Islas Canarias
Del 15 al 19 de noviembre de 2021

REGISTRO INTEGRADO DE EMPRESAS DE CANARIAS

Jesús Alberto González Yanes
Instituto Canario de Estadística, ISTAC
jgonyanp@gobiernodecanarias.org

Domingo Jesús Lorenzo Díaz
Instituto Canario de Estadística, ISTAC
dlordia@gobiernodecanarias.org

Bencomo Domínguez Martín
Ecosistemas Virtuales y Modulares, EVM
bencomo.dominguez@evm.net

Índice

Introducción

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 Instituto Canario de Estadística los ficheros administrativos de los que sean titulares y que sean necesarios para el ejercicio de la función estadística. Los referidos ficheros deberán adecuarse, para su remisión, a los requisitos técnicos que establezca el Instituto. Asimismo se señala que el banco de datos administrativos para fines estadísticos deberá facilitar la fusión de los ficheros para fines estadísticos.

El banco de datos tiene como objetivo 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 se impulsará el Sistema de Datos Integrados (iDatos). Para enmarcar iDatos dentro de la arquitectura del banco de datos es necesario entender la organización del mismo y los componentes que lo conforman, lo cual se muestra en la siguiente figura:

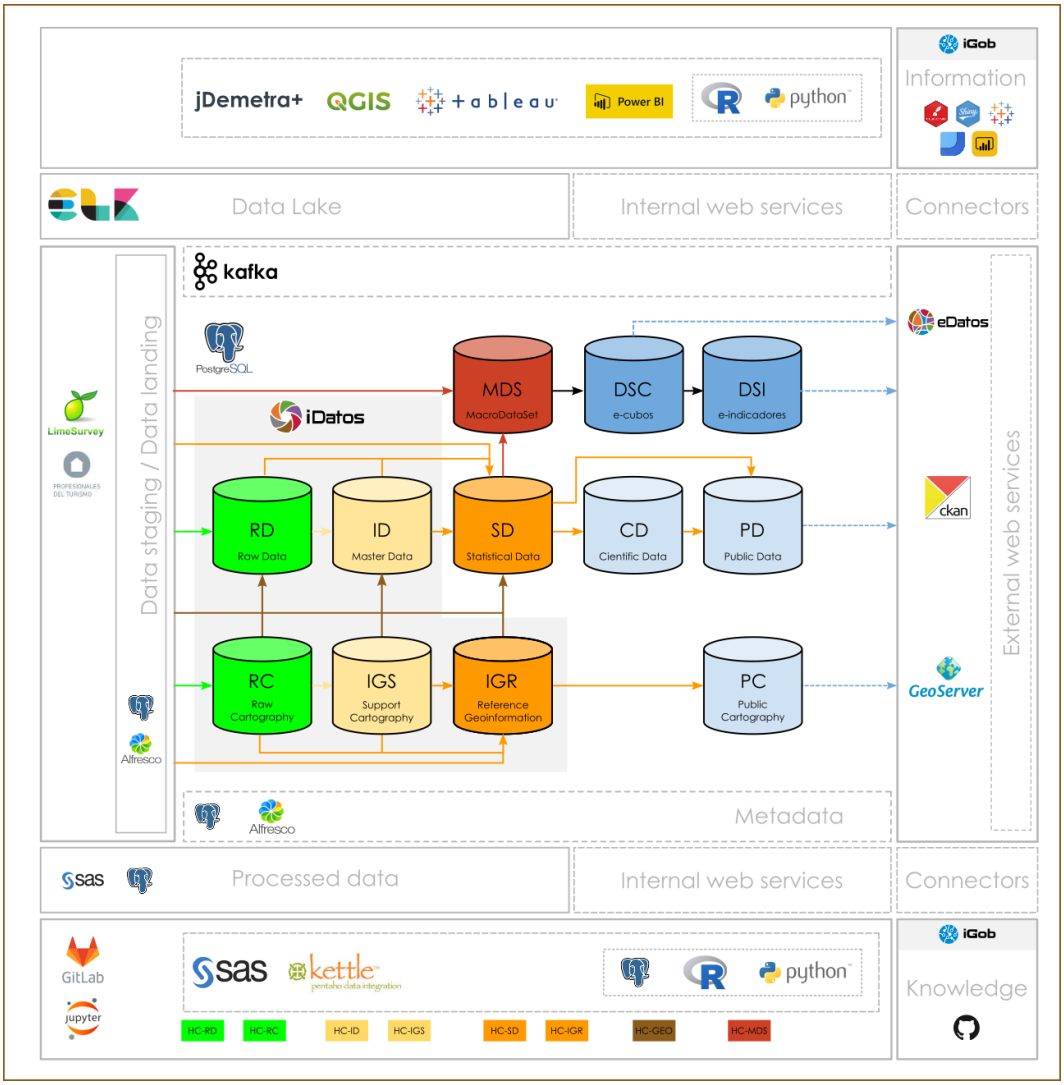


Figura 1: Modelo de arquitectura del banco de datos

El banco de datos está dividido en los siguientes entornos:

- **Entorno de recepción o captura (Data Collection and Staging Area):** Es el área en la que se lleva a cabo la recopilación y preparación de los datos (Staging Area) y se considera como el principal punto de inclusión de datos. Estos datos de entradas pueden clasificarse

según su estructura en datos no estructurados, semiestructurados y estructurados, y estadísticamente en microdatos o macrodatos.

- **Entorno repositorio:** Este entorno se organiza según la naturaleza de los ficheros en esquemas-tipo, que a su vez se clasifican en tres grupos según almacenen cartografías, microdatos o macrodatos, dentro de los cuales se encuentran los siguientes tipos de datos.

Grupo	Tipo de datos
Cartografía	Raw Cartography (RC)
	Support Cartography (IGS)
	Geographic Information Reference (IGR)
	Public Cartography (PC)
Microdatos	Raw Data (RD)
	Master Data (ID)
	Statistical Data (SD)
	Scientific Data (CD)
	Public Data (PD)
Macrodatos	MacroDataSet (MDS)
	DataSetCube (DSC)
	IndicatorsCube (DSI)

- **Entorno computacional:** Es el entorno encargado del inicio, control, seguimiento y registro de los procesos en la capa de datos del Sistema Integrado.
- **Entornos analíticos:** El entorno analítico contiene los almacenes de datos de forma optimizada para el procesamiento analítico en línea (OLAP), de forma que junto al repositorio de metadatos, contiene toda la información necesaria para las operaciones de análisis.
- **Entornos de difusión:** El entorno de difusión proporciona toda la funcionalidad necesaria para proporcionar los productos finales del entorno analítico a usuarios externos. Dicho entorno debe preservar tanto como sea posible el contenido informativo de los datos recopilados y procesados, garantizando al mismo tiempo que no se pueda volver a identificar a personas concretas, lo que se conoce como el problema del control de divulgación estadísticas (SDC).

Los esquemas-tipo sobre los que trabaja el Sistema de Datos Integrados (iDatos), dentro del entorno repositorio son:

- Raw Cartography (RC)
- Support Cartography (IGS)
- Geographic Information Reference (IGR)
- Raw Data (RD)

- Master Data (ID)

En particular, los Master Data (ID) representan los esquemas de microdatos maestros, en los cuales podemos enmarcar los cuatro grandes directorios: Lugares, Edificios, viviendas y locales (DEV), Población y hogares (DPH) y Empresas y Establecimientos (DUE). Existen hilos computacionales entre los diferentes directorios y también entre los distintos registros que conforman cada directorio. El Registro Integrado de Empresas está enmarcado en el directorio de Empresas y Establecimientos.

Objetivos

El Registro Integrado de Empresas se plantea como una infraestructura estadística integradora de las distintas fuentes de datos relacionadas con empresas de Canarias, conformada principalmente por registros administrativos y con el objetivo de permitir identificar a las empresas según un conjunto básico de variables. Estos datos se someten a un proceso de normalización y se establecen relaciones entre ellos por medio de enlaces determinísticos y probabilísticos (Record Linkage).

Metodología

Estructura del Registro de Empresas

El Registro Integrado de Empresas cuenta con tres grandes conjuntos de datos.

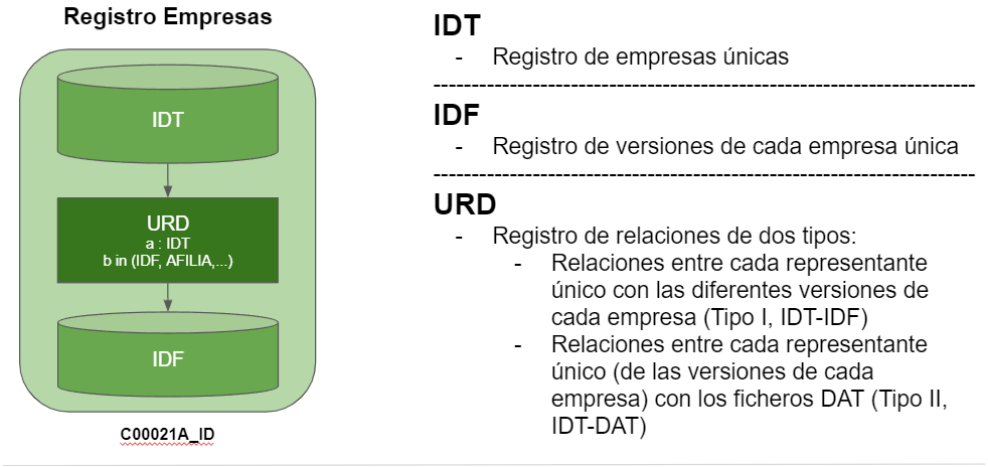


Figura 2: Componentes del Registro Integrado de Empresas

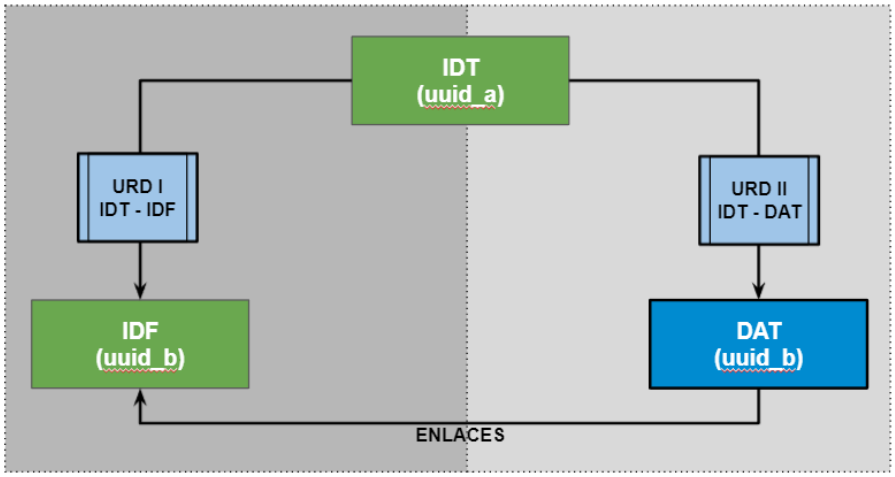


Figura 3: Tipos de relaciones en el URD

Diseño de registros del IDF

Tras un estudio de las diferentes fuentes de microdatos se decidió que las variables básicas que definían una empresa y que iban a estar presentes en el IDF son las siguientes:

BLOCK_1	CONCEPT_ID	TECH_TYPE	TECH_SIZE	LABEL
IDENTIFICADORES	UUID	varchar	36	Identificador Único Universal
	LUID	serial		Identificador Único Local
	STID	varchar	61	Identificador de esquema y tabla
	MARCATIEMPO	date	9	Sello de tiempo de creación de la observación
VARIABLES COMUNES	EMPRESA_CCOTIZA	varchar	15	Código de la Cuenta de Cotización de la empresa
	EMPRESA_PERSONA_FISICA	varchar	2	Persona Física (Sí/No)
	EMPRESA_IDENTIFICADOR_NIF	varchar	8	Número de identificación fiscal
	EMPRESA_NIF_LIDEN	varchar	2	Carácter de control del NIF
	EMPRESA_NIF_LIDEN_RC	varchar	2	Carácter de control del NIF (recalculado)
VARIABLES PERSONA FÍSICA	EMPRESA_NOMB_APES	varchar	110	Nombre y apellidos de la persona física
	EMPRESA_NOMBRE	varchar	45	Nombre completo de la persona física
	EMPRESA_NOMBRE_1	varchar	20	Primer nombre de la persona física
	EMPRESA_NOMBRE_2	varchar	20	Resto de nombres de la persona física
	EMPRESA_APELLIDOS	varchar	65	Apellidos de la persona física
	EMPRESA_PARTICULA_APELLIDO1	varchar	6	Partícula apellido 1
	EMPRESA_APELLIDO_1	varchar	25	Primer apellido de la persona física
	EMPRESA_PARTICULA_APELLIDO2	varchar	6	Partícula apellido 2
	EMPRESA_APELLIDO_2	varchar	25	Segundo apellido de la persona física
	EMPRESA_FNAC	date	10	Fecha nacimiento

	EMPRESAR_SEXO	varchar	2	Sexo
	EMPRESA_CODMUNNAC	varchar	5	Municipio de nacimiento
	EMPRESA_PAISNAC	varchar	13	País de nacimiento
VARIABLES PERSONA JURÍDICA	EMPRESA_RAZON_SOCIAL	varchar	170	Razón social de la empresa
	EMPRESA_NOMBRE_COMERCIAL	varchar	170	Nombre comercial de la empresa
	EMPRESA_TELEFONO	varchar	9	Teléfono de la empresa
	EMPRESA_EMAIL	varchar	80	Email de la empresa
	EMPRESA_WEB	varchar	90	Página web de la empresa
	EMPRESA_LONGITUD	varchar	20	Longitud de la ubicación de la empresa
	EMPRESA_LATITUD	varchar	20	Latitud de la ubicación de la empresa
	EMPRESA_GEOHASH	varchar	10	Geohash de la ubicación de la empresa

Figura 4: Diseño de registros del IDF

Dado que en el registro de empresas se incluyen los trabajadores autónomos, se han dividido las variables que conforman el IDF en cuatro grupos: identificadores, variables comunes, variables de persona física (autónomos sin empleados) y variables de persona jurídica (empresas y autónomos con empleados). En la actualidad, la única fuente de administrativa de la que se obtienen datos de autónomos sin empleados son los ficheros de Afiliaciones a la Seguridad Social.

Diseño de registros del URD

Como se ha comentado anteriormente, en el URD se almacenan dos tipos de relaciones (IDT-IDF e IDT-DAT), que aunque comparten el mismo diseño de registros, no almacenan la misma información.

Relaciones del Tipo I (IDT-IDF)

Variable	Tipo	Descripción
uuid	*norma, varchar (36)*	*(UUID de la relación, se utiliza en la parte II cuando IDT-DAT [1])*
stid	*norma, varchar (61)*	c00021a_id.urd_empresas ([2])
luid	*norma, integer*	*(Identificador único local de la relación)*
marcatiempo	*norma, date*	*(Fecha de creación de la relación, DDMMMAAAA)*
uuid_a	*norma, varchar (36)*	*(UUID del IDT: puede repetirse para un conjunto de IDF´s)*
stid_a	*norma, varchar (61)*	c00021a_id.idt_empresas
uuid_b	*norma, varchar (36)*	*(UUID del IDF: diferentes nombres que tiene un mismo registro)*
stid_b	*norma, varchar (61)*	c00021a_id.idf_empresas
rel_type	varchar (30)	SOURCE
rel_period	norma, date	*(Fecha de referencia del .DAT que generó el IDF)*
rel_event	norma, date	(Vacío)
rel_description	varchar (255)	*(stid del .DAT que generó la relación)*

link_type	varchar (30)	RL_DETERMINISTIC
link_description	varchar (255)	TX_ENLACE
link_quality	varchar (255)	“Enlace de n IDF a 1 IDT”
link_quality_rank	varchar (10)	MUY ALTA, ALTA, MEDIA, BAJA
Link_active	varchar (1)	Indicador de actividad del enlace.
Link_active_in	*norma, date*	Fecha de alta del enlace IDT-IDF
Link_active_out	*norma, date*	Fecha de baja del enlace IDT-IDF

Figura 5: Diseño de registros URD - I

Relaciones del Tipo II (IDT-DAT)

Variable	Tipo	Descripción
uuid	*norma, varchar (36)*	*(UUID de la relación)*
stid	*norma, varchar (61)*	c00021a_id.urd_empresas
luid	*norma, integer*	*(Identificador único local de la relación)*
marcatiempo	*norma, date*	*(Fecha de creación de la relación, DDMMMAAAA)*
uuid_a	*norma, varchar (36)*	*(UUID del IDT: puede repetirse para un conjunto de DAT’s)*
stid_a	*norma, varchar (61)*	c00021a_id.idt_empresas
uuid_b	*norma, varchar (36)*	*(UUID de origen del fichero .DAT)*
stid_b	*norma, varchar (61)*	*(stid del fichero .DAT)*
rel_type	varchar (30)	TRACE_STOCK; TRACE_EVENT
rel_period	norma, date	*(Fecha de referencia de inventario o evento del .DAT)*
rel_event	*norma, date*	Obligatorio cuando rel_type es: a) EVENT; b) STOCK y tiene fecha de alta administrativa
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 n DAT a 1 IDT”
link_quality_rank	varchar (10)	MUY_ALTA
Link_active	varchar (1)	Indicador de actividad del enlace.
Link_active_in	*norma, date*	Fecha de alta del enlace IDT-DAT
Link_active_out	*norma, date*	Fecha de baja del enlace IDT-DAT

Figura 6: Diseño de registros URD-II

Enlaces

Cada vez que se incluye en el Registro Integrado de Empresas una nueva fuente de datos, se realiza una actualización del mismo que consiste en buscar enlaces entre las empresas que ya están incluidas en el registro y las del nuevo fichero DAT. Los enlaces los clasificamos principalmente en dos grupos.

- Enlaces de igualdad (Enlaces E00A y E00B): Este enlace se produce cuando una empresa, de un nuevo fichero DAT, ya está incluida en el registro con, exactamente, los mismos datos para cada una de las variables básicas que definen la empresa. En ese caso, no se genera un nuevo representante único (IDT) ni una nueva versión (IDF), y únicamente se generan las relaciones del Tipo II en el URD, es decir, relaciones de los representantes únicos ya existentes con los registros del nuevo fichero DAT.

- Enlaces de similitud (Resto de enlaces): Las empresas del nuevo fichero DAT que no están almacenadas con la misma información en sus variables básicas, pero que se tienen evidencias más que suficientes para asegurar que son la misma empresa que otra que ha sido almacenada previamente, generan un enlace de similitud. Con este tipo de enlaces no se generan nuevos representantes únicos (IDT), sí se crea una nueva versión de la empresa y se almacenan los dos tipos de relaciones en el URD.

- Sin enlace: No se trata de un enlace como tal, pero aquellas empresas del fichero DAT que no enlazan ni por igualdad ni por similitud con las empresas anteriormente incluidas, se consideran nuevas empresas. En este caso, se almacena un nuevo representante único (IDT), una nueva versión (IDF) y los dos tipos de relaciones en el URD.

Los enlaces utilizados en la actualidad y las variables que se tienen en cuenta en cada uno de ellos son los siguientes:

Enlace	LINK_QUALITY_RANK
E00A: NIF + CCC + SEXO + FNAC(DD/MM/AAAA) + NOMB + APE1 + APE2	MUY_ALTA
E02A: NIF + (APE1 + APE2 + NOMB) + SEXO + FNAC*	MUY_ALTA
E02B: NIF + (APE1 + APE2 + NOMB) + (SEXO OR FNAC*)	ALTA
E02C: NIF +COMPGED(SOUNDEX(APE1 || APE2 || NOMB)) + FNAC/FNAC*	ALTA
E02D: (APE1 + APE2 + NOMB) + SEXO + FNAC	MEDIA
E02E: (NIF or TIPO_NIF1 <> TIPO_NIF2) + (APE1||APE2||NOMB)* + FNAC*	MEDIA
E01A: NIF + (APE1 + APE2 + NOMB) = NIF + RAZON_SOCIAL	ALTA
E01B: NIF + COMPGED(SOUNDEX(APE1 || APE2 || NOMB),SOUNDEX(RAZON_SOCIAL))<= 400	ALTA
E01C: NIF	MEDIA
E00B: NIF + RAZON_SOCIAL + (CCC + TELEFONO + EMAIL + WEB + LONGITUD + LATITUD)	MUY_ALTA
E03A: NIF + RAZON_SOCIAL + (CCC OR TELEFONO)	MUY_ALTA
E03B: NIF + RAZON_SOCIAL	ALTA
E03C: NIF + COMPGED(SOUNDEX(RAZON_SOCIAL)) <= 500	ALTA
E03D: TIPO_NIF1 <> TIPO_NIF2 + CCC + RAZON_SOCIAL	MEDIA
E03E: TIPO_NIF1 <> TIPO_NIF2 + CCC + COMPGED(SOUNDEX(RAZON_SOCIAL)) <= 500	MEDIA
E04A: NIF + TELEFONO + EMAIL	ALTA
E04B: NIF + (TELEFONO OR EMAIL)*	ALTA
E04C: COMPGED(SOUNDEX(RAZON_SOCIAL)) <= 200 + TELEFONO + EMAIL	BAJA

E05A: NIF + RAZON_SOCIAL + (TELEFONO OR WEB)	MUY_ALTA
E05B: NIF + COMPGED(SOUNDEX(RAZON_SOCIAL)) <= 500 + (TELEFONO OR WEB)	ALTA
E05C: NIF + (TELEFONO OR WEB)	ALTA
E05D: COMPGED(SOUNDEX(RAZON_SOCIAL)) <= 100 + TELEFONO + WEB	MEDIA

Figura 7: Enlaces de igualdad y similitud

La variable LINK_QUALITY_RANK, presente en el URD, nos indica la calidad del enlace.

Actualización del Registro Integrado de Empresas

La actualización del Registro Integrado de Empresas se realiza por familia de ficheros o fuente. El procedimiento seguido para incluir cada una de las fuentes en el registro es similar, únicamente variando los enlaces de similitud que se realizan para cada una de ellas según las variables básicas de la empresa que estén presentes cada fichero DAT. Por cada fuente administrativa se desarrolla una ETL con la que se realiza la carga en el registro. En caso de modificaciones en el DSD de la fuente de datos, si alguna de las nuevas variables incluidas en los ficheros se encontrara entre las que incluimos en el IDF, sería necesario una nueva ETL para realizar la carga o modificar la existente.

Pasos para la actualización del Registro Integrado de Empresas

El primer fichero DAT que se incluye en el Registro Integrado Empresas genera los primeros identificadores únicos de las empresas (IDT), sus correspondientes versiones (IDF) y los dos tipos de relaciones en el URD, cargándose cada uno de ellos en la tabla de datos que le corresponde. En este punto de partida es importante realizar un proceso de deduplicación en caso de que en este fichero inicial hubiera empresas que cuentan con distintas versiones, para identificar que hacen referencia al mismo representante único.

Una vez han sido almacenados los primeros IDT, IDF y URD, el proceso para cargar una nueva fuente en el registro cuenta con los siguientes pasos:

Paso 1: Selección y normalización de las variables a incluir

En primer lugar, se establece el periodo de referencia del fichero y se ejecuta la macro del SOUNDEX, una herramienta utilizada para buscar similitudes fonéticas entre diferentes cadenas de caracteres y que se aplicará en algunos enlaces. Una vez realizada esta tarea, se toman del fichero fuente (DAT) que se va a incluir en el registro aquellas variables identificativas que han sido seleccionadas para formar parte del IDF. A estas variables se les realiza un proceso de normalización para que el formato coincida con el de las variables ya almacenadas en el IDF con el objetivo de poder realizar comparaciones entre ellas.

En los siguientes pasos se van a comparar las variables de nuestro fichero DAT con las del IDF para buscar relaciones de igualdad o similitud. Para ahorrar tiempo de computación se seleccionan solo las variables del IDF que están presentes en el fichero que vamos a incluir en el registro, creando una copia reducida del mismo. Además, en caso de que se utilice la macro del SOUNDEX en alguno de los enlaces, se creará otra copia del IDF en la que se incluirán, además de las variables que aparecen en el fichero fuente, UUID, STID y las variables a las que se aplicará el SOUNDEX.

Paso 2: Enlace de igualdad

Una vez llevadas a cabo estas tareas de normalización, se comienzan a realizar comparaciones entre el fichero DAT normalizado y la copia reducida del IDF. En primer lugar, se buscan los registros

del fichero de microdatos que son copias exactas de los que ya están incluidos en el IDF, es decir, enlaces de igualdad. Se enfrentan ambos ficheros y una vez que se han detectado los registros que cumplen el enlace de igualdad, estos son descartados para pasar a la siguiente fase en la que buscaremos enlaces de similitud y, los que no lo cumplen (Residuo DAT_0), sí pasan a esta fase. El enlace E00A se utiliza para obtener relaciones de igualdad entre los ficheros de Afiliaciones a la Seguridad Social y el IDF, está destinado a identificar relaciones de igualdad entre trabajadores autónomos. El enlace E00B está orientado a buscar relaciones de igualdad entre el resto de ficheros y el IDF, es decir, establecer relaciones de igualdad entre empresas.

En este punto se generan las relaciones en el URD entre los identificadores únicos universales (UUID) de aquellas empresas que cumplen el enlace de igualdad y los identificadores únicos (Tipo II), a partir de la información que ya estaba incluida en el registro. El proceso a seguir se describe en la siguiente figura:

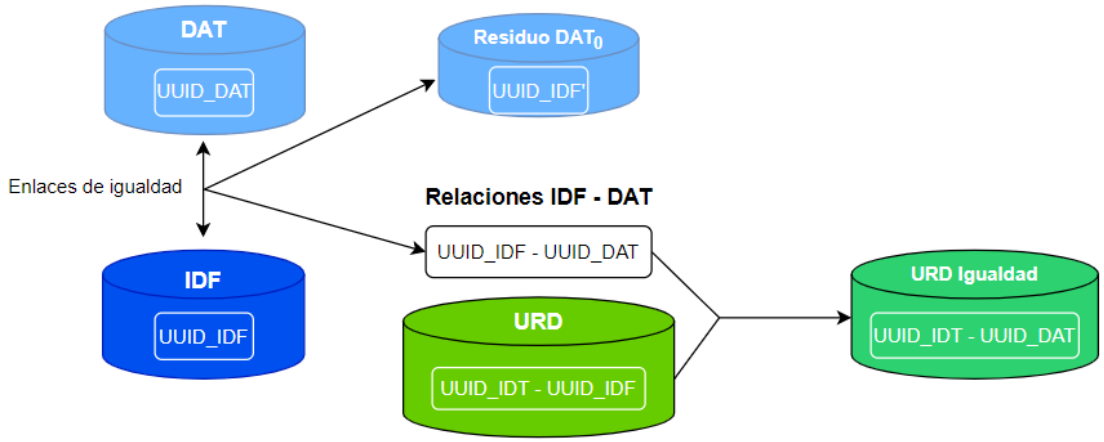


Figura 8: Enlaces de igualdad DAT-IDF

Paso 3: Nuevos IDF y enlaces de similitud

Los registros del fichero DAT que no han enlazado con el IDF (Residuo DAT_0) mediante el enlace de igualdad pasan a la siguiente fase del proceso de carga en el Registro de Integrado de Empresas. Estos registros van a generar nuevas versiones de empresas, es decir, nuevos IDF, por lo que ya se les asigna un nuevo UUID del IDF y se guarda la relación con el UUID del DAT.

El proceso a seguir es similar al realizado para aplicar el enlace de igualdad. Se compara el denominado Residuo DAT_0 mediante el primer enlace de similitud con el IDF, se genera la parte del URD que le corresponde a ese enlace y el resto de registros del DAT que no ha enlazado, comienza un proceso de comparación con el IDF en el que el residuo de un enlace de similitud se utiliza para compararse con el IDF en el siguiente enlace de similitud, lo que se denota como Residuo DAT_i y se generarán tantos como enlaces se apliquen. Normalmente este proceso se realiza de una manera gradual, comenzando primero con enlaces de similitud más fuertes y que involucran a un mayor número de variables, y dejando paso luego a enlaces de similitud más débiles.

Cada vez que se aplica un enlace de similitud, se genera una nueva parte del URD que se va almacenando en un contenedor (Pre-URD). En el Pre-URD se van guardando las relaciones que surgen de aplicar todos los enlaces de similitud y se realiza una única consulta al URD para establecer con qué identificadores únicos (IDT) se relacionan tanto las nuevas versiones (IDF), como los registros del DAT (Figura 10). En las siguientes figuras, UUID_IDF’ representa al UUID

de las nuevas versiones generadas y UUID_IDF al UUID de las versiones ya presentes en el Registro Integrado de Empresas.

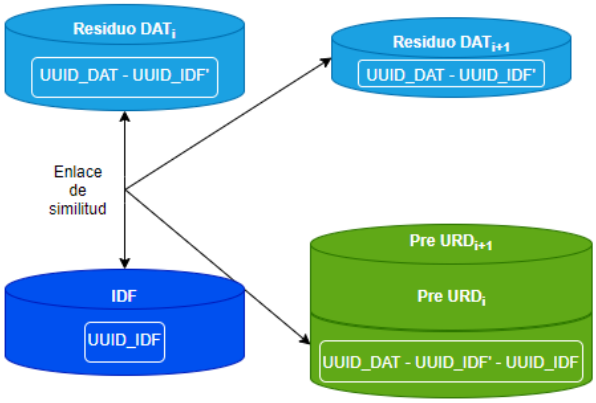


Figura 9: Enlaces de similitud Residuo DAT$_i$ - IDF

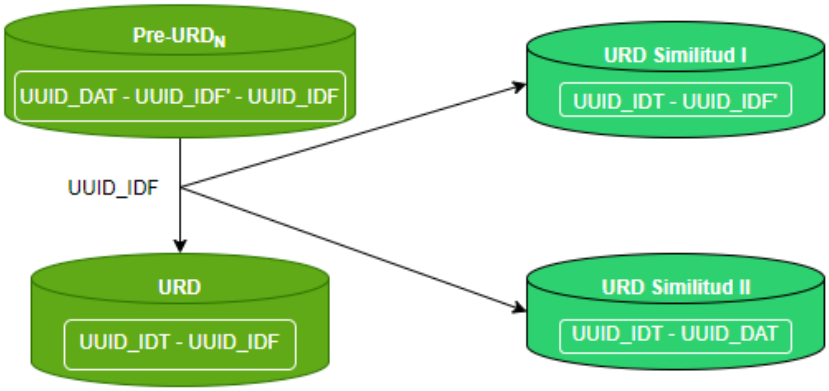


Figura 10: URD de similitud

El procedimiento descrito en el esquema se repite tantas veces como enlaces de similitud se apliquen (N), y el residuo que queda (Residuo DAT$_N$) tras aplicar el último de ellos se considera que son nuevas empresas a las que se debe asignar un nuevo IDT.

Paso 4: Nuevos IDT

Las empresas presentes en el Residuo DAT$_N$, no han sido incluidas previamente en el Registro Integrado de Empresas, por tanto, van a originar un nuevo representante único de la empresa. Al ser residuos que provienen de Residuo DAT$_0$ ya tienen asignados los UUID de las nuevas versiones (UUID_IDF) y están relacionados con los UUID del DAT, en este punto se le asignan los UUID de los nuevos representantes únicos (UUID_IDT), y se generan los dos tipos de relaciones que almacenaremos en el URD.

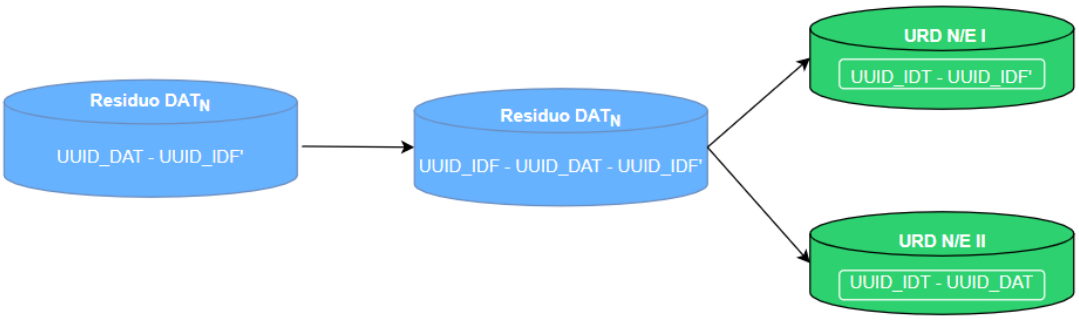
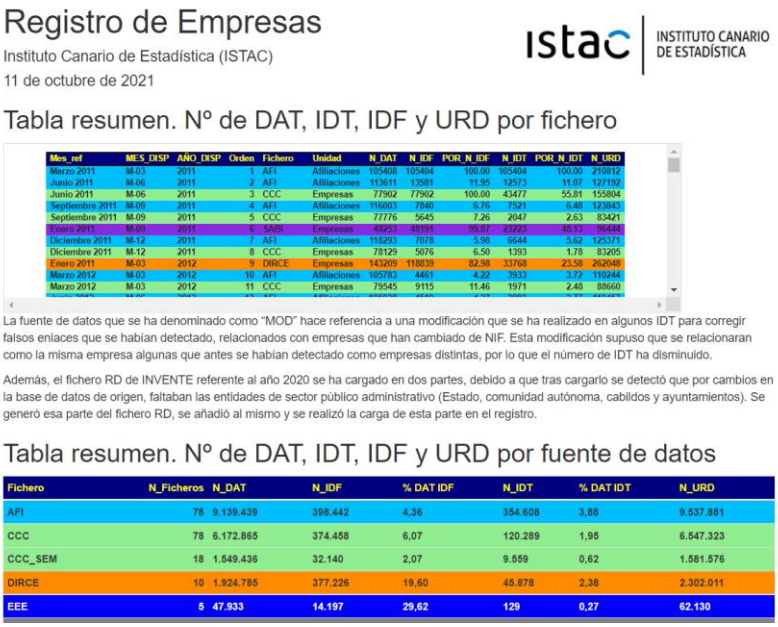


Figura 11: URD de los registros que no enlazan (N/E)

Paso 5: Informe y comprobaciones

Durante todo el proceso se almacenan datos que nos servirán para detectar si podría haber alguna circunstancia excepcional que influya en la carga del fichero DAT en el registro. A partir de estos datos se genera un informe automatizado en R Markdown que permite realizar comprobaciones antes de cargar los conjuntos de datos que hemos generado en el registro.



Registro de Empresas

Instituto Canario de Estadística (ISTAC)

11 de octubre de 2021

istac | INSTITUTO CANARIO DE ESTADÍSTICA

Tabla resumen. Nº de DAT, IDT, IDF y URD por fichero

Mes_ref	MES_DISP	AÑO_DISP	Orden	Fichero	Unidad	N_DAT	N_IDF	POR_N_IDF	N_IDT	POR_N_IDT	N_URD
Marzo 2011	M-03	2011	1	AFI	Afiliaciones	105408	105404	100.00	105404	100.00	210812
Junio 2011	M-06	2011	2	AFI	Afiliaciones	113611	13581	11.95	12573	11.07	127192
Junio 2011	M-06	2011	3	CCC	Empresas	77902	77902	100.00	43477	55.81	155804
Septiembre 2011	M-09	2011	4	AFI	Afiliaciones	116003	7840	6.76	7521	6.48	123843
Septiembre 2011	M-09	2011	5	CCC	Empresas	77776	5645	7.26	2047	2.63	83421
Enero 2011	M-01	2011	6	AFI	Empresas	42423	40918	96.45	42423	100.00	83341
Diciembre 2011	M-12	2011	7	AFI	Afiliaciones	118293	7078	5.98	6644	5.62	125371
Diciembre 2011	M-12	2011	8	CCC	Empresas	78129	5076	6.50	1393	1.78	83205
Enero 2011	M-03	2012	9	DIRCE	Empresas	143209	118839	82.98	33768	23.58	262048
Marzo 2012	M-03	2012	10	AFI	Afiliaciones	105783	4461	4.22	3933	3.72	110244
Marzo 2012	M-03	2012	11	CCC	Empresas	79545	9115	11.46	1971	2.48	88660

La fuente de datos que se ha denominado como "MOD" hace referencia a una modificación que se ha realizado en algunos IDT para corregir falsos enlaces que se habían detectado, relacionados con empresas que han cambiado de NIF. Esta modificación supuso que se relacionaran como la misma empresa algunas que antes se habían detectado como empresas distintas, por lo que el número de IDT ha disminuido.

Además, el fichero RD de INVENTE referente al año 2020 se ha cargado en dos partes, debido a que tras cargarlo se detectó que por cambios en la base de datos de origen, faltaban las entidades de sector público administrativo (Estado, comunidad autónoma, cabildos y ayuntamientos). Se generó esa parte del fichero RD, se añadió al mismo y se realizó la carga de esta parte en el registro.

Tabla resumen. Nº de DAT, IDT, IDF y URD por fuente de datos

Fichero	N_Ficheros	N_DAT	N_IDF	% DAT IDF	N_IDT	% DAT IDT	N_URD
AFI	78	9.139.439	398.442	4,36	354.608	3,88	9.537.881
CCC	78	6.172.865	374.458	6,07	120.289	1,95	6.547.323
CCC_SEM	18	1.549.436	32.140	2,07	9.559	0,62	1.581.576
DIRCE	10	1.924.785	377.226	19,60	45.878	2,38	2.302.011
EEE	5	47.933	14.197	29,62	129	0,27	62.130

Figura 12: Informe automatizado del Registro Integrado de Empresas

Paso 6: Carga en el registro

El último paso, una vez realizadas las comprobaciones pertinentes, consiste en incluir los recursos que hemos ido generando en el Registro Integrado de Empresas. Del Residuo DAT_0 se obtienen las nuevas versiones que se incluyen en la tabla de versiones (IDF). En el caso del URD se toman las distintas porciones que se han generado con relaciones delos tipos II y II y las anexamos al URD. Esto proceso de carga lo ilustra la Figura 13. En el caso del Registro Integrado de Empresas los IDT no se almacenan en una tabla propia, sino que podemos obtenerlos a partir de la columna UUID_A del URD.

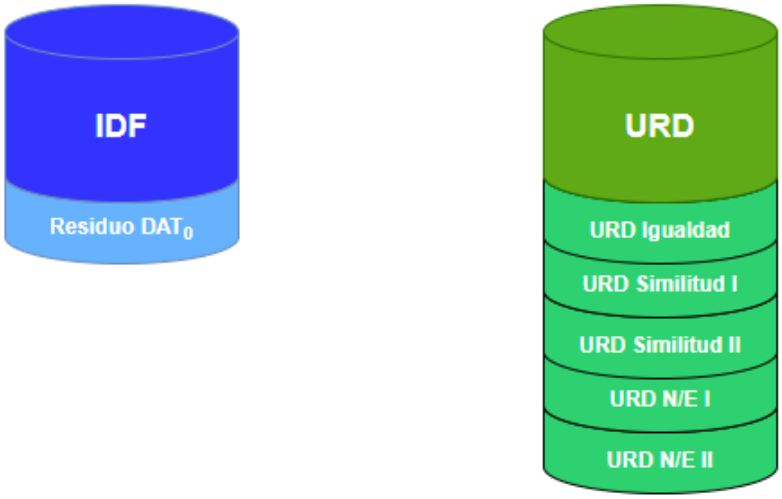


Figura 13: Carga de nuevos registros en IDF y URD

Resultados

En la actualidad el Registro Integrado de Empresas cuenta con las siguientes fuentes administrativas:

1. Ficheros RD de Cuentas de Cotización a la Seguridad Social (mensuales y trimestrales) (Desde 2005 hasta 2021)
2. Ficheros RD de Cuentas de Cotización a la Seguridad Social (semestrales) (Desde 2005 hasta 2013)
3. Trabajadores autónomos de los ficheros RD de Afiliaciones a la Seguridad Social (mensuales y trimestrales) (Desde 2005 hasta 2021)
4. Ficheros RD del Directorio Central de Empresas del INE (DIRCE) (Desde 2011 hasta 2020)
5. Ficheros RD de la Estadística Estructural de Empresas (EEE) (Desde 2015 hasta 2019)
6. Ficheros RD de la Encuesta Industrial Anual de Productos (EIAP) (Desde 2004 hasta 2019)
7. Ficheros RD del Inventario de Entidades del Sector Público Estatal, Autonómico y Local de la Intervención General de la Administración del Estado (INVENTE) (Desde 2014 hasta 2020)
8. Ficheros RD de la base de datos SABI (Desde 2008 hasta 2020)
9. Ficheros SD de Empresas Inscritas en la Seguridad Social (EISS) (Desde 2005 hasta 2021)

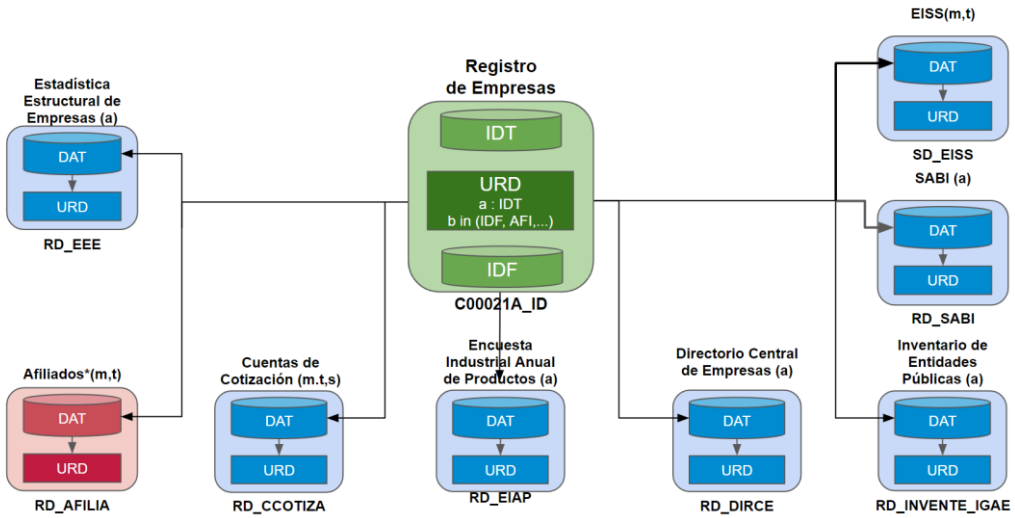


Figura 14: Fuentes administrativas incluidas en el Registro Integrado de Empresas

Si se clasifican por fuente administrativa, los datos obtenidos de la inclusión de ficheros en el Registro Integrado de Empresas en relación al número de ficheros incluidos (N_ficheros), número de registros DAT incluidos (N_DAT), número de versiones de empresas generadas (N_IDF), porcentaje de registros DAT que generan IDF (% DAT IDF), número de empresas únicas detectadas (N_IDT), porcentaje de registros DAT que generan IDT (% DAT IDT) y número de relaciones en el URD generadas (N_URD), se presentan en la siguiente tabla:

Fichero	N_Ficheros	N_DAT	N_IDF	% DAT IDF	N_IDT	% DAT IDT	N_URD
AFI	78	9.139.439	398.442	4,36	354.608	3,88	9.537.881
CCC	78	6.172.865	374.458	6,07	120.289	1,95	6.547.323
CCC_SEM	18	1.549.436	32.140	2,07	9.559	0,62	1.581.576
DIRCE	10	1.924.785	377.226	19,6	45.878	2,38	2.302.011
EEE	5	47.933	14.197	29,62	129	0,27	62.130
EIAP	16	17.833	556	3,12	38	0,21	18.389
EISS	78	5.228.936	0	0	0	0	5.228.936
INVENTE	8	5.631	1.974	35,06	547	9,71	7.605
MOD	1	0	0	0	-2.314	0	0
SABI	13	659.569	150.581	22,83	38.732	5,87	810.150
Total	**305**	**24.746.427**	**1.349.574**	**5,45**	**567.466**	**2,29**	**26.096.001**

Figura 15: Datos del Registro Integrado de Empresas por fuente administrativa

En total se han incluido 305 ficheros en el Registro Integrado de Empresas que contaban con 24.746.427 registros, generándose 1.349.574 versiones de empresas, 567.466 empresas únicos y 26.096.001 relaciones en el URD. La fuente administrativa etiquetada como MOD hace referencia a una modificación que se realizó de los IDT, ya que se detectó un caso en el que empresas que cambiaban de NIF se identificaban como empresas diferentes cuando no lo eran. Al subsanar dicho error, el número de representantes únicos de las empresas disminuyó.

En cuanto al porcentaje total de IDF, IDT y URD aportados al Registro Integrado de Empresas por cada fuente de datos, los resultados son los siguientes:

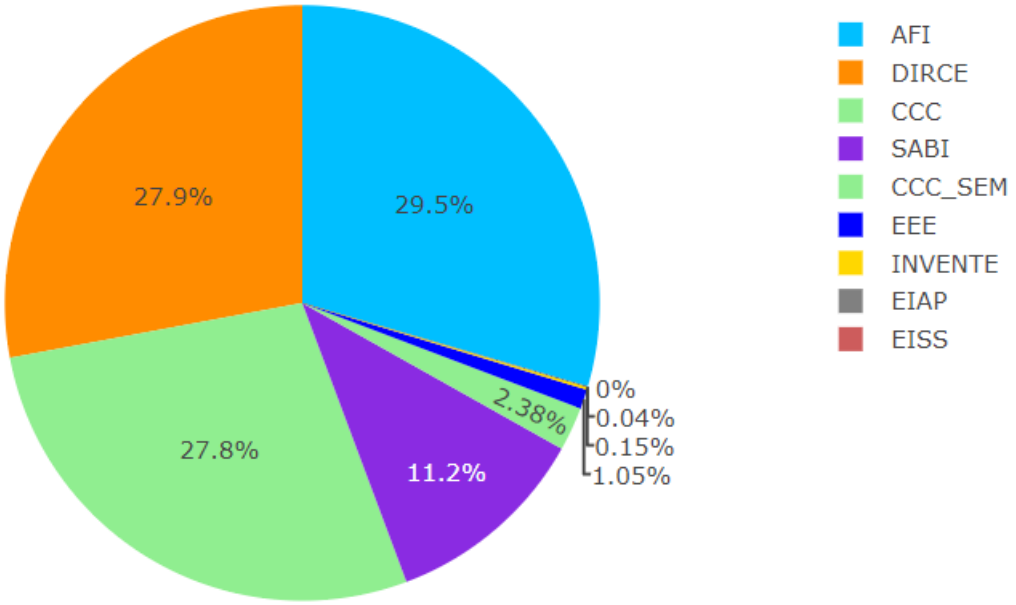


Figura 16: Porcentaje del total de IDF generados por fuente de datos

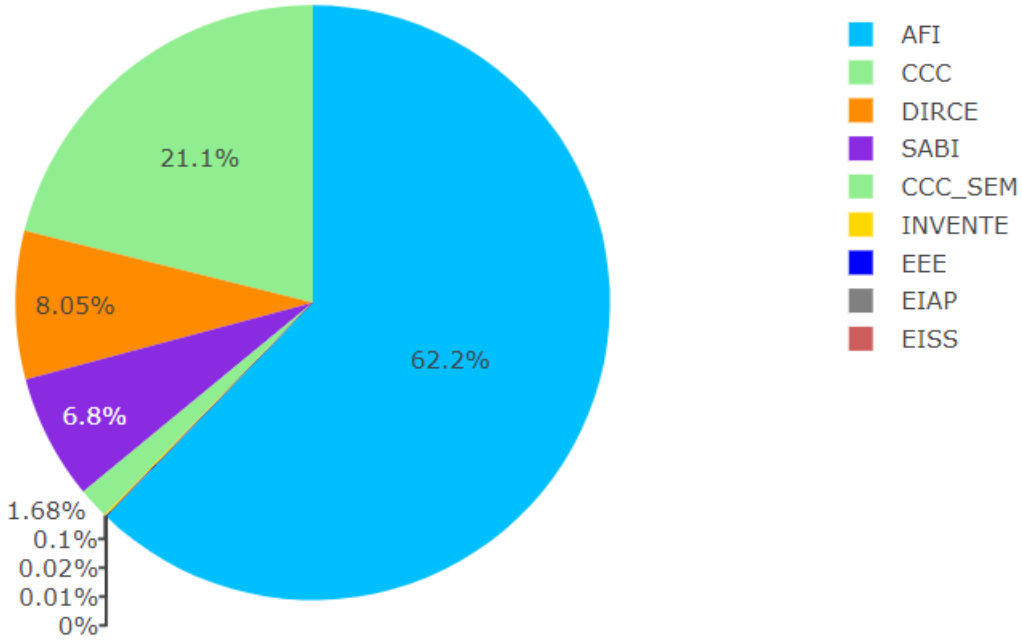


Figura 17: Porcentaje del total de IDT generados por fuente de datos

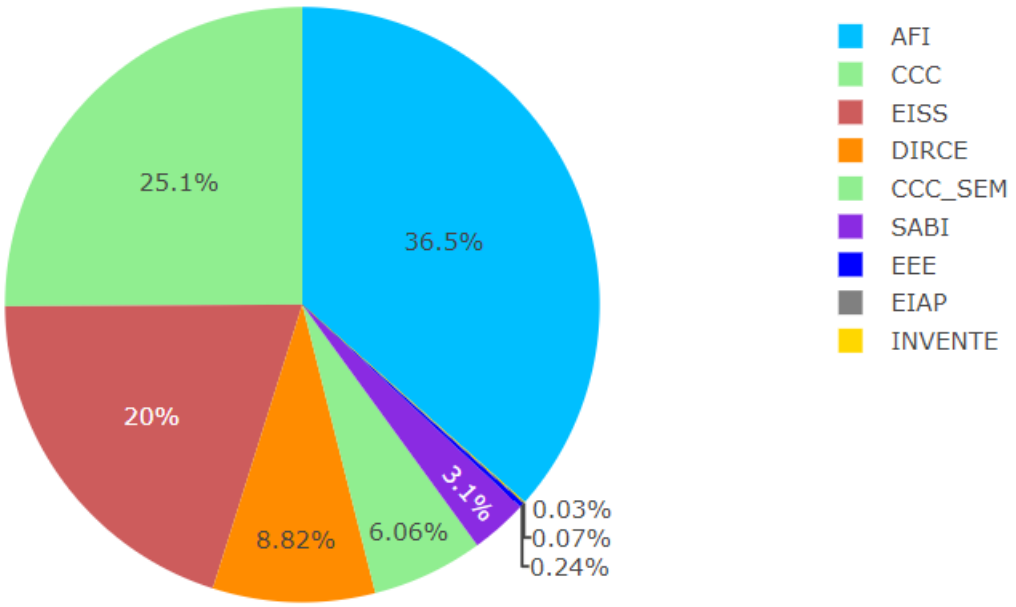


Figura 18: Porcentaje del total de URD generados por fuente de datos

La fuente administrativa que aporta un mayor porcentaje de IDT, IDF y URD es Afiliaciones a la Seguridad Social (sus trabajadores autónomos, en este caso).

Conclusión

En la actualidad, el Registro Integrado de Empresas se muestra como una herramienta potente e innovadora para el almacenamiento de los datos de las empresas de Canarias y, sobre todo, para establecer relaciones entre los ficheros de las distintas fuentes administrativas a lo largo del tiempo. En este sentido, los datos incluidos en el registro podrían servir en un futuro como punto de partida para realizar estudios longitudinales que analicen la vida de las empresas en un periodo determinado de tiempo y para distintos niveles territoriales. En estos momentos, a partir del análisis que realizamos en el informe automatizado del registro, ya se pueden detectar acontecimientos sociales y económicos que han ocurrido a lo largo de la serie temporal que abarca y comprobar cómo han afectado estos a la información almacenada en el Registro Integrado de Empresas, proporcionando un mayor conocimiento sobre las fuentes administrativas incluidas en el mismo.

Principales referencias bibliográficas

- «BOE» núm. 68, de 20 de marzo de 1991, páginas 8892 a 8897 (6 págs.)
- «BOC» núm. 103, de 29 de mayo de 2018
- Christen, P. (2006) “A Comparison of Personal Name Matching: Techniques and Practical Issues”, Sixth IEEE International Conference on Data Mining, pp. 290-294.
- European business statistics methodological manual for statistical business registers: https://ec.europa.eu/eurostat/web/statistical-business-registers/methodology
- Ivan P. Fellegi & Alan B. Sunter (1969) A Theory for Record Linkage, Journal of the American Statistical Association, 64:328, 1183-1210, DOI: 10.1080/01621459.1969.10501049

- 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