Seminario sobre GeoKettle

Segunda actividad

Enunciado del problema.

Se dispone de ***varios archivos de texto con el mismo formato de datos***. Se desea ***fundir en un solo archivo*** todo el contenido eliminando la primera línea que contiene los nombres de los campos (columnas) en todos los archivos.

Solución inmediata:

Utilizar un editor de texto y quitar de todos los archivos excepto del primero la primera fila. Una vez hecho esto usar la orden del sistema operativo que permite concatenar todos los archivos en uno de salida. Si los archivos iniciales tiene la extensión csv, por ejemplo, y el de salida txt, la orden podría ser la siguiente: COPY *.csv salida.txt

Requiere por tanto abrir uno a uno los archivos iniciales y eliminar la primera línea excepto en el primero y luego ejecutar la orden COPY.

Solución basada en GeoKettle:

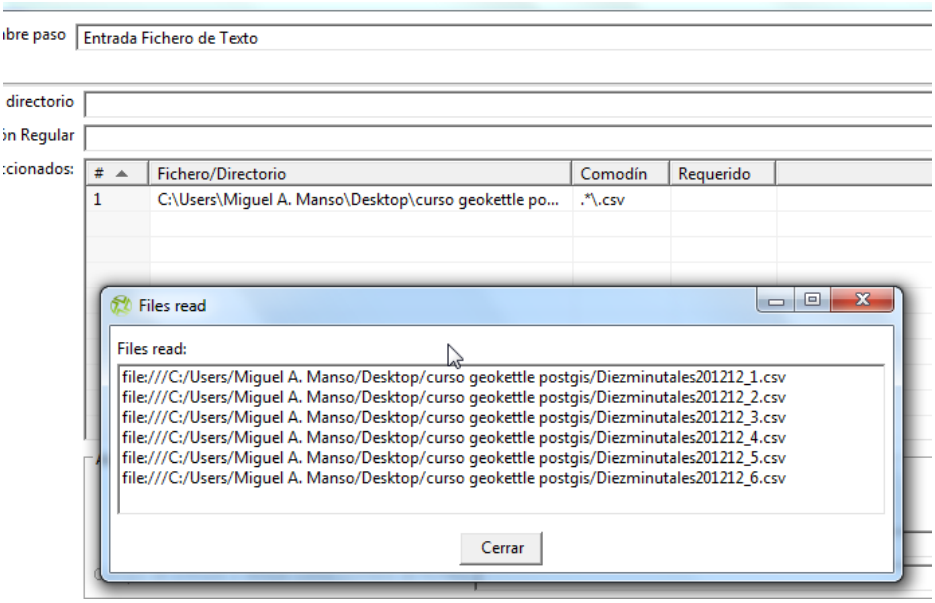
Crear una transformación simple que abra de forma secuencial todos los archivos de un directorio con la extensión seleccionada (CSV) y que genere un archivo de salida único en el que aparecerá solo una vez la línea con los nombres de las columnas.

En la primera etapa de la transformación, se ha de indicar el patrón de la extensión de los archivos a procesar. Esto se realiza con una ***expresión regular*** como la siguiente: .*\.CSV.

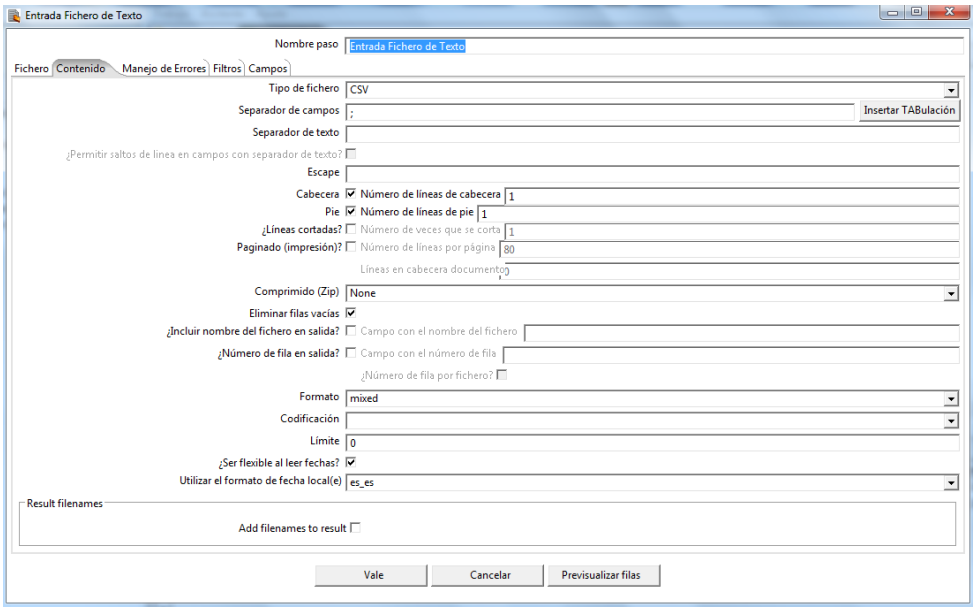
Un pequeño tutorial de expresiones regulares puede verse aquí:

http://misc.yarinareth.net/regex.html

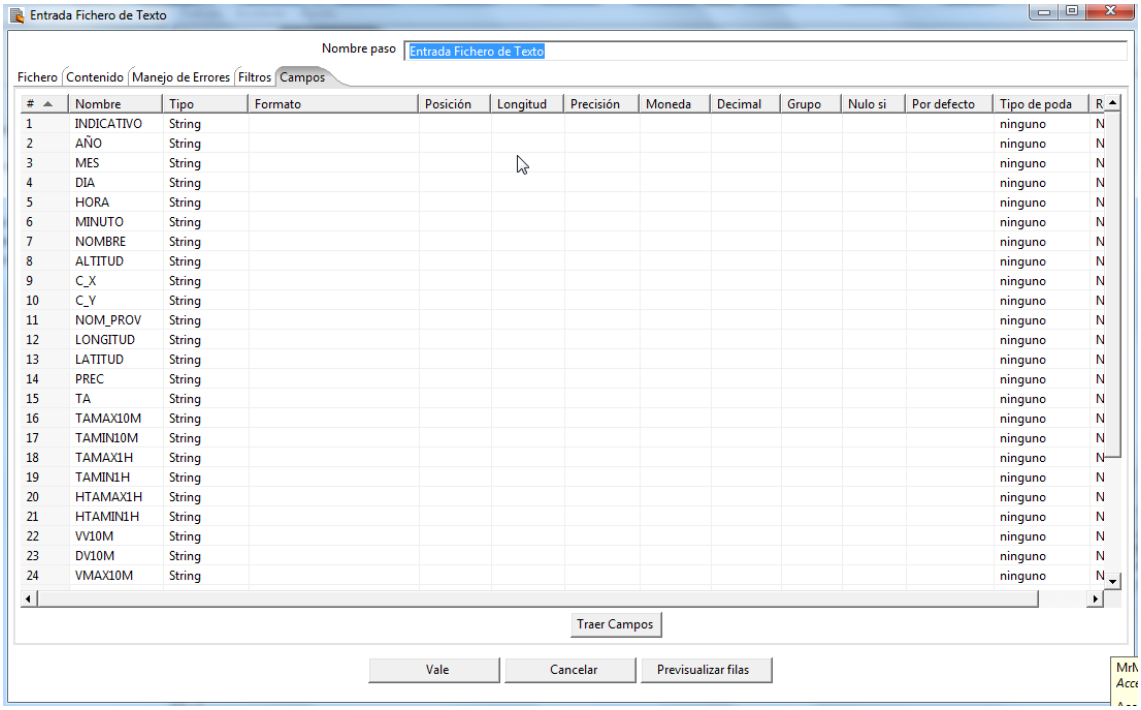
Después seleccionar el directorio donde se encuentran los archivos con el botón ***Examinar..*** Finalmente añadir dicha entrada en la lista (Añadir). Hecho esto se puede comprobar si se listan los archivos con la extensión seleccionada en el directorio seleccionado. Se hace con el botón ***Mostrar Fichero(s)***.



Dentro de la misma entapa o entrada de la transformación en la solapa ***Contenido***, se puede indicar cuál es el formato del archivo, si hay cabecera y pie en el archivo, etc. Ver ilustración para indicar que hay una línea de cabecera y pie por archivo.



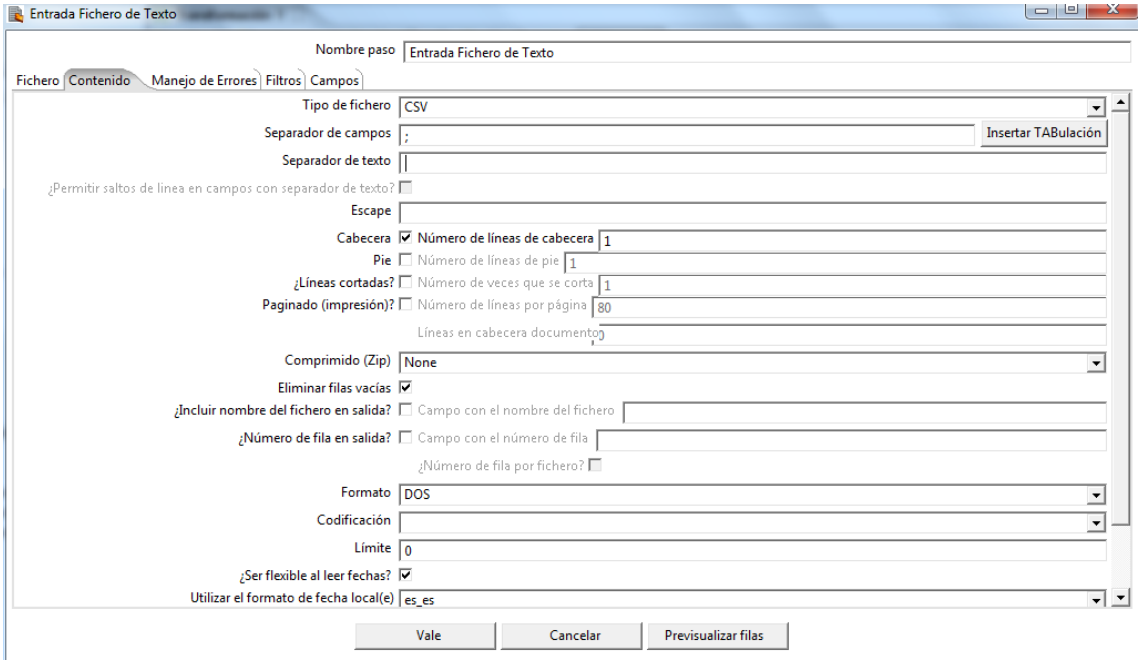
También con la pestaña ***Campos*** se puede hacer que la herramienta nos detecte los campos y los mapee automáticamente. Esto se hace con el botón ***Traer Campos***. Sin embargo para evitar que los transforme a formato entero, fecha, etc. es mejor que al aparecer la ventana emergente al Traer Campos se diga cancelar. Debería de aparecer del siguiente modo los campos.



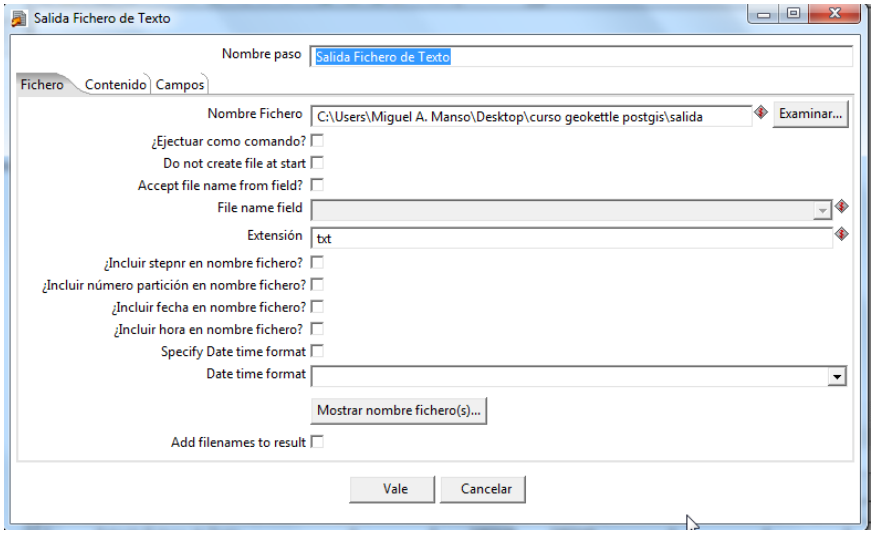
#	Nombre	Tipo	Formato	Posición	Longitud	Precisión	Moneda	Decimal	Grupo	Nulo si	Por defecto	Tipo de poda	R
1	INDICATIVO	String										ninguno	N
2	AÑO	String										ninguno	N
3	MES	String										ninguno	N
4	DIA	String										ninguno	N
5	HORA	String										ninguno	N
6	MINUTO	String										ninguno	N
7	NOMBRE	String										ninguno	N
8	ALTITUD	String										ninguno	N
9	C_X	String										ninguno	N
10	C_Y	String										ninguno	N
11	NOM_PROV	String										ninguno	N
12	LONGITUD	String										ninguno	N
13	LATITUD	String										ninguno	N
14	PREC	String										ninguno	N
15	TA	String										ninguno	N
16	TAMAX10M	String										ninguno	N
17	TAMIN10M	String										ninguno	N
18	TAMAX1H	String										ninguno	N
19	TAMIN1H	String										ninguno	N
20	HTAMAX1H	String										ninguno	N
21	HTAMIN1H	String										ninguno	N
22	VV10M	String										ninguno	N
23	DV10M	String										ninguno	N
24	VMAX10M	String										ninguno	N

Como puede apreciarse en la ilustración, el juego de datos de prueba contiene muchas columnas y cada una de ellas de un tipo de datos y longitudes. El problema de hacer que se identifique los nombres de las columnas y se mapeen en campos individuales es que posteriormente hay que concatenarlos para escribirlos en el único archivo de salida. Por esta razón se descarta esta opción. Para eliminar las filas, se las selecciona todas y con el botón derecho sobre el menú contextual emergente se selecciona ***Borrar filas seleccionadas***.

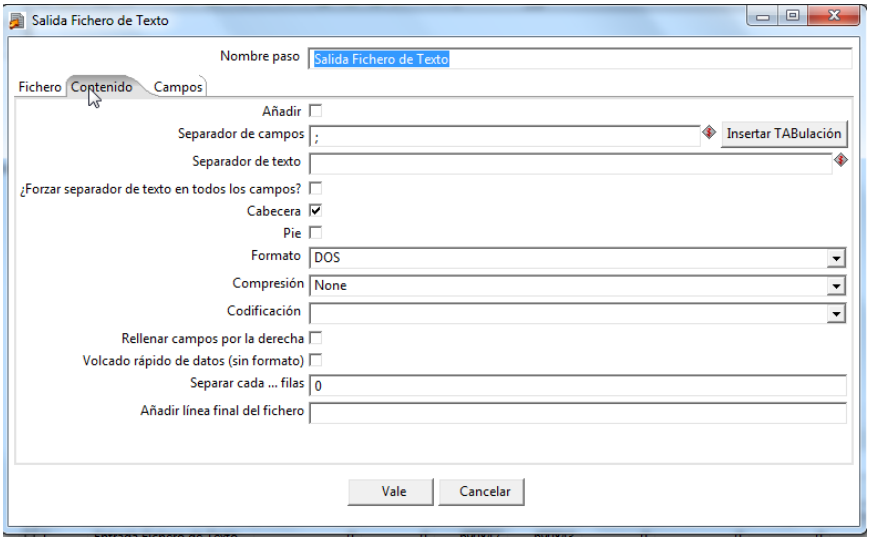
Para eliminar la primera línea con la cabecera de cada archivo accedemos a la pestaña Contenido. En ella podemos seleccionar el formato del archivo (CSV), el separador entre columnas, el tipo de codificación de los caracteres, y más opciones. La que nos interesa marcar es ***Cabecera, Nº de filas de cabecera***, indicando que es 1. En nuestro caso, los archivos no estaban comprimidos, si lo hubieran estado se podría haber indicado el tipo de compresión (zip) por ejemplo para que lo descomprima, lo lea y lo procese.



En último lugar se guardar los datos en el archivo de salida. Ver ilustración. En la solapa ***Fichero***, se ha de indicar el lugar y nombre del archivo a generar. Importante si se usa el mismo directorio de los datos de entrada que el archivo de salida no tenga la misma extensión porque se produciría una especie de bucle ya que el archivo de salida que se va generando formaría parte de la entrada.



En la solapa Contenido, se indican algunas características del contenido, como el tipo de separador a usar, si se entrecomillan los textos, si se pone cabecera y pie en el archivo, si se incluye la fecha y hora en el nombre del archivo, etc… ***Ver ilustración***.



Finalmente, en los campos, seleccionando ***Traer campos***, nos aparecerán todos los campos que teníamos en el paso anterior, con el tipo de datos que habíamos definido “String”.