fecha desde la que el dato es valido	Fecha de carga	Fecha_Dato
2	Fecha hasta	Fecha hasta la que el dato es valido	Fecha de carga anterior	Fecha_Dato anterior
3	Cod_subsidio	Código subsidio	Autogenerado	Detalle en la tabla

4	Des_subsidio	Descripción subsidio	Autogenerado	

Tabla de dimensiones: Causal subsidio				
#	**Dato**	**Comentario**	**Tabla de origen**	**Observaciones**
1	Fecha desde	Fecha desde la que el dato es valido	DT33_BPS_Causal_Subsidios	Fecha_Dato
2	Fecha hasta	Fecha hasta la que el dato es valido	DT33_BPS_Causal_Subsidios	Fecha_Dato anterior
3	Cod_Causal_Subsidio	Código de Causal de Subsidio	DT33_BPS_Causal_Subsidios	
4	Desc_Causal_Subsidio	Causal de Subsidio	DT33_BPS_Causal_Subsidios	

Tabla de dimensiones: Mutualista SNIS				
#	**Dato**	**Comentario**	**Tabla de origen**	**Observaciones**
1	Fecha desde	Fecha desde la que el dato es valido	DT48_BPS_Mutualista_SNIS	Fecha_Dato
2	Fecha hasta	Fecha hasta la que el dato es valido	DT48_BPS_Mutualista_SNIS	Fecha_Dato anterior
3	Cod_Mutualista_SNIS	Código de Mutualista del SNIS	DT48_BPS_Mutualista_SNIS	
4	Desc_Mutualista_SNIS	Descripción de Mutualista del SNIS	DT48_BPS_Mutualista_SNIS	

Discapacidad

HECHOS

Tabla de hechos: Discapacidad				
#	Dato	Comentario	Tabla de origen	Observaciones
1	Fecha desde	Fecha desde la que el dato es valido	Fecha_Dato	
2	Fecha hasta	Fecha hasta la que el dato es valido		
3	Version	version de actualización		
4	Id_persona	Id estadísitico de la persona	Tablas que se mencionan en la codiguera	Documento & Cod_Tipo_Documento & Cod_Pais_Documento Se hace una correspondencia y se crea el Id de persona
5	Programa_discapacidad	Variable que indica el programa de discapacidad del cual la persona es beneficiaria		Variable generada

DIMENSIONES

Tabla de dimensiones: Programa discapacidad				
#	Dato	Comentario	Tabla de origen	Observaciones
1	Fecha desde	Fecha desde la que el dato es valido	Fecha de carga	Fecha_Dato
2	Fecha hasta	Fecha hasta la que el dato es valido	Fecha de carga anterior	Fecha_Dato anterior
3	Cod_discapacidad	Código de la discapacidad	Tabla + Programa	Detalle en la tabla
4	Des_discapacidad	Descripción de la discapacidad	Autogenerado	
5	Orden	Numerico ascendente	Autogenerado incremental	
6	Origen	Fuente del dato		

Programas sociales

HECHOS

Tabla de hechos: Programas Sociales				
#	Dato	Comentario	Tabla de origen	Observaciones
1	Fecha desde	Fecha desde la que el dato es valido	Fecha_Dato	
2	Fecha hasta	Fecha hasta la que el dato es valido		
3	Version	version de actualización		
4	Id_persona	Id estadísitico de la persona	Todas las tablas que se detallan la codiguera	Documento & Cod_Tipo_Documento & Cod_Pais_Documento Se hace una correspondencia y se crea el Id de persona
7	Programa_social	Variable que indica el programa social del cual la persona es beneficiaria		Variable generada Armar codiguera para unir MIDES e INAU. Tabla + MIDES o INAU. Se marca en cual programa social está.
8	Prestación_INDA	Descripción de la prestación INDA	FT05_INDA.txt	Unir por el ID de persona
9	Es_Titular	Indica si es o no es titular de la tarjeta (1=Sí, 0=No)	FT05_MIDES_Tarj_Alim.txt	
10	TUS_Doble	Indica si cobra o no TUS doble (1=Doble, 0=Simple)	FT05_MIDES_Tarj_Alim.txt	
11	Monto_Mensual_Hogar	Monto total mensual pagado al hogar por TUS	FT05_MIDES_Tarj_Alim.txt	

DIMENSIONES

Tabla de dimensiones: Programa discapacidad				
#	Dato	Comentario	Tabla de origen	Observaciones
1	Fecha desde	Fecha desde la que el dato es valido	Fecha de carga	Fecha_Dato

2	Fecha hasta	Fecha hasta la que el dato es valido	Fecha de carga anterior	Fecha_Dato anterior
3	Cod_prog_social	Código del programa social	Tabla + Programa	Detalle en la tabla
4	Des_prog_social	Descripción del programa social	Autogenerado	
5	Orden	Numerico ascendente	Autogenerado incremental	
6	Origen	Fuente del dato		

Residencia

HECHOS

Tabla de hechos: Residencia				
#	Dato	Comentario	Tabla de origen	Observaciones
1	Fecha desde	Fecha desde la que el dato es valido	Fecha_Dato	
2	Fecha hasta	Fecha hasta la que el dato es valido		
3	Version	version de actualización		
4	Id_persona	Id estadísitico de la persona		
14	Cod_Departamento	Departamento	DT01_IM_NN_Hogar.txt DT01_MVOTMA_Hogar.txt DT02_ASSE_Persona.txt DT01_INAU_Hogar.txt DT01_CEIP_Hogar.txt DT02_CETP_Persona.txt DT01_BPS_Hogar DT01_MIDES_Hogar	Dimensiones generales
15	Cod_Localidad	Localidad		Dimensiones generales
16	Cod_Calle	Calle		
17	Numero_Puerta	Número de Puerta		
18	Bis	Bis		
19	Apto	Apartamento		
20	Telefono_Hogar	Número de Teléfono del Hogar	DT01_MVOTMA_Hogar.txt DT01_IM_NN_Hogar.txt DT02_MVOTMA_Persona.txt DT02_ASSE_Persona.txt DT01_CEIP_Hogar.txt	
21	Cod_Seccion	Sección	DT02_IM_NN_Persona.txt DT01_MVOTMA_Hogar.txt DT01_IM_NN_Hogar.txt DT02_ASSE_Persona.txt DT02_CETP_Persona.txt DT01_INAU_Hogar.txt DT01_BPS_Hogar DT02_BPS_Persona DT01_MSP_Hogar.txt DT02_MSP_Persona.txt DT01_MIDES_Hogar	Codiguera?
22	Cod_Segmento	Segmento		Codiguera?
23	Cod_Zona	Zona		Zona geográfica?
24	Coordenada_X	Coordenada Geográfica X		
25	Coordenada_Y	Coordenada Geográfica Y		
26	Texto_Complemento_Direccion	Texto que complementa los datos de la dirección		

27	Manzana	Manzana	DT01_MVOTMA_Hogar.txt DT01_IM_NN_Hogar.txt DT02_MVOTMA_Persona.txt DT02_ASSE_Persona.txt DT01_CEIP_Hogar.txt	
28	Solar	Solar		
29	Complejo	Complejo		
30	Block	Block		
31	Padron	Padrón		
32	Cod_Entre_Calle_1	Código de entre calle 1		
33	Cod_Entre_Calle_1	Código de entre calle 2		
34	Codigo_Postal	Código Postal	DT01_MVOTMA_Hogar.txt DT01_IM_NN_Hogar.txt DT02_MVOTMA_Persona.txt DT02_ASSE_Persona.txt DT01_CEIP_Hogar.txt	
35	Latitud	Latitud	DT01_MVOTMA_Hogar.txt	
36	Longiud	Longiud		

DIMENSIONES

Tabla de dimensiones: Calle

#	Dato	Comentario	Tabla de origen	Observaciones
1	Fecha desde	Fecha desde la que el dato es valido	DT10_MVOTMA_Calle.txt	Fecha_Dato
2	Fecha hasta	Fecha hasta la que el dato es valido	DT10_MVOTMA_Calle.txt	Fecha_Dato anterior
3	Cod_Calle	Código de Calle	DT10_MVOTMA_Calle.txt	
4	Desc_Calle	Descipción de Calle	DT10_BPS_Calle	

Educación

Tabla de hechos: Educación				
	Dato	**Comentario**	**Tabla de origen**	**Observaciones**
	Fecha desde	Fecha desde la que el dato es valido	Fecha_Dato	
	Fecha hasta	Fecha hasta la que el dato es valido	Fecha_desde anterior	
	Version	version de actualización		
	Id_persona	Id estadísitico de la persona		Documento & Cod_Tipo_Documento & Cod_Pais_Documento Se hace una correspondencia y se crea el Id de persona
	Año_Lectivo	Año lectivo		
	Cod_Departamento_Escuela	Código de departamento de la escuela	FT01_CEIP_GURI_Inscriptos.txt	
	Numero_Escuela	Número de la escuela	FT01_CEIP_GURI_Inscriptos.txt	
	Cod_Grado_Escolar	Código de grado escolar	FT01_CEIP_GURI_Inscriptos.txt	
	Cod_Nivel_Educ	Código de nivel educativo (Inicial, Primaria, Especial, Media Básica)	FT01_CEIP_GURI_Inscriptos.txt	
	Cod_Asistencia	Código de asistencia (1 = Asiste, 0 = No asiste)	FT01_CEIP_GURI_Inscriptos.txt	
	Cantidad_Inasist_Justif	Cantidad de inasistencias justificadas	FT03_CEIP_GURI_Eval_Final.txt	
	Cantidad_Inasist_Injustif	Cantidad de inasistencias injustificadas	FT03_CEIP_GURI_Eval_Final.txt	
	Cod_Evaluacion_Final	Código de evaluación final (promovido o no)	FT03_CEIP_GURI_Eval_Final.txt	
	Cod_Motivo_Repeticion	Código de motivo de repetición (si el alumno aprueba, va vacío).	FT03_CEIP_GURI_Eval_Final.txt	
	Programa_PMC	Código de participación (1 = Asiste, 0 = No asiste)	FT04_CEIP_GURI_PMC.txt	Variable nueva si se encuentra o no en la tabla FT04_CEIP_GURI_PMC.txt
	Eval_Alfab_Hogares	Asistencia (cantidad) a Alfabetización en Hogares	FT04_CEIP_GURI_PMC.txt	
	Eval_Grupos_Familia	Asistencia (cantidad) a Grupos con las Familias	FT04_CEIP_GURI_PMC.txt	
	Forma_Ingreso	Escuela Primaria de la cual proviene	FT01_CES_Secund_Inscriptos.txt	
	Cod_Liceo	Código de liceo	FT01_CES_Secund_Inscriptos.txt	
	Cod_Plan	Código de plan de estudios	FT01_CES_Secund_Inscriptos.txt	

	Grado	Número de grado	FT01_CES_Secund_Inscriptos.txt	
	Cod_Orientacion	Código de orientación	FT01_CES_Secund_Inscriptos.txt	
	Cod_Opcion	Código de opción	FT01_CES_Secund_Inscriptos.txt	
	Cod_Turno	Código de turno	FT01_CES_Secund_Inscriptos.txt	
	Grupo	Grupo	FT01_CES_Secund_Inscriptos.txt	
	Cod_Asistencia	Código de asistencia (1 = Asiste, 0 = No asiste)	FT01_CES_Secund_Inscriptos.txt	¿? El dato de esta variable puede ir en la variable 8
	Calificacion	Calificación	FT02_CES_Secund_Evaluaciones.txt	
	Cantidad_Inasist_Justif	Cantidad de inasistencias justificadas	FT02_CES_Secund_Evaluaciones.txt	
	Cantidad_Inasist_Injustif	Cantidad de inasistencias injustificadas	FT02_CES_Secund_Evaluaciones.txt	
	Cod_Reparticion	Código de repartición	FT01_CETP_UTU_Inscriptos.txt	
	Cod_Plan	Código de plan	FT01_CETP_UTU_Inscriptos.txt	
	Cod_Tipo_Curso	Código de tipo de curso	FT01_CETP_UTU_Inscriptos.txt	
	Cod_Curso	Código de curso	FT01_CETP_UTU_Inscriptos.txt	
	Cod_Nivel	Código de nivel	FT01_CETP_UTU_Inscriptos.txt	
	Grado	Grado	FT01_CETP_UTU_Inscriptos.txt	
	Modulo	Módulo	FT01_CETP_UTU_Inscriptos.txt	
	Trayecto	Trayecto	FT01_CETP_UTU_Inscriptos.txt	
	Semestre	Semestre	FT01_CETP_UTU_Inscriptos.txt	
	Grupo	Grupo	FT01_CETP_UTU_Inscriptos.txt	
	Turno	Turno	FT01_CETP_UTU_Inscriptos.txt	
	Cod_Asistencia	Código de asistencia (1 = Asiste, 0 = No asiste)	FT01_CETP_UTU_Inscriptos.txt	¿? El dato de esta variable puede ir en la variable 8
	Experiencia_Comunitaria	Si es una experiencia comunitaria (1 = Sí, 0 = No), las cuales son focalizadas	FT01_CETP_UTU_Inscriptos.txt	
	Cantidad_Inasist_Justif	Cantidad de inasistencias justificadas	FT02_CETP_UTU_Evaluaciones.txt	
	Cantidad_Inasist_Injustif	Cantidad de inasistencias injustificadas	FT02_CETP_UTU_Evaluaciones.txt	
	Cantidad_Inasist_Susp	Cantidad de inasistencias por suspensión	FT02_CETP_UTU_Evaluaciones.txt	
	Cod_Asistencia	Código de asistencia (si asiste o no, y por qué)	FT02_CETP_UTU_Evaluaciones.txt	

23	Cantidad_Asignat_Aprob	Cantidad de asignaturas aprobadas	FT02_CETP_UTU_Evaluaciones.txt	

Registros de Empresas

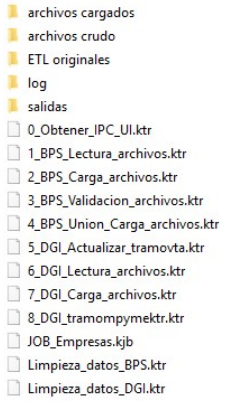
1. Convención de nombres y buenas prácticas

1.1. Estructura de carpetas

Se propone una estructura de carpetas siguiendo la siguiente jerarquía:

2. **ETL Empresas:** repositorio donde se almacenarán los ETL. Dentro del mismo se crearán 4 carpetas:
    a. **archivos cargados**: repositorio donde se almacenarán todos los archivos que se ejecutaron correctamente y sus datos fueron cargados al DW. Los archivos se moverán automáticamente en el último paso del JOB.
    b. **archivos crudos:** repositorio donde se almacenarán los archivos a cargar en el DW.
    c. **salidas:** repositorio donde se almacenarán los archivos que se fueron creando con la ejecución de los ETL.
    d. **log:** repositorio donde se almacenarán el log de las ejecuciones.

A continuación, se detalla un ejemplo:

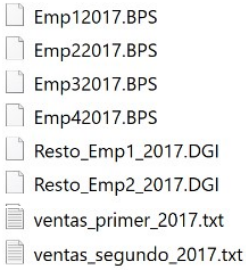


1.2. Nombre y formato de los archivos

Para la ejecución de los ETL se requieren de 8 archivos de entrada:

1. Archivos BPS: Se esperan 4 archivos de BPS cuyo nombre sea: EmpN°TrimestreAño.BPS
2. Archivos DGI: Se esperan 2 archivos de DGI cuyo nombre sea: Resto_EmpN°Semetre_Año.DGI
3. Archivos de ventas: Se esperan 2 archivos de ventas cuyo nombre sea: ventas_semetre_Año.txt (Semestre debe detallarse primer o segundo)

A continuación, se detalla un ejemplo:



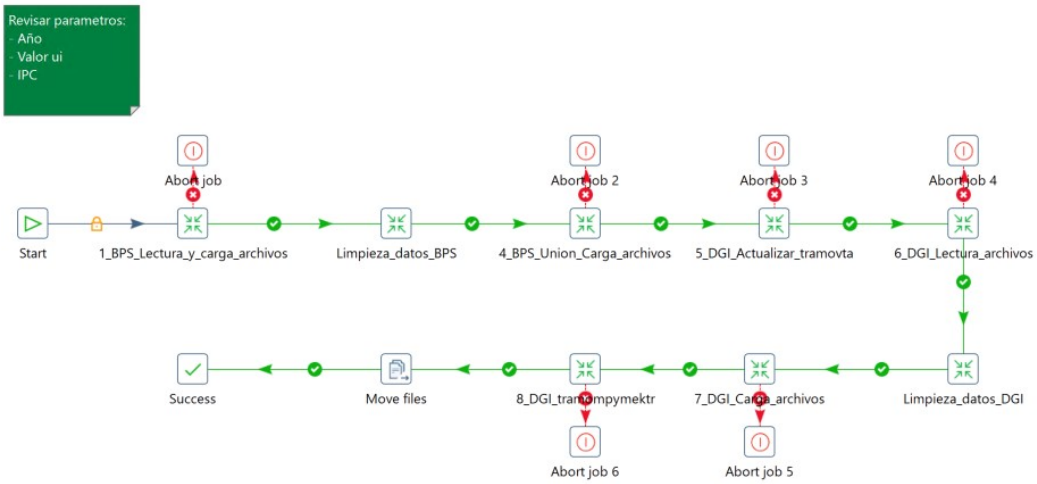
2. Flujo de trabajo

2.1. Comentarios generales

Para el presente módulo, se trabajó con ETL previamente desarrollados. Se observaron los mismos y se realizaron las mejoras pertinentes para generalizar los ETL, mejorar la performance o diseño de los mismos.

2.2. ETL en Pentaho PDI

JOB_empresas



El JOB de empresas tienen definido todos los ETL que se describirán a continuación. Se ejecutan uno a continuación del otro. De presentar algún error se detalla el mensaje de error. El último paso del JOB consiste en mover los archivos de la carpeta “archivos_crudos” a “archivos_cargados”.

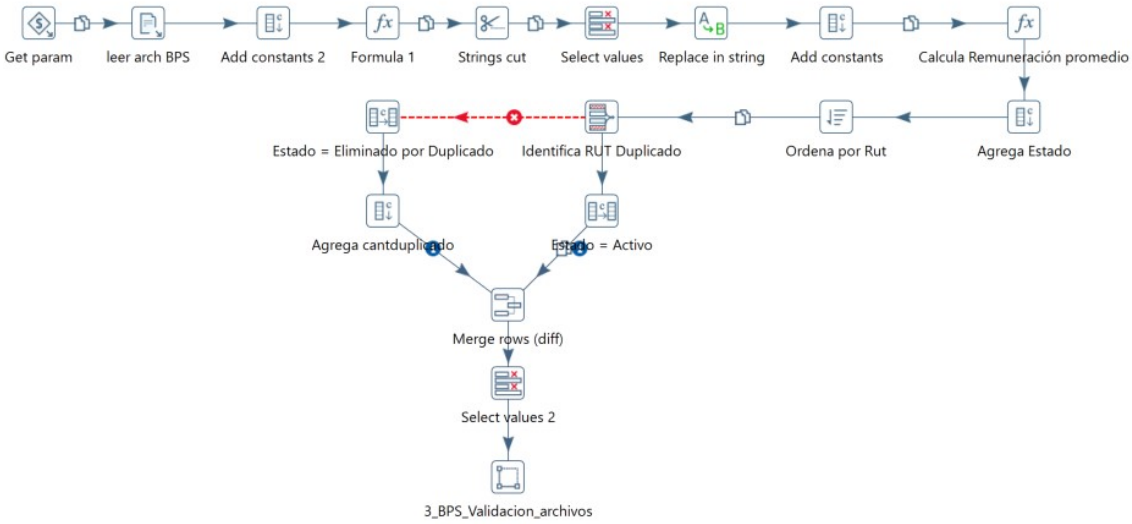
Se deben definir o revisar los parámetros de Esquema, Directorio de archivos crudos, Directorio de las salidas, Año, Valor UI y Valores de IPC.

1_BPS_Lectura_archivos

La finalidad de este ETL es obtener la url de todos los archivos cuya terminación es ".BPS" para luego ejecutar la transformación 2_BPS_Carga_archivos.

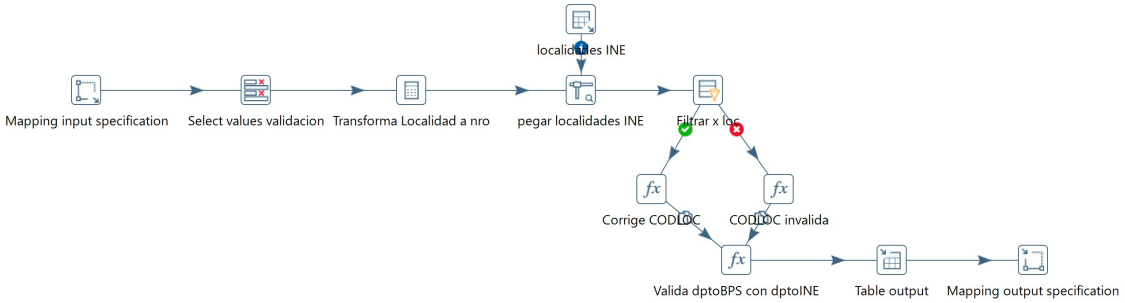
Como parámetros del JOB recibe el nombre del esquema, el año y el directorio donde se encuentran los archivos crudos.

2_BPS_Carga_archivos



La finalidad de este ETL es cargar los archivos cuya terminación es ".BPS". Las url de los mismos se obtuvieron en el paso anterior. Posteriormente a cargar los archivos, se pasa a un ETL de validación (3_BPS_Validacion_archivos).

3_BPS_Validacion_archivos



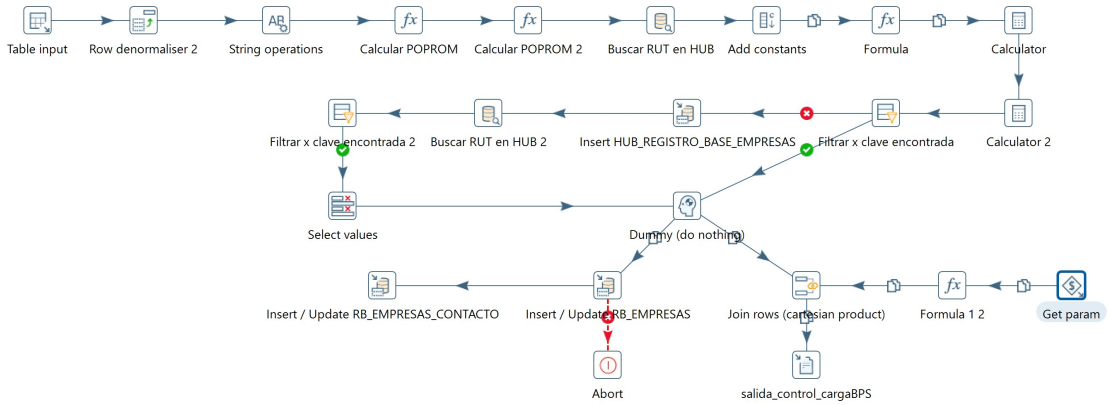
La finalidad de este ETL es validar las localidades de INE. A diferencia del ETL original, se decidió reemplazar el archivo de salida por una tabla llamada ODS_BPS.

Limpieza_datos_BPS

Al cargarse los datos de forma incremental, la finalidad con la que se incluye este ETL es eliminar valores repetidos o duplicados producto de un error de carga.

Como parámetros del JOB recibe el nombre del esquema.

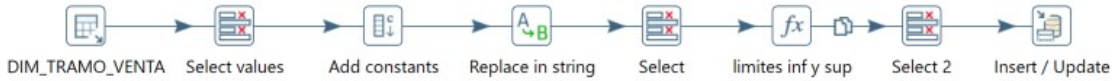
4_BPS_Union_Carga_archivos



La finalidad de este ETL es cargar todos los datos de BPS correspondiente al año determinado en el parámetro y actualizar las tablas de RB EMPRESAS y RB EMPRESAS CONTACTO.

Como parámetros del JOB recibe el nombre del esquema, el año y el directorio de salida.

5_DGI_Actualizar_tramovta

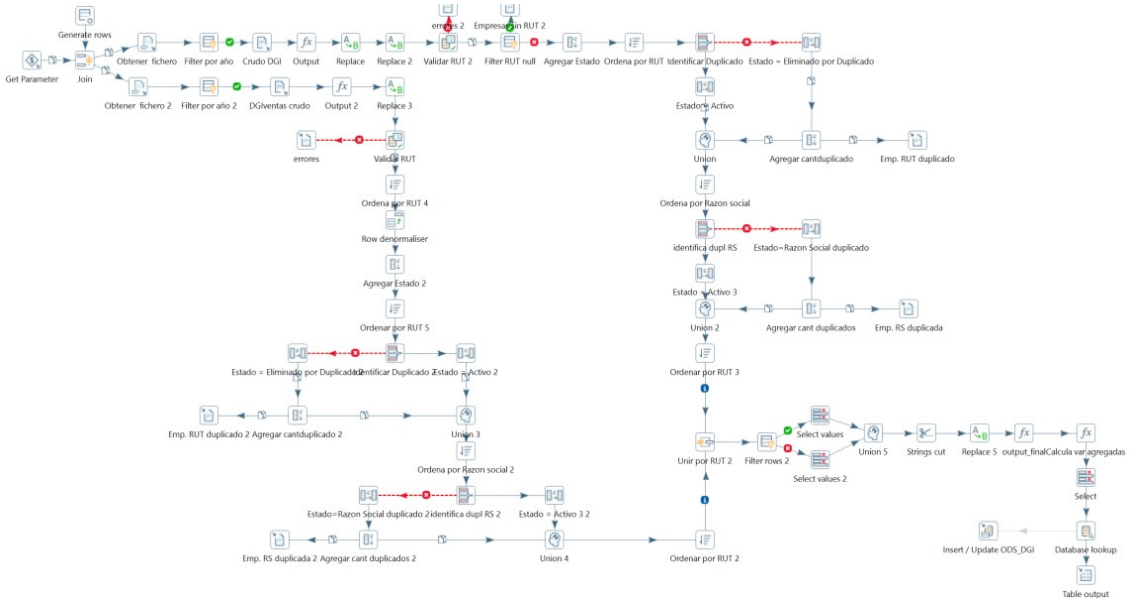


La finalidad de este ETL es calcular el tramo de venta del año en el que se está trabajando. Tomando el tramo de ventas del año anterior generar el nuevo tramo de ventas, cargar el tramo de venta para las empresas del año y actualizar los datos para el año siguiente.

A diferencia del ETL original, se decidió reemplazar el archivo csv de entrada por una consulta a la tabla DIM_TRAMO_VENTA.

Como parámetros del JOB recibe el nombre del esquema, el año y el valor de IPC.

6_DGI_Lectura_archivos



La finalidad de este ETL es leer todos los archivos de DGI que se encuentren en la carpeta de archivos crudos y cargarlos en una tabla. A diferencia del ETL original, se decidió reemplazar el archivo de salida por una tabla llamada ODS_DGI. Posteriormente ejecuta el ETL 7_DGI_Carga_archivos.

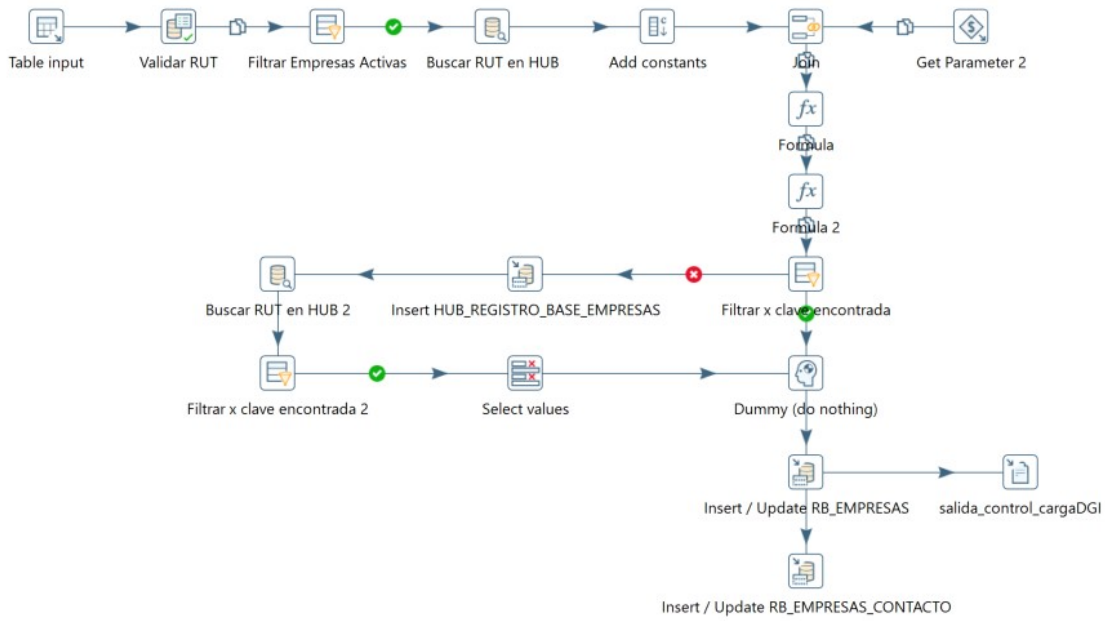
Como parámetros del JOB recibe el nombre del esquema, el año y los directorios de entrada y de salida.

Limpieza_datos_DGI

Al cargarse los datos de forma incremental, la finalidad con la que se incluye este ETL es eliminar valores repetidos o duplicados producto de un error de carga.

Como parámetros del JOB recibe el nombre del esquema.

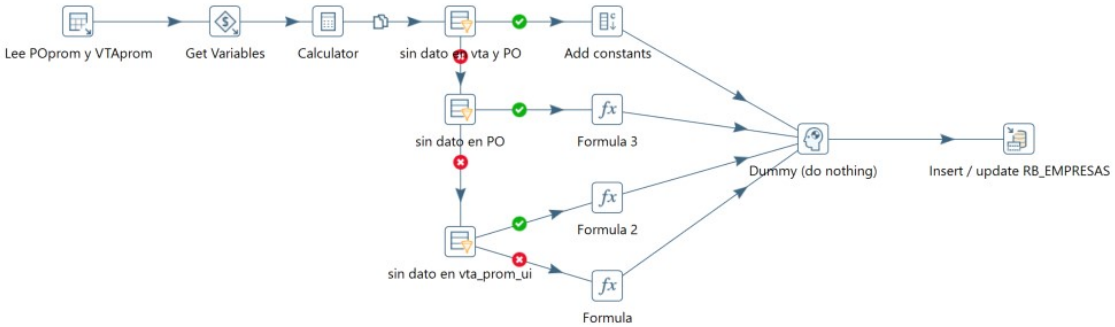
7_DGI_Carga_archivos



La finalidad de este ETL es cargar todos los datos de DGI correspondiente al año determinado en el parámetro y actualizar las tablas de RB EMPRESAS y RB EMPRESAS CONTACTO.

Como parámetros del JOB recibe el nombre del esquema, el año y el directorio de salida.

8_DGI_tramompymektr



La finalidad de este ETL es calcular si la empresa es una Micro, pequeña, mediana o gran empresa.

Como parámetros del JOB recibe el nombre del esquema, el año y el valor UI.

2.3. Nuevas tablas del DW

Se deben crear las tablas:

- ODS_BPS

- ODS_DGI

3. Mejoras sugeridas

A continuación, se detallan algunas mejoras sugeridas para realizar una vez finalizado el proyecto o como versión 2 del mismo.

- Obtener el valor de UI y de IPC directamente desde la página web del INE utilizando Web Scraping (se entrega un ETL modelo para su implementación: 0_Obtener_IPC_UI.ktr).
- Utilizar una herramienta centralizada de información.
- Armar un esquema de visualización.
- Armar módulo de carga que permita cargar los archivos a través del módulo y no de forma manual.
- Realizar analítica avanzada.

Metadatos (diccionario de variables) del Registro Base de Empresas

Variables ORIGINALES						Transformación	TABLA	Variables SIRE				
Fuente	Variable original	Etiqueta	Descripción	Tipo	Categorías			Variable original	Etiqueta	Descripción	Tipo	Categorías
BPS	Nro_cont_int	Nro. de Contribuyente Int.	Número de Contribuyente I+C5:C35NTERNO	Numeric			Tabla RB_empresas_contacto	NRO_CONT_INT	Nro. de Contribuyente Int.	Número de Contribuyente I+C5:C35NTERNO	String	
BPS	Nro_emp_int	Nro de Empresa	Número Interno de Empresa	Numeric			Tabla RB_empresas_contacto	NRO_EMP_INT	Nro de Empresa	Número Interno de Empresa	String	
BPS	Caja	Tipo de Aporte	Tipo de Aporte de Datos	Numeric			Tabla RB_empresas	CAJA	Tipo de Aporte	Tipo de Aporte de Datos	String	
BPS	Ruc	Nro. de Contribuyente Ext.	Número de Contribuyente Externo (RUC si es contribuyente en DGI)	String			Tabla RB_empresas_contacto	RUT	Registro único Tributario	Registro único Tributario	String	
BPS	Nro_emp	Nro. Afiliación	Número De Afiliación	String			Tabla RB_empresas_contacto	NRO_EMP	Nro. Empresa	Número De Afiliación de la Empresa	String	
BPS	nombre_emp	Nombre Empresa	Nombre Comercial de la empresa	String			Tabla RB_empresas_contacto	NOMBRE_EMPRESA	Nombre Empresa	Nombre Comercial de la empresa	String	
BPS	razon_soci	Razon Social	Nombre de Razón Social	String			Tabla RB_empresas_contacto	RAZON_SOCIAL	Razon Social	Nombre de Razón Social	String	
BPS	Cod_calle_emp	Cod. Calle	Código de Calle de la Empresa	Numeric			Tabla RB_empresas_contacto	COD_CALLE_EMP	Cod. Calle	Código de Calle de la Empresa	String	
BPS	Nombrecall	Nombre de Calle Emp.	Nombre de la Calle de la Empresa	String			Tabla RB_empresas_contacto	NOMBRE_CALLE	Nombre de Calle Emp.	Nombre de la Calle de la Empresa	String	
BPS	nro_puerta	Nro. Puerta Emp.	Número de Puerta de la Empresa	Numeric			Tabla RB_empresas_contacto	NRO_PUERTA	Nro. Puerta Emp.	Número de Puerta de la Empresa	String	
BPS	bis_emp	Bis Emp.	Bis de la Empresa	String			Tabla RB_empresas_contacto	BIS_EMP	Bis Emp.	Bis de la Empresa	String	
BPS	apto_emp	Apartamento Emp.	Apartamento de la Empresa	String			Tabla RB_empresas_contacto	APTO_EMP	Apartamento Emp.	Apartamento de la Empresa	String	
BPS	Localidad	Cod. Loc. Emp.	Código de Localidad de la Empresa	Numeric			Tabla RB_empresas	LOCALIDAD	Cod. Loc. Emp.	Código de Localidad de la Empresa	String	
BPS	Departam	Cod. Depto Emp.	Código de Departamento de la Empresa	Numeric			Tabla RB_empresas	DEPTO	Cod. Depto Emp.	Código de Departamento de la Empresa	String	
BPS	telefono_e	Telefono Emp.	Teléfono de la Empresa	Numeric			Tabla RB_empresas_contacto	TELEFONO_E	Telefono Emp.	Teléfono de la Empresa	String	
BPS	Codigopost	Cod. Postal Emp.	Código Postal de la Empresa	Numeric			Tabla RB_empresas_contacto	CODIGOPOST	Cod. Postal Emp.	Código Postal de la Empresa	String	
BPS	ciiu_princ	CIIU Principal	CIIU de la empresa Principal (Ciiu Rev2) ***o Rev 4***	Numeric	codificador CIIU		Tabla RB_empresas	CIIU_PRINC	CIIU Principal	CIIU de la empresa Principal (Ciiu Rev2) *o Rev 4*	String	
BPS	Apertura_ciiu1	Cod. Apertura Ppal.	Código de Apertura de la Empresa Principal	Numeric	codificador CIIU		Tabla RB_empresas	APERTURA_CIIU1	Cod. Apertura Ppal.	Código de Apertura de la Empresa Principal	String	
BPS	ciiu_secun	CIIU Secundaria	CIIU de la Empresa Secundaria (Ciiu Rev2)	Numeric	codificador CIIU		Tabla RB_empresas	CIIU_SECUN	CIIU Secundaria	CIIU de la Empresa Secundaria (Ciiu Rev2)	String	
BPS	Apertura_ciiu2	Cod. Apertura Sec.	Código de Apertura de la empresa Secundaria	Numeric	codificador CIIU		Tabla RB_empresas	APERTURA_CIIU2	Cod. Apertura Sec.	Código de Apertura de la empresa Secundaria	String	
BPS	nat_jur	Nat. Jurídica	Naturaleza Jurídica	Numeric			Tabla RB_empresas	NAT_JUR	Nat. Jurídica	Naturaleza Jurídica	String	
BPS	Cod_calle_contr	Cod. Calle Contr.	Código de Calle del Contribuyente	Numeric			Tabla RB_empresas_contacto	COD_CALLE_CONTR	Cod. Calle Contr.	Código de Calle del Contribuyente	String	

BPS	nombre_cal	Nombre de Calle Contr.	Nombre de la Calle del Contribuyente	String			Tabla RB_empresas_contacto	NOMBRE_CAL	Nombre de Calle Contr.	Nombre de la Calle del Contribuyente	String	
BPS	numero_pu	Nro. Puerta Contr.	Número de Puerta del Contribuyente	Numeric			Tabla RB_empresas_contacto	NUMERO_PU	Nro. Puerta Contr.	Número de Puerta del Contribuyente	String	
BPS	bis_contr	Bis Contr.	Bis del Contribuyente	String			Tabla RB_empresas_contacto	BIS_CONTR	Bis Contr.	Bis del Contribuyente	String	
BPS	apto_contr	Apartamento Contr.	Apartamento del Contribuyente	String			Tabla RB_empresas_contacto	APTO_CONTR	Apartamento Contr.	Apartamento del Contribuyente	String	
BPS	localidad_	Cod. Loc. Contr.	Código de Localidad del Contribuyente	Numeric			Tabla RB_empresas	LOCALIDAD_1	Cod. Loc. Contr.	Código de Localidad del Contribuyente	String	
BPS	Departamen	Cod. Depto Contr.	Código Departamento del Contribuyente	Numeric			Tabla RB_empresas	DEPARTAMEN	Cod. Depto Contr.	Código Departamento del Contribuyente	String	
BPS	telefono_c	Telefono Contr.	Teléfono del Contribuyente	Numeric			Tabla RB_empresas_contacto	TELEFONO_C	Telefono Contr.	Teléfono del Contribuyente	String	
BPS	codigo_pos	Cod. Postal Contr.	Código Postal del Contribuyente	Numeric			Tabla RB_empresas_contacto	CODIGO_POS	Cod. Postal Contr.	Código Postal del Contribuyente	String	
BPS	fecha_ini_	Fecha Inicio Contr.	Fecha de Inicio del Contribuyente	dd/mm/yyyy			Tabla RB_empresas	FECHA_INI	Fecha Inicio Contr.	Fecha de Inicio del Contribuyente	Timestamp	
BPS	fecha_ini_emp	Fecha Inicio Emp.	Fecha de Inicio de la Empresa	dd/mm/yyyy			Tabla RB_empresas	FECHA_INI_EMP	Fecha Inicio Emp.	Fecha de Inicio de la Empresa	Timestamp	
BPS	Anio	Anio de fuente	Año de Fuente	Numeric			Tabla RB_empresas	ANIO	Anio de fuente	Año de Fuente	String	
BPS	Trim	Trimestre	Trimestre Informado	Numeric			Tabla RB_empresas	TRIM_1	Trimestre 1	Primer Trimestre	BigNumber	
							Tabla RB_empresas	TRIM_2	Trimestre 2	SegundoTrimestre	BigNumber	
							Tabla RB_empresas	TRIM_3	Trimestre 3	Tercer Trimestre	BigNumber	
							Tabla RB_empresas	TRIM_4	Trimestre 4	Cuarto Trimestre	BigNumber	
BPS	RTot_1	PO Total mes1	Personal Ocupado Total del Trimestre Informado Correspondiente al Mes 1	Numeric			Tabla RB_empresas	RTOT_1	PO Total mes1	Personal Ocupado Total del Trimestre Informado Correspondiente al Mes 1	String	
							Tabla RB_empresas	RTOT_4	PO Total mes4	Personal Ocupado Total del Trimestre Informado Correspondiente al Mes 4	String	
							Tabla RB_empresas	RTOT_7	PO Total mes7	Personal Ocupado Total del Trimestre Informado Correspondiente al Mes 7	String	
							Tabla RB_empresas	RTOT_10	PO Total mes10	Personal Ocupado Total del Trimestre Informado Correspondiente al Mes 10	String	
BPS	RMay_1	PO Monto Imponible mayor a Cero Mes 1	Personal Ocupado con Monto Imponible mayor a Cero en el Mes 1	Numeric			Tabla RB_empresas	RMAY_1	PO Monto Imponible mayor a Cero Mes 1	Personal Ocupado con Monto Imponible mayor a Cero en el Mes 1	String	
							Tabla RB_empresas	RMAY_4	PO Monto Imponible mayor a Cero Mes 4	Personal Ocupado con Monto Imponible mayor a Cero en el Mes 4	String	
							Tabla RB_empresas	RMAY_7	PO Monto Imponible mayor a Cero Mes 7	Personal Ocupado Total del Trimestre Informado Correspondiente al Mes 7	String	
							Tabla RB_empresas	RMAY_10	PO Monto Imponible mayor a Cero Mes 10	Personal Ocupado con Monto Imponible mayor a Cero en el Mes 10	String	

BPS	MImp_1	Monto Imponible Mes 1	Monto Imponible correspondiente al Mes 1	Numeric			Tabla RB_empresas	REM_1	Monto Imponible Mes 1	Monto Imponible Correspondiente al Mes 1	String	
							Tabla RB_empresas	REM_4	Monto Imponible Mes 4	Monto Imponible Correspondiente al Mes 4	String	
							Tabla RB_empresas	REM_7	Monto Imponible Mes 7	Monto Imponible Correspondiente al Mes 7	String	
							Tabla RB_empresas	REM_10	Monto Imponible Mes 10	Monto Imponible Correspondiente al Mes 10	String	
BPS	RTot_2	PO Total mes 2	Personal Ocupado Total del Trimestre Informado Correspondiente al Mes 2	Numeric			Tabla RB_empresas	RTOT_2	PO Total mes2	Personal Ocupado Total del Trimestre Informado Correspondiente al Mes 2	String	
							Tabla RB_empresas	RTOT_5	PO Total mes5	Personal Ocupado Total del Trimestre Informado Correspondiente al Mes 5	String	
							Tabla RB_empresas	RTOT_8	PO Total mes8	Personal Ocupado Total del Trimestre Informado Correspondiente al Mes 8	String	
							Tabla RB_empresas	RTOT_11	PO Total mes11	Personal Ocupado Total del Trimestre Informado Correspondiente al Mes 11	String	
BPS	RMay_2	PO Monto Imponible mayor a Cero Mes 2	Personal Ocupado con Monto Imponible mayor a Cero en el Mes 2	Numeric			Tabla RB_empresas	RMAY_2	PO Monto Imponible mayor a Cero Mes 2	Personal Ocupado con Monto Imponible mayor a Cero en el Mes 2	String	
							Tabla RB_empresas	RMAY_5	PO Monto Imponible mayor a Cero Mes 5	Personal Ocupado con Monto Imponible mayor a Cero en el Mes 5	String	
							Tabla RB_empresas	RMAY_8	PO Monto Imponible mayor a Cero Mes 8	Personal Ocupado Total del Trimestre Informado Correspondiente al Mes 8	String	
							Tabla RB_empresas	RMAY_11	PO Monto Imponible mayor a Cero Mes 11	Personal Ocupado con Monto Imponible mayor a Cero en el Mes 11	String	
BPS	MImp_2	Monto Imponible Mes 2	Monto Imponible Mes 2	Numeric			Tabla RB_empresas	REM_1	Monto Imponible Mes 1	Monto Imponible Correspondiente al Mes 2	String	
							Tabla RB_empresas	REM_4	Monto Imponible Mes 4	Monto Imponible Correspondiente al Mes 5	String	
							Tabla RB_empresas	REM_7	Monto Imponible Mes 7	Monto Imponible Correspondiente al Mes 8	String	
							Tabla RB_empresas	REM_10	Monto Imponible Mes 10	Monto Imponible Correspondiente al Mes 11	String	
BPS	RTot_3	PO Total mes 3	Personal Ocupado Total del Trimestre Informado Correspondiente al Mes 3	Numeric			Tabla RB_empresas	RTOT_3	PO Total mes2	Personal Ocupado Total del Trimestre Informado Correspondiente al Mes 2	String	
							Tabla RB_empresas	RTOT_6	PO Total mes5	Personal Ocupado Total del Trimestre Informado Correspondiente al Mes 5	String	
							Tabla RB_empresas	RTOT_9	PO Total mes8	Personal Ocupado Total del Trimestre Informado	String	

										Correspondiente al Mes 8		
							Tabla RB_empresas	RTOT_12	PO Total mes11	Personal Ocupado Total del Trimestre Informado Correspondiente al Mes 11	String	
BPS	RMay_3	PO Monto Imponible mayor a Cero Mes 3	Personal Ocupado con Monto Imponible mayor a Cero en el Mes 3	Numeric			Tabla RB_empresas	RMAY_3	PO Monto Imponible mayor a Cero Mes 3	Personal Ocupado con Monto Imponible mayor a Cero en el Mes 3	String	
							Tabla RB_empresas	RMAY_6	PO Monto Imponible mayor a Cero Mes 6	Personal Ocupado con Monto Imponible mayor a Cero en el Mes 6	String	
							Tabla RB_empresas	RMAY_9	PO Monto Imponible mayor a Cero Mes 9	Personal Ocupado Total del Trimestre Informado Correspondiente al Mes 9	String	
							Tabla RB_empresas	RMAY_12	PO Monto Imponible mayor a Cero Mes 12	Personal Ocupado con Monto Imponible mayor a Cero en el Mes 12	String	
BPS	MImp_3	Monto Imponible Mes 3	Monto Imponible Mes 3	Numeric			Tabla RB_empresas	REM_3	Monto Imponible Mes 3	Monto Imponible Correspondiente al Mes 3	String	
							Tabla RB_empresas	REM_6	Monto Imponible Mes 6	Monto Imponible Correspondiente al Mes 6	String	
							Tabla RB_empresas	REM_9	Monto Imponible Mes 9	Monto Imponible Correspondiente al Mes 9	String	
							Tabla RB_empresas	REM_12	Monto Imponible Mes 12	Monto Imponible Correspondiente al Mes 12	String	

Metadatos (diccionario de variables) del Registro Base de Empresas

Variables ORIGINALES						Variables SIRE				
Fuente	**Variable original**	**Etiqueta**	**Descripción**	**Tipo**	**Categorías**		**Variable original**	**Etiqueta**	**Descripción**	**Tipo**
DGI	ruc	No de Contribuyente	Número de Contribuyente			Tabla RB_empresas_contacto	RUT	Registro único Tributario	Registro único Tributario	String
DGI	nombre	Razon Social	Razón Social			Tabla RB_empresas_contacto	RAZON_SOCIAL	Razon Social	Razón Social	String
DGI	ccalle	Cod. Calle	Código de Calle de la Empresa			Tabla RB_empresas_contacto	COD_CALLE_EMP	Cod. Calle	Código de Calle de la Empresa	String
DGI	ncalle	Nombre de Calle Emp.	Nombre de la Calle de la Empresa			Tabla RB_empresas_contacto	NOMBRE_CAL	Nombre de Calle Emp.	Nombre de la Calle de la Empresa	String
DGI	puerta	Nro. Puerta Emp.	Número de Puerta de la Empresa			Tabla RB_empresas_contacto	NUMERO_PU	Nro. Puerta Emp.	Número de Puerta de la Empresa	String
DGI	bis	Bis Emp.	Bis de la Empresa			Tabla RB_empresas_contacto	BIS_EMP	Bis Emp.	Bis de la Empresa	String
DGI	clocalid	Cod. Loc. Emp.	código localidad			Tabla RB_empresas	CODLOC	Cod. Loc. Emp.	código localidad	String
DGI	Cdpto	Cod. Depto Emp.	Código departamento			Tabla RB_empresas	DPTO_OK	Cod. Depto Emp.	Código departamento	String
DGI	giro	Clase de Actividad	Clasificador de clase de actividad			Tabla RB_empresas	CIIU	Clase de Actividad	Clasificador de clase de actividad	String
DGI	Finsc	Fecha de Inscripción	fecha de inscripción	dd/mm/aaaa		Tabla RB_empresas	FECHA_INI	Fecha de Inscripción	fecha de inscripción	Date
DGI	fclausura	Fecha de Clausura	fecha de clausura de actividades	dd/mm/aaaa		Tabla RB_empresas	FECHA_CLAUSURA	Fecha de Clausura	fecha de clausura de actividades	Date
DGI	freinicio	Fecha de Reinicio	Ultima fecha de reinicio de actividades	dd/mm/aaaa		Tabla RB_empresas	FECHA_REINICIO	Fecha de Reinicio	Ultima fecha de reinicio de actividades	Date
DGI	grupo	Cod. Grupo	Código de grupo (CEDE, Serv Per, etc)			Tabla RB_empresas	GRUPO	Cod. Grupo	Código de grupo (CEDE, Serv Per, etc)	String
DGI	ddbce	Dia Balance	Día de balance	dd		Tabla RB_empresas	DIA_BAL	Dia Balance	Día de balance	Date
DGI	mmdce	Mes Balance	Mes de balance	mm		Tabla RB_empresas	MES_BAL	Mes Balance	Mes de balance	Date
DGI	cnatjur	Cod. Naturaleza Jurídica	Código de Naturaleza jurídica			Tabla RB_empresas	COD_NATJUR	Cod. Naturaleza Jurídica	Código de Naturaleza jurídica	String
DGI	Fecha_i	Fecha Informada	Fecha informada	01/mm/aaaa		Tabla RB_empresas	fecha	Fecha Informada	Fecha informada	Date

Derivada	Tramovta	Tramo de Ventas	Tramo de ventas del mes fecha_i (según tabla de promedios mensuales)	String	0 -Sin ventas 1- Hasta $ 25.000 2 - De $ 25,001 a $ 50.000 3 - De $ 50.001 a $ 100.000 4 - De $ 100.001 a $ 250.000 5 - De $ 250.001 a $ 500.000 6 - De $ 500.001 a $ 1.000.000 7 - De $ 1.000.001 a $ 5.000.000 8 - De $ 5.000.001 a $ 10.000.000 9 - De $ 10.000.001 a $ 20.000.000 10 - De $ 20.000.001 a $ 50.000.000 11 - De $ 50.000.001 a $ 100.000.000 12 - De $ 100.000.001 a $ 200.000.000 13 - De $ 200.000.001 a $ 300.000.000 14 - Más de $ 300.000.000	Tabla RB_empresas	VTATRAMO	Tramo de Ventas	Tramo de ventas del mes fecha_i (según tabla de promedios mensuales)	String

Registros de Inmuebles

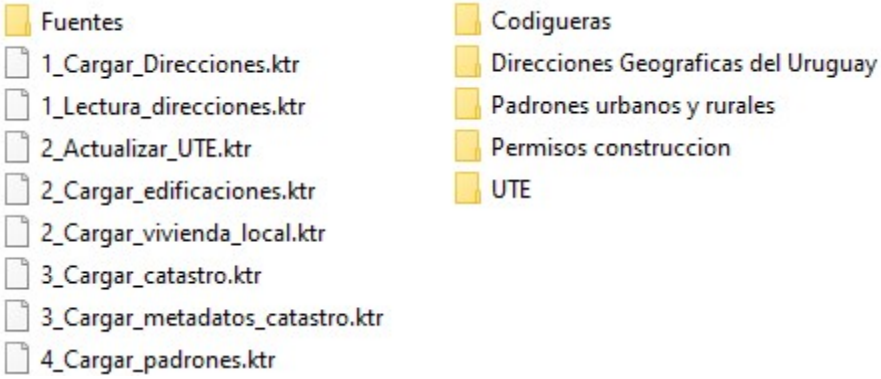
1. Convención de nombres y buenas prácticas

1.1. Estructura de carpetas

Se propone una estructura de carpetas siguiendo la siguiente jerarquía:

3. **ETL Inmueble:** repositorio donde se almacenarán los ETL. Dentro del mismo se crearán la carpeta Fuentes.
    a. **Fuentes:** repositorio donde se almacenarán los archivos a cargar en el DW. Dentro del mismo se crearán 5 carpetas:
        i. **Codigueras:** repositorio donde se almacenarán todos los archivos utilizados como codigueras.
        ii. **Direcciones Geográficas del Uruguay:** repositorio donde se almacenarán todos los archivos obtenidos de Direcciones Geográficas del Uruguay - Conjuntos de Datos - Catálogo de Datos Abiertos (catalogodatos.gub.uy)
        iii. **Padrones urbanos y rurales**: repositorio donde se almacenarán todos los archivos obtenidos de Padrones urbanos y rurales - Conjuntos de Datos - Catálogo de Datos Abiertos (catalogodatos.gub.uy)
        iv. **Permisos construcción**: repositorio donde se almacenarán todos los archivos obtenidos de Permisos de construcción aprobados - Conjuntos de Datos - Catálogo de Datos Abiertos (catalogodatos.gub.uy)
        v. **UTE:** repositorio donde se almacenarán todos los archivos de UTE comercial y UTE no comercial

A continuación, se detalla un ejemplo:

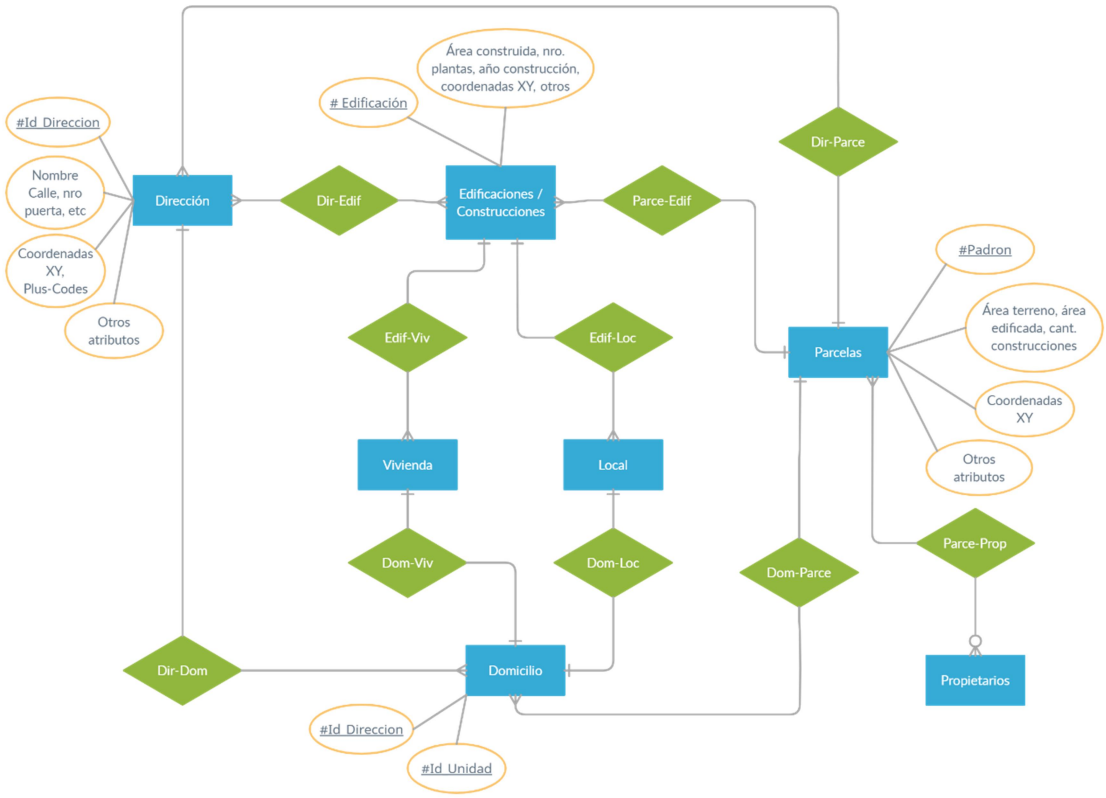


2. Flujo de trabajo

2.1. Comentarios generales

El módulo de inmuebles se desarrolló a partir del esquema que se detalla a continuación.

Esquema 1: Estructura propuesta para el módulo de inmuebles.



La fuente de datos que nutren dichas tablas son las siguientes:

- La tabla de DIRECCIONES se nutre de los 19 archivos de direcciones por departamento (Direcciones Geográficas del Uruguay - Conjuntos de Datos - Catálogo de Datos Abiertos (catalogodatos.gub.uy)).
- La tabla de EDIFICACIONES se nutre del archivo permisos de construcción (Permisos de construcción aprobados - Conjuntos de Datos - Catálogo de Datos Abiertos (catalogodatos.gub.uy)).
- Las tablas de VIVIENDA y LOCAL se nutren de la tabla EDIFICACIONES filtrada por el destino de la construcción:
  - Si el destino incluye Vivienda entonces se carga en VIVIENDA.
  - Si el destino incluye comercio o industria, se carga en LOCAL.
- La tabla de PARCELAS se nutre de los archivos de padrón rural y padrón urbano (Padrones urbanos y rurales - Conjuntos de Datos - Catálogo de Datos Abiertos (catalogodatos.gub.uy)). Se definió una

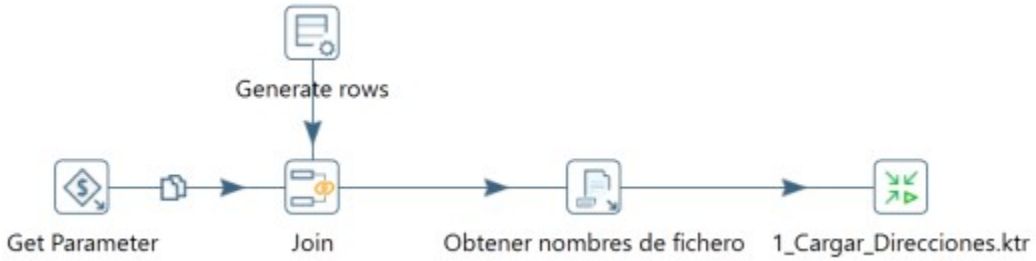
única tabla que contenga ambos archivos, donde se especifica el tipo de padrón.

Comentarios generales:

- La tabla de DOMICILIO no fue creada ya que la misma se nutre de la unión de la tabla de direcciones y de la tabla de edificaciones para definir el número de unidad, pero a la fecha no hay posibilidad de unión entre dichas tablas. El INE está trabajando en una tabla de direcciones normalizada la cual permitirá la creación de la tabla DOMICILIO.
- La tabla de PROPIETARIOS no fue creada ya que no se contaba con los datos.
- Se crearon todas las tablas de codiguera de los archivos de catastro.
- Se cuenta con archivos de UTE comercia y no comercial que se utilizaran para actualizar las tablas de VIVIENDA y LOCAL indicando el consumo de electricidad de las mismas.

2.2. ETL en Pentaho PDI

1_Lectura_direcciones

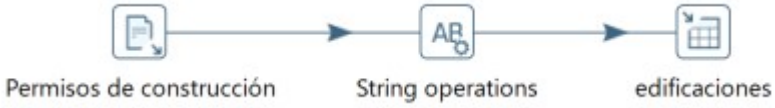


La finalidad de este ETL es obtener la url de todos los archivos fuentes de Direcciones Geográficas del Uruguay para luego ejecutar la transformación 1_Cargar_direcciones.

1_Cargar_direcciones

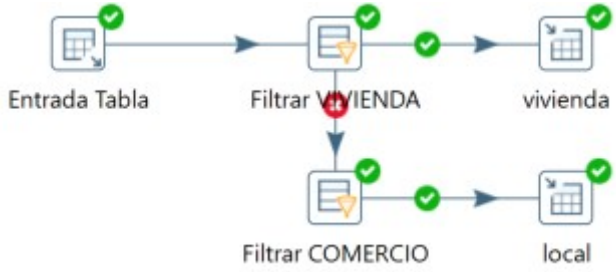
La finalidad de este ETL es cargar los archivos de Direcciones Geográficas del Uruguay. Las url de los mismos se obtuvieron en el paso anterior. Se cargan todas las direcciones por departamento en la tabla DIRECCIONES.

2_Cargar_edificaciones



La finalidad de este ETL es carga el archivo permisos de construcción (Permisos de construcción aprobados - Conjuntos de Datos - Catálogo de Datos Abiertos (catalogodatos.gub.uy)) obtenido de datos abiertos en la tabla EDIFICACIONES.

2_Cargar_vivienda_local



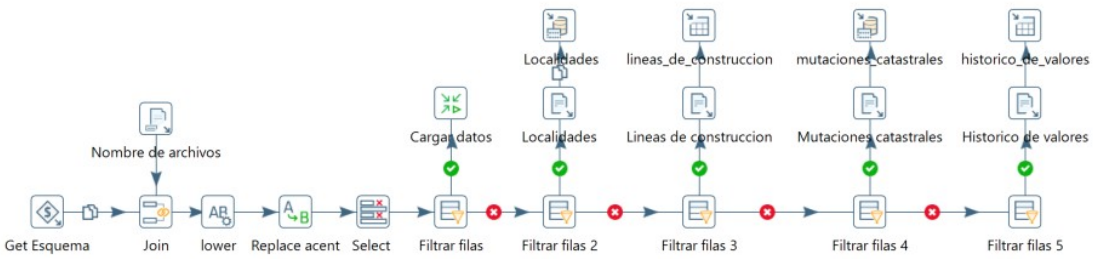
La finalidad de este ETL es cargar las tablas de VIVIENDA y LOCAL a partir de la tabla EDIFICACIONES cargada previamente. Se realiza la separación a través del campo que detalla el destino de la construcción:

- Si el destino incluye Vivienda entonces se carga en VIVIENDA.
- Si el destino incluye comercio o industria, se carga en LOCAL.

Para los destinos Otros o Desconocido, se van a analizar en qué consisten dichos campos para ver si a futuro se crea o no alguna otra tabla auxiliar.

Hay registros que van a quedar cargados dobles ya que existen combo de destino de vivienda que incluyen a ambos.

3_Cargar_catastro



La finalidad de este ETL es cargar las codigueras de catastro. Para aquellas tablas cuya estructura consiste en código y descripción, se utiliza el ETL

2_Cargar_metadata_catastro. Para aquellas tablas con una estructura más compleja, las mismas se cargan en el ETL en cuestión.

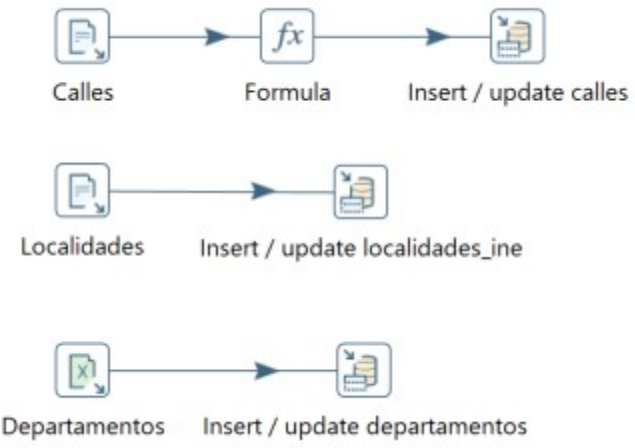
3_Cargar_metadata_catastro

La finalidad de este ETL es cargar las tablas de las codigueras de catastro que consisten en código y descripción.

4_ Cargar_parcelas

La finalidad de este ETL es cargar la tabla de PARCELAS a partir de los archivos de padrón rural y padrón urbano (Padrones urbanos y rurales - Conjuntos de Datos - Catálogo de Datos Abiertos (catalogodatos.gub.uy)). Se definió una única tabla que contenga ambos archivos con una columna auxiliar donde se especifica el tipo de padrón. El archivo de padrones rurales contiene menos columnas que el archivo de urbanos, una de ellas es el código de localidad, la cual fue seteada en 900 para los padrones rurales (es el que utiliza el INE).

Codigueras



La finalidad de este ETL es cargar las tablas de DEPARTAMENTOS, LOCALIDADES_INE y CALLES. La tabla de departamento contiene los

códigos del SIIAS, de catastro y del INE unificados en la misma tabla. Con dicha tabla se podrían normalizar todos los códigos de todos los módulos o relacionarlos con la tabla sin modificar las fuentes.
La tabla de localidades contiene los códigos INE de localidad, descripción y código de departamento.
La tabla de calle contiene los id de las calles, las descripciones y la relación con el departamento y localidad.

3. Problemas encontrados

Si bien las tablas de inmuebles fueron creadas, no se ha podido establecer un criterio o variables de unión entre las mismas, producto que los archivos fuentes no contienen los datos.
Este módulo quedará en funcionamiento cuando el INE logre realizar la tabla de normalización de direcciones. Con dicha tabla, se podrán obtener como claves de unión entre las tablas de inmueble el número de padrón o la latitud y longitud. También se podrá usar para realizar la unión con el módulo de personas.

4. Mejoras sugeridas

A continuación, se detallan algunas mejoras sugeridas para realizar una vez finalizado el proyecto o como versión 2 del mismo.

- Utilizar webservice para la carga de datos.
- Automatizar la descarga y actualización de datos desde la página web de catálogo datos abiertos (Bienvenida - Catálogo de Datos Abiertos (catalogodatos.gub.uy).
- Utilizar una herramienta centralizada de información.
- Armar un esquema de visualización.
- Realizar analítica avanzada.