Resumen

El presente proyecto terminal propone diseñar, implementar y caracterizar un módulo de consultas federadas de datos geoespaciales para un triple store que no tenga implementado dicho módulo, en específico, a la plataforma Apache Marmotta cuya arquitectura y funcionamiento están basados en los estándares SPARQL y GeoSPARQL. Además, con el propósito de que los usuarios finales visualicen e interactúen con los resultados desplegados por el módulo de consultas federadas geoespaciales, se usará una aplicación Web para que los resultados recuperados de la Web de *Linked Data* puedan ser visualizados y explorados.

Palabras clave: Módulo, consultas federadas, triple store, datos geoespaciales, Apache Marmotta, Linked Data, SPARQL, GeoSPARQL, Web Semántica, aplicación web.

Contenido

[Introducción 3](#_Toc18501831)

[Justificación 3](#_Toc18501832)

[Planteamiento del problema 4](#_Toc18501833)

[Propuesta de solución 6](#_Toc18501834)

[Alcances (Resultados esperados) 9](#_Toc18501835)

[Objetivo general 10](#_Toc18501836)

[Objetivos específicos 10](#_Toc18501837)

[Referencias 11](#_Toc18501838)

# Introducción

La creación de la Web, llevada a cabo por Tim Berners Lee, y la popularidad que alcanzó provocó que los usuarios se interesaran en aportar contenido de toda índole en poco tiempo sin prestar atención a desarrollar un conjunto de buenas prácticas la cuales sirvieran como referencia para los usuarios al momento de crear y subir contenido a la Web. Debido a esta omisión, la posibilidad de tener una web inteligente se volvería difícil de lograr, esto a consecuencia de que las computadoras no son capaces de interpretar ni de hacer inferencias en el contenido de la Web [1]. Sin embargo, se propuso una evolución que le permitiría a la Web tener un contexto y significado en el contenido que alberga en ella; es aquí donde surge la Web Semántica. Con esta propuesta se pretende que el contenido en Web pueda ser interpretada por las computadoras a nivel semántico [2]. A partir de este acontecimiento surgen servidores de triple store cuya información que almacenan son tripletas en documentos del tipo Marco de Descripción de Recursos (RDF, por sus siglas en inglés) que describen entidades y relaciones en la Web Semántica a través de grafos y a su vez, surgen las plataformas de *Linked Data* (LDP, por sus siglas en inglés) las cuales son herramientas que son capaces de manipular dichas entidades y las relaciones existentes entre ellas. Las *triple stores* se basan en SPARQL [3] que es el lenguaje estandarizado de consultas para bases de datos de tipo RDF. Este concepto también es utilizado en el dominio de la Web Semántica para datos geoespaciales donde para la realización de consultas se utiliza el estándar GeoSPARQL [4]. Este estándar, en conjunto con tecnologías propias de la Web Semántica, ha sido aplicado en problemas de logística, hidrología, turismo, entre otros [5]. Estos ejemplos, con frecuencia, presentan propuestas donde se realizan consultas únicamente a un triple store y éste se encarga de devolver la información almacenada con características geoespaciales. No obstante, existen algunas propuestas donde se han realizado algunos ejemplos ad hoc de consultas federadas en el ámbito de los datos geoespaciales [6] [7]. Sin embargo, el estado del arte actual presenta una importante limitación, ya que existen *triple stores* que no permiten la realización de consultas federadas a través de múltiples *triple stores* que presenten de información geoespacial (conforme a GeoSPARQL) en el contexto de la nube de *Linked Data*.

# Justificación

Existen diversas empresas y organizaciones que se encargan de desarrollar herramientas para la Web Semántica y *Linked Data* para la manipulación y almacenamiento de datos semánticos. Una de estas organizaciones es la organización sin fines de lucro *Apache Software Foundation* (ASF) la cual ofrece diversas herramientas para diferentes necesidades en cuanto a software se refiere. Ejemplos de estas herramienta son servidores Web, *frameworks*, bases de datos, entre otras. Para el mundo del *Linked Data* y Web Semántica, ASF también tiene su plataforma y se le conoce como Apache Marmotta. La plataforma Apache Marmotta entre las diversas características que este software tiene, las 3 relevantes para este proyecto terminal son: es una LDP, un SPARQL *endpoint* y también es una base de datos para tripletas RDF, triple store. Para contextualizar a Apache Marmotta, una analogía con bases de datos SQL se muestra en la figura 1.



Fig. 1 Analogía entre MySQL y Apache Marmotta

A pesar de que es un software desarrollado por Apache, aún tiene características por incorporar al sistema. Ejemplo de estas características es que al estar basado en el estándar SPARQL 1.1, este documento estipula que hay 11 características por cumplir para que un triple store se considere completo conforme a esta versión del estándar (SPARQL 1.1). Apache Marmotta carece de 2 características, consultas federadas y regímenes de vinculación; las consultas federadas será la característica que será implementada en el presente proyecto terminal [8] con el fin de ofrecer una herramienta *open source* a desarrolladores e investigadores que usen la Web Semántica en sus trabajos e investigaciones.

# Planteamiento del problema

Se ha demostrado que aproximadamente el 80% de los datos tienen relación con una ubicación geográfica [8]. Esta es una de las razones por la que las herramientas que traten con datos geoespaciales estén en constante actualización. El caso de la implementación del *Linked Data*, no es la excepción. Investigadores, empresas y organizaciones gubernamentales usan la nube del *Linked Data* para el estudio y administración de su información [9]. Actualmente existen diversas motores de *triple store* que no son capaces de hacer consultas a diversas fuentes de información geográfica a la vez, es decir, consultas federadas a *triple stores* de datos geoespaciales conforme a los estándares SPARQL 1.1 y GeoSPARQL.

Adicionalmente, otro aspecto a considerar para justificar el desarrollo del módulo propuesto en este proyecto es que las bases de datos no solo son del tipo distribuidas, sino que también existen aquellas que están bajo el esquema de bases de datos federadas. En el escenario de la Web Semántica, a este tipo de bases de datos se les denomina *federated* *triple store*. Con esta característica también se involucra el hecho de que los servidores estén en diferentes localizaciones geográficas, por lo que si no se realizan consultas federada a una *triple store* federada, se estaría obteniendo una porción de información. Por el contrario, si la consulta realizada es federada, se obtendría una respuesta completa que agrupa los resultados de las múltiples *triple store*. Además, considerando la característica de que la información proviene de distintos proveedores de información, el desarrollo de este módulo podría permitir que los expertos en las múltiples disciplinas que tiene asociada la información geográfica [6] puedan abordar sus problemáticas con una nueva herramienta que permita integrar y enriquecer sus análisis y estudio.

Ante este escenario, desarrollar e implementar un módulo que solucione dicha problemática sería clave para que los motores de *triple store* puedan ser usados de una mejor manera y obtener mejores resultados en las búsquedas. En el caso particular de este proyecto, el trabajo se realizará sobre Apache Marmotta. Tal y como se indica en la figura 2, entre los elementos que contiene este *triple store,* no se encuentra un módulo de consultas federadas de datos geoespaciales.



y SPARQL *endpoint.*

Fig. 2 Elementos que contiene Apache Marmotta a considerar en el proyecto

La plataforma funciona con los estándares SPARQL y GeoSPARQL para establecer la manipulación correcta de datos en la nube del *Linked Data* y datos geoespaciales respectivamente. Esto quiere decir que tales características ya existen, pero lo que aún no se ha implementado es la capacidad de hacer consultas federadas. Como prueba, la figura 3 muestra una captura de pantalla del sitio oficial de Apache Marmotta[[1]](#footnote-1). encerrado en un rectángulo rojo, indicando que la plataforma aún no es capaz de hacer consultas federadas.



Fig. 3 Apache Marmotta no soporta consultas federadas.

Sin embargo, hay aspectos a cumplir para que este propósito sea realidad. El primer punto por tratar es que Apache Marmotta [10] está escrito en Java por lo que hay que comprender cómo es que está diseñado y construido. Las librerías, objetos y el paradigma de programación implementados son ejemplos de elementos a estudiar. De igual forma se debe de abordar los protocolos que son indispensables para que Apache Marmotta funcione.

Así mismo, puesto que se escribirá código y algoritmos, dicho código debe de ser eficiente y de fácil mantenimiento. Esto se irá mejorando con las múltiples pruebas descritas en escenario de pruebas.

Otros aspectos por considerar es la funcionalidad del módulo la cual consiste en permitir que las consultas federadas se desplieguen en la nube de *Linked Data* y que el usuario final pueda recuperar información geoespacial de forma distribuida. De esta manera, el módulo deberá ser capaz de combinar las respuestas de las diversas fuentes consultadas sin que existan resultados repetidos u omitidos, y cómo es que se manipulará los datos geográficos para poderlos desplegar en una aplicación web. En este sentido, existen diversas alternativas para lograr esta meta, un ejemplo es la herramienta *Map4RDF* cuya función es visualizar y explorar *datasets* RDF cuya información sea geométrica [11]. Por tanto, cabe aclarar que se reutilizará una herramienta de este estilo solo para visualizar los datos recuperados por el módulo de consultas federadas geoespaciales.

De esta forma, el proyecto a desarrollar pretende contribuir a las herramientas usadas en la Web Semántica. Con la intención de cumplir tal propósito la pregunta que surge es: ¿Cuánta información extra revelará una consulta geoespacial federada frente a una consulta a un único repositorio?

# Propuesta de solución

En la actualidad, existen diversos motores *triple store,* aunque gran parte de ellos [12] solo pueden hacer consultas de forma individual sobre aquellos conjuntos de datos con características geoespaciales que tienen almacenados. Así, la herramienta a desarrollar en el proyecto terminal pretende solucionar este problema brindando la posibilidad de realizar consultas distribuidas a conjuntos de datos geoespaciales presentes en la nube de *Linked Data* para el *triple store* Apache Marmotta. De esta manera, se pretende aprovechar los beneficios que dicha plataforma ofrece para avanzar en el estado del arte y contribuir con el desarrollo y progreso de la Web Semántica geoespacial.

En la figura 4 se muestra en línea negra una consulta simple a un solo *triple store*, mientras que, en la línea roja punteada, funcionalidad que proporcionará este proyecto, representa la capacidad de consultar diversos *triple store* a la vez.



Fig. 4 Diagrama de Apache Marmotta haciendo una consulta individual (línea negra) y consultas federadas (líneas punteadas rojas)

Descomponiendo la figura 4, en la figura 5 se muestra un diagrama a bloques que detalla el proceso de una consulta geoespacial simple, es decir, una consulta a un solo *triple store*. La consulta, en función de los operadores usados, retornará una geometría o un booleano.



Fig. 5 Consulta geoespacial simple

La intención del módulo a desarrollar es que la consulta en Apache Marmotta no se centre en un solo *triple store,* sino que consulte a todas las *triple store* qué estén disponibles y que presenten información con características geoespaciales. Una vez que se consulten a todas las bases de datos geoespaciales, las respuestas tendrán que ser unidas de tal forma que no exista datos repetidos. Este resultado será íntegro, ya que contendrá información no solo de un *triple store,* sino que la consulta será respondida con información de diferentes fuentes. La figura 6 muestra un diagrama a bloques de lo explicado.



Fig. 6 Diagrama a bloques de una consulta geoespacial federada

El desarrollo e implementación del módulo de consultas federadas geoespaciales serán basados en los estándares SPARQL y GeoSPARQL, es decir, su funcionamiento, desarrollo y resultados deberán cumplir con las características estipuladas en los documentos *SPARQL 1.1 Federated Query* [3] y *A Geographic Query Language for RDF Data* [13].

Apache Marmotta está basado en el lenguaje de consultas de la Web Semántica SPARQL y es capaz de hacer consultas geográficas SPARQL cuya tecnología es mejor conocida como GeoSPARQL. El módulo que será desarrollado se integraría al código fuente del motor para lograr que éste sea capaz de hacer múltiples consultas en vez de solo una. En el proyecto se plantea desarrollar el módulo de consultas federadas para que funcione con GeoSPARQL. La figura 7 muestra una representación de los módulos de Apache Marmotta. En verde se encuentra el lenguaje y protocolo SPARQL mientras que en amarillo la extensión GeoSPARQL; en rojo se muestra el módulo de consultas federadas a implementar.



Fig. 7 Diagrama a bloques de la implementación del módulo de consultas federadas

Para caracterizar el módulo desarrollado, se propone llevar a cabo un conjunto de pruebas descritas en el escenario de pruebas. Esto a su vez permitirá hacer un *benchmarking* contra otras plataformas similares para determinar sus fortalezas y debilidades respecto dichas plataformas.

Los parámetros para tomar en cuenta en la caracterización estarán completamente determinados después de haber concluido el *benchmarking,* pero entre estos podrían estar los siguientes:

* Tiempo de carga.
* Tiempo de consulta.
* Consultas por segundo.
* Consultas mezcladas por segundo.
* Uso de operadores, modificadores y acceso a datos

Estos parámetros están detallados en los siguientes 3 documentos:

* *An Evaluation of Approaches to Federated Query* *Processing over Linked Data* [14]
* *SP2Bench: A SPARQL Performance Benchmark* [15]
* *The Berlin SPARQL Benchmark* [16]

Con los datos recabados y una vez desarrollado el módulo, se podrá redactar la documentación de dicha herramienta que permita su correcto uso, así como su compartición con la comunidad relacionada con esos temas.

Además, el módulo se complementará con una herramienta para que los usuarios finales puedan visualizar los resultados de las consultas geoespaciales federadas. En la actualidad ya existen herramientas que permiten visualizar datos geoespaciales, entonces con el único fin de demostrar que el módulo funciona correctamente, por ende, otra alternativa de comprobación de funcionamiento, este trabajo presentará los resultados obtenidos usando uno de esos recursos que permitan desplegar dicha información y analizarla de forma amigable como puede ser una aplicación web. Un resultado, similar al esperado, se muestra en la figura 8 donde se aprecian puntos en color azul que representan la información geoespacial recuperada de múltiples *triple store*.

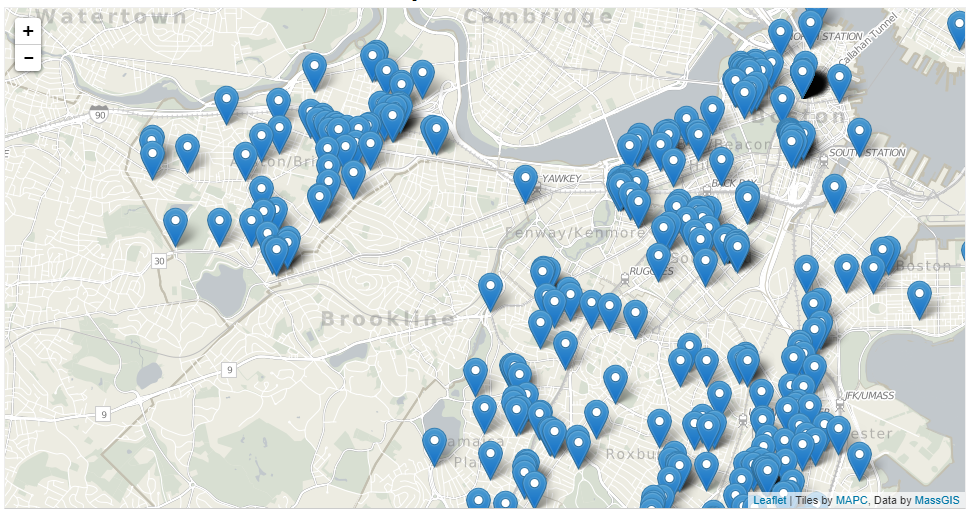


Fig. 8 Propuesta de visualización

## Alcances (Resultados esperados)

Al no existir un módulo de consultas federadas geoespaciales para Apache Marmotta, ningunos de los alcances propuestos han sido elaborados hasta el momento, Así, los alcances fijados para este trabajo son:

* El *triple store* Apache Marmotta en conjunto con el módulo será capaz de hacer consultas federadas geoespaciales con base a los estándares SPARQL y GeoSPARQL.
* El *benchmarking* estará inspirado en trabajos presentes en el estado del arte, tales como [14]*,* [15] y[16].
* La caracterización será delimitada con los resultados arrojados por el *benchmarking*.
* La comprobación y explotación de datos geoespaciales devueltos por el módulo de consultas federadas en Apache Marmotta serán mediante una aplicación Web.
* Una vez que se tenga el módulo desarrollado, implementado y después de haberlo puesto a prueba con las pruebas descritas en el escenario de pruebas, se propondrá a la organización *Apache Software Foundation* que el módulo sea incorporado en el sistema Apache Marmotta de forma oficial.

**Limitaciones**

* Se necesita una computadora con conexión a Internet, sistema operativo Unix-Linux Ubuntu para instalar Apache Marmotta y el módulo y un navegador web para visualizar la aplicación web.
* El módulo estará integrado al código fuente del motor Marmotta por lo que no será un *framework* ni una API. Su uso será completamente dentro del entorno Apache Marmotta.
* La herramienta por usar para la exploración de datos geoespaciales desplegados por el módulo de consultas federadas no será desarrollada, será seleccionada de las opciones actuales y usada.
* Se busca incorporar la funcionalidad a Apache Marmotta mas no competir contra otros *triple stores*.
* El proyecto no trata sobre una búsqueda semántica interespacial.

# Objetivo general

Desarrollar un módulo de consultas geoespaciales federadas para el *triple store* Apache Marmotta, con el propósito de contribuir al avance de las tecnologías usadas en la Web Semántica y proveer una alternativa *open source* diferente a los *triple store* existentes.

### Objetivos específicos

* Implementar, con base en los estándares SPARQL, GeoSPARQL, así como auxiliándose de otras tecnologías involucradas en la Web Semántica y *Linked Data,* un módulo de consultas federadas para el *triple store* Apache Marmotta*.*
* Evaluar el rendimiento de las consultas hechas por el módulo desarrollado (*benchmarking*).
* Comparar el *triple store* Apache Marmotta con otros *triple store* auxiliándose de la caracterización y *benchmarking* del módulo.
* Caracterizar el módulo de consultas federadas.
* Explotar lar características geoespaciales de los datos obtenidos del despliegue de consultas federadas mediante una aplicación Web para poder visualizarlos y explorarlos.

# Referencias

|  |  |
| --- | --- |
| [1] | T. Berners Lee, J. Hendler y O. Lassila, «The Semantic Web,» *Scientific American,* vol. 284, nº 5, pp. 34-43, 2001. |
| [2] | C. Bizer, T. Heath y T. Berners-Lee, «Linked Data - The Story So Far,» de *Semantic services, interoperability and web applications: emerging concepts*, USA, Information Science Reference, 2009, pp. 205-227. |
| [3] | W3C, «SPARQL 1.1 Overview,» 21 Marzo 2013. [En línea]. Available: https://www.w3.org/TR/sparql11-overview/. [Último acceso: 18 Marzo 2019]. |
| [4] | O. G. Consortium, «GeoSPARQL - A Geographic Query Language for RDF Data,» 7 Julio 2001. [En línea]. Available: https://www.opengeospatial.org/standards/geosparql. [Último acceso: 18 Marzo 2019]. |
| [5] | R. Battle y D. Kolas, «Linking Geospatial Data With GeoSPARQL,» de *Semant Web J Interoperability, Usability, Appl. Accessed*, vol. 24, Arlington, 2011. |
| [6] | L. Lupercio, F. Baculima, M. Espinoza y V. Saquicela, «Explotación de información en el dominio geo-hídrico ecuatoriano utilizando tecnología semántica,» *Maskana,* vol. 6, pp. 69-77, 2015. |
| [7] | N. Charlampos, D. Kallirroi, K. Kostis y K. Manolis, «Sextant: Browsing and Mapping the Ocean of Linked Geospatial Data,» de Extended Semantic Web Conference, Grecia, 2013. |
| [8] | A. Marmotta, «Apache Marmotta Platform: SPARQL,» 30 Abril 2014. [En línea]. Available: http://marmotta.apache.org/platform/sparql-module.html. [Último acceso: 24 Abril 2019]. |

1. <http://marmotta.apache.org/platform/sparql-module.html> [↑](#footnote-ref-1)