Generación de Datos Semánticos a partir de una Base de Datos Relacional de una Institución de Educación Superior

Freddy Tapia León, Walter Fuertes
Universidad de las Fuerzas Armadas - ESPE, Quito, Pichincha, Ecuador, {fmtapia, wmfuertes}@espe.edu.ec

ABSTRACT

This research handles the analysis and migration of data from a public university, which are stored in relational databases RDB to a Resource Description Framework RDF metadata. For this purpose, we have mapped each of the tables that comprised the database deployment to RDF, by mean of standardized vocabularies or related to the Semantic Web development, allowing its interpretation and reuse in order to publish readable data by Web applications.

1. INTRODUCCIÓN

De acuerdo con IBM (IBM, 2011), cada día se generan alrededor de 2.5 quintillones de bytes de datos (10^{30}), donde el 90% han sido creados en los últimos tres años. A su gestión eficiente se le conoce como Big Data, que es una nueva oportunidad para proponer ideas, que facilitan el analisis de datos y sus contenidos, para una toma de decisión más ágil. Por este motivo, algunos gobiernos están abriendo sus plataformas de información con el propósito de publicar y transparentar su gestión, lo que da lugar al término Open Data (De la Fuente, 2013).

Esta investigación se enfoca en el análisis y migración de información académica, almacenada en bases de datos relacionales (RDB) hacia metadatos tipo Resource Description Framework (RDF). Para llevarlo a cabo se han desarrollado mapeos a cada una de las tablas posibilitando su interpretación y reutilización (i.e., dotarles de contenido semántico), por medio de vocabularios estandarizados o relacionados al desarrollo de la Web Semántica. Como contribución se ha implementando un proceso inédito, combinando técnicas para mejorar los datos de la Web tradicional.

El resto del artículo ha sido organizado como sigue: La sección 2 describe el fundamento teórico. La sección 3 muestra los componentes del experimiento, asi como el proceso de implementación relacionado a Open Data. La sección 4 ilustra los resultados obtenidos. Finalmente en la sección 5 se exponen las conclusiones y trabajo futuro.

2. MARCO REFERENCIAL

2.1 WEB SEMÁNTICA

Es un gran repositorio, que permite la interoperabilidad entre diferentes sistemas, debido al significado semántico que ésta contiene. Esta caracterísitca facilita la ubicación e identificación de los datos dentro de esta Web (W3C, 2009). En esta investigación, se han combinado novedosas técnicas como: *i) RDF*, el cual es un modelo de datos para la representación de metadatos en la Web Semántica, proporcionando información descriptiva sobre los recursos que se encuentran en la Web; *ii) Uniform Resource Identifier (URI)*, que son cadenas de texto que proporcionan la identificación de recursos; *iii) RDF Schema*, que es una ontología básica la cual define propiedades y clases de recursos RDF (W3C, 2004); *iv) EL Lenguaje de Ontologías Web (OWL)*, que está construido sobre RDF y RDF Schema, con la finalidad de dotarle de contenido semántico a las ontologías descritas por este lenguaje (W3C, 2004); *v) SPARQL*, que es un lenguaje de consultas para datos RDF; *vi) LINKED DATA*, que son las diferentes interconexiones que tienen los nodos dentro de la Web Semántica.

2.2 ONTOLOGÍA DBPEDIA

Esta ha permitido garantizar una adecuada reutilización de los datos, por cuanto utiliza una representación formal y consensuada de conceptos y relaciones para la extracción de los datos (DBpedia), por esta razón el posicionamiento que ocupa dentro de Linked Data. Para el desarrollo del presente trabajo se utilizó esta ontología, por el número de propiedades y clases definidas, así como también por el nivel de estandarización y métodos de extracción que permite.

3. MATERIALES Y MÉTODOS

La Fig. 1, muestra la topología experiencial. Inicia con el análisis de los datos (RDB), y su subsecuente articulación con la ontología DBpedia por medio de los vocabularios y parámetros de mapeos establecidos para cada tabla, posibilitando el mapeo manual y su posterior obtención de datos abiertos

(RDF). En este análisis se determina que tablas y datos pueden ser migrados y cuáles no. Luego, conforme a lo establecido por el movimiento Open Data, se pueden reutilizar los metadatos por parte de todos los actores de la comunidad de Internet.

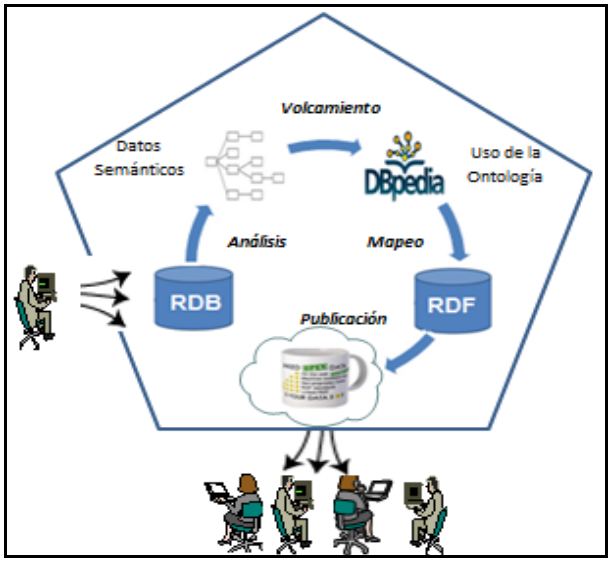

Figura 1. Topología de Implantación

3.1 Método De Implementación

El método utilizado para la obtención de datos abiertos RDF, consta de cuatro etapas:

a) *Análisis de los Datos:* Se usó un gestor de Base de Datos (MySQL), para el proceso de migración y almacenamiento de los nuevos datos.
b) *Generación de mapeos:* Se usaron las clases y propiedades definidas por la ontología DBpedia, así como también otros estándares como: Dublin Core (DC), Friend of a Friend (FOAF), entre otros.
c) *Obtención de datos RDF:* Para esta etapa se realizó el volcamiento de los datos a formato RDF, lo que permitió su posterior publicación y reutilización. Estos estarían representados en forma de tripletas y contienen toda la información de la RDB original.
d) *Visualización de los datos:* Se ejecutó consultas por medio de lenguaje SPARQL EndPoint (FrontEnd), que permite a los usuarios acceder y consultar datos RDF mediante el lenguaje SPARQL.

Para una prueba de concepto e información complementaria, como evidencia del proceso realizado favor acceder al siguiente enlace: http://es.dbpedia.org/Wiki.jsp?page=Informaci%C3%B3n%20t%C3%A9cnica, último apartado.

4. Evaluación de Resultados

La Fig. 2 muestra los resultados generados por medio de búsquedas secuenciales. Como se puede apreciar, el bloque superior contiene todos los vocabularios definidos en la etapa de mapeo inicial. En el bloque central se ingresan las instrucciones SPARQL (búsqueda y filtrado). Finalmente, en el bloque inferior se muestran los resultados obtenidos.

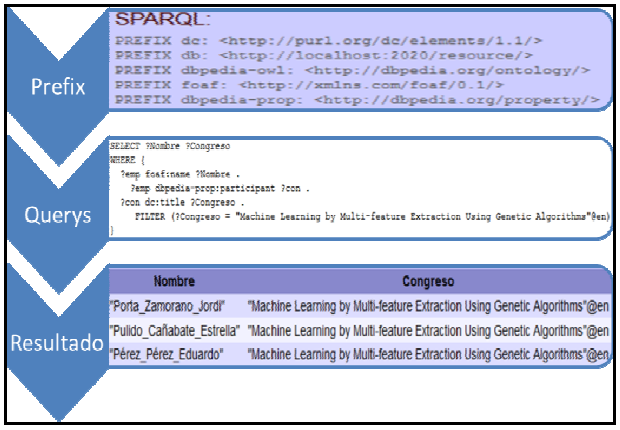

Figura 2: Publicación de datos legibles y reusables.

5. Conclusiones

En este proyecto se aplicó las tecnologías de la Web Semántica, obteniendo como resultado la generación de datos semánticos a partir de una RDB de una universidad pública. Se modeló adecuadamente los datos, factor clave para una apropiada migración de las tablas. La ontología utilizada y mapeos definidos garantizaron la reutilización y obtención de datos abiertos. Los resultados fueron publicados como metadatos legibles en la Web.

Como trabajo futuro se planean probar otros métodos de extracción, usabilidad y seguridad.

Referencias

IBM (2011). "Big Data at the Speed of Business", http://www-01.ibm.com/software/data/bigdata/, 04/28/14 (date accessed).

De la Fuente and Álvarez (2013). "La Promesa del Gobierno Abierto". Article Open linked data: http://www.lapromesadelgobiernoabierto.info/lpga.pdf, 04/28/14 (date accessed).

W3C (2004). "OWL Web Ontology Language Overview", http://www.w3.org/TR/owl-features/, 04/28/14 (date accessed).

W3C (2009), "Guía Breve de la Web Semántica", http://www.w3c.es/Divulgacion/GuiasBreves/WebSemantica, 04/28/14 (date accessed).