UNIVERSIDAD TÉCNICA PARTICULAR DE LOJA

La Universidad Católica de Loja

ÁREA TECNICA

TÍTULO DE INGENIERO EN SISTEMAS INFORMATICOS Y COMPUTACION

Publicación de Datos Estadísticos en Linked Data

TRABAJO DE TITULACIÓN.

AUTOR: Jaramillo Espinoza, Víctor Manuel

DIRECTOR: Morocho Yunga, Juan Carlos, Ing

LOJA - ECUADOR

2016

2016

APROBACIÓN DEL DIRECTOR DEL TRABAJO DE TITULACIÓN

Ingeniero.
Juan Carlos Morocho Yunga
DOCENTE DE LA TITULACIÓN

De mi consideración:

El presente trabajo de titulación: Publicación de Datos Estadísticos en Linked Data realizado por Víctor Manuel Jaramillo Espinoza, ha sido orientado y revisado durante su ejecución, por cuanto se prueba la presentación del mismo.

Loja, Octubre de 2016

f) . . . . . .

DECLARACIÓN DE AUTORÍA Y CESIÓN DE DERECHOS

“Yo Jaramillo Espinoza Víctor Manuel declaro ser autor del presente trabajo de titulación: Publicación de Datos Estadísticos en Linked Data, de la Titulación de Sistemas Informáticos y Computación, siendo el Ing. Juan Carlos Morocho Yunga director del presente trabajo; y eximo expresamente a la Universidad Técnica Particular de Loja y a sus representantes de posibles reclamos o acciones legales. Además certifico que las ideas, conceptos, procedimientos y resultados vertidos en el presente trabajo investigativo, son de mi exclusiva responsabilidad.

Adicionalmente declaro conocer y aceptar la disposición del Art. 88 del Estatuto Orgánico de la Universidad Técnica Particular de Loja que en su parte pertinente textualmente dice: “Forman parte del patrimonio de la Universidad la propiedad intelectual de investigaciones, trabajos científicos o técnicos y tesis de grado o trabajos de titulación que se realicen con el apoyo financiero, académico o institucional (operativo) de la Universidad”

f…….

Autor: Jaramillo Espinoza Victor Manuel

Cedula: 1105232688

DEDICATORIA

Ha sido un camino largo que he recorrido y aun cuando este lejos de terminarlo nada de lo que hoy he logrado lo habría conseguido solo.

Dedico este trabajo a todos quienes en estos años han confiado en mí y han sido mi apoyo para permitirme llegar a donde estoy ahora, pero sobre todo dedico este trabajo a mis padres Victor Aurelio y Jesusita Espinoza quienes con su cariño me enseñaron el valor de trabajo y a nunca rendirme sin importar el reto que tuviese delante.

Se lo dedico a mis queridos hermanos Anabel Estefanía, Jehny María y Leonardo Andrés quienes fueron mi inspiración y en más de una ocasión la razón de mi esfuerzo y superación.

Pero sobre todo y con el más profundo amor dedico este trabajo a mi amado abuelo el sr Victor Eduardo Jaramillo quien me enseño, no solo lo importante de saber aprovechar las oportunidades sino que también me brindo un sinfín de ellas y me mostro que para alcanzar un futuro brillante lo único que se necesita es esfuerzo y dedicación.

AGRADECIMIENTO

Agradezco a la Universidad Técnica Particular de Loja por abrirme sus puertas y acogerme en sus aulas, de las que me llevo grandes enseñanzas y recuerdos.

Agradezco a todos mis profesores que con infinita paciencia supieron brindarme sus conocimientos y guiarme con sus valores, pero sobre todo agradezco al Ing. Juan Carlos Morocho, mi director de tesis, quien me brindó la oportunidad de realizar este trabajo y con sus enseñanzas, empeño y motivación me impulso a conseguir este logro. Gracias por su confianza y amistad brindadas.

Agradezco a mi familia por el apoyo recibido, a mis compañeros y amigos pero en especial, a aquellos que supieron brindarme su ayuda a lo largo de todo este trayecto ofreciéndome no solo sus conocimientos sino también sus diversos puntos de vista y pensamientos.

ÍNDICE DE CONTENIDOS

ÍNDICE DE FIGURAS

INDICE DE TALBAS

RESUMEN

La información es uno de los recursos más importantes de nuestra generación, y uno de sus grandes representantes es la Web de Datos.

No obstante la creciente cantidad de información ha ocasionado problemas de interoperabilidad obligando a idear nuevas formas de emplear los datos; siendo así como nace este proyecto, que se encuentra orientado a trabajar con los datos de índole estadístico, por ser de vital importancia tanto para personas particulares e instituciones.

En este proyecto se tratan temas referentes al ciclo de publicación de "linked data" incluyendo el refinamiento, depuración, modelamiento, transformación y explotación de la información. Además se realiza una evaluación a los vocabularios que permiten modelar datos estadísticos considerando sus principales características.

Para llegar al objetivo propuesto se realiza una publicación de datos estadísticos referentes a revistas científicas a nivel mundial, basándose principalmente en información brindada por Scopus, la cual fue conciliada y normalizada partiendo de archivos en diferentes formatos (Excel y CSV) para finalmente ser explotada mediante un sitio web que permitiera comprobar la eficiencia y valides del modelo realizado utilizando RDF Data Cube.

PALABRAS CLAVES: Web Semántica, Estadística, Data Cube, Vocabulario, Revistas científicas.

ABSTRACT

The information is one of the most important resources of our generation, and one of its great representatives is the Web of Data.

Despite the increasing amount of information has led to interoperability problems forcing devise new ways to use the data; being as well as this project, which is designed to work with the data of statistical nature, being of vital importance to both individuals and institutions born people.

In this project cycle issues relating to the publication of "linked data" including refinement, debugging, modeling, processing and exploitation of information they are discussed. In addition an assessment is made vocabularies that allow statistical data modeling considering its main features.

To reach the proposed objective publication of statistical data is made to scientific journals worldwide, based primarily on information provided by Scopus, which was reconciled and standardized starting from files in different formats (Excel and CSV) to finally be exploited by a website that would prove the efficiency and validate the model made using RDF Data Cube.

KEYWORDS: Semantic Web, Statistics, Data Cube, Vocabulary, scientific journals

INTRODUCCIÓN

En la actualidad la web se ha convertido en una importante herramienta para el desarrollo de nuestras actividades y tareas diarias, siendo la principal plataforma para compartir y publicar información.

La cantidad de datos que se generan en la web crece a pasos agigantados a tal punto que esto ha provocado el desarrollo de nuevos conceptos e ideas, enfocados en el empleo y utilización de dicha información.

No obstante, gracias a la llegada de la Web Semántica y al fuerte trabajo realizado por parte de organizaciones como la W3C , es posible dar significado a los datos mediante el uso de conceptos tales como: metadata, “linked data”, vocabularios, entre otros.

En el tema referente al uso de datos estadísticos, si realizamos una búsqueda en la web notaremos que los trabajos que explican la forma de modelar estos datos son muy puntuales, siempre realizando la implementación de un determinado vocabulario sin profundizar en las opciones como son el uso de RDF Data Cube, SKOVO o RDF-SDMX, como ejemplo de esto tenemos la iniciativa “Integrating Serbian Public Data into the LOD cloud” y la investigación “Representing verifiable statistical index computations as linked data” de las cuales se habla en el CAPÍTULO II en la sección “2.5 Trabajos Relacionados”, no obstante en ninguna de ellas, ni en otros trabajos se realiza una evaluación profunda de los vocabularios, explicando que opciones existen, las ventajas o desventajas que poseen y como se comparan con los demás.

En este proyecto se realiza dicha tarea, evaluándose de manera objetiva los principales vocabularios disponibles que permiten el modelamiento de la información estadística.

Nos centraremos en esta información debido a que es uno de los principales problemas que se afronta a nivel de todo el mundo, y que en muchos de los casos se traduce en mejoras de procesos, incremento de la eficiencia y predicciones más acertadas tanto climatológicas, económicas, ambientales, científicas y de otras áreas.

Para esta tarea se pretende cumplir con los siguientes objetivos:

- Realizar un análisis detallado de los vocabularios existentes que permitan publicar datos estadísticos en la web, en formato estándar, que resulte entendible por humanos y máquinas, con la finalidad de mejorar su procesamiento.

- Desarrollar una guía procedimental y aplicada, que permita la publicación de datos estadísticos, empleando uno de los vocabularios consensuados previamente analizados.
- Aplicar los procedimientos anteriormente desarrollados, para publicar información estadística referente a las revistas científicas y sus diversos indicadores, como una forma de validar el trabajo desarrollado.

Con el fin de lograr estos objetivos el presente documento se encuentra conformado por cinco capítulos principales siendo su contenido el siguiente:

Capitulo l.- En esta primera sección del documento se explican algunos conceptos introductorios para entender la web semántica, su funcionamiento y como esta evolucionó a partir de las primeras versiones. También se explica el concepto fundamental de “Linked Data”, conjuntamente con las principales tecnologías utilizadas para almacenar y extraer la información. Además se realiza un análisis de los vocabularios que permiten la representación de datos estadísticos, puntualizándose sus orígenes, evolución y características

Capitulo ll.- Se analizan las principales motivaciones para realizar este proyecto, encontrándose apartados tales como: planteamiento del problema, justificación y la solución propuesta. También se analizan algunos trabajos relaciones y se muestra ejemplo de implementaciones exitosas de vocabularios que permiten representar datos estadísticos.

Capitulo lll.- Se efectúan las primeras actividades concernientes a los datos preparándolos para la publicación, por cuanto se realizan actividades referentes a la obtención, depuración y preparación de los datos.

Capitulo IV.- Se realizan las tareas referentes al ciclo de vida de Linked Data contemplándose las etapas de especificación donde se trabaja el diseño de los URIs, modelado de los datos donde se consideran aspectos de la reutilización de recursos ontológicos y el diseño de un vocabulario que permita modelar las nuevas entidad a ser utilizadas en el modelo, y finalmente la generación de tripletas donde se explica como transformar la información a formato RDF utilizado el modelo de la etapa anterior.

Capitulo V.- Se explican temas referentes al almacenamiento de la información y la construcción de una aplicación web que permita utilización y explotación de dichos datos. También se explican temas relacionados a las pruebas realizadas para medir el rendimiento de la aplicación y la retroalimentación de los datos que se realizada con información del año 2014.

Las últimas tres secciones de este documento son las Conclusiones, las Recomendaciones y los Trabajos Futuros. En estos apartados se resume de manera objetiva los resultados obtenidos, la experiencia adquirida, problemas afrontados y posibles propuestas para continuar con el presente proyecto.

CAPITULO I
ESTADO DEL ARTE

1.1 Introducción

En este capítulo se hablara acerca de las diferentes versiones que han existido de la web, así como también de las principales tecnologías que hoy en día conforman la web semántica, prestando especial atención en los vocabularios. Todo esto con la finalidad de comprender el contexto actual en el cual se sitúa este proyecto y facilitar la comprensión del lector referente a algunos conceptos y temas que se hablan más a profundidad en los capítulos siguientes.

1.1.1. Orígenes y evolución de la web.

La historia de la web comienza en la década de los 90. En estos años el crecimiento del internet se había producido de manera vertiginosa, incrementándose en gran medida el número de computadores que se conectaban a internet (Vences & Segura, 2011). Sin embargo este incremento también dejo en evidencia la existencia de múltiples problemas al momento de poder acceder a la información, debido a que no existían estándares para acceder a estos datos y mucho menos organizarlos.

Es en este contexto, desde donde Tim Berners-Lee comienza su investigación para desarrollar un sistema de información descentralizado, mediante el uso de hipertexto que una vez implementado permitió conectar varias piezas de información, resolviendo así varios de los problemas de acceso a los datos y también creando lo que hoy en día conocemos como la World Wide Web (para el resto del documento se utilizará el término “web” para identificar a la World Wide Web) en su primera versión (Universidad Pompeu Fabra, n.d.).

Esta primera versión de la web fue llama web 1.0 y a pesar de que su implementación significaba grandes adelantos para el campo de la computación(Lozada, 2014), aún se encontraba lejos de ser perfecta y tenía mucho por mejorar.

La web 1.0 era una versión de la web que se encontraba caracterizada por búsquedas de texto sumamente rápidas, pero cuyos resultados obtenidos eran simples documentos de texto, esto debido a que la web era utilizada solamente para mostrar información, ignorando la forma de presentación y la interacción con el usuario en el sentido de que este pudiese generar o enriquecer la información existente mediante comentarios, citas, etc. Un esquema simple de su funcionamiento se puede observar en la Figura № 1.

Figura 1 Web 1.0
Fuente: Recuperado de: http://goo.gl/RxIL7i
Elaborado por: (Lozada, 2014)

Para comprender el porqué de esta web, debemos recordar que en un comienzo tanto el internet como la web eran tecnologías relativamente nuevas y por tales motivos los únicos que accedían a ellos eran las universidades e instituciones gubernamentales con fines investigativos y que poseían grandes recursos económicos para realizarlo, por tanto en un comienzo el número inicial de usuarios de la web era sumamente reducido y las búsquedas de texto plano era más que suficiente.

Entre las principales tecnologías que implementaba la web 1.0 podemos destacar el sistema de hipertexto llamado Enquire[1], la aparición del lenguaje Hypertext Markup Language[2] (HTML), el perfeccionamiento de múltiples protocolos para el envío de información en la web entre otros(Vences & Segura, 2011).

Con el trascurrir del tiempo poco a poco la interacción con el usuario, que en un principio había sido dejada de lado, comenzó a tomar importancia debido al creciente número de usuarios que ahora tenían acceso a la web.

Este incremento de usuarios se debió a que los equipos computacionales se volvieron más pequeños y potentes, reduciendo considerablemente sus costos de producción y de venta. Resultando que las familias, empresas, escuelas y personas particulares tuvieron acceso a computadoras y posteriormente al internet, la Figura № 2 muestra de manera gráfica este incremento, el mismo que daría comienzo a una nueva etapa para la web que se encontraría caracterizada por el aumento sustancial de usuarios y sus nuevas exigencias.

[1] https://padillayolanda2.wordpress.com/tag/enquire/
[2] http://www.w3.org/MarkUp/

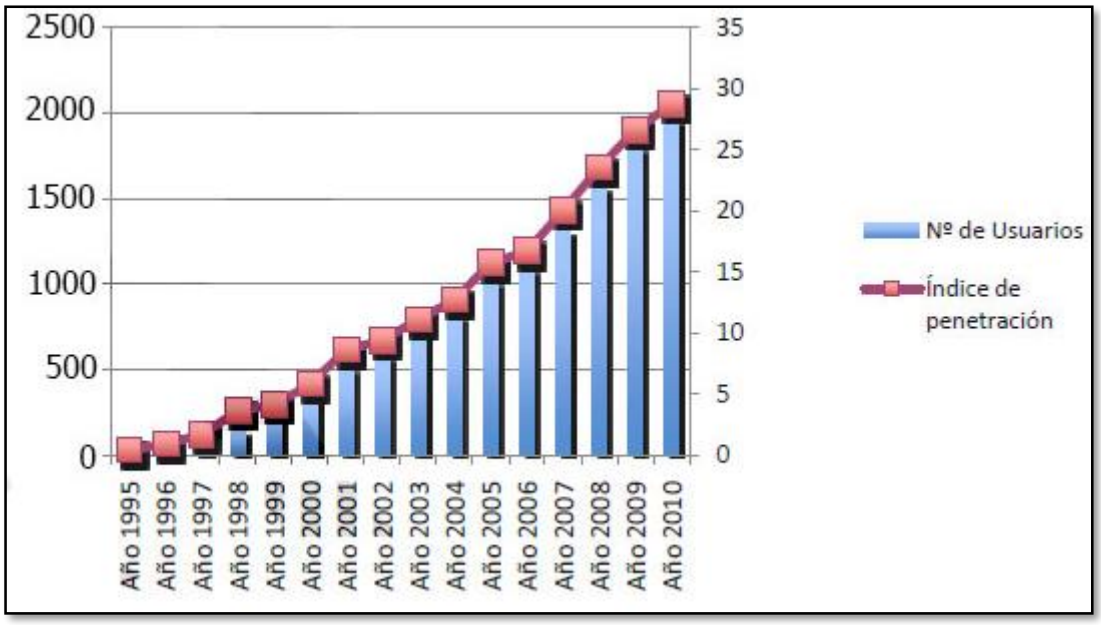


Figura 2 Evolución del número de internautas en el mundo.
Fuente: Recuperado de: http://goo.gl/j5zXqo
Elaborado por: (Vences & Segura, 2011)

Es bajo estas condiciones que se impulsa el surgimiento de la siguiente versión de la web llamada web 2.0. Considerando al usuario en un papel más relevante, siendo este capaz de generar su propio contenido y de esta forma convertir a la web, que en un principio fuese netamente estática, en una web dinámica e interactiva, en la que cualquier persona podía participar y compartir no solamente sus conocimientos sino también sus opiniones (Lozada, 2014).

Gracias a este cambio en el enfoque, los recursos y páginas web también se vieron alterados, comenzando a considerarse en mayor medida el aspecto estético e interactivo de una página web, tratando de incluir mayor contenido multimedia como son imágenes, sonido y video. Al mismo tiempo se comenzaban a implementar aspectos de usabilidad[3] dentro de las páginas web, permitiendo así que las personas sin amplios conocimientos de informática pudiesen gestionar sus propios sitios web, utilizando nuevas tecnologías y lenguajes de programación como:

- ***CMS:*** Significa "Content Management System[4]" y es un sistema que permite la creación y administración de información de manera fácil y sin ser necesarios conocimientos amplios de informática.

[3] http://www.nosolousabilidad.com/manual/1.htm
[4] https://goo.gl/3muXjq

- ***CSS:*** Son las siglas en inglés de "Cascading Style Sheets[5]". Uno de los principales leguajes utilizados para poder modificar la presentación de una página web y mejorar el aspecto de los resultados obtenidos en las búsquedas.
- ***XML:*** Es un conjunto de tecnologías[6] utilizadas para el manejo de datos e información.
- ***AJAX:*** Son las siglas en ingles de "Asynchronous JavaScript And XML[7]". Una técnica de desarrollo web que resultó fundamental para mejorar y añadir funcionalidades a los diferentes sitios en internet.

Estos son solo algunos ejemplos de los cambios en las tecnologías y cómo estos afectaron a la web (Lozada, 2014).

La versión 2.0 de la web se caracterizó también por el surgimiento de nuevos conceptos en los cuales primaba la cooperación entre usuarios y el aprendizaje colaborativo. Además de dar gran importancia a los usuarios finales, considerándolos consumidores y creadores tanto de conocimientos como contenidos, esto resultó fundamental para lograr la expansión y crecimientó de una web social abierta que estuviera en constante actualización, esta filosofía se muestra en la Figura № 3 donde se observa el nuevo flujo de información ocasionado por este cambio. Varios de estos conceptos se plasmaron en lo que hoy conocemos como wikis[8], blogs[9], RSS[10], videocasts[11], potscast[12], redes sociales[13] entre otros.

[5] https://goo.gl/ctnvII
[6] https://goo.gl/e2SGPA
[7] https://goo.gl/L96YW8
[8] http://www.maestrosdelweb.com/queeswiki/
[9] http://es.slideshare.net/killerlusca/definicin-de-blog
[10] http://www.rss.nom.es/
[11] http://es.slideshare.net/mariaparavariar/breve-definicion-de-videocast
[12] http://es.slideshare.net/gustavo0311/podcast-4020062
[13] http://es.scribd.com/doc/24658747/Redes-sociales-definicion#scribd

Figura 3 Web 2.0
Fuente: Recuperado de: http://goo.gl/5eCym8
Elaborado por: (Lozada, 2014)

Con el establecimiento de la web 2.0 comenzó la era de la información, llamada así debido a la facilidad para crear, publicar y compartir información dentro de la web (Belloch, Mide, & Valencia, 2002).

Estas tareas se volvieron mucho más fáciles, a tal punto que hoy en día miles de personas publican constantemente nueva información en internet. Esto podría sonar como algo grandioso desde el punto de vista de la creación de nueva información, pero como suele suceder en el área de la informática, todo adelanto tiene su precio. En este caso tener tan inimaginable cantidad de información acarrea dos nuevos problemas:

> *"¿Cómo realizar consultas precisas de un tema específico entre tal cantidad de datos? y ¿Cómo procesar los diferentes formatos en los cuales se encontraba la información en internet?"*
> (Rodríguez Méndez, n.d.)

Con estas premisas nace no solo la idea, sino también la necesidad de tener un nuevo cambio, si la web 2.0 se había centrado en la relación entre usuarios para la cooperación y creación de información, ahora la web 3.0 o también llamada web semántica (Lozada, 2014) se tendría que centrar en nuevas formas para procesar esa información y permitir que dichos datos resultaran de utilidad.

Uno de los primeros en abordar este nuevo problema fue el propio creador de la web

> *"Tim Berners-Lee presentó en colaboración con Hendler y Lassila en el año 2001 en un artículo de la revista Scientific American una posible solución."*
> (Corchuelo, 2008)

En aquel artículo se propuso una web donde los diferentes programas y software de internet, se encargaran de realizar las búsquedas por los usuarios. Este principio se ve

reflejado en la definición de web semántica proporcionada por la W3C, el cual dice lo siguiente:

> *"La Web Semántica es una Web extendida, dotada de mayor significado en la que cualquier usuario en Internet podrá encontrar respuestas a sus preguntas de forma más rápida y sencilla gracias a una información mejor definida sobre lo que busca…."*
> (W3C, 2010)

Este concepto puede aparentar ser sencillo, pero al momento de su implementación se torna evidente que es una tarea sumamente compleja. Esto en gran medida se debe a que, un cambio como el que se plantea, significa convertir a la web en algo más similar a una gran base de conocimiento con un modelo muy similar al que se muestra en la Figura № 4, diferente a lo que existe actualmente, que es simplemente un conjunto de páginas web publicadas en internet.

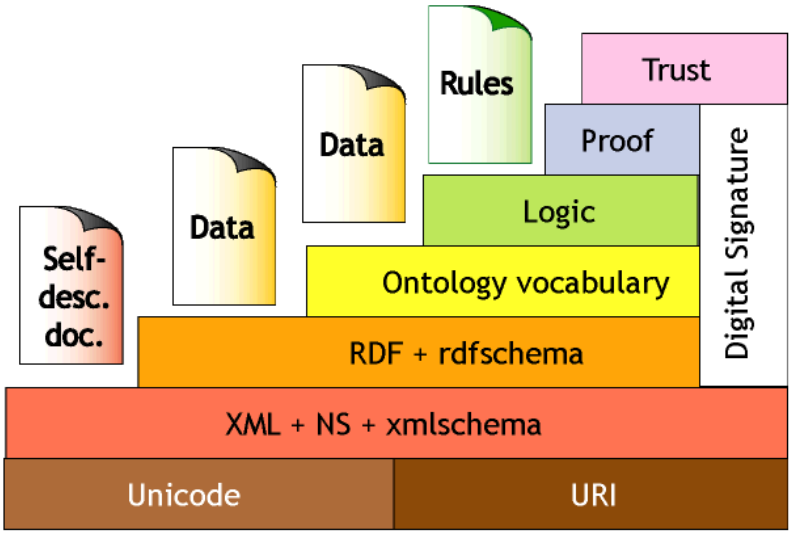


Figura 4 Modelo de capas para la web semántica.
Fuente: Recuperado de: http://goo.gl/NnKIEu
Elaborado por: (Merino, n.d.)

El principal motivo para realizar este cambio resulta ser la mejora en la eficiencia al ejecutar búsquedas. Al momento que se lleva a cabo una consulta en una base de conocimiento, podemos obtener mejores resultados y de manera más rápida que si se realizara una consulta de texto plano, es decir la basada en búsqueda por coincidencia de palabras como se efectúa en la mayoría de los buscadores actuales.

Un ejemplo de esto sería realizar la búsqueda de la oración "naranja grande", donde sí se realizara en la web 2.0 daría como resultado una confusión en los resultados entre el color y la fruta "naranja". En cambio la web semántica debería de poder deducir que

al estar hablando de “naranja grande” se refiere a un fruto de gran tamaño y no de un color.

En la web semántica lo que se buscan son conceptos, por cuanto se podría decir que la misma web entiende lo que se está buscando de forma similar a como lo haría una persona (W3C, 2010). Pero para que esto sea posible significa que las máquinas deben ser capaces de comprender el conocimiento y si bien esto no es posible, por el momento, en un sentido estricto como tal, sí se puede dotar a las máquinas con herramientas para que manejen de mejor forma la información y por ende los datos en general.

Estas herramientas son los metadatos[14], que aplicados de manera correcta y en combinación con nuevas formas de estructurar la información, permiten realizar definiciones de los datos, para luego haciendo uso del enlazado de contenidos, las máquinas puedan realizar combinaciones y deducciones lógicas correctas, todo esto apoyado en las nuevas tecnologías y conceptos semánticos, los mismos que se muestran en la Figura № 4 y se explican a mayor detalle en el apartado “1.2 Tecnologías”

1.2. Tecnologías de la web semántica

Las tecnologías y conceptos que se utilizan en la web semántica son muy variadas pero todas ellas tienen como objetivo ayudar a solucionar los dos problemas que surgieron en la web 2.0 y que fueron mencionados en la sección “1.1.1 Orígenes y Evolución de la web”. En este apartado vamos a hablar un poco acerca de cada una de estas tecnologías y conceptos intentando dar una idea clara de su utilidad y el aporte que estas representan para la web 3.0.

1.2.1 Ontologías.

Para comenzar a hablar de que es una ontología, primero se debe entender que no son un concepto nuevo, pues ya existían desde hace mucho tiempo, sin embargo la novedad esta en aplicarlo al área de la informática como tal, puesto que las ontologías en un principio eran usadas netamente en la filosofía y la metafísica.

Con la aplicación de las ontologías al campo de la informática, surgió un nuevo concepto que fue cambiando con el tiempo. En nuestro caso utilizaremos la definición brindada por Thomas Gruber la cual dice:

> *“Una Ontología es una especificación formal y explícita de una conceptualización compartida”*

[14] http://www.infor.uva.es/~sblanco/Tesis/Metadatos.pdf

(Giraldo, Acevedo, & Moreno, 2011)

Esta definición a primera vista pudiese parecer bastante compleja y difícil de comprender, pero no tiene por qué ser así, como ya se ha mencionado la web semántica es similar a una base de conocimiento, por tanto resulta entendible que para cada área de conocimiento exista un determinado dominio, una ontología es la definición de términos que nos permiten describir ese domino o área de conocimiento específico como puede ser: la robótica, el cáncer, la cocina, las finanzas o cualquier otra área de interés. Para poder representar un dominio, la ontología debe contener no solo las definiciones de los conceptos relevantes, sino también las relaciones existentes entre ellos. Con la finalidad de poder reutilizarlos e incrementar en si el conocimiento sin repetir definiciones ya existentes.

Debemos considerar que los términos de una ontología tienen que ser consensuados. Además podemos mencionar que una ontología posee ciertas partes de las cuales se encuentra compuesta, incluyendo los conceptos, relaciones y otras nociones tales como:

- *"Funciones: relaciones en las cuales el elemento n-ésimo es único para los n-1 anteriores. Por ejemplo una relación serPadreDe se puede modelar como una función ya que el atributo que evalúa es único para cada caso."*
  (Rodríguez Álvarez, 2012)

- *"Axiomas: modelan "verdades" que siempre se cumplen en el modelo.*
  (Rodríguez Álvarez, 2012)

- *Instancias: representan realizaciones específicas del dominio de la ontología."*
  (Rodríguez Álvarez, 2012)

Las ontologías pueden ser representadas gráficamente mediante el uso de grafos que permiten un mayor entendimiento y facilitan la construcción de modelos. Para un caso práctico podemos definir a los grafos como:

> *"Herramientas que nos permiten la representación de problemas o información de manera gráfica mediante el uso de vértices y aristas que los conectan."*
> (Villalobos, 2006)

A continuación en la Figura № 5 se muestra un ejemplo de una ontología donde se representa de manera muy sencilla el dominio referente a al software libre.

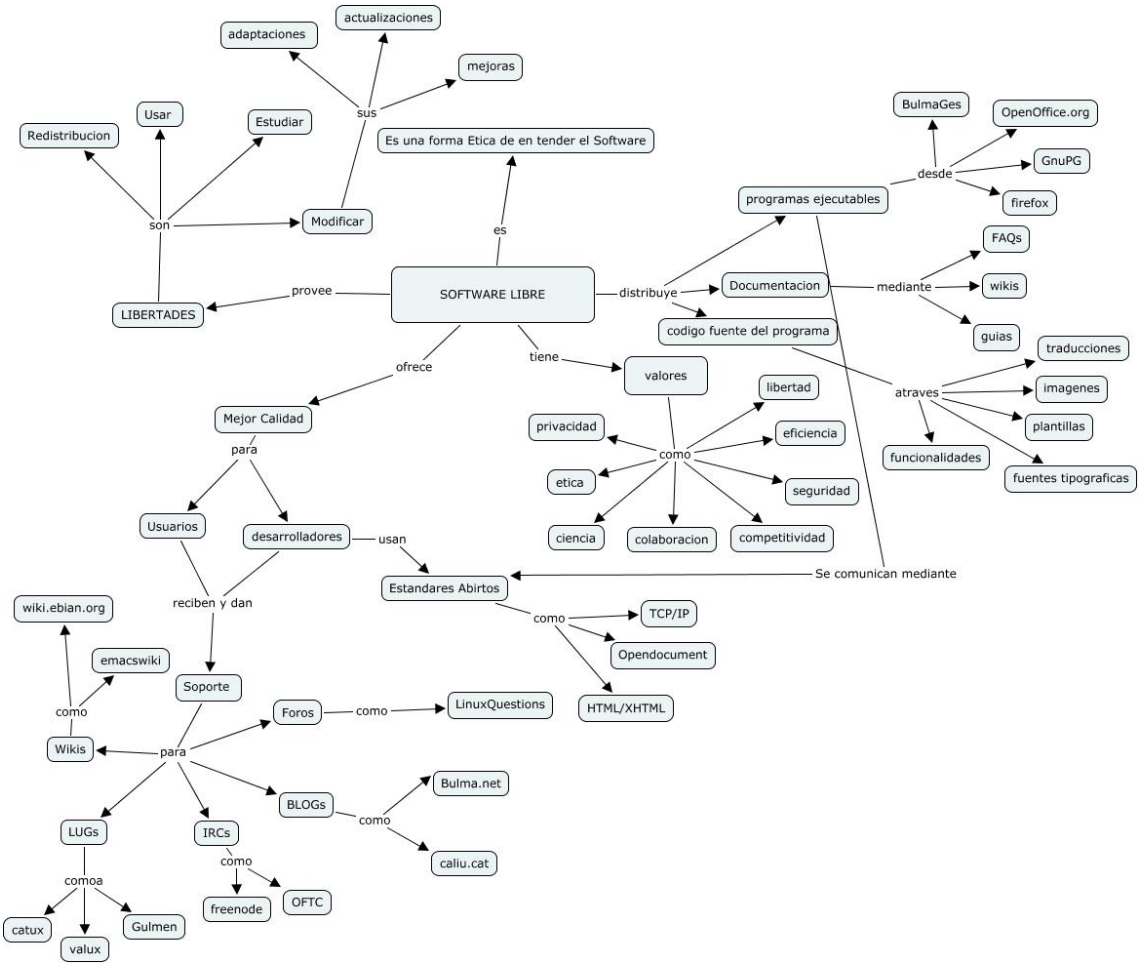


Figura 5 Ontología.
Fuente: Recuperado de: https://goo.gl/YeuSY2
Elaborado por: (Jara, 2012)

1.2.2 URI.

Son las siglas en ingles de “Uniform Resource Identifiers”. El concepto de URI ha estado presente desde 1990 con la iniciativa Mundial de la información, que se dio en ese mismo año y ha sido uno de los principales conceptos sobre los cuales se ha apoyado la web semántica. Podríamos definir el URI como:

> *“Identificadores uniformes de recursos (URI) que proporcionan una forma sencilla y extensible de medios para identificar un recurso.”*
> (Berners-Lee, 1998)

Un recurso puede ser cualquier cosa que posea identidad dentro de la web. Puede ser una imagen, un video, un documento electrónico, un servicio o hasta un concepto como el que sería la descripción de una persona específica o un lugar (Lee, n.d.). Por ejemplo Ecuador, Quito o cualquier lector de este documento podría llegar a ser considerado un

recurso, pues cada uno de ellos puede ser buscado y encontrado en la web, ya sea que se encuentre en una red social o en una página web específica.

Por tanto dentro de la web semántica cuando realizamos una búsqueda lo que obtendremos como resultado será un recurso el cual se encuentra identificado por su correspondiente URI mediante el cual se puede hacer referencia a él. Cada URI es único e irrepetible (es muy parecido al concepto de una cédula, en donde cada persona posee un número de cédula único e irrepetible. En la web también cada recurso posee un URI único que lo identifica).

Los URI se encuentran conformados por dos tipos de identificadores, estos se clasifican tomando en cuenta el aspecto al cual hagan referencia, de tal forma que un URI puede ser del tipo URL o URN tal como se muestra en la Figura № 6.

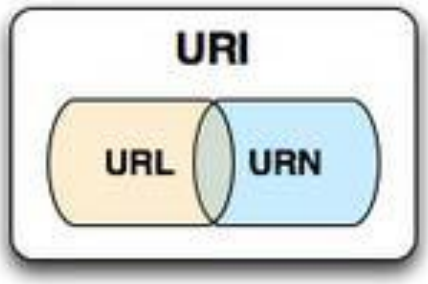


Figura 6 URI.
Fuente: Recuperado de: http://goo.gl/6jCKy8
Elaborado por: (Lee, n.d.)

Un URL (Uniform Resource Locator) dentro del contexto de la web semántica no permite la identificación de recursos, sin embargo como su nombre lo indica lo que hacen es servir como un medio que permita la localización de dichas entidades dentro de la web. (Universidad Nacional Abierta y a Distancia, n.d.). El URL puede contener los siguientes componentes mostrados en la Figura № 7:

Figura 7 Ejemplo URL.
Fuente: Basado en: http://goo.gl/VZ0xE4
Elaborado por: El Autor

En cambio los URN (Uniform Resource Name) en el contexto de la web semántica sirven para hacer referencia a un nombre que necesita permanecer global, permitiendo la identificación recursos, pero sin que impliquen su disponibilidad. Esto significa que no se pueden eliminar aun cuando el recurso no se encuentra disponible o haya dejado de existir. A continuación en la Figura № 8 se muestra como luciría un hipotético URN.

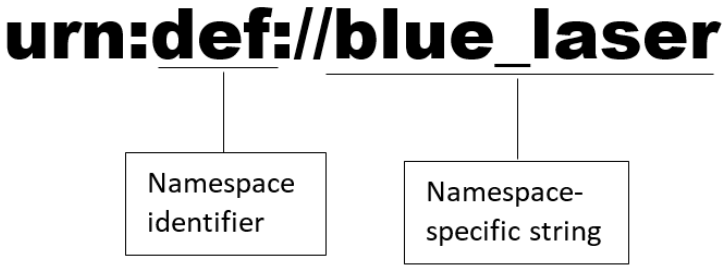


Figura 8 Ejemplo URN.
Fuente: Recuperado de: http://goo.gl/UgOk4W
Elaborado por: (Universidad Nacional Abierta y a Distancia, n.d.)

Cabe destacar que los URI se encuentran organizados de forma jerárquica, desde los más globales a los más específicos de manera similar a como se estructuran las ontologías, permitiendo que mediante el uso de caracteres especiales (mayormente "/") sumados a una sintaxis correcta se pueda utilizar los URI para dar referencia a un subconcepto específico. De esta manera se logra que todo el conocimiento que se encuentra en la web semántica modelado mediante el uso de ontologías pueda ser fácilmente ubicado u accedido mediante el uso correcto de URI. Un ejemplo de esto se muestra en la Figura № 9 donde el concepto especifico de "mamíferos" es accedido mediante URI correspondiente:

Figura 9 Ejemplo URI.
Fuente: Recuperado de: http://goo.gl/8gnQ9Z
Elaborado por: (Rouse, 2005)

1.2.3 RDF.

Al adentrarnos en el funcionamiento la web semántica resulta fundamental hablar de RDF y sus conceptos. Estas son las siglas en inglés de "Resource Description

Framework" y es una tecnología que tiene sus orígenes en el XML[15], esto debido a que originalmente fue pensado como una forma de codificación de metadatos para la parte superior de los documentos XML (Fernández, n.d.).

En un principio su uso se limitaba a la descripción de la información que se encontraba en un documento, permitiendo indicar aspectos tales como: la fecha de publicación, idioma, autor, entre otros.

Luego en el año 2004 el concepto y funcionalidad de RDF se ven alterados cuando se realiza una actualización, pasando se ser un medio para el manejo de metadatos de documentos a un modelo estándar para el intercambio de datos, obteniendo así una mayor importancia, tal como se menciona en (Tauberer, 2006). Define a RDF como:

> *"un modelo estándar para el intercambio de datos en la Web… que facilitan la fusión de los datos incluso si los esquemas subyacentes difieren, y soporta específicamente la evolución de los esquemas..."*
> (W3C -c, 2014)

RDF intenta resolver el problema subyacente en la web semántica de como representar la información, para esto hace uso de metadatos, que literalmente son datos acerca de datos. Según como lo menciona (Tauberer, 2006) RDF realiza dos tareas específicas.

La primera tarea es la descomposición de la información que se encuentra en la web semántica en estructuras pequeñas y manejables llamadas triple RDF, como lo menciona (Ruiz, 2008) cada triple RDF corresponde a una sentencia RDF, que en conjunto sirven para expresar el significado de los datos. El principal motivo para la utilización de triples RDF es que permiten una forma sencilla y entendible de representar el conocimiento, muy parecida a como lo hacemos las personas.

Estas estructuras se encuentran conformadas por tres partes fundamentales las cuales son: sujeto, predicado y objeto. La representación gráfica de una tripleta se realiza mediante grafos y su representación más simple se ve en la Figura № 10:

Figura 10 Modelo de Tripletas RDF.
Fuente: Recuperado de: http://goo.gl/IUpf8A
Elaborado por: (Rodríguez Álvarez, 2012)

[15] http://www.w3.org/XML/

Cada una de las partes que conforman una sentencia RDF es una representación de algo. Bajo esta premisa podemos decir que los sujetos y objetos son recursos dentro de la web semántica que representan a una entidad. Esta entidad puede ser un objeto, una persona, un lugar, una imagen, etc. Cada una de estas entidades se encuentran representadas por un identificador llamado URI (concepto mencionado en la sección "1.2.2 URI").

Así mismo como los sujetos y objetos son representaciones de algo, estos se encuentran relacionados por alguna característica común, esta característica o propiedad corresponde al predicado dentro de la tripleta y al igual que los casos anteriores también se encuentra representado por un URI, esto debido a que el predicado también es un recurso, cuyo objeto seria el valor asignado a él, ya sea un conjunto de caracteres o números, estos valores son llamados literales (Emilio & Gayo, n.d.).

La segunda tarea que realiza es la descripción de cómo codificar los triples RDF dentro de la web semántica. Para esto se han definido distintas formas de notación dependiendo de la situación, documento o información de la cual se trate. Entre algunas de las sintaxis que RDF nos ofrece podemos encontrar las siguientes: RDF/XML, Notación 3, Turtle, entre otros. (Para este caso no entraremos mucho en detalle acerca de los formatos, para mayor información se recomienda visitar la página de la W3C en su apartado de estándares[16])

A continuación se muestran algunos ejemplos de los diferentes tipos de sintaxis utilizados en la web semantica:

Notación 3:

En el siguiente ejemplo se describe la relación entre algunas películas y series de televisión.

```
@prefix rdf: <http://www.w3.org/1999/02/22-rdf-syntax-ns#> .
@prefix ex: <http://www.example.org/> .

ex:vincent_donofrio ex:starred_in ex:law_and_order_ci .
ex:law_and_order_ci rdf:type ex:tv_show .
ex:the thirteenth floor ex:similar plot as ex:the matrix .
```

Ejemplo tomado de:(Tauberer, 2006)

RDF/XML:

[16] https://www.w3.org/

A continuación se muestra un ejemplo de RDF/XML que muestra la relación entre películas y series de televisión

```
<rdf:RDF xmlns:rdf="http://www.w3.org/1999/02/22-rdf-syntax-ns#"
    xmlns:ex="http://www.example.org/">
    <rdf:Description
rdf:about="http://www.example.org/vincent_donofrio">
        <ex:starred_in>
            <ex:tv_show
rdf:about="http://www.example.org/law_and_order_ci" />
        </ex:starred_in>
    </rdf:Description>
    <rdf:Description
rdf:about="http://www.example.org/the_thirteenth_floor">
        <ex:similar_plot_as
rdf:resource="http://www.example.org/the_matrix" />
    </rdf:Description>
</rdf:RDF>
```

Ejemplo tomado de:(Tauberer, 2006)

Turtle:

Finalmente se muestra un ejemplo de Turtle usando la relación que existe entre los personajes el duende verde (Green Goblin) y el hombre araña (Spiderman) del comic estadounidense Spiderman:

```
@base <http://example.org/> .
@prefix rdf: <http://www.w3.org/1999/02/22-rdf-syntax-ns#> .
@prefix rdfs: <http://www.w3.org/2000/01/rdf-schema#> .
@prefix foaf: <http://xmlns.com/foaf/0.1/> .
@prefix rel: <http://www.perceive.net/schemas/relationship/> .

<#green-goblin>
    rel:enemyOf <#spiderman> ;
    a foaf:Person ;
    foaf:name "Green Goblin" .

<#spiderman>
    rel:enemyOf <#green-goblin> ;
    a foaf:Person ;
    foaf:name "Spiderman".
```

Ejemplo tomado de: (W3C -b, 2014)

1.2.4 OWL.

Una vez ya tenemos claro que es RDF podemos hablar de OWL que son las siglas en inglés de “Web Ontology Language”. OWL tal como se menciona en (W3C -b, 2012) es un lenguaje que nos permite representar el conocimiento abundante y complejo de las

cosas, llegando a ser considerado en ciertos casos como una extensión de los vocabularios que ofrece RDF.

Entrando en temas históricos OWL nace en el año 2004 de la necesidad de lenguajes ontológicos más expresivos, puesto que en la web semántica las búsquedas se realizan por contexto y no por coincidencia de palabras, era necesario un lenguaje que estuviera diseñado para procesar el contenido de la información en lugar de únicamente mostrarlo (W3C, 2004), siendo así que hoy en día OWL es uno de los lenguajes más utilizados gracias a sus múltiples características como son la cardinalidad, relación entre clases, igualdad y desigualdad tanto de clases como de instancias, entre otros (Fernández, n.d.).

Desde su aparición OWL fue actualizado y publicada una primera edición en el año 2009 que llegaría a conocerse como OWL 2, posteriormente se realizaría una publicación de la segunda edición en el año 2012.

Uno de los puntos fuertes de OWL son la gran cantidad de características y posibilidades de uso, pero esa también es una de las razones por las cuales fue necesario separarlo en sub-lenguajes que se adaptaran de mejor manera a las situaciones específicas en las cuales se los tendría que llegar a utilizar. Estos sub-lenguajes de OWL tal como se menciona en (W3C, 2004) son:

OWL Lite: es la opción más sencilla, ofrece la posibilidad de realizar jerarquías, clasificaciones y restricciones simples. La idea de OWL Lite es ofrecer las características mínimas de OWL permitiendo así una fácil implementación y manteniendo la eficiencia en casos en los cuales una implementación de los otros sub-lenguajes resultaría demasiado excesiva.

OWL DL: Es el segundo sub-lenguaje de OWL, se encuentra orientado a situaciones en las cuales se requiere un alto nivel de expresividad, pero implementando ciertas restricciones como la separación de pares entre clases, tipos de datos, ausencia de restricciones de cardinalidad, entre otros. Con estas características se logra un nivel de OWL bastante completo garantizando que todos los cálculos terminaran y que todas las conclusiones son entendibles por las máquinas.

OWL Full: Es la versión de sub-lenguaje más extensa y completa. No implementa ninguna restricción de los anteriores sub-lenguajes, además ofrece toda la libertad sintáctica permitida por RDF. Sin embargo su desventaja se encuentra en que debido a su gran nivel de complejidad resulta imposible garantizar que todas las conclusiones sean computables.

Los sub-lenguajes están organizados de forma jerárquica desde el más simple al más completo. Es por eso que se debe destacar que las implementaciones realizadas en los lenguajes más simples también son correctas para los más complejos. Esto se puede observar de manera gráfica en la Figura № 11:

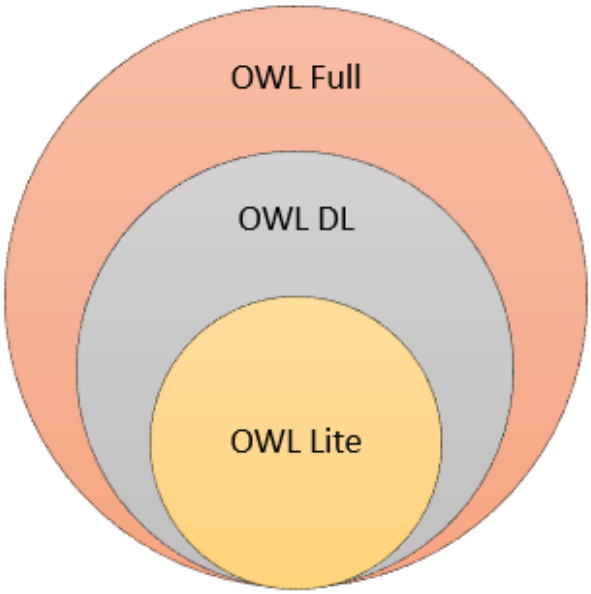


Figura 11 Versión de Lenguaje OWL.
Fuente: Recuperado de: http://goo.gl/uBvRol
Elaborado por: (Ruiz, 2008)

1.3. Extracción de datos

Al momento de pensar en la forma de extraer los datos de la web semántica resulta fundamental recordar uno de sus principales conceptos, el cual dice que esta versión de la web es muy similar a una “base de datos de conocimiento”, siendo así que la forma más fácil de extraer información es mediante la realización de consultas tal como seria en cualquier base de datos (W3C, 2008). Pero debido al gran tamaño y estructura de la web, realizar una consulta requiere de un lenguaje de consultas con características únicas, bajo este contexto que surge SPARQL como una solución ante estos problemas. Tal como lo explicó Tim Berners-Lee creador de la web

> *"Pretender usar la Web Semántica sin SPARQL es como pretender usar una base de datos relacional sin SQL"*
> (W3C, 2008)

Si bien SPARQL es un lenguaje de consulta, presenta muchas diferencias con otros lenguajes más tradicionales como pueden ser SLQ[17] y XQuery[18]. Sus diferencias se deben principalmente al momento en el cual fueron creados y a las necesidades que debían resolver cada uno.

Los lenguajes más tradicionales fueron diseñados para realizar búsquedas en almacenes de datos locales y como tal no afrontaron el problema de tener que manejar múltiples formatos, debido a que en la mayoría de los casos las consultas se limitaban a un formato específico. Además por su forma de diseño en muchas ocasiones resultaban ineficientes, provocando que una consulta debiera ser reformulada dependiendo de la aplicación o del modelo de base de datos relacional sobre la cual se estuviera trabajando.

Posteriormente con el advenimiento de la web semántica esto cambió, los datos ya se encontraban relacionados, por cuanto al momento de realizar una consulta se debía manejar no solo diferentes formatos, si no también distintos orígenes, debido a que la información podía encontrarse distribuida en múltiples ubicaciones y no siempre con las mismas propiedades.

Los lenguajes tradicionales no se encontraban preparados para este nuevo tipo de consultas, y para la obtención de datos de múltiples orígenes debían de ser realizadas numerosas consultas cuyos resultados serían unidos mediante declaraciones lógicas.

Ante todos estos problemas surgió SPAQRL como una opción ante los lenguajes de consulta tradicionales y con nuevas ideas de cómo realizar las tareas que hasta ese momento resultaban ser un gran problema para los desarrolladores. Tomando las palabras dichas por Tim Berners-Lee

> *"SPARQL hace posible consultar información desde bases de datos y otros orígenes de datos en sus estados primitivos a través de la Web"*
> (W3C, 2008)

SPARQL fue diseñado para trabajar a nivel web resolviendo muchos de los problemas de sus antecesores, permite realizar consultas de datos que no fuesen uniformes, además permite estructurar consultas a través de múltiples orígenes y como no se encuentra ligado a un formato de base de datos especifico resulta útil para la creación de aplicaciones, por todos estos motivos y muchos más SPARQL resulta no solo ser

[17] http://ocw.uoc.edu/computer-science-technology-and-multimedia/bases-de-datos/bases-de-datos/P06_M2109_02149.pdf
[18] http://www.w3.org/TR/2001/WD-xquery-20010215/

una herramienta fundamental para la extracción de datos dentro de la web, sino que además permite mejorar la eficiencia y calidad de los resultados que hasta ese momento se habían venido manejando.

1.4. Motores de representación

Si bien SPARQL ha significado muchos cambios en comparación con los modelos tradicionales, cualquier lenguaje de consulta resultaría inútil si no existiese una base de datos sobre la cual consultar, en contraste cada base de datos necesita de una forma de ser gestionada o controlada y es aquí donde aparecen los gestores de base de datos[19].

En la actualidad existen muchos diferentes gestores de base de datos, sin embargo los que se encuentran en la web semántica poseen algunas características que los diferencian de los demás, entre ellas podemos mencionar: mayor estabilidad, se encuentran orientados a objetos y diferente forma de estructurar los datos (The National Center for Biomedical Ontology, 2009).

Los gestores de base de datos tradicionales o SQL, por lo regular suelen poseer una estructura relacional donde las tablas de una base de datos se relacionan unas con otras mediantes diferentes atributos, en cambio en los gestores de base de datos usados en web semántica, existen diferentes formas de clasificar y estructurar las información, entre ellas podemos mencionar las basadas en Columnas, documentos, valores clave y las fundamentadas en Grafos (acens, 2014). Entre algunos de los principales gestores de base de datos podemos mencionar:

Cassandra.- Es una base de datos de licencia libre, la cual como se menciona en (acens, 2014) se encuentra orientada a un modelo de almacenamiento clave-valor. Entre sus principales características podemos resaltar su escalabilidad lineal y gran rendimiento, además posee gran tolerancia a fallos convirtiéndose en una de las más usadas actualmente, e implementada en importantes compañías tales como: Netflix, GitHub, Instagram etc.

Virtuoso Univesal Server.- Tal como se menciona en (The National Center for Biomedical Ontology, 2009) Virtuoso es un middleware hibrido que posee dos diferentes versiones una de pago y una versión gratuita. Entre algunas de las características que

[19] http://ocw.uoc.edu/computer-science-technology-and-multimedia/bases-de-datos/bases-de-datos/P06_M2109_02147.pdf

posee podemos destacar que combina funcionalidades de una base de datos objeto relacional, de objetos, RDF, XML y documentos. Además resulta ser sumamente escalable y mediante la utilización de conectores puede trabajar con Jena, Oracle, entre otros.

MongoDB.- Al igual que en los casos anteriores MongoDB es una base de datos no relacional y de código abierto. Tal como se expone en (Mongo DB, 2015) MongoDB resalta por su agilidad y estabilidad. Su estructura se basa en guardar la información en documentos de tipo JSON[20] llamados BSON, estos son agrupados en colecciones, permitiendo así gran velocidad y sencillez en sus consultas. Además en (The National Center for Biomedical Ontology, 2009) se destaca que gracias a la utilización de este formato para almacenar la información es uno de los sistemas de base de datos más adaptables y que ofrecen mayor integración con las aplicaciones.

En la Tabla 1 se muestra parte de un estudio realizado por (DB-Engines, n.d.) donde nos muestran algunas de las características que poseen estos motores de representación, para mayor detalle revisar la tabla completa en el Anexo 1.

Tabla 1 Comparación de Propiedades del Sistema Casandra vs. MongoDB vs. Virtuoso

Nombre	**Cassandra**	**MongoDB**	**Virtuoso**
Descripción	Almacén de datos basada en ideas de BigTable y DynamoDB	Almacén de datos basado en documentos	Servidor de datos multi-modelo
Licencia	Código Abierto	Código Abierto	Código Abierto
Lenguaje de implementación	Java	C ++	C
Modelo de base de datos	Almacén de columna	Almacén de documentos	XML nativo DBMS, relacional y almacén RDF
API y otros métodos de acceso	Protocolo propietario	Protocolo propietario mediante JSON	- OLE DB - ADO.NET - JDBC - ODBC

Fuente: Recuperado de: http://goo.gl/obRXvV
Elaborado por: (DB-Engines, n.d.)

Como podemos apreciar existen muchas características que diferencian a un sistema de gestión de base de datos de otro, sin embargo debido a que distintos entornos requieren diferentes características resulta imposible decir que existe un sistema de gestión que sea aplicable a todos los casos.

[20] http://json.org/

Otra gran opción que cabe incluir cuando de gestores de base de datos se trata es Oracle. La popular compañía de software como se menciona en (Oracle, 2015) posee una distribución de gestor de base de datos NoSQL, esta utiliza el formato clave-valor y fue diseñada para la web semántica. Sus principales características son la fiabilidad, escalabilidad y disponibilidad.

Oracle gestor de base de datos pagado puesto que brinda soporte técnico y otras comodidades. Sin embargo también posee una distribución para desarrolladores que permite realizar pruebas de concepto y ayudar con la construcción de las aplicaciones. Esta distribución es llamada KVLite y es una versión simplificada que permite la ejecución de un único proceso sin necesidad de una interfaz administrativa.

Finalmente una opción que resulta muy popular hoy en día es la utilización de servidores SPARQL que trabajan conjuntamente con bases de datos relacionales para realizar el almacenamiento de información.

En este contexto resulta particularmente notable la herramienta llamada "Apache Marmotta[21]" por ser una de las más utilizadas y completas actualmente. Su popularidad de debe principalmente a la gran cantidad de módulos y funciones que posee, permitiendo así que mediante su utilización se encierre la mayoría de las actividades referentes a la publicación de datos en enlazados.

Entre sus principales características podemos denotar la integración de módulos de seguridad básicos, permite la realización de consultas SPARQL ya sea mediante el uso de la interfaz gráfica de la aplicación o con el uso de servicios web, se encuentra adaptado para trabajar con tres diferentes bases de datos como son MySQL[22], PostgreSQL[23], o H2[24], posee módulos para la realizar la carga de información, entre otras características.

1.5. Linked data

Este es uno de los principales conceptos de la web semántica y el tema principal sobre el que se sustenta este proyecto de tesis. Si bien ya hemos hablado de las tecnologías y diversos componentes de la web semántica, podríamos decir que "Linked Data" (o

[21] http://marmotta.apache.org/
[22] https://www.mysql.com/
[23] http://www.postgresql.org.es/
[24] http://www.h2database.com/html/main.html

también llamados “Datos Enlazados” como es su traducción al español) es el corazón y centro de la misma.

La web semántica pretende ser una base de conocimiento, pero para lograrlo debemos conseguir que la información se relacione y tenga sentido (Muchas de estas relaciones se logran mediante la utilización de Triples RDF mencionados en el apartado “1.2.3 RDF.”). Con esto se consigue al enriquecer el conocimiento aislado y utilizarlo de mejor manera. Para comprender mejor la importancia de “linked data” supongamos que existe una empresa que comercializa cierto tipo de plantas, ellos emplean información del crecimiento de las plantas de la región en relación con el clima, y así utilizan estos datos para poder fijar los precios a los cuales vender las plantas, sin embargo sus proyecciones se basan en estimaciones climáticas a corto plazo, a ellos les convendría relacionar su información con la que posee un centro climatológico, con la finalidad de tener proyecciones más precisas y poder estar preparados ante una eventual caída de precios.

En este caso se ha relacionado netamente la información climatológica y económica; sin embargo en la realidad, se puede relacionar casi cualquier información mediante los datos que tengan en común o a los cuales pertenezcan.

Un ejemplo de esto sería el concepto de la fruta manzana, que puede estar relacionado con el de la plantas, que a su vez se relaciona con el de seres vivos y así sucesivamente con múltiples conceptos (W3C, 2010).

Estas relaciones e integraciones a gran escala es sobre lo que se trata “linked data” y resulta el principal motivo por el cual podemos crear una web semántica y expandirla continuamente mediante la conexión de información, a tal punto que una representación gráfica de estos datos al año 2014 serían muy similar a lo que se observa en la Figura № 13.

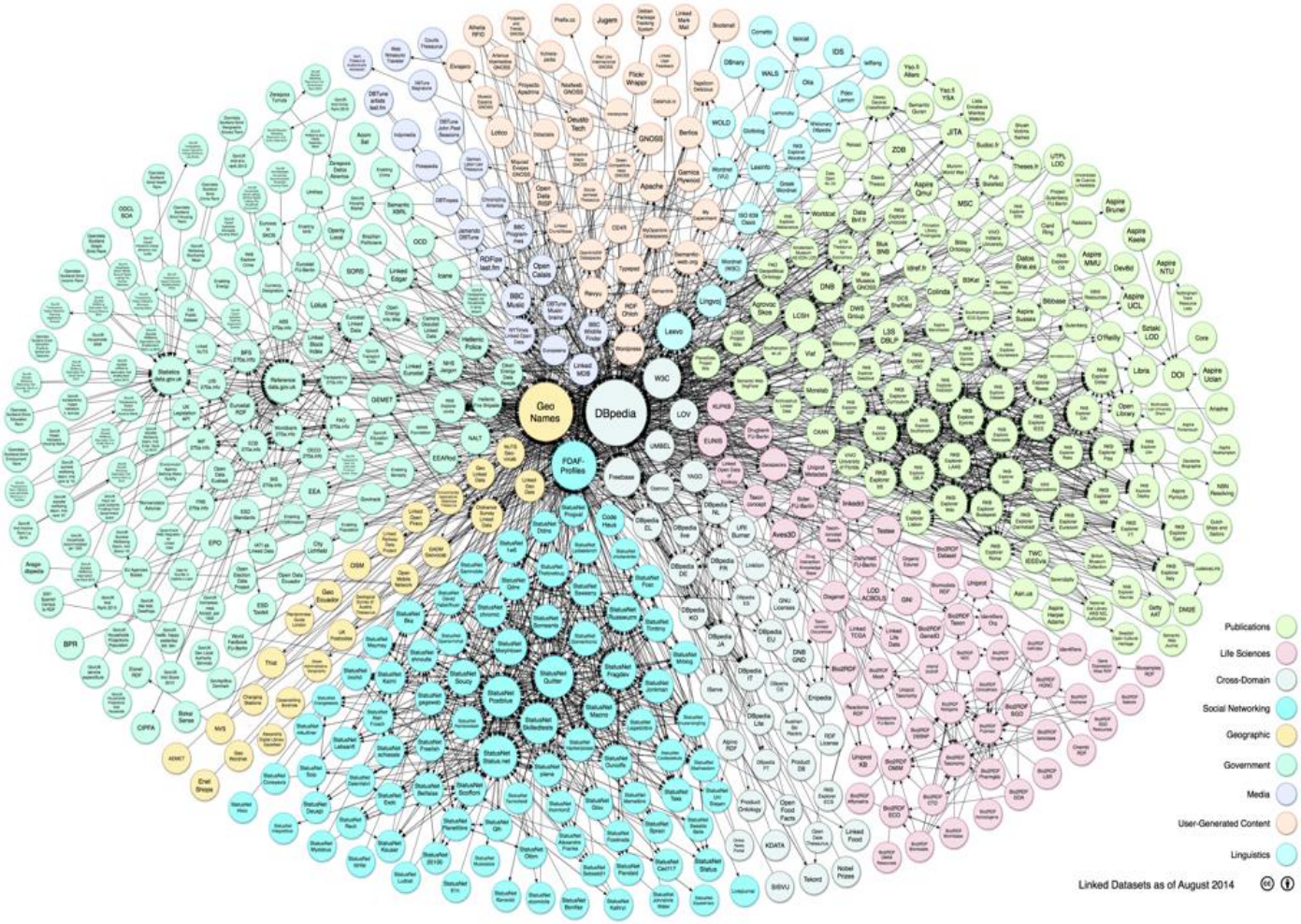


Figura 12 Grafo Linked Data.
Fuente: Recuperado de: http://goo.gl/HcTLC9
Elaborado por: (Cyganiak & Jentzsch, 2014)

1.6. Iniciativas de linked data

“Linked data” se encentra en constante desarrollo e investigación mediante iniciativas, tesis y demás proyectos. Hablar de “linked data” es también hablar de todas estas iniciativas, pues todas ellas tienen algo que aportar y conocimiento que puede ayudar a los demás. Es bajo esta premisa que en este apartado se hablan a breves rasgos de algunas iniciativas importantes con “linked data”.

1.6.1. Linking open data.

Es un proyecto impulsado por la W3C y nuestro primer punto de referencia en este apartado. El objetivo de este proyecto es extender la Web semántica con la publicación de diversos conjuntos de datos en formatos RDF y el establecimiento de enlaces entre ellos, de tal forma que la información se encuentre relacionada, pudiendo así ser utilizada por diversos buscadores web semánticos.

Actualmente existen varios importantes conjuntos de datos formando parte de este proyecto como son Wikipedia, Wikibooks, Geonames, MusicBrainz, WordNet, entre otros. Un aspecto importante de este proyecto es que ha servido para impulsar múltiples iniciativas de publicación de "linked data" por parte de muchos países europeos los cuales comienzan a entender la importancia de publicar sus datos en la web, tomando especial importancia los datos estadísticos y geográficos (Heath & Bizer, 2011).

Entre los principales vocabularios resaltan el vocabulario RDF Data Cube y SDMX-RDF los cuales son ampliamente usados por los proyecto de "linked data" en especial la organización europea Eurostat. Hablaremos un poco más de cómo los vocabularios son utilizados en la sección "1.7 Origen de los Vocabularios Estadísticos".

1.6.2. DBpedia-Latinoamérica.

Uno de los proyectos más representativos del movimiento de "Linked Data" dentro de Latinoamérica es el llamado DBpedia-Latinoamérica, desarrollado en colaboración por la Universidad Técnica Particular de Loja en donde se encuentra alojada la información, la Universidad de Cuenca y la Asociación de Linked Data – Latinoamérica.

Esta iniciativa busca apoyar y enriquecer la información que se encuentra en la web además de servir de inspiración para la creación de nuevos mecanismos para navegación y enlazado de datos, todo esto realizado mediante la publicación de artículos escritos en español y provenientes de la Wikipedia, los que se relacionan con los datos existentes en la nube de datos abiertos enlazados y pueden ser recuperados mediante la ejecución de consultas (Dbpedia, n.d.).

1.6.3. Repositorio ecuatoriano de datos enlazados geoespaciales.

En un ámbito más regional podemos resaltar la iniciativa del Repositorio Ecuatoriano De Datos Enlazados Geoespaciales la cual tiene como objetivo la generación, publicación y consumo de colecciones de datos abiertos y enlazados dentro de nuestro país. La iniciativa se encuentra compuesta por varios grupos de investigación, organizaciones gubernamentales y algunas universidades entre las cuales podemos mencionar la participación de la Universidad Técnica Particular de Loja, la Universidad de Cuenca y múltiples investigadores independientes relacionados con este campo (Geo Linked Data Ecuador, 2014).

1.6.4. Plataforma de integración, publicación y consulta integrada de recursos bibliográficos en la web semántica.

Es un proyecto impulsado por la REDCEDIA (Red Nacional de Investigación y Educación del Ecuador) tiene como objetivo realizar el análisis, generación y publicación de Linked Data de las bibliotecas de las instituciones que participan. Actualmente este proyecto está siendo impulsado por la Universidad Técnica Particular de Loja, la Universidad de Cuenca y la Escuela Politécnica Nacional.

Para conseguir dicho objetivo han efectuado el diseño e implementación de un framework que permita un acceso unificado centrado en el usuario, en cualquier lugar y a cualquier hora dichos recursos, todo esto basándose en tecnologías de servicios Web, Web Semántica, integración de información basada en ontologías y librerías que permitan crear interfaces amigables. (REDCEDIA, n.d.)

1.6.5. Proyectos dentro de la universidad técnica particular de loja.

Ya orientados en un aspecto institucional cabe destacar que la Universidad Técnica Particular de Loja ha venido desarrollado múltiples investigaciones y proyectos en el área de la web semántica, dando como resultado importantes publicaciones en diferentes revistas científicas las cuales se recomienda revisar ingresando al “Sistema de Información Académica Científica” de la Universidad Técnica Particular de Loja (Universidad Técnica Particular de Loja, n.d.).

Si bien estos son solo algunos de los proyectos de “linked data” que podemos encontrar a nivel mundial, podemos notar como la web semántica cada vez toma más fuerza e importancia no solo en el ámbito informático sino también en el manejo de los datos.

En contraste en Ecuador los proyectos de este tipo son cada vez más impulsados por parte de universidades que han optado por la realizar investigaciones en esta área, con la finalidad de promover el desarrollo de estas tecnologías tanto dentro como fuera del país.

1.7. Origen de los vocabularios estadísticos

Actualmente la información que se publica en la web semántica crece día tras día, encontrándose toda clase de datos en ella. Uno de los tipos más relevantes de información que existen en la web semántica son las publicaciones estadísticas, que resultan ser de sumo interés no solo para personas particulares sino también para múltiples instituciones y empresas que las usan constantemente para poder tomar decisiones críticas con respecto a cómo dirigir sus negocios.

Hoy en día el principal formato utilizado en la web semántica para gestión la información es el RDF, por tanto para poder publicar nuevos datos en dicha web; éstos primero deben ser transformados a ese formato. Esta tarea es una de las más importantes en la publicación de la información y por lo regular implica la utilización de vocabularios que permitan modelar los datos, sin embargo en un principio existieron diferentes posturas de cómo se debía realizar esta transformación.

En nuestro caso nos vamos a centrar en una de las propuestas realizada por Eurostat[25], específicamente a la que originó la iniciativa Riese (W3C -a, n.d.), esto debido a ser una de las primeras y más ampliamente utilizadas en la actualidad.

Riese son las siglas de “RDFising and Interlinking the Eurostat Data Set Effort” y su principal objetivo era dar a conocer los datos de Eurostat en la web semántica. La más impórtate contribución de esta iniciativa fue la aceptación de que los datos debían transformarse a RDF, en lugar de como sugerían otras posturas realizar un empaquetamiento o algún otro tipo de tratamiento a los datos para su posterior publicación, con Riese se consiguieron gran cantidad de datos estadísticos de las diversas ciudades de Europa.

[25] http://ec.europa.eu/eurostat/about/overview

Una vez Riese sentara las bases para la transformación de la información comenzaron a surgir diversos vocabularios cada uno orientado a ciertos tipos de datos. En el aspecto estadístico surgieron vocabularios como: SKOS[26], SCOVO[27], SCOVOLink[28], entre otros.

SKOS son las siglas en inglés de la iniciativa "Simple Knowledge Organization System" que dio origen al vocabulario que lleva su nombre. Tal como se menciona en ("Introduction to SKOS," 2012) este vocabulario permitía representar los sistemas de organización del conocimiento utilizando el formato RDF. Si bien este se orientaba al conocimiento en general era sumamente útil en ciertos casos para poder representar información estadística.

Por su parte SCOVO ya era un vocabulario orientado netamente a la información estadística. En éste ya se comenzaba a incluir al tiempo como un aspecto importante de la representación de datos estadísticos. Además uno de sus aspectos fundamentales como se menciona en (W3C -a, n.d.) fue la definición de su estructura considerando tres conceptos básicos los cuales son: el conjunto de datos, los elementos y las dimensiones. Posteriormente este vocabulario sería extendido dando origen a SCOVOLink, que como se menciona en (Mynarz, Cyganiak, Iqbal, & Hausenblas, n.d.) aborda el dominio de la semántica, encargándose de la manera en la cual un conjunto de datos se refiere a las cosas sobre las cuales trata.

A pesar de todo esto SCOVO aun poseía serias limitaciones, entre las más importantes destaca la forma de describe su contenido, siendo muy limitada y dando como resultado que la recuperación de la estructura fuera muy compleja.

Otra iniciativa que influyó de gran manera a los vocabularios estadísticos fue la iniciativa SDMX que son las siglas de "Statistical Data and Metadata Exchange". De esta iniciativa surgieron múltiples estándares tanto para el intercambio como para manejo de datos y metadatos estadísticos, siendo una de las más importantes la SDMX/RDF, que tal como se menciona en (W3C -a, 2012) consistía en traducir el estándar SDMX para el formato RDF y al final conformar su propio vocabulario.

Posteriormente surgiría el vocabulario RDF Data Cube cuyo núcleo podemos decir se encontraba basado en la iniciativa SDMX (específicamente SDMX 2.0). Este nuevo vocabulario en algunos casos suele ser considerado como una evolución de SCOVO desde el punto de vista que adoptaba su estructura multidimensional, pero como se menciona en (Mynarz et al., n.d.) el RDF Data Cube añade una mayor fuerza expresiva

[26] https://www.w3.org/2004/02/skos/
[27] http://sw.joanneum.at/scovo/schema.html
[28] http://vocab.deri.ie/scovolink

cuando se trata de describir la estructura interna de un conjunto de datos, solucionando así el problema fundamental de SCOVO.

Sin embargo es más correcto considerar al RDF Data Cube como el agrupamiento no solo de los vocabularios e iniciativas hasta aquí mencionadas sino también de algunos otros como son: FOAF, DCMI y VoID.

El vocabulario FOAF, tiene su nombre debido las siglas en inglés de la frase “Friend of a Friend”. Tal como se menciona en (“FOAF Vocabulary Specification,” 2014) FOAF sirve como diccionario de propiedades y clases con nombre propio utilizando la tecnología RDF.

Los vocabularios DCMI y VoID son utilizados normalmente para la gestión de metadatos. El primero debe su nombre a las siglas de la iniciativa “Dublin Core Metadata Initiative” la cual dio como resultado un modelo para el manejo de metadatos. Mientras que el vocabulario VoID sirve para expresar metadatos de conjuntos de datos RDF. Este vocabulario como se explica en (W3C, 2011) pretende ser un puente entre los editores y usuarios de datos RDF. Ambos son utilizados hasta cierta medida por el vocabulario RDF Data Cube para el manejo de metadatos.

1.8. RDF data cube

Llegados a este punto podemos decir que conocemos a breves rasgos los diversos cambios y adaptaciones que han sufrido los vocabularios estadísticos a través del tiempo desde sus orígenes hasta nuestros días.

Es por ese motivo que este resulta ser el momento perfecto para profundizar en la estructura del vocabulario RDF Data Cube. Esto debido a que será el vocabulario utilizado durante la parte práctica del presente proyecto de tesis. Las motivaciones para su utilización se encuentran detalladas en la sección “4.2 ¿Qué Vocabulario Utilizar?” del capítulo IV.

1.8.1. Estructura.

Uno de los primeros puntos a ser considerado para poder entender este vocabulario es su estructura, la misma que se muestra en la Figura № 13:

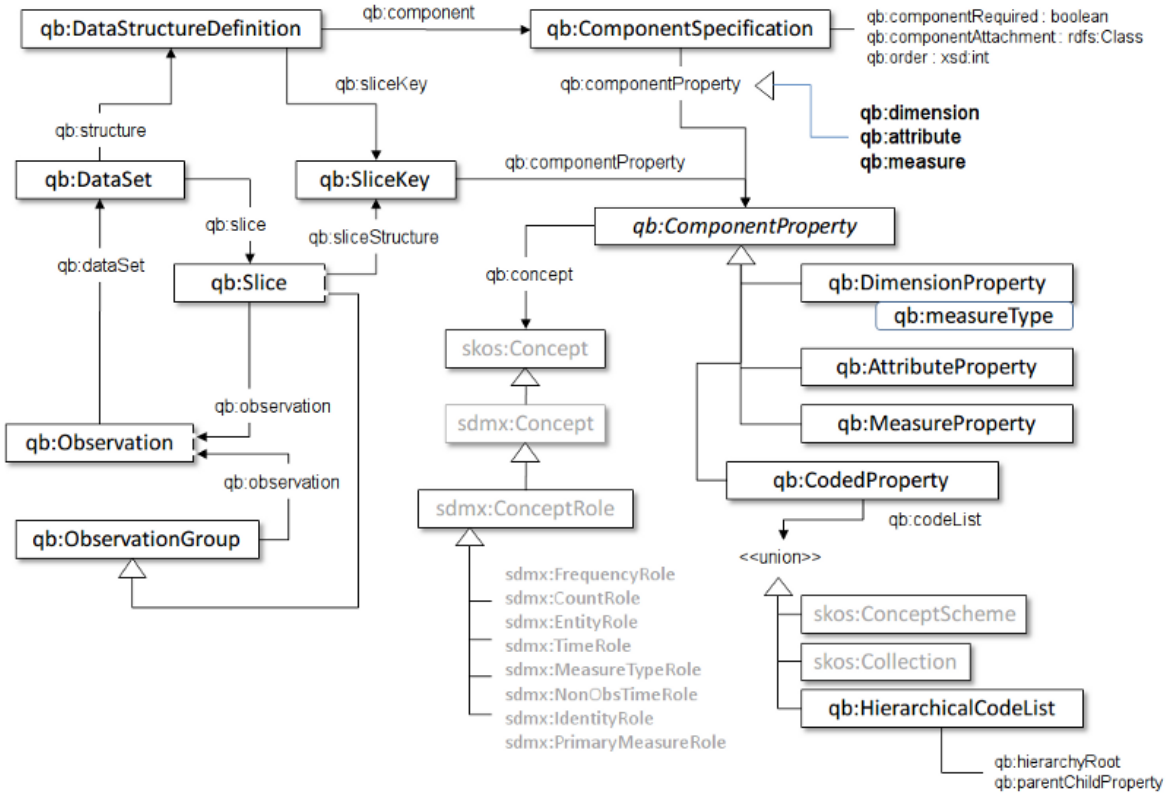
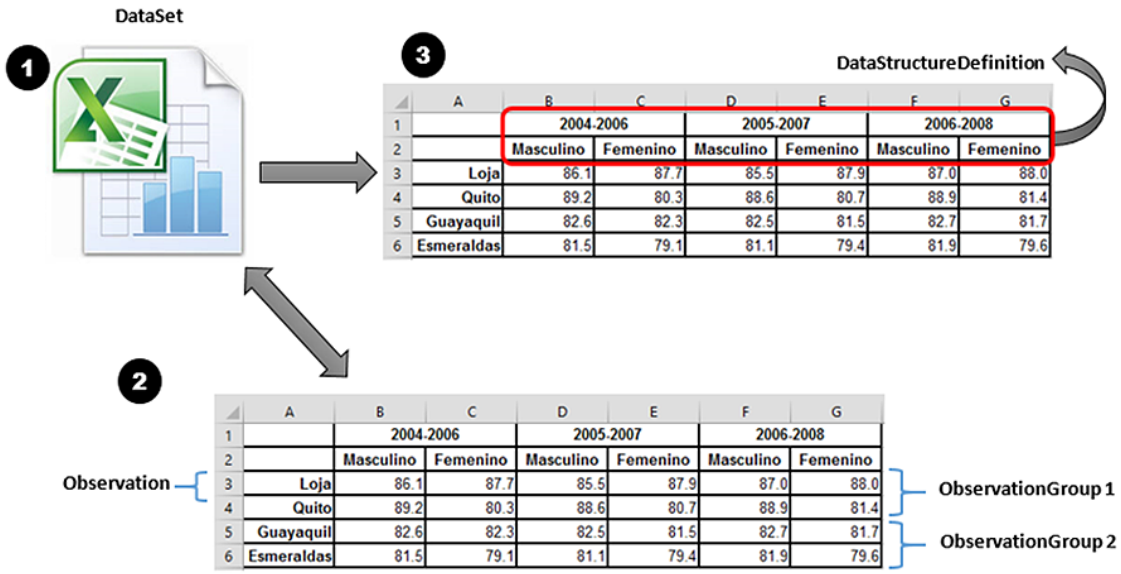


Figura 13 Estructura Data Cube.
Fuente: Recuperado de: http://goo.gl/W9je1W
Elaborado por: (W3C -d, 2014)

Podemos observar múltiples relaciones entre los componentes del vocabulario. En este caso se explica las principales relaciones existentes partiendo desde el qb:Dataset, sin entrar mucho en detalle en algunos conceptos. Para mayor información se recomienda visitar el sitio web de la W3C en su apartado “The RDF Data Cube Vocabulary”.

El qb:DataSet hace referencia al conjunto de datos sobre los cuales se está trabajando, y su representación gráfica se encuentra en la Figura № 14 apartado 1 y en el mundo real equivaldrían al conjunto de archivos documentos donde se encuentra la información.

Estos datos son definidos como un conjunto de observaciones (qb: Observation) de algún fenómeno las cuales por lo regular son almacenadas con el pasar del tiempo, su equivalente se muestra en la Figura № 14 apartado 2 donde se observa que cada observación sería un registro de la información que se posee.



	A	B	C	D	E	F	G
1		2004-2006		2005-2007		2006-2008	
2		Masculino	Femenino	Masculino	Femenino	Masculino	Femenino
3	Loja	86.1	87.7	85.5	87.9	87.0	88.0
4	Quito	89.2	80.3	88.6	80.7	88.9	81.4
5	Guayaquil	82.6	82.3	82.5	81.5	82.7	81.7
6	Esmeraldas	81.5	79.1	81.1	79.4	81.9	79.6

Figura 14 Estructura de Data Cube en Formato Físico
Fuente: El Autor
Elaborado por: El Autor

Es de aquí de donde nace una relación entre los datos y su proyección en un tiempo, llegándose a obtener una relación de tres dimensiones, la misma que suele estar representada mediante un cubo, en esta ocasión la Figura № 15 nos muestra una representación gráfica. (Cabe destacar que dependiendo de los datos con los que se trabaje pueden existir más o menos dimensiones).

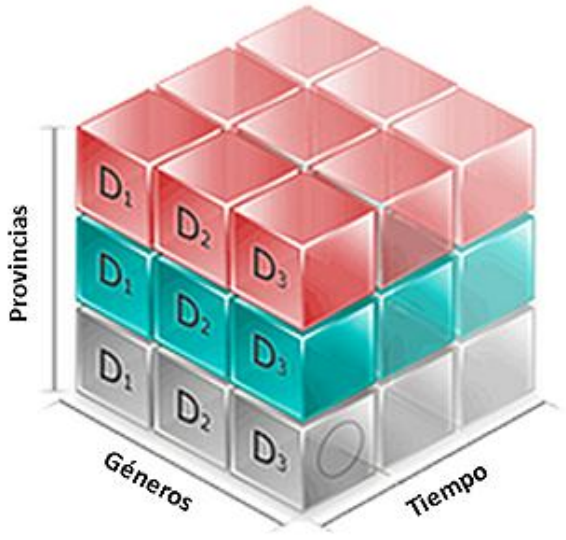


Figura 15 Cubo de datos.
Fuente: Recuperado de: https://goo.gl/gm5kNr
Elaborado por: (Tiger Logic, 2013)

Del conjunto de observaciones que conforman el cubo de datos podemos obtener fragmentos los cuales representan información considerando grupos específicos de observaciones (qb: ObservationGroup) en la Figura № 14 apartado 2, se encuentra un ejemplo de esto conformándose dos diferentes grupos de observaciones, estas porciones de datos son representadas por la propiedad qb:Slice. A su vez estos grupos de observaciones necesitan ser identificadas, por cuanto cada qb:Slice se encuentra ligada a un identificador denominado qb:SliceKey.

Continuando con la estructura del vocabulario, debemos recalcar que por lo regular cada qb:DataSet posee una definición de su estructura, identificada por la propiedad qb:DataStructureDefinition que en muchos casos también es utilizado para definir las qb:Slice. Gran parte de la importancia de la qb:DataStructureDefinition se encuentra en que nos ayuda con la verificación de información incoherente, al mismo tiempo que nos brinda la posibilidad de reutilizar esta misma estructura para cada publicación y así tener una serie regular, mostrado de manera gráfica esto equivaldría a los encabezados de los archivos de los documentos de donde se obtienen los datos (en algunos casos se contemplan más aspectos). Esto se observa gráficamente en la Figura № 14 apartado 3.

El qb:DataStructureDefinition como se menciona en (W3C -d, 2014) hace referencia a un conjunto de especificaciones de componentes, los cuales son identificados por la propiedad qb:ComponentSpecification. Estas permiten definir si los componentes son obligatorios, requeridos, opcionales, etc.

Por otro lado para poder hacer referencia a los componentes dentro del vocabulario se utiliza la propiedad qb:ComponentProperty, la misma comprende el concepto a ser representado (población, región geográfica, etc.), la naturaleza del componente (dimensiones, medidas, atributos) y la lista de códigos que serán utilizados para representar el valor.

En situaciones en las cuales la definición propia del concepto a ser representado resalta de valor, se recomienda la utilización de los vocabularios SKOS y SDMX. Ambos facilitan no solo la definición del concepto, sino que también proveen la posibilidad de asignar roles según sea necesario, pudiendo elegir FrequencyRole, TiemRole, CountRole, EntityRole, entre otros.

Cabe destacar que cada qb:ComponentProperty se encuentra a su vez estructurado por tres elementos principales los cuales son: qb:DimensiónProperty que hace referencia a las dimensiones, qb:MeasureProperty que hace referencia a las medidas y qb:AttributeProperty que se refiere a los atributos. Estos elementos son la base

fundamental sobre la cual se sustenta la estructura de este vocabulario y por tanto vale la pena explicarlos un poco más a fondo.

Dimensiones.- Permiten identificar el fenómeno que se está observando o estudiando.

Medidas.- Son las que permiten representar al fenómeno como tal, son los valores observados.

Atributos.- Estos permiten especificar las unidades de medida, escalas y nos ayudan a interpretar los valores. Son metadatos de la información con la cual se está trabajando y su utilización puede llegar a ser opcional.

Para poder entender mejor cada uno de estos conceptos y sus diferencias presentamos el siguiente ejemplo basado en el presentado en (W3C -d, 2014). En la Tabla 2 observamos un conjunto de datos imaginarios los cuales hacen referencia a la esperanza de vida desglosada por regiones, la edad y el tiempo.

Tabla 2 Datos esperanza de vida Ecuador

	2004-2006		**2005-2007**		**2006-2008**	
	Masculino	**Femenino**	**Masculino**	**Femenino**	**Masculino**	**Femenino**
Loja	86.1	87.7	85.5	87.9	87.0	88.0
Quito	89.2	80.3	88.6	80.7	88.9	81.4
Guayaquil	82.6	82.3	82.5	81.5	82.7	81.7
Esmeraldas	81.5	79.1	81.1	79.4	81.9	79.6

Fuente: El Autor
Elaborado por: El Autor

Partiendo de estos datos podemos identificar que existen tres dimisiones en este conjunto de datos, dichas dimensiones son: periodos de tiempo, género y las regiones. Existe una medida correspondiente a la esperanza de vida de las diferentes poblaciones y finalmente un atributo referente a la escala que se está utilizado, en este caso puntual años.

Como podemos observar el vocabulario RDF Data Cube consta de múltiples partes las cuales se interrelacionan, sin embargo sus relaciones no se limitan solo con sus propias definiciones, en muchos de los casos también se incluyen referencias a otros vocabularios, todo esto basado en la idea de poder reutilizar vocabularios y adaptar la solución dependiendo del problema o los datos que se utilizan. Entre las relaciones más importantes podemos caracterizar las siguientes

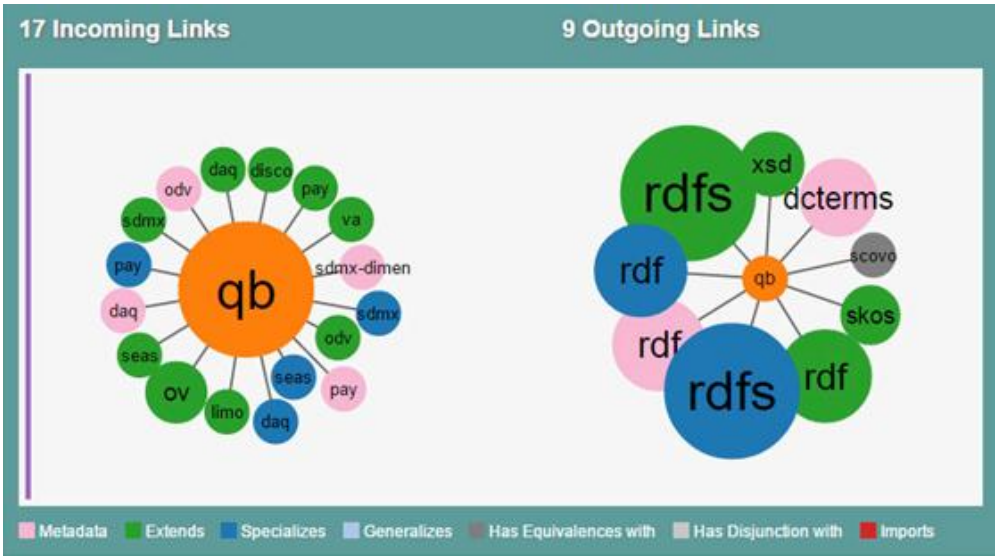


Figura 16 Relación del Data Cube con Otros Vocabularios.
Fuente: Recuperado de: http://goo.gl/XicPE9
Elaborado por: (Linked Open Vocabularies, n.d.)

1.8.2. Data cube frente a otros vocabularios.

Tal como se apreció en la sección "1.7 Origen de los Vocabularios Estadísticos", los vocabularios han evolucionado y cambiando con el paso del tiempo, sin embargo han conservado ciertas características, mientras que otras han cambiado radicalmente. En este apartado resaltaremos las principales características que posee el vocabulario Data Cube que lo diferencia de los demás.

Una de esas características a resaltar es su carácter multidimensional con lo que se puede obtener fragmentos de información, que con otros vocabularios como el SCOVO resulta imposible. Además una de las grandes ventajas que ofrece es su adaptabilidad frente a diversos entornos y exigencias, puesto que al estar basado netamente en el núcleo de la iniciativa SDMX y no abarcar completamente su extensión, puede realizar la tarea de modelar datos estadísticos de mejor manera, dejando la definición de conceptos y otros aspectos del modelado a vocabularios especializados como son el SKOS, DCMI, entre otros.

También otra importante característica que lo diferencia de los demás vocabularios es que brinda la posibilidad de definir sus respectivas propiedades y clases según sea necesario, todo esto gracias al uso de sus diferentes componentes como son el qb: DimensionProperty, qb:MeasureProperty, entre otros.

Finalmente al hablar del vocabulario Data Cube no podemos dejar de mencionar su definición estructural interna, que resulta ser sumamente sólida al hacer uso de la

propiedad qd:DataStructure y que a su vez también permite la reducción sustancial en la complejidad de las consultas, algo que resultaba imposible con otros vocabularios estadísticos como el SCOVO. Es por todas estas características que hoy en día el Data Cube resulta ser la mejor opción cuando de modelado de datos estadísticos se refiere y motivos por los cuales es el vocabulario elegido para presente proyecto.

No obstante aún existen determinados casos en los que opciones de vocabularios más sencillas como el SCOVO resultan ser suficientes. En estas situaciones por lo regular la información a ser modelada es sumamente sencilla y no presenta mayores complicaciones. A pesar de eso la elección de un vocabulario para la representación y modelado de datos resulta ser una tarea de crucial importancia, por tanto se debe analizar cuidadosamente si los datos son lo suficientemente sencillos como para usar opciones parecidas a SCOVO u optar por una más completa como sería el vocabulario Data Cube.

1.9. Ciclo de vida

Comenzando a adentrarnos un poco más en aspectos prácticos, resulta de interés conocer uno de los conceptos más importantes dentro de la web semántica al momento de trabajar con datos como es el ciclo de vida de "Linked Data".

Este nos permite estructurar y separar las actividades necesarias para poder publicar los datos, actualmente existen algunas opciones entre las cuales podemos elegir, las más importantes son las propuestas por Bernadette Hyland[29], Michael Hausenblas[30] y Villazón-Terrazas[31].

Si bien cada opción posee sus propias características y estructuran las actividades de diferente forma, el trabajo que se realiza es mismo por cuanto es evidente que comparten e incluso superponen muchas de las tareas que se realizan.

En este caso nos vamos a centrar en el modelo propuesto en (Villazón-Terrazas, Vilches-Blázquez, Corcho, & Gómez-Pérez, 2011[32]) debido a que es el que mejor especifica las diferentes etapas que comprende la publicación de "linked data", además de establecer las actividades específicas a desarrollarse en cada una de ellas, permitiendo así especificar el trabajo, facilitar la planeación y organización de las mismas.

[29] https://www.w3.org/2011/gld/wiki/Linked_Data_Cookbook
[30] http://linked-data-life-cycles.info/
[31] http://link.springer.com/chapter/10.1007/978-1-4614-1767-5_2
[32] http://www.w3.org/2011/gld/wiki/GLD_Life_cycle#Villazon-Terrazas_et_al.

El modelo se encuentra conformado por cinco etapas, donde cada una de ellas cuenta con diferentes actividades a ser desarrolladas y que se relacionan directamente con la etapa siguiente, a continuación en la Figura № 17 se muestra una ilustración gráfica del modelo y se explica brevemente cada una de ellas.

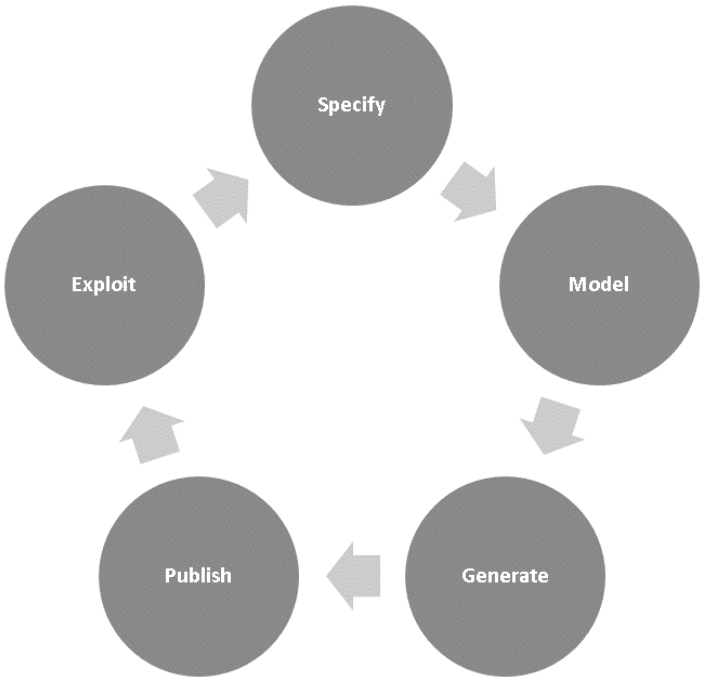


Figura 17 Ciclo de Vida de Linked Data.
Fuente: Recuperado de: http://goo.gl/R712Fl
Elaborado por: (W3C -c, 2012)

- **Especificación.-** En esta etapa tal como su nombre lo sugiere se realizan todas las actividades de especificación referentes a los datos. Se determina las fuentes de información a ser utilizadas, se obtienen los datos y finalmente se analizan para determinar su factibilidad de ser usados en el proyecto.
- **Modelado.-** Durante esta etapa se trabajan todos los aspectos relacionados al modelamiento de los datos, se defienden ontologías a ser desarrolladas o reutilizadas dependiendo del caso y las formas en las cuales se organizara la información una vez que se haya convertido a formato RDF.
- **Generación.-** Se trabajan todos los aspectos referentes a la generación de los archivos RDF, partiendo de los datos obtenidos en la etapa de especificación y basándose en las ontologías trabajadas durante la etapa de modelado.

- **Publicación.-** Una vez obtenidos los archivos RDF, estos deben de ser publicados y puestos a disposición de la comunidad para ser utilizados y consultados, todos estos aspectos son trabajados en esta etapa.
- **Explotación.-** Se trabajan todos los temas relacionados al uso y visualización de los datos.

1.10. Comentarios finales

Terminado este primer capítulo, se puede concluir que las tecnologías y conceptos que conforman la web semántica son muy diversos y complejos por lo menos desde el punto de vista teórico, dificultándose en muchos casos su correcto entendimiento y su aplicación. Sin embargo su comprensión se facilita enormemente cuando se contrapone los conceptos contra grupos de datos que permitan su aplicación especialmente en el contexto estadístico.

En definitiva las personas que se encuentran realizando sus primeras aproximaciones a este tipo de proyectos deben tomarse su tiempo para comprender correctamente cada uno de los temas y así evitar problemas futuros al momento de aplicarlos.

CAPITULO II
PROBLEMÁTICA

2.1. Introducción

En este capítulo se explican las principales motivaciones para el desarrollo de este proyecto, abordándose temas referentes a la problemática que se intenta resolver, la justificación para el desarrollo del proyecto, una propuesta de solución y se habla brevemente de algunos proyectos relacionados con la publicación de datos estadísticos que sirvieron como referencia para el desarrollo de esta investigación y de los cuales se sugiere su lectura.

2.2. Planteamiento del problema

En la actualidad el poder publicar datos en la web semántica cada día cobra más importancia, sin embargo esta no resulta ser una tarea sencilla en especial cuando se trabajan con datos estadísticos, en este contexto existen trabajos que ilustran la implementación puntal de uno u otro vocabulario específico. Sin embargo ninguno realiza una evaluación profunda de los vocabularios, explicando que opciones existen, las ventajas o desventajas que poseen y como se comparan con los demás

En este contexto el presente proyecto se encarga de realizar dicho análisis al mismo tiempo que afronta la multitud de problemas que surgen al momento de efectuar una publicación de datos estadísticos, considerándose aspectos desde la depuración, normalización y conciliación de los datos, hasta el modelamiento de la información, su publicación y explotación.

2.3. Justificación

Si bien actualmente se están desarrollando muchos proyectos relacionados a la web semántica tanto en la Universidad Técnica Particular de Loja como en otras instituciones, muy pocos de ellos se encuentran orientados a trabajar con datos estadísticos.

Es por tanto que nace este proyecto, como una oportunidad para impulsar el trabajo en esta área y aportar con un proyecto de fin de titulación que permita facilitar las tareas que implica la publicación de datos estadísticos y sirva de guía para aquellas personas que se encuentran realizando sus primeras aproximaciones a estos temas.

2.4. Solución propuesta

Para lograr resolver el problema planteado se pretende realizar una investigación a profundidad acerca de los temas referentes a la web semántica, considerando su origen, evolución, tecnologías, y principales conceptos prestando especial atención en los vocabularios que permiten modelar datos estadísticos.

Para luego poder efectuar una publicación de datos estadísticos que empleando información brindada por Scopus referente a revistas científicas a nivel mundial y sus diversos indicadores, muestren como resolver los diversos problemas que surgen durante cada una de las fases de la publicación de datos, al mismo tiempo que se crea una guía procedimental que permita orientar a las personas con poco conocimiento en esta área acerca de los principales aspectos y procedimientos a tener en cuenta en cada proceso.

2.5. Trabajos relacionados

Finalmente en esta sección se habla acerca de algunos trabajos relacionados con la publicación de datos estadísticos en los que se ha empleado de manera muy puntual los vocabularios RDF Data Cube, SDMX-RDF y SKOS, con la finalidad de servir de referencia para futuras investigaciones y ejemplificar el uso exitoso de esos vocabularios en un ambiente del mundo real.

2.5.1. Integrating serbian public data into the LOD cloud.

Es un proyecto de la república Serbia impulsado por su Oficina de Estadística. Dicho proyecto surge como un medio para que Serbia se una a las iniciativas implementadas por varios países de la Unión Europea (España, Reino Unido, Alemania, etc.) y de esta forma exponer, compartir y articular sus datos en la Web Semántica.

Tal como se menciona en (Janev & Milosević, 2012) actualmente “La Oficina de Estadística de la República de Serbia” se encuentra en un proceso de armonización de normas, clasificaciones y desarrollo de metodologías en cooperación con el sistema de estadísticas oficiales de la Unión Europea.

Los datos estadísticos de Serbia se representan usando el vocabulario RDF Data Cube, seleccionado no solo por ser uno de los más completos, sino también por ser compatible con la norma ISO SDMX que está siendo utilizada por la Junta de la Reserva Federal

Estadounidense, el Banco Central Europeo, Eurostat, la OMS, el FMI y el Banco Mundial. También se está empleando el vocabulario SDMX-RDF para la semántica de dominio, los metadatos del conjunto de datos y otra información crucial necesaria en el proceso de intercambio de datos estadísticos. El proceso de trasformación de los datos que realiza "La Oficina de Estadística de la República de Serbia" consta de primero tomar sus datos en formato XML, los cuales son pasados como entrada a un procesador XSLT y transformados posteriormente en RDF utilizando los vocabularios mencionados y esquemas de conceptos.

2.5.2. ICANE.

Esta es una iniciativa puesta en marcha por el Instituto Cántabro de Estadística que tiene como objetivo poner a disposición de sus usuarios la totalidad de su Banco de datos, adhiriéndose a las directrices del World Wide Web Consortium en materia de datos abiertos y enlazados, para de esta forma permitir que tanto sus datos como sus respectivos metadatos se puedan consultar mediante herramientas automatizadas.

Para lograr esto se han utilizado diferentes herramientas y vocabularios entre los cuales podemos resaltar la utilización del vocabulario SKOS[33] para la descripción de secciones y subsecciones, esto con la finalidad de que las relaciones jerárquicas entre términos sean fáciles de interpretar. En el caso de las series de datos, ellos utilizan el vocabulario RDF Data Cube y finalmente para el resto de metadatos han decidido utilizar vocabularios de uso común, tales como Dublin Core Metadata Terms (DCTerms)[34], Friend of a Friend (FOAF)[35] o Web Ontology Language (OWL)[36]. Además de un vocabulario propio, para identificar los diferentes componentes del banco de datos fácilmente.

2.5.3. Representing verifiable statistical index computations as linked data.

Es un proyecto investigativo presentado en el SemStats del 2014, donde se describe el desarrollo de un portal web que muestre datos estadísticos referentes al Web Index, que

[33] https://www.w3.org/2004/02/skos/
[34] http://dublincore.org/documents/dcmi-terms/
[35] http://xmlns.com/foaf/spec/
[36] https://www.w3.org/TR/owl-features/

es una medida de la contribución de la World Wide Web de al progreso social, económico y político en los países de todo el mundo. Este proyecto reúne información de 81 países, que ha sido reunida durante 5 años y estructurada en 116 indicadores diferentes (Emilio, Gayo, Farham, Fern, & Mar, 2014).

Para lograr esto realizaron la transformación de la información que se encontraba recopilada en archivos Excel, convirtiéndola a RDF mediante el uso un vocabulario propio llamado Computex, que se encuentra basado principalmente en el RDF Data Cube por cuanto ambos vocabularios son compatibles. Además de eso utilizaron bases de datos RDF para el almacenamiento de la información y herramientas basadas en Pubby[37] para la visualización de los datos.

2.6. Comentarios finales

Finalizado este capítulo de puede observar algunos de los proyectos estadísticos que se trabajan a nivel mundial, con lo cual se resalta de manera especial que existen diversas opciones al momento de elegir un vocabulario para representar datos estadísticos, pero cada uno de ellos es válido para un determinado escenario y puede ser empleado exitosamente siempre y cuando las necesidades del proyecto se ajusten a sus características.

[37] http://wifo5-03.informatik.uni-mannheim.de/pubby/

CAPITULO III
PREPARACION DE DATOS

3.1. Introducción

En este capítulo se comienzan a desarrollar los aspectos prácticos referentes a la preparación de los datos la cual es una de las tareas más críticas y complejas a ser realizadas, se trabajan temas puntuales referentes a la especificación de datos como son: la proveniencia de datos y la preparación de la información.

Durante este capítulo se mencionan problemas comunes que existen al trabajar con datos en distintos formatos y se muestran distintas opciones para estandarizar la información, todas con sus ventajas y desventajas propias, dependido así de las características del proyecto para optar por una de ellas.

3.2. Procedencia de datos

Para el proyecto como ya se mencionó antes se utilizan datos estadísticos referentes a la clasificación y relevancia de múltiples revistas científicas en todo el mundo. Dichos datos serán recolectados de distintos repositorios, con la finalidad de mejorar la fiabilidad de la información al mismo tiempo que se incrementa su cantidad.

El primer sitio del que se extrajo información fue Journal Metrics[38], perteneciente a la editorial Elseiver, este ofrece información referente a revistas científicas y tal como se menciona en (Metrics, n.d.) se encuentra constantemente actualizando sus datos. También se debe mencionar que Journal Metrics ofrece una categorización y evaluación del impacto que poseen cada una de las revistas científicas de las que posee información.

Posteriormente esta información fue complementada con los datos brindados por SCImago Journal & Country Rank[39], el cual como explica en su descripción oficial (Scimago Journal & Country Rank, n.d.) es un portal web que brinda información de revistas científicas y los indicadores de los países, estos datos se encuentran desarrollados a partir de la información contenida en la base de datos Scopus.

Finalmente la información se completó con datos extraídos directamente del sitio web oficial de Scopus[40], la mayor base de datos de resúmenes y citas de la literatura por pares de indexación de revistas científicas según como se menciona en (“Scopus,” n.d.).

[38] http://www.journalmetrics.com/
[39] http://www.scimagojr.com/
[40] http://www.scopus.com/

No obstante a pesar de la utilización de múltiples repositorios de información para la extracción de dados, en algunos casos la información obtenida de los activos descargados se encontraba incompleta, por ejemplo en ningún repositorio se encontraban datos referente a los cuartiles considerando sus valoración por categoría y año, por tanto fue necesario la realización de Scrapy sobre los mencionados sitios web para poder obtener todos los datos necesarios.

En lo referente a los términos de uso de información y licencias intelectuales de los datos, debemos destacar que debido a que la información recolectada se encuentran de alguna manera relacionada con la base de datos de Scopus, se respetaran los términos y condiciones especificados por Elseiver, quien es el dueño del dominio de la información, y que menciona los límites de uso de los mismos en su página oficial (“Privacy Policy Elseiver,” n.d.).

Elseiver permiten hacer uso de todos sus datos siempre se sea para uso personal, no comercial, informativo o académico y siempre manteniendo intactos todos los derechos de autor y otros avisos de propiedad. Por cuando en este caso al ser utilizados en un proyecto de fin de titulación no existiría ningún problema o complicación legal.

Además la información y datos generados en el presente proyecto, serán publicados de acuerdo a las políticas de proyectos de fin de titulación establecidos por la Universidad Técnica Particular.

3.2.1. Grafo de datos.

Como ya se mencionó los datos a ser utilizados en este proyecto provienen de las páginas Scopus, SCImago Journal & Country Rank y Journal Metrics, complementadas con la información extraída mediante Scrapy (esta se compone de dos documentos uno de categorías y otro de información de revistas). La información obtenida fue posteriormente filtrada de tal forma que solo se consideró los datos relevantes para el proyecto, obteniendo así que la información seleccionada de cada origen se muestra en la Figura № 18.

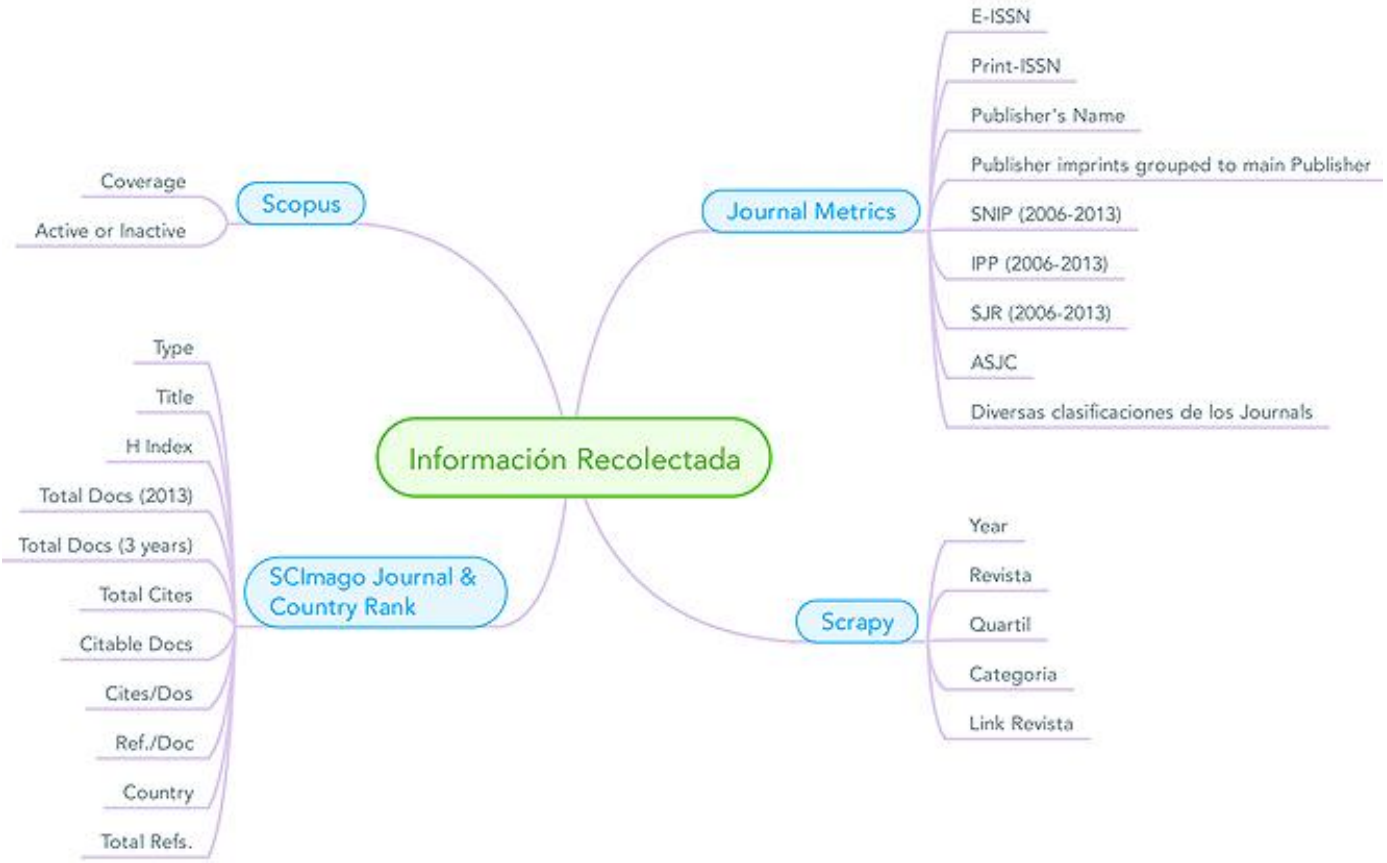


Figura 18 Estructura y Fuentes de Extracción de los Datos.
Fuente: El Autor
Elaborado por: El Autor

3.3. Depuración y preparación de datos

Una vez establecidos los distintos orígenes de donde se extraerá la información y definido las licencias para los datos que se utilizará durante el desarrollo del proyecto, el siguiente paso fue llevar acabo la depuración y preparación de los mismos. Esto fue necesario debido a que los datos poseían ciertas inconsistencias y se encontraban en formatos diferentes, los cuales al momento de ser consumidos entorpecían el proceso, siendo un impedimento para continuar el proceso.

La depuración de los datos comprendió tanto el filtrado de la información útil, como la unificación de la misma dentro de un solo repositorio, para la unificación se consideró el emparejamiento por campos con valores unívocos y los cuales se encontraban presentes en cada uno de los distintos orígenes de datos, en esta tarea se utilizó la herramienta desarrollada por Google llamada “Open Refine” por ser de código libre y una de las más poderosas para el tratamiento de datos.

Debido a la gran cantidad de información con la que se debía trabajar, fue necesario ampliar la cantidad de memora RAM configurado por defecto para esta herramienta, incrementándose su valor original de 1 Gb hasta ampliarlo a 6 Gb.

Antes de realizar este cambio en la configuración se debe considerar el sistema operativo que se está utilizando y la capacidad de memoria RAM disponible en el equipo, puesto que si ingresamos un valor demasiado alto podemos afectar de manera negativa el funcionamiento del computador. Los pasos necesarios para realizar la ampliación de la memora RAM se detallan en el Anexo 2.

El proceso de depuración se realizó en tres diferentes etapas más un pre procesamiento y una etapa de normalización que resulta opcional. Durante estas etapas en algunos casos fue necesario ejecutar más de un script desarrollado en GREL[41], estas etapas se describen a continuación y los scripts utilizados pueden ser encontrados en el repositorio de Git de la Universidad Técnica Particular de Loja o mediante siguiente URL: https://git.taw.utpl.edu.ec/vmjaramillo1/PublicacionEstadisticaEnLinkedData

3.3.1. Pre procesamiento.

En esta etapa de la depuración se realizó la primera ronda de eliminación de datos duplicados e información no relevante para el proyecto (registros de datos anteriores al año 2006), además se colocaron los formatos adecuados y estandarizaron los nombres de las columnas de los archivos, con la finalidad de que estos pudieran ser procesados de mejor manera en las etapas siguientes. En el pre procesamiento se emplearon un total de cinco scripts, cada uno con un objetivo diferente.

Lo primero a realizarse fue la depuración de los datos obtenidos mediante scrapy y que se encontraban en dos archivos separados, uno referente a las revistas y otro que poseía la información de las categorías. Luego de realizadas algunas pruebas se determinó que era necesario realizar primero la conversión del archivo de revistas que se encontraba en formato CVS y pasarlo a Excel, para esta tarea se empleó la herramienta “Libre Office” de Linux por ser la que mejor resultado obtuvo.

Posterior a la conversión se procedió a la carga del archivo convertido a Excel en la herramienta Open Refine para de esta forma ejecutar el Script número 1 y así eliminar los errores. Sin embargo se debe destacar que la depuración final en este archivo es

[41] https://github.com/OpenRefine/OpenRefine/wiki/GREL-Functions

una tarea que se debe realizar manualmente, puesto que la naturaleza de los errores hace que resulte sumamente difícil la automatización de esta sección del proceso.

Referente al archivo de categorías, solo es necesario procesarlo directamente en la herramienta Open Refine en su formato nativo (CSV) y depurado con la ejecución del script dos. (Para más información de los scripts revisar el Anexo 3).

En esta etapa también se procedió a realizar la unión de los documentos descargados desde la página web SCImago Journal & Country Rank y que contenían información estadística de dos años diferentes, esto se efectuó con la finalidad conciliar datos he ir unificando la información en un solo archivo, para esta tarea se procesó de manera conjunta dichos archivos utilizando la herramienta Open Refine y ejecutando el script tres. (Para más información de los scripts revisar el Anexo 3)

Finalmente se ejecutaron los scripts cuatro y cinco para preparar los datos obtenidos desde Jounral Metrics y Scopus correspondientemente. Un gráfico que resume esta etapa se muestra en la Figura № 19.

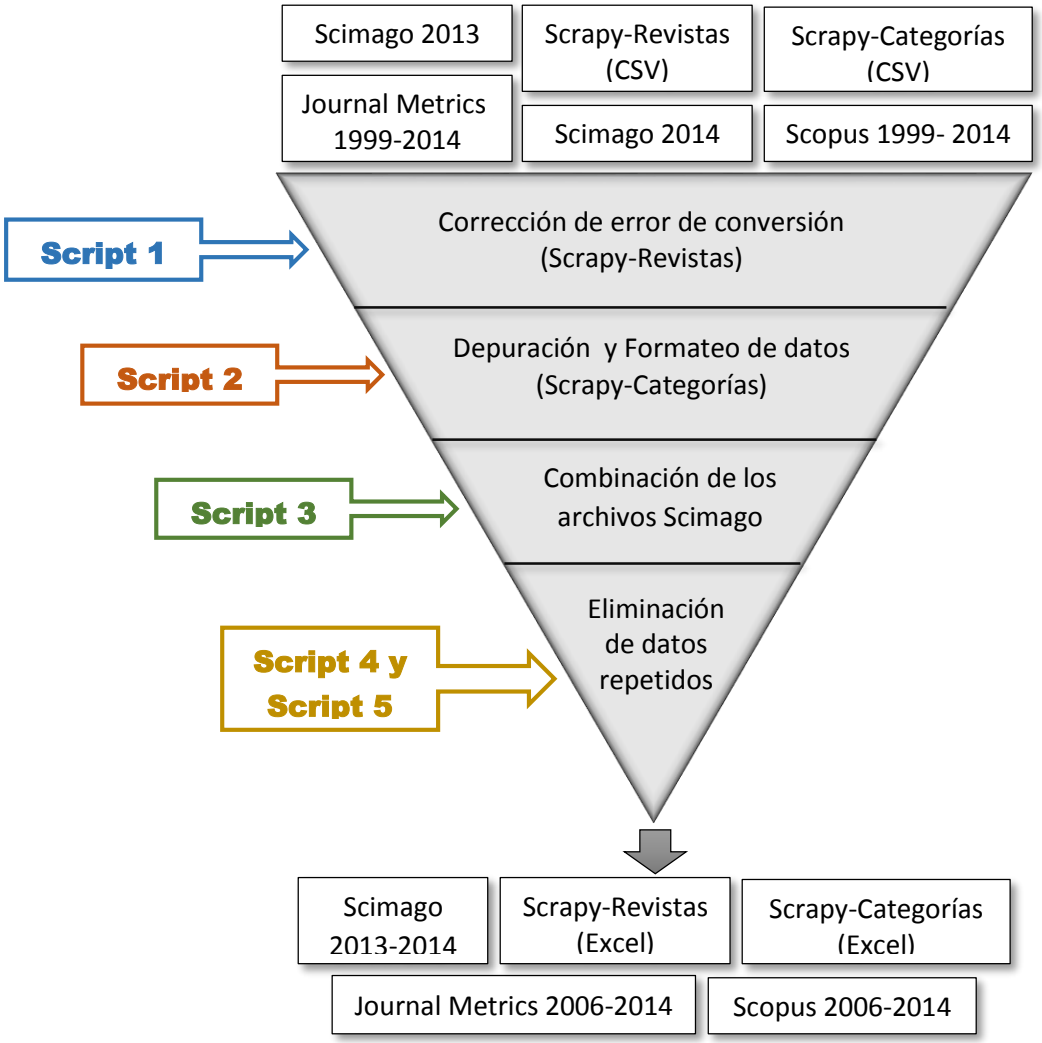


Figura 19 Actividades del Pre procesamiento
Fuente: El Autor
Elaborado por: El Autor

3.3.2. Etapa № 1.

Se realiza la unión de los datos obtenidos mediante scrapy y que se encontraban en dos diferentes archivos (revistas y categorías), conformándose así un solo documento que llamaremos “rev-cat” y que contendrá toda la información, esta unión se llevó acabo en la herramienta Open Refine mediante la ejecución del script seis. (Para más información de los scripts revisar el Anexo 3)

3.3.3. Etapa № 2.

Durante esta etapa se efectúan dos unificaciones con los archivos restantes, lo primero en realizarse es la combinación de los archivos procedentes de Journal Metrics y Scopus, para esto se empleó el script número siete que utiliza principalmente la columna “Sourcerecord id” que comparten ambos documentos, como campo principal para unificar los datos.

En la segunda parte de esta etapa se combinan los datos obtenidos mediante scrapy y que actualmente se encontraban en el documento llamado “rev-cat” procedente de la etapa № 1 junto con obtenidos en el pre procesamiento, referente a la información obtenida de SCImago Journal & Country Rank. En este caso la información se unifico considerando coincidencias en los títulos de las revistas y en los identificaciones ISSN. Se utilizó el script número ocho y el archivo resultante fue llamado “rev-cat-mas-scimago”.

3.3.4. Etapa № 3.

En esta etapa se realizó la unión y depuración final de los datos hasta ahora trabajados y que en este momento se encontraban en dos archivos diferentes, uno referente a la información de scrapy y la descargada desde de SCImago Journal & Country Rank (este documento se denomina "rev-cat-mas-scimago") y otro referente a la información obtenida de Journal Metrics más Scopus (este documento se denomina "jMetrics-scopus").

Debido a la cantidad de información se efectuó dos diferentes fases de unificación, con la finalidad de obtener el mayor número de coincidencias posibles y reducir la cantidad de errores, una primera fase se realizó por coincidencia de títulos utilizando el script nueve y otra por coincidencia de ISSN ejecutando el script número diez. Cabe destacar que se utilizaron estos atributos debido a que luego de efectuarse varias pruebas fueron los que mejores resultados obtuvieron, con el mayor número de aciertos y menor cantidad de errores, esto en gran parte a que la proveniencia de los datos corresponden a un mismo dominio. Finalmente los dos resultados se unieron y formatearon utilizando el script once que dio como resultado el archivo final. (Para más información de los scripts revisar el Anexo 3)

El resultado obtenido de esta etapa es un documento Excel con la estructura que se muestra en la Tabla 3.

Tabla 3 Estructura Dase de Datos Depurados.

Title	Source Type	Print-ISSN	E-ISSN	link_revista	Country	Categoría	2006			
							SNIP	IPP	SJR	quartil
Molecular Systems Biology	Journal	17444292	NULL	http://www.scimagojr.com/journalsearch.php?q=4700152228&tip=sid&clean=0	United Kingdom	Agricultural and Biological Sciences (miscellaneous)	NULL	NULL	2,283	1
Proceedings of the Royal Society B: Biological Sciences	Journal	9628452	NULL	http://www.scimagojr.com/journalsearch.php?q=130030&tip=sid&clean=0	United Kingdom	Agricultural and Biological Sciences (miscellaneous)	1,712	4,121	2,869	1

Fuente: El Autor
Elaborado por: El Autor

3.3.5. Normalización.

Esta es una etapa opcional por cuanto su ejecución no es obligatoria, lo que se realiza en esta sección es la normalización de los datos para estandarizarlos de un formato de orientado a columnas como es el que se obtiene al final de la etapa № 3 a uno orientado a filas. Ambos formatos son válidos y se puede trabajar con cualquiera de ellos, sin embargo, cada uno presenta sus propias ventajas y desventajas, por cuanto en este caso se realizó la normalización de los datos para obtener una idea más clara de su estructura y los retos que este conlleva al momento de ser efectuado.

Para la normalización se empleó de entrada el archivo resultante de la Etapa № 3, el mismo que fue sometido a la ejecución del script doce, que se encarga de efectuar la simplificación del formato y realizar las transiciones de la información que se encontraba originalmente en columnas para luego pasarlas a filas. Un extracto del documento resultante se puede apreciar en la Tabla 4:

Tabla 4 Modelo Normalizado de Datos

Title	Type	ISSN	E-ISSN	link_revista	Pais	Year	Categoria	SNIP	IPP	SJR	Quartil
21st Century Music	Journals	15343219	NULL	http://www.scimagojr.com/journalsearch.php?q=18500162600&tip=sid&clean=0	United States	2006	Music	0	0	0.101	3
21st Century Music	Journals	15343219	NULL	http://www.scimagojr.com/journalsearch.php?q=18500162600&tip=sid&clean=0	United States	2007	Music	0	0	0.101	3
21st Century Music	Journals	15343219	NULL	http://www.scimagojr.com/journalsearch.php?q=18500162600&tip=sid&clean=0	United States	2008	Music	0	0	0.104	3

Fuente: El Autor
Elaborado por: El Autor

Al finalizar con la depuración podemos apreciar como este proceso actualmente resulta sumamente tedioso y extenso de realizar, aun cuando la mayor parte del trabajo se encuentra encapsulada en scripts, sin embargo en un futuro haciendo uso de los scripts ya mencionados y en cohesión con el API proporcionado por “Open Refine” se podría realizar un programa computacional que automatizara dicho proceso. Sin embargo esta tarea supondría un extenso tiempo de análisis y un considerable trabajo de desarrollo, por cuanto esta actividad no forma parte del desarrollo de este proyecto. La dificultad de desarrollar dicho programa radica mayormente en a la necesidad de elaborar algoritmos que reconozcan los patrones en los cuales los datos son publicados, además se tendrían que elaborar mecanismo que permitan la identificación de errores que suceden al momento de la trasformación de datos o durante la conciliación de la información, por cuanto actualmente estas son tareas que requieren de la supervisión humana.

3.4. Comentarios finales

Luego de terminadas las actividades de este capítulo se puede concluir que la actividad de conciliar múltiples datos de distintos orígenes resulta ser una de las complicadas durante el proceso de publicación de datos, además tal como se menciona en (Cadme & Piedra, 2014) la tarea de conocer y documentar la procedencia de los datos debe ser tomada muy seriamente pues es un aspecto crucial que además de facilitar en algunos casos el análisis, puede ayudar a entender cual fue el propósito original para la recolección de esa información y en que nuevas formas se puede utilizar en el proyecto en curso.

Sin embargo el aspecto más sorprendente de este capítulo resulta ser es la gran importancia que tiene la organización final de los datos, puesto que elegir entre la opción basada en columnas u optar por un esquema normalizado basado en filas, puede influir de manera drástica en el desarrollo del proyecto, siendo esta una decisión crucial y que no debe ser tomada a la ligera.

CAPITULO IV
DESARROLLO DEL CICLO DE VIDA DE LINKED DATA

4.1. Introducción

En este capítulo se parte de los datos ya depurados anteriormente para poder explicar los temas referentes al desarrollo del ciclo de vida de Linked Data, trabajándose puntualmente las fases de especificación donde se diseñaran los URIs para los recursos, el modelado de los datos en el cual se explican los aspectos más importantes a considerarse y la generación de las tripletas que muestra cómo realizar la trasformación de los datos a RDF mediante la utilización de una aplicación desarrollada en el lenguaje de programación Java.

4.2. Primera etapa, especificación

Siguiendo el ciclo de vida propuesto en la sección “1.9 Ciclo de Vida” del capítulo I, durante esta primera etapa se realizan las tareas de la especificación de los datos que quedaron pendientes en el capítulo anterior, siendo así que se efectua el diseño de URIs que serán utilizadas para identificar a los recursos durante todo el proyecto.

4.2.1. Diseño de URIs.

En este caso para la realización del diseño de URIs se ha preferido utilizar el concepto de URIs Cool[42] por cuanto partiendo de un URI base se han determinado dos conjuntos de URIs específicas, una para el vocabulario y otra para los recursos. La estructura de estos dos conjuntos se muestra en la Figura № 20 y la Figura № 21.

Vocabulario

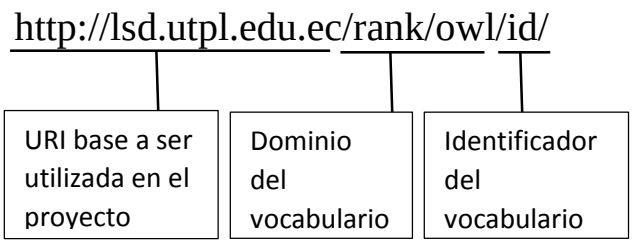


Figura 20 Estructura del URI del Vocabulario.
Fuente: El Autor
Elaborado por: El Autor

[42] http://www.w3.org/TR/cooluris/

Recursos

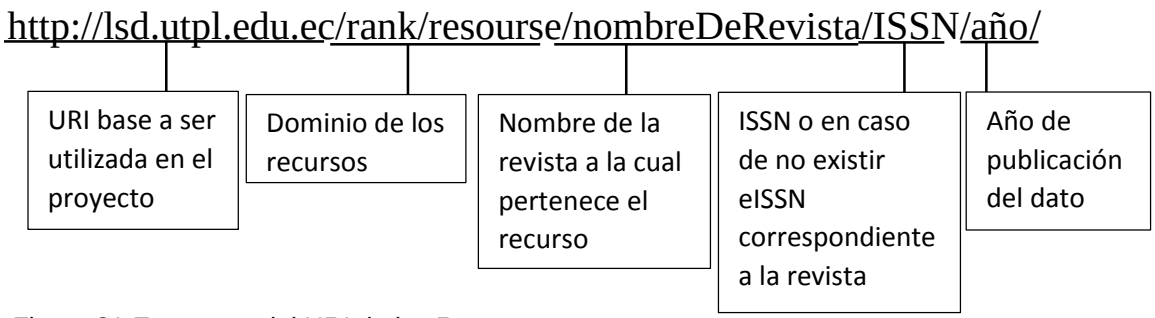


Figura 21 Estructura del URI de los Recursos.
Fuente: El Autor
Elaborado por: El Autor

4.3. Segunda etapa, modelado

Una vez terminada la especificación de los datos y el diseño de las URIs la siguiente etapa consta en realizar el modelado de esta información, para esto se efectúa un análisis de que vocabularios utilizar para representar la información estadística, se aplica una metodología para poder modelar la información al mismo tiempo que se realiza el diseño y validación de un vocabulario que permita representar las entidades y relaciones propias del proyecto, que actualmente no podrían ser modeladas de otra forma.

4.3.1. ¿Qué vocabulario utilizar?

Como se ha mencionado en la sección “1.7 Origen de los Vocabularios Estadísticos” existen algunas propuestas de vocabularios orientados al índole estadístico, pero nosotros nos vamos a centrar en los tres más recientes, debido a ser los más utilizadas y que mejores características poseen.

SCOVO.- Este vocabulario resulta ser el menos recomendado actualmente, salvo muy puntuales excepciones. Debido a sus falencias al momento de describir su contenido, solo se recomienda su utilización en casos donde la información que se está modelando sea sumamente sencilla, de lo contrario existirían muchos problemas al momento de recuperar determinados datos. Cabe destacar que actualmente varias organizaciones, entre ellas la W3C consideran a este vocabulario como obsoleto y en su lugar recomiendan la utilización de alguna de las opciones que se detallan a continuación.

SDMX-RDF.- Al estar vasado en los estándares de la “Statistical Data and Metadata Exchange” posiblemente este sea uno de los vocabularios más completos cuando se habla de modelado de datos y de metadatos estadísticos. Consta de múltiples definiciones que permiten modelar todo tipo de información estadística, prestando especial importancia en la estructura interna de los datos. Sin embargo paradójicamente su complejidad que lo convierte en un vocabulario sumamente robusto también puede llegar a ser un problema, en muchos de casos las clases y relaciones que ofrece resultan no ser utilizadas, dificultando la tarea de modelado para personas con poca experiencia o que han elegido erróneamente este vocabulario para representar datos demasiado sencillos. Su utilización se sugiere mayormente en conjuntos de datos sumamente grandes o para personas que ya poseen experiencia en el uso de este vocabulario.

RDF Data Cube.- Considerado por muchos como la mejor opción al momento de realizar el modelado de datos estadísticos, podemos decir que este vocabulario reúne las mejores partes de sus antecesores y las combina de manera que se crea un balance ideal. Siendo sumamente robusto pero sin llegar a los excesos, permite el modelado de cualquier información estadística sin importar lo compleja o sencilla que esta sea. Es sumamente adaptable y cabe destacar que actualmente es el vocabulario recomendado por la W3C.

Al considerar las características de los vocabularios anteriormente mencionados podemos decir que para este proyecto en particular el RDF Data Cube es el vocabulario que mejor se adapta con las necesidades del proyecto, al considerar que la información a ser modelada se refiere a datos estadísticos de revistas científicas cuya complejidad es intermedia resulta difícil pensar en otras opciones, que aun cuando su utilización puede ser posible, no resulta recomendable puesto que solo ocasionaría mayor dificultad al momento de realizar el modelado provocando mucho más trabajo sin que se obtuvieran mejores resultados.

4.3.2. Metodología.

Para la etapa de modelado se ha decidido utilizar la metodología NeOn, dicha metodología tal como se menciona en (Ontology Engineering Group, n.d.) se encuentra basada en los aspectos de colaboración de desarrollo de ontologías y su reutilización, características fundamentales a ser aplicadas en este proyecto.

El principal motivo por el cual se decidió utilizar esta metodología tal como se describe en (Ontology Engineering Group, n.d.) es que cuenta con un conjunto de nueve

escenarios que permiten la construcción de ontologías, la reingeniería y la fusión de las mismas, abarcando perfectamente el trabajo que se pretende desarrollar, ofreciéndonos así un conjunto de directrices metodológicas para diferentes procesos y actividades que permiten el desarrollo de la ontología. Los principales pasos para poder desarrollar una ontología según NeOn son:

1. Especificación de la Ontología.
2. Reutilización y Reingeniería de recursos No-Ontológicos.
3. Reutilización de recursos Ontológicos.

Dichos pasos se encuentran comprendidos en dos diferentes ciclos de vida que ofrece NeOn, y brinda la posibilidad de elegir entre el Modelo en cascada y el Modelo iterativo-incremental. Para el presente proyecto se utilizó el segundo debido a que es el que mejor se adapta a las necesidades del proyecto, dicho ciclo de vida se muestra en la Figura № 22.

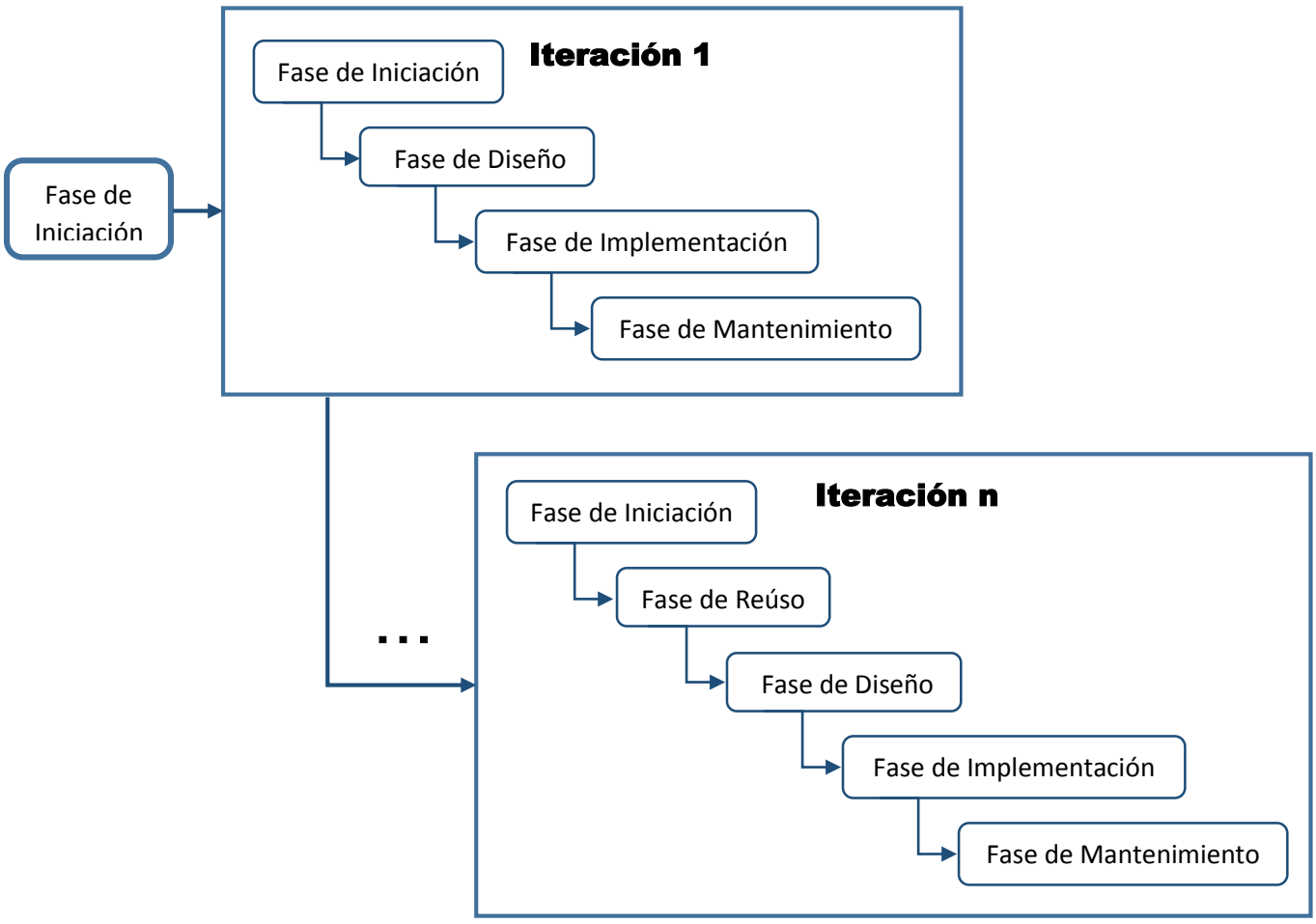


Figura 22 Ciclo de Vida de Metodología NeOn.
Fuente: Recuperado de: http://goo.gl/TAOCsN
Elaborado por: (Gómez-pérez & Suárez-figueroa, 2005)

4.3.2.1. *Especificación de la ontología*

En este paso se especifican los requerimientos y principales objetivos de la ontología, como resultado a continuación se muestran el propósito y el alcance que se definieron, para mayor información se recomienda revisar el Anexo 4.

1) Propósito

Poder representar el estado histórico y actual de las diferentes revistas científicas, permitiendo así que el público en general y personal académico (docentes, investigadores y estudiantes) tengan un referente al momento de seleccionar las revistas en las cuales publicar sus descubrimientos y realizar sus consultas e investigaciones

2) Alcance

La presente ontología se limitará a modelar revistas científicas considerando su tópico, publicita e indicadores clasificados por año.

4.3.2.2. *Reutilización y reingeniería de recursos No-Ontológicos.*

En este paso se procede a la búsqueda y reutilización de recursos no ontológicos, en nuestro caso se utilizó la clasificación de las revistas científicas ofrecidas por SCImago Journal Metrics para la creación de una taxonomía que se pudiera utilizar en la ontología, obteniéndose así como resultado la Tabla que se encuentra en Anexo 4, a continuación en la Figura № 23 se muestra una pequeña representación gráfica de la clasificación taxonómica de las revistan científicas y las categorías en las cuales pueden ser clasificadas.

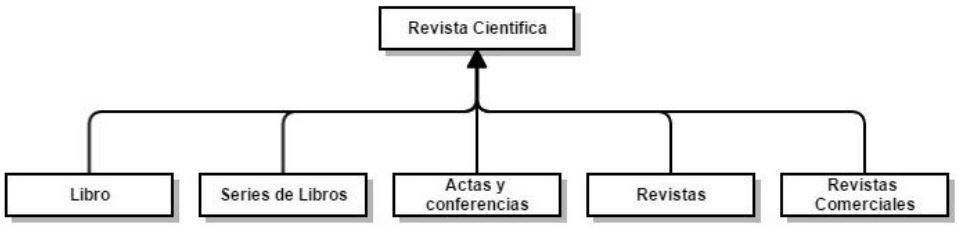


Figura 23 Clasificación taxonómica de las revistan científicas.
Fuente: El Autor
Elaborado por: El Autor

En este caso por tratarse de una taxonomía la parte de reingeniería se omitió, debido a que esta ya se encuentra comprendida dentro de la ontología final que se ha desarrollado.

4.3.2.3. *Reutilización de recursos ontológicos.*

Llegados al último paso se debe reflexionar que recursos ontológicos pueden ser reutilizados con la finalidad de poder reutilizar lo mayor posible y no crear nada que no sea necesario. Para lograr esto lo primero que se realizo fue un modelado a alto nivel, donde se consideró las especificaciones de primer paso y además se identificó la estructura fundamental de los datos, lo que permitió dar un primer aproximamiento a las distintas entidades y relaciones que debían ser modeladas, expresándose gráficamente el modelado puede ser observado en la Figura № 24:

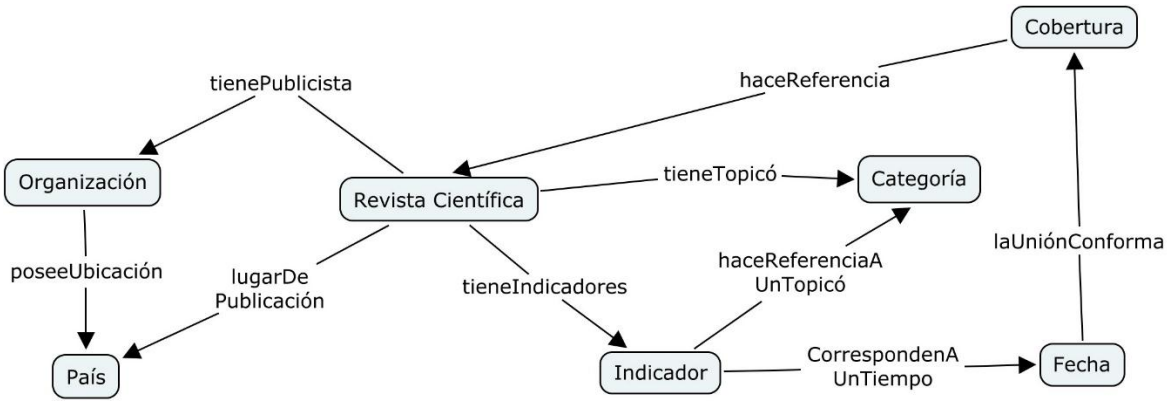


Figura 24 Modelado Alto Nivel de Ontología de Revistas Científicas.
Fuente: El Autor
Elaborado por: El Autor

Lo siguiente fue efectuar la búsqueda e identificación de vocabularios auxiliares utilizando como herramienta de búsqueda principal al sitio web “Linked Open Vocabularies”, recordando que los vocabularios seleccionados debían trabajar conjuntamente con el Data Cube para representar cada una de las entidades modeladas anteriormente. Los vocabularios seleccionados se explican a continuación:

The Bibliographic Ontology: También conocido con el prefijo “Bibo” este es uno de los principales vocabularios usados actualmente para describir recursos bibliográficos, por cuanto se lo utilizo para describir las principales características de la entidad revista científica, incluyendo aspectos como el ISSN, eISSN, entre otros.

DCMI Metadata Terms: También conocido con el prefijo “Dcterms” este es uno de los vocabularios más utilizados en web semántica debido a que permite la inclusión de metadatos dentro de la información, por cuanto su utilización resulta fundamental para la correcta estructuración de los datos y registros que se pretende representar.

OWL-Time: Cuando se necesita modelar conceptos referentes a fechas en un tiempo específico uno de los principales vocabularios a ser utilizados es el owl-Time, en este caso su uso permitió modelar los conceptos referentes a fechas de publicación y especificar las fechas correspondientes a los datos.

Core organization ontology: Es un vocabulario normalmente empleado para modelar estructuras organizacionales y que en nuestro caso permitió representar la organización o entidad encargada de realizar la publicación de las revistas científicas

SKOS: Es un vocabulario utilizado para modelar estructuras de conocimiento ya sean tesauros, conceptos, jerarquías, entre otros. En el actual proyecto fue empleado para poder representar diferentes conceptos relacionados a los indicadores de las revistas científicas.

4.3.3 Diseño del vocabulario.

Una vez definidos los recursos ontológicos a ser utilizados en el nuevo vocabulario la siguiente tarea que se realizo fue su elaboración, en esta ocasión se utilizó la herramienta “Protégé” desarrollada por la Universidad de Stanford, que además de adaptarse perfectamente a las necesidades del proyecto, aportaba características únicas que facilitarían algunas de las siguientes tareas de evaluación y corrección.

Se puede agrupar el trabajo realizado en “Protégé” puntualizándolo de la siguiente manera:

Importación de recursos ontológicos.- Lo primero a realizarse fue la importación de los distintos vocabularios que serían reutilizados dentro de la nueva ontología, todo esto con la finalidad de evitar duplicidad al momento de realizar las definiciones de las distintas clases y propiedades. En este momento también se realiza la inclusión de metadata dentro del vocabulario, asignándose nombres alternativos, anotaciones, comentarios, se especifican derechos de autoría, entre otros. Estas actividades se pueden apreciar en la Figura № 25.

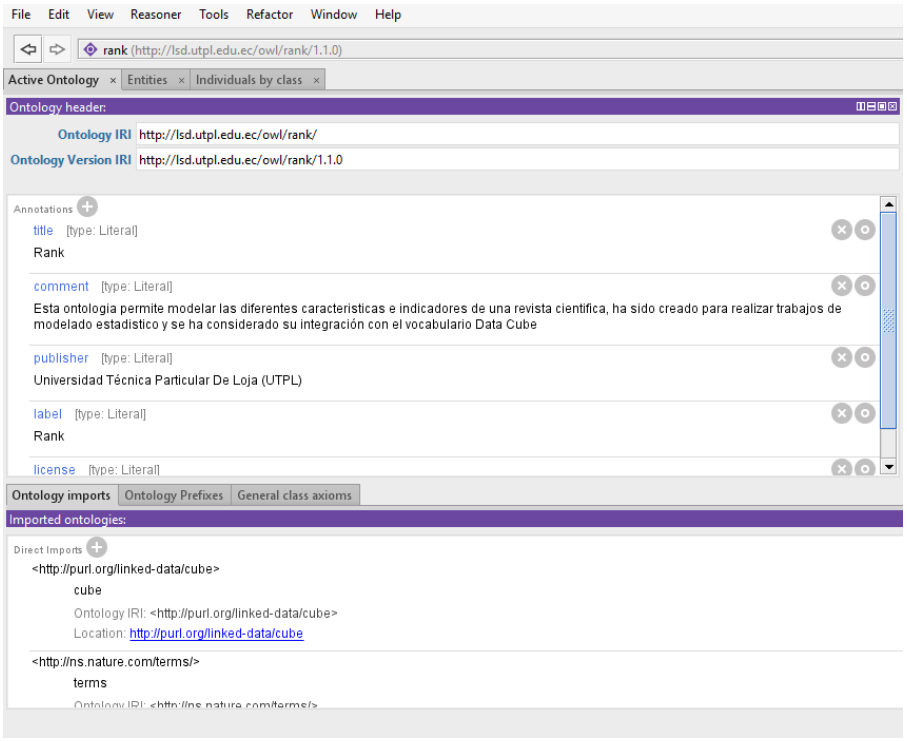


Figura 25 Importación de recursos ontológicos y metadata.
Fuente: El Autor
Elaborado por: El Autor

Definición de nuevos recursos ontológicos.- Se crean las entidades que conformarán el nuevo vocabulario definiéndose: clases, propiedad de objeto y propiedades de datos. También se especifican los nombres, comentarios y relaciones existentes entre las distintas entidades tal como se muestra en la Figura № 26.

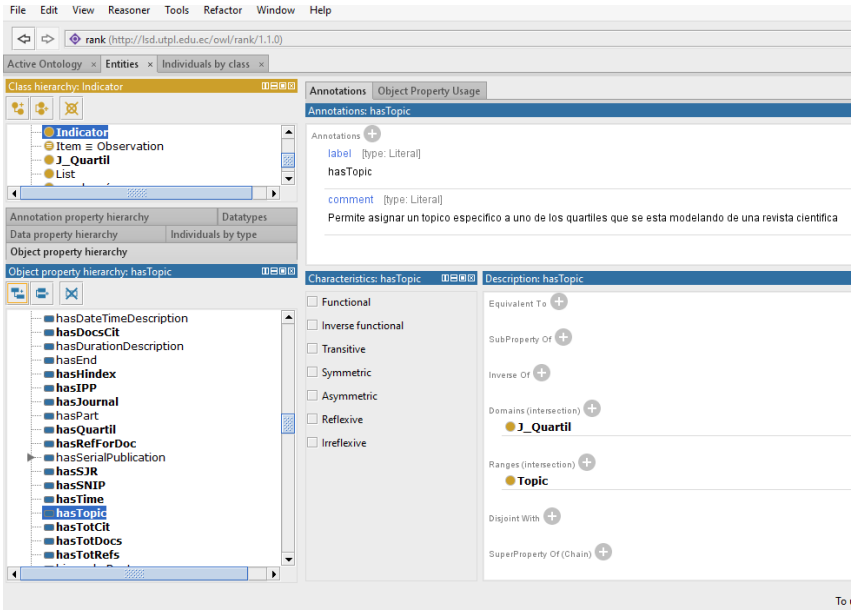


Figura 26 Definición de nuevos recursos ontológicos.
Fuente: El Autor
Elaborado por: El Autor

Pruebas del vocabulario.- Finalmente se realizan las primeras pruebas del vocabulario ejecutándose una instanciación de las principales entidades dentro de la herramienta Protégé, obteniéndose un pequeño banco de datos para ser consultado. Desde el punto de vista gráfico los datos tendrían el aspecto mostrado en la Figura № 27.

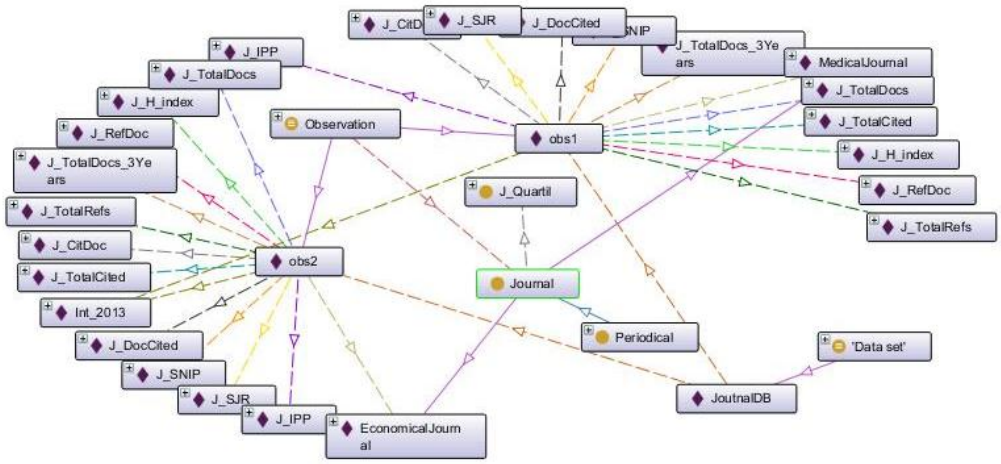


Figura 27 Instanciación de entidades.
Fuente: El Autor
Elaborado por: El Autor

Con los datos ya preparados lo siguiente a realizarse fue una serie de consultas utilizando el SPARQL Query proporcionado por la herramienta con la finalidad de verificar valores, clases y relaciones. Los datos para la realización de estas pruebas fueron directamente tomados de los documentos Excel e ingresados manualmente en Protégé esto a causa de que los archivos en formato RDF aún no se encontraban disponibles, además debido a que estas eran pruebas para verificar la estructura y funcionamiento del vocabulario estos conjuntos de datos eran suficiente. Algunos fragmentos de los resultados de estas consultas se muestran en las Figuras № 28 y № 29:

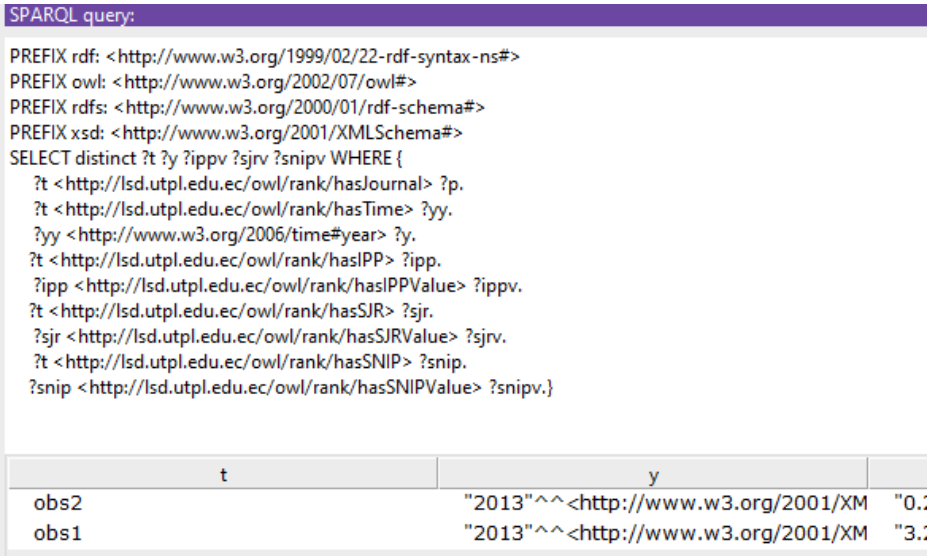


Figura 28 Consulta de propiedades en vocabulario Rank.
Fuente: El Autor
Elaborado por: El Autor

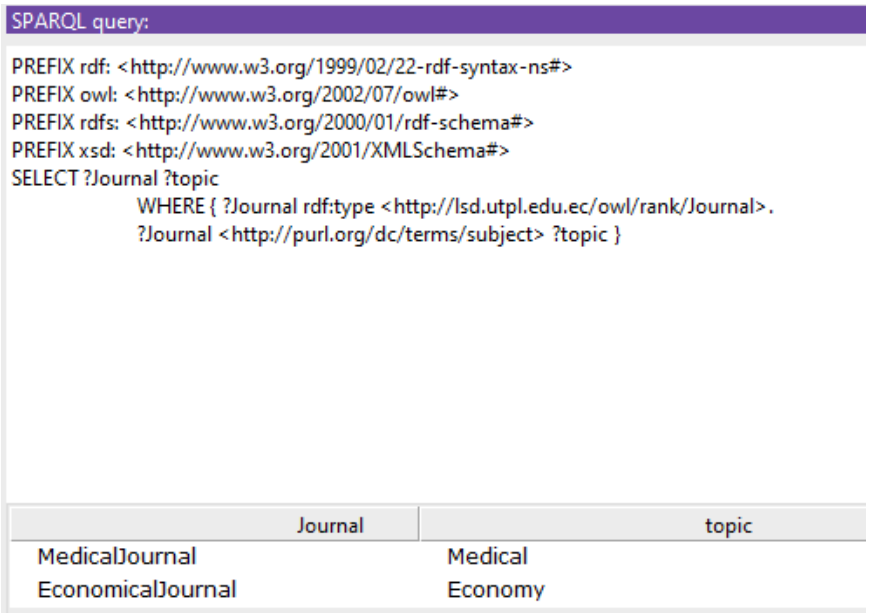


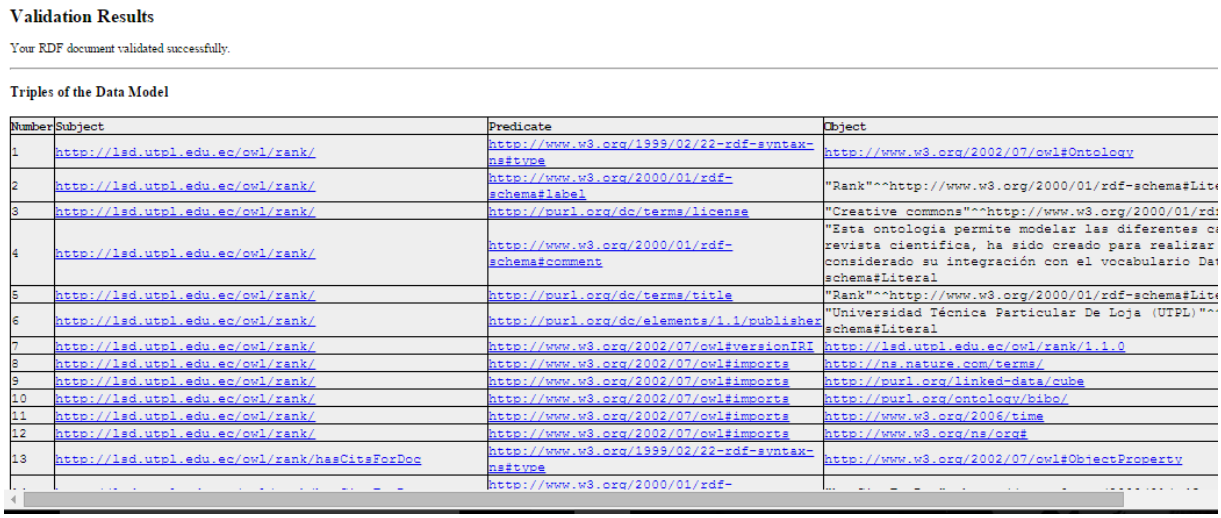
Figura 29 Consulta de Journal y Topic en vocabulario Rank.
Fuente: El Autor
Elaborado por: El Autor

4.3.4 Validación del vocabulario.

Llegados a este punto resulta primordial poder validar el vocabulario desarrollado con la finalidad de descubrir posibles errores cometidos en las fases anteriores y poder efectuar las correcciones pertinentes en caso de ser necesario, las tareas de validación

fueron efectuadas mediante la utilización de dos herramientas distintas llamadas "Validatos Service[43]" y "OOPS! (OntOlogy Pitfall Scanner!)[44]".

La primera es parte de un servicio de validación ofrecido por la W3C dentro de su sitio oficial en el apartado RDF y permite verificar la relación tanto en tripletas como en el grafo obtenido de sus relaciones. En esta validación se descubrieron problemas de redacción en algunas descripciones y nombres cambiados los cuales fueron adecuadamente corregidos. A continuación en la Figura № 30 se muestra una imagen obtenida al momento de ejecutar esta validación.



Validation Results

Your RDF document validated successfully.

Triples of the Data Model

Number	Subject	Predicate	Object
1	http://lsd.utpl.edu.ec/owl/rank/	http://www.w3.org/1999/02/22-rdf-syntax-ns#type	http://www.w3.org/2002/07/owl#Ontology
2	http://lsd.utpl.edu.ec/owl/rank/	http://www.w3.org/2000/01/rdf-schema#label	"Rank"^^http://www.w3.org/2000/01/rdf-schema#Lit
3	http://lsd.utpl.edu.ec/owl/rank/	http://purl.org/dc/terms/license	"Creative commons"^^http://www.w3.org/2000/01/rd
4	http://lsd.utpl.edu.ec/owl/rank/	http://www.w3.org/2000/01/rdf-schema#comment	"Esta ontologia permite modelar las diferentes c revista cientifica, ha sido creado para realizar considerado su integración con el vocabulario Da schema#Literal
5	http://lsd.utpl.edu.ec/owl/rank/	http://purl.org/dc/terms/title	"Rank"^^http://www.w3.org/2000/01/rdf-schema#Lit
6	http://lsd.utpl.edu.ec/owl/rank/	http://purl.org/dc/elements/1.1/publisher	"Universidad Técnica Particular De Loja (UTPL)"^ schema#Literal
7	http://lsd.utpl.edu.ec/owl/rank/	http://www.w3.org/2002/07/owl#versionIRI	http://lsd.utpl.edu.ec/owl/rank/1.1.0
8	http://lsd.utpl.edu.ec/owl/rank/	http://www.w3.org/2002/07/owl#imports	http://ns.nature.com/terms/
9	http://lsd.utpl.edu.ec/owl/rank/	http://www.w3.org/2002/07/owl#imports	http://purl.org/linked-data/cube
10	http://lsd.utpl.edu.ec/owl/rank/	http://www.w3.org/2002/07/owl#imports	http://purl.org/ontology/bibo/
11	http://lsd.utpl.edu.ec/owl/rank/	http://www.w3.org/2002/07/owl#imports	http://www.w3.org/2006/time
12	http://lsd.utpl.edu.ec/owl/rank/	http://www.w3.org/2002/07/owl#imports	http://www.w3.org/ns/org#
13	http://lsd.utpl.edu.ec/owl/rank/hasCitsForDoc	http://www.w3.org/1999/02/22-rdf-syntax-ns#type	http://www.w3.org/2002/07/owl#ObjectProperty
		http://www.w3.org/2000/01/rdf-	

Figura 30 Validación de Vocabulario Rank en "Validator Service".
Fuente: El Autor
Elaborado por: El Autor

La segunda es una herramienta desarrollada por el "Ontology Engineering Group" que tiene como finalidad ayudar a detectar inconciencias existentes dentro de las ontologías. Al utilizar esta herramienta se pudo apreciar la existencia de ciertas imprecisiones al momento de realizar la inclusión de la metadata dentro del vocabulario, dichas inconsistencias fueron inmediatamente corregidas, sin embargo una característica a resaltar es que ciertas advertencias permanecieron debido a que la herramienta incluye los errores encontrados en los vocabularios reutilizados y por tanto realizar dichas correcciones resulta imposible puesto que aquellos recursos se encuentran fuera de nuestro control. Una imagen de la validación en esta herramienta se muestra en la Figura № 31.

[43] https://www.w3.org/RDF/Validator/
[44] http://oops.linkeddata.es/

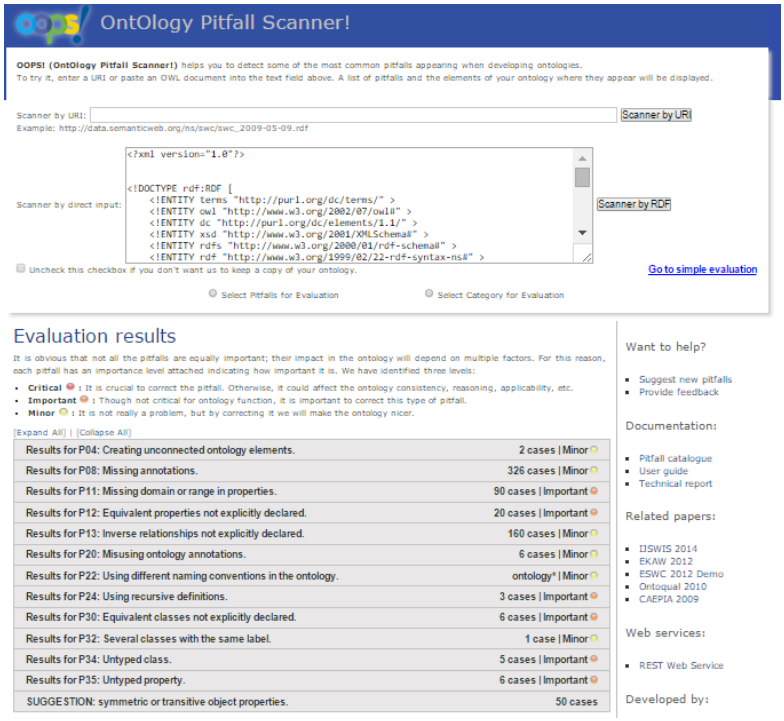


Figura 31 Validación de Vocabulario Rank en " OOPS! (OntOlogy Pitfall Scanner!)".
Fuente: El Autor
Elaborado por: El Autor

4.3.5 Vocabulario Rank.

El resultado final de esta etapa fue la culminación del vocabulario “Rank” el cual cumple con las especificaciones propuestas al comienzo. Cabe destacar que siguiendo los principios de la web semántica, para la realización del modelado de datos se intentó en lo posible la reutilización de la mayor cantidad de clases y atributos por cuanto mucho de lo que se explica a continuación pertenece al vocabulario RDF Data Cube, sin embargo en estas secciones se explica y detalla su utilización dentro proyecto, centrándose de manera directa en la relación que existe entre las clases y los datos que se está modelando.

En varias de estas secciones también se explica la estructura fundamental del vocabulario “Rank” y como sus clases se relacionan con las clases del vocabulario RDF Data Cube, para facilitar su entendimiento se ha decido explicar todos estos conceptos en secciones, puesto que el diagrama completo puede resultar confuso y difícil de entender.

4.3.5.1 Clases DataSet, DataStructureDefinition, Slice y Observation.

En esta sección se detallan cuatro de los componentes fundamentales que permiten realizar la organización de los diferentes datos e información de las revistas científicas de manera estructurada. Estas clases han sido tomadas directamente del vocabulario RDF Data Cube por cuanto poseen todas sus propiedades y características.

Para poder explicar de manera correcta su funcionamiento se utilizara de referencia la Figura № 32, que corresponde con un fragmento del modelado en donde ya se encuentran colocados ciertos datos.

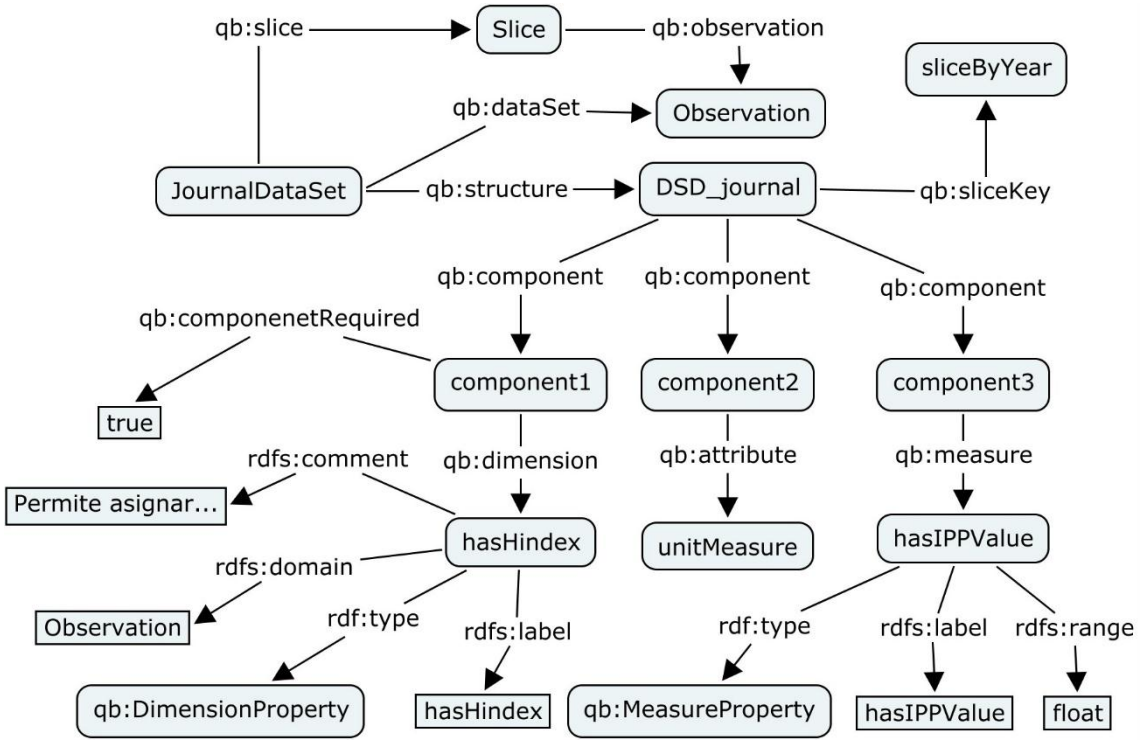


Figura 32 Estructura superior del modelo.
Fuente: El Autor
Elaborado por: El Autor

El primer elemento a destacarse resulta ser JournalDataSet, una instancia de la clase DataSet, este permite modelar todo el conjunto de datos, englobando toda la información recolectada de las diferentes revistas científicas. A su vez estos pueden ser estructurados de diferente forma dependiendo del uso que se les desee dar, en este contexto se emplean de los elementos Slice y Observation, los cuales se relacionan con el DataSet mediante las propiedades “slice” y “dataSet”.

El primero es una instanciación de la clase Slice que se encarga de agrupar la información del DataSet que se encuentra divida en Observations, considerando rasgos

o características comunes que posean, pudiéndose agrupar de acuerdo a las diferentes dimensiones o la de mayor relevancia. En este caso debido a que la información se encuentra dividida por años se decidió realizar los fragmentos de la información por dicho atributo, de tal forma que la relación existente entre las Slices y las Observations se realiza mediante la propiedad “observation”.

El elemento Observation como era de esperarse se instancia directamente de la clase Observation del vocabulario RDF Data Cube, este modela cada registro de información recopilada, por cuanto posee múltiples relaciones con otras clases que permiten representar las características y medidas que encontramos en los diferentes registros. Cada nuevo registro que se desee agregar estaría representado por una Observation.

Finalmente el último componente a ser descrito resulta no solo uno de los más importantes, sino también el que caracteriza y diferencia al vocabulario RDF Data Cube de los demás. El DSD_journal es una instanciación de la clase DataStructureDefinition y su función es la de definir la estructura interna del conjunto de datos que se está modelando mediante el uso de components y especificando la clave (sliceKey) mediante la cual se agruparan las Observations en Slices, en este caso los ejemplos que se observan son propiedades que luego en la inclusión de datos permitirán asignar los indicadores y sus valores correspondientes a las diferentes instancias que se crearan de la clase Observation. Esto al mismo tiempo permite definir a que hace referencia cada propiedad, pudiendo clasificarse en una de tres posibles opciones: Atributos, Medidas o Dimensiones. Otra característica importante del DataStructureDefinition es que por cada componente se puede definir su obligatoriedad o no, con esto se puede establecer un nivel mínimo de datos para cada Observation.

4.3.5.2 Clase Journal, Concept y Organization.

Adentrándonos más en el modelo, resulta esencial explicar la forma en la que cada Observation organiza internamente los datos de las revistas científicas. Para modelar esta información se hace uso de tres clases diferentes: Journal, Concept y Organization. Cada una de ellas posee diferentes relaciones las que se pueden observar en la Figura № 33.

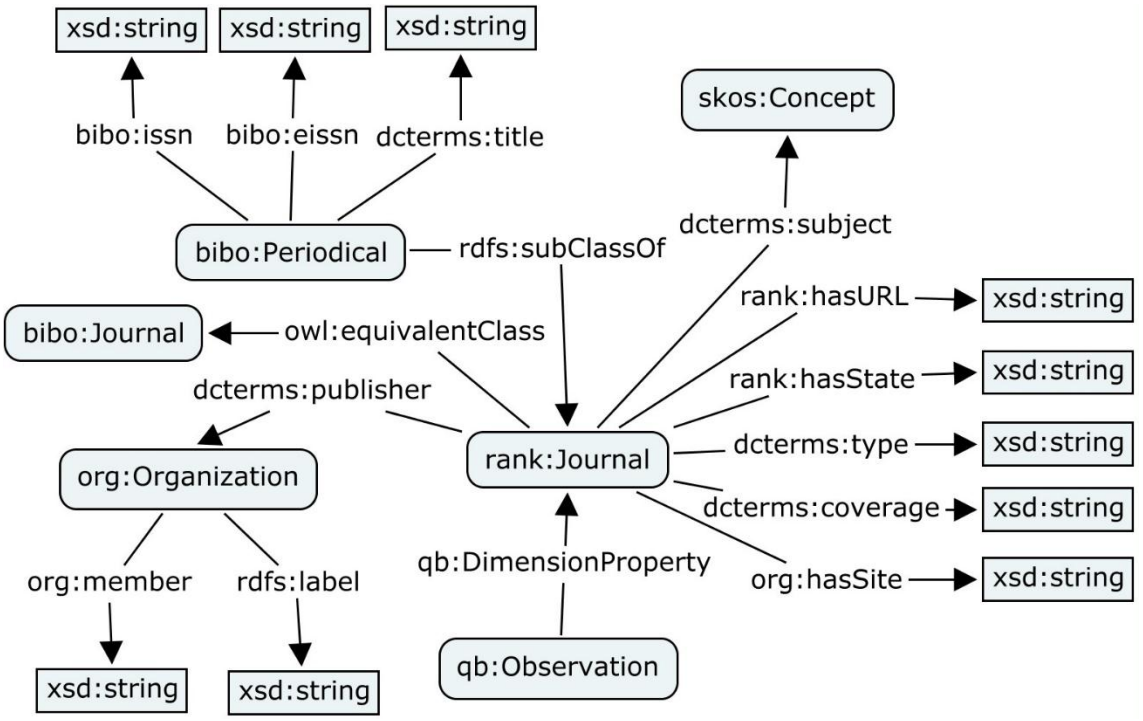


Figura 33 Estructura Interna del Modelo.
Fuente: El Autor
Elaborado por: El Autor

Journal es la primera de las clase creada en el vocabulario Rank y su principal objetivo es modelar las características fundamentales de las revistas científicas, tal como podemos observar en la Figura № 33. Se puede decir que es una especialización directa de la clase Periodical del vocabulario Bibo por compartir algunas similitudes, sin embargo las características principales son diferentes, además se han incluido algunas propiedades específicas que permitan representar de manera fidedigna la información recolectada, entre los principales campos agregados encontramos el tipo de revista científica, el estado, la URL, cobertura, y otros indicadores que fueron creados específicamente en el vocabulario Rank.

Debemos destacar que la clase Journal se encuentra representando una dimensión dentro de modelado y además posee importantes relaciones con las clases Concept del vocabulario SKOS, que es la encargada de modelar de las áreas de investigación y Organization del vocabulario Organization, que permite modelar las editoriales de las revistas.

4.3.5.3 Clase Indicator y J_Quartil.

La siguiente sección a ser detallada corresponde a las clases “Indicator” y “J_Quartil” pertenecientes al vocabulario Rank, son clases creadas específicamente para este modelado, la primera permite crear instancias referentes a los indicadores que posean las revistas y la segunda se refiere específicamente al indicador del cuartil de la revista, sus estructuras se puede visualizar en la Figura № 34.

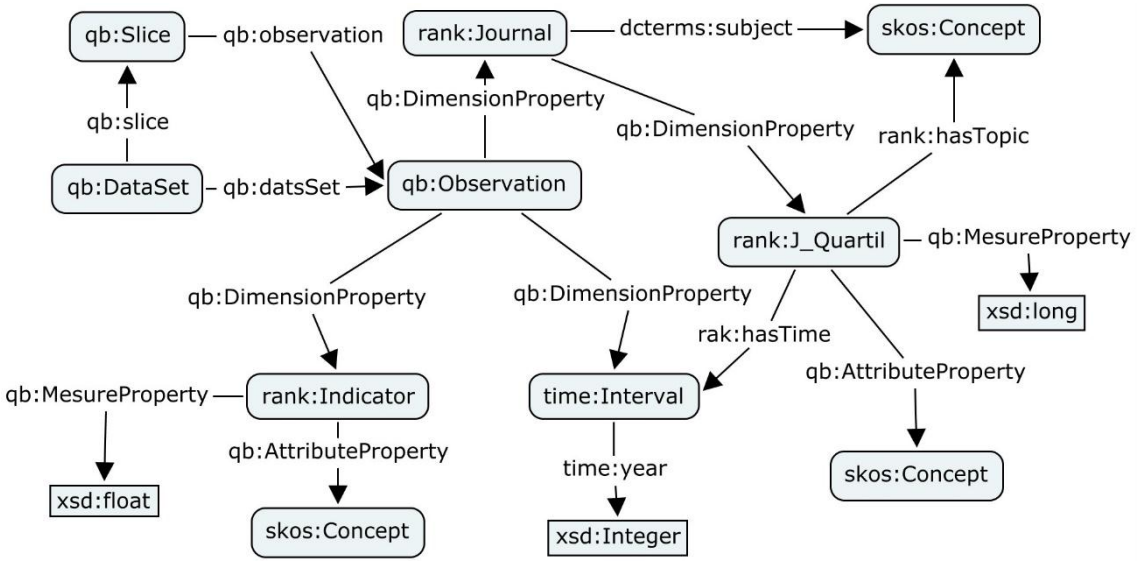


Figura 34 Estructura de la Observation.
Fuente: El Autor
Elaborado por: El Autor

La J_Quartil fue creada debido a que las relaciones que modela son muy específicas por cuando se necesitaba de una clase muy especial para representarlas, entre las principales clases con las cuales se relaciona podemos mencionar por ejemplo que cada Journal posee una valoración en cuartiles (J_Quartil) de las diferentes áreas de investigación que posee y que cada una de esta valoraciones se actualizan anualmente, obteniéndose así por ejemplo que una en revista un año determinado tenga uno o varios indicadores J_Quartil con los cuales se relacione.

En cambio en la clase Indicator se debe destacar que un indicador no necesariamente debe encontrarse presente en una revista científica o que incluso puede no poseer valor para un determinado año, por ese motivo en la Figura № 34 solo se coloca uno de forma demostrativa, pero en la realidad existirán muchos más. Dentro de esta clase encontramos no solo el valor de cada indicador si no también la unidad de medida que utiliza.

Cada indicador representa una dimensión según los conceptos del vocabulario Data Cube, por ese motivo también la propiedad mediante la cual se relaciona cada indicador con una Observation o Journal según sea el caso, es mediante una instanciación de la clase DimencionProperty, que permite crear determinadas propiedades, las mismas que se encuentran previamente especificadas en la DataStructureDefinition del modelado. Lo mismo sucede con el valor de las medidas donde se utiliza MesureProperty y con los atributos empleándose el AttributeProperty.

Finalmente cabe destacar que todas las secciones hasta aquí explicadas forman parte de un solo modelado El cual puede ser encontrado en el apartado Anexo 6, donde se evidencia de manera gráfica las relaciones de los componentes previamente expuestos.

4.4. Tercera etapa, generación.

Una vez obtenido el modelado de los datos de la etapa anterior lo siguiente a realizarse es la trasformación de la información según ese modelo, estas son las actividades que se explican en esta etapa, donde se detallan los principales aspectos para la trasformación de los datos que hasta ahora se encontraban en archivos Excel o CSV para ser convertidos a formato RDF y así ser utilizados en las etapas siguientes.

4.4.1 Generación de las tripletas

Para esta etapa debido a la complejidad de la estructura a ser modelada se decidió utilizar Jena, un Framework de código abierto que trabaja sobre el lenguaje de programación Java y que fue implementado en el entorno de desarrollo de NetBeans para poder realizar la generación de las tripletas. De esta forma se dejó de lado otras opciones como podrían ser la utilización software especializado (Esxcel2rdf, NOR2O, RDF skeleton, etc.) optándose así por desarrollar un algoritmo computacional que emplease dicho framework y permitiera un mayor control al momento de la generación, facilitando la depuración y corrección de errores.

El algoritmo recibe de entrada toda la información de las revista científicas contendidas en un archivo de Excel (o en el txt si se realizó la etapa de normalización) y dando como resultado la generación de un archivo RDF donde la información ya se encontrara modelada según lo indicado en Capitulo IV. El procedimiento seguido en esta etapa fue el siguiente:

1. Creación de un nuevo proyecto java del tipo “Java Aplication”
2. Crear una carpeta dentro del directorio raíz del proyecto y colocar el archivo (Excel o txt) con los datos de las revistas para poder tener acceso a la información.
3. Importación de las librerías necesarias, en este caso se utilizó específicamente las siguientes:
    - Jena .- Permite la codificación y creación del modelado
    - Poi.- Se encarga de la manipulación de archivos Excel en este caso específicamente para la lectura del archivo que contiene la información
    - XMLBeans.- Brinda la posibilidad de realizar la gestión de archivos XML en este caso se refiere al archivo RDF resultante.
    - Commons Lang.- Para manipulación de datos.
4. Codificación y ejecución del algoritmo de trasformación.

Al finalizar se obtuvo un programa que permitió el procesamiento de 22073 revistas científicas de las cuales se poseía información almacenada en 171 columnas y que fueron trasformadas en 3891066 tripletas RDF que posteriormente fueron cargadas en los servidores de base de datos.

Cabe destacar que el número de tripletas obtenido por cada revista varía, debido principalmente por la diferencia que existe en el número de áreas de investigación con las cuales se relacionan (categorías). Estas diferencias van desde revistas con trece distintas áreas de investigación hasta algunas que poseían solo una, esto influencia en la creación no solo de dichas instancias sino también de algunos indicadores que se encuentran relacionados a ellas, además dependiendo de si el registro de información ya fue leído anteriormente el programa la instancia no será creada si no simplemente enlazada.

A pesar de lo extenso y complejo del algoritmo utilizado se puede observar su funcionamiento de manera muy sencilla y resumida en la Figura № 35. Para conocer el código implementado en el algoritmo se recomienda visitar el Anexo 7.

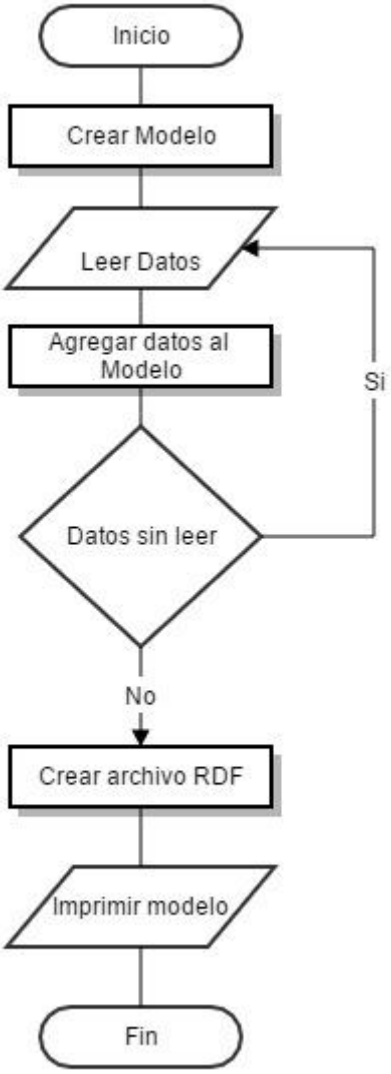


Figura 35 Diagrama de Flujo del Algoritmo de Transformación a Tripletas
Fuente: El Autor
Elaborado por: El Autor

Finalmente resta explicar que el algoritmo fue desarrollado para trabar directamente con los archivos resultantes del capítulo III debido a que esa era la manera sencilla de realizarlo, puesto que transferir la información a una base de datos para ser consumidos desde ahí hubiera supuesto un trabajo extra que no supondría mayores beneficios.

4.5. Comentarios finales

Completado este capítulo cabe resaltar que una de las etapas más importantes e influyentes para el proyecto es el modelado de datos, puesto que para su realización se debe entender los vocabularios que están utilizando, comprendiendo de manera muy clara sus características y limitaciones.

Esto resulta de especial interés debido a que en la etapa de la generación de tripletas poder encajar los datos con los modelos suele resultar una tarea muy ardua y difícil, en especial cuando no se posee practica en este tema, lo cual conlleva a la experimentación mediante prueba y error para obtener los mejores resultados. Viéndose así la importancia de la experiencia y la práctica que implican estas etapas de la publicación de datos.

CAPITULO V
PUBLICACION Y EXPLOTACION DE DATOS

5.1. Introducción

En este capítulo de detallan las últimas etapas necesarias para la publicación de datos, considerándose aspectos referentes a como almacenar la información ya transformadas a RDF y la construcción de un sitio web que ayude a su visualización y explotación permitiendo así que puedan ser de utilidad no solo para personas del área de la informática sino que también puedan ser aprovechados por el público en general.

5.2. Cuarta etapa, publicación

Terminada la generación de tripletas del capítulo anterior, la información debe ser almacenada para posteriormente poder ser utilizada, este aspecto es el que se trabaja en la etapa de publicación.

La primera tarea a efectuarse fue la selección de los servidores y bases de datos usados para el almacenamiento de la información. En un principio se decidió trabajar utilizando dos servidores diferentes, esto con motivo de poder realizar pruebas de rendimiento del modelo y su funcionamiento, permitiendo así al final seleccionar uno de ellos para ser servidor principal y seguir trabajando en futuras extensiones del proyecto.

La primera herramienta que se utilizó fue Apache Jena Fuseki, un servidor SPARQL de código libre, sumamente sencillo de utilizar y que no requiere instalación para trabajar, lo único que necesita es realizar su descarga, descomprimirlo y mandarlo a ejecutar mediante una línea de comandos en la consola.

El segundo servidor empleado fue Apache Marmotta, es de código libre y posee las siguientes características: integración con Base de datos, ejecución de consultas mediante el uso de servicios web, mecanismos de seguridad básicos integrados, apartado para graficación y visualización de datos, entre otras. Siendo este último seleccionado como el servidor principal debido a ser el más completo y que mejor se adapta al proyecto.

La instalación de Apache Marmotta resulta sumamente sencilla siendo necesario únicamente seguir las instrucciones que vienen incluidas cuando se descarga el archivo comprimido, sin embargo para realizar las configuraciones en algunos casos si no funciona mediante la interfaz del aplicativo puede resultar necesario efectuar los cambios directamente en los archivos, un ejemplo de esta situación resulta ser el cambio de la configuración de base de datos que viene por defecto. En el proyecto este cambio

se realizó para sustituir la base de datos “H2” que resulta ser sumamente sencilla por una más robusta y con mejor rendimiento como es la base de datos “PostgresSQL”.

Para efectuar esta tarea lo primero a realizarse luego de la instalación de “PostgresSQL” es la creación de una nueva base de datos con el nombre “marmotta”, en esta será donde se almacenara la información.

Posteriormente es necesario realizar algunos cambios en el archivo “system-config.properties” que se encuentra dentro de la ruta de instalación de Apache Marmota en la carpeta “marmotta-home”.

Los apartados a ser cambiados en el caso de que se trabaje con los parámetros por defecto son los siguientes, aunque estos pueden cambiar dependiendo de la configuración pudiendo personalizarse el usuario y la contraseña.

```
database.url =
jdbc:postgresql://localhost:5432/marmotta?prepareThreshol
d=3;MVCC=true;DB_CLOSE_ON_EXIT=FALSE;DB_CLOSE_DELAY=10
database.type = postgres
database.user = postgres
database.password = *****
```

Figura 36 Ejemplo configuración de Marmotta.
Fuente: El Autor
Elaborado por: El Autor

5.3. Quinta etapa, explotación de datos

Terminadas las configuraciones lo siguiente es la explotación de los datos, para esto se decidió realizar la construcción de una página web que permitiera no solo la búsqueda de la información, sino que además esta fuera sencilla de entender para el usuario siendo utilizadas gráficas comparativas para facilitar su entendimiento.

La arquitectura utilizada para la construcción del sitio web tal como se muestra en la Figura № 37 es del tipo cliente-servidor de tres capas. Se decidió utilizar este modelo por ser el que mejor se adaptaba con las necesidades del proyecto, permitiendo una clara división entre la capa de presentación, capa de negocios y la capa de datos de las cuales se habla a continuación.

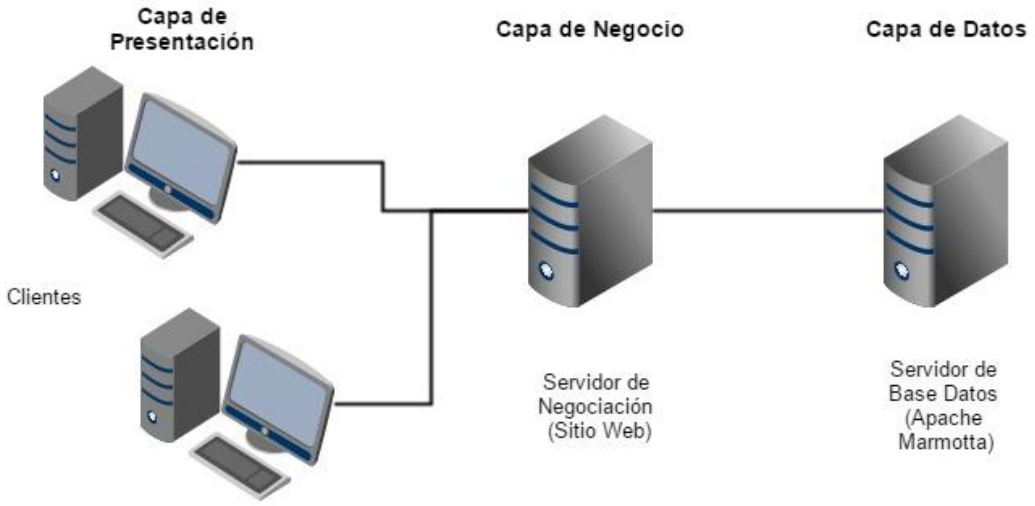


Figura 37 Arquitectura 3 Capas
Fuente: El Autor
Elaborado por: El Autor

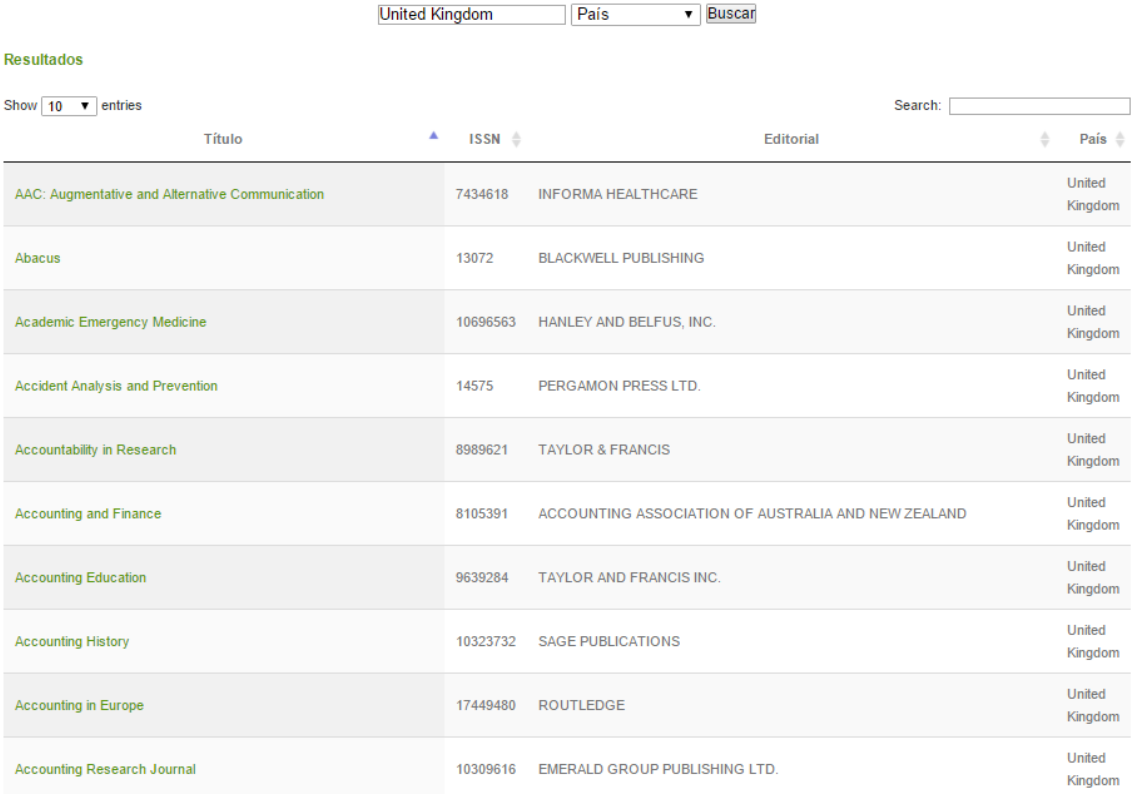
Capa de Presentación.- Se refiere a quienes serán los que soliciten los recursos, normalmente mediante un navegador web como pueden ser: Opera, Internet Explorer, Google Chrome entre otros

Capa de Negocio.- Este es un nivel intermedio el cual tiene como objetivo la entrega de los recursos solicitados al cliente. Para nuestro caso se refiere al sitio web desarrollado y que interactúa directamente con la capa de datos para acceder a la información que necesite.

Capa de Datos.- Es donde se encuentra almacenada la información, en este caso se refiere al gestor de base de dados que aloja a las tripletas RDF y que al momento de recibir una solicitud del sitio web será el encargado de enviar la información requerida.

Explicado a mayor detalle el sitio web consulta los datos que se encuentran almacenados en el servidor de base de datos mediante el uso de servicios web que permiten la ejecución de consultas SPARQL estructuradas. La información es obtenida en formato JSON para posteriormente ser procesada y mostrada en uno de los tres diferentes apartados de consulta que brinda el sitio web, los indicadores son mostrados en gráficas realizadas mediante el uso de la librería “google charts”, mientras que los demás datos son expuestos en formato de texto. (Para mas detalles técnicos revisar el Anexo 8) A continuación se muestra un ejemplo de los diferentes apartados y la forma en la que se muestra la información con la revista “Journal of Geophysics and Engineering” como referencia.

El primer apartado es el de búsqueda de una revista científica, en esta sección se puede seleccionar entre cuatro diferentes criterios para realizar la búsqueda los cuales son: “Titulo”, “País”, “Número ISSN” y “Editorial”. Los resultados de la búsqueda se presentan en una tabla que permite ordenar los resultados de manera descendente o ascendente tal como se muestra en la Figura № 38.



Título	ISSN	Editorial	País
AAC: Augmentative and Alternative Communication	7434618	INFORMA HEALTHCARE	United Kingdom
Abacus	13072	BLACKWELL PUBLISHING	United Kingdom
Academic Emergency Medicine	10696563	HANLEY AND BELFUS, INC.	United Kingdom
Accident Analysis and Prevention	14575	PERGAMON PRESS LTD.	United Kingdom
Accountability in Research	8989621	TAYLOR & FRANCIS	United Kingdom
Accounting and Finance	8105391	ACCOUNTING ASSOCIATION OF AUSTRALIA AND NEW ZEALAND	United Kingdom
Accounting Education	9639284	TAYLOR AND FRANCIS INC.	United Kingdom
Accounting History	10323732	SAGE PUBLICATIONS	United Kingdom
Accounting in Europe	17449480	ROUTLEDGE	United Kingdom
Accounting Research Journal	10309616	EMERALD GROUP PUBLISHING LTD.	United Kingdom

Figura 38 Resultado de la Búsqueda “United Kingdom” en el Sitio Web Desarrollado.
Fuente: El Autor
Elaborado por: El Autor

Luego de seleccionar la revista que se está buscando (Journal of Geophysics and Engineering) se mostraran todos los datos que se poseen de la misma, algunos de ellos en formato texto como es el caso del título de la revista, ISSN, eISSN, Editorial, grupo de la editorial y país.

En cambio los periodos de tiempo en los cuales ha permanecido activa, las áreas de investigación y calificación en cuartiles por cada área se muestran en gráficos. Un ejemplo de esto se muestra en la Figura № 39.

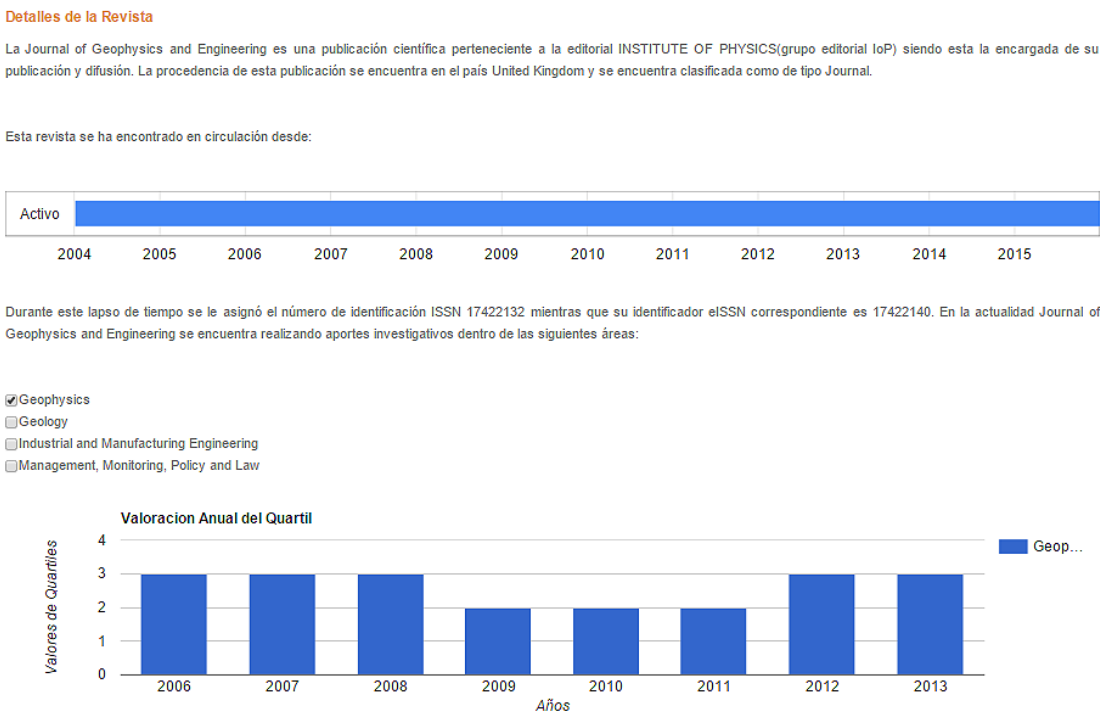


Figura 39 Detalles de la Revista "Journal of Geophysics and Engineering".
Fuente: El Autor
Elaborado por: El Autor

Para la información referente a los indicadores se incluyeron gráficas comparativas entre el SNIP, SJR e IPP de la revista y una tabla que muestra el índice H, el total de documentos publicados, entre otros datos, considerando todo esto la vista obtenida para nuestro ejemplo se observa en la Figura 40.

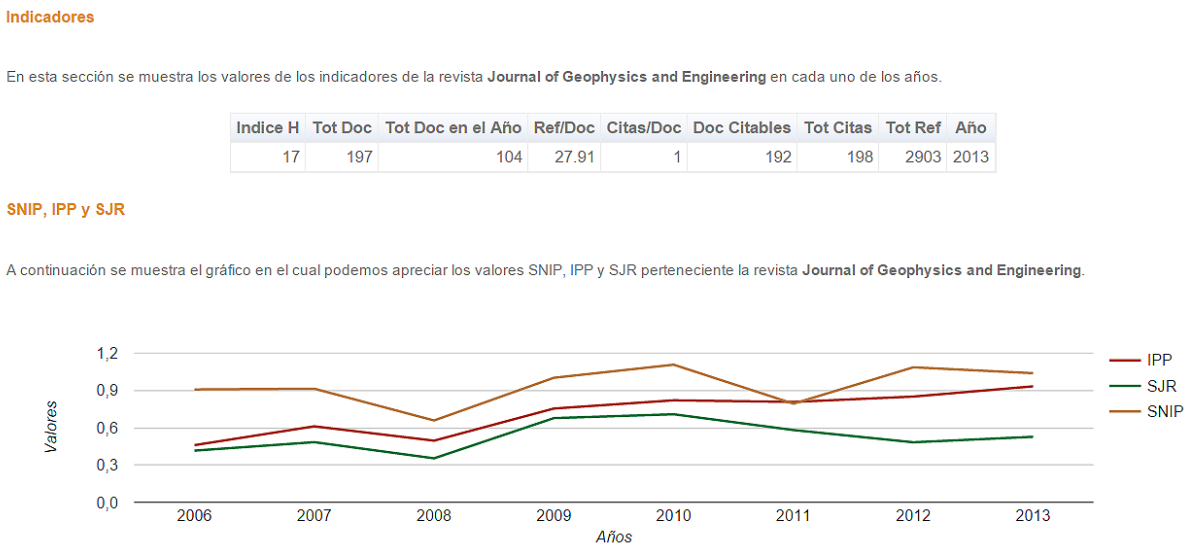


Figura 40 Indicadores de la Revista "Journal of Geophysics and Engineering".
Fuente: El Autor
Elaborado por: El Autor

El segundo apartado corresponde a la sección de comparativas, esta consiste en la búsqueda paralela de hasta cuatro revistas científicas con la finalidad de examinar sus características e indicadores al mismo tiempo. Esta sección comprende similitudes entre las áreas de investigación de las revistas, por cuanto para su correcto funcionamiento se deben comparar revistas que compartan áreas de investigación y limitar la comparativa a un año específico, en este caso se ha limitado el tiempo para trabajar netamente con los años 2013 y 2014 debido a que se posee más información en estas fechas. De igual forma se muestran datos de la valoración de los Cuartiles, documentos publicados, H – index, entre otros.

A continuación se muestran cuatro diferentes gráficos obtenidos al realizar una comparativa éntre las revistas “Journal of Economic Studies” y “Local Economy” en el área de “Economics, Econometrics and Finance”.

La primera es la Figura № 41 donde se muestra la comparativa entre los cuartiles de las revistas, aqui resalta una leve supremacía de la revista “Local Economy” en los años 2010 – 2012.

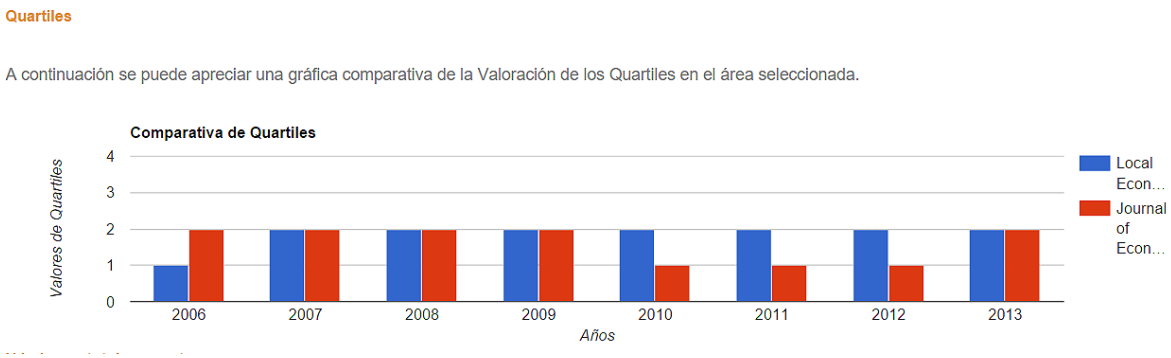


Figura 41 Cuartiles de las Revistas “Journal of Economic Studies” y “Local Economy”.
Fuente: El Autor
Elaborado por: El Autor

La Figura № 42 muestra la relación entre el H - index y el total de documentos publicados, resaltando que en este caso el H - index de la revista “Journal of Economic Studies” es mayor, aun cuando el número de publicaciones que realiza es menor que el realizado por su competidor.

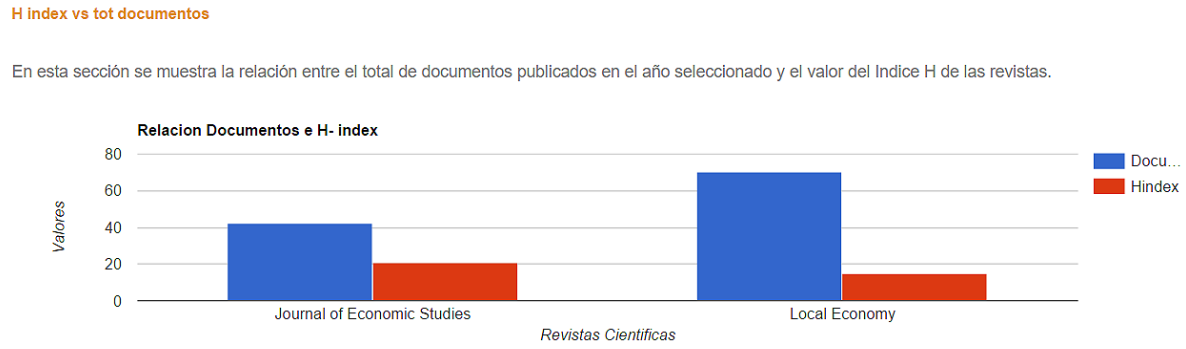


Figura 42 H - index vs Total Documentos de las Revistas “Journal of Economic Studies” y “Local Economy”.
Fuente: El Autor
Elaborado por: El Autor

La Figura № 43 muestra los indicadores IPP, SNIP o SJR de acuerdo a como se los seleccione efectuándose una comparativa de ellos, en este caso nos centraremos únicamente en el SJR donde se observa que ambas revistas poseen valores relativamente cambiantes, “Local Economy” se encuentra con un valor en alza, mientras que para “Journal of Economic Studies” el valor esta en declive.

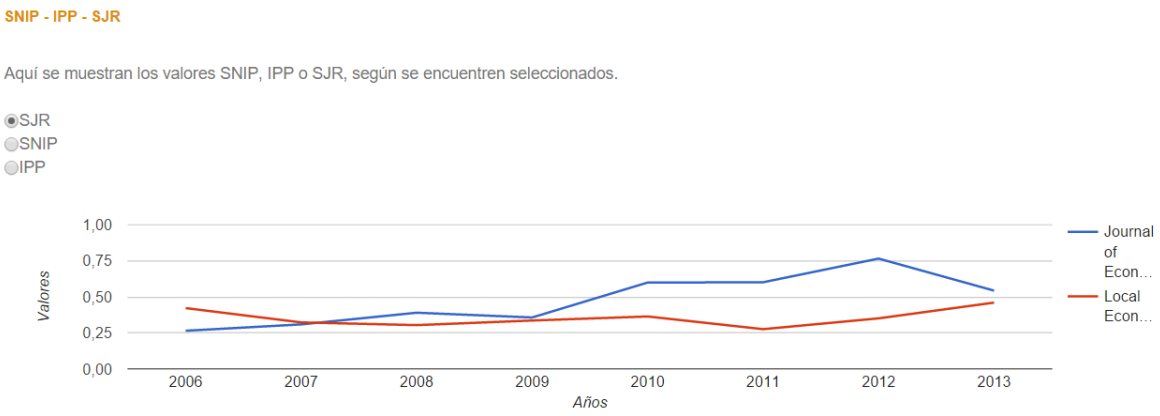


Figura 43 Relación del SJR de las Revistas “Journal of Economic Studies” y “Local Economy”.
Fuente: El Autor
Elaborado por: El Autor

La Figura № 44 muestra la relación entre el total de publicaciones realizadas y cuáles de estas han resultado ser documentos citables. Aquí se resalta que “Local Economy” publica más y mayor variedad de documentos.

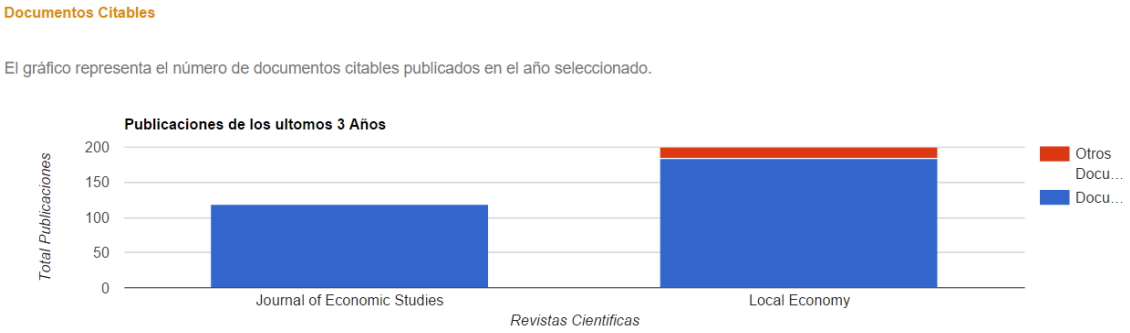


Figura 44 Documentos Citables de las Revistas “Journal of Economic Studies” y “Local Economy”.
Fuente: El Autor
Elaborado por: El Autor

Finalmente el último apartado que compone el sitio web es el de visualizaciones generales, aquí se muestra la información acumulada de toda la base de datos considerando las principales editoriales, número de revistas científicas por país, ente otros. Las Figuras № 45, № 46 y № 47 muestran una serie de imágenes de cómo se observa en un navegador:

Figura 45 Mapa General de la Distribución de las Revistas en Todo el Mundo.
Fuente: El Autor
Elaborado por: El Autor

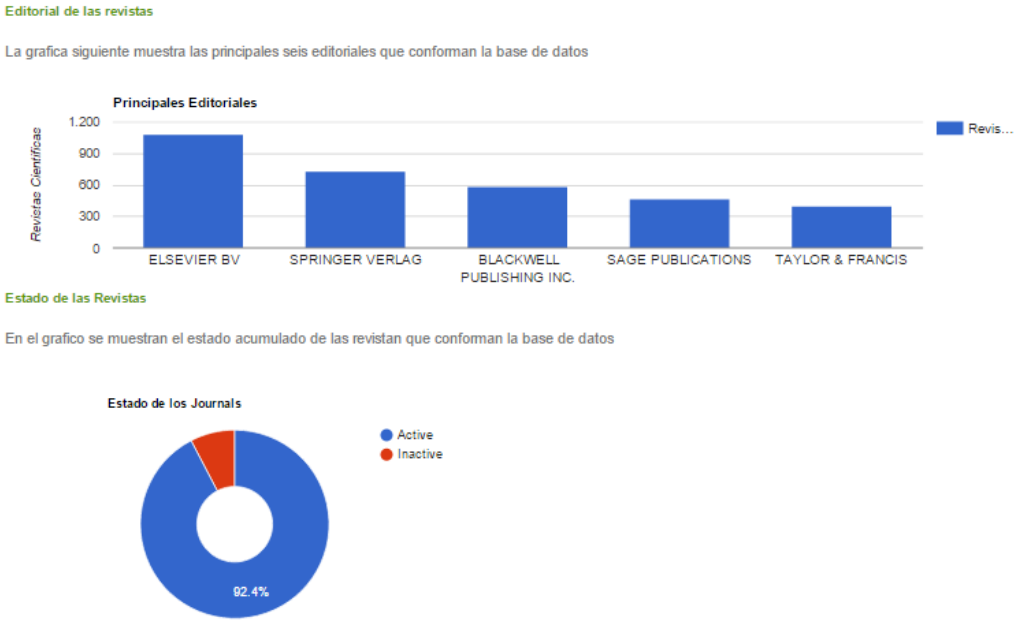


Figura 46 Graficas de Principales Editoriales y Estado de las Revistas
Fuente: El Autor
Elaborado por: El Autor

Figura 47 Tipos de Publicaciones.
Fuente: El Autor
Elaborado por: El Autor

5.4. Análisis de rendimiento

Una vez terminada la explotación de los datos lo siguiente a realizarse correspondía a la ejecución de una serie de mediciones y pruebas que permitieran conocer tanto los tiempos de ejecución como la veracidad de los datos presentados en el sitio web, esta información era sumamente importante puesto que en caso de obtenerse resultados negativos permitirá realizar los ajustes pertinentes y así corregir dicho errores.

El primer aspecto que las pruebas examinaron fueron los tiempos de respuesta de las consultas, para esta tarea se decidió utilizar un complemento llamado “Advanced REST client” y que se encuentra disponible para el navegador Google Chrome. En la Figura № 48 se muestra una ilustración de la herramienta para tener una mejor idea de su funcionamiento.

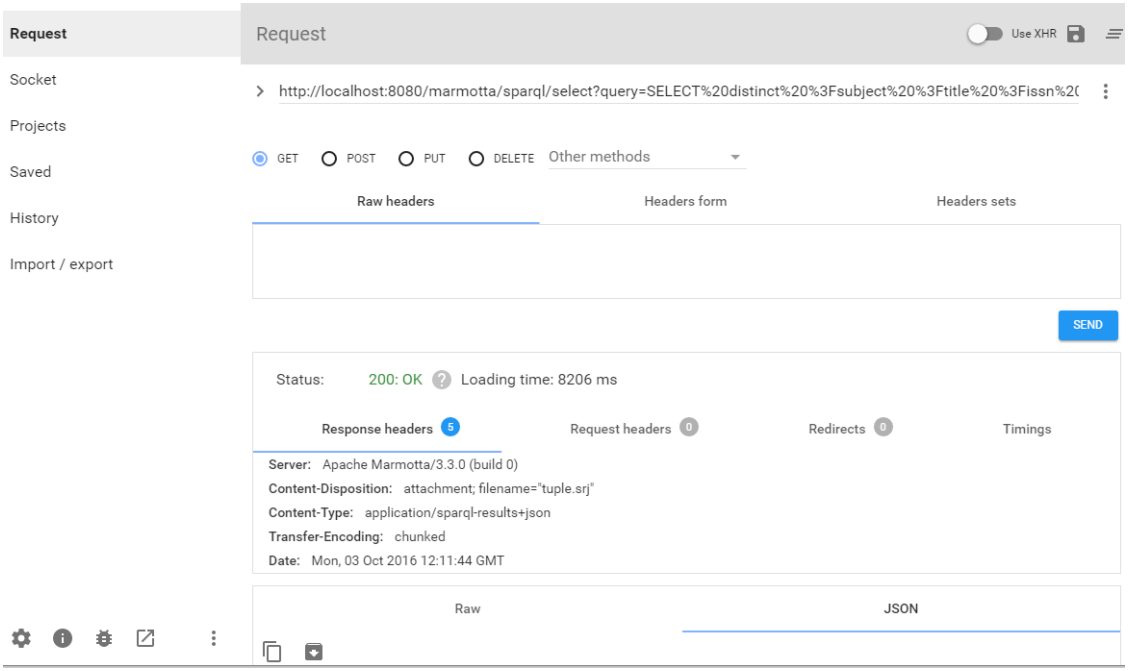


Figura 48 Herramienta “Advanced REST client”
Fuente: El Autor
Elaborado por: El Autor

Este complemento fue utilizado debo a ser el que mejor se adaptaba a las necesidades, permitiendo interactuar de manera directa con el servidor de base de datos Apache Marmotta brindándonos información relacionada con los tiempos de respuesta al ejecutar consultas SPAQL, el tipo de respuesta devuelta por el servidor, posibles problemas en la ejecución, entre otros datos.

Para realizar las pruebas solo es necesario colocar el url correspondiente al Endpoint de Apache Marmotta en la arte donde indica la herramienta y enviar a ejecutar, cabe destacar que el url que se envía a probar ya contiene la sentencia SPARQL que se desea probar. A continuación se muestran algunas de las sentencias SPAQL probadas y sus distintos tiempos de respuestas

- Consulta para obtener el número de revistas científicas por país. Tiempo de ejecución 1000 ms.

```
SELECT ?pais (COUNT(?pais) AS ?count)
WHERE {
?subject <http://www.w3.org/ns/org#hasSite> ?pais.
}
group by ?pais order by desc (?count)
```

Figura 49 Consulta número de revistas por país (SPARQL).
Fuente: El Autor
Elaborado por: El Autor

- Consulta para obtener el título, número ISSN, país y nombre de la editorial de todas las revistas que posean la palabra “Journal” en su nombre. Tiempo de ejecución empleado 7386 ms.

```
SELECT distinct ?subject ?title ?issn ?pn ?site
WHERE {
?subject  <http://purl.org/dc/terms/title> ?g.
?subject  <http://purl.org/dc/terms/title> ?title.
?subject <http://purl.org/ontology/bibo/issn> ?issn.
?subject <http://www.w3.org/ns/org#hasSite> ?site.
?subject <http://purl.org/dc/terms/publisher> ?publisher.
?publisher <http://www.w3.org/2000/01/rdf-schema#label> ?pn.
FILTER regex(?g, "Journal ","i")
}
```

Figura 50 Consulta para buscar revistas con la palabra Journal (SPARQL)..
Fuente: El Autor
Elaborado por: El Autor

- Consulta para obtener el título, número ISSN, cobertura, número e-ISSN, tipo, categoría, editorial, grupo al cual pertenece la editorial, país y estado de la revista “4OR”. Tiempo de ejecución 41 ms.

```
SELECT distinct ?estado ?title ?issn ?coverage ?eissn ?type ?site
?categoryname ?pn ?pg
WHERE {
<http://lsd.utpl.edu.ec/rank/resourse/Or/16194500>
<http://lsd.utpl.edu.ec/owl/hasState> ?estado.
<http://lsd.utpl.edu.ec/rank/resourse/Or/16194500>
<http://purl.org/dc/terms/title> ?title.
<http://lsd.utpl.edu.ec/rank/resourse/Or/16194500>
<http://purl.org/ontology/bibo/issn> ?issn.
<http://lsd.utpl.edu.ec/rank/resourse/Or/16194500>
<http://purl.org/dc/terms/coverage> ?coverage.
<http://lsd.utpl.edu.ec/rank/resourse/Or/16194500>
<http://purl.org/ontology/bibo/eissn> ?eissn.
<http://lsd.utpl.edu.ec/rank/resourse/Or/16194500>
<http://purl.org/dc/terms/type> ?type.
<http://lsd.utpl.edu.ec/rank/resourse/Or/16194500>
<http://www.w3.org/ns/org#hasSite> ?site.
<http://lsd.utpl.edu.ec/rank/resourse/Or/16194500>
<http://purl.org/dc/terms/subject> ?category.
?category <http://www.w3.org/2004/02/skos/core#prefLabel> ?categoryname.
<http://lsd.utpl.edu.ec/rank/resourse/Or/16194500>
<http://purl.org/dc/terms/publisher> ?publisher.
?publisher <http://www.w3.org/2000/01/rdf-schema#label> ?pn.
?publisher <http://www.w3.org/ns/org#member> ?pg.
}
```

Figura 51 Consulta obtener datos de la revistas 4OR (SPARQL).
Fuente: El Autor
Elaborado por: El Autor

Otro aspecto examinado fue los tiempos de carga del sitio web, los cuales fueron medidos utilizando las herramientas para desarrolladores ofrecidas por los navegadores, puntualmente se utilizó el apartado de “Timeline” para medir los tiempos de cada una de la diferentes páginas del sitio web. Los promedios de los resultados se muestran en la Tabla № 5.

Tabla 5 Tiempo de carga del sitio web

Página	**Tiempo de carga**
Inicio	3 s
Búsquedas	4 s
Comparativas	5 s
Visualizaciones	5 s
Acerca de	3 s

Contacto	3 s

Fuente: El Autor
Elaborado por: El Autor

Si analizamos los resultados se observa que los tiempos de carga se encuentran en un rango aceptable puesto que si los comparamos contra un sitio web similar como es Scimago obtenemos un promedio de tiempo general del sitio de 2-5 segundos, viéndose una diferencia de no más de 1 segundo con respecto al sitio web desarrollado.

Finalmente la última prueba realizada se refiere a la fiabilidad de la información, para esto se realizaron una serie de búsquedas en el sitio web desarrollado y se compararon con los datos ofrecidos por Scimago y Journal Metrics. En la Tabla № 6 se presentan algunos de los resultados obtenidos

Tabla 6 Comparativa de Datos

Fuente de Datos	Nombre	ISSN	Tipo	Categoría № 1	SNIP 2014	IPP 2014	SJR 2014
Sitio web	Abacus	00013072	Journal	Accounting	0,767	1,119	0,707
Scimago	Abacus	00013072	Journal	Accounting	X	X	0,707
Journal Metrics	Abacus	00013072	X	X	0.767	1.119	0.707
Sitio web	21st Century Music	15343219	Journal	Music	X	X	0.113
Scimago	21st Century Music	15343219	Journal	Music	X	X	0.113
Journal Metrics	21st Century Music	15343219	X	X	X	X	0.113
Sitio web	Hearing Journal	07457472	Journal	Speech and Hearing	0.151	0.189	0.125
Scimago	Hearing Journal	07457472	Journal	Speech and Hearing	X	X	0.125
Journal Metrics	Hearing Journal	07457472	X	X	0.151	0.189	0.125

Fuente: El Autor
Elaborado por: El Autor

5.5. Ventajas de la implementación

Como se puede observar en el apartado anterior los tiempos de respuesta y resultados de las pruebas fueron positivos, sin embargo estos no son los únicos aspectos que se puede analizar.

En este caso debido a que la implementación realizada se encuentra modelada según los conceptos de web semántica tenemos varias ventajas con respecto a esquemas más tradicionales, entre ellas las principales son:

Escalabilidad y Flexibilidad.- Se encuentra bien preparada para gestionar el crecimiento en la cantidad de datos y los nuevos requerimientos que pudiesen surgir. Esto se debe principalmente a la forma en la cual se encuentra modelados y almacenados los datos, puesto que permiten la inclusión de nueva información sin afectar al modelo que ya se encurta en funcionamiento, algo que resultaba muy difícil de conseguir con el modelo relacional que en muchos casos necesitaba de una reingeniería completa en sus tablas para poder incluir nueva información.

Información mejor Estructurada.- Debido a que la información se encurta estructurada en tripletas resulta sumamente sencilla entender la forma en la cual se organizan los datos, esto a su vez permite que la información pueda ser utilizada de manera más fácil y eficiente, eliminando la necesidad de tener que adaptar las aplicación directamente a un esquema de tablas como el que se usa en los modelos relacionales, en contraparte con el modelo implementado simplemente se debe conocer a breves rasgos el modelado para poder realizar una consulta.

Mejora en la Eficiencia.- Gracias al uso de RDF Stores, el modelo implementado se encuentra mejor preparado para gestionar grandes volúmenes de información, lo cual en la práctica se traduce en menos tiempo de ejecución, resultados más acertados en las búsquedas y aplicaciones más eficientes.

Optimización en consultas.- En este caso las consultas son realizadas mediante el uso de SPAQL lo cual permite eliminar varios de los problemas que se presentan en los modelos relacionales como son los diferentes tipos de formatos en los cuales se encuentran los datos, las diferentes estructuras y consultas poco eficientes al momento de tener que consultar de varios orígenes, un claro ejemplo de esto se muestra en la Figura № 52 donde observamos una consulta realizada sobre los datos del proyecto y que además obtiene información desde la Dbpedia para poder consultar el tipo de moneda que se maneja en los países a los cuales pertenece cada revista.

```
SELECT ?titulo ?pais ?pais2 ?m
WHERE {
  ?subject <http://purl.org/dc/terms/title> ?titulo.
  ?subject <http://www.w3.org/ns/org#hasSite> ?pais.
  service <https://dbpedia.org/sparql> {
    SELECT DISTINCT ?pais2 ?m
      WHERE
      {
      ?b <http://www.w3.org/1999/02/22-rdf-syntax-ns#type>
<http://dbpedia.org/class/yago/Economy108366753>.
      ?b <http://www.w3.org/2000/01/rdf-schema#label> ?p.
      ?b <http://dbpedia.org/property/currency> ?m.
      FILTER (lang(?p) = "en")
      BIND(REPLACE(?p, "@en", "", "i") AS ?pais2)
      }
 }
 FILTER (?pais = ?pais2)
}
```

Figura 52 Consulta obtener Titulo de la revista, país y moneda (SPARQL).
Fuente: El Autor
Elaborado por: El Autor

5.6. Retroalimentación e inserción de nuevos datos

Terminada la inserción del primer grupo de datos correspondientes a los años 2006-2013 se procedió a efectuar una etapa adicional en la cual se incluyó la información perteneciente al año 2014.

Esto permitió corroborar el funcionamiento de los procedimientos utilizados y obtener una referencia de los pasos mínimos a seguir para efectuar una inserción de datos. Siendo así en relación al primer grupo de datos se obtener un decremento importante en los pasos necesarios a seguir, esto se puede apreciar en la Figura № 53.

Figura 53 Primera inserción vs Segunda inserción de datos.
Fuente: El Autor
Elaborado por: El Autor

No obstante este no fue el único cambio puesto que también se apreció una reducción sustancial del tiempo y esfuerzo empleados, que en comparación solo se necesitó la mitad del tiempo que la vez anterior, esto es debido principalmente a que gracias al proceso ya establecido el trabajo resulta mucho más sencillo, y aún cuando existen ciertas partes que deben realizarse de manera manual, mediante al uso de ejemplos (en este caso los scripts realizados en la primera inserción de datos) esta tarea también se simplifica en gran medida.

Con esto podemos evidenciar la factibilidad de usar Data Cube como vocabulario principal para realizar publicación de datos estadísticos, además comprobamos la valía no solo de ciertas herramientas utilizadas en este proyecto, sino también de las técnicas y procedimientos empleados los mismos que fueron depurados y corregidos en cada paso hasta conseguir su correcto funcionamiento tal como se esperaba en un principio.

5.7. Comentarios finales

Terminada la publicación de datos se puede resaltar lo complicadas y abrumadoras que pueden llegar a ser algunas de sus fases, en este capítulo se puede apreciar lo compleja que se presenta la tarea de la explotación de los datos. Esto debido principalmente a que se deben considerar diferentes aspectos como son la eficiencia, eficacia, utilidad y visualización de la información.

Sin embargo también resulta ser una de las tareas más gratificantes siendo la primera ocasión en la cual se puede apreciar el trabajo de manera estética, y aunque para la mayoría de los usuarios resulta invisible, el sitio web es la culminación de mucho tiempo y esfuerzo por parte de las personas implicadas en su desarrollo.

Pero no debemos olvidar que realmente el trabajo nunca termina y aun cuando se podría dar por finalizada la publicación de datos, siempre se debe estar en constante búsqueda de mejoras y pensando en nuevas de utilizar e interpretar la información.

CONCLUSIONES

Una vez terminado el proyecto y habiéndose culminado el ciclo de vida de publicación de dados enlazados, lo que resta por realizar es un análisis crítico de los resultados obtenidos, los mismos que se encuentran condensados en las siguientes conclusiones:

- Los vocabularios que actualmente permiten la publicación de datos estadísticos (SKOVO, RDF-SDMX y RDF Data Cube) son válidos en diferentes situaciones, el más importante actualmente es el Data Cube porque adopta las principales características de sus antecesores como son su actitud multidimensional, fuerte definición de la estructura de los datos y gran adaptabilidad con otros vocabularios, siendo por estos motivo el vocabulario utilizado en el desarrollo de este proyecto y recomendando ampliamente su uso para proyectos similares.

- Se ha evidenciado que utilizando los métodos y procedimientos aquí detallados, es posible realizar la publicación de datos estadísticos, empleando datos en formatos Excel y CSV, que transformándolos a RDF y consolidándolos dentro de un solo repositorio, provean mayor interoperabilidad facilitando el mantenimiento, gestión y recuperación de la información.

- Terminado el proyecto se pudo constatar los beneficios de realizar la publicación de datos estadísticos, obteniéndose una base datos conformada por información de múltiples repositorios, que gracias al uso de RDF Stores resulta ser muy escalable y flexible, con un buen tiempo de respuesta a las búsquedas y en un formato entendible tanto por humanos y máquinas como es RDF.

- Se ha demostrado que mediante la utilización del vocabulario Data Cube y en combinación con herramientas “Open Source” es posible conformar una plataforma de trabajo que soporte la publicación de datos en la web semántica.

- Realizada la publicación de los datos estadísticos acerca de las revistas científicas y posteriormente efectuada una nueva actualización con los datos correspondientes al año 2014, se ha comprobado la reusabilidad del modelo, mejora en la eficiencia con relación a una disminución importante en el tiempo requerido y el correcto funcionamiento de los procedimientos, demostrando que es posible efectuar una implementación que combine no solo el aspecto de ingeniería sino también una interfaz amigable, intuitiva y de fácil uso para el usuario.

- Al finalizar el proyecto puedo valorar lo mucho que he aprendido en el transcurso del mismo, puesto que además de aplicar todos los conocimientos que he adquirido en mis años de estudio, he podido experimentar de primera mano los retos reales que implica un proyecto de esta envergadura, teniendo que comprender muchos conceptos para mi desconocidos, además he entendido el aspecto cambiante que puede tener un proyecto, viendo la necesidad de tener que adaptar las soluciones ante un determinado problema o imprevisto.

RECOMENDACIONES

Uno de los puntos más importantes de todo proyecto es reconocer los principales problemas afrontados y tomar medidas preventivas para que estos no se repitan en un futuro en proyectos de similares características. Con esta premisa en mente en esta sección se recapitulan una serie de recomendaciones que deben ser consideradas al momento de realizar la publicación de "linked data".

- Definir cuidadosamente el formato con el que se trabajaran los datos ya sea uno simple o normalizado, una elección equivocada puede afectar de manera negativa el proyecto, incrementado la cantidad de recursos necesarios, tiempo empleado, dificultar el procesamiento de datos e incluso ocasionar problemas al momento de incluir nuevos datos.

- Antes de emplear la herramienta "Open Refine" con los datos propios del proyecto, realizar pruebas de rendimiento con la finalidad de determinar la cantidad máxima de información procesable con la configuración establecida. Así en caso de ser necesario se incremente o disminuya la cantidad de memoria RAM asignada, puesto que una mala configuración de la misma puede provocar fallas y ocasionar que el programa deje de funcionar, perdiéndose el trabajo elaborado.

- Al momento de instalar la herramienta "Apache Marmotta" y en caso de que el computador se encuentre conectado a internet, verificar que se cuenta con todos los permisos de red necesarios, puesto que de no ser así la herramienta no se podrá instalar correctamente, en cuyo caso se recomienda realizar la instalación con el adaptador de red apagado.

- Al momento de la construcción del modelado se recomienda descomponer el problema en secciones pequeñas y que resulten de fácil comprensión, con la finalidad de facilitar el proceso, evitar futuros errores y disminuir su complejidad.

- Para la validación de los datos convertidos a RDF se recomienda utilizar "OOPS! (OntOlogy Pitfall Scanner!)" para verificar el dominio, rango, relaciones y errores sintácticos, conjuntamente con el "Validator Service"[45] ofrecido por la W3C para la comprobación visual del grafo y las tripletas generadas.

[45] http://www.w3.org/RDF/Validator/

- En la generación de tripletas considerar varias alternativas como son Jena, el complemento RDF para Google Refine, NOR2O u otras opciones antes de decidirse a usar una en concreto, puesto que la elección de la herramienta correcta supone un ahorro importante de tiempo y esfuerzo.
- Realizar pruebas de rendimiento que midan el tiempo de respuesta de las consultas SPARQL en el triple store implementado, puesto que en algunos casos la realización inadecuada de las mismas puede resultar en tiempos de espera muy extensos para el usuario final, y crear la ilusión de que el problema se encuentra en los datos o en el modelado.

TRABAJOS FUTUROS

Si bien el trabajo realizado en el presente proyecto se ha culminado satisfactoriamente, el mismo ha sido muy arduo y extenso, más esto no significa que no se pueda seguir avanzando, pues siempre existen nuevos horizontes que alcanzar, futuras metas que realizar e increíbles proyectos por efectuar. Por tal motivo a continuación se plantean una serie de posibles trabajos a ser realizados en un futuro con la finalidad de seguir mejorando y nunca detener el avance del conocimiento.

- Automatizar la preparación de los datos mediante el desarrollo de un sistema que implemente y ejecute los scripts desarrollados en el presente proyecto.
- Enlazar el repositorio de información de revistas científicas creado en este proyecto con uno que posea datos de los diferentes artículos publicados por cada revista, con el fin de obtener una mayor fuente de comparación y extender la información ofrecida al usuario.
- Ampliar el sitio web para que permita mayor interacción con el usuario, soportando valoraciones, opiniones y criterios de las revistas científicas. Además ofrecer la posibilidad de generar reportes del crecimiento con cada nueva actualización de datos.
- Extender el sitio web para poder incluir apartado de retroalimentación donde los usuarios puedan brindar sus opiniones y se permita de esta forma la mejora continua del sitio web y sus servicios.
- Incrementar la información de las revistas científicas con la finalidad de incluir aspectos referentes a reglas de publicación, URL a sitios web oficiales de las revistas, afiliaciones con otras entidades u organizaciones, cambios históricos de los nombres y otros aspectos relevantes.

BIBLIOGRAFÍA

acens. (2014). Bases de datos NoSQL . Qué son y tipos que nos podemos encontrar. Retrieved from https://www.acens.com/wp-content/images/2014/02/bbdd-nosql-wp-acens.pdf

Belloch, C., Mide, D., & Valencia, U. De. (2002). Las Tecnologías de la Información y Comunicación en el aprendizaje, 1–9.

Berners-Lee, T. (1998). Uniform Resource Identifiers (URI): Generic Syntax. Retrieved March 2, 2015, from https://www.ietf.org/rfc/rfc2396.txt

Cadme, E., & Piedra, N. (2014). Una aproximación para la descripción semántica de requisitos para categorización docentes de investigadores Ecuatorianos. *Tic Ec*, (2006), 97–110.

Corchuelo, R. (2008). Introduccion a la Web Semantica. *Sevilla: Universidad de Sevilla*, 18. Retrieved from http://www2.tdg-seville.info/projects/Integraweb/Seminars/semi-19-01-07-material.pdf

Cyganiak, R., & Jentzsch, A. (2014). The Linking Open Data cloud diagram. Retrieved March 4, 2015, from http://lod-cloud.net/

DB-Engines. (n.d.). Cassandra vs. MongoDB vs. Virtuoso Comparison. Retrieved January 27, 2015, from http://db-engines.com/en/system/Cassandra%3BMongoDB%3BVirtuoso

Dbpedia. (n.d.). Sobre DBpedia-Latinoamérica. Retrieved January 27, 2016, from http://es-la.dbpedia.org/home/

Emilio, J., & Gayo, L. (n.d.). Web Semántica RDF.

Emilio, J., Gayo, L., Farham, H., Fern, J. C., & Mar, J. (2014). Representing verifiable statistical index computations as linked data.

Fernández, R. C. (n.d.). Representación del Conocimiento . Web Semántica.

FOAF Vocabulary Specification. (2014). Retrieved March 6, 2015, from http://xmlns.com/foaf/spec/#sec-intro

Geo Linked Data Ecuador. (2014). Linked Data Ecuador. Retrieved March 19, 2015, from http://linkeddata.ec/?q=es/viewaboutGeo2014

Giraldo, G., Acevedo, J., & Moreno, D. (2011). An ontology for the representation of software design concepts, *8*, 103–110. Retrieved from http://www.redalyc.org/articulo.oa?id=133122679013

Gómez-pérez, A. A., & Suárez-figueroa, M. C. (2005). Scheduling using gOntt Scheduling using gOntt.

Heath, T., & Bizer, C. (2011). Linked Data: Evolving the Web into a Global Data Space. *Synthesis Lectures on the Semantic Web: Theory and Technology*, *1*(1), 1–136. http://doi.org/10.2200/S00334ED1V01Y201102WBE001

Introduction to SKOS. (2012). Retrieved March 5, 2015, from http://www.w3.org/2004/02/skos/intro

Janev, V., & Milosević, U. (2012). Integrating Serbian Public Data into the LOD cloud | Linked Data Stack. Retrieved March 19, 2015, from http://stack.linkeddata.org/stack-in-use/integrating-serbian-public-data-into-the-

lod-cloud/

Jara, J. E. (2012). Ontología. Retrieved from http://es.slideshare.net/juanemiliano/ontologa-14011321

Lee, M. (n.d.). URI (uri).

Linked Open Vocabularies. (n.d.). Linked Open Vocabularies (LOV). Retrieved April 14, 2015, from http://lov.okfn.org/dataset/lov/vocabs/qb

Lozada, P. (2014). Microsoft PowerPoint - Evolucion_Web [Modo de compatibilidad] - Evolucion_Web.pdf. Retrieved March 27, 2015, from http://julionica.udem.edu.ni/wp-content/uploads/2014/01/Evolucion_Web.pdf

Merino, P. J. M. (n.d.). Web Semántica Web Semántica :

Metrics, J. (n.d.). Journal Metrics. Retrieved May 26, 2015, from http://www.journalmetrics.com/about-journal-metrics.php

Mongo DB. (2015). Reinventando la gestión de datos | MongoDB. Retrieved March 19, 2015, from http://www.mongodb.com/es

Mynarz, J., Cyganiak, R., Iqbal, A., & Hausenblas, M. (n.d.). Modelling of Statistical Linked Data.

Ontology Engineering Group. (n.d.). La Metodología NeOn. Retrieved August 5, 2015, from http://mayor2.dia.fi.upm.es/oeg-upm/index.php/es/methodologies/59-neon-methodology

Oracle. (2015). Oracle NoSQL Database Technical Overview. Retrieved February 13, 2015, from http://www.oracle.com/technetwork/database/database-technologies/nosqldb/overview/index.html?ssSourceSiteId=ocomen

Privacy Policy Elseiver. (n.d.). Retrieved May 26, 2015, from http://www.elsevier.com/legal/privacy-policy

REDCEDIA. (n.d.). Plataforma de Integración, Publicación y Consulta Integrada de Recursos Bibliográficos en la Web Semántica. Retrieved from https://www.cedia.org.ec/proyecto-plataforma-de-integracion-publicacion-y-consulta-integrada-de-recursos-bibliograficos-en-la-web-semantica

Rodríguez Álvarez, J. M. (2012). Métodos Semánticos de Reutilización de Datos Abiertos Enlasados en las Licitaciones Públicas, 355.

Rodríguez Méndez, S. J. (n.d.). Web Semántica (Parte I): Vista General De Rdf, (5), 1–12.

Rouse, M. (2005). What is URN (Uniform Resource Name)? - Definition from WhatIs.com. Retrieved January 20, 2014, from http://searchsoa.techtarget.com/definition/URN

Ruiz, R. G. (2008). TFC : XML y Web Semántica, 235.

Scimago Journal & Country Rank. (n.d.). Scimago Journal & Country Rank. Retrieved January 27, 2014, from http://www.scimagojr.com/

Scopus. (n.d.). Retrieved May 26, 2015, from http://www.journalmetrics.com/about-scopus.php

Tauberer, J. (2006). What Is RDF. Retrieved March 2, 2015, from http://www.xml.com/pub/a/2001/01/24/rdf.html?page=1

The National Center for Biomedical Ontology. (2009). Comparison of Triple Stores, 1–

8. Retrieved from http://www.google.com.br/url?sa=t&rct=j&q=triplestore&source=web&cd=2&ved=0CHMQFjAB&url=http://www.bioontology.org/wiki/images/6/6a/Triple_Stores.pdf&ei=vdkWUJbbNIG-8ATrtYHICQ&usg=AFQjCNENyqLeZqhTXipCKMy1_lKn5t6itQ&sig2

Tiger Logic. (2013). D3 Data Model. Retrieved March 20, 2015, from https://www.tigerlogic.com/tigerlogic/pick/database/d3datamodel.jsp

Universidad Nacional Abierta y a Distancia. (n.d.). Formatos de dirección electrónica. Retrieved May 27, 2014, from http://datateca.unad.edu.co/contenidos/100201/HT2013Exe/leccin__13_formatos_de_direccin_electrnica.html

Universidad Pompeu Fabra. (n.d.). Tema 1 : Historia y evolución de Internet.

Universidad Técnica Particular de Loja. (n.d.). Sistema de Información Académica Científica. Retrieved from https://sica.utpl.edu.ec

Vences, N., & Segura, R. (2011). El desarrollo de la World Wide Web en España: Una aproximación teórica desde sus orígenes hasta su transformación en un medio semántico. *Razón Y Palabra*. Retrieved from http://wwww.razonypalabra.org.mx/N/N75/varia_75/varia3parte/31_Avuin_V75.pdf

Villalobos, A. (2006). Grafos: herramienta informática para el aprendizaje y resolución de problemas reales de teoría de grafos. *X Congreso de Ingeniería de Organización*. Retrieved from http://adingor.es/congresos/web/articulo/detalle/a/908

Villazón-Terrazas, B., Vilches-Blázquez, L. M., Corcho, O., & Gómez-Pérez, A. (2011). Methodological Guidelines for Publishing Government Linked Data. *Linking Government Data.*, 27–49. http://doi.org/10.1007/978-1-4614-1767-5

W3C. (2004). OWL Web Ontology Language Overview. Retrieved March 3, 2015, from http://www.w3.org/TR/2004/REC-owl-features-20040210/

W3C. (2008). El W3C expone los datos en la Web con SPARQL. Retrieved March 2, 2015, from http://www.w3c.es/Prensa/2008/nota080115_sparql

W3C. (2010). Guía Breve de Web Semántica. Retrieved March 2, 2015, from http://www.w3c.es/Divulgacion/GuiasBreves/WebSemantica

W3C. (2011). Describing Linked Datasets with the VoID Vocabulary. Retrieved March 6, 2015, from http://www.w3.org/TR/void/

W3C -a. (n.d.). The Statistical Core Vocabulary (scovo). Retrieved March 5, 2015, from http://sw.joanneum.at/scovo/schema.html

W3C -a. (2012). Data Cube Vocabulary. Retrieved March 5, 2015, from http://www.w3.org/2011/gld/wiki/Data_Cube_Vocabulary#The_history_of_Data_Cube.2C_SDMX-RDF_and_SCOVO

W3C -b. (2012). OWL - Semantic Web Standards. Retrieved March 3, 2015, from http://www.w3.org/2001/sw/wiki/OWL

W3C -b. (2014). RDF 1.1 Turtle. Retrieved March 2, 2015, from http://www.w3.org/TR/turtle/

W3C -c. (2012). GLD Life cycle - Government Linked Data (GLD) Working Group

Wiki. Retrieved September 12, 2015, from http://www.w3.org/2011/gld/wiki/GLD_Life_cycle

W3C -c. (2014). RDF - Semantic Web Standards. Retrieved March 2, 2015, from http://www.w3.org/RDF/

W3C -d. (2014). The RDF Data Cube Vocabulary. Retrieved March 20, 2015, from http://www.w3.org/TR/vocab-data-cube/

GLOSARIO DE TÉRMINOS

En esta sección encontraremos un glosario de términos con algunas de las palabras utilizadas en este documento, con la finalidad de facilitar al lector su entendimiento y mejorar la comprensión de algunos conceptos.

A

Apache: Se refiere al software desarrollado por la Apache Software Foundation, organización que distribuye y desarrolla software libre.

API: Interfaz de programación de aplicaciones

C

Commons Lang: Librería Java que provee métodos para la manipulación de datos

Cuartil: Medidas de posición no centrales.

E

eISSN: Número Digital Internacional Normalizado de Publicaciones Seriadas

F

Framework: Se refiere a una estructura conceptual y tecnológica de soporte definido.

G

Grafo: Estructura de datos compuesta por un conjunto de vértices o nodos unidos mediante aristas o arcos.

H

HTML: Siglas de HyperText Markup Language, es un lenguaje de marcado utilizado normalmente para estructurar texto en el desarrollo de sitios web y presentarlo en forma de hipertexto.

I

Indicadores: Dato que permite conocer la valoración de una característica específica y su intensidad.

Índice H: Medición de la calidad profesional que considera la cantidad de citas que recién los artículos científicos.

IPP: Indicador que permite medir la proporción de citas en un año (Y) para artículos académicos publicados en los tres años anteriores (Y-1, Y-2, Y-3), dividido por el número de artículos académicos publicados en los mismos año (Y-1, Y-2, Y-3).

ISSN: Número Internacional Normalizado de Publicaciones Seriadas

J

Java: Lenguaje de programación concurrente, orientado a objetos y capaz de ejecutarse en diferentes tipos de arquitecturas, sistemas computacionales y dispositivos.

JavaScript: Lenguaje de programación interpretado que se ejecuta del lado del cliente en aplicaciones web y sosamente presido a Java.

Jena: Es un framework Java desarrollado por HP Labs en el año 2000 y que provee un API para construir aplicaciones basadas en Ontologías.

JSON: Siglas de JavaScript Object Notation, es un formato utilizado para el intercambio de datos.

L

Linked Data: Revisar el apartado 1.5

M

Metadato: Formato utilizado para organizar información referente a un archivo dentro de él.

N

NeOn: Metodología utilizada para la construcción de redes ontológicas

NetBeans: IDE utilizado principalmente para el desarrollo integrado en java.

Notación 3: También conocido como N3 es una forma abreviada de socialización no-XML de modelos en RDF

O

Ontología: Revisar el apartado 1.2.1

Open Refine: Herramienta libre y de código abierta utilizada para trabajar y depurar gran cantidad datos.

Owl: Revisar el apartado 1.2.4

P

Poi: Librería Java que provee métodos para trabajar con archivos Excel.

PostgresSQL: Gestor de base de datos relacional y orientado a objetos.

R

RDF: Revisar el apartado 1.2.3

RDF/XML: Sintaxis utilizada para expresar un grafo RDF como un documento XML

Recurso Ontológico: Son los diferentes vocabularios y ontologías existentes en la web.

Revista Científica: Publicación serializada y de carácter científico en la cual se publican investigación, descubrimientos y entre otros

S

Scrapy: Es un framework de Phyton utilizado para rastrear sitios web y extraer datos estructurados de sus páginas.

SJR: Es un valor que permite medir el prestigio científico de fuentes académicas. SJR asigna puntuaciones relativas a todas las fuentes en una red de citas.

SNIP: Es un indicador que mide el impacto de la citación del contexto de una fuente.

SPARQL: Sigas de Protocol and RDF Query Language, es un lenguaje estandarizado para la consulta de grafos RDF.

SQL: Son las siglas de Structured Query Language, es un lenguaje declarativo de acceso a bases de datos relacionales.

Store RDF: Revisar el apartado 1.4

T

Tripleta RDF: Se refiere a la unión de tres recursos denominados Sujeto, Objeto y Predicado.

Turtle: Formato para serializar RDF.

U

URI: Revisar el apartado 1.2.2

URL: Siglas de Uniform Resource Locator, significa Identificador de Recursos Uniforme.

V

Vocabularios: Conjunto de términos con un significado conocido que se utilizan para describir recursos.

W

Web Semántica: También conocida como web 3.0 es una versión de la web en la cual se busca por coincidencia de conceptos y no por coincidencia de palabras.

W3C: Siglas de World Wide Web Consortium, es una comunidad internacional que desarrolla estándares que aseguran el crecimiento de la Web a largo plazo.

X

XML: Es un conjunto de tecnologías utilizadas para el manejo de datos e información.

XMLBeans: Librería Java utilizada para la gestión de archivos XML.

ANEXOS

ANEXO 1: Tabla comparativa de las propiedades de los sistemas Cassandra vs. MongoDB vs. Virtuoso

Nombre	Cassandra	MongoDB	Virtuoso
Descripción	Almacén de datos basada en ideas de BigTable y DynamoDB	Almacén de datos basado en documentos	Servidor de datos multi-modelo
Sitio web	cassandra.apache.org	www.mongodb.org	virtuoso.openlinksw.com
Documentación técnica	www.datastax.com/docs	docs.mongodb.org/manual	docs.openlinksw.com/virtuoso
Desarrollado por	Apache Software Foundation	MongoDB, Inc	OpenLink Software
Versión inicial	2008	2009	1998
Licencia	Código Abierto	Código Abierto	Código Abierto
Lenguaje de implementación	Java	C ++	C
Sistemas operativos de servidor	- BSD - Linux - OS X - Windows	- Linux - OS X - Solaris - Windows	- AIX - FreeBSD - HP-UX - Linux - OS X - Solaris - Windows
Modelo de base de datos	Almacén de columna	Almacén de documentos	XML nativo DBMS, relacional y almacén RDF
Esquema de datos	Sin esquema	Sin esquema	Si
SQL	No	No	Si
API y otros métodos de acceso	Protocolo propietario	Protocolo propietario mediante JSON	- OLE DB - ADO.NET - JDBC - ODBC
Lenguajes de programación	- C # - C ++ - Clojure - Erlang - Go - Haskell - Java - JavaScript - Perl - PHP - Python - Ruby	- Actionscript - C - C# - C++ - Clojure - ColdFusion - D - Dart - Delphi - Erlang - Go - Groovy - Haskell	- .Net - C - # C - C ++ - Java - JavaScript - Perl - PHP - Python - Ruby - Visual Basic

	- Scala	- Java - JavaScript - Lisp - Lua - MatLab - Perl - PHP - PowerShell - Prolog - Python - Ruby - Scala - Smalltalk	
Scripts de servidor	No	JavaScript	Si
Disparadores	Si	No	Si
Métodos de particionamiento	Sharding	Sharding	Si
Métodos de replicación	Factor de replicación seleccionable	Replicación maestro-esclavo	- Replicación maestro-maestro - Replicación maestro-esclavo
Conceptos de consistencia	- Consistencia eventual - Consistencia Inmediata	- Consistencia eventual - Consistencia Inmediata	Consistencia Inmediata
Claves externas	No	No	Si
Conceptos de transacción	No	No	ACID
Concurrencia	Si	Si	Si
Durabilidad	Si	Si	Si
Usuario	Derechos de acceso para los usuarios pueden ser definidos por objeto	Derechos de acceso para usuarios y roles	Derechos de acceso de grano fino de acuerdo con SQL estándar

Nota: Recuperado de DB-Engines.

ANEXO 2: Configuración de memoria en Open Refine.

Dependiendo del sistema operativo en el que se trabaje, los pasos son los siguientes:

Windows:

Para realizar la ampliación de memoria RAM se debe acceder al archivo google-refine.l4j.ini que se encuentra dentro de la ruta de instalación del programa. Una vez abierto este archivo se debe ampliar el número referente al tamaño máximo de memoria. Este valor se encuentra en megabytes. A continuación en la Figura № 21 encontramos un ejemplo, donde se encuentra subrayada la línea que debe ser editada.

```
# Launch4j runtime config
# initial memory heap size
-Xms256M
# max memory memory heap size
-Xmx4096M
# Use system defined HTTP proxies
-Djava.net.useSystemProxies=true
#-XX:+UseLargePages
#-Dsomevar="%SOMEVAR%"
```

Figura 54 Ejemplo archivo de configuracion Windows
Fuente: El Autor
Elaborado por: El Autor

Mac:

En caso de utilizar una computadora Mac se debe ir a la carpeta de aplicaciones, realizar clic en el icono de Open Refine para mostrar las opciones, seleccionar "Mostar contenido del paquete" y luego se debe acceder a la carpeta Contents. Ya en este nivel se edita el archivo "Info.plist", centrándonos en la sección referente a las opciones de java, donde se realiza el incremento del valor referente al máximo tamaño de memoria RAM que puede usar, dicho campo se encuentra especificado en megabytes. A continuación en la Figura 22 se muestra un ejemplo en el que se encuentra subrayado el campo que debe ser cambiado.

```
<key>JVMMainClassName</key>
<string>com/google/refine/Refine</string>
<key>JVMOptions</key>
<array>
<string>-Xms256M</string>
<string>-Xmx6144M</string>
<string>-Drefine.version=2.6-beta.1</string>
<string>-
Drefine.webapp=$APP_ROOT/Contents/Resource/webapp</string>
```

Figura 55 Ejemplo archivo de configuración Mac.
Fuente: El Autor
Elaborado por: El Autor

Linux:

Finalmente en caso de tener que realizar la ampliación de memoria RAM en un sistema operativo Linux, se debe realizar la edición del archivo llamado “refine.ini” en este se especificará el nuevo tamaño de memoria asignado. El archivo se encuentra ubicado en la ruta de instalación del programa y un ejemplo del mismo se muestra en la Figura 23, donde la parte a ser editada durante la ampliación de memoria se encuentra subrayada.

```
#REFINE_PORT=3333
#REFINE_HOST=127.0.0.1
#REFINE_WEBAPP=main\webapp
#REFINE_MEMORY=1024M
# Some sample configurations. These have no defaults.
#ANT_HOME=C:\grefine\tools\apache-ant-1.8.1
#JAVA_HOME=C:\Program Files\Java\jdk1.6.0_25
#JAVA_OPTIONS=-XX:+UseParallelGC -verbose:gc -
Drefine.headless=true
```

Figura 56 Ejemplo archivo de configuración Linux
Fuente: El Autor
Elaborado por: El Autor

ANEXO 3: Scripts de Open Refine (proceso se depuración)

Debido a lo extenso de los scripts su inclusión dentro del documento resulta muy poco eficiente, entorpeciendo la búsqueda de información y extendido de manera excesiva esta sección, sin embargo si se desea revisar a profundidad los scripts se los puede encontrar en el repositorio Git de la Universidad Técnica Particular de Loja o accediendo mediante el enlace:

https://git.taw.utpl.edu.ec/vmjaramillo1/PublicacionEstadisticaEnLinkedData/tree/master/Proceso%20de%20Depuracion

ANEXO 4: Detalles técnicos de la especificación de la ontología

En esta sección se encuentran los detalles técnicos completos referentes a la especificación de la ontología, realizados siguiendo la metodología NeOn y utilizando las plantillas que esta ofrece.

1) Propósito

Poder representar el estado histórico y actual de las diferentes revistas científicas, permitiendo así que el público en general y personal académico (docentes, investigadores y estudiantes) tengan un referente al momento de seleccionar las revistas en las cuales publicar sus descubrimientos y realizar sus consultas e investigaciones

2) Alcance

La presente ontología se limitará a modelar revistas científicas considerando su tópico, publicita e indicadores clasificados por año.

3) Lenguaje de implementación

Para la implementación de la ontología se utilizara RDF como lenguaje formal.

4) Usuarios finales Destinados

Usuario 1: Investigador que desea publicar su trabajo y busca una referencia de cual revista seria su mejor opción, esto debido a que la elección podría influir en el impacto de su publicación.

Usuario 2: Persona que desea conocer el estado actual de una revista para realizar una investigación basándose en sus artículos, elegir una de mayor o menor posición en el ranking podría afectar de manera significativa la relevancia o calidad del estudio.

Usuario 3: Desde el punto de vista de las revistas serviría a los editores para saber el estado actual de su revista, sería una referencia que indica el nivel de las publicaciones que realiza y así saber si se deben realizar cambios o no en las políticas de aceptación de las revistas

Usuario 4: Para un investigador inexperto serviría para elegir una revista que sea menos exigente con las normas de las investigaciones que publica y así tener mayores posibilidades de que su trabajo sea aceptado por los editores.

5) Usos destinados

Búsqueda de estado de una revista

Búsqueda de histórico de la revista

Categorías en las cuales se encuentra clasificada una revista

6) Grupos de Cuestiones de Competencia

a) Requerimientos no Funcionales

La ontología debe ser compatible con el vocabulario Data Cube y estar preparada para trabajar con observaciones.
La ontología debe adaptarse a la información de cada revista soportando que en algunos casos no toda la información se encuentre disponible.

b) Requerimientos Funcionales: Agrupados por preguntas de competencia

Tabla 7 Requerimientos Ontológicos.

Grupo	N°	Pregunta	Respuesta
PCG1 Revista	PC1	¿Cuál es el título de la revista?	PLoS Biology
	PC2	¿Cuál es el ISSN de la revista?	15449173
	PC3	¿Cuál es el eISSN de la revista?	15457885
	PC4	¿Cuál es el estado de la revista?	Active
	PC5	¿De qué URL se obtuvo información de la revista?	http://www.scimagojr.com/journalsearch.php?q=12977&tip=sid&clean=0
	PC6	¿De qué tipo es la revista?	Journal
	PC7	¿La revista se ha encontrado inactiva en algún momento desde 2006?	No
	PC8	¿Cuál es porcentaje de revistas inactivas es el mundo?	7.4
	PC9	¿Desde qué año se encuentra en circulación la revista?	2003
PCG2 Editorial	PC10	¿Cuál es el nombre de la editorial?	Public Library of Science
	PC11	¿De qué grupo es miembro la editorial?	Public Library of Science
	PC12	¿En qué país se ubica de la editorial?	United States

	PC13	¿Cuál es la principal editorial de revistas científicas los según los datos recolectados?	Elsevier
	PC14	¿Su editorial se encuentra entre las primeras 5 editoriales con más publicaciones?	No
PCG3 Indicador	PC15	¿A qué año corresponde el indicador?	2013
	PC16	¿Cuál es su H index de la revista?	149
	PC17	¿Cuál es el total de referencias de la revista?	12886
	PC18	¿Cuál es el total de citas de la revista?	6828
	PC19	¿Cuantos documentos citados posee la revista?	788
	PC20	¿Cuantas citas por documento posee de la revista?	8.08
	PC21	¿Cuantas referencias por documento posee de la revista?	44.74
	PC22	¿Cuál es el valor SJR de la revista?	5,372
	PC23	¿Cuál es el valor SNIP de la revista?	2,038
	PC24	¿Cuál es el valor IPP de la revista?	8,683
	PC25	¿Cuál es el total de documentos de la revista ese año?	288
	PC26	¿Cuál es el total de documentos de la revista en los 3 últimos años?	829
	PC27	¿La calidad de la revista ha mejorado, empeorado o se ha mantenido con el pasar del tiempo?	Se ha mantenido
	PC28	¿Cuál es el mayor número de revistas existentes en un país según los datos consultados?	5897
	PC29	¿Cuál es el país con mayor cantidad de revistas científicas según los datos recolectados?	United States
PCG4 Categoría	PC30	¿Cuántas categorías posee la revista?	4
	PC31	¿A qué año corresponden las categorías?	2013
	PC32	¿Mejor valor de cuartil que ha obtenido?	1
	PC33	¿Cuál es la categoría con mejor valoración?	Todas tienen la misma valoración
	PC34	¿La revista ha incrementado sus categorías investigativas en los últimos años?	No

Fuente: El autor
Elaborado por: El Autor

7) Pre-Glosario de términos

a) Términos de preguntas de competencia

Tabla 8 Preguntas competencia.

Termino	Frecuencia
➢ Revista	19
o Titulo	1
o ISSN	1
o eISSN	1
o Estado	1
o URL	1
o Tipo	1
o Cobertura	1
➢ Editorial	4
o Nombre	2
o Grupo miembro	1
o Ubicación	1
➢ Categoría	3
o Quartil	1
➢ Indicador	2
o Año	2
o H index	1
o Total Referencias	1
o Total de Citas	1
o Total Documentos citados	1
o Citas por Documento	1
o Referencias por Documento.	1
o SJR	1
o SNIP	1
o IPP	1
o Total Documentos del año	1
o Total Documentos de los últimos 3 años	1

Fuente: El Autor
Elaborado por: El Autor

b) Términos de respuestas

Tabla 9 Términos de respuesta

Termino	Frecuencia
➢ Active	1
➢ Journal	1

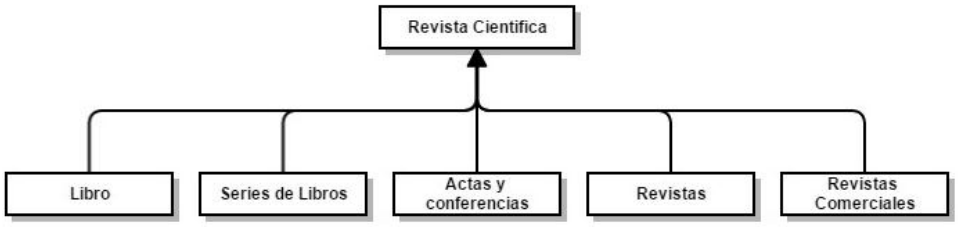
Fuente: El Autor
Elaborado por: El Autor

c) Objetos incluidos en las cuestiones de competencia y en sus respuestas.

United States, Public Library of Science

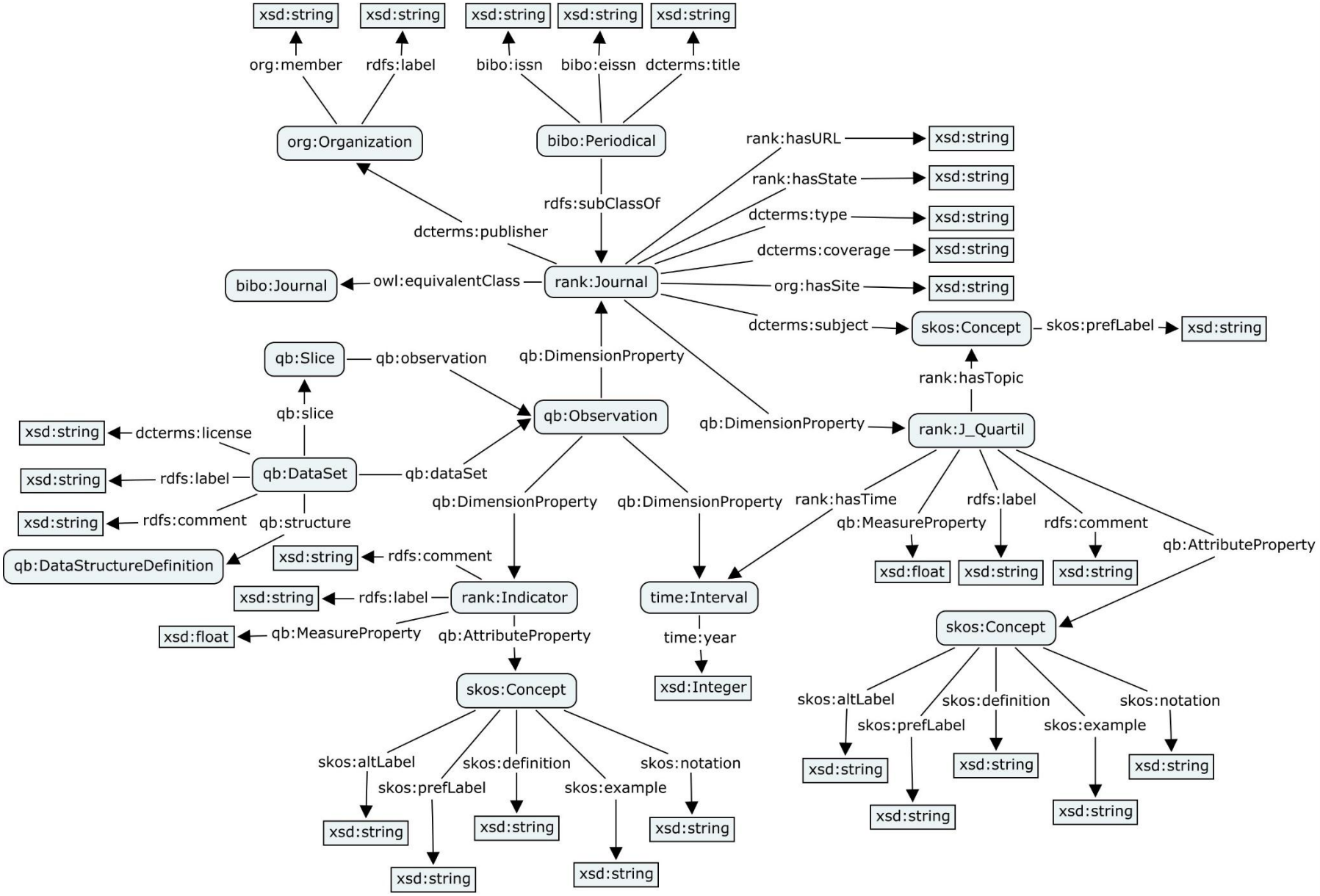
ANEXO 5: Plantilla NeOn para elaboración de taxonomías

Tabla 10 Plantilla NeOn para Elaboración de Taxonomías

<table>
<tr><th colspan="2">Información general</th></tr>
<tr><td>Nombre</td><td>Taxonomía de clasificación de revistas</td></tr>
<tr><td>Identificador</td><td>Tipo-Revista-TX01</td></tr>
<tr><td>Tipo de componente</td><td>Modelo para la reingeniería de recursos no Ontológico</td></tr>
<tr><th colspan="2">Caso de Uso</th></tr>
<tr><td colspan="2">Se requiere un esquema de clasificación de revistas científicas para diseñar una taxonomía, en el caso que se desea clasificar una revista científica dentro de un conjunto específico de publicaciones previamente establecidas a las cuales se pueda corresponder.</td></tr>
<tr><th colspan="2">Modelo para la reingeniería de recursos no Ontológicos</th></tr>
<tr><th colspan="2">Recurso a ser Reusado</th></tr>
<tr><td colspan="2">La clasificación de revistas científicas se realizan considerando aspectos como el número de volúmenes, tiempo de publicación, etc. En este caso se tomara de referencia la clasificación utilizada por SCImago Journal Metrics en la cual se ofrecen las clasificaciones: Libro, Serie de Libros, Actas y conferencias, Revistas, Revistas Comerciales.</td></tr>
<tr><th colspan="2">Representación gráfica</th></tr>
<tr><td colspan="2">
<table>
<tr><th>Id</th><th>Nombre de Tipo Revista</th><th>Padre</th></tr>
<tr><td>1</td><td>Revista Científica</td><td>Null</td></tr>
<tr><td>2</td><td>Libro</td><td>1</td></tr>
<tr><td>3</td><td>Serie de Libros</td><td>1</td></tr>
<tr><td>4</td><td>Actas y conferencias</td><td>1</td></tr>
<tr><td>5</td><td>Revistas</td><td>1</td></tr>
<tr><td>6</td><td>Revistas Comerciales</td><td>1</td></tr>
</table>
</td></tr>
<tr><th colspan="2">Diseño ontológico</th></tr>
<tr><td colspan="2">Cada una de las clasificaciones de revistas corresponden a un tipo específico de publicación</td></tr>
<tr><th colspan="2">Representación grafica</th></tr>
<tr><td colspan="2"> 
</td></tr>
</table>

Fuente: El Autor
Elaborado por: El Autor

ANEXO 6: Modelo completo del vocabulario Rank



ANEXO 7: Código de la generación de tripletas

Debido a que la implementación del código resulta ser sumamente extensa para poder colocarlo en un apartado, se ha decidido almacenarlo en un repositorio git junto con las demás especificaciones del proyecto. Para su consulta se recomienda visitar el enlace https://git.taw.utpl.edu.ec/vmjaramillo1/PublicacionEstadisticaEnLinkedData/tree/master/Generacion_Tripletas

ANEXO 8: Implementación del sitio web

Debido a la extensión del código utilizado para el desarrollo del sitio web se consideró ineficiente su colocación dentro de este documento, sin embargo en caso de que se dese revisar de primera mano la aplicación el código fuente puede ser consultado en el enlace:
https://git.taw.utpl.edu.ec/vmjaramillo1/PublicacionEstadisticaEnLinkedData/tree/master/Sitio%20web/journalv3

ANEXO 9: Guía Procedimental

Introducción

En esta sección se detallan las pautas principales a seguirse durante el proceso de publicación de datos estadísticos, esto con la finalidad de servir de herramienta y orientar a las personas con poco o ningún conocimiento en esta área.

Para poder entender completamente algunos de los temas que se explican en esta guía, resulta fundamental que la persona tenga nociones básicas de las tecnologías empleadas en la web semántica, de no ser así se recomienda la lectura del Capítulo I donde se explican estos temas.

Finalmente cabe destacar que en esta guía se explican solo los puntos más esenciales del proceso, por cuanto se recomienda la lectura de los Capítulos III, IV y V, donde se encuentra la aplicación y detalles técnicos descritos más a fondo.

Los contenidos que se explican en esta guía se encuentran dividas en siete secciones, cada una de las cuales comprende diferentes aspectos de la publicación de datos y las actividades que deben ser realizadas.

Contenido

Planificación del proyecto.

La planificación es la primera etapa a ser considerada dentro de la publicación de datos estadísticos, esta resulta de vital importancia para el desarrollo de las actividades, permitiendo llevar un control en todo momento del estado actual del proyecto.

Las principales actividades que se debe comprender en esta etapa son:

Objetivos.- Se deben definir los objetivos que se pretenden conseguir con el proyecto de publicación de datos estadísticos.

Alcance.- Se delimita el trabajo que se realizara en el proyecto.

Ciclo de vida de linked data.- Se selecciona un ciclo de vida de linked Data que se ajuste a las necesidades del proyecto y ayude a estructurar las diferentes actividades que comprenden la publicación de datos estadísticos.

Recursos.- Se gestionan los recursos disponibles para el proyecto considerándose tiempo, equipos y personal humano a disposición.

Definición de recursos.

Esta etapa comprende la definición de los URIs bajo los cuales se publicara la información. Para esto se debe considerar que en muchas ocasiones se requerirán por lo menos dos estructuras de URIs diferentes para ser utilizadas dentro del proyecto. (Revisar Capítulo III apartado 3.3)

La primera comprenderá a los datos que serán publicados, es decir los diferentes recursos que serán creados utilizando los vocabularios, mientras que la segunda

estructura se refiere al caso en el cual el proyecto de publicación de datos estadísticos comprenda la creación de un vocabulario. Para la definición de la estructura de los URIs se recomienda considerar lo siguiente:

- De ser posible utilizar el concepto de URIs Cool para obtener una estructura entendible y fácil de manejar.
- Si el proyecto se está desarrollando en una organización que ya posea una estructura de URIs, se recomienda su utilización base de las nuevas URIs para mantener el dominio de la información y evitar la creación de estructuras innecesarias.

Preparación de los datos.

Obtención de datos.

Esta etapa se refiere a la recolección de la información que se empleara durante todo el proyecto. (Revisar Capítulo III apartado 3.2) Siendo esta una etapa crucial se recomienda tener en cuenta lo siguiente:

- En caso de no ser los dueños de la información considerar los derechos de uso de los datos para no infringir leyes de derechos de autor.
- Intentar obtener la información de distintas fuentes con la finalidad de evitar errores e incrementar la fiabilidad de los datos
- En caso que la información proporcionada por las empresas no sea suficiente considerar otras formas de obtener dados como pueden ser: realizar encuestas o la efectuar Scrapy en las distintas páginas web siempre y cuando con esto no se infrinjan leyes de derechos de autor.
- Si la información se encuentra disponible en distintos formatos, se recomienda la utilización de un estándar para así poder evitar errores de transformación de datos y facilitar el trabajo.
- En caso que la información provenga de distintos orígenes se recomienda la utilización de diagramas para mostrar la estructura y uso final que se les dará a los datos.

Depuración de datos.

En esta etapa se comienza con el refinamiento, depuración y estructuración de los datos. (Revisar Capítulo III apartado 3.3) Las actividades que se comprende son:

Depuración de Datos incensarios.- Se elimina información duplicada, redundante o que no es de interés para el proyecto.

Conciliación de la Información.- En caso que la información se encuentre dispersa en múltiples archivos o repositorios se la unifica dentro de uno solo para poder trabarla mejor y así posteriormente poder estructurarla.

Estructuración de los datos.- Se da una estructura a los datos que facilite su procesamiento, se recomienda elegir entre una estructura orientada a filas o columnas. Se debe considerar muy bien la elección y elegir la que mejor se ajuste al proyecto pues esto puede afectar significativamente en etapas posteriores.

Validación.- Se valida que no haya existido perdida de datos o errores que afecten la veracidad de la información. Para esto se comparan los datos ya estructurada con los datos originales.

Diseño y modelado.

El diseño y modelado comprende todos los aspectos referentes a la forma en la cual se utilizaran los distintos vocabularios para representar los datos de la etapa anterior. Se debe tener especial cuidado en esta etapa puesto que varias de las actividades son sumamente complejas. (Revisar Capítulo IV sección 4.3) Las principales actividades a realizarse son:

Investigación de Vocabularios.- Se debe investigar profundamente los distintos vocabularios que permiten el modelamiento de los datos.

Diseño del Vocabulario.- En caso de que se necesite crear un vocabulario para modelar algún aspecto específico de los datos se debe pasar por una etapa de diseño, que mediante la utilización de una metodología permita determinar las clases y propiedades necesarias a ser creadas en el nuevo vocabulario.

Validación.- Luego de creado el nuevo vocabulario este debe ser validado, con la finalidad de encontrar y corregir posibles errores en el caso de que fuese necesario.

Transformación de datos.

Lo siguiente a realizarse es efectuar la transformación de los datos utilizando los vocabularios previamente seleccionados. Sin embargo la etapa de transformación debe ser trabajada con mucho cuidado para evitar la realización de trabajos innecesarios. (Revisar Capítulo IV apartado 4.4) Se debe considerar lo siguiente:

Análisis de las Herramientas.- Se debe efectuar el análisis y elección de una herramienta que permita la transformación de los datos, sin embargo esta elección no debe ser tomada a la ligera, lo mejor es realizar pruebas con pequeñas cantidades de datos para poder elegir el que mejor se adapte a las necesidades del proyecto.

Conversión de los datos.- Efectuar la conversión de los datos y posteriormente analizar los datos convertidos en RDF para asegurar su integridad y estructura.

Almacenamiento de la información.

Los datos en formato RDF deben ser almacenados para su utilización, este es el aspecto trabajado en esta etapa. (Revisar Capítulo V apartado 5.2) Para ello se debe:

Selección de Stores RDF.- Se debe analizar las posibles opciones de Stores y elegir el que sea más acorde con las necesidades

Almacenamiento de la Información RDF.- Aun cuando parezca una actividad muy sencilla primero se debe asegurar de que el store RDF seleccionado posea las configuraciones necesarias puesto de no ser así, en muchos casos si los archivos RDF son muy encontrara problemas durante el almacenamiento y recuperación de la información.

Presentación de datos.

Finalmente la última etapa a ser trabajada es la presentación de los datos, para esto existen muchas opciones como la utilización de herramientas especializadas o el desarrollo de sitios web que presenten los datos. En este caso vamos a profundizar en la segunda opción debido a ser la que se empleó en el proyecto. (Revisar Capítulo V apartados 5.3, 5.4, 5.5 y 5.6) Para esto se debe considerar lo siguiente:

Selección de arquitectura.- Se debe seleccionar la arquitectura con la cual él se desarrollara el sitio web con la finalidad de poder estructurar mejor el trabajo e identificar los componentes a ser desarrollados.

Desarrollo de consultas.- Este aspecto se lo ha considerado por separado debido a que es un punto crítico en esta etapa, las consultas desarrolladas deben ser probadas por separado para asegurar su integridad y tiempos de respuesta puesto que un error en las mismas afectaría de manera muy negativa el rendimiento del sitio web.

Construcción del sitio web.- A viendo analizado todos los aspecto anteriores se comienza con el desarrollo del sitio web considerando aspectos tanto aspectos funcionales como estéticos para obtener los mejores resultados. Se recomienda la utilización de gráficos para la presentación de la información y así facilitar el entendimiento de la misma por parte del usuario.

Análisis de Rendimiento.- Se deben efectuar pruebas para comprobar el funcionamiento del sitio web desarrollado con la finalidad de comprobar tiempos de respuesta, integridad de la información presentada y corrección de errores en caso de ser necesario.