Islas Canarias
Del 15 al 19 de noviembre de 2021

INSTITUTO CANARIO DE ESTADÍSTICA

REGISTRO INTEGRADO DE POBLACIÓN DE CANARIAS

Noelia Martín Morales
Instituto Canario de Estadística (ISTAC)
noelia.martinmorales@gobiernodecanarias.org

Rafael Betancor Villalba
Instituto Canario de Estadística (ISTAC)
rbetvilb@gobiernodecanarias.org

Johana Valera Puche
Ecosistemas Virtuales Modulares (EVM)
johana.valera@evm.net

Bencomo Domínguez Martín
Ecosistemas Virtuales Modulares (EVM)
bencomo.dominguez@evm.net

1. INTRODUCCIÓN

El Instituto Canario de Estadística (ISTAC) considera que es necesario fortalecer, estructurar, sistematizar, documentar y automatizar los sistemas de producción de operaciones estadísticas y en concordancia con el artículo 32 de la Ley 1/1991 de Estadística de la Comunidad Autónoma de Canarias, indica que se constituirá un banco de datos administrativos para fines estadísticos, que se nutrirá prioritariamente de los ficheros administrativos de la Comunidad Autónoma de Canarias, a cuyo fin todos los departamentos deberán remitir al ISTAC los ficheros administrativos de los que sean titulares y que sean necesarios para el ejercicio de la función estadística.

Esto es especialmente importante para cualquier institución que esté orientada a la producción de estadísticas y que debido a incrementos de demandas de información se ha visto en la necesidad de avanzar hacia procesos de estandarización e industrialización y que además pueda garantizar:

1. Sistematización del conocimiento
2. Continuidad en el tiempo frente a la movilidad de las unidades de análisis
3. Análisis, evaluación y mejora de los procesos de producción

Asimismo se señala que el banco de datos administrativos deberá facilitar la fusión de los ficheros para fines estadísticos a fin de mejorar la eficiencia de la actividad estadística de interés de la Comunidad Autónoma de Canarias y para su desarrollo durante la ejecución del Plan Estadístico de Canarias 2018-2022 que impulsará el Sistema de Datos Integrados (iDatos).

Dicho sistema debe permitir la producción de estadísticas multifuentes mediante la elaboración de datos maestros organizados en directorios y registros que faciliten el enlazamiento de fuentes diversas. A su vez el Sistema de Datos Integrados debe facilitar la obtención de estadísticas espaciales, mediante la georreferenciación de la información dentro de un Marco de Estadística Espacial de Canarias alineado con el Global Statistical Geospatial Framework elaborado por el Grupo de Expertos de Integración de Información Estadística y Geoespacial de Naciones Unidas (UN EG-ISGI. United Nations Expert Group on the integration of Statistical and Geospatial Information).

Para llevar a cabo esta estrategia, la arquitectura del banco de datos se ha dirigido hacia el procesamiento supervisado por lotes enmarcado en 5 tipos de entornos:

1. Entorno de recepción, captura y prealmacenamiento: Son entornos donde se almacenan y catalogan previamente los inputs del banco de datos.
2. Entorno de repositorio: Es un entorno de base de datos, donde se almacenan los microdatos o macrodatos que son el resultado final de procesos de tratamiento generadores de un cambio de naturaleza de los mismos.
3. Entorno computacional: Son entornos para el procesamiento eficiente de datos y su objetivo es el cambio de naturaleza de un fichero de datos. Los hilos computacionales se representan con flechas de colores en el esquema de la arquitectura del banco de datos.

4. Entorno analítico: Son entornos para el análisis exploratorio de datos. Estos entornos requieren acceso al entorno repositorio, o también necesitan de copias o transformaciones de datos para su función analítica.

5. Entorno de difusión: No se trata de un entorno en sí mismo, sino de una capa de servicios web para la difusión de datos y cartografía.

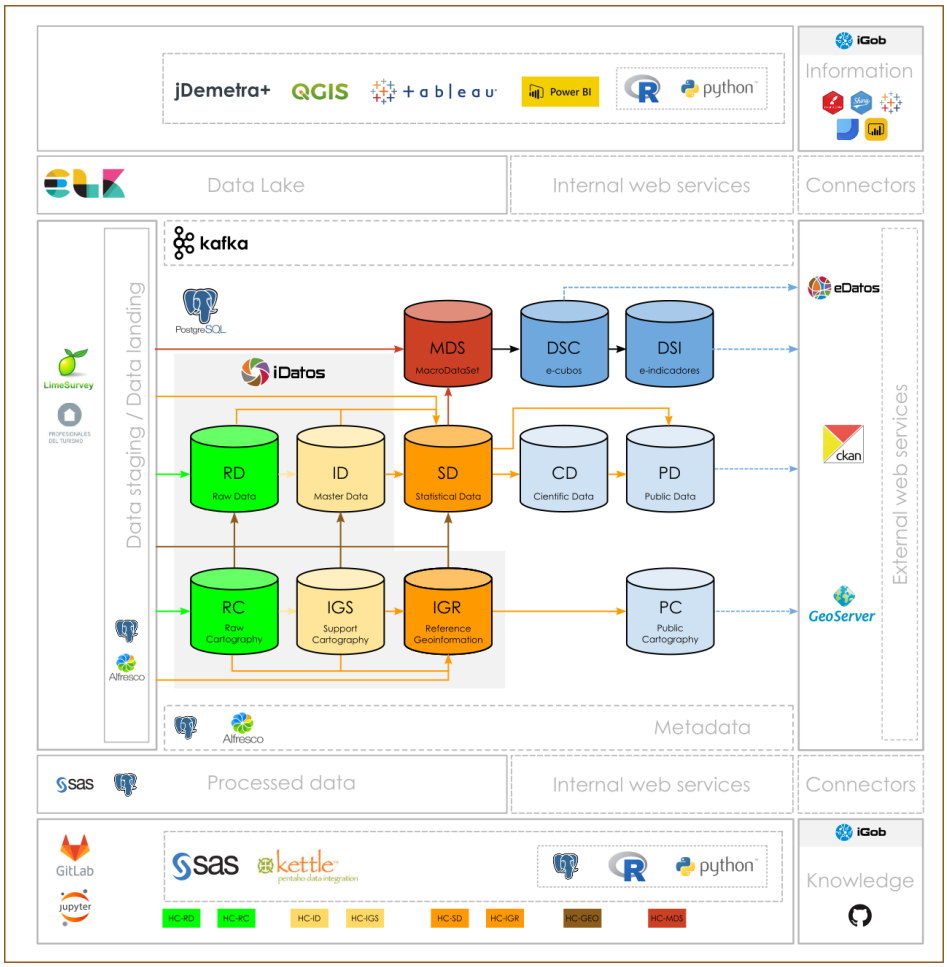

Figura 1. Modelo de arquitectura del banco de datos

Específicamente, el entorno repositorio de iDatos se organiza en esquemas-tipo sobre los que trabaja el Sistema de Datos Integrados: Raw Cartography (RC), Support Cartography (IGS), Geographic Information Reference (IGR), Raw Data (RD) y Master Data (ID).

El proceso de integración se contempla en el Plan Estadístico de Canarias 2018-2022 y en el Programa Estadístico Anual 2021 según el boc-a-2021-073-1949 de 12 de abril de 2021 donde se almacenan los principales directorios dentro del esquema ID:

1. C00047A. Directorio de Calles y Direcciones
2. C00111A. Directorio de Edificios, Viviendas y Locales
3. C00021A. Directorio de Unidades Económicas (DUE)
4. C00063A. Directorio de Población y Hogares

2. OBJETIVO

El Registro de Población tiene el objetivo de integrar las distintas fuentes de información de Canarias que poseen la misma unidad de análisis, en este caso, los datos sobre personas, mediante procesos de diseño, normalización, transformación de campos y programación de enlaces entre los ficheros de microdatos estadísticos según procedimientos determinísticos y probabilísticos (Record Linkage) para identificar a cada unidad (IDT), sus diferentes versiones de escritura (IDF) y almacenar su participación en las distintas fuentes administrativas (DAT) en una tabla de relaciones (URD).

3. METODOLOGÍA

El Registro de Población forma parte del Directorio de Población y Hogares dentro del Sistema de Datos Integrados para fines estadísticos (iDatos) y se plantea como una infraestructura estadística integradora de las diferentes fuentes de datos, que permite numerar e identificar a las personas de Canarias según un conjunto básico de variables sociodemográficas en los registros administrativos y en las diversas versiones de formatos y escrituras.

Los datos que entran provienen de una mezcla de fuentes internas o externas y que mediante el uso de programas de extracción, transformación y carga (ETL) se puede identificar, emparejar y vincular microdatos que representan la vida administrativa del conjunto de personas que lo componen.

La estrategia de construcción de los registros es siempre similar: se construyen registros de referencia (IDT) a partir de una fuente básica que se enriquece con otras fuentes auxiliares (IDF).

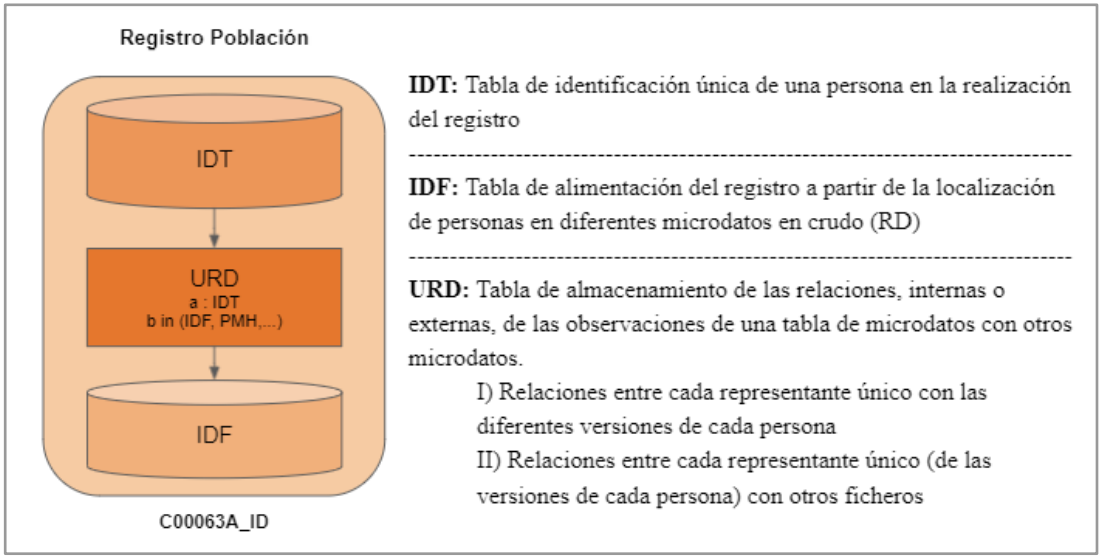

Figura 2. Componentes del Registro de Población

3.1. Diseño de Registros del IDF (Versión simplificada)

Para integrar y almacenar relaciones entre fuentes de información heterogéneas, se requieren campos normalizados y depurados que representen la identidad de cada

persona, haciendo referencia al conjunto de características propias y que permitan reconocerlo como un individuo independientemente de los demás.

BLOCK	CONCEPT_ID	LABEL
CLAVE	UUID	Identificador Único Universal
	LUID	Identificador Único Local
	STID	Identificador de esquema y tabla
	MARCATIEMPO	Sello de tiempo de creación de la observación
IDENTIFICADORES	NOMEPER_TIPO_IPF	Tipo de documento
	NOMEPER_IDENTIFICADOR_IPF	Número de identificación de la persona física
	NOMEPER_IPF_LIDEN	Letra de control del IPF
	NOMEPER_IPF_LIDEN_RC	Letra de control del IPF (recalculada)
PERSONALES	NOMEPER_NOMBRE_APELLIDOS	Nombre y apellidos de la persona física
	NOMEPER_NOMBRE	Nombre completo de la persona física
	NOMEPER_NOMBRE_1	Primer nombre de la persona física
	NOMEPER_NOMBRE_2	Resto de nombres de la persona física
	NOMEPER_APELLIDOS	Apellidos de la persona física
	NOMEPER_PARTICULA_APELLIDO1	Partícula apellido 1
	NOMEPER_APELLIDO1	Primer apellido de la persona física
	NOMEPER_PARTICULA_APELLIDO2	Partícula apellido 2
	NOMEPER_APELLIDO2	Segundo apellido de la persona física
DEMOGRÁFICAS	NOMEPER_FNAC	Fecha nacimiento
	NOMEPER_SEXO	Sexo
	NOMEPER_CODMUNNAC	Municipio de nacimiento
	NOMEPER_PAISNAC	País de nacimiento

Tabla 1. Variables del diseño de registros del IDF

3.2. Trabajos previos

Por lo general, los datos provienen de diversas fuentes y en consecuencia existen diferencias en formas de escritura, longitudes de campos y calidad de los datos, generando así escenarios que deben ser considerados al construir algoritmos de vinculación.

Una vez examinada la naturaleza de cada fuente se puede estandarizar y almacenar la información disponible, prestando especial atención a los campos que son comunes en el conjunto de ficheros mediante reglas técnicas y semánticas para la normalización y depuración de campos.

3.2.1. Normalización de los ficheros de entrada

- El primer campo que pasa por el proceso de normalización es el identificador de persona física (IPF) y su tipo de identificación (TIPO_IPF). Para realizar enlaces se ha decidido que se conserva el número de identificación si la persona tiene DNI o NIE, mientras que otro tipo de documentos como pasaportes, identificaciones

de la UE o identificaciones del resto del mundo pasan a la categoría TIPO_IPF= "_O" (Otros) y su IPF= "_X" (fuera de rango).

- El tratamiento que reciben los campos de nombres y apellidos se muestra en la siguiente tabla:

Descripción	Nombres y apellidos		
	Entorno de repositorio IDF	Entorno computacional	
		Función SOUNDEX (var1=* var2)	Función SOUNDEX_ ES
Funciones de depuración: upcase, compress, substr, compbl, tranwrd y scan	✓	✓	✓
Quitar tildes, diéresis, puntuación, espacios múltiples	✓	✓	✓
Quitar "DE", "DEL", "DE LA", "DE LOS", "DE LAS"		✓	✓
Sustituir "Ñ", "Ç"		✓	
Se contrae el nombre de "MARIA" por "M"		✓	✓

Tabla 2. Depuración de cadenas de texto

3.2.2. Operadores fonéticos entre cadenas

- Similitud fonética con el operador SOUNDEX (var1=* var2): Es un operador que se encarga de realizar un apareamiento entre cadenas de caracteres.

- Similitud fonética a partir de los valores SOUNDEX_ES: es un algoritmo fonético que ha sido preparado para encontrar un nombre por su sonido, al ser pronunciado en español. El objetivo básico de este algoritmo es codificar la cadena de texto a un código numérico para encontrar similitud de nombres con la misma pronunciación. Donde:

  - Las vocales, las letra H y W no se recodifican

  - Las que tienen similitud fonética tendrán el mismo code

  - Si dos o más letras codificadas adyacentes tienen el mismo code entonces se descartan todas menos la primera. Ej: Anna tiene soundex_es= 6

- Distancia de edición generalizada COMPGED: esta función mide la diferencia entre dos cadenas usando una variación de Levenshtein que calcula el "costo" de realizar cierta cantidad de ediciones necesarias para cambiar una cadena X en otra cadena Y.

  - Por ser asimétrica se calcula el promedio (COMPGED(X,Y) + COMPGED(Y, X))/2 sobre los nombres y apellidos previos al cálculo del SOUNDEX_ES o sobre las cadenas de texto que forman los códigos SOUNDEX_ES calculados.

3.3. Relaciones URD

Cuando se identifica una persona (IDT), sus diferentes versiones de escritura (IDF) y su participación en las distintas fuentes administrativas (DAT) de Canarias, estos datos se almacenan en una tabla de relaciones (URD) que se concentran en dos bloques:

- URD I: se refiere a la relación entre la tabla IDT y la tabla de formas de escritura o capa fuente IDF
- URD II: se refiere a la relación entre la tabla IDT y las observaciones del fichero de entrada DAT.

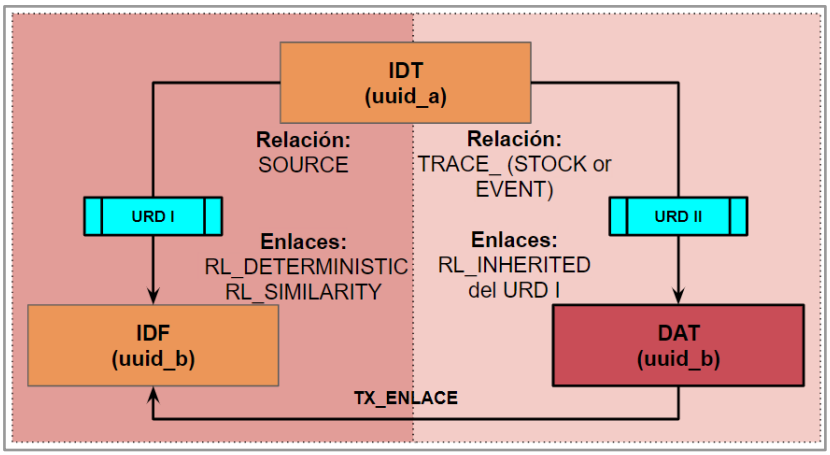

Figura 3. Partición de las relaciones URD

Además de identificar las relaciones entre fuentes, el URD permite almacenar información complementaria sobre el tipo de relación encontrada, la calidad de los enlaces derivados durante el proceso, permite reconocer el origen de los enlaces y la vida administrativa de la persona a través de evolución longitudinal y transversal de sus datos.

Variable	**Tipo**	**Descripción**
uuid	*norma, varchar (36)*	*(UUID de la relación, se reutiliza en la parte II cuando IDT-DAT [1])*
stid	*norma, varchar (61)*	c00063a_id.urd_poblacion (*se reutiliza en la parte II [2]*
luid	*norma, integer*	*(Identificador único local de la relación)*
marcatiempo	*norma, date*	*(Fecha de creación de la relación, AAAA/MM/DD)*
uuid_a	*norma, varchar (36)*	*(UUID del IDT: puede repetirse para un conjunto de IDF´s)*
stid_a	*norma, varchar (61)*	c00063a_id.idt_poblacion
uuid_b	*norma, varchar (36)*	*(UUID del IDF : diferentes nombres que tiene un mismo registro)*
stid_b	*norma, varchar (61)*	c00063a_id.idf_poblacion

rel_type	varchar (30)	SOURCE
rel_period	*norma, date*	(Fecha de referencia del .DAT que generó el IDF, surge otra versión, la fecha debe actualizarse)
rel_event	*norma, date*	(Vacío, se declara en URD-II)
rel_description	varchar (255)	(stid del .DAT que generó la relación)
link_type	varchar (30)	RL_DETERMINISTIC
link_description	varchar (255)	(“Algoritmo de deduplicación de personas del PMH”o TX_ENLACE entre DAT e IDF)
link_quality	varchar (255)	“Enlace de **n1** IDF a 1 IDT” *(se calcula **n1**= count(distinct(uuid_idf)) by uuid_idt)*
link_quality_rank	varchar (10)	(link_quality_rank cuando se enlaza DAT e IDF)
link_active	varchar (1)	Indicador de actividad del enlace. 1=Sí (activo); 6=No (inactivo) ISTAC:CL_DICOTOMICA_NUMERICA(01.000)
link_active_in	*norma, date*	Fecha de alta del enlace IDT-IDF. Si está activo, coincide con la MARCATIEMPO
link_active_out	*norma, date*	Fecha de baja del enlace IDT-IDF. Obligatorio cuando está inactivo

Tabla 3 .Diseño de registro del URD-I: idt_poblacion - idf_poblacion

Variable	**Tipo**	**Descripción**
uuid	*norma, varchar (36)*	*(UUID de la relación)*
stid	*norma, varchar (61)*	c00063a_id.urd_poblacion
luid	*norma, integer*	*(Identificador único local de la relación)*
marcatiempo	*norma, date*	*(Fecha de creación de la relación, AAAA/MM/DD)*
uuid_a	*norma, varchar (36)*	*(UUID del IDT: puede repetirse para un conjunto de DAT´s)*
stid_a	*norma, varchar (61)*	c00063a_id.idt_poblacion
uuid_b	*norma, varchar (36)*	*(UUID de origen del fichero DAT. IMPORTANTE: todas las filas del fichero DAT deben estar registradas)*
stid_b	*norma, varchar (61)*	[schema.table]
rel_type	varchar (30)	TRACE_STOCK; TRACE_EVENT *(Si es un fichero de inventario o evento)*
rel_period	*norma, date*	(Fecha de referencia de inventario o evento del .DAT)
rel_event	*norma, date*	Si rel_type es:

		a)TRACE_ EVENT la fecha de alta del evento es obligatoria *b)*TRACE_ STOCK se coloca fecha de alta administrativa o cambio de domicilio según el tipo de fuente.
rel_description	varchar (255)	"Identificación de personas según enlace IDT-IDF"
link_type	varchar (30)	RL_INHERITED
link_description	varchar (255)	"*[1]:[2]" (relación IDT-IDF a la que pertenece el DAT)*
link_quality	varchar (255)	"Enlace de **n2** DAT a 1 IDT" *(se calcula **n2**= count(uuid_dat) by uuid_idt)*
link_quality_rank	varchar (10)	(link_quality_rank cuando se enlaza DAT e IDF)
link_active	varchar (1)	Indicador de actividad del enlace. 1=Sí (activo); 6=No (inactivo) ISTAC:CL_DICOTOMICA_NUMERICA(01.000)
link_active_in	*norma, date*	Fecha de alta del enlace IDT-DAT. Si está activo, coincide con la MARCATIEMPO
link_active_out	*norma, date*	Fecha de baja del enlace IDT-DAT. Obligatorio cuando está inactivo

Tabla 4 .Diseño de registro del URD-II: idt_poblacion - <<Table>>

3.3.1. Tipo de enlaces

La identidad de una persona se basa en las siguientes variables:

- NEWIPF: Tipo de ipf + identificador+letra recalculada. Si el NEWIPF empieza por "1" se trata de DNI, si empieza por "3" se trata de un NIE.
- SEXO: 1=Hombre; 2=Mujer (normalizado bajo la norma SEX_ISO_5218).
- FNAC: Fecha de nacimiento de la persona física.
- NOMBRE: Nombre de la persona física.
- APELLIDO1: Primer apellido de la persona física (contiene partículas).
- APELLIDO2: Segundo apellido de la persona física (contiene partículas).
- PAISNAC: País de nacimiento de la persona física (normalizado a la clasificación M49_GEO_REGIONS de las Naciones Unidas).
- CODMUNNAC: Municipio de nacimiento de la persona física (normalizado a la clasificación interna ISTAC:CL_AREA_ES(01.001)).

Además:

- A*= Cuando el NEWIPF es del tipo DNI o NIE
- C*= Contiene reglas de similitud entre fechas de nacimiento (ejemplo: "DDMM" por "MMDD")

Tipo de enlace	Genera			TIPO_IPF	Elimina partículas "DE", etc	Contrae el nombre "MARIA"	Aplica en la fuente
E00A: A+B+C+D+E+F+G+H	Solo URD II	No genera IDF	No genera IDT	Todos	No	No	Solo PMH
E00A: A+B+C+D+E+F							Todas exc.PMH
E01A: A*+ C* + (D||E||F)	URD II y II	IDF		Solo DNI,NIE	Sí	Sí	Todas
E01B: A* +C* + (E||F) + [D=*D or COMPGED(D)]							
E01C: A* +C*+ (D||E||F)				Solo DNI,NIE (Dat)			
E01D: A* + C*+ (E||F) + [D=*D and COMPGED(D)]							
E01E: A*+C + COMPGED (E||F||D<=500)							Solo AFI
E02A: A* + [INDEX(SOUNDEX_ES(E) and SOUNDEX_ES(F)) or (C)]							Todas
E03A: A*+INDEX (SOUNDEX_ES(E) and SOUNDEX_ES(F)) + COMPGED (SOUNDEX_ES(D) ||SOUNDEX_ES(E) ||SOUNDEX_ES(F) <=300)				_X (Dat)			
E04A: [C* + (D||E||F)=* (D||E||F) or INDEX(D||E||F) and COMPGED(D||E||F)] or [A* and (D||E||F)=* (D||E||F) or INDEX(D||E||F)]				_X or DNI,NIE			
E04B: C* + A* + [(D||E||F)=*(D||E||F) or COMPGED(D||E||F)]							PMH <=2007
E05A: A*+ B+ C*+(D||E||F) donado por el IDF		No genera IDF		Solo DNI,NIE			DEM CON LLAM
E05B: A* + (D||E||F) =" _U"		IDF					
Sin enlace - IDF e IDT nuevos			IDT	N/A	N/A	N/A	Todas

Tabla 5. Enlaces según variables y características de integración

3.4. Integración de ficheros al Registro de población

El ingreso de nuevos datos al Registro de Población depende del calendario de planificación de ficheros que irán entrando en función del período de referencia y disponibilidad de las BDDD durante el período de referencia.

Fuente	Recepción: Año aaaa											
	Ene	Feb	Mar	Abr	May	Jun	Jul	Ago	Sep	Oct	Nov	Dic
Padrón Munic. de Hab.		RD PMH aaaa 0101						RD PMH aaaa 0731				
Afiliados a la SS	RD AFI (aaaa -1) 1231			RD AFI aaaa 0331			RD AFI aaaa 0630			RD AFI aaaa 0930		
Afiliados a MUFACE	RD MUF (aaaa -1) 1231			RD MUF aaaa 0331			RD MUF aaaa 0630			RD MUF aaaa 0930		
Demandantes	RD DEM (aaaa -1) 1231	RD DEM aaaa 0101	RD DEM aaaa 0228	RD DEM aaaa 0331	RD DEM aaaa 0430	RD DEM aaaa 0531	RD DEM aaaa 0630	RD DEM aaaa 0731	RD DEM aaaa 0831	RD DEM aaaa 0930	RD DEM aaaa 1031	RD DEM aaaa 1130
Contratos Laborales	RD CON (aaaa -1) 1231	RD CON aaaa 0101	RD CON aaaa 0228	RD CON aaaa 0331	RD CON aaaa 0430	RD CON aaaa 0531	RD CON aaaa 0630	RD CON aaaa 0731	RD CON aaaa 0831	RD CON aaaa 0930	RD CON aaaa 1031	RD CON aaaa 1130
LLamamientos	RD LLAM (aaaa -1) 1231	RD LLAM aaaa 0101	RD LLAM aaaa 0228	RD LLAM aaaa 0331	RD LLAM aaaa 0430	RD LLAM aaaa 0531	RD LLAM aaaa 0630	RD LLAM aaaa 0731	RD LLAM aaaa 0831	RD LLAM aaaa 0930	RD LLAM aaaa 1031	RD LLAM aaaa 1130
Egresados Univ.	Cuando esté disponible											

Figura 4. Calendario de ficheros administrativos según período de recepción y período de referencia

Para monitorear la influencia cronológica de los ficheros, graficar las tendencias de cada fuente administrativa y calcular el porcentaje de versiones aportadas, se realizará la integración hacia adelante partiendo del año piloto 2017, siguiendo el calendario de ficheros según su fecha de referencia y disponibilidad durante el año (Forward).

Para integrar la serie 2011-2016 y evitar saltos bruscos en la incorporación de nuevas versiones o revisar cambios de diseño de registros, se realizará la carga del año anterior empezando por el segundo semestre y luego por el primer semestre (Backward).

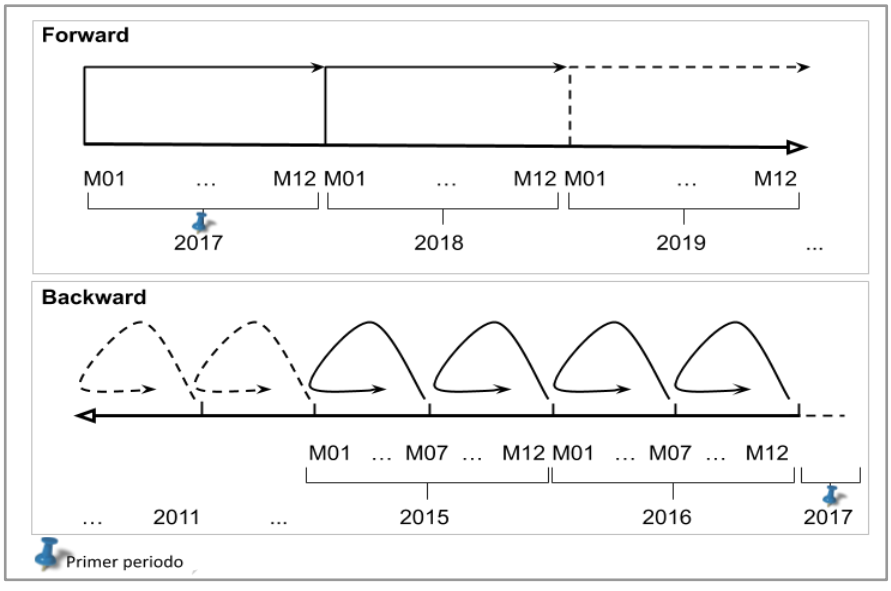

Figura 5. Integración de ficheros según período

3.4.1. Año piloto 2017

Se eligió el Padrón Municipal de Habitantes 01/01/2017 como primer fichero administrativo del Registro de Población ya que combina la mayor cantidad de observaciones y es uno de los primeros que llega en el año. Su información permitió realizar el alta de las personas mediante enlaces para identificar registros duplicados IDF y asignarles un IDT. Estas relaciones se almacenan y se cargan los campos demográficos en el primer registro de versiones de personas.

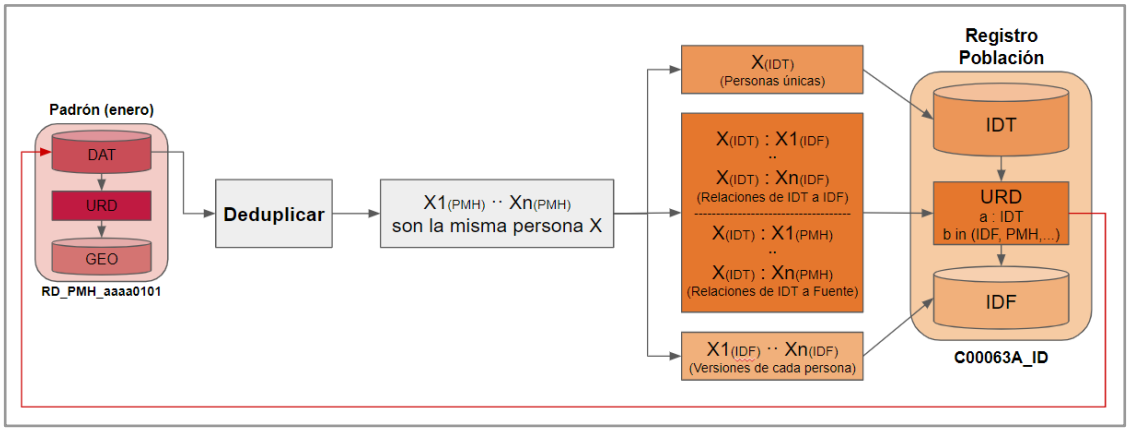

Figura 6. Deduplicación e integración del primer fichero del Padrón

En esta primera carga se obtuvo que el 99,38% de las observaciones del Padrón de enero 2017 eran de personas distintas (fase 1). El resto de observaciones eran personas que aparecían más de una vez y que debían deduplicarse en la fase 2. Esto obedece a posibles cambios de residencia: personas que están de alta en un municipio y que aún no se les ha dado de baja en el municipio anterior pero que por métodos de enlazamientos y aplicación de funciones se han podido detectar a las personas (IDT), sus versiones (IDF) y se han almacenado las relaciones (URD).

Fase	Registros	Obs	% Obs
1	Únicos	2.174.703	99,38%
2	Únicos	12.363	0,56%
	Duplicados	1.244	0,06%
	Triplicados	2	0,00%
Total		**2.188.312**	**100%**

Tabla 6. Deduplicación del Padrón Municipal de Habitantes 01/01/2017

3.4.2. Integración y carga del resto de fuentes administrativas

El fichero del padrón de 2017 inició el reconocimiento de las primeras unidades de personas IDT y se han almacenado los primeros ficheros de versiones IDF y de relaciones URD en el repositorio ID. A partir de ahora, los ficheros (DAT) de entrada se enlazarán

con los datos de los ficheros del Registro de Población IDF y URD. Al momento de enlazar pueden surgir tres casos:

- Enlaces por igualdad: cuando algunas observaciones del DAT enlazan perfectamente con el Registro de Población disponible. E00A, E05A son los enlaces de mejor calidad que no generan nuevos IDF ni IDT.
- Enlaces por similitud: cuando algunas observaciones del DAT de entrada tienen características comunes pero no son exactamente iguales con respecto al Registro de Población disponible. E01A, E01B, E01C, E01D, E01E, E02A, E04A, E04B, E05B son enlaces que generan una versión de IDF, pero no generan nuevo IDT.
- Sin enlace: cuando algunas observaciones del DAT de entrada no enlazan con el Registro de Población disponible. Se generan nuevos IDF y nuevos IDT.

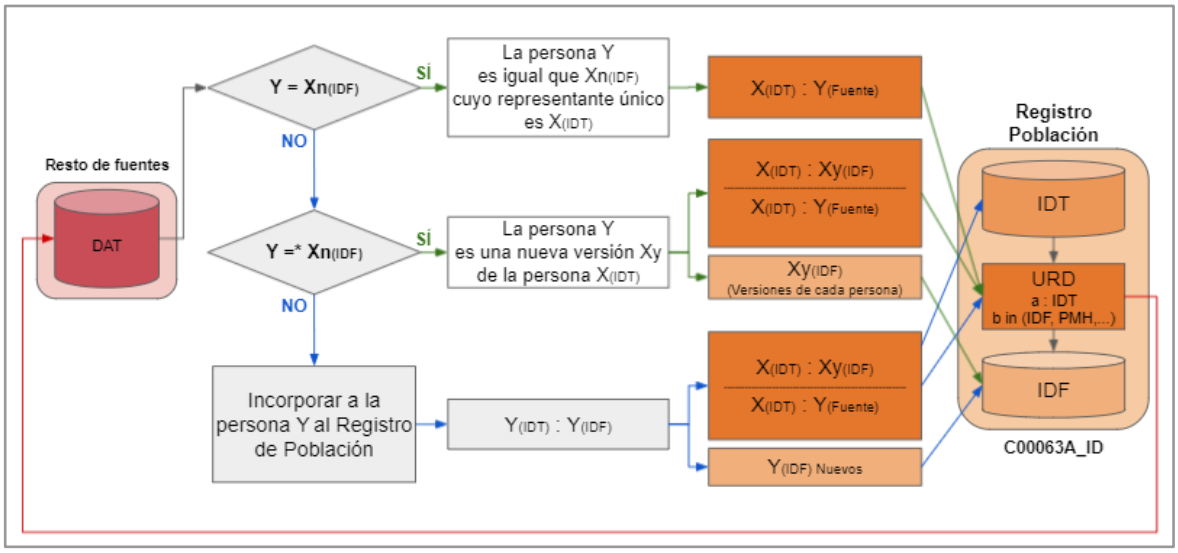

Figura 7. Integración de fuentes administrativas al Registro de Población

3.4.3. Procesos ETL

Los procesos ETL se ejecutan dentro de hilos computacionales capaces de realizar tratamiento sobre los microdatos que son generadores de un cambio de naturaleza de los mismos. Para el Registro de Población se ha generado una ETL por cada fuente administrativa debido a la naturaleza del dato en crudo que proviene de los ficheros de microdatos.

Sin embargo, hay fuentes que han tenido cambios en la forma de recolección de los datos o cambios coyunturales en algún momento y para solventarlo se han creado versiones de ETL.

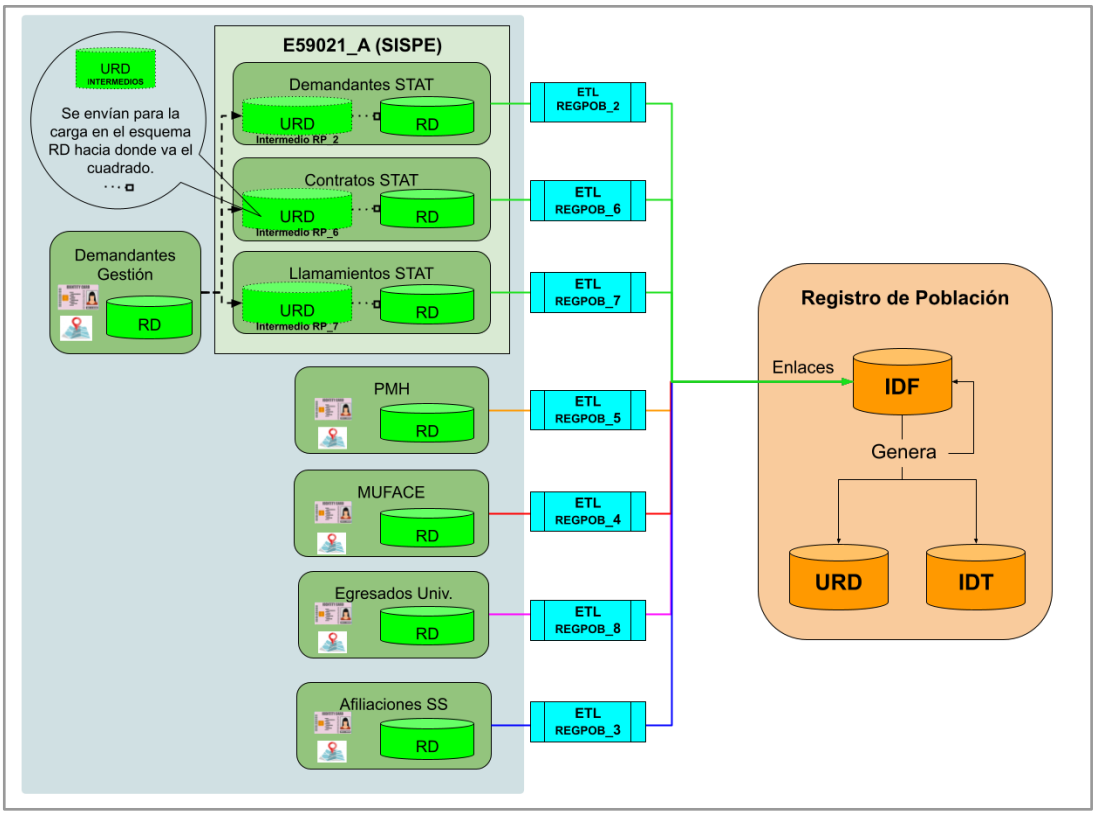

Figura 8. Proceso ETL de integración de ficheros

Los prototipos de ETL de integración de ficheros al Registro de Población por ahora se ejecutan en el software SAS, con miras a su automatización en HOP y se dividen en cuatro fases. La figura 9 representa un ejemplo de ETL de integración del Padrón al Registro de Población

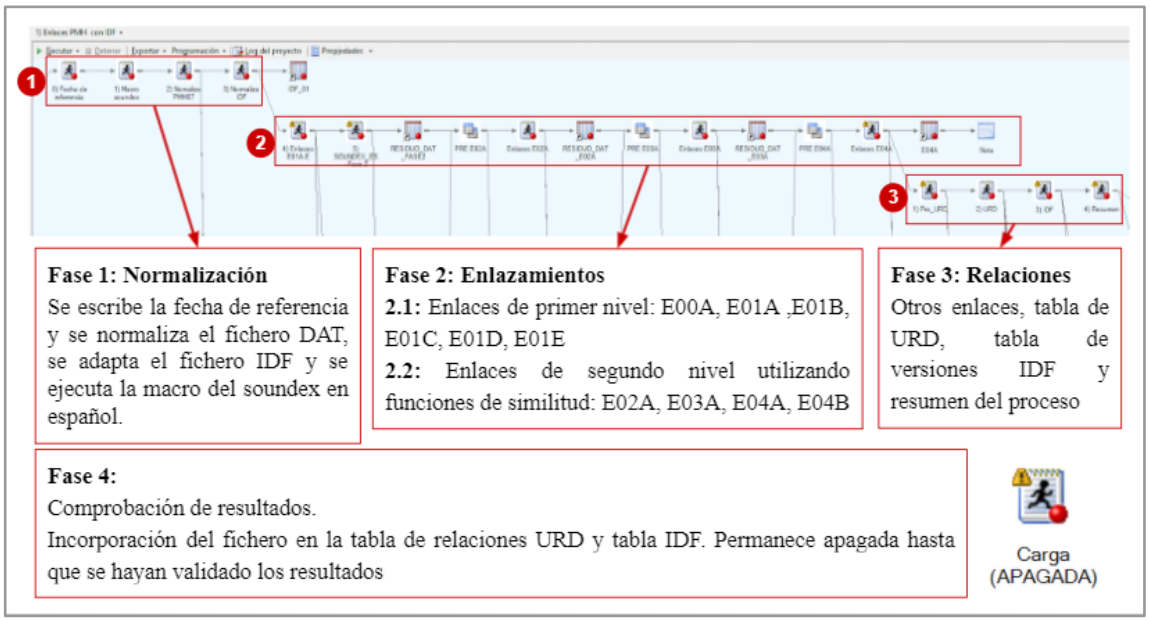

Figura 9. Prototipo de ETL de integración del padrón al Registro de Población

3.5. Resultados Generales. Serie 2003-2021

Hasta el momento se han integrado al Registro de Población 7 fuentes administrativas:

- Padrón Municipal de Habitantes. Serie anual 2003-2009 y serie semestral 2010-2021(jul)
- Demandantes de empleo del Sistema de Información de los Servicios Públicos de Empleo (SISPE). Serie mensual 2011-2021(sep)
- Contratos laborales registrados del SISPE. Serie mensual 2011-2021(sep)
- Llamamientos laborales del SISPE. Serie mensual 2011-2021(sep)
- Afiliaciones a la Seguridad Social. Serie trimestral 2011-2020 (abr) y serie mensual 2020(may) -2021(sep)
- Afiliaciones a la MUFACE. Serie trimestral 2011-2021(jun)
- Egresados universitarios. Serie anual 2015-2018

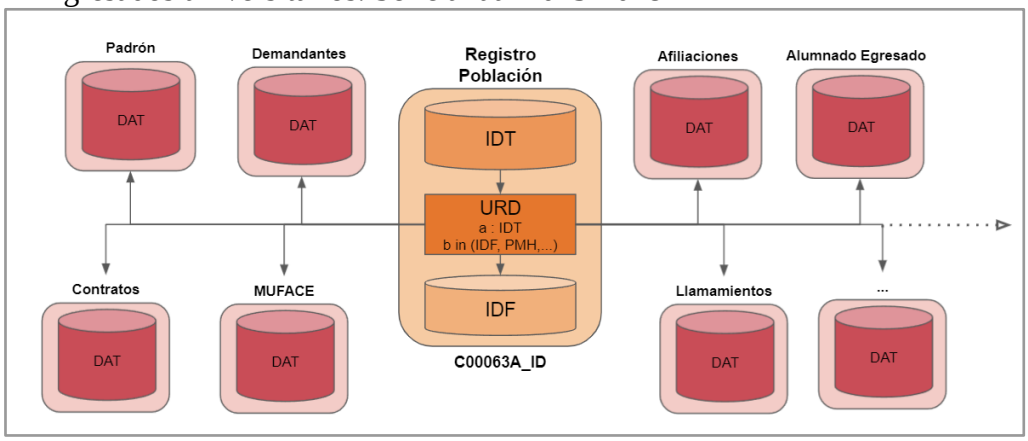

Figura 10. Ecosistema del Registro de Población

La tabla 7 muestra que durante los años 2003 y 2021 se han encontrado 3.3 millones de personas distintas en los registros administrativos antes descritos de Canarias (IDT), que se identifican a través 7.3 millones de versiones (IDF) y las relaciones se han almacenado en una tabla URD con más 169 millones de observaciones.

Fuente	Ficheros analizados	Observaciones	Nro. IDF	Nro. IDT	Nro. URD
Padrón municipal de hab.	31	67.031.646	6.397.617	3.133.164	73.429.263
Afiliaciones a la SS	54	40.073.268	487.120	102.946	40.560.388
Contratos	129	7.546.002	65.648	49.694	7.611.650
Demandantes	129	45.194.052	371.086	47.266	45.565.138
Afiliaciones a MUFACE	43	1.706.324	15.568	1.416	1.721.892
Egresados universitarios	4	27.125	4.975	242	32.100
Llamamientos	129	154.138	128	86	154.266
Total general	**519**	**161.732.555**	**7.342.142**	**3.334.814**	**169.074.697**

Tabla 7. Ficheros incluidos en el Registro de Población. Serie 2003-2021 sep

4. CONCLUSIONES

El Registro de Población permite conocer la vida administrativa de las personas y su participación en los distintos ficheros integrados en la tabla de relaciones URD y a su vez permite introducir buenas prácticas que hagan más eficiente la producción de estadísticas del ISTAC.

La infraestructura del Registro de Población está en constante aprendizaje cada vez que se integra una fuente administrativa ya que aporta nuevas versiones del individuo y a medida que se integren más ficheros o fuentes, el aprendizaje de versiones (IDF) irá decreciendo.

Al mismo tiempo ésta es la vía para utilizar sus datos y construir un Marco de Población Presente en Canarias que serviría de apoyo para las encuestas y resto de proyectos a fin de reducir el número de variables ya que podrían obtenerse haciendo uso del URD y consultando los registros administrativos necesarios.

5. REFERENCIAS BIBLIOGRÁFICAS

Amón, Iván; Moreno, Francisco; Echeverri, Jaime (2012). Algoritmo fonético para detección de cadenas de texto duplicadas en el idioma español. Revista Ingenierías Universidad de Medellín, vol. 11, No. 20 pp. 127-138 - ISSN 1692-3324. Disponible en: http://www.scielo.org.co/pdf/rium/v11n20/v11n20a11.pdf

Christen, P. (2006) "A Comparison of Personal Name Matching: Techniques and Practical Issues", Sixth IEEE International Conference on Data Mining, pp. 290-294.

Pinto, David; et al (2012). The Soundex Phonetic Algorithm Revisited for SMS-based Information Retrieval. II Congreso Español de Recuperación de la Información. Disponible en: https://books.google.es/books?id=je7VCgAAQBAJ&pg=PA99&lpg=PA99&dq=Soundex+R+en+espa%C3%B1ol&source=bl&ots=PjOLsulabA&sig=ACfU3U3fm5_Sr6dR6ZomdAXD3TPDkxP4WQ&hl=es&sa=X&ved=2ahUKEwjbhYHTg7LgAhVLxxoKHWv_BOo4ChDoATAAegQIAhAB#v=onepage&q=Soundex%20R%20en%20espa%C3%B1ol&f=false

SAS (2015). *How to perform a fuzzy match using SAS functions*. Disponible en: https://blogs.sas.com/content/sgf/2015/01/27/how-to-perform-a-fuzzy-match-using-sas-functions/

SAS (2011). *SOUNDEX Function*. Language Reference: Dictionary, Fourth Edition. Disponible en: http://support.sas.com/documentation/cdl/en/lrdict/64316/HTML/default/viewer.htm#a000245948.htm

Yefim, Gershteyn (2000). Use of SPEDIS Function in Finding Specific Values. Coders' Corner, Paper 86-25. Disponible en: https://support.sas.com/resources/papers/proceedings/proceedings/sugi25/25/cc/25p086.pdf