Universidad Politécnica
de Madrid

Escuela Técnica Superior de Ingenieros Informáticos

Grado en Ingeniería Informática

Trabajo Fin de Grado

Introducir Soporte al Formato GeoPackage en Herramientas de Linked Data Geográfico Desarrolladas por el Grupo de Ingeniería Ontológica

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

Madrid, Marzo 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: Introducir Soporte al Formato GeoPackage en Herramientas de Linked Data Geográfico Desarrolladas por el Grupo de Ingeniería Ontológica

Marzo 2021

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

Resumen

«Aquí va el resumen del TFG. Extensión máxima 2 páginas.»

Abstract

«Abstract of the Final Degree Project. Maximum length: 2 pages.»

Tabla de contenidos

Capítulo 1

Introducción

La introducción del TFG debe servir para que los profesores que evalúan el Trabajo puedan comprender el contexto en el que se realiza el mismo, y los objetivos que se plantean.

1.1. Objetivos

El objetivo principal del trabajo es introducir soporte al formato GeoPackage en 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]

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 objetivo de este trabajo es el de dar soporte GeoPackage para las herramientas normalmente utilizadas para este tipo de tareas.

1. Dar soporte GeoPackage a la herramienta Map4RDF.
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.

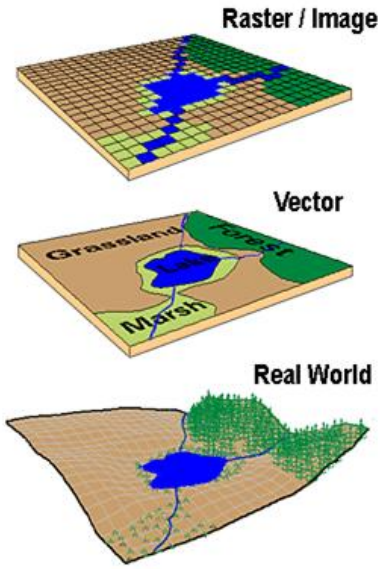

Figura 1.1: Representación del terreno mediante vectores y raster

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.
- 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] 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]

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.

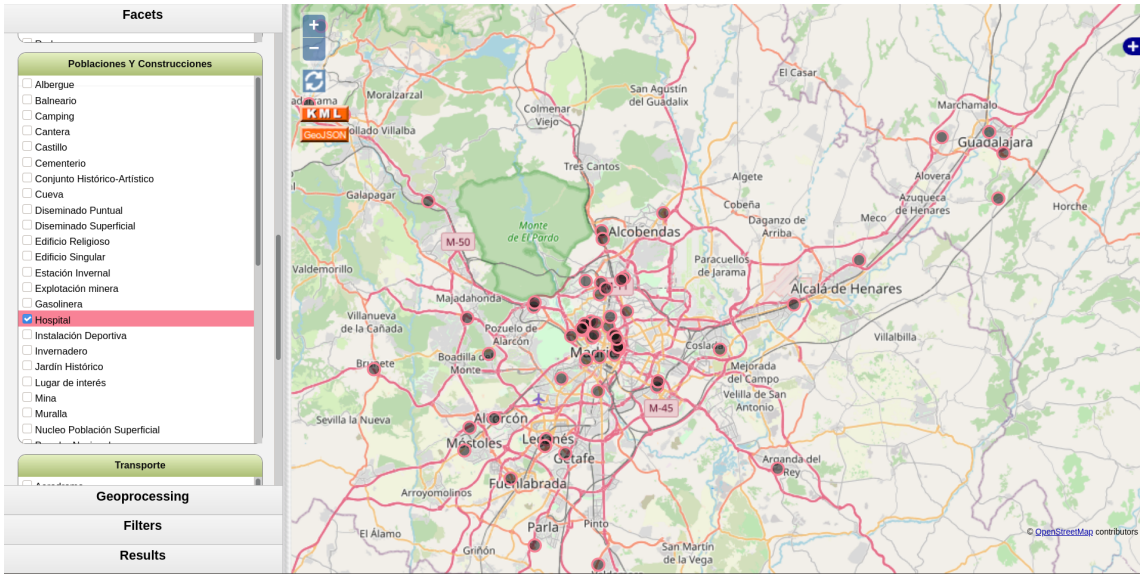

Figura 1.2: http://certidatos.ign.es/map/ que implementa Map4Rdf

1.2.5. GeoKettle

GeoKettle es una 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]

[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

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

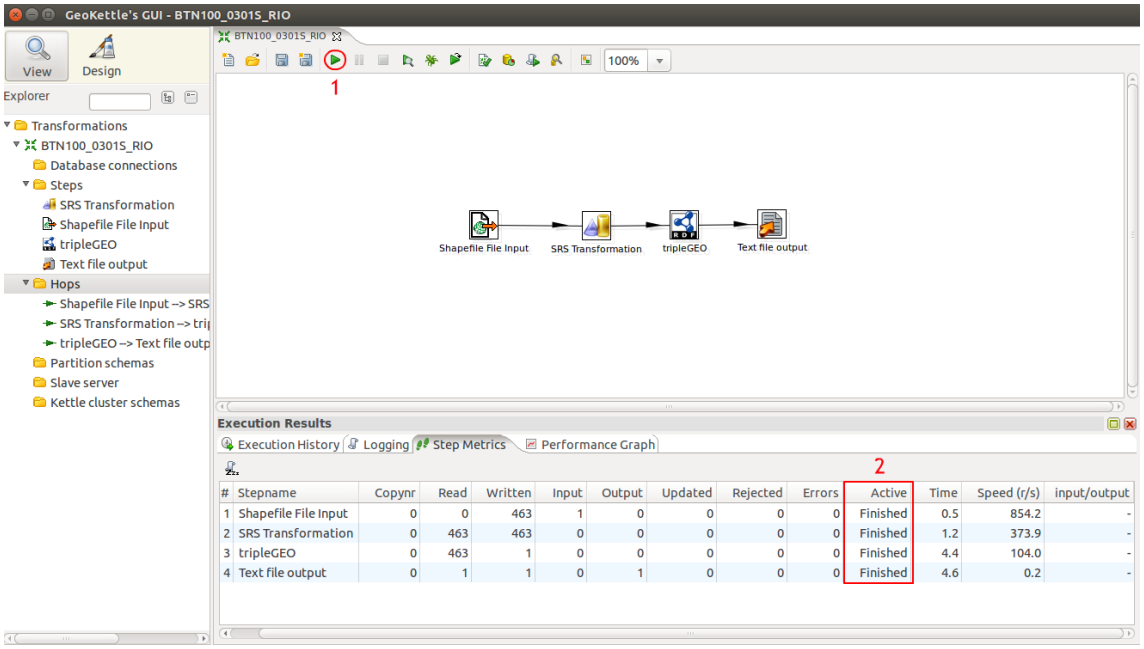

Figura 1.3: TripleGeoKettle en funcionamiento, TripleGeoKettle wiki

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. Por tanto, se comenzará actualizando 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 disponible en el marketplace llamado pentaho-gis-plugins. Se actualizará la primera fase a: **"replicar la funcionalidad y las transformaciones de GeoKettle + TripleGeo en la suite PDI."** Si es sencillo, se considerará también dar soporte a GeoPackage.

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.

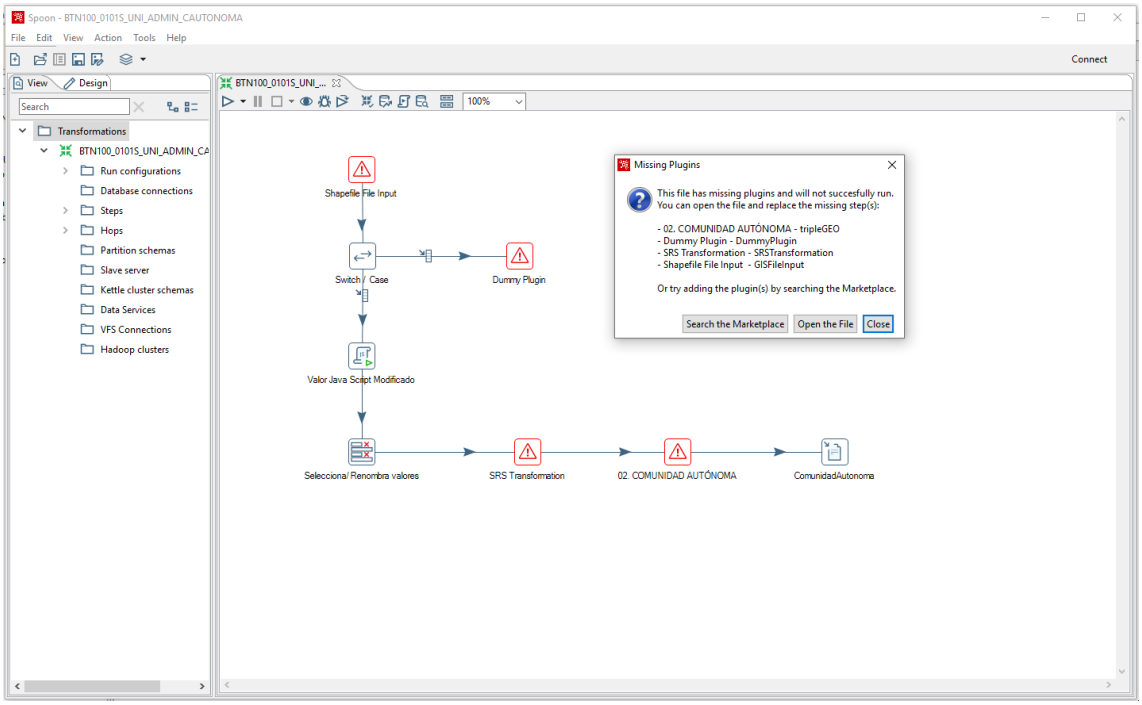

Figura 2.1: Workflow importado en la nueva suite

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] 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