Junta de Andalucía | Consejería de Economía, Hacienda y Fondos Europeos | Instituto de Estadística y Cartografía de Andalucía

MEMORIA TÉCNICA DE LA ACTIVIDAD

"MÉTODOS AUTOMÁTICOS DE ENLACE DE REGISTROS"

Cofinanciado por la Unión Europea | MINISTERIO DE HACIENDA Y FUNCIÓN PÚBLICA | Fondos Europeos

0. IDENTIFICACIÓN

- **Código y denominación de la actividad:** 14.00.16 Métodos automáticos de enlace de registros
- **Organismo responsable**: Instituto de Estadística y Cartografía de Andalucía
- **Unidad ejecutora:** Servicio de Planificación y Coordinación
- **Organismos colaboradores y convenio:**

  -

1. INTRODUCCIÓN

- **Objetivos**

La información disponible ha crecido de forma exponencial en los últimos tiempos. Esta información puede proceder de una o más fuentes de datos, tan distintas como censos, encuestas o fuentes administrativas y, a menudo, es necesario integrarla para poder llevar a cabo su aprovechamiento estadístico o cartográfico exhaustivo.

En este contexto, las técnicas de enlace de registros juegan un papel importante ya que además de enlazar registros de uno o dos ficheros para intentar determinar qué parejas de ellos se refieren a una misma entidad, mejoran la integridad y la calidad de los datos, permitiendo reutilizar fuentes de información ya existentes y reducir costes y esfuerzo en la adquisición de información para realizar nuevos estudios.

Los métodos que permiten enlazar registros o encontrar duplicados son variados. El caso más sencillo es aquel en el que se dispone de un único identificador sobre la entidad de interés, común a todos los conjuntos de datos que se van a enlazar. En esta situación, el problema es trivial ya que el enlace se puede realizar mediante dicho identificador, utilizando algún lenguaje de programación (por ejemplo, en SQL el operador “join”).

Así pues, la finalidad de esta actividad es generar una metodología adecuada que permita implementar métodos automáticos para el enlace de registros o fusión de ficheros. Además, se proporciona una herramienta informática orientada a este fin.

El objetivo es disminuir las solicitudes de información a las personas físicas y jurídicas mediante el aprovechamiento de las técnicas de fusión de ficheros y mejorar los aspectos de coordinación y los procedimientos metodológicos del Sistema Estadístico y Cartográfico de Andalucía.

Proporcionar un modelo para identificar registros referidos a la misma unidad poblacional en dos o más ficheros distintos cuando no se disponga de identificadores únicos, de manera que permita:

- Limpiar y estandarizar la información contenida en el fichero.
- Localizar duplicados en un mismo fichero.
- Aumentar la cantidad de información disponible acerca de los registros incluidos en los ficheros.
- Construir o mantener actualizado el marco de una población.
- Completar la información de encuestas con datos administrativos.
- Disminuir las solicitudes de información a las personas físicas y jurídicas.
- Promover el tratamiento conjunto la información estadística y cartográfica 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.

- **Marco conceptual**

El marco de trabajo en el que se desarrolla este proyecto hace que no sea posible disponer de una terminología estandarizada, de manera que puede haber más de un concepto que haga referencia a una misma acepción.

En este contexto, se habla indistintamente de fusión de ficheros o de enlace de registros (record linkage) al referirnos al proceso de comparación de los registros de dos ficheros para intentar determinar qué pares corresponden a la misma entidad o unidad poblacional (individuo, organización, empresa…). Si lo que se pretende es enlazar dos ficheros con información común y donde uno de ellos contiene coordenadas geográficas que permiten posicionar en el territorio los registros, el proceso de fusión de ficheros se convierte en un proceso de geocodificación.

Cuando la comparación tiene lugar entre los registros del mismo fichero el proceso se denomina búsqueda de duplicados. El concepto se puede generalizar a tres o más ficheros, aunque la metodología existente suele trabajar con pares de ficheros.

En un sentido amplio, entenderemos por fichero un conjunto de datos en soporte electrónico que ofrece información relativa a una serie de atributos correspondientes a un colectivo de entidades o unidades poblacionales. La información recogida en él puede proceder de una fuente o registro administrativo o de una operación estadística (censo, encuesta…).

Se denomina registro al subconjunto de datos extraído del fichero que contiene información relativa a una entidad o unidad poblacional. Esta información puede encontrarse estructurada o no en campos o variables.

Por otro lado, se definen los conceptos de enlace o coincidencia, no enlace o no coincidencia y posible enlace:

- Un enlace o coincidencia es un par de registros que hace referencia a la misma entidad o unidad poblacional.
- Un no enlace o no coincidencia es aquel par de registros que hace referencia a entidades distintas.
- Un posible enlace es aquel par de registros del que no se tiene la seguridad de que sea un enlace o un no enlace.

Si en los ficheros de trabajo se dispone de un mismo y único identificador común entonces el problema de la fusión se reduce a una simple operación de unión, a través de dicho identificador, utilizando algún lenguaje de programación (por ejemplo, “join” en SQL). La problemática surge cuando los ficheros que se van a enlazar no comparten el mismo identificador común, en este caso, es necesario utilizar otro tipo de métodos de enlace como los determinísticos o probabilísticos. Los primeros utilizan un conjunto de reglas para llevar a cabo el enlace, las cuales son muy dependientes de los conjuntos de datos que se van a enlazar y en la práctica están limitados a conjuntos de datos pequeños;

mientras que los probabilísticos utilizan modelos estadísticos para llevar a cabo el proceso. Estos últimos se pueden subdividir a su vez en aquellos basados en la teoría probabilística clásica de enlace de registros, como la desarrollada por Fellegi y Sunter, y los enfoques más nuevos que usan técnicas de aprendizaje automatizado y de minería de datos.

En concreto, los métodos de enlace determinísticos usan conjuntos de reglas para llevar a cabo el enlace clasificando los pares de registros como enlaces y no enlaces. Estos presentan los siguientes inconvenientes:

- El conjunto de enlaces y no enlaces depende de la regla elegida.
- El conjunto de reglas utilizadas para enlazar dos ficheros va ser muy dependiente de las características de estos, con lo cual si se pretende enlazar ficheros distintos a los anteriores, el conjunto de reglas predefinido probablemente no va a servir.
- A menudo, los conjuntos de reglas son complejos.
- En la práctica estos métodos están limitados a conjuntos de datos pequeños.

Por otro lado, los métodos de enlace probabilísticos usan modelos de decisión estadísticos para clasificar los pares de registros en enlaces, no enlaces y posibles enlaces. Estos, a su vez, se pueden subdividir en aquellos basados en la teoría probabilística clásica de enlace de registros, como la desarrollada por Fellegi y Sunter, que a pesar de su relativa complejidad proporcionan numerosas ventajas comparadas con otros procedimientos ad hoc usados para este propósito y los enfoques más nuevos que usan técnicas de aprendizaje automatizado y de minería de datos, tanto para mejorar el proceso de enlace como para permitir enlazar grandes conjuntos de datos.

- **Marco jurídico**

Para la fase de normalización se usan las recomendaciones del ***Manual de buenas practicas para la normalización de fuentes y registros administrativos de la Junta de Andalucía*** *y* en el proceso de enlaces para la geocodificación de ficheros se siguen las recomendaciones de la Guia de Geocodifiación de Fuentes de Información Administrativ***a***.

La normativa aplicable es la siguiente:

- Ley 4/1989, de 12 de diciembre, de Estadística de la Comunidad Autónoma de Andalucía.
- Ley 9/2023, de 25 de septiembre, por la que se aprueba el Plan Estadístico y Cartográfico de Andalucía 2023-2029 y sus programas estadísticos y cartográficos de desarrollo.

- **Antecedentes**

A nivel andaluz, se incorporó por primera vez a la programación estadística oficial la actividad *Métodos Automáticos de Enlace de Registros* en el *Programa Estadístico Anual 2008*, en el marco del Plan Estadístico de Andalucía 2007-2012.

Por otro lado, el *Plan Estadístico y Cartográfico de Andalucía 2013-2020* 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. Por ello es esencial disponer 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. Por todo ello, el Instituto de Estadística y Cartografía de Andalucía (IECA) ha desarrollado la aplicación aLink: herramienta de fusión de ficheros.

Debe señalarse que el recurso computacional fundamental sobre el que se trabaja para este proyecto andaluz es el sistema Febrl que, por tratarse de una aplicación de código abierto, ha permitido modificar y adaptar el código fuente a las necesidades del Sistema Estadístico y Cartográfico de Andalucía.

- **Justificación y utilidad**

En el mundo en el que nos movemos la calidad de la información con la que se va a trabajar se convierte en un tema clave. El aumento de la demanda de información estadística, los recursos limitados de las oficinas estadísticas y el propósito de evitar una excesiva carga de respuesta, tanto a las personas físicas como jurídicas, hace necesario el uso eficiente así como la integración de la información proveniente tanto de censos y encuestas como de fuentes administrativas.

Es por esta razón por la que las técnicas de enlace de registros adquieren un papel relevante, ya que nos van a permitir construir o mantener actualizado un fichero maestro de una población, aumentar la cantidad de información disponible acerca de las unidades de la población y reducir la demanda de información a la ciudadanía.

- **Restricciones y alternativas**

El éxito de los resultados en la fusión de ficheros va a depender de como esté recogida la información de las variables de los ficheros. Cuanto mejor sea dicha información, los procesos de enlaces y los resultados que se obtengan serán más óptimos.

Los recursos personales disponibles en cada momento y el tamaño de los ficheros pueden ralentizar los procesos de fusión de ficheros. El uso de la herramienta requiere un conocimiento amplio de la herramienta por lo que es imprescindible contar con personal cualificado para soportar suministro técnico.

- **Comparabilidad territorial**

El enlace de registros asistido por ordenador se remonta al año 1950. En ese momento, muchos proyectos de enlace estaban basados en métodos ad hoc heurísticos. El

inconveniente que presentan estos métodos es que utilizan reglas que son dependientes de los conjuntos de datos a enlazar, por lo que puede que dichas reglas no sean aplicables a pares de registros distintos de los usados en la definición de éstas.

Las ideas básicas del enlace probabilístico de registros fueron introducidas por Newcombe y Kennedy en 1962, mientras el fundamento teórico fue proporcionado por Fellegi y Sunter en 1969, siendo el modelo matemático propuesto por estos el que básicamente subyace en todos los proyectos de enlace. En la actualidad, otros autores como Christen, Winkler o Yancey están siguiendo líneas de investigación similares a las de este proyecto.

De forma independiente, en el sector informático se han desarrollado técnicas similares en el área de indexación y recuperación de documentos. No obstante, hasta hace poco no se han encontrado muchas referencias cruzadas entre el enfoque estadístico e informático.

En España, la aplicación de métodos automáticos de fusión de registros comenzó en el Instituto Vasco de Estadística (EUSTAT) en los años 90. En primer lugar aplicaron métodos determinísticos y desde el año 2002 han ido desarrollando métodos probabilísticos. Concretamente, para llevar a cabo la fusión este organismo ha desarrollado una metodología y un programa informático en lenguaje SAS siguiendo las directrices marcadas por la metodología basada en el artículo de Fellegi-Sunter, “A theory for a record linkage”. De esta forma en 2006 disponen de una aplicación específica denominada Modulo de Fusión que flexibiliza e independiza el procedimiento de fusión de forma que acepta diferentes tipos de entradas (texto plano, Access…) y mejora la normalización de los identificadores. En 2009 disponen de una segunda versión de esta herramienta con mejoras funcionales y de rendimiento en el proceso de fusión.

A nivel europeo, el Instituto de Estadística Italiano, teniendo como apoyo metodológico un grupo de profesionales de distintos ámbitos (estadística e informática), ha desarrollado una herramienta de enlace de registros denominada RELAIS (REcord Linkage at IStat). Se trata de un proyecto de código abierto implementado usando dos lenguajes de programación JAVA y R, elegidos en línea con la filosofía de código abierto del proyecto RELAIS.

A nivel internacional, destaca la aplicación Febrl desarrollada por la Universidad Nacional de Australia. Se trata de una herramienta orientada tanto al enlace de registros como a la búsqueda de duplicados, desarrollada por Peter Christen y Tim Churches, con la característica notable de trabajar con código abierto. Está escrita en lenguaje de programación Python.

También se debe hacer referencia a la aplicación BigMatch creada por la Oficina del Censo de los Estados Unidos de América y desarrollada por William Yancey y William E. Winkler. Presenta el problema de no ser una aplicación de código abierto, por lo que no es posible acceder a la implementación de los métodos que son utilizados en el enlace de los registros, imposibilitando su adaptación a otros proyectos como el del Instituto de Estadística y Cartografía de Andalucía.

2. ÁMBITO DE ESTUDIO

- **Objeto de estudio.** Cualquier fichero de datos con información procedente de fuentes tan distintas como censos, encuestas o fuentes administrativas.

- **Resolución, escala o desagregación del objeto de estudio.** La desagregación máxima alcanzada es puntual.

- **Fenómenos o variables.** Las variables objeto de estudio en este proyecto son de diversa naturaleza en el sentido de que son múltiples los ámbitos para los que se dispone de registros, ya sean administrativos o no, que pueden ser susceptibles de enlace.

3. RECOGIDA O CAPTURA DE DATOS

- **Sujeto informante.** Instituto de Estadística y Cartografía de Andalucía, Consejerías, organismos y agencias administrativas de la Junta de Andalucía.

- **Tipología de datos a suministrar.** Los datos recogidos son de personas físicas, jurídicas y/o direcciones postales.

- **Periodicidad.** Continua.

- **Método de obtención.** Entre los objetivos de esta actividad se contempla la fusión de ficheros administrativos o registros con información para la normalización de algunas de las variables y el enlace con otros registros o fuentes administrativas. Por ello, el tipo de recogida de esta actividad se considera fuente administrativa.

  Sin embargo, no podemos dar la información exhaustiva de las fuentes o registros administrativos que se van a a utilizar en esta actividad ya que pueden ser cualquiera de los que se recogen en el Inventario de fuentes administrativas de Andalucía. La herramienta *aLink* podrá usarse con cualquier otro tipo de fichero o registro de cualquier otra naturaleza o ámbito territorial.

  En definitiva, esta actividad trabajará con las fuentes administrativas o cualquier otro tipo de fichero que se requieran dependiendo del proceso que se lleve a cabo y de las necesidades que surjan en cada momento.

4. FLUJO O PROCESO DE TRABAJO

- **Preparación y tratamiento base de la información**

La metodología bajo la que se desarrolla el proceso de fusión de ficheros llevado a cabo en el Instituto de Estadística y Cartografía de Andalucía se sintetiza básicamente en el siguiente esquema:

Figura 1: Etapas del proceso de fusión de ficheros

Previamente, se debe realizar de modo manual una revisón de los ficheros para analizar el diseño de registro y detectar campos que ambos registros tienen en común.

A continuación es imprescindible, y se debe hacer de forma obligatoria, un *Tratamiento previo* con la herramienta *aLink* a todos los ficheros de datos con los que se va a trabajar. El tratamiento previo permite transformar el fichero de trabajo en un formato compatible con la herramienta además de establecerle una codificación estándar y limpiar caracteres que puedan ser extraños. Esta tratamiento está embutido en la misma herramienta.

A continuación, se explica brevemente en que consiste cada una de estas fases descritas en la *Figura 1*:

Fase de normalización

Mucha de la información contenida en los ficheros que se pretenden enlazar contiene errores, está incompleta, se ha codificado de forma diferente de un fichero a otro, etc. Es por este motivo por lo que es necesario transformar los datos originales en otros que corrijan estas situaciones.

La fase de normalización es de suma importancia ya que su correcta ejecución ayudará a obtener mejores resultados en el proceso de enlace. Comprende las tareas de:

- *Limpieza y estandarización*. Su objetivo es transformar los datos originales brutos en otros con formatos consistentes y bien definidos, así como la resolución de inconsistencias sobre la forma en que se representa y codifica la información.
- *Segmentación*. El objetivo es separar las entidades presentes en un campo para facilitar las comparaciones. Por ejemplo, un campo que contiene una dirección postal puede ser separado en tres nuevos campos: tipo de vía, nombre de vía y número de la vía. No siempre es evidente cómo aislar la descripción clara de una dirección o un nombre. Para extraer los distintos descriptores se han empleado los Modelos Ocultos de Markov. Esta metodología parte de una muestra de registros que contienen valores del campo a normalizar y una vez analizada la estructura seguida por los elementos contenidos en la muestra se construirá propiamente el Modelo Oculto de Markov, que será el reflejo de las diferentes estructuras que siguen los elementos del campo a normalizar.

Fase de agrupación de registros

Una vez efectuada la normalización de los ficheros de datos, el principal obstáculo computacional que se presenta es el tamaño de los ficheros de datos A y B a enlazar, ya que es frecuente trabajar con bases de datos públicas con miles o incluso millones de registros. A fin de reducir el número de comparaciones a realizar, es conveniente aplicar técnicas de agrupación de registros. El objetivo de estas técnicas es reducir el número de comparaciones mediante la formación de grupos. Los grupos se forman de acuerdo a algún criterio (variables de agrupación), teniendo que ser el mismo en ambos ficheros. De esta forma los registros que se encuentran en grupos que no tengan su grupo equivalente en el otro fichero se considerarían directamente como no enlaces, aunque habría que analizarlos posteriormente puesto que podrían existir errores de normalización o bien podría haberse producido una mala elección del criterio de agrupación. Entre los métodos de agrupación analizados se han considerado dos, las técnicas de bloqueo estándar o blocking tradicional y el método de los vecinos ordenados.

Fase de comparación de pares de registros

En esta fase se parte de los grupos que se han formado anteriormente en ambos ficheros, de forma que cada uno de ellos tiene su equivalente en el otro. En este caso se comparan los registros de cada grupo con los de su grupo equivalente, de forma que para cada par de registros comparados debe obtenerse un vector de comparaciones o de pesos a partir del cual se pueda tomar la decisión final de clasificarlo como enlace o no enlace. En general, se obtienen vectores cuyas componentes resultan de la aplicación de alguna medida de similitud (funciones de comparación), y en las que el valor peso de coincidencia (habitualmente 1) corresponde a una coincidencia exacta, mientras que el valor peso de no coincidencia (en general 0) se asigna a discrepancias totales. Estos vectores tendrán tantas componentes como campos se hayan comparado. Las medidas utilizadas en esta fase permiten comparar, de forma exacta o aproximada, tanto valores numéricos como cadenas de caracteres.

Fase de clasificación

Cada par de registros comparado tiene asociado un vector de pesos calculado mediante alguna de las funciones de comparación y son los que se utilizan para clasificar los pares de registros como enlaces, no enlaces y posibles enlaces. Se distinguen dos grandes grupos de métodos de clasificación, los supervisados y los no supervisados, es decir, métodos que necesitan un conocimiento previo acerca del verdadero estado de los enlaces y los que no lo necesitan.

Debido a que en la mayoría de las situaciones no se dispone de ese conocimiento previo, el proyecto de fusión de ficheros desarrollado en el Instituto de Estadística y Cartografía de Andalucía se centró en el estudio e implementación de métodos de clasificación no supervisados.

En concreto, *aLink: Herramienta de Fusión de Ficheros* tiene implementados los siguientes métodos de clasificación:

- Clasificador basado en la metodología de Fellegi y Sunter: este método suma las componentes de los vectores de pesos obtenidos tras comparar los distintos campos de los registros. El valor obtenido se denomina peso total o de enlace y se va a comparar con dos valores umbral, establecidos por el usuario, de forma que aquellos pares cuyo peso total sea menor que el valor umbral inferior serán clasificados como no enlaces, los que estén por encima del valor umbral superior se clasificarán como enlaces y los que estén entre ambos umbrales como posibles enlaces.
- Clasificador de dos pasos (*TwoSteps*): el clasificador de dos pasos se basa en las dos siguientes hipótesis: los vectores de pesos obtenidos en la etapa de comparación que tienen valores altos en sus componentes tienen una probabilidad alta de representar a un par de registros que sea un verdadero enlace, esto es, tienen una alta probabilidad de que representen a la misma entidad, mientras que los vectores de pesos obtenidos en la etapa de comparación con valores bajos en sus componentes tienen una probabilidad alta de que representen a entidades distintas.

En este sentido, el clasificador de dos pasos se basa en la idea de construir en un primer paso dos conjuntos de entrenamiento formados por vectores de pesos, de forma que cada uno de ellos contenga un determinado número de vectores de pesos que con una alta probabilidad den lugar a enlaces y a no enlaces. La elección del número de vectores que formará parte de cada conjunto se lleva a cabo por parte del usuario, siendo uno de los métodos usados para determinar dicho valor la expresión dada por Peter Christen:

$$r = \frac{\min\left(|A|, |B|\right)}{|W| - \min\left(|A|, |B|\right)}$$

que representa una razón que estima el número de enlaces sobre el de no enlaces, donde |A| y |B| representan el tamaño de los ficheros A y B respectivamente y |W| representa el tamaño del conjunto de vectores de pesos o lo que es lo mismo el número de comparaciones realizadas en función de la variable de agrupación utilizada. Por ejemplo, si r=0.05 por cada par de registros que no sea un enlace habrá que incluir 0.05 enlaces o lo que es lo mismo por cada 100 no enlaces habrá 5 enlaces.

En un segundo paso y una vez construidos los conjuntos de entrenamiento, éstos se utilizarán para clasificar el conjunto completo de vectores de pesos. Para ello se usa alguno de los clasificadores implementados en la herramienta: máquina-vector-soporte y k-medias. El resultado de tal clasificación da lugar a dos ficheros que contienen los pares de registros enlazados y los no enlazados. En este caso el conjunto de posibles enlaces no tiene sentido ya que siempre se clasificarán los pares de registros comparados como enlaces o no enlaces.

Tras la fase de clasificación de los pares de registros comparados, el proceso de fusión se ha de centrar en analizar aquellos registros que no se han enlazado y en los que se han clasificado como posibles enlaces (nótese que este grupo solo se tendrá cuando se haya usado el clasificador basado en la metodología de Fellegi y Sunter). Si el conjunto de posibles enlaces es relativamente pequeño como para ser tratado manualmente, se analizaría el mismo y se clasificarían los pares de registros allí contenidos como enlaces o no enlaces. En caso contrario, se realizará un nuevo proceso de fusión para todos aquellos registros que no se han enlazado utilizando otras variables de agrupación, funciones de comparación y clasificadores, así como diferentes parámetros para cada uno de ellos.

El procedimiento que soportan los datos a través de *aLink: Herramienta de Fusión de Ficheros*, se puede consultar en el ***Manual de la herramienta aLink***.

Como fase final se realiza un análisis de los resultados de los ficheros de salida. Tras realizar este análisis se tiene un fichero con un enlace para cada registro y es el momento de incorporar, mediante estos enlaces entre ambos ficheros, campos del fichero B al A.

Para llevar a cabo este proceso se hará uso de la opción 'Incluir campos a enlaces' del menú Herramientas de la *Herramienta de Enlace*: En ella el usuario podrá:

- Incluir el fichero del que se extraerá la información o campos para incluir en el fichero de enlaces (Fichero 1).
- Incluir el fichero en el que se incorporará la información o campos (Fichero 2).

Por último, y con el fin de obtener nuevos enlaces, se eliminarán de los ficheros de partida, los registros considerados como enlaces para quedarnos con los registros que no han enlazado en esta primera fase y volver a repetir el proceso.

Para eliminar los registros enlazados se hace uso de la opción que nos proporciona la Herramienta de Enlace llamada *Eliminar registros enlazados.*

En el ***Manual de la herramienta aLink*** puede consultarse más profundamente el desarrollo de este apartado.

- **Garantía del secreto estadístico y protección de datos personales**

En esta actividad se trabaja con datos personales y por tanto está sometida a toda la regulación europea, nacional y autonómica sobre secreto estadístico y protección de datos, el objetivo principal de esta actividad es fusionar o enlazar ficheros con registros

referentes a una misma unidad común. Para garantizar el secreto estadístico y protección de datos se realizan una serie de actuaciones que se detallan a continuación:

Para la recepción y descarga de la información existe un protocolo que contiene, entre otras, las siguientes actuaciones:

- Los ficheros se intercambian a través de carpetas compartidas a las que solo tiene acceso el personal que trabaja con la herramienta aLink.

- Una vez recepcionados, los ficheros se almacenan en carpetas y bases de datos pgAdmin de acceso restringido y con control de las entradas, que solo permiten acceder a través de los equipos autorizados e introduciendo la correspondiente clave. Estos ficheros se cargan directamente con la herramienta aLink para la normalización y enlace.

En cuanto al tratamiento de dicha información, los mecanismos establecidos para garantizar el secreto y la protección de los datos son los siguientes:

- Los ficheros que tienen información de carácter personal se trabajan sin hacer uso de las variables que identifican a personas, es decir asigna un id que se utiliza para identificar de forma univoca a cada registro.

Y respecto a la publicación de resultados, esta actividad no difunde los ficheros fruto de la normalización e integración, sino que se remiten a la unidad responsable para el fin que tengan establecido, normalmente su aprovechamiento estadístico o cartográfico, siendo esta unidad responsable la que a partir de ese momento debe velar por la garantía del secreto estadístico y protección de datos personales.

- **Codificación, estándares, nomenclaturas y clasificaciones utilizadas**

Las variables que se van a utilizar para la fusión de ficheros, va a depender de cada proceso y de cada caso. Si, por ejemplo, se utilice los métodos de enlace de registros para la geocodificación de ficheros usando el Callejero Digital de Andalucía Unificado (CDAU), las variables relativas a la dirección postal pueden ser fundamentales para el proceso de enlaces.

- **Mantenimiento, conservación y actualización**

La aplicación *aLink: Herramienta de Fusión de Ficheros* está implementada actualmente en Python, recientemente ha sido migrada de la versión Python 2.7 a la versión Python 3.9. Dicha aplicación está sometida a procesos de mejora continua con idea de incorporar técnicas que complementen y aporten una mayor versatilidad a las herramientas de normalización y enlace.

5. PLAN DE DIFUSIÓN

- **Producto:** aLink: Herramienta de Fusión de Ficheros.
- **Tipo de resultados y formatos:** Aplicación informática.
- **Periodicidad:** continua.
- **Usuarios:** no se prevén procedimientos para evaluar la satisfacción y calidad percibida por los usuarios.

6. CALIDAD

- Respecto al **productor** de los datos:

  - *Reproducibilidad del proceso:* Se elaboran documentos que facilita el posterior mantenimiento y la transferencia del conocimiento sobre la aplicación, dichos documentos son:

    - *Manual de usuario*: cuyo objeto es describir de manera sencilla la interfaz gráfica de usuario junto con sus diferentes opciones de configuración, sin profundizar demasiado en las técnicas y algoritmos subyacentes que se han empleado.

    - *Documentación técnica*: aporta una descripción detallada del código fuente de la aplicación.

    - *Guía rapida*: el propósito de esta guía es proporcionar una visión general y directa sobre los pasos fundamentales que deben seguirse para realizar, un proceso de normalización y geocodificación de un fichero, que disponga de direcciones postales.

    - *Documento interno*: en el que se describen de forma detallado los trabajos que se realizan actualmente de forma periódica .

  - *Oportunidad*: Al tratarse de una herramienta que da soporte a otras actividades este concepto aplicará en cada una de las actividades que apoya.

  - *Puntualidad:* La aplicación se actualiza esporádicamente con alguna mejora de funcionalidades y cuando se prevé tener algún tipo de actualización previsto se cumple con dicho calendario.

  - *Disposición y disponibilidad:* La aplicación *aLink: Herramienta de Fusión de Ficheros* está disponible en la siguiente dirección web (aquí), en ella se dispone de un descargable de la aplicación junto con información adicional de gran importancia. Ademas está previsto, desarrollar unos videos tutoriales.

- Respecto a los **procesos**: En cada proceso (tanto en la normalización como en el proceso de enlaces) se realiza un análisis exhaustivo y manual de cada uno de ellos. Para garantizar que los enlaces son correctos se estudian también los enlaces duplicados para obtener sólo aquel enlace de registro de mayor exactitud.

- Respecto a los **resultados**:

  - *Relevancia y utilidad: aLink: herramienta de fusión de ficheros,* permite el aprovechamiento de las fuentes, registros e infraestructuras de información para la producción de datos estadísticos y geoespaciales, con el objetivo de favorecer la disminución de la carga de respuesta de las personas encuestadas y la mejora en el uso de los recursos públicos.

  - *Precisión y confiabilidad:* Dependiendo del fichero de origen y los parámetros utilizados para la normalización y enlace, se obtendrán los porcentaje de éxito necesarios para su fiabilidad y/o precisión.

- *Nivel de estandarización o conformidad:* La aplicación permite seleccionar los campos de salida en los que se quiere segmentar la normalización, basándose en el *Manual de buenas prácticas para la normalización de fuentes y registros administrativos de la Junta de Andalucía.*

- *Esquema de calidad:* No se sigue ningún estándar.