Conduciendo la información estadística georreferenciada del DANE a otra dimensión

Leading to DANE´s georeferenced statistic information to other dimension

Luis M. Vilches-Blázquez[1], Julián Mauricio Alvarado Torres[2]

Cómo citar este artículo: Vilches-Blázquez, L. y Alvarado Torres, J. (2017). Conduciendo la información estadística georreferenciada del DANE a otra dimensión. *Análisis Geográficos, 52,* 63-73.

Resumen

La web Linked Data supone un nuevo paradigma que pretende explotar la web como un espacio global de información. La aplicación de los principios de esta nueva web a la información estadística georreferenciada del DANE permitirá superar las barreras actuales en los procesos de publicación ortodoxa, logrando una integración e interoperabilidad semántica de los datos.

En este trabajo se presentan las características de un caso de estudio para la generación y publicación de datos estadísticos georreferenciados conforme a Linked Data. Este prototipo representa los cimientos para la integración de datos abiertos gubernamentales, conjuntos de datos heterogéneos de la web Linked Data y los procesos de innovación que van a caracterizar al DANE moderno.

Palabras clave: información estadística georreferenciada, Web 3.0, integración, Linked Data, innovación.

Abstract

The Web of Linked Data is a new paradigm that enables to explode the Web as a global information space. The principles of Linked Data associated with DANE's geo-statistical information will enable to overcome current barriers in the information delivering process, and will reach a semantic integration and interoperability of data.

In this paper we present the characteristics of a study case for the generation and publication of geo-statistical data according to Linked Data. This prototype is the starting point for integrating Open Government Data, heterogeneous datasets of the Linked Data Web and the innovation processes of the modern DANE.

Keywords: *georeferenced statistical information, Web 3.0, integration, Linked Data, innovation.*

[1] *Departamento Administrativo Nacional de Estadística (DANE), Dirección de Geoestadística, Bogotá, Colombia. Correo: lmvilches@dane.gov.co.*
[2] *Departamento Administrativo Nacional de Estadística (DANE), Dirección de Geoestadística, Bogotá, Colombia. Correo: jmalvaradot@dane.gov.co.*

Introducción

La demanda de información estadística se está convirtiendo en una necesidad de primer orden como consecuencia del potencial de estos datos para facilitar los procesos de análisis vinculados a la toma de decisiones por parte de un amplio espectro de usuarios. Sin embargo, la débil estructuración y heterogeneidad en el nivel de detalle, formatos y vocabularios utilizados son problemas que trascienden cuando la producción de información estadística está centrada en las necesidades específicas de cada investigación. Ante esta situación, hoy en día resulta imperioso disponer de modelos comunes y fácilmente procesables de información.

El Departamento Administrativo Nacional de Estadística (DANE) está empeñado en aprovechar y explotar las ventajas y posibilidades de los más recientes y poderosos paradigmas de información. En este sentido, ha decidido poner en marcha un proyecto para combinar la información estadística georrefenciada con la iniciativa Linked Data. Esta iniciativa se refiere a una nueva forma de publicar y enlazar datos para representar información en la Web 3.0 (también conocida como web semántica), utilizando Resource Description Framework (RDF), un lenguaje para presentar información sobre recursos, propuesto por el Consorcio de la World Wide Web en el área de la web semántica. Así, la iniciativa Linked Data supone un nuevo paradigma que pretende explotar la web como un espacio global de información en el que la navegación se realiza a través de datos estructurados enlazados (Linked Data), en lugar de realizarse a través de documentos. De esta manera, los servicios de información estadística georreferenciada institucionales superan las barreras actuales en los procesos de publicación ortodoxa, logrando una integración e interoperabilidad semántica de los datos.

En este trabajo se presentan las características de un caso de estudio para la generación y publicación de datos estadísticos georreferenciados conforme a Linked Data. Este prototipo constituye los cimientos para la integración de datos abiertos gubernamentales, conjuntos de datos heterogéneos de la web de Linked Data y los procesos de innovación que van a caracterizar al DANE moderno. Así, primero se presenta el *workflow* del proceso de transformación a Linked Data de la información estadística georreferenciada con la que se trabaja en este proyecto. Luego se describen los detalles del proceso de generación y publicación de Linked Data del DANE, y finalmente se exponen las conclusiones y trabajos futuros.

Workflow del proceso de transformación de la información estadística georreferenciada

El desarrollo de *workflow* para la transformación de la información estadística georreferenciada del DANE a Linked Data se basa en la propuesta metodológica para la generación y publicación de datos estructirados entrelazados descrita por Sauermann *et al.* (2008). Esta guía metodológica propone un modelo de ciclo de vida incremental iterativo basado en continuas mejoras y extensiones del Linked Data generado. La referida metodología contempla las siguientes actividades: 1) especificación, 2) modelado, 3) generación de RDF, 4) generación de links, 5) publicación y 6) explotación. Cada una de estas actividades está compuesta de una o más tareas. La figura 1 muestra una visión

general de *workflow* del proyecto, asociado con las actividades que recoge la metodología mencionada.

A continuación se detallan los componentes que conforman el *workflow* del proyecto presentado en la figura 1. Cabe anotar que los componentes que se describen en este artículo son los que se han abordado en el contexto del proyecto hasta la fecha de presentación de este trabajo.

Generación y publicación de Linked Data del DANE

Esta sección describe las diferentes actividades y tareas propuestas en el *workflow*, asociadas con la metodología utilizada por Vilches-Blázquez *et al*. (2014), realizadas durante el proceso de generación y publicación del proyecto Statistical Linked Data del DANE.

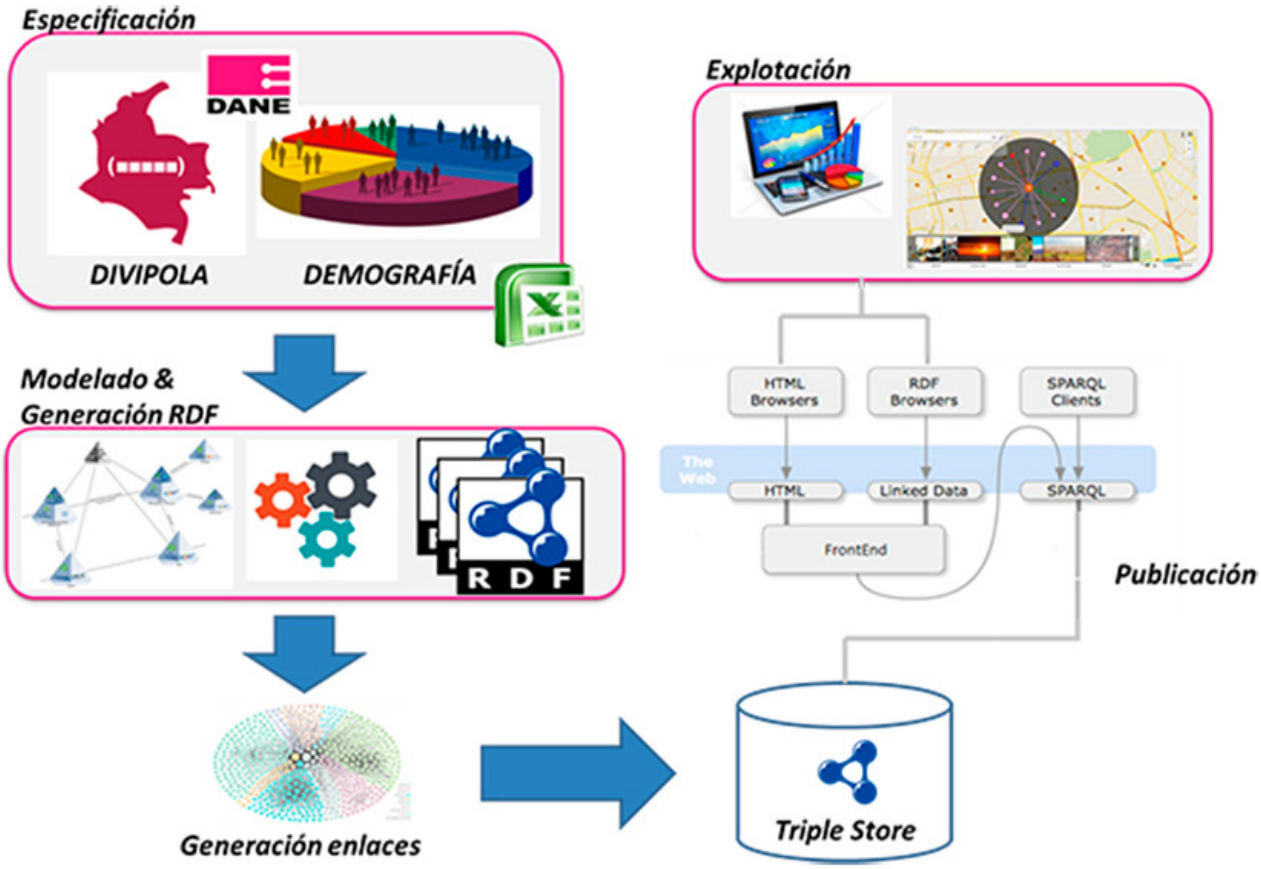

Figura 1. *Workflow* del proyecto Statistical Linked Data
Fuente: elaboración propia.

Especificación

En esta sección se describe el análisis de las fuentes de datos que forman parte del proyecto, los patrones adoptados en el diseño de los identificadores de recursos (URI) y la definición de la licencia asociada a los datos.

Análisis de las fuentes de datos

La información del DANE, utilizada para el desarrollo de este trabajo, está conformada por un conjunto de ficheros en formato Microsoft Excel que contienen datos sobre la división político-administrativa de Colombia (Divipola) y estadísticas relacionadas con demografía y población.

La Divipola es un estándar de codificación que permite contar con un listado organizado y actualizado de la totalidad de unidades en las que está dividido el territorio nacional, dándole a cada departamento, municipio, corregimiento departamental y centro poblado el máximo de estabilidad en su identificación.

Esta codificación, acorde con la dinámica territorial del país, está disponible en la web (DANE, 2015) y es actualizada periódicamente por el DANE de acuerdo con la información suministrada por las administraciones municipales y departamentales, constituyéndose en una fuente de consulta sobre la organización administrativa y política del país.

En relación con la información estadística asociada al tema demografía y población, se trabaja con datos asociados al último censo de población publicado hasta la fecha (2005), series y proyecciones de población. Entre los diversos conjuntos de datos, se encuentran variables estadísticas como: población censada, tasa de natalidad, tasa de fecundidad, esperanza de vida al nacer, viviendas ocupadas, hogares, etc.

Diseño de URI

Con el objetivo de llevar a cabo la transformación de los datos del DANE a Linked Data, una de las principales decisiones, previas al proceso de transformación de las fuentes de información a RDF, es el formato o patrón en que los identificadores de las instancias (URI) van a ser generados. Las URI son extremadamente relevantes en este proceso, ya que contribuirán de manera clave en el alineamiento de instancias provenientes de diferentes fuentes de información. Por ello, en esta tarea se genera un patrón de URI para el conjunto de datos estudiados. La realización de esta tarea se da de conformidad con el principio de Linked Data señalado por Sauermann *et al.* (2008), que propone la utilización de URI para identificar recursos. Para el diseño del patrón de las URI que identifican los datos en el contexto de este trabajo, también se adoptaron las recomendaciones y buenas prácticas señaladas por Sauermann *et al.* (2008) y por Davidson (2009). A continuación se recogen los principales detalles del patrón:

Raíz de las URI. Se adopta como raíz de las URI *<<http://geoportal.dane.gov.co/linkedstat/>>.*

A su vez, este será el dominio donde se publicará toda la información generada en el marco de este trabajo.

Ontología (modelo). El patrón adoptado para la identificación de un recurso (fenómeno geográfico) modelado en las diferentes ontologías utilizadas es el siguiente:

Patrón:
http://geoportal.dane.gov.co/linkedstat/voc/{nombre ontología}/{concepto|propiedad}
Ejemplo:
http://geoportal.dane.gov.co/linkedstat/voc/DIVIPOLA#Departamento

Datos (instancias). Para identificar los recursos asociados a los datos (instancias) se adopta el siguiente patrón:

Patrón:
http://geoportal.dane.gov.co/linkedstat/recurso/{tipo de recurso}/{nombre de recurso}
Ejemplo:
http://geoportal.dane.gov.co/linkedstat/recurso/Departamento/25

Así mismo, para identificar la geometría asociada a los diferentes recursos relacionados con la Divipola (centroides) se adopta el siguiente patrón:

Patrón:
http://geoportal.dane.gov.co/linkedstat/recurso/lat_long
Ejemplo:
http://geoportal.dane.gov.co/linkedstat/recurso/-74,6695456055_4,37585313348

Sobre este patrón de URI asociado a la información geométrica, merece ser destacado que los recursos se caracterizan por presentar la información geométrica conforme a GeoSPARQL serializada como Well-Known Text (WKT) o Geography Markup Language (GML). Asimismo, la geometría de estos recursos se representa en el sistema de referencia WGS84.

Definición de la licencia

Según acuerdo, y previo estudio de las características de la puesta a disposición de la información al ciudadano por parte del DANE, la licencia que se utiliza en la publicación del Linked Data del DANE es del tipo Creative Commons en su acepción *by,* es decir, por reconocimiento. Esta licencia permite cualquier explotación de la obra, incluyendo una finalidad comercial, así como la creación de obras derivadas, la distribución de las cuales también está permitida sin ninguna restricción (Creative Commons, 2015).

Modelado

Para modelar la información contenida en los conjuntos de datos se ha creado una red de ontologías, que es una colección de ontologías unidas a través de una variedad de diferentes relaciones. Esta red se ha desarrollado siguiendo la metodología de NeOn señalada por Suárez-Figueroa (2010), mediante la reutilización de ontologías y vocabularios existentes. Este trabajo está asociado con la actividad de *modelado* propuesta en la metodología mencionada. La figura 2 presenta el modelo de alto nivel de la red de ontologías DANE.

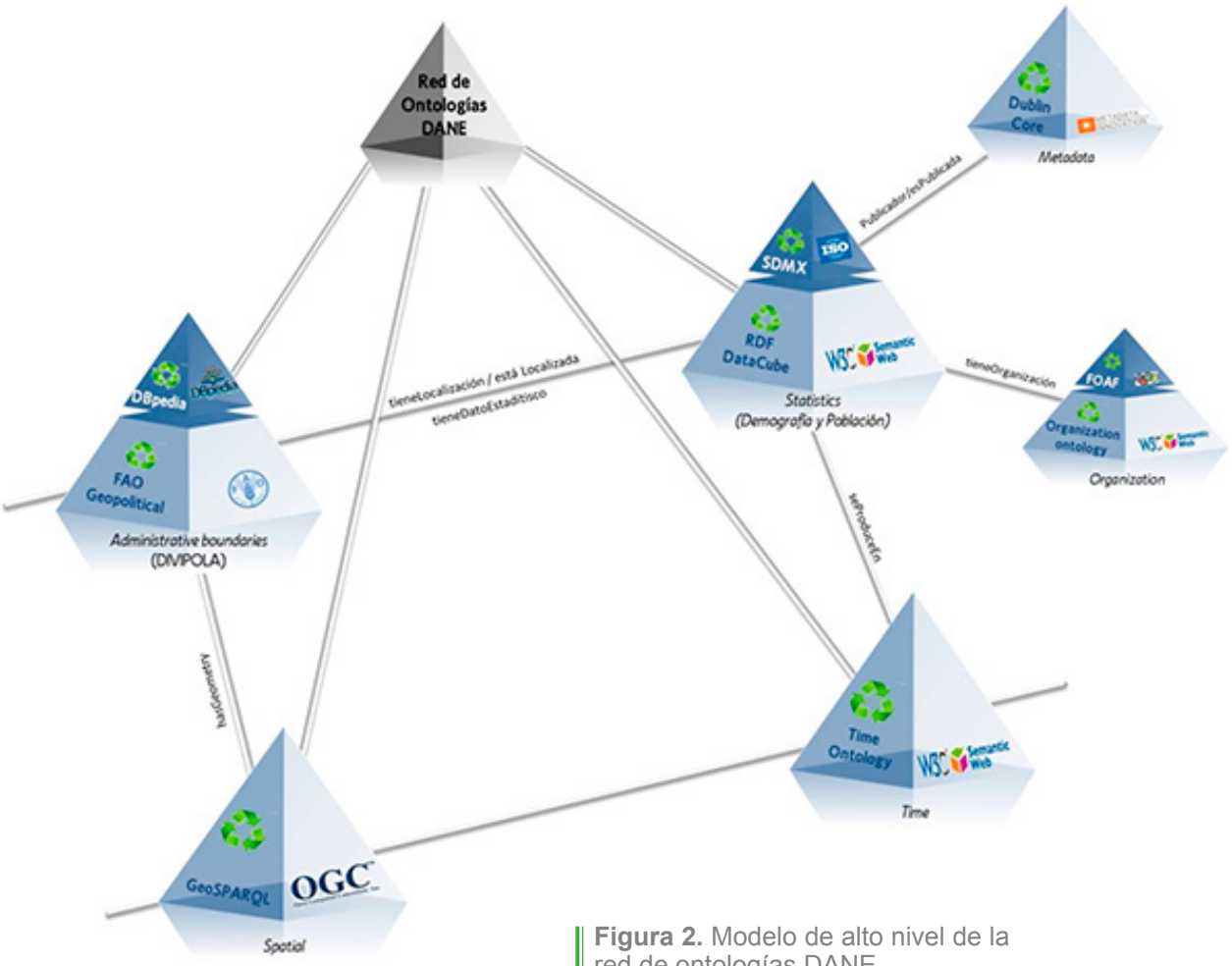

Figura 2. Modelo de alto nivel de la red de ontologías DANE
Fuente: elaboración propia.

En esta figura se puede observar que la red de ontologías está compuesta por cuatro módulos diferentes que se corresponden con los principales temas identificados en el proyecto, es decir, división administrativa, estadística, (geo)espacial y tiempo.

El módulo asociado con la *división político-administrativa* del territorio *(administrative boundaries)* está compuesto por las siguientes ontologías:

- FAO geopolitical. Esta ontología ha sido desarrollada por la Organización de las Naciones Unidas para la Agricultura y la Alimentación (FAO) para facilitar el intercambio y distribución de datos de manera estandarizada entre sistemas de gestión de información sobre países o regiones; para el efecto incluye información en inglés acerca de continentes, regiones, países, etc.
- DBpedia recoge un esfuerzo de la comunidad por extraer información estructurada de Wikipedia y disponer de esta información en la web. Esta iniciativa pretende inspirar nuevos mecanismos para la navegación, la vinculación y la mejora de la propia enciclopedia.

Sobre la reutilización de estos recursos ontológicos, en este módulo se añade el conocimiento asociado a la división político-administrativa de Colombia, proveniente de los recursos no ontológicos asociados a la Divipola.

El módulo asociado con la *información estadística* y, más concretamente, con la información relacionada con el tema demografía y población, reutiliza los siguientes recursos:

- RDF Data Cube. Este vocabulario, propuesto como un estándar de W3C[1], permite describir y publicar datos estadísticos multidimensionales haciendo uso del estándar Resource Description Framework (RDF) del W3C.

  El modelo subyacente al vocabulario Data Cube es compatible con el modelo que subyace en Statistical Data and Metadata eXchange (SDMX). El vocabulario Data Cube es una base fundamental que apoya la extensión de vocabularios para permitir la publicación de otros aspectos de los flujos de datos estadísticos u otros conjuntos de datos multidimensionales.

- SDMX. La iniciativa SDMX fue creada para tratar con mayor eficiencia la práctica estadística. En la actualidad esta iniciativa es un estándar de la International Organization for Standardization (ISO) para intercambiar y compartir datos y metadatos estadísticos entre organizaciones.

Sobre este estándar existe un vocabulario desarrollado para tratar los cubos de datos y apoyar la publicación de datos estadísticos en RDF utilizando un modelo de información basado en el estándar mencionado.

Junto a la reutilización de estos recursos ontológicos, en este módulo se añade el conocimiento asociado a la información estadística relacionada con el tema demografía y población, objeto de trabajo en este proyecto.

El módulo asociado con la *información espacial* se caracteriza por la reutilización de la propuesta proveniente del Open Geospatial Consortium (OGC). Este consorcio es el encargado de liderar los temas relacionados con la información geográfica y su interoperabilidad. En este sentido, se procede a reutilizar la ontología GeoSPARQL en su totalidad.

- GeoSPARQL. Es una especificación propuesta por el OGC, orientada a la representación y consulta de los datos espaciales en la web semántica. Esta propuesta define un vocabulario para la representación de información espacial en RDF/OWL.

[1] *W3C (World Wide Web Consortium) es un consorcio internacional que genera recomendaciones y estándares que aseguran el crecimiento de la World Wide Web a largo plazo (tomado de https://es.m.wikipedia.org).*

El módulo asociado con la *información temporal* se caracteriza por la reutilización de la propuesta proveniente del World Wide Web Consortium (W3C).

- W3C Time Ontology. Esta ontología de conceptos temporales desarrollada en el contexto del W3C proporciona un vocabulario para expresar hechos sobre las relaciones topológicas entre los instantes y los intervalos, junto con información sobre duración y fecha-hora.

Junto a los módulos anteriores, propiamente vinculados con los diferentes dominios de este proyecto, se añaden dos módulos adicionales a la red de ontologías del DANE que sirven para proporcionar descripciones adicionales sobre la información, que será objeto de su publicación conforme a los principios de Linked Data y sobre la organización que los proporciona, en este caso, el DANE.

Con respecto al módulo que recoge los elementos *(metadatos)* para la descripción de la información en la red de ontologías, este reutiliza la propuesta de Dublin Core.

- Dublin Core. Esta iniciativa recoge un modelo de metadatos elaborado y auspiciado por la Dublin Core Metadata Initiative (DCMI), una organización dedicada a fomentar la adopción extensa de los estándares interoperables de los metadatos y a promover el desarrollo de los vocabularios especializados de metadatos para describir recursos que permitan a los sistemas el descubrimiento de recursos.

El módulo que recoge los elementos para la *descripción de la organización,* en este caso, el DANE, está compuesto por las siguientes ontologías:

- The organization ontology. Esta propuesta del W3C recoge una ontología fundamental para estructuras, destinadas a apoyar la publicación de Linked Data sobre la información de las organizaciones a través de diferentes dominios. Esta ontología está diseñada para permitir las extensiones específicas de dominio y posibilitar la adición de información relacionada con clasificación, actividades y roles organizacionales.

- FOAF vocabulary. Esta especificación describe el lenguaje Friend of a Friend (FOAF), definido como un diccionario de propiedades y clases nombradas utilizando la tecnología RDF del W3C.

Generación de RDF

El objetivo de esta actividad es la generación de RDF de las fuentes de información asociadas a este proyecto para transformar los datos originales a un formato estándar e interoperable en el contexto de la web semántica. Además, este proceso permite transformar los datos del DANE en datos estructurados y en formato no propietario. Para ello, se ha utilizado Open Refine, una aplicación de escritorio de código abierto que admite la limpieza y transformación de datos a otros formatos, entre ellos RDF.

En la figura 3 se recoge una representación gráfica de los diferentes componentes del RDF generado de la Divipola conforme a GeoSPARQL y, por tanto, a la red de ontologías desarrollada en el contexto del proyecto.

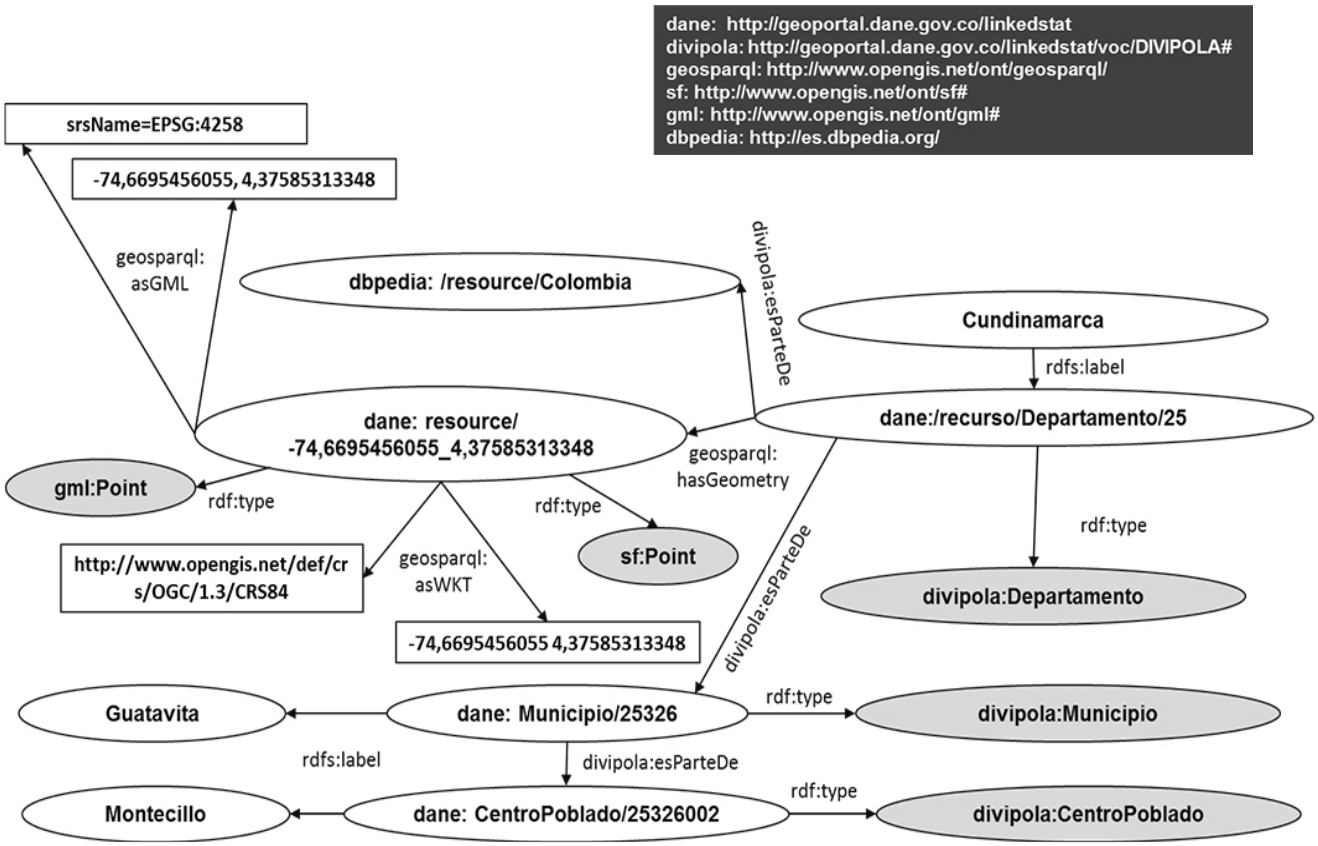

Figura 3. Ejemplo de grafo RDF generado de Divipola
Fuente: Elaboración propia.

Este trabajo va a permitir, por un lado, que el RDF generado en el contexto del proyecto sea conforme a los diferentes vocabularios y especificaciones para el tratamiento de la información estadística georreferenciada en el contexto de la web semántica. Además, dicho RDF es consistente con las buenas prácticas existentes en la comunidad geoespacial, al incluir en el RDF generado la geometría en formato WKT y GML.

Generación de enlaces

El objetivo de esta actividad es la generación de relaciones (enlaces) entre los datos del DANE en formato RDF y DBpedia (2015). Este trabajo se relaciona con la actividad de generación de links de la metodología utilizada y va a permitir que los datos del DANE aumenten su navegabilidad en la web de Linked Data, así como un enriquecimiento de los mismos mediante la incorporación de descripciones adicionales.

Para la generación de enlaces entre los datos del DANE y DBpedia se utilizan las funciones de reconciliación proporcionadas por Open Refine. Esto permite conectar los datos originales con los suministrados por DBpedia a través de su SPARQL Endpoint y, de esta manera, enriquecer los datos del DANE.

Conclusiones y trabajo futuro

En este artículo se describen los principales detalles del proceso en curso para la generación y publicación de Linked Data de datos del DANE. Asimismo, en este trabajo se muestra cómo la información estadística georreferenciada puede aprovechar las ventajas y posibilidades de los más recientes y poderosos paradigmas de información mediante su interacción con la iniciativa Linked Data.

En definitiva, el trabajo realizado permite que los datos del DANE con los que se ha trabajado adquieran mayor expresividad y significado, gracias a la publicación de los mismos conforme a los principios de Linked Data. Esto supone que el proyecto lleva los datos del DANE hacia una nueva dimensión, caracterizada por superar las barreras actuales en los procesos de publicación ortodoxa, logrando una integración e interoperabilidad semántica de los datos.

En cuanto al trabajo futuro, el proyecto se va a centrar en culminar las fases de publicación y explotación para poner a disposición y permitir al usuario acceder a estos datos de una manera accesible y amigable. En este sentido, los resultados actuales de estos avances se encuentran en el sitio web Geoportal del DANE.

Bibliografía

Creative Commons (2015). *Explicación de las licencias.* Recuperado de http://es.creativecommons.org/license/.

Davidson, P. (2009) *Designing URI Sets for the UK Public Sector. A report from the Public Sector Information Domain of the CTO Council's cross-Government Enterprise Architecture UK Chief Technology Officer Council.* Recuperado de https://www.gov.uk/government/uploads/system/uploads/attachment_data/file/60975/designing-URI-sets-uk-public-sector.pdf.

DBpedia (2015). *About.* Recuperado de http://wiki.dbpedia.org/about.

Departamento Administrativo Nacional de Estadística (DANE) (2015). *Codificación de la División Político-administrativa de Colombia (Divipola).* Disponible en http://geoportal.dane.gov.co:8084/Divipola/.

Departamento Administrativo Nacional de Estadística (DANE) (2016). *Geoportal DANE.* Disponible en: https://geoportal.dane.gov.co/v2.

Sauermann, L., Cyganiak, R., Ayers, D. & Völkel, M. (2008). *Cool URIs for the Semantic Web.* W3C Interest Group Note 20080331. Recuperado de https://www.w3.org/TR/2008/NOTE-cooluris-20080331/.

Suárez-Figueroa, M. C. (2010). *NeOn Methodology for Building Ontology Networks: Specification, Scheduling and Reuse* (PhD Thesis). Universidad Politécnica de Madrid, España.

Vilches-Blázquez, L. M., Villazón-Terrazas, B., Corcho, O. & Gómez-Pérez, A. (2014). Integrating geographical information in the Linked Digital Earth. *International Journal of Digital Earth, 7*(7), 554-575. doi: http://dx.doi.org/10.1080/17538947.2013.783127.