Memoria de seguimiento

Beñat Agirre Arruabarrena
b.agirre@alumnos.upm.es

30 de abril de 2021

Índice

1. Resumen del trabajo realizado

1. Analizar el funcionamiento de las herramientas de Linked Data Geográfico utilizadas por el Grupo de Ingeniería Ontológica y determinar si existen nuevas alternativas mejores.
2. Preparar el nuevo entorno de desarrollo con el SDK de PDI 9 y crear el fichero pom.xml de maven para gestión automática de dependencias de tripleGeoKettle.
3. Comenzar a portar tripleGeoKettle y la toolchain de transformaciones desde GEOKettle a PDI 9.
4. Escribir primeros apartados de la memoria

2. Explicación y justificación de las modificaciones al Plan de Trabajo

Desde que se publicó "A sustainable process and toolbox for geographical linked data generation and publication: a case study with BTN100" en 2019, GeoKettle ha dejado de estar soportado. La pagina oficial y de documentación ya no están disponibles. Un objetivo de este TFG era dar soporte GeoPackage a GeoKettle. No tiene sentido desarrollar soluciones de "modernización" sobre software abandonado. Por tanto, se ha optado por actualizar la herramienta.

Algunas funcionalidades de GeoKettle se integraron en PDI directamente y otras desaparecieron. Actualmente, el soporte GIS de Pentaho está dentro de PDI Spoon y además hay algunas funcionalidades más en el plugin llamado pentaho-gis-plugins. Se actualizará la fase de añadir soporte GeoPackage a GEOKettle a: "replicar la funcionalidad y las transformaciones de GeoKettle y TripleGeo en la suite PDI." Si es sencillo, se considerará también dar soporte a GeoPackage.

2.1. Revisión de la lista de objetivos del trabajo

- Replicar la funcionalidad y las transformaciones de GeoKettle y TripleGeo en la suite PDI
- Dar soporte GeoPackage a la herramienta GeoKettle y su plugin para transformar a RDF
- Realizar un procesado completo de todos los datos del IGN para generar este tipo de formato.

2.2. Revisión de la lista de tareas

La lista de tareas no ha cambiado ya que era bastante general.

- Análisis del formato GeoPackage y herramientas asociadas: 20 %
- Diseño de soluciones: 20 %
- Implementación de soluciones: 40 %
- Evaluación: 10 %
- Documentación: 10 %

2.3. Revisión del Diagrama de Gantt

El diagrama de Gantt tampoco cambia porque no ha cambiado la lista de tareas.

DIAGRAMA DE GANTT

INICIO	08/03/2021
FIN	31/05/2021

ESFUERZO EN HORAS	324
SEMANAS DISPONIBLES	12
HORAS/SEMANA	27.0
HORAS/DIA	3.9

TAREAS	INICIO	FIN	PORCENTAJE	DURACION (h)
Análisis del formato GeoPackage y herramientas asociadas	8/3/2021	3/30/2021	20%	64.8
Diseño de soluciones	20/3/2021	12/4/2021	20%	64.8
Implementación de soluciones	10/4/2021	23/5/2021	40%	129.6
Evaluación	08/03/2021	31/05/2021	10%	32.4
Documentación	8/3/2021	31/05/2021	10%	32.4

MARZO: 8 9 10 11 12 13 14 15 16 17 18 19 20 21 22 23 24 25 26 27 28 29 30 31
M T W R F S S M T W R F S S M T W R F S S M T W R

ABRIL: 1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 21 22 23 24 25 26 27 28 29 30
T W R F S S M T W R F S S M T W R F S S M T W R F S

MAYO: 1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 21 22 23 24 25 26 27 28 29 30 31
S S M T W R F S S M T W R F S S M T W R F S S M T W R F S

3. Borrador de la Memoria Final

A continuación se adjunta el borrador completo de la memoria final:

Universidad Politécnica de Madrid

Escuela Técnica Superior de Ingenieros Informáticos

Grado en Ingeniería Informática

Trabajo Fin de Grado

Actualizar las Herramientas de Linked Data Geográfico utilizadas por el Grupo de Ingeniería Ontológica

Autor: Beñat Agirre Arruabarrena
Tutor(a): Oscar Corcho García

Madrid, Mayo 2021

Este Trabajo Fin de Grado se ha depositado en la ETSI Informáticos de la Universidad Politécnica de Madrid para su defensa.

Trabajo Fin de Grado
Grado en Ingeniería Informática

Título: Actualizar las Herramientas de Linked Data Geográfico utilizadas por el Grupo de Ingeniería Ontológica

Mayo 2021

Autor: Beñat Agirre Arruabarrena
Tutor: Oscar Corcho García
Departamento de Inteligencia Artificial
ETSI Informáticos
Universidad Politécnica de Madrid

Resumen

El Ontology Engineering Group lleva más de una década trabajando con datos geográficos enlazados españoles. En 2010[2] se definió un caso de uso y en 2019[3] se refinó el proceso de generación y publicación de los datos abiertos utilizando el dataset BTN100 como caso de estudio. En los últimos años han se han popularizado nuevas herramientas y formatos que ofrecen ventajas no disponibles en las usadas hasta el momento. Entre ellas se encuentran el programa de transformaciones de datos PDI9 Kettle, que reemplaza a GEOKettle; el formato Geopackage, más práctico que el shapefile; y Apache Maven, más extenso que Apache Ant. Por consiguiente, también será necesario actualizar las herramientas propias desarrolladas por el OEG (tripleGeoKettle, Map4RDF) para integrarlas en el nuevo toolbox.

Abstract

The Ontology Engineering Group has been working with Spanish geographical linked data for over a decade. In 2010[2] a use case was defined, and in 2019[3] the open data generation and publication process was refined in a case estudy with BTN100. New tools and formats have gained popularity over the past few years, which offer more advantages over the ones used to date. This includes the data transformation tool PDI9 Kettle, which replaces GEOKettle; Geopackage, which is more practical than the Shapefile format; and Apache Maven, more fully-featured than Apache Ant. Thus, the related tools developed by the OEG, namely tripleGeokettle and Map4RDF, will also need an update.

Tabla de contenidos

Capítulo 1

Introducción

1.1. Objetivos

El objetivo principal del trabajo es modernizar las herramientas de Linked Data Geográfico desarrolladas por el Grupo de Ingeniería Ontológica. En el OEG se ha venido tradicionalmente trabajando con el Instituto Geográfico Nacional para la exportación de algunos de sus datos geográficos a formato Linked Data. Un ejemplo se puede encontrar en la web del Instituto Geográfico Nacional. [7]

Desde que se publicó "A sustainable process and toolbox for geographical linked data generation and publication: a case study with BTN100" en 2019[3], GeoKettle ha dejado de estar soportado. La pagina oficial y de documentación ya no están disponibles. Algunas funcionalidades de GeoKettle se integraron en PDI directamente y otras desaparecieron. Actualmente, el soporte GIS de Pentaho está dentro de PDI Spoon y además hay algunas funcionalidades más en el plugin llamado pentaho-gis-plugins[25].

Por otro lado, recientemente, el Open Geospatial Consortium ha publicado el formato GeoPackage, que tiene el objetivo de convertirse en un estándar para la representación de datos geográficos. El siguiente objetivo de este trabajo es el de dar soporte GeoPackage para las herramientas normalmente utilizadas para este tipo de tareas. Ya que el 3 de abril de 2021 pentaho-gis-plugins añadió soporte geopackage a su plugin, parte del trabajo está hecho.

En resumen:

1. Replicar la funcionalidad y las transformaciones de GeoKettle + TripleGeo en la nueva suite PDI.

2. Dar soporte GeoPackage a la herramienta GeoKettle y su plugin para transformar a RDF.

3. Realizar un procesado completo de todos los datos del IGN para generar este tipo de formato.

1.2. Estado del Arte

1.2.1. GIS

Los sistemas de información geográfica son herramientas que permiten almacenar y analizar datos geoespaciales. Los sistemas digitales actuales permiten realizar consultas interactivas, añadir entradas a las bases de datos y visualizarlos de manera intuitiva. La información geográfica se puede aplicar a todo tipo de áreas, entre las que se encuentran la ingeniería, transporte, telecomunicación, economía, sociologia... Debido a la gran importancia tanto en el sector público como el privado[5], los estándares abiertos cobran importancia por estar disponibles al público, no tener que pagar licencias y ser consensuados por organizaciones de estándares internacionales. Entre ellas se encuentra el Open Geospatial Consortium(OEG) que se creó en 1994 y agrupa a 521 (en marzo de 2021) miembros de organizaciones públicas y privadas.[6] El OGC trabaja junto con las principales organizaciones de estándares de su ámbito (ISO/TC 211, W3C, IETF...) [4]

Existen diversos formatos de fichero GIS, divididos en **raster** y **vector**. La diferencia es equivalente a la que existe entre imágenes con resolución limitada por el número de píxeles (raster) y las imágenes vectoriales formadas por puntos, líneas y polígonos; con resoluciones infinitas. Cada tipo de formato tiene sus ventajas y desventajas y la elección dependerá del caso de uso. Existen varios formatos vectoriales pero para este trabajo sólo se se considerarán el formato *shapefile* y el *GeoPackage* para cumplir los objetivos.

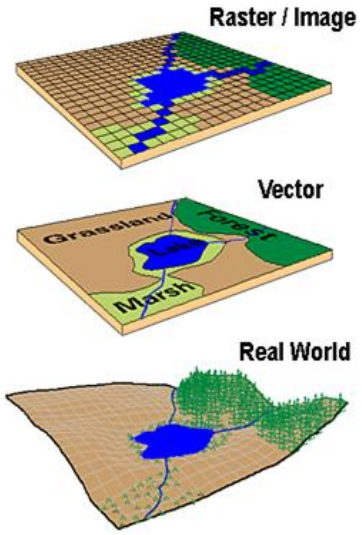


Figura 1.1: Representación del terreno mediante vectores y raster

1.2.1.1. Shapefile

El formato ESRI Shapefile (SHP) es un formato de archivo de datos espaciales vectorial desarrollado por la compañía ESRI a principios de la década de 1990. A pesar de ser propietario, la especificación es abierta, y se considera un estandar de facto. Debido a su popularidad, goza de grán compatibilidad con sig. Gracias al uso de un fichero índice, se obtiene una velocidad de lectura alta, y su eficiencia de tamaño produce archivos relativamente pequeños.

Sin embargo, tiene varias desventajas, algunas derivadas del uso del estándar dBase [11] [14]:

1. No tiene definición de sistema de referencia de coordenadas [1] , se puede usar uno pero no es parte estándar de la especificación.
2. Se reparte en múltiples ficheros: es incómodo y lleva a errores al compartirlos.
3. Los nombres de atributos están limitados a 10 caracteres ASCII
4. El número máximo de campos de atributo es 255.
5. Solo admite float, integer, date y text con un máximo de 254 caracteres.
6. No se puede especificar conjunto de caracteres de la BBDD.
7. El tamaño está limitado a 4GB.
8. No admite valores NULL
9. No hay forma de describir las relaciones topológicas en el formato.
10. Solamente puede almacenar una geometría por archivo.
11. Utiliza una estructura de datos de tabla plana, sin jerarquías, relaciones ni estructura en árbol.
12. El soporte 3D es muy limitado.

1.2.1.2. GeoPackage

GeoPackage es un formato GIS implementado en SQLite publicado por el OEG en 2014. [12]

El formato geopackage tiene las siguientes ventajas [14]:

- Es abierto, no propietario, basado en estándares, independiente de plataformas, portable y compacto.
- Gracias a SQLite puede almacenar datos grandes (hasta 140TB)[17] y los atributos de las geometrías pueden contentener nombres muy largos.

[1] Un sistema de coordenadas es un sistema de referencia que se utiliza para representar la ubicación de entidades geográficas, imágenes y observaciones (como las localizaciones GPS) dentro de un marco geográfico común. Los sistemas de coordenadas permiten a los datasets geográficos utilizar ubicaciones comunes para la integración de datasets. [15]

- Dispone de índices espaciales basados en R-trees [16] que incrementan la velocidad de búsquedas espaciales y su visualización en los SIG de escritorio.
- Todo el contenido se almacena en un único archivo .gpkg que puede almacenar multitud de tipos de geometrías
- Soporta el uso directo, para acceder a los datos de GeoPackage de forma «nativa» sin traducciones de formato intermedio.
- GeoPackage es soportado por GDAL[18], la librería de conversión de datos utilizada por multitud de programas GIS (incluído GeoKettle), y los principales programas GIS.

1.2.2. Datos enlazados

El objetivo de los datos enlazados es utilizar la web como una única base de datos global. Tim Berners Lee, creador de la World Wide Web, quien acuñó el término linked data[13], definió sus 4 principios fundamentales:

1. Utilizar URIs para identificar los recursos publicados en la Web.
2. Utilizar URIs HTTP para que las personas puedan consultar esos recursos.
3. Cuando alguien acceda a una URI, proporcionar información útil mediante estándares (RDF*, SPARQL).
4. Incluir enlaces a otras URIs para facilitar el descubrimiento de más información relacionada.

Los datos enlazados posibilitan la web semántica, extensión de la web tradicional en la que la información tiene significado bien definido[19] y fundamentada en:

- URIs: cadena de caracteres que identifica los recursos de una red de forma unívoca.
- RDF: método para la descripción conceptual o modelado de la información.
- HTTP: protocolo de comunicación.

RDF modela información mediante triples o tripletas de sujeto-predicado-objeto. El sujeto hace referencia al recurso y el predicado a sus rasgos o aspectos y relación entre el sujeto y el objeto. SPARQL es el lenguaje para la consulta de grafos RDF.

1.2.3. Portales de Datos abiertos

Los datos abiertos parten de la idea de que los datos deberían estar disponibles de forma libre para todo el mundo, libre de derechos de autor, patentes o de otros mecanismos de control. Los portales de datos abiertos proporcionan una manera sencilla de buscar y obtener estos datos. Los datos pueden tener cualquier procedencia, pero han cobrado especial importancia los datos ligados a las

políticas de Gobierno abierto, que persigue que los datos y la información, especialmente las que poseen las administraciones públicas, se publiquen de forma abierta. [20]

El tercer objetivo de este trabajo se centra en realizar un procesado completo de todos los datos del IGN. *datos.ign.es* es una iniciativa del Instituto Geográfico Nacional (IGN) para la generación de la información semántica de sus recursos[21]. Actualmente el dataset disponible es la Base Topográfica Nacional 1:100.000 (BTN100), un catálogo de datos geográficos agrupados por temáticas.

1.2.4. Map4RDF

Map4rdf es una herramienta para la navegación y visualización de datasets RDF con información geoespacial mediante facetas[8]. Algunos ejemplos de las facetas y sus contenidos que permiten clasificar los elementos del BTN100:

- Altimetria: Cerro, Cordillera, Montaña...
- Hidrografía: Bahía, Cabo, Playa...
- Transporte: Aeropuerto, Calle, Faro...
- ...

El funcionamiento de Map4rdf es el siguiente:

1. El componente *DAO* [2] se conecta a una *triplestore* [3] mediante el *endpoint SPARQL* [4] para responder a las consultas de facetas.
2. La interfaz de navegación facetada obtiene la lista de facetas y las visualiza.
3. El usuario selecciona una faceta y el componente DAO realiza una consulta en el triplestore mediante el endpoint SPARQL para recuperarlas la información pedida.
4. La interfaz recibe toda esta información y la visualiza en el mapa.

[2]Data Acces Object: propociona una interfaz abstracta a una base de datos.
[3]Triplestore: base de datos de tripletas
[4]SPARQL endpoint: url capaz de recibir y procesar peticiones del protocolo SPARQL

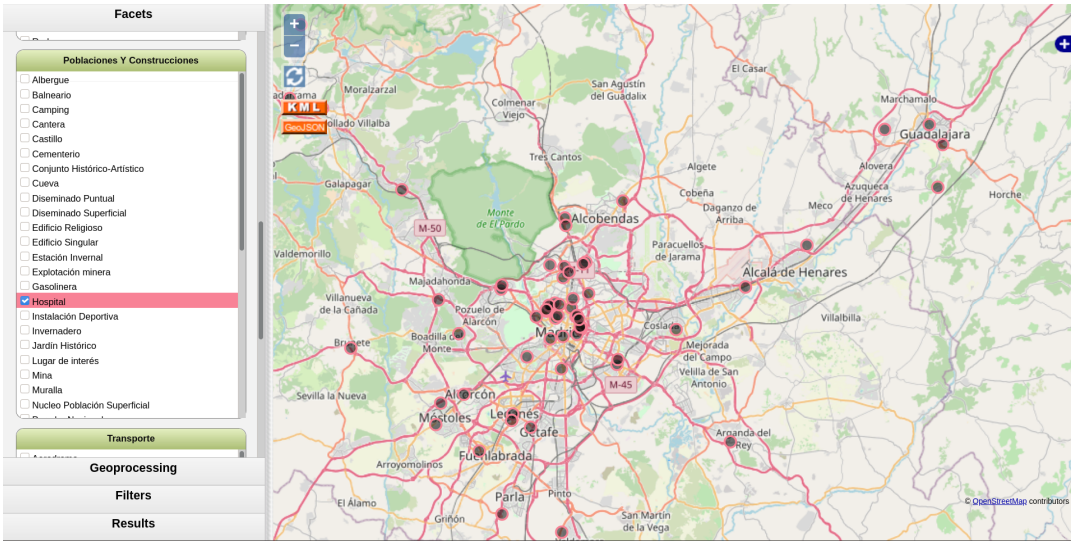


Figura 1.2: http://certidatos.ign.es/map/ que implementa Map4Rdf

1.2.5. GeoKettle

GeoKettle es una (antigua) versión de Pentaho Data Integration (Kettle)[23] con capacidad de tratamiento de datos espaciales. Es una potente herramienta ETL: extracción, transformación y carga orientada al uso de metadatos y con funcionalidades espaciales dedicada a la integración de diversos orígenes de datos para la construcción y/o actualización de bases de datos espaciales y almacenes de datos espaciales. [22]

TripleGeo es un plugin para GeoKettle que transforma datos geoespaciales en tripletas RDF siguiendo el standar GeoSPARQL [24]

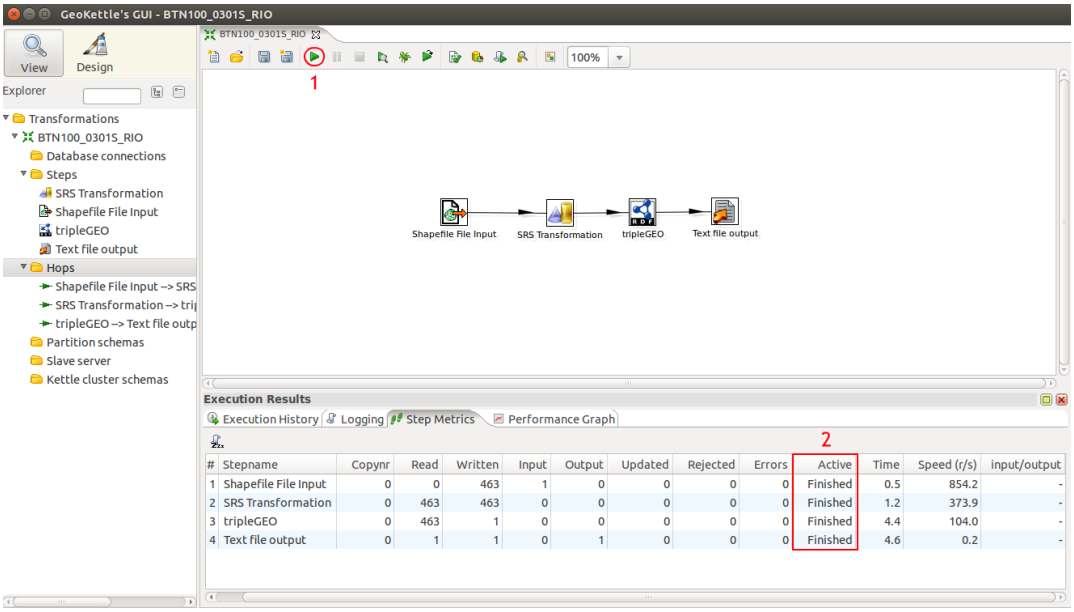


Figura 1.3: TripleGeoKettle en funcionamiento, TripleGeoKettle wiki

1.3. Pentaho Data Integration Spoon

PDI Spoon[27] es la herramienta que reemplaza a GEOKettle. La GUI, el funcionamiento y el SDK para desarrollar plugins es parecido. Sin embargo, en cuanto al desarrollo de plugins, cambia la manera de gestionar las dependencias, las cabeceras de algunas interfaces que se deben implementar, los iconos, metadatos, estructura de carpetas...

1.3.1. Pentaho GIS Plugins

Es un plugin desarrollado por Atol Conseils et Développements[26] para PDI9. Proporciona parte de la funcionalidad GIS que tenia GEOKettle. fig.1.4

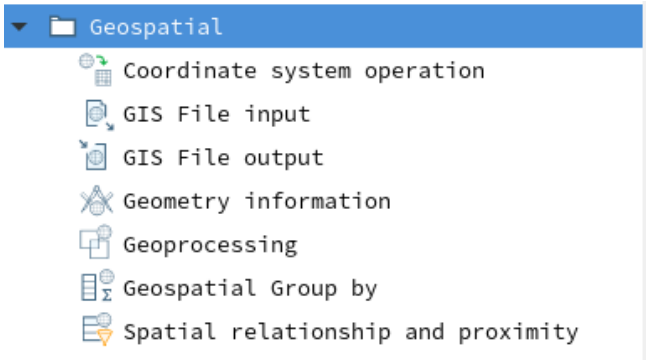


Figura 1.4: Steps incluidos en gis-plugins

1.3.2. Apache Ant

Apache Ant[28] es una libreria Java y herramienta de línea de comandos para construir aplicaciones java (compilar, ensamblar y ejecutarlas). Los scripts de configuración se escriben en formato XML y son muy flexibles.

1.3.3. Apache Maven

Apache Maven [29] es una herramienta de gestión de proyectos y dependencias. Está basado en el concepto Project Object Model (POM). Maven es capaz de construir las aplicaciones, descargar las dependencias y gestionarlas, ejecutar tests, crear documentación...

El fichero principal, pom.xml detalla la configuración. Se pueden incluir repositorios externos y dependencias. Maven se encarga de descargar las dependencias y sus subdependencias desde los repositorios para no tener que hacerlo a mano.

Capítulo 2

Desarrollo

2.1. Map4rdf

2.1.1. Instalación

Se ha optado por utilizar la máquina Virtual proporcionada en la Wiki del proyecto para minimizar la posibilidad de incompatibilidades.

2.2. GeoKettle

Desde que se publicó "A sustainable process and toolbox for geographical linked data generation and publication: a case study with BTN100" en 2019, GeoKettle ha dejado de estar soportado. La pagina oficial y de documentación ya no están disponibles. Un objetivo de este TFG es dar soporte GeoPackage a GeoKettle. No tiene sentido desarrollar soluciones de "modernización" sobre software abandonado.

Algunas funcionalidades de GeoKettle se integraron en PDI directamente y otras desaparecieron. Actualmente, el soporte GIS de Pentaho está dentro de PDI Spoon y además hay algunas funcionalidades más en el plugin llamado pentaho-gis-plugins[25]. Se actualizará la primera fase a: **"replicar la funcionalidad y las transformaciones de GeoKettle + TripleGeo en la suite PDI."**. Dado que se trata de replicar la funcionalidad anterior, se analizarán las tranformaciones realizadas por el OEG en el repositorio de GitHub BTN100. Como se puede ver en la figura 2.1, partes del workflow fallan. Es lo que se pretende solucionar.

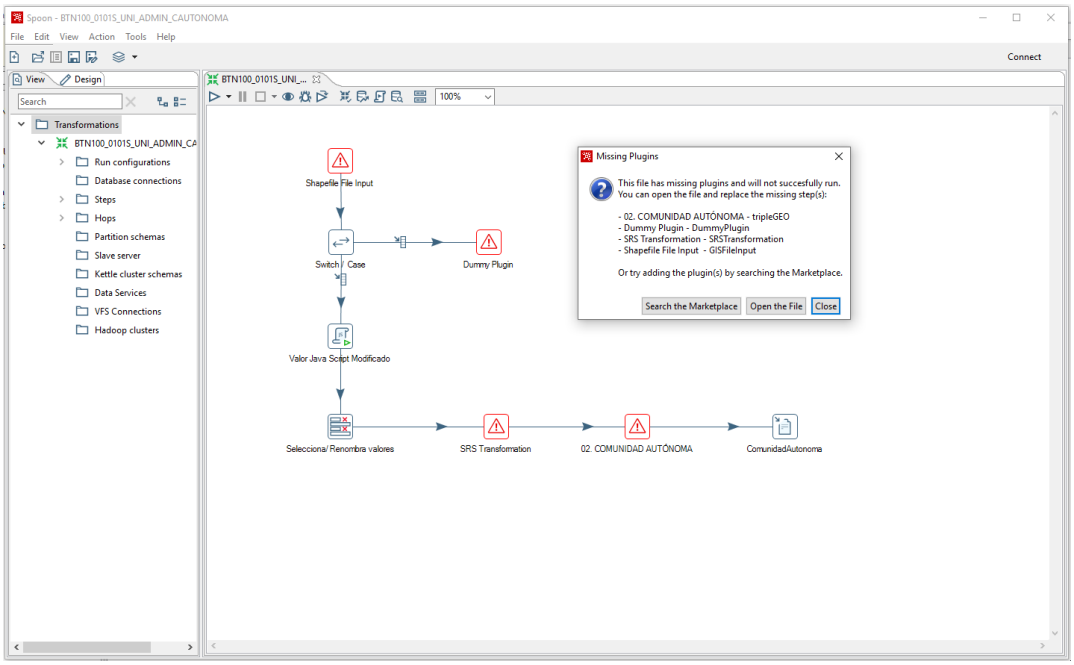


Figura 2.1: Workflow importado en la nueva suite

2.2.1. Funcionamiento

Para poder realizar el "port" de GeoKettle a Spoon, primero es necesario entender el funcionamiento y transformaciones actuales de GeoKettle observando la entrada y salida de cada paso. Además esta manera se observará mejor el flujo de datos y será más facil añadir soporte a GeoPackage en el futuro. No todas las transformaciones tienen los mismos pasos, pero son parecidas. Como ejemplo se utilizarán los datos de BTN100_0101S_UNI_ADMIN y la transformación BTN100_0101S_UNI_ADMIN_CAUTONOMA. La transformación contiene los siguientes pasos:

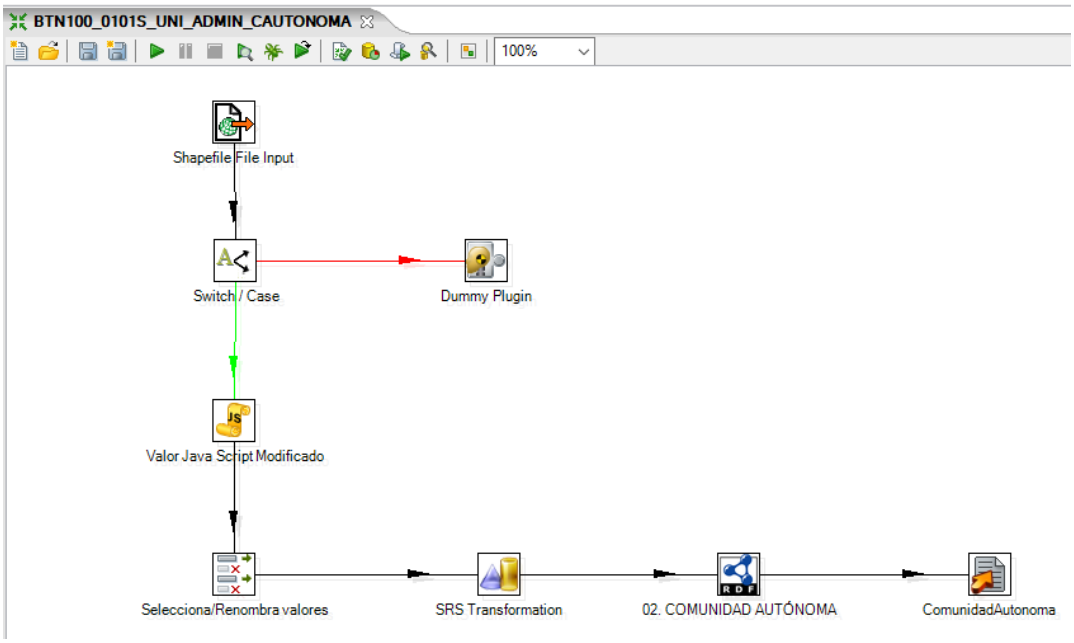


Figura 2.2: Transformación BTN100_0101S_UNI_ADMIN_CAUTONOMA

1. Shapefile File Input
2. Switch Case
3. Dummy Plugin
4. Valor Java Script Modificado
5. Selecciona/Renombra valores
6. SRS Transformation
7. TripleGeo
8. Text Output

2.2.1.1. Shapefile File Input

Lee el fichero shapefile.

1. *.shp*: Geometría fig.2.3

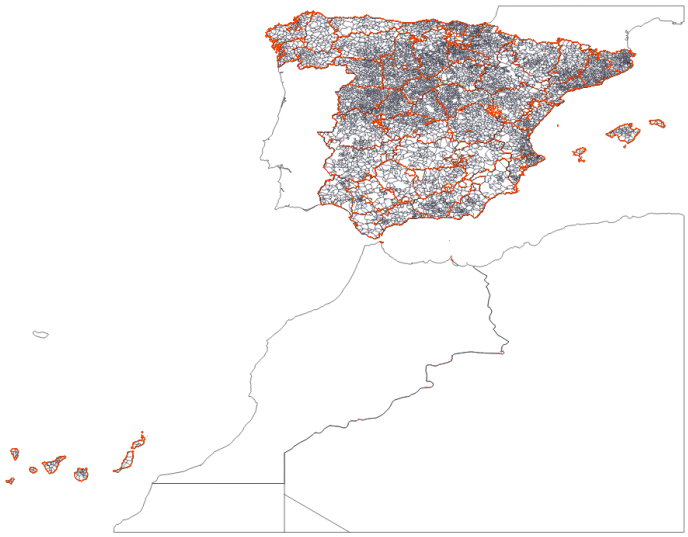
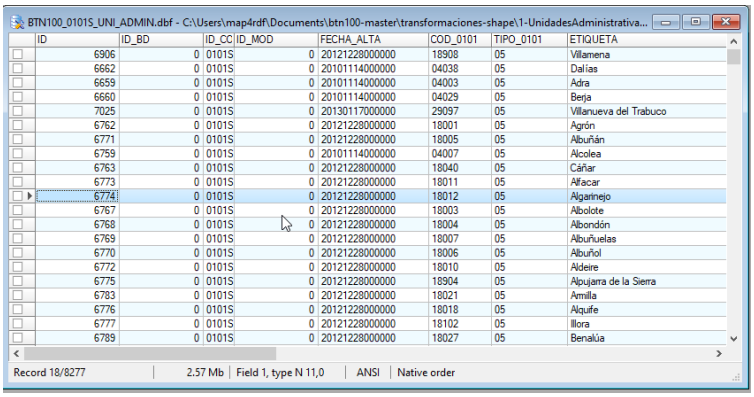


Figura 2.3: Geometría contenida en el shapefile

2. *.dbf*: Datos asociados en columnas fig.2.4



BTN100_0101S_UNI_ADMIN.dbf - C:\Users\map4rdf\Documents\btn100-master\transformaciones-shape\1-UnidadesAdministrativa...

	ID	ID_BD	ID_CC	ID_MOD	FECHA_ALTA	COD_0101	TIPO_0101	ETIQUETA
	6906	0	0101S	0	20121228000000	18908	05	Villamena
	6662	0	0101S	0	20101114000000	04038	05	Dalías
	6659	0	0101S	0	20101114000000	04003	05	Adra
	6660	0	0101S	0	20101114000000	04029	05	Berja
	7025	0	0101S	0	20130117000000	29097	05	Villanueva del Trabuco
	6762	0	0101S	0	20121228000000	18001	05	Agrón
	6771	0	0101S	0	20121228000000	18005	05	Albuñán
	6759	0	0101S	0	20101114000000	04007	05	Alcolea
	6763	0	0101S	0	20121228000000	18040	05	Cáñar
	6773	0	0101S	0	20121228000000	18011	05	Alfacar
	6774	0	0101S	0	20121228000000	18012	05	Algarinejo
	6767	0	0101S	0	20121228000000	18003	05	Albolote
	6768	0	0101S	0	20121228000000	18004	05	Albondón
	6769	0	0101S	0	20121228000000	18007	05	Albuñuelas
	6770	0	0101S	0	20121228000000	18006	05	Albuñol
	6772	0	0101S	0	20121228000000	18010	05	Aldeire
	6775	0	0101S	0	20121228000000	18904	05	Alpujarra de la Sierra
	6783	0	0101S	0	20121228000000	18021	05	Amilla
	6776	0	0101S	0	20121228000000	18018	05	Alquife
	6777	0	0101S	0	20121228000000	18102	05	Illora
	6789	0	0101S	0	20121228000000	18027	05	Benalúa

Record 18/8277 | 2.57 Mb | Field 1, type N 11,0 | ANSI | Native order

Figura 2.4: Datos columnares dbf asociados a la geometría

3. *.shx*: Índice para acelerar búsquedas
4. *.prj*: Sistema de coordenadas

2.2.1.2. Switch case

El switch case se encarga de filtrar y seleccionar las comunidades autónomas, identificadas por el valor 02 del Campo TIPO_0101. Si son una CCAA, se envían al paso 4, e.o.c. se envían al paso 3. fig.2.5 y 2.6

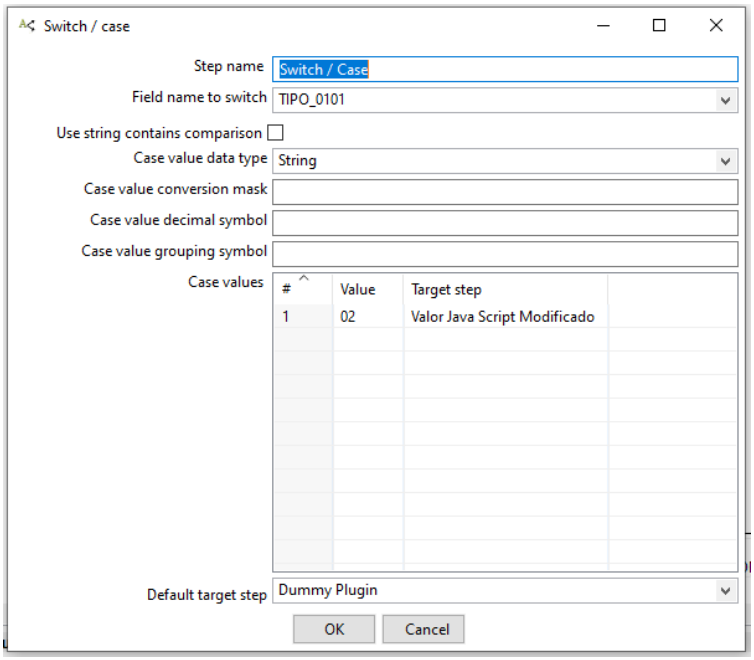
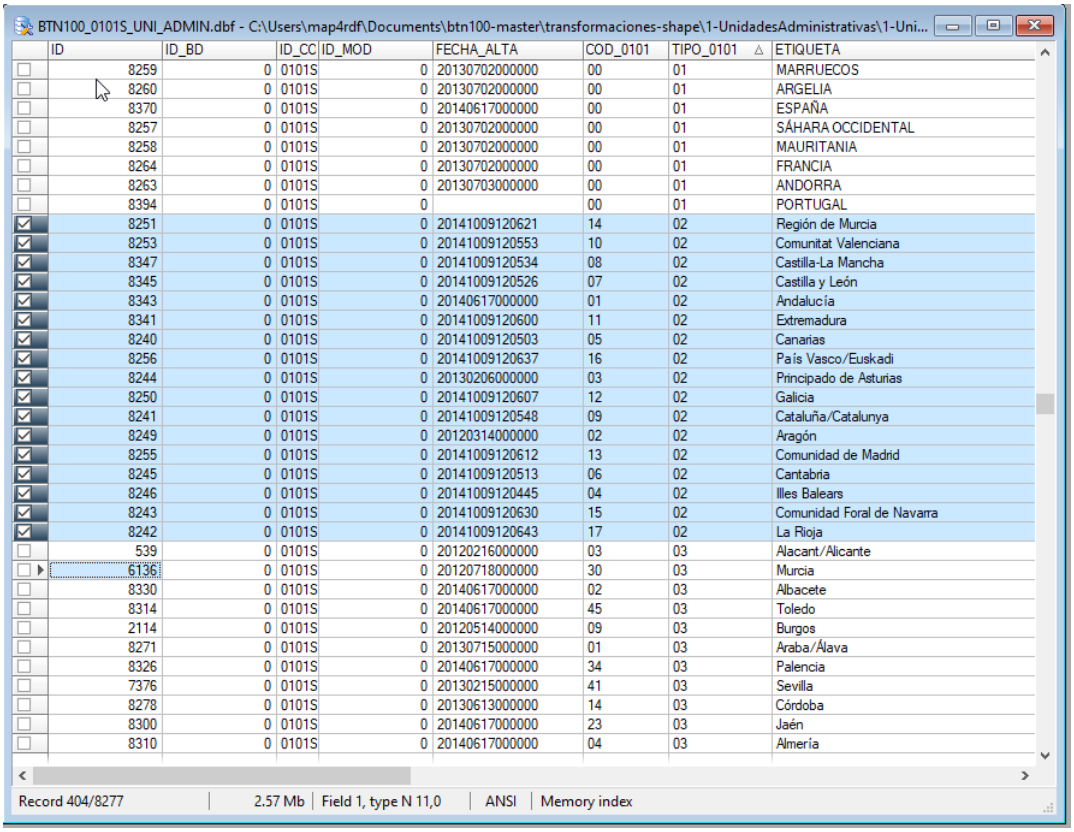


Figura 2.5: Paso switch



ID	ID_BD	ID_CC	ID_MOD	FECHA_ALTA	COD_0101	TIPO_0101	ETIQUETA
8259	0	0101S	0	20130702000000	00	01	MARRUECOS
8260	0	0101S	0	20130702000000	00	01	ARGELIA
8370	0	0101S	0	20140617000000	00	01	ESPAÑA
8257	0	0101S	0	20130702000000	00	01	SÁHARA OCCIDENTAL
8258	0	0101S	0	20130702000000	00	01	MAURITANIA
8264	0	0101S	0	20130702000000	00	01	FRANCIA
8263	0	0101S	0	20130703000000	00	01	ANDORRA
8394	0	0101S	0		00	01	PORTUGAL
8251	0	0101S	0	20141009120621	14	02	Región de Murcia
8253	0	0101S	0	20141009120553	10	02	Comunitat Valenciana
8347	0	0101S	0	20141009120534	08	02	Castilla-La Mancha
8345	0	0101S	0	20141009120526	07	02	Castilla y León
8343	0	0101S	0	20140617000000	01	02	Andalucía
8341	0	0101S	0	20141009120600	11	02	Extremadura
8240	0	0101S	0	20141009120503	05	02	Canarias
8256	0	0101S	0	20141009120637	16	02	País Vasco/Euskadi
8244	0	0101S	0	20130206000000	03	02	Principado de Asturias
8250	0	0101S	0	20141009120607	12	02	Galicia
8241	0	0101S	0	20141009120548	09	02	Cataluña/Catalunya
8249	0	0101S	0	20120314000000	02	02	Aragón
8255	0	0101S	0	20141009120612	13	02	Comunidad de Madrid
8245	0	0101S	0	20141009120513	06	02	Cantabria
8246	0	0101S	0	20141009120445	04	02	Illes Balears
8243	0	0101S	0	20141009120630	15	02	Comunidad Foral de Navarra
8242	0	0101S	0	20141009120643	17	02	La Rioja
539	0	0101S	0	20120216000000	03	03	Alacant/Alicante
6136	0	0101S	0	20120718000000	30	03	Murcia
8330	0	0101S	0	20140617000000	02	03	Albacete
8314	0	0101S	0	20140617000000	45	03	Toledo
2114	0	0101S	0	20120514000000	09	03	Burgos
8271	0	0101S	0	20130715000000	01	03	Araba/Álava
8326	0	0101S	0	20140617000000	34	03	Palencia
7376	0	0101S	0	20130215000000	41	03	Sevilla
8278	0	0101S	0	20130613000000	14	03	Córdoba
8300	0	0101S	0	20140617000000	23	03	Jaén
8310	0	0101S	0	20140617000000	04	03	Almería

Figura 2.6: La filas correspondientes a las CCAA

2.2.1.3. Dummy Plugin

No hace ninguna transformación, su propósito es recoger los datos innecesarios del switch.

2.2.1.4. Valor Java Script Modificado

El script cambia el formato de la fecha para facilitar la lectura: de YYYYMMDDHHMMSS a YYYY-MM-DD. También crea un nuevo campo llamado identificador a partir del campo etiqueta, cambiando espacios por barras bajas, mayúsculas por minúsculas, quitando tildes y signos de puntuación. fig.2.7 y 2.8

Script 1

```
//Script here

var aux = FECHA_ALTA.match(/^(\d{4})(\d{2})(\d{2})/);

//Here we fix the date
if(!aux || aux.length < 4){
    FECHA_ALTA = "";
}
else{
    FECHA_ALTA = new Date(aux[1], aux[2]-1, aux[3]);
    FECHA_ALTA = date2str(FECHA_ALTA, "yyyy-MM-dd");
}


var etiq = ETIQUETA.toLowerCase().replace(' ','-');
etiq = etiq.replace(/\s/g, '-').replace('á', 'a').replace('Á', 'A').replace('é', 'e').replace('É', 'E').replace('í', 'i')
.replace('Í', 'I').replace('ó', 'o').replace('Ó', 'O').replace('ú', 'u').replace('Ú', 'U').replace('ñ', 'n');
etiq = etiq.replace(',','').replace(';','').replace('.','');
identificador =  etiq
```

Figura 2.7: Script Javascript

Examine preview data

Standard view | Geographic view

Rows of step: Valor Java Script Modificado (17 rows)

#	the_geom	ID	ID_BD	ID_CODIGO	ID_MOD	FECHA_ALTA	COD_0101	TIPO_0101	ETIQUETA	identificador
1	MULTIPO...	8251	0	0101S	0	2014-10-09	14	02	Región de Murcia	region-de-murcia
2	MULTIPO...	8253	0	0101S	0	2014-10-09	10	02	Comunitat Valenciana	comunitat-valenciana
3	MULTIPO...	8347	0	0101S	0	2014-10-09	08	02	Castilla-La Mancha	castilla-la-mancha
4	MULTIPO...	8345	0	0101S	0	2014-10-09	07	02	Castilla y León	castilla-y-leon
5	MULTIPO...	8343	0	0101S	0	2014-06-17	01	02	Andalucía	andalucia
6	MULTIPO...	8341	0	0101S	0	2014-10-09	11	02	Extremadura	extremadura
7	MULTIPO...	8240	0	0101S	0	2014-10-09	05	02	Canarias	canarias
8	MULTIPO...	8256	0	0101S	0	2014-10-09	16	02	País Vasco/Euskadi	pais-vasco/euskadi
9	MULTIPO...	8244	0	0101S	0	2013-02-06	03	02	Principado de Asturias	principado-de-asturias
10	MULTIPO...	8250	0	0101S	0	2014-10-09	12	02	Galicia	galicia
11	MULTIPO...	8241	0	0101S	0	2014-10-09	09	02	Cataluña/Catalunya	cataluna/catalunya
12	MULTIPO...	8249	0	0101S	0	2012-03-14	02	02	Aragón	aragon
13	MULTIPO...	8255	0	0101S	0	2014-10-09	13	02	Comunidad de Madrid	comunidad-de-madrid
14	MULTIPO...	8245	0	0101S	0	2014-10-09	06	02	Cantabria	cantabria
15	MULTIPO...	8246	0	0101S	0	2014-10-09	04	02	Illes Balears	illes-balears
16	MULTIPO...	8243	0	0101S	0	2014-10-09	15	02	Comunidad Foral de Navar...	comunidad-foral-de-nav...
17	MULTIPO...	8242	0	0101S	0	2014-10-09	17	02	La Rioja	la-rioja

Figura 2.8: Resultado del cambio de formato de fecha

2.2.1.5. Selecciona/Renombra valores

Cambia los metadatos de la columna FECHA_ALTA para que sea reconocida como fecha. fig.2.9

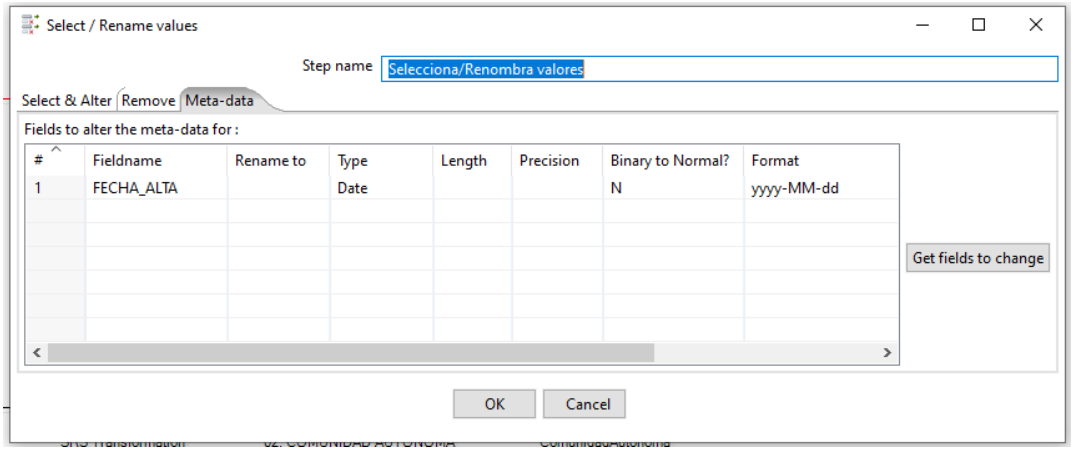


Figura 2.9: Selecciona/renombra valores

2.2.1.6. SRS Transformation

Realiza la reproyección del sistema de coordenadas, en este caso de ETRS89 a WGS84. fig.2.10

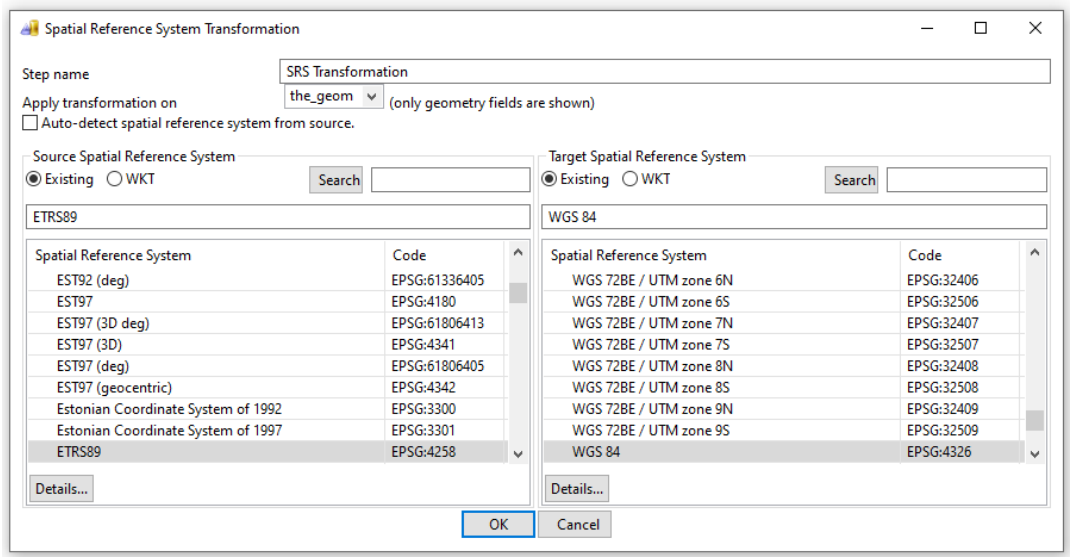


Figura 2.10: Transformacion SRS

2.2.1.7. TripleGeoKettle

Transforma el shapefile del paso anterior en RDF en formato ttl. Se pueden configurar los parámetros asociados a la ontologia y decidir si se muestran ciertas columnas o no. fig.2.11

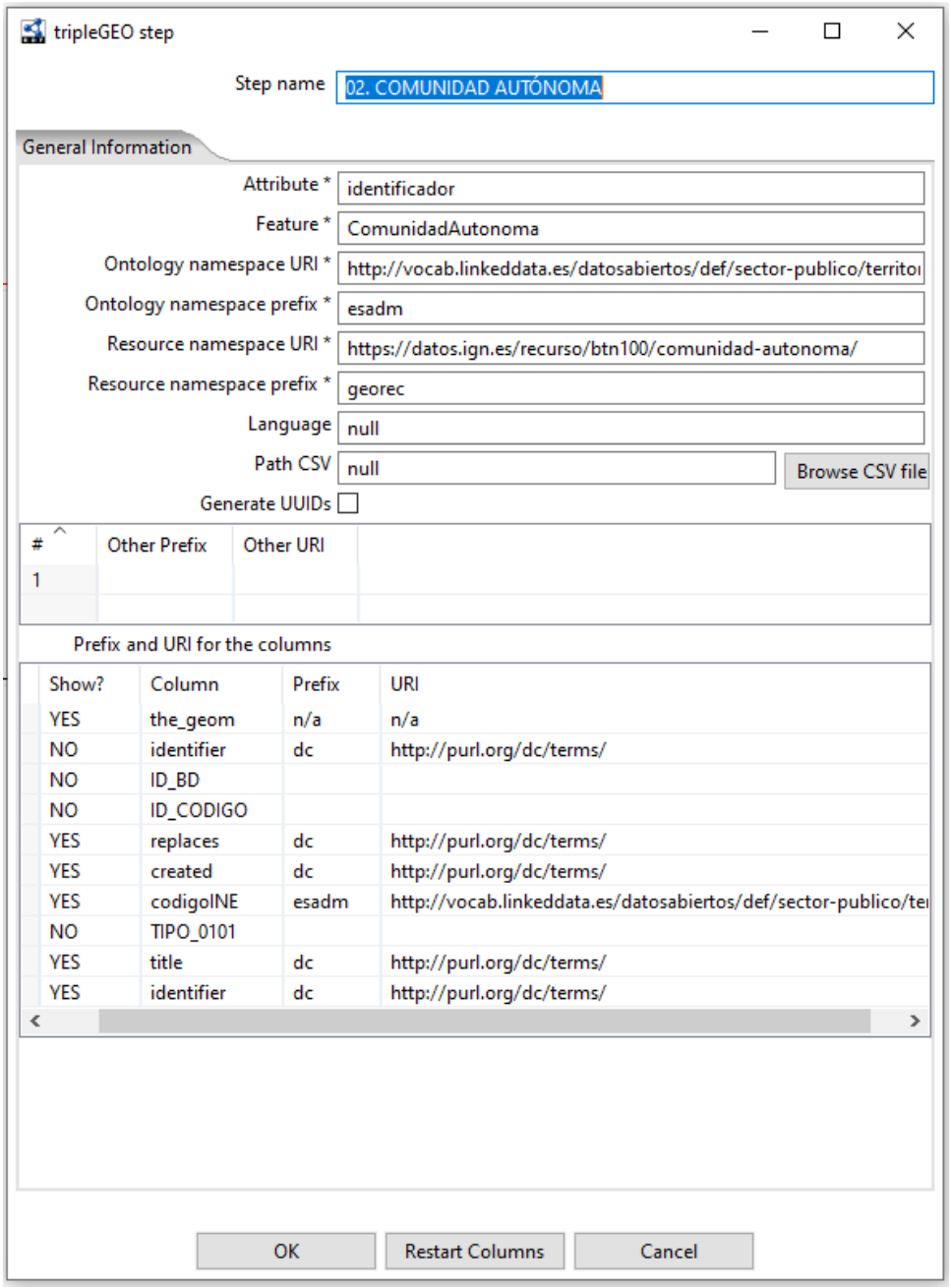


Figura 2.11: tripleGeoKettle

2.2.1.8. Text file output

Escribe los datos RDF en un fichero de texto con formato .ttl. fig.2.12

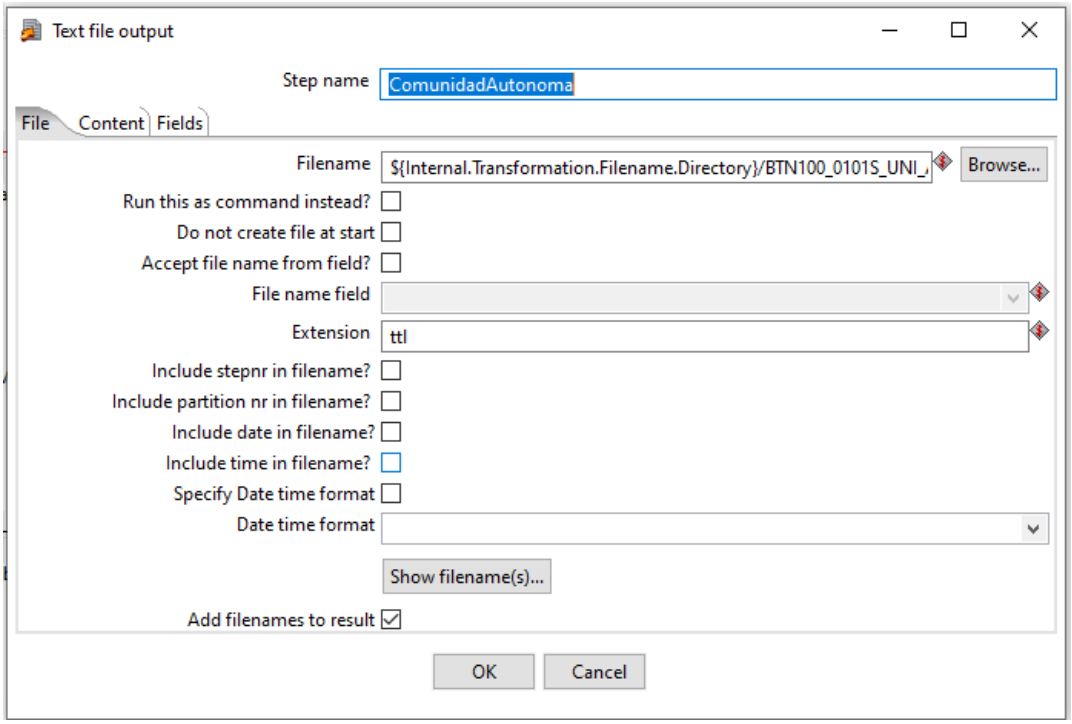


Figura 2.12: text-file-output

2.2.1.9. Resultado de la transformación

```
@prefix geo:   <http://www.w3.org/2003/01/geo/wgs84_pos#> .
@prefix geosparql: <http://www.opengis.net/ont/geosparql#> .
@prefix sf:    <http://www.opengis.net/ont/sf#> .
@prefix rdf:   <http://www.w3.org/1999/02/22-rdf-syntax-ns#> .
@prefix owl:   <http://www.w3.org/2002/07/owl#> .
@prefix xsd:   <http://www.w3.org/2001/XMLSchema#> .
@prefix georec: <https://datos.ign.es/recurso/btn100/comunidad-autonoma/> .
@prefix esadm: <http://vocab.linkeddata.es/datosabiertos/def/sector-publico/territorio#> .
@prefix rdfs:  <http://www.w3.org/2000/01/rdf-schema#> .
@prefix foaf:  <http://xmlns.com/foaf/0.1/> .
@prefix dc:    <http://purl.org/dc/terms/> .

georec:comunidad-de-madrid
        a                       esadm:ComunidadAutonoma ;
        rdfs:label              "comunidad-de-madrid" ;
        dc:created              "2014-10-09"^^xsd:date ;
        dc:identifier           "comunidad-de-madrid" ;
        dc:title                "Comunidad de Madrid" ;
        esadm:codigoINE         "13"^^xsd:int ;
        geosparql:hasGeometry   <https://datos.ign.es/recurso/btn100/comunidad-autonoma/...

georec:region-de-murcia
        a                       esadm:ComunidadAutonoma ;
        rdfs:label              "region-de-murcia" ;
        dc:created              "2014-10-09"^^xsd:date ;
        dc:identifier           "region-de-murcia" ;
```

```
        dc:title                "Región de Murcia" ;
        esadm:codigoINE         "14"^^xsd:int ;
        geosparql:hasGeometry   <https://datos.ign.es/recurso/btn100/comunidad-autonoma/...

<https://datos.ign.es/recurso/btn100/comunidad-autonoma/aragon/geometry>
        a                   sf:Polygon ;
        geosparql:asWKT   "POLYGON ((-1.6174492000010632 40.94373283914169, -1.62366030000...
        ...
```

2.3. Port a PDI9

2.3.1. Tests con pentaho-gis-plugins

El 3 de Marzo el plugin añadió soporte para el formato GeoPackage. Con el siguiente test se comprueba que los componentes que se utilizarán para reemplazar las transformaciones antiguas funcionan correctamente.

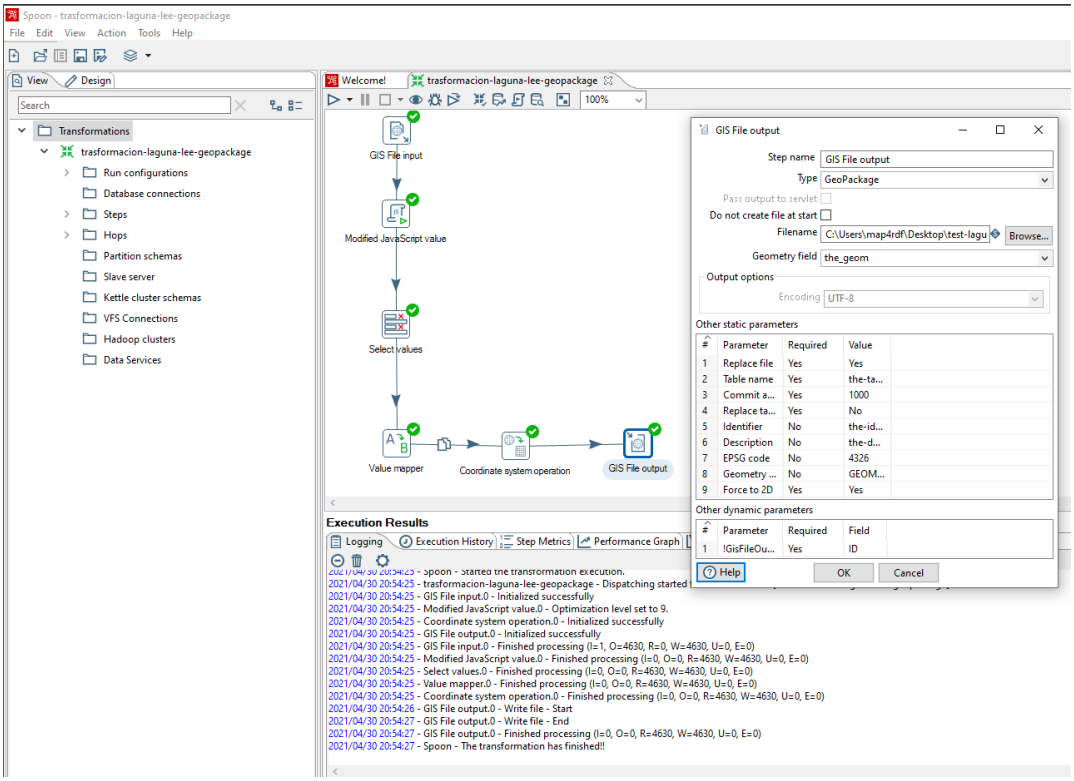


Figura 2.13: test-laguna-geopackage

2.3.2. Entorno de desarrollo

2.3.2.1. Dependencias e instalación

En los ultimos años Pentaho ha pasado de utilizar Apache Ant a utilizar Apache Maven. TripleGeoKettle también utilizaba Ant, y se ha decidido utilizar Maven

por las ventajas que ofrece. El proyecto TripleGeoKettle contenía una carpeta lib en la que se encontraban los .jar con las dependencias necesarias. Ahora, las dependencias se administran con Maven y el pom.xml. Se incluyen los repositorios de Pentaho y OSGeo y las dependencias que se quieren incluir. Las properties funcionan a modo de "variable" para poder cambiar la versión de pdi fácilmente en un solo lugar.

```
<?xml version="1.0"?>
<project xsi:schemaLocation="http://maven.apache.org/POM/4.0.0 http://maven.apache.org/
    xsd/maven-4.0.0.xsd" xmlns="http://maven.apache.org/POM/4.0.0"
        xmlns:xsi="http://www.w3.org/2001/XMLSchema-instance">
    <modelVersion>4.0.0</modelVersion>
    <groupId>oeg-upm</groupId>
    <artifactId>tripleGeoKettle-oeg</artifactId>
    <version>1</version>
    <name>tripleGeoKettle</name>

    <properties>
        <pentaho-metadata.version>9.1.0.0-324</pentaho-metadata.version>
        <pdi.version>9.1.0.0-324</pdi.version>
    </properties>

    <repositories>
        <repository>
            <id>pentaho-releases</id>
            <url>https://nexus.pentaho.org/content/groups/omni</url>
        </repository>
        <repository>
            <id>osgeo</id>
            <url>https://repo.osgeo.org/repository/release/</url>
        </repository>
    </repositories>

    <dependencies>
        <dependency>
            <groupId>org.pentaho</groupId>
            <artifactId>pentaho-metadata</artifactId>
            <version>${pentaho-metadata.version}</version>
            <scope>provided</scope>
        </dependency>
        <dependency>
            <groupId>pentaho-kettle</groupId>
            <artifactId>kettle-core</artifactId>
            <version>${pdi.version}</version>
            <scope>provided</scope>
        </dependency>
        <dependency>
            <groupId>pentaho-kettle</groupId>
            <artifactId>kettle-engine</artifactId>
            <version>${pdi.version}</version>
            <scope>provided</scope>
        </dependency>
        <dependency>
            <groupId>pentaho-kettle</groupId>
            <artifactId>kettle-ui-swt</artifactId>
            <version>${pdi.version}</version>
            <scope>provided</scope>
        </dependency>

        <dependency>
            <groupId>org.apache.jena</groupId>
            <artifactId>jena-arq</artifactId>
            <version>3.0.0</version>
        </dependency>
        <dependency>
```

```
            <groupId>com.vividsolutions</groupId>
            <artifactId>jts</artifactId>
            <version>1.11</version>
        </dependency>
        <dependency>
            <groupId>org.geotools</groupId>
            <artifactId>gt-geometry</artifactId>
            <version>2.7-M0</version>
        </dependency>

    </dependencies>

    <build>
        <plugins>
            <plugin>
                <artifactId>maven-assembly-plugin</artifactId>
                <executions>
                    <execution>
                        <id>distro-assembly</id>
                        <phase>package</phase>
                        <goals>
                            <goal>single</goal>
                        </goals>
                        <configuration>
                            <appendAssemblyId>false</appendAssemblyId>
                            <descriptors>
                                <descriptor>src/main/assembly/assembly.xml</descriptor>
                            </descriptors>
                        </configuration>
                    </execution>
                </executions>
            </plugin>
        </plugins>
    </build>
</project>
```

A continuación se muestran los pasos para configurar el entorno de desarrollo en Arch Linux:

```
# Instalar PDI desde AUR (los mirrors son mas rapidos que los de SourceForge)
yay -S pdi-ce
# Instalar Java 8
sudo pacman -S jdk8-openjdk
# Cambiar la version de java con
sudo archlinux-java set java-8-openjdk
# Instalar el Plugin de desarrollo
https://sourceforge.net/projects/pentaho/files/Pentaho%209.1/plugins/kettle-sdk-plugin-assembly-9.1.0.0-324.zip/download
# Instalar Maven
sudo pacman -S maven
# Descargar el settings.xml del repositorio de github en ~/.m2
https://raw.githubusercontent.com/pentaho/maven-parent-poms/master/maven-support-files/settings.xml
# Desde el directorio del plugin compilar y empaquetar el plugin
mvn clean package
# Instalarlo en PDI9 copiando el contenido del .zip generado en el directorio target
sudo cp target/tripleGeoKettle-oeg-1.jar /opt/pdi/plugins/steps/tripleGeoKettle-oeg-1/
```

2.3.2.2. Pentaho sdk plugins

Pentaho ofrece plugins[30] muy sencillos de ejemplo que muestran cómo implementar un plugin. Para el port de TripleGeoKettle se ha utilizado como referencia

kettle-sdk-step-plugin que simplemente escribe Hello World. Con él se ha probado la compilación, el pom.xml customizado, y pruebas varias como cambiar el icono o el nombre.

2.4. Proceso de porteo

Se comienza con el sdk de Pentaho y se realizan los cambios necesarios para integrar TripleGeoKettle en el nuevo entorno. Se muestran las imagenes del antes y después de la estructura de carpetas. fig.2.14 y 2.15

Figura 2.14: Estructura de carpetas del sdk

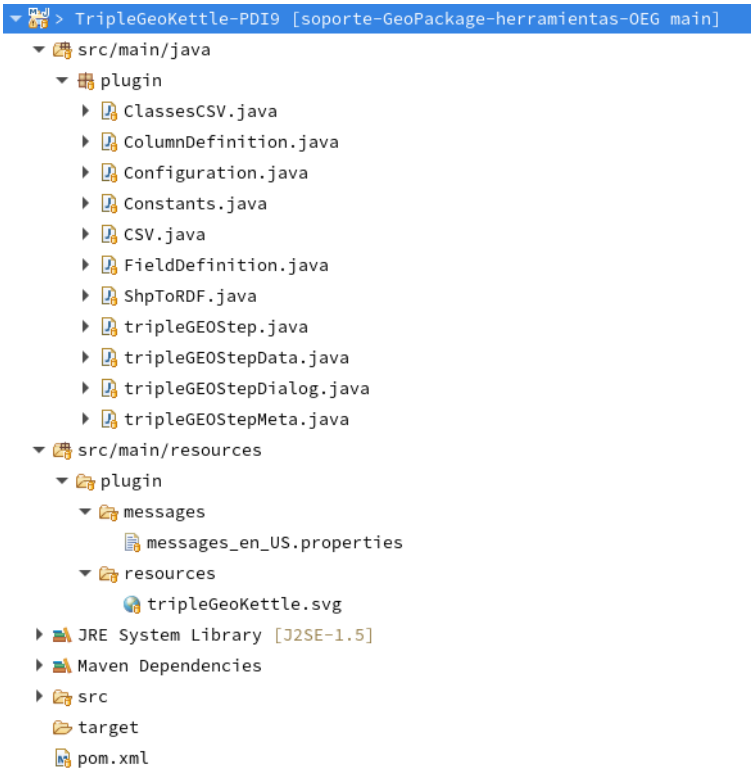


Figura 2.15: Estructura de carpetas de TripleGeoKettle portado

Pasos seguidos:

1. Cambiar el nombre del paquete a “plugin”. Es importante que el de java y el de resources tengan el mismo nombre para que los dialogos lean el fichero properties correctamente.

2. Cambiar messages.properties

```
#sdk properties
tripleGEO.FieldName.Label=Output field name
tripleGEO.CheckResult.ReceivingRows.OK=Step is receiving input from other steps.
tripleGEO.CheckResult.ReceivingRows.ERROR=No input received from other steps!

tripleGEOStep.Name=TripleGeoKettle
tripleGEOStep.TooltipDesc=An ETL Tool for Transforming Geospatial Data into RDF
    under the GeoSPARQL standard.
tripleGEOStep.DocumentationURL=https://github.com/oeg-upm/geo.linkeddata.es-
    TripleGeoKettle/wiki
tripleGEOStep.CasesURL=https://github.com/oeg-upm/geo.linkeddata.es-TripleGeoKettle
    /issues
tripleGEOStep.ForumURL=https://github.com/oeg-upm/geo.linkeddata.es-TripleGeoKettle
    /issues
tripleGEOStep.Linenr=Linenr {0}
tripleGEOStep.Error.NoOutputField=Could not find Output Field in row

# tripleGeoKettle custom properties
tripleGEOStepDialog.Shell.Title=tripleGEO step
tripleGEOStepDialog.Tab.MainTab=General Information
```

```
tripleGEOStepDialog.AttributeName.Label=Attribute *
tripleGEOStepDialog.Feature.Label=Feature *
tripleGEOStepDialog.OntologyNS.Label=Ontology namespace URI *
tripleGEOStepDialog.OntologyNSPrefix.Label=Ontology namespace prefix *
tripleGEOStepDialog.ResourceNS.Label=Resource namespace URI *
tripleGEOStepDialog.ResourceNSPrefix.Label=Resource namespace prefix *
tripleGEOStepDialog.Language.Label=Language
tripleGEOStepDialog.PathCSV.Label=Path CSV
tripleGEOStepDialog.PathCSVButton.Label=Browse CSV file
tripleGEOStepDialog.PathCSVButtonTooltip.Label=Browse for CVS file
tripleGEOStepDialog.uuids.Label=Generate UUIDs
tripleGEOStepDialog.Fields.Label=Other prefix and URI
tripleGEOStepDialog.other.Label=Other URI
tripleGEOStepDialog.otherPrefix.Label=Other Prefix
tripleGEOStepDialog.Column.Label=Column
tripleGEOStepDialog.Columns.Label=Prefix and URI for the columns
tripleGEOStepDialog.otherColumns.Label=URI
tripleGEOStepDialog.otherPrefixColumns.Label=Prefix
tripleGEOStepDialog.ShowColumns.Label=Show?
tripleGEOStepDialog.RestartFields.Button=Restart Columns
```

3. Modificar la annotation @Step de DemoStepMeta.java. Se utiliza para que el programa reconozca y categorice el plugin correctamente.

```
@Step(
   id = "TripleGeoKettle",
   name = "tripleGEOStep.Name",
   description = "tripleGEOStep.TooltipDesc",
   image = "plugin/resources/tripleGeoKettle.svg",
   categoryDescription = "i18n:org.pentaho.di.trans.step:BaseStep.Category.
       Transform",
   i18nPackageName = "tripleGeoKettle",
   documentationUrl = "tripleGEOStep.DocumentationURL",
   casesUrl = "tripleGEOStep.CasesURL",
   forumUrl = "tripleGEOStep.ForumURL"
   )
```

4. Cambiar los nombres de las clases de Demo a TripleGeoKettle.

5. Importar las clases java auxiliares de tripleGeoKettle.

6. Hay un metodo deprecado en tripleGeoStepMeta.java: Valuemeta(). Se utiliza el constructor sin parámetros y luego se le asigna el tipo 2. Para sustituirlo, como el tipo es 2, que segun esta tabla[31] significa string, es necesario cambiarlo a new ValueMetaString.

7. Error en los metodos readRep y saveRep, hay que importar ObjectId y cambiar long por ObjectId en la cabecera de la función.

8. Cambiar la siguiente línea para que la clase Dialog lea correctamente los contenidos del fichero properties.

```
main/java/plugin/tripleGEOStepDialog.java
69: private static String PKG = tripleGEOStepDialog.class.getPackage().getName();
```

9. PDI9 utiliza iconos .svg y proporciona una guía de diseño[32]. Por tanto se ha actualizado el icono antiguo .png a .svg con los nuevos colores fig.2.16

Figura 2.16: Nuevo icono svg

2.4.1. Resultado parcial

La base del porteo se ha realizado correctamente como se puede ver en la figura 2.17. El dialogo se abre y se pueden cambiar los parámetros de configuración. También es capaz de leer la salida del paso anterior de SRS. Sin embargo, todavía no funciona correctamente. Probablemente se debe a que el paso anterior (transformación de coordenadas) pasa su información de manera distinta al de GeoKettle (en PDI le llegan 9 campos y en GeoKettle 8). Será necesario seguir los pasos de la documentación para conectar el debugger y solucionar el problema de NullPointerException.

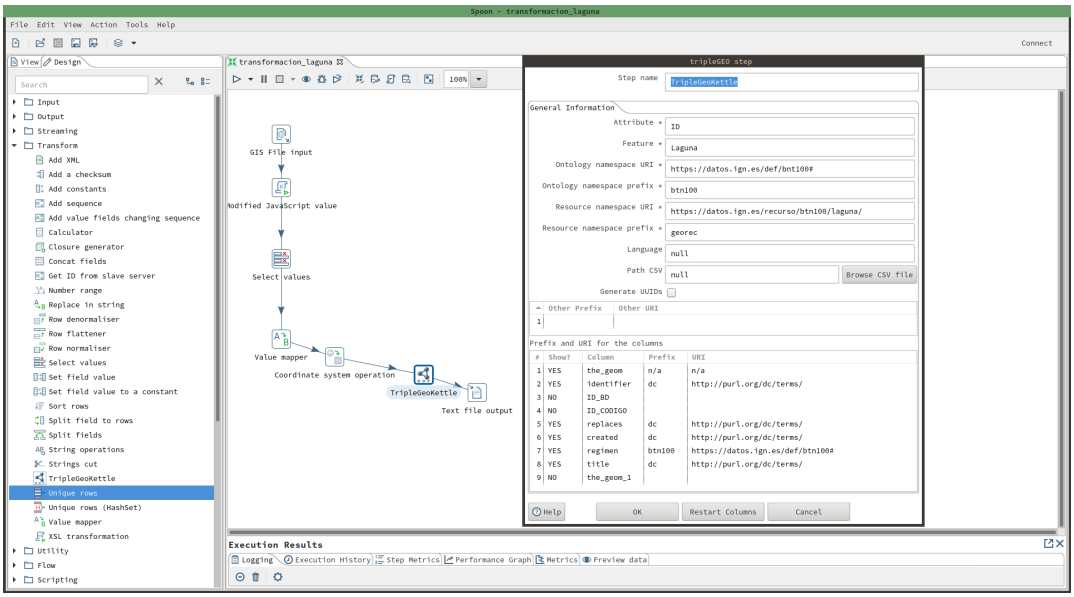


Figura 2.17: TripleGeoKettle corriendo en PDI9

Capítulo 3

Resultados y conclusiones

Resumen de resultados obtenidos en el TFG. Y conclusiones personales del estudiante sobre el trabajo realizado.

Capítulo 4

Análisis de impacto

En este capítulo se realizará un análisis del impacto potencial de los resultados obtenidos durante la realización del TFG, en los diferentes contextos para los que se aplique:

- Personal
- Empresarial
- Social
- Económico
- Medioambiental
- Cultural

En dicho análisis se destacarán los beneficios esperados, así como también los posibles efectos adversos.

Se recomienda analizar también el potencial impacto respecto a los Objetivos de Desarrollo Sostenible (ODS), de la Agenda 2030, que sean relevantes para el trabajo realizado (ver enlace)

Además, se harán notar aquellas decisiones tomadas a lo largo del trabajo que tienen como base la consideración del impacto.

Bibliografía

[1] Publicaciones utilizadas en el estudio y desarrollo del trabajo. Hay que utilizar un sistema internacional para referencias bibliográficas, de acuerdo con las indicaciones del tutor. Por ejemplo, el **sistema de IEEE**.

[2] A. de León, V. Saquicela, LM. Vilches-Blázquez, B. Villazón-Terrazas, F. Priyatna and Oscar Corcho, "Geographical Linked Data: a Spanish Use Case", in *Proceedings of the In I-SEMANTICS '10 6th International Conference on Semantic Systems*

[3] Espinoza-Arias, P., García-Delgado, M., Corcho, O. et al. "A sustainable process and toolbox for geographical linked data generation and publication: a case study with BTN100.", in *Open geospatial data, softw. stand. 4, 2 (2019).*

[4] OGC, "OGC's Role in the Spatial Standards World", in *An Open GIS Consortium (OGC) White Paper*

[5] OGC, ISO, TC/ 211, IHO, "A Guide to the Role of Standards in Geospatial Information Management", in *Fifth Session of the United Nations Committee of Experts on Global Geospatial Information Management (UN-GGIM). Held from 3-7 August 2015 at the United Nations Headquarters in New York.*

[6] Miembros del OGC, "https://www.ogc.org/ogc/members",

[7] Web de datos del Instituto Geográfico Nacional, http://datos.ign.es/

[8] Leon, Alexander de; Wisniewki, Filip; Villazón-Terrazas, Boris y Corcho, Oscar " Map4rdf - Faceted Browser for Geospatial Datasets", in *PMOD workshop, 2012*

[9] Web de Map4rdf "https://oeg-upm.github.io/map4rdf/",

[10] "ESRI Shapefile Technical Description", *An ESRI White Paper—July 1998*

[11] ESRI, "Geoprocessing considerations for shapefile output", *Arcgis Desktop 9.3 Help*

[12] OGC® GeoPackage Encoding Standard 1.3 "http://www.geopackage.org/spec/",

[13] Tim Berners-Lee "Linked Data Design Issues", in *W3C 2009*

[14] Aurelio Morales “Di no al Shapefile y sí al GeoPackage”, in *mappingis.com 2018*

[15] ESRI ArcMap 10.8 Docs, “¿Qué son las proyecciones cartográficas?”, in *https://desktop.arcgis.com/es/arcmap/latest/map/projections/what-are-map-projections.htm*

[16] Antonin Guttman “1984. R-trees: a dynamic index structure for spatial searching”, in *Proceedings of the 1984 ACM SIGMOD*

[17] “Limits In SQLite”, in *https://www.sqlite.org/limits.html*

[18] “GDAL”, in *https://gdal.org/*

[19] Tim Berners-Lee, James Hendler, Eric Miller, “Integrating Applications on the Semantic Web”, in *Journal of the Institute of Electrical Engineers of Japan, Vol 122(10), October, 2002, p. 676-680.*

[20] “¿Qué son Datos Abiertos?”, in *https://datos.madrid.es*

[21] “http://datos.ign.es/”,

[22] “OSGeo Live Wiki”,

[23] “Pentaho Data Integration 9.1 Documentation”,

[24] “TripleGeo Documentation”, in *Github*

[25] “https://github.com/atolcd/pentaho-gis-plugins”,

[26] “https://www.atolcd.com/expertise/solutions-geographiques-open-source-sig”,

[27] “https://help.pentaho.com/Documentation/8.2/Products/Data_Integration/PDI_Client”,

[28] “https://ant.apache.org/”,

[29] “https://maven.apache.org/”,

[30] “https://github.com/pentaho/pdi-sdk-plugins”,

[31] https://javadoc.pentaho.com/kettle/constant-values.html#org.pentaho.di.core.row.ValueM

[32] https://help.pentaho.com/Documentation/8.2/Developer_Center/PDI/Extend/035ns,

[33] ab “cd”, in *ef*

Apéndice A

Anexos

A.1. Glosario de términos

- *OGC*: Open Geospatial Consortium
- *GIS*: Geographic Information System
- *ESRI*: Environmental Systems Research Institute
- *IGN*: Instituto Geográfico Nacional
- *GDAL*: Geospatial Data Abstraction Library
- *URI*: Uniform Resource Identifier
- *RDF*: Resource Description Framework
- *SPARQL*: SPARQL And Rdf Query Language
- *ETL*: Extract, Transform and Load
- *KETTLE*: Kettle Extraction Transformation Transport Load Environment
- *PDI*: Pentaho Data Integration
- *ab*: cd