Estado del arte

**Internacional**

***Querying Geospatial Data over the Web: a GeoSPARQL Interface***

En este artículo se describe cómo es que Nancy, Ralhp y Dave crearon e implementaron una interfaz para datos GeoSPARQL llamada *GeoQuery TOOL*. Esta interfaz intuitiva pretende hacer que las consultas geoespaciales sean más fáciles de hacer al implementar listas en su interfaz para poder escoger atributos y operadores espaciales. Con base en los datos de entrada que haya ingresado el usuario, *GeoQuery* genera código GeoSPARQL automáticamente, realiza la consulta usando el *triple store* Parliament y es desplegado en una aplicación web en vez de utilizar el Sistema de Información Geográfica (GIS, por sus siglas en inglés) [17]

***Enabling the Geospatial Semantic Web with Parliament and GeoSPARQL***

Este trabajo de Robert y Dave se presentan razones por las cuáles hay que usar GeoSPARQL, su estado del arte en la industria y en la investigación, y su implementación de GeoSPARQL en el *triple store* Parliament. Explican conceptos geoespaciales tales como las diferencia entre una característica y una geometría, qué es un sistema de referencia de coordenadas (CRS, por sus siglas en inglés) y las relaciones topológicas que existen. En esta última mencionan 8 operaciones básicas y sus 2 variantes Egenhofer y RCC8. Ambas expresan las mismas operaciones al ser equivalentes. En general, este documento da las herramientas para comprender GeoSPARQL, así como su uso e implementación con la intención de que empresas u organizaciones consideren adoptar esta tecnología [18].

***Strategies for Executing Federated Queries in SPARQL1.1***

En esta propuesta se analizan diferentes estrategias para implementar consultas federadas con la intención de evitar los límites que un *endpoint* presenta. Las estrategias que proponen están basadas en la versión de SPARQL 1.1 mediante descomposición de consultas federadas. En este artículo se describe la sintaxis de SPARQL mediante teoría de conjuntos, así como la evaluación de las estrategias propuestas probando los teoremas propuestos. Por último, los autores muestran la mejora de resultados habiendo implementado sus estrategias [19].

***Linking UK Gorvernment Data***

En este trabajo se establecen los casos de uso para la adopción de los principios de *Linked* Data para la publicación de datos públicos del gobierno de Reino Unido. Además, los autores plantean los beneficios de usar *Linked Data*. En sí, el trabajo pretende convencer a empresas, centros de estudio y a desarrolladores a empezar a usar *Linked Data*. En el documento se abordan los temas de datos públicos del gobierno y la responsabilidad que deben existir para su publicación, patrones de diseño, tópicos imprescindibles para abordar la publicación de datos relacionados con información estadística y geoespacial. Por último, presentan las tecnologías disponibles en el contexto *Linked Data* para que desarrolladores de software puedan crear nuevas herramientas [20].

***A parallel approach for improving Geo-SPARQL query performance***

Esta investigación expone el problema actual que existe en las consultas geoespaciales que involucran complejas relaciones topológicas y que en conjunto a las bases de conocimiento las cuales no están indexadas, generan ineficiencia en consultas de sus datos. Entonces, en el documento a parte de exponer el problema, proponen una estrategia para disminuirlo, que consiste en el uso de cómputo en paralelo y la creación virtual de índices de la base de conocimiento. En el desarrollo del documento, se muestran el antes y el después de los resultados obtenidos al usar su estrategia sobre consultas en la ciudad de Connecticut, EUA [21].

***DBpedia SPARQL Benchmark – Performance Assessment with Real Queries on Real Data***

El documento propone una nueva forma de hacer un *benchmarking* con el propósito de demostrar que los *triples stores* existentes no son tan homogéneos como los otros *benchmarking* lo muestran, es decir, el rendimiento depende de los parámetros usados por lo que una consulta no puede mostrar la misma eficiencia respecto a otra consulta.

Se ponen a prueba 4 *triple store* populares Virtuoso, Sesame, Jena-TDB y BigOWLIM. A lo largo del documento muestran el desarrollo del *benchmarking*, desde la generación del conjunto de datos de DBpedia, hasta los resultados obtenidos. En este trabajos los autores miden la cantidad de consultas por segundo (QpS, por sus siglas en inglés) en 4 casos: 10%, 50%, 100% y 200% del conjunto de datos generados previamente.

A parte del *benchmarking*, se presenta una discusión donde se habla sobre las fortalezas y debilidades de cada *triple store* al hacer múltiples y simples consultas con ellos. También se presentan trabajos relacionados y los retos por vencer en el campo de los *benchmarking* en *triple stores* [22].

**Explotación de información en el dominio geo-hídrico ecuatoriano utilizando tecnología semántica**

El documento presenta una propuesta de cómo aprovechar los datos geográficos en el dominio geo hídrico. Se apoyan en la única herramienta *open source,* Parliament, para llevar a cabo las consultas geográficas de forma federada.

Presentan el escenario de el dominio geo hídrico ecuatoriano y la manera de cómo explotan los datos mediante 3 repositorios de *triple store*.

Cuando terminan de obtener los resultados de las consultas, estos son visualizados con el apoyo de una herramienta llamada MAP4RDF

Cabe decir que en este trabajo no se hace ninguna implementación o desarrollo sobre un *triple store*, solo utilizan ejemplos de consultas y obtienen datos [6].

***Sextant: Browsing and Mapping the Ocean of Linked Geospatial Data***

Es este documento presentan una herramienta Web llamada *Sextant* que permite la exploración de datos enlazados geoespaciales para crear, compartir y edición colaborativa mediante la combinación de datos geoespaciales enlazados y otro tipo de información disponible en archivos de formato OGC [23].

**Nacional**

***Linked Open Data* en la Biblioteca Digital Semántica Académica**

En este trabajo se describe y analiza cómo es que el *Linked Open Data* es aplicado en las bibliotecas digitales semánticas. En el documento se presenta un marco teórico explicando las tecnologías que son necesarias para el trabajo tales como XML, RDF, SPARQL, *triple store.* Además, se describen las herramientas que cuentan en la UNAM para el desarrollo de la biblioteca digital. También presentan iniciativas de *Linked Open Data* en Alemania, Reino Unido y en México. Para concluir su documento, dan una justificación del porque su proyecto es innovador al usar nuevas alternativas de publicación, búsqueda y recuperación de información [24].

**Enfoque semántico para el descubrimiento de recursos sensible al contexto sobre contenidos académicos estructurados con OAI-PMH**

En este trabajo describen un enfoque que considera los recursos de información estructurados con el Protocolo para Cosecha de Metadatos de la Iniciativa de Archivos Abiertos, OAI-PMH por sus siglas en inglés, representación ontológica y su uso en aplicaciones de recuperación de información. Los autores presentan los conceptos a tomar en cuenta como son OAI-PMH, Dublin-Core, sensibilidad al contexto, ontologías. En el trabajo se usó Apache Jena el cual es un *triple store* donde llevaron a cabo sus pruebas. Se presenta una metodología para consultar y obtener información. Después muestran sus resultados mostrando un grafo que describe la relación entre las instancias de ejemplo que propusieron. Para finalizar, abordan trabajos relacionados, conclusiones y el trabajo a futuro para mejorar su propuesta [25].

**Facilitador de contenido móvil para el viajero basado en servicios de localización y Web Semántica**

Esta tesis desarrollada en el CIC del IPN donde proponen el diseño e implementación de un facilitador de contenido móvil usando técnicas de la Web Semántica y datos geográficos con GeoSPARQL para recomendaciones de alimentación, hospedaje y sitios de interés a turistas. El proyecto se basó en lenguajes de la Web Semántica como RDF, OWL y SPARQL. Fue desarrollado sobre un *servlet* de Java en conjunto a servicios basado en geolocalización. Se basó en una metodología dividida en 5 etapas: conceptualización, recuperación de términos, recuperación de información turística y su presentación. Al ser un proyecto que pretende estar disponibles en cualquier dispositivo móvil, se hicieron pruebas experimentales de su servicio web, de la aplicación móvil, de la recuperación de información turística y la forma de presentar los datos. Por último, en la conclusión hace una comparación de su proyecto respecto a una popular aplicación llamada *TripAdvisor* con la principal característica de que en el proyecto desarrollado muestra resultados precisos, limitados, pero correspondían a la zona donde el usuario se encuentra, mientras que la aplicación comercial en diversos casos devolvía resultados que no estaban relacionados a la consulta hecha [26].

**UPIITA**

**Recuperación de información geográfica utilizando similitud semántica**

Trabajo hecho por 2 estudiantes donde proponen el uso de un sistema de información geográfica (GIS) y un sistema de geoposicionamiento global (GPS) para identificar sitios de interés alrededor de la zona de donde se encuentre el usuario usando técnicas de similitud semántica, en específico la teoría de confusión. El proyecto fue probado en los alrededores de la IPN Zacatenco para encontrar banco, hospitales y lugares para comer y entretenerse [27].

**Software similar**

A continuación, se muestran 5 *triple store* que existen en la actualidad. Cabe decir que ninguno de ellos tiene la capacidad que se pretende implementar en Apache Marmotta.

* *Oracle Spatial* [28]
* *ClioPatria* [29]
* *Mulgara* [30]
* *Marklogic* [31]
* *Blazegraph* [32]

# Marco teórico

Los conceptos descritos en esta sección son fundamentales en el presente proyecto terminal para mostrar qué compone este proyecto. RDF, URI, *triple store* y SPARQL *endpoint* son conceptos que se deben manejar sí o sí para el desarrollo del módulo ya que estos son elementos indispensables para su desarrollo; tales conceptos explican los objetos con los que se trabajaran a lo largo de este proyecto terminal.

*Linked Data* y Web Semántica son conceptos por tener presentes ya que no se debe de perder de vista sobre qué contexto se está desarrollando el módulo, porque bien se podría hablar de datos cualquiera pero no es así, el proyecto terminal estará construido bajo estos conceptos.

SPARQL y GeoSPARL son los estándares por seguir y cumplir. Si no fuera así, simplemente no se podrá construir nada que funcione con el *triple store* Apache Marmotta o cualquier otra plataforma que se basen en cualquiera de los 2 estándares.

***Linked Data***

El *Linked Data* es un conjunto de buenas prácticas para publicar y conectar datos estructurados en la web. Los principios que tiene asociado son URI para identificar entidades en el mundo, HTTP que es el mecanismo por donde se recuperan recursos o descripciones de recursos y RDF para estructurar y enlazar datos que describen cosas en el mundo [33].

Ya que el *Linked* Data está basado en web, la diferencia entre sitios de Internet comunes y los basados en *Linked* Data es que mientras que los HTML simples en la web son conectados mediante hipervínculos comunes, Linked Data se basa en documentos que albergan datos en formato RDF.

**Web Semántica**

Es una web de datos de cualquier índole. En la figura 9 se muestra la pila que Tim Berners-Lee propone para describir la Web Semántica en cuanto a elementos que lo conforman.



Fig. 9 Diagrama propuesto por Tim Berners-Lee mostrando las tecnologías que conforma la Web Semántica

Describiendo de abajo hacia arriba, en el primer nivel se encuentran las tecnologías que permiten la identificación de cada entidad (Unicode y URI). En el segundo nivel se encuentran las tecnologías que le dan estructura, pero no significado, al documento publicado en la web (XML, NS, XML *schema*). A partir del tercer nivel a la sexto, la pila muestra una firma digital (color blanco) para garantizar la autenticidad de cada entidad. En el tercer nivel están las tecnologías que describen el contenido del documento tal que permiten a la mayoría de las computadoras entender el significado de la entidad en cuestión (RDF y RDF schema). En el cuarto nivel, si bien no es una tecnología, se encuentra un documento o archivo que define las relaciones entre entidades mediante ontologías y reglas de inferencia (comúnmente se implementa el lenguaje OWL para lograrlo). En el quinto nivel se encuentra la lógica la cual reúne las diversas ontologías usadas al igual que las reglas de los lenguajes que se usaron para describir y estructurar la entidad; en este nivel se llevan a cabo las inferencias y se les da un significado a los datos. En el sexto nivel, se realiza la prueba, es decir, el cómo se hicieron las inferencias y el origen de los datos. Por último, se encuentra la confianza, la confianza de que el sistema es capaz de funcionar correctamente, de que el sistema pueda explicar que hace, del origen de las fuentes de datos y servicios, así como la tecnología e interfaz de usuario.

La Web Semántica se puede ver como un todo y el *Linked Data* como los elementos que la conforman puesto que proporciona sustento y las bases para que la Web Semántica sea construida correctamente [34].

**RDF (*Resource Description Framework*)**

El *framework* de descripción de recurso, RDF por sus siglas en inglés, es un modelo estándar para el intercambio de datos en la web.

RDF extiende las estructuras de enlaces de la web al usar URI tanto para nombrar relaciones entre cosas como para los puntos finales de las relaciones, a veces referido como “triple”. El modelo RDF permite representar los datos como sujeto, predicado y objecto. El sujeto y el predicado de un *triple* son URI que identifican a cada uno. El predicado especifica como el sujeto y el objeto están relacionados, y también es representado por un URI. Esta característica provee un modelo de datos basado en grafos [35].

**URI (*Uniform Resource Identifier*)**

Identificador de recursos uniforme, URI por sus siglas en inglés, es una cadena ASCII que identifica recursos de información en la Web Semántica.

Tal y como se observa en la figura 10, una URI puede estar compuesto de un localizador de recursos uniforme (URL, por sus siglas en inglés), de un nombre de recursos uniforme (URN, por sus siglas en inglés) o de ambos [36].



Fig. 10 Diagrama de cómo está compuesto un URI

**RDF *Triple Store***

Es un tipo de base de datos basada en grafos de tripletas RDF.

Al ser una base de datos basada en grafos, el *triple store* puede ser vista como una red de objetos enlazados. El *triple store* al ser una herramienta de la Web Semántica, las entidades que conforman a la base de datos, tripletas RDF, son representadas como sujeto, predicado y objeto o también puede ser considerada como sujeto, predicado y etiqueta [37].

**SPARQL**

SPARQL es un acrónimo para el Protocolo SPARQL y Lenguaje de Consultas RDF, por sus siglas en inglés, y es un protocolo y lenguaje de consultas para *Linked Data* en la web o bases de datos semánticas basadas en grafos (*RDF triple stores*). SPARQL está diseñado y respaldado por el consorcio de la web (W3C).

El siguiente código muestra cómo hacer una consulta SPARQL de los músicos mexicanos famosos que ya han muerto y que están sobre el *triple store* de DBpedia.

PREFIX dcterms: <http://purl.org/dc/terms/>

PREFIX rdfs: <http://www.w3.org/2000/01/rdf-schema#>

PREFIX dbp: <http://dbpedia.org/ontology/>

SELECT ?musico ?nombreMusico ?fechaFallecimiento

WHERE {

?musico dcterms:subject

<http://dbpedia.org/resource/Category:Mexican\_musicians>;

rdfs:label ?nombreMusico ;

dbp:birthDate ?fechaNacimiento ;

dbp:deathDate ?fechaFallecimiento .

FILTER (LANG(?nombreMusico) = "es")

}

La figura 11 muestra el resultado de la consulta



Fig. 11 Resultado de la consulta en SPARQL

Actualmente se encuentra en su segunda versión, SPARQL 1.1, y ya describe una extensión para explícitamente delegar subconsultas a diferentes SPARQL *endpoint*. Esta característica es conocida como consulta federada.

**SPARQL *endpoint***

Un SPARQL *endpoint* es un servicio web que acepta consultas SPARQL [38].

DBpedia es quizá el SPARQL *endpoint* más famoso, el cual aloja la información que se ve, a veces, en la parte derecha de las páginas de Wikipedia, estas son conocidas como *infoboxes*. En la figura 12 se ven datos extraídos desde DBpedia (cuadro rojo) y usados por Wikipedia.



Fig. 12 Datos de DBpedia usados por Wikipedia

**GeoSPARQL**

Es un lenguaje de consultas geográficas para datos RDF en la Web Semántica estandarizado por el *Open Geospatial Consortium* (OGC)*.* GeoSPARQL define un vocabulario para la representación de datos en RDF y también define una extensión para el lenguaje de consultas geoespaciales SPARQL.

Un ejemplo de consulta es el siguiente: Determinar los objetos que estén contenidos en la figura A de la figura 13



Fig. 13 Representación de datos espaciales

A continuación, la consulta SPARQL asociada para resolver ese interrogante:

PREFIX my: <http://example.org/ApplicationSchema#>

PREFIX geo: <http://www.opengis.net/ont/geosparql#>

PREFIX geof: <http://www.opengis.net/def/function/geosparql/>

SELECT ?f

WHERE { my:A my:hasExactGeometry ?aGeom .

?aGeom geo:asWKT ?aWKT .

?f my:hasExactGeometry ?fGeom .

?fGeom geo:asWKT ?fWKT .

FILTER (geof:sfContains(?aWKT, ?fWKT) &&

!sameTerm(?aGeom, ?fGeom))

}

El resultado se muestra en la figura 14.



Fig. 14 Resultado de la consulta en GeoSPARQL

Como era de esperarse, el resultado dice que ‘A’ contiene tanto a ‘B’ como al punto ‘F’