TÍTULO: NORMALIZACIÓN, GEOCODIFICACIÓN Y ENLACE DE FICHEROS CON DIRECCIONES POSTALES CON LA HERRAMIENTA aLink Y SU APLICACIÓN AL DIRECTORIO DE EMPRESAS Y ESTABLECIMIENTOS CON ACTIVIDAD ECONÓMICA EN ANDALUCÍA

Autores:

Caballero Ruiz, Elisa Isabel (elisai.caballero@juntadeandalucia.es)

Galera Pozo, Ana Gema (gema.galera@juntadeandalucia.es)

Organismo: Instituto de Estadística y Cartografía de Andalucía (IECA)

Resumen: El aprovechamiento de fuentes, registros e infraestructuras de información, así como la normalización y garantía de la calidad; la difusión, el acceso y reutilización de la información son estrategias esenciales para el seguimiento de políticas europeas, nacionales y autonómicas, y contribuye a la toma de decisiones participativas por la sociedad andaluza y para el desarrollo de la sociedad del conocimiento.

Además, la integración entre la información estadística y cartográfica permite dar valor añadido a ambas, así como proporcionar mayor utilidad de los datos.

Para la representación espacial de la información estadística sigue siendo fundamental el uso de la dirección postal como base territorial para conseguir geocodificar dichos datos.

El objetivo de esta ponencia es presentar aLink, la herramienta de la que dispone el Instituto de Estadística y Cartografía de Andalucía (IECA) para conseguir la normalización y la geocodificación masiva de datos a través de la dirección postal, así como presentar la metodología que se sigue para la elaboración del Directorio de Empresas y Establecimientos con Actividad Económica en Andalucía aprovechando el uso de registros administrativos y aLink.

Palabras claves: Normalización, geocodificación, direcciones postales, aLink, Directorio.

1. INTRODUCCIÓN

El territorio está jugando un papel bastante importante en los últimos años dentro de las estadísticas oficiales del Sistema Estadístico y Cartográfico de Andalucía. Ya en el Plan Estadístico de Andalucía 2007-2012 se incorporó el eje transversal referido al territorio que introducía una nueva dimensión en la práctica estadística, integrando la territorialización de la información. Siguiendo esta línea, el Plan Estadístico y Cartográfico de Andalucía 2013-2020 integra por primera vez la información estadística y cartográfica, siendo en este sentido un plan innovador tanto en el ámbito español como europeo. En él se promueve el tratamiento conjunto de ambos tipos de información con el fin de seguir avanzando en la georreferenciación de las estadísticas aprovechando el potencial de la información territorial que aportan muchas de ellas y ofreciendo estadísticas con el máximo nivel de desagregación territorial. Esta sinergia entre ambos datos refuerza además el valor de la información tanto estadística como cartográfica, ya que la territorialización de la estadística ayuda a la interpretación de los datos y contribuye al desarrollo de la sociedad del conocimiento, necesidad esencial de esta nueva sociedad.

Por otro lado, este último Plan define el aprovechamiento de las fuentes, registros e infraestructuras de información, la normalización y garantía de la calidad y la difusión, el acceso y reutilización de la información como estrategias esenciales para la consecución de sus objetivos. En relación a estos registros y fuentes de información administrativa, no hay que perder de vista que las mismas se crean para fines de gestión, por lo que no siempre la información está recogida de manera normalizada o siguiendo criterios de buenas prácticas.

El Instituto de Estadística y Cartografía de Andalucía (IECA) dispone de herramientas para tratar que la información que pueda ser aprovechable de manera estadística y/o cartográfica, sea de mejor calidad para que finalmente sea mucho más fiable, comparable e integrada. Concretamente, resulta fundamental que la información relativa a la dirección postal esté lo mejor normalizada posible para después conseguir un éxito mejor en la geocodificación o cualquier otro proceso de enlace en el que se desee utilizar.

En este documento se va a presentar una de las herramientas desarrollada por el IECA para la normalización, enlace y geocodificación de la información que tiene asignada una dirección postal. La geocodificación de la dirección postal se lleva a cabo a través de enlaces probabilísticos de ficheros con direcciones postales, aprovechando a su vez, el potencial de la información alfanumérica que ofrece el Callejero Digital de Andalucía Unificado (CDAU).

Esta herramienta es ***aLink: Herramienta de Fusión de Ficheros*** que además de permitir enlazar ficheros con procesos probabilísticos a través de variables comunes, permite también normalizar direcciones postales, nombres y/o apellidos de personas físicas, e identificadores de personas físicas y/o jurídicas como NIF, DNI y NIE. Se trata de una aplicación libre y gratuita generada en Python a partir del proyecto **F**reely **E**xtensible **B**iomedical **R**ecord **L**inkage (FEBRL), desarrollo de software libre de la Universidad Nacional de Australia y si se quiere hacer uso de la misma se puede descargar a través del servicio de Descarga de Software del IECA. A continuación, se muestra una imagen de la

interfaz principal de la aplicación, así como del esquema de las distintas etapas en las que se fundamenta su metodología:

Imagen 1: Interfaz principal de aLink: Herramienta de Fusión de Ficheros.

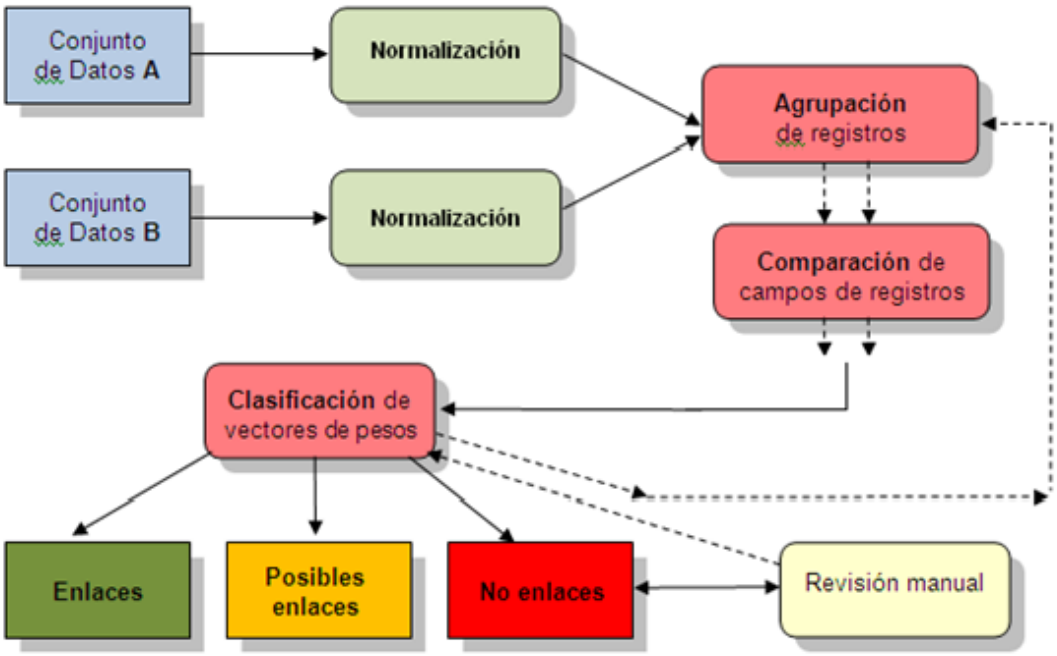

Imagen 2: Etapas del proceso de enlace de ficheros.

En el IECA dicha aplicación se está utilizando en distintos proyectos, entre ellos en la creación y geocodificación del Directorio Empresas y Establecimientos con Actividad Económica en Andalucía, a partir de la normalización y del enlace de fuentes administrativas.

2. HERRAMIENTA DE NORMALIZACIÓN aLink

Para la geocodificación de la información estadística y de los registros y/o fuentes administrativas sigue siendo la dirección postal la variable que más se utiliza como base territorial para dicho procedimiento. La geocodificación puntual de direcciones postales es el proceso de asignar coordenadas geográficas (X e Y) a dichas direcciones. Es por ello que cuanto mejor se tenga dicha información mayor éxito se obtendrá en dicho proceso.

Por ello, en la actualidad es imprescindible incidir en la mejora y eficacia en la recogida de información que pueda ser susceptible de utilización de manera estadística, y en concreto en la recogida de la información de la dirección postal que será la base de la información cartográfica. La normalización y el uso de buenas prácticas son fundamentales para la mejora de la calidad de los datos.

En este sentido, para normalizar la recogida de información en las fuentes de información administrativa, el IECA ha elaborado un **Manual de buenas prácticas para la normalización de fuentes y registros administrativos de la Junta de Andalucía.** Este manual describe para un conjunto de variables frecuentemente utilizadas en registros y en encuestas, recomendaciones para la recogida, codificación en los sistemas de información y difusión de las mismas. En concreto también se describen variables en materia de información geográfica. Tener bien recogidas las variables relacionadas con la dirección postal es condición necesaria para geocodificar fuentes administrativas.

Cuando la información postal ya viene recogida en fuentes o registros administrativos y no está lo suficientemente normalizada, existen herramientas que ayudan a estandarizar y segmentar en tantos campos como se quiera dicha dirección, consiguiendo así su normalización. Este paso es fundamental para enlazar registros con variables comunes. Esto permitirá añadir, complementar o actualizar información de un fichero a otro, de forma que cuanto mejor sea la normalización, mayor será el éxito del proceso de enlace porque las variables comunes tendrán mayor similitud.

En concreto, si se dispone de un fichero con información alfanumérica que tenga direcciones postales normalizadas, y de otro que también tenga direcciones postales normalizadas junto con sus coordenadas geográficas X e Y, la dirección postal puede usarse como variable de enlace para cruzar los dos ficheros y añadir al primero las coordenadas geográficas del segundo, de manera que el primer fichero quedaría geocodificado.

Para realizar un proceso de normalización con aLink se utiliza la Herramienta de Normalización de la aplicación (ver imagen 3).

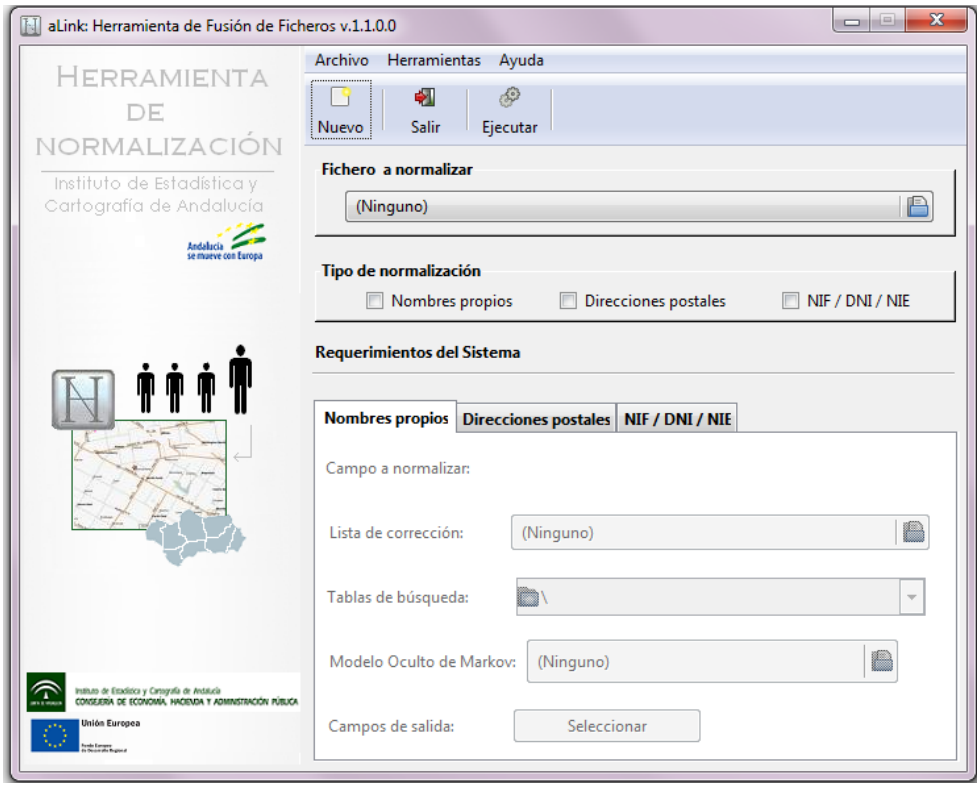

Imagen 3: Herramienta de Normalización de aLink.

Con esta herramienta se podrán transformar los datos originales brutos en otros con formato consistente y comparable: limpiando, estandarizando y segmentando los mismos. Además, con ella se tiene cubierta la primera etapa del proceso de enlace de ficheros: la normalización (ver imagen 2). Como ya se ha comentado anteriormente esta herramienta permite la normalización de variables que contienen nombres de personas, direcciones postales e identificadores de personas físicas y/o jurídicas. Así si lo que se pretende normalizar son direcciones postales visualmente se querría obtener lo siguiente:

	A	B	C	D	E	F	G	H	I	J	K	L	M
1	DIRECCIÓN SIN NORMALIZAR	DIRECIÓN NORMALIZADA											
2	"direccion"	tipo_de_via	nombre_de_via	identificador_de_numeracion	ein	bloque	portal	escalera	planta	puerta	tipo_de_agrupacion	agrupacion	odub
3	pg calonge c/ rodio numero 2	calle	rodio	numero	2						poligono_industrial	calonge	
4	pg amate c/ estornino 26	calle	estornino		26						poligono_industrial	amate	
5	po hytasa c/ lino numero 13	calle	lino	numero	13						poligono_industrial	hytasa	
6	pg hytasa. c/ algodon s/ n	calle	algodon	sin_numero							poligono_industrial	hytasa	
7	pg ind su eminencia c/ c s/ n	calle	c	sin_numero							poligono_industrial	su eminencia	
8	c/ luis montoto 122. planta baja edif corte ingles	calle	luis montoto		122				baja				edificio corte ingles
9	pg la negrilla c/ 4. nave 1	calle	4								poligono_industrial	la negrilla	
10	po la negrilla. c/ imprenta	calle	imprenta								poligono_industrial	la negrilla	
11	av san francisco javier numero 24 primera a edificio sevilla 1	avenida	san_francisco_javier	numero	24				primera	a			edificio sevilla
12	pg navisa c/ nana de espinar 8	calle	nana de espinar		8						poligono_industrial	navisa	
13	pg navisa. c/ los palos s/ n	calle	los palos	sin_numero							poligono_industrial	navisa	
14	pg navisa. c/ quejio. 2	calle	quejio		2						poligono_industrial	navisa	
15	cl luis de morales 32. tercera c edificio forum	calle	luis de morales		32				tercera	c			edificio forum
16	pg aerpto c/ trebedes nave 3	calle	trebedes								poligono_industrial	aeropuerto	nave 3
17	ur al alba-c/ cielo num 7	calle	cielo	numero	7						urbanizacion	al alba	
18	c/ aviacion numero 14 - pol calonge nave 8	calle	aviacion	numero	14						poligono_industrial	calonge	
19	pg calonge c/ metalurgia 8	calle	metalurgia		8						poligono_industrial	calonge	
20	pg calonge c/ automocion 5	calle	automocion		5						poligono_industrial	calonge	

Imagen 4: Fichero antes y después de la normalización.

El proceso de normalización de datos a través de la herramienta de normalización de aLink consta de las siguientes fases:

- **Tratamiento previo del fichero de datos.** Esta fase es obligatoria en cualquier proceso de normalización o enlace que se vaya a realizar con *aLink*, ya que la aplicación trabaja solamente con ficheros en formato .csv cuyos elementos estén separados por el carácter ';'. Dado que en la mayoría de los casos los ficheros de trabajo no se encuentran en dicho formato, *aLink* dispone de una herramienta que transforma el fichero a normalizar o enlazar en un fichero de texto con el formato adecuado. A continuación, se muestran los distintos formatos de partida que transforma *aLink*: csv, tabulador, texto plano, Excel, Access, Mysql, ods y dbf. No obstante, hay que indicar que actualmente se está trabajando en una nueva versión de *aLink* que permite convertir a csv tablas de PostgreSQL y Oracle. Además, con este tratamiento también se recodifican los datos del fichero de trabajo al sistema de codificación de caracteres UTF-8, por eso también es necesario realizar el tratamiento incluso cuando el fichero está en formato .csv. Por otro lado, los datos que están en mayúscula se transforman a minúscula y también se eliminan o sustituyen algunos símbolos o caracteres que por su codificación pueden provocar errores, por ejemplo, el carácter 'ñ' se sustituye por los caracteres 'kk' o se eliminan tildes. Por último, hay que indicar que este tratamiento previo se realiza a todas las variables del fichero a normalizar o enlazar.

- **Normalización del fichero de datos tratado**. En esta fase se normaliza alguna variable del fichero de datos tratado anteriormente que contenga direcciones postales, nombres de personas físicas o NIF, DNI o NIE, de manera que una vez esté normalizada la información aparecerá desagregada en tantos campos de salida como especifique el usuario. Por ejemplo, para la dirección postal es posible desagregar la información en tantos campos como los que componen el CDAU (tipo de vía, nombre de la vía, entidad inferior de numeración, calificador de la entidad inferior de numeración, tipo de agrupación, agrupación, otros datos de ubicación, etc.). Para realizar la normalización se utilizan las siguientes herramientas:

  - **Listas de corrección**: son ficheros que permiten limpiar el campo a normalizar del fichero de datos, es decir, contienen los caracteres que el usuario considera oportuno eliminar o sustituir en dicho campo. Por ejemplo, se eliminan caracteres del tipo: '|', '$','[',']', '(', ')', '.', etc. y se sustituyen términos del tipo 'drcha' por su valor estandarizado 'derecha'. Estos ficheros son editables por el usuario e intervienen en la fase de limpieza del campo a normalizar.

  - **Tablas de búsqueda**: son ficheros que contienen un listado de valores que permiten sustituir cada elemento del campo a normalizar por su valor estandarizado y, además, le asignan una etiqueta. Por ejemplo, si en el campo a normalizar aparece el elemento 'c/' se sustituye por 'calle' y se le asigna la etiqueta 'TV' que significa 'Tipo de Vía'. Al igual que en el caso anterior, estos ficheros son editables por el usuario e intervienen en la fase de estandarización y segmentación del campo a normalizar.

  - **Modelo Oculto de Markov** (en inglés Hidden Markov Models ó HMM): son ficheros que intervienen en la fase de segmentación de los datos del campo a normalizar tratando de reconocer el patrón o estructura que es más probable que sigan los mismos. Esa información va a servir posteriormente para segmentar el contenido del mismo en los distintos elementos que lo componen. Por ejemplo, en el caso de las direcciones postales tratan de analizar si el

primer elemento de la dirección es el tipo de vía, si a continuación aparece el nombre de la vía, si lo siguiente es el número de policía de la vía, etc., o si por el contrario la dirección postal tiene otra estructura distinta y todas comienzan por el nombre de la vía, luego el número de policía, etc. Para generar estos ficheros se utilizan técnicas de aprendizaje supervisado, por lo que se requiere seleccionar una muestra aleatoria con reposición de los datos del campo a normalizar y una vez conocida la estructura o patrón de los datos contenida en la misma, se extrapola dicho conocimiento a la totalidad de datos a normalizar.

Es necesario indicar que aLink pone a disposición del usuario un primer modelo oculto de Markov para cada uno de los campos que puede normalizar y en concreto para direcciones postales ofrece dos modelos HMM, uno para poder llevar a cabo una desagregación de la dirección postal de acuerdo a CDAU y otro para realizar una desagregación a medida. Ambos modelos permiten normalizar un número bastante elevado de direcciones pero si todavía quedaran algunas sin normalizar, que tuvieran una estructura similar, la aplicación ofrece la posibilidad de crear un nuevo modelo HMM para volver a normalizarlas, y así sucesivamente se continuaría hasta que normalizaran todas las direcciones.

- **Validación del proceso de normalización.** El fichero de salida resultante ofrece la información normalizada y estructurada de la manera que el usuario haya especificado. Además, muestra un campo de salida que se llama 'validación' que toma un valor 0 cuando la aplicación cree que ese registro ha sido bien normalizado y 1 cuando cree que algo ha fallado. Este resultado debe revisarse por el usuario para validar la normalización.

3. HERRAMIENTA DE ENLACE DE aLink

El proceso de enlace permite como su nombre indica enlazar dos ficheros de datos a partir de uno o varios campos que incluyen información común. Se comparan dos ficheros (fichero A y fichero B) para detectar aquellos registros que corresponden a una misma entidad o unidad poblacional (individuos, establecimientos, direcciones postales, etc.), incluso en aquellos casos en los que los ficheros no dispongan de identificadores únicos o se vean afectados por algún tipo de error. Para detectar aquellos registros que son comunes en ambos ficheros es necesario que tengan un campo o varios en común. Para comparar los campos en común y detectar los que son iguales se utilizan diversas medidas de similitud que permiten realizar comparaciones de forma exacta o aproximada. Con esta herramienta se cubren tres etapas del proceso de enlace de ficheros (ver imagen 2):

- **Agrupación**, se toman los registros del fichero A y del fichero B que pertenecen a un mismo grupo para realizar las comparaciones sólo dentro de cada agrupación. Por ejemplo, cuando se quieren comparar direcciones postales, sería fundamental realizar agrupaciones por municipios para que sólo se comparasen las direcciones dentro de ese mismo municipio
- **Comparación**, en esta etapa se realiza la comparación de cada campo o campos a confrontar. La herramienta dispone de distintas funciones según lo que se quiera comparar y

cómo se quiera comparar. Así por ejemplo, si se compara el nombre de la vía, es posible usar una función de comparación de cadena aproximada, de tal modo que si la cadena de caracteres a comparar está escrita de igual modo la función devolverá el valor máximo de similitud e irá disminuyendo ese valor en función de que la cadena de caracteres se parezca más o menos. De este modo, se podrán unir registros a través de campos que sean iguales pero que pueden estar escritos de modo diferente en cada uno de los ficheros, es lo que se llama enlaces probabilísticos. La posibilidad de realizar estos enlaces probabilísticos es lo que le da un valor añadido a la aplicación *aLink: Herramienta de Fusión de Ficheros.*

- **Clasificación**, en este paso el usuario podrá decidir a partir de los valores de similitud obtenidos en las comparaciones, aquellos pares de registros que son enlaces, no enlaces e incluso los posibles enlaces.

Cada una de estas etapas están definidas en la interfaz de la Herramienta de Enlace de *aLink*, siendo la imagen inicial de la misma la que se muestra a continuación:

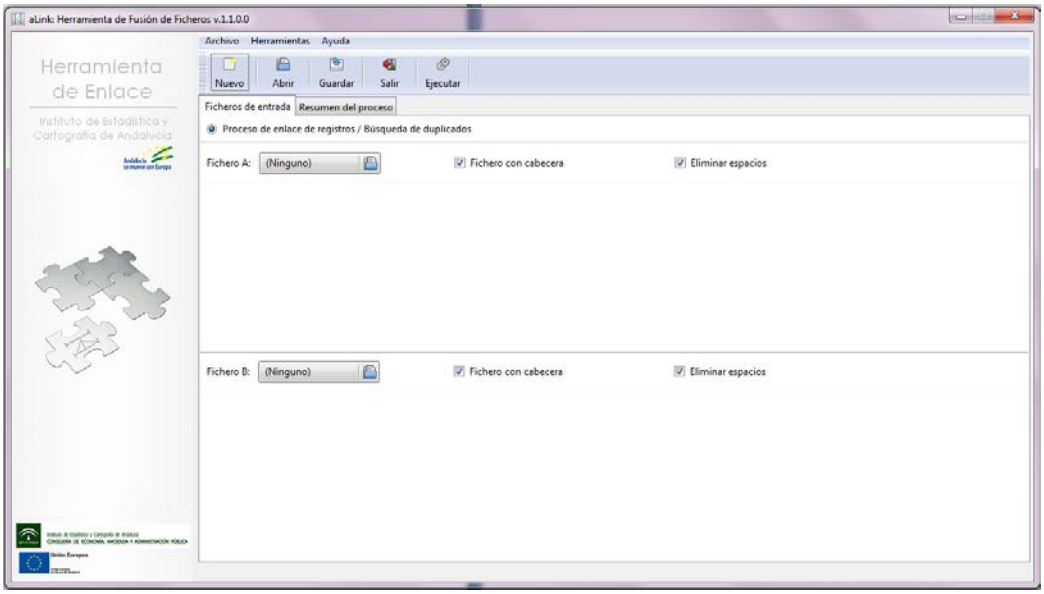

Imagen 5: Herramienta de Enlace de aLink.

Entre las funcionalidades de la herramienta de enlace cabe destacar:

a. La posibilidad de actualizar o completar la información de uno de los ficheros a enlazar con la información contenida en el otro.
b. La posibilidad de realizar un proceso de geocodificación de un fichero de datos cuando uno de los ficheros a enlazar contenga las coordenadas geográficas X e Y que permiten localizar una dirección en un mapa. Con ello se abre todo un abanico de posibilidades de tratamiento de esta información geocodificada.

Tanto el proceso de normalización como el de enlace permiten interactuar con *aLink* para crear los procesos más adecuados en cada momento y adaptarse a las necesidades de los usuarios y de los propios ficheros. Además, ambos procesos se mejoran de forma iterativa, es decir, después de la primera normalización o enlace, se puede lanzar un nuevo proceso y así sucesivamente hasta conseguir normalizar o enlazar la mayoría de los registros.

4. NORMALIZACIÓN Y ENLACE. APLICACIÓN AL DIRECTORIO DE EMPRESAS Y ESTABLECIMIENTOS CON ACTIVIDAD ECONÓMICA EN ANDALUCÍA

4.1 DIRECTORIO DE EMPRESAS Y ESTABLECIMIENTOS CON ACTIVIDAD ECONÓMICA EN ANDALUCÍA

El **Directorio de Empresas y Establecimientos con Actividad Económica en Andalucía**, en adelante Directorio, pone a disposición de la sociedad andaluza información relativa a las empresas y establecimientos que desarrollan su actividad económica en Andalucía a través de características fundamentales como son: el número, el tamaño, el sector al que pertenecen y la forma jurídica. Además se ofrece la ubicación física de los establecimientos.

Cada año es necesario actualizarlo de tal modo que se tenga un Directorio con el número de establecimientos que están de alta a 1 de enero de dicho año. Para su actualización se han usado principalmente dos fuentes de información administrativas: el **Impuesto de Actividades Económicas (IAE)** y las **Cuentas de Cotización a la Seguridad Social (CCSS).**

La creación de este nuevo Directorio de Empresas y Establecimientos viene dado por la comparación del Directorio del año anterior con la información de las nuevas fuentes administrativas del año a actualizar. Así, a partir del Directorio anterior se va obteniendo el universo de establecimientos que siguen estando en alta un año después, aquellos que han sufrido alguna modificación, aquellos que han cerrado (están en baja) y aquellos otros que entran como nuevas altas o reactivaciones con respecto al año anterior (no aparecían o aparecían en baja en el Directorio de un año antes).

Para tratar de entender mejor la metodología que se sigue con la actualización del Directorio es necesario tener claro conceptos fundamentales:

- Un establecimiento es una unidad productora de bienes o/y servicios que desarrolla una o más actividades de carácter económico o social, bajo la responsabilidad de un titular o empresa, en un local situado en un emplazamiento fijo y permanente. Cuando la actividad no se ejerce en un local fijo se considera como establecimiento el lugar desde donde se organiza la citada actividad. Por tanto, un mismo establecimiento puede ejercer varias actividades económicas. Sólo y cuando un establecimiento tenga en baja todas sus actividades, se considera que éste está de baja o ha cerrado.
- Empresa: Toda organización definida jurídicamente, con contabilidad independiente, sometida a una autoridad rectora que puede ser, según los casos, una persona jurídica, o una persona física y constituida con miras a ejercer en uno o varios lugares, una o varias actividades de producción de bienes o prestación de servicios. Una empresa puede tener varios establecimientos. Todos los establecimientos de una misma empresa tiene en común el mismo identificador jurídico.

Por tanto, es necesario señalar que dos establecimientos de una misma empresa tendrán en común el mismo identificador jurídico (NIF o DNI), y sólo se distinguirán en la dirección postal del

establecimiento, ya que si se encuentran situados en distintos lugares, territorialmente hablando, entonces se puede afirmar que se tratan de dos establecimientos diferentes.

Concretamente, la dirección postal es una característica indispensable, junto con el identificador de la empresa, para buscar los establecimientos diferentes con actividad económica en Andalucía.

Por lo cual, para encontrar los establecimientos con actividad económica en alta en Andalucía, hay que partir de Directorio del año anterior y cruzarlo con las fuentes administrativas IAE y CCSS que ofrecen información sobre las actividades económicas en altas y bajas en el último año y también ofrecen información sobre las cuentas de cotización y el empleo.

En los cruces, para identificar a un mismo establecimiento es indispensable que tenga la misma dirección. Por este motivo es necesario que estos se hagan siempre considerando la dirección postal del establecimiento y su NIF. De esta forma se van a poder detectar aquellos establecimientos que permanecen igual o que han podido cambiar de estado o de actividad económica principal.

A continuación se muestran las variables utilizadas en los cruces del diseño de registro de Directorio de base (tabla 1) que se usa para cruzar y obtener el del año siguiente, y los diseños de registros de las dos fuentes administrativas que se usan para la actualización (tablas 2 y 3):

VARIABLE	TIPO	DESCRIPCIÓN
ID	N	IDENTIFICADOR IECA DE CADA ESTABLECIMIENTO DEL DIRECTORIO
ESTADO	A1	SITUACIÓN DEL ESTABLECIMIENTO: A=ACTIVO; B=BAJA
FECHA_CESE	A8	FECHA DE CESE DEL ESTABLECIMIENTO
NIF	A9	NIF DEL ESTABLECIMIENTO
RAZON_SOCIAL	A150	RAZÓN SOCIAL DE LA EMPRESA
DENOESTA	A150	DENOMINACIÓN DEL ESTABLECIMIENTO
TVIA	A5	TIPO DE VÍA A 5 POSICIONES
TIPO_VIA	A2	TIPO DE VÍA A 2 POSICIONES
NOMBRE_VIA	A75	NOMBRE DE LA VÍA
EIN	A4	NÚMERO DE LA VÍA O TAMBIÉN ENTIDAD INFERIOR DE NUMERACIÓN
CEIN	A1	CALIFICADOR DE LA ENTIDAD INFERIOR DE NUMERACIÓN
ESN	A4	ENTIDAD SUPERIOR DE NUMERACIÓN
CESN	A1	CALIFICADOR DE LA ENTIDAD SUPERIOR DE NUMERACIÓN
KM	A10	KILOMETRO DE LA VIA
ODUB	A100	OTROS DATOS DE UBICACIÓN
BLOQUE	A4	BLOQUE
PORTAL	A5	PORTAL
ESCALERA	A2	ESCALERA
PLANTA	A3	PLANTA
PUERTA	A3	PUERTA
CODPROV	A2	CÓDIGO INE DE LA PROVINCIA DONDE SE UBICA EL ESTABLECIMIENTO
CODMUN	A5	CÓDIGO INE DEL MUNICIPIO DONDE SE UBICA EL ESTABLECIMIENTO
CODPOS	A5	CÓDIGO POSTAL DONDE SE UBICA EL ESTABLECIMIENTO
ENTIDAD	A100	NÚCLEO DE POBLACIÓN O DISEMINADO DONDE SE UBICA EL ESTABLECIMIENTO
CNAE2009	A4	CÓDIGO DE CNAE 2009 PRINCIPAL
CNAE2009_S	A4	CÓDIGO DE CNAE 2009 SECUNDARIA
NUMTRA	N	NÚMERO DE TRABAJADORES
EMPLEO	N	EMPLEO ASOCIADO AL ESTABLECIMIENTO
ASALARIADO	N	NÚMERO DE ASALARIADOS ASOCIADOS AL ESTABLECIMIENTO
NO_ASALARIADO	N	NÚMERO DE NO ASALARIADOS ASOCIADOS AL ESTABLECIMIENTO
FECHA_INICIO	A10	FECHA DE INICIO DE LA ACTIVIDAD
ORIGEN	A2	FUENTE DE ORIGEN DEL ESTABLECIMIENTO
CCC1	A15	CÓDIGO DE CUENTA DE COTIZACIÓN

VARIABLE	TIPO	DESCRIPCIÓN
CCC2	A15	CÓDIGO DE CUENTA DE COTIZACIÓN

Tabla 1: Diseño de registro del Directorio

VARIABLE	DESCRIPCIÓN
ID_NUM	Identificador de los registros de IAE 2016 incluido por Sv. Económicas. ¡OJO! no coincide con el de 2015
NIF	Número de identificador fiscal
RAZ_SOC	Razón social
ESTADO	Estado del registro (A:Alta, B:Baja)
ANNO_I	Año de inicio
MES_I	Mes de inicio
DIA_I	Día de inicio
FECHA_I	Fecha de inicio (dd/mm/aaaa)
ANNO_C	Año de cese
MES_C	Mes de cese
DIA_C	Día de cese
FECHA_C	Fecha de cese (dd/mm/aaaa)
TIPO2	Tipo de vía del domicilio tributario (establecimiento)
NOMBRE2	Nombre de vía del domicilio tributario (establecimiento)
NUM2	Número de portal del domicilio tributario (establecimiento)
ESCALERA2	Escalera del domicilio tributario (establecimiento)
PISO2	Piso del domicilio tributario (establecimiento)
PUERTA2	Puerta del domicilio tributario (establecimiento)
PUNTOKM	Punto kilométrico del domicilio tributario (establecimiento)
CODMUN2	Código del municipio del domicilio tributario (establecimiento) Necesita depuración
CODMUN2R	Código intermedio del municipio del domicilio tributario (establecimiento) Necesita depuración
CODMUN_EST	Código depurado del municipio del domicilio tributario (establecimiento)
MUNIC2	Nombre del municipio domicilio tributario (establecimiento). Necesita depuración
MUNIC2R	Nombre depurado del municipio domicilio tributario (establecimiento)
CODPOS2	Código postal del domicilio tributario (establecimiento). Necesita depuración
CODPOS2R	Código postal depurado del domicilio tributario (establecimiento)
CNAE2009	Código de la CNAE-2009 obtenido a partir de la variable EPIGRA (epígrafe fiscal)

Tabla 2: Diseño de registro del IAE

VARIABLE	DESCRIPCIÓN
REG1	Código de registro de la cuenta de cotización
TES1	Código de provincia de la cuenta de cotización
NUM1	Código de la cuenta de cotización
REG2	Código de registro de la cuenta de cotización principal de la empresa
TES2	Código de la provincia de la cuenta de cotización principal de la empresa
NUM2	Código de la cuenta de cotización principal de la empresa
NIF	Código de identificación fiscal
RAZON	Razón Social
ANAGRAMA	Anagrama de la empresa
TIPOV	Tipo de vial
NOMBREV	Nombre de vial
NUM	Número de vial
BISMUN	Calificador del número de vial
BLOQ	Bloque
ESC	Escalera
PISO	Planta
PUERTA	Puerta
CODPOS	Código postal
CODINE	Código INE de provincia-municipio
CODLOCAL	Código de localidad
LOCALIDAD	Literal de localidad
ACTIV	Código de actividad según CNAE2009
NUMTRA	Número de trabajadores de la cuenta de cotización
ANNO_A	Año de inicio de la cuenta de cotización
MES_A	Mes de inicio de la cuenta de cotización
DIA_A	Día de inicio de la cuenta de cotización
TRELAC	Tipo de relación
ENTIDAD	Código de entidad
EMPCOL	Empresa colaboradora
SITU	Situación de la cuenta de cotización

VARIABLE	DESCRIPCIÓN
ANNO_S	Año de situación de la cuenta de cotización
MES_S	Mes de situación de la cuenta de cotización
DIA_S	Día de situación de la cuenta de cotización
CCC1	REG1 + TES1 + NUM1
CCC2	REG2 + TES2 + NUM2

Tabla 3: Diseño de registro del CCSS

Como ya se viene diciendo durante todo este apartado, la dirección postal va a jugar un papel muy importante en los enlaces de las tres fuentes, por ello es fundamental tratar de tener del modo más adecuado dicha dirección y bien normalizada, y es preciso que esa normalización sea lo más parecida posible en las tres fuentes para así garantizar un mejor éxito en la fase de enlaces que se realizará a posteriori. Por ello, antes de comenzar los trabajos de cruces entre las tres fuentes, se realiza un paso previo que consiste en la normalización de la dirección postal con la que se va a trabajar en cada una de las fuentes (Directorio del año n-1, IAE año n y CCSS año n).

4.2 NORMALIZACIÓN DE LA DIRECCIÓN POSTAL DEL DIRECTORIO Y DE LAS FUENTES ADMINISTRATIVAS

Cada uno de los establecimientos incluidos en el Directorio de Empresas y Establecimientos con Actividad Económica en Andalucía están identificados a través de una variable identificadora denominada '*ID*', su NIF y su dirección postal, de forma que si un establecimiento del Directorio que está en proceso de actualización cambia de dirección postal se considera un establecimiento nuevo. Igualmente para las altas nuevas de establecimientos que se proponen para entrar en cada Directorio resulta necesario comprobar si ya existe un establecimiento con ese mismo NIF y en esa misma dirección postal. Por tanto, la dirección postal juega un papel crucial a la hora de configurar el universo de establecimientos que componen del Directorio de cada año.

En este sentido, la normalización de las direcciones postales contenidas tanto en el Directorio como en las dos fuentes administrativas principales utilizadas para su actualización, IAE y CCSS, resulta fundamental para evitar que se produzcan entradas de establecimientos que ya están en el Directorio, e incluso para detectar posibles duplicados que ya estuvieran incluidos en el mismo de años anteriores.

Para la normalización de las direcciones postales de las fuentes administrativas que permiten la elaboración del Directorio (IAE y CCSS), así como las del propio Directorio, se ha usado tanto la *Herramienta de Normalización* como la *Herramienta de Enlace* de ***aLink***, así como el sistema gestor de base de datos ***PostgreSQL***. Por otro lado, las fuentes de referencia utilizadas para obtener los códigos de las vías han sido:

- El portalero y vialero del Callejero Digital de Andalucía Unificado del IECA.
- El tramero del Callejero del Censo Electoral del INE.

El objetivo de la normalización es asignar a los literales de las direcciones postales de Directorio, de IAE o de CCSS un código de vía, ya sea el definido en el Callejero Digital de Andalucía Unificado (CDAU) o el asignado por el Instituto Nacional de Estadística (INE) a los viales. Disponer de los códigos de vía asociados a los literales de las direcciones postales permite trabajar de forma más eficiente y precisa tanto en los procesos de elaboración del Directorio como en la geocodificación de las direcciones postales del mismo.

Con la *Herramienta de Enlace de aLink* se obtienen directamente dichos códigos de vía simplemente comparando, dentro del mismo municipio, con funciones de comparación aproximadas (como la distancia de edición o la función de Levenshtein), una serie de campos como son: el código INE del municipio, el código postal, el tipo de vía y el nombre de la vía. Además, es necesario indicar, que se realizan procesos de enlace sin tener en cuenta el código postal dado que se detectó que algunos de ellos están desactualizados o son erróneos.

Por otro lado, la *Herramienta de Normalización* de *aLink* se utiliza para segmentar, en un primer paso, los literales de las direcciones postales que no habían enlazado en las primeras fases de enlace. La segmentación se realiza en una serie de campos de salida o componentes que definen una dirección postal. La partición se asemeja a la estructura de las direcciones en CDAU, las cuales siguen la normativa INSPIRE. En un segundo paso, cuando ya se dispone de las direcciones normalizadas, se realizan enlaces haciendo uso de las comparaciones aproximadas con las distintas fuentes de referencia que tienen los códigos de vías utilizando la herramienta de enlace de *aLink*, del mismo modo que se ha contado en el párrafo anterior.

Además, mediante programación en código sql se realizan comparaciones de las direcciones postales de Directorio, IAE o CCSS con las del portalero y vialero de CDAU y el tramero del INE, de manera que ambas coinciden de forma exacta.

Dado que las direcciones postales que participan en la elaboración del Directorio son bastante parecidas de un año a otro, para reutilizar los trabajos de normalización ya realizados se ha generado un repositorio de direcciones postales distintas tal y como aparecen en el Directorio y en las fuentes administrativas que participan en su elaboración. A continuación, se muestra una imagen del mismo:

Imagen 6: Repositorio de direcciones postales normalizadas

Como se puede ver en el repositorio no solo se han incluido los códigos de vía sino que se intenta completar la información con los códigos INE del núcleo de población y el código postal correcto en caso de que esté desactualizado o sea erróneo.

Así al disponer de este repositorio se facilitarán las tareas de normalización de los próximos Directorios, ya que solo tendremos que comparar las direcciones postales tal y como aparecen de inicio en las fuentes y en Directorio con las direcciones postales de este repositorio y en el caso de que coincidan asignar los códigos de vía a los literales de las direcciones postales.

Esta tarea de normalización que parece relativamente sencilla no lo es tanto debido a la variabilidad de direcciones postales del Directorio y de las otras dos fuentes de información. De hecho en algunos casos se requiere de una revisión manual para poder asignar el código de vía a las direcciones postales. Este es el caso de direcciones postales genéricas del tipo: urbanizaciones, barrios o barriadas, conjuntos residenciales, polígonos industriales, centros comerciales, mercados de abastos, etc.

El tratamiento dado a estos tipos de direcciones es diferente dependiendo de si el establecimiento puede ser localizado en una ubicación exacta y única, como es el caso de los centros comerciales y los mercados de abastos en los que aparece el nombre de los mismos o si el establecimiento se encuentra dentro de una agrupación de viales y no sabemos exactamente cuál podría ser su dirección postal exacta, este sería el caso de los polígonos industriales, barrios o barriadas y urbanizaciones o centros comerciales y mercados o plazas de abastos para los que no se dispone del nombre de los mismos y pueden existir varios de ellos en un mismo municipio.

Así pues, si la dirección postal de la que se dispone es Mercado de Abastos de Triana (Sevilla) ó Centro Comercial Airesur (Castilleja de la Cuesta), aunque va a ser imposible localizar estas direcciones postales entre los viales del CDAU o en el tramero del Callejero del Censo Electoral del INE, sí que se puede realizar una búsqueda manual de estas localizaciones a través de Internet para conseguir su dirección postal adecuada, por ejemplo:

Dirección postal en Directorio, IAE o CCSS	Dirección postal localizada en Internet
Mercado de Abastos de Triana	Plaza del Altozano
Centro Comercial Airesur	Carretera SE-3403

Tabla 4: Direcciones postales genéricas

Y con estas direcciones ya sí que se puede realizar un proceso de normalización y/o enlace con ***aLink*** para asignar el código del vial y el resultado obtenido se puede incluir en el repositorio de direcciones postales distintas normalizadas.

Diferente sería si la dirección postal de la que se dispone es aún más genérica, como por ejemplo, Polígono Industrial Juncaril (Peligros-Granada), Barrio Santa Cruz (Sevilla), o Mercado de Abastos (Sevilla). En estos casos no se podría saber qué vial corresponde realmente a estas direcciones postales, ya que puede existir más de un vial en este tipo de zonificaciones. Por tanto, en estas

situaciones se requiere buscar manualmente por Internet el establecimiento en sí para intentar localizarlo en el vial correcto. Por ejemplo, un establecimiento con dirección 'Polígono Industrial Juncaril' podría estar realmente ubicado en el vial Calle Lanjarón y a posteriori se le asignaría con *aLink* o *PostgreSQL* el código de vía correspondiente al literal de este vial.

Hay que tener en cuenta que estos casos en los que la dirección es tan genérica no se van a incluir en el repositorio de direcciones postales normalizadas puesto que se estaría dando por hecho que todas las direcciones del tipo 'Polígono Industrial Juncaril' tienen asignado el código de vía de la Calle Lanjarón, cuando en realidad otros establecimientos podrían estar ubicados en otros viales del polígono distintos a éste. Igual ocurriría con los barrios o urbanizaciones.

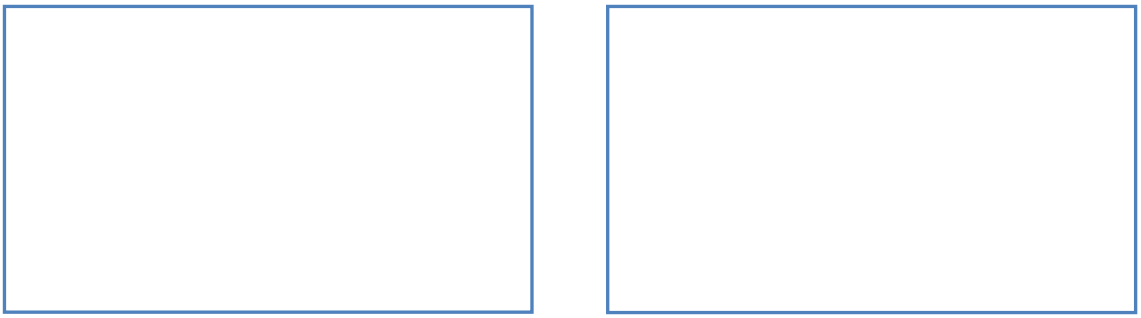

Imagen 7: Viales del Polígono Industrial Juncaril y del Barrio Santa Cruz

En el caso de mercados de abastos o centros comerciales puede darse el caso de que existan varios en el mismo municipio. En esta situación, igual que en los casos anteriores, se ha tenido que buscar manualmente por Internet el establecimiento en sí, así como utilizar información auxiliar como por ejemplo el código postal u otras direcciones existentes en los ficheros referidas al mismo establecimiento que dispongan de más información, ya que en algunos casos esta información ha permitido ubicar el establecimiento en una zona más concreta y diferenciar el centro comercial mercado de abastos del que se trata.

Data Output | Explain | Messages | History

	id_dom bigint	tipo2 charact	nombre2 character varying(80)	codmun_est character(5)	codpos2r character(5)	munic2r character varyi	tviac character varyi	nviac character varying(100)	tipo_de_edificio character varying(100)	edificio character varyi
113	297600	me	de abastos	04101	04240	viator	calle	gregorio torres	mercado de abastos	viator
114	337769	pz	del mercado abastos	04101	04240	viator	calle	gregorio torres	mercado de abastos	viator
115	339504	pz	mercado de abastos	04101	04240	viator	calle	gregorio torres	plaza de abastos	viator
116	335587	pz	abastos	04102	04738	vicar			mercado de abastos	
117	335685	pz	abastos-la gangosa-barrac	04102	04738	vicar	calle	jaspe	mercado de abastos	la gangosa
118	297601	me	de abastos ¿Gangosa o Puebla Vicar?	04102	04738	vicar			mercado de abastos	
119	297241	mc	de abastos de la gangosa	04102	04738	vicar	calle	jaspe	mercado de abastos	la gangosa
120	297656	me	de abastos la gangosa	04102	04738	vicar	calle	jaspe	mercado de abastos	la gangosa
121	144550	ca	mercado abastos	04102	04738	vicar			mercado de abastos	
122	144574	ca	mercado de abastos	04102	04738	vicar			mercado de abastos	
123	339505	pz	mercado de abastos	04102	04738	vicar			mercado de abastos	
124	234530	cl	mercado de abastos de vic	04102	04738	vicar	calle	calderon de la barca	mercado de abastos	puebla de vic
125	321304	pl	mercado de abastos la gan	04102	04738	vicar	calle	jaspe	mercado de abastos	la gangosa
126	322147	pl	plaza abastos gangosa	04102	04738	vicar	calle	jaspe	mercado de abastos	la gangosa
127	313883	pl	abastos el ejido pto.11	04902	04700	ejido (el)	plaza	chica	mercado de abastos	el ejido
128	335704	pz	abastos las norias	04902	04700	ejido (el)	carretera	mojonera	mercado de abastos	las norias
129	335717	pz	abastos sto domingo	04902	04700	ejido (el)	avenida	oasis	mercado de abastos	santo domingo

Imagen 8: Ejemplos direcciones postales de mercados de abastos.

4.3 ENLACE DE FUENTES PARA LA CREACIÓN DEL DIRECTORIO

Finalizados los procesos de normalización de la dirección postal de las fuentes administrativas y del Directorio año n-1, se está en disposición de comenzar la fase de enlace del mismo con las fuentes administrativas, IAE y CCSS, que aportan la información actualizada para el año n.

La dirección postal (normalizada o no) y el NIF van a determinar los registros que pertenecen a un mismo establecimiento. Por este motivo van a ser las variables utilizadas en los procesos de enlace. Así si un establecimiento ha cambiado dicha dirección con respecto al año anterior, entonces se dará de baja con la información que tenía en Directorio n-1 y entrará como nuevo establecimiento (alta nueva) con la nueva dirección.

El objetivo de los enlaces son los siguientes:

a) Detectar aquellos establecimientos que se encuentran en Directorio del año n-1 y también en las fuentes administrativas del año n para así actualizar su información, ya que han podido sufrir algún cambio: modificación del estado (de alta a baja o de baja a alta), cambio de NIF o cambio de la actividad económica principal que venía ejecutando. También puede darse el caso de que el establecimiento no haya sufrido ningún cambio, por lo que toda la información se mantendrá igual que el año anterior.
b) Detectar aquellos establecimientos que estando en Directorio del año n-1, no figuran para el año n en ninguna de las fuentes administrativas, ni en IAE ni en CCSS. Por tanto, se entiende que estos establecimientos que no figuran en el año a actualizar han debido de darse de baja por no aparecer en dichas fuentes administrativas. Estos establecimientos se darán de baja por omisión.
c) Detectar aquellos establecimientos que no estaban en Directorio del año n-1 pero, que sin embargo, figuran de “alta” en su estado en alguna o en ambas fuentes administrativas del año n. Estos establecimientos entran como altas nuevas, con respecto al año anterior, al aparecer en las fuentes administrativas del año a actualizar.

4.3.1 Fase de Enlaces

En este proceso se realizan varios enlaces para ir obteniendo los registros de cada fuente que pertenecen a un mismo establecimiento de los que figuran en el Directorio del año anterior. Los enlaces se realizan siempre buscando la misma dirección y el mismo NIF, pero en cada cruce se utilizan distintas variables que hacen referencia a la dirección postal. En la fase de normalización se consiguió añadir el ***código INE de vía*** o el ***código de vía de CDAU (ID_VIAL)*** a muchas de las direcciones de Directorio del año anterior, de IAE y de CCSS por lo que dicho código se ha utilizado en los cruces. En cuanto al NIF se utilizan los 8 últimos dígitos por si la empresa ha cambiado en el último año de forma jurídica.

Los cruces que se realizan se han llevado a cabo con el **gestor de base de datos *PostgreSQL*** cuando las comparaciones de las variables son exactas, y con ***aLink*** cuando las comparaciones son aproximadas para aprovechar el potencial de la aplicación al permitir utilizar enlaces probabilísticos

para encontrar direcciones escritas de modo diferente pero que se refieren a la misma. Exactamente las variables utilizadas en los cruces realizados son:

- Los 8 últimos dígitos del NIF, ID_VIAL (que es el código de vía de CDAU), CODMUN (que es el código INE del municipio) de manera exacta con el gestor de base de datos ***PostgreSQL***.
- Los 8 últimos dígitos del NIF, INE_VIA (que es el código INE de vía), CODMUN (que es el código INE del municipio) de manera exacta con el gestor de base de datos ***PostgreSQL.***
- Los 8 últimos dígitos del NIF, TIPO_VIA, NOMBRE_VIA, CODMUN (que es el código INE del municipio) de manera exacta con el gestor de base de datos ***PostgreSQL***
- Los 8 últimos dígitos del NIF, TIPO_VIA, NOMBRE_VIA, CODMUN (que es el código INE del municipio) comparando de manera aproximada el nombre de la vía con la herramienta ***aLink***.

Y los procesos de enlace que se llevaron a cabo son:

- **Primer proceso** que compara el fichero de Directorio de establecimientos del año n-1 con el fichero de IAE del año n. En este proceso se detectan aquellos establecimientos que han cambiado alguna variable respecto al Directorio del año anterior, y aquellos establecimientos que permanecen invariables en la información que se ha comparado.
- **Segundo proceso** que compara el fichero de Directorio de establecimientos del año n-1 con el fichero de CCSS del año n. En este proceso se detectan aquellos establecimientos que han cambiado alguna variable respecto al Directorio del año anterior, y aquellos establecimientos que permanecen invariables en la información que se ha comparado.

Al finalizar estos dos procesos se unen todos los enlaces de Directorio del año n-1 con ambas fuentes, obteniéndose por un lado el universo de registros tanto de IAE como de CCSS que pertenecen a cada establecimiento del Directorio del año anterior, y por otro el universo de registros de Directorio del año anterior que no cruzaron con ninguna de las fuentes y que por tanto se darán de baja ‘por omisión’ por no aparecer en las mismas.

Por otro lado, existen registros de las fuentes administrativas del año n que no han cruzado con el Directorio del año n-1 pero que están en alta en dichas fuentes. Estos deben entonces entrar a formar parte del Directorio de actualización como nuevas altas. No obstante, antes de incluirlas hay que realizar un tratamiento a las mismas para evitar entradas de duplicados en el Directorio y por este motivo se realiza el siguiente proceso de enlace:

- **Tercer proceso** que compara los registros de IAE del año n que no cruzaron en el primer proceso con Directorio, con los registros de CCSS del año n que tampoco cruzaron en el segundo proceso con Directorio. Para ello se utilizan las mismas variables de comparación que en el primer y segundo proceso, pero realizando el cruce entre IAE y CCSS. En este proceso se consiguen incorporar al Directorio del año n establecimientos considerados como altas nuevas y que tienen información en ambas fuentes administrativas por haber cruzado entre sí.

Finalizados estos trabajos, la fase de cruces con *aLink* y *PostgreSQL* ha terminado. Sin embargo, el tratamiento para la inclusión de altas nuevas no ha finalizado. A continuación se tienen que seguir

realizando trabajos para detectar registros asociados a un mismo establecimiento que se encuentra en alta, bien sea en IAE o bien sea en CCSS pero que no estaban en el Directorio anterior ni han cruzado entre estas fuentes. Como en el tercer proceso, con esta tarea se pretende evitar que se incluyan duplicados de altas nuevas en el Directorio a actualizar.

4.3.2 Incorporación de altas nuevas sólo en IAE o sólo en CCSS

En este apartado se indican los dos procesos a realizar para incorporar altas nuevas que proceden sólo de IAE o sólo de CCSS:

- **Primer proceso**. En este paso hay que incorporar aquellos establecimientos en alta en IAE que no cruzaron en los procesos anteriores, y que por tanto no se encontraron en Directorio del año anterior, y tampoco entre las nuevas altas de CCSS. En este proceso se agruparon todos los registros pertenecientes a un mismo establecimiento puesto que en algunos casos para un mismo establecimiento aparecía la misma actividad económica en alta y baja, o varias actividades distintas en alta y/o baja. Para realizar las agrupaciones se buscaron coincidencias de las siguientes variables:

    - NIF, CODMUN_EST (código INE del municipio del establecimiento), ID_VIAL
    - NIF, CODMUN_EST (código INE del municipio del establecimiento), NOMBRE DE LA VIA DEL ESTABLECIMIENTO

Tras las agrupaciones anteriores, se obtienen el universo de registros de IAE que pertenecen a un mismo establecimiento.

- **Segundo proceso**, que incorpora establecimientos de CCSS en alta que no cruzaron con Directorio del año anterior, que tampoco cruzaron con IAE y que tiene trabajadores asociados a la cuenta de cotización. El objetivo de este segundo proceso es el mismo que el anterior proceso, sólo que en este caso se trabaja con la información de CCSS. De igual forma tras las agrupaciones anteriores, se obtienen el universo de registros de CCSS que pertenecen a un mismo establecimiento y que tiene al menos un trabajador asociado a la cuenta de cotización.

4.3.3 Selección del estado final y actividad principal y secundaria del establecimiento

Tras realizar la fase de enlaces se consigue el universo de registros que pertenecen a un mismo establecimiento y a partir de dichos registros se va a obtener el estado final del establecimiento. Además, en el caso de que esté en alta y ejerza más de una actividad económica, entonces se debe elegir cuál es la actividad económica principal y cuál es la secundaria.

En términos generales un establecimiento se considera que está en baja si todas las actividades económicas asociadas al mismo se encuentran en baja a fecha de actualización, sin embargo, si una sola actividad económica está en alta, entonces el establecimiento se encuentra en alta.

En la teoría estos conceptos son fáciles de entender, pero cuando se trabaja con el grupo de registros que hay en cada establecimiento, es necesario tener en cuenta varias cuestiones a la hora de llegar a la conclusión final. Por ejemplo, en el conjunto de los enlaces realizados entre Directorio y las fuentes administrativas, una misma actividad asociada a un mismo establecimiento puede figurar varias veces, en alta y baja. Para saber cómo se encuentra a fecha de actualización será necesario ordenar dichos registros por fecha de estado para ir viendo si finalmente la actividad está en alta o baja.

A continuación se muestran dos ejemplos:

	id integer	nif_fuente character va	raz_soc_dir character varying(150)	codmun characte	codpos charac	tvia_ char	nvia_fuente character varyin	numero integer	localidad characte	estado_dir character(1)	estado_fuente character(1)	fecha_estado_fuente character varying(10)	cnae09_dir character v	cnae09_fuente character varyi	
1	225383	B04214417	CALZADOS SCARPI ALMERIA SL	04013	04004	CA	ALCALDE MUÑOZ	32	ALMERIA	A	A	2016-11-11	4772	4772	*
2	225383	B04214417	CALZADOS SCARPI ALMERIA SL	04013	04004	CA	ALCALDE MUÑOZ	32	ALMERIA	A	B	2016-10-11	4772	4772	
3	225383	B04214417	CALZADOS SCARPI ALMERIA SL	04013	04004	CL	ALCALDE MUÑOZ	32	ALMERIA	A	A	1995-11-21	4772	4772	

Imagen 9: Establecimiento con una única actividad económica asociada y diferentes estados de alta y baja

	id integer	nif_fuente character va	raz_soc_dir character varying(150)	codmun characte	codpos_ charact	tvia_ char	nvia_fu charact	numero integer	localidad_fuente character varying(15	estado_dir character(1)	estado_fuente character(1)	fecha_estado_fuente character varying(10)	cnae09_dir character v	cnae09_fuente character varyi	
1	162804	51831241M	REQUENA SEGADO ANTONIO	23063	23240	CA	REAL	83	NAVAS DE SAN JUAN	A	B	2015-12-31	4321	4321	*
2	162804	51831241M	REQUENA SEGADO ANTONIO	23063	23240	CA	REAL	83	NAVAS DE SAN JUAN	A	B	2015-12-31	4321	4322	*
3	162804	51831241M	REQUENA SEGADO ANTONIO	23063	23240	CA	REAL	83	NAVAS DE SAN JUAN	A	B	2015-12-31	4321	4322	
4	162804	51831241M	REQUENA SEGADO ANTONIO	23063	23240	CA	REAL	76	NAVAS DE SAN JUAN	A	B	2015-12-31	4321	4729	*
5	162804	51831241M	REQUENA SEGADO ANTONIO	23063	23240	CA	REAL	83	NAVAS DE SAN JUAN	A	B	2015-12-31	4321	4759	*
6	162804	51831241M	REQUENA SEGADO ANTONIO	23063	23240	CA	REAL	76	NAVAS DE SAN JUAN	A	B	2015-12-31	4321	4759	
7	162804	51831241M	REQUENA SEGADO ANTONIO	23063	23240	CA	REAL	83	NAVAS DE SAN JUAN	A	A	2016-01-01	4321	6820	*

Imagen 10: Establecimiento con diferentes actividades económicas asociadas en distintos estados de alta y baja

En ambos casos, el proceso de actualización del Directorio se quedaría únicamente con los registros marcados con asterisco por ser los que determinan el estado final actual de las actividades económicas asociadas a estos establecimientos.

Seguidamente se muestran los pasos a seguir para actualizar las variables estado final y actividad económica de los establecimientos que formarán parte del Directorio a actualizar. En la práctica para llevar a cabo esta tarea se tienen que ir tomando una serie de decisiones por lo que se organizan los establecimientos del siguiente modo:

1.- Establecimientos con registros procedentes de IAE y de CCSS

Se van a ordenar los registros dentro de un mismo establecimiento y de una misma actividad y de una misma fuente, es decir, se van a tener todos los registros pertenecientes a un mismo establecimiento y actividad y que proceden de IAE; y todos aquellos que proceden de una cuenta de cotización en concreto.

En el siguiente esquema se muestra la organización que se realiza en la práctica, y que se ha indicado anteriormente:

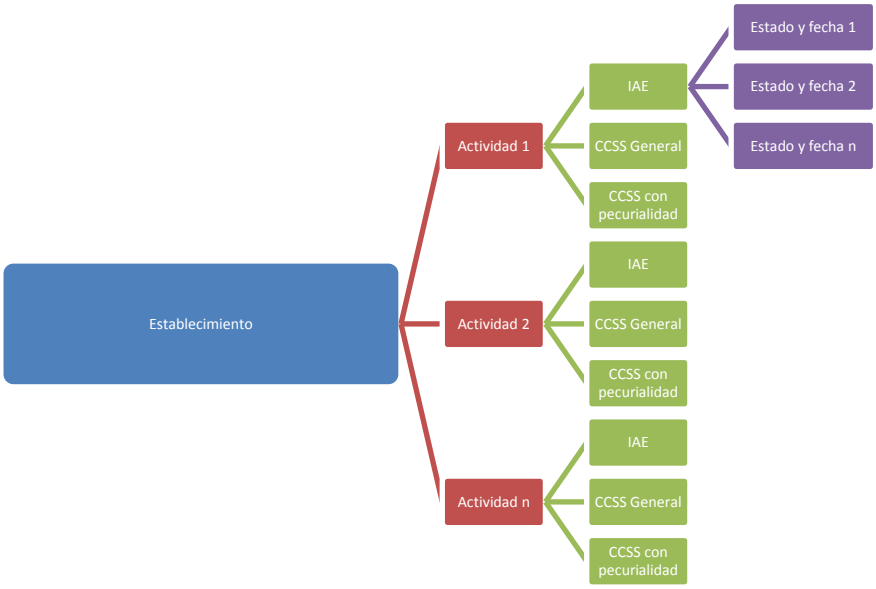

Imagen 11: Esquema de organización de los registros pertenecientes a un mismo establecimiento

Para obtener el estado final de la actividad económica en cada fuente de procedencia se ordenan los registros por fecha de estado para finalmente obtener el registro que da el estado final (ver imagen 13)

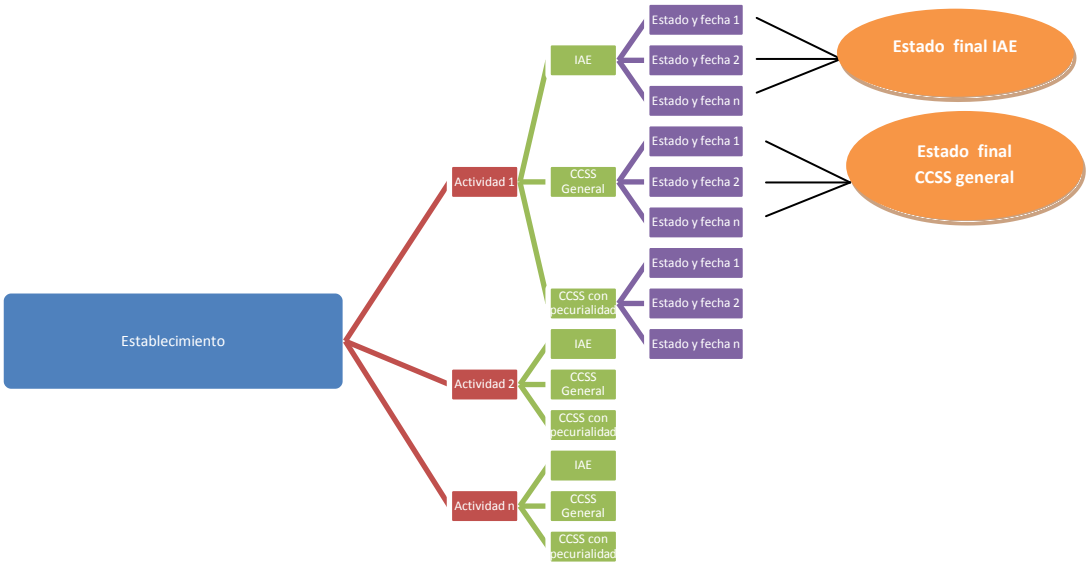

Imagen 12: Esquema de organización de los registros pertenecientes a un mismo establecimiento y que da los estados finales de la actividad en cada fuente

Posteriormente, es necesario obtener un registro único con el estado final de cada actividad económica del establecimiento. Por ejemplo, si la actividad económica 4730 aparece en IAE con estado final en alta, y también tiene asociada una cuenta de cotización en baja, entonces el estado final de la actividad económica 4730 va a ser alta, porque en una de las fuentes sigue en alta. Por tanto sólo se darán de baja las actividades que estén en baja en todas las fuentes. Por ello, en este paso se ordena por la variable 'estado de la fuente'. El segundo criterio de ordenación se basa en dejar siempre como registro prioritario el de IAE frente a la de CCSS, y entre distintas CCSS, a la cuenta general sobre las peculiaridades

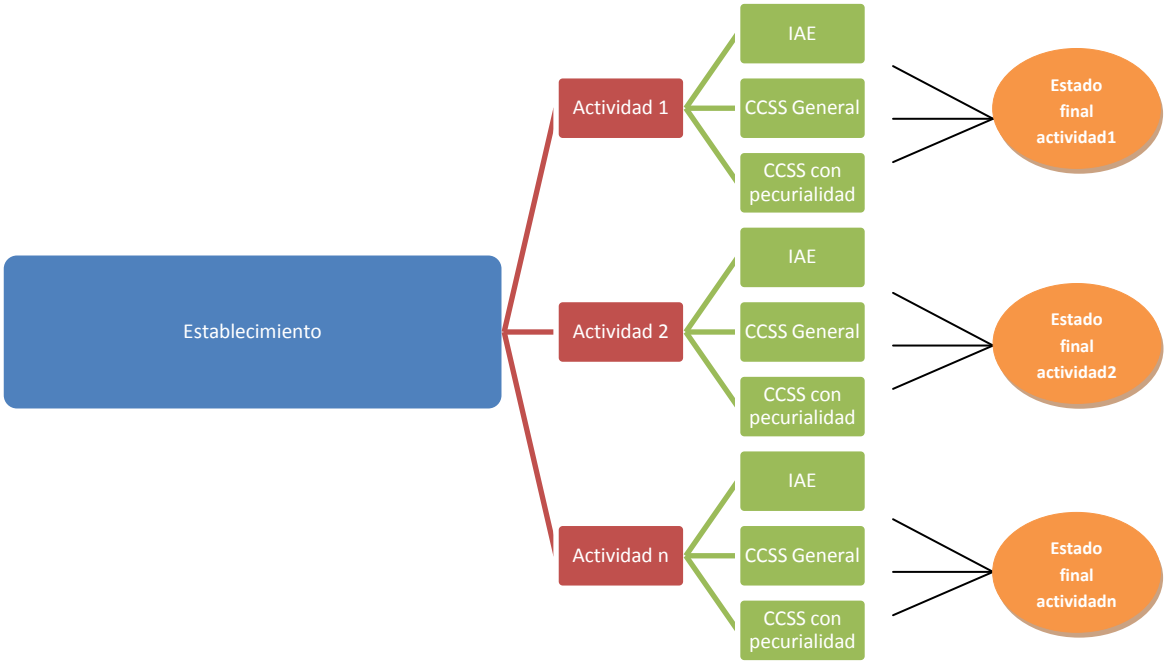

Imagen 13: Esquema de organización de los registros pertenecientes a un mismo establecimiento y que da los estados finales de la actividad

En este paso ya solo se tiene un registro por cada actividad, que nos dice el estado final de dicha actividad y la información actualizada. Lo siguiente sería elegir la actividad principal, pero sólo entre aquellas que a 1 de enero del año a actualizar estén de alta. En el caso de que tenga dos o más actividades en alta, también será necesario elegir la actividad secundaria. Si un establecimiento está compuesto por *n* actividades y todas en baja, entonces el estado final de este establecimiento es 'Baja'. En el caso contrario, si al menos el establecimiento tiene una actividad en alta, entonces el estado final del establecimiento será 'Alta'.

Los criterios usados para la elección de las actividades principales y secundarias se han fundamentado en aplicar al conjunto de actividades económicas asociadas a un establecimiento, una serie de criterios de selección aportados por el Servicio de Estadísticas Económicas del Instituto de Estadística y Cartografía.

2.- Establecimientos con registros procedentes sólo de IAE

En este caso como los registros pertenecientes a un mismo establecimiento sólo proceden de IAE entonces, únicamente es necesario elegir el estado final de cada actividad económica. Por tanto, se ordenan las fechas de estado de los registros en cada actividad económica para obtener el estado final de cada una de ellas.

Llegado a este punto, sólo falta decidir el estado final del establecimiento según si todas las actividades están en baja (estado final baja) o si al menos alguna actividad ha quedado en alta

(estado final alta). Por último, se establece la actividad principal y secundaria de cada establecimiento siguiendo las indicaciones del Servicio de Estadísticas Económicas del IECA.

3.- Establecimientos con registros procedentes sólo de CCSS

En este caso los registros que pertenecen a un mismo establecimiento sólo proceden de CCSS, pero pueden existir varias cuentas de cotización asociadas a una misma actividad, por ejemplo: una cuenta de cotización general, una cuenta de cotización en prácticas, etc. Por tanto, con que una de las cuentas esté en alta, la actividad económica estaría de alta. Por esa razón, en este caso, dentro de cada actividad es necesario ordenar por los estados de las cuentas de cotización, para que siempre se tome en primera opción una cuenta que esté en alta. A continuación se ordena por número de trabajadores, porque si hay dos cuentas de cotización en alta se elegirá la que da mayor empleo. En tercer lugar si coincide el número de trabajadores en ambas cuentas, se ordena por el tipo de relación laboral dando prioridad a las cuentas de cotización generales frente a las que presentan algún tipo de peculiaridad. La actividad sólo estará en baja si todas las cuentas de cotización asociadas a la misma actividad se encuentran de baja.

Llegado a este punto, sólo falta decidir el estado final del establecimiento según si todas las actividades están en baja (estado final baja) o si al menos alguna actividad ha quedado en alta (estado final alta). Por último, se establece la actividad principal y secundaria de cada establecimiento siguiendo las indicaciones del Servicio de Estadísticas Económicas del IECA.

5. GEOCODIFICACIÓN DE DIRECCIONES POSTALES. APLICACIÓN AL DIRECTORIO DE EMPRESAS Y ESTABLECIMIENTOS CON ACTIVIDAD ECONÓMICA EN ANDALUCÍA.

El IECA para geocodificar cualquier fichero que contenga direcciones postales con *aLink*, está utilizando como fuente principal de referencia la información alfanumérica de CDAU. Exactamente, está usando dos ficheros, uno que contiene la información alfanumérica de los portales de CDAU con las coordenadas exactas a cada portal (*fichero de portales*) y otro que contiene la información alfanumérica de las vías de CDAU con las coordenadas puntuales al centro de la vía (*fichero de viales*). Ambos ficheros se pueden descargar desde la sección de Recursos del Callejero de la página web del IECA. Además, el Instituto está siguiendo las recomendaciones de la ***Guía de georreferenciación de fuentes administrativas*** para realizar los procesos de geocodificación, por lo tanto las prioridades a la hora de geocodificar un fichero con estas fuentes son:

1º. Geocodificación a portales exactos de la vía. Se encuentra el mismo número de portal.
2º. Geocodificación a portales cercanos de la vía. En el caso de que no se encuentre la numeración exacta del portal, pero sí se encuentre un número cercano a ese portal.
3º. Geocodificación a un punto central de la vía. Si no se encuentra ni exacto ni cercano se usará el fichero de viales de CDAU.

A continuación, se muestra en la **imagen 6** cómo quedaría un fichero geocodificado con *aLink* si éste se enlaza con el fichero de portales de CDAU a través de los campos tipo de vía, nombre de vía y número. Obsérvese que en el fichero geocodificado no solo se han incluido las coordenadas X e Y sino que además se han añadido otras variables contenidas en CDAU que complementan la información del fichero original, como por ejemplo, la denominación oficial de la vía, el código INE de la vía, la referencia catastral, etc.:

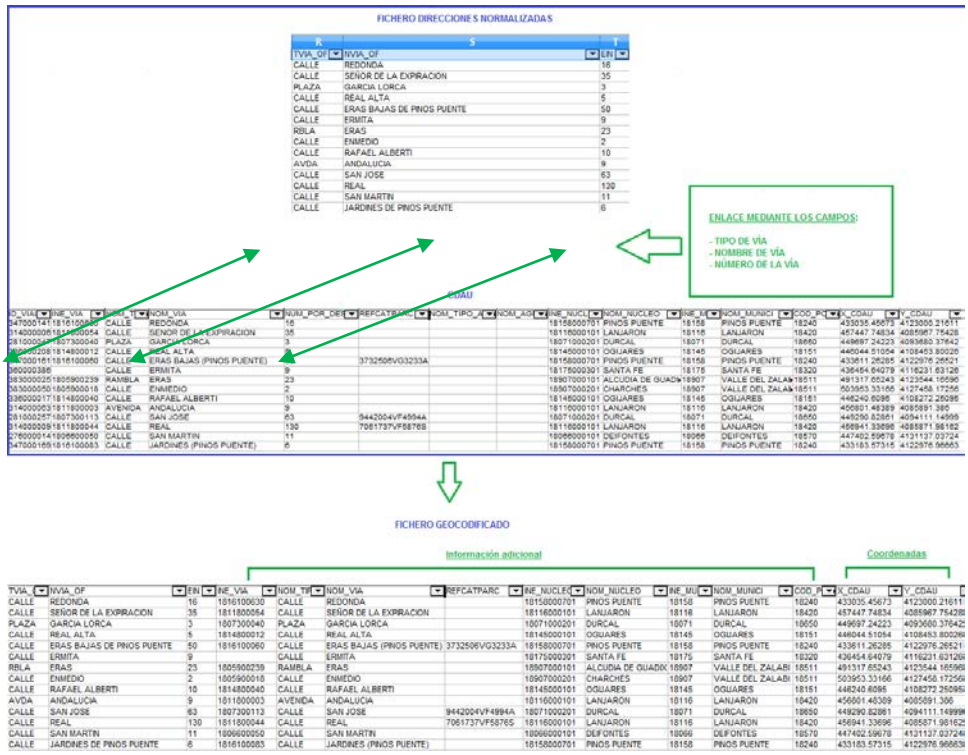

Imagen 14: Fichero antes y después de la geocodificación a través de la fusión de ficheros.

Con esta información geocodificada se podrían realizar diversos tipos de análisis espaciales con aplicaciones en múltiples campos.

Para realizar la geocodificación del Directorio de Empresas y Establecimientos con Actividad Económica en Andalucía, se ha partido de la normalización previa de las direcciones postales del mismo.

Tal y como se contaba en el punto 3, para normalizar las direcciones postales se hacía uso del portalero y vialero de CDAU, así como del tramero del Callejero del Censo Electoral del INE, para así obtener el código de vía del CDAU (ID_VIAL) o el código INE de la vía que puede encontrarse en ambas fuentes. Estos códigos son únicos para cada vía dentro de cada municipio, por lo que en la fase de geocodificación son los primeros candidatos a utilizar para dicho proceso y por eso se aprovechan de la fase de normalización.

El proceso de geocodificación se lleva a cabo principalmente, comparando las direcciones postales de Directorio con las direcciones postales de base de CDAU. En algunos casos también se ha utilizado como información base el Censo de Edificios y Viviendas 2011.

En general, los enlaces que se realizan suelen ir de más estrictos a menos, por ejemplo, se empieza usando el código postal y el código INE del municipio en un primer cruce y después se prescinde del primero y se enlaza sólo por código INE de municipio. También se comienza usando el código vía de CDAU o el código INE de vía para que los enlaces sean más rápidos y eficaces, y después aquellos que no se han enlazado por estos códigos o que no tenían rellenos estos campos (porque no se habían normalizado), se intentan cruzar usando el nombre de la vía y los tipos de vía. Cuando se hace uso de los nombres de vía es muy útil el uso de la herramienta *aLink*, para así obtener enlaces de nombres escritos de manera diferente referidos a una misma vía.

También es necesario comenzar los procesos de geocodificación buscando en primer lugar una geocodificación a portal exacto, por lo que es esencial hacer uso del número de portal y que coincida de manera exacta.

Cuando ya no es posible la geocodificación a portal exacto, entonces se intenta una geocodificación a portal cercano dejando un margen de 5 portales por encima o por debajo. En este caso se busca que al menos el portal esté en la misma acera, por lo que se suele añadir la condición de que el portal cercano sea par, si el número de portal de la vía a geocodificar es par; o impar si el número de portal de la vía a geocodificar es impar.

Por último, cuando ya no se ha podido geocodificar ni a portal exacto ni cercano, o la vía no dispone de número de portal (sin número o campo vacío), entonces se realiza una geocodificación al centro de vía. Para ello se usa el fichero de viales de CDAU que tiene las coordenadas geográficas X e Y al centro de la vía.

6. CONCLUSIONES

Como ya se viene contando a lo largo de este documento, el éxito de que se enlacen satisfactoriamente dos ficheros con variables comunes, y concretamente del proceso de geocodificación, va a depender en gran medida de la normalización de los campos a enlazar. En concreto, cuando se habla de direcciones postales es si cabe aún más importante porque dicha información suele venir escrita de múltiples maneras existiendo una gran variabilidad de posibilidades a la hora de añadir la dirección postal. Por este motivo es fundamental que la dirección postal se recoja lo mas normalizada posible, de manera que tenga una estructura similar a las direcciones del fichero con coordenadas geográficas X e Y, que se va a usar como base para el proceso de geocodificación.

El IECA utiliza el potencial de la aplicación ***aLink*** para realizar enlaces usando los campos de la dirección postal frente a los ficheros con información alfanumérica de CDAU (portales y viales). No obstante esta herramienta permite utilizar cualquier otro fichero que tenga coordenadas geográficas y que se pueda usar como base para geocodificar direcciones postales.

La versatilidad de la herramienta hace posible que se puedan enlazar registros con campos escritos de manera diferente pero que se refieren a la misma entidad, por ejemplo, el nombre de la vía que puede venir escrito de forma distinta. Así se consiguen enlaces probabilísticos.

De todos modos, cuanto más se parezca el campo en ambos ficheros mejor será el proceso de enlace y por tanto mejor será la geocodificación. Por ello, se insiste tanto en que el paso previo para cualquier proceso de geocodificación, en general para los procesos de enlace, es la normalización de los campos que se quieren usar para el proceso o procesos.

En particular, los resultados obtenidos en cuanto a normalización del último Directorio de Empresas y Establecimientos con Actividad Económica en Andalucía son los siguientes:

	IAE		CCSS		DIRECTORIO	
	valores absolutos	valores porcentuales (%)	valores absolutos	valores porcentuales (%)	valores absolutos	valores porcentuales (%)
REPOSITORIO	847.654	74,33	207.422	67,68	570.252	78,43
OTRO	80.356	7,05	17.180	5,61	34.802	4,79
SIN NORMALIZAR	212.340	19	81.893	27	121.987	0
TOTAL NORMALIZADO	928.010	81,38	224.602	73,28	605.054	83,22
TOTAL	**1.140.350**	**100,00**	**306.495**	**100,00**	**727.041**	**83,22**

Tabla 5: Resultados de la normalización de las direcciones postales del Directorio y las fuentes administrativas que se usaron para la actualización del Directorio 2017

Como puede observarse en la tabla 5, los resultados de la normalización que se había obtenido en años anteriores (fila 'Repositorio') son los responsables de normalizar la mayor parte de la información en las tres fuentes. Por ello, si cada año se consigue aumentar el repositorio de direcciones normalizadas, también irá aumentando el número de registros normalizados. En un futuro se espera que con los trabajos previos de normalización se consiga obtener casi la totalidad de la normalización, llegando a un número cercano del 100%.

La tabla 6 muestra un ejemplo de resultados obtenidos en cuanto a la última geocodificación masiva de Directorio:

	TOTAL		PORTAL EXACTO		PORTAL CERCANO		CENTRO DE VÍA	
	Valor absoluto	Valor porcentual (%)	Valor absoluto	Valor porcentual (%)	Valor absoluto	Valor porcentual (%)	Valor absoluto	Valor porcentual (%)
Sí Geocodificado	435.593	78,34	350.130	62,97	38.315	6,89	47.148	8,48
No Geocodificado	120.463	21,66						

Tabla 6: Resultados de la geocodificación de los establecimientos del Directorio 2015 que tienen menos de 50 empleados

La geocodificación de direcciones postales suele dar en general buenos resultados, en la mayoría de los casos, en torno al 80% de los registros consiguen tener una coordenada geográfica. Como puede comprobarse en la tabla anterior, el porcentaje de registros geocodificados a centro de vía suele ser mayor que a portal cercano porque en bastantes casos las direcciones vienen sin número de portal o este campo está vacío. Por ello, estos registros sólo pueden ser geocodificados a centro de vía.