Normalización en la recogida de la información de registros administrativos y geocodificación a partir de la dirección postal

Martínez Escriche, Fernando
fernando.martinez.escriche@juntadeandalucia.es
Ruiz Gabaldón, Mónica
monica.ruiz@juntadeandalucia.es
Caballero Ruiz, Elisa Isabel
elisai.caballero@juntadeandalucia.es
Galera Pozo, Ana Galera
gema.galera@juntadeandalucia.es
Instituto de Estadística y Cartografía de Andalucía

Resumen: El Plan Estadístico y Cartográfico de Andalucía 2013-2017 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 este sentido, con el objetivo de normalizar la recogida de información en las fuentes de información administrativa, el Instituto de Estadística y Cartografía de Andalucía (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 la redacción de estas reglas de normalización se ha utilizado como referencia fundamental, el informe "Task Force on Core Social Variables" publicado por Eurostat.

Las variables seleccionadas, sobre las que se proponen reglas, son la mayor parte de las denominadas "variables sociales nucleares". Éstas recogen información demográfica (sexo, edad, nacionalidad, país de nacimiento, estado civil y composición del hogar), información geográfica para su localización (país, región, provincia, localidad, dirección y coordenadas geográficas) e información socioeconómica (situación laboral, situación profesional, ocupación, sector de actividad y nivel más alto de estudios terminados).

Este manual se utiliza como herramienta para el desarrollo de la función legal del IECA sobre el informe de las normas de creación, modificación o supresión de registros administrativos.

En materia de información geográfica, tener bien recogida las variables relacionadas con la dirección postal es condición necesaria para georreferenciar fuentes administrativas.

Por ello y como ayuda en la georreferenciación de dichas fuentes, el IECA también ha elaborado una **Guía de Georreferenciación de Fuentes de Información Administrativa**. La guía pretende orientar al usuario sobre los procesos a seguir para georreferenciar fuentes de información a partir de una dirección postal y/o, en su caso, de la referencia catastral. En concreto, en ella se recogen los pasos para normalizar direcciones postales, se indican los sistemas de referencia de coordenadas, bases cartográficas e información de referencia a usar en el proceso de geocodificación, así como herramientas para realizar tal proceso. También, se señalan los procedimientos de control de calidad, se describen las principales herramientas SIG y se proporciona la forma de visualizar la información georreferenciada.

Entre las herramientas indicadas en la Guía para llevar a cabo un proceso de georreferenciación, se encuentra la aplicación informática **aLink: Herramienta de Fusión de Ficheros**, libre y gratuita y desarrollada por el IECA a partir de FEBRL (desarrollo de software libre de la Universidad Nacional de Australia). La aplicación combina una serie de técnicas en distintas etapas para llevar a cabo un proceso de fusión probabilística con ficheros con gran volumen de datos. En particular, si uno de los ficheros contiene las coordenadas geográficas X e Y asociadas a una dirección postal, es decir, está geocodificado, se podría enlazar con cualquier otro fichero que contenga direcciones postales. De esa manera éste último quedaría georreferenciado.

El Instituto de Estadística y Cartografía de Andalucía (IECA) para llevar a cabo la georreferenciación de cualquier fichero con aLink, utiliza como fuente principal de referencia la información alfanumérica del Callejero Digital Unificado de Andalucía (CDAU) y está siguiendo las recomendaciones de la Guía de Georreferenciación, en la que se indica que las prioridades a la hora de geocodificar un fichero con esta fuente son:

1º. Geocodificación a portales exactos de la vía

2º. Geocodificación a portales cercanos de la vía

3º. Geocodificación a un punto central de la vía

Para finalizar, indicar que este proceso de georreferenciación realizado en el IECA está dando buenos resultados prácticos con ficheros de grandes volúmenes de datos como, por ejemplo, el Directorio de Empresas y Establecimientos con Actividad Económica en Andalucía, del que se han obtenido una georreferenciación del 90% de sus registros.

1. Introducción

El Plan Estadístico y Cartográfico de Andalucía 2013-2017 es el instrumento de ordenación y planificación de la actividad estadística y cartográfica de la Comunidad Autónoma para sus propios fines. Es un Plan innovador, que por primera vez integra la información estadística y la cartografía.

El Plan promueve un tratamiento conjunto de ambos tipos de información, con la finalidad de conseguir que la cartografía se alimente de fuentes estadísticas y que las estadísticas avancen en su georreferenciación. La integración de los datos estadísticos con los espaciales refuerza además el valor de ambos, enriqueciéndolos mutuamente y abriendo nuevas posibilidades de utilización; de tal forma que la estadística demanda la territorialización de la información, al tiempo que la cartografía ha ampliado su alcance hacia el nuevo concepto de datos espaciales, que implica la incorporación de información georreferenciada de origen estadístico.

En este sentido, el Plan establece que la estadística y la cartografía constituyen un elemento crucial en el desarrollo de la sociedad de la información y del conocimiento, proporcionando una información que resulta imprescindible para la Administración Pública, los agentes económicos y sociales y para la ciudadanía en general. Así, la disponibilidad de datos estadísticos y cartográficos se encuentra entre las necesidades esenciales en la nueva sociedad del conocimiento.

Por otro lado, el 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.

Esta ponencia se van a presentar varios proyectos que el Instituto de Estadística y Cartografía de Andalucía está desarrollando en relación con la normalización de la recogida de información en fuentes administrativas y la georreferenciación de dichas fuentes a partir de una dirección postal.

2. Normalización en la recogida de información de fuentes y registros administrativos de la Junta de Andalucía

En la actualidad es imprescindible incidir en la mejora de la eficacia y el ahorro en los costes de la gestión. En este sentido y en materia estadística y cartográfica, la

utilización de registros y fuentes de información administrativa para el desarrollo de actividades estadísticas y cartográficas permite ampliar el campo de información para conseguir estadística y cartografía utilizables a bajo costo y, por otra parte, sustituir o complementar la recogida de datos por muestreo con su consiguiente ahorro.

En relación con estos registros y fuentes de información administrativa no hay que perder de vista que las mismas se crean para fines de gestión, pero si conseguimos introducir en el procedimiento regulador de la norma que lo crea o modifica y en el mantenimiento y actualización del sistema de información que contiene los datos recogidos, criterios normalizadores y buenas prácticas para la mejora de la calidad podremos disponer de registros con información mucho más fiable, comparable e integrada, con datos menos redundantes y previamente depurados y con alto potencial para su aprovechamiento estadístico y cartográfico.

En esta materia el Instituto de Estadística y Cartográfica de Andalucía (IECA) lleva trabajando más de 10 años. Así, en el último trimestre de 2005 comenzó a elaborar el Inventario de fuentes de información administrativa de Andalucía, una herramienta estadística, que nos permite conocer las características básicas de las fuentes administrativas de la Administración Autonómica Andaluza. Desde 2009 el Inventario se mantiene actualizado y publicado en la web del IECA http://www.juntadeandalucia.es/institutodeestadisticaycartografia/bd/infadWeb/ con un formato en el que se pueden conocer las principales características de cada fuente.

Por otra parte, la nueva redacción, tras la publicación de la Ley 4/2007, de la Ley 4/1989 de Estadística de la Comunidad Autónoma de Andalucía, estableció en su artículo 30.h) que el Instituto deberá "*Informar preceptivamente los proyectos de normas por las que se creen, modifiquen o supriman registros administrativos en lo relativo a su aprovechamiento estadístico*".

De igual forma, las Unidades Estadísticas y Cartográficas de las Consejerías y Agencias dependientes tienen entre sus funciones, según el artículo 35.c) de la citada Ley 4/1989, de 12 de diciembre, la de "*participar en el diseño y, en su caso, en la implantación, de registros o ficheros de información administrativa que sean susceptibles de posterior tratamiento estadístico, velando de manera especial por la compatibilidad de las clasificaciones utilizadas en aquellos con las clasificaciones estadísticas de uso obligatorio, así como organizar la incorporación de información*

de origen administrativo a la actividad estadística, garantizando la eficiencia, la integridad de su contenido y el respecto al secreto estadístico".

La función que la legislación atribuye al Instituto, en el artículo 30h de la Ley 4/2007, permite conocer los registros administrativos antes de su publicación (estructura, datos recogidos….), facilita la actualización del Inventario de Fuentes de Información Administrativa y lo que es más relevante, permite incidir en el contenido del registro administrativo para la mejora de su calidad encaminada a su aprovechamiento estadístico y cartográfico.

Por otra parte, en el artículo 18 del Plan Estadístico y Cartográfico 2013-2017 se establece que con la finalidad de asegurar la comparabilidad y facilitar la integración de las fuentes, los registros administrativos y los sistemas de información, el Instituto de Estadística y Cartografía de Andalucía elaborará y publicará las reglas para la normalización en la codificación de variables, siguiendo estándares nacionales e internacionales, de acuerdo con el código de buenas prácticas de las estadísticas europeas, los Reglamentos de desarrollo de la Directiva 2007/2/EC y el Esquema Nacional de Interoperabilidad.

En este sentido, con el objetivo de 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***, que se utiliza como herramienta para el desarrollo de la función legal del IECA sobre el informe de las normas de creación, modificación o supresión de registros administrativos.

Este Manual describe para un conjunto de variables frecuentemente utilizadas en registros y en encuestas, recomendaciones para la recogida, la codificación en los sistemas de información y la difusión de las mismas.

En la redacción de estas reglas de normalización se ha utilizado como referencia fundamental, el informe "Task Force on Core Social Variables" publicado por Eurostat.

Las variables seleccionadas, sobre las que se proponen reglas, son la mayor parte de las denominadas "variables sociales nucleares" elegidas por:

- Su relevancia y el uso potencial de la información recogida en las variables.
- La simplicidad de sus definiciones para la recogida de la información.

- La viabilidad de su aplicación.
- La armonización en el input que permita la recogida de la información de forma normalizada en los registros y otras fuentes de información administrativa.
- La utilización de definiciones ya existentes y suficientemente conocidas.
- El uso de estándares internacionales.

Estas variables recogen información relativa a las personas que residen en un territorio determinado, y más concretamente en la Comunidad Autónoma de Andalucía. En ellas se puede distinguir entre:

- Las que describen características sociodemográficas: sexo, edad, país de nacimiento, nacionalidad, estado civil y composición del hogar.
- Las que aportan información geográfica para la localización: país, región y provincia, localidad, dirección y coordenadas geográficas.
- Las que aportan información socioeconómica: situación laboral, situación profesional, ocupación, sector de actividad y nivel más alto de estudios terminados.

Además, las variables las podemos clasificar en función de las buenas prácticas que se recomiendan en el manual en:

- si se sugiere una clasificación estadística oficial como la ocupación, sector de actividad en el empleo y nivel más alto de estudios terminados
- si usan un estándar establecido por el Instituto Nacional de Estadística (INE) al efecto como el país de nacimiento, nacionalidad, país de residencia, región y provincia, municipio y entidad de población

- según una norma legal: sexo y coordenadas geográficas
- según su uso estadístico: edad, estado civil, composición del hogar, dirección postal, situación laboral, situación profesional

Así, por ejemplo, es necesario que se recoja el sexo de las personas, para el cumplimiento de la Ley 12/2007 de 26 de noviembre para la promoción de la igualdad de género en Andalucía, que indica que los poderes públicos de Andalucía, para garantizar de modo efectivo la integración de la perspectiva de género en su ámbito de actuación, deberán incluir sistemáticamente la variable sexo en las estadísticas, encuestas y recogida de datos que realicen. De esta manera, en las bases de datos, encuestas y fuentes administrativas debe incluirse un ítem que contemple la recogida del sexo, no siendo válido codificarlo a partir del nombre de la persona.

En el caso de la variable edad, esta es un parámetro básico en las encuestas y en los registros debido a que las diferencias existentes entre los distintos grupos de edad es una información muy relevante para el diseño de las políticas y programas públicos. Por tanto, en las bases de datos, encuestas y fuentes administrativas deben incluirse tres ítems que recojan la fecha exacta de nacimiento (día, mes y año), de manera que la información de edad se obtenga a partir de la fecha completa de nacimiento.

La versión actualizada del manual, que puede ser muy útil a las Unidades Estadísticas y Cartográficas en sus funciones de participación en el diseño e implantación de registros o ficheros de información administrativa, se puede localizar en la url:

http://www.juntadeandalucia.es/institutodeestadisticaycartografia/ieagen/instituto.html

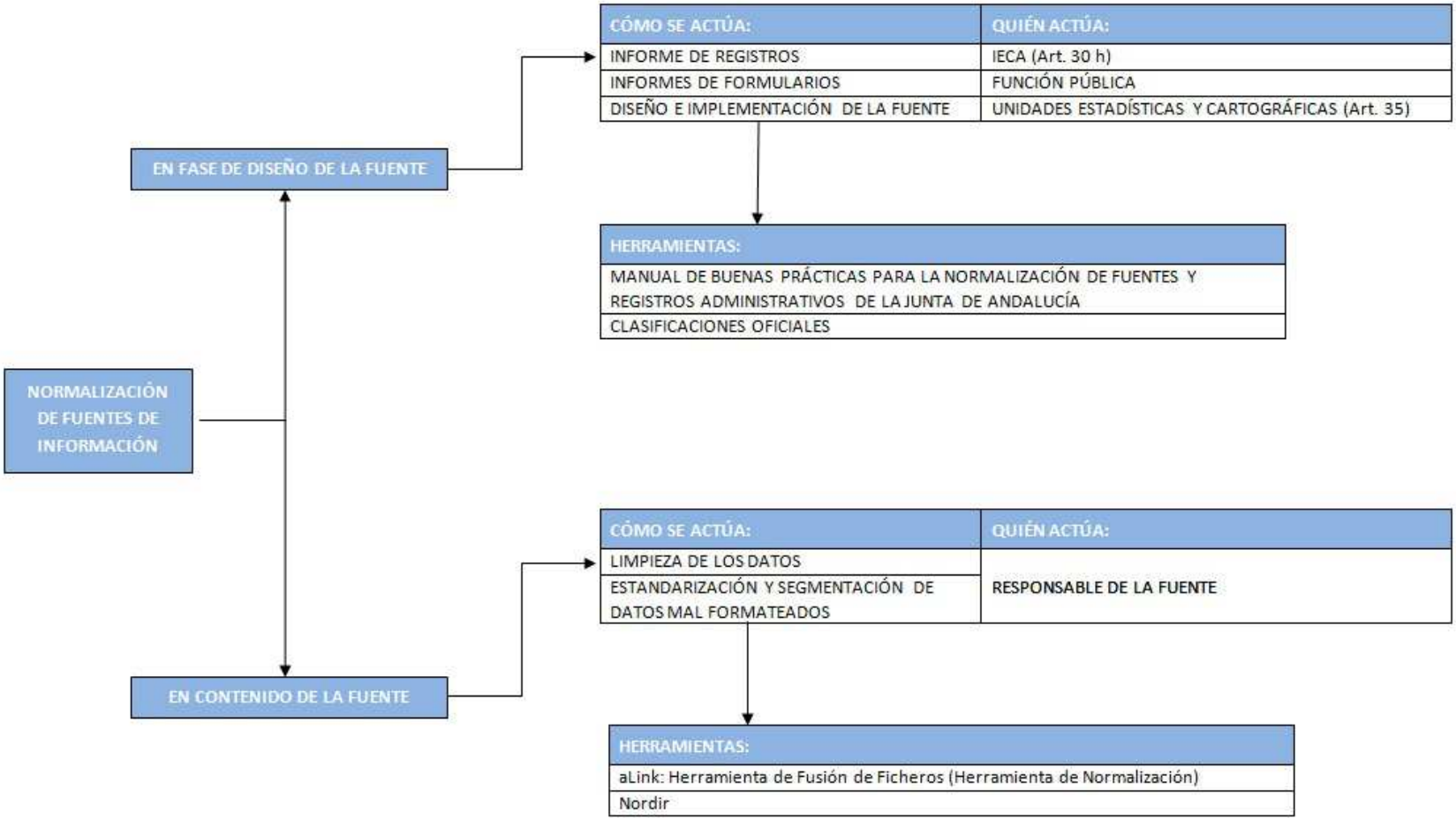
Por otra parte, una vez superada la fase de normalización en el diseño e implementación de la fuente, hay que tener presente que la información con la que nos encontramos en el mundo real, provenga o no de fuentes o registros administrativos, puede contener errores, estar incompleta o incorrectamente formateada. Por este motivo, es necesario transformar los datos originales brutos en otros datos con formatos consistentes y bien definidos, así como resolver las posibles inconsistencias sobre la forma en la que se representa y codifica la información.

Así, en relación con la información contenida en las fuentes, que no se ajuste a lo indicado en el Manual de buenas prácticas, es preciso aplicar un conjunto de técnicas encaminadas a la obtención de datos consistentes (esta fase no será precisa o necesitará el empleo de menos recursos cuanto mejor haya sido el proceso de diseño e implementación de la fuente y la automatización en los procesos para incorporar información en las fuentes) que redundarán en una mejor calidad y fiabilidad en posteriores análisis de esos datos. Este proceso es de suma importancia ya que su correcta ejecución ayudará a obtener mejores resultados con las distintas herramientas o procesos que utilicemos para la georreferenciación, ya sea ante un proceso de enlace de registros o mediante el uso de otros procesos para la geocodificación de los datos.

En el proceso de normalización de la información contenida en las fuentes se establecen dos fases principales. Una primera fase de limpieza donde no importa el contenido semántico del fichero de datos, y se realizan tareas de codificación del fichero así como de eliminación de abreviaturas y signos de puntuación en los datos contenidos en él. La segunda fase es la de estandarización del conjunto de datos, en este caso se analiza el contenido semántico del fichero y se clasifica el contenido de este según el valor de sus componentes. En esta fase, en su caso, se realizará la segmentación de los datos en cada una de las componentes que los forman.

Para finalizar, en el siguiente esquema se resumen las distintas fases del proceso de normalización, que desde el Sistema Estadístico y Cartográfico de Andalucía se puede actuar en las mismas a través de mecanismos y utilizando una serie de herramientas.

Esquema 1. Normalización de fuentes de información: fases, posibilidades de actuación, agentes y herramientas



3. Georreferenciación de fuentes administrativas

La geocodificación puntual es el proceso de asignar coordenadas geográficas (X e Y) a puntos del espacio, frecuentemente direcciones postales. Las coordenadas obtenidas posibilitan la ubicación de los elementos en un mapa en un Sistema de Información Geográfico y estos pueden ser almacenados, manipulados y analizados para poder resolver problemas y satisfacer unas necesidades concretas de información.

Con el objetivo de georreferenciar las fuentes de información administrativa a partir de una dirección postal y, en su caso, de la referencia catastral, se ha elaborado una ***Guía de georreferenciación de fuentes de información administrativa*** que pretende orientar al usuario de los pasos a seguir para georreferenciar fuentes de información así como difundir los procedimientos y las herramientas necesarias para llevar a cabo dicho proceso.

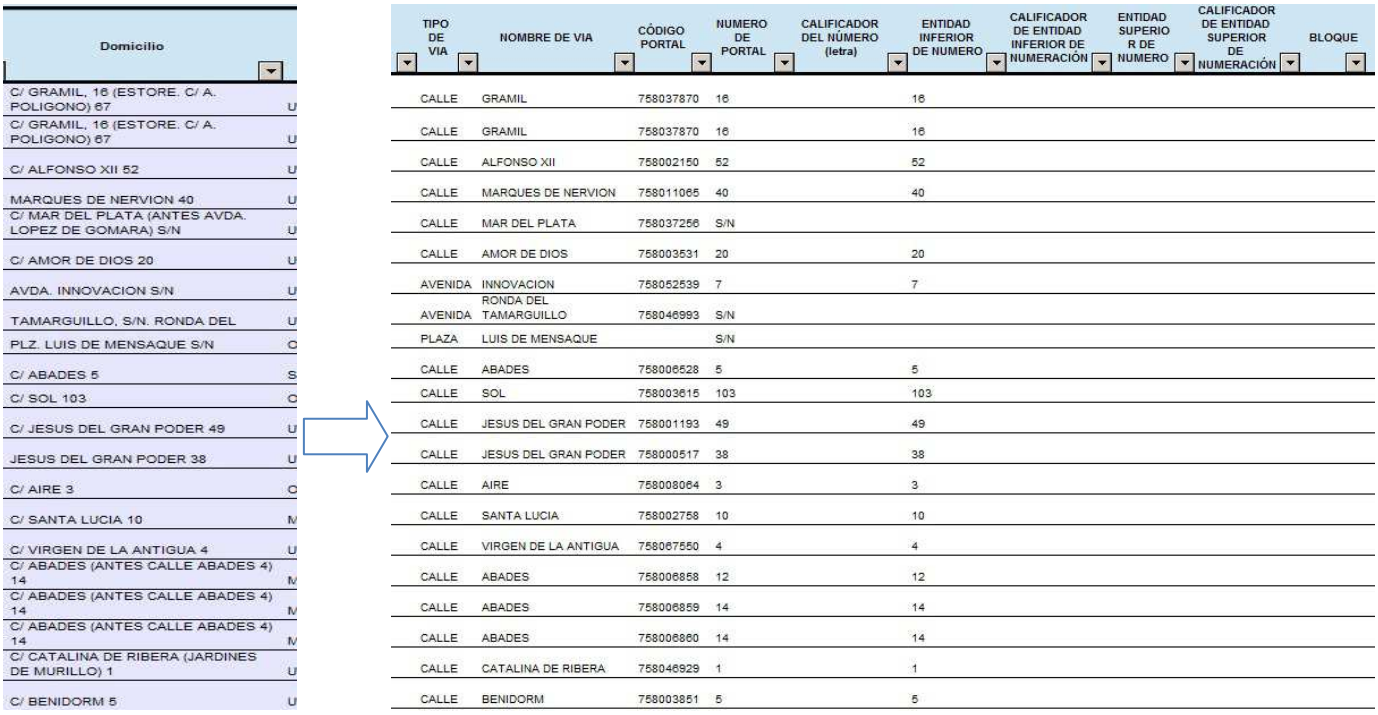
El contenido de la Guía se basa en la descripción y desarrollo de los siguientes temas:

- Recogida de información geográfica: se describe cómo se tiene que recopilar la información de la dirección postal y de la referencia catastral para poder geocodificar la fuente o registro administrativo cuando los datos de las coordenadas geográficas no figuren entre la información que ofrece la fuente.
- Herramientas de normalización: se indica el proceso de normalización de la dirección postal y se describen distintas aplicaciones informáticas para realizarlo. Entre estas aplicaciones que realizan de forma sencilla un proceso de normalización de direcciones postales a partir de un fichero cuya estructura de datos conocemos están: NorDir y la Herramienta de Normalización de la aplicación informática *aLink: Herramienta de Fusión de Ficheros*.

    - NorDir es el cliente web del servicio de normalización del Callejero Digital de Andalucía. Forma parte del proyecto del Sistema de Información Geográfica Corporativo de la Junta de Andalucía (SIGC) y permite normalizar un fichero de direcciones, generando un fichero csv con el resultado de la normalización.
    - La Herramienta de Normalización permite normalizar direcciones postales, nombres de personas e identificadores de personas físicas y

jurídicas. Esta herramienta se incluye en el programa aLink: Herramienta de Fusión de Ficheros, aplicación de software libre, desarrollada por el IECA a partir de FEBRL, accesible a todos los usuarios y descargable desde la sección de Descarga de software de la página web del Instituto de Estadística y Cartografía de Andalucía: http://www.juntadeandalucia.es/institutodeestadisticaycartografia/ieagen/otrosServidores/software/index.htm#uno

Ambas herramientas nos devuelven un fichero con los datos de la dirección postal normalizados y desagregados en varios campos (véase Imagen 1).



Domicilio	
C/ GRAMIL, 16 (ESTORE. C/ A. POLIGONO) 67	U
C/ GRAMIL, 16 (ESTORE. C/ A. POLIGONO) 67	U
C/ ALFONSO XII 52	U
MARQUES DE NERVION 40	U
C/ MAR DEL PLATA (ANTES AVDA. LOPEZ DE GOMARA) S/N	U
C/ AMOR DE DIOS 20	U
AVDA. INNOVACION S/N	U
TAMARGUILLO, S/N. RONDA DEL	U
PLZ. LUIS DE MENSAQUE S/N	C
C/ ABADES 5	S
C/ SOL 103	C
C/ JESUS DEL GRAN PODER 49	U
JESUS DEL GRAN PODER 38	U
C/ AIRE 3	C
C/ SANTA LUCIA 10	M
C/ VIRGEN DE LA ANTIGUA 4	U
C/ ABADES (ANTES CALLE ABADES 4) 14	M
C/ ABADES (ANTES CALLE ABADES 4) 14	M
C/ ABADES (ANTES CALLE ABADES 4) 14	M
C/ CATALINA DE RIBERA (JARDINES DE MURILLO) 1	U
C/ BENIDORM 5	U

TIPO DE VIA	NOMBRE DE VIA	CÓDIGO PORTAL	NUMERO DE PORTAL	CALIFICADOR DEL NÚMERO (letra)	ENTIDAD INFERIOR DE NUMERO	CALIFICADOR DE ENTIDAD INFERIOR DE NUMERACIÓN	ENTIDAD SUPERIOR DE NUMERO	CALIFICADOR DE ENTIDAD SUPERIOR DE NUMERACIÓN	BLOQUE
CALLE	GRAMIL	758037870	16		16				
CALLE	GRAMIL	758037870	16		16				
CALLE	ALFONSO XII	758002150	52		52				
CALLE	MARQUES DE NERVION	758011065	40		40				
CALLE	MAR DEL PLATA	758037256	S/N						
CALLE	AMOR DE DIOS	758003531	20		20				
AVENIDA	INNOVACION	758052539	7		7				
AVENIDA	RONDA DEL TAMARGUILLO	758046993	S/N						
PLAZA	LUIS DE MENSAQUE		S/N						
CALLE	ABADES	758006528	5		5				
CALLE	SOL	758003615	103		103				
CALLE	JESUS DEL GRAN PODER	758001193	49		49				
CALLE	JESUS DEL GRAN PODER	758000517	38		38				
CALLE	AIRE	758008064	3		3				
CALLE	SANTA LUCIA	758002758	10		10				
CALLE	VIRGEN DE LA ANTIGUA	758067550	4		4				
CALLE	ABADES	758006858	12		12				
CALLE	ABADES	758006859	14		14				
CALLE	ABADES	758006860	14		14				
CALLE	CATALINA DE RIBERA	758046929	1		1				
CALLE	BENIDORM	758003851	5		5				

I

Imagen 1. Resultado del campo domicilio tras su normalización

- Sistema de referencia de coordenadas y coordenadas geográficas: se indican los sistemas de referencia de coordenadas, las bases cartográficas e información de referencia a usar en el proceso de geocodificación. Es fundamental que se defina, además del sistema de referencia que se usa cuando se trabaja con coordenadas geográficas proyectadas, el huso en la que se trabaja y tratar toda la información en el mismo sistema de referencia.

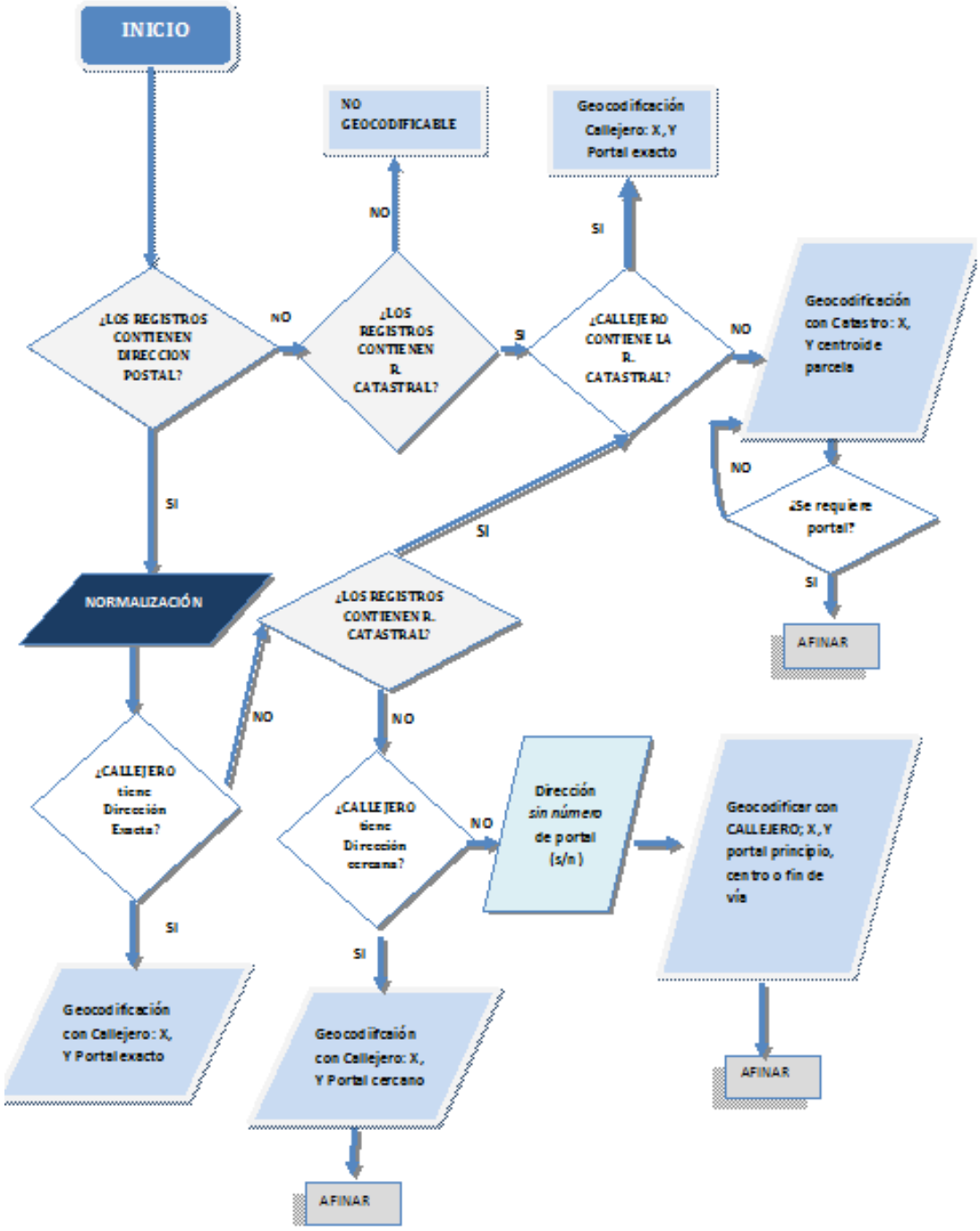
El Real Decreto 1071/2007 establece el ETRS89 como sistema de referencia geodésico oficial en España para la referenciación geográfica y cartográfica en el ámbito de la Península Ibérica y las Islas Baleares. Cada Sistema de Referencia de Coordenadas (CRS) tiene asociado un código numérico denominado EPSG y que es usado ampliamente. Los más habituales son:

Sistema de referencia	Sistema de coordenadas	Código EPSG
ETRS89	Geográficas	4258
	UTM huso 30	25830
ED50	Geográficas	4230
	UTM huso 30	23030
WGS84	Geográficas	4326
	UTM huso 30	32630

Tabla 1. Sistemas de referencia, sistemas de coordinadas y códigos EPSG

Además, se describe la herramienta CALAR, desarrollada por la Junta de Andalucía, que facilita la transformación de la información espacial contenida en ficheros en los formatos más comunes entre los diversos sistemas de referencia oficiales en Andalucía.

- Proceso de geocodificación: se explica el proceso de geocodificación para fuentes y registros administrativos. Dicho proceso se resume en el siguiente esquema:



Esquema 2. Proceso de georreferenciación

- Herramientas para geocodificar: se señalan las distintas herramientas para el proceso de geocodificación dependiendo del modelo de datos que tengamos y de la información de localización. Entre ellas se mencionan herramientas como Geodir, Geocoder, aLink o Telegeo.

- Control de calidad de la información: en el caso de la georreferenciación puntual de registros, la garantía de que el producto final es el adecuado y que se ajusta a las necesidades del usuario depende, por un lado, de que la información que se utiliza como base sea lo más óptima posible, y por otro, que el resultado tras el proceso de georreferenciación ha dado los resultados adecuados. Por ello, el control de calidad viene determinado por dos procesos, por una parte, la información georreferenciada depende de la fuente de base que se utiliza para la geocodificación o la obtención de las coordenadas X e Y, y por otra, debe someterse a un control de calidad los datos resultantes de la georreferenciación para garantizar que el proceso ha dado resultados óptimos.

  Para el caso de los datos geocodificados los controles de calidad a aplicar así como la exigencia de los mismos dependerán del uso que se vaya a dar a la información cartográfica. El dato geográfico se caracteriza por una posición espacial, sus atributos y el tiempo en el que suceden. En el caso de la geocodificación hay que tener en cuenta fundamentalmente la exactitud posicional y completitud, ya que las demás características han quedado ya establecidas en el control de calidad de la fuente de base utilizada.

- Herramientas de la información georreferenciada: se mencionan algunos software SIG para el tratamiento de la información georreferenciada: ARCGIS, GVSIG, KOSMO y QGIS.
- Difusión de los datos georreferenciados: se indican como publicar los datos espaciales. Está se hará a partir de un servicio web estándar, que se basa en un sistema distribuido con una arquitectura cliente/servidor. Para servir información espacial por la red o a través de internet, es necesario: un Servidor Web y un Servidor de Mapas.

  Algunos de los servidores de mapas de código abierto (Open Source) que cumplen uno o varios de los principales estándares en relación con el acceso a datos espaciales son: Mapserver y Geoserver.

  Los estándares de servicios geográficos más usuales son: WMS y WFS.

  Para poder visualizar la información espacial que generan los servicios web existentes es necesario la utilización de visores. Estos deben ser capaces de consultar servicios estándares de mapas cumpliendo las especificaciones de la OGC y las normativas ISO. Un visor de servicios desarrollado por el SIG

Corporativo es Mapea que es una herramienta para la inserción de visores cartográficos en páginas web (Mapshup) pero que a su vez permite generar directamente una dirección o URL e insertarla en cualquier navegador.

Por otro lado, la geocodificación depende de la información alfanumérica que nos indica su localización. En el caso de direcciones postales, ésta puede realizarse al punto donde se encuentra el portal. Sin embargo, la referencia catastral de Catastro realiza la geocodificación al centroide de la parcela donde se encuentra el inmueble.

Las dos informaciones de referencia usadas para obtener las coordenadas de la georreferenciación, son:

a. Callejero Digital de Andalucía Unificado (CDAU): coordenadas a portales de las vías y coordenadas a los centros de las vías
b. Catastro: coordenadas al centroide de la parcela

En la guía se definen los pasos a seguir, así como las prioridades a tener en cuenta en el proceso:

1. Geocodificación a portales exactos -> CDAU -> información que podemos usar: dirección y referencia catastral
2. Geocodificación al centroide de las parcelas -> Catastro -> información que podemos usar: referencia catastral
3. Geocodificación a portales cercanos -> CDAU -> información que podemos usar: dirección postal
4. Geocodificación al punto central de la vía -> CDAU -> dirección postal sin portal

Cada prioridad está basada en la información de localización que aporta el registro: por ejemplo, si un registro no aporta información de referencia catastral, entonces si éste no puede georreferenciarse a portal exacto, se intentará georreferenciar a portal cercano y si no al centro de la vía.

4. Herramienta para geocodificar

Para realizar la geocodificación de un fichero, existen distintas herramientas cuya utilización va a depender del modelo de datos que tengamos y de la información de localización.

Entre las herramientas indicadas en la Guía de georreferenciación de fuentes de información administrativa se encuentra ***aLink: Herramienta de Fusión de Ficheros***.

La aplicación combina una serie de técnicas en distintas etapas para llevar a cabo un proceso de fusión de ficheros de grandes volúmenes de datos. En el siguiente esquema se pueden visualizar cada una de las etapas:

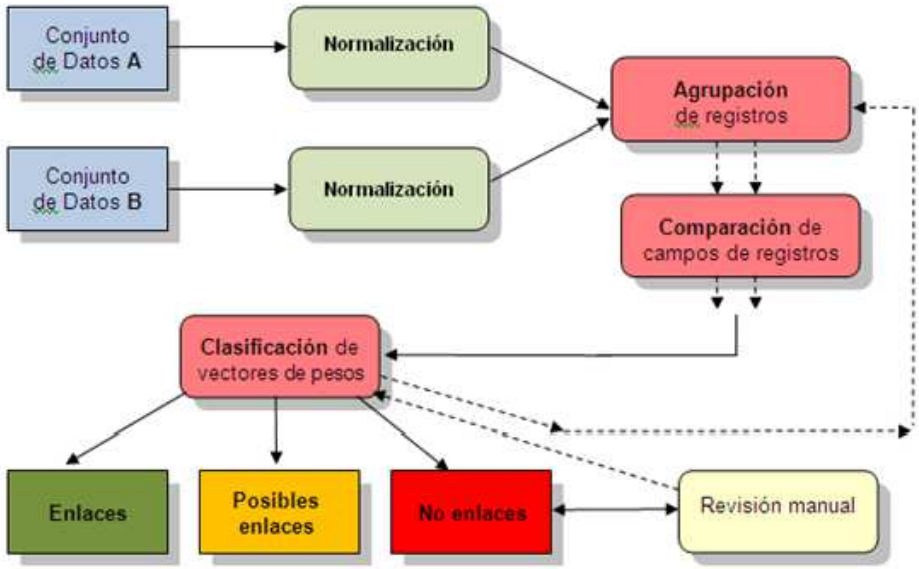


Imagen 2: Etapas del proceso de fusión de ficheros

aLink: Herramienta de Fusión de Ficheros está compuesta por dos herramientas, siendo su interfaz principal la que se muestra a continuación:

Imagen3: Interfaz principal de aLink: Herramienta de Fusión de Ficheros.

Herramienta de Normalización

El proceso de normalización se considera fundamental para obtener buenos resultados en un proceso de enlace. Para el mismo se requiere que existan campos en los ficheros que se van a enlazar con la misma estructura e información que sea comparable. Con esta herramienta se podrán transformar los datos originales brutos en otros con formato consistente: limpiando, estandarizando y segmentando los mismos. Además, con ella tenemos cubierta la primera etapa del proceso de fusión de ficheros: la normalización.

En concreto, esta herramienta permite la normalización de nombres de personas, direcciones postales e identificadores de personas físicas o jurídicas. Constituye la evolución de ADYN Herramienta de Normalización pero incluyendo nuevas funcionalidades y mejoras, sobre todo en lo relativo a direcciones postales. Por ejemplo, la aplicación ofrece ahora al usuario la posibilidad de desagregar una dirección postal de acuerdo a los campos de salida que ofrece el Callejero Digital de Andalucía Unificado (CDAU).

Por tanto, si se dispone de un fichero con campos que contienen información sobre nombres de personas, direcciones postales y DNI o NIF, con la Herramienta de Normalización podremos normalizar el contenido de los mismos de la siguiente forma:

*Campo **NOMBRE***:

NOMBRE
MARIA DEL CARMEN DEL MORAL RUIZ
ANTONIO GARCIA MARTINEZ
IRINA PETROV
FRANCOIS DUBOIS
ANA MARIA DIAZ MARTINEZ
FCO MARTIN VELASCO
JUAN DE DIOS LOPEZ HIDALGO
MARIA FERNANDEZ JIMENEZ

nombre1	particula_nombre1	nombre2	particula_nombre2	nombre3	preparticula_apellido1	apellido1	particula_apellido1	preparticula_apellido2	apellido2	particula_apellido2
maria	del	carmen			del	moral			ruiz	
antonio						garcia			martin	
irina						petrov				
francois						dubois				
ana		maria				diaz			martinez	
francisco						martin			velasco	
juan	de	dios				lopez			hidalgo	
maria						fernandez			jimenez	

*Campo **DIRECCIÓN***:

DIRECCIÓN POSTAL
C/ DIVINA PASTORA Nº 47 PISO 2 EDIFICIO ALAMEDILLA
C/ LOJA 13 POLIG JUNCARIL SECTOR A MANZ 2 PARC 15
AVDA DE CÁDIZ, 145
CTRA A 92 SEVILLA ALMERIA KM 375
PARAJE CAMINO VIEJO S/N
CL EJIDO PLAZA DE TOROS
CTRA N 340, PK 50
C/ EL SALVADOR

tipo_de_via	nombre_de_via	identificador_de_numeracion	ein	cein	esn	cesn	bloque	portal	escalera	planta	puerta	entidad_singular	municipio	provincia	codigo_postal	tipo_de_agrupacion	agrupacion	odub
calle	divina pastora	numero	47							2								edificio alamedilla
calle	loja		13													poligono_industrial	juncaril	sector a manzana 2
avenida	de cadiz		145															
carretera	a 92 sevilla almeria	kilometro	375															
camino	viejo	sin_numero																paraje camino viejo
calle	ejido plaza de toros																	
carretera	n 340	kilometro	50															
calle	el salvador																	

*Campo **NIF***:

A
NIF
74639267n
a88092546
w76092176

A	B	C	D
nif	letra_inicio	numero_id	caracter_control
74639267n		74639267	n
a88092546	a	8809254	6
w76092176	w	7609217	6

Herramienta de Enlace

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. Consiste en comparar dos ficheros para detectar aquellos registros que corresponden a una misma entidad o unidad poblacional (individuos, establecimientos, 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 ello, se utilizan diversas medidas de similitud que permiten generar enlaces exactos o aproximados. Con esta herramienta se tienen cubiertas las tres etapas restantes del proceso de fusión de ficheros: agrupación, comparación y clasificación.

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.

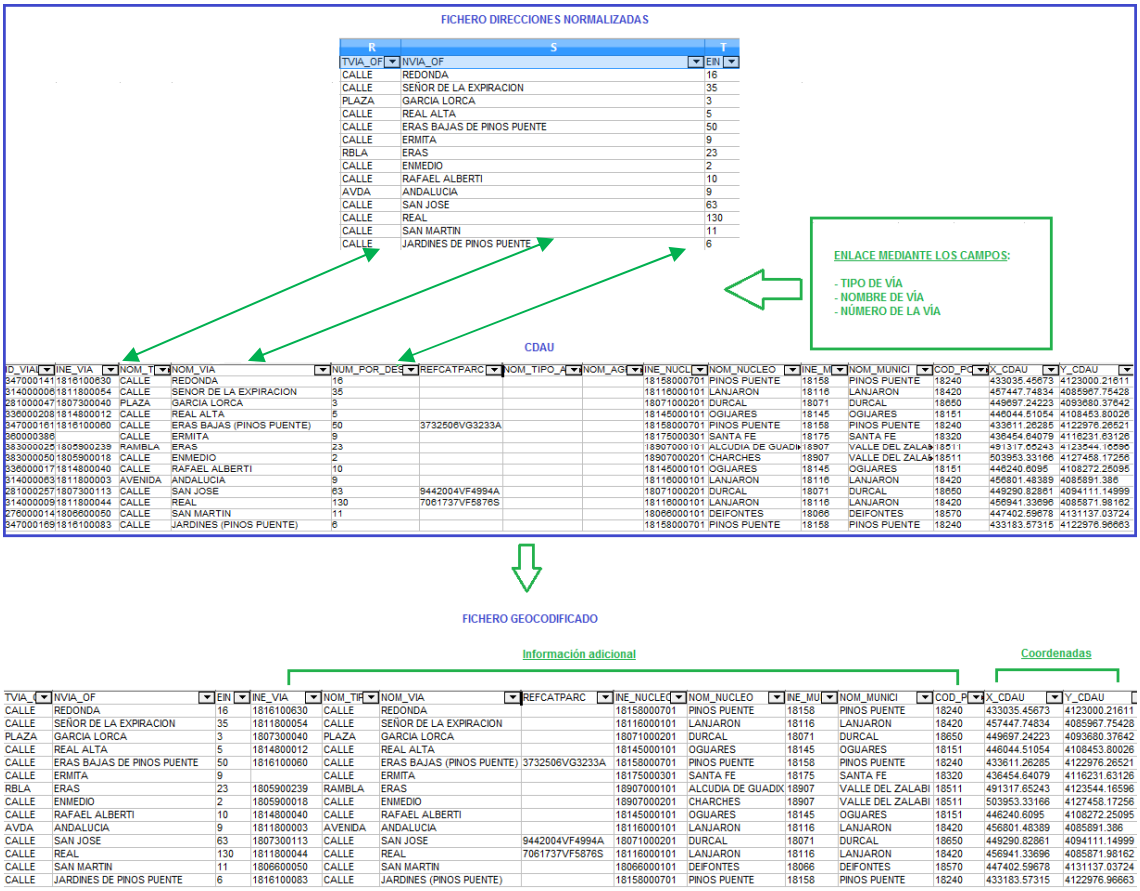
En particular, el IECA para geocodificar cualquier fichero que contenga direcciones postales con aLink, está utilizando como fuente principal de referencia la información alfanumérica de CDAU. Exactamente, está usando dos ficheros, uno que contiene la información alfanumérica de los portales de CDAU 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 Utilidades del Callejero de la página web del IECA. Además, el Instituto está siguiendo las recomendaciones de la Guía de georreferenciación de fuentes administrativas, por lo tanto las prioridades a la hora de geocodificar un fichero con estas fuentes son:

1º. Geocodificación a portales exactos de la vía. Si un registro contiene información de la entidad de numeración que coincide exactamente con el número de portal de un registro del fichero de portales de CDAU, conseguiremos realizar una geocodificación a portal exacto.

2º. Geocodificación a portales cercanos de la vía. En el caso de que no encontramos la numeración exacta del portal, pero si encontramos un número cercano a ese portal entre los registros del fichero de portales de CDAU, entonces se realizará una geocodificación a portal cercano.

3º. Geocodificación a un punto central de la vía. Si tampoco se encuentra el portal cercano, se realizará la geocodificación al centro de la vía de la dirección postal. En este caso se usará el fichero de viales de CDAU.

A continuación, se muestra 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.:



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

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.

Caso práctico con aLink: Herramienta de Fusión de Ficheros

Para finalizar vamos a mostrar algunos de los resultados obtenidos al geocodificar con aLink el Directorio de Empresas y Establecimientos con Actividad Económica en Andalucía. Como se ha comentado anteriormente, para llevar a cabo este proceso se han utilizadolos ficheros de portales y viales de CDAU. Hay que indicar además, que en este proceso solamente se han geocodificado los establecimientos en alta y no las empresas puesto que muchas de ellas están ubicadas fuera de la Comunidad Autónoma de Andalucía. En total el fichero a geocodificar contenía 546.118 establecimientos y aunque el proceso de geocodificación podría haberse llevado a cabo con el fichero completo, se decidió por operabilidad dividirlo por provincias.

En cuanto a los resultados obtenidos en la etapa de normalización hay que indicar que prácticamente el 96% de los registros de cada provincia se han normalizado al realizar el primer proceso de normalización. El porcentaje restante se ha conseguido normalizar realizando nuevos procesos o incluso normalizándolos manualmente si su número era reducido.

Por otro lado, en relación con el proceso de enlace, el Directorio contiene información de la localización del establecimiento a través de la dirección postal. Los campos que han intervenido en los procesos de enlace han sido principalmente los que se muestran en la siguiente tabla:

DIRECTORIO	CDAU PORTALES	CDAU VIALES	DESCRIPCIÓN
INEVIA	INEVIA	INEVIA	CODIGO INE DE LA VIA
TIPO_VIA	NOM_TIP_VIA	NOM_TIP_VIA	TIPO DE VIA NORMALIZADO
NOMBRE_VIA	NOM_VIA	NOM_VIA	NOMBRE DE LA VIA NORMALIZADO
EIN	NUM_POR_DESDE	NUM_POR_DESDE	ENTIDAD INFERIOR DE NUMERACIÓN
TIPO_AGRUPACION	NOM_TIP_AGRUPACION	-	TIPO DE AGRUPACIÓN NORMALIZADO (POLIG IND, BARRIO, URB., ETC.)
AGRUPACION	NOM_AGRUPACION	-	NOMBRE DE LA AGRUPACIÓN NORMALIZADO

ODUB	TXT_APP	-	OTROS DATOS DE UBICACIÓN
CODMUN	INEMUN	INEMUN	CODIGO MUNICIPIO
CODPOS	COD_POSTAL	-	CODIGO POSTAL

En cuanto a los resultados obtenidos en los diferentes procesos de enlace, se puede decir que alrededor del 90% de los registros de cada provincia se encuentran geocodificados. La mayor parte de los registros sin geocodificar corresponden a direcciones ubicadas en carreteras, lo que permite deducir que es preciso mejorar la recogida de información de éstas en el fichero del Directorio.

	Nº registros	**Geocodificados**					**No geocodificados**
Provincia		**Exacto**	**Cercano**	**Centro vía**	**Zona cercana**	**Total**	
Almería	45.166	21.950 (48.60%)	6.595 (14.60%)	4.243 (9.39%)	575 (1.27%)	**33.363 (73.87%)**	11.803 (26.13%)
Cádiz	68.948	42.174 (61.17%)	7.830 (11.36%)	11.083 (16.07%)	895 (1.30%)	**61.982 (89.90%)**	6.966 (10.10%)
Córdoba	54.383	35.093 (64.53%)	6.843 (12.58%)	7.699 (14.16%)	85 (0.16%)	**49.720 (91.43%)**	4.663 (8.57%)
Granada	62.200	41.194 (62.23%)	7.218 (11.6%)	9.197 (14.79%)	372 (0.6%)	**57.981 (93.22%)**	4.219 (6.78%)
Huelva	29.328	17.845 (60.85%)	3.553 (12.11%)	3.900 (13.30%)	790 (2.69%)	**26.088 (88.95%)**	3.240 (11.05%)
Jaén	39.126	26.207 (66.98%)	4.546 (11.62%)	4.010 (10.25%)	772 (1.97%)	**35.535 (90.82%)**	3.591 (9.18%)
Málaga	120.913	75.699 (62.61%)	10.389 (8.59%)	22.420 (18.54%)	611 (0.51%)	**109.119 (90.25%)**	11.794 (9.75%)
Sevilla	126.054	87.603 (69.50%)	18.877 (14.98%)	10.696 (8.49%)	180 (0.14%)	**117.356 (93.10%)**	8.698 (6.90%)
Total	**546.118**	**347.765 (63.68%)**	**65.851 (12.06%)**	**73.248 (13.14%)**	**4.280 (0.78%)**	**491.144 (89.93%)**	**54.974 (10.07%)**

Tabla 2.- Resultados del proceso de geocodificación con aLink del fichero del Directorio de Empresas y Establecimientos con actividad económica en Andalucía

Conclusiones

Para obtener una información de calidad que permita su aprovechamiento estadístico y cartográfico es esencial normalizar los datos recogidos en las fuentes de información administrativa.

Esta normalización es el paso previo para el uso de fuentes administrativas y nos va a permitir ampliar el campo de las estadísticas y cartografías generadas así como avanzar en la desagregación territorial de los datos, ya que las fuentes de información públicas, suelen disponer de la dirección postal. Por otra parte, la utilización de fuentes permitirá mejorar la eficiencia administrativa así como disminuir las solicitudes de información a las personas físicas y jurídicas mediante encuestación.

El Sistema Estadístico y Cartográfico de Andalucía, dispone de mecanismos legales, para informar todos los registros que se creen, modifiquen o supriman en la Junta de Andalucía, así como un manual de buenas prácticas para la recogida de información en el que propone como recoger la información para las variables que aparecen más frecuentemente en los registros. Adicionalmente ha desarrollado herramientas para la normalización de información como aLink o Nordir.

Con los instrumentos antes mencionados, es posible normalizar la información de las fuentes de información administrativa tanto actuando antes de la publicación de la norma que de soporte al registro como a posteriori en relación con los datos incorporados en el sistema de información en el que se implemente la fuente.

La normalización de la información permitirá, mediante un proceso de enlace, la integración de fuentes y por tanto la posibilidad de añadir datos de unas fuentes en otras y disponer de información muy rica para el desarrollo de actividades estadísticas y cartográficas.

Un caso relevante en materia de normalización e integración de fuentes, consiste en la recogida normalizada de las direcciones postales, que permitirá a su vez, enlazando éstas con el callejero, en el caso de Andalucía, el Callejero Digital de Andalucía Unificado (CDAU), la geocodificación de la información y por tanto su representación en el territorio. Para realizar estas tareas, el Sistema Estadístico y

Cartográfico de Andalucía dispone de una guía para la geocodificación, información actualizada de CDAU y herramientas como aLink para realizar procesos de enlace.

Disponer de información georreferenciada nos va a permitir realizar análisis espaciales con aplicaciones en múltiples campos como el urbanismo, el geomarketing, el desarrollo rural, la realización de obras públicas, ubicación óptima de elementos, prevención de riesgos naturales, turismo...

Utilizando la herramienta aLink, aplicando el procedimiento marcado en la guía de geocodificación y utilizando como información de referencia CDAU, se ha llevado a cabo, entre otras, la geocodifcación del Directorio de Empresas y Establecimientos de Andalucía, una base de datos con más de 500.000 registros y para la que se ha obtenido un porcentaje de geocodificación del 90%.