aLink: Herramienta de Fusión de Ficheros

Manual de Usuario

Versión: 0300

Fecha: 29/04/2021

Versión 01.0.3

Queda prohibido cualquier tipo de explotación y, en particular, la reproducción, distribución, comunicación pública y/o transformación, total o parcial, por cualquier medio, de este documento sin el previo consentimiento expreso y por escrito de la Junta de Andalucía.

HOJA DE CONTROL

Organismo	Instituto de Estadística y Cartografía de Andalucía		
Proyecto	aLink: Herramienta de Fusión de Ficheros		
Entregable	Manual de Usuario		
Autor	Elisa Isabel Caballero Ruíz		
Versión/Edición	0300	**Fecha Versión**	13/05/2014
Aprobado por		**Fecha Aprobación**	13/05/2014
		Nº Total de Páginas	177

REGISTRO DE CAMBIOS

Versión	Causa del Cambio	Responsable del Cambio	Fecha del Cambio
0100	Versión inicial	Elisa Isabel Caballero Ruíz	13/05/2014
0200	Versión modificada	Ana Gema Galera Pozo	22/12/2015
0300	Versión modificada	Elisa Isabel Caballero Ruíz	29/04/2021

CONTROL DE DISTRIBUCIÓN

Nombre y Apellidos
Elisa Isabel Caballero Ruíz

ÍNDICE

1 Introducción....6

2 Proceso de fusión de ficheros con aLink: Herramienta de Fusión de Ficheros....8

3 Formato de los ficheros de trabajo....10

4 Instalación....11

4.1 Entorno Windows: requerimientos de software incluidos en el paquete de instalación....11

4.2 Entorno Windows: instalación, ejecución y desinstalación....11

4.2.1 Instalación en Windows....11

4.2.2 Ejecución en Windows....13

4.2.3 Desinstalación en Windows....13

4.3 Entorno Linux: requerimientos....14

4.4 Entorno Linux: instalación y ejecución....14

4.4.1 Instalación en Debian Wheezy....14

4.4.2 Ejecución en Debian Wheezy....16

4.4.3 Instalación en Ubuntu 12.04 LTS....17

4.4.4 Ejecución en Ubuntu 12.04 LTS....19

5 Proceso de normalización de un fichero de datos....20

6 Herramienta de Normalización....24

6.1 Nuevas funcionalidades de la Herramienta de Normalización respecto a *ADYN: Herramienta de Normalización v 2.0*....24

6.2 Descripción general de la Herramienta de Normalización....26

6.2.1 Barra de menú de la Herramienta de Normalización....26

6.2.2 Barra de herramientas de la Herramienta de Normalización....27

6.2.3 Área de normalización....28

6.3 Menú Herramientas de la Herramienta de Normalización....37

6.3.1 Tratamiento previo....37

6.3.1.1 Tratamiento de un fichero CSV....43

6.3.1.2 Tratamiento de un fichero TAB....47

6.3.1.3 Tratamiento de un fichero PLANO....47

6.3.1.4 Tratamiento de un fichero EXCEL....51

6.3.1.5 *Tratamiento* de un fichero MySQL....52

6.3.1.6 Tratamiento de un fichero ACCESS....54

6.3.1.7 Tratamiento de un fichero ODS....55

6.3.1.8 Tratamiento de un fichero DBF....55

6.3.2 HMM: Selección de la muestra....57

6.3.3 HMM: Entrenamiento de la muestra....72

6.3.4 Editor de listas de corrección....78

6.3.5 Editor de tablas de búsqueda....85

6.4 Validación del proceso de normalización....92

7 Herramienta de Enlace....100

7.1 Descripción general de la Herramienta de Enlace....100

7.1.1 Barra de menú de la Herramienta de Enlace....101

7.1.2 Barra de herramientas de la Herramienta de Enlace....102
7.1.3 Área de enlace....103
7.1.3.1 Pestaña Ficheros de entrada....104
7.1.3.2 Pestaña Resumen del proceso....105
7.1.3.3 Pestaña Análisis exploratorio....106
7.1.3.4 Pestaña Agrupación....108
7.1.3.5 Pestaña Comparación....113
7.1.3.6 Pestaña Clasificación....117
7.1.3.7 Pestaña Salida....122
7.1.3.8 Pestaña Evaluación....125
7.1.3.9 Pestaña Resultados....126
7.2 Menú Herramientas de la Herramienta de Enlace....127
7.2.1 Tratamiento previo....127
7.2.2 Insertar índices....127
7.2.3 Incluir campos a enlaces....128
7.2.4 Eliminar registros enlazados....129
8 FAQ....132
9 ANEXOS....139
Anexo I: Instalación manual de *aLink: Herramienta de Fusión de Ficheros* en un entorno Windows....139
Anexo II: Modelos Ocultos de Markov disponibles en *aLink: Herramienta de Fusión de Ficheros*....148
Anexo III: Campos de salida para nombres de personas e identificadores de personas físicas y/o jurídicas....150
Anexo IV: Campos de salida del fichero normalizado....151
Anexo V: Etiquetas usadas en el proceso de normalización para construir un modelo HMM....154
Anexo VI: Usar HMM anterior....157
Anexo VII: Estados usados en el proceso de normalización para construir un modelo HMM....161
Anexo VIII: Métodos de suavizado....164
Anexo IX: Listas de corrección....166
Anexo X: Tablas de búsqueda....167
Anexo XI: Métodos de agrupación y proceso "full index"....169
Anexo XII: Funciones de comparación....171
Anexo XIII: Métodos de clasificación....176
10 GLOSARIO....177
11 BIBLIOGRAFÍA....178

1 Introducción

El Instituto de Estadística y Cartografía de Andalucía (IECA) dispone de una gran cantidad de información procedente de diversas fuentes, tan distintas como censos, encuestas o fuentes administrativas. Para poder llevar a cabo un aprovechamiento estadístico y cartográfico exhaustivo de la información, así como reducir costes y esfuerzos en la adquisición de la misma, lo ideal sería poder integrar la información procedente de las distintas fuentes.

Así se pretende dar respuesta a uno de los objetivos generales de la Ley del Plan Estadístico y Cartográfico de Andalucía 2013-2017, aprobada el 17 de julio de 2013, que es el de aprovechar el potencial que genera la integración de la información estadística y cartográfica para contribuir al desarrollo de la sociedad del conocimiento.

En este sentido, el IECA ha desarrollado una aplicación informática libre y gratuita basada en FEBRL (desarrollo de software libre de la Universidad Nacional de Australia) y liberada bajo la licencia ANOUS (Australian National University Open Source License) versión 1.3. Dicha aplicación se denomina ***aLink: Herramienta de Fusión de Ficheros*** y ha sido diseñada y desarrollada para combinar una serie de técnicas en distintas etapas, las cuales van a permitir llevar a cabo un proceso de fusión de ficheros completo con grandes volúmenes de datos. A grandes rasgos, *aLink: Herramienta de Fusión de Ficheros* consta de dos herramientas, una de normalización y otra de enlace. La Herramienta de Normalización constituye la última versión de ***ADYN: Herramienta de Normalización***, por lo que a partir de este momento ésta última dejará de distribuirse de forma independiente y lo hará integrada como parte de *aLink: Herramienta de Fusión de Ficheros*, mientras que la Herramienta de Enlace como su nombre indica permitirá enlazar dos ficheros de datos.

El objetivo de un proceso de fusión de ficheros es detectar aquellos registros de dos ficheros de datos A y B 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. De esta forma no solo se mejorará la integridad y calidad de los datos sino que se podrá enriquecer la información de la que se dispone para la actualización de estudios anteriores o la realización de nuevos.

Para poder realizar de forma eficiente un proceso de fusión de ficheros se presenta este **Manual de Usuario** de la aplicación *aLink: Herramienta de Fusión de Ficheros*, 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.

La estructura de este manual es la que sigue:

- Proceso de fusión de ficheros con *aLink: Herramienta de Fusión de Ficheros.* Descripción de la metodología en la que se basa este proceso: normalización y enlace.
- Formato de los ficheros de trabajo. Esta sección analiza los distintos formatos de ficheros con los que se puede trabajar en *aLink: Herramienta de Fusión de Ficheros.*
- Instalación. Muestra los detalles de instalación de *aLink: Herramienta de Fusión de Ficheros*, tanto en un entorno Windows como en un entorno Linux.
- Herramienta de Normalización. Descripción de cómo llevar a cabo la normalización de un fichero de datos, así como de la

interfaz gráfica que permite llevar a cabo dicho proceso.

- Herramienta de Enlace. Descripción de cómo llevar a cabo el enlace de dos ficheros de datos, así como de la interfaz gráfica que permite llevar a cabo dicho proceso.
- FAQ.
- Anexos.
- Glosario de términos.
- Bibliografía.

2 Proceso de fusión de ficheros con aLink: Herramienta de Fusión de Ficheros

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 en el siguiente esquema:

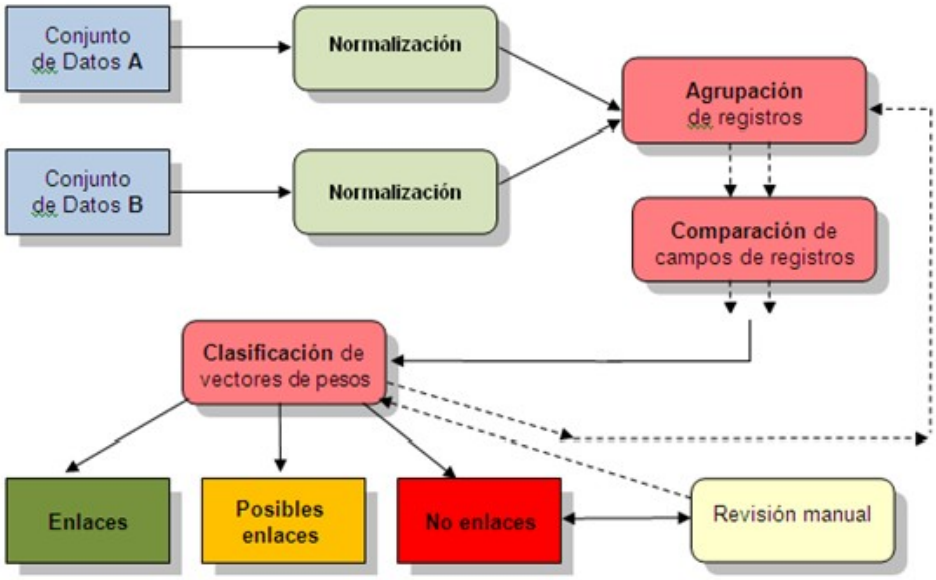


Imagen 0. Etapas del proceso de fusión de ficheros

A continuación, se explica brevemente en qué consiste cada una de estas fases o etapas.

Fase de normalización

Mucha de la información contenida en los ficheros que se pretenden enlazar contiene errores, está incompleta, se codifica 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, que 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 el nombre y apellidos puede ser separado en tres nuevos campos: nombre, primer apellido y segundo apellido. No siempre es evidente determinar los elementos en los que se segmenta una dirección o un nombre. Para extraer los distintos elementos se han empleado *Modelos Ocultos de Markov.* Esta metodología parte de una muestra de registros del fichero de datos que contiene valores del campo a normalizar y una vez analizada la estructura seguida por los elementos contenidos en la muestra se construye el Modelo Oculto de Markov, que servirá para normalizar el fichero de datos completo.

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 a enlazar, ya que es frecuente trabajar con bases de datos públicas que contienen 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 (BlockingIndex)* y el *método de los vecinos ordenados (SortingIndex)*.

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, no enlace o posible 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 (en general 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. Además, las funciones de comparación utilizadas en esta fase permiten comparar, tanto de forma exacta como aproximada, valores numéricos y cadenas de caracteres.

Fase de clasificación

Cada par de registros comparados 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 ha centrado en el estudio e implementación de métodos de clasificación no supervisados. En concreto, la aplicación dispone de los siguientes métodos: clasificador basado en la metodología de Fellegi y Sunter (*FellegiSunter*) y clasificador de dos pasos (*TwoSteps*).

3 Formato de los ficheros de trabajo

Para normalizar o enlazar dos ficheros de datos con la aplicación informática *aLink: Herramienta de Fusión de Ficheros*, se requiere que los ficheros de trabajo tengan formato CSV y sus elementos estén separados por el carácter ";". Dado que esta situación no va a ser la habitual en la mayoría de entornos de trabajo, en la aplicación *aLink: Herramienta de Fusión de Ficheros* se dispone de una herramienta auxiliar que permite transformar al formato requerido (ficheros CSV separados por ";") los ficheros que se encuentran en los siguientes formatos:

- CSV
- TAB
- PLANO
- EXCEL
- MySQL
- ACCESS
- ODS
- DBF

Respecto a alguno de estos formatos existen restricciones que se analizarán más detenidamente en el apartado 6.3.1 de este Manual.

Con esta herramienta, además de transformar los ficheros originales a ficheros de texto CSV separados por ";" se realiza automáticamente una recodificación de los datos así como la eliminación o sustitución de símbolos, elementos o caracteres que por su codificación pueden provocar fallos en un posterior proceso de normalización o enlace.

4 Instalación

En este apartado se presenta de forma detallada cómo llevar a cabo el proceso de instalación de la aplicación *aLink: Herramienta de Fusión de Ficheros* tanto en un entorno Windows como en Linux.

4.1 Entorno Windows: requerimientos de software incluidos en el paquete de instalación

La instalación de *aLink: Herramienta de Fusión de Ficheros* en un entorno Windows requiere tener instalados previamente los requisitos de software que se indican abajo. Todos ellos se encuentran incluidos en el paquete de instalación completo y se instalan al ejecutar el mismo.

- **Python-XY** (distribución científica de Python)
- **PyGTK All-in-one** (librerías GTK Windows para Python)
- **Python-matplotlib** (módulo de generación de gráficos para Python)
- **MySQL-Python** (módulo de bases de datos para Python)
- **Gawk** (lenguaje AWK para Windows)
- **Coreutils** (colección de utilidades GNU para Windows)
- **Xlrd-0.9.2-3_py27** (complemento de Python(x,y) que permite al usuario trabajar con ficheros MSExcel 2007)
- **Pyodbc-3.0.7** (programa que permite realizar conexiones, mediante ODBC con diferentes formatos de bases de datos)
- **Dbfpy-2.2.5.win32** (programa permite realizar conexiones con bases de datos DBF)
- **Notepad2** (editor de texto mejorado)

No obstante, si el usuario ya tuviera instalados tales requisitos o algunos de ellos, podría proceder a la instalación manual de los restantes y descargar únicamente el código fuente de la aplicación. En el Anexo I se indica cómo llevar a cabo la instalación manual de los mismos, así como las urls de descarga.

4.2 Entorno Windows: instalación, ejecución y desinstalación

Antes de instalar *aLink: Herramienta de Fusión de Ficheros* el usuario debe comprobar si tiene instalada alguna versión de Python. En caso de que la tenga y sea distinta de la versión Python 2.7.2 deberá desinstalarla y seguir con el proceso de instalación de *aLink: Herramienta de Fusión de Ficheros* que se detalla a continuación.

4.2.1 Instalación en Windows

Para instalar *aLink: Herramienta de Fusión de Ficheros*, el usuario tiene que hacer doble click sobre el ejecutable que se descargará al solicitar la aplicación a través de la sección de *Servicios*, *Descarga de Software* de la página web del Instituto de Estadística y Cartografía de Andalucía.

El ejecutable es un paquete msi que contiene la información necesaria para automatizar la instalación de *aLink* junto con todos los requisitos de software indicados anteriormente y que son necesarios para que funcione correctamente en un entorno Windows. No obstante, en caso de que se produjera algún error al ejecutar dicho instalador el usuario podría llevar a cabo la instalación de los prerequisitos tal y como se indica en el Anexo I.

Una vez finalizado el proceso de instalación, al usuario le aparecerá una ventana con el siguiente mensaje: "*Debe reiniciar el sistema para que los cambios de configuración efectuados surtan efecto. Haga click en Sí para reiniciar el sistema ahora o elija No si tiene previsto reiniciarlo manualmente más tarde*".

Tras reiniciar el sistema el usuario deberá tener en la unidad C:\ una carpeta denominada **alink** con la siguiente información:

- *app*: carpeta que contiene el código fuente de *aLink: Herramienta de Fusión de Ficheros*.
- *listas_tablas*: carpeta que contiene una serie de ficheros necesarios para llevar a cabo un proceso de normalización de un fichero de datos con aLink. El contenido de la misma se describe en este Manual más adelante.
- *muestras_modelos*: al igual que la carpeta listas_tablas, contiene otra serie de ficheros necesarios para llevar a cabo un proceso de normalización de un fichero de datos. Como en el caso anterior, más adelante se analizará su contenido.
- *manuales*: carpeta que contiene el Manual de Usuario de *aLink: Herramienta de Fusión de Ficheros* y una guía de estilo de Python.
- El archivo *aLink.bat* que permite abrir la aplicación *aLink: Herramienta de Fusión de Ficheros* sin más que hacer doble click sobre el.
- Un acceso directo y el icono de la aplicación *aLink: Herramienta de Fusión de Ficheros*.

Además, en el Escritorio aparecerá un acceso directo a la aplicación *aLink: Herramienta de Fusión de Ficheros*, así como en el menú de Inicio.

Nota: Si el instalador descargado desde la Sección de Descarga de Software de la página web del IECA es el referido al Sistema Operativo Windows 10, puede que aparezcan los siguientes mensajes de aviso durante el proceso de instalación del mismo:

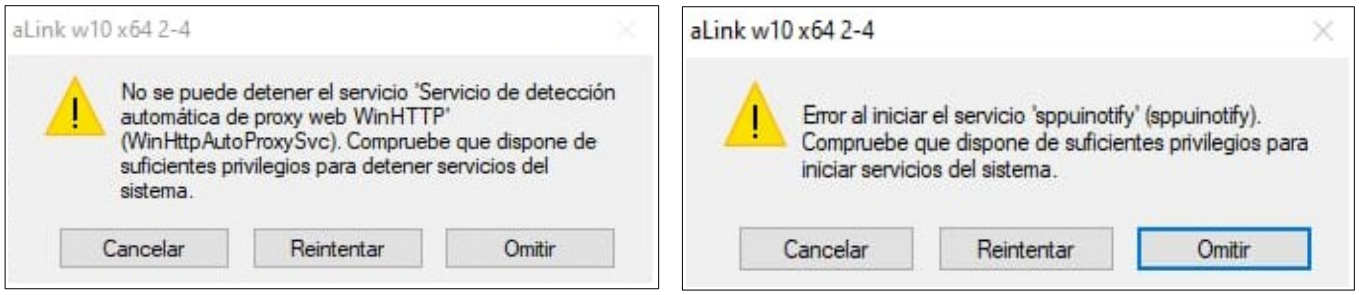


Imagen 1. Mensajes aviso instalación aLink Windows 10

En este caso lo que hay que hacer es *Omitir* ambos avisos y continuar con la instalación.

4.2.2 Ejecución en Windows

Para ejecutar la aplicación *aLink: Herramienta de Fusión de Ficheros* en Windows el usuario tendrá que ir al Escritorio y hacer doble click en el acceso directo que se genera al instalar la aplicación o también podrá hacerlo desde el Menú Inicio de Windows. De esta forma se abrirá la aplicación tal y como se muestra en la siguiente imagen:

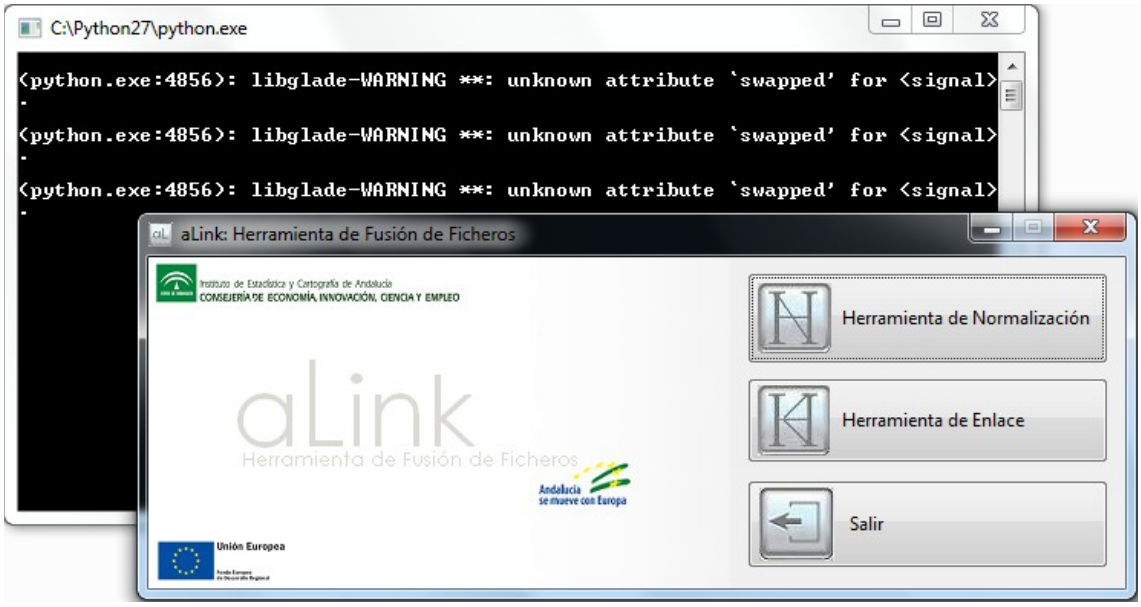


Imagen 2. Interfaz gráfica inicial de *aLink: Herramienta de Fusión de Ficheros* para Windows

4.2.3 Desinstalación en Windows

Para desinstalar la aplicación *aLink: Herramienta de Fusión de Ficheros* en Windows el usuario tendrá que utilizar la opción desinstalar programas del Panel de Control de Windows. Por ejemplo, si aLink: Herramienta de Fusión de Ficheros se ha instalado en un equipo con sistema operativo Windows 7 a 64 bits, se buscaría "aLink w7 x64" en el listado de programas y se desinstalaría. Obsérvese que en caso de haber realizado la instalación manual del software necesario para que *aLink* funcione correctamente en un entorno Windows, habría que ir desinstalando uno a uno los siguientes programas:

- **GnuWin32: Coreutils version 5.3.0**
- **GnuWin32: Gawk-3.1.6-1**
- **Python 2.7 matplotlib-1.1.0**
- **Python MySQL-python-1.2.3**
- **Python 2.7 PyGTK 2.24.1**
- **Python 2.7.2**
- **Python (x,y)**

4.3 Entorno Linux: requerimientos

A continuación, se describen los requisitos que se necesitan para instalar *aLink: Herramienta de Fusión de Ficheros* en los entornos Linux: **Debian Wheezy** y **Ubuntu 12.04 LTS**.

Como en el caso de Windows, para el correcto funcionamiento de la aplicación es necesario tener instalados previamente el siguiente software:

- **Python 2.7**
- **matplotlib** (módulo de generación de gráficos para Python)
- **Pytables** (módulo que permite manejar grandes conjuntos de datos)
- **xlrd** (complemento de Python que permite al usuario trabajar con ficheros MSExcel 2007)
- **pyodbc** (programa que permite realizar conexiones, mediante ODBC con diferentes formatos de bases de datos)
- **pymysqldb** (módulo de bases de datos para Python)
- **dbfpy** (programa que permite realizar conexiones con bases de datos DBF)
- **glade-gtk2** (módulo gráfico, SOLO SE REQUIERE PARA EL SISTEMA OPERATIVO UBUNTU 12.04 LTS)
- **myunity** (módulo para poder configurar el entorno de escritorio de este sistema operativo, SOLO SE REQUIERE PARA EL SISTEMA OPERATIVO UBUNTU 12.04 LTS)

4.4 Entorno Linux: instalación y ejecución

4.4.1 Instalación en Debian Wheezy

Para realizar la instalación de *aLink: Herramienta de Fusión de Ficheros* en este sistema operativo se deberán seguir los siguientes pasos:

0. Descomprimir archivo con la aplicación

El usuario deberá descomprimir el archivo que se descargará al solicitar la aplicación a través de la sección de Servicios, Descarga de Software de la página web del Instituto de Estadística y Cartografía de Andalucía. Al descomprimirlo, tendrá que copiar la carpeta "alink" en la ubicación que desee. Esta carpeta contiene la siguiente información:

- *app*: carpeta que contiene el código fuente de *aLink: Herramienta de Fusión de Ficheros*.
- *listas_tablas*: carpeta que contiene una serie de ficheros necesarios para llevar a cabo un proceso de normalización de un fichero de datos con aLink. El contenido de la misma se describe en este Manual más adelante.
- *muestras_modelos*: al igual que la carpeta listas_tablas, contiene otra serie de ficheros necesarios para llevar a cabo un proceso de normalización de un fichero de datos. Como en el caso anterior, más adelante se analizará su contenido.
- *manuales*: carpeta que contiene el Manual de Usuario de *aLink: Herramienta de Fusión de Ficheros* y una guía de estilo de Python.

1. Actualización del sistema

Abrir una terminal y escribir en la misma: #apt-get update. El objetivo es actualizar el sistema. A continuación, se instalarán una serie de programas necesarios para el correcto funcionamiento de aLink en Debian Wheezy.

2. Instalación de matplotlib

Escribir en la terminal: #apt-get install python-matplotlib

3. Instalación de Pytables

Escribir en la terminal: #apt-get install python-tables

4. Instalación de xlrd

Escribir en la terminal: #apt-get install python-xlrd

5. Instalación de pyodbc

Escribir en la terminal: #apt-get install python-pyodbc

6. Instalacion de pymysqldb.

Escribir en la terminal: #apt-get install python-mysqldb

7. Instalación de dbfpy

Esta herramienta no se encuentra en el repositorio de Wheezy, así pues habrá que situarse en la web del proveedor y descargarla a través del siguiente enlace:

http://sourceforge.net/projects/dbfpy/files/dbfpy/

En el mismo aparecerá un listado de carpetas (folders) con distintas versiones de la herramienta. De entre ellas el usuario elegirá la última estable, que en este caso es 2.2.5. Accederá a dicha carpeta y descargará el archivo "dbfpy-2.2.5.tar.gz". A continuación, el usuario descomprimirá dicho archivo mediante la orden:

#tar -xvzf dbfpy-2.2.5.tar.gz

Una vez descomprimido, entrará en el directorio creado y situado en el mismo escribirá en la terminal las siguientes órdenes para finalizar con la instalación:

#python setup.py build

#python setup.py install

8. Configuración visual de *aLink: Herramienta de Fusión de Ficheros*

Con el fin de que la visualización de la Herramienta sea similar a la del presente Manual se recomienda cambiar la apariencia de Debian Wheezy de acuerdo a lo siguiente:

Themes: Clearlooks

Default fonts: Sans 8

Para ello el usuario se dirigirá al menú 'Preferencias'-> 'Personalizar apariencia' y seleccionará los parámetros anteriores.

Finalmente, el usuario podrá modificar el tamaño de la fuente con el fin de que la Herramienta de Enlace presente la siguiente apariencia:

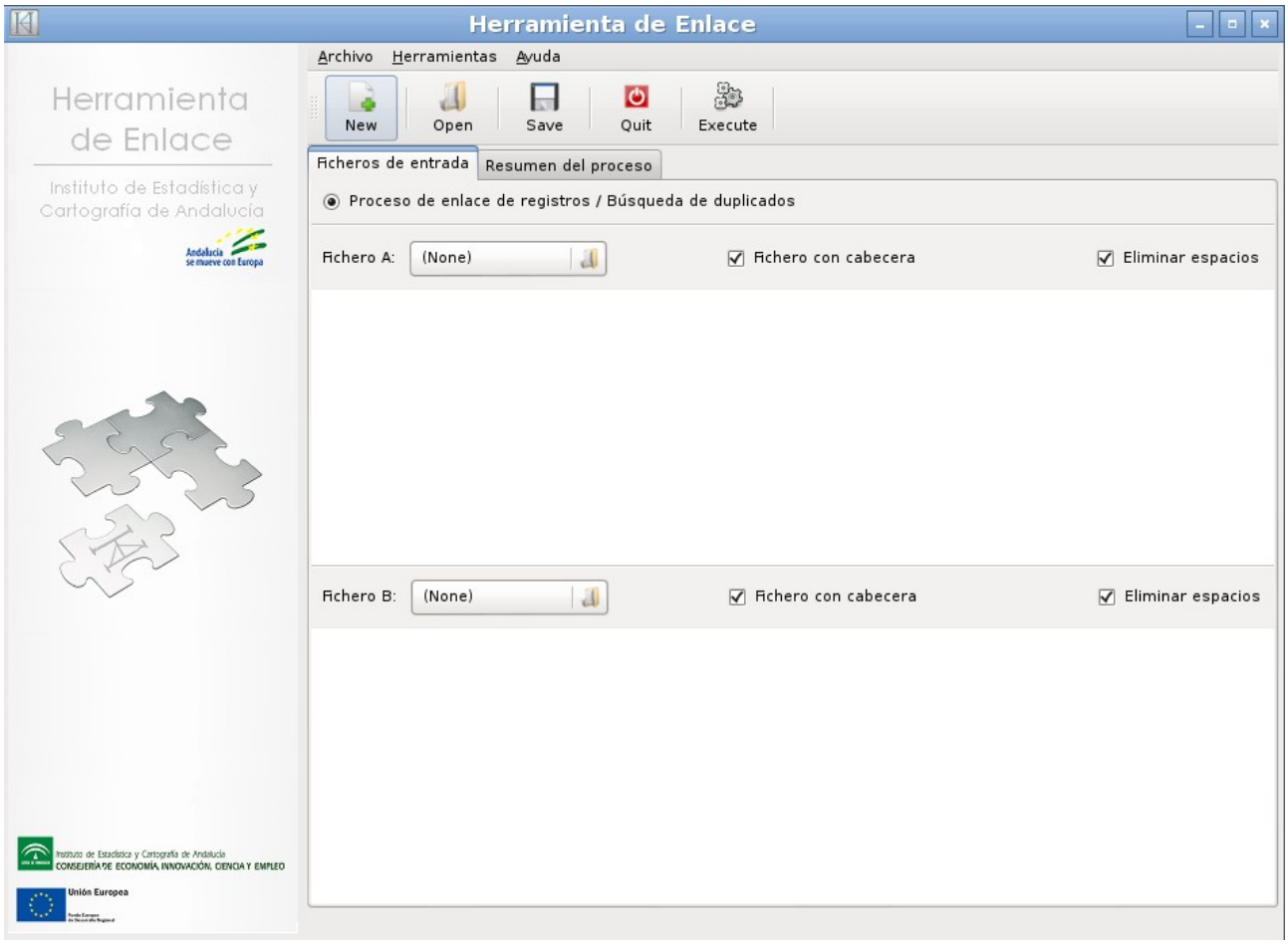


Imagen 3. Apariencia de la interfaz de aLink en Debian Wheezy

Obsérvese que en este caso no ha sido necesario instalar Python 2.7, ya que esta distribución de Linux lo tiene preinstalado por defecto.

4.4.2 Ejecución en Debian Wheezy

Una vez instalados todos los elementos necesarios se procederá a la ejecución de la aplicación. Para ello se accederá a través de la terminal al directorio **app** de *aLink: Herramienta de Fusión de Ficheros*, que se encontrará dentro de la carpeta alink, donde el usuario la haya guardado. Una vez dentro de la carpeta app habrá que escribir en la terminal la orden: # python inicio.py

De esta forma se abrirá la ventana inicial de *aLink: Herramienta de Fusión de Ficheros*, cuya apariencia es prácticamente igual a la de la Imagen 2 del entorno Windows.

4.4.3 Instalación en Ubuntu 12.04 LTS

Para realizar la instalación de *aLink: Herramienta de Fusión de Ficheros* en este sistema operativo se deberán seguir los siguientes pasos:

0. Descomprimir archivo con la aplicación

El usuario deberá descomprimir el archivo que se descargará al solicitar la aplicación a través de la sección de Servicios, Descarga de Software de la página web del Instituto de Estadística y Cartografía de Andalucía. Al descomprimirlo, tendrá que copiar la carpeta "alink" en la ubicación que desee. Esta carpeta contiene la siguiente información:

- *app*: carpeta que contiene el código fuente de *aLink: Herramienta de Fusión de Ficheros*.
- *listas_tablas*: carpeta que contiene una serie de ficheros necesarios para llevar a cabo un proceso de normalización de un fichero de datos con aLink. El contenido de la misma se describe en este Manual más adelante.
- *muestras_modelos*: al igual que la carpeta listas_tablas, contiene otra serie de ficheros necesarios para llevar a cabo un proceso de normalización de un fichero de datos. Como en el caso anterior, más adelante se analizará su contenido.
- *manuales*: carpeta que contiene el Manual de Usuario de *aLink: Herramienta de Fusión de Ficheros* y una guía de estilo de Python.

1. Actualización del sistema

Abrir una terminal y escribir en la misma: #apt-get update. El objetivo es actualizar el sistema. A continuación, se instalarán una serie de programas necesarios para el correcto funcionamiento de aLink en Ubuntu 12.04 LTS.

1. Actualización del sistema

Escribir en la terminal: #apt-get update

2. Instalación de matplotlib

Escribir en la terminal: #apt-get install python-matplotlib

3. Instalación de Pytables

Escribir en la terminal: #apt-get install python-tables

4. Instalación de xlrd

Escribir en la terminal: #apt-get install python-xlrd

5. Instalación de pyodbc.

Escribir en la terminal: #apt-get install python-pyodbc

6. Instalacion de pymysqldb

Escribir en la terminal: #apt-get install python-mysqldb

7. Instalación de dbfpy

Esta herramienta no se encuentra en el repositorio de UBUNTU 12.04 LTS, así pues habrá que situarse en la web del proveedor y descargarla a través del siguiente enlace:

http://sourceforge.net/projects/dbfpy/files/dbfpy/

En el mismo aparecerá un listado de carpetas (folders) con distintas versiones de la herramienta. De entre ellas el usuario elegirá la última estable, que en este caso es 2.2.5. Accederá a dicha carpeta y descargará el archivo "dbfpy-2.2.5.tar.gz". A continuación, el usuario descomprimirá dicho archivo mediante la orden:

#tar -xvzf dbfpy-2.2.5.tar.gz

Una vez descomprimido, entrará en el directorio creado y situado en el mismo escribirá en la terminal las siguientes órdenes para finalizar con la instalación:

#python setup.py build

#python setup.py install

8. Instalación de glade-gtk2

Escribir en la terminal: #apt-get install glade-gtk2

9. Configuración visual de *aLink: Herramienta de Fusión de Ficheros*

Con el fin de que la visualización de la Herramienta sea similar a la del presente Manual habría que instalar en Ubuntu 12.04 LTS el paquete *myunity* de la siguiente forma:

#apt-get install myunity

Y una vez instalado y abierto, se recomienda cambiar la apariencia de acuerdo a los siguientes parámetros:

Themes: Clearlooks

Font: System → Sans 6

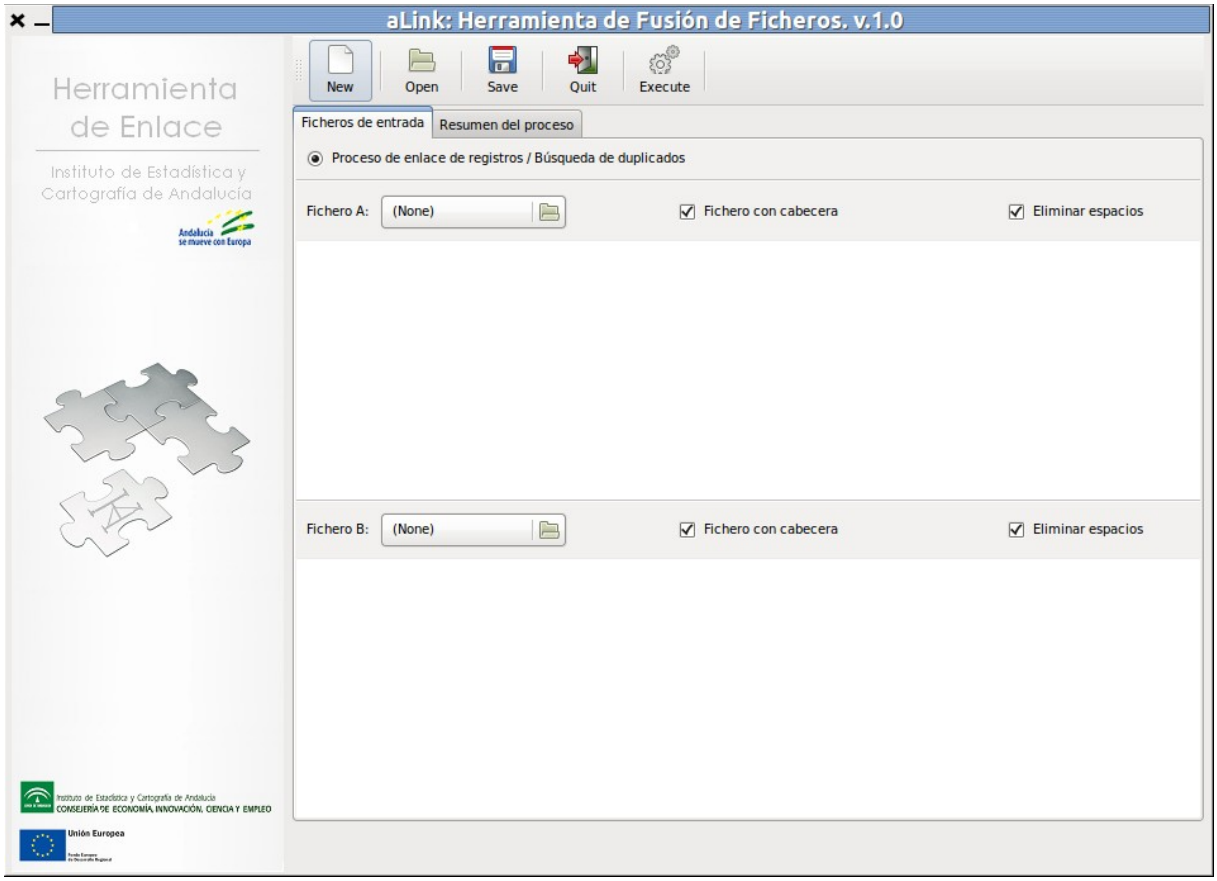


Imagen 4. Apariencia de la interfaz de aLink en Ubuntu 12.04 LTS

Al igual que en el entorno Debian Wheezy, se puede observar que para este entorno no ha sido necesario instalar Python 2.7, ya que como en el caso anterior esta distribución de Linux lo tiene preinstalado por defecto.

4.4.4 Ejecución en Ubuntu 12.04 LTS

Se realiza de la misma forma que en Debian Wheezy.

5 Proceso de normalización de un fichero de datos

Normalmente, la mayoría de la información con la que se trabaja contiene errores, está incompleta o incorrectamente formateada, se codifica de manera distinta de una fuente a otra, etc. Es por ello por lo que es necesario dar solución a esta situación.

El conjunto de técnicas encaminadas a la obtención de datos consistentes se engloban en el llamado proceso de normalización de datos y redundará en una mejor calidad y fiabilidad en posteriores análisis de dichos datos.

En el proceso de normalización, llevado a cabo con la Herramienta de Normalización, se establecen dos fases principales:

- Limpieza y estandarización. Su objetivo es transformar los datos originales brutos en otros con formatos consistentes y bien definidos, así como resolver las inconsistencias sobre la forma en que se representa y codifica la información. En el proceso de limpieza no importa el contenido semántico del campo del fichero de datos a normalizar, y se realizan tareas de codificación del mismo, así como de eliminación de abreviaturas y signos de puntuación. Mientras que en el proceso de estandarización, sí que se analiza el contenido semántico del campo, modificando algunos de sus valores por valores normalizados y etiquetando o clasificando el contenido de este según el valor de sus componentes.
- Segmentación. El objetivo de esta fase es separar las entidades presentes en el campo a normalizar para, posteriormente en un proceso de integración o de enlace de la información, facilitar las comparaciones. Por ejemplo, un campo que contiene una dirección postal puede ser separado en tres nuevos campos que contengan el tipo de vía, el nombre de la vía y el número de la vía.

El objetivo de la Herramienta de Normalización es la limpieza, estandarización y segmentación de los siguientes campos de un fichero de datos:

- Nombres de personas
- Direcciones postales
- Identificadores de personas físicas y/o jurídicas (NIF, DNI y NIE)

Hay que indicar que la aplicación no permite la normalización conjunta de dichos campos, es decir, no se podrán normalizar los tres campos a la vez. Por tanto, si se desean normalizar todos ellos, habrá que realizar la normalización de cada campo por separado. Tampoco es posible normalizar a la vez dos o más campos del mismo fichero que contengan direcciones postales, o nombres de personas o DNIs, NIFs, etc.

Para llevar a cabo la normalización de cualquiera de estos campos se utilizan tres herramientas:

- Las **listas de corrección**: son ficheros que permiten limpiar el fichero de datos a normalizar, es decir, contienen los caracteres o cadenas que el usuario considera oportuno eliminar o sustituir en el fichero. Por ejemplo, con ellas se pueden sustituir caracteres del tipo '|', '$', etc. por espacios en blanco o sustituir vocales con tildes por vocales sin tildes.
- Las **tablas de búsqueda**: son ficheros que sustituyen cada elemento del campo a normalizar por su valor estandarizado y, además, le asignan una etiqueta. Por ejemplo, si se está normalizando el campo “dirección postal” y en el mismo se

encuentra el elemento 'c/', éste se sustituye por 'calle' y se le asigna la etiqueta 'TV' que significa Tipo de Vía.

- Los **Modelos Ocultos de Markov (modelos HMM)**: son ficheros que tratan de reconocer el patrón o estructura que con más probabilidad siguen los datos del campo a normalizar, permitiendo segmentar dichos datos. Esta metodología se ha utilizado ya que no siempre es evidente cómo aislar la descripción clara de una dirección o un nombre. En concreto, esta técnica parte de una muestra de registros del fichero de datos que contiene el campo a normalizar y analiza los valores de dicho campo, intentando detectar el patrón o estructura que siguen los datos. Una vez analizados y adquirido dicho conocimiento, este queda recogido en el Modelo Oculto de Markov, que servirá para normalizar el fichero de datos completo describiendo el proceso que con más probabilidad ha generado los datos y segmentando a estos en los distintos elementos que lo componen.

  El proceso de segmentación también se puede realizar mediante técnicas basadas en reglas en lugar de usar Modelos Ocultos de Markov. Sin embargo, y a pesar de que para nombres de personas e identificadores de personas físicas o jurídicas los resultados de utilizar estas técnicas ofrecen resultados similares a los de aplicar técnicas basadas en Modelos Ocultos de Markov, para direcciones postales debido a la gran complejidad que muestran estos datos, sería muy complicado disponer de reglas que abarcasen la mayoría de los casos a segmentar. De esa forma el usuario tendría que crear una nueva regla conforme apareciera un caso nuevo que no haya sido tratado anteriormente. Así, usando los Modelos Ocultos de Markov para direcciones postales, el proceso se simplifica ya que a través de una muestra del campo a normalizar y estableciendo las estructuras y patrones de esos datos se puede estandarizar y segmentar todo el fichero.

La Herramienta de Normalización incluye de partida un conjunto de listas de corrección y tablas de búsqueda para nombres de personas, direcciones postales e identificadores de personas físicas y/o jurídicas. También dispone de algunos modelos HMM para nombres de personas y direcciones postales, que se han creado al normalizar algunos ficheros con los que se ha trabajado en el Instituto de Estadística y Cartografía de Andalucía. Para el caso de las direcciones postales estos modelos permiten llevar a cabo tanto la desagregación de la dirección postal de acuerdo al modelo de datos del Callejero Digital de Andalucía Unificado, como una desagregación a medida. No obstante, si el usuario comprueba que los modelos HMM proporcionados en la aplicación no responden a la estructura de los datos de su fichero de trabajo, podrá crear su propio modelo HMM.

En secciones posteriores de este Manual se mostrará cómo las listas de corrección, las tablas de búsqueda y los modelos HMM podrán ser editados o creados en su totalidad, de tal forma que el usuario podrá ir enriqueciendo y personalizando la información contenida en ellos.

A continuación, se muestra el esquema general de la metodología bajo la que se desarrolla un proceso de normalización de un fichero de datos:

Imagen 5. Proceso general de normalización

Como se puede comprobar, para el proceso de limpieza se necesitan las listas de corrección, mientras que para el proceso de estandarización y segmentación son necesarias las tablas de búsqueda y los Modelos Ocultos de Markov. Para este último proceso, tal y como se ha comentado anteriormente, se podrán utilizar los modelos que se proporcionan junto con la aplicación o habrá que generarlos previamente. Así, en caso de que haya que generar tales modelos, el esquema general del proceso de normalización quedaría configurado de la siguiente manera:

Imagen 6. Proceso general de normalización usando un modelo HMM generado a partir de los datos

En este caso, antes de realizar el proceso de normalización se extraerá automáticamente una muestra del fichero de datos que contiene el campo a normalizar. Al seleccionar la muestra, los elementos que componen los valores de dicho campo quedarán etiquetados o clasificados de forma automática y a continuación, será el usuario el que manualmente tendrá que asignarle a cada etiqueta un estado.

Una vez asignados los estados, la aplicación llevará a cabo el entrenamiento de la muestra, obteniéndose así el Modelo Oculto de Markov que se usará en el proceso de normalización.

6 Herramienta de Normalización

La Herramienta de Normalización es una aplicación informática que permite realizar de forma sencilla un proceso de normalización de nombres de personas, direcciones postales o identificadores de personas físicas y/o jurídicas, de tal forma que a partir del conocimiento de la estructura o patrón que presentan los datos contenidos en una muestra se puede normalizar la totalidad del fichero de datos. Para llevar a cabo un eficiente proceso de normalización con la Herramienta de Normalización, se necesitan realizar los siguientes pasos:

1. **Tratamiento previo del fichero de datos**: este paso es obligatorio en cualquier proceso de normalización o de enlace ya que permite transformar el fichero a normalizar o enlazar en un fichero de texto CSV cuyos elementos estén separados por ";", así como recodificar los datos y eliminar algunos símbolos o caracteres que por su codificación pueden provocar fallos en el proceso de normalización o enlace.

2. **Normalización del fichero de datos**: como su nombre indica en este paso se normalizará el fichero de datos. Para ello se utilizarán las herramientas indicadas en el apartado 6 de este Manual: listas de corrección, tablas de búsqueda y Modelos Ocultos de Markov. En el caso de que el Modelo Oculto de Markov proporcionado por la aplicación no se adecue al fichero de trabajo habrá que crear un nuevo modelo.

3. **Validación del proceso de normalización**: una vez finalizada la normalización es necesario comprobar la bondad de la misma.

Para acceder a la Herramienta de Normalización se hará a través del botón correspondiente de la interfaz inicial de *aLink: Herramienta de Fusión de Ficheros*:

Imagen 7. Interfaz inicial de *aLink: Herramienta de Fusión de Ficheros*

6.1 Nuevas funcionalidades de la Herramienta de Normalización respecto a *ADYN: Herramienta de Normalización v 2.0*

La Herramienta de Normalización incluye las siguientes funcionalidades con respecto a la última versión de *ADYN: Herramienta de*

Normalización:

- Trabaja con ficheros CSV cuyos elementos están separados por el carácter ";". En la anterior versión de *ADYN: Herramienta de Normalización* se trabajaba con ficheros CSV cuyos elementos estaban separados por el carácter ",". El motivo de realizar este cambio es que es más frecuente encontrar el carácter "," que el carácter ";" entre la información recogida en un fichero.
- Se ha incluido en la interfaz gráfica de la aplicación barras de menús y submenús para poder acceder a las herramientas auxiliares que permiten complementar el proceso de normalización de un fichero de datos.
- Se ha incluido una herramienta de tratamiento inicial de los ficheros de trabajo. Esta herramienta recodifica y elimina símbolos o elementos que por su codificación podrían provocar problemas en la normalización y transforma a formato *csv* determinados ficheros que tienen un formato distinto al requerido para trabajar con la *Herramienta de Normalización* (*csv* separado por ";").
- Se ha modificado la denominación del directorio que contiene el código de la aplicación así como del directorio en el que se encuentran las listas de corrección, las tablas de búsqueda y los Modelos Ocultos de Markov. Anteriormente se denominaban respectivamente "codigo" y "datos" y ahora "app" y "listas_tablas".
- En lo que respecta a nombres de personas, se ha incluido un nuevo Modelo Oculto de Markov así como la muestra a partir de la cual se ha construido. Dicho modelo permite segmentar campos con nombres de personas que contengan el nombre de pila y los apellidos conjuntamente.
- Por otro lado, en cuanto a direcciones postales, se ha incorporado una nueva desagregación de la dirección postal y se ha modificado en cierta medida la ya existente. La nueva desagregación muestra nuevos campos de salida en el fichero normalizado y se ha llevado a cabo de acuerdo con la desagregación que utiliza el modelo de datos del Callejero Digital de Andalucía Unificado (CDAU). Por tanto, la aplicación permite ahora dos tipos de desagregaciones, la basada en el CDAU y otra similar a la ofrecida en la última versión de *ADYN: Herramienta de Normalización*, que también es a libre elección del usuario.

  En concreto, con motivo de la inclusión de la desagregación CDAU, se han modificado e incluido algunos campos de salida en relación con la numeración de las vías, con los que se pretende cubrir numeraciones de vía del tipo: 17A-21C. Además, se ha incluido otro campo de salida denominado *agrupacion*, que hace referencia a un conjunto de construcciones no consideradas como un núcleo de población en el Nomenclátor del Instituto Nacional de Estadística (INE) tales como barrios, polígonos industriales, urbanizaciones, etc. y se han realizado análisis y algunas modificaciones de los campos de salida ya existentes. Debido a todos estos cambios, ha sido necesario construir un nuevo Modelo Oculto de Markov para direcciones postales y alguna nueva tabla de búsqueda.

  Por otro lado, se ha realizado un análisis de las tablas de búsqueda ya existentes para direcciones postales y se han añadido nuevos elementos a las mismas. También se han modificado e incluido algunos estados asociados a dichas tablas.

6.2 Descripción general de la Herramienta de Normalización

La interfaz gráfica inicial de la *Herramienta de Normalización* se visualiza en la siguiente imagen:

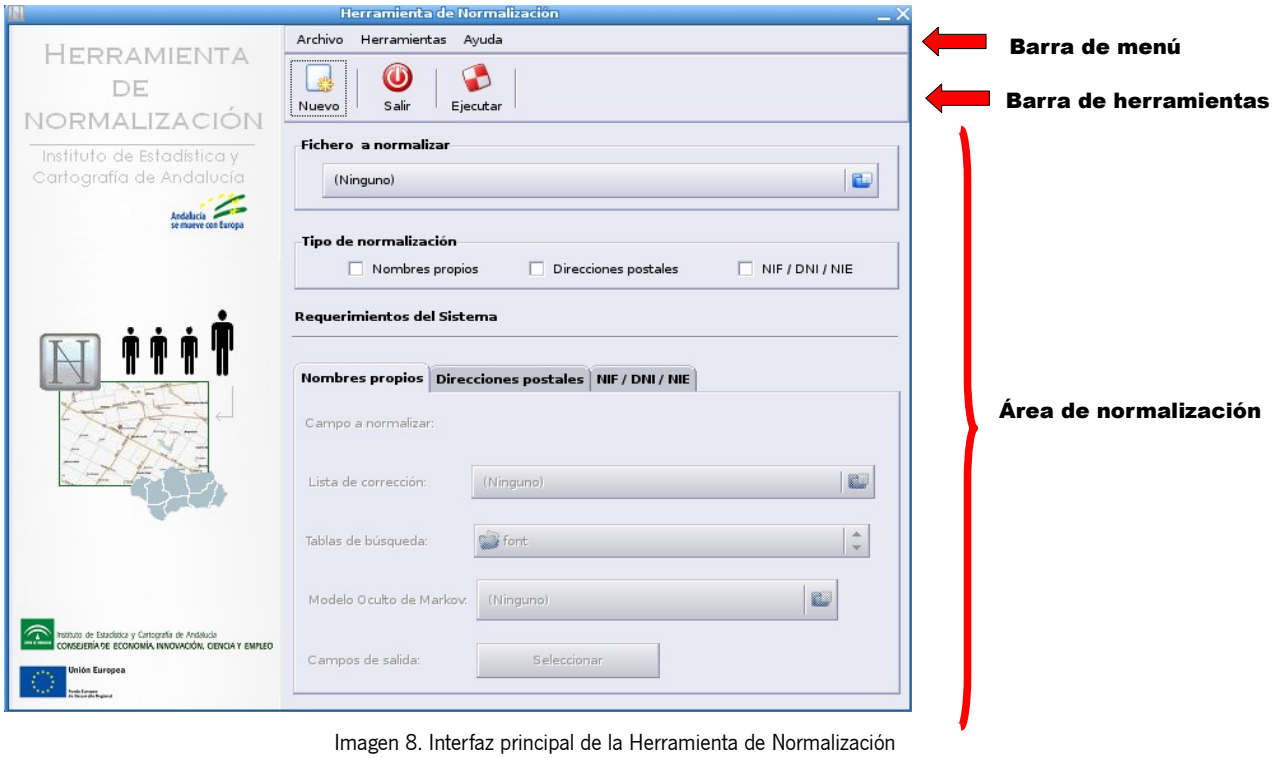


Imagen 8. Interfaz principal de la Herramienta de Normalización

Como se puede observar la interfaz está estructurada en tres partes: la parte superior de la ventana contiene una barra de menú y justo debajo de ella aparece una barra de herramientas. El resto de la ventana constituye la parte más importante de la interfaz, el área de normalización. En ella se especificarán los parámetros y requisitos del sistema necesarios para llevar a cabo un proceso de normalización.

6.2.1 Barra de menú de la Herramienta de Normalización

La barra de menú contiene las opciones:

- **Archivo**: a través de este menú se puede inicializar un nuevo proceso de normalización, seleccionando **Nuevo** o salir de la herramienta de normalización, seleccionando **Salir**.

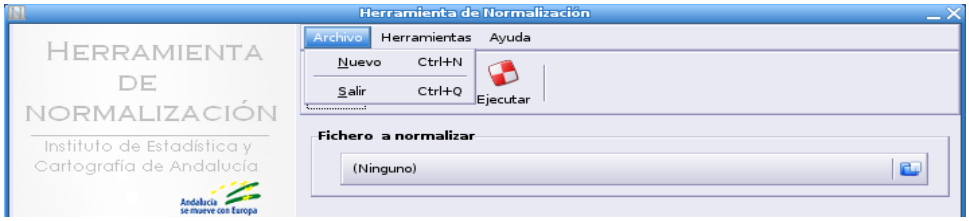


Imagen 9. Menú Archivo de la Herramienta de Normalización

- **Herramientas**: este menú permite al usuario realizar un tratamiento previo del fichero a normalizar (**Tratamiento previo**), seleccionar una muestra del campo a normalizar y etiquetarla automáticamente como paso previo a la creación de un Modelo Oculto de Markov (**HMM: Selección de la muestra**), entrenar dicha muestra o lo que es lo mismo generar el Modelo Oculto de Markov (**HMM: Entrenamiento de la muestra**) y editar algunos de los ficheros utilizados en el proceso de normalización (**Editor de listas de corrección** y **Editor de tablas de búsqueda**).

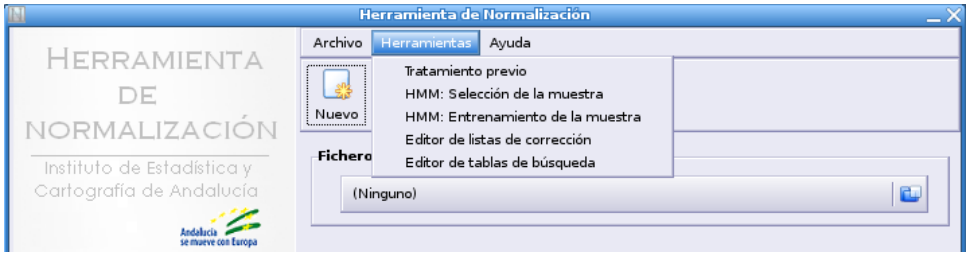


Imagen 10. Menú Herramientas de la Herramienta de Normalización

Las herramientas que se incluyen en este menú se analizarán con más detalle en el apartado 6.3 de este Manual.

- **Ayuda**: este menú ofrece al usuario una serie de información sobre *aLink: Herramienta de Fusión de Ficheros* que le puede ser útil, como por ejemplo, información sobre la licencia bajo la que se desarrolla esta herramienta.

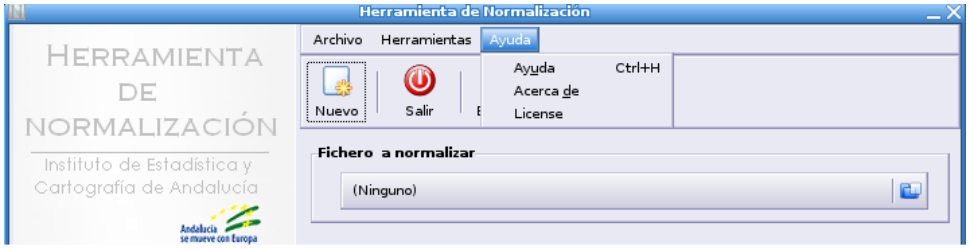


Imagen 11. Menú Ayuda de la Herramienta de Normalización

6.2.2 Barra de herramientas de la Herramienta de Normalización

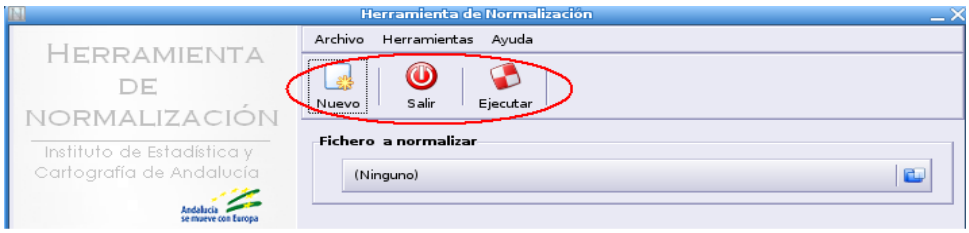


Imagen 12. Barra herramientas de la Herramienta de Normalización

La barra de herramientas cuenta con los siguientes botones:

- **Nuevo**: su funcionamiento es equivalente a la opción 'Nuevo' del menú Archivo.
- **Salir**: su funcionamiento es equivalente a la opción 'Salir' del menú Archivo.
- **Ejecutar**: al pulsar este botón se lleva a cabo el proceso de normalización de cualquiera de los campos que permite normalizar la herramienta: nombres de personas, direcciones postales o identificadores de personas físicas y/o jurídicas.

6.2.3 Área de normalización

Imagen 13. Área de normalización de la Herramienta de Normalización

El área de normalización contiene los siguientes elementos:

- **Fichero a normalizar**: botón donde el usuario especificará la ruta en la que se ubica el fichero que desea normalizar.
- **Tipo de normalización**: permite al usuario seleccionar el tipo de campo que va a normalizar. Sus posibles valores son: Nombres propios, Direcciones postales y NIF/DNI/NIE. No se deberá seleccionar más de un campo a la vez para evitar problemas al ejecutar el proceso de normalización.
- **Requerimientos del sistema**: los parámetros requeridos por la aplicación en esta sección son:
  - Nombres propios/ Direcciones postales / NIF/DNI/NIE: pestaña que el usuario tendrá que seleccionar para indicar el tipo de normalización que va a realizar.
  - Campo a normalizar: combo que contiene todas las variables o campos del fichero a normalizar. De entre ellos, el usuario seleccionará el que desee normalizar.

- Lista de corrección: en este botón el usuario indicará la ubicación del directorio en el que se encuentra la lista de corrección. En la aplicación se proporcionan listas de corrección para nombres de personas, direcciones postales e identificadores de personas físicas y/o jurídicas. Estas están ubicadas en el directorio *alink\app\listas_tablas\listas_de_correccion*. De entre ellas se seleccionará la correspondiente al tipo de normalización que se va a realizar, esto es, *nombres_correccion.lst* para nombres de personas, *direcciones_correccion.lst* para direcciones postales o *idpersona_correccion.lst* para identificadores de personas físicas y/o jurídicas. La estructura de las mismas se muestra en el apartado 6.3.4 de este Manual.

- Tablas de búsqueda: en este botón el usuario indicará la ubicación del directorio en el que se encuentran las tablas de búsqueda. En la aplicación se proporcionan tablas de búsqueda para nombres de personas, direcciones postales e identificadores de personas físicas y/o jurídicas. Estas están ubicadas en el directorio *alink\app\listas_tablas\tablas_de_busqueda*. De entre ellas se seleccionará la correspondiente al tipo de normalización que se va a realizar, esto es, *tbl_nombre* para nombres de personas, *tbl_direccion* para direcciones postales o *tbl_idpersona* para identificadores de personas físicas y/o jurídicas. La estructura de las mismas se muestra en el apartado 6.3.5 de este Manual.

- Modelo Oculto de Markov: en este botón el usuario indicará la ubicación del directorio en la que se encuentran los Modelos Ocultos de Markov. En la aplicación se proporcionan modelos para nombres de personas, direcciones postales e identificadores de personas físicas y/o jurídicas. Estos están ubicados en el directorio *alink\app\muestras_modelos*. De entre los modelos disponibles se seleccionará el correspondiente al tipo de normalización que se va a realizar, esto es, nombres de personas, direcciones postales o identificadores de personas físicas y/o jurídicas. En el Anexo II se indica más detalladamente todos los modelos HMM disponibles para cada uno de estos tres casos. No obstante, si los modelos proporcionados no se adecuan al fichero de datos a normalizar, el usuario tendría que construir uno nuevo. En los apartados 6.3.2 y 6.3.3 de este Manual se verá con más detalle cómo construir estos modelos.

- Campos de salida: este botón permite al usuario seleccionar los campos de salida en los que desea segmentar los valores del campo a normalizar. La elección de los mismos está basada en el Manual de buenas prácticas para la normalización de fuentes y registros administrativos de la Junta de Andalucía [6]. Una vez elegidos solamente tendrá que pulsar el botón **OK** para confirmar la selección. Por ejemplo, si el usuario va a normalizar un campo que contiene direcciones postales, los campos de salida en los que se puede segmentar la dirección son los que se muestran a continuación:

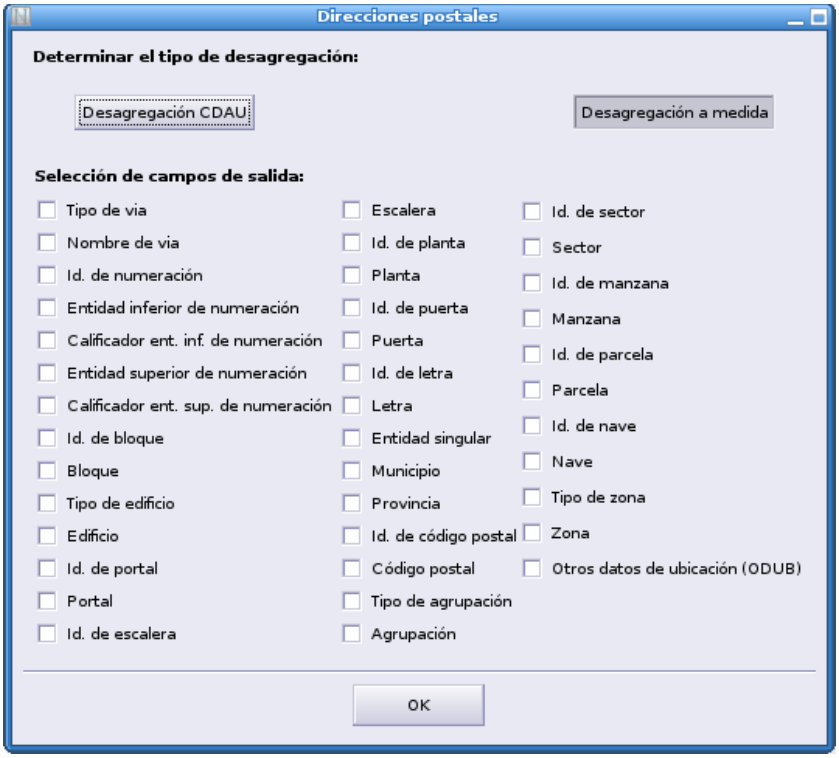


Imagen 14. Campos de salida para direcciones postales. Desagregación a medida

Como se puede observar en la ventana aparecen dos botones:

- *Desagregación a medida*: este botón es el que aparece marcado por defecto y permite una desagregación de los valores del campo a normalizar a libre elección del usuario. Así, si se elige esta opción se podrán seleccionar todos los campos de salida (39 campos) o solamente algunos de ellos.

- *Desagregación CDAU*: esta opción permite una desagregación del campo a normalizar de acuerdo con la desagregación usada en el Callejero Digital de Andalucía Unificado. Si el usuario la selecciona, automáticamente se marcarán los campos relativos a esta desagregación, no siendo posible seleccionar ningún otro campo. Los campos que conforman la desagregación CDAU son los que se muestran marcados en la siguiente imagen:

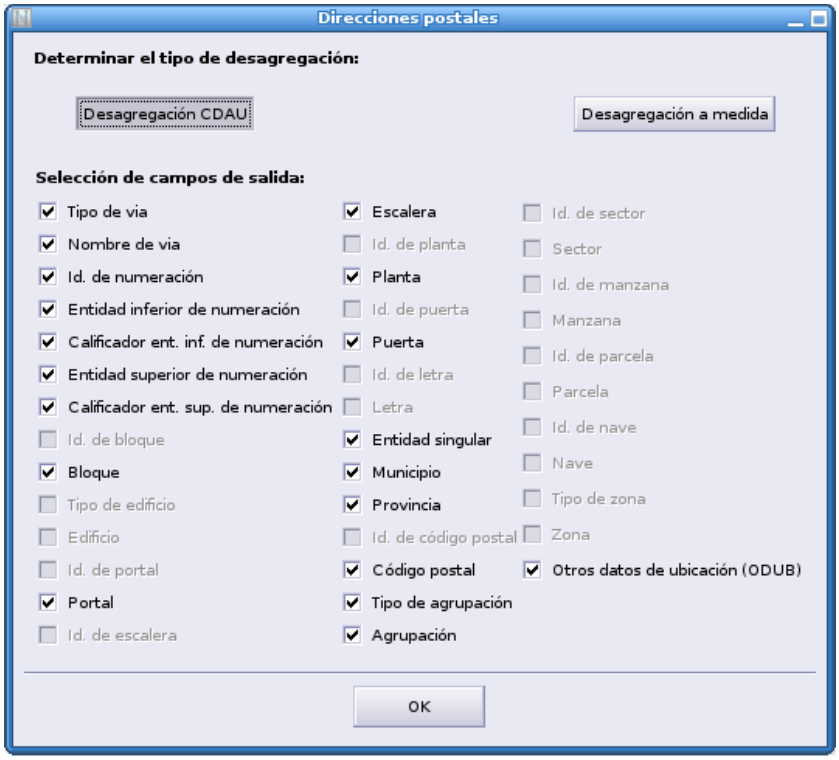


Imagen 15. Campos de salida para direcciones postales. Desagregación CDAU

Como se puede comprobar, si se comparan los campos de salida de la nueva versión de la Herrramienta de Normalización con los de la anterior versión de *ADYN: Herramienta de Normalización*, se han producido inclusiones, modificaciones y eliminaciones de algunos de ellos.

Por ejemplo, tal y como se comentó al principio de este Manual, en esta nueva versión de la Herramienta de Normalización algunos de los campos de salida relacionados con la numeración de las vías han cambiado su denominación y se han incluido otros nuevos. Así por ejemplo, el campo de salida denominado "Id. de número" ha pasado a denominarse "Id. de numeración" mientras que el campo "Número" ha pasado a denominarse "Entidad inferior de numeración" (ein). Además, se han incluido otros tres nuevos campos en relación con la numeración de las vías, en concreto, "Entidad superior de numeración" (esn), "Calificador ent. inf. de numeración" (cein) y "Calificador ent. sup. de numeración" (cesn). La inclusión de este tipo de campos dará cobertura a los casos en los que el número de la vía es el tipo: 17A-21C, en donde el número 17 corresponderá a la entidad inferior de numeración (ein), la letra A al calificador de la entidad inferior de numeración (cein), el número 21 a la entidad superior de numeración (esn) y la letra C al calificador de la entidad superior de numeración (cesn).

También se ha cambiado la denominación del campo "Localidad" e "Informacion adicional", pasando ahora a denominarse "Municipio" y "Otros datos de ubicación (ODUB)".

Por otro lado, se han incluido nuevos campos de salida como "Tipo de agrupación", "Agrupación" y "Provincia". En concreto, los dos primeros hacen referencia a un conjunto de construcciones no consideradas como núcleos de

población en el Nomenclátor del INE, tanto si se corresponde a una agrupación aislada, como si es una parte integrante de un núcleo de población. Estos conjuntos son: barrios, barriadas, polígonos industriales, parques comerciales y urbanizaciones. Esta situación ha generado que campos de salida de la anterior versión de *ADYN: Herramienta de Normalización* como "Id. de barriada", "Barriada", "Id. de complejo" y "Complejo" hayan desaparecido y los valores relativos a estos campos se hayan incluido en los campos de salida de la nueva versión "Tipo de agrupación" y "Agrupación".

Otra de las modificaciones llevadas a cabo en los campos de salida respecto de la última versión de *ADYN: Herramienta de Normalización*, hace referencia a los campos de salida "Id. de edificio singular", "Edificio singular", "Tipo de comercio" y "Comercio", que al igual que los cuatro anteriores han desaparecido, incorporándose los valores relativos a los mismos en los campos de salida "Tipo de edificio" y "Edificio". Igual ha sucedido con los campos "Id. de kilómetro" y "Kilómetro", que tras su desaparición, los valores de los mismos se incluyen en los campos "Id. de numeración" y "Entidad inferior de numeración".

Por último, en relación a los campos de salida hay indicar que en el Anexo III de este Manual se muestran los relativos a nombres de personas e identificadores de personas físicas y jurídicas.

Para finalizar con este apartado de normalización se presenta a modo de ejemplo cómo quedaría configurada la interfaz de la Herramienta de Normalización al normalizar el campo *direccion* de un fichero de datos que contiene direcciones postales de establecimientos comerciales. En este proceso se usará además, la desagregación CDAU:

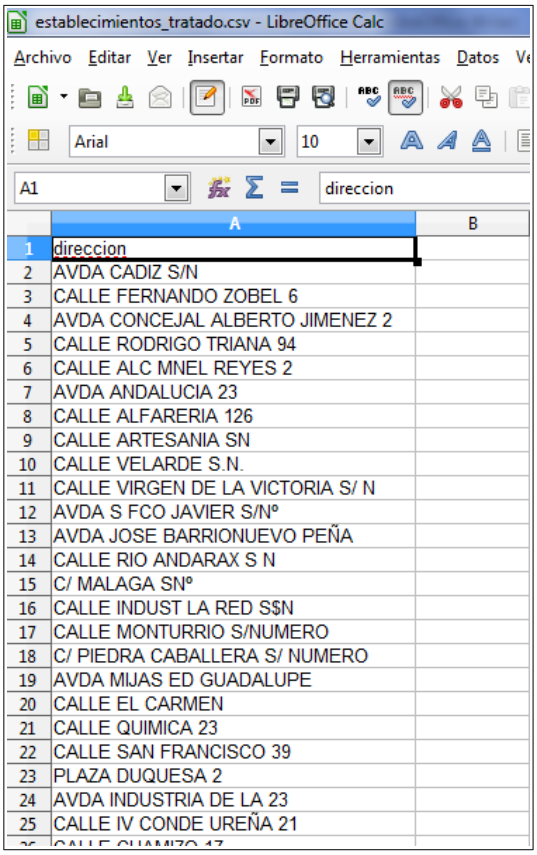


Imagen 16. Detalle del fichero de establecimientos a normalizar

Imagen 17. Interfaz de la Herramienta de Normalización con parámetros establecidos

Establecidos los elementos y seleccionados los campos de salida, que como se ha indicado serán los relativos a la Desagregación CDAU, el usuario deberá pulsar el botón **Ejecutar** para llevar a cabo la normalización.

El proceso de normalización generará cuatro ficheros de salida que se guardarán en la carpeta donde se encuentra el fichero a normalizar, 'Empresas_tratado.csv'. Estos serán:

- Fichero **'NORM_<fecha_creación>-<hora_creación>_<nombre_fichero_a_normalizar>.csv'**: contiene todos los campos del fichero original, junto con los campos de salida que ofrece la desagregación CDAU. Si el usuario hubiera seleccionado una desagregación a medida, aparecerían los campos de salida que éste hubiera marcado. Además, se incluye una columna adicional en el fichero denominada 'validacion' y que servirá para analizar la bondad del proceso de normalización. La estructura de este fichero se observa en la siguiente imagen:

NORM_20140114-2108_establecimientos_tratado.csv - LibreOffice Calc

Archivo Editar Ver Insertar Formato Herramientas Datos Ventana Ayuda

Arial 10

A61 paraje cakkada del moro. numero 109

	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S	T	U
1	'direccion'	tipo_de_via	nombre_de_via	identificador_de_numeraci	ein	cein	esn	cesn	bloque	portal	escalera	planta	puerta	entidad_singular	municipio	provinci	codigo_postal	tipo_de_agrupacion	agrupacion	odub	validacion
2	avda cadiz s/ n	avenida	cadiz	sin_numero																	0
3	calle fernando zobel 6	calle	fernando zobel		6																0
4	avda concejal alberto jimenez 2	avenida	concejal alberto jimenez		2																0
5	calle rodrigo triana 94	calle	rodrigo triana		94																0
6	calle alc mnel reyes 2	calle	alc mnel reyes		2																0
7	avda andalucia 23	avenida	andalucia		23																0
8	calle alfareria 126	calle	alfareria		126																0
9	calle artesania sn	calle	artesania	sin_numero																	0
10	calle velarde s.n.	calle	velarde	sin_numero																	0
11	calle virgen de la victoria s/ n	calle	virgen de la_victoria	sin_numero																	0
12	avda s fco javier s/ numero	avenida	s fco javier	sin_numero																	0
13	avda jose barrionuevo pekka	avenida	jose barrionuevo pekka																		0
14	calle rio andarax s n	calle	rio andarax	sin_numero																	0
15	c/ malaga snumero	calle	malaga snumero																		0
16	calle indust la red s5n	calle	indust la red	sin_numero																	0
17	calle monturrio s/ numero	calle	monturrio	sin_numero																	0
18	c/ piedra caballera s/ numero	calle	piedra caballera	sin_numero																	0
19	avda mijas ed guadalupe	avenida	mijas																	edificio guadalu	0
20	calle el carmen	calle	el_carmen																		0
21	calle quimica 23	calle	quimica		23																0
22	calle san francisco 39	calle	san_francisco		39																0
23	plaza duquesa 2	plaza	duquesa		2																0
24	avda industria de la 23	avenida	industria de la		23																0
25	calle iv conde urekka 21	calle	iv conde urekka		21																0
26	calle chamizo 17	calle	chamizo		17																0
27	calle jose diaz 1	calle	jose diaz		1																0
28	calle tocina 30	calle	tocina		30																0
29	avda greco 1	avenida	greco		1																0
30	calle capitan 35	calle	capitan		35																0
31	calle cesteria 10	calle	cesteria		10																0
32	c/ cascada 10		c/ cascada		10																0
33	calle alfonso xii 81	calle	alfonso xii		81																0
34	avda de las ciencias 20	avenida	de las ciencias		20																0
35	avda principes de espakka 9	avenida	principes de espakka		9																0
36	calle virgen milagrosa 4	calle	virgen milagrosa		4																0

Imagen 18. Detalle del fichero con direcciones normalizado

Al observar el fichero se puede comprobar que la denominación de los campos de salida de acuerdo a CDAU (Imagen 15) no se corresponden exactamente con la denominación de los campos del fichero de salida normalizado. Por ejemplo, el campo **Otros datos de ubicación (ODUB)** aparece en el fichero normalizado como **odub** y lo mismo ocurre con algunos otros. En el Anexo IV se puede consultar la denominación exacta de los campos de salida del fichero normalizado, tanto para direcciones postales como para nombres de personas e identificadores de personas físicas y jurídicas.

- Fichero de proyecto **'proy<fecha_creación>-<hora_creación>_<nombre_fichero_origen>.py'**: contiene el código en el que se indica el conjunto de parámetros con los que se ha realizado el proceso de normalización, permitiendo reproducir o modificar este proceso posteriormente. Para ello deberá guardarse en la carpeta 'app' de la aplicación *aLink: Herramienta de Fusión de Ficheros* y dependiendo del entorno en el que se trabaje la forma de ejecutarlo será la siguiente:
    - Si se trabaja en un entorno Windows se hará doble click sobre el o se abrirá el mismo con algún programa como IDLE o Geany y se ejecutará.
    - Si se trabaja en un entorno Linux habrá que acceder al directorio app en el que se ha copiado el fichero y escribir la orden #python *denominación_del_fichero_de_proyecto*.py.

La estructura de este fichero es:

proy_20140114-...ntos_tratado.py

```python
#!/usr/bin/env python
# -*- coding: utf-8 -*-
# ==========================================================================
# Comienzo del módulo: "proy_20140114-2108_establecimientos_tratado.py"
#
# Generado: Tue Jan 14 21:18:41 2014
# ==========================================================================

# Importar módulos (primero módulos estándar de Python y luego los módulos de Febrl)

import logging
import time
import os

import dataset
import lookup
import simplehmm
import standardisation

# --------------------------------------------------------------------------
# Intializar registro
#
log_level = logging.WARNING # logging.WARNING

my_logger = logging.getLogger()
my_logger.setLevel(log_level)

tiempo_de_comienzo = time.time()
print "La ejecución comenzó: %s" % time.asctime(time.localtime(tiempo_de_comienzo))

# --------------------------------------------------------------------------
# Tipo de proyecto: Normalización
# --------------------------------------------------------------------------

# --------------------------------------------------------------------------

# Definición del juego de datos de entrada A
#
data_set_a = dataset.DataSetCSV(description="Datos de entrada",
                                access_mode="read",
                                strip_fields=True,
                                miss_val=[],
                                rec_ident="__rec_id_a__",
                                file_name="D:"+os.sep+"Trabajo_IECA"+os.sep+"Fusion_ficheros"+os.sep+"Ficheros pruebas"+os.sep+"establecimientos_tratado.utf-8.limpio.csv",
                                header_line=True,
                                delimiter=",",
                                field_list = [("`direccion`",0)])

# --------------------------------------------------------------------------

# Definición de los componentes de normalización
#
addr_corr_list_0 = lookup.CorrectionList(descr = "Lista de corrección para direcciones")
addr_corr_list_0.load("D:"+os.sep+"Trabajo_IECA"+os.sep+"Fusion_ficheros"+os.sep+"aLink_Herramienta_Fusion_Ficheros_v1_0_win7_x64b"+os.sep+"aLink_Herramienta_Fusion_Ficheros_v1_0_win7_x64b"+os.sep+"alink"+

addr_tag_table_0 = lookup.TagLookupTable(descr = "Tablas de búsqueda para direcciones")
addr_tag_table_0.load(["D:"+os.sep+"Trabajo_IECA"+os.sep+"Fusion_ficheros"+os.sep+"aLink_Herramienta_Fusion_Ficheros_v1_0_win7_x64b"+os.sep+"aLink_Herramienta_Fusion_Ficheros_v1_0_win7_x64b"+os.sep+"alink"+
                       "D:"+os.sep+"Trabajo_IECA"+os.sep+"Fusion_ficheros"+os.sep+"aLink_Herramienta_Fusion_Ficheros_v1_0_win7_x64b"+os.sep+"aLink_Herramienta_Fusion_Ficheros_v1_0_win7_x64b"+os.sep+"alink"+
                       "D:"+os.sep+"Trabajo_IECA"+os.sep+"Fusion_ficheros"+os.sep+"aLink_Herramienta_Fusion_Ficheros_v1_0_win7_x64b"+os.sep+"aLink_Herramienta_Fusion_Ficheros_v1_0_win7_x64b"+os.sep+"alink"+
                       "D:"+os.sep+"Trabajo_IECA"+os.sep+"Fusion_ficheros"+os.sep+"aLink_Herramienta_Fusion_Ficheros_v1_0_win7_x64b"+os.sep+"aLink_Herramienta_Fusion_Ficheros_v1_0_win7_x64b"+os.sep+"alink"+
                       "D:"+os.sep+"Trabajo_IECA"+os.sep+"Fusion_ficheros"+os.sep+"aLink_Herramienta_Fusion_Ficheros_v1_0_win7_x64b"+os.sep+"aLink_Herramienta_Fusion_Ficheros_v1_0_win7_x64b"+os.sep+"alink"+
                       "D:"+os.sep+"Trabajo_IECA"+os.sep+"Fusion_ficheros"+os.sep+"aLink_Herramienta_Fusion_Ficheros_v1_0_win7_x64b"+os.sep+"aLink_Herramienta_Fusion_Ficheros_v1_0_win7_x64b"+os.sep+"alink"+
                       "D:"+os.sep+"Trabajo_IECA"+os.sep+"Fusion_ficheros"+os.sep+"aLink_Herramienta_Fusion_Ficheros_v1_0_win7_x64b"+os.sep+"aLink_Herramienta_Fusion_Ficheros_v1_0_win7_x64b"+os.sep+"alink"+
                       "D:"+os.sep+"Trabajo_IECA"+os.sep+"Fusion_ficheros"+os.sep+"aLink_Herramienta_Fusion_Ficheros_v1_0_win7_x64b"+os.sep+"aLink_Herramienta_Fusion_Ficheros_v1_0_win7_x64b"+os.sep+"alink"+
                       "D:"+os.sep+"Trabajo_IECA"+os.sep+"Fusion_ficheros"+os.sep+"aLink_Herramienta_Fusion_Ficheros_v1_0_win7_x64b"+os.sep+"aLink_Herramienta_Fusion_Ficheros_v1_0_win7_x64b"+os.sep+"alink"+
                       "D:"+os.sep+"Trabajo_IECA"+os.sep+"Fusion_ficheros"+os.sep+"aLink_Herramienta_Fusion_Ficheros_v1_0_win7_x64b"+os.sep+"aLink_Herramienta_Fusion_Ficheros_v1_0_win7_x64b"+os.sep+"alink"+
                       "D:"+os.sep+"Trabajo_IECA"+os.sep+"Fusion_ficheros"+os.sep+"aLink_Herramienta_Fusion_Ficheros_v1_0_win7_x64b"+os.sep+"aLink_Herramienta_Fusion_Ficheros_v1_0_win7_x64b"+os.sep+"alink"+
                       "D:"+os.sep+"Trabajo_IECA"+os.sep+"Fusion_ficheros"+os.sep+"aLink_Herramienta_Fusion_Ficheros_v1_0_win7_x64b"+os.sep+"aLink_Herramienta_Fusion_Ficheros_v1_0_win7_x64b"+os.sep+"alink"+
                       "D:"+os.sep+"Trabajo_IECA"+os.sep+"Fusion_ficheros"+os.sep+"aLink_Herramienta_Fusion_Ficheros_v1_0_win7_x64b"+os.sep+"aLink_Herramienta_Fusion_Ficheros_v1_0_win7_x64b"+os.sep+"alink"+
                       "D:"+os.sep+"Trabajo_IECA"+os.sep+"Fusion_ficheros"+os.sep+"aLink_Herramienta_Fusion_Ficheros_v1_0_win7_x64b"+os.sep+"aLink_Herramienta_Fusion_Ficheros_v1_0_win7_x64b"+os.sep+"alink"+
                       "D:"+os.sep+"Trabajo_IECA"+os.sep+"Fusion_ficheros"+os.sep+"aLink_Herramienta_Fusion_Ficheros_v1_0_win7_x64b"+os.sep+"aLink_Herramienta_Fusion_Ficheros_v1_0_win7_x64b"+os.sep+"alink"+
                       "D:"+os.sep+"Trabajo_IECA"+os.sep+"Fusion_ficheros"+os.sep+"aLink_Herramienta_Fusion_Ficheros_v1_0_win7_x64b"+os.sep+"aLink_Herramienta_Fusion_Ficheros_v1_0_win7_x64b"+os.sep+"alink"+
                       "D:"+os.sep+"Trabajo_IECA"+os.sep+"Fusion_ficheros"+os.sep+"aLink_Herramienta_Fusion_Ficheros_v1_0_win7_x64b"+os.sep+"aLink_Herramienta_Fusion_Ficheros_v1_0_win7_x64b"+os.sep+"alink"+
                       "D:"+os.sep+"Trabajo_IECA"+os.sep+"Fusion_ficheros"+os.sep+"aLink_Herramienta_Fusion_Ficheros_v1_0_win7_x64b"+os.sep+"aLink_Herramienta_Fusion_Ficheros_v1_0_win7_x64b"+os.sep+"alink"+
                       "D:"+os.sep+"Trabajo_IECA"+os.sep+"Fusion_ficheros"+os.sep+"aLink_Herramienta_Fusion_Ficheros_v1_0_win7_x64b"+os.sep+"aLink_Herramienta_Fusion_Ficheros_v1_0_win7_x64b"+os.sep+"alink"+
                       "D:"+os.sep+"Trabajo_IECA"+os.sep+"Fusion_ficheros"+os.sep+"aLink_Herramienta_Fusion_Ficheros_v1_0_win7_x64b"+os.sep+"aLink_Herramienta_Fusion_Ficheros_v1_0_win7_x64b"+os.sep+"alink"+

adr_hmm = simplehmm.hmm("HMM para direcciones ", [], [])
adr_hmm.load_hmm("D:"+os.sep+"Trabajo_IECA"+os.sep+"Fusion_ficheros"+os.sep+"aLink_Herramienta_Fusion_Ficheros_v1_0_win7_x64b"+os.sep+"aLink_Herramienta_Fusion_Ficheros_v1_0_win7_x64b"+os.sep+"alink"+os.sep

address_comp_std_0 = standardisation.AddressStandardiser(input_fields = ["`direccion`"],
                                output_fields = ["tipo_de_via","nombre_de_via","identificador_de_numeracion","ein","cein","esn","cesn",None,"bloque",None,None,None,"portal",None,"escalera",None,"planta",None,
                                field_separator = " ",
                                check_word_spill = False,
                                corr_list = addr_corr_list_0,
                                tag_table = addr_tag_table_0,
                                address_hmm = adr_hmm,)

# --------------------------------------------------------------------------

# Definición de los datos de salida normalizados y del registro de normalización
#
data_set_a_std = dataset.DataSetCSV(description="Datos de salida normalizados",
                                access_mode="write",
                                delimiter=",",
                                file_name="D:"+os.sep+"Trabajo_IECA"+os.sep+"Fusion_ficheros"+os.sep+"Ficheros pruebas"+os.sep+"NORM_20140114-2108_establecimientos_tratado.csv",
                                field_list = [("`direccion`",0),
                                              ("tipo_de_via",1),
                                              ("nombre_de_via",2),
                                              ("identificador_de_numeracion",3),
                                              ("ein",4),
                                              ("cein",5),
                                              ("esn",6),
                                              ("cesn",7),
                                              ("bloque",8),
                                              ("portal",9),
                                              ("escalera",10),
                                              ("planta",11),
                                              ("puerta",12),
                                              ("entidad_singular",13),
                                              ("municipio",14),
                                              ("provincia",15),
                                              ("codigo_postal",16),
                                              ("tipo_de_agrupacion",17),
                                              ("agrupacion",18),
                                              ("odub",19),
                                              ("validacion",20)],
                                rec_ident="__rec_id",
                                strip_fields=True,
                                header_line=True,
                                write_header=True)

# Definición del registro de normalización
#
rec_std = standardisation.RecordStandardiser(descr = "Registro de normalización",
                                input_data =  data_set_a,
                                output_data = data_set_a_std,
                                comp_stand_list = [address_comp_std_0],
                                progress_report = 10,
                                pass_field_list = [("`direccion`","`direccion`")])

# Comenzar la normalización
#
rec_std.standardise()

tiempo_de_finalizacion = time.time()
print "La ejecución terminó: %s" % time.asctime(time.localtime(tiempo_de_finalizacion))
print "Tiempo total de ejecución: %s segundos" % (tiempo_de_finalizacion - tiempo_de_comienzo)


# ==========================================================================
# Fin del módulo: "proy_20140114-2108_establecimientos_tratado.py"
# ==========================================================================
```

Imagen 19. Fichero de proyecto de un proceso de normalización

- Dos ficheros de codificación: ***'nombre_fichero_origen.utf-8.csv'*** y ***'nombre_fichero_origen.utf-8.limpio.csv'***: son ficheros intermedios útiles para la generación del fichero normalizado final pero no tienen ninguna otra utilidad para el

usuario. Por tanto, este podría eliminarlos sin ningún problema.

6.3 Menú Herramientas de la Herramienta de Normalización

6.3.1 Tratamiento previo

Antes de llevar a cabo un proceso de normalización es necesario realizar un tratamiento inicial de los ficheros de trabajo. El principal motivo es que la mayoría de los ficheros con los que se trabaja no se encuentran en el formato requerido por *aLink: Herramienta de Fusión de Ficheros*, es decir, en formato CSV con elementos separados por ";". Luego para poder normalizar o enlazar un fichero es necesario transformarlo. No obstante, incluso si el fichero que se va a normalizar se encontrara en el formato requerido, podría ocurrir que contuviera símbolos o elementos que por su codificación pudieran provocar errores en la normalización.

Para solucionar esta situación, el proceso de tratamiento recodifica automáticamente los datos, así por ejemplo, convierte a minúscula todo el fichero de datos, el carácter ñ que en bastantes ocasiones presenta problemas de codificación se ha sustituido automáticamente por los caracteres "kk", mientras que los caracteres "," y ";" se han sustituido automáticamente por un espacio en blanco para evitar que se produzca una incorrecta segmentación de la información una vez tratados los ficheros. Hay que indicar que la sustitución de los caracteres ";" se lleva a cabo en todos los formatos de los ficheros salvo en aquellos que ya tienen formato CSV separado por el carácter ";", ya que si en estos casos se eliminaran los ";" el fichero se quedaría sin separador de elementos.

Un tratamiento especial requieren los valores numéricos que presentan decimales y utilizan como separador decimal el carácter ",", como por ejemplo los puntos kilométricos. En estos casos, el carácter "," se ha sustituido automáticamente por el carácter "+". Es decir, si en un campo se tienen valores del tipo "427,500", tras el tratamiento previo del fichero, pasarán a mostrarse como "427+500". Si no se hubiera realizado este cambio y se hubiera conservado la sustitución generalizada del carácter "," por el espacio en blanco, el valor "427,500" se mostraría como "427 500", situación que puede llevar a confusión a la hora de segmentar la información del fichero ya que en lugar de tener un valor numérico con esta sustitución se tienen dos. Ni que decir tiene que esta sustitución automática puede provocar errores en algunos otros valores del fichero, por ejemplo, si se tiene la dirección "C/LEONARDO DA VINCI, 35,37" tras el tratamiento este valor se mostrará como "C/LEONARDO DA VINCI, 35+37". Con lo cual en estos casos el usuario deberá valorar, teniendo en cuenta el contenido de su fichero de trabajo, si tras el tratamiento del fichero decide sustituir o no el carácter "+", por otro que considere más adecuado. **¡OJO!** Si decide sustituirlo por otro carácter no deberá hacerlo lógicamente ni por el carácter "," ni por ";". Podría por ejemplo, utilizar el carácter "." y luego tener la precaución de que dicho carácter no se encuentre dentro de las listas de corrección.

A continuación, se presenta la interfaz gráfica que permite llevar a cabo el tratamiento inicial de un fichero de datos:

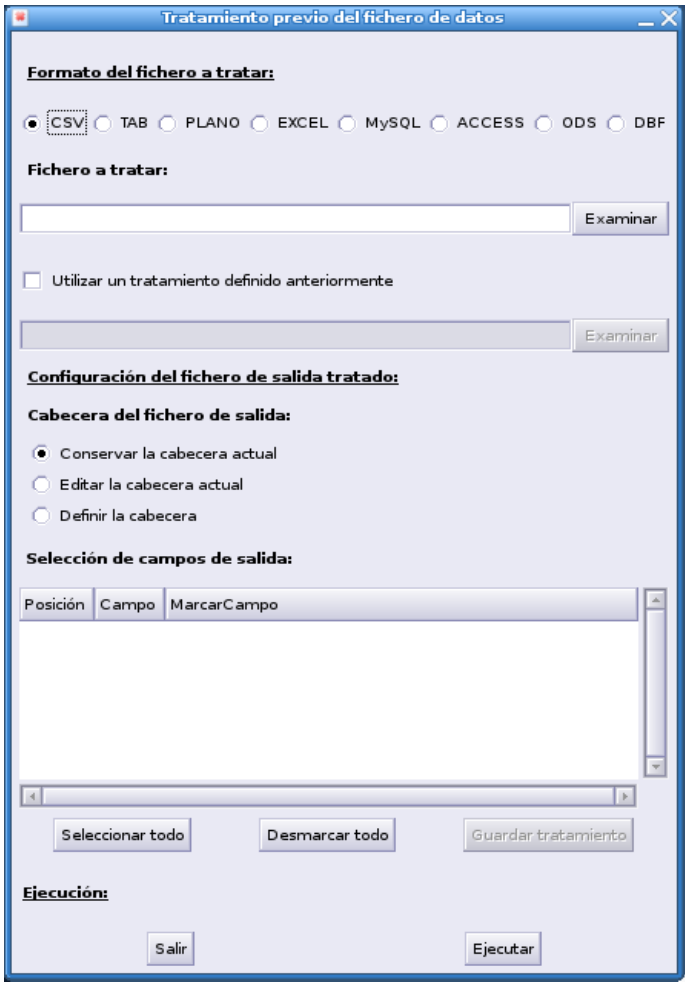


Imagen 20. Interfaz de tratamiento previo

Como se puede comprobar, la primera parte de la ventana recoge el formato del fichero a tratar y su ubicación, a continuación se muestra la sección donde se establece la configuración o estructura del fichero de salida tratado y por último se presentan los botones para ejecutar el tratamiento o salir de la herramienta. A continuación, se analiza más detalladamente cada elemento:

- **Formato del fichero de datos a tratar**: en esta sección el usuario indicará el formato del fichero de datos original que va a tratar. Los formatos con los que permite trabajar esta herramienta, así como algunas de las restricciones que presentan son:
  - CSV: ficheros de texto cuyos campos o variables están separados por algún signo de puntuación o símbolo especial, por ejemplo “;”, “,”, “%”, etc.
  - TAB: ficheros de texto cuyos campos o variables se encuentran separados por tabulaciones.
  - PLANO: ficheros de texto cuyos campos o variables se encuentran en una posición determinada. Lo habitual es que para este tipo de ficheros exista un diseño de registro en el que se indique, entre otra información, la posición y número de caracteres (de texto o numéricos) que conforman cada variable del fichero.
  - EXCEL: ficheros de MS Excel v.2007 y anteriores.
  - MySQL: tablas presentes en bases de datos MySQL.

- ACCESS: ficheros de MS Access v.2000-2003-2007.
- ODS: ficheros de Libre Office Calc v3.6 y anteriores.
- DBF: tablas presentes en bases de datos DBF.

- **Fichero a tratar**: en este apartado el usuario indicará la ruta en la que se ubica el fichero que desea tratar. Para ello tendrá que pulsar el botón **Examinar**. Una vez seleccionado el fichero, en el área "Selección de campos de salida" de la interfaz se visualizarán todos las variables o campos del mismo.

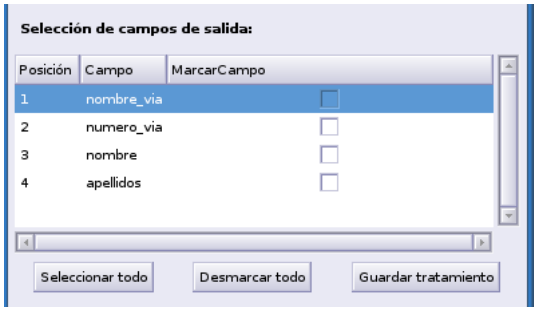


Imagen 21. Campos del fichero original a tratar

- **Utilizar un tratamiento definido previamente**: seleccionando esta opción el usuario podrá utilizar un tratamiento que haya realizado anteriormente, ya sea a una versión anterior del mismo fichero o a otro con un diseño de registro equivalente. Si se marca esta opción, habrá que especificar la ubicación del fichero en el que se encuentra guardado dicho tratamiento. Al cargarse el tratamiento predefinido, en el área "Selección de campos de salida" de la interfaz se visualizarán todos las variables o campos tal y como se guardaron en su momento.

  Esta función resulta útil cuando, por ejemplo, se reciben actualizaciones periódicas de un fichero de datos, ya que en este caso se reduce la carga de trabajo del usuario, no teniendo éste que especificar nuevamente los parámetros del proceso de tratamiento.

- **Configuración del fichero de salida tratado**: en esta sección se define la estructura que va a tener el fichero de salida tratado. Así, el usuario puede decidir entre mantener la misma estructura que la del fichero original, esto es, mantener los mismos campos, en el mismo orden y con la misma denominación, o establecer una nueva estructura, es decir, seleccionar un número menor de campos para el fichero de salida tratado, modificar su orden y denominación, o en el caso de que el fichero original no tuviera cabecera el usuario podría definirla. Exactamente las posibilidades que ofrece esta sección son:

  - Cabecera del fichero de salida:

    - *Conservar la cabecera actual*: es la opción que la aplicación tiene marcada por defecto. Si se deja seleccionada el fichero de salida tratado mantendrá la misma cabecera que el fichero original. **¡OJO!** Si se opta por esta

opción, el usuario deberá comprobar antes de ejecutar el proceso que la denominación de los campos de la cabecera no coinciden con los del fichero de salida normalizado. Estos últimos se pueden consultar en el Anexo IV. Si coinciden obligatoriamente deberá modificarlos para que no se produzca un error al normalizar el fichero. La manera de proceder en este caso se explica justo debajo.

- *Editar la cabecera actual*: si el usuario selecciona esta opción las variables o campos del fichero a tratar mostradas en el área "Selección de campos de salida" son editables, pudiéndose modificar la denominación de todas o alguna de ellas. Para editar las variables basta con hacer doble clic sobre la que se desee modificar y especificar la nueva denominación. **¡OJO!** Para que el cambio sea efectivo es necesario pulsar ENTER o sobre cualquier otra de las variables del fichero. Se aconseja que en la nueva denominación no se usen tildes para evitar problemas de codificación.

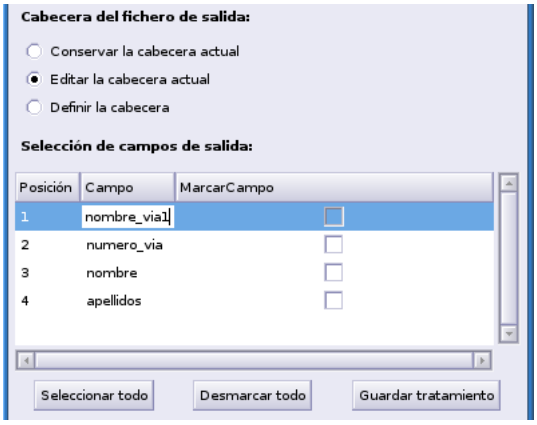


Imagen 22. Edición de uno de los campos del fichero original

- *Definir la cabecera*: esta opción se utilizará cuando el fichero original a tratar no disponga de cabecera. En este caso en el área "Selección de campos de salida" las variables o campos que se muestran corresponden a la primera fila del fichero original y será el usuario el que observando los valores de las variables o con el diseño de registro del fichero original, el que establezca su denominación. Al igual que en el caso anterior para que los cambios en la denominación de las variables sean efectivos, es necesario pulsar ENTER o sobre cualquier otra variable del fichero y nuevamente se aconseja que al indicar la denominación no se usen tildes.

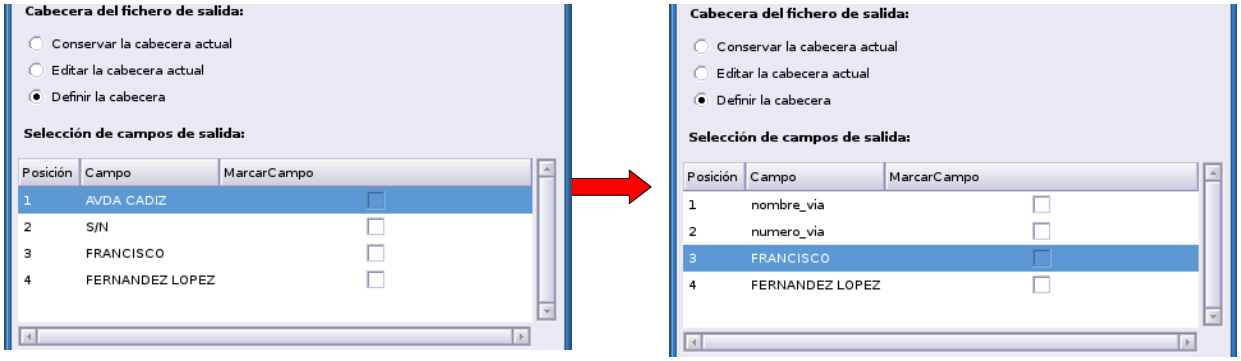


Imagen 23. Definición de campos de un fichero de datos sin cabecera

– Selección de campos de salida: esta sección muestra todas las variables o campos que componen el fichero original. Para cada una de ellas se muestra el orden en el que aparecen en el fichero original (***Posición***), su denominación o en el caso de que el fichero original no tenga cabecera el valor de la variable (***Campo***) y una casilla de selección que permite al usuario decidir si incluye o no dicho campo en el fichero de salida tratado (***MarcarCampo***).

Si el usuario desea que todos los campos del fichero original estén en el fichero de salida tratado, simplemente tendrá que pulsar el botón **Seleccionar todo**. Por el contrario, si tiene todas las variables seleccionadas pero solamente desea marcar algunas de ellas, puede pulsar el botón **Desmarcar todo** y a continuación seleccionar una a una cada variable. Se recomienda incluir alguna variable más aparte de la que se va a normalizar, si es posible una que identifique unívocamente a cada registro. El motivo es que podría darse el caso de que al tratar el fichero se desordenaran los registros, con lo cual sería más complicado añadirle la información posteriormente.

Por otra parte, en esta sección también se permite al usuario modificar el orden en el que van a aparecer las variables en el fichero de salida tratado. Para llevar a cabo este proceso tendrá que pulsar sobre la variable o campo que desee modificar de posición y arrastrarla hasta la posición en la que quiera ubicarla.

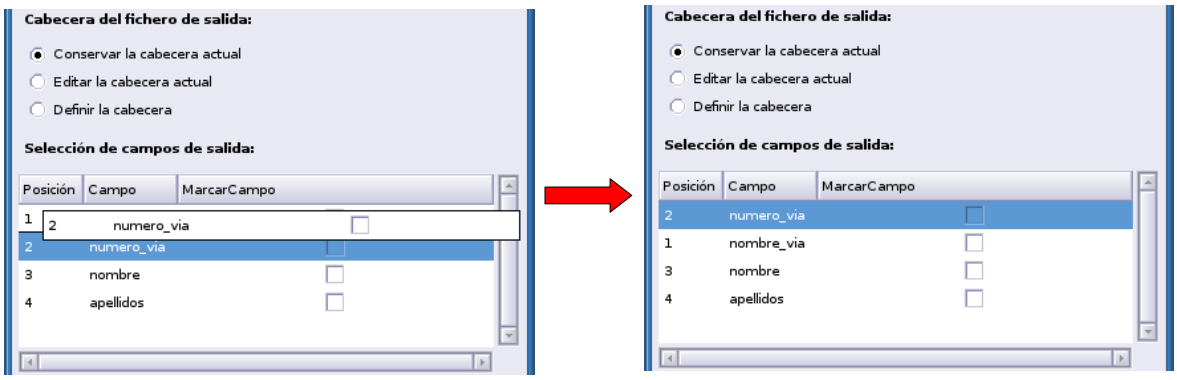


Imagen 24. Ordenación de campos

Además, en esta sección se incluye un botón que permite guardar los parámetros establecidos en el proceso de

tratamiento. De esta forma, si posteriormente se tiene que tratar un fichero similar se podrá cargar el tratamiento predefinido marcando el botón al que se ha hecho referencia anteriormente: **Utilizar un tratamiento definido previamente**.

Así, si se pulsa el botón **Guardar tratamiento** se abrirá una ventana de navegación del tipo:

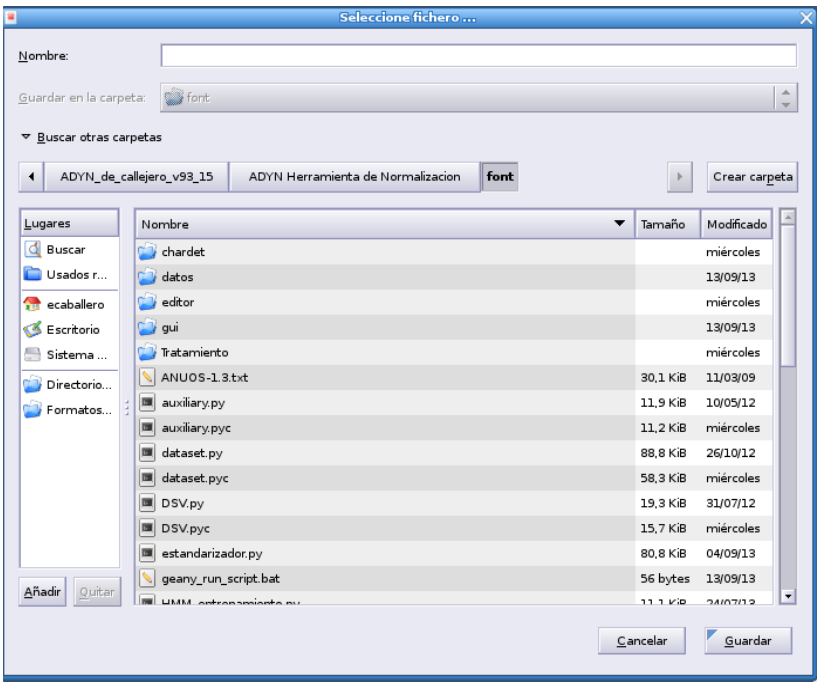


Imagen 25. Ventana para guardar tratamiento

en la que el usuario especificará la denominación con la que quiere guardar el fichero y la ruta donde desea ubicarlo. La denominación del fichero no requiere ninguna estructura ni extensión específica, así, por ejemplo, si se está tratando el fichero *direcciones_centroseducativos.csv*, el fichero de tratamiento podría denominarse *tratamiento_direcciones_centroseducativos*. **¡OJO!** Si no se especifica la ruta donde se quiere guardar el tratamiento, el fichero se guardará en la carpeta app que contiene el código de la aplicación.

- **Ejecución:** con este último apartado el usuario puede ejecutar el tratamiento o salir de la interfaz. Si elige ejecutar el tratamiento le aparecerá una ventana de navegación de archivos donde debe indicar la denominación con la que desea guardar el fichero tratado. En cuanto a la denominación del fichero no existe ninguna restricción pero su **extensión** tiene que ser **obligatoriamente .csv**.

Tras el análisis de la interfaz de tratamiento se procede a explicar cómo se lleva a cabo la transformación de cada fichero a uno de tipo CSV cuyos campos o variables se encuentran separados por el carácter ";".

Para finalizar se vuelve a hacer hincapié en que este proceso de tratamiento es **OBLIGATORIO** para todos los ficheros, independientemente de que el fichero con el que se trabaje ya tenga formato CSV y sus elementos estén separados por ";".

6.3.1.1 Tratamiento de un fichero CSV

La Herramienta de Normalización y la de Enlace trabaja con ficheros de texto CSV separados por el carácter ";". De esta manera se podría pensar que los ficheros que ya se encuentran en dicho formato no necesitan un tratamiento inicial. Este planteamiento no es totalmente cierto ya que aunque el fichero utilice como separador el carácter ";", podría ocurrir que tuviera algún problema de codificación de caracteres que interesara corregir. Así que esta tarea es necesaria realizarla para cualquier tipo de fichero CSV.

Para tratar ficheros de este tipo con la interfaz de tratamiento, en *Formato del fichero de datos a tratar* el usuario seleccionará la opción CSV y en *Fichero a tratar* pulsará el botón **Examinar** para incluir la ubicación del mismo.

Al especificar estos elementos, aparecerá la ventana que se muestra debajo en la que el usuario deberá especificar el separador de campo que se está utilizando para separar las variables o campos del fichero original (";", ",", "%", etc.). Si se desconoce el símbolo o carácter que se está utilizando como separador se podrá abrir el fichero utilizando algún editor de texto, como por ejemplo, 'Notepad2' para Windows o 'Gedit' o cualquier editor similar para Linux.

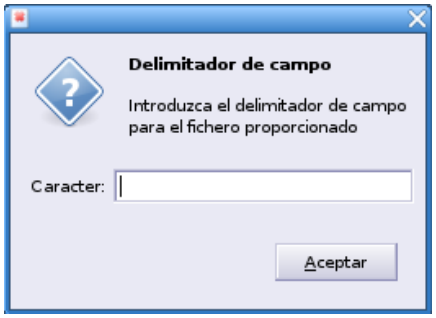


Imagen 26. Ventana de delimitador de campos del fichero original

Una vez incluido el separador y pulsando **Aceptar**, en el área de "Selección de campos de salida" se mostrarán todas las variables o campos del fichero a tratar, tal y como se muestra en la siguiente imagen:

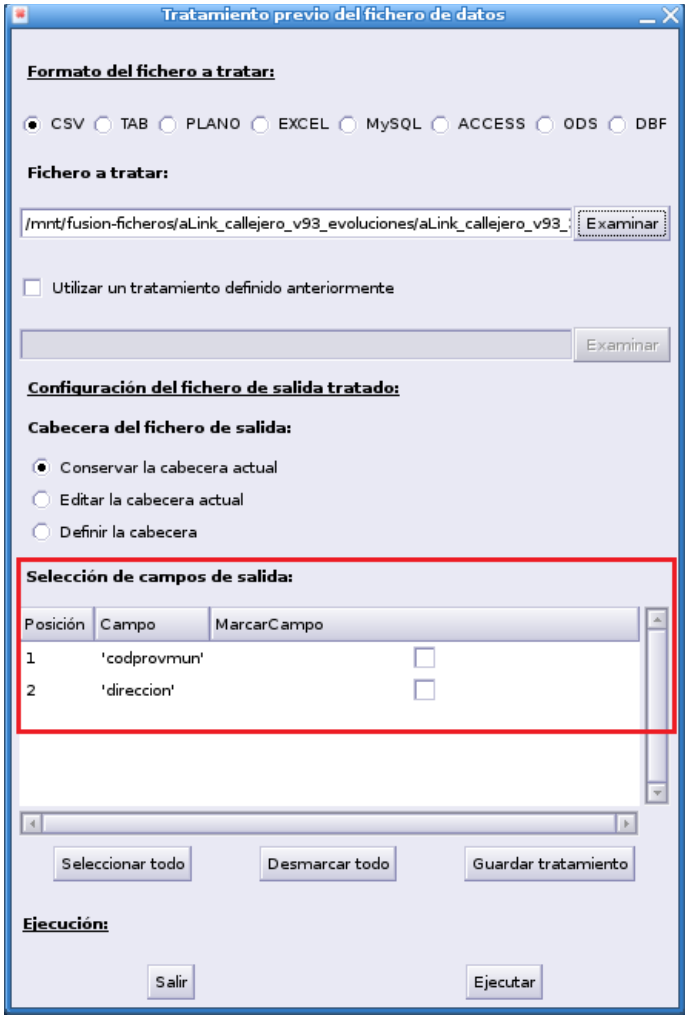


Imagen 27. Interfaz de tratamiento de un fichero CSV

En este momento, será el usuario el que decida si todos los campos del fichero original van a pasar a formar parte del fichero tratado, sin más que pulsar el botón **Seleccionar todos** o si por el contrario únicamente selecciona determinados campos. También decidirá entre realizar una nueva ordenación de los campos del fichero tratado o no realizarla y/o realizar un cambio de denominación de los mismos o no llevarlo a cabo. Tras realizar todas las operaciones que considere oportunas, el usuario podrá guardar los parámetros establecidos por si desea utilizarlos posteriormente, para ello debe pulsar el botón **Guardar tratamiento**.

Por ejemplo, si el usuario desea que el fichero de salida tratado contenga solamente el campo 'direccion' y decide guardar el tratamiento, entonces deberá marcar únicamente la celdilla relativa a este campo y a continuación pulsar el botón **Guardar tratamiento** tal y como se observa en la siguiente imagen:

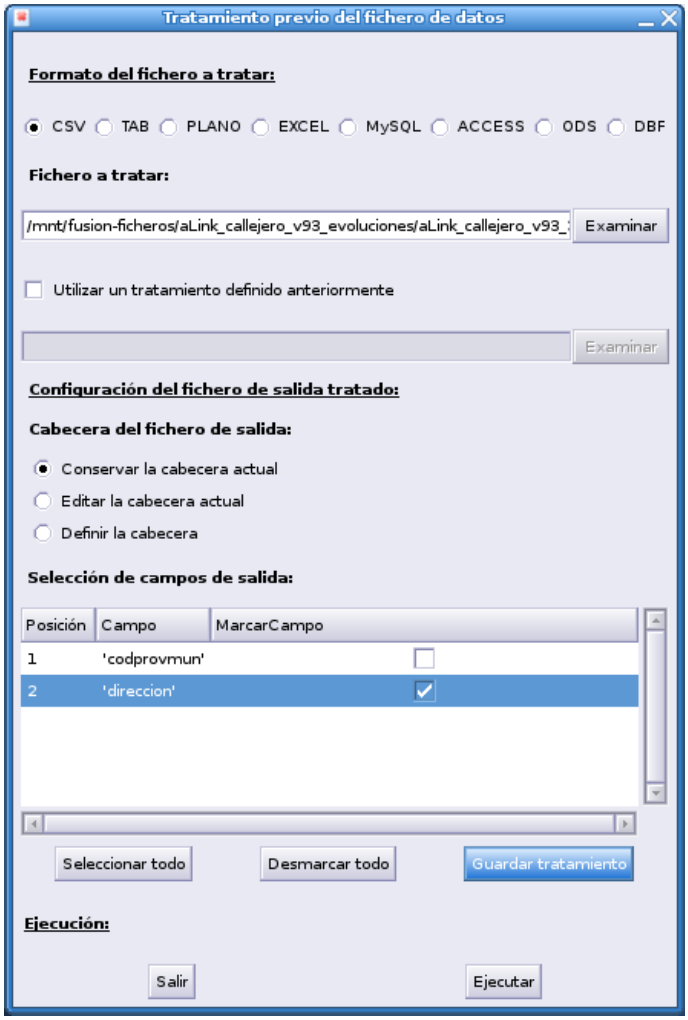


Imagen 28. Interfaz de tratamiento de un fichero CSV. Guardar tratamiento

Automáticamente se le abrirá una ventana de navegación de archivos en la que el usuario indicará el nombre y la ruta con la que quiere guardar este tratamiento. Tal y como se explicó en el apartado 6.3.1 de este Manual, la denominación y ubicación de este fichero es a libre elección del usuario y **no debe tener** ningún tipo de extensión. Por ejemplo, si se está tratando el fichero *direcciones_establecimientos.csv*, el fichero tratado podría denominarse *tratamiento_direcciones_establecimientos*.

Por último, el usuario debe pulsar el botón **Ejecutar** para llevar a cabo el tratamiento. En este caso, le aparecerá una nueva ventana de navegación de archivos como la que se muestra abajo y en ella deberá indicar el nombre con el que se va a guardar el fichero de datos tratado así como su ubicación.

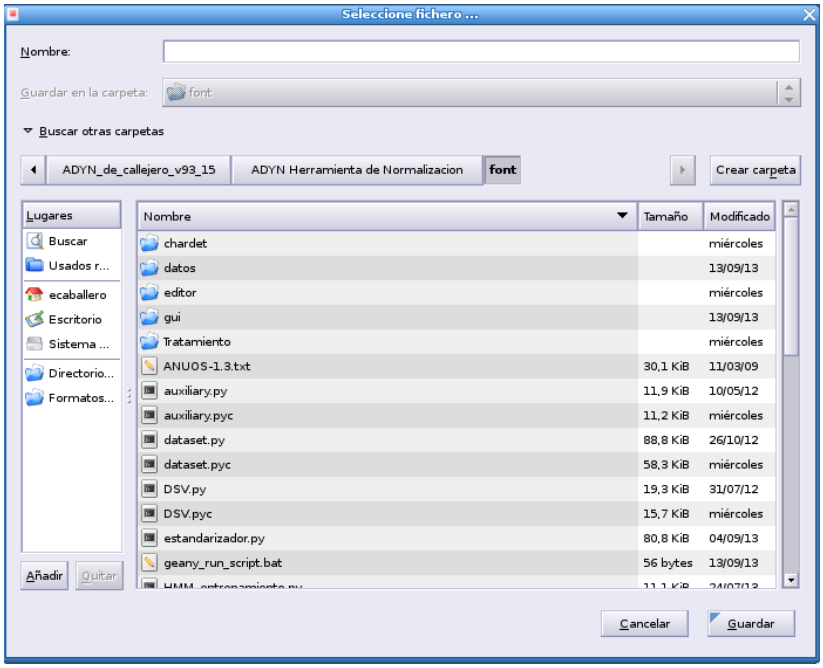


Imagen 29. Ventana para guardar el fichero tratado

Hay que decir que en cuanto a la denominación del fichero tratado no existen restricciones pero sí en lo que se refiere a su extensión, que **obligatoriamente** tiene que ser **".csv"**. Por ejemplo, siguiendo con el ejemplo anterior, si se está tratando el fichero *direcciones_establecimientos.csv*, el fichero tratado podría denominarse *direcciones_establecimientos_tratado.csv*.

Tras indicar la denominación y ubicación del fichero tratado y pulsar el botón **Guardar**, se abrirá la ventana de delimitador de campos que se muestra a continuación:

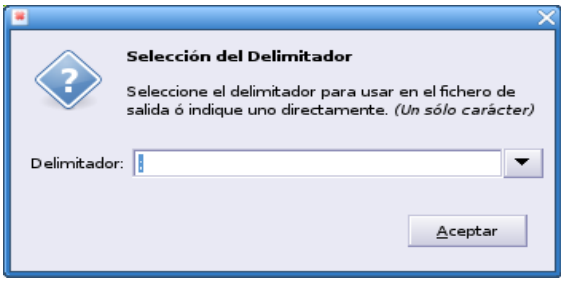


Imagen 30. Ventana de delimitador de campos del fichero tratado

En ella el usuario **obligatoriamente** tendrá que seleccionar el separador de campo **";"**, que es el requerido para trabajar con la Herramienta de Normalización. No obstante, para darle mayor versatilidad a la herramienta se han incluido otros separadores e incluso se le ofrece al usuario la posibilidad de especificar uno propio. Así, esta funcionalidad puede serle útil si necesita trabajar con algún otro programa que requiera trabajar con ficheros de tipo CSV en los que el delimitador de campo sea un carácter distinto al ";".

Finalizado el tratamiento del fichero se mostrará la siguiente ventana:

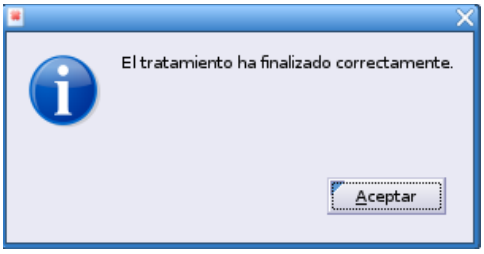


Imagen 31. Ventana de finalización de tratamiento

Por último, indicar que si el usuario no hubiera querido guardar el tratamiento entonces tras realizar todas las operaciones que considerara oportunas con las variables del fichero original (seleccionar todas o solo unas cuantas, ordenarlas, editarlas o definir su denominación), tendría que haber pulsado el botón **Ejecutar** para llevar a cabo el tratamiento.

6.3.1.2 Tratamiento de un fichero TAB

Los ficheros TAB son ficheros de texto cuyos campos o variables están separados por tabulaciones. Para tratar ficheros de este tipo, en *Formato del fichero de datos a tratar* se seleccionará la opción TAB y en *Fichero a tratar* se pulsará el botón **Examinar** para indicar la ubicación del mismo. Al establecer estos elementos, en el área de "Selección de campos de salida" se mostrarán todas las variables o campos del fichero a tratar.

A partir de aquí la forma de proceder con este tipo de ficheros será equivalente a la realizada en el tratamiento de ficheros con formato CSV.

6.3.1.3 Tratamiento de un fichero PLANO

Los ficheros de texto PLANO son aquellos formados exclusivamente por texto (únicamente caracteres) sin ningún formato. En este tipo de ficheros la información referida a cada registro se encuentra en un línea y no existe ningún tipo de separador de campo para las variables del fichero. Un ejemplo de fichero de este tipo sería:

Imagen 32. Fichero de texto plano

En estos casos se requiere el diseño de registro del fichero para conocer la posición en la que se encuentran los caracteres en los que comienza y termina cada campo. Para el ejemplo de arriba, este podría ser el diseño:

Campo	Descripción	Tipo de campo	Nº de caracteres
tipo_via	Tipo de vía	Carácter	5
nombre_via	Nombre de vía	Carácter	30
numero_via	Número de vía	Carácter	3

Tabla 1. Diseño de registro de un fichero de texto plano

Para tratar ficheros de este tipo, en *Formato del fichero de datos a tratar* se seleccionará la opción PLANO y en *Fichero a tratar* se indicará la ubicación del mismo. Al establecer estos elementos, se abrirá la siguiente ventana:

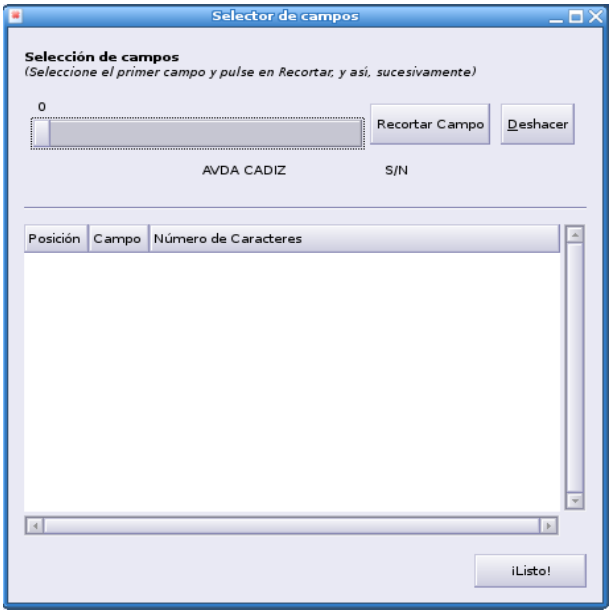


Imagen 33. Selector de campos en fichero de texto plano

Como se puede observar la ventana muestra una barra que comienza en la posición 0. Junto a ella aparecen dos botones, **Recortar Campo** y **Deshacer**. Justo debajo de estos elementos se muestra la primera fila del fichero de texto plano a tratar (AVDA CADIZ S/N) y el área donde se van a mostrar las variables que formarán el fichero original. En este área se incluyen los elementos:

- **Posición**: indica la posición que va a ocupar cada una de las variables o campos seleccionados dentro del fichero original.
- **Campo**: muestra el valor de la variable o campo seleccionado del fichero original.
- **Número de Caracteres**: indica la longitud del campo o variable, es decir, el número de caracteres que ocupa cada variable o campo dentro del fichero original.

Tras la definición de los elementos de la ventana, se indica el funcionamiento de los mismos:

La barra permite al usuario seleccionar el número de caracteres que conforman cada campo. Si el usuario pincha con el ratón sobre ella y se desplaza hacia la derecha puede ir seleccionando campos de acuerdo con el diseño de registro del fichero. Además, a la vez que se produce el desplazamiento se resaltan en color naranja los caracteres de la primera fila del fichero que corresponden a cada campo. Así, siguiendo con nuestro ejemplo, si se arrastra el ratón hasta la posición 5 que según el diseño de registro del fichero constituye el primer campo, se observará en la ventana lo siguiente:

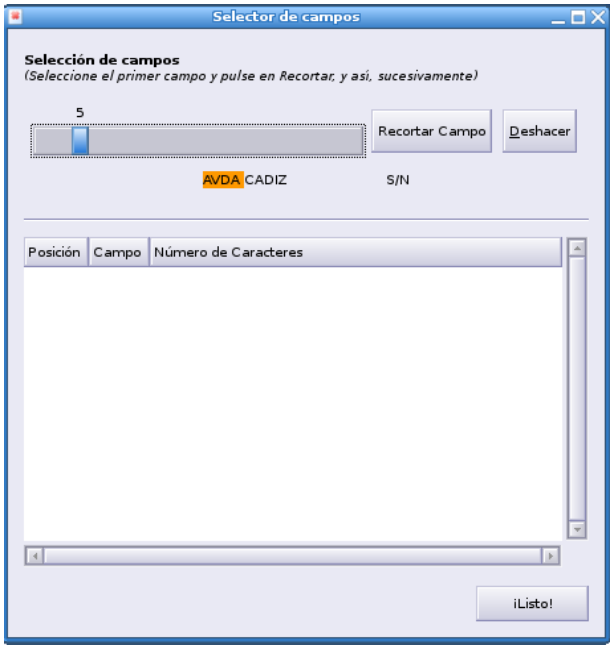


Imagen 34. Selección del primer campo de un fichero de texto plano

Si a continuación se pulsa el botón **Recortar Campo** se tiene el siguiente resultado:

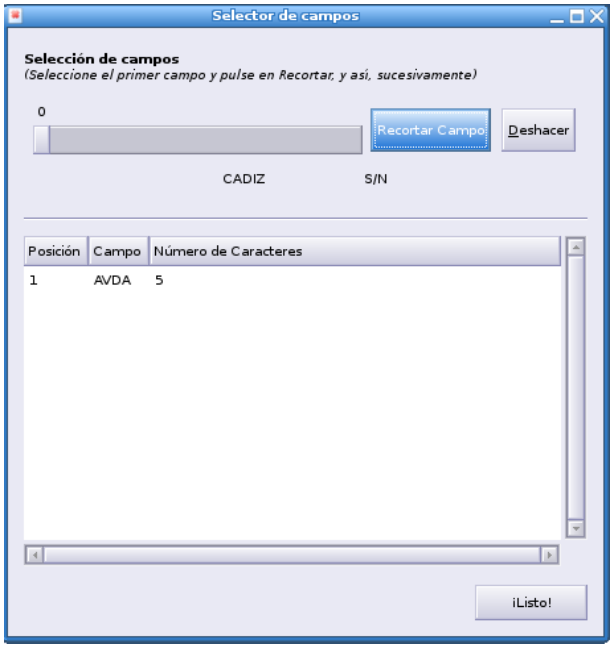


Imagen 35. Selección del primer campo del fichero de texto plano

Esto es, se ha creado la primera variable que formará parte del fichero original (ver ***Posición***), dicha variable se ha denominado automáticamente por la herramienta como AVDA (ver ***Campo***) y tiene cinco caracteres (ver ***Número de Caracteres***).

Seguidamente se recortará el siguiente campo y así sucesivamente, de forma que cuando estén segmentados todos los campos la ventana debería quedar configurada de la siguiente forma:

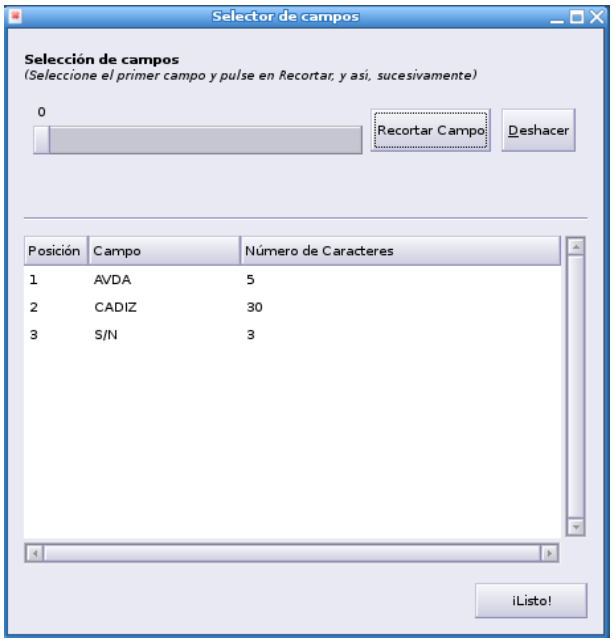


Imagen 36. Selección de todos los campos del fichero de texto plano

Si a la hora de recortar un campo se produce una equivocación en cuanto al número de caracteres seleccionado, el usuario podrá deshacer esta operación pulsando el botón **Deshacer**. Pero si por ejemplo, se hubiera producido un error al especificar el número de caracteres que ocupa el campo referido al nombre de vía, en este caso la variable CADIZ, se tendría que deshacer el recorte de la variable referida al número de la vía y posteriormente la referida al nombre de la vía.

Tras la segmentación del fichero, pulsando el botón **¡Listo!** aparecerá la interfaz de tratamiento previo con la siguiente información:

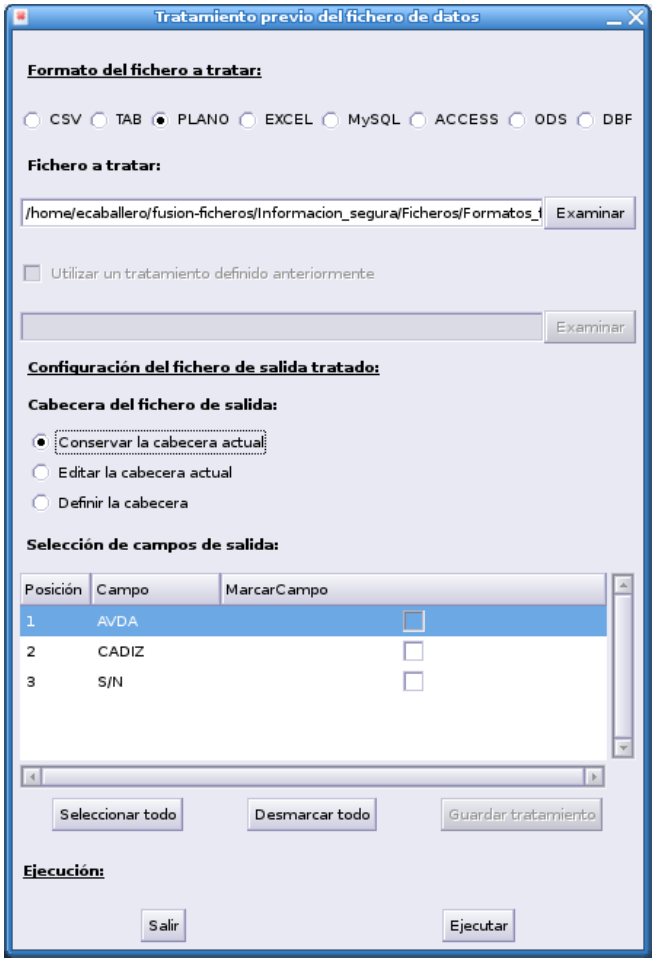


Imagen 37. Interfaz de tratamiento para un fichero de texto plano

A partir de este momento, el usuario debe definir la cabecera del fichero tal y como se ha explicado en el apartado 6.3.1 de este Manual y el resto de operaciones que se pueden realizar y ventanas que van a aparecer son equivalentes a las de los tratamientos explicados anteriormente.

Es importante resaltar que al tratar ficheros de este tipo no es posible guardar el tratamiento realizado, por lo tanto el botón **Guardar tratamiento** está deshabilitado. El motivo se debe a la manera de trabajar que tiene la herramienta de tratamiento previo con este tipo de ficheros. Nótese que en cuanto se selecciona en *Formato del fichero a tratar* la opción PLANO y se indica la ubicación del fichero aparece directamente la ventana del selector de campos, con lo cual hay que seleccionar obligatoriamente cada uno de ellos cada vez que se abre un fichero de este tipo y no se tiene opción de utilizar un tratamiento anterior.

6.3.1.4 Tratamiento de un fichero EXCEL

Para tratar ficheros de tipo EXCEL, en *Formato del fichero de datos a tratar* el usuario seleccionará la opción EXCEL y en *Fichero a tratar* incluirá la ubicación del mismo. Al especificar estos elementos, aparecerá una ventana en la que el usuario deberá especificar

la hoja del fichero Excel en la que se encuentran los datos. Tal ventana se muestra a continuación:

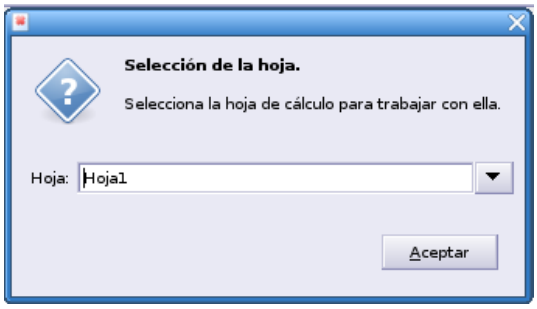


Imagen 38. Ventana de selección de hoja de datos en Excel

Una vez indicada la hoja y pulsando **Aceptar**, en el área de "Selección de campos de salida" se mostrarán todas las variables o campos del fichero a tratar. A partir de aquí la forma de proceder con este tipo de ficheros será equivalente a la realizada en el tratamiento de ficheros con formato CSV.

6.3.1.5 Tratamiento de un fichero MySQL

Para tratar ficheros de tipo MySQL, en *Formato del fichero de datos a tratar* se seleccionará la opción MySQL y en *Fichero a tratar* al pulsar en el botón **Examinar** aparecerá la siguiente ventana:

Imagen 39. Ventana de conexión al servidor MySQL

En ella el usuario deberá especificar:

- **Host**: nombre del servidor MySQL en donde se encuentran los datos.
- **Usuario**: nombre del usuario para acceder al servidor MySQL.
- **Contraseña**: contraseña del usuario para acceder al servidor MySQL.

Tras especificar estos requerimientos y pulsar el botón **Aceptar**, se abrirá la siguiente ventana que permite indicar la base de datos

del servidor en la que se encuentra la información a tratar:

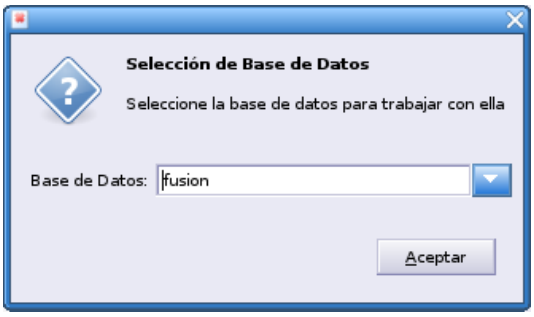


Imagen 40. Ventana de selección de base de datos de MySQL

Una vez seleccionada la base de datos y pulsado el botón **Aceptar**, aparecerá una nueva ventana para seleccionar la tabla de la base de datos en la que se encuentra la información. Dicha ventana tiene el siguiente aspecto:

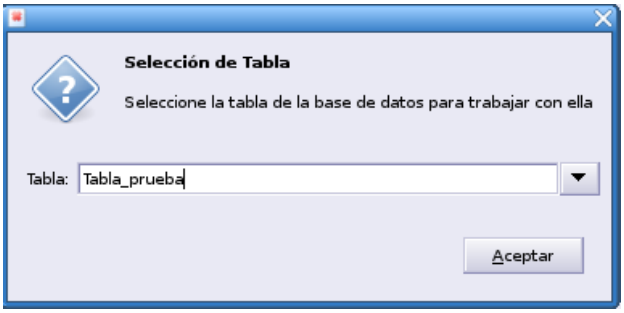


Imagen 41. Ventana de selección de tabla de base de datos de MySQL

A continuación, pulsando **Aceptar** en el área de "Selección de campos de salida" de la interfaz de tratamiento previo se mostrarán todas las variables o campos del fichero a tratar.

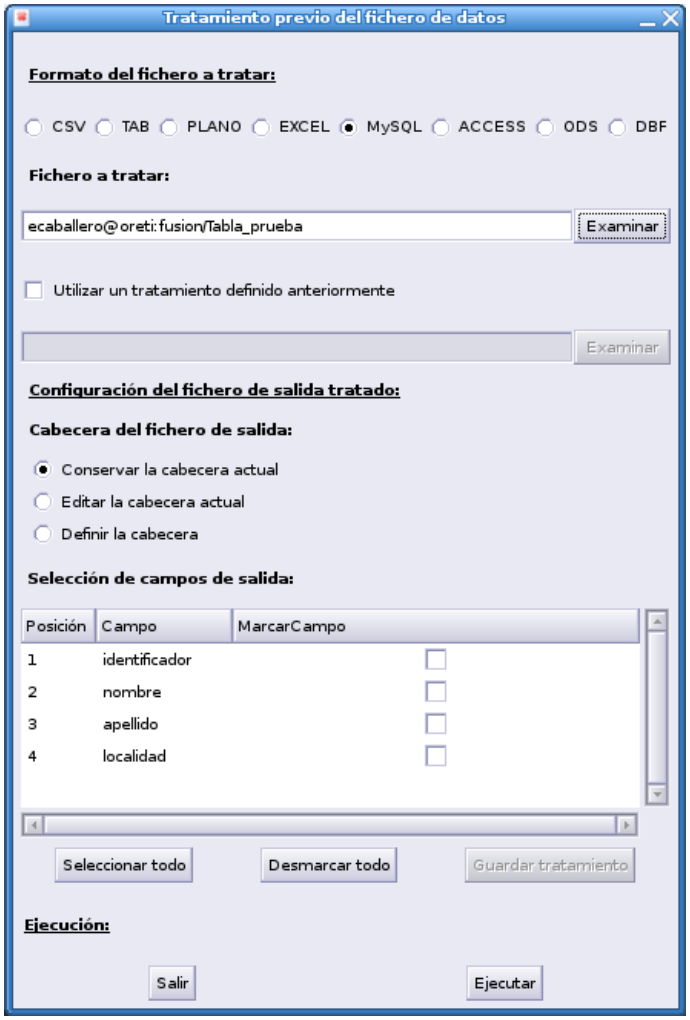


Imagen 42. Interfaz de tratamiento de tabla de base de datos de MySQL

A partir de aquí la forma de proceder con este tipo de ficheros será equivalente a la realizada en el tratamiento de ficheros con formato CSV, es decir, se puede conservar o editar la denominación de los campos o variables del fichero original, se pueden seleccionar todas o solo algunas de las variables del fichero original para que formen parte del fichero de salida tratado, se pueden ordenar las variables y ejecutar el tratamiento. La única diferencia que existe con respecto al tratamiento de los ficheros con formato CSV es que en este caso, al igual que para los ficheros de texto plano, el botón **Guardar tratamiento** está deshabilitado. El motivo se debe a cómo se ha configurado la herramienta de tratamiento previo para acceder a una tabla de este tipo de bases de datos.

6.3.1.6 Tratamiento de un fichero ACCESS

Para tratar ficheros de tipo ACCESS con la *Herramienta de Normalización* es **obligatorio** que **todos los campos** de la tabla de la base de datos en la que se encuentra la información sean de **tipo texto**.

Una vez controlada esta situación, en *Formato del fichero de datos a tratar* el usuario seleccionará la opción ACCESS y en *Fichero a tratar* incluirá la ubicación del mismo. Al especificar estos elementos, aparecerá una ventana en la que el usuario deberá especificar

la tabla de la base de datos en la que se encuentra la información. Tal ventana se muestra a continuación:

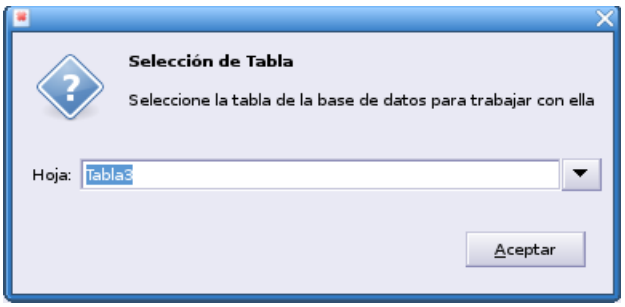


Imagen 43. Ventana de selección de tabla de una base de datos ACCESS

Una vez indicada la tabla y pulsando **Aceptar**, en el área de "Selección de campos de salida" se mostrarán todas las variables o campos del fichero a tratar. A partir de aquí la forma de proceder con este tipo de ficheros será equivalente a la realizada en el tratamiento de ficheros con formato CSV.

6.3.1.7 Tratamiento de un fichero ODS

Para tratar ficheros de tipo ODS, en *Formato del fichero de datos a tratar* se seleccionará la opción ODS y en *Fichero a tratar* se incluirá la ubicación del mismo. Al especificar estos elementos, aparecerá una ventana en la que el usuario deberá especificar la hoja del fichero ODS en la que se encuentran los datos. Tal ventana se muestra a continuación:

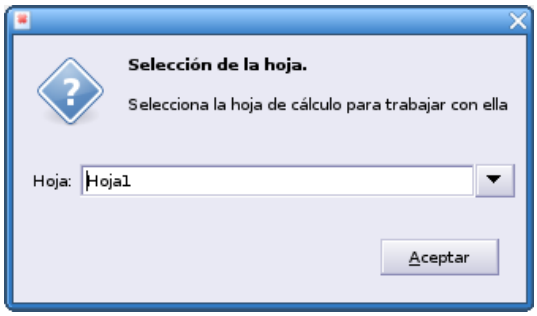


Imagen 44. Ventana de selección de hoja dedatos en ODS

Una vez indicada la hoja y pulsando **Aceptar**, en el área de "Selección de campos de salida" se mostrarán todas las variables o campos del fichero a tratar. A partir de aquí la forma de proceder con este tipo de ficheros seráequivalente a la realizada en el tratamiento de ficheros con formato CSV.

6.3.1.8 Tratamiento de un fichero DBF

Los ficheros DBF son ficheros de tipo dBASE. Para tratar ficheros de este tipo, en *Formato del fichero de datos a tratar* se seleccionará la opción DBF y en *Fichero a tratar* se pulsará el botón **Examinar** para indicar la ubicación del mismo. Al establecer estos elementos, en el área de "Selección de campos de salida" se mostrarán todas las variables o campos del fichero a tratar.

A partir de aquí la forma de proceder con este tipo de ficheros será equivalente a la realizada en el tratamiento de ficheros con formato CSV.

6.3.2 HMM: Selección de la muestra

La Herramienta de Normalización incluye Modelos Ocultos de Markov para nombres de personas, direcciones postales e identificadores de personas físicas y/o jurídicas (como se comentó anteriormente en el Anexo II se pueden consultar los mismos). No obstante, cuando el usuario considere que los Modelos Ocultos de Markov disponibles no realizan una correcta segmentación de los elementos que componen el campo a normalizar de su fichero de datos, deberá crear su propio modelo. Hay que aclarar que para identificadores de personas físicas y/o jurídicas no sería necesario construir el modelo HMM, ya que para este caso particular se suministra un modelo útil para cualquier fichero de datos.

Para crear un Modelo Oculto de Markov el usuario deberá realizar los siguientes pasos:

1. Selección y etiquetado de una muestra del fichero de datos que contenga el campo a normalizar. El campo debe contener nombres de personas, direcciones postales o identificadores de personas físicas y/o jurídicas.
2. Asignación manual de estados a cada uno de los elementos que componen el campo a normalizar.
3. Entrenamiento de la muestra para conocer la estructura de segmentación de los elementos del campo a normalizar y extrapolar ese conocimiento al fichero completo.

Para llevar a cabo el primer paso, el usuario tendrá que utilizar la herramienta *HMM: Selección de la muestra*. A continuación se muestran su interfaz y los elementos de la misma:

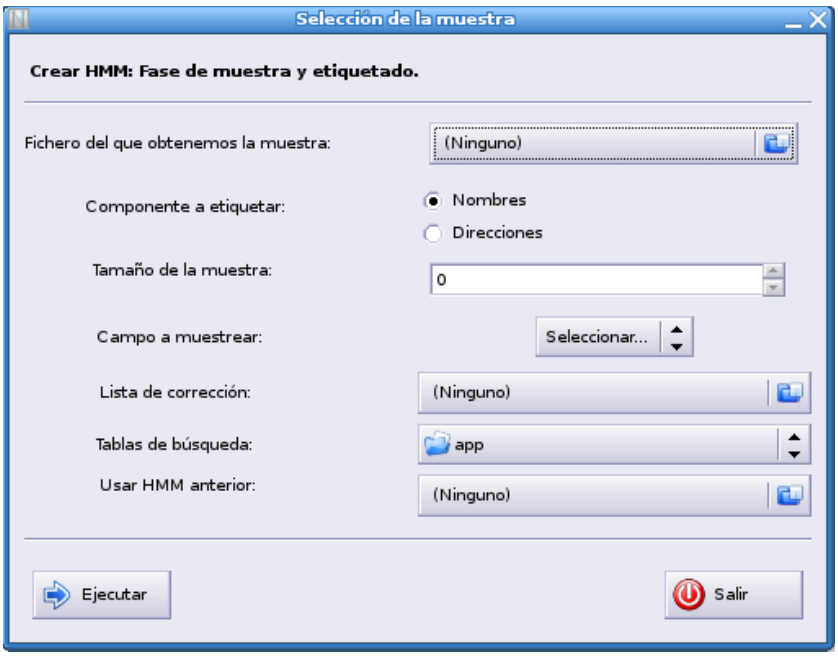


Imagen 45. Interfaz de selección de la muestra

- **Fichero del que obtenemos la muestra**: al pulsar este botón el usuario indicará la ruta en la que se encuentra el

fichero de datos del que se va a extraer la muestra. Dicho fichero será el que se ha tratado inicialmente con la herramienta de *Tratamiento previo* y tiene que tener extensión *.csv*. También se puede utilizar otro previamente tratado con un diseño de registro exactamente igual al que se quiere normalizar.

- **Componente a etiquetar**: aquí el usuario indicará si va a etiquetar un campo de la muestra que contiene nombres de personas o direcciones postales. No se incluye la opción de identificadores de personas físicas y/o jurídicas porque para este caso la aplicación ya ofrece un modelo HMM definitivo y por tanto no habría que construirlo.

  El proceso de etiquetado consiste en lo siguiente: la herramienta irá buscando en la muestra cada uno de los elementos del campo a normalizar y los irá etiquetando de acuerdo con las etiquetas asociadas a cada una de las tablas de búsqueda. Si los encuentra en alguna de ellas, les asignará la etiqueta correspondiente a la tabla en la que se ha localizado y si no la aplicación les asignará alguna de las etiquetas no asociadas a ninguna tabla de búsqueda. Estas últimas, se asignan a elementos como valores numéricos, palabras de una sola letra y elementos no encontrados en ninguna de las tablas de búsqueda. En el Anexo V se muestran las etiquetas usadas en el proceso de normalización para nombres de personas y direcciones postales.

  Por ejemplo, en el caso de direcciones postales, suponiendo que se extrayera una muestra con una única dirección postal del tipo: *C/ Leonardo Da Vinci 22 San Fernando Cádiz*, la herramienta intentará detectar los valores de la misma (C/, Leonardo, Da, etc.) dentro de las tablas de búsqueda de direcciones postales de la Herramienta de Normalización. Así, el elemento 'C/' lo localizará en la tabla de búsqueda de tipos de vía, con lo cual la herramienta le asignará la etiqueta TV, los elementos 'Leonardo', 'Da' y 'Vinci' no los va a detectar en ninguna tabla de búsqueda de direcciones postales, por lo tanto les asignará a cada uno de ellos la etiqueta UN (*unknown*, de desconocido en inglés), al elemento 22 le asignará la etiqueta correspondiente a valores numéricos, que es NU, el elemento 'San Fernando' lo localizará en la tabla de búsqueda de municipios, con lo cual le asignará la etiqueta MU mientras que al elemento 'Cádiz' lo podrá localizar tanto en la tabla de búsqueda de municipios como en la de provincias, con lo cual le asignará respectivamente las etiquetas MU y PR que son las correspondientes a las mismas.

- **Tamaño de la muestra**: en esta sección el usuario indicará el tamaño de la muestra que va a seleccionar. La selección de la muestra se lleva a cabo automáticamente por la aplicación utilizando un muestreo aleatorio simple con reposición, por lo que podría darse el caso de que un mismo elemento aparezca más de una vez en la muestra.

  El valor por defecto que tiene este campo es cero pero al especificarse la ubicación del fichero de datos del que se va a extraer la muestra (fichero de salida tratado) cambia al valor uno. En cuanto al número de elementos a incluir en la misma dependerá de lo heterogéneos que sean los valores del campo a normalizar. A mayor heterogeneidad mayor tiene que ser el tamaño de la muestra, teniendo en cuenta que éste como máximo será igual al tamaño del fichero de datos menos uno. No obstante, independientemente del tipo de componente que se esté etiquetando (nombres de personas o direcciones postales), se recomienda empezar con un tamaño de muestra no muy elevado, por ejemplo, entre 10 y 20 elementos para posteriormente, si es necesario, ir enriqueciendo la muestra con nuevos elementos.

- **Campo a muestrear**: este combo contiene todos los campos del fichero tratado con la herramienta de *Tratamiento*

previo. Estos se cargan directamente cuando el usuario especifica la ruta de ubicación del fichero tratado. De entre ellos, el usuario seleccionará el campo del que desea extraer la muestra.

- **Lista de corrección**: en este botón el usuario indicará la ubicación en la que se encuentran las listas de corrección proporcionadas en la aplicación. En concreto, están ubicadas en el directorio *alink\app\listas_tablas\listas_de_correccion*. De entre ellas se seleccionará la correspondiente al tipo de componente que se va a etiquetar: *nombres_correccion.lst* para nombres de personas o *direcciones_correccion.lst* para direcciones postales.
- **Tablas de búsqueda**: en este botón el usuario indicará la ubicación del directorio en el que se encuentran las tablas de búsqueda proporcionadas en la aplicación. Exactamente, están ubicadas en el directorio *alink\app\listas_tablas\tablas_de_busqueda*. De entre ellas se seleccionará la correspondiente al tipo de componente que se va a etiquetar: *tbl_nombre* para nombres de personas o *tbl_direccion* para direcciones postales.
- **Usar HMM anterior**: permite al usuario indicar la ubicación de un Modelo Oculto de Markov ya creado a partir de un fichero que tenga un diseño de registro idéntico al fichero a normalizar. En el Anexo VI se pueden consultar más detenidamente las ventajas de utilizar esta opción.
- **Ejecutar**: con este botón el usuario puede ejecutar el proceso de selección y etiquetado de la muestra.
- **Salir**: este botón proporciona al usuario la posibilidad de salir de la herramienta *HMM: Selección de la muestra*.

Para realizar el proceso de selección de la muestra es obligatorio especificar todos los elementos anteriores salvo el relativo a **Usar HMM anterior**.

Así por ejemplo, si para direcciones postales como las que aparecen en la siguiente imagen:

	A
1	'direccion'
2	c/ juan ramon jimenez. 34
3	avda. de andalucia. 36. bj. pta. drcha.
4	c/ barrionuevo. numero 8
5	alameda del parral. s/n
6	c/corredera. numero 26
7	ctra. alanis-cazalla. s/n
8	juan de leon. 25
9	ctra alanis fuenteobejuna km
10	carretera de la ermita. s/n
11	autovia a-376. km. 11.210
12	centro comercial los alcores planta baja
13	c/ arahal. 7
14	urb. sevilla golf. c/ eagle. 9. planta 1. pta. c
15	orense. s/n
16	garci perez de vargas. numero 13
17	cl mairena 8
18	ctra. alcala-utrera. km 2+5
19	c/ malasmakkanas. 74
20	c/ arahal. 16 -bj. b
21	cr alcala-utrera. km 9
22	c/ juez perez diaz. s/n
23	av de portugal.s/n 0
24	cl silos. s/n
25	c/ benagila. 33
26	cl pepe luces 00004 b
27	c/ pepe luces. esq. c/ luna 1
28	cl cervantes 00003
29	c/ joaquin vals sevillano. 10
30	c/ silos. numero 63
31	cl ntra.sra.aguila 00014
32	c/ manuel de falla. 30 local
33	cl pescaderia 00002
34	c/ mairena. 46 local 6
35	pl de los pescadores 1
36	c/ paseo juan carlos i. s/n
37	av andalucia 113
38	plaza dr ramon alvarez. s/n
39	ctra a-8006 km. 18+900
40	c/ pedro cano amores. 8
41	c/ manuel moreno geniz.39

Imagen 46. Fichero tratado del que se extrae la muestra

se establecen los elementos obligatorios en la interfaz *HMM: Selección de la muestra*, tal y como se observa en la siguiente imagen:

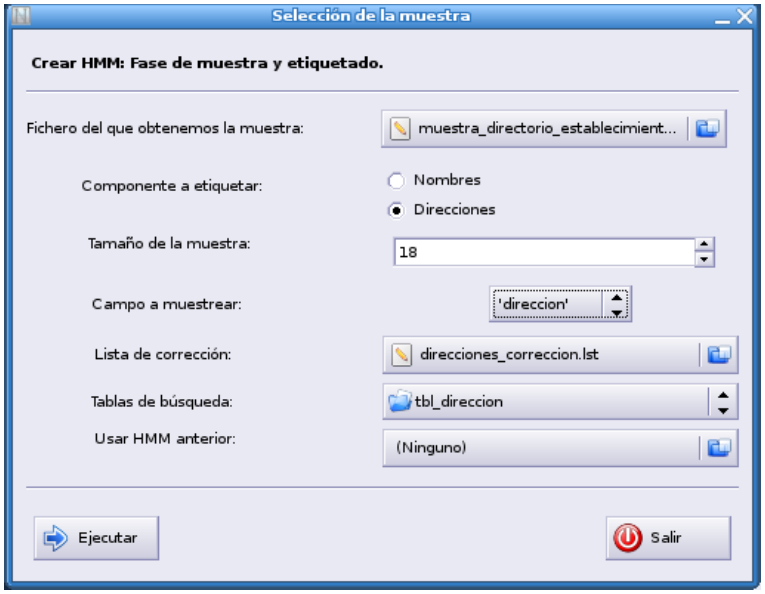


Imagen 47. Fichero tratado del que se extrae la muestra

y se pulsa el botón **Ejecutar**, al usuario le aparecerá la siguiente ventana de salida:

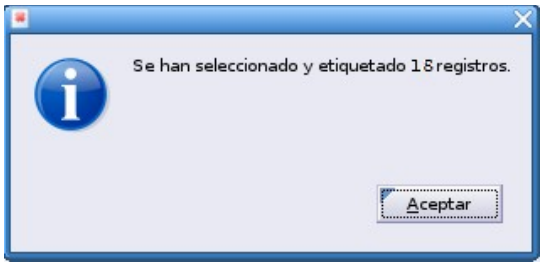


Imagen 48. Finalización del proceso de selección de la muestra

En ella se indica el número de registros del campo a normalizar que se han seleccionado y etiquetado para formar parte de la muestra.

Como resultado del proceso de selección de la muestra se generará un fichero de extensión *.csv* que contiene la muestra etiquetada. La ubicación del mismo coincide con la del fichero tratado y su denominación sigue el formato:

muestra_etiquetada_<fecha_creación>-<hora_creación>_<denominación_fichero_tratado>.csv

Por ejemplo, si el día 2 de octubre de 2013 a las 13:07 horas se hubiera extraido una muestra del fichero tratado *ejemplo_tratado.csv*, se generaría un fichero denominado *muestra_etiquetada_20131002-1307_ejemplo_tratado.csv*.

No obstante, una vez generado el fichero el usuario podría dirigirse a la ubicación en la que se encuentra el mismo para renombrarlo o cambiarlo de ubicación si así lo desea. Para visualizar el contenido del fichero se recomienda utilizar el editor de texto Notepad2 que se suministra con *aLink: Herramienta de Fusión de Ficheros* cuando se trabaje en un sistema operativo Windows o con Gedit cuando se trabaje en Linux. El motivo es que permite que la codificación de los ficheros con los que se trabaja sea la correcta (UTF-8) y de esa forma se evita la inserción de caracteres propios de otras codificaciones.

La estructura del fichero generado para este ejemplo de direcciones postales se visualiza en las imágenes 61 y 62:

```
1 ##########################################################################################################
2 #
3 # Creado Mon Dec  2 13:12:57 2013
4 #
5 # Fichero de entrada: /home/ecaballero/fusion-ficheros/aLink_callejero_v93_33/Ejemplo/direcciones_establecimientos_tratado.csv
6 # Fichero de salida:  /home/ecaballero/fusion-ficheros/aLink_callejero_v93_33/Ejemplo/muestra_etiquetada_20131202-1312_direcciones_establecimientos_tratado.c
7 # Componente: direccion
8 # Parámetros:
9 # - Posición del primer registro: 0
10 # - Posición del ultimo registro:    1230
11 # - Numero de registros seleccionados y etiquetados: 18
12 #
13 #
14 #################### Descripcion de las etiquetas:
15 #
16 # NOTA:
17 # Incluiremos la etiqueta y la tabla de busqueda a la que corresponde, por ejemplo, TV significa que la palabra
18 # identificada con esta etiqueta se encuentra en la tabla de busqueda de tipos de via.
19 #
20 # Etiquetas  y tablas de busqueda asociadas:
21 #
22 # AG corresponde a agrupaciones                NU corresponde a numeros (*)
23 # BL corresponde a bloques                     N5 corresponde a numeros de 5 digitos (*)
24 # CP corresponde a codigos postales            NV corresponde a naves
25 # ED corresponde a edificios                   PA corresponde a parcelas
26 # EG corresponde a entidades singulares        PR corresponde a provincias
27 # ES corresponde a escaleras                   PT corresponde a portales
28 # LE corresponde a letras (*)                  PU corresponde a puertass
29 # MU corresponde a municipios                  ST corresponde a sectores
30 # MZ corresponde a manzanas                    TV corresponde a tipos de via
31 # NM corresponde a identificadores de numeros  UN no se encuentra incluida en ninguna tabla de busqueda (*)
32 # NP corresponde a numeros de planta           ZO corresponde a zonas
33 #
34 #
35 #################### Listado de posibles estados para direcciones:
36 #
37 # tipo_de_via                          nombre_de_via
38 # identificador_de_numeracion
39 # ein                                  cein
40 # esn                                  cesn
41 # identificador_de_bloque              bloque
42 # tipo_de_edificio                     edificio
43 # identificador_de_portal              portal
44 # identificador_de_escalera            escalera
45 # identificador_de_planta              planta
46 # identificador_de_puerta              puerta
47 # identificador_de_letra               letra
48 # entidad_singular                     municipio
49 # provincia                            identificador_de_codigo_postal
50 # codigo_postal                        tipo_de_agrupacion
51 # agrupacion                           identificador_de_sector
52 # sector                               identificador_de_manzana
53 # manzana                              identificador_de_parcela
54 # parcela                              identificador_de_nave
55 # nave                                 identificador_de_zona
56 # zona                                 odub
57 ##########################################################################################################
```

Imagen 49. Cabecera del fichero con la muestra seleccionada y etiquetada

Como se puede comprobar, la parte superior del fichero está rodeada de almohadillas #. Todo lo que se escribe a la derecha de las almohadillas son comentarios con información acerca del fichero, por lo tanto dicha información no será leída ni tenida en cuenta por la herramienta por ser comentarios. En concreto, se muestra la siguiente información:

- Fecha de creación del fichero con la muestra etiquetada.
- Fichero de entrada: indica la ruta del fichero tratado del que se va a extraer la muestra. Nótese que se ha realizado automáticamente un cambio de codificación a UTF-8 para subsanar posibles problemas que se pueden presentar con la codificación de caracteres.
- Fichero de salida: indica la ruta donde se encuentra almacenado el fichero con la muestra etiquetada.
- Componente: hace referencia al tipo de campo del que se ha extraído y etiquetado la muestra, en este caso, *direccion*.
- Parámetros: indica la posición o fila en la que se encuentra el primer y último registro del fichero de datos tratado, así como el tamaño de la muestra seleccionada, en este caso 18. Nótese que la posición del primer registro se inicia en el valor 0.
- Descripción de las etiquetas: muestra todas las etiquetadas asociadas a las tablas de búsqueda, en este caso de

direcciones, así como aquellas otras que no están asignadas a ninguna tabla. En concreto, las que tienen (*) no están asociadas a ninguna tabla.

- Listado de posibles estados: muestra la lista de posibles estados que se pueden asignar a cada una de las etiquetas. En este caso el listado es para direcciones postales.

Tras toda esta información, se muestran en la siguiente imagen los registros que conforman la muestra con sus elementos etiquetados. **¡OJO!** Las etiquetas que la aplicación ha asignado automáticamente a cada elemento **NO** se modifican en ningún caso.

```
58 ##########################################################################################
59
60 # 72 (0): |camino de monasterejo. 3|
61 #         |camino de monasterejo 3|
62    TV:, UN:, UN:, NU:
63
64 # 156 (1): |valdes leal. 22|
65 #          |valdes leal 22|
66    EG:, UN:, NU:
67
68 # 418 (2): |ctra. lora del rio-la campana km 16+5|
69 #          |carretera lora_del_rio la_campana kilometro 16+5|
70    TV:, MU:, MU:, NM:, UN:
71    TV:, EG:, MU:, NM:, UN:
72
73 # 472 (3): |cl suarez trasierra 6|
74 #          |calle suarez trasierra 6|
75    TV:, UN:, UN:, NU:
76
77 # 516 (4): |avd. de sevilla. 22|
78 #          |avenida de sevilla 22|
79    TV:, UN:, MU:, NU:
80    TV:, UN:, PR:, NU:
81
82 # 876 (5): |cl doctor pedro de castro 1|
83 #          |calle doctor pedro de castro 1|
84    TV:, UN:, UN:, UN:, UN:, NU:
85
86 # 57  (6): |c/ perez galdos.4|
87 #         |calle perez galdos 4|
88    TV:, UN:, UN:, NU:
89
90 # 473 (7): |cl suarez trasierra 6|
91 #          |calle suarez trasierra 6|
92    TV:, UN:, UN:, NU:
93
94 # 132 (8): |finca los nietos. ctra nacional iv. km 509|
95 #          |finca los nietos carretera nacional iv kilometro 509|
96    ED:, UN:, UN:, TV:, UN:, UN:, NM:, NU:
97
98 # 142 (9): |ctr nacional iv madrid cadiz km 523|
99 #           |carretera nacional iv madrid cadiz kilometro 523|
100    TV:, UN:, UN:, UN:, MU:, NM:, NU:
101    TV:, UN:, UN:, UN:, PR:, NM:, NU:
102
103 # 168 (10): |centro comercial airesur. plta. baja. local lb17b|
104 #           |centro_comercial airesur planta baja local lb17b|
105    ED:, UN:, PL:, NP:, NM:, UN:
106    ED:, UN:, PL:, NP:, PU:, UN:
107
108 # 193 (11): |ctra a-455. km 4+100|
109 #           |carretera a 455 kilometro 4+100|
110    TV:, LE:, NU:, NM:, UN:
111
112 # 190 (12): |finca la cartuja. s/n.|
113 #           |finca la_cartuja sin_numero|
114    ED:, EG:, NM:
115
116 # 17  (13): |paraje cakkada del moro. numero  109|
117 #           |paraje cakkada_del_moro numero  109|
118    ZO:, EG:, NM:, NU:
119
120 # 373 (14): |poligono industrial servialsa. calle b. nave numero  9.|
121 #           |poligono_industrial servialsa calle b nave numero  9|
122    AG:, UN:, TV:, LE:, NV:, NM:, NU:
123
124 # 42  (15): |c/ camino de ronda 124 urb los pinares|
125 #           |calle camino de ronda 124 urbanizacion los pinares|
126    TV:, TV:, UN:, MU:, NU:, AG:, UN:, UN:
127
128 # 742 (16): |c/ paseo federico garcia lorca s/n polig. indust. juncaril|
129 #           |calle paseo federico garcia lorca sin_numero poligono_industrial juncaril|
130    TV:, ZO:, UN:, UN:, UN:, NM:, AG:, UN:
131
132 #38   (17): |c/ carril|
133 #           |calle carril|
134    TV:, TV:
```

Imagen 50. Registros que forman la muestra con elementos etiquetados

Para cada uno de ellos se tiene la siguiente información:

- Primera línea: comienza con una almohadilla, con lo cual todo lo que se escribe a la derecha de la misma se considera un

comentario y no se tiene en cuenta por la herramienta. A continuación, se indica la posición o fila en la que se encuentra el registro seleccionado dentro del fichero tratado (72, 156, 418, etc.), así como el número que se le asigna a ese registro en la muestra (se trata de un listado secuencial iniciado en 0 y representado por valores numéricos entre paréntesis). Nótese que como la selección de la muestra se realiza usando un muestreo aleatorio simple con reposición, los registros (3) y (7) están repetidos. Por último, se muestra entre barras el valor del campo a normalizar, por ejemplo para el primer registro: |camino de monasterejo. 3|.

- Segunda línea: al igual que la anterior comienza con una almohadilla, por lo tanto tiene la misma consideración que la primera, es decir, es un comentario. En ella se muestran entre barras los valores del campo a normalizar una vez que se le han aplicado la lista de corrección y las tablas de búsqueda, es decir, los valores estandarizados. Así, para el primer registro de este ejemplo se tiene: |camino de monasterejo 3|. Nótese que en este caso el carácter "." se ha sustituido por un espacio en blanco porque en la lista de corrección de direcciones postales se ha establecido que cuando la aplicación encuentre un punto lo sustituya por espacio en blanco. No obstante, si el usuario decide que el carácter "." no debe ser sustituido por un espacio en blanco puede modificar esta situación utilizando los editores de las listas de corrección tal y como se verá en el apartado 6.3.4 de este Manual.

- Tercera línea: esta a diferencia de las dos anteriores no comienza por el carácter #, por lo tanto, toda la información que se indique aquí va a ser tenida en cuenta por la herramienta. En concreto, esta línea contiene las etiquetas asignadas a cada uno de los elementos que componen el campo a normalizar. Como se ha comentado anteriormente, estas etiquetas **NO** se deben modificar nunca.

  Nótese, que por ejemplo, a los registros (2), (4) y (10) de la muestra se le han asignado dos posibles secuencias de etiquetas. En el primer caso esto se debe a que el valor *'lora del rio'*, que forma parte del nombre de la vía, se ha encontrado tanto en la tabla de búsqueda de municipios (etiqueta MU) como en la de entidades singulares (etiqueta EG). En el segundo caso, el valor '*sevilla*', que también forma parte del nombre de la vía, se ha encontrado tanto en la tabla de búsqueda de municipios (etiqueta MU) como en la de provincias (etiqueta PR). Mientras que para el tercer caso, el valor '*local*' se ha etiquetado como NM y PU ya que se ha localizado tanto en la tabla de búsqueda de identificadores de numeración como en la de identificadores de puerta. Ante esta situación el usuario deberá quedarse con la secuencia más probable teniendo en cuenta cómo está estructurada la información del fichero del que se ha extraido la muestra.

  Así, para los dos primeros casos el hecho de que los valores *'lora del rio'* y '*sevilla*' se etiqueten respectivamente como municipio o entidad singular o como municipio o provincia es indiferente ya que, teniendo en cuenta cómo se estructura la información del campo a normalizar, ambos elementos hacen referencia al nombre de la vía. Luego, en estos dos casos el usuario podría quedarse con cualquiera de las secuencias o mantener ambas.

  Sin embargo en el tercer caso, teniendo en cuenta cómo está estructurada la dirección, el valor '*local*' tiene más sentido que corresponda a un identificador de puerta que a un identificador de numeración, por lo que el usuario debería quedarse con la segunda secuencia de etiquetas.

Asignación manual de estados a las etiquetas

Una vez seleccionada y etiquetada la muestra el usuario tendrá que asignar a cada etiqueta su estado correspondiente. Por **estado**, se entiende el valor que identifica realmente a cada uno de los elementos del campo a normalizar. Al asignar los estados a las etiquetas lo que se pretende es que todos los elementos que tengan el mismo estado vayan al mismo campo de salida del fichero normalizado. En el Anexo VII se pueden consultar todos los estados posibles asociados a nombres de personas y direcciones postales.

La asignación de estados se realizará manualmente por el usuario y se recomienda que utilice el editor de texto Notepad2 o Gedit para evitar problemas de codificaciones. Para ello deberá tener en cuenta cómo está estructurada la información que aparece en el campo que se va a normalizar. Además, para el caso de direcciones postales se deberá tener en cuenta el tipo de desagregación de la dirección postal que se va a realizar, es decir, a medida o de acuerdo a CDAU.

Por ejemplo, si va a normalizar un campo con direcciones postales y desagregación a medida tendrá que ver cómo se estructuran las mismas, es decir, si tienen un patrón del tipo:

C/ Leonardo Da Vinci 32 San Fernando Cádiz

en donde primero aparece el tipo de vía, luego el nombre de la vía, posteriomente el número, a continuación el municipio y luego la provincia, o si por el contrario el patrón es:

Leonardo Da Vinci 32 Cádiz

Federico García Lorca 22 Granada

Juan XXIII 1 Sevilla

...

en donde lo primero que aparece es el nombre de la vía, luego el número y por último la provincia. Esto le dará al usuario una idea de cómo debería asignar los estados a las etiquetas.

De igual forma si se va a normalizar un campo con nombres de personas tendrá que ver cómo se estructuran éstos, esto es, si en el campo a normalizar aparecen nombres y apellidos, si solo aparecen nombres de pila, etc.

Para explicar cómo se lleva a cabo la asignación manual de estados se van a utilizar a modo de ejemplo los registros que aparecen en la Imagen 46. En este caso la **asignación de estados** se va a realizar **considerando** tanto **una desagregación a medida de direcciones postales** como **una desagregación de acuerdo a CDAU**.

1. Asignación de estados cuando se usa una desagregación a medida

Para el registro (4), al elemento *avd.* que la aplicación ha normalizado como 'avenida' y ha etiquetado como TV por estar dentro de la tabla de búsqueda de tipos de vía, el usuario le va asignar el estado *tipo_de_via* por ser un identificador del tipo de vía. Al elemento *de* normalizado como 'de' y etiquetado como UN por no localizarse en ninguna tabla de búsqueda se le va a asignar el estado *nombre_de_via* por ser parte del nombre de la vía. Al elemento *sevilla* normalizado como 'sevilla' y al que se le ha asignado la etiqueta MU y PR por estar dentro de las tablas de búsqueda de municipios y provincias, se le va a asignar el estado

nombre_de_via ya que se entiende que este elemento también forma parte del nombre de la vía. Por otro lado, al elemento *22* normalizado como '22' y al que se le ha asignado la etiqueta NU por ser un valor numérico se le va a asignar el estado 'ein' porque se entiende que este valor corresponde al número de la vivienda o local. De esta forma, el registro (4) quedaría como:

```
77 # 516 (4): |avd. de sevilla. 22|
78 #          |avenida de sevilla 22|
79    TV:tipo_de_via, UN:nombre_de_via, MU:nombre_de_via, NU:ein
80 #  TV:, UN:, PR:, NU:
```

Imagen 51. Registro (4) con estados asignados. Desagregación a medida

Nótese que para este registro el usuario se va a quedar con la primera secuencia de etiquetas, por lo que la segunda la podría eliminar o insertarle el carácter "#" al inicio de la misma para que no sea tenida en cuenta por la herramienta de normalización. Esta última opción es la que se muestra en la anterior imagen.

Procediendo de igual manera para el resto de registros de la muestra, esta podría quedar tal como se observa en la siguiente imagen:

```
59 ##########################################################################################
60
61 # 72 (0): |camino de monasterejo. 3|
62 #         |camino de monasterejo 3|
63    TV:tipo_de_via, UN:nombre_de_via, UN:nombre_de_via, NU:ein
64
65 # 156 (1): |valdes leal. 22|
66 #          |valdes leal 22|
67    EG:nombre_de_via, UN:nombre_de_via, NU:ein
68
69 # 418 (2): |ctra. lora del rio-la campana km 16+5|
70 #          |carretera lora_del_rio la_campana kilometro 16+5|
71    TV:tipo_de_via, MU:nombre_de_via, MU:nombre_de_via, NM:identificador_de_numeracion, UN:ein
72 #  TV:, EG:, MU:, NM:, UN:
73
74 # 472 (3): |cl suarez trasierra 6|
75 #          |calle suarez trasierra 6|
76    TV:tipo_de_via, UN:nombre_de_via, UN:nombre_de_via, NU:ein
77
78 # 516 (4): |avd. de sevilla. 22|
79 #          |avenida de sevilla 22|
80    TV:tipo_de_via, UN:nombre_de_via, MU:nombre_de_via, NU:ein
81 #  TV:, UN:, PR:, NU:
82
83 # 876 (5): |cl doctor pedro de castro 1|
84 #          |calle doctor pedro de castro 1|
85    TV:tipo_de_via, UN:nombre_de_via, UN:nombre_de_via, UN:nombre_de_via, UN:nombre_de_via, NU:ein
86
87 # 57  (6): |c/ perez galdos.4|
88 #         |calle perez galdos 4|
89    TV:tipo_de_via, UN:nombre_de_via, UN:nombre_de_via, NU:ein
90
91 # 473 (7): |cl suarez trasierra 6|
92 #          |calle suarez trasierra 6|
93    TV:tipo_de_via, UN:nombre_de_via, UN:nombre_de_via, NU:ein
94
95 # 132 (8): |finca los nietos. ctra nacional iv. km 509|
96 #          |finca los nietos carretera nacional iv kilometro 509|
97    ED:tipo_de_edificio, UN:edificio, UN:edificio, TV:tipo_de_via, UN:nombre_de_via, UN:nombre_de_via, NM:identificador_de_numeracion, NU:ein
98
99 # 142 (9): |ctr nacional iv madrid cadiz km 523|
100 #           |carretera nacional iv madrid cadiz kilometro 523|
101    TV:tipo_de_via, UN:nombre_de_via, UN:nombre_de_via, UN:nombre_de_via, MU:nombre_de_via, NM:identificador_de_numeracion, NU:ein
102 #  TV:, UN:, UN:, UN:, PR:, NM:, NU:
103
104 # 168 (10): |centro comercial airesur. plta. baja. local lb17b|
105 #           |centro_comercial airesur planta baja local lb17b|
106 #  ED:, UN:, PL:, NP:, NM:, UN:
107    ED:tipo_de_edificio, UN:edificio, PL:identificador_de_planta, NP:planta, PU:identificador_de_puerta, UN:puerta
108
109 # 193 (11): |ctra a-455. km 4+100|
110 #           |carretera a 455 kilometro 4+100|
111    TV:tipo_de_via, LE:nombre_de_via, NU:nombre_de_via, NM:identificador_de_numeracion, UN:ein
112
113 # 190 (12): |finca la cartuja. s/n.|
114 #           |finca la_cartuja sin_numero|
115    ED:tipo_de_edificio, EG:nombre_de_edificio, NM:identificador_de_numeracion
116
117 # 17  (13): |paraje cakkada del moro. numero  109|
118 #           |paraje cakkada_del_moro numero  109|
119    ZO:identificador_de_zona, EG:zona, NM:identificador_de_numeracion, NU:ein
120
121 # 373 (14): |poligono industrial servialsa. calle b. nave numero  9.|
122 #           |poligono_industrial servialsa calle b nave numero  9|
123    AG:tipo_de_agrupacion, UN:agrupacion, TV:tipo_de_via, LE:nombre_de_via, NV:identificador_de_numeracion, NM:identificador_de_numeracion, NU:ein
124
125 # 42  (15): |c/ camino de ronda 124 urb los pinares|
126 #           |calle camino de ronda 124 urbanizacion los pinares|
127    TV:tipo_de_via, TV:nombre_de_via, UN:nombre_de_via, MU:nombre_de_via, NU:ein, AG:tipo_de_agrupacion, UN:agrupacion, UN:agrupacion
128
129 # 742 (16): |c/ paseo federico garcia lorca s/n polig. indust. juncaril|
130 #           |calle paseo federico garcia lorca sin_numero poligono_industrial juncaril|
131    TV:tipo_de_via, ZO:nombre_de_via, UN:nombre_de_via, UN:nombre_de_via, UN:nombre_de_via, NM:identificador_de_numeracion, AG:tipo_de_agrupacion, UN:agrupacion
132
133 #38   (17): |c/ carril|
134 #           |calle carril|
135    TV:tipo_de_via, TV:nombre_de_via
```

Imagen 52. Registros con estados asignados. Desagregación a medida

En esta asignación de estados se puede observar que al elemento 'camino' de los registros (0) y (15) el usuario le ha asignado dos estados distintos, *tipo_de_via* y *nombre_de_via* respectivamente. En el primer caso le ha asignado el estado *tipo_de_via* debido a

que la mayoría de las direcciones postales del fichero de la imagen 46 comienzan con un identificador del tipo de vía, con lo cual 'camino' se entiende que es un tipo de vía. Por el contrario en el segundo caso el elemento 'camino' se entiende que forma parte del nombre de la vía ya que la dirección ya comienza por un identificador del tipo de vía que es 'calle'. Por este motivo le ha asignado el estado *nombre_de_via*.

Por otro lado, en el registro (15) al elemento *urb* que la aplicación ha normalizado como 'urbanizacion' y ha etiquetado como AG por estar dentro de la tabla de búsqueda de tipos de agrupación, el usuario le ha asignado el estado *tipo_de_agrupacion* por ser uno de los conjuntos no considerados dentro de los del Nomenclátor del INE (barrios, barriadas, urbanizaciones, polígonos industriales, etc.). Mientras que a los dos elementos *los* y *pinares*, etiquetados por la aplicación como UN por no haberlos detectado en ninguna tabla de búsqueda, el usuario les ha asignado el estado *agrupacion* por considerar que ambos forman parte del nombre de la urbanización. En la siguiente imagen enmarcado en color rojo se muestra lo comentado:

```
125 # 42  (15): |c/ camino de ronda 124 urb los pinares|
126 #           |calle camino de ronda 124 urbanizacion los pinares|
127   TV:tipo_de_via, TV:nombre_de_via, UN:nombre_de_via, NU:nombre_de_via, NU:ein, AG:tipo_de_agrupacion, UN:agrupacion, UN:agrupacion
```

Imagen 53. Detalle registro (15). Desagregación a medida

También cabe resaltar la asignación de estados realizada en los registros (8), (10) y (12). En estos tres casos aparecen tipos de edificios como por ejemplo, fincas o un centro comercial así como la denominación de los mismos ('los nietos', 'airesur' y 'la cartuja'). En esta situación como el usuario está realizando una desagregación a medida, les ha asignado los estados *tipo_de_edificio* y *edificio* respectivamente tal y como se puede observar en la siguiente imagen:

```
94
95 # 132 (8): |finca los nietos. ctra nacional iv. km 509|
96 #          |finca los nietos carretera nacional iv kilometro 509|
97   ED:tipo_de_edificio, UN:edificio, UN:edificio, TV:tipo_de_via, UN:nombre_de_via, UN:nombre_de_via, NM:identificador_de_numeracion, NU:ein
98                ...
99
100 # 168 (10): |centro comercial airesur. plta. baja. local 1b17b|
101 #           |centro_comercial airesur planta baja local 1b17b|
102 #  ED:, UN:, PL:, NP:, NM:, UN:
103   ED:tipo_de_edificio, UN:edificio, PL:identificador_de_planta, NP:planta, PU:identificador_de_puerta, UN:puerta
104               ...
105
106 # 190 (12): |finca la cartuja. s/n.|
107 #           |finca la_cartuja sin_numero|
108   ED:tipo_de_edificio, EG:nombre_de_edificio, NM:identificador_de_numeracion
```

Imagen 54. Detalle registros (8), (10) y (12). Desagregación a medida

Para finalizar comentar la asignación realizada al registro (13). En este caso el usuario se encuentra con un tipo de zona como es un paraje y la denominación de la misma 'cakkada del moro'. Obsérvese además, que el elemento 'cakkada' que hace referencia a 'cañada' aparece de esta forma debido al tratamiento previo realizado al fichero, es decir, se ha sustituido el carácter 'ñ' por 'kk'. En esta situación como el usuario está realizando una desagregación a medida y existe un campo de salida específico para tipos de zonas y sus denominaciones, éste les ha asignado los estados *identificador_de_zona* y *zona* respectivamente tal y como se puede

observar en la siguiente imagen:

```
117 # 17  (13): |paraje cakkada del moro. numero  109|
118 #           |paraje cakkada_del_moro numero  109|
119   ZO:identificador_de_zona, EG:zona, NM:identificador_de_numeracion, NU:ein
```

Imagen 55. Detalle registro (13). Desagregación a medida

2. Asignación de estados cuando se usa una desagregación de acuerdo a CDAU

Si el usuario hubiera elegido realizar una **desagregación de las direcciones postales de acuerdo a CDAU**, la forma de actuar sería diferente ya que en esta situación no se contempla la existencia de campos de salida específicos para recoger la información relativa a tipos de edificios y a su denominación ni a otro tipo de zona y a su denominación (parajes, cerros, lomas, etc.), con lo cual el usuario deberá indicar en la muestra que esta información va a ir a un campo de salida que se ha denominado Odub (Otros datos de ubicación). Esto lo conseguirá asignando el estado *odub* a dichos elementos, tal y como se observa en las siguientes imágenes:

```
94
95 # 132 (8): |finca los nietos. ctra nacional iv. km 509|
96 #          |finca los nietos carretera nacional iv kilometro 509|
97   ED:odub, UN:odub, UN:odub, TV:tipo_de_via, UN:nombre_de_via, UN:nombre_de_via, NM:identificador_de_numeracion, NU:ein
98          ...
99
100 # 168 (10): |centro comercial airesur. plta. baja. local 1b17b|
101 #           |centro_comercial airesur planta baja local 1b17b|
102 #  ED:, UN:, PL:, NP:, NM:, UN:
103   ED:odub, UN:odub, PL:identificador_de_planta, NP:planta, PU:identificador_de_puerta, UN:puerta
104          ...
105 # 190 (12): |finca la cartuja. s/n.|
106 #           |finca la_cartuja sin_numero|
107   ED:odub, EG:odub, NM:identificador_de_numeracion
```

Imagen 56. Detalle registros (8), (10) y (12). Desagregación CDAU

```
117 # 17  (13): |paraje cakkada del moro. numero  109|
118 #           |paraje cakkada_del_moro numero  109|
119   ZO:odub, EG:odub, NM:identificador_de_numeracion, NU:ein
```

Imagen 57. Detalle registro (13). Desagregación CDAU

Por último, se puede visualizar cómo quedaría el **fichero** con la **muestra etiquetada** si el usuario quisiera realizar una **desagregación de las direcciones postales de acuerdo a CDAU**:

```
59 ##########################################################################################
60
61 # 72 (0): |camino de monasterejo. 3|
62 #         |camino de monasterejo 3|
63    TV:tipo_de_via, UN:nombre_de_via, UN:nombre_de_via, NU:ein
64
65 # 156 (1): |valdes leal. 22|
66 #          |valdes leal 22|
67    EG:nombre_de_via, UN:nombre_de_via, NU:ein
68
69 # 418 (2): |ctra. lora del rio-la campana km 16+5|
70 #          |carretera lora_del_rio la_campana kilometro 16+5|
71    TV:tipo_de_via, MU:nombre_de_via, MU:nombre_de_via, NM:identificador_de_numeracion, UN:ein
72 #  TV:, EG:, MU:, NM:, UN:
73
74 # 472 (3): |cl suarez trasierra 6|
75 #          |calle suarez trasierra 6|
76    TV:tipo_de_via, UN:nombre_de_via, UN:nombre_de_via, NU:ein
77
78 # 516 (4): |avd. de sevilla. 22|
79 #          |avenida de sevilla 22|
80    TV:tipo_de_via, UN:nombre_de_via, MU:nombre_de_via, NU:ein
81 #  TV:, UN:, PR:, NU:
82
83 # 876 (5): |cl doctor pedro de castro 1|
84 #          |calle doctor pedro de castro 1|
85    TV:tipo_de_via, UN:nombre_de_via, UN:nombre_de_via, UN:nombre_de_via, UN:nombre_de_via, NU:ein
86
87 # 57  (6): |c/ perez galdos.4|
88 #         |calle perez galdos 4|
89    TV:tipo_de_via, UN:nombre_de_via, UN:nombre_de_via, NU:ein
90
91 # 473 (7): |cl suarez trasierra 6|
92 #          |calle suarez trasierra 6|
93    TV:tipo_de_via, UN:nombre_de_via, UN:nombre_de_via, NU:ein
94
95 # 132 (8): |finca los nietos. ctra nacional iv. km 509|
96 #          |finca los nietos carretera nacional iv kilometro 509|
97    ED:odub, UN:odub, UN:odub, TV:tipo_de_via, UN:nombre_de_via, UN:nombre_de_via, NM:identificador_de_numeracion, NU:ein
98
99 # 142 (9): |ctr nacional iv madrid cadiz km 523|
100 #           |carretera nacional iv madrid cadiz kilometro 523|
101    TV:tipo_de_via, UN:nombre_de_via, UN:nombre_de_via, UN:nombre_de_via, MU:nombre_de_via, NM:identificador_de_numeracion, NU:ein
102 #  TV:, UN:, UN:, UN:, PR:, NM:, NU:
103
104 # 168 (10): |centro comercial airesur. plta. baja. local lb17b|
105 #           |centro_comercial airesur planta baja local lb17b|
106 #  ED:, UN:, PL:, NP:, NM:, UN:
107    ED:odub, UN:odub, PL:identificador_de_planta, NP:planta, PU:identificador_de_puerta, UN:puerta
108
109 # 193 (11): |ctra a-455. km 4+100|
110 #           |carretera a 455 kilometro 4+100|
111    TV:tipo_de_via, LE:nombre_de_via, NU:nombre_de_via, NM:identificador_de_numeracion, UN:ein
112
113 # 190 (12): |finca la cartuja. s/n.|
114 #           |finca la_cartuja sin_numero|
115    ED:odub, EG:odub, NM:identificador_de_numeracion
116
117 # 17  (13): |paraje cakkada del moro. numero  109|
118 #           |paraje cakkada_del_moro numero  109|
119    ZO:odub, EG:odub, NM:identificador_de_numeracion, NU:ein
120
121 # 373 (14): |poligono industrial servialsa. calle b. nave numero  9.|
122 #           |poligono_industrial servialsa calle b nave numero  9|
123    AG:tipo_de_agrupacion, UN:agrupacion, TV:tipo_de_via, LE:nombre_de_via, NV:identificador_de_numeracion, NM:identificador_de_numeracion, NU:ein
124
125 # 42  (15): |c/ camino de ronda 124 urb los pinares|
126 #           |calle camino de ronda 124 urbanizacion los pinares|
127    TV:tipo_de_via, TV:nombre_de_via, UN:nombre_de_via, MU:nombre_de_via, NU:ein, AG:tipo_de_agrupacion, UN:agrupacion, UN:agrupacion
128
129 # 742 (16): |c/ paseo federico garcia lorca s/n polig. indust. juncaril|
130 #           |calle paseo federico garcia lorca sin_numero poligono_industrial juncaril|
131    TV:tipo_de_via, ZO:nombre_de_via, UN:nombre_de_via, UN:nombre_de_via, UN:nombre_de_via, NM:identificador_de_numeracion, AG:tipo_de_agrupacion, UN:agrupacion
132
133 #38   (17): |c/ carril|
134 #           |calle carril|
135    TV:tipo_de_via, TV:nombre_de_via
```

Imagen 58. Registros con estados asignados. Desagregación CDAU

Una vez asignados los estados se guardarán los cambios realizados teniendo en cuenta que el fichero tiene que tener **obligatoriamente extensión .csv**. A continuación, el usuario procederá a realizar el entrenamiento de la muestra, el cual se va a explicar en el siguiente apartado de este Manual.

6.3.3 HMM: Entrenamiento de la muestra

Para llevar a cabo el entrenamiento de la muestra que dará lugar a la generación del Modelo Oculto de Markov se utilizará la herramienta ***HMM: Entrenamiento de la muestra***. La interfaz y los elementos de la misma se muestran a continuación:

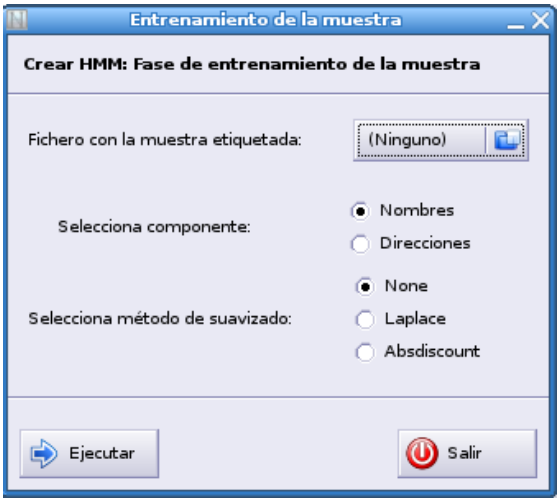


Imagen 59. Interfaz de entrenamiento de la muestra

- **Fichero con la muestra etiquetada**: en este botón el usuario indicará la ubicación en la que se encuentra el fichero con la muestra etiquetada y a la que se le asignaron los estados, ya fuese manualmente por el usuario o automáticamente utilizando un modelo HMM anterior (ver Anexo VI). Este fichero tiene que tener extensión .csv.
- **Selecciona componente**: este elemento permite al usuario indicar el tipo de campo que se va a entrenar. El tipo de campo puede ser nombres de personas o direcciones postales. No se muestra la opción de identificadores de personas físicas y/o jurídicas ya que para este caso no es necesario construir el modelo debido a que la aplicación proporciona un modelo HMM definitivo.
- **Selecciona método de suavizado**: si el usuario no selecciona ningún método de suavizado, la segmentación de los valores que componen el campo a normalizar se realizará en base a la muestra de entrenamiento seleccionada por el mismo. Nótese que en esta muestra no estarán representadas todas las estructuras o patrones del campo a normalizar, por lo tanto puede que los valores del campo a normalizar no se segmenten correctamente. Si por el contrario el usuario elige alguno de los métodos de suavizado (**Laplace** ó **Absdiscount**), en este caso la aplicación proporcionará una determinada probabilidad a las estructuras no presentes en la muestra, que se segmentarán de acuerdo a ésta. En el Anexo VIII se explica con más detalle cada uno de los métodos de suavizado.
- **Ejecutar**: este botón permite al usuario llevar a cabo el entrenamiento de la muestra.
- **Salir**: pulsando este botón el usuario puede salir de la herramienta *HMM: Entrenamiento de la muestra*.

Para el entrenamiento de la muestra es obligatorio indicar todos los elementos anteriores, así que una vez establecidos y pulsando el botón **Ejecutar**, aparecerá la siguiente ventana de salida:

Imagen 60. Entrenamiento de la muestra realizado

Como resultado del entrenamiento de la muestra se generará un fichero de extensión *.hmm* que contiene el Modelo Oculto de Markov. La ubicación del mismo coincide con la del fichero con la muestra etiquetada y estados asignados y su denominación sigue el formato:

<denominación_fichero_muestra_etiquetada>_<fecha_creación>-<hora_creación>.hmm

Por ejemplo, si el día 2 de octubre de 2013 a las 13:07 horas se hubiera entrenado el fichero *muestra_etiquetada_ejemplo.csv*, se generaría un modelo HMM denominado *muestra_etiquetada_ejemplo_20131002-1307.hmm*

No obstante, una vez generado el fichero el usuario podría renombrarlo o cambiarlo de ubicación si así lo desea. Para visualizar el contenido del fichero se recomienda utilizar el editor de texto Notepad2 que se suministra con *aLink: Herramienta de Fusión de Ficheros* si se trabaja en un entorno Windows o Gedit si se trabaja en Linux. El motivo es al igual que en los casos anteriores, evitar poblemas de codificación de caracteres.

El contenido del fichero con el Modelo Oculto de Markov es el siguiente:

- Fecha de creación del fichero con el Modelo Oculto de Markov.
- Fichero: indica la ruta en la que se ubica el fichero con el Modelo Oculto de Markov.
- HMM descripción: indica que el fichero contiene un Modelo Oculto de Markov.
- HMM contador: hace referencia al número de registros o elementos de la muestra a partir de los cuales se ha construido el modelo HMM.
- HMM estados: contiene un listado con los estados asociados a una dirección postal o nombre de persona.
- HMM etiquetas: muestra el conjunto de etiquetas asociadas a los elementos del campo a normalizar (direcciones postales o nombres de personas).
- HMM probabilidades iniciales: se trata de un vector de probabilidades iniciales que indica la probabilidad de que una dirección postal o un nombre de persona comience por cada uno de los estados. Sus componentes suman uno en total y

tiene tantas componentes como estados asociados a una dirección postal o a un nombre de persona existen.

- HMM probabilidades de transición: se trata de una matriz de probabilidades de transición entre estados. Esta indicará la probabilidad de pasar de un estado a otro según lo establecido manualmente por el usuario en la muestra etiquetada. Es una matriz cuadrada donde el número de filas y columnas coincide con el número de estados. Por ejemplo, la primera columna y primera fila corresponden al estado 'tipo_de_via', la segunda columna y segunda fila corresponden al estado nombre_de_via' y así sucesivamente.
- HMM probabilidades etiquetas: se trata de una matriz de probabilidades de observación, o de etiquetas, es decir, muestra la probabilidad de que una etiqueta tenga asociado un estado determinado. Esta matriz tiene tantas columnas como etiquetas y tantas filas como número de estados.

A continuación, se analiza a modo de ejemplo un caso particular de Modelo Oculto de Markov construido para un fichero que contiene direcciones postales con la siguiente estructura o patrón: TIPO DE VÍA, NOMBRE DE VÍA Y NÚMERO DE VÍA O IDENTIFICADOR DE NUMERACIÓN (Nº, S/N, etc.)

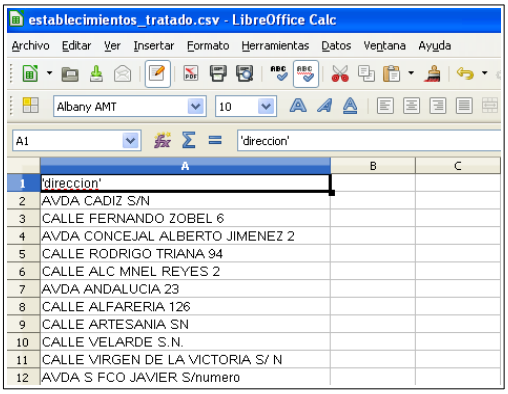


Imagen 61. Ejemplo de fichero con direcciones postales

Para este fichero una posible muestra etiquetada y con estados asignados podría ser:

```
# 12 (0): |AVDA JOSE BARRIONUEVO PEKKA|
#         |avenida jose barrionuevo pekka|
   TV:tipo_de_via, UN:nombre_de_via, UN:nombre_de_via, UN:nombre_de_via

# 30 (1): |CALLE MALAGA S/N|
#         |calle malaga sin_numero|
   TV:tipo_de_via, MU:nombre_de_via, NM:identificador_de_numeracion
#  TV:, PR:, NM:

# 7 (2): |CALLE ALFARERIA 126|
#        |calle alfareria 126|
   TV:tipo_de_via, UN:nombre_de_via, NU:ein

# 15 (3): |CALLE EL CARMEN|
#         |calle el carmen|
   TV:tipo_de_via, UN:nombre_de_via, UN:nombre_de_via

# 24 (4): |PLZA DUQUESA 21|
#         |plaza duquesa 21|
   TV:tipo_de_via, UN:nombre_de_via, NU:ein
```

Imagen 62. Posible muestra etiquetada con estados asignados

Al entrenar esta muestra el Modelo Oculto de Markov generado es el que se muestra a continuación:

```
# Creacion Thu Dec 12 11:42:45 2013
#
# Fichero: /home/ecaballero/fusion-ficheros/Ficheros/practicas_usuarios/ecaballero/muestra_eustat_con_estados_asignado_20131212-1142.hmm
#-----------------------------------------------------------------------

# HMM descripcion
#
Modelo Oculto de Markov HMM

# HMM contador
#
5

# HMM estados
#
tipo_de_via, nombre_de_via, identificador_de_numeracion, ein, cein, esn, cesn, identificador_de_bloque, bloque, tipo_de_edificio, edificio, identificador_de_portal, portal, identificador_de_escalera, escalera, identificador_de_planta, p

# HMM etiquetas
#
TV, NM, NU, BL, ED, PT, NP, ES, PL, PU, LE, ST, AG, CP, NS, MU, PR, ZO, MZ, PA, NV, UN, EG

# HMM probabilidades iniciales
#
1.000000, 0.000000, 0.000000, 0
```

uno). En la siguiente imagen se muestra más claramente esta situación:

```
14 # HMM estados
15 #
16 tipo_de_via, nombre_de_via, identificador_de_numeracion, ein, cein, esn, cesn, identificador_de_bloque, bloque, tipo_de_edificio, edificio,
17
18 # HMM etiquetas
19 #
20 TV, NM, NU, BL, ED, PT, NP, ES, PL, PU, LE, ST, AG, CP, N5, MU, PR, ZO, MZ, PA, NV, UN, EG
21
22 # HMM probabilidades iniciales
23 #
24 1.000000, 0.000000, 0.000000, 0.000000, 0.000000, 0.000000, 0.000000, 0.000000, 0.000000, 0.000000, 0.000000, 0.000000, 0.000000, 0.000000,
```

Imagen 64. Vector de probabilidades iniciales

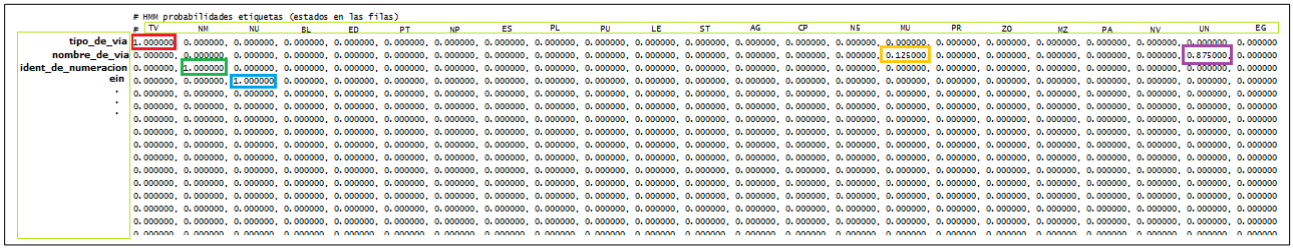
Por otro lado, observando la matriz de probabilidades de transición se tiene que:

```
                                  # HMM probabilidades de transicion (desde los estados en la filas)
                                  #
                                                                 identificador_de
                                  tipo_de_via  nombre_de_via numeracion      ein      ...
                                  ------------------------------------------------------------------------------------------------------------------------
                tipo_de_via       0.000000, 1.000000, 0.000000, 0.000000, 0.000000, 0.000000, 0.000000, 0.000000, 0.000000, 0.000000, 0.000000, 0.000000, 0.0
            nombre_de_via         0.000000, 0.500000, 0.166667, 0.333333, 0.000000, 0.000000, 0.000000, 0.000000, 0.000000, 0.000000, 0.000000, 0.000000, 0.0
identificador_de_numeracion       0.000000, 0.000000, 0.000000, 0.000000, 0.000000, 0.000000, 0.000000, 0.000000, 0.000000, 0.000000, 0.000000, 0.000000, 0.0
                      ein         0.000000, 0.000000, 0.000000, 0.000000, 0.000000, 0.000000, 0.000000, 0.000000, 0.000000, 0.000000, 0.000000, 0.000000, 0.0
                                  0.000000, 0.000000, 0.000000, 0.000000, 0.000000, 0.000000, 0.000000, 0.000000, 0.000000, 0.000000, 0.000000, 0.000000, 0.0
                                  0.000000, 0.000000, 0.000000, 0.000000, 0.000000, 0.000000, 0.000000, 0.000000, 0.000000, 0.000000, 0.000000, 0.000000, 0.0
                                  0.000000, 0.000000, 0.000000, 0.000000, 0.000000, 0.000000, 0.000000, 0.000000, 0.000000, 0.000000, 0.000000, 0.000000, 0.0
                                  0.000000, 0.000000, 0.000000, 0.000000, 0.000000, 0.000000, 0.000000, 0.000000, 0.000000, 0.000000, 0.000000, 0.000000, 0.0
                                  0.000000, 0.000000, 0.000000, 0.000000, 0.000000, 0.000000, 0.000000, 0.000000, 0.000000, 0.000000, 0.000000, 0.000000, 0.0
                                  0.000000, 0.000000, 0.000000, 0.000000, 0.000000, 0.000000, 0.000000, 0.000000, 0.000000, 0.000000, 0.000000, 0.000000, 0.0
                                  0.000000, 0.000000, 0.000000, 0.000000, 0.000000, 0.000000, 0.000000, 0.000000, 0.000000, 0.000000, 0.000000, 0.000000, 0.0
                                  0.000000, 0.000000, 0.000000, 0.000000, 0.000000, 0.000000, 0.000000, 0.000000, 0.000000, 0.000000, 0.000000, 0.000000, 0.0
                                  0.000000, 0.000000, 0.000000, 0.000000, 0.000000, 0.000000, 0.000000, 0.000000, 0.000000, 0.000000, 0.000000, 0.000000, 0.0
                                  0.000000, 0.000000, 0.000000, 0.000000, 0.000000, 0.000000, 0.000000, 0.000000, 0.000000, 0.000000, 0.000000, 0.000000, 0.0
                                  0.000000, 0.000000, 0.000000, 0.000000, 0.000000, 0.000000, 0.000000, 0.000000, 0.000000, 0.000000, 0.000000, 0.000000, 0.0
                                  0.000000, 0.000000, 0.000000, 0.000000, 0.000000, 0.000000, 0.000000, 0.000000, 0.000000, 0.000000, 0.000000, 0.000000, 0.0
                                  0.000000, 0.000000, 0.000000, 0.000000, 0.000000, 0.000000, 0.000000, 0.000000, 0.000000, 0.000000, 0.000000, 0.000000, 0.0
                                  0.000000, 0.000000, 0.000000, 0.000000, 0.000000, 0.000000, 0.000000, 0.000000, 0.000000, 0.000000, 0.000000, 0.000000, 0.0
```

Imagen 65. Matriz de probabilidades de transición

- La probabilidad de pasar del estado 'tipo_de via' al estado 'nombre_de_via' es 1.000000 (cuadro rojo). Esto es, la probabilidad de que después de 'calle' aparezca el 'nombre de la calle' o que después de 'avenida' aparezca el 'nombre de la avenida' es uno.
- La probabilidad de pasar del estado 'nombre_de_via' al estado 'nombre_de_via' es de 0.500000 (cuadro verde). Esta situación se presentará en aquellos casos en los que el nombre de la dirección postal sea compuesto, es decir, del tipo 'Manuel Siurot'.
- La probabilidad de pasar del estado 'nombre_de_via' al estado 'identificador_de_numeracion' es de 0.166667 (cuadro azul), es decir, la probabilidad de que después del nombre de la vía aparezca un elemento del tipo: n°, num, numero, km, p.k., punto kilométrico, s/n, sin número, etc. es 0.166667.
- La probabilidad de pasar del estado 'nombre_de_via' al estado 'ein' es de 0.333333 (cuadro naranja), es decir, la probabilidad de que después del nombre de la vía aparezca el número de la vivienda o local es 0.333333.

Por último, en relación a las probabilidades de las etiquetas o matriz de probabilidades de observación se puede ver en la siguiente imagen que:



HMM probabilidades etiquetas (estados en las filas)

	TV	NM	NU	BL	ED	PT	NP	ES	PL	PU	LE	ST	AG	CP	NS	MU	PR	ZO	MZ	PA	NV	UN	EG
tipo_de_via	1.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000
nombre_de_via	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.125000	0.000000	0.000000	0.000000	0.000000	0.000000	0.875000	0.000000
ident_de_numeracion	0.000000	1.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000
ein	0.000000	0.000000	1.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000
.	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000
.	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000
.	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000
	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000
	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000
	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000
	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000
	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000
	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000
	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000
	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000
	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000

Imagen 66. Matriz de probabilidades de observación

- La probabilidad de que la etiqueta 'TV' tenga asociado el estado 'tipo_de_via' es 1.000000 (cuadro rojo).
- La probabilidad de que la etiqueta 'NM' que hace referencia a un identificador de enumeración del tipo nº, s/n, sin/num, km, p.k., etc., tenga asociado el estado 'identificador_de_numeración', es decir, sea realmente un elemento de este tipo es 1.000000 (cuadro verde).
- La probabilidad de que la etiqueta 'NU', correspondiente a valores numéricos, tenga asociado el estado 'ein', que hace referencia al número de la vía o local, es 1.000000 (cuadro azul).
- La probabilidad de que la etiqueta 'MU' tenga asociado el estado 'nombre_de_via' es 0.125000 (cuadro naranja), es decir, la probabilidad de que el nombre de un municipio sea realmente el nombre de la vía es 0.125000.
- La probabilidad de que la etiqueta 'UN' tenga asociado el estado 'nombre_de_via' es 0.875000 (cuadro morado), es decir, la probabilidad de que un elemento desconocido forme parte del nombre de la vía es 0.875000.

6.3.4 Editor de listas de corrección

Las **listas de corrección** son archivos con extensión '.lst' que permiten limpiar el fichero de datos a normalizar de caracteres o cadenas que el usuario considere oportuno sustituir. Se utilizan al comienzo del proceso de normalización de un conjunto de datos y contienen caracteres o cadenas y sus correspondientes correcciones, todos ellos **obligatoriamente entrecomillados con comillas simples**. Por ejemplo, con las listas de corrección el usuario podría sustituir caracteres del tipo: '|', '$', '.', etc. por el elemento vacío '' (lo que equivaldría a eliminar dichos caracteres), o sustituirlos por espacios en blanco utilizando el carácter espacio en blanco ' ', o por cualquier otro carácter o cadena que considere. También podría sustituir cadenas de caracteres del tipo 'hnos' por 'hermanos', 'ntra' por 'nuestra', etc.

Por su función, se trata de ficheros que se deberían actualizar de forma continua ya que a medida que se van realizando procesos de normalización van a ir apareciendo nuevos elementos no recogidos en las listas o va a ser necesario eliminar algunos de los ya existentes porque podrían generar sustituciones incorrectas en el fichero. Por ejemplo, si la lista de corrección de direcciones postales contuviera el elemento 'dr' a sustituir por el elemento 'doctor', al aplicar dicha lista a direcciones postales del tipo:

'C/Leonardo Da Vinci 22 1° dr' estas se transformarían en 'C/Leonardo Da Vinci 22 1° doctor' y esto no sería lo correcto ya que en este caso 'dr' parece que está haciendo referencia a la cadena de caracteres 'derecha'. Con lo cual habría que modificar o eliminar este elemento de la lista de corrección.

En la Herramienta de Normalización existen listas de corrección para nombres de personas, direcciones postales e identificadores de personas físicas y/o jurídicas las cuales pueden ser personalizadas por el usuario. Nótese que los valores contenidos en estas listas hacen referencia a caracteres o cadenas que se pueden encontrar en ficheros con nombres de personas, direcciones postales y/o identificadores de personas físicas y jurídicas españoles, por lo que estos ficheros tendrían que ser modificados o ajustados si se usan en otros países.

Para visualizar las listas de correción, así como para insertar, editar o eliminar elementos de las mismas, la Herramienta de Normalización dispone de un editor al que se accede a través de su menú Herramientas. La interfaz de dicho editor se muestra en la siguiente imagen:

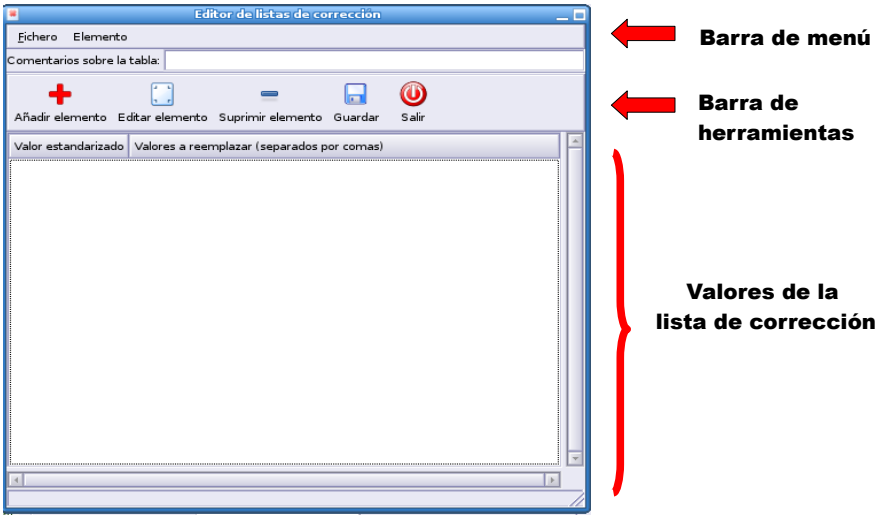


Imagen 67. Interfaz del editor de las listas de corrección

Tal y como se puede observar la interfaz está estructurada básicamente en tres partes: la parte superior de la ventana contiene una barra de menú, justo debajo de ella aparece un comentario sobre la lista y a continuación se visualiza una barra de herramientas. El resto de la ventana constituye el área en la que se muestran los valores de las listas de corrección al abrir cada una de ellas. Este área comienza con dos pestañas 'Valor estandarizado' y 'Valores a reemplazar (separados por comas)'.

A continuación, se muestra a modo de ejemplo el contenido de la lista de corrección para direcciones postales. En el Anexo IX, se muestran las correspondientes a las listas de corrección de nombres de personas y de identificadores de personas físicas y jurídicas:

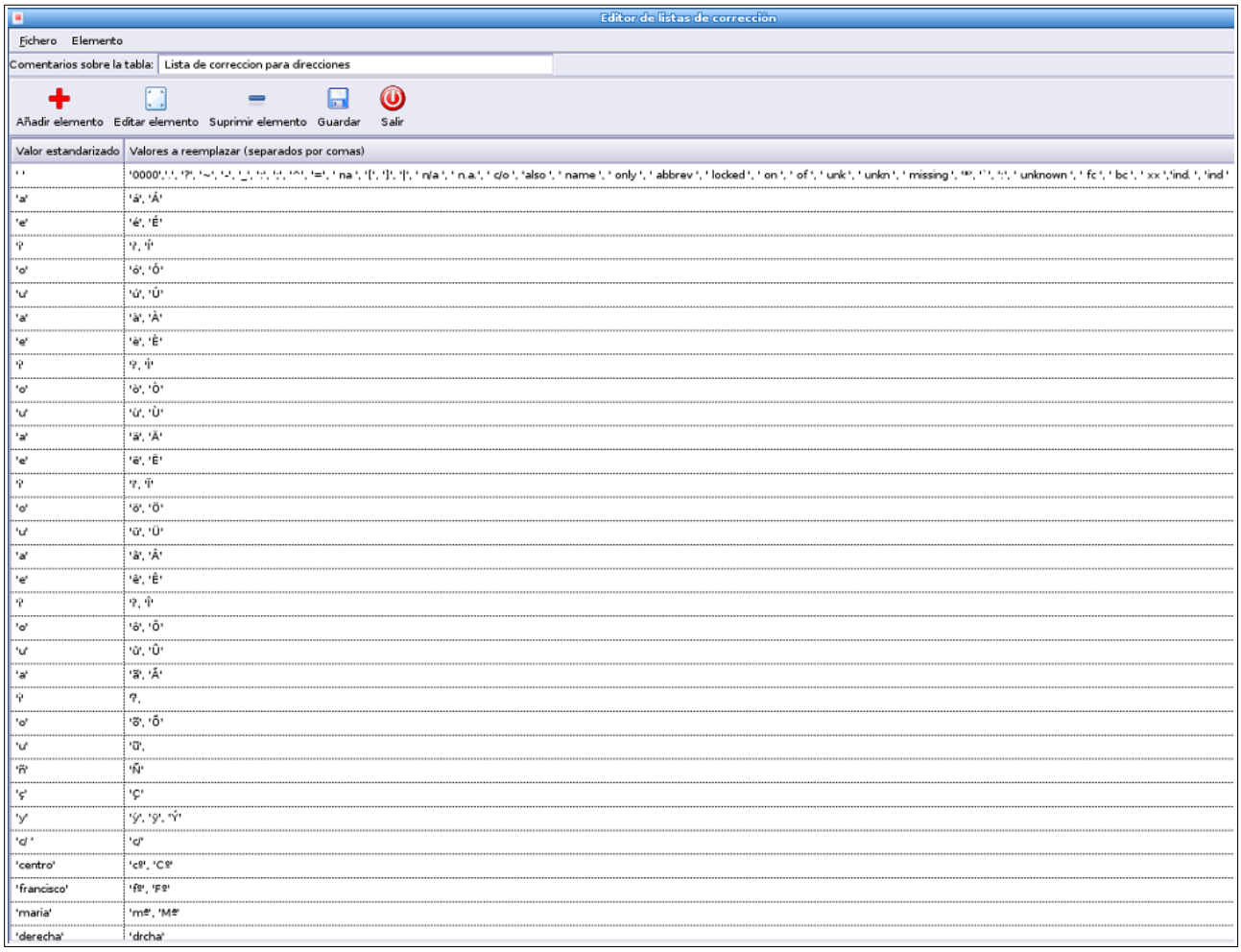


Imagen 68. Lista de corrección para direcciones postales

La barra de menú contiene los siguientes elementos:

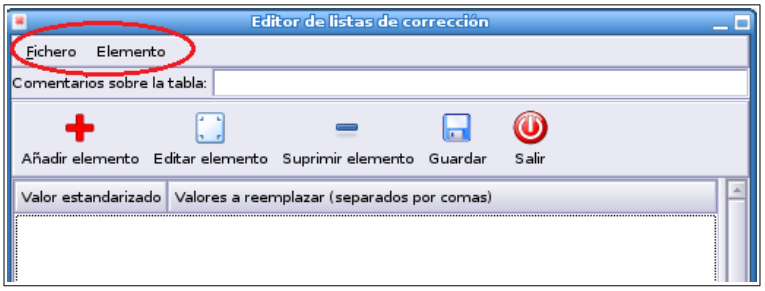


Imagen 69. Barra de menú. Editor lista de corrección

- **Fichero**: este menú permite al usuario indicar la ubicación de las listas de corrección para abrirlas, para ello tendría que pulsar **Abrir**. En concreto, éstas se encuentran en *alink\listas_tablas\listas_de_correccion*. De entre ellas se seleccionará

la que se desee editar, esto es, *nombres_correccion.lst* para nombres de personas, *direcciones_correccion.lst* para direcciones postales e *idpersona_correccion.lst* para identificadores de personas físicas y/o jurídicas. También permite guardar la lista de corrección una vez que el usuario haya realizado alguna modificación de la misma, pulsando **Guardar** o salir del editor de listas de corrección seleccionando la opción **Salir**.

- **Elemento**: menú que permite al usuario añadir un elemento a la lista de corrección (para ello tendrá que seleccionar la opción **Añadir elemento**), editar un elemento de la lista de corrección (seleccionando **Editar elemento**) o suprimir un elemento de la lista de corrección (escogiendo **Suprimir elemento**). Justo debajo se indica con más detalle cómo funcionan estas opciones, las cuales están disponibles a su vez en la barra de herramientas.

Por otro lado, la barra de herramientas está formada por los siguientes elementos:

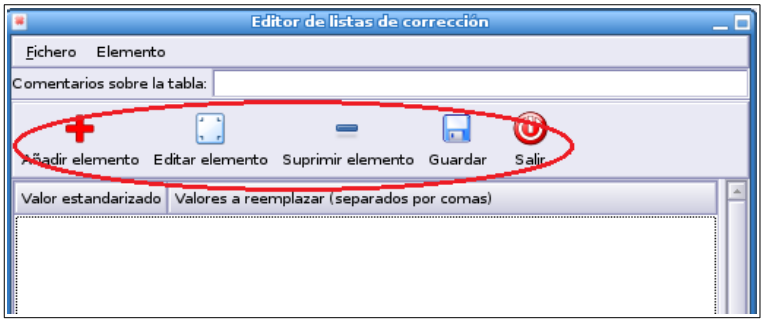


Imagen 70. Barra de herramientas. Editor lista de corrección

- **Añadir elemento**: permite al usuario añadir un elemento a la lista de corrección seleccionada. Antes de añadirlo sería recomendable que el usuario comprobara si dicho elemento ya está incluido. Para ello tiene dos opciones, la primera de ellas sería buscarlo directamente entre los valores de la lista de corrección y la segunda realizar una ordenación de la misma para buscarlo por orden alfabético. Si decide buscarlo directamente, el usuario deberá situarse sobre cualquiera de los elementos de la lista y comenzar a escribir el valor que desea incluir. **¡OJO!** siempre comenzando con una comilla simple ya que los elementos de las listas de corrección van entrecomillados con comillas simples. Si por el contrario desea ordenar la lista deberá pulsar en la pestaña *Valor estandarizado* tal y como se observa en la siguiente imagen y a continuación buscar el valor en la lista ordenada:

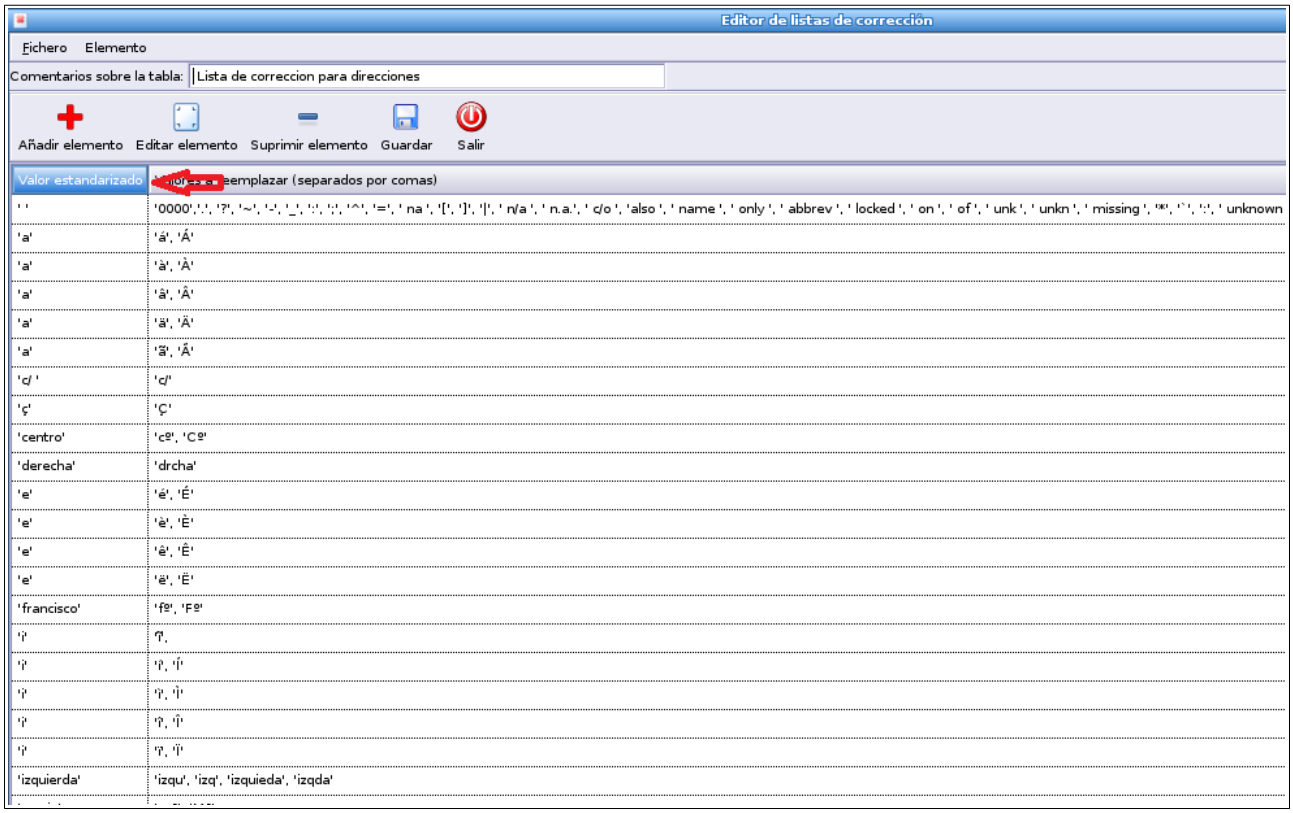


Imagen 71. Ordenación de lista de corrección de direcciones postales

La ordenación de la lista se realiza en orden ascendente pero si se pulsa en una segunda ocasión sobre dicha pestaña la lista se ordena en orden descendente.

Una vez comprobado que el elemento no se encuentra en la lista de corrección, el usuario pulsará el botón **Añadir elemento** y se abrirá la siguiente ventana:

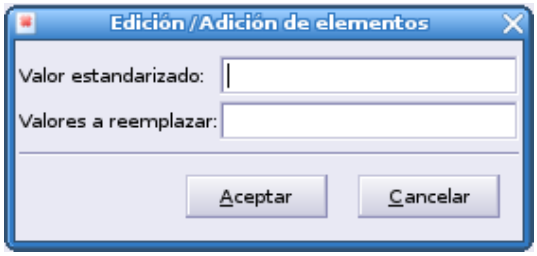


Imagen 72. Ventana de añadir elementos a la lista de corrección

En ella el usuario deberá especificar en el campo *Valor estandarizado* el valor que desea añadir a la lista de corrección. Este puede ser un solo carácter o una cadena de caracteres y **obligatoriamente** debe ir entrecomillado con **comillas simples**. Por otro lado, en el campo *Valores a reemplazar* especificará una lista de una o más cadenas de caracteres separadas por comas e igualmente entrecomilladas con comillas simples. Así, al aplicar la lista de corrección al fichero a normalizar cada valor de esta lista será reemplazado por el valor estandarizado.

Por ejemplo, si tras normalizar un fichero con direcciones postales se observa que aparecen valores en el fichero del tipo: *izqu*, *izq* o *izquieda* el usuario podría decidir que dichos valores sean reemplazados por el valor *izquierda*. Para ello debería realizar lo siguiente: abrir el editor de listas de corrección, a continuación, abrir la lista de corrección de direcciones postales y pulsar el botón **Añadir elemento**. En *Valor estandarizado* debería incluir el valor: 'izquierda' y en *Valores a reemplazar*: 'izqu', 'izq', 'izquieda'. Esto es:

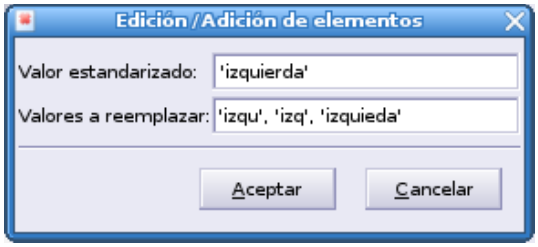


Imagen 73. Elemento añadido a la lista de corrección de direcciones postales

A continuación, se pulsaría el botón **Aceptar** y dicho elemento aparecerá en la lista de corrección. **¡OJO!** Esta inclusión no estará guardada hasta que se pulse el botón **Guardar**.

Tras guardar los cambios si el usuario vuelve a normalizar el fichero con esta nueva lista de corrección todos los valores 'izqu', 'izq', 'izquieda' serán sustituidos por 'izquierda'.

- **Editar elemento**: permite al usuario modificar un elemento de la lista de corrección seleccionada. Funciona de una manera similar a la de añadir un elemento solo que en este caso el usuario tendría que situarse sobre el elemento de la lista de corrección que desea modificar y a continuación tendría que pulsar el botón **Editar elemento**.

  Por ejemplo, si el usuario detectara este nuevo valor en el fichero con direcciones postales: *izqda*, entonces podría editar el elemento 'izquierda' añadido anteriormente. Para ello tendría que situarse en la lista de corrección de direcciones postales sobre el valor estandarizado 'izquierda' y pulsar el botón **Editar elemento**. La ventana que aparecerá en este caso es:

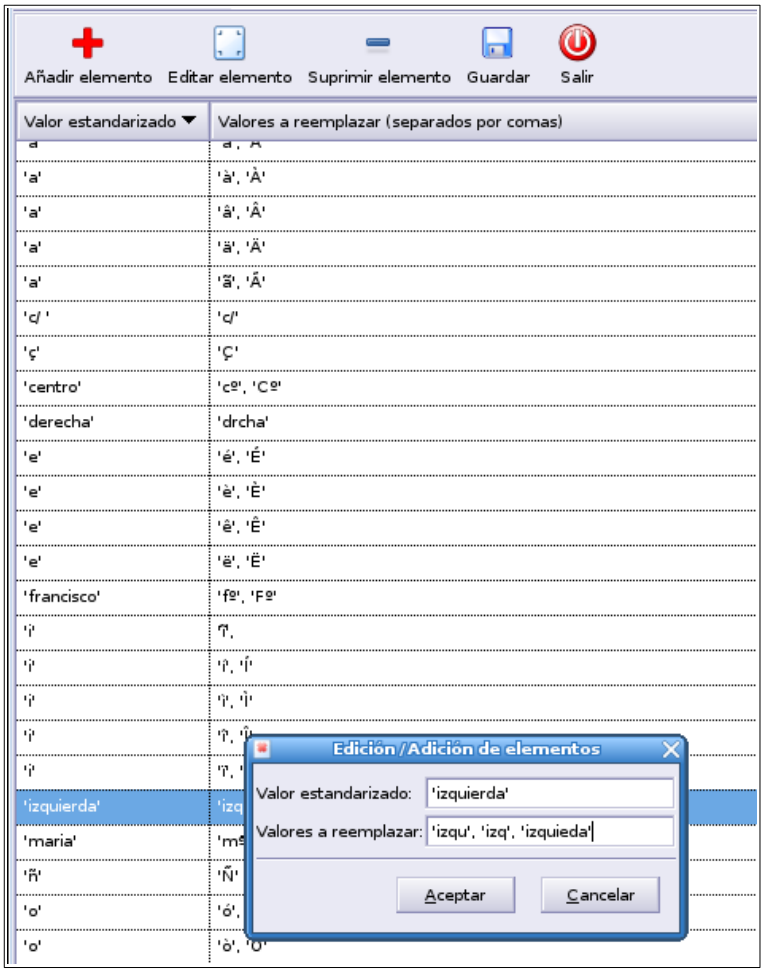


Imagen 74. Elemento modificado en la lista de corrección de direcciones postales

A continuación, en *Valores a reemplazar* el usuario debería incluir el valor 'izqda' tal y como aparece en la siguiente imagen:

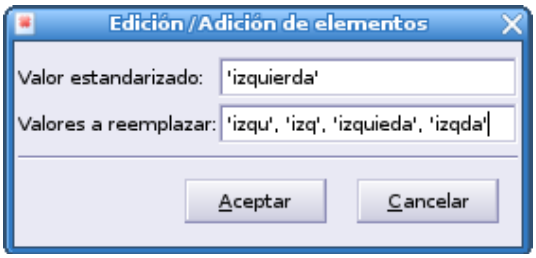


Imagen 75. Ventana de edición de elemento

Para finalizar debe pulsar **Aceptar** y **Guardar** el elemento modificado. **¡OJO!** Si no se guardan los cambios no se lleva a cabo su modificación.

- **Suprimir elemento**: permite al usuario eliminar o suprimir un elemento de la lista de corrección seleccionada. Para eliminar un elemento de la lista el usuario tiene que situarse sobre dicho elemento y pulsar el botón **Suprimir elemento**.

Por ejemplo, si en la lista de corrección de direcciones postales se quisiera eliminar el elemento 'izquierda' añadido y modificado anteriormente, el usuario se tendría que situar sobre el mismo y pulsar el botón **Suprimir elemento**. En este caso la ventana que le aparecerá es del tipo:

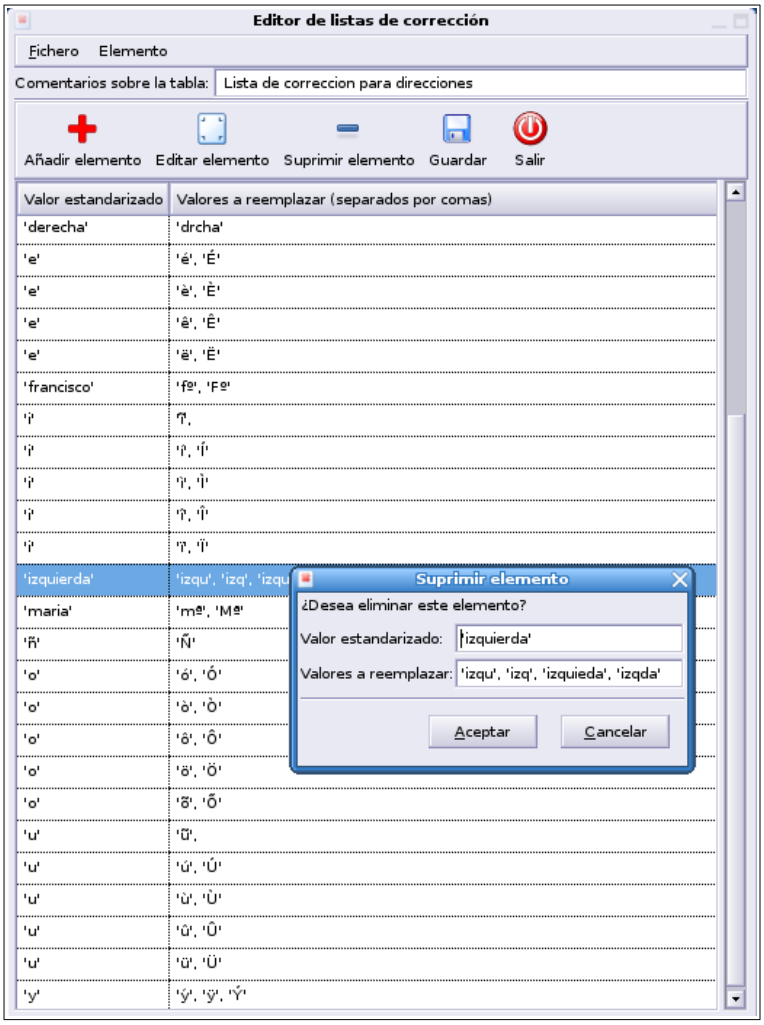


Imagen 76. Elemento a suprimir en la lista de corrección de direcciones postales

Para confirmar que desea eliminar el elemento debe pulsar **Aceptar** y posteriormente **Guardar**. **¡OJO!** Si no se guardan los cambios no se lleva a cabo su eliminación.

- **Guardar**: como ya se ha comentado en párrafos anteriores, este botón del editor de listas de corrección permite al usuario guardar los cambios realizados en las mismas, ya sean inserciones de elementos, modificaciones o eliminaciones. Si no se pulsa cada vez que se realiza una de estas operaciones los cambios no serán considerados.
- **Salir**: botón que permite al usuario salir del editor de las listas de corrección.

6.3.5 Editor de tablas de búsqueda

Las **tablas de búsqueda** son ficheros con extensión '.tbl' que contienen cadenas de caracteres que hacen referencia a un mismo

elemento común, como por ejemplo, tipos de vías, municipios, provincias, nombres masculinos, femeninos, neutros, etc., junto con las cadenas o valores estandarizados por los que se desean reemplazar. La diferencia que existe entre estos ficheros y los de las listas de corrección es que todos los elementos de una misma tabla de búsqueda tienen asignada la misma etiqueta común y los valores de las tablas van sin entrecomillar. Así, las tablas de búsqueda van a sustituir cada elemento o valor del campo a normalizar por su valor estandarizado y, además, le van a asignar una etiqueta. Por ejemplo, si se está normalizando el campo "dirección postal" y en el mismo se encuentra el valor *c/sol*, entonces el elemento *c/* se sustituirá por *calle* y se le asignará la etiqueta TV que significa Tipo de Vía, mientras que el elemento *sol* no se reemplazará por ningún valor puesto que no se va a localizar en ninguna tabla de búsqueda y por tanto se le asignará la etiqueta UN de desconocido.

Nótese, que al igual que las listas de corrección, las tablas de búsqueda se utilizan al inicio del proceso de normalización de un fichero de datos y se deberían actualizar de forma continua a medida que se realizan procesos de normalización si fuese necesario.

La Herramienta de Normalización dispone de tablas de búsqueda para nombres de personas, direcciones postales e identificadores de personas físicas y jurídicas, las cuales pueden ser personalizadas por el usuario (en el Anexo X se pueden consultar las mismas). Obsérvese, que al igual que en el caso de las listas de corrección, los valores que contienen las tablas de búsqueda hacen referencia a elementos localizados en ficheros con nombres de personas, direcciones postales e identificadores de personas físicas y/o jurídicas españoles, por lo que estos ficheros tendrían que ser modificados o ajustados si se usan en otros países. Además, para el caso concreto de direcciones postales se han incluido tres tablas de búsqueda, que son las de entidades singulares, municipios y provincias que contienen valores exclusivamente de la Comunidad Autónoma de Andalucía, con lo cual si la Herramienta de Normalización es usada por otra comunidad autónoma debería modificar los valores de las mismas.

Para visualizar las tablas de búsqueda, así como para insertar, editar o eliminar elementos de las mismas, la Herramienta de Normalización dispone de un editor al que se accede a través de su menú Herramientas. La interfaz de dicho editor se muestra en la siguiente imagen y como se puede observar presenta una estructura exactamente igual a la del editor de listas de corrección, es decir, tiene tres partes: la parte superior de la ventana contiene una barra de menú, justo debajo de ella aparece un comentario sobre la tabla y a continuación se visualiza una barra de herramientas. El resto de la ventana la constituye el área en la que se muestran los valores de las tablas de búsqueda cuando se abre cada una de ellas. Este área contiene dos pestañas *Valor estandarizado* y *Valores a reemplazar (separados por comas)*:

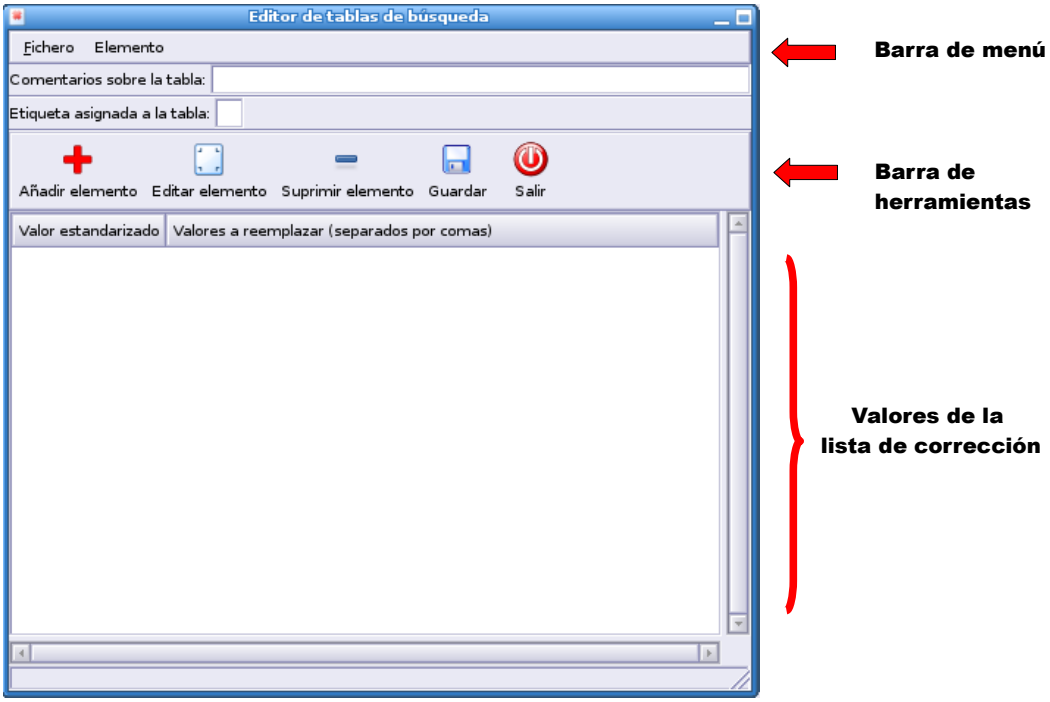
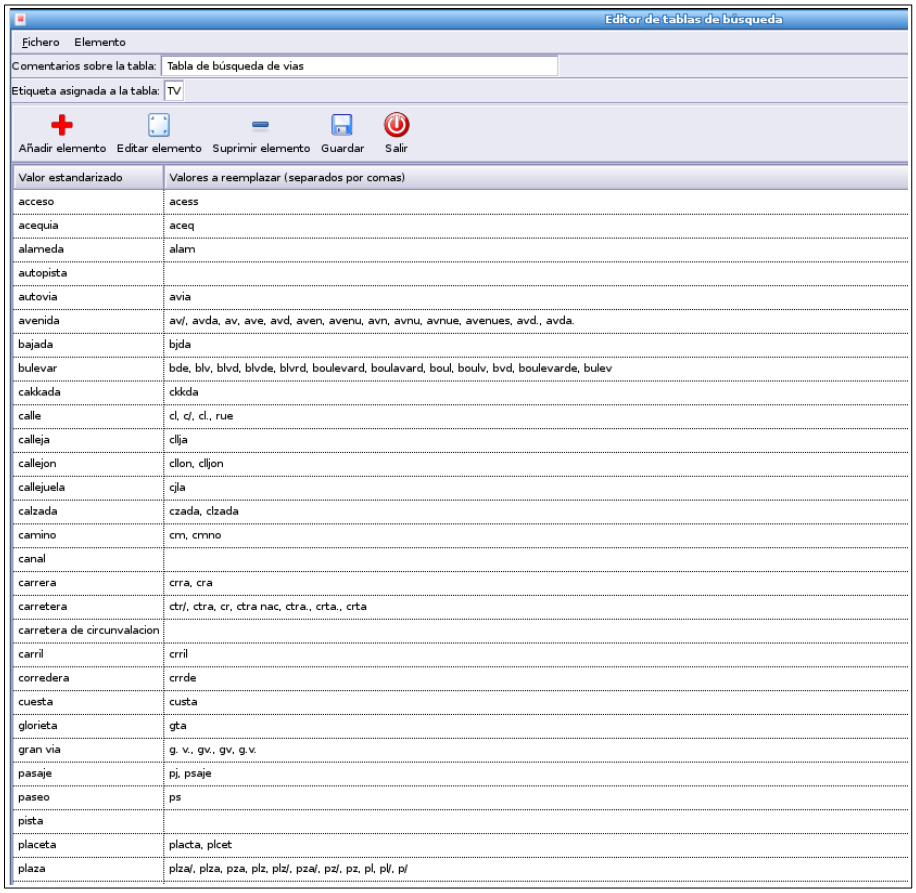


Imagen 77. Interfaz del editor de las tablas de búsqueda

A continuación, se muestra a modo de ejemplo el contenido de la tabla de búsqueda de tipos de vía para direcciones postales.



Valor estandarizado	Valores a reemplazar (separados por comas)
acceso	acess
acequia	aceq
alameda	alam
autopista	
autovia	avia
avenida	av/, avda, av, ave, avd, aven, avenu, avn, avnu, avnue, avenues, avd., avda.
bajada	bjda
bulevar	bde, blv, blvd, blvde, blvrd, boulevard, boulavard, boul, boulv, bvd, boulevarde, bulev
cakkada	ckkda
calle	cl, c/, cl., rue
calleja	clja
callejon	cllon, cljon
callejuela	cjla
calzada	czada, clzada
camino	cm, cmno
canal	
carrera	crra, cra
carretera	ctr/, ctra, cr, ctra nac, ctra., crta., crta
carretera de circunvalacion	
carril	crril
corredera	crrde
cuesta	custa
glorieta	gta
gran via	g. v., gv., gv, g.v.
pasaje	pj, psaje
paseo	ps
pista	
placeta	placta, plcet
plaza	plza/, plza, pza, plz, plz/, pza/, pz/, pz, pl, pl/, p/

Imagen 78. Tabla de búsqueda de tipos de vías para direcciones postales

A continuación, se analizan las funcionalidades del editor de las tablas de búsqueda, las cuales son equivalentes a las de las listas de corrección.

Así, la barra de menú contiene los siguientes elementos:

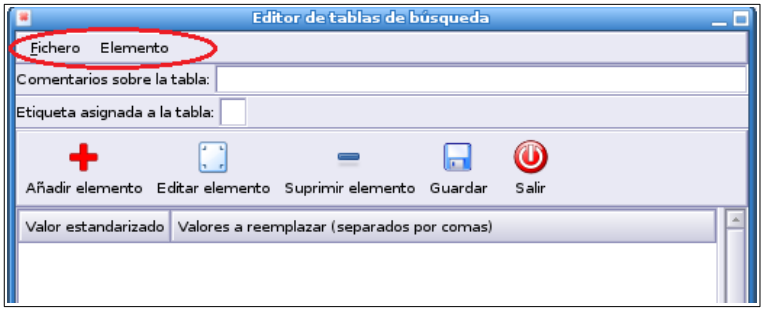


Imagen 79. Barra de menú. Editor tabla de búsqueda

- **Fichero**: este menú permite al usuario indicar la ubicación de la tabla de búsqueda que desee abrir, para ello tendría que pulsar **Abrir**. En concreto, éstas se encuentran en *alink\listas_tablas\tablas_de_busqueda*. De entre ellas se seleccionará la que se desee editar, esto es, *tbl_nombre* para nombres de personas, *tbl_direccion* para direcciones postales y *tbl_idpersona* para identificadores de personas físicas y/o jurídicas. También permite guardar la tabla de búsqueda una vez que el usuario haya realizado alguna modificación de la misma pulsando **Guardar** o salir del editor de tablas de búsqueda seleccionando la opción **Salir**.
- **Elemento**: menú que permite al usuario añadir un elemento a la tabla de búsqueda (para ello tendrá que seleccionar la opción **Añadir elemento**), editar un elemento de la tabla de búsqueda (seleccionando **Editar elemento**) o suprimir un elemento de la tabla (escogiendo **Suprimir elemento**). A continuación, se indica más detalladamente cómo funcionan estas opciones, las cuales están disponibles a su vez en la barra de herramientas.

Por otro lado, las funcionalidades de la barra de herramientas del editor de tablas de búsqueda son:

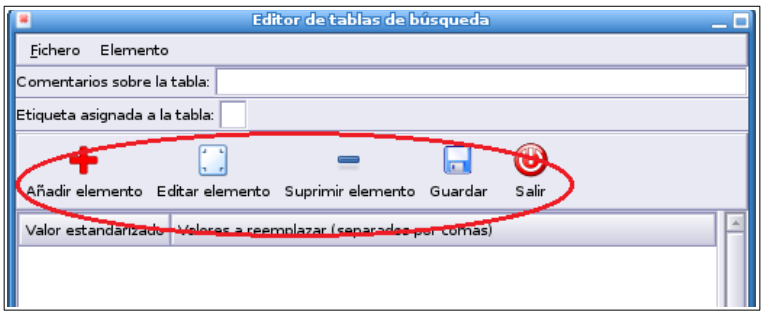
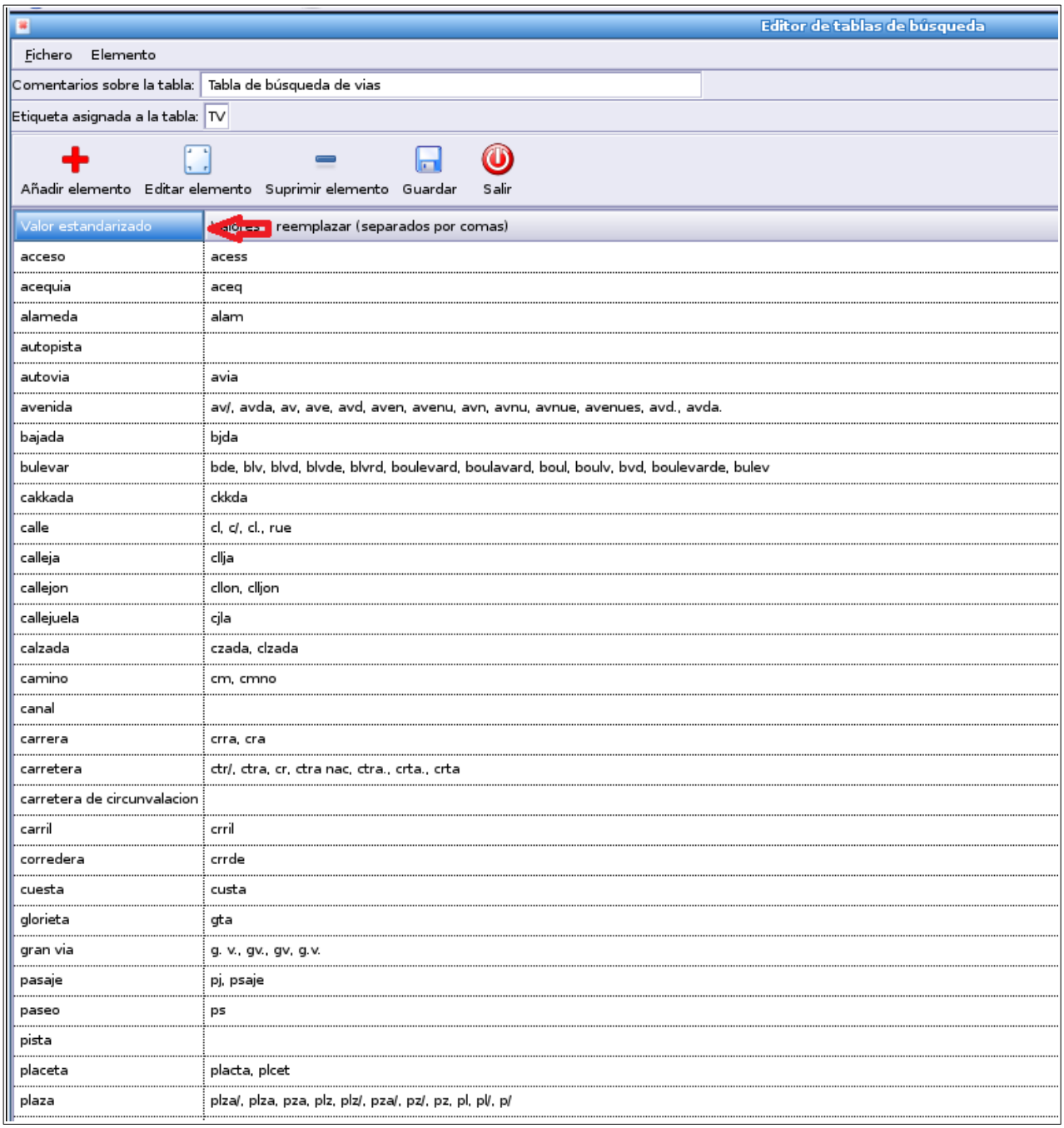


Imagen 80. Barra de herramientas. Editor tabla de búsqueda

- **Añadir elemento**: permite al usuario añadir un elemento a la tabla de búsqueda seleccionada. Antes de añadirlo sería

recomendable que el usuario comprobara si dicho elemento ya está incluido. Para ello tiene dos opciones, la primera de ellas sería buscarlo directamente entre los valores de la tabla de búsqueda y la segunda realizar una ordenación de la misma y buscarlo por orden alfabético. Si opta por buscarlo directamente, el usuario deberá situarse sobre cualquiera de los elementos de la tabla y comenzar a escribir el valor que desea incluir. Si por el contrario desea ordenar los valores de la tabla deberá pulsar en la pestaña *Valor estandarizado* tal y como se observa en la siguiente imagen y a continuación buscar el elemento que desea añadir:



Valor estandarizado	reemplazar (separados por comas)
acceso	acess
acequia	aceq
alameda	alam
autopista	
autovia	avia
avenida	av/, avda, av, ave, avd, aven, avenu, avn, avnu, avnue, avenues, avd., avda.
bajada	bjda
bulevar	bde, blv, blvd, blvde, blvrd, boulevard, boulavard, boul, boulv, bvd, boulevarde, bulev
cakkada	ckkda
calle	cl, c/, cl., rue
calleja	cllja
callejon	cllon, clljon
callejuela	cjla
calzada	czada, clzada
camino	cm, cmno
canal	
carrera	crra, cra
carretera	ctr/, ctra, cr, ctra nac, ctra., crta., crta
carretera de circunvalacion	
carril	crril
corredera	crrde
cuesta	custa
glorieta	gta
gran via	g. v., gv., gv, g.v.
pasaje	pj, psaje
paseo	ps
pista	
placeta	placta, plcet
plaza	plza/, plza, pza, plz, plz/, pza/, pz/, pz, pl, pl/, p/

Imagen 81. Ordenación de tabla de búsqueda de tipos de vía

Como se puede observar, la ordenación de la tabla se realiza en orden ascendente pero si se pulsa en una segunda ocasión sobre dicha pestaña la lista 2e ordena en orden descendente. Así, una vez comprobado que el elemento no se encuentra en la tabla de búsqueda, el usuario pulsará el botón **Añadir elemento** y se abrirá la siguiente ventana:

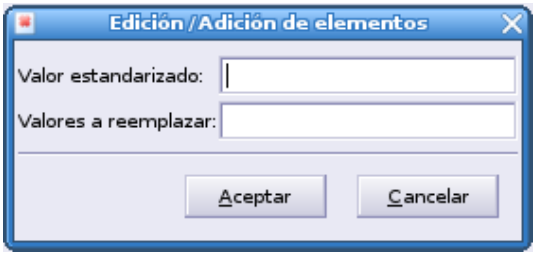


Imagen 82. Ventana de añadir elementos a la tabla de búsqueda

En ella el usuario deberá especificar en el campo *Valor estandarizado* el valor que desea añadir a la tabla de búsqueda. Este puede ser una **cadena o varias cadenas de caracteres pero NO deben ir entrecomilladas**.

Por otro lado, en el campo *Valores a reemplazar* especificará una lista con ninguna, una o más cadenas de caracteres separadas por comas pero como se ha comentando anteriormente estos valores se introducen sin entrecomillar. Así, cada uno de los valores de esta lista va a ser reemplazado por el valor estandarizado.

Por ejemplo, si se desea añadir el elemento *plazoleta* a la tabla de búsqueda de tipos de vía, el usuario tendría que proceder de la siguiente forma: abrir el editor de las tablas de búsqueda, abrir la tabla de búsqueda de tipos de vía para direcciones postales y pulsar el botón **Añadir elemento**. En *Valor estandarizado* debe incluir el valor: plazoleta y en *Valores a reemplazar* no incluirá ninguna cadena. Esto es:

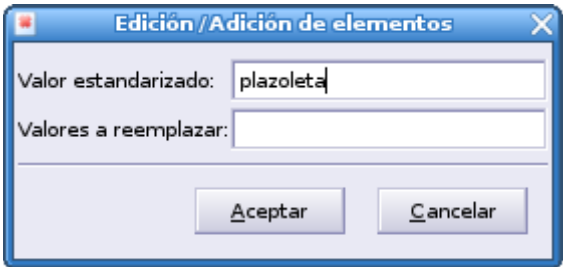


Imagen 83. Elemento añadido a la tabla de búsqueda de tipos de vía de direcciones postales

A continuación, se pulsaría el botón **Aceptar** y dicho elemento aparecerá entre los valores de la tabla. **¡OJO!** Esta inclusión no estará guardada hasta que se pulse el botón **Guardar**.

Además, una vez añadido el elemento y guardado se puede ordenar la tabla de búsqueda pulsando en la pestaña 'Valor estandarizado' tal y como se indicó anteriormente.

- **Editar elemento**: permite al usuario modificar un elemento de la tabla de búsqueda seleccionada. Funciona de una

manera similar a la de añadir un elemento solo que en este caso el usuario tendría que situarse sobre el elemento de la tabla de búsqueda que desea modificar y a continuación tendría que pulsar el botón **Editar elemento**.

Por ejemplo, si el usuario detectara los elementos *pzta* y *plzoleta* en un fichero con direcciones postales, entonces podría editar el elemento *plazoleta* añadido anteriormente para incluir estas dos opciones. Para ello tendría que abrir la tabla de búsqueda de tipos de vías y situarse sobre el elemento *plazoleta* y pulsar el botón **Editar elemento**. La ventana que aparecerá en este caso es:

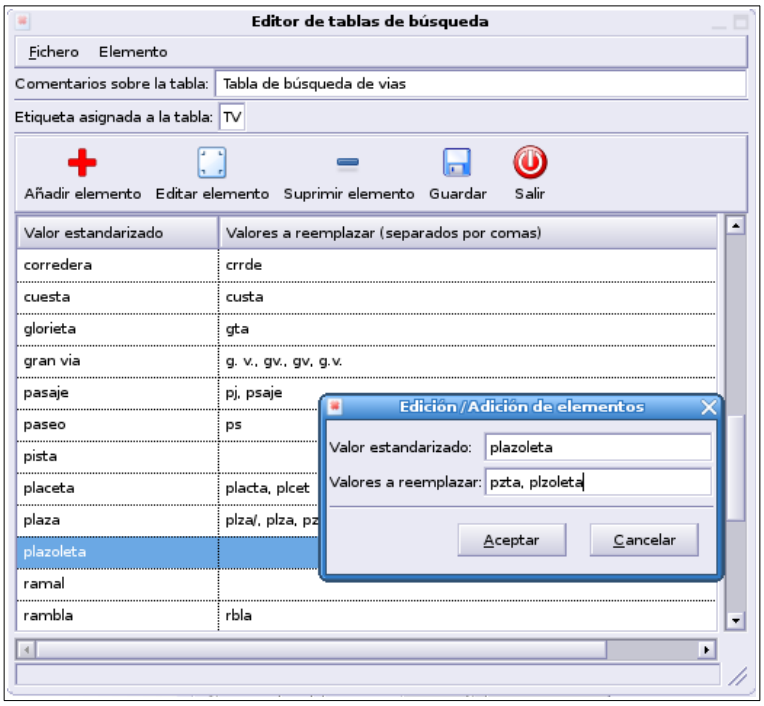


Imagen 84. Elemento modificado en la tabla de búsqueda de tipos de vías

A continuación, en *Valores a reemplazar* el usuario debería incluir los valores *pzta* y *plzoleta* tal y como aparece en la imagen de arriba.

Para finalizar debe pulsar **Aceptar** y **Guardar** el elemento modificado. **¡OJO!** Si no se guardan los cambios no se lleva a cabo su modificación.

- **Suprimir elemento**: permite al usuario eliminar o suprimir un elemento de la tabla de búsqueda seleccionada. Para eliminar un elemento de la tabla el usuario tiene que situarse sobre dicho elemento y pulsar el botón **Suprimir elemento**.

  Por ejemplo, si en la tabla de búsqueda de tipos de vías para direcciones postales se quisiera eliminar el elemento *plazoleta* añadido y modificado anteriormente, el usuario se tendría que situar sobre el mismo y pulsar el botón **Suprimir elemento**. En este caso la ventana que le aparecerá es del tipo:

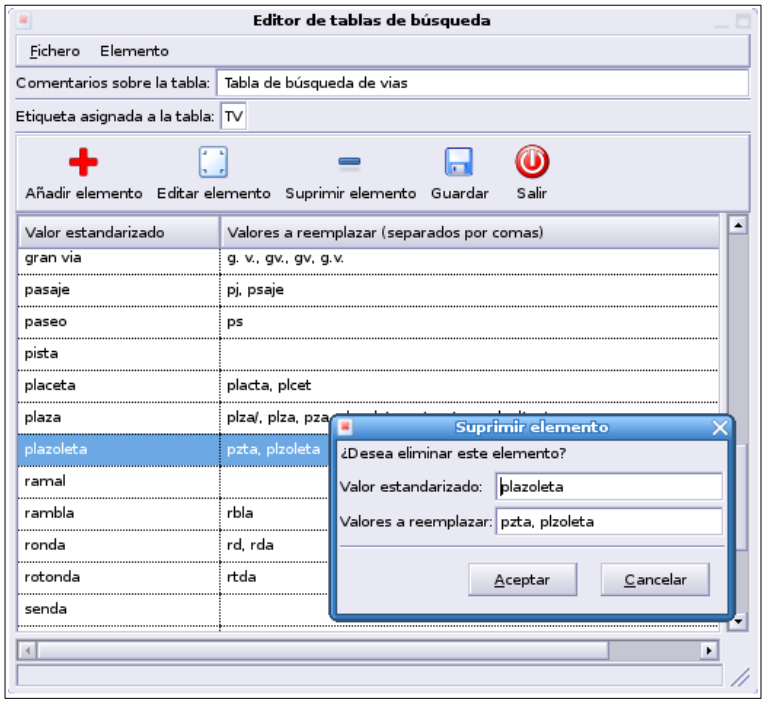


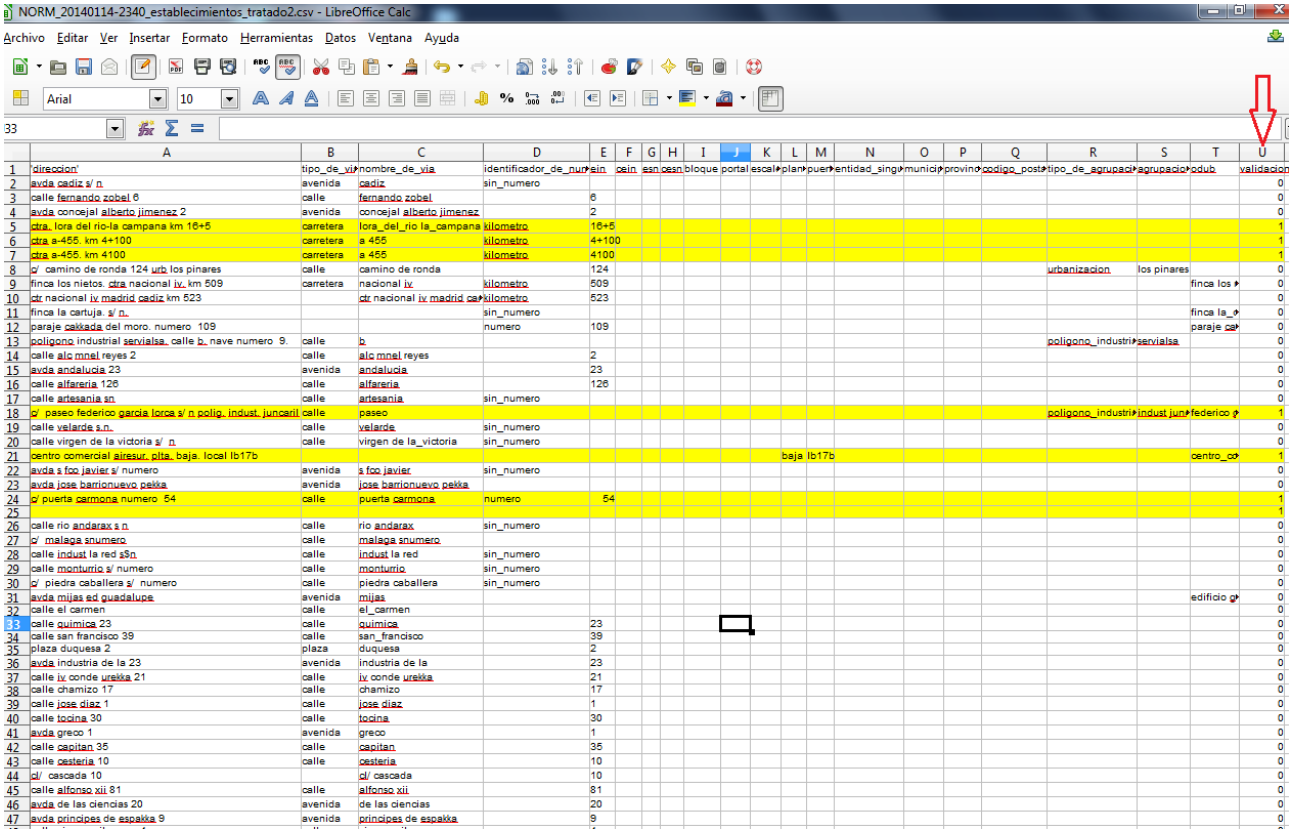
Imagen 85. Elemento a suprimir de la tabla de búsqueda de tipos de vía

Para confirmar que desea eliminar el elemento debe pulsar **Aceptar** y posteriormente **Guardar**. **¡OJO!** Si no se guardan los cambios no se lleva a cabo su eliminación.

- **Guardar**: al igual que para las listas de corrección, este botón permite al usuario guardar los cambios realizados en las tablas de búsqueda, ya sean inserciones de elementos, modificaciones o eliminaciones. Si no se pulsa cada vez que se realiza una de estas operaciones los cambios no serán considerados.
- **Salir**: botón que permite al usuario salir del editor de tablas de búsqueda.

6.4 Validación del proceso de normalización

Una vez realizado el proceso de normalización de un fichero de datos, el usuario debe comprobar la bondad del mismo. Para ello deberá abrir el fichero de datos normalizado y consultar la variable o campo 'validacion' que aparece en la última columna de dicho fichero. Se recomienda abrir el fichero normalizado con el programa LibreOffice Calc para evitar problemas de codificación. A continuación, se muestra a modo de ejemplo el resultado de normalizar el campo 'direccion' del fichero de la imagen 16, en el que se ha utilizado la desagregación CDAU y el modelo HMM asociado a este tipo de desagregación y disponible en la Herramienta de Normalización:



	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S	T	U
1	'direccion'	tipo_de_vi	nombre_de_via	identificador_de_nur	ein	cein	esn	cesn	bloque	portal	escal	plan	puer	entidad_sing	municip	provinc	codigo_post	tipo_de_agrupaci	agrupacio	odub	validacion
2	avda cadiz s/ n	avenida	cadiz	sin_numero																	0
3	calle fernando zobel 6	calle	fernando zobel		6																0
4	avda concejal alberto jimenez 2	avenida	concejal alberto jimenez		2																0
5	ctra. lora del rio-la campana km 16+5	carretera	lora_del_rio la_campana	kilometro	16+5																1
6	ctra a-455. km 4+100	carretera	a 455	kilometro	4+100																1
7	ctra a-455. km 4100	carretera	a 455	kilometro	4100																1
8	c/ camino de ronda 124 urb los pinares	calle	camino de ronda		124													urbanizacion	los pinares		0
9	finca los nietos. ctra nacional iv. km 509	carretera	nacional iv	kilometro	509															finca los	0
10	ctr nacional iv madrid cadiz km 523		ctr nacional iv madrid ca	kilometro	523																0
11	finca la cartuja. s/ n.			sin_numero																finca la_	0
12	paraje cakkada del moro. numero 109			numero	109															paraje ca	0
13	poligono industrial servialsa. calle b. nave numero 9.	calle	b															poligono_industri	servialsa		0
14	calle alc mnel reyes 2	calle	alc mnel reyes		2																0
15	avda andalucia 23	avenida	andalucia		23																0
16	calle alfareria 126	calle	alfareria		126																0
17	calle artesania sn	calle	artesania	sin_numero																	0
18	c/ paseo federico garcia lorca s/ n polig. indust. juncaril	calle	paseo															poligono_industri	indust jun	federico g	1
19	calle velarde s.n.	calle	velarde	sin_numero																	0
20	calle virgen de la victoria s/ n	calle	virgen de la_victoria	sin_numero																	0
21	centro comercial airesur. plta. baja. local lb17b											baja	lb17b							centro_co	1
22	avda s fco javier s/ numero	avenida	s fco javier	sin_numero																	0
23	avda jose barrionuevo pekka	avenida	jose barrionuevo pekka																		0
24	c/ puerta carmona numero 54	calle	puerta carmona	numero	54																1
25																					1
26	calle rio andarax s n	calle	rio andarax	sin_numero																	0
27	c/ malaga snumero	calle	malaga snumero																		0
28	calle indust la red s$n	calle	indust la red	sin_numero																	0
29	calle monturrio s/ numero	calle	monturrio	sin_numero																	0
30	c/ piedra caballera s/ numero	calle	piedra caballera	sin_numero																	0
31	avda mijas ed guadalupe	avenida	mijas																	edificio g	0
32	calle el carmen	calle	el_carmen																		0
33	calle quimica 23	calle	quimica		23																0
34	calle san francisco 39	calle	san_francisco		39																0
35	plaza duquesa 2	plaza	duquesa		2																0
36	avda industria de la 23	avenida	industria de la		23																0
37	calle iv conde urekka 21	calle	iv conde urekka		21																0
38	calle chamizo 17	calle	chamizo		17																0
39	calle jose diaz 1	calle	jose diaz		1																0
40	calle tocina 30	calle	tocina		30																0
41	avda greco 1	avenida	greco		1																0
42	calle capitan 35	calle	capitan		35																0
43	calle cesteria 10	calle	cesteria		10																0
44	cl/ cascada 10		cl/ cascada		10																0
45	calle alfonso xii 81	calle	alfonso xii		81																0
46	avda de las ciencias 20	avenida	de las ciencias		20																0
47	avda principes de espakka 9	avenida	principes de espakka		9																0

Imagen 86. Ejemplo de salida de un fichero con direcciones postales normalizado

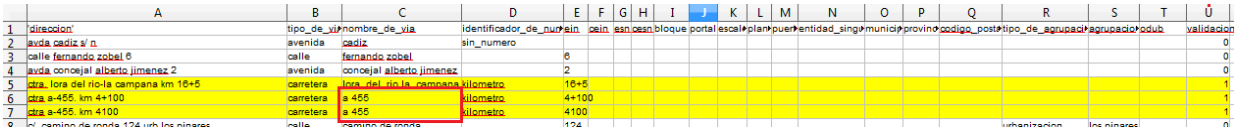
Como se puede observar la columna 'validacion' toma los valores 0 y 1 que representan lo siguiente:

- El valor 1 advierte al usuario de que el valor del campo a normalizar podría estar incorrectamente normalizado. Para conocer cuantos registros están incorrectamente normalizados el usuario puede situarse en la columna 'validacion' y sumar los unos que tiene dicha columna. En este ejemplo se han resaltado en color amarillo para una rápida visualización.
- El valor 0 indica que el algoritmo de validación no ha encontrado nada incoherente que haga pensar que el registro está mal normalizado.

A continuación, se indican los motivos por los que el algoritmo de validación muestra el valor 1:

1) Existen valores en los campos de salida que son incoherentes con el tipo de campo de salida. No obstante, esto no quiere decir que la normalización haya sido incorrecta. Por ejemplo, para el fichero de direcciones anterior, el valor 1 aparece cuando:

En el nombre de la vía aparecen elementos de tipo numérico:



	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S	T	U
1	'direccion'	tipo_de_vi	nombre_de_via	identificador_de_nur	ein	cein	esn	cesn	bloque	portal	escal	plan	puer	entidad_sing	municip	provinc	codigo_post	tipo_de_agrupaci	agrupacio	odub	validacion
2	avda cadiz s/ n	avenida	cadiz	sin_numero																	0
3	calle fernando zobel 6	calle	fernando zobel		6																0
4	avda concejal alberto jimenez 2	avenida	concejal alberto jimenez		2																0
5	ctra. lora del rio-la campana km 16+5	carretera	lora_del_rio la_campana	kilometro	16+5																1
6	ctra a-455. km 4+100	carretera	a 455	kilometro	4+100																1
7	ctra a-455. km 4100	carretera	a 455	kilometro	4100																1
8	c/ camino de ronda 124 urb los pinares	calle	camino de ronda		124													urbanizacion	los pinares		0

Imagen 87. Ejemplo de registros incorrectamente normalizados por existir un valor numérico en el nombre de la vía

O cuando en el nombre de la vía aparecen elementos que identifican la puerta de una vivienda, local, etc.:

	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S	T	U
1	'direccion'	tipo_de_vi	nombre_de_via	identificador_de_nur	ein	cein	esn	cesn	bloque	portal	escal	plan	puer	entidad_sing	municip	provinc	codigo_post	tipo_de_agrupaci	agrupacio	odub	validacion
23	avda jose barrionuevo pekka	avenida	jose barrionuevo pekka																		0
24	c/ puerta carmona numero 54	calle	puerta carmona	numero	54																1
25																					1
26	calle rio andarax s n	calle	rio andarax	sin_numero																	0
27	c/ malaga snumero	calle	malaga snumero																		0

Imagen 88. Ejemplo de registro incorrectamente normalizado por existir un identificador de puerta en el nombre de la vía

Obsérvese que en este último ejemplo, al usuario se le advierte de que uno de los elementos usados para identificar la puerta de una vivienda o local (puerta) forma parte del nombre de la vía y es por ello por lo que en el campo 'validacion' aparece el valor 1. No obstante, la normalización de esta dirección postal es correcta.

2) Existen estructuras o patrones de nombres de personas o direcciones postales que no se han incluido en la muestra de entrenamiento que generará el Modelo Oculto de Markov o bien existen elementos no incluidos en las tablas de búsqueda y por tanto no van a ser reconocidos por dicho Modelo.

En el siguiente ejemplo de direcciones postales se puede comprobar cómo la estructura o patrón referida a esta dirección no se encuentra dentro del fichero con la muestra etiquetada, además el elemento 'polig. indus.' no se está normalizando de la manera adecuada, con lo cual habría que echar un vistazo a la tabla de búsqueda de agrupaciones que es donde se encuentra el mismo y si no está añadirlo o editarlo. La manera de proceder en estos casos se explicará con más detalle más abajo.

	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S	T	U
1	'direccion'	tipo de vi	nombre de via	identificador de nur	ein	cein	esn	cesn	bloque	portal	escal	plan	puer	entidad sing	municip	provinc	codigo post	tipo de agrupaci	agrupacio	odub	validacion
17	calle artesania sn	calle	artesania	sin_numero																	0
18	c/ paseo federico garcia lorca s/ n polig. indust. juncaril	calle	paseo															poligono_industri	indust jun	federico g	1
19	calle velarde s n.	calle	velarde	sin_numero																	0
20	calle virgen de la victoria s/ n	calle	virgen de la victoria	sin numero																	0

Imagen 89. Ejemplo de registro incorrectamente normalizado por no existir este patrón de dirección en la muestra de entrenamiento

3) El campo a normalizar tiene valores perdidos, es decir, a todos aquellos registros que tengan un valor perdido en el campo a normalizar se les asignará el valor 1. En el ejemplo que se muestra a continuación se puede comprobar.

	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S	T	U
1	'direccion'	tipo_de_vi	nombre_de_via	identificador_de_nur	ein	cein	esn	cesn	bloque	portal	escal	plan	puer	entidad_sing	municip	provinc	codigo_post	tipo_de_agrupaci	agrupacio	odub	validacion
25																					1
26	calle rio andarax s n	calle	rio andarax	sin_numero																	0

Imagen 90. Ejemplo de registro incorrectamente normalizado por no existir valor para el mismo

En consecuencia, la importancia del proceso de validación reside fundamentalmente en que va a permitir detectar estructuras o patrones de datos que no han sido correctamente normalizados.

A continuación, se explica con más detalle cómo se procedería en el caso 2), es decir, cuando existen estructuras o patrones de datos no presentes en la muestra de entrenamiento que generará el Modelo Oculto de Markov. En estas situaciones se puede trabajar de dos formas dependiendo de si la estructura o patrón faltante difiere bastante con respecto a los patrones habituales del

campo a normalizar. Para ello se hará uso de la dirección postal de la imagen 89, que parece diferir un poco de la estructura habitual del campo a normalizar, que es: tipo de vía, nombre de vía, identificador de enumeración y número:

c/ paseo federico garcia lorca s/ n polig. indust. juncaril

En este caso existe además, una incorrecta estandarización del elemento *polig. indust.*, hecho que habría que corregir. En concreto, lo que está sucediendo es que el elemento *polig.* lo ha encontrado dentro de la tabla de búsqueda de tipos de agrupación y lo ha sustituido por *poligono_industrial*, con lo cual el valor *polig. indust. juncaril* ha pasado a ser *poligono_industrial indust juncaril* (ver de nuevo imagen 89).

Así pues, en esta situación el usuario actuará de la siguiente manera:

1. Accederá al editor de tablas de búsqueda y abrirá la tabla de búsqueda relativa a tipos de agrupación (*kagrupacion.tbl*), que es donde se encuentra este elemento. Localizará el valor *poligono industrial* y lo editará, tal y como se ha explicado en el apartado 6.3.5. En este caso añadirá el elemento *polig. indust.* a *Valores a reemplazar* y se guardarán los cambios. Asi, cuando se vuelva a normalizar el fichero el valor *polig. indust. juncaril* será sustituido por *poligono_industrial juncaril* .

2. A continuación, aunque la estructura de la dirección postal de ejemplo difiere un poco de las habituales del fichero a normalizar, el usuario abrirá la muestra de entrenamiento a partir de la cual se generó el Modelo Oculto de Markov con el que se normalizó el fichero. Para ello usará un editor de texto como Notepad2 o Gedit y por el conocimiento que tiene sobre los registros del fichero a normalizar y sobre las etiquetas y sus posibles estados introducirá la siguiente estructura:

   TV:tipo_de_via, ZO:nombre_de_via, UN:nombre_de_via, UN:nombre_de_via, NM:identificador_de_numeracion, AG:tipo_de_agrupacion, UN:agrupacion

   Guardará los cambios realizados en la muestra y la entrenará con la herramienta **HMM: Entrenamiento de la muestra**. Una vez obtenido el nuevo modelo HMM lo utilizará para volver a normalizar el fichero de datos completo.

   Tal y como se puede comprobar en la siguiente imagen, la dirección aparece correctamente normalizada y no ha habido direcciones anteriores que con esta nueva inclusión aparezcan ahora incorrectamente normalizadas. Sin embargo, siguen apareciendo registros con 'validacion' igual a 1 porque las estructuras o patrones correspondientes a estas direcciones no se han incluido en la muestra:

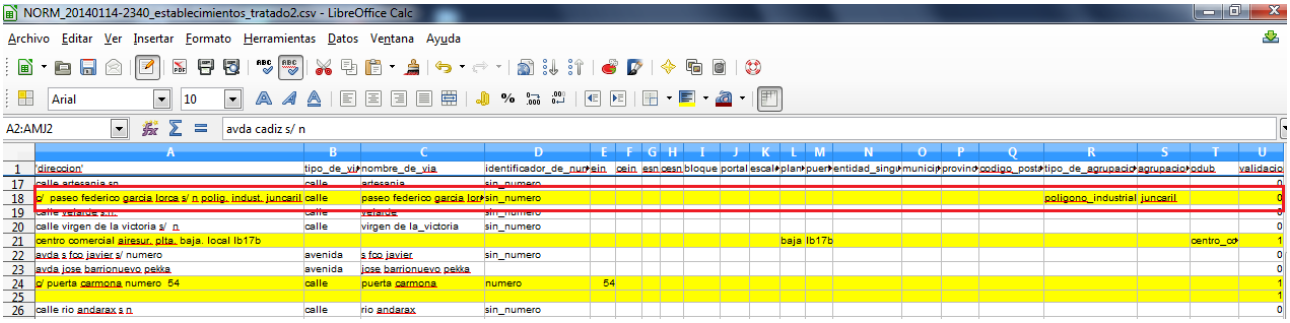


Imagen 91. Ejemplo de registro correctamente normalizado tras introducir patrón

No obstante, hay que advertir al usuario que al enriquecer la muestra con nuevas estructuras muy diferentes de las ya incluidas podría ocurrir que registros que estaban correctamente normalizados ahora no lo estén. Esto se debe a que al incluir nuevos patrones o estructuras, las probabilidades de las que ya estaban en la muestra se ven modificadas. Con lo cual habría que tener cuidado con esta operación. Además, en el caso de que existieran muchos patrones o estructuras a incluir en la muestra, esta tarea podría ser bastante tediosa para el usuario.

Por tanto, para evitar estos problemas se recomienda que se trabaje de la siguiente forma:

Seleccionar del fichero de datos normalizado aquellos registros en los que el valor de la variable 'validacion' es 1. Si se trabaja con LibreOffice Calc, se puede realizar un filtro sobre esta variable y seleccionar los casos en los que el valor es 1 tal y como se observa en la siguiente imagen:

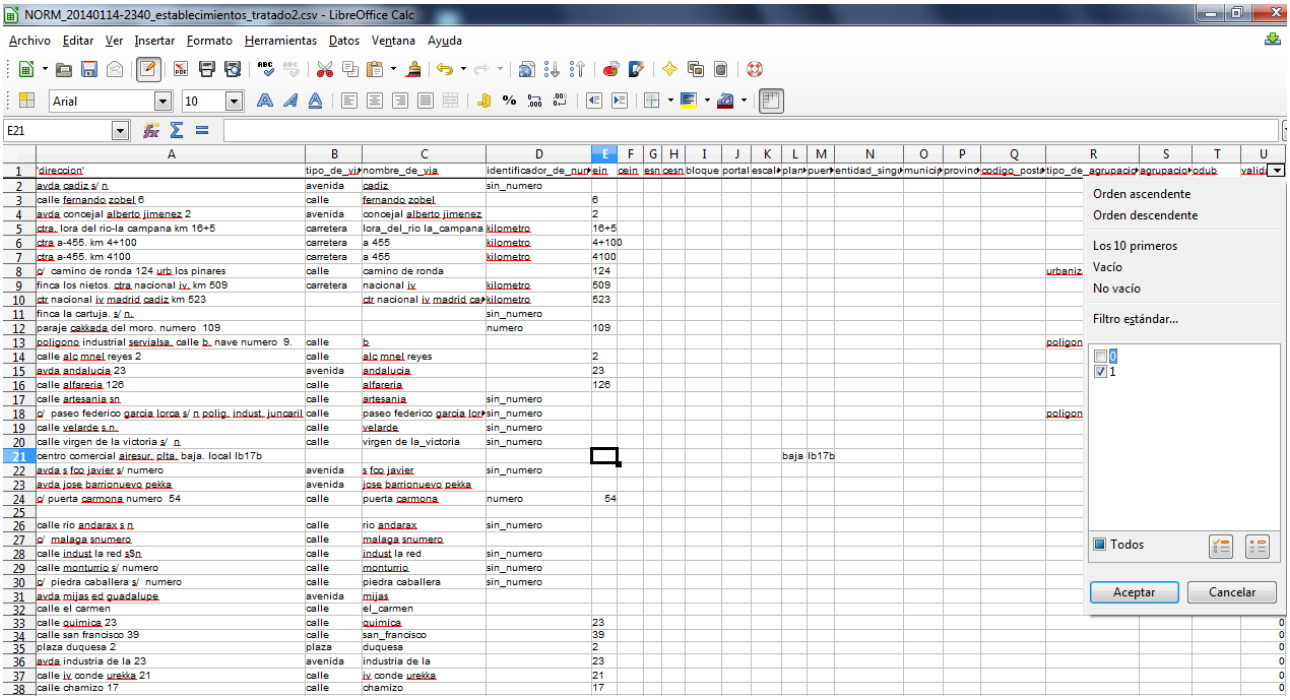


Imagen 92. Filtrado variable 'validacion'

A continuación, se copiarán todos estos registros en un nuevo fichero de LibreOfficeCalc. Para ello se seleccionarán todas las filas filtradas y se pulsará con el botón derecho del ratón sobre la columna de registros enumerados. De entre las opciones que aparecen se elegirá la de **Copiar** y seguidamente se abrirá una nueva hoja de cálculo pulsando en la opción **Nuevo** del menú **Archivo**.

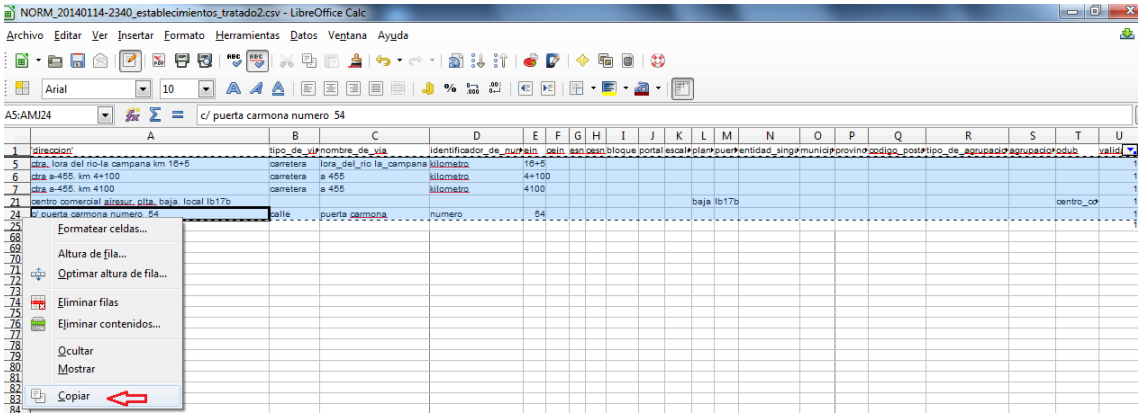


Imagen 93. Copiar registros con valor 1 en variable *validacion*

Todos estos registros constituirán un nuevo fichero que el usuario podría guardar con el nombre de "establecimientos_mal_normalizados.csv".

Análogamente del fichero normalizado se eliminarán todos los registros con validación 1. Para ello, se pulsará sobre las filas filtradas y pulsando con el botón derecho del ratón sobre la columna de registros enumerados se elegirá la opción **Eliminar filas**. A continuación, se quitará el filtro sobre la variable 'validacion' y se guardarán los cambios realizados. Será el usuario el que decida si guarda el fichero normalizado (que ya no tiene registros con valor 1) con el mismo nombre o con otro, pero **siempre** tendrá que tener **extensión .csv** y sus **elementos** tendrán que estar **separados por ";"**. Igualmente el nuevo fichero con registros incorrectamente normalizados tendrá que guardarse con **formato .csv** y sus **elementos** tendrán que estar **separados por ";"**. En las siguientes ventanas se muestra cómo realizar este proceso exactamente:

Para guardar el fichero el usuario indicará la denominación y ubicación del mismo y en el botón **Tipo** seleccionará la opción **Texto CSV (.csv)**:

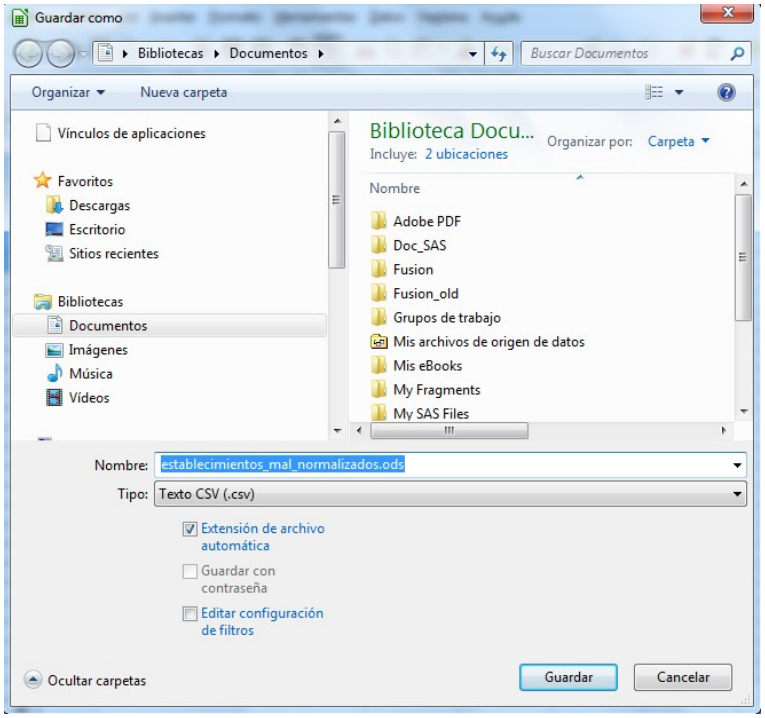


Imagen 94. Guardar registros con valor 1 en variable *validacion*

Una vez indicado pulsará el botón **Guardar** y aparecerá una ventana como la de abajo, en la que deberá pulsar el botón **Usar el formato Texto CSV**:

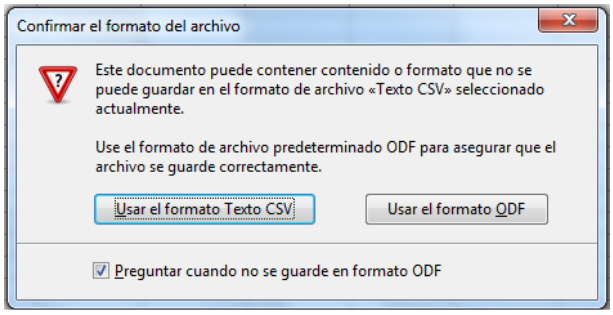


Imagen 95. Ventana de confirmación de formato CSV

Para finalizar deberá especificar los siguientes parámetros en esta nueva ventana:

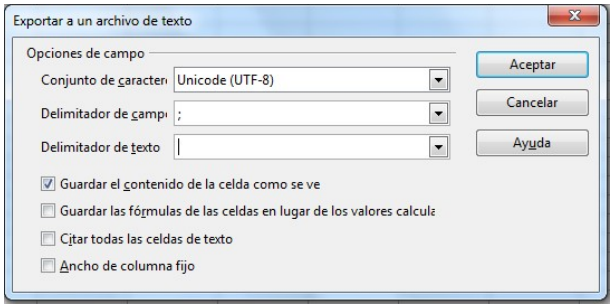


Imagen 96. Establecimiento de parámetros para guardar el fichero

De esta manera el usuario tendrá correctamente guardado su fichero con registros incorrectamente normalizados con la denominación "establecimientos_mal_normalizados.csv". A continuación, el usuario podrá seleccionar una muestra de este nuevo fichero utilizando la herramienta **HMM: Selección de la muestra** y entrenarla con la herramienta **HMM: Entrenamiento de la muestra** tal y como se ha indicado en los apartados 6.3.2 y 6.3.3. Con el nuevo modelo HMM obtenido el usuario normalizará el fichero "establecimientos_mal_normalizados.csv". Si todos los registros se han normalizado correctamente entonces el usuario solamente tendría que unir los dos ficheros normalizados. En caso de que no fuese así, tendría que volver a extraer los mal normalizados y realizar el mismo proceso de antes.

7 Herramienta de Enlace

La Herramienta de Enlace de *aLink: Herramienta de Fusión de Ficheros*, permite realizar de forma completa un proceso de enlace probabilístico de ficheros de datos. Para acceder a la misma se pulsará el botón correspondiente de la interfaz inicial de *aLink: Herramienta de Fusión de Ficheros*:

Imagen 97. Interfaz inicial de *aLink: Herramienta de Fusión de Ficheros*

7.1 Descripción general de la Herramienta de Enlace

La interfaz gráfica inicial de la Herramienta de Enlace se visualiza en la siguiente imagen:

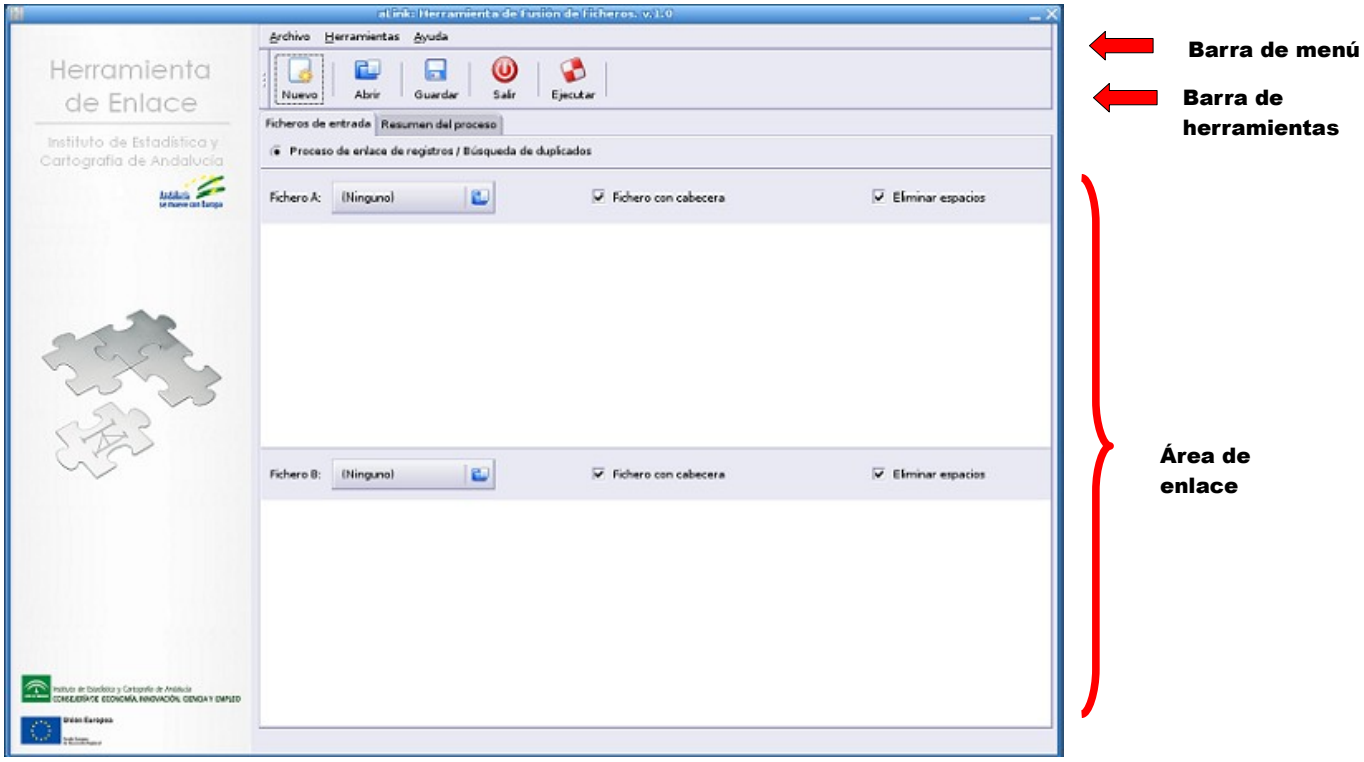


Imagen 98. Interfaz principal de la Herramienta de Enlace

Tal y como se aprecia, la interfaz está estructurada en tres partes: la parte superior de la ventana contiene una barra de menú y

justo debajo de ella aparece una barra de herramientas.

El resto de la ventana constituye la parte más importante de la interfaz, el área de enlace. En ella se especificarán los parámetros y requisitos del sistema necesarios para llevar a cabo un proceso de enlace en cada una de sus fases. En concreto, la herramienta consta de una serie de ventanas, exactamente nueve, en las que o bien se van a ir introduciendo los distintos parámetros necesarios para realizar un proceso de búsqueda de duplicados o de enlace de registros, o bien van a ir proporcionando información al usuario acerca del proceso realizado.

7.1.1 Barra de menú de la Herramienta de Enlace

La barra de menú contiene las opciones:

- **Archivo**: a través de este menú se puede iniciar un nuevo proceso de enlace, seleccionando **Nuevo**, salir de la Herramienta de Enlace, seleccionando **Salir** o bien, abrir o guardar un proyecto de enlace, seleccionando su correspondiente opción (**Abrir** o **Guardar como**).

Imagen 99. Menú Archivo de la Herramienta de Enlace

Por guardar un proyecto se entiende, almacenar en un fichero todos los parámetros del proceso de enlace que se esté llevando a cabo en ese momento, de tal forma que pueda ser cargado o abierto en la Herramienta posteriormente con el fin de repetir o modificar el proceso de enlace.

Para guardarlo se pulsará sobre la opción **Guardar como** y aparecerá una ventana similar a la siguiente:

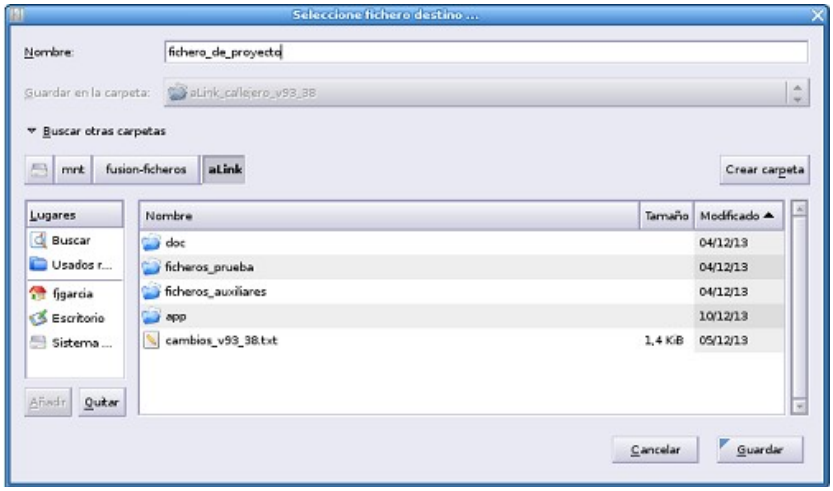


Imagen 100. Ventana Guardar fichero proyecto de enlace

El nombre del proyecto se escribirá sin ningún tipo de extension y se almacenará en la ubicación que desee el usuario para su posterior uso a través de la opción **Abrir**.

- **Herramientas**: esta opción permite al usuario, al igual que en el caso de la normalización, realizar un tratamiento previo de los ficheros a enlazar (**Tratamiento previo**), insertar índices a los ficheros que se van a enlazar (**Insertar índices**), incluir al fichero de enlaces los campos que se deseen de los ficheros que se han enlazando (**Incluir campos a enlaces**) y eliminar de los ficheros que se están enlazando los registros que han enlazado en alguno de los anteriores procesos de enlace (**Eliminar registros enlazados**).

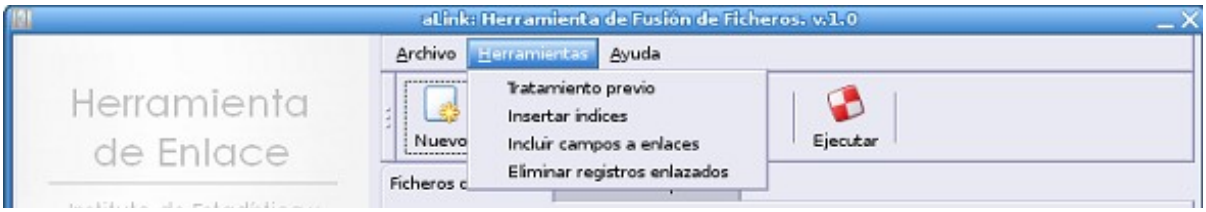


Imagen 101. Menú Herramientas de la Herramienta de Enlace

Las herramientas que se incluyen en este menú se analizarán con más detalle en el apartado 7.2 de este Manual.

- **Ayuda**: este menú ofrece al usuario la misma información que el menú Ayuda de la Herramienta de Normalización.

Imagen 102. Menú Ayuda de la Herramienta de Enlace

7.1.2 Barra de herramientas de la Herramienta de Enlace

Imagen 103. Barra herramientas de la Herramienta de Enlace

La barra de herramientas cuenta con los siguientes botones:

- **Nuevo**: su funcionamiento es equivalente a la opción 'Nuevo' del menú Archivo.
- **Abrir**: su funcionamiento es equivalente a la opción 'Abrir' del menú Archivo.
- **Guardar**: su funcionamiento es equivalente a la opción 'Guardar como' del menú Archivo.
- **Salir**: su funcionamiento es equivalente a la opción 'Salir' del menú Archivo.
- **Ejecutar**: al pulsar este botón la aplicación cargará la información que el usuario haya introducido en la pestaña del área de enlace, generará una nueva pestaña para la siguiente fase de enlace y en el caso de que el usuario se encuentre dentro de la última pestaña ('Salida') se ejecutará el proceso de enlace. **¡OJO!** Hay que remarcar que para avanzar en el proceso de enlace y aparezcan el resto de pestañas es necesario pulsar siempre el botón **Ejecutar**.

7.1.3 Área de enlace

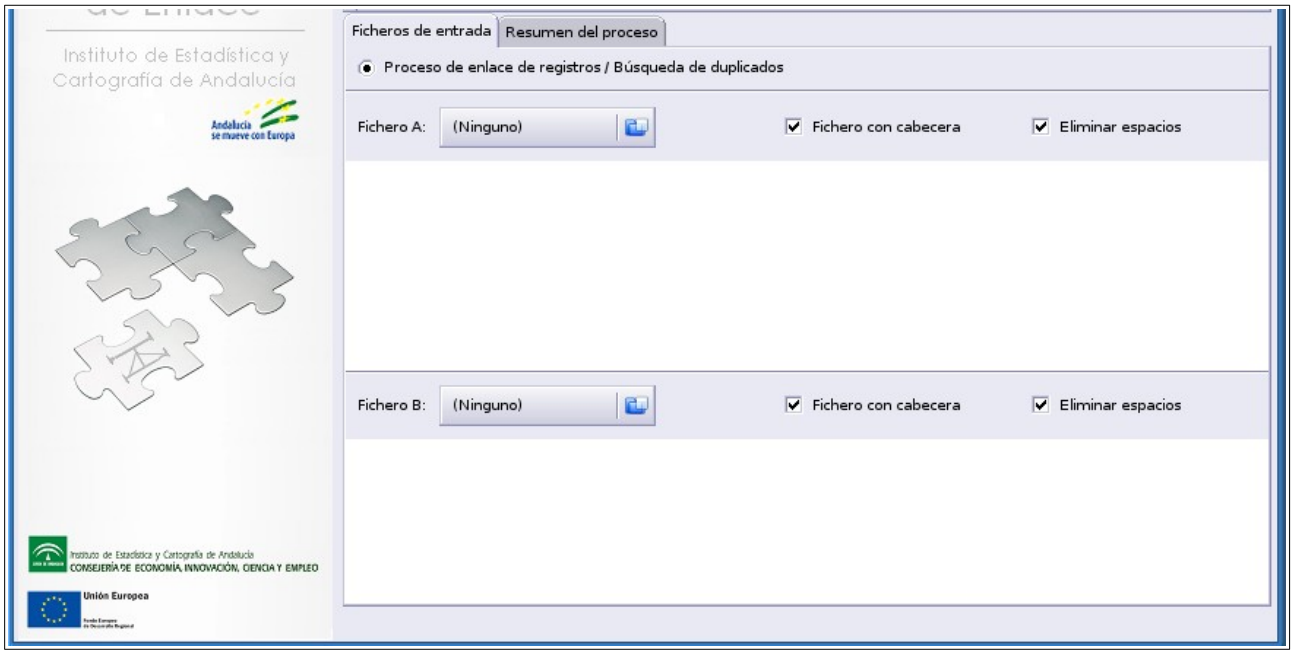


Imagen 104. Área de enlace de la Herramienta de Enlace

El área de enlace consta en un principio de dos pestañas, 'Ficheros de entrada', que es la que está activada por defecto y 'Resumen del proceso'. El resto de pestañas aparecerán conforme el usuario vaya introduciendo la información de cada una de las fases del proceso de enlace y pulse el botón 'Ejecutar'. A continuación, se analizan detenidamente.

7.1.3.1 Pestaña Ficheros de entrada

En esta pestaña se introducirán los ficheros que se desean enlazar o bien, si se trata de un proceso de búsqueda de duplicados, se introducirá el mismo fichero como 'Fichero A' y como 'Fichero B'. **Es imprescindible que los ficheros que se introduzcan estén tratados e indexados con las herramientas que *aLink: Herramienta de Fusión de Ficheros* proporciona al respecto y que tengan formato CSV separados por punto y coma.** Estas herramientas son la de ***Tratamiento previo*** y la de ***Insertar índices***, las cuales se encuentran en el menú Herramientas de la aplicación.

Tanto a la hora de introducir el 'Fichero A', como el 'Fichero B' se aprecian dos casillas de verificación marcadas que el usuario puede desmarcar, en función de sus necesidades. Estas son:

- **Fichero con cabecera:** al marcar esta opción el usuario indica que el fichero que va a enlazar posee una primera fila con las denominaciones de los campos. Esta información aparecerá en negrita dentro de la ventana de previsualización del fichero de datos.

Imagen 105. Ventana de previsualización de datos del Fichero A con la cabecerá del fichero en negrita

En el caso de que el fichero no tenga cabecera el usuario desmarcará esta opción y de forma automática aparecerán unas denominaciones de campos por defecto ('field- '). El usuario puede modificarlas haciendo doble click sobre cada una de ellas.

Imagen 106. Ventana de previsualización de datos del Fichero A. En este caso el fichero no tiene cabecera

- **Eliminar espacios:** al marcar esta opción se eliminan los espacios en blanco que existen al principio y al final de los campos que se van a utilizar para el proceso de enlace o búsqueda de duplicados. Es de gran utilidad por lo que se recomienda dejarla marcada.

Finalmente, cuando el usuario haya introducido los ficheros a enlazar deberá pulsar el botón 'Ejecutar' con el fin de confirmar dicha selección.

7.1.3.2 Pestaña Resumen del proceso

Esta pestaña contiene un resumen de todos los parámetros que se van estableciendo en el proceso de enlace o de búsqueda de duplicados, así como de las funciones que se utilizan y de los resultados obtenidos. Así pues, cuando se introduzcan los valores de los parámetros en cada una de las pestañas del proceso de enlace y se pulse el botón 'Ejecutar' estos cambios quedarán reflejados en esta pestaña.

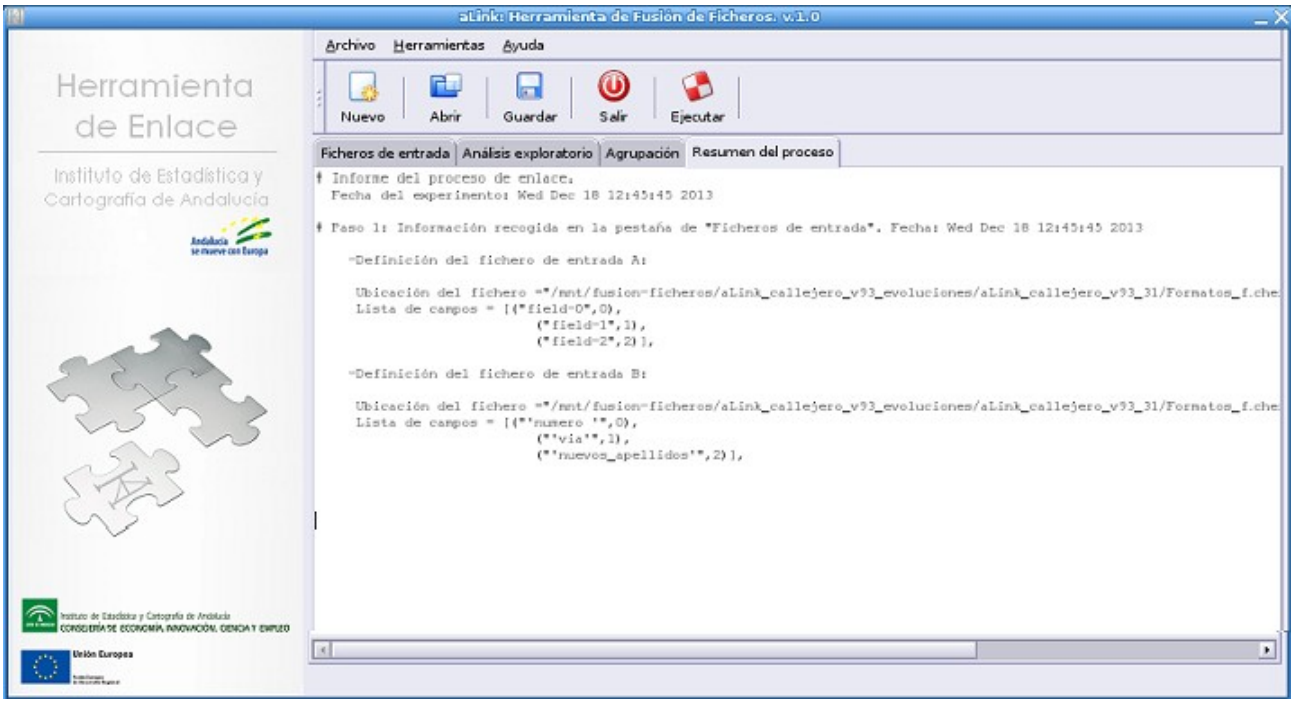


Imagen 107. Pestaña de Resumen del proceso una vez cargados los Ficheros A y B

7.1.3.3 Pestaña Análisis exploratorio

Esta pestaña surge cuando el usuario introduce los ficheros a enlazar en la pestaña 'Ficheros de entrada' y pulsa el botón 'Ejecutar'. Junto a ella también aparece la pestaña 'Agrupación', que se estudiará en el siguiente apartado. La pestaña 'Análisis exploratorio' presenta el siguiente aspecto:

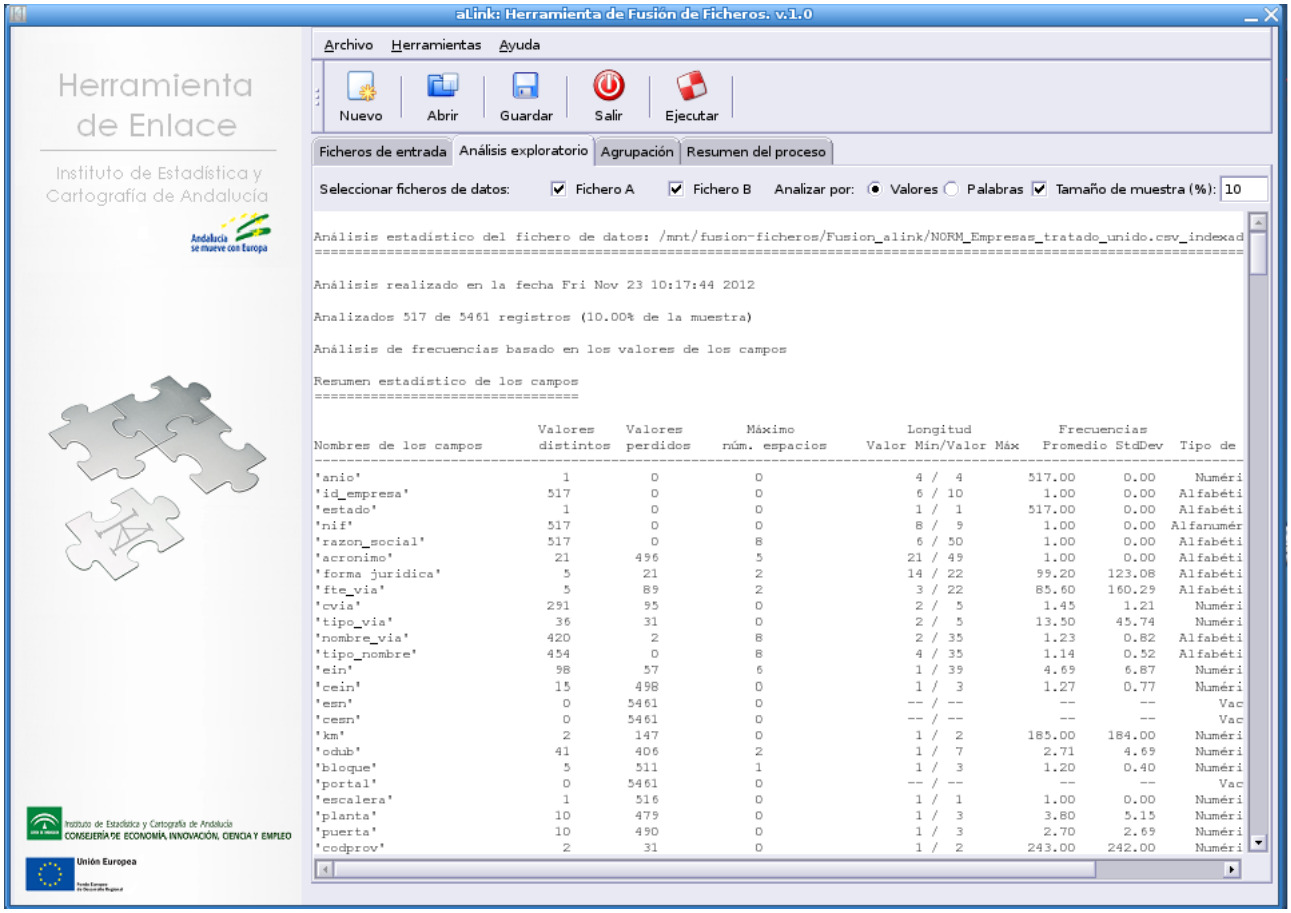


Imagen 108. Vista de la pestaña Análisis exploratorio

Esta ventana permite al usuario analizar una muestra de registros de los ficheros de trabajo o bien los ficheros en su totalidad (campo 'Tamaño de la muestra (%)'). El análisis se puede llevar a cabo de dos formas:

- Considerando los valores de los campos como palabras separadas (Palabras)
- Considerando los valores de los campos como valores completos (Valores)

Por ejemplo, suponiendo que se analizan los valores de la variable nombre; si dicha variable contiene el valor 'maria del carmen' y se activa la casilla de verificación de 'análisis de palabras' se realizará un análisis de frecuencias considerando las palabras: 'maria', 'del' y 'carmen'. Si no se activa, el análisis de frecuencias se realizará considerando el valor 'maria del carmen'.

Para cada campo (columna/atributo) se recoge, en una tabla, la siguiente información en función del tamaño de muestra seleccionado:

- El número de valores únicos o de valores distintos.
- La frecuencia media y la desviación estándar de los valores.
- La longitud del valor mínimo y máximo.
- Si el campo es numérico, alfabético o alfanumérico.
- El número máximo de espacios en blanco dentro de los valores.
- El número de registros con valores perdidos.

Posteriormente, se genera una tabla que resume las estadísticas de los cuantiles y además, se genera otra tabla que contiene detalles de la idoneidad de los campos para la agrupación (de acuerdo a su número de valores y a la proporción de valores perdidos).

Así pues, a través de esta ventana el usuario puede hacerse una primera idea acerca de las variables candidatas a formar parte de la siguiente etapa de agrupación.

7.1.3.4 Pestaña Agrupación

Como se ha indicado en el apartado anterior esta pestaña surge a la vez que la de Análisis exploratorio. En esta pestaña se seleccionan las variables por las que se realizará la agrupación dentro del proceso de enlace. Cada valor de la variable de agrupación dará lugar a un grupo, de manera que al agrupar los registros según los valores de las variables de agrupación se busca reducir el número de comparaciones a realizar. De esta forma en la siguiente fase del proceso de enlace, que es la de comparación, únicamente se compararan aquellos registros de ambos ficheros que coincidan en los valores de la variable de agrupación. La ventana de la etapa de agrupación presenta la siguiente información:

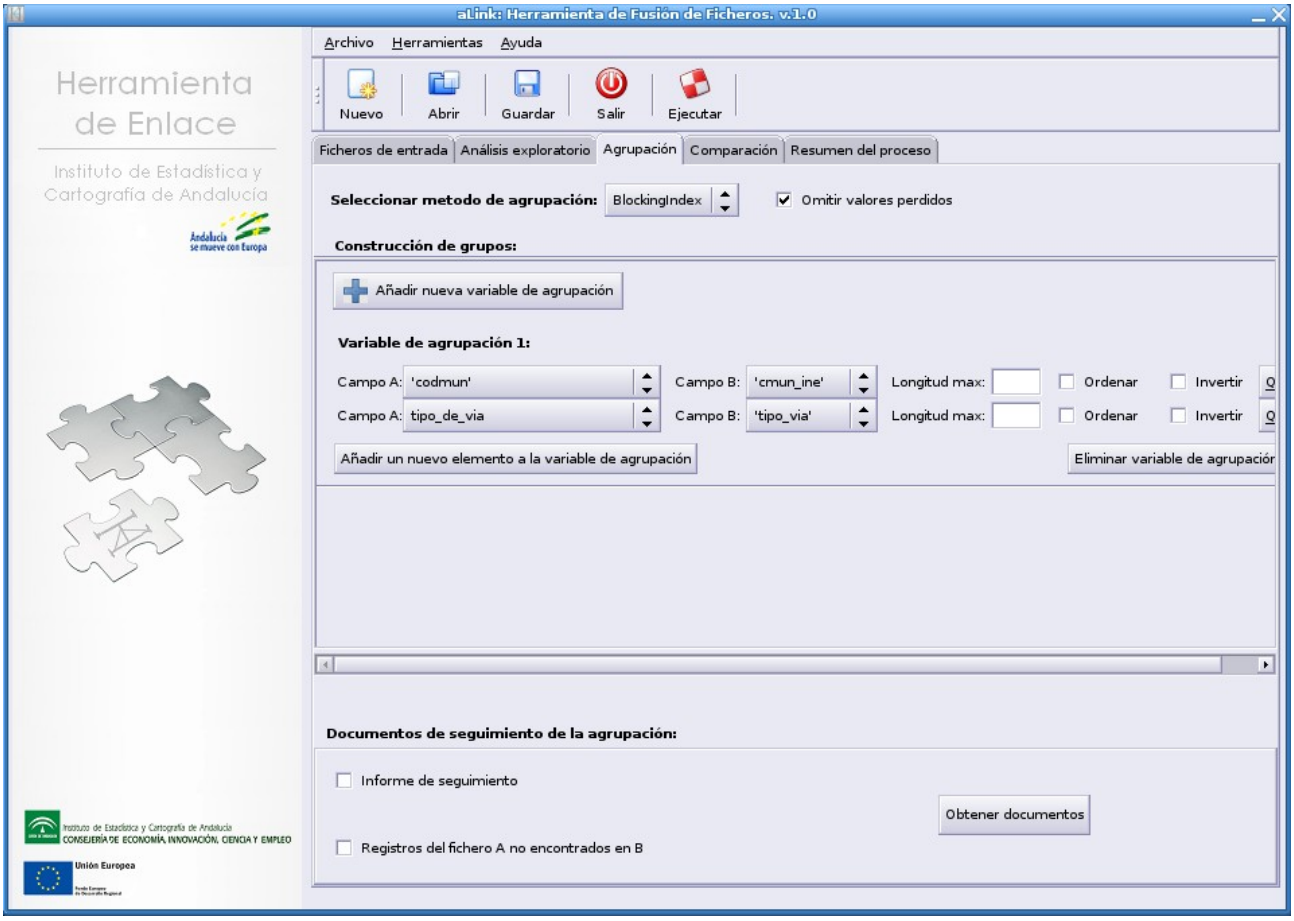


Imagen 109. Vista de la pestaña Agrupación

En esta pestaña el usuario deberá elegir entre las siguientes opciones:

- **Seleccionar método de agrupación:** puede elegir entre dos métodos de agrupación (blocking index y sorting index) o no utilizar ninguno de ellos (full index). (Ver Anexo XI para descripción más detallada de estos métodos y del proceso 'full index').
- **Omitir valores perdidos:** si está activada la casilla, los valores perdidos que existan en la variable de agrupación definida no serán tenidos en cuenta a la hora de realizar los grupos, es decir, no habrá un grupo de valores perdidos.
- **Construcción de grupos:** en este apartado se definen las variables de agrupación que se van a utilizar, éstas pueden

ser una o varias. Para la definición de cada una de ellas podemos utilizar un solo campo o variable, así como la concatenación de varios de ellos usando el botón 'Añadir un nuevo elemento a la variable de agrupación'. Además, en la definición de la variable de agrupación se pueden utilizar operaciones como: el truncado de los valores de un campo, sin más que indicar la **longitud máxima de caracteres** que se desean considerar del campo en cuestión, la **ordenación alfabética** de las palabras que forman el campo (ordenar) y la **inversión alfabética** de los valores del campo (invertir). Para el caso que se quiera utilizar más de una variable de agrupación se definirá la primera y se utilizará el botón 'Añadir nueva variable de agrupación' para definir la segunda y así sucesivamente si se decidiera añadir una tercera, cuarta, etc. Igualmente, si se decide eliminar alguna de las variables de agrupación definidas, el usuario podrá hacerlo pulsando el botón 'Eliminar variable de agrupación'.

Por ejemplo, supongamos que se disponen de dos ficheros A y B que se desean enlazar. Cada uno de estos ficheros presenta la siguiente información:

Fichero A				
DNI_A	**Nombre_A**	**Apellidos_A**	**Ciudad_A**	**Pais_A**
75098174W	antonio	galera	Cordoba	Espana
75098175K	maria	gomez	Cordoba	Espana
75098176P	marta	gomez	Cordoba	Argentina
75098177R	jose	aimar	Cordoba	Argentina

Tabla 2. Ejemplo Fichero A

Fichero B			
DNI_B	**Nombre_B**	**Apellidos_B**	**Direccion_B**
75098174W	antonio jesús	galera	C/ sol 24
	Maria	gomez	C/ luna 3 puerta 1
	marta	gomez	Avenida astronomia 14
75098177R	Jose luis	aimar	Avd de la libertad 3, plta 2, puerta 1

Tabla 3. Ejemplo Fichero B

Suponiendo que vamos a utilizar los campos DNI y Apellidos de cada fichero para la fase de agrupación podremos crear lo siguiente:

a) Una variable de agrupación como concatenación de los campos DNI y los dos primeros caracteres del campo Apellidos de cada fichero. En este caso, una vez que se formen los grupos, se compararán en la siguiente fase aquellos registros de ambos ficheros que tengan el mismo DNI y los mismos dos primeros caracteres del campo Apellidos. El aspecto de la pestaña de agrupación sería el siguiente:

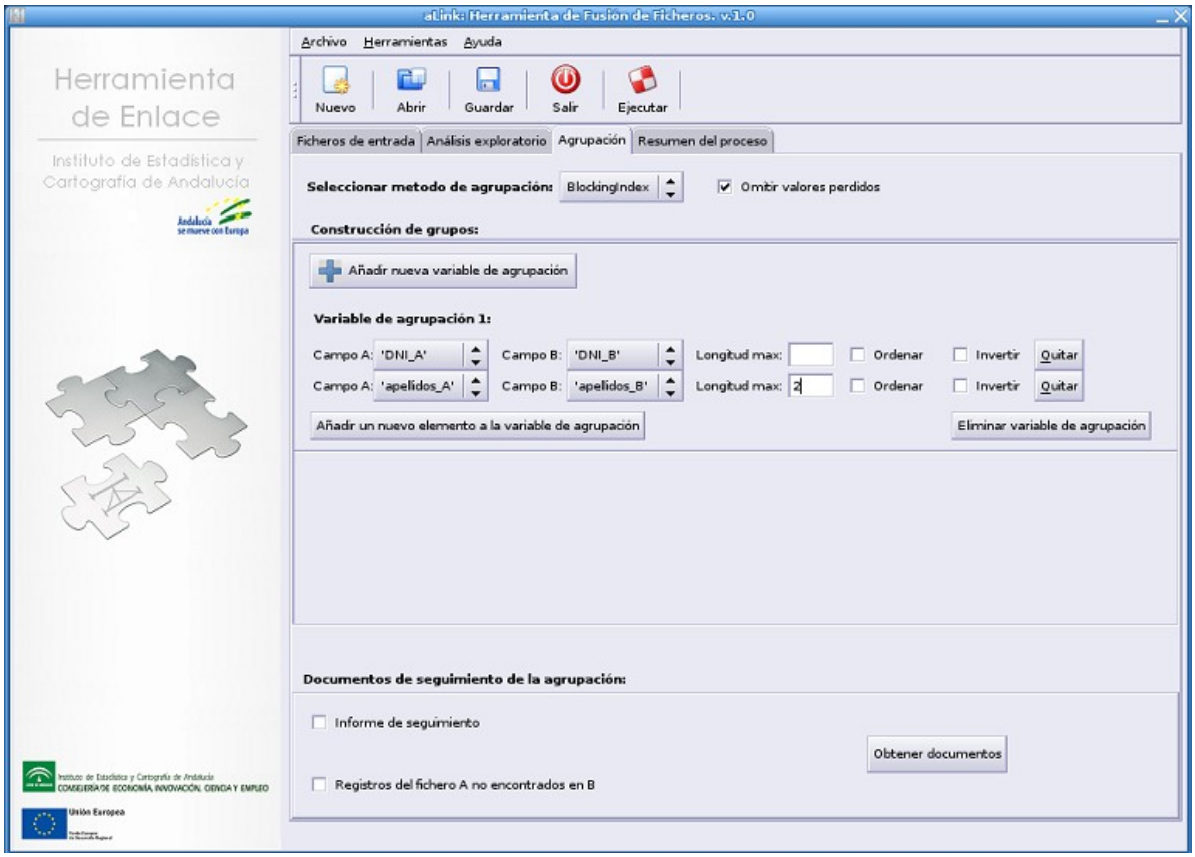


Imagen 110. Creación de una variable de agrupación

Obsérvese que al incluir los dos campos que forman la variable de agrupación han aparecido en la interfaz, al lado de la opción 'Invertir' dos nuevos botones denominados 'Quitar'. Estos permiten al usuario eliminar alguno de ellos si lo considera necesario.

b) Dos variables de agrupación donde la primera será el DNI y la segunda el campo Apellidos de cada fichero. De esta forma se comparará, en primer lugar, todos los registros que coincidan en el DNI. Seguidamente para todos aquellos registros que no coincidan en el DNI, se compararán en función de que coincidan en el campo Apellidos. Para este segundo ejemplo, el aspecto de la pestaña de agrupación sería el siguiente:

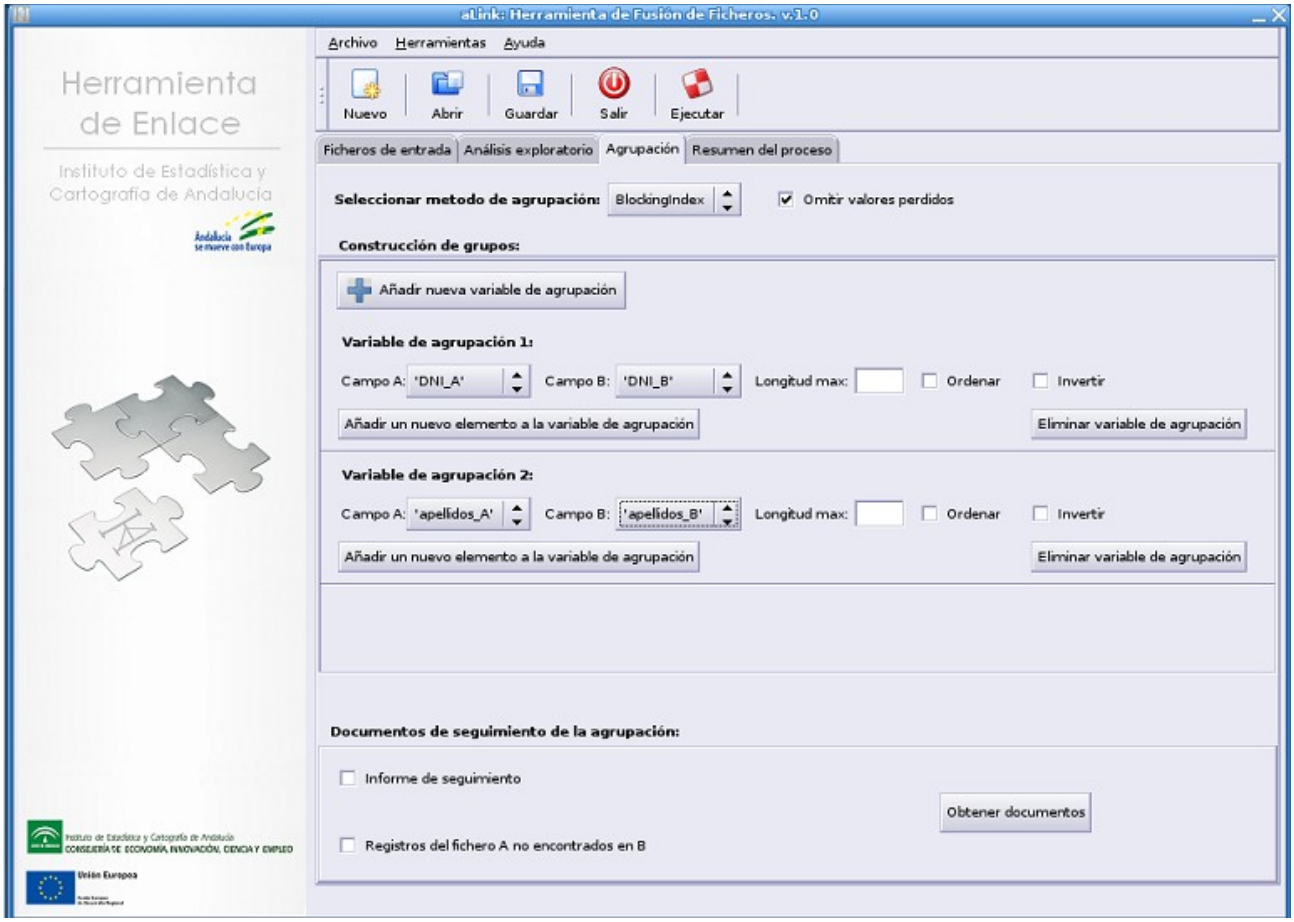


Imagen 111. Creación de dos variables de agrupación

- **Documentos de seguimiento de la agrupación:** en esta sección se obtiene diversa información acerca de la etapa de agrupación, con el objeto de que antes de pasar a la siguiente etapa (comparación), el usuario, pueda decidir sobre la idoneidad del método y el criterio o variable de agrupación que ha elegido. En concreto ahora mismo la información que aparece en relación a la variable de agrupación elegida es:

  - Informe de seguimiento: se trata de un fichero con extensión .txt que contiene la definición de la variable de agrupación utilizada tanto en el fichero A como en el B, los grupos que se han formado en los ficheros A y B al elegir dicha variable de agrupación, los grupos que son comunes en ambos ficheros, los grupos del fichero A que no están en el fichero B y los grupos del fichero B que no están en A. También se ofrecen una serie de indicadores sobre la bondad del proceso de agrupación. Visualmente presenta la siguiente forma:

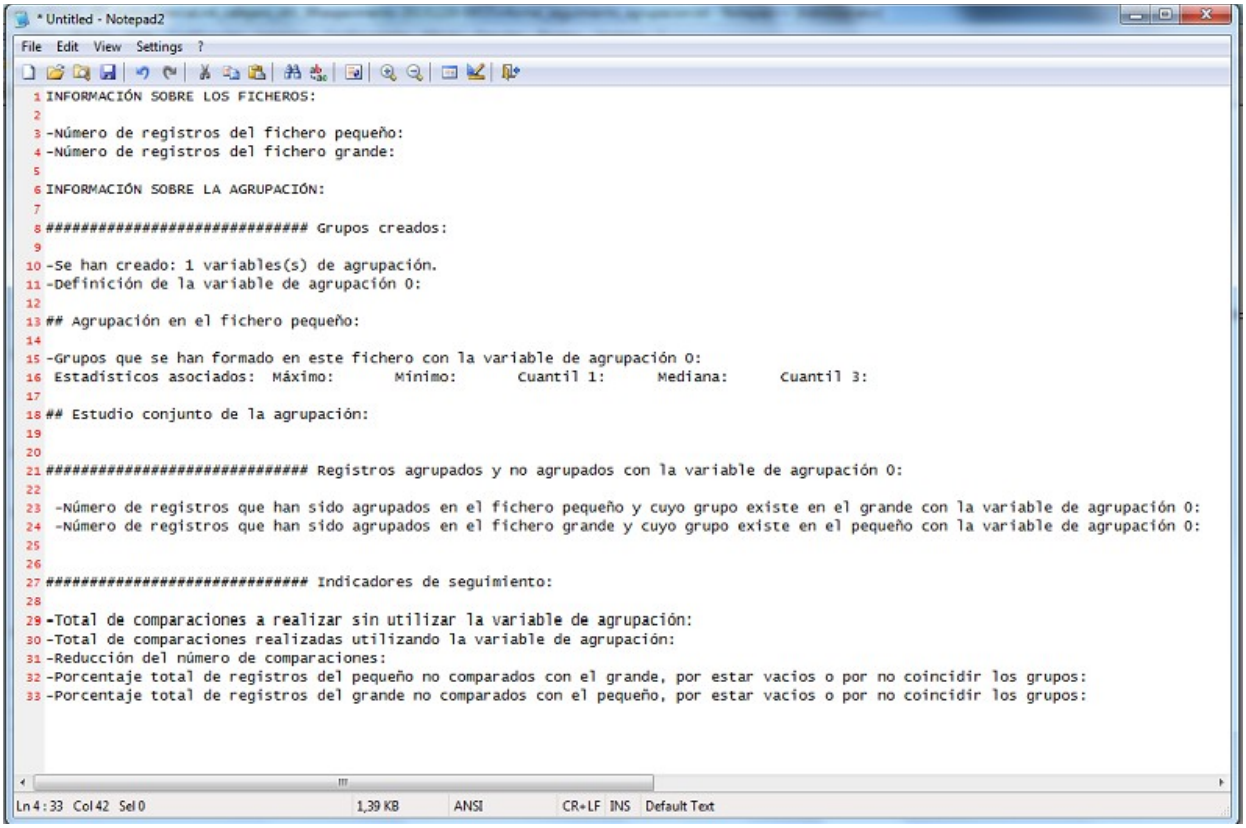


Imagen 112. Estructura del informe de seguimiento

- Registros del fichero A no encontrados en B: es un fichero que contiene los registros del fichero A cuyo grupo no se encuentra en B y que por tanto, no van a ser comparados. Este fichero tiene extensión .csv.

Para obtener dichos ficheros el usuario deberá pulsar el botón **Obtener documentos**. Estos quedarán almacenados en una carpeta cuya denominación sigue el formato: *experimento_<fecha_creación><hora_creación>* y que se ubica dentro del directorio donde se encuentran los ficheros a enlazar.

7.1.3.5 Pestaña Comparación

En esta pestaña se llevará a cabo la etapa de comparación del proceso de enlace de registros. Al ya tener agrupados los registros de ambos ficheros en función a las variables de agrupación establecidas en la fase anterior, se compararan únicamente aquellos registros de ambos ficheros que coincidan en los valores de dichas variables de agrupación. Para ello se utilizan una serie de funciones de comparación que permiten comparar cadenas de caracteres y cadenas numéricas, tanto de forma exacta como aproximada.

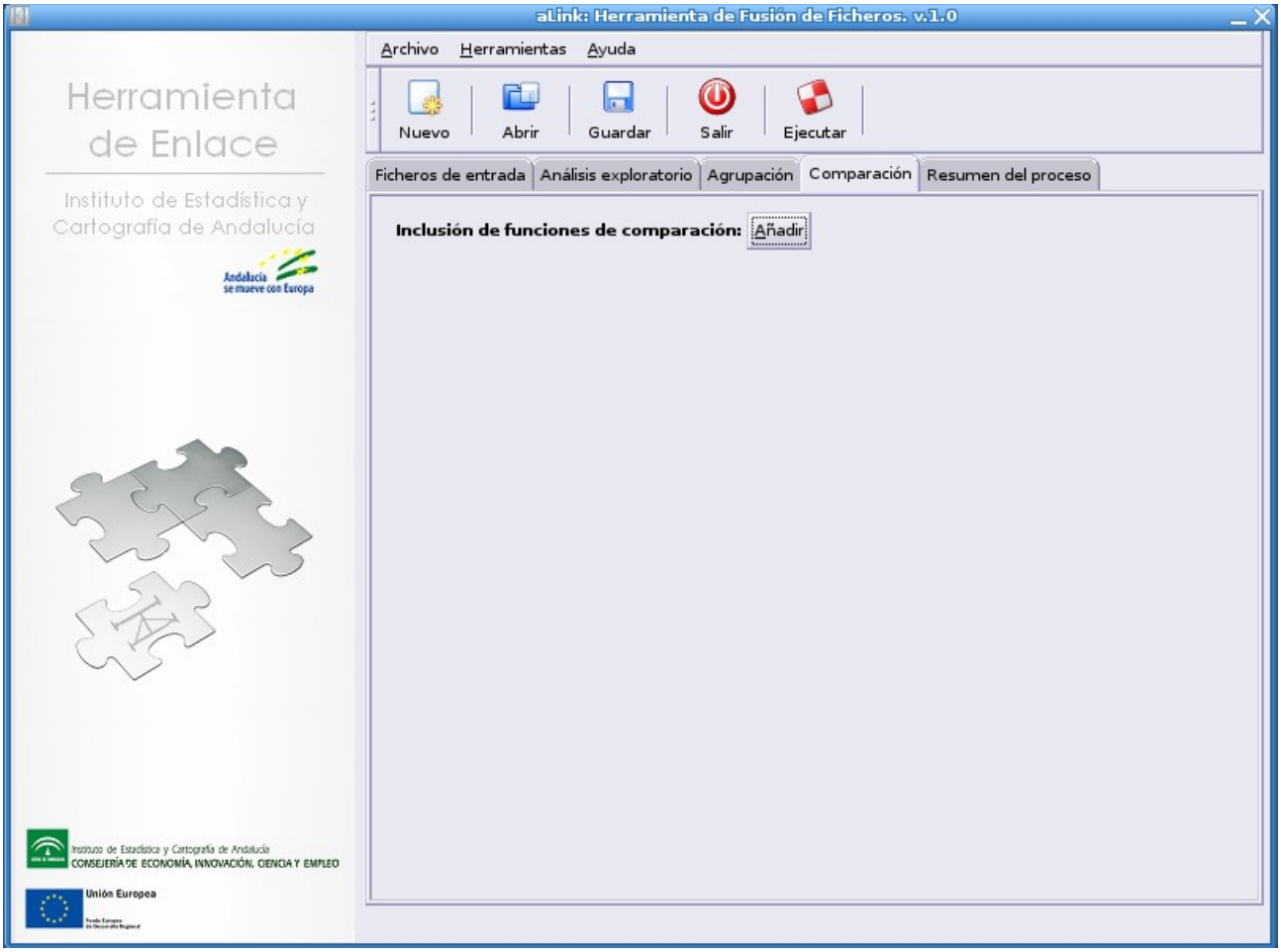


Imagen 113. Aspecto inicial de la pestaña Comparación

El usuario puede incluir tantas funciones de comparación como considere necesario para su proceso de enlace, para ello bastará con pulsar el botón 'Añadir'. Para cada inclusión se muestra la siguiente información:

- **Función de comparación**: mide la similitud entre dos cadenas de caracteres o valores numéricos. El usuario puede elegir entre las siguientes:

  - Función de comparación de cadena exacta (Str-Exact).

  - Función de comparación de cadena contenida (Str-Contains).

  - Función de comparación de cadena truncada (Str-Truncate).

- Función de comparación de cadena aproximada de Jaro (Jaro).
- Comparación de cadena aproximada de Winkler (Winkler).
- Comparación de cadena aproximada con la distancia de edición (Edit-Dist).
- Comparación de cadena aproximada con la distancia de Damerau-Levenshtein (Dam-Le-Edit-Dist).
- Comparación de cadena aproximada con la distancia Bag (Bag-Dist).
- Comparación de cadena aproximada con la distancia de Smith-Waterman (Smith-Water-Dist).
- Comparación de cadena aproximada Seq-Match (Seq-Match).
- Comparación de porcentaje numérico (Num-Perc).
- Comparación numérica absoluta (Num-Abs).
- Función de comparación Key-diff (Key-Diff).

En el Anexo XII se muestra una descripción más detallada de las mismas.

- **Campo A y Campo B:** en los combos aparecen los nombres de todos los campos de los ficheros A y B respectivamente, por lo que el usuario solo tendrá que seleccionar el que desee usar para comparar. Por ejemplo, si se quieren comparar direcciones postales, el usuario comparará los campos que tengan información sobre viales de ambos ficheros.
- **Peso valor perdido:** peso establecido por el usuario para aquellos casos en los que se comparan los valores de dos campos y uno de ellos o los dos es un valor perdido. Por defecto su valor es 0.0.
- **Peso coincidencia:** peso establecido por el usuario cuando coinciden exactamente los campos comparados. Por defecto su valor es 1.0.
- **Peso mínimo:** peso mínimo de comparación. Si en la pestaña de Clasificación se escoge el método *FellegiSunter,* ningún par de registros se clasificará como enlace si, de manera individual, en esa variable tiene un peso por debajo del mínimo establecido, independientemente del peso total de todas las variables.
- **Peso no coincidencia:** peso establecido por el usuario cuando no coinciden exactamente los campos comparados. Por defecto su valor es 0.0.
- **Eliminar función de comparación**: botón que permite eliminar la función de comparación elegida por el usuario.

Existen otra serie de parámetros en relación con la función de comparación utilizada. Por ejemplo, en aquellos casos en los que la función compara de forma aproximada se puede dar un valor umbral para determinar a partir de qué valor se va considerar una coincidencia.

Finalmente, en esta etapa de comparación se tiene que hacer especial mención a los casos en los que se comparan nombres y apellidos de personas y los ficheros no disponen de un identificador único de las mismas. La aplicación de cada función de comparación da como resultado un peso o puntuación según la concordancia entre los caracteres de los valores comparados. Para ello hay que tener en cuenta lo comunes o raros que pueden ser dichos valores en la población, es decir, tener en cuenta la frecuencia de aparición de estos nombres y apellidos.

Supongamos el siguiente ejemplo al comparar el campo ‘primer apellido’ de dos ficheros, A y B.

FICHERO A	FICHERO B
Apellido1	**Apellido1**
goikoetxea	goikoetxea
Peso de concordancia utilizando una función de comparación: 1	

Tabla 4. Comparación de dos apellidos raros

FICHERO A	FICHERO B
Apellido1	**Apellido1**
garcia	garcia
Peso de concordancia utilizando una función de comparación: 1	

Tabla 5. Comparación de dos apellidos no raros

Ambos tienen el mismo peso de concordancia independientemente de la función de comparación usada, pero sin embargo las frecuencias de aparición en la población no son las mismas en el total de España. Así se tiene:

Apellido1	Frecuencia (Total de España)
garcia	1.481.923
goikoetxea	2.221

Tabla 6. Frecuencia de aparición de apellidos

Esta situación hace intuir que los apellidos o nombres raros o pocos comunes tienen mayor poder de discriminación a la hora de determinar si un par de registros corresponde al mismo individuo. Por ejemplo, teniendo en cuenta las frecuencias anteriores, es más probable que de los 2.221 individuos con primer apellido goikoetxea, dos de ellos sean el mismo, que de los 1.481.923 que se apellidan garcia lo sean, a pesar de que en ambos casos el peso de coincidencia es 1.

Por lo tanto, se tiene que ponderar ese peso según la aparición del apellido o nombre en una tabla de frecuencias previamente construida. En la Herramienta de Enlace esta situación está prevista y es por ello por lo que se dispone de un fichero auxiliar que tiene la siguiente estructura:

apellido1, frecuencia
garcia, 1.481.923
gonzalez, 933.497
..., ...
pardo, 44.037
..., ...
goikoetxea, 2.221
..., ...

Tabla 7. Detalle del fichero de frecuencias

y es el que se utiliza en el cálculo del peso de concordancia.

Luego, en el caso de no disponer de campos que en teoría identifiquen unívocamente al individuo como por ejemplo, DNI, NIF, número de afiliación a la Seguridad Social, etc., se tendrá que recurrir a otro tipo de campos para realizar las comparaciones, como por ejemplo: nombres y apellidos de los individuos, fechas y lugar de nacimiento, domicilio, etc. En estos casos, el uso de tablas de frecuencias solamente tiene sentido cuando se tengan coincidencias exactas que es donde se necesita tener alguna herramienta que permita discriminar para decidir si dos entidades son la misma. Para coincidencias aproximadas, en principio, ya se sabe que los dos individuos no coinciden (aunque puedan ser el mismo), con lo cual ya no habría que discriminar sino que el estudio en este caso sería otro.

Por estos motivos, se ha implementado en la Herramienta de Enlace la metodología relativa a tablas de frecuencias para la función de comparación de cadena exacta. Para ello se ha incluido en la interfaz de la pestaña 'Comparación', un combo que permite añadir un fichero auxiliar que contiene las tablas de frecuencias relativas a nombres y apellidos cuando se elige la función de comparación de 'Cadena exacta' (Str-Exact).

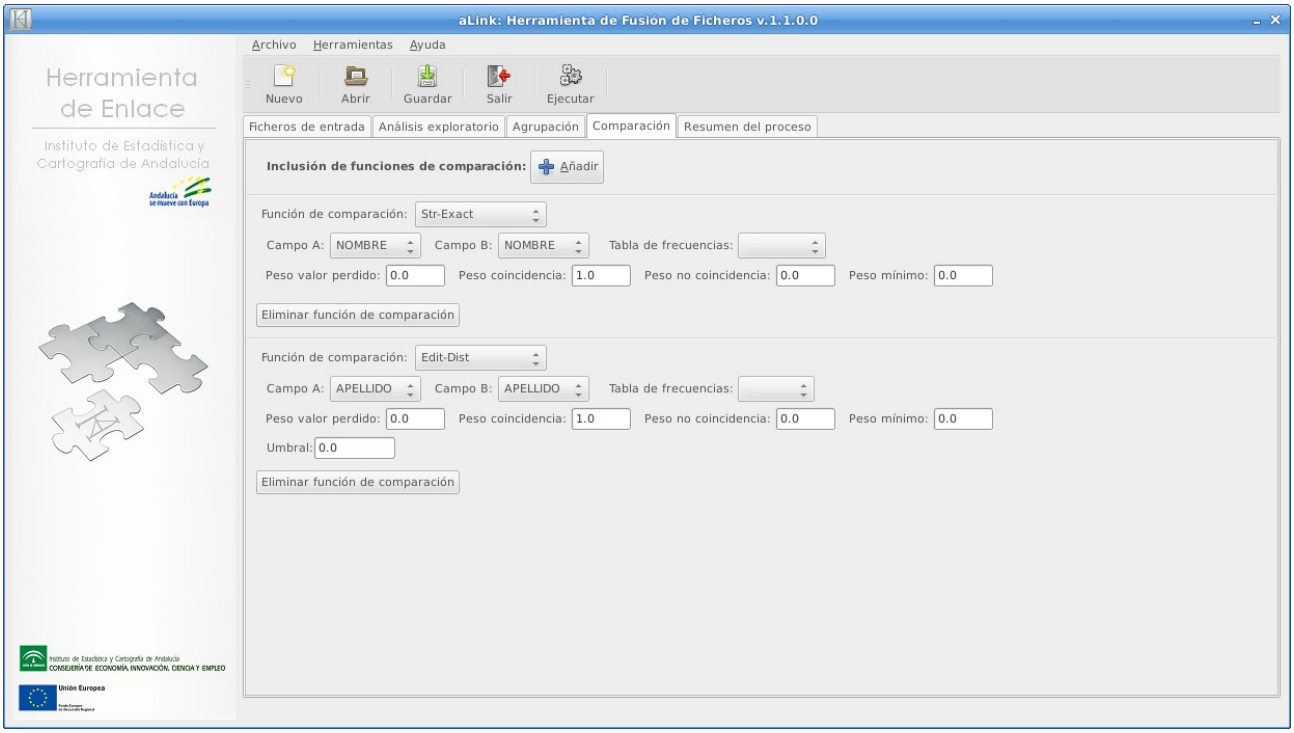


Imagen 114. Aspecto de la pestaña Comparación con dos funciones de comparación (Str-Exact y Edit-Dist)

7.1.3.6 Pestaña Clasificación

En esta ventana el usuario va poder establecer el método mediante el cual se van a clasificar los pares de registros comparados.

La aplicación permite trabajar con dos métodos de clasificación: el basado en la metodología de Fellegi y Sunter (*FellegiSunter*) y el basado en el clasificador de dos pasos (*TwoSteps*). En el Anexo XIII se realiza una breve descripción de los mismos.

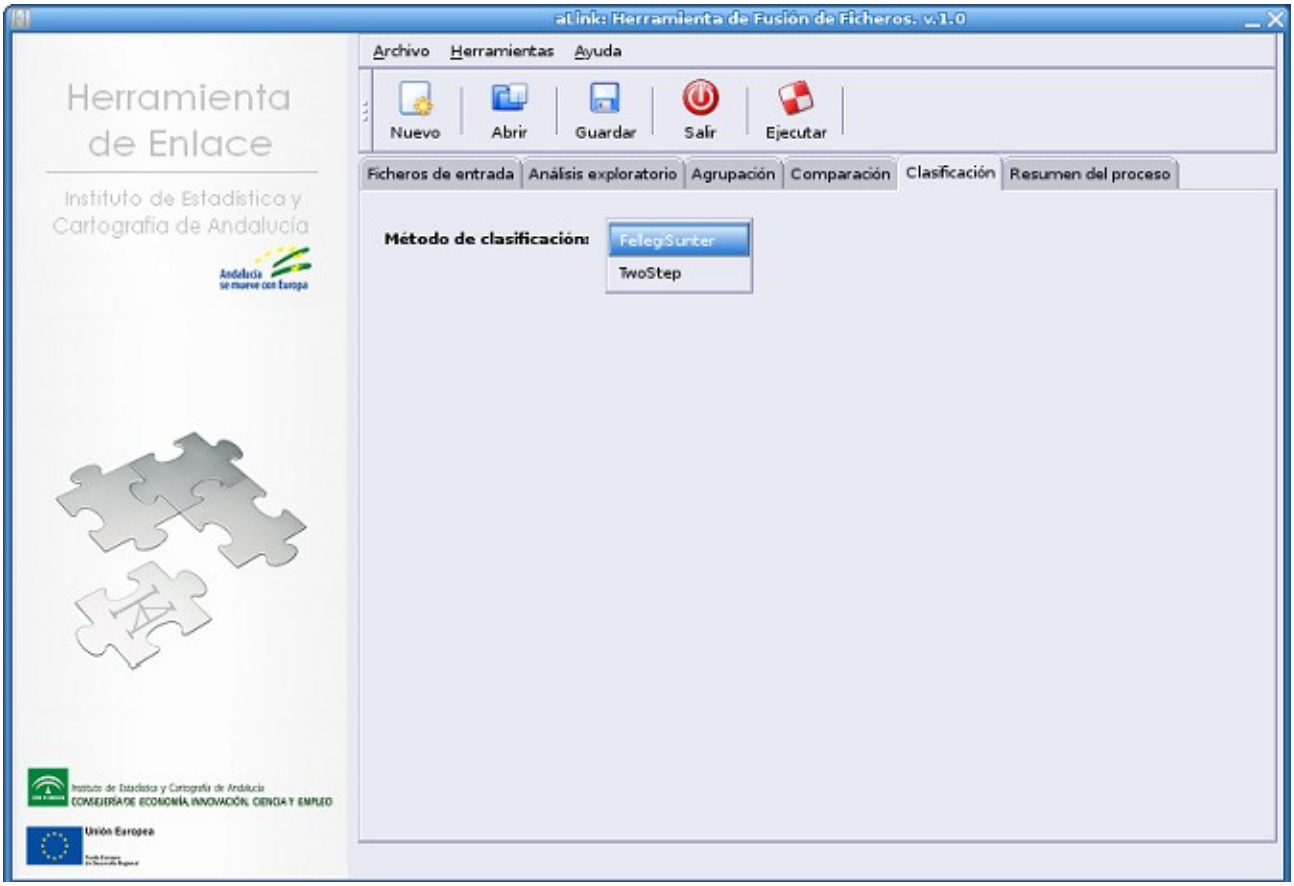


Imagen 115. Selección del método de clasificación en la pestaña Clasificación

Detalladamente:

a) Si se selecciona FellegiSunter como método de clasificación, el usuario tiene que especificar dos valores umbral, uno superior y otro inferior, que servirán para clasificar los pares de registros comparados en enlaces, no enlaces y posibles enlaces. El valor umbral inferior debe ser más pequeño o igual que el umbral superior, en caso contrario aparecerá un mensaje de error al pulsar el botón 'Ejecutar' del proceso de enlace. Además, el valor umbral superior será como máximo igual a la suma de los pesos de coincidencia que el usuario haya establecido en la fase anterior de comparación, es decir, si el usuario deseara comparar tres campos de una dirección postal como por ejemplo, el tipo de vía, el nombre de la vía y el número y los pesos que ha establecido para una coincidencia exacta son 1, 3 y 4 respectivamente, entonces como máximo el valor umbral superior que podrá incluir es 8 (8 = 1+3+4). Igual pasa con el valor umbral inferior, es decir, el valor mínimo que tomará será la suma de los pesos establecidos para un desacuerdo total (pesos de no coincidencia).

Así, la clasificación usando esta metodología se realizará en el siguiente sentido:

- Los pares de registros comparados cuyo peso esté por encima del umbral superior se clasificarán como enlaces.
- Los pares de registros comparados cuyo peso esté por debajo del umbral inferior se clasificarán como no enlaces.
- Los pares de registros comparados cuyo peso esté entre los umbrales inferior y superior se clasificarán como posibles enlaces.

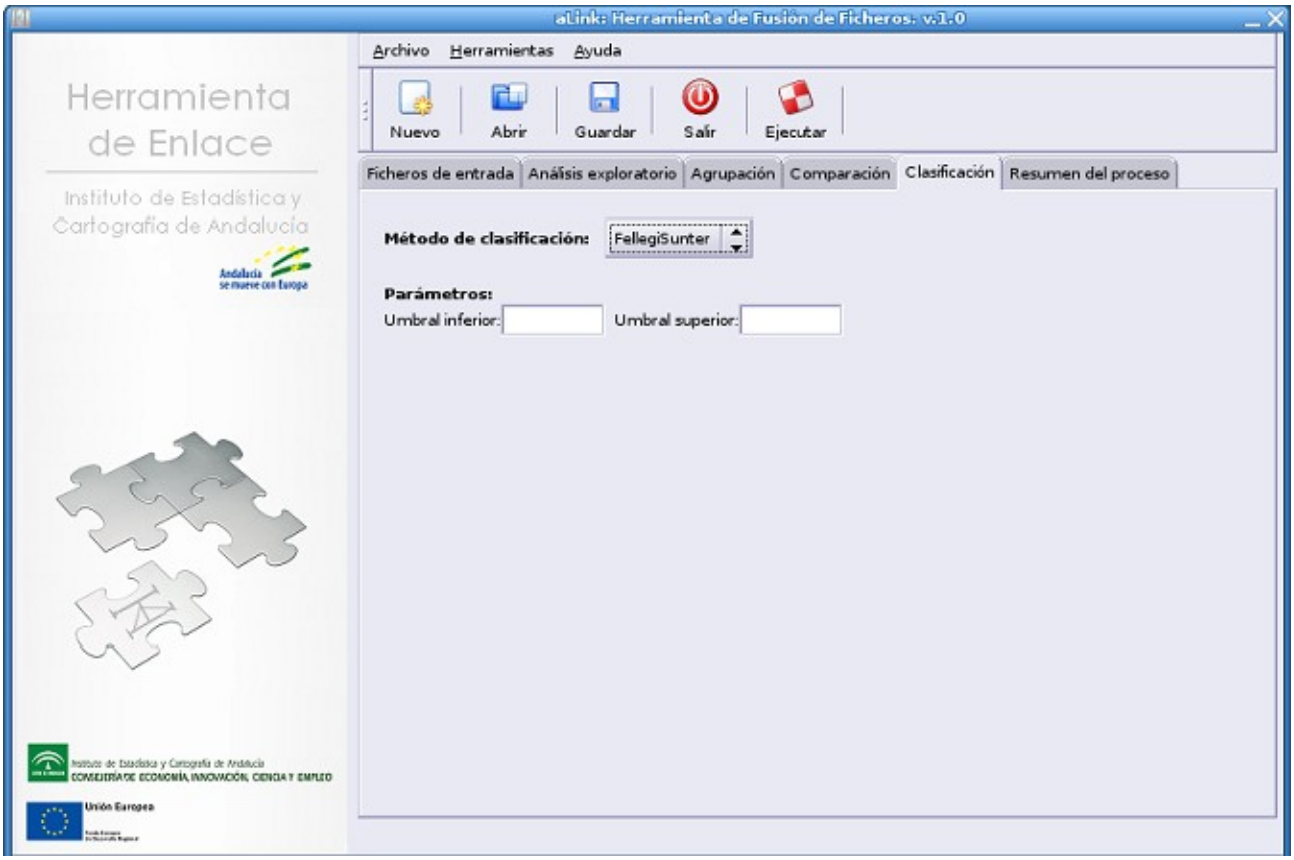


Imagen 116. Introducción de parámetros para el clasificador de Fellegi y Sunter

b) Si se elige TwoSteps, en un primer paso se seleccionarán automáticamente un conjunto de vectores de pesos cuyo número será determinado por el usuario (ver Anexo XIII para más detalle). Los vectores seleccionados serán aquellos que con una alta probabilidad darán lugar a verdaderos enlaces y a verdaderos no enlaces. Posteriormente, en un segundo paso se utilizarán dichos vectores para clasificar los pares de registros comparados mediante alguno de los métodos de clasificación supervisados implementados en la Herramienta de Enlace (máquina vector soporte y k-medias).

En la interfaz gráfica esta información queda recogida en los siguientes apartados:

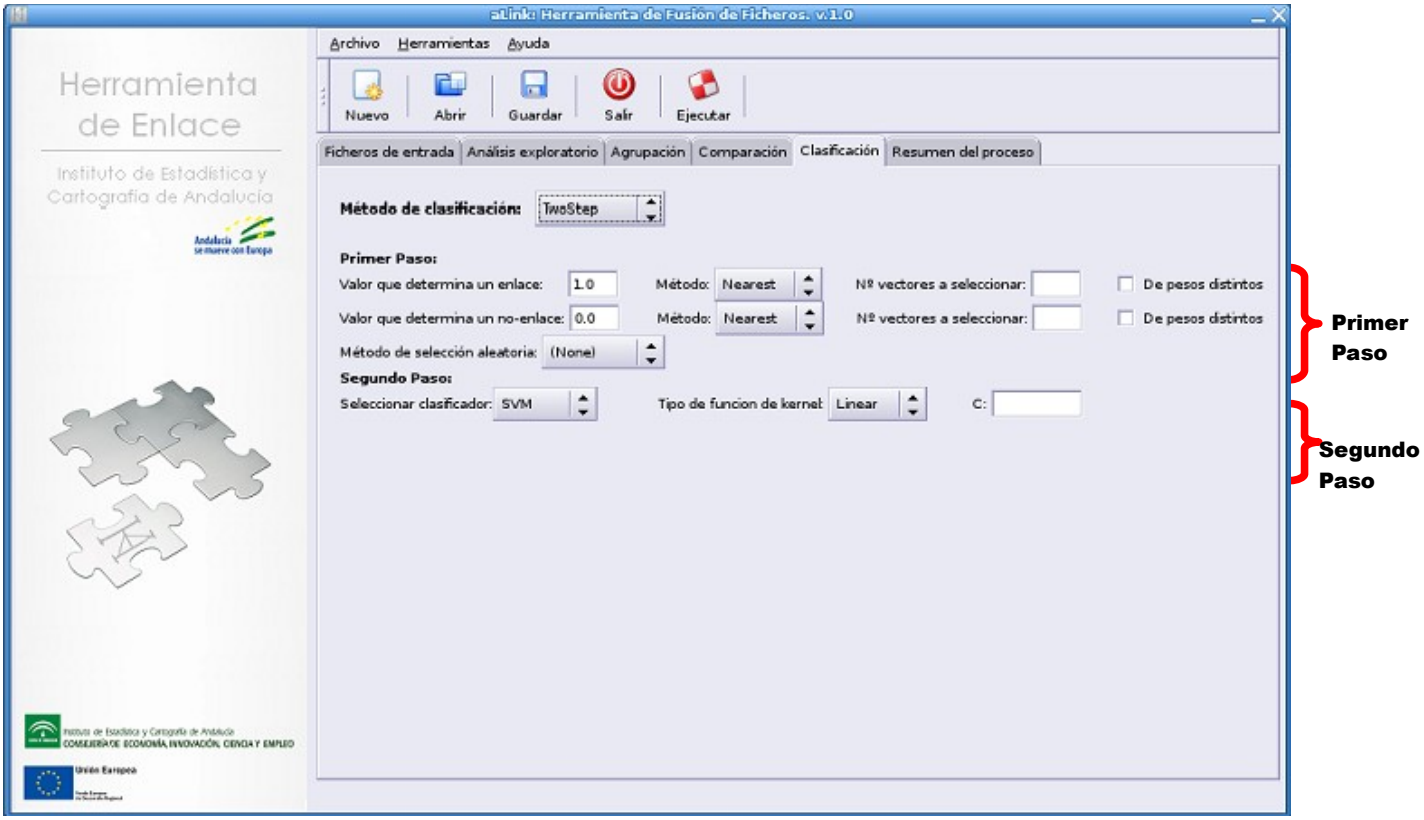


Imagen 117. Introducción de parámetros para el clasificador TwoStep

Primer Paso

Tanto para enlaces como no enlaces se determinan:

- **Valores que determinan un enlace o un no-enlace**: son los valores que el usuario ha establecido para definir a un verdadero enlace o un verdadero no enlace. Por defecto toman los valores 1.0 y 0.0.
- **Método de selección**: establece cómo se van a seleccionar los vectores de pesos para formar parte de los conjuntos de entrenamiento de enlaces o de no enlaces. El método puede ser Nearest o Threshold.

  Nearest: selecciona los vectores de pesos que van a formar parte del conjunto de entrenamiento de enlaces y del de no enlaces usando el método basado en los vecinos más cercanos. La idea de este método es seleccionar un cierto número de vectores de pesos, que estén lo suficientemente cerca de los vectores que representan a un verdadero enlace o a un verdadero no enlace e incluirlos en los respectivos conjuntos de entrenamiento de enlaces y no enlaces. Para ello se utilizará una distancia apropiada, como por ejemplo, la distancia euclídea o la distancia Manhattan. Si se selecciona este método, el usuario puede decidir además, si los vectores de pesos que formarán parte de los conjuntos de entrenamiento van a ser distintos o pueden repetirse. Se aconseja que estos sean distintos, con lo cual el usuario deberá marcar la opción 'De pesos distintos', ya que si se repiten podría darse el caso de llegar a una situación en la que todos los vectores de pesos seleccionados para construir el conjunto de entrenamiento de enlaces y el de no enlaces fuesen iguales al vector de pesos ideal para un verdadero enlace o un verdadero no enlace y esta situación no sería buena para entrenar el clasificador en el segundo paso.

  Threshold: selecciona los vectores de pesos que van a formar parte del conjunto de entrenamiento de enlaces y del de

no enlaces usando valores umbral. Para ello se calcula la diferencia en valor absoluto entre cada componente de los vectores de pesos y los valores que el usuario ha establecido para un verdadero enlace y un verdadero no enlace. Si dichos valores son mayores que los valor umbrales fijados respectivamente para el conjunto de enlaces y de no enlaces, entonces el par de registros asociado al vector de pesos no se considera bueno para el conjunto de entrenamiento de enlaces ni para el de no enlaces.

- **Número de vectores a seleccionar**: estos valores se establecen a libre elección del usuario pero este deberá tener en cuenta que en general es mucho más probable que tras el proceso de comparación de pares de registros el número de verdaderos enlaces sea mucho menor que el de verdaderos no enlaces y que por tanto, el número de vectores a seleccionar para el conjunto de entrenamiento de enlaces será mucho menor que el del conjunto de no enlaces. Christen propone una forma de calcular tales valores (ver Anexo XIII para más detalles).
- **Método de selección aleatoria**: permite la inclusión adicional, mediante selección aleatoria, de vectores de pesos para ser incluidos en el conjunto de entrenamiento de enlaces o en el de no enlaces. La opción por defecto es 'None' (Ninguno) pero en caso de que se decida incluir alguno, los vectores candidatos se elegirán de entre aquellos que no han sido asignados a ningún conjunto de entrenamiento. Los posibles métodos de selección aleatoria son: Uniforme, Lineal y Exponencial. Tras la elección del método, el usuario deberá indicar qué porcentaje de vectores de pesos se van a seleccionar aleatoriamente para incluir en los conjuntos de entrenamiento de enlaces y no enlaces. Por defecto el proceso está construido de forma que si se selecciona la opción 'Método de selección aleatoria' por lo menos se van a incluir un par de registros en ambos conjuntos, por lo que el usuario tendrá que tener en cuenta que la suma de ambos porcentajes no debe ser superior al 80%.

Segundo Paso:

Seleccionados los vectores comienza el proceso de entrenamiento y clasificación. Para ello se ha de completar la siguiente información:

- **Seleccionar clasificador**: el usuario puede elegir entre utilizar un clasificador de máquina vector soporte (*SVM*) o un clasificador de K-medias (*K-means*).

  Si elige el clasificador 'SVM' deberá especificar los siguientes parámetros requeridos:

  - **Tipo de función de kernel**: las funciones kernel o funciones núcleo son funciones matemáticas que se emplean en las Máquinas de Soporte Vectorial y que permiten convertir lo que sería un problema de clasificación no lineal en el espacio dimensional original, a un sencillo problema de clasificación lineal en un espacio dimensional mayor. En este caso el usuario podrá elegir entre las siguientes funciones núcleo: lineal (Linear), polinómicas (Poly), funciones de base radial (RBF) y sigmoidal (Sigmoid).
  - **Parámetro C**: se trata de un parámetro de penalización que se establece para contrarestar el efecto de vectores de pesos mal clasificados. Un valor grande de *C* equivale a una mayor penalización de los errores. Este parámetro lo establece el usuario pero según la información consultada del proyecto Febrl el valor 10 podría ser adecuado.

  En cambio, si elige el clasificador 'K-means', este funciona realizando un solo paso del algoritmo de *k_medias*, es decir, se va a calcular el centroide de los vectores de pesos del conjunto de entrenamiento de enlaces y el de no enlaces y a

continuación, se calcularán las distancias entre tales centroides y los vectores de pesos obtenidos en el proceso de comparación de registros. Para medir dicha distancia el usuario podrá elegir entre distintas medidas, como por ejemplo: euclídea, L-infinito, Caberra ó Manhattan.

7.1.3.7 Pestaña Salida

En esta ventana el usuario puede decidir, entre otras cosas, qué ficheros de resultados desea obtener:

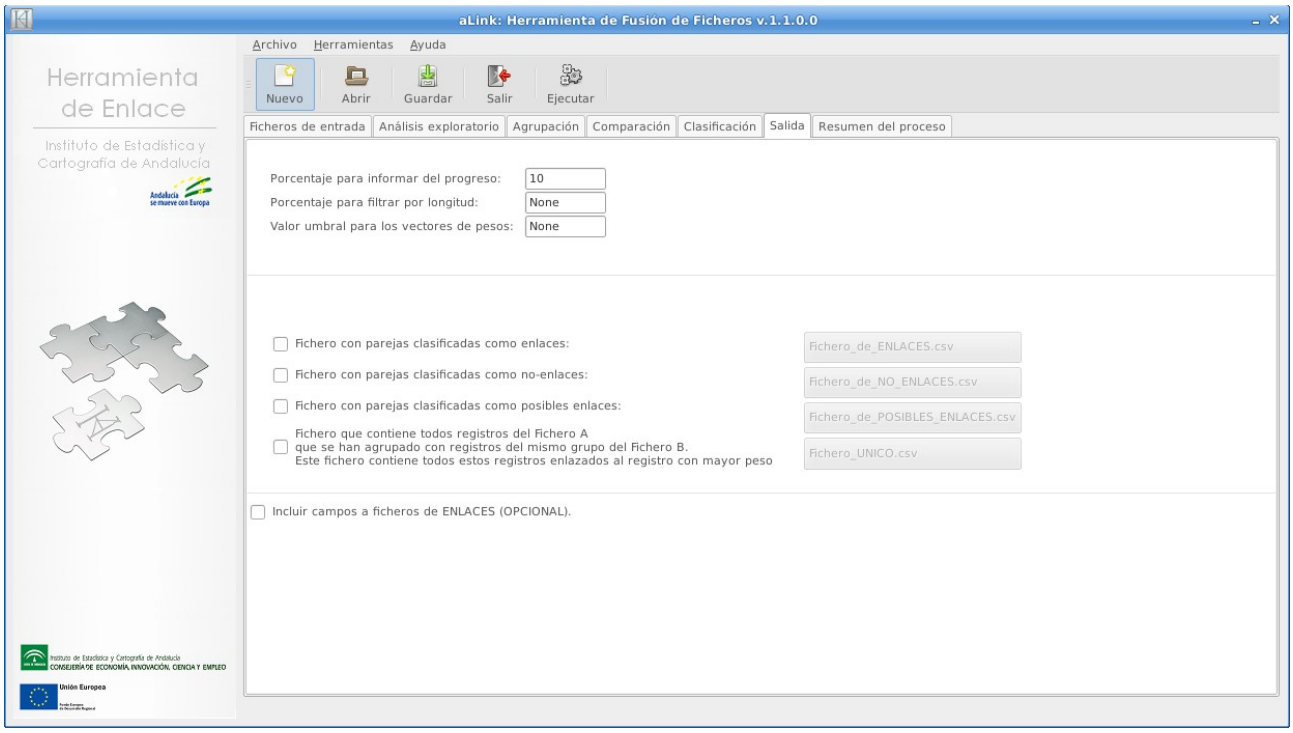


Imagen 118. Pestaña de Salida

La selección de ficheros de salida se realizará en la parte inferior de la pantalla. Estos son:

- **Fichero con parejas clasificadas como enlace**: contiene los pares de registros clasificados como enlaces, los valores y los pesos individualizados de cada campo comparado y el peso total.
- **Fichero con parejas clasificadas como no enlaces**: contiene los pares de registros clasificados como no enlaces, los valores y los pesos individualizados de cada campo comparado y el peso total.
- **Fichero con parejas clasificadas como posibles enlaces**: contiene los pares de registros clasificados como posibles enlaces, los valores y los pesos individualizados de cada campo comparado y el peso total.
- **Fichero que contiene todos los registros del Fichero A que se han agrupado con registros del mismo grupo del Fichero B:** Este fichero contiene todos estos registros enlazados al registro con mayor peso.

El usuario podrá elegir la ubicación de estos ficheros sin más que seleccionar el archivo correspondiente y pulsar sobre las pestañas que aparecen a la derecha. No obstante, se recomienda que se seleccionen los mismos y no se indique la ubicación. El motivo es que tras ejecutar un proceso de enlace se generará automáticamente una carpeta con la denominación *experimento-<AAAAMMDD-*

HHMM> con dichos ficheros. Esta carpeta se ubica en la misma ruta que el fichero que se ha considerado como Fichero A en el proceso de enlace.

Opcionalmente, se podrá activar la opción **Incluir campos a enlaces** para mejorar la legibilidad de los ficheros de salida. En este apartado, el usuario podrá elegir qué campos añadir del Fichero A y/o del Fichero B a los ficheros de salida *Fichero_de_ENLACES.csv* y *Fichero_de_POSIBLES_ENLACES.csv.* Esto, sin embargo, aumentará el tiempo de ejecución del programa.

Los campos por los que se unirán los ficheros serán "rec_id1" para el Fichero A y "rec_id2" para el Fichero B. De esta manera, sólo será posible activar esta opción si tanto el Fichero A como el Fichero B están indexados adecuadamente mediante la herramienta Insertar índices.

La estructura de estos ficheros es la que se observa en la siguiente imagen:

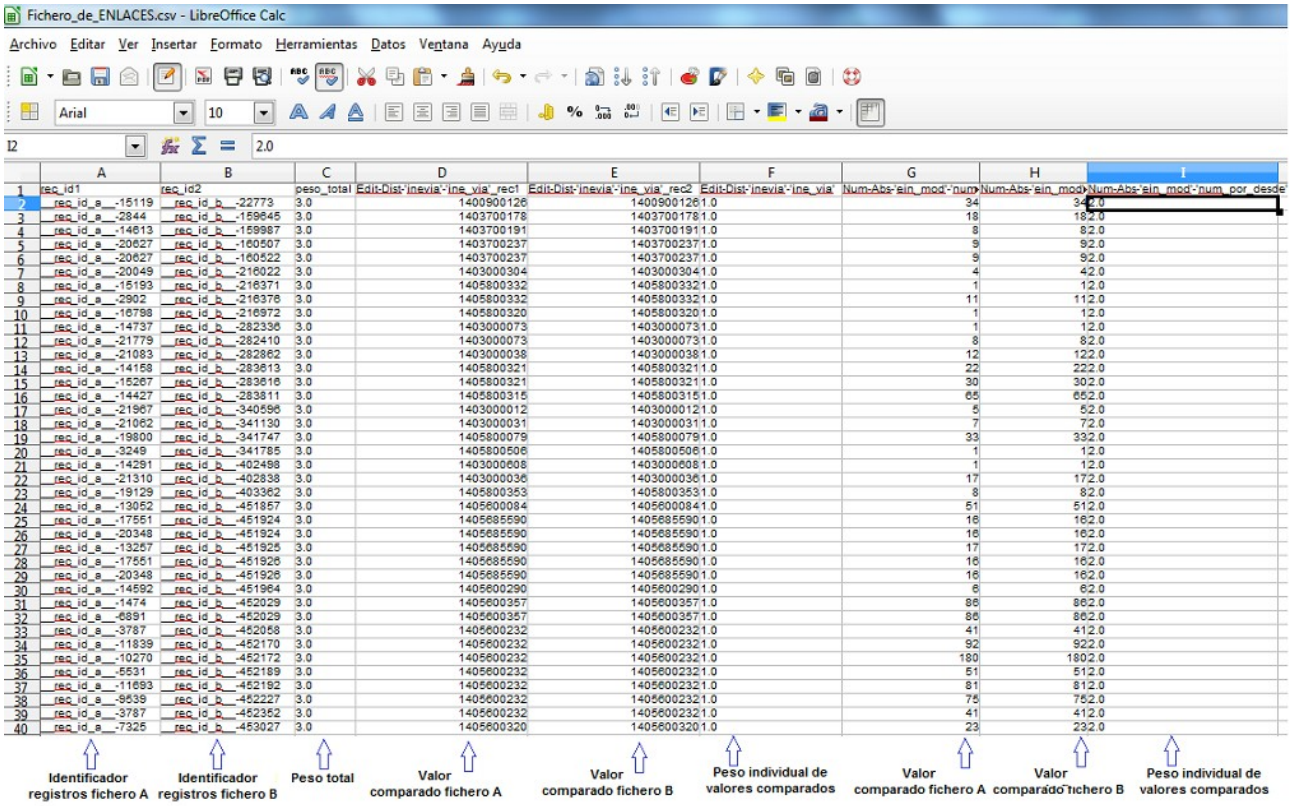


Imagen 119. Estructura del Fichero_de_Enlaces.csv

Además, en la parte superior de la ventana, el usuario puede establecer otra serie de parámetros, como por ejemplo:

- **Porcentaje para informar del progreso**: se trata del porcentaje de registros leídos a partir de los cuales se va a ir mostrando un mensaje de informe de progreso.
- **Porcentaje para filtrar por longitud**, que permite establecer un porcentaje comprendido entre 1 y 100 de manera que antes de realizar la comparación de pares de registros mediante las funciones de comparación elegidas se va a

comparar la longitud, en caracteres, de los valores de los campos. Si la diferencia porcentual en longitud es mayor que el porcentaje fijado para este parámetro, entonces los dos registros no serán comparados mediante las funciones de comparación y tendrán peso 0. La idea básica de este parámetro es comparar campos de registros que tengan longitudes parecidas ya que si estas son muy distintas probablemente no se van a referir a la misma entidad.

Veamos un ejemplo para comprobar cómo funciona. Para ello, supongamos que se tienen dos registros del tipo:

Nº registro	Registro	Longitud (en caracteres)
1	JOSE DEL CASTILLO FERNÁNDEZ	27
2	JOSE RUIZ GARCIA	16

Tabla 8. Ejemplo porcentaje filtrado por longitud

En este caso la diferencia porcentual vendría dada por:

$$diferencia\ \ porcentual = \frac{|27-16|}{\max(27,16)} = \frac{11}{27} = 0{,}4047 \approx 41$$

Así pues, si el parámetro para filtrar por longitud se ha fijado al 30% entonces como la diferencia porcentual obtenida (41%) es superior a este valor los dos registros no se van a comparar y su peso será 0.

Hay que indicar que esta opción no se aplica cuando en la fase de agrupación se elige la opción FullIndex.

- **Valor umbral de corte para los vectores de pesos**: este parámetro permite filtrar aquellos pares de registros cuyo peso total sea inferior al valor dado para este parámetro. Así pues todos los vectores de pesos que verifiquen esta propiedad serán considerados como no-enlaces y no se van a almacenar en memoria.

  Al igual que en el caso anterior, esta opción no se aplica si en la fase de agrupación se elige la opción FullIndex.

7.1.3.8 Pestaña Evaluación

Ofrece una evaluación de los resultados obtenidos en el proceso de enlace. Muestra un histograma con el número de pares de registros en función del peso total del par comparado, dando una idea acerca de dónde se concentran los pesos. **¡OJO!** Indicar que el valor representado por cada barra del histograma es aquel presente en su extremo inferior. Así, la observación del histograma puede ser orientativa para establecer los pesos a partir de los cuales los registros serán considerados como enlaces o no enlaces.

La situación teórica ideal estaría compuesta por dos secciones, una alrededor de un peso alto que correspondería a los pares de registros que representan a la misma entidad y otra sección alrededor de un peso bajo que correspondería a los pares de registros que representan a entidades diferentes.

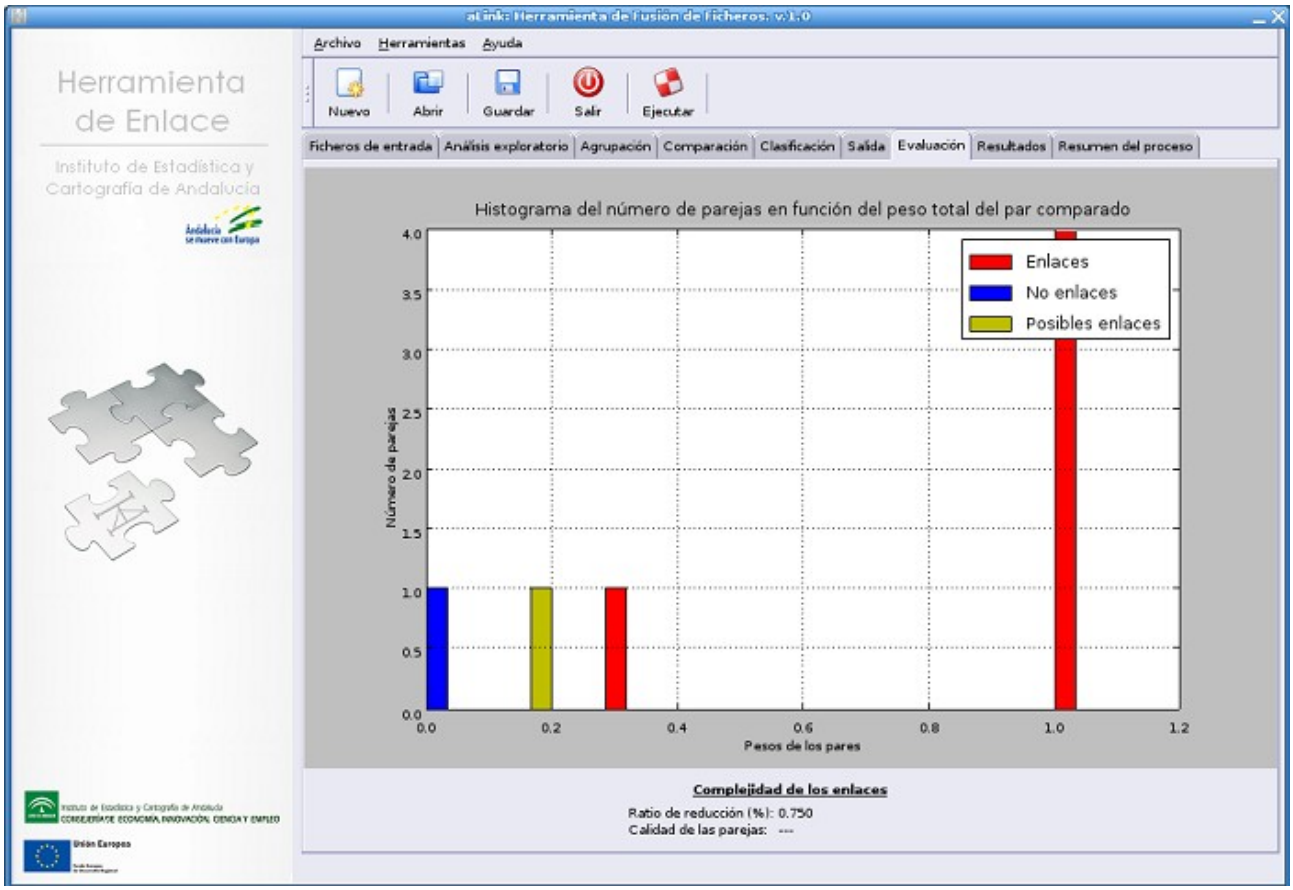


Imagen 120. Pestaña Evaluación

Finalmente aparece un indicador sobre la complejidad de los enlaces:

- Ratio de reducción (%): indica la disminución, en porcentaje, del número de pares de registros comparados al utilizar un método de agrupación en comparación a un proceso donde no se utiliza ninguno.

7.1.3.9 Pestaña Resultados

Contiene un diagrama de sectores en el que se indica la distribución de los pares de registros que se han enlazado, los que no y los que son posibles enlaces de entre todas las comparaciones que se han realizado. En esta pestaña se ofrece además, información sobre tiempos de ejecución del proceso de enlace, así como el número de comparaciones realizadas, el número de enlaces, no enlaces y posibles enlaces.

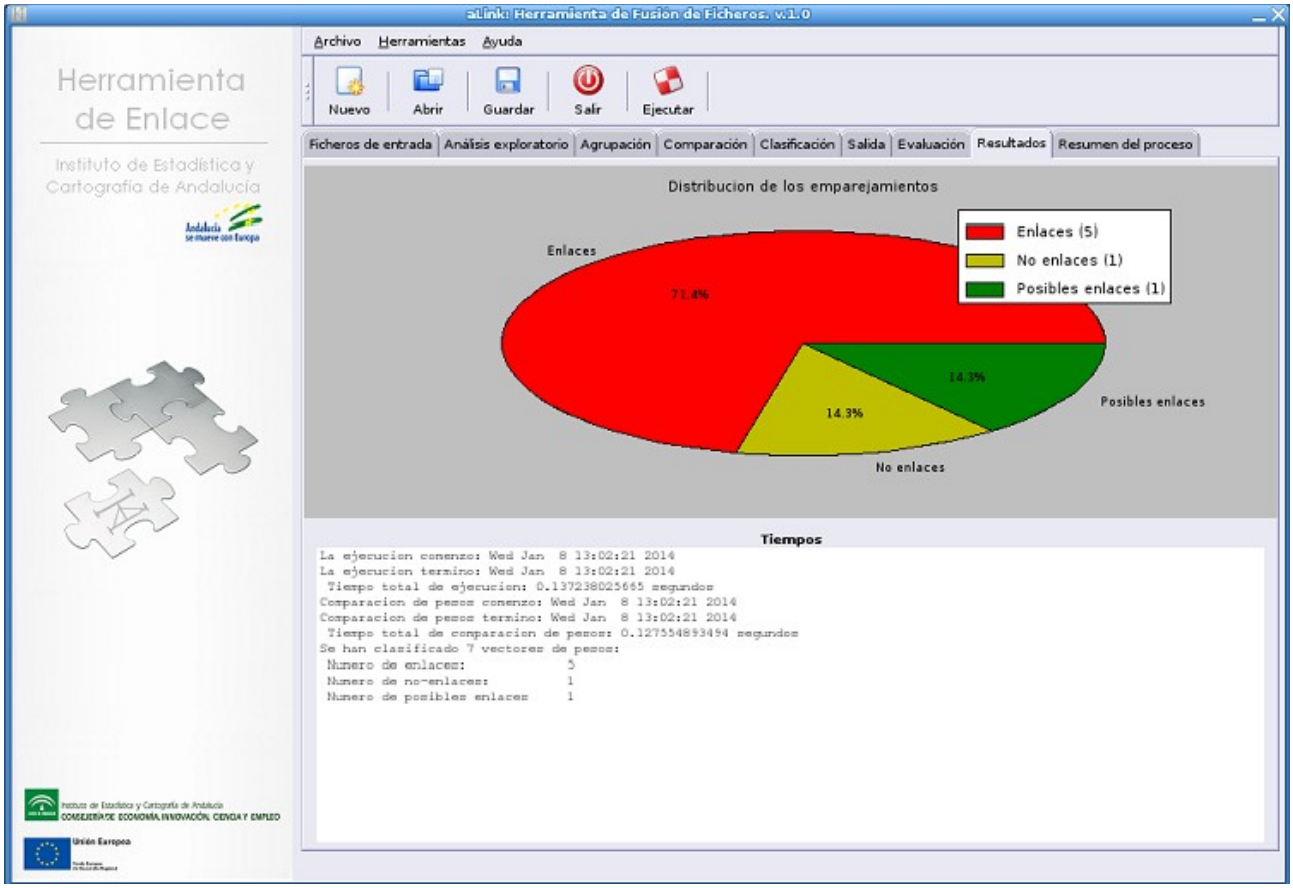


Imagen 121. Pestaña Resultados

7.2 Menú Herramientas de la Herramienta de Enlace

7.2.1 Tratamiento previo

Esta herramienta es exactamente igual que la descrita en la Herramienta de Normalización (ver apartado 6.3.1 de este Manual).

7.2.2 Insertar índices

Esta herramienta permite al usuario incluir en los ficheros que se van a enlazar un campo índice que va a identificar de forma unívoca a cada uno de los registros de los ficheros de datos que va a permitir hacer referencia a los mismos de manera rápida. Esta tarea es obligatoria antes de llevar a cabo un proceso de enlace.

A continuación, se presenta su interfaz así como los elementos de la misma:

Imagen 122. Interfaz de indexación

- **Fichero a indexar**: al pulsar este botón el usuario indicará la ubicación en la que se encuentra el fichero que quiere indexar.
- **Asignación de índice**: permite asignar el formato del índice que se va a incluir en cada fichero. Los valores son: *__rec_id_a__-* y *__rec_id_b__-*. En la práctica se asigna el índice *__rec_id_a__-* al fichero que se ha introducido como Fichero A en la pestaña de Ficheros de entrada y el índice *__rec_id_b__-* al Fichero B.
- **Salir**: este botón permite al usuario salir de la herramienta de idexación.
- **Ejecutar**: botón que permite al usuario realizar la indexación de cada uno de los ficheros a enlazar. Tras ejecutar el proceso se generan dos ficheros situados en la misma ubicación que los ficheros normalizados a enlazar. La denominación de los mismos sigue la estructura:

  nombre_del_fichero_normalizado.csv_indexado.csv

  Por ejemplo, si se hubiera indexado el fichero normalizado *Norm_empresas.csv*, la denominación del fichero indexado será: *Norm_empresas.csv_indexado.csv*.

7.2.3 Incluir campos a enlaces

Con la herramienta ***Incluir campos a enlaces*** el usuario puede incorporar al fichero con pares de registros clasificados como enlaces información de los ficheros que se están enlazando. La incorporación se realizará utilizando los índices que se definieron en el apartado anterior, motivo por el cual la tarea de indexación es obligatoria y bastante importante. La interfaz de esta herramienta es la que se muestra a continuación:

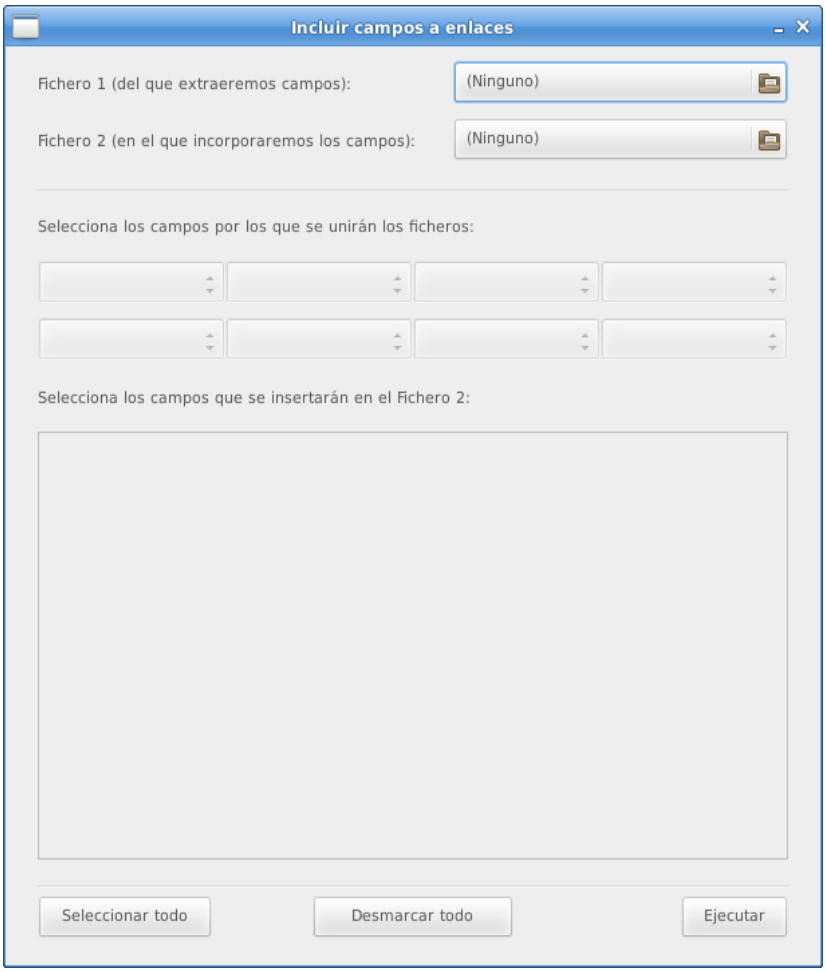


Imagen 123. Interfaz de incluir campos a enlaces

En ella el usuario podrá:

- Indicar la ruta en la que se encuentra el fichero del que se extraerá la información o campos para incluir en el fichero que contiene verdaderos enlaces (**Fichero 1 (del que extraeremos los campos)**).
- Indicar la ruta en la que se encuentra el fichero en el que se incorporará la información o campos, es decir, la ruta en la que se encuentra el fichero que contiene verdaderos enlaces (**Fichero 2 (en el que incorporaremos los campos)**).
- Seleccionar el campo o campos por los que se va a llevar a cabo la unión de los dos ficheros. Lo normal será utilizar los campos índice que se han creado en el apartado anterior junto a los campos equivalentes existentes en el fichero que contiene verdaderos enlaces. (**Selecciona los campos por los que se unirán los ficheros**).
- Seleccionar el campo o campos del fichero del que se extraerá la información (Fichero 1) para incluirlos en el fichero que

contiene verdaderos enlaces (Fichero 2) (**Selecciona los campos que se insertarán en el Fichero 2**).

- **Seleccionar todo**: al pulsar este botón se seleccionarán todos los campos.
- **Desmarcar todo**: al pulsar este botón se desmarcarán todos los campos.
- **Ejecutar**: al pulsar este botón el usuario ejecutará el proceso de incluir información al fichero que contiene verdaderos enlaces. El resultado es un fichero de extensión .csv, cuya denominación sigue el formato:

denominación_del_Fichero2salida.csv

A continuación, se muestra el contenido de la interfaz usando dos ficheros:

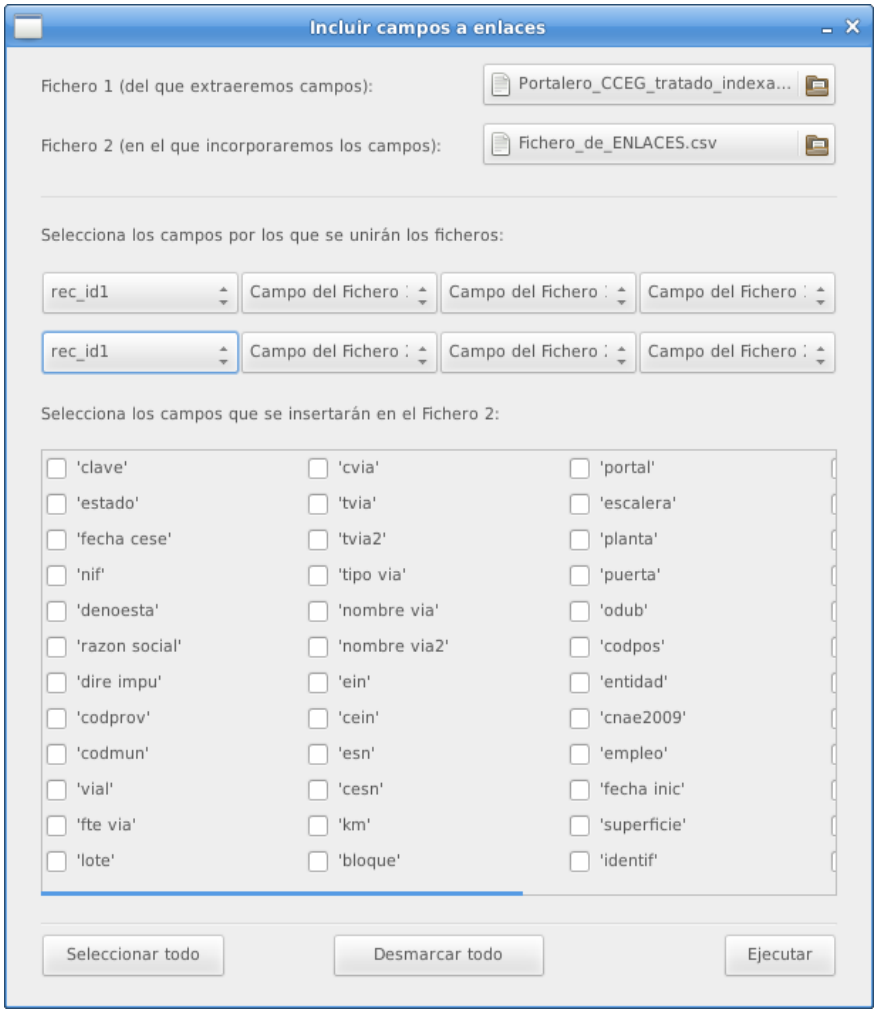


Imagen 124. Ejemplo interfaz incluir campos a enlaces

7.2.4 Eliminar registros enlazados

Una vez realizado un proceso de enlace concreto, lo normal es que queden registros sin enlazar. Por tanto, para llevar a cabo un nuevo proceso de enlace, el usuario deberá eliminar del fichero de menor tamaño aquellos registros que han enlazado previamente. Para ello utilizará la herramienta ***Eliminar registros enlazados***, cuya interfaz se muestra a continuación:

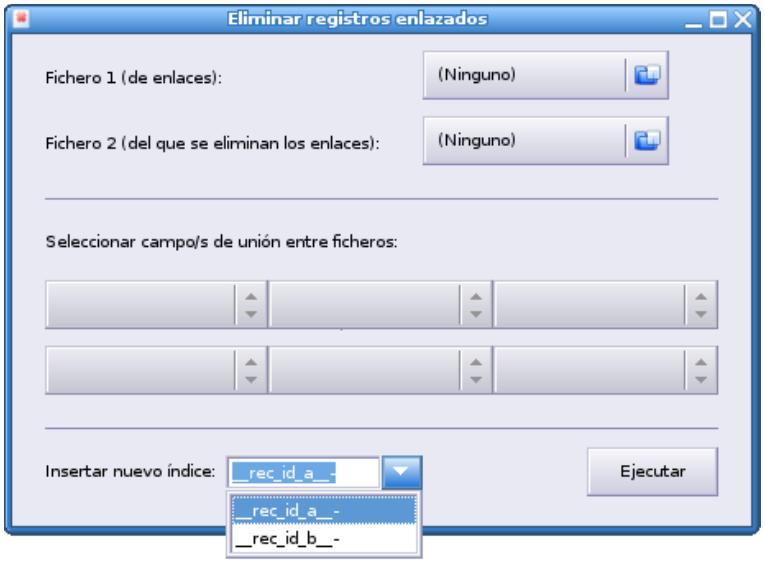


Imagen 125. Interfaz eliminar registros enlazados

En ella el usuario podrá:

- Indicar la ruta en la que se encuentra el fichero que contiene los pares de registros considerados como enlaces (**Fichero1 (de enlaces)**).
- Especificar la ruta en la que se encuentra el fichero del que se eliminarán los pares de registros enlazados (**Fichero 2 (del que se eliminan los enlaces)**).
- Seleccionar el campo o campos por los que se van a unir los dos ficheros anteriores. Lo normal será utilizar los campos índice que se han creado en el apartado de indexación junto a los campos equivalentes existentes en el fichero que contiene verdaderos enlaces. (**Selecciona campo/s de unión entre ficheros**).
- Insertar un nuevo índice al fichero que se generará al ejecutar el proceso de eliminar los pares de registros enlazados (**Insertar nuevo índice**). En la imagen anterior se pueden ver los valores del mismo (*__rec_id_a__-* y *__rec_id_b__-*). Si el fichero ya estaba indexado, se guardará esa información, añadiéndole en la cabecera la fecha y hora del proceso de actual. Este antiguo índice, sin embargo, no se usará posteriormente para futuras ejecuciones del programa.

Así, una vez completada la información, al pulsar el botón **Ejecutar** se generará un nuevo fichero con extensión .csv, que se guardará en la misma ubicación que el fichero del que se han eliminado los pares de registros enlazados. La denominación del mismo sigue el siguiente formato: *denominación_del_Fichero2salida.csv*

Un ejemplo de cómo quedaría configurada la interfaz de eliminar registros enlazados se muestra a continuación:

Imagen 126. Ejemplo interfaz eliminar registros enlazados

8 FAQ

1. **¿De entre los sistemas operativos en los que se puede instalar *aLink: Herramienta de Fusión de Ficheros* con cuál de ellos se obtiene un mayor rendimiento computacional a la hora de realizar un proceso de fusión de ficheros?**

   Si se tienen dos equipos con idéntico hardware pero con distinto sistema operativo, se recomienda optar siempre por el equipo que tenga un sistema operativo Linux ya que además de ser un sistema operativo libre ofrece muchas ventajas respecto a la hora de trabajar con grandes volúmenes de datos debido a que en este tipo de sistemas no hay tantas restricciones de uso de memoria como se tendrían en Windows.

2. **He instalado en un entorno Windows *aLink: Herramienta de Fusión de Ficheros* pero no consigo abrirla, ¿a qué puede ser debido?**

   Esto puede deberse a dos motivos:

   – A la existencia en tu equipo de alguna versión de Python distinta a la 2.7.2

   – A que ya se tiene instalada la versión de Python 2.7.2 en el equipo pero no ha sido seleccionada durante el proceso de instalación.

   En ambos casos se recomienda desinstalar todas las versiones de Python, así como los programas asociados a *aLink: Herramienta de Fusión de Ficheros* usando la opción desinstalar programas del Panel de Control de Windows y volver a realizar la instalación.

3. **¿Se pueden normalizar ficheros de gran tamaño sin tener que segmentarlos en varias partes?**

   Sí, para ello se utilizará el fichero de proyecto que se genera al normalizar un fichero de datos. En concreto, los pasos a seguir son los siguientes:

   1. Extraer una muestra del fichero de datos que se va a normalizar, fichero que ha tenido que ser tratado previamente. La muestra por tanto va a tener formato .csv con elementos separados por ";".

      Para seleccionar la muestra se puede usar algún editor de texto como Notepad2 o Gedit u algún programa de hoja de cálculo como LibreOffice Calc. Nótese que al intentar abrir el fichero de gran tamaño con alguno de estos programas es posible que no lo haga completamente. Esto no es un problema ya que solamente se necesita seleccionar una muestra del mismo. En cuanto a su tamaño, con un único registro bastaría, ya que lo que se necesita conocer es la estructura del fichero a normalizar, es decir, el número de campos que componen su cabecera, la denominación de los mismos, etc.

   2. Utilizar la Herramienta de Normalización para normalizar la muestra seleccionada.

   3. Tras la normalización se generan varios ficheros, entre ellos, uno con extensión .py (fichero de proyecto) que será el que se va a modificar para poder normalizar el fichero completo. El fichero de proyecto se abrirá con algún editor de texto y bastará con realizar dos sustituciones:

- Primero se sustituirá el nombre del fichero de entrada por el nombre del fichero completo a normalizar (en la imagen de abajo se indica el lugar del fichero de proyecto en el que se tiene que realizar el cambio). Nótese que el fichero completo a normalizar deberá encontrarse en la misma ruta que el fichero de muestra para que no surjan problemas de no encontrar al fichero. Si no es así, habrá que modificar además la ruta del fichero de entrada. **¡Ojo!** Dependiendo de si se trabaja en un entorno Windows o en un entorno Linux la forma de indicar los directorios es distinta. Por ejemplo, si se trabaja en Windows las rutas se indicarán como sigue: "C:"+os.sep+"ficheros"+os.sep+"establecimientos.csv", mientras que si se trabaja en Linux se indicarán como: "home/ecaballero/ficheros/establecimientos.csv".

```
31  # ----------------------------------------------------------------------
32  # Tipo de proyecto: Normalización
33  # ----------------------------------------------------------------------
34
35  # ----------------------------------------------------------------------
36
37  # Definición del juego de datos de entrada A
38  #
39  data_set_a = dataset.DataSetCSV(description="Datos de entrada",
40                                  access_mode="read",
41                                  strip_fields=True,
42                                  miss_val=[],
43                                  rec_ident="__rec_id_a__",
44                                  file_name="D:"+os.sep+"Trabajo_IECA"+os.sep+"Fusion_ficheros"+os.sep+"Ficheros pruebas"+os.sep+"establecimientos_tratado.utf-8.limpio.csv",
45                                  header_line=True,
46                                  delimiter=";",
47                                  field_list = [("'direccion'",0))
48
49
50  # ----------------------------------------------------------------------
```

Imagen 127. Ruta donde se encuentra el fichero a normalizar

- A continuación, se sustituirá el nombre del fichero de salida (en la imagen de abajo se indica el lugar del fichero de proyecto en el que se tiene que realizar el cambio). Si no se cambia la ruta en donde se quiere guardar el fichero completo normalizado, se guardará en el mismo lugar donde se guardó la muestra normalizada.

```
90  # ----------------------------------------------------------------------
91
92  # Definición de los datos de salida normalizados y del registro de normalización
93  #
94  data_set_a_std = dataset.DataSetCSV(description="Datos de salida normalizados",
95                                      access_mode="write",
96                                      delimiter=";",
97                                      file_name="D:"+os.sep+"Trabajo_IECA"+os.sep+"Fusion_ficheros"+os.sep+"Ficheros pruebas"+os.sep+"NORM_20140114-2108_establecimientos_tratado.csv",
98                                      field_list = [("'direccion'",0),
99                                                    ("tipo_de_via",1),
100                                                   ("nombre_de_via",2),
101                                                   ("identificador_de_numeracion",3),
102                                                   ("ein",4),
103                                                   ("cein",5),
104                                                   ("esn",6),
105                                                   ("cesn",7),
106                                                   ("bloque",8),
107                                                   ("portal",9),
108                                                   ("escalera",10),
109                                                   ("planta",11),
110                                                   ("puerta",12),
111                                                   ("entidad_singular",13),
112                                                   ("municipio",14),
113                                                   ("provincia",15),
114                                                   ("codigo_postal",16),
115                                                   ("tipo_de_agrupacion",17),
116                                                   ("agrupacion",18),
117                                                   ("odub",19),
118                                                   ("validacion",20)],
119                                      rec_ident="__rec_id",
120                                      strip_fields=True,
121                                      header_line=True,
122                                      write_header=True)
123
```

Imagen 128. Ruta donde se desea guardar el fichero normalizado

4. Copiar el fichero de proyecto .py en la carpeta donde se encuentra el código de *aLink: Herramienta de Fusión de Ficheros*, en concreto dentro de la carpeta 'app'. Una vez ahí, se procederá de la siguiente forma:

- Si se trabaja en un entorno Linux habrá que acceder al directorio app en el que se ha copiado el fichero y

escribir la orden #python *proyecto*.py.

- Si se trabaja en un entorno Windows se hará doble click sobre el o se abrirá el mismo con algún programa como IDLE o Geany y se ejecutará. Por ejemplo, si se utiliza IDLE los pasos a seguir son:

  1. Ir al botón de Inicio de Windows.
  2. Seleccionar de entre todos los programas, el denominado Python(x,y). Dentro del mismo, seleccionar la opción IDLE, tal y como se observa en la siguiente imagen:

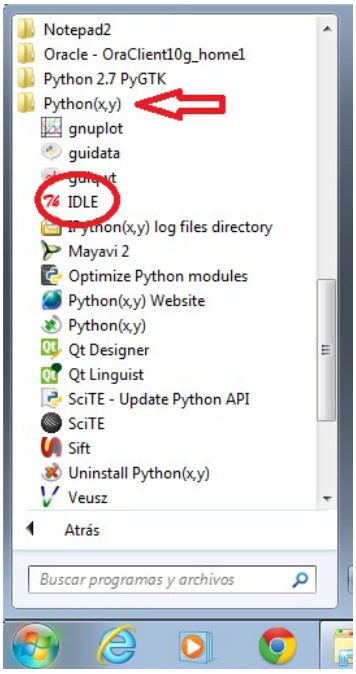


Imagen 129. Selección de IDLE a través del botón de inicio de Windows

  3. Una vez abierto, el usuario deberá ir a la carpeta donde se encuentra el código de la aplicación (C:\alink\app) y donde debe haber copiado el fichero de proyecto que va a usar para la normalización del fichero completo. Para ello deberá elegir la opción **Open** del menú **File**, de tal forma que llegará a algo similar a lo que se observa en la siguiente imagen:

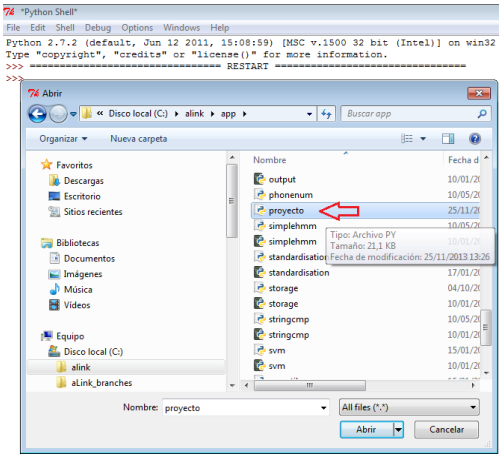


Imagen 130. Abrir proyecto en IDLE

4. Tras pulsar el botón **Abrir**, se abrirá el fichero de proyecto. Para ejecutarlo habría que elegir del menú **Run** la opción **Run Module** o bien pulsar **F5** tal y como se muestra en la siguiente imagen:

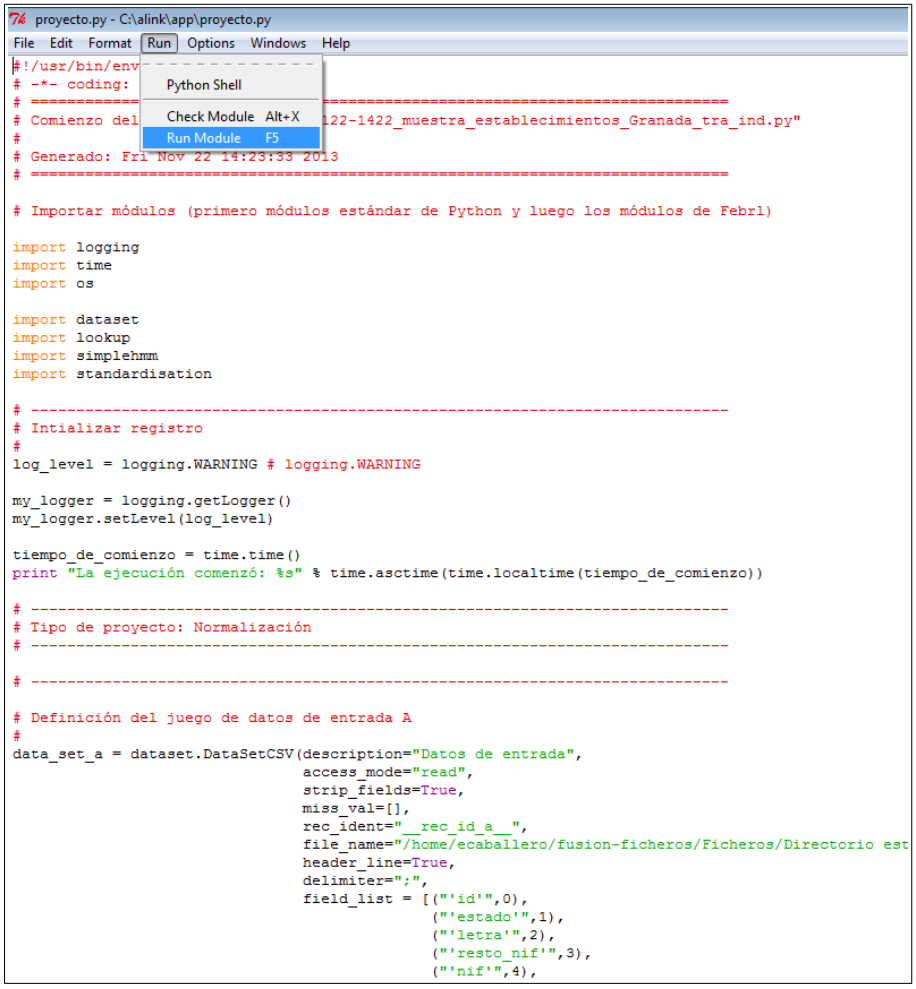


Imagen 131. Ejecutar proyecto en IDLE

Si por el contrario se decidiera trabajar con Geany, habría que realizar los siguientes pasos:

1. Abrir el fichero de proyecto con el que se va a llevar a cabo la normalización del fichero completo, que debe encontrarse en la carpeta C:\\alink\app. Para ello se pulsará la opción Abrir.
2. A continuación, se comprobará que la ruta que permite ejecutar los comandos de Python está bien establecida. Dado que Python 2.7.2 va a estar instalado en C:\\Python27, habría que pulsar la pestaña del menú **Construir** tal y como se observa en la imagen de abajo. En la ventana que le aparece al usuario debería mostrarse en la sección de **Ejecutar comandos** la información que está recuadrada en rojo (C://Python27/python "%f"). Si no es así, el usuario deberá cambiarla por esta.

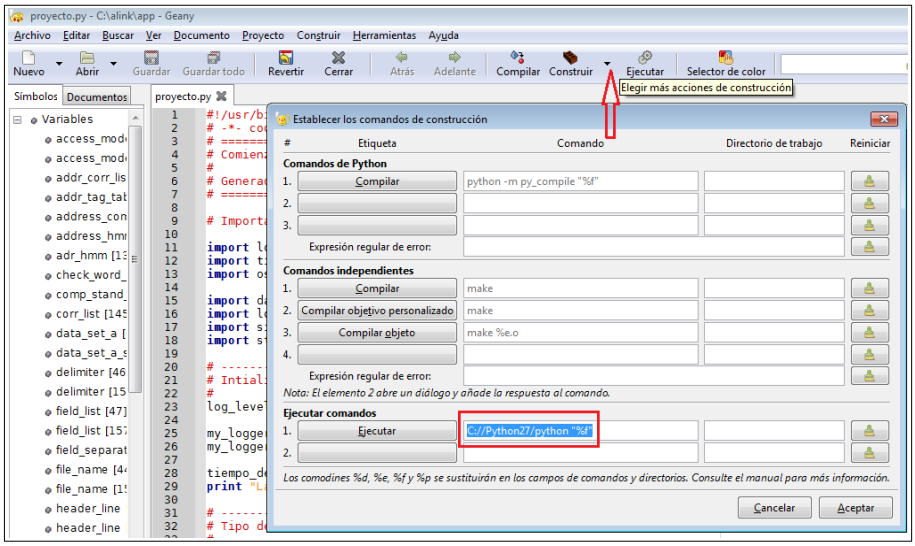


Imagen 132. Comprobación de ruta de ejecución de comandos en Geany

3. Tras esta comprobación se procederá a ejecutar el fichero de proyecto pulsando el botón **Ejecutar**.

4. ¿Qué tratamiento se realiza al carácter ñ?

Con la herramienta de tratamiento previo el carácter ñ pasa a ser sustituido automáticamente por los caracteres kk.

5. Estoy tratando un fichero de ACCESS y no funciona correctamente la herramienta de tratamiento previo, ¿a qué puede deberse?

Esto puede ser debido a que los campos que componen la tabla de ACCESS no son todos de tipo texto. Una vez modificados todos los campos a este formato se recomienda compactar la base de datos.

6. Al normalizar un fichero de datos obtengo algunos valores unidos con el carácter guión bajo, ¿a qué es debido?

Esto se debe a que la aplicación une mediante este carácter todos aquellos elementos que se encuentren en alguna tabla de búsqueda y que estén formados por valores compuestos. Por ejemplo, si se está normalizando un fichero con direcciones postales y la aplicación se encuentra con el valor *san antolin*, este va a ser sustituido por *san_antolin* por haber sido localizado en la tabla de búsqueda de entidades singulares y por ser un elemento compuesto por el valor *san* y *antolin*. En cambio, si se localiza el elemento *san agustín*, dado que este no está dentro de ninguna tabla de búsqueda aparecerá como *san agustin*. En estos casos, si el usuario lo considera conveniente podría eliminar tal carácter usando una función de buscar- reemplazar.

7. ¿Qué puedo hacer para que el identificador de numeración S/N, sin/número, etc. sea considerado como un número y tome un valor concreto?

En este caso la manera de proceder sería la siguiente: abrir el editor de tablas de búsqueda y abrir la tabla de búsqueda correspondiente a *knumero_local.tbl*. A continuación, se editaría el elemento *sin numero* y este podría ser el resultado suponiendo que los valores s/n, sin numero, s.n, etc. se quisieran sustituir por el número 0:

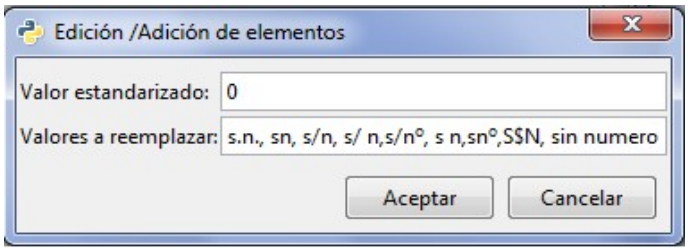


Imagen 133. Edición del elemento *sin numero*

A continuación, habría que entrenar una muestra que contuviera direcciones con la siguiente estructura o patrón:

```
#C/ SOL S/N
TV:tipo_de_via, UN:nombre_de_via, NM:ein
#AVD MIRAFLORES Nº 3
TV:tipo_de_via, UN:nombre_de_via, NM:identificador_de_numeracion, NU:ein
 #AVD CADIZ SIN NUMERO
TV:tipo_de_via, MU:nombre_de_via, NM:ein
#TV:tipo_de_via, PR:nombre_de_via, NM:ein
...
```

Una vez entrenada la muestra y generado el Modelo Oculto de Markov, el usuario podrá aplicarlo para normalizar el fichero completo. De esta forma las anteriores direcciones deberían aparecer segmentadas de la siguiente manera:

tipo_ de_via	nombre_de_via	identificador_de_numeracion	ein
calle	sol		0
avenida	miraflores	numero	3
avenida	cadiz		0
...	...	...	...

Tabla 9. Ejemplo de direcciones postales segmentadas

8. ¿Se puede eliminar una tabla de búsqueda porque su contenido no sea válido para un proceso de normalización?

No. En este caso la forma de proceder sería la siguiente: abrir el editor de tablas de búsqueda, abrir la tabla de búsqueda cuyo contenido no sea válido para el proceso de normalización y suprimir todos sus elementos. A continuación, sería necesario guardar los cambios para que estos fueran efectivos.

9. ¿Es posible que al enriquecer una muestra con nuevas estructuras o patrones y generar el Modelo Oculto de Markov correspondiente, se obtenga un peor resultado que en un proceso de normalización anterior?

Sí. Esto se debe a que al introducir nuevas estructuras en la muestra y generar el modelo HMM, las probabilidades de las estructuras ya existentes en la misma se han visto modificadas, con lo cual registros que estaban bien normalizados en procesos de normalización anteriores ahora no lo están.

En esta situación se aconseja no enriquecer la muestra sino crear un fichero de datos que contenga aquellos registros que no se han normalizado correctamente y a partir del mismo extraer una muestra de registros del campo a normalizar para generar un nuevo modelo HMM.

10. La aplicación no es capaz de detectar un fichero en la ruta en la que el usuario le indique.

Si el usuario especifica la ruta de un fichero pero la aplicación no lo detecta y aparece en mensaje de error similar al de la imagen de abajo puede ser debido a que la ruta en la que se ubica sea muy larga. Para solucionarlo modificar la ubicación del fichero.

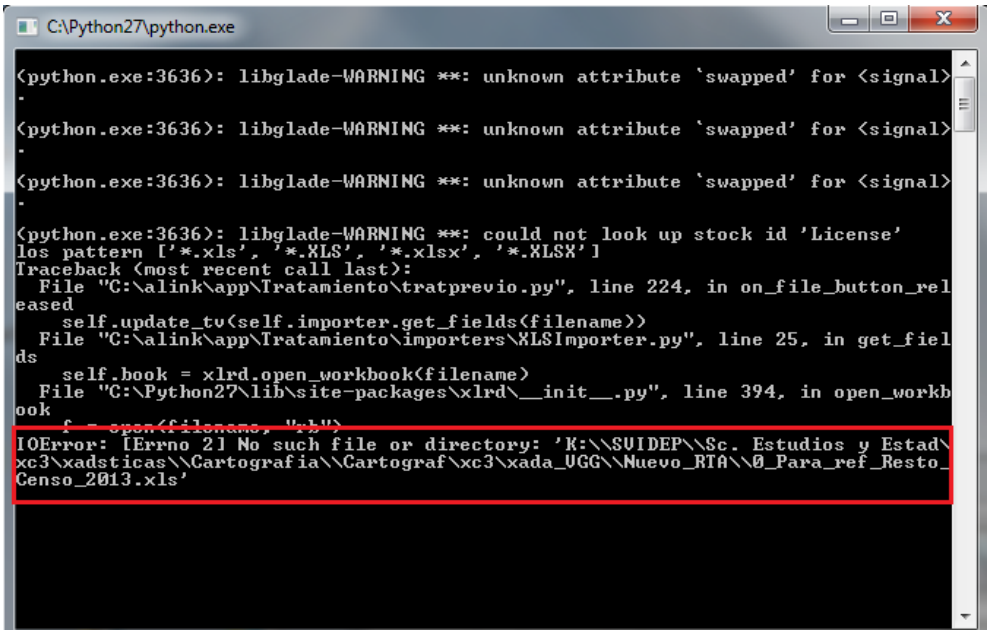


Imagen 134. No detección de fichero o directorio

9 ANEXOS

Anexo I: Instalación manual de *aLink: Herramienta de Fusión de Ficheros* en un entorno Windows

En caso de que no haya sido posible la instalación de aLink mediante el fichero de instalación proporcionado en la Sección de Descarga de Software de la página web del Instituto de Estadística y Cartografía de Andalucía, el usuario podrá instalar manualmente la aplicación. Para ello deberá proceder de la siguiente manera:

1. Solicitar al Instituto de Estadística y Cartografía de Andalucía los siguientes ficheros de sistema de Windows que son necesarios para el proceso de instalación:

- MSVCP71.DLL
- msvcr71.dll

y copiarlos en la siguiente ubicación:

- Para sistemas a 32 bits, deberá copiar los anteriores archivos en la siguiente ubicación: C:\Windows\System32.
- Para sistemas a 64 bits, deberá copiarlos en las dos ubicaciones siguientes: C:\Windows\System32 y C:\Windows\SysWOW64.

2. Solicitar al Instituto de Estadística y Cartografía de Andalucía los siguientes programas que se indican a continuación e instalarlos uno a uno con los parámetros que se especifican más abajo:

- Python-xy-2.7.2.3.exe
- Pygtk-all-in-one-2.24.1.win32-py2.7.msi
- Matplotlib-1.1.0.win32-py2.7.exe
- MySQL-python-1.2.3.win32-py2.7.exe
- Gawk-3.1.6-1-setup.exe
- Coreutils-5.3.0.exe
- Xlrd-0.9.2-3_py27.exe
- Pyodbc-3.0.7.win32-py2.7.exe
- Dbfpy-2.2.5.win32.exe
- Notepad2_x64.exe ó Notepad2_x32 en función de si el equipo es de 64 bits o de 32 bits.

3. También es necesario solicitar al Instituto de Estadística y Cartografía de Andalucía la siguiente librería y paquetes de Python:

- libsvm.dll
- svm.py y svmutil.py

Python XY

Al instalar este programa, el usuario desplegará en la ventana que aparece la **rama Python** para configurar los módulos que se quieren instalar con esta distribución.

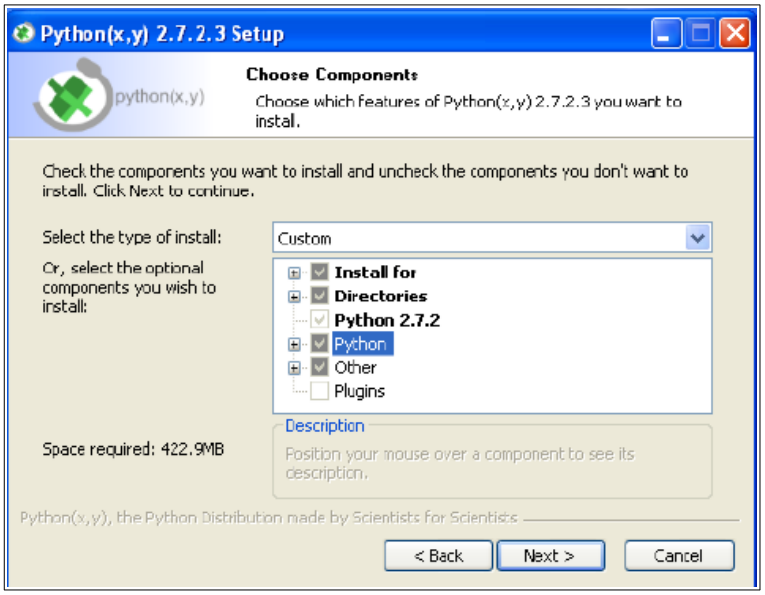


Imagen 135. Ventana inicial para la instalación de Python(x,y) 2.7.2.3

Una vez desplegada la rama, debe recorrer la lista de módulos **asegurando DESMARCAR** los siguientes:

- guiqwt 2.1.6.3
- xy 1.2.14.3
- PyOpenGL 3.0.2a6
- ETS 4.1.0.2
- gnuplot 1.8.0.3

A continuación, el usuario recorrerá de nuevo la lista de módulos y manteniendo los ya marcados, **SELECCIONARÁ** el siguiente:

- Cython 0.16

Una vez seleccionados los módulos de la rama Python, se pliega esta rama y se despliega la **rama Other.** Aquí solamente se marcan los siguientes programas:

- Console 2.0.148.5
- MinGW 4.5.2.2

A continuación, solo es necesario pulsar **Next** e **Install** hasta que finalice el proceso de instalación.

Seguidamente, se procede a la instalación de PyGTK para Windows.

PYGTK ALL-IN-ONE

Esta versión incluye todo lo necesario para instalar los bindings de la biblioteca gráfica GTK para Python 2.7 en Windows. Su instalación sólo requerirá pulsar el botón **Next** hasta que finalice el proceso. No es necesario tocar ningún ítem de configuración.

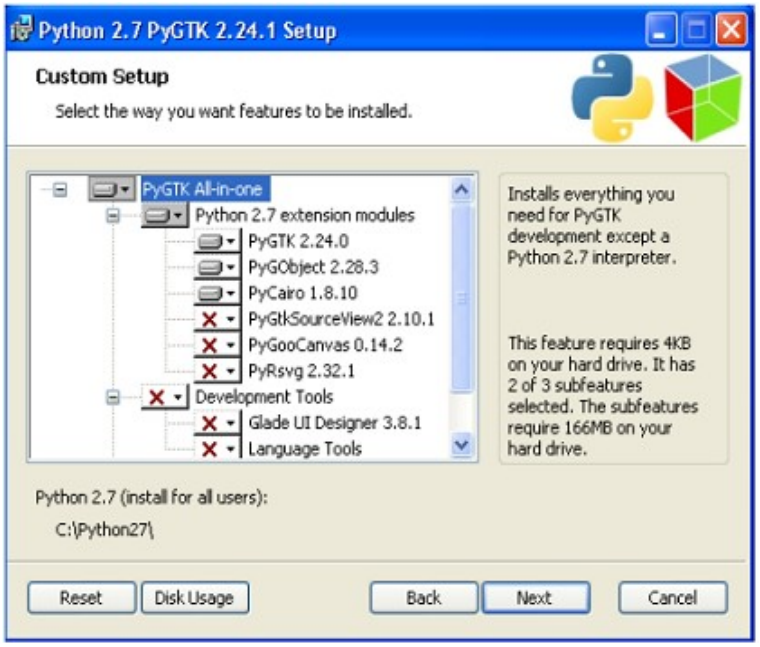


Imagen 136. Ventana inicial para la instalación de PyGTK 2.24.1

El siguiente elemento a instalar es matplotlib.

MATPLOTLIB

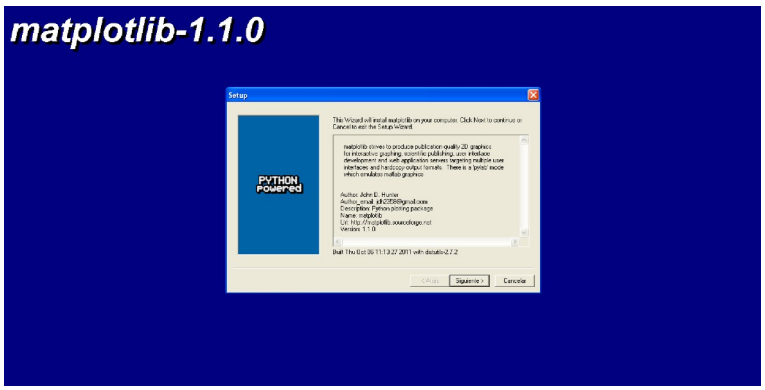


Imagen 137. Ventana inicial para la instalación de matplotlib-1.1.0

Se trata de un módulo de generación de gráficos para Python. La instalación es igualmente sencilla y sólo es necesario pulsar el botón **Siguiente** hasta que el proceso finalice.

MYSQL-PYTHON

A continuación, se instala este módulo de gestión de bases de datos MySQL para Python. Como en el caso anterior, la instalación es igualmente sencilla y sólo es necesario pulsar el botón **Siguiente** hasta que el proceso finalice.

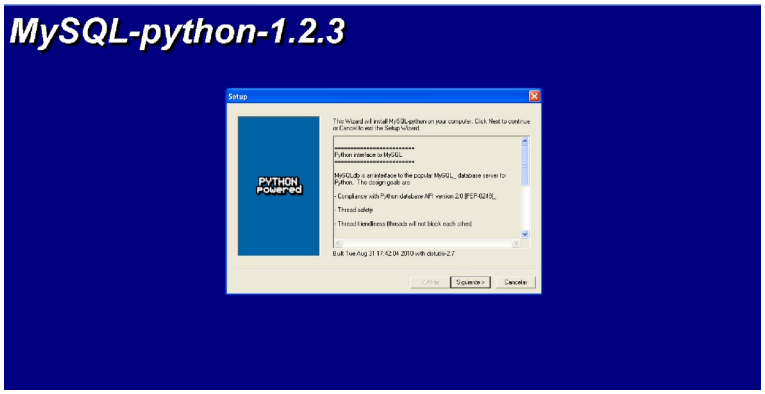


Imagen 138. Ventana inicial para la instalación de MySQL-python-1.2.3

GAWK

El siguiente programa a instalar es *Gawk*, la implementación GNU del lenguaje de programación AWK. Su instalación consistirá en pulsar el botón **Next** hasta que el proceso de instalación finalice, sin tocar ningún ítem de configuración.

Imagen 139. Ventana inicial para la instalación de Gawk

COREUTILS

Imagen 140. Ventana inicial para la instalación de CoreUtils

El siguiente programa a instalar es una colección de utilidades GNU para Windows. Como el programa anterior, la única acción a realizar para la instalación de esta herramienta es pulsar el botón **Next** hasta que termine el proceso.

XLRD-0.9.2-3_py27

A continuación, se instala el programa Xlrd-0.9.2-3, complemento de Python(x,y) que permite al usuario trabajar con ficheros MSExcel 2007. Para su instalación, únicamente se tiene que pulsar **Next** en las ventanas de navegación.

Imagen 141. Ventana inicial para la instalación de xlrd 0.9.2-3

PYODBC-3.0.7

El siguiente programa a instalar es Pyodbc-3.0.7, el cual permite realizar conexiones mediante ODBC con diferentes formatos de bases de datos. Como en el caso anterior, para su instalación únicamente se tiene que pulsar **Siguiente** en las diferentes ventanas de navegación.

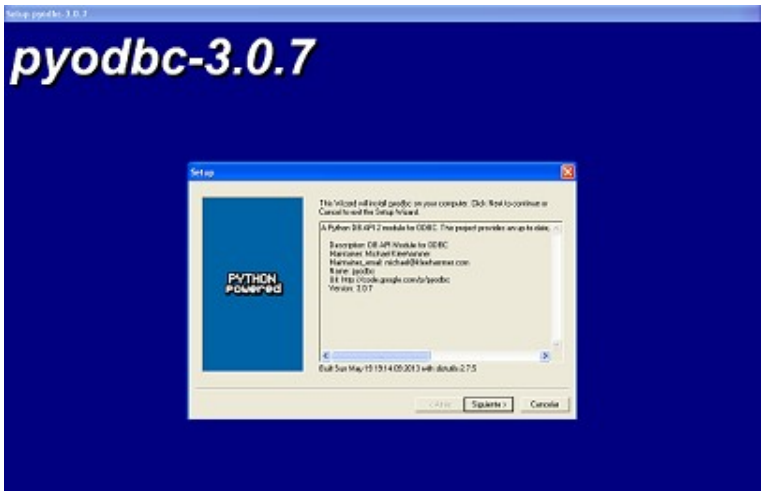


Imagen 142. Ventana inicial para la instalación de pyodbc-3.0.7

DBFPY-2.2.5.win32

Este programa permite realizar conexiones con bases de datos DBF. Para su instalación, el usuario tendrá que pulsar **Siguiente** en las ventanas de navegación que aparecen.

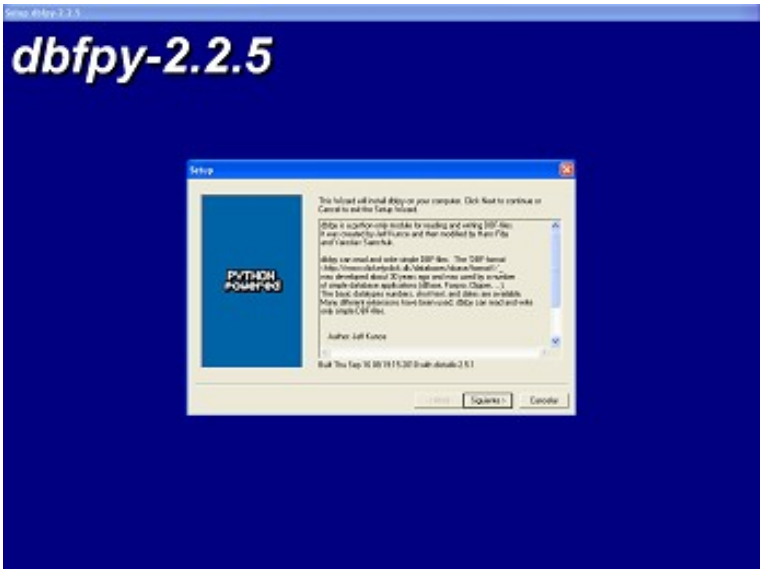


Imagen 143. Ventana inicial para la instalación de dbfpy-2.2.5

Si durante la instalación de este programa aparece el siguiente mensaje:

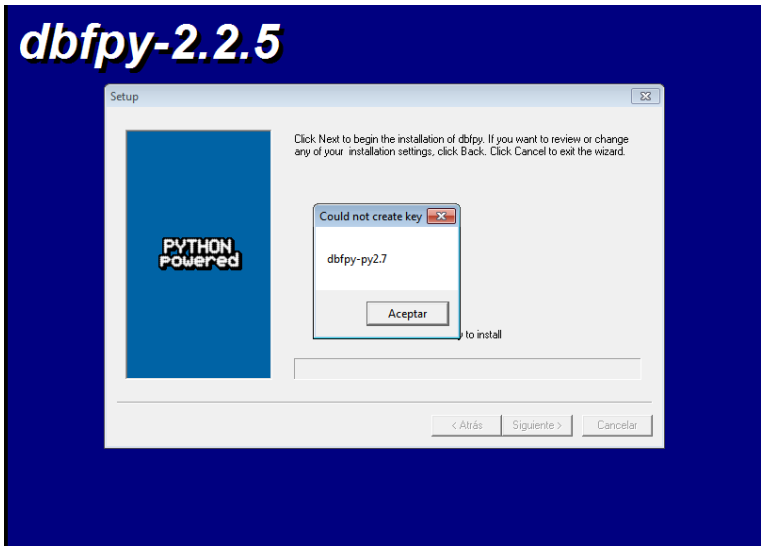


Imagen 144. Mensaje de No instalación de dbfpy-2.2.5

el usuario deberá aceptarlo, así como los dos siguientes mensajes que aparecerán y continuar con la instalación del resto de programas. Esto significa que el programa no se ha instalado, por tanto, el usuario deberá dirigirse a la carpeta donde tenga el ejecutable de este programa (dbfpy-2.2.5.win32.exe), situarse sobre el mismo y pulsando con el botón derecho del ratón elegir la opción ejecutar como administrador.

NOTEPAD2

El último programa a instalar es un editor de texto mejorado similar al bloc de notas que trae instalado por defecto Windows. Este programa permite la edición de textos evitando que ocurran problemas de codificación. Pulsando **Instalar** comienza su proceso de instalación.

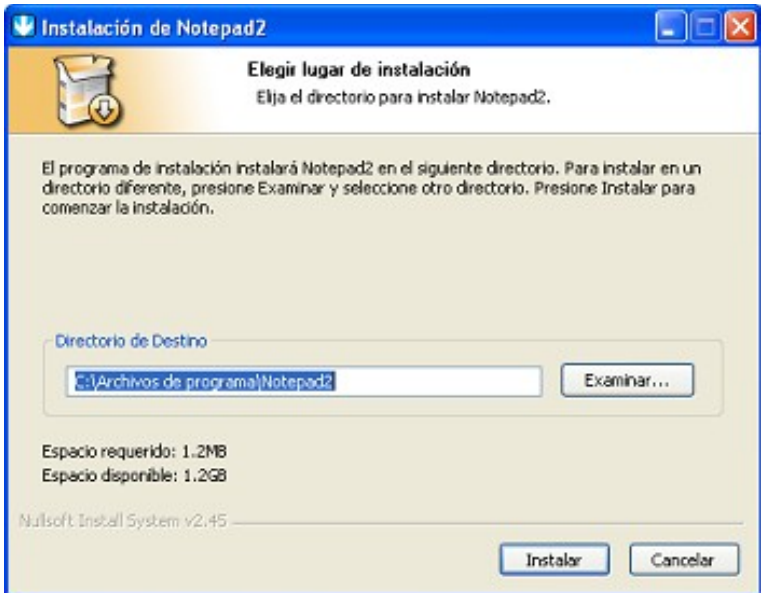


Imagen 145. Ventana inicial para la instalación de Notepad2

A continuación, es necesario instalar los paquetes *chardet* y *odfpy* de Python 2.7.2. Para ello, el usuario deberá ir al Menú Inicio y acceder a la línea de comandos de Windows ejecutando el comando *cmd*, tal y como se observa en la siguiente imagen:

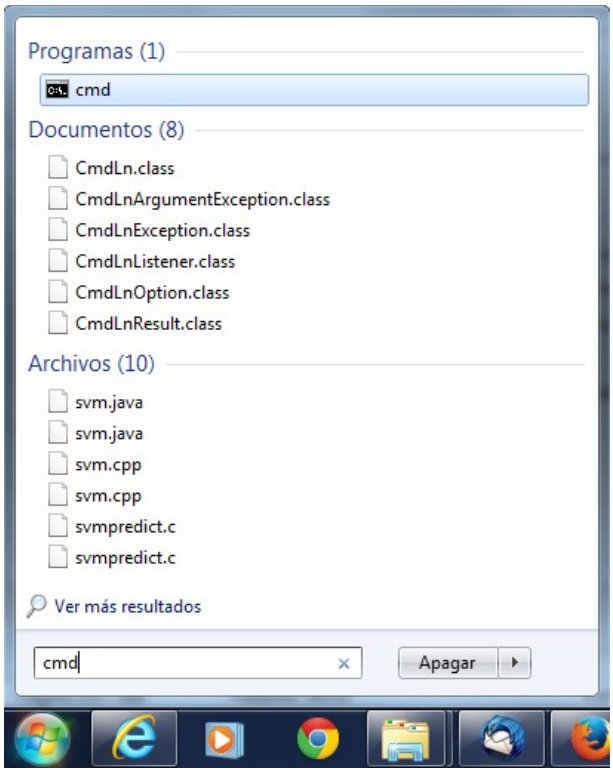


Imagen 146. Ventana menú Inicio comando cmd

En la ventana que aparece se ejecutarán las siguientes órdenes:

- C:\Python27\Scripts\easy_install-2.7.exe chardet
- C:\Python27\Scripts\easy_install-2.7.exe odfpy

3. Tras la instalación de estos paquetes, el usuario deberá copiar respectivamente los siguientes ficheros:

- libsvm.dll
- svm.py y svmutil.py

en la siguiente ubicación:

- libsvm.dll en C:\Python27\DLLs
- svm.py y svmutil.py en C:\Python27\Lib

Por último, el usuario deberá añadir a la variable PATH del sistema la siguiente ruta:

- C:\Program Files (x86)\GnuWin32\bin para sistemas de 64 bits
- C:\Archivos de programa\GnuWin32\bin para sistemas de 32 bits

Para ello deberá acceder a las propiedades del sistema para modificar la variable de entorno PATH. La manera de proceder es la que se observa en la siguiente imagen:

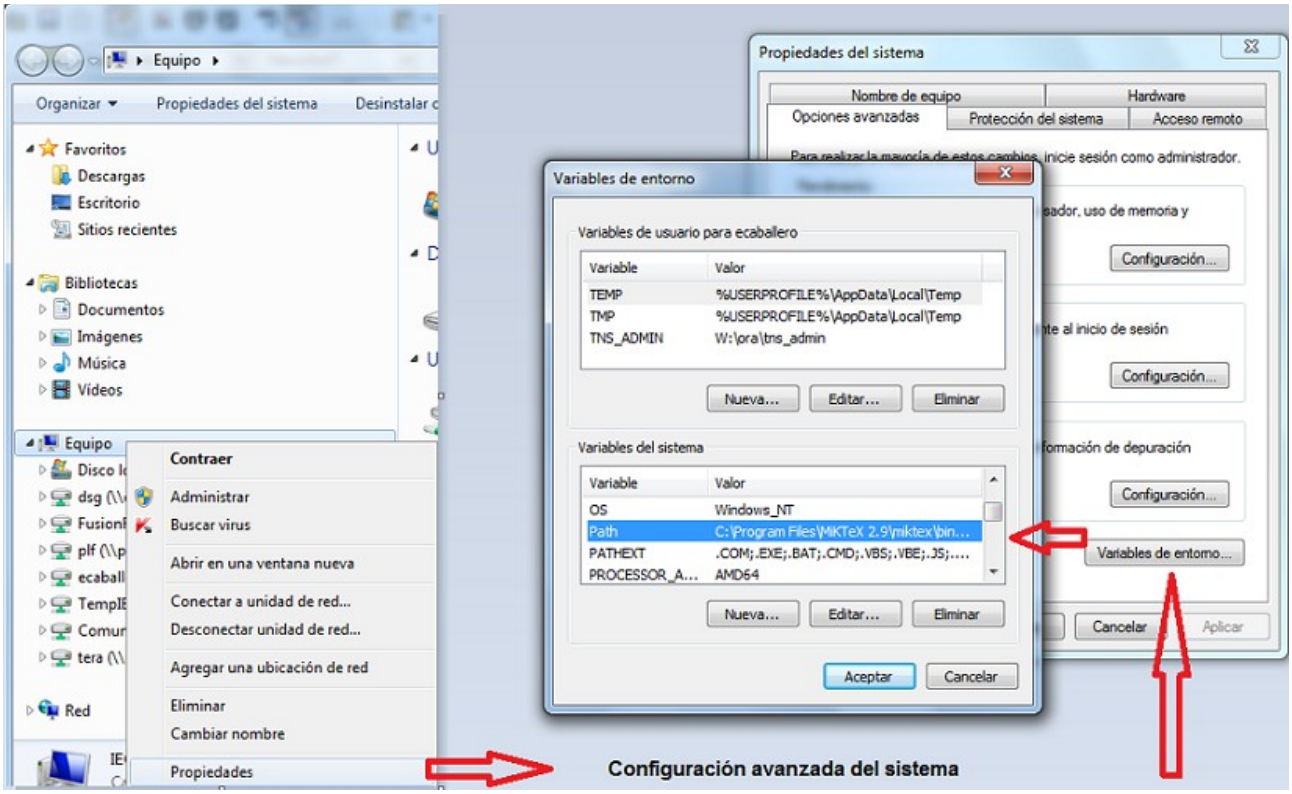


Imagen 147. Establecimiento de la variable PATH del sistema

Anexo II: Modelos Ocultos de Markov disponibles en *aLink: Herramienta de Fusión de Ficheros*

En *aLink: Herramienta de Fusión de Ficheros* se proporciona al usuario una serie de modelos HMM de partida que podría utilizar para normalizar de primeras un fichero de datos. Además de los modelos se proporcionan las muestras a partir de las cuales se han generado los mismos, así como su ubicación. No obstante, si el modelo HMM utilizado no resulta del todo adecuado se podrá modificar la muestra con la que se generó para disponer de un modelo más eficiente. En concreto se tiene:

Para nombres de personas:

Ubicación: aLink/app/muestras_modelos/nombres/muestras_modelos		
Denominación del modelo HMM para nombres de personas	**Muestra de la que procede el modelo**	**Descripción**
modeloHMM_nombres_pila.hmm	muestra_nombres_pila.csv	Modelo HMM generado a partir de una muestra que contiene estructuras relativas a nombres de persona simples y compuestos, del tipo: Juan, Marta, José Manuel, María del Carmen, etc.
modeloHMM_ape1.hmm	muestra_ape1.csv	Modelo HMM generado a partir de una muestra que contiene estructuras relativas a primer apellido de persona, tanto simples como compuestos: García, Fernández de la Vega, etc.
modeloHMM_ape2.hmm	muestra_ape2.csv	Modelo HMM generado a partir de una muestra que contiene estructuras relativas a segundo apellido de personas, tanto simples como compuestos: García, Fernández de la Vega, etc.
modeloHMM_dos_apellidos.hmm	muestra_dos_apellidos.csv	Modelo HMM generado a partir de una muestra que contiene estructuras relativas a dos apellidos de persona, tanto simples como compuestos, del tipo: García Fernández de la Vega, González Martín, etc.
modeloHMM_nombresyapellidos.hmm	muestra_nombresyapellidos.csv	Modelo HMM generado a partir de una muestra que contiene estructuras relativas a nombres y dos apellidos de persona, tanto simples como compuestos, del tipo: Antonio García Fernández de la Vega, Marta González Martín, María del Pilar Gámiz Luque, etc.

Tabla 10. Modelos HMM para nombres de personas disponibles en *aLink:Herramienta de Fusión de Ficheros*

Para direcciones postales:

Ubicación: aLink/app/muestras_modelos/direcciones/muestras_modelos		
Denominación del modelo HMM para direcciones postales	**Muestra de la que procede el modelo**	**Descripción**
modeloHMM_direcciones_amedida.hmm	muestra_direcciones_amedida.csv	Modelo HMM generado a partir de una muestra que contiene estructuras relativas a direcciones postales completas, es decir, tipo de vía, nombre de vía y número de vía, escalera, bloque, planta, puerta, etc. y que permite desagregar los campos de la dirección postal de acuerdo a libre elección del usuario
modeloHMM_direcciones_CDAU.hmm	muestra_direcciones_CDAU.csv	Modelo HMM generado a partir de una muestra que contiene estructuras relativas a direcciones postales completas, es

Ubicación: aLink/app/muestras_modelos/direcciones/muestras_modelos		
		decir, tipo de vía, nombre de vía y número de vía, escalera, bloque, planta, puerta, etc. y que permite desagregar los campos de la dirección postal de acuerdo a CDAU

Tabla 11. Modelos HMM para direcciones postales disponibles en *aLink:Herramienta de Fusión de Ficheros*

Para identificadores de personas físicas y/o jurídicas:

Ubicación: aLink/app/muestras_modelos/idpersonas/modelo_propuesto		
Denominación del modelo HMM para identificadores de personas físicas y/o jurídicas	**Muestra de la que procede el modelo**	**Descripción**
modeloHMM_idpersona.hmm	- -	Modelo HMM generado a partir de una muestra que contiene estructuras de DNI, NIF y NIE

Tabla 12. Modelo HMM para identificadores de personas físicas y/o jurídicas disponible en *aLink:Herramienta de Fusión de Ficheros*

Nótese que para el caso de identificadores de personas físicas y/o jurídicas solo se proporciona el modelo HMM definitivo sin la muestra de partida. El motivo es que al contrario de los casos anteriores, la muestra a partir de la cual se ha construido este modelo contiene todas las estructuras posibles de segmentación de un DNI, NIF o NIE y no tendría que ser alimentada con ninguna estructura más, con lo cual se ha decidido proporcionar solo el modelo HMM.

Anexo III: Campos de salida para nombres de personas e identificadores de personas físicas y/o jurídicas

Campos de salida para nombres de personas:

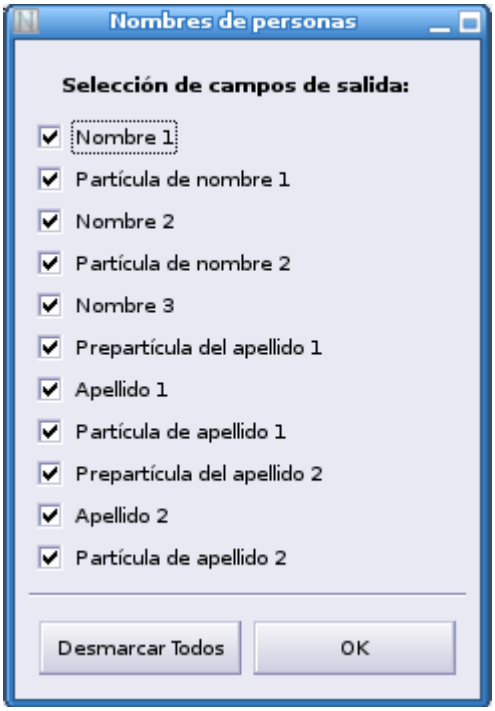


Imagen 148. Campos de salida nombres de personas

Campos de salida para identificadores de personas físicas y/o jurídicas:

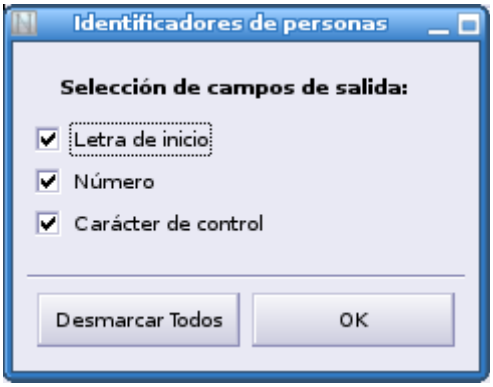


Imagen 149. Campos de salida identificadores de personas físicas y/o jurídicas

Anexo IV: Campos de salida del fichero normalizado

En las siguientes tablas se muestran los campos de salida en los que la Herramienta de Normalización puede desagregar un campo de un fichero que contenga nombres de personas, direcciones postales o identificadores de personas físicas y jurídicas. Junto a estos campos se indica la denominación con la que aparecen en el fichero normalizado ya que en algunos casos no es la misma. Obsérvese además, que la denominación de los campos del fichero normalizado ni contienen tildes ni espacios en blanco.

Campos de salida para nombres de personas:

Denominación del campo de salida en interfaz de aLink	Denominación del campo de salida en el fichero normalizado
Nombre 1	nombre1
Partícula de nombre 1	particula_nombre1
Nombre 2	nombre2
Partícula de nombre 2	particula_nombre2
Nombre 3	nombre3
Prepartícula del apellido 1	preparticula_apellido1
Apellido 1	apellido1
Partícula del apellido 1	particula_apellido1
Prepartícula del apellido 2	preparticula_apellido2
Apellido 2	apellido2
Partícula del apellido 2	particula_apellido2

Tabla 13. Campos de salida del fichero normalizado. Nombres de personas

Campos de salida para direcciones postales:

En este caso se añade a la tabla una nueva columna que recoge si el campo de salida se muestra en el fichero normalizado al usar la desagregación CDAU o no.

Denominación del campo de salida en interfaz de aLink	Denominación del campo de salida en el fichero normalizado	¿En desagregación CDAU?
Tipo de vía	tipo_de_via	Sí
Nombre de vía	nombre_de_via	Sí
Id. de numeración	identificador_de_numeracion	Sí
Entidad inferior de numeración	ein	Sí
Calificador ent. inf. de numeración	cein	Sí
Entidad superior de numeración	esn	Sí
Calificador ent. sup. de numeración	cesn	Sí
Id. de bloque	identificador_de_bloque	No
Bloque	bloque	Sí
Tipo de edificio	tipo_de_edificio	No
Edificio	edificio	No
Id. de portal	identificador_de_portal	No
Portal	portal	Sí
Id. de escalera	identificador_de_escalera	No
Escalera	escalera	Sí
Id. de planta	identificador_de_planta	No
Planta	planta	Sí
Id. de puerta	identificador_de_puerta	No
Puerta	puerta	Sí
Id. de letra	identificador_de_letra	No
Letra	letra	No
Entidad singular	entidad_singular	Sí
Municipio	municipio	Sí
Provincia	provincia	Sí

Denominación del campo de salida en interfaz de aLink	Denominación del campo de salida en el fichero normalizado	¿En desagregación CDAU?
Id. de código postal	identificador_de_codigo_postal	No
Código postal	codigo_postal	Sí
Tipo de agrupación	tipo_de_agrupacion	Sí
Agrupación	agrupacion	Sí
Id. de sector	identificador_de_sector	No
Sector	sector	No
Id. de manzana	identificador_de_manzana	No
Manzana	manzana	No
Id. de parcela	identificador_de_parcela	No
Parcela	parcela	No
Id. de nave	identificador_de_nave	No
Nave	nave	No
Tipo de zona	identificador_de_zona	No
Zona	zona	No
Otros datos de ubicación (ODUB)	odub	Sí

Tabla 14. Campos de salida del fichero normalizado. Direcciones postales

Campos de salida para identificadores de personas físicas y/o jurídicas:

Denominación del campo de salida en interfaz de aLink	Denominación del campo de salida en el fichero normalizado
Letra de inicio	letra_inicio
Número	numero_id
Carácter de control	caracter_control

Tabla 15. Campos de salida del fichero normalizado. Id. personas físicas y/o jurídicas

Anexo V: Etiquetas usadas en el proceso de normalización para construir un modelo HMM

En este Anexo se especifican las etiquetas usadas en el proceso de normalización realizado con la Herramienta de Normalización, las cuales son necesarias para construir el Modelo Oculto de Markov. Se muestran etiquetas para nombres de personas y direcciones postales. Junto a las etiquetas se indica una descripción de las mismas así como la tabla de búsqueda a través de la que están definidas. Nótese que existen etiquetas que no van a tener una tabla de búsqueda asociada, estas son las referidas a valores numéricos, palabras de una sola letra o elementos no encontrados en las tablas de búsqueda.

Etiquetas para nombres de personas

Etiqueta	Descripción	Tabla de búsqueda
NF	Etiqueta asociada a un nombre femenino	knombres_femeninos.tbl
NM	Etiqueta asociada a un nombre masculino	knombres_masculinos.tbl
NN	Etiqueta asociada a un nombre neutro	knombres_neutros.tbl
PS	Etiqueta asociada a partículas ligadas a nombres y/o apellidos	kparticulas.tbl
LE	Etiqueta asociada a palabras de una sola letra	No tiene tabla de búsqueda asociada
UN	Etiqueta asociada a valores no encontrados en ninguna de las tablas de búsqueda de nombres de personas	No tiene tabla de búsqueda asociada

Tabla 16. Etiquetas para normbres de personas

Etiquetas para direcciones postales

Etiquetas	Descripción	Tabla de búsqueda
AG	Etiqueta asociada a conjuntos de construcciones no consideradas como núcleos de población en el Nomenclátor del INE tales como: barrios, barriadas, urbanizaciones, polígonos industriales y parques comerciales. Además, también tiene asociada elementos referidos a la identificación de complejos, conjuntos o grupos, como por ejemplo, complejos hoteleros, residenciales, etc.	kagrupacion.tbl
BL	Etiqueta asociada a elementos referidos a la identificación de un bloque (bloque, blq, bl,	kbloque.tbl

Etiquetas	Descripción	Tabla de búsqueda
	etc.)	
CP	Etiqueta asociada a elementos referidos a la identificación de códigos postales (codigo postal, CP, C.P., etc.)	kcodigo_postal.tbl
ED	Etiqueta asociada a elementos referidos a tipos de edificios, como edificio, casa, finca, chalet, caserío, cortijo, etc. También se asocia a edificios singulares tales como como ayuntamientos, bibliotecas, aeropuertos, puertos, estaciones de ferrocarril o de autobuses, colegios, institutos, etc., así como a comercios, por ejemplo, mercados, plazas de abastos, supermercados, centros comerciales, gasolineras, hoteles, campings, etc.	kedificio.tbl
EG	Etiqueta asociada a las entidades singulares existentes en la Comunidad Autónoma andaluza	kentidad_singular.tbl
ES	Etiqueta asociada a elementos referidos a la identificación de la escalera de un bloque o edificio (escalera, esc, esca, etc.)	kescalera.tbl
LE	Etiqueta asociada a elementos referidos a la identificación de la letra de la puerta de una vivienda (letra, letr)	kletra.tbl
	Etiqueta asociada a palabras de una sola letra	No tiene tabla de búsqueda asociada
MU	Etiqueta asociada a los municipios de la Comunidad Autónoma andaluza	kmunicipio.tbl
MZ	Etiqueta asociada a elementos que identifican manzanas (manzana, mzna)	kmanzana.tbl
N5	Etiqueta asociada a valores numéricos con cinco dígitos	No tiene tabla de búsqueda asociada
NM	Etiqueta asociada a elementos que identifican el número de la vía, local, punto kilométrico etc. (numero, nº, num, local, sin número, s/n, kilometro, km, pk, etc.)	knumero_local.tbl
NP	Etiqueta asociada a elementos que identifican el número de la planta de un bloque o edificio (1º, 2º, 3º, ático, bajo, etc.)	kplanta_numero.tbl
NU	Etiqueta asociada a valores numéricos. Se excluyen los números con cinco dígitos	No tiene tabla de búsqueda asociada
NV	Etiqueta asociada a elementos que identifican naves industriales (nave, nav)	knave.tbl
PA	Etiqueta asociada a elementos que identifican parcelas (parcela, parc)	kparcela.tbl
PL	Etiqueta asociada a elementos que identifican la planta de un bloque o edificio (planta, plt, plnt, etc.)	kplanta.tbl
PR	Etiqueta asociada a las ocho provincias de la Comunidad Autónoma andaluza	kprovicia.tbl
PT	Etiqueta asociada a elementos que identifican a un portal (portal, prtal, ptal, etc.)	kportal.tbl

Etiquetas	Descripción	Tabla de búsqueda
PU	Etiqueta asociada a elementos que identifican la puerta de una vivienda (puerta, prta, pu, puert, etc.)	kpuerta.tbl
ST	Etiqueta asociada a elementos identificativos de sectores (sector, sect)	ksector.tbl
TV	Etiqueta asociada a elementos identificativos del tipo de vía (calle, c/, avenida, avda, plz, carretera, etc.)	kvia.tbl
UN	Etiqueta asociada a elementos no incluidos en ninguna tabla de búsqueda de direcciones	No tiene tabla de búsqueda asociada
ZO	Etiqueta asociada a elementos que identifican zonas tales como parques, jardines, paseos, parajes, arboledas, pagos, lugares, etc.	kzona.tbl

Tabla 17. Etiquetas para direcciones postales

Anexo VI: Usar HMM anterior

Cuando en un proceso de normalización de un fichero de datos se dispone de un Modelo Oculto de Markov creado previamente a partir de un fichero con diseño de registro similar a este, el proceso de selección y asignación de estados de la muestra se puede simplificar. El motivo se debe a que la opción **Usar HMM anterior** de la herramienta ***HMM: Selección de la muestra*** permite asignar estados automáticamente a las etiquetas de la muestra seleccionada. Así, la muestra de entrenamiento obtenida va a contener para cada registro las etiquetas y estados que el modelo HMM le asigne automáticamente. No obstante, se aconseja realizar una revisión manual del fichero con la muestra etiquetada con el fin de corregir posibles errores en el proceso de asignación de estados. A continuación, se explica mediante un ejemplo cómo funciona este proceso. El fichero que se utilizará para ello contiene un campo denominado *direccion* con direcciones postales de una serie de establecimientos comerciales del tipo:

	A
1	'direccion'
2	AVDA CADIZ S/N
3	CALLE FERNANDO ZOBEL 6
4	AVDA CONCEJAL ALBERTO JIMENEZ 2
5	CALLE RODRIGO TRIANA 94
6	CALLE ALC MNEL REYES 2
7	AVDA ANDALUCIA 23
8	CALLE ALFARERIA 126
9	CALLE ARTESANIA SN
10	CALLE VELARDE S.N.
11	CALLE VIRGEN DE LA VICTORIA S/ N
12	AVDA S FCO JAVIER S/numero
13	AVDA JOSE BARRIONUEVO PEKKA
14	CALLE RIO ANDARAX S N
15	C/ MALAGA Snumero
16	CALLE INDUST LA RED S$N
17	CALLE MONTURRIO S/NUMERO
18	C/ PIEDRA CABALLERA S/ NUMERO
19	AVDA MIJAS ED GUADALUPE
20	CALLE EL CARMEN
21	CALLE QUIMICA 23
22	CALLE SAN FRANCISCO 39
23	PLAZA DUQUESA 2
24	AVDA INDUSTRIA DE LA 23
25	CALLE IV CONDE UREKKA 21
26	CALLE CHAMIZO 17
27	CALLE JOSE DIAZ 1
28	CALLE TOCINA 30
29	AVDA GRECO 1

Imagen 150. Fichero con direcciones postales de establecimientos comerciales

Así por ejemplo, si se hubiera seleccionado una muestra de cinco registros del campo a normalizar y no se hubiera utilizado un modelo HMM anterior, un posible fichero de salida sería:

```
1 ##########################################################################################
2 # Creado Mon Dec  2 13:12:57 2013
3 #
4 # Fichero de entrada: /home/ecaballero/fusion-ficheros/aLink_callejero_v93_33/Ejemplo/direcciones_establecimientos_tratado.csv
5 # Fichero de salida:  /home/ecaballero/fusion-ficheros/aLink_callejero_v93_33/Ejemplo/muestra_etiquetada_20131202-1312_direcciones_establecimientos_tratado.csv
6 # Componente: direccion
7 # Parámetros:
8 # - Posición del primer registro: 0
9 # - Posición del ultimo registro:    1230
10 # - Numero de registros seleccionados y etiquetados: 5
11 #
12 ##################### Descripcion de las etiquetas:
13 #
14 # NOTA:
15 # Incluiremos la etiqueta y la tabla de busqueda a la que corresponde, por ejemplo, TV significa que la palabra
16 # identificada con esta etiqueta se encuentra en la tabla de busqueda de tipos de via.
17 #
18 # Etiquetas  y tablas de busqueda asociadas:
19 #
20 # AG corresponde a agrupaciones                  NU corresponde a numeros (*)
21 # BL corresponde a bloques                       N5 corresponde a numeros de 5 digitos (*)
22 # CP corresponde a codigos postales              NV corresponde a naves
23 # ED corresponde a edificios                     PA corresponde a parcelas
24 # EG corresponde a entidades singulares          PR corresponde a provincias
25 # ES corresponde a escaleras                     PT corresponde a portales
26 # LE corresponde a letras (*)                    PU corresponde a puertas
27 # MU corresponde a municipios                    ST corresponde a sectores
28 # MZ corresponde a manzanas                      TV corresponde a tipos de via
29 # NM corresponde a identificadores de numeros    UN no se encuentra incluida en ninguna tabla de busqueda (*)
30 # NP corresponde a numeros de planta             ZO corresponde a zonas
31 #
32 ##################### Listado de posibles estados para direcciones:
33 #
34 # tipo_de_via                          nombre_de_via
35 # identificador_de_numeracion
36 # ein                                  cein
37 # esn                                  cesn
38 # identificador_de_bloque              bloque
39 # tipo_de_edificio                     edificio
40 # identificador_de_portal              portal
41 # identificador_de_escalera            escalera
42 # identificador_de_planta              planta
43 # identificador_de_puerta              puerta
44 # identificador_de_letra               letra
45 # entidad_singular                     municipio
46 # provincia
47 # identificador_de_codigo_postal       codigo_postal
48 # tipo_de_agrupacion                   agrupacion
49 # identificador_de_sector              sector
50 # identificador_de_manzana             manzana
51 # identificador_de_parcela             parcela
52 # identificador_de_nave                nave
53 # identificador_de_zona                zona
54 # odub
55 ##########################################################################################
56
57 # 12 (0): |AVDA JOSE BARRIONUEVO PEKKA|
58 #         |avenida jose barrionuevo pekka|
59   TV:, UN:, UN:, UN:
60
61 # 30 (1): |CALLE MALAGA S/N|
62 #         |calle malaga sin_numero|
63   TV:, MU:, NM:
64   TV:, PR:, NM:
65
66 # 7 (2): |CALLE ALFARERIA 126|
67 #        |calle alfareria 126|
68   TV:, UN:, NU:
69
70 # 15 (3): |CALLE EL CARMEN|
71 #         |calle el carmen|
72   TV:, UN:, UN:
73
74 # 24 (4): |PLZA DUQUESA 21|
75 #         |plaza duquesa 21|
76   TV:, UN:, NU:
```

Imagen 151. Fichero con muestra etiquetada sin usar HMM anterior

Por el contrario, si se hubiera utilizado un modelo HMM creado previamente un posible fichero de salida sería:

```
1 ##########################################################################################
2 #
3 # Creado Tue Dec 10 13:50:57 2013
4 #
5 # Fichero de entrada: /home/ecaballero/fusion-ficheros/Ficheros/practicas_usuarios/ecaballero/muestra_directorio_establecimientos_EUSTAT_tratado.utf-8.csv
6 # Fichero de salida:  /home/ecaballero/fusion-ficheros/Ficheros/practicas_usuarios/ecaballero/muestra_etiquetada_20131210-1350_muestra_directorio_establecimientos_EUSTAT_tratac
7 # Componente: direccion
8 # Parámetros:
9 # - Posición del primer registro: 0
10 # - Posición del ultimo registro:    1230
11 # - Numero de registros seleccionados y etiquetados: 5
12 # - HMM usado: /home/ecaballero/fusion-ficheros/Ficheros/practicas_usuarios/ecaballero/muestra_eustat_con_estados_asignado_20131210-1347.hmm
13 # Fichero de frecuencia de patrones: /home/ecaballero/fusion-ficheros/Ficheros/practicas_usuarios/ecaballero/FPP_20131210-1349.txt
14 #
15 #
16 #
17 ##################### Descripcion de las etiquetas:
18 #
19 # NOTA:
20 # Incluiremos la etiqueta y la tabla de busqueda a la que corresponde, por ejemplo, TV significa que la palabra
21 # identificada con esta etiqueta se encuentra en la tabla de busqueda de tipos de via.
22 #
23 # Etiquetas  y tablas de busqueda asociadas:
24 #
25 # AG corresponde a agrupaciones                 NU corresponde a numeros (*)
26 # BL corresponde a bloques                      N5 corresponde a numeros de 5 digitos (*)
27 # CP corresponde a codigos postales             NV corresponde a naves
28 # ED corresponde a edificios                    PA corresponde a parcelas
29 # EG corresponde a entidades singulares         PR corresponde a provincias
30 # ES corresponde a escaleras                    PT corresponde a portales
31 # LE corresponde a letras (*)                   PU corresponde a puertass
32 # MU corresponde a municipios                   ST corresponde a sectores
33 # MZ corresponde a manzanas                     TV corresponde a tipos de via
34 # NM corresponde a identificadores de numeros   UN no se encuentra incluida en ninguna tabla de busqueda (*)
35 # NP corresponde a numeros de planta            ZO corresponde a zonas
36 # (*) Estas etiquetas no estan asociadas a ninguna tabla de busqueda, se asignan automaticamente.
37 #
38 #
39 ##################### Listado de posibles estados para direcciones:
40 #
41 # tipo_de_via                          nombre_de_via
42 # identificador_de_numeracion
43 # ein                                  cein
44 # esn                                  cesn
45 # identificador_de_bloque              bloque
46 # tipo_de_edificio                     edificio
47 # identificador_de_portal              portal
48 # identificador_de_escalera            escalera
49 # identificador_de_planta              planta
50 # identificador_de_puerta              puerta
51 # identificador_de_letra               letra
52 # entidad_singular                     municipio
53 # provincia
54 # identificador_de_codigo_postal       codigo_postal
55 # tipo_de_agrupacion                   agrupacion
56 # identificador_de_sector              sector
57 # identificador_de_manzana             manzana
58 # identificador_de_parcela             parcela
59 # identificador_de_nave                nave
60 # identificador_de_zona                zona
61 # odub
62 ##########################################################################################
63
64 # 4 (0): |CALLE RODRIGO TRIANA 94|
65 #        |calle rodrigo triana 94|
66    TV:tipo_de_via, UN:tipo_de_via, EG:tipo_de_via, NU:tipo_de_via
67 # Probabilidad máxima de Viterbi: 0.00000
68
69 # 5 (1): |CALLE ALC MNEL REYES 2|
70 #        |calle alc mnel reyes 2|
71    TV:tipo_de_via, UN:nombre_de_via, UN:nombre_de_via, UN:nombre_de_via, NU:ein
72 # Probabilidad máxima de Viterbi: 0.05583
73
74 # 9 (2): |CALLE VELARDE S.N.|
75 #        |calle velarde sin_numero|
76    TV:tipo_de_via, UN:nombre_de_via, NM:identificador_de_numeracion
77 # Probabilidad máxima de Viterbi: 0.14583
78
79 # 13 (3): |CALLE RIO ANDARAX S N|
80 #         |calle rio andarax sin_numero|
81    TV:tipo_de_via, EG:tipo_de_via, UN:tipo_de_via, NM:tipo_de_via
82 # Probabilidad máxima de Viterbi: 0.00000
83
84 # 16 (4): |CALLE MONTURRIO S/NUMERO|
85 #         |calle monturrio sin_numero|
86    TV:tipo_de_via, UN:nombre_de_via, NM:identificador_de_numeracion
87 # Probabilidad máxima de Viterbi: 0.14583
```

Imagen 152. Fichero con muestra etiquetada usando HMM anterior

Como se puede observar al utilizar un modelo HMM anterior se han asignado estados a las etiquetas automáticamente sin necesidad de hacerlo de forma manual, no obstante en el primer registro se ha producido un error, ya que al elemento '94' le ha asignado el estado *tipo_de_via* y el correcto sería *ein.* Además para cada registro se muestra la probabilidad máxima que tiene cada dirección postal de la muestra de seguir el patrón o secuencia de etiquetas y estados asignado. Dicha probabilidad se calcula mediante el algoritmo de Viterbi.

A parte del fichero anterior, en este proceso de selección de la muestra se genera un fichero adicional con extensión *.txt*, el fichero de frecuencia de patrones. Dicho fichero contiene la frecuencia con la que aparecen los patrones en la muestra. Este se encuentra

ubicado en la misma ruta que el fichero tratado del que se extrae la muestra y su denominación sigue el formato:

FFP_20131210-1349.txt

El contenido del mismo se muestra en la siguiente imagen:

```
1 ##############################################################################################
2 # Fichero de frecuencia de patrones.
3 #
4 # Creado Tue Dec 10 13:50:57 2013
5 #
6 # Fichero de entrada: /home/ecaballero/fusion-ficheros/Ficheros/practicas_usuarios/ecaballero/muestra_directorio_establecimientos_EUSTAT_tratado.ut
7 # Fichero de salida:  /home/ecaballero/fusion-ficheros/Ficheros/practicas_usuarios/ecaballero/muestra_etiquetada_20131210-1350_muestra_directorio_e
8 # Parametros:
9 # - Posicion del primer registro: 0
10 # - Posicion del ultimo registro: 1230
11 # - Numero de registros seleccionadosy etiquetados: 5
12 # Modelo HMM usado:  /home/ecaballero/fusion-ficheros/Ficheros/practicas_usuarios/ecaballero/muestra_eustat_con_estados_asignado_20131210-1347.hmm
13 #
14 ##############################################################################################
15
16 # Patrón:    TV:tipo_de_via, EG:tipo_de_via, UN:tipo_de_via, NM:tipo_de_via
17 # Frecuencia: 1
18 # Probabilidad máxima de Viterbi: 0.0
19 # Ejemplos:
20 #    |calle rio andarax sin_numero|
21    TV:tipo_de_via, EG:tipo_de_via, UN:tipo_de_via, NM:tipo_de_via
22
23 # Patrón:    TV:tipo_de_via, UN:nombre_de_via, UN:nombre_de_via, UN:nombre_de_via, NU:ein
24 # Frecuencia: 1
25 # Probabilidad máxima de Viterbi: 0.0558267670898
26 # Ejemplos:
27 #    |calle alc mnel reyes 2|
28    TV:tipo_de_via, UN:nombre_de_via, UN:nombre_de_via, UN:nombre_de_via, NU:ein
29
30 # Patrón:    TV:tipo_de_via, UN:tipo_de_via, EG:tipo_de_via, NU:tipo_de_via
31 # Frecuencia: 1
32 # Probabilidad máxima de Viterbi: 0.0
33 # Ejemplos:
34 #    |calle rodrigo triana 94|
35    TV:tipo_de_via, UN:tipo_de_via, EG:tipo_de_via, NU:tipo_de_via
36
37 # Patrón:    TV:tipo_de_via, UN:nombre_de_via, NM:identificador_de_numeracion
38 # Frecuencia: 2
39 # Probabilidad máxima de Viterbi: 0.145833625
40 # Ejemplos:
41 #    |calle velarde sin_numero|
42 #    |calle monturrio sin_numero|
43    TV:tipo_de_via, UN:nombre_de_via, NM:identificador_de_numeracion
```

Imagen 153. Fichero de frecuencia de patrones

En la imagen se observa, a modo de comentario, las secuencias de etiquetas y estados asociados al patrón o estructura que sigue cada una de las direcciones postales de la muestra, la frecuencia de aparición de ese patrón en la muestra, la probabilidad máxima que tiene cada dirección postal de la muestra de tener el patrón o secuencia de etiquetas y estados asignado, así como ejemplos de la muestra para los que se ha encontrado dicho patrón. Por último, se muestra la secuencia de etiquetas y estados asociados, que es la misma que la que aparece en el fichero '*.csv*' utilizado por la aplicación.

Anexo VII: Estados usados en el proceso de normalización para construir un modelo HMM

En este Anexo se especifican los estados usados en el proceso de normalización realizado con la Herramienta de Normalización. Estos son necesarios para construir los Modelos Ocultos de Markov. Se muestran estados para nombres de personas y direcciones postales. Junto a ellos se indica la utilidad de los mismos.

Estados para nombres de personas

Estado	Utilidad
nombre1	Para identificar el primer nombre de pila
particula_nombre1	Para identificar partículas que siguen al primer nombre si es compuesto (Ejemplo: María **del** Carmen, María **de los** Ángeles)
nombre2	Para identificar el segundo nombre de pila
particula_nombre2	Para identificar partículas que siguen al segundo nombre si es compuesto (Ejemplo: María Jesús **de los** Ángeles)
nombre3	Para identificar el tercer nombre de pila
preparticula_apellido1	Para identificar prepartículas que preceden al primer apellido si es compuesto (Ejemplo: **del** Rosal, **de la** Vega)
apellido1	Para identificar el primer apellido
particula_apellido1	Para identificar partículas que siguen al primer apellido si es compuesto (Ejemplo: Fernández **de la** Vega, López **del** Moral)
apellido2	Para identificar el segundo apellido
particula_apellido2	Para identificar partículas que siguen al segundo apellido si es compuesto (Ejemplo: Martín **de la** Rosa, López **del** Moral)

Tabla 18. Estados para nombres de personas

Estados para direcciones postales

Estado	Utilidad
tipo_de_via	Para identificar el tipo de vía
nombre_de_via	Para indicar el nombre de la vía
identificador_de_numeracion	Para identificar elementos relacionados con identificadores del número de la vía, por ejemplo, n°, num, local, s/n, etc. o con el punto kilométrico de la carretera, esto es, km, pk, etc.
ein	Para identificar la entidad inferior de numeración de la vivienda o local
cein	Para identificar el calificador de la entidad inferior de numeración
esn	Para identificar la entidad superior de numeración de la vivienda o local
cesn	Para identificar el calificador de la entidad superior de numeración
tipo_de_edificio	Para identificar elementos relacionados con tipos de edificios, por ejemplo, edificio, edif, caserio, cortijo, finca, ayuntamientos, bibliotecas, aeropuertos, mercados, plazas de abastos, centros comerciales, hoteles, etc.
edificio	Para identificar la denominación del edificio
identificador_de_bloque	Para identificar elementos relacionados con identificadores de bloques, por ejemplo, bloque, bloq, blq, etc.
bloque	Para identificar la denominación del bloque. Puede ser un número, una letra o cualquier otro nombre.
identificador_de_portal	Para identificar elementos relacionados con identificadores del portal de una vivienda, por ejemplo, portal, port, etc.
portal	Para identificar el nombre o número del portal
identificador_de_escalera	Para identificar elementos relacionados con identificadores de escaleras, por ejemplo, escalera, esca, esc, etc.
escalera	Para identificar el nombre o número de la escalera
identificador_de_planta	Para identificar elementos relacionados con identificadores de plantas, por ejemplo, planta, plta, plant, etc.
planta	Para identificar el nombre o número de la planta
identificador_de_puerta	Para identificar elementos relacionados con identificadores de puerta, por ejemplo, puerta, prta, etc.
puerta	Para identificar el nombre o número de la puerta
identificador_de_letra	Para identificar elementos relacionados con identificadores de letras, por ejemplo, letra, letr, etc.
letra	Para identificar elementos asociados a la letra de la puerta de una vivienda o local
entidad_singular	Para identificar las entidades singulares de la Comunidad Autónoma andaluza
municipio	Para identificar los municipios de la Comunidad Autónoma andaluza

Estado	Utilidad
provincia	Para identificar las provincias de la Comunidad Autónoma andaluza
identificador_de_codigo_postal	Para identificar elementos relacionados con identificadores de códigos postales, por ejemplo, codigo postal, cp, etc.)
codigo_postal	Para identificar el valor numérico correspondiente al código postal
tipo_de_agrupacion	Para identificar conjuntos de construcciones no considerados como núcleos de población en el Nomenclátor del INE, que ahora mismo son: urbanizaciones, barrios, barriadas, polígonos industriales y parques comerciales, así como otros elementos que se considera que contienen una serie de vías, como por ejemplo aldeas, poblados, conjuntos, grupos, complejos, etc.
agrupacion	Para identificar la denominación de la agrupación
identificador_de_sector	Para identificar elementos relacionados con identificadores de sectores, por ejemplo, sector, sect, etc.
sector	Para identificar la denominación del sector
identificador_de_manzana	Para identificar elementos relacionados con identificadores de manzanas, por ejemplo, manzana, manz, etc.
manzana	Para identificar la denominación de la manzana
identificador_de_parcela	Para identificar elementos relacionados con identificadores de parcelas, por ejemplo, parcela, parc, etc.
parcela	Para identificar la denominación de la parcela
identificador_de_nave	Para identificar elementos relacionados con identificadores de naves, por ejemplo, nave, nav, etc.
nave	Para identificar la denominación de la nave
identificador_de_zona	Para identificar elementos relacionados con identificadores de zonas como parajes, jardines, parques, paseos, pagos, etc.
zona	Para identificar la denominación de la zona
odub	Para identificar todos aquellos elementos que no se han identificado por alguno de los estados anteriores

Tabla 19. Estados para direcciones postales

Anexo VIII: Métodos de suavizado

A la hora de construir el Modelo Oculto de Markov se debe tener en cuenta que se parte de una muestra aleatoria del conjunto de datos que vamos a normalizar. Por lo tanto, se pueden quedar fuera elementos cuya estructura sea diferente a los que se encuentran en la misma y por consiguiente las probabilidades de observación de esas etiquetas y estados asociados serán nulas.

Para solucionar este problema y que todas las etiquetas junto con sus estados asociados tengan una determinada probabilidad se utilizan los llamados MÉTODOS DE SUAVIZADO. En concreto se han implementado dos de las técnicas que ofrecen mejores resultados, que son el **suavizado de Laplace** y la **técnica de descuento absoluto** (en inglés, *absolute discounting*).

El suavizado de Laplace es un método de suavizado básico consistente en asignar una cierta probabilidad al espacio de sucesos no conocidos mediante la aplicación de la Ley de Laplace, también conocida como Añadir Uno (en inglés, *Adding One*) (Jeffreys, 1948). En este método se incrementa la frecuencia de todos los sucesos en una unidad y la probabilidad de observación se define como:

$$P^{suavizada}(e_k \mid c_i) = \begin{cases} \dfrac{f(e_k, c_i)+1}{f(e_i)+|V|} & \text{si } f(e_k, c_i) \text{ existe} \\[2ex] \dfrac{1}{f(e_i)+|V|} & \text{si } f(e_k, c_i) \text{ no existe} \end{cases}$$

donde V es el número de etiquetas que aparecen en el conjunta de entrenamiento, $f(e_k,ci)$ es la cantidad de veces que el estado e_k está etiquetado con ci y $f(e_k)$ es el número total de etiquetas que tiene asociado el estado e_k en el conjunto de entrenamiento.

En la técnica de descuento absoluto (*absolute discounting*) se sustrae un valor pequeño, digamos 'x', de la probabilidad de todas las etiquetas c_j conocidas en el estado j (probabilidad ≠ 0). Entonces se distribuye la probabilidad acumulada equitativamente entre los sucesos no conocidos. Así la probabilidad de una etiqueta no conocida es:

$$\frac{c_j x}{c - c_j}$$

donde 'c' es el número total de etiquetas, mientras que para una etiqueta conocida, su probabilidad será:

$$b_{jk}-x$$

donde b_{jk} es el cociente entre el número de veces que el estado 'k' tiene asociada la etiqueta 'j' y el total de estados asociados a la

etiqueta 'j'.

No hay ninguna teoría sobre cómo seleccionar el mejor valor para x, por lo que se ha decidido tomar el siguiente valor:

$$X = \frac{1}{f(e_k) + |V|}$$

donde V es el número de etiquetas que aparecen en el conjunto de entrenamiento y $f(e_k)$ es el número total de etiquetas que tiene asociado el estado e_k en el conjunto de entrenamiento.

Anexo IX: Listas de corrección

Lista de correccion para nombres de personas

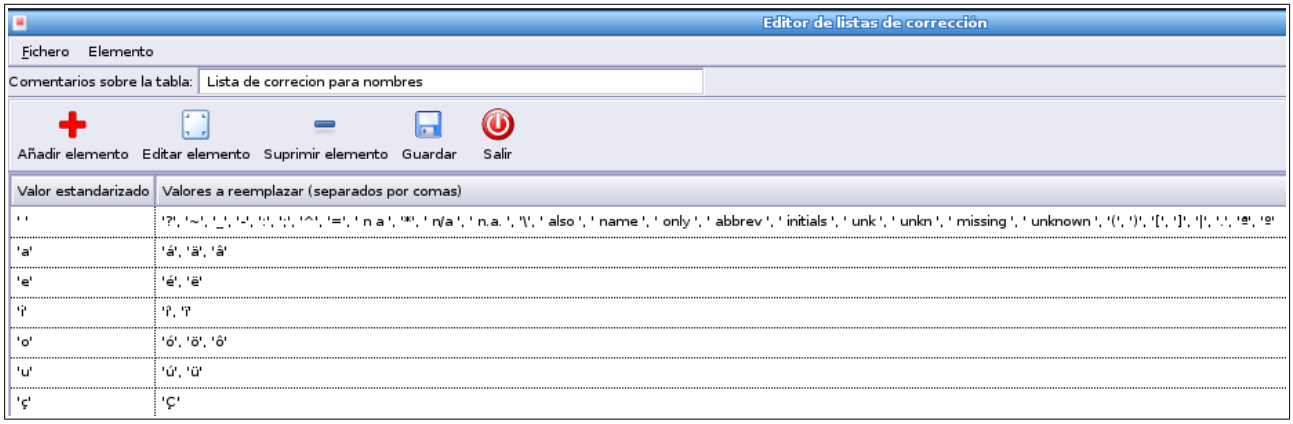


Imagen 154. Listas de corrección para nombres de personas

Lista de corrección para identificadores de personas físicas y/o jurídicas

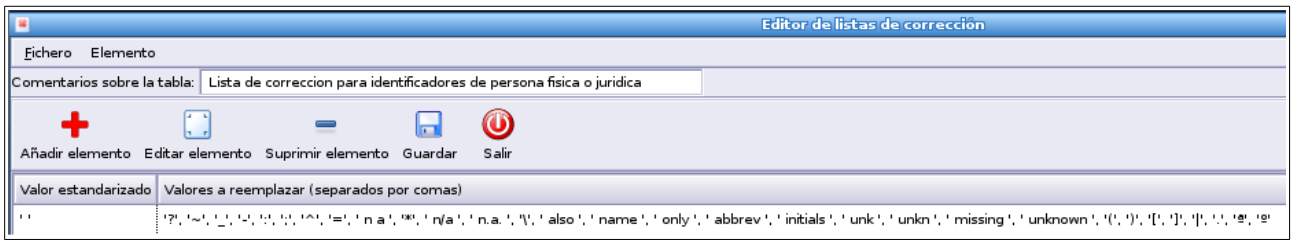


Imagen 155. Listas de corrección para identificadores de personas físicas y/o jurídicas

Anexo X: Tablas de búsqueda

Tablas de búsqueda para nombres de personas

Tabla de búsqueda	Etiqueta asociada	Descripción
knombres_femeninos.tbl	NF	Contiene valores asociados a un nombre femenino
knombres_masculinos.tbl	NM	Contiene valores asociados a un nombre masculino
knombres_neutros.tbl	NN	Contiene valores asociados a un nombre neutro
kparticulas.tbl	PS	Contiene valores asociados a partículas ligadas a nombres y/o apellidos

Tabla 20. Tablas de búsqueda para nombres de personas

Tablas de búsqueda para direcciones postales

Tabla de búsqueda	Etiqueta asociada	Descripción
kagrupacion.tbl	AG	Contiene valores asociados a conjuntos de construcciones no consideradas como núcleos de población en el Nomenclátor del INE tales como: barrios, barriadas, urbanizaciones, polígonos industriales y parques comerciales. Además, también contiene valores asociados a elementos referidos a la identificación de complejos, conjuntos o grupos, como por ejemplo, complejos hoteleros, residenciales, etc.
kbloque.tbl	BL	Contiene valores asociados a elementos referidos a la identificación de un bloque (bloque, blq, bl, etc.)
kcodigo_postal.tbl	CP	Contiene valores asociados a elementos referidos a la identificación de códigos postales (codigo postal, CP, C.P., etc.)
kedificio.tbl	ED	Contiene valores asociados a elementos referidos a tipos de edificios, como edificio, casa, finca, chalet, caserío, cortijo, etc. También contiene valores asociados a edificios singulares tales como como ayuntamientos, bibliotecas, aeropuertos, puertos, estaciones de ferrocarril o de autobuses, colegios, institutos, etc., así como a comercios, por ejemplo, mercados, plazas de abastos, supermercados, centros comerciales, gasolineras, hoteles, campings, etc.
kentidad_singular.tbl	EG	Contiene valores asociados a las entidades singulares existentes en la Comunidad Autónoma andaluza
kescalera.tbl	ES	Contiene valores asociados a elementos referidos a la identificación de la escalera de un bloque o edificio (escalera, esc, esca, etc.)

Tabla de búsqueda	Etiqueta asociada	Descripción
letra.tbl	LE	Contiene valores asociados a elementos referidos a la identificación de la letra de la puerta de una vivienda (letra, letr)
kmunicipio.tbl	MU	Contiene valores asociados a los municipios de la Comunidad Autónoma andaluza
kmanzana.tbl	MZ	Contiene valores asociados a elementos que identifican manzanas (manzana, mzna)
knumero_local.tbl	NM	Contiene valores asociados a elementos que identifican el número de la vía, local, punto kilométrico etc. (numero, nº, num, local, sin número, s/n, kilometro, km, pk, etc.)
kplanta_numero.tbl	NP	Contiene valores asociados a elementos que identifican el número de la planta de un bloque o edificio (1º, 2º, 3º, ático, bajo, etc.)
knave.tbl	NV	Contiene valores asociados a elementos que identifican naves industriales (nave, nav)
kparcela.tbl	PA	Contiene valores asociados a elementos que identifican parcelas (parcela, parc)
kplanta.tbl	PL	Contiene valores asociados a elementos que identifican la planta de un bloque o edificio (planta, plt, plnt, etc.)
kprovicia.tbl	PR	Contiene valores asociados a las ocho provincias de la Comunidad Autónoma andaluza
kportal.tbl	PT	Contiene valores asociados a elementos que identifican a un portal (portal, prtal, ptal, etc.)
kpuerta.tbl	PU	Contiene valores asociados a elementos que identifican la puerta de una vivienda (puerta, prta, pu, puert, etc.)
ksector.tbl	ST	Contiene valores asociados a elementos identificativos de sectores (sector, sect)
kvia.tbl	TV	Contiene valores asociados a elementos identificativos del tipo de vía (calle, c/, avenida, avda, plz, carretera, etc.)
kzona.tbl	ZO	Contiene valores asociados a elementos que identifican zonas tales como parques, jardines, paseos, parajes, arboledas, pagos, lugares, etc.

Tabla 21. Tablas de búsqueda para direcciones postales

Tabla de búsqueda para identificadores de personas físicas y/o jurídicas

Tabla de búsqueda	Etiqueta asociada	Descripción
kidpersona.tbl	TC	Contiene valores sobre las claves que indican la forma jurídica o tipo de entidad

Tabla 22. Tabla de búsqueda para identificadores de personas físicas y/o jurídicas

Anexo XI: Métodos de agrupación y proceso "full index"

BlockingIndex

Agrupa los registros en función de los distintos valores de la variable de agrupación. Por ejemplo, si la variable de agrupación elegida para los ficheros A y B es el nombre de pila de la persona, los grupos formados podrían ser:

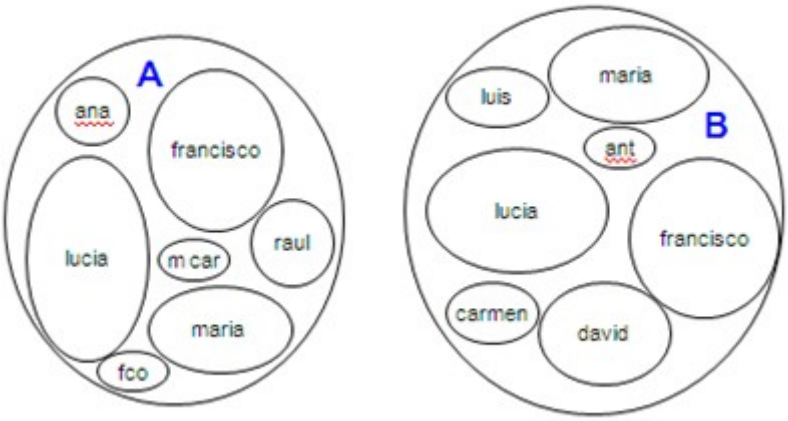


Imagen 156. Agrupación *Blocking Index*

Donde el grupo del fichero A denominado 'lucia' contendrá todos aquellos registros cuyo nombre de pila sea Lucía, el denominado 'francisco' contendrá todos aquellos registros cuyo nombre de pila sea Francisco, y así sucesivamente para el resto de grupos y para el fichero de datos B.

SortingIndex

Ordena los registros alfabéticamente, y a continuación se inspeccionan mediante una ventana de tamaño fijo w, de modo que se comparan aquellos pares que están incluidos dentro de la ventana. La ventana se va desplazando hasta recorrer todos los registros. Nótese que cuando el tamaño de la ventana es 1, ambos métodos coinciden.

A continuación se puede ver gráficamente cómo funciona este método:

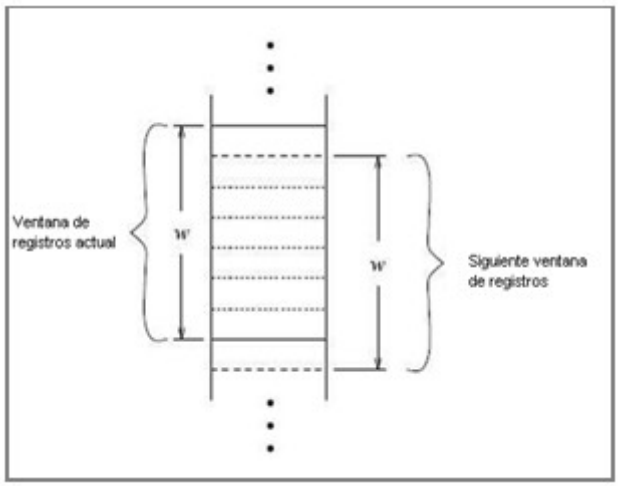


Imagen 157. Agrupación *Sorting Index*

FullIndex

Con esta opción se comparan los registros de un fichero con todos los registros del otro fichero. El motivo de su presencia en la aplicación se debe a que permite obtener resultados los cuales podrían ser comparados posteriormente con resultados en los que sí se ha usado alguno de los procedimientos de agrupación. Hay que aclarar que este método de agrupación sólo se utilizará cuando se trabaje con ficheros de datos de tamaño pequeño, ya que si no esta fase sería muy costosa tanto desde el punto de vista computacional como del temporal, por la imposibilidad de reducir el número de comparaciones a realizar.

Anexo XII: Funciones de comparación

A continuación, se realiza una breve descripción de cada una de las funciones de comparación implementadas en *aLink: Herramienta de Fusión de Ficheros*. Para una descripción más detallada de las mismas consultar [3].

Función de comparación de cadena exacta (Str-Exact)

Compara los valores de la variable y si son iguales se devuelve el valor que se haya establecido para una coincidencia exacta o acuerdo total. Dicho valor se denomina *peso de coincidencia* o *peso de acuerdo* y por defecto en la Herramienta de Enlace será 1. Si los valores son distintos se devolverá el valor de desacuerdo establecido o también denominado *peso de desacuerdo* o *de no coincidencia*, que por defecto en la Herramienta de Enlace será 0.

Si alguno o ambos de los valores comparados es un valor perdido, es decir, el campo comparado está vacío, se devolverá el peso establecido para un valor perdido. Por defecto, en la Herramienta de Enlace es 0.

Función de comparación de cadena contenida (Str-Contains)

Comprueba si la cadena más corta de las dos cadenas a comparar está totalmente contenida en la más larga. Si es así, se devuelve el peso establecido para una coincidencia (por defecto, en la Herramienta de Enlace, 1), en otro caso se devuelve el peso de desacuerdo (por defecto, en la Herramienta de Enlace, 0).

Si alguno o ambos de los valores comparados es un valor perdido se devolverá el peso establecido para un valor perdido. Por defecto, en la Herramienta de Enlace, dicho valor es 0.

Función de comparación de cadena truncada (Str-Truncate)

Compara un número determinado de caracteres iniciales de las cadenas a comparar, mediante la función comparación de cadena exacta. La determinación del número de caracteres iniciales a comparar se lleva a cabo estableciendo el parámetro 'Número de caracteres a verificar'. Si los caracteres comparados coinciden se devuelve el peso establecido para una coincidencia (por defecto , en la Herramienta de Enlace, 1) y si no coinciden se devuelve el peso de desacuerdo establecido (por defecto, en la Herramienta de Enlace, 0).

Si alguno o ambos de los valores comparados es un valor perdido se devolverá el peso establecido para un valor perdido. Por defecto, en la Herramienta de Enlace, dicho valor es 0.

Función de comparación de cadena aproximada de Jaro (Jaro)

Esta función, como todas las siguientes funciones de comparación de cadenas aproximadas, calculá un valor de similaridad entre 0.0 (cadenas distintas) y 1.0 (cadenas iguales), el cual se encuentra dentro de un rango establecido por el usuario mediante los parámetros peso de coincidencia y peso de no coincidencia. Además, todas las funciones de comparación aproximadas tienen asociado un valor umbral (Threshold) que tomará un valor entre 0.0 y 1.0, de forma que si el valor de similaridad calculado es mayor que el valor umbral establecido, entonces la función de comparación devolverá un peso de coincidencia parcial calculado

mediante la siguiente expresión:

$$Peso_coincidencia_parcial = peso_coincidencia - \frac{1 - val_cad_aprox}{1 - \text{umbral}} * (peso_coincidencia + |peso_nocoincidencia|)$$

donde *val_cad_aprox* es el valor de similaridad devuelto por el comparador de cadena aproximada.

En cambio, si el valor de similaridad calculado es menor que el umbral, entonces la función de comparación devolverá el peso de desacuerdo.

En concreto, el algoritmo de comparación de Jaro calcula el valor de similitud de la siguiente manera: este contabiliza el número de inserciones, eliminaciones y transposiciones que se llevan a cabo para que una cadena sea lo más similar a otra. Dicho algoritmo calcula el número de caracteres comunes en ambas cadenas y el número de trasposiciones, entendiendo por caracteres comunes aquellos que coinciden y que ocupan en ambas cadenas la misma posición o se encuentran como mucho a una distancia inferior a la mitad de la longitud de la cadena más larga.

Para más detalles sobre la función de comparación de Jaro véase [3].

Comparación de cadena aproximada de Winkler (Winkler)

También se denomina función de *Jaro-Winkler*. Es una mejora de la función de comparación de Jaro y se basa en la idea de que con frecuencia se comenten más errores tipográficos al final de las cadenas de caracteres, dando un mayor peso a los caracteres que coinciden al inicio de las cadenas, considerándose como máximo hasta cuatro caracteres iniciales.

Esta función incluye los tres parámetros siguientes:

- Comprobación de caracteres similares: comprueba si existen pares de caracteres similares, tales como 'a' y 'e', 'i' y 'j', 's' y 'z'..., de manera que aumenta el valor de similaridad si tales caracteres son encontrados en las cadenas de entrada.
- Comprobación de caracteres iniciales iguales: aumenta el valor de similaridad cuando dos cadenas tienen los mismos caracteres iniciales (hasta 4).
- Comprobación de cadenas largas: esta opción permite comprobar más caracteres coincidentes en cadenas largas y modifica el valor de similaridad en consecuencia.

Comparación de cadena aproximada con la distancia de edición (Edit-Dist)

Se basa en la *distancia de edición* o *Levenshtein*. La distancia de edición entre dos cadenas es el número mínimo de operaciones requeridas para transformar una cadena en otra. Se entiende por operación una inserción, eliminación o sustitución de un carácter.

Así, calculada la distancia de edición entre dos cadenas con longitudes l_1 y l_2, se calculará un valor de similaridad entre 0.0 y 1.0 mediante la siguiente expresión:

$$Sim = 1 - \frac{dis\tan cia\ \ edición}{\max(l_1\ ,\ l_2)}$$

Para esta función de comparación no es necesario establecer parámetros específicos, por lo que solo es necesario establecer el valor umbral y los valores de los pesos de coincidencia, no coincidencia y el de valor perdido.

Comparación de cadena aproximada con la distancia de Damerau-Levenshtein (Dam_Le_Edit-Dist)

Esta función de comparación es similar a la anterior, con la única diferencia de que las trasposiciones son contadas como una operación elemental en vez de como dos (una inserción y una eliminación). El valor de similaridad es calculado de la misma manera que para la distancia de edición. No se especifica ningún parámetro. Al igual que en el caso anterior, tampoco necesita que se establezcan parámetros específicos.

Comparación de cadena aproximada con la distancia Bag (Bag-Dist)

El inconveniente de la distancia de edición y de la de Damerau-Levenshtein es que son de complejidad cuadrática en la longitud de las dos cadenas que va a ser comparadas. Por ejemplo, para dos cadenas de longitud l_1 y l_2, se tienen que realizar $l_1 \times l_2$ cálculos. Esto se convierte en un problema cuando el conjunto de datos contiene cadenas muy largas.

La distancia Bag es un método barato para calcular la distancia entre dos cadenas. Puede ser usada como una aproximación de la distancia de edición, ya que siempre es menor o igual que esta y por tanto la medida de similaridad calculada por la distancia Bag es siempre igual o mayor que la medida de similaridad obtenida con la distancia de edición.

Para esta función de comparación el usuario no necesita indicar ningún parámetro específico y si desea más información sobre la misma puede consultar [2] y [3].

Comparación de cadena aproximada con la distancia Smith-Waterman (Smith-Water-Dist)

Esta función de comparación se basa en la distancia de Smith-Watermann, la cual se usa comúnmente en la alineación local de secuencias biológicas, como ADN (Ácido desoxiribonucleico) o ARN (Ácido ribonucleico), es decir determina si dos secuencias biológicas tienen algún fragmento en común, o son de una gran similitud biológica teniendo en cuenta las posibles mutaciones, inserciones o eliminaciones de elementos dentro de una cadena. Para más información sobre ella consultar [3]. Nótese que esta función de comparación es de complejidad cúbica (n^3) en la longitud de las cadenas que van a ser comparadas.

Comparación de cadena aproximada Seq-Match (Seq-Match)

Se basa en la función de Python ***SequenceMatcher*** la cual está disponible en el módulo ***difflib***. Para más detalles sobre este módulo se puede consultar la url:

http://www.python.org/doc/current/lib/module-difflib.html

Esta función de comparación aproximada no requiere de ningún parámetro específico, a parte del parámetro umbral (***Threshold***) requerido para las mismas.

Comparación de porcentaje numérico (Num-Perc)

Compara campos numéricos tolerando una diferencia en porcentaje dada. Una vez calculada, devuelve el peso de coincidencia si los números son los mismos, y el peso de no coincidencia si la diferencia de porcentaje entre los dos números es más grande que el valor del máximo porcentaje tolerado. En caso de que cualquiera de los valores comparados no sean numéricos también se devolverá el peso de no coincidencia.

Para indicar el porcentaje de tolerancia permitido tendremos que establecer el parámetro 'Diferencia máxima de porcentaje', que será un valor entre 0 y 100. Si es establece a 0, la función de comparación se reduce a una comparación numérica exacta.

La diferencia de porcentaje entre dos valores numéricos, valor1 y valor2, se calcula como:

$$diferencia_porcentaje\,f = 100 * \frac{|\,valor1 - valor\,2\,|}{\max(|\,valor1\,|, |\,valor\,2\,|)}$$

Si este valor calculado es menor que el máximo porcentaje permitido, entonces se calcula un peso de coincidencia parcial de acuerdo a la siguiente fórmula:

$$Peso_coincidencia_parcial = peso_coincidencia - \frac{diferencia_porcentaje\,f}{\max_diferencia_porcentaje\,f + 1} * (peso_coincidencia + |\,peso_nocoincidencia\,|)$$

Comparación numérica absoluta (Num-Abs)

Esta función va a comparar campos numéricos de forma que se tolerará una diferencia numérica absoluta dada. Devolverá el peso de coincidencia si los números comparados son los mismos y el peso de no coincidencia si la diferencia absoluta entre los dos valores es más grande que el parámetro 'Diferencia máxima absoluta' establecido por el usuario. El peso de no coincidencia también será devuelto si cualquiera de los valores introducidos no son numéricos.

Si la diferencia absoluta es igual o menor que la máxima diferencia permitida, entonces se devolverá un peso de coincidencia parcial calculado mediante la expresión:

$$Peso_coincidencia_parcial = peso_coincidencia - \frac{|\,valor1 - valor\,2\,|}{\max\,|\,dif\,f\,| + 1} * (peso_coincidencia + |\,peso_nocoincidencia\,|)$$

Función de comparación Key-diff (Key-Diff)

Compara valores de tipo cadena y valores numéricos del tipo: números de teléfono o códigos postales. Cuenta el número de caracteres distintos entre ambos valores estableciendo un número máximo de caracteres diferentes a tolerar. Por este motivo es necesario indicar un número máximo de caracteres diferentes que se pueden tolerar (parámetro 'Máxima diferencia de caracteres'). Si el número de caracteres diferentes es mayor que el máximo especificado, entonces se devolverá el peso de no coincidencia y si es menor se utilizará la siguiente expresión para calcular un peso parcial de coincidencia:

$$Peso_coincidencia_parcial = peso_coincidencia - \frac{f}{\max f + 1} * (peso_coincidencia + |peso_nocoincidencia|)$$

donde *f* es el número de caracteres diferentes entre dos valores, y *max f* es el número máximo de caracteres diferentes tolerado.

Anexo XIII: Métodos de clasificación

Clasificador basado en la metodología de Fellegi y Sunter (FellegiSunter)

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 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, es decir, de representar 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 representar a pares de registros que no corresponden a la misma entidad.

En este sentido, el clasificador de dos pasos se basa en la idea de seleccionar automáticamente en un primer paso aquellos vectores de pesos que con una alta probabilidad van a dar lugar a *verdaderos enlaces* y a *verdaderos no_enlaces*. Estos vectores formarán dos conjuntos de entrenamiento que posteriormente se utilizarán para clasificar los pares de registros comparados mediante alguno de los métodos de clasificación supervisados implementados en la Herramienta de Enlace (máquina vector soporte y k-medias). La elección del número de vectores que formará parte de cada conjunto de entrenamiento se lleva a cabo por parte del usuario. Christen [5] propone el uso de una estimación de la razón de enlaces sobre los no-enlaces, ***r***, cuyo valor se calcula mediante la expresión:

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

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 a realizar en función de la variable de agrupación utilizada. Por ejemplo, si r=0.05 entonces por cada 100 no enlaces habrá 5 enlaces.

10 GLOSARIO

Término	Descripción
Proceso de normalización	Conjunto de técnicas que transforman los datos originales brutos en otros con formatos consistentes y corrigen las posibles inconsistencias sobre como se representa y codifica la información. Tiene dos fases principales: limpieza y estandarización y segmentación. Se lleva a cabo con la Herramienta de Normalización de *aLink: Herramienta de Fusión de Ficheros*
Proceso de enlace de registros	Consiste en detectar aquellos registros de dos ficheros de datos que corresponden a una misma entidad o unidad poblacional (individuos, establecimientos, etc.). Se lleva a cabo con la Herramienta de Enlace de *aLink: Herramienta de Fusión de Ficheros*
Proceso de fusión de ficheros	Proceso que engloba las etapas de normalización y enlace de registros de un fichero de datos. Se lleva a cabo con *aLink: Herramienta de Fusión de Ficheros*
Estado	Valor que se asigna manualmente por el usuario a cada uno de los elementos del campo a normalizar
Etiqueta	Valor que se asigna automáticamente por la Herramienta de Normalización a los valores del campo a normalizar

11 BIBLIOGRAFÍA

[1]	P. Christen. Febrl-Freely extensible biomedical record linkage. Release0.4.1
[2]	I. Bartolini, P. Ciaccia, and M. Patella. String matching with metric trees using an approximate distance. In SPIRE, LNCS 2476, pages 271–283, Lisbon, Portugal, 2002.
[3]	P. Christen. A comparison of personal name matching: Techniques and practical issues. In Workshop on Mining Complex Data (MCD), held at IEEE ICDM'06, Hong Kong, 2006
[4]	P. Christen. Febrl – A freely available record linkage system with a graphical user interface. In Australasian Workshop on Health Data and Knowledge Management (HDKM'08), CRPIT. 80, Wollongong, Australia, 2008.
[5]	P. Christen and K. Goiser. Quality and complexity measures for data linkage and deduplication. In F. Guillet and H. Hamilton, editors, Quality Measures in Data Mining, volume 43 of Studies in Computational Intelligence, pages 127–151. Springer, 2007.
[6]	Manual de buenas prácticas para la normalización de fuentes y registros administrativos de la Junta de Andalucía. Instituto de Estadística y Cartografía de Andalucía. 2013.