PONTIFICIA UNIVERSIDAD CATÓLICA DEL PERÚ

FACULTAD DE CIENCIAS E INGENIERÍA

IMPLEMENTACIÓN DE UN BUSCADOR SEMÁNTICO DE DOCUMENTOS EN EL DOMINIO DE LA LINGÜÍSTICA

Tesis para optar el Título de **Ingeniero Informático**, que presenta el bachiller:

Diego Andrés Malpartida Valverde

ASESOR: Dr. Héctor Andrés Melgar Sasieta

Lima, junio de 2017

AGRADECIMIENTOS

Al Dr. Andrés Melgar, por compartir el conocimiento obtenido mediante su trayectoria profesional y brindar su constante apoyo durante todo el proyecto.

El autor reconoce también el apoyo del "Programa Nacional de Innovación para la Competitividad y Productividad", Perú, en el marco del contrato 124-PNICP-PIAP-2015.

RESUMEN

La World Wide Web (WWW) ha mejorado considerablemente el acceso a la información digital. La búsqueda y navegación en la Web se han convertido en parte de nuestras vidas diarias, siendo los motores de búsquedas y herramientas de navegación Web un estándar que ha cambiado la forma en la que buscamos e interactuamos con la información. Sin embargo, la Web como la conocemos hoy está diseñada para que la información contenida en las páginas o documentos sea entendible por las personas y no por las computadoras. Es decir, las computadoras no poseen de una manera para procesar la semántica o significado de la información.

Esto ocasiona que solo se puedan realizar búsquedas sintácticas de la información, en lugar de búsquedas semánticas. Las búsquedas sintácticas consisten en la recuperación de aquellos documentos cuyo contenido posee las palabras o frases ingresadas por el usuario en la consulta. Se basan en la similitud de cadenas de caracteres (las ingresadas por el usuario y las que contiene el documento). El problema con las búsquedas sintácticas es que se limitan a esta coincidencia de palabras y no consideran el significado de la información, lo que ha sido demostrado que genera imprecisión (mucha información irrelevante) en los resultados.

En este contexto, en el Departamento de Humanidades de la universidad existe la necesidad de recuperar información de aproximadamente 2000 documentos lingüísticos para fines académicos. Una búsqueda convencional o sintáctica no sería una buena solución, ya que como se mencionó anteriormente retorna mucha información irrelevante. Entonces, se puede definir el problema central del proyecto como la dificultad para obtener información relevante de documentos en el dominio de la lingüística.

Como alternativa de solución, el presente proyecto de fin de carrera implementa un buscador que emplee los conceptos y principios de la Web Semántica. Este tipo de buscador se basa en el análisis semántico de la consulta ingresada por el usuario y del contenido de los documentos, recuperando aquellos cuya representación semántica coincide con la de la consulta. A diferencia de la búsqueda sintáctica, este enfoque analiza el significado de las palabras o frases y no solo su representación sintáctica. El beneficio de las búsquedas semánticas es que permiten alcanzar una mayor precisión en los resultados obtenidos; es decir, brindan resultados de mayor relevancia para el usuario.

FACULTAD DE
CIENCIAS E INGENIERÍA
ESPECIALIDAD DE
INGENIERÍA INFORMÁTICA

PONTIFICIA
UNIVERSIDAD CATÓLICA
DEL PERÚ

TEMA DE TESIS PARA OPTAR EL TÍTULO DE INGENIERO INFORMÁTICO

TÍTULO: **Implementación de un buscador semántico de documentos en el dominio de la lingüística**

ÁREA: Ciencias de la Computación

ASESOR: Dr. Héctor Andrés MELGAR SASIETA

ALUMNO: Diego Andrés MALPARTIDA VALVERDE

CÓDIGO: 20100952

TEMA N°: # 650

PONTIFICIA UNIVERSIDAD CATÓLICA DEL PERÚ
Facultad de Ciencias e Ingeniería
20 MAY. 2016
MIGUEL MEJIA PUENTE
DECANO

FECHA: San Miguel, 15 de octubre de 2016

DESCRIPCIÓN

La World Wide Web (WWW) ha mejorado considerablemente el acceso a la información digital. La búsqueda y navegación en la Web se han convertido en parte de nuestras vidas diarias, siendo los motores de búsquedas y herramientas de navegación Web un estándar que ha cambiado la forma en la que buscamos e interactuamos con la información. Sin embargo, la Web como la conocemos hoy está diseñada para que la información contenida en las páginas o documentos sea entendible por las personas y no por las computadoras. Es decir, las computadoras no poseen de una manera para procesar la semántica o significado de la información.

Esto ocasiona que solo se puedan realizar búsquedas sintácticas de la información, en lugar de búsquedas semánticas. Las búsquedas sintácticas consisten en la recuperación de aquellos documentos cuyo contenido posee las palabras o frases ingresadas por el usuario en la consulta. Se basan en la similitud de cadenas de caracteres (las ingresadas por el usuario y las que contiene el documento). El problema con las búsquedas sintácticas es que se limitan a esta coincidencia de palabras y no consideran el significado de la información, lo que ha sido demostrado que genera imprecisión (mucha información irrelevante) en los resultados.

En este contexto, en el Departamento de Humanidades de la universidad existe la necesidad de recuperar información de aproximadamente 2000 documentos

Av. Universitaria 1801
San Miguel, Lima – Perú

Apartado Postal 1761
Lima 100 – Perú

Teléfono:
(511) 626 2000 Anexo 4801

lingüísticos para fines académicos. Una búsqueda convencional o sintáctica no sería una buena solución, ya que como se mencionó anteriormente retorna mucha información irrelevante. Entonces, se puede definir el problema central del proyecto como la dificultad para obtener información relevante de documentos en el dominio de la lingüística.

Como alternativa de solución, el presente proyecto de fin de carrera implementará un buscador que emplee los conceptos y principios de la Web Semántica. Este tipo de buscador se basa en el análisis semántico de la consulta ingresada por el usuario y del contenido de los documentos, recuperando aquellos cuya representación semántica coincide con la de la consulta. A diferencia de la búsqueda sintáctica, este enfoque analiza el significado de las palabras o frases y no solo su representación sintáctica (cadenas de caracteres). El beneficio de las búsquedas semánticas es que permiten alcanzar una mayor precisión en los resultados obtenidos; es decir, brindan resultados de mayor relevancia para el usuario.

OBJETIVO GENERAL

Implementar un buscador semántico de documentos digitales en el dominio de la lingüística.

OBJETIVOS ESPECÍFICOS

Los objetivos específicos son:

OE1. Modelar el dominio de la lingüística empleando tecnologías para representar y codificar conocimiento.

OE2. Desarrollar una herramienta de software que permita anotar semánticamente y de forma manual la información de los documentos con elementos del dominio de la lingüística.

OE3. Desarrollar una aplicación en el ámbito de la Web Semántica que permita la recuperación de documentos en el dominio de la lingüística.

ALCANCE

El presente proyecto de fin de carrera implementará una aplicación en el ámbito de la Web Semántica para la recuperación de documentos en el dominio de la lingüística.

FACULTAD DE
CIENCIAS E INGENIERÍA
ESPECIALIDAD DE
INGENIERÍA INFORMÁTICA

PONTIFICIA
UNIVERSIDAD CATÓLICA
DEL PERÚ

Este proyecto pertenece al área de las Ciencias de la Computación, empleando conceptos como ontología, anotación semántica y recuperación de información.

Se desarrollará una ontología que represente el conocimiento en el dominio de la lingüística. Dado que el dominio de la lingüística es bastante amplio, no es factible desarrollar una ontología que represente todo este dominio. Con apoyo del profesor de lingüística ayudante en el proyecto, se analizaron los documentos y se determinó que los temas y conceptos que tratan cubren lo que es lingüística descriptiva. Por lo tanto, se definió que el alcance del proyecto no sería todo el dominio de la lingüística, sino la lingüística descriptiva.

Luego del desarrollo de la ontología, se realizará la anotación semántica de los documentos, los cuales estarán almacenados en un repositorio digital. Para esto, se escogerán aproximadamente 50 documentos que servirán como data de prueba para la aplicación. Se definirá un proceso de anotación de los documentos, el cual consistirá en una anotación manual y externa. Los documentos serán anotados con conceptos de la ontología lingüística, no pudiendo incluir términos de otros dominios.

Finalmente, se desarrollará una aplicación que permita recuperar los documentos mediante búsquedas semánticas. Al recibir consultas por parte de los usuarios, esta aplicación explorará la ontología lingüística desarrollada y buscará anotaciones para así recuperar los documentos relevantes.

Máximo: 100 páginas

Av. Universitaria 1801
San Miguel, Lima – Perú

Apartado Postal 1761
Lima 100 – Perú

Teléfono:
(511) 626 2000 Anexo 4801

Tabla de contenido

Índice de Figuras

Índice de Tablas

1 DEFINICIÓN DEL PROBLEMA

1.1 PROBLEMÁTICA

La *World Wide Web* (WWW) ha mejorado considerablemente el acceso a la información digital (Wahlster, 2006). La búsqueda y navegación en la Web se han convertido en parte de nuestras vidas diarias, siendo los motores de búsquedas y herramientas de navegación Web un estándar que ha cambiado la forma en la que buscamos e interactuamos con la información (Levene, 2011). Sin embargo, la Web como la conocemos hoy está diseñada para que la información contenida en las páginas o documentos sea entendible por las personas y no por las computadoras. Es decir, las computadoras no poseen una manera de procesar la semántica o significado de la información (Berners-Lee, Hendler & Lassila, 2001).

Esto ocasiona que solo se puedan realizar búsquedas sintácticas, en lugar de búsquedas semánticas. Las búsquedas sintácticas consisten en la recuperación de aquellos documentos cuyo contenido posee las palabras o frases ingresadas por el usuario en la consulta (Giunchiglia, Kharkevich & Zaihrayeu, 2009). Se basan en la similitud de cadenas de caracteres (las ingresadas por el usuario y las que contiene el documento). El problema con las búsquedas sintácticas es que se limitan a esta coincidencia de palabras y no consideran el significado de la información, lo que ha sido demostrado que genera imprecisión (mucha información irrelevante) en los resultados. (Giunchiglia et al., 2009).

Ante estas limitaciones de la Web; Berners-Lee, Hendler y Lassila (2001) crearon la Web Semántica. La Web Semántica es una extensión de la web actual en la que el significado de la información de los documentos está estructurada de tal forma que las computadoras puedan procesar y actuar sobre ésta de forma útil (Berners-Lee et al., 2001). Las búsquedas en el ámbito de la Web Semántica consisten en búsquedas semánticas. Las búsquedas de este tipo se basan en el análisis semántico de la consulta ingresada por el usuario y del contenido de los documentos, recuperando aquellos cuya representación semántica coincide con la de la consulta (Giunchiglia et al., 2009). A diferencia de la búsqueda sintáctica, este enfoque analiza el significado

de las palabras o frases y no solo su representación sintáctica (cadenas de caracteres). Las búsquedas semánticas, por lo general, permiten alcanzar una mayor precisión en los resultados obtenidos (Giunchiglia et al., 2009).

En este contexto, en el Departamento de Humanidades de la PUCP existe la necesidad de recuperar información de aproximadamente 2000 documentos lingüísticos para fines académicos. Se tienen cinco tipos diferentes de documentos: artículos, diccionarios, gramáticas, tesis y libros. Una búsqueda convencional o sintáctica no sería una buena solución, ya que como se mencionó anteriormente retorna mucha información irrelevante. Por ejemplo, si se ingresa la consulta "sujeto" queriendo buscar aquellos documentos que traten sobre el concepto de sujeto del área de la gramática (subárea de la lingüística), esta consulta recuperará los documentos que usen la palabra "sujeto" en todos los contextos (como adjetivo, por ejemplo). De esta manera, el usuario tendría que revisar una gran cantidad de documentos con el fin de encontrar cuáles son en realidad de su interés. Esta tarea de revisar los documentos puede considerarse como una búsqueda manual posterior a la búsqueda realizada por el software, lo que significaría un doble esfuerzo. Esto genera problemas como pérdida de tiempo o de información relevante.

Con el fin de mejorar los resultados de las búsquedas, una alternativa de solución es realizar una búsqueda en el ámbito de la Web Semántica; es decir, una búsqueda semántica. En una búsqueda de este tipo, el usuario plantea una consulta y el buscador puede "entender" el significado de ésta. Para que el buscador pueda realizar esto, se necesita proveerle conocimiento del dominio correspondiente (en este caso, del dominio de la lingüística). Esto se puede realizar a través de las ontologías, las cuales permiten modelar un dominio de conocimiento a través de un conjunto de representaciones (Lim, Liu & Lee, 2011). Estas representaciones contienen clases, atributos o propiedades, y relaciones entre las clases (Lim, Liu & Lee, 2011). Con ayuda de las ontologías, el computador puede explorar los conceptos, atributos y relaciones entre conceptos del dominio que se está representando para así ofrecer resultados de mayor calidad.

Farrar y Langendoen (2003) desarrollaron una ontología en el dominio de la lingüística denominada GOLD. Sin embargo, esta ontología no es considerada para

este proyecto, ya que no es una ontología diseñada para búsquedas semánticas. La ontología GOLD fue desarrollada para solucionar problemas encontrados en proyectos de bases de datos tipológicas y de procesamiento de lenguaje natural (NLP) (Farrar & Langendoen, 2003).

Entonces, para el caso de este proyecto, se enfrenta el problema de que no se cuenta con un mecanismo que provea al buscador conocimiento sobre el dominio de la lingüística. Este dominio es bastante amplio ya que abarca diversas áreas de estudio, y no se tiene definido en una estructura adecuada los conceptos y temas que tratan estas áreas ni las relaciones entre éstos. Es decir, no se cuenta con una ontología lingüística para búsquedas semánticas. Sin tal ontología no es posible realizar esta búsqueda semántica que mejore los resultados obtenidos frente a un buscador convencional.

Otro problema es que los documentos no se encuentran estructurados adecuadamente para que puedan ser recuperados por un buscador en el ámbito de la Web Semántica. El Departamento de Humanidades de la universidad simplemente tiene almacenado estos documentos. No existe un mecanismo que ofrezca información acerca de lo que contienen los documentos; es decir, estos documentos no poseen metadatos asociados que describan el significado de su contenido. Sin una forma de reconocer el contenido, un buscador no podrá saber si el documento es de importancia para el usuario o no.

Para poder describir el significado de la información de los documentos, existe un método conocido como anotación semántica. La anotación semántica es un esquema de generación y uso de metadatos específicos sobre un dominio, que tiene como objetivo permitir nuevos métodos de acceso a la información y extender los existentes (Kiryakov et al., 2004). Este esquema permite realizar búsquedas semánticas en lugar de búsquedas sintácticas. Las anotaciones pertenecerán a un documento en particular y permiten relacionarlo con las clases y propiedades de la ontología, de manera que un buscador explorando esta ontología identifique por medio de inferencias una anotación semántica como resultado, para luego recuperar el documento asociado a dicha anotación.

Las anotaciones semánticas permiten la búsqueda de documentos por distintos criterios. Estas anotaciones permiten, con ayuda de las ontologías, realizar búsquedas por temas, por conceptos que tratan los documentos, o por cualquier otro tipo de significado que se pueda identificar en el contenido de estos documentos. De esta manera, se puede tener una búsqueda que profundice en el significado de la información, en lugar de búsquedas convencionales directas.

En conclusión, se puede definir el problema central como la dificultad para obtener información relevante de documentos en el dominio de la lingüística. Este problema se da debido a los mecanismos usados por las búsquedas convencionales (sintácticas). Como alternativa de solución, el presente proyecto de fin de carrera plantea implementar un buscador que emplee los conceptos y principios de la Web Semántica.

1.2 Objetivo general

Implementar un buscador semántico de documentos digitales en el dominio de la lingüística.

1.3 Objetivos específicos

O1: Modelar el dominio de la lingüística empleando tecnologías para representar y codificar conocimiento.

O2: Desarrollar una herramienta de software que permita anotar semánticamente y de forma manual la información de los documentos con elementos del dominio de la lingüística.

O3: Desarrollar una aplicación en el ámbito de la Web Semántica que permita la recuperación de documentos en el dominio de la lingüística.

1.4 Resultados esperados

Resultados del O1 (Modelar el dominio de la lingüística empleando tecnologías para representar y codificar conocimiento):

- Ontología que representa y provee conocimiento en el dominio de la lingüística.
- Pruebas de consistencia de la ontología.

Resultados del O2 (Desarrollar una herramienta de software que permita anotar semánticamente y de forma manual la información de los documentos con elementos del dominio de la lingüística):

- Definición del proceso manual de anotación semántica externa de los documentos.
- Estructura para la persistencia de las anotaciones semánticas.
- Aplicación que permite anotar manualmente los documentos con elementos de la ontología.

Resultados del O3 (Desarrollar una aplicación en el ámbito de la Web Semántica que permita la recuperación de documentos en el dominio de la lingüística):

- Indexación de los documentos mediante elementos de la ontología.
- Herramienta de software que permite realizar búsquedas basadas en texto previas a las búsquedas semánticas.
- Aplicación que permite realizar búsquedas semánticas de documentos en el dominio de la lingüística.
- Evaluación de los resultados obtenidos por las búsquedas semánticas usando precisión y *recall*.

1.5 Herramientas, métodos, metodologías y procedimientos

En esta sección se presentarán y describirán las herramientas, métodos, metodologías y procedimientos que se usarán en este proyecto de fin de carrera. La Tabla 1 muestra el mapeo de cada una de estas con los resultados esperados del proyecto.

Tabla 1: Mapeo de resultados esperados y herramientas.

Resultado esperado	**Herramientas**
1. Ontología que representa y provee conocimiento en el dominio de la lingüística.	OWL, Protégé, *Ontology Development 101*
2. Pruebas de consistencia de la ontología.	Protégé
3. Definición del proceso manual de anotación semántica externa de los documentos.	Jena, MySQL
4. Estructura para la persistencia de las anotaciones semánticas.	MySQL
5. Aplicación que permite anotar manualmente los documentos con elementos de la ontología.	Jena, MySQL
6. Indexación de los documentos mediante elementos de la ontología.	Jena, Lucene
7. Herramienta de software que permite realizar búsquedas basadas en texto previas a las búsquedas semánticas.	Lucene
8. Aplicación que permite realizar búsquedas semánticas de documentos en el dominio de la lingüística.	Jena, MySQL
9. Evaluación de los resultados obtenidos por las búsquedas semánticas usando precisión y *recall*.	Precisión y *Recall*

1.5.1 Herramientas

1.5.1.1 OWL[1]

Ontology Web Language (OWL) es un lenguaje de Web Semántica desarrollado por la W3C y diseñado para representar conocimiento rico y complejo sobre cosas, grupos de cosas y relaciones entre cosas. OWL es un lenguaje basado en lógica computacional, tal que el conocimiento expresado por éste puede ser aprovechado por programas de computadora para, por ejemplo, verificar la consistencia del conocimiento o para hacer explícito el conocimiento implícito. Los documentos de OWL, conocidos como ontologías, pueden ser publicados en la *World Wide Web* y pueden referenciar o ser referenciados por otras ontologías OWL.

El presente proyecto de fin de carrera establece como uno de sus resultados esperados el desarrollo de una ontología en el dominio de la lingüística, por lo que este lenguaje es el que se usará para su modelado y construcción.

OWL forma parte de una familia de tecnologías de la Web Semántica, la cual incluye RDF, RDFS, SPARQL, entre otras. Es una extensión del lenguaje RDF y dependiendo del nivel de expresividad se divide en OWL Lite, OWL DL y OWL FULL. El lenguaje que se usará para el proyecto es el OWL DL, ya que provee máxima expresividad garantizando integridad de información para modelar clases y subclases en la ontología.

1.5.1.2 Protégé[2]

Es un editor de ontologías de código abierto y un *framework* para la construcción de sistemas inteligentes desarrollado por la Universidad de Stanford. Es mantenido por una fuerte comunidad de académicos, gubernamentales y usuarios corporativos que lo utilizan para construir soluciones basadas en conocimiento en áreas tan diversas como la biomedicina, el comercio electrónico y el modelado organizacional. Proporciona un conjunto de herramientas para la construcción de modelos de dominio y aplicaciones basadas en conocimiento con ontologías. Protégé implementa

[1] http://www.w3.org/2001/sw/wiki/OWL
[2] http://protegewiki.stanford.edu/wiki/Protege

un amplio conjunto de estructuras y funciones para el modelado de conocimiento que soportan la creación, visualización y manipulación de ontologías en diversos formatos de representación. Para todas estas funciones relacionadas al manejo de ontologías, Protégé provee una interfaz gráfica, lo que facilita la ejecución de estas tareas. Además, es totalmente compatible con las últimas especificaciones de OWL y RDF de la *World Wide Web Consortium*.

Esta herramienta será usada en el proyecto para el desarrollo de la ontología, ya que facilita esta tarea mediante su interfaz gráfica y las funciones que proporciona. Protégé permitirá construir la ontología en lenguaje OWL. Además, también será usada para las pruebas de consistencia de la ontología, mediante consultas a ésta con el uso de razonadores que provee la herramienta.

1.5.1.3 Jena[3]

Es un *framework* de código abierto escrito en Java para la construcción de aplicaciones de Web Semántica y *Linked Data*. Proporciona un entorno de programación con librerías para manejar tecnologías como RDF, RDFS y OWL, SPARQL, entre otros. La API principal de Jena permite crear y leer grafos RDF, mientras que la API de Ontologías permite añadir semántica extra a los datos RDF mediante OWL. Además, incluye un motor de inferencia basado en reglas el cual razona sobre ontologías OWL. Este *framework* será usado en el presente proyecto para la implementación del buscador en el ámbito de la Web Semántica, proporcionando las herramientas necesarias para la interacción con la ontología lingüística escrita en OWL.

1.5.1.4 Lucene[4]

Es una librería de motor de búsqueda de alto rendimiento y funcionalidades completas escrita enteramente en Java. Es una tecnología adecuada para casi cualquier aplicación que requiera búsqueda basada en texto, especialmente multiplataforma. El presente proyecto establece en sus resultados esperados la implementación de un buscador basado en texto, por lo que esta librería será utilizada, proporcionando funcionalidades para la recuperación de información e indexación.

[3] https://jena.apache.org/about_jena/about.html
[4] https://lucene.apache.org/

1.5.1.5 MySQL *Community Edition*[5]

MySQL *Community Edition* es la versión de descarga gratuita de la base de datos de código abierto MySQL. Está disponible bajo la licencia GPL y es mantenida por una gran comunidad activa de desarrolladores de código abierto. Se encuentra disponible para más de 20 plataformas y sistemas operativos, incluyendo Linux, Unix, Mac y Windows. Esta base de datos será usada para la definición de la estructura que almacenará los metadatos relacionados a las anotaciones semánticas de los documentos.

1.5.2 Métodos

1.5.2.1 Precisión y *Recall*

La precisión y el *recall* son dos medidas estándar para estimar la efectividad de la recuperación de información (Manning & Raghavan, 2009). Estas medidas serán utilizadas en este proyecto para la evaluación de los resultados obtenidos por el buscador semántico una vez que esté implementado.

La Precisión es la fracción de documentos recuperados que son relevantes. Se calcula mediante la división entre la cantidad de documentos relevantes recuperados y el total de documentos recuperados (Manning & Raghavan, 2009). El *Recall*, en cambio, es la fracción de documentos relevantes que son recuperados. Se calcula mediante la división entre la cantidad de documentos relevantes recuperados y el total de documentos relevantes (Manning & Raghavan, 2009).

1.5.3 Metodologías

1.5.3.1 Ontology Development 101

Es un enfoque iterativo para el desarrollo de ontologías. Posee tres reglas fundamentales para el diseño de estas ontologías (Noy & McGuinness, 2001):

[5] http://www.mysql.com/products/community

- No existe solo una manera correcta para modelar un dominio, siempre existen alternativas viables. La mejor solución casi siempre depende de la aplicación que se tiene en mente y las extensiones que se anticipan.
- El desarrollo de ontologías es necesariamente un proceso iterativo.
- Los conceptos de la ontología deben estar estrechamente relacionados a objetos (físicos o lógicos) y a relaciones en el dominio de interés. Estos probablemente serán sustantivos (objetos) o verbos (relaciones) en oraciones que describen el dominio.

Además, esta metodología consiste en 7 pasos (Noy & McGuinness, 2001):

- Determinar el dominio y alcance de la ontología
- Considerar reusar ontologías existentes
- Enumerar términos importantes en la ontología
- Definir las clases y la jerarquía de clases
- Definir las propiedades de las clases (ranuras)
- Definir las facetas de las ranuras
- Crear instancias

Esta metodología será usada en el presente proyecto para el desarrollo de la ontología lingüística, guiando el diseño y construcción mediante sus reglas y pasos.

1.6 Alcance

El presente proyecto de fin de carrera implementará una aplicación en el ámbito de la Web Semántica para la recuperación de documentos en el dominio de la lingüística. Este proyecto pertenece al área de las Ciencias de la Computación, empleando conceptos como ontología, anotación semántica y recuperación de información.

Dado que el dominio de la lingüística es bastante amplio, no es factible desarrollar una ontología que represente todo este dominio. Entonces, con el objetivo de determinar el alcance que realmente se va a representar, se necesitó conocer qué partes o subáreas de la lingüística cubren los temas y conceptos que tratan los documentos para los que el buscador será implementado (estos documentos

pertenecen al Departamento de Humanidades de la universidad). En una reunión con el profesor de lingüística ayudante en el proyecto, se determinó que estos documentos cubren lo que es lingüística descriptiva. La lingüística descriptiva es una parte de la lingüística que estudia la descripción de las estructuras fonológicas, gramaticales y semánticas de las lenguas en un momento determinado de la historia [27]. Por lo tanto, se definió que el alcance de la ontología no es todo el dominio de la lingüística, sino la lingüística descriptiva.

Luego del desarrollo de la ontología, se realizará la anotación semántica manual de los documentos (artículos, diccionarios, gramáticas, tesis y libros), los cuales estarán almacenados en un repositorio digital. Para esto, se escogerán aproximadamente 50 documentos que servirán como data de prueba para la aplicación. Se escogerán 50 documentos ya que se considera una cantidad que significa un esfuerzo aceptable en el contexto de un proyecto de fin de carrera.

Los documentos serán anotados con conceptos de la ontología lingüística. No se podrá incluir términos de otros dominios. Para esto, antes de realizar una anotación se verificará que el término sea algún concepto de la ontología. Finalmente, se desarrollará una herramienta que permita recuperar los documentos mediante búsquedas semánticas y se evaluará su performance.

Un buscador semántico puede ser desarrollado para otros dominios, habiéndose elegido para este proyecto el de la lingüística. Por esta razón, la ontología que se desarrollará será del dominio de la lingüística con el alcance definido y no incluirá conceptos de otros dominios. De esta manera, no se trata de un buscador general de documentos, sino de documentos científicos relacionados a la lingüística.

Por último, la alternativa de solución que se implementará consistirá en una aplicación web. Si bien el proyecto surgió para solucionar los problemas del Departamento de Humanidades de la universidad, se quiso generalizar la solución, por lo que se tomó esta decisión.

1.7 Justificación

El presente proyecto de fin de carrera implementa una aplicación que permite buscar documentos del dominio de la lingüística almacenados en un repositorio digital. Se eligió el dominio de la lingüística ya que en la especialidad de Humanidades de la PUCP se poseen aproximadamente 2000 documentos lingüísticos, y no se cuenta con un medio adecuado para su búsqueda. Es decir, este proyecto surge como una solución a los problemas actuales que presenta la especialidad de Humanidades de la PUCP para la búsqueda de sus documentos lingüísticos.

En general, cualquier tipo de centro de estudio que cuente con una facultad de lingüística podría beneficiarse con esta aplicación. A menudo se cuenta con grandes cantidades de documentos pero no se cuenta con un medio adecuado para su consulta o recuperación. Estos centros de estudios podrían poner a disposición sus documentos académicos a través de esta aplicación. De esta manera, se les facilitaría a sus alumnos y profesores la recuperación de información relacionada al área o dominio de la lingüística.

Es posible que existan centros de estudios que cuenten con buscadores convencionales (sintácticos) para la recuperación de sus documentos académicos. Sin embargo, la diferencia del buscador que implementa este proyecto con los buscadores convencionales basados en texto es que es mejor en cuanto a precisión. Es decir, brinda resultados de mayor relevancia para el usuario. Al facilitar a los usuarios la obtención de información relevante se les está brindando mayores capacidades para sus fines académicos, contribuyendo a sus requerimientos de información al momento de realizar trabajos, proyectos o investigaciones.

1.8 Alternativas de solución

El buscador semántico de documentos en el dominio de la lingüística que implementará el presente proyecto de fin de carrera es una alternativa de solución que se ha planteado para el problema central, el cual se define como la dificultad para obtener información relevante de documentos digitales en el dominio de la

lingüística. Siendo una alternativa de solución, esto quiere decir que pueden existir muchas otras.

La ontología lingüística se desarrollará luego de un análisis previo que se realizará a este dominio con el objetivo de determinar su alcance. Esta ontología es solo una alternativa a muchas otras ontologías que se puedan desarrollar en el mismo dominio. Otras ontologías podrían tener un menor o mayor alcance, incluir otros conceptos o modelar los conceptos y sus relaciones de una manera distinta. Estas diferentes ontologías ocasionarían que los resultados de las búsquedas semánticas varíen.

De la misma manera, las anotaciones semánticas a los documentos podrían realizarse de muchas otras formas. En el presente proyecto, se realizarán anotaciones manuales, pudiendo ser otra alternativa las anotaciones automáticas. La ventaja de las anotaciones manuales es que el usuario elige directamente qué términos desea usar; en las automáticas, en cambio, algún mecanismo automatizado es el que decide qué términos se van a usar, existiendo la probabilidad de que no se anoten algunos términos deseados por el usuario. Para resolver este problema, se pueden usar anotaciones semiautomáticas. Por el otro lado, la ventaja de las anotaciones automáticas es que no se requiere de la intervención del usuario para realizar las anotaciones; sin embargo, desarrollar una herramienta de anotación automática (o semiautomática) es complejo e implica un esfuerzo que no se considera adecuado para un proyecto de fin de carrera.

Además, las anotaciones se almacenarán separadas a los documentos; es decir, se tratan de anotaciones externas. Otra alternativa de solución sería realizar anotaciones empotradas, las cuales son anotaciones que forman parte del contenido del archivo o documento. Se eligieron anotaciones externas ya que los documentos con los que se cuenta son en formato pdf, lo que no permite que se modifique su contenido.

2 MARCO CONCEPTUAL

2.1 Introducción

En esta sección, se presentarán y explicarán los conceptos de Web Semántica, Ontologías, Metadatos y Anotación Semántica; con el objetivo de entender el funcionamiento de un buscador en el ámbito de la Web Semántica. Estos conceptos están relacionados entre sí y cada uno de ellos ayuda a un mejor entendimiento del concepto principal, el cual es Web Semántica. De esta manera, se obtienen los conocimientos necesarios y completos sobre lo que implica un buscador semántico.

2.2 Web Semántica

Desde su creación, la *World Wide Web* ha permitido que las computadoras entiendan únicamente la estructura y diseño de las páginas Web con el objetivo de mostrar su información, sin tener acceso a su significado. El contenido de la Web actual está diseñado para ser leído y entendido por las personas, no para que las computadoras puedan manipularlo significativamente. Las computadoras pueden hábilmente analizar sintácticamente las páginas Web para el procesamiento de su diseño y estructura (identifica y muestra cabeceras, enlaces, entre otros); sin embargo, no poseen una forma de procesar el significado de la información (Berners-Lee, 2001).

Es en este contexto que surge la Web Semántica como una extensión de la Web actual, en donde la información tiene un significado bien definido, permitiendo a las personas y computadoras trabajar cooperativamente (Berners-Lee, 2001). Consiste en estructurar el significado (semántica) de la información contenida en los documentos de una manera en que las computadoras puedan procesar, transformar, reunir e incluso actuar sobre ésta de forma útil (Yu, 2007). Esta estructuración del significado es realizada por una capa de metadatos añadida a los datos existentes, la cual es "entendible" por las máquinas, lo que le permite efectuar las tareas antes mencionadas.

Una característica importante de esta descripción o estructuración de la semántica de los recursos es que permite a las computadoras razonar por inferencia. Una vez que los recursos son descritos usando hechos, asociaciones y relaciones; motores de

inferencia, llamados también razonadores, pueden derivar nuevo conocimiento y obtener conclusiones lógicas de la información existente (Cardoso, 2007). Por ejemplo, en el dominio de la lingüística, se tiene el concepto de Morfología asociado el concepto de Lexemas; es decir, la morfología estudia, entre muchas otras cosas, a los lexemas. Si un documento tiene asociado el concepto de Lexemas (el documento tiene como uno de sus temas a los lexemas), entonces, por inferencia, se puede concluir que dicho documento trata sobre morfología. De esta manera, el uso de motores de inferencia en la Web Semántica permite a aplicaciones como buscadores preguntarse cómo es que una conclusión lógica en particular ha sido trazada (Cardoso, 2007); es decir, los buscadores semánticos pueden dar prueba de sus resultados explicando los pasos involucrados en el razonamiento lógico que se realizó sobre los datos hasta llegar a dicho resultado. Siguiendo el ejemplo anterior, si se le consulta a un buscador semántico por todos los documentos que traten sobre morfología, mostrará entre sus resultados a los documentos que traten sobre lexemas. El buscador llegará a esta conclusión ya que le fue provista la información de que los Lexemas son un campo de estudio de la Morfología.

2.3 Ontologías

El término “Ontología” se origina en la filosofía, y ha estado siendo cada vez más objeto de estudio en las ciencias de la computación y sistemas de información (Lim, Liu & Lee; 2011). Desde el punto de vista filosófico (Aristóteles y Kant, por ejemplo), ontología es el estudio de la existencia, es una forma fundamental de representar conocimiento sobre el mundo real (Lim, Liu & Lee; 2011). Desde el punto de vista de las ciencias de la computación, se define como un conjunto de representaciones con las cuales modelar un dominio de conocimiento (Lim, Liu & Lee; 2011). Estas representaciones contienen clases, atributos o propiedades, y relaciones entre las clases (Lim, Liu & Lee; 2011).

Las ontologías hacen posible el desarrollo de sistemas basados en conocimiento a través de especificaciones formales que permiten que ingenieros del conocimiento desarrollen su propia ontología por medio del reúso e intercambio entre ellos (Lim, Liu & Lee; 2011). Gracias a las ontologías, estos sistemas permiten que las

computadoras puedan intercambiar conocimiento, estableciéndose así una comunicación inteligente, como la que se da en los agentes de software (Lee, 2007).

Para un adecuado intercambio de conocimiento, debe existir un adecuado modelamiento de las ontologías. Este modelamiento en las ciencias de la computación, denominado ontología computacional, es menos complejo que en la filosofía. Consiste en una representación simbólica de objetos de conocimiento, clases de objetos, propiedades de objetos y las relaciones entre estos objetos para explícitamente representar conocimiento sobre un dominio (Lim, 2011). Este modelamiento es usualmente simplificado en diferentes tipos de definiciones matemáticas, definiciones lógicas o lenguaje estructural. A continuación, se presenta la Figura 1 en la que se muestra un ejemplo simple de ontología del dominio de la Lingüística.

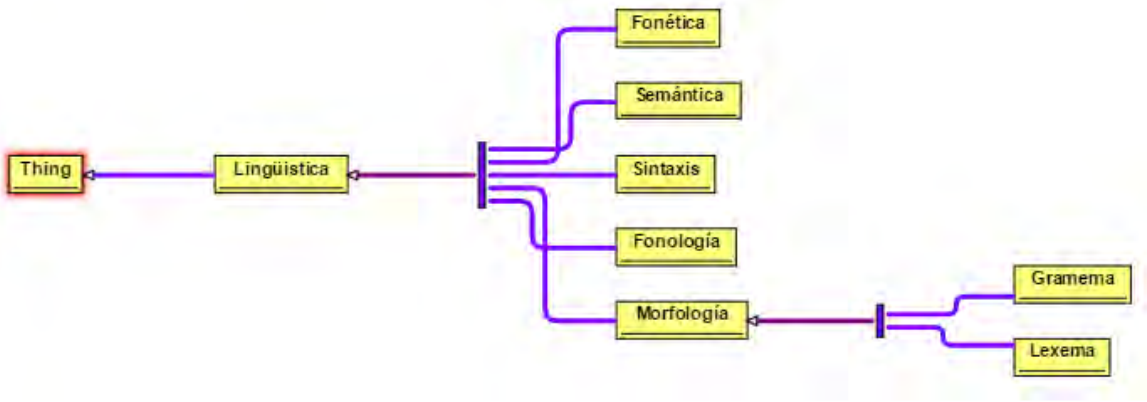

Figura 1: Ejemplo de ontología del dominio Lingüística. Autoría propia.

Las ontologías son uno de los componentes básicos de la Web Semántica (Berners-Lee, Hendler & Lassila; 2001). Para investigadores del campo de Inteligencia Artificial, una ontología es un documento o archivo que define formalmente las relaciones entre términos o conceptos. El tipo más común de ontología posee una taxonomía y un conjunto de reglas de inferencia (Berners-Lee, Hendler & Lassila; 2001).

Una taxonomía define clases de objetos y las relaciones entre ellos. Por ejemplo, se tiene la clase *tema* con una de sus instancias siendo *morfología*, y la clase *concepto*

con una de sus instancias siendo *lexemas*. Además, se tiene la relación *tieneConcepto*, la cual relaciona un tema a los conceptos que trata. De esta manera, se podría tener la siguiente relación: *morfología tieneConcepto lexemas*. Las clases, subclases y las relaciones entre entidades son una herramienta muy poderosa para el uso de la Web (Berners-Lee, Hendler & Lassila; 2001). Se pueden expresar un gran número de relaciones entre entidades mediante la asignación de propiedades a las clases y permitiendo a las subclases heredar tales propiedades.

Las ontologías pueden mejorar el funcionamiento de la Web de muchas formas. Una de ellas es que pueden ser usadas de manera sencilla para mejorar la precisión de las búsquedas en la Web (Berners-Lee, Hendler & Lassila; 2001). Un programa de búsqueda puede buscar sólo aquellos documentos que hacen referencia a un concepto preciso en lugar de todos los que utilizan palabras clave ambiguas. Por ejemplo, se quiere buscar todos aquellos documentos lingüísticos que traten sobre el concepto de "sujeto", del área de la gramática. Gracias a las ontologías, un buscador semántico examinaría todos los documentos dejando de lado aquellos que hacen uso de la palabra en otros contextos (el adjetivo "sujeto", por ejemplo), mejorando así la precisión de la búsqueda.

Además, las páginas Web enriquecidas con metadatos relacionados a ontologías hacen que sea mucho más fácil desarrollar programas que puedan hacer frente a preguntas complicadas cuyas respuestas no residen en una sola página (Berners-Lee, Hendler & Lassila; 2001). Por ejemplo, se desea encontrar una de las obras del famoso lingüista Andrés Bello. No se conoce el nombre exacto de la obra, pero se sabe que fue publicada entre los años 1880 y 1890, y tiene como tema central a la Gramática. Un programa inteligente de búsqueda puede examinar todas las páginas sobre el lingüista Andrés Bello en donde se mencionan sus obras, dejando de lado aquellas publicadas fuera del rango especificado, y seguir los enlaces a las páginas de sus obras para examinar si tienen como tema a la Gramática.

2.4 Metadatos

El término *metadatos* se originó en las ciencias de la información y en las comunidades de datos geoespaciales antes de ser adoptado y redefinido parcialmente

por bibliotecas, archivos y comunidades de museos a finales del siglo XX (Baca, 2008). Hoy en día, el término ha sido extensamente adoptado por diversas audiencias. La definición concisa es "datos que describen otros datos" (Baca, 2008). Un ejemplo común es una tarjeta de catálogo de una biblioteca, el cual contiene datos sobre el contenido y ubicación de un libro.

Una propiedad importante de los metadatos es que son normalmente estructurados para modelar los atributos más importantes del tipo de objeto que describe (Gill, 2008). Al modelar con exactitud los atributos más esenciales de la clase a la que pertenecen los objetos de información siendo descritos, los metadatos se convierten en una herramienta útil para usar y administrar dicha clase. En este contexto, entonces, se puede definir a los metadatos como "una descripción estructurada de los atributos esenciales de un objeto de información" (Gill, 2008).

Los metadatos desempeñan un papel muy importante en lo que es la Web Semántica, ya que definen el significado de los objetos de información con cierta precisión y sus relaciones estructurales son explícitas (Wittenburg & Broeder, 2015). Es decir, se encargan de la anotación semántica de la información, de manera que sea "entendible" por las computadoras. Una de las tecnologías más importantes usadas para esto es el *Resource Description Framework* (RDF), el cual es un modelo de datos para metadatos que se usa como método general para la descripción conceptual o modelamiento del significado de la información; es decir, para el modelado de ontologías en la Web Semántica.

El RDF es un candidato prometedor para realizar algunos de los retos de la Web Semántica (Wittenburg & Broeder, 2015). Este *framework* ha sido desarrollado por expertos en metadatos y representación del conocimiento (Wittenburg & Broeder, 2015). Se basa en XML para crear descripciones complejas de recursos. Ofrece un conjunto de reglas para definir elementos y para la creación de relaciones semánticas. Las relaciones son definidas con un mecanismo muy simple, que permite su procesamiento automático.

En un entorno RDF, cada recurso tiene que tener un identificador único (URI). Pueden tener propiedades y estas propiedades pueden tener valores. Una afirmación

simple sería la siguiente: "la obra *Gramática de la lengua castellana* tiene como autor a *Andrés Bello*" (ver Figura 2), en donde la obra sería un recurso con un identificador URI único y tendría la propiedad *autor* cuyo valor sería *Andrés Bello*. El autor podría ser un literal o también un recurso con su respectivo URI.

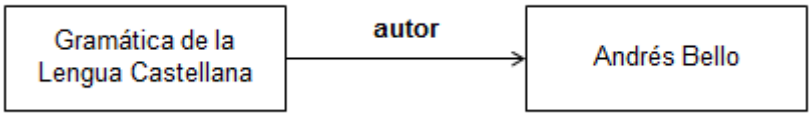

Figura 2: Mecanismo simple de afirmación RDF donde un objeto se caracteriza por la propiedad "autor", el cual toma el valor "Andrés Bello". Autoría propia.

Los metadatos diseñados cuidadosamente basados en repositorios abiertos pueden ser vistos como representaciones de pedazos de ontologías de diferentes dominios. De esta manera, las discusiones que se tienen acerca de estos metadatos son una gran contribución a tales ontologías. Por lo tanto, se puede decir que las iniciativas actuales sobre metadatos constituyen pasos importantes hacia la realización de la Web Semántica (Wittenburg & Broeder, 2015).

2.5 Anotación Semántica

Supongamos que añadimos una etiqueta "`<p>`" a una parte de un documento de la siguiente manera: "`<p> Andrés Bello </p> es un famoso lingüista…`". ¿Los metadatos introducidos de esta manera resultan útiles?, ¿se puede decir que están representado algún tipo de semántica? Sin consideraciones adicionales, la respuesta a estas preguntas es no.

Para que los metadatos sean útiles en el contexto de la Web Semántica, éstos deben significar algo; es decir, los símbolos que constituyen estos metadatos deben permitir una interpretación adicional de la información (Kiryakov, Popov, Terziev, Manov, & Ognyanoff, 2004). Esta interpretación se puede realizar como resultado de la asignación de significado o semántica a los metadatos con respecto a un determinado contexto o dominio. El proceso de introducción de metadatos, los cuales tienen asociado semántica sobre un dominio, a todo o parte del documento es lo que se conoce como anotación semántica.

El dominio puede ser modelado por una ontología, la cual resulta un factor clave para el funcionamiento de las anotaciones semánticas. La Web Semántica propone anotar el contenido de los documentos utilizando información semántica de ontologías (Berners-Lee et al. 2001). Estas anotaciones identifican formalmente conceptos y relaciones entre conceptos en los documentos (Uren, Cimiano, Iria, Handschuh, Vargas-Vera, Motta, & Ciravegna, 2006), relacionando su contenido con las entidades y propiedades de una ontología, asignando así significado a la información. Es esta relación la que permite la "interpretación adicional" de la que se hablaba en el párrafo anterior. Un buscador, por ejemplo, puede realizar inferencias sobre la información a través de la explotación de las ontologías, brindando mayores capacidades a la recuperación de la información (Uren et. al., 2006), Estas búsquedas obtendrían como resultado anotaciones semánticas, recuperando los documentos asociados a estas anotaciones.

En otras palabras, la anotación semántica es un esquema de generación y uso de metadatos específicos, que tiene como objetivo permitir nuevos métodos de acceso a la información y extender los existentes (Kiryakov, Popov, Terziev, Manov, & Ognyanoff, 2004). Este esquema está basado en el concepto de *named entities*, el cual constituye una parte importante de la semántica de los documentos (Kiryakov, Popov, Terziev, Manov, & Ognyanoff, 2004). En el campo de Procesamiento de Lenguaje Natural (NLP), las *named entities* son entidades que pueden ser referenciadas por nombres, como por ejemplo personas, organizaciones, lugares, entre otros (Kiryakov, Popov, Terziev, Manov, & Ognyanoff, 2004). También están incluidos valores numéricos como fechas, números, cantidades. Estas entidades pueden ser asociadas a sus descripciones formales, proporcionando así semántica sobre un determinado dominio (Kiryakov, Popov, Terziev, Manov, & Ognyanoff, 2004).

La anotación semántica se basa en asociar las entidades con sus respectivas descripciones o definiciones semánticas a través de los metadatos (Kiryakov, Popov, Terziev, Manov, & Ognyanoff, 2004). Consiste en etiquetar las instancias de las clases de la ontología que se presentan en los documentos y mapearlas a esta ontología, de donde se puede explorar sus definiciones; es decir, sus propiedades y relaciones con otras clases o entidades (Reeve & Han, 2005). De esta manera, se

puede extraer conocimiento basándose en el análisis de las definiciones de las instancias etiquetadas.

Por último, existen dos tipos de anotaciones semánticas: empotradas y externas. En las anotaciones empotradas, las anotaciones forman parte del contenido del archivo; es decir, los metadatos están introducidos dentro del documento. Las anotaciones externas consisten en anotaciones almacenadas separadas del contenido del documento. Esto permite que los usuarios puedan agregar sus propias anotaciones y compartirlas, siendo más dinámicas.

2.6 Conclusión

La Web Semántica consiste en una Web de Datos en la que los datos de todas las diferentes fuentes se encuentran correctamente integrados y su semántica se encuentra definida explícitamente de manera que es "entendible" por las computadoras. La semántica es definida por medio de las anotaciones semánticas y las ontologías. Estas anotaciones enlazan o relacionan el contenido de los documentos con las respectivas ontologías a través de metadatos, los cuales permiten que la semántica pueda ser procesada automáticamente por las computadoras, para lo cual la tecnología usada es el *Resource Description Framework* (RDF).

Se han revisado los conceptos considerados más importantes relacionados a la Web Semántica, que es el concepto principal que rodea a la problemática. Un buscador semántico hace uso de las tecnologías que ofrece la Web Semántica para encontrar documentos basándose en el procesamiento del significado de la información. De esta manera, luego de esta revisión, se comprende cómo es que funciona un buscador semántico.

3 ESTADO DEL ARTE

3.1 Introducción

En esta sección, se presentarán algunos proyectos relacionados a la implementación de buscadores semánticos de documentos, describiendo cómo hacen uso de las tecnologías de la Web Semántica.

El objetivo de esta revisión es conocer los diferentes avances en este tema realizados en los últimos años. De este modo, estos avances pueden ser tomados en cuenta para el desarrollo del presente proyecto de fin de carrera.

3.2 Método usado en la revisión del estado del arte

El método usado en la revisión del estado del arte fue la revisión sistemática. Este método consiste en identificar, evaluar e interpretar toda investigación disponible relevante a una pregunta de revisión en particular, área temática o fenómeno de interés (Kitchenham, 2004).

3.2.1 Preguntas de revisión

Las preguntas de revisión que se formularon para guiar la revisión fueron las siguientes:

[P1] ¿De qué manera el buscador semántico obtiene conocimiento sobre un determinado dominio usando ontologías?

[P2] ¿Cómo deben estar estructurados los documentos para poder ser recuperados por un buscador semántico?

[P3] ¿Qué métodos/tecnologías/herramientas/mecanismos se han empleado en las implementaciones de buscadores semánticos de documentos digitales?

[P4] ¿Qué estudios o proyectos de buscadores semánticos en el dominio de la lingüística se han realizado?

3.2.2 Selección de las fuentes

Se usaron distintas fuentes para la búsqueda de estudios que ayuden a responder las preguntas de revisión. Se eligieron fuentes recomendadas y confiables para el área de las ciencias de la computación. Éstas se pueden apreciar en la Tabla 2.

Tabla 2: Fuentes usadas en la revisión sistemática

Scopus	http://www.scopus.com
ACM	http://dl.acm.org/
IEEE	http://ieeexplore.ieee.org

Para las búsquedas en estas fuentes se usaron distintas cadenas para cada preguntas de revisión (ver Tabla 3). De esta manera, los estudios encontrados mediante una cadena de búsqueda se enfocan en responder a su pregunta asociada.

Tabla 3: Cadenas de búsqueda utilizadas

Pregunta de revisión	**Cadena de búsqueda**
P1	(semantic web OR web) AND semantic search AND ontology
P2	(semantic web OR web) AND semantic annotation AND (documents OR digital documents)
P3	(semantic web OR web) AND semantic search AND (method OR tools OR frameworks)
P4	semantic search AND (linguistics domain OR linguistic ontology)

Además, con el objetivo de obtener mejores resultados en la selección de la información encontrada, se establecieron criterios de inclusión y exclusión. El

criterio de inclusión considera todos aquellos estudios o proyectos que traten meramente sobre implementación y no otros aspectos o temas con respecto a buscadores semánticos. El de exclusión pretende no tomar en cuenta aquellos proyectos que tengan un enfoque distinto al de la Web Semántica. Se aceptaron los proyectos encontrados que cumplieron con ambos criterios.

La Tabla 4 muestra la cantidad de estudios encontrados para cada pregunta de revisión y por cada fuente antes y después de aplicar los criterios de inclusión y exclusión.

Tabla 3: Número de documentos encontrados por fuente y pregunta.

Fuente	**Número de documentos encontrados**							
	Antes de aplicar criterios				**Luego de aplicar criterios**			
	P1	**P2**	**P3**	**P4**	**P1**	**P2**	**P3**	**P4**
Scopus	6	4	3	0	5	2	2	0
ACM	4	2	2	0	2	1	2	0
IEEE	5	3	7	1	4	3	4	1
Subtotal	**15**	**9**	**12**	**1**	**11**	**6**	**8**	**1**
Total	**37**				**26**			

3.3 Investigaciones acerca del tema

En esta sección se presentan evidencias en forma de estudios o proyectos con el objetivo de responder a las preguntas de revisión planteadas anteriormente. Se describen los proyectos considerados más relevantes entre los encontrados en la revisión sistemática, enfocándose en los aspectos que tengan relación con las preguntas.

3.3.1 P1: Obtención de conocimiento sobre un determinado dominio usando ontologías

[P1] ¿De qué manera el buscador semántico obtiene conocimiento sobre un determinado dominio usando ontologías? La Tabla 5 muestra el resumen de los estudios más relevantes encontrados que responden a esta pregunta.

Tabla 5: Resumen de estudios más relevantes que responden a la pregunta de revisión [P1]

Estudio	Dominio de la ontología	Consideraciones importantes
[17]	Periodismo	- Dominio amplio - Tecnología usada: RDF
[19]	Leyes y administración pública	- Desarrollo de un conjunto de ontologías en dominios particulares - Cada concepto de las ontologías fue asociado a un conjunto de sinónimos
[20]	Informática	- Creación de la ontología en dos pasos: integración de palabras claves y formalización

El propósito de esta pregunta fue conocer cómo es que un buscador obtiene o se le provee conocimiento sobre un determinado dominio mediante ontologías, ya que esto le permite realizar búsquedas semánticas. Para esto, los proyectos de buscadores semánticos siguen un proceso de definición y construcción de sus ontologías (en caso no reutilicen de terceros), para luego integrarlas a sus procesos de búsqueda. Sin embargo, si bien todas emplean ontologías, cada proyecto define el dominio de ésta de acuerdo a su contexto y necesidades.

Por ejemplo, se tiene el proyecto Neptuno [17], el cual propuso la introducción de las tecnologías de la Web Semántica para mejorar los procesos de creación, mantenimiento y explotación de los archivos de la hemeroteca digital del diario español SEGRE. El primer paso para la implementación de su buscador semántico

fue la definición de una ontología para representar y proveer conocimiento del área del periodismo. Uno de los aspectos importantes al momento de definir una ontología es determinar su alcance. En el caso de este proyecto, se observó que el área periodística y de noticias tenía la peculiaridad de tratar con temas de todas las áreas del conocimiento humano y de la actualidad (política, cultura, leyes, ciencia, deporte, artes, economía, etc.), por lo que se vio necesario establecer cuidadosamente un límite en el dominio a representar. De lo contrario, daría lugar a todo un proyecto para cada área en particular. Luego, en la construcción de la ontología, la tecnología que se utilizó fue el *Resource Description Framework* (RDF), ya que se consideró el estándar más maduro, usado y estable en los últimos proyectos sobre Web Semántica.

Por otro lado; Berrueta, Labra y Polo (2006) aplicaron las tecnologías de la Web Semántica para la creación de un motor de búsqueda de documentos legales y de administración pública para el BOPA[6] (Boletín Oficial del Principado de Austrias). Luego de un análisis extensivo del alcance del dominio, vieron necesario la construcción de dos tipos de ontologías (desarrolladas en lenguaje OWL-DL) con diferentes propósitos:

1. Una ontología jurídica y administrativa. Esta ontología formaliza la estructura básica del BOPA y de la Administración Pública Regional del Principado de Austrias. Captura los conocimientos del área legislativa y de administración pública.

2. Un conjunto de ontologías de dominios particulares. Cada una captura un área pequeña y bien definida de conocimiento general, llevando el conocimiento experto humano al sistema.

Cabe resaltar este proyecto, ya que en la construcción de sus ontologías se tuvo una consideración adicional. Cada concepto de estas ontologías fue asociado a un conjunto de sinónimos, empleando un mecanismo similar al presentado en la arquitectura WordNet[7]. El objetivo fue brindar mayor transparencia para los usuarios; es decir, que no necesariamente tengan que conocer exactamente los términos empleados por las ontologías. Además, se puede considerar que este mecanismo permite que el buscador "entienda" un rango más amplio de términos.

[6] https://www.asturias.es/**bopa**/
[7] https://**wordnet**.princeton.edu/

Otro proyecto a destacar que provee conocimiento semántico mediante ontologías es el denominado *"Language Technology for eLearning"* (LT4eL) [20]. Este proyecto implementa un Sistema de Gestión de Aprendizaje para mejorar la gestión, distribución y sobre todo la recuperación en diferentes idiomas de material de aprendizaje. Este material de aprendizaje consiste en documentos digitales abarcando el dominio de la informática (*computing*). Uno de los resultados obtenidos en este proyecto fue la construcción de una ontología de dicho dominio.

La creación de esta ontología se puede resumir en dos pasos: integración de palabras claves y formalización. En el primer paso, las palabras clave fueron desambiguadas y clasificadas en el espacio conceptual del dominio. En la etapa de formalización, las definiciones de los conceptos extraídos y sus relaciones fueron formuladas utilizando el lenguaje OWL-DL. Finalmente, se añadieron nuevos conceptos (no representados por ninguna palabra clave) con el fin de mejorar la cobertura del dominio.

3.3.2 P2: Estructuración de los documentos para poder ser recuperados por un buscador semántico

[P2] ¿Cómo deben estar estructurados los documentos para poder ser recuperados por un buscador semántico? La Tabla 6 muestra el resumen de los estudios más relevantes encontrados que responden a esta pregunta.

Tabla 6: Resumen de estudios más relevantes que responden a la pregunta de revisión [P2]

Estudio	Anotación semántica de los documentos	Mecanismo empleado
[18]	- Externa - Manual y automática	- Anotación a todo o partes del documento - Soporta documentos en varios formatos
[17]	- Externa - Manual	- Vocabularios controlados (IPTC *Subject Reference System*) - Dos formas alternativas para clasificar más a fondo el contenido: por género y por contenido

[20]	- Externa y empotrada - Manual	- En la anotación externa se incluyen todos los conceptos y relaciones que se consideren importantes para la clasificación del documento - En la anotación empotrada se consideran las apariciones de los conceptos dentro de todo el contenido del documento o de sus partes (párrafos, oraciones)

El significado o semántica de la información contenida en los documentos deber estar definido y estructurado explícitamente para poder ser procesado por un buscador semántico. De esta manera, los buscadores pueden reconocer el contenido de estos documentos y determinar si es de relevancia para el usuario o no. Todos los estudios y proyectos encontrados realizan esto empleando el concepto de anotaciones semánticas. Sin embargo, cada uno define y emplea un proceso de anotación distinto de acuerdo a sus necesidades y capacidades. Esta situación se evidencia a continuación, mediante la descripción de proyectos relevantes.

Uno de los proyectos que resalta en este tema es GoNTogle [18], el cual consistió en la implementación de un *framework* para la anotación semántica de documentos y su recuperación, construido en base a las tecnologías de la Web Semántica (OWL, RDF, entre otros) e IR (Recuperación de Información).

Este *framework* soporta anotación semántica basada en ontologías para documentos en varios formatos (doc, pdf, txt, rtf, odt, sxw, entre otros). Permite la anotación de todo el documento o partes de éste, y provee mecanismos de anotación manual y automática. La anotación automática se basa en un método de aprendizaje que explota el historial de anotaciones de los usuarios e información textual para sugerir automáticamente anotaciones para los nuevos documentos. Todas las anotaciones son almacenadas en un servidor centralizado, separados de los documentos originales. Esto permite un entorno colaborativo en donde los usuarios pueden anotar y buscar documentos. La Figura 3 muestra el modelo de anotación basado en ontologías desarrollado en GoNTogle.

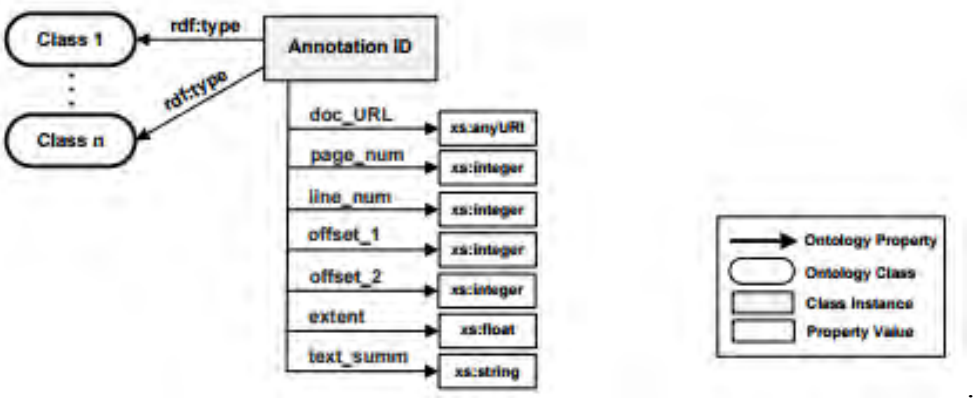

Figura 3: Modelo de anotación basado en ontología. Imagen recuperada de [18]

En el proyecto Neptuno [17], las anotaciones semánticas se realizaron usando vocabularios controlados. Los vocabularios controlados consisten en un esquema que exige el uso de términos predefinidos, autorizados por el creador del vocabulario (Noruzi, 2007). Estos términos predefinidos se usaron para clasificar el contenido de los documentos. Existen diversos estándares de vocabularios controlados en el campo periodístico (dominio en el que desarrolló su ontología) como NewsML, NITF, XMLNews, el IPTC *Subject Reference System* y PRISM; la mayoría basados en XML.

Luego de una evaluación de todos estos estándares, el proyecto decidió adoptar el IPTC *Subject Reference System*, el cual es un sistema de clasificación temática para contenido de archivos periodísticos y de noticias. Este estándar se planteó como una solución a los problemas que ocasionaban el sistema de clasificación anterior.

En adición al estándar IPTC, se incluyeron dos formas alternativas para clasificar más a fondo el contenido de los documentos. La primera se realiza de acuerdo al género, el cual tiene que ver con la naturaleza de una noticia (noticia de última hora, resumen, entrevista, opinión, encuesta, pronóstico, etc.) en lugar de su contenido específico. La segunda se realiza de acuerdo a algunas palabras claves que describan el contenido.

Por su parte, el proyecto LT4eL [20] implementó un proceso con dos tipos de anotaciones semántica: *inline* y a través de metadatos. En este último, la anotación es externa y es almacenada para su posterior uso al momento de indizar los

documentos. La persona encargada de esta anotación puede incluir todos los conceptos y relaciones que considere importantes para la clasificación del documento. Esta anotación se utiliza para la recuperación de documentos desde el repositorio. Por otro lado, la anotación *inline* es usada como una extensión de esta recuperación donde se consideran las apariciones de los conceptos dentro de todo el contenido del documento o de sus partes (párrafos, oraciones).

3.3.3 P3: Métodos/tecnologías/herramientas/mecanismos empleados en las implementaciones de buscadores semánticos de documentos digitales

[P3] ¿Qué métodos/tecnologías/herramientas/mecanismos se han empleado en las implementaciones de buscadores semánticos de documentos digitales? La Tabla 7 muestra el resumen de los estudios más relevantes encontrados que responden a esta pregunta.

Tabla 7: Resumen de estudios más relevantes que responden a la pregunta de revisión [P3]

Estudio	Arquitectura desarrollada	Mecanismos y herramientas en los procesos de búsqueda
[17]	- Base de conocimiento basada en su ontología - Módulo de búsqueda semántica - Módulo de visualización y navegación de la ontología	- Búsqueda semántica y textual por clases y campos - Uso del *framework* Jena en los módulos de búsqueda y visualización - Consultas RDQL a la base de conocimiento (ontología)
[18]	- Componente de Anotación Semántica - Servidor de Ontología - Componente de Indexación - Componente de Búsqueda	- Búsqueda textual (librería Lucene) - Búsqueda semántica - Búsqueda híbrida (resultó la más eficaz)

[19]	Sin información relevante	- Enfoque híbrido de búsqueda - Empleo de un algoritmo de "activación de propagación"
[20]	Sin información relevante	- Consultas en diferentes idiomas - Ontología con léxicos de ocho idiomas vinculados a ésta

A pesar de que todos los estudios y proyectos de buscadores semánticos están dentro del contexto de la Web Semántica, se observó que cada uno de ellos puede utilizar diferentes métodos, tecnologías, herramientas, mecanismos de acuerdo a sus necesidades y capacidades. Implementan buscadores sobre arquitecturas distintas, emplean algoritmos particulares en el proceso de búsqueda o introducen alguna funcionalidad extra para mejorar las búsquedas, entre otras cosas. El objetivo de esta sección es comparar distintos enfoques para poder evaluar cuál es el que más se adecúa a este proyecto de fin de carrera.

3.3.3.1 Arquitecturas desarrolladas para soportar búsquedas semánticas

La implementación de un buscador semántico no solo implica el desarrollar el proceso de búsqueda, sino además la construcción y almacenamiento de las ontologías, el proceso de anotación semántica, entre otras cosas. Se necesita desarrollar una arquitectura completa que contemple todos estos aspectos para así permitir una búsqueda semántica.

El proyecto Neptuno [17], por ejemplo, implementó una plataforma (ver Figura 4) que consiste en:

- Una base de conocimiento basada en una ontología para la descripción de información periodística. Antes del proyecto, se tenían bases de datos con millones de noticias acumuladas en los últimos años, por lo que se necesitó integrar estos documentos al nuevo sistema.

- Un módulo de búsqueda semántica.

- Un módulo para la visualización y navegación de contenido basado en ontologías.

Figura 4: Navegación y búsqueda en la plataforma Neptuno. Imagen tomada de [17]

GoNTogle [18], en cambio, desarrolla una arquitectura dividida en 4 componentes básicos (ver Figura 5):

- Componente de Anotación Semántica: Proporciona facilidades en cuanto a la anotación semántica de documentos. Se compone de 3 módulos: Visor de documentos, Visor de ontología y Editor de anotación.

- Servidor de Ontología: Almacena las anotaciones semánticas de los documentos en la forma de instancias de clases. Consiste de 2 módulos: Gestor de ontología y Base de Conocimiento.

- Componente de Indexación: Es el responsable de la indexación de los documentos mediante índices invertidos.

- Componente de Búsqueda: Permite a los usuarios buscar documentos usando los tres tipos de búsqueda mencionados.

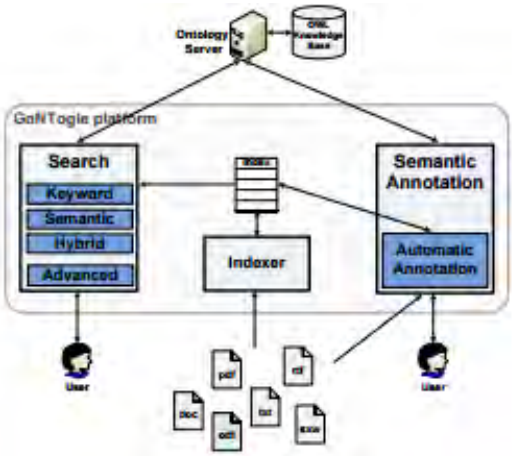

Figura 5: Arquitectura de GoNTogle. Imagen tomada de [18]

Como se puede apreciar en estos dos ejemplos, las arquitecturas son distintas. Sin embargo, ambas poseen todo lo necesario para una búsqueda semántica de documentos digitales.

3.3.3.2 Mecanismos y herramientas usados en los procesos de búsqueda

Se identificó que los distintos estudios y proyectos revisados usan diferentes mecanismos y herramientas en sus procesos de búsqueda. Algunos desarrollan búsquedas puramente semánticas, mientras otros búsquedas hibridas (combinación entre semántica y sintáctica); algunos utilizan algún algoritmo en particular para realizar sus búsquedas; algunos introducen funcionalidades extras en las consultas para mejorar los resultados. A continuación, se describirá el funcionamiento de las búsquedas en algunos proyectos relevantes para así evidenciar la diversidad de mecanismos y herramientas que se pueden desarrollar o utilizar.

En el proyecto Neptuno [17], el módulo de búsqueda fue desarrollado siguiendo los principios de búsqueda semántica. Sin embargo, el módulo también combina búsqueda directa por clases y campos, con la posibilidad de navegar la taxonomía IPTC, según la cual los archivos de noticias y documentos son clasificados.

Este módulo opera directamente sobre la base de conocimientos en RDF. El usuario plantea solicitudes de búsqueda a través de una interfaz web en la que selecciona la clase de contenido que desea buscar (Noticia, Fotografía, Gráficos o Página), y especifica palabras clave para los campos deseados (título, autor, sección, fecha,

tema, etc.). Esta información es enviada al servidor de Neptuno donde la solicitud es formalizada como una consulta RDQL, la cual se ejecuta en la base de conocimientos. El acceso a la base de conocimientos desde los módulos de búsqueda y visualización se lleva a cabo por medio de la librería Jena para RDF.

El proyecto GoNTogle [18] también proporciona servicios de búsqueda más allá de la búsqueda tradicional basada en palabras clave. Propone una combinación flexible de búsqueda basada en palabras clave y búsqueda semántica junto con operaciones avanzadas basadas en ontologías. GoNTogle soporta tres tipos de búsquedas:

- Búsqueda basada en palabras clave: Este es el modelo de búsqueda tradicional. Se adopta la métrica de similitud textual utilizado en el motor de IR Lucene.

- Búsqueda semántica: Este tipo de búsqueda permite al usuario navegar a través de las clases de una ontología y enfocar su búsqueda en una o más de estas clases.

- Búsqueda híbrida: El usuario puede buscar documentos utilizando palabras clave y las clases de la ontología. Puede, también, determinar si el resultado de su búsqueda será la intersección o la unión de las dos búsquedas. Las evaluaciones experimentales del proyecto validan que este método es el más eficaz en comparación con los otros dos.

Al igual que en GoNTogle; Berrueta, Labra y Polo (2006) también desarrollaron un enfoque híbrido de búsqueda, pero de una manera distinta. En primer lugar, se tiene la consulta compuesta de un conjunto de conceptos elegidos por el usuario a través de una interfaz que oculta la complejidad de las ontologías subyacentes (véase el paso 1 en la Figura 6). Luego, el proceso de búsqueda transforma automáticamente la consulta semántica en una consulta sintáctica equivalente, mediante un algoritmo denominado de “activación de propagación”. Este algoritmo recorre la ontología a modo de grafo, explotando las relaciones en ésta y retornando una lista de conceptos estrechamente relacionados semánticamente a los conceptos ingresados en la consulta.

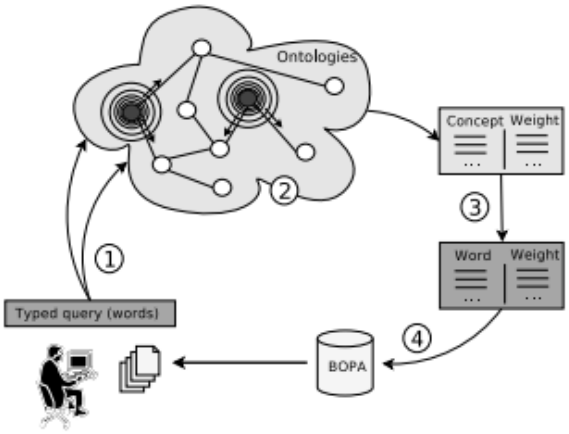

Figura 6: Representación del proceso de búsqueda semántica. Imagen tomada de [19]

En el tercer paso, la lista de conceptos es transformada a una lista de palabras, teniendo en cuenta que cada concepto tiene un conjunto de sinónimos asociado. Finalmente en el cuarto y último paso, una consulta de búsqueda sintáctica es construida y ejecutada sobre los documentos XML usando un motor de búsqueda convencional.

Otro proceso de búsqueda semántica destacable es el del proyecto LT4eL [20]. En esta búsqueda, las consultas pueden ser realizadas en diferentes idiomas. Esto es posible ya que el proyecto desarrolló una ontología independiente del lenguaje, con léxicos de ocho idiomas vinculados a ésta.

El flujo de datos desde la consulta del usuario hasta la recuperación de los documentos se da de la siguiente manera: a) Los términos de búsqueda son buscados en el léxico del idioma elegido. b) Los elementos léxicos encontrados son mapeados a los correspondientes conceptos de la ontología. c) Una vez que se elige un conjunto de conceptos, aquellos documentos que los contienen se presentan al usuario.

3.3.4 P4: Estudios y proyectos de buscadores semánticos en el dominio de la lingüística

[P4] ¿Qué estudios o proyectos de buscadores semánticos en el dominio de la lingüística se han realizado? No se encontraron estudios o proyectos de buscadores

semánticos; sin embargo, se encontró un proyecto el cual ha desarrollado una ontología en el dominio de la lingüística.

Como parte del proyecto denominado *Electronic Metastructure for Endangered Languages Data* (EMELD), se desarrolló una ontología de conceptos denominada GOLD [25] que abarca una amplia gama de fenómenos lingüísticos. GOLD es una ontología que formaliza las categorías y relaciones más básicas utilizadas en la descripción científica del lenguaje humano. Está diseñada para capturar el conocimiento de un lingüista bien entrenado, y por lo tanto, puede ser vista como un intento de codificar el conocimiento general del campo. Además, facilitará el razonamiento automatizado sobre datos lingüísticos con el objetivo de solucionar problemas encontrados en proyectos de bases de datos tipológicas y de procesamiento de lenguaje natural. Finalmente, GOLD pretende ser compatible con los objetivos generales de la Web Semántica.

3.4 Conclusiones sobre el estado del arte

Luego de describir en la sección anterior los distintos aspectos implicados en los buscadores semánticos de documentos digitales, se tiene una idea más clara de los avances que se han realizado en los últimos años en este tema.

Una de las maneras en que los buscadores semánticos obtienen conocimiento de un determinado dominio es a través de ontologías. Para el caso de este proyecto de fin de carrera, el dominio será el de la lingüística. Luego de esta revisión, los diferentes puntos o aspectos considerados por proyectos pasados al momento de construir sus ontologías pueden ser tomados en cuenta para el modelamiento de una ontología en dicho dominio.

En cuanto a los documentos digitales, su contenido es estructurado por medio de las anotaciones semánticas. Se necesita definir el proceso de anotación semántica de documentos lingüísticos que se empleará en este proyecto, para lo cual se puede aprovechar los métodos utilizados en proyectos pasados.

Además, se pudo apreciar que en el proceso de búsqueda se pueden emplear distintos mecanismos, herramientas o algunas consideraciones adicionales. Éstos pueden ser

evaluados con el objetivo de determinar si alguno de ellos puede ser utilizado situándolo al contexto de este proyecto.

En general, las distintas soluciones planteadas en contextos diferentes pueden ahora ser analizadas para aprovecharlas adaptándolas de la mejor manera a la problemática de este proyecto.

4 ONTOLOGÍA EN EL DOMINIO DE LA LINGÜÍSTICA

4.1 Ontología que representa y provee conocimiento en el dominio de la lingüística

La construcción de la ontología se realizó mediante la metodología *Ontology Development 101*, consistente en 7 pasos, los cuales se desarrollarán a continuación. Además, se usó la herramienta Protégé, la cual permitió desarrollar la ontología mediante una interfaz gráfica y finalmente guardarla en lenguaje OWL.

4.1.1 Paso 1: Determinar el dominio y alcance de la ontología

Se recomienda iniciar el desarrollo de una ontología mediante la definición de su dominio y alcance (Noy & McGuinness, 2001). Esto se puede realizar respondiendo a preguntas básicas como las siguientes:

- ¿Cuál es el dominio que la ontología cubrirá? (Noy & McGuinness, 2001)
  El dominio que se cubrirá es el de la lingüística. Sin embargo, este dominio es muy amplio por lo que se necesita limitar su alcance; es decir, reducir considerablemente la cantidad de conceptos o partes de la lingüística que se van a cubrir. Esto se logra respondiendo a las siguientes preguntas.

- ¿Para qué se usará la Ontología? (Noy & McGuinness, 2001)
  La ontología se usará para la búsqueda semántica de documentos en el dominio de la lingüística. Los documentos para los que el buscador será implementado pertenecen al Departamento de Humanidades de la universidad. Con el objetivo de limitar el dominio a representar, se necesitó conocer qué partes o subáreas de la lingüística cubren los temas y conceptos que tratan estos documentos. En una reunión con el profesor de lingüística ayudante en el proyecto, se determinó que estos documentos cubren lo que es lingüística descriptiva. La lingüística descriptiva es una parte de la lingüística que estudia la descripción de las estructuras fonológicas, gramaticales y semánticas de las lenguas en un momento determinado de la historia [27]. Por lo tanto, se definió que el alcance de la ontología no es todo el dominio de la lingüística, sino la lingüística descriptiva.

- ¿Para qué tipo de preguntas la ontología debe proveer respuestas? (Noy & McGuinness, 2001)

  La ontología debe poder responder a consultas que hagan referencia a temas de la lingüística descriptiva, por lo cual la ontología debe contener conceptos de esta área y modelar sus relaciones.

- ¿Quién podrá usar y mantener la ontología? (Noy & McGuinness, 2001)

  El uso y mantenimiento de la ontología será hecho por el autor de esta.

Estas preguntas fueron respondidas mediante entrevistas que se programaron con un profesor de lingüística de la universidad. Estas entrevistas permitieron recabar la información necesaria de las necesidades de búsqueda del Departamento de Humanidades con respecto a sus documentos lingüísticos. Luego de responder estas preguntas, se tiene definido el dominio y alcance de la ontología que se desarrollará en este proyecto.

4.1.2 Paso 2: Considerar reusar ontologías existentes

Casi siempre vale la pena considerar lo que otras personas han hecho y verificar si podemos refinar y extender las fuentes existentes para nuestro domino y tarea particular (Noy & McGuinness, 2001). Para este paso de la metodología, se buscaron ontologías existentes que pudieran ser reusadas. Esto se realizó mediante la revisión del estado del arte descrita en el capítulo anterior.

La única ontología que se encontró fue la ontología GOLD. Sin embargo, no se consideró relevante usar esta ontología, ya que modela conceptos de toda el área de la lingüística en general y no profundiza en temas de la lingüística descriptiva, lo cual es necesario para los fines de este proyecto (los documentos profundizan en temas pertenecientes a lingüística descriptiva). Además, no es una ontología diseñada para búsquedas semánticas. La ontología GOLD fue desarrollada para solucionar problemas encontrados en proyectos de bases de datos tipológicas y de procesamiento de lenguaje natural (NLP) (Farrar & Langendoen, 2003).

4.1.3 Paso 3: Enumerar términos importantes en la ontología

Es útil escribir una lista de todos los términos acerca de los que nos gustaría declarar o explicar al usuario. ¿Cuáles son los términos de los que nos gustaría hablar? ¿Qué propiedades tienen estos términos? ¿Que nos gustaría decir sobre estos términos? (Noy & McGuinness, 2001)

En este paso, se realizó una lista de los términos más importantes que representen conceptos o temas de la lingüística descriptiva. Esto se realizó mediante las entrevistas que se programaron con un profesor de lingüística de la universidad. Estas entrevistas permitieron recabar la información necesaria para identificar términos importantes de la ontología. Algunos de estos términos fueron los siguientes:

- Morfología
- Sintaxis
- Fonética
- Fonología
- Pragmática
- Semántica
- Léxico
- Verbo
- Clases de palabras
- Clases abiertas
- Clases cerradas
- Adjetivo
- Adverbio
- Verbo
- Pronombre
- Determinante
- Conjunción
- Interjección
- Relación gramatical
- Manipulación de la valencia
- Posesión
- Imperativo
- Negación
- Transitividad
- Entonación
- Inventario fonológico
- Procesos fonológicos
- Morfofonémica
- Lista léxica
- Vocabulario

Estos son algunos de los conceptos que se identificaron para modelar el dominio. Hacer una lista de estos conceptos ayudó a tener una idea de las clases que se tendrán en la ontología y luego desarrollar la taxonomía, lo cual se realiza en el siguiente paso.

4.1.4 Paso 4: Definir las clases y la jerarquía de clases

Para el desarrollo de la jerarquía de clases, se hizo uso del proceso de desarrollo de arriba hacia abajo o *top-down*, el cual es un enfoque que comienza con la definición de los conceptos más generales en el dominio para su posterior especialización. (Noy and McGuiness, 2001). La Figura 7 muestra la jerarquía de clases que se desarrolló en la ontología, la cual es una taxonomía de conceptos pertenecientes a la lingüística descriptiva. Esta jerarquía presenta un gran número de clases (175). Algunas de éstas se muestran en la figura.

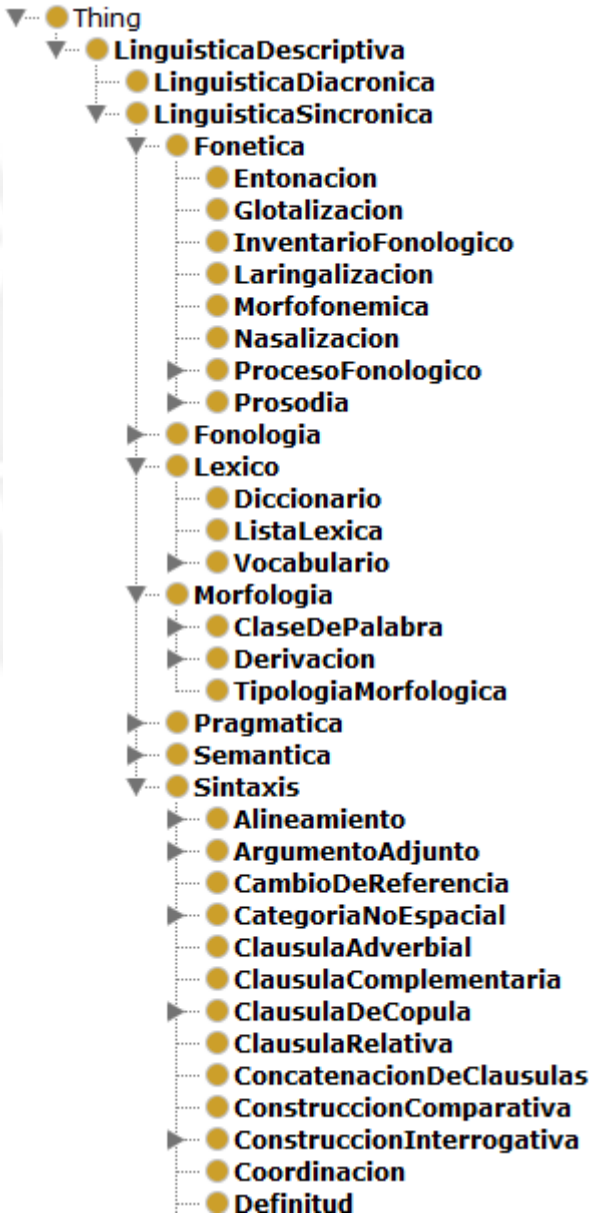

Figura 7: Jerarquía de clases de la ontología. Autoría propia.

4.1.5 *Paso 5 y 6: Definir las propiedades de las clases (ranuras) y Definir las facetas de las ranuras*

Para los Pasos 5 y 6, se analizaron los conceptos definidos en la jerarquía de clases con el objetivo de asignarles propiedades (las cuales son llamadas "ranuras" en esta metodología) de manera que puedan potenciar aún más las búsquedas semánticas. Las propiedades de las clases pueden tener los siguientes valores (los cuales son llamados "facetas" en esta metodología): cadenas de caracteres, números o instancias de otras clases (lo que define la relación de un clase con otra). Sin embargo, no se vio necesaria la creación de propiedades en ninguna clase ya que se determinó que la taxonomía desarrollada era suficiente para cumplir con los objetivos para los que se creó la ontología, los cuales fueron planteados en las respuestas a las preguntas que se mencionaron en el Paso 1.

Sin embargo, se identificó que algunas relaciones de clase y subclase en la ontología podían ser modeladas también a través de propiedades. Por ejemplo, se tiene la clase "Adjetivo" el cual tiene como subclases "CategoriaDeAdjetivo" y "ClaseDeAdjetivo". Estas dos subclases podrían haber sido modeladas como propiedades de la superclase, cuyos nombres podrían ser "tieneCategoria" y "tieneClase" respectivamente. Este enfoque en el modelado también se consideró válido. Sin embargo, se eligió modelarlas con una relación jerárquica debido a que brinda un mejor entendimiento de los conceptos, siendo la categoría y clase de un adjetivo subconceptos o conceptos más pequeños del concepto más grande que es Adjetivo. Además, un modelado en esta forma facilita las tareas de búsqueda semántica, las cuales son el propósito de la ontología. Como este ejemplo, se encontraron varios casos similares.

4.1.6 *Paso 7: Crear instancias*

Este último paso consiste en la creación de las instancias para las clases definidas en los pasos anteriores. Se analizaron cada una de estas clases para evaluar la relevancia de la creación de sus instancias. Se determinó que era posible crear instancias para las clases pero que no resultaría relevante para la aplicación; es decir, para las búsquedas semánticas. Por ejemplo, se tiene la clase "Adverbio" para la cual sus instancias serían la infinidad de adverbios que existen en el lenguaje. Como este

ejemplo, se identificó la misma característica para la mayoría de clases. Estas instancias no brindan ningún valor agregado a la aplicación, ya que no resulta relevante anotar los documentos ni realizar consultas con éstas. Por lo tanto, no se crearon instancias en esta ontología.

4.2 Pruebas de consistencia de la ontología

Este resultado esperado pretende verificar la consistencia de la ontología en lenguaje OWL desarrollada en la sección anterior. Para esto, se hizo uso de la herramienta Protégé. Esta herramienta permite realizar consultas a la ontología mediante un razonador. Un razonador semántico, motor de razonamiento, motor de reglas, o simplemente un razonador, es una pieza de software capaz de inferir consecuencias lógicas a partir de un conjunto de hechos afirmados o axiomas (Russell & Cohn, 2012). De esta manera, mediante un razonador se puede comprobar que las inferencias en la ontología se están realizando de manera correcta.

Protégé permite usar diferentes razonadores. El que se usó para esta ocasión es el razonador Pellet. Las consultas a la ontología se realizaron usando la sintaxis denominada Manchester. Se realizaron diferentes consultas que permitieron verificar la consistencia de la ontología. La Figura 8 muestra una de estas consultas y el resultado que se obtuvo. Se consultó la clase “Morfología” y se infirieron correctamente todas sus subclases.

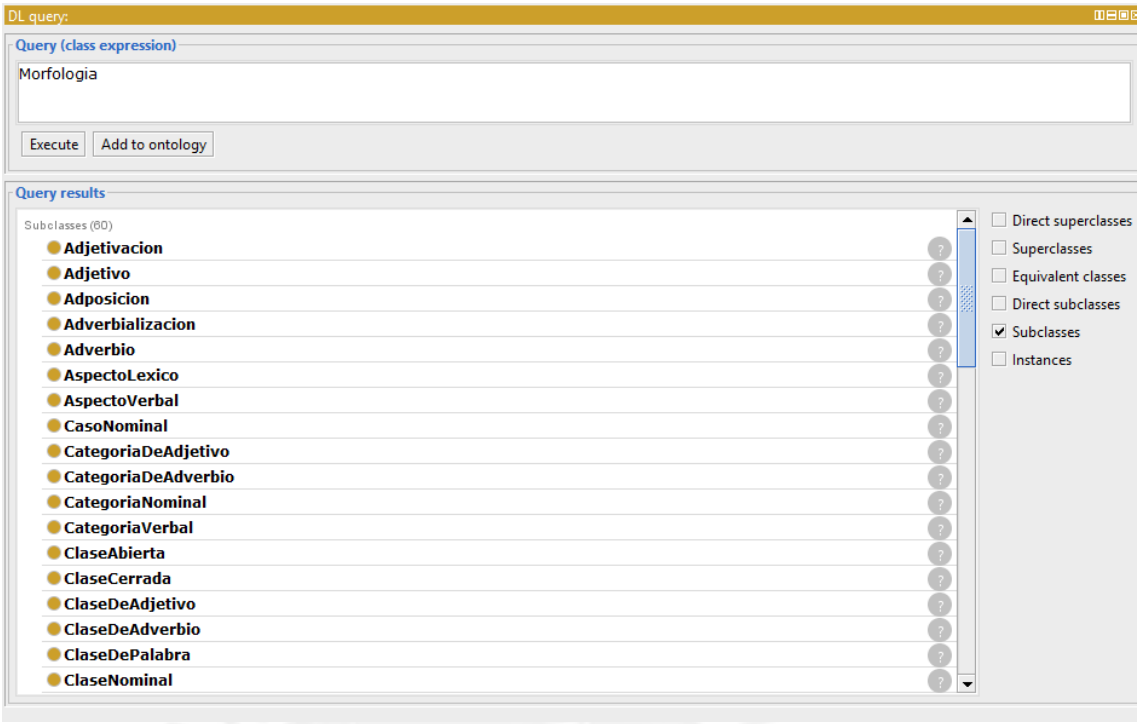

Figura 8: Consulta de las subclases de la clase "Morfología". Autoría propia.

Finalmente, se obtuvo la ontología con un total de 175 clases. Las Figuras 9, 10 y 11 muestran partes de la ontología.

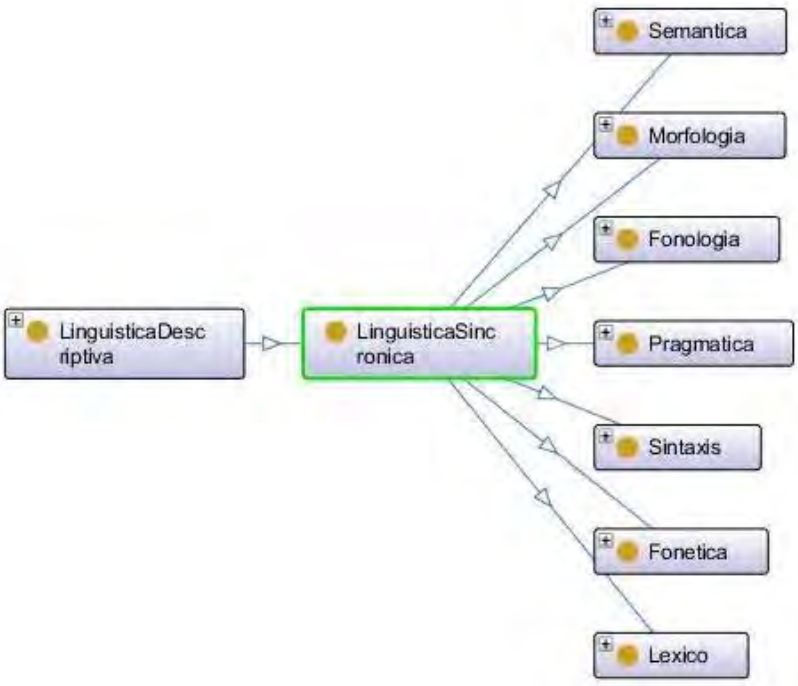

Figura 9: Subclases directas de la clase "Lingüística Sincrónica". Autoría propia.

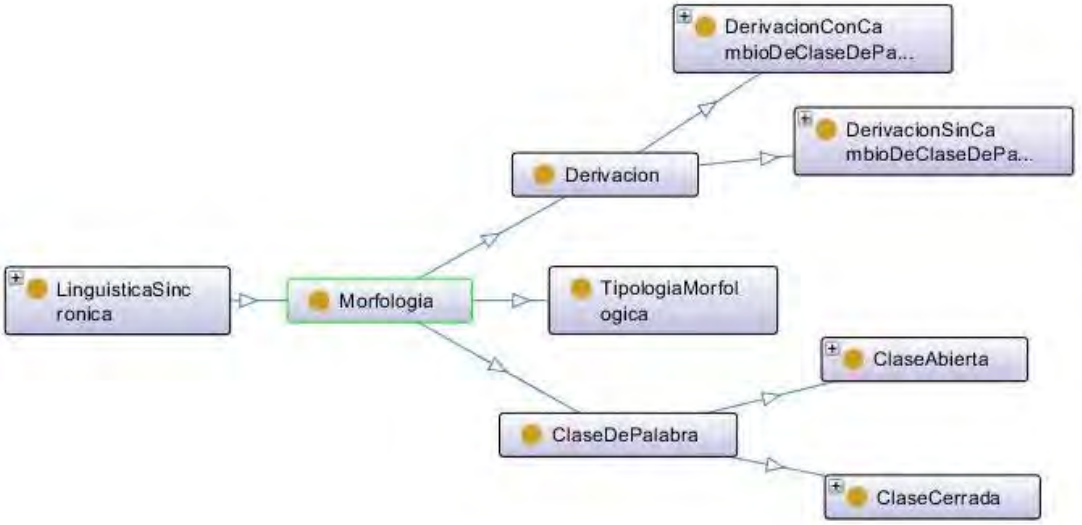

Figura 10: Algunas subclases de la clase "Morfología". Autoría propia.

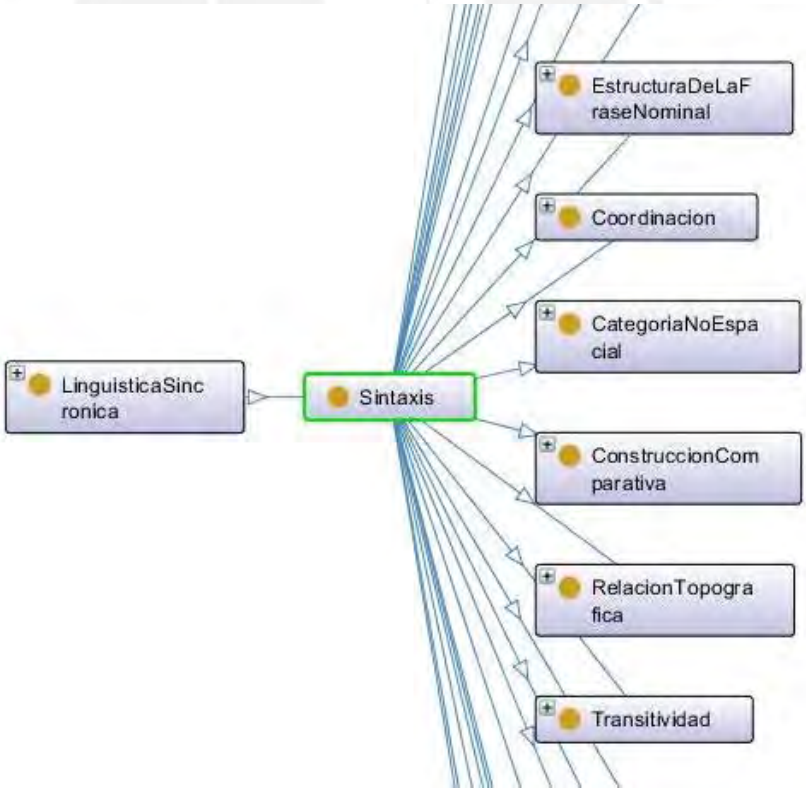

Figura 11: Subclases directas de la clase "Sintaxis". Autoría propia.

5 ANOTACIÓN SEMÁNTICA DE LOS DOCUMENTOS

5.1 Definición del proceso manual de anotación semántica externa de los documentos

En este resultado esperado se definió el proceso de anotación semántica que se seguirá e implementará el presente proyecto. Este proceso permite que las anotaciones se realicen de dos maneras distintas: a través de un archivo en formato csv o ingresando directamente las anotaciones para un determinado documento. Es decir, el proceso puede recibir dos tipos de entradas o *inputs*. A continuación, se describe cómo se reciben las anotaciones en cada uno de estos dos casos:

- Mediante un archivo csv: Las anotaciones se registran junto con el registro de un nuevo documento. El proceso de anotación recibirá un archivo en formato csv, el cual contiene por cada fila la información necesaria para realizar las anotaciones a un determinado documento. Esta información consistirá en la URL del documento, título, autor, tipo de documento (1 = artículo, 2 = diccionario, 3 = gramática, 4 = tesis, 5 = libro), y seguido de los diferentes conceptos que se quieren anotar a este documento. La Figura 12 muestra un ejemplo del contenido que puede tener este tipo de archivo.

- El proceso también permitirá para un determinado documento, el cual ya ha sido registrado, recibir las anotaciones que se le desean realizar ingresándolas directamente una por una. Esta forma de anotación resulta útil cuando se quiere anotar un documento de manera rápida.

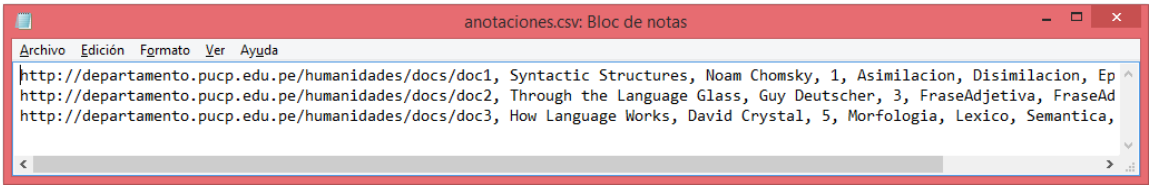

Figura 12: Ejemplo de archivo csv para realizar anotaciones semánticas. Autoría propia.

Luego de recibir las anotaciones semánticas en cualquiera de las dos formas, se verificará que estas anotaciones sean conceptos de la ontología desarrollada. En caso no lo sean, no serán anotadas a los documentos.

Finalmente, las anotaciones serán almacenadas en una base de datos separada de los documentos (anotación externa), la cual relacionará la información de los documentos con sus respectivas anotaciones. Esto permitirá que las búsquedas semánticas que se realizarán posteriormente puedan recuperar estas anotaciones y los documentos relacionados. En la Figura 13 se puede apreciar de manera resumida el proceso manual de anotación semántica externa que se ha definido.

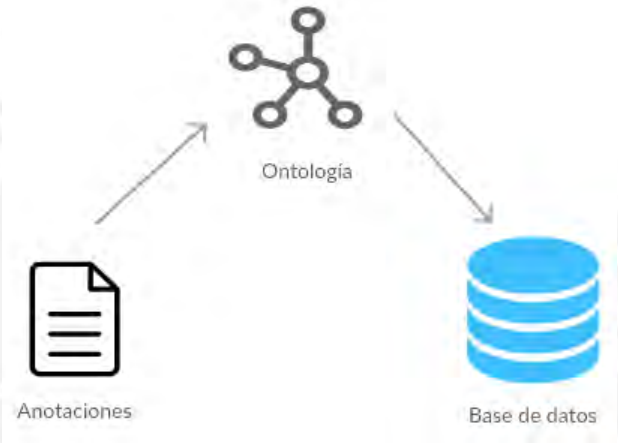

Figura 13: Proceso de anotación semántica. Autoría propia.

5.2 Estructura para la persistencia de las anotaciones semánticas

Para la persistencia de las anotaciones semánticas se hará uso de la base de datos MySQL *Community Edition*. Esta base de datos almacenará la información necesaria requerida por el buscador semántico que le permita recuperar las anotaciones y finalmente la URL de los documentos relacionados. La estructura que se definió consiste en tres tablas: TipoDocumento, Documento y Anotación. En la Figura 14 se pueden ver las tablas y sus relaciones.

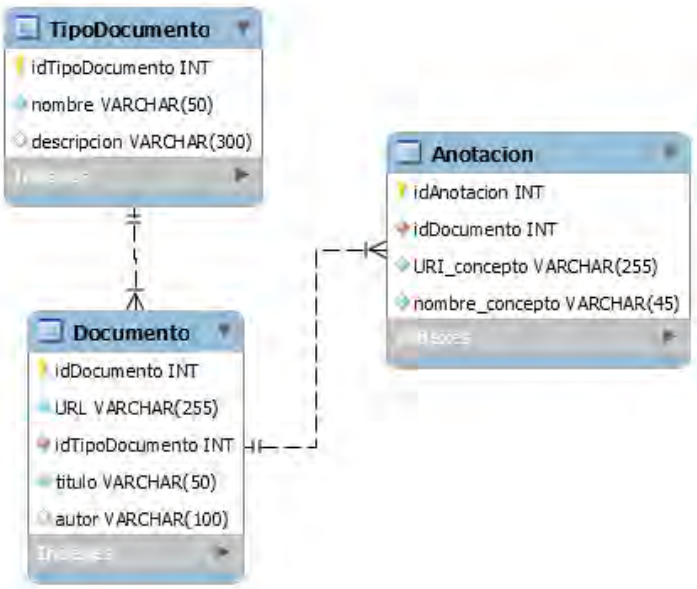

Figura 14: Base de datos para la persistencia de las anotaciones semánticas. Autoría propia.

La tabla TipoDocumento almacenará la información de los tipos de documentos que pueden existir. Estos son los siguientes: artículo, diccionario, gramática, tesis y libro. La tabla Documento almacenará la siguiente información de los documentos: la URL del documento, el tipo de documento e información básica como título y autor. La última tabla, Anotación, almacenará la relación entre un documento y un concepto de la ontología, pudiendo un documento tener muchos conceptos asociados. Los conceptos son representados mediante su URI en la ontología, ya que esto permite identificarlos de manera única. Esta tabla es la más relevante, ya que es la que representa las anotaciones semánticas.

5.3 Aplicación que permite anotar manualmente los documentos con elementos de la ontología

La aplicación o herramienta que soporta el proceso de anotación semántica definido anteriormente y usa la base de datos definida para la persistencia de estas anotaciones fue desarrollada como un módulo aparte al de búsqueda semántica. El objetivo de desarrollar esta herramienta como un módulo es que permite mantener la información de las anotaciones en el tiempo; es decir, las anotaciones de los documentos pueden ser añadidas y eliminadas en cualquier momento por usuarios de

la aplicación designados para estas tareas, y no solo por el administrador del sistema. Este enfoque brinda tres principales ventajas:

- Permite que el o los usuarios puedan ir afinando las anotaciones con el tiempo; es decir, en caso cambien de opinión, pueden modificar las anotaciones de un documento según sus conocimientos del dominio de la lingüística. De esta manera, se puede mejorar la precisión de las búsquedas semánticas.
- En caso la aplicación sea usada por varios usuarios, permite un trabajo conjunto y cooperativo para las anotaciones de los documentos, lo que puede también mejorar la precisión de las búsquedas debido a los conocimientos de varias personas en el dominio.
- En caso se añadan nuevos documentos al repositorio en donde se almacenan, permite que los usuarios puedan agregar estos nuevos documentos a la aplicación y anotarlos, evitando así la necesidad de que algún administrador del sistema tenga que realizar estas tareas. De esta manera, se puede ir extendiendo la cantidad inicial de documentos, permitiendo que el sistema crezca.

Este enfoque requiere que solo usuarios designados puedan tener acceso a esta aplicación y no cualquier usuario. Esto debido a que las tareas de anotación semántica deben ser realizadas de manera consiente y cuidadosa, ya que afectan directamente la precisión de las búsquedas semánticas. Debido a esto, se implementó un mecanismo de autenticación basado en usuario y contraseña para acceder al módulo de anotación semántica. De esta manera, para acceder a este módulo, el usuario debe primero autenticarse.

Una vez que el usuario se autentica y accede al módulo; la aplicación permite, como se definió en el proceso, anotar los documentos mediante un archivo en formato csv o ingresar directamente las anotaciones para un determinado documento que ya se encuentra registrado.

El primer modo de anotación se implementó mediante un formulario consistente en un *input* para seleccionar la ruta del archivo que se desea usar (ver Figura 15). Este formulario solo permite archivos con extensión ".csv".

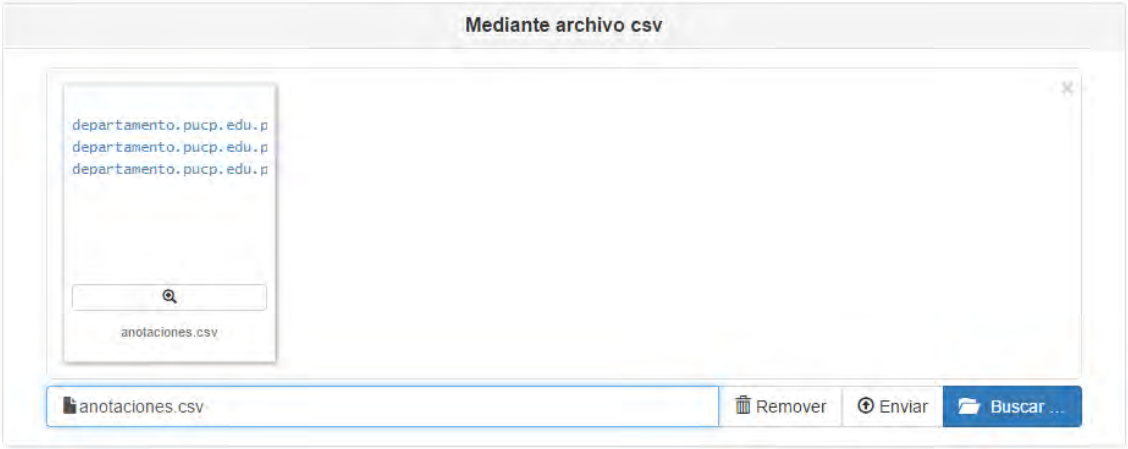

Figura 15: Anotación semántica mediante un archivo en formato csv. Autoría propia.

Para la otra forma de anotación, se implementó una interfaz la cual se puede apreciar en la Figura 16. Esta interfaz consiste de un *input* para la búsqueda mediante la URL del documento que se quiere anotar. En caso se encuentre un documento registrado con dicha URL, se muestran las anotaciones del documento y otro *input* para ingresar una nueva anotación. Al momento de ingresar una nueva anotación, se le ayuda al usuario mostrándole una lista con las entidades de la ontología que coinciden con el texto ingresado por éste, de manera que pueda elegir una de estas entidades (ver Figura 17). Esta forma de anotación resulta útil cuando se quiere ver y modificar la información de un solo documento de manera gráfica y rápida.

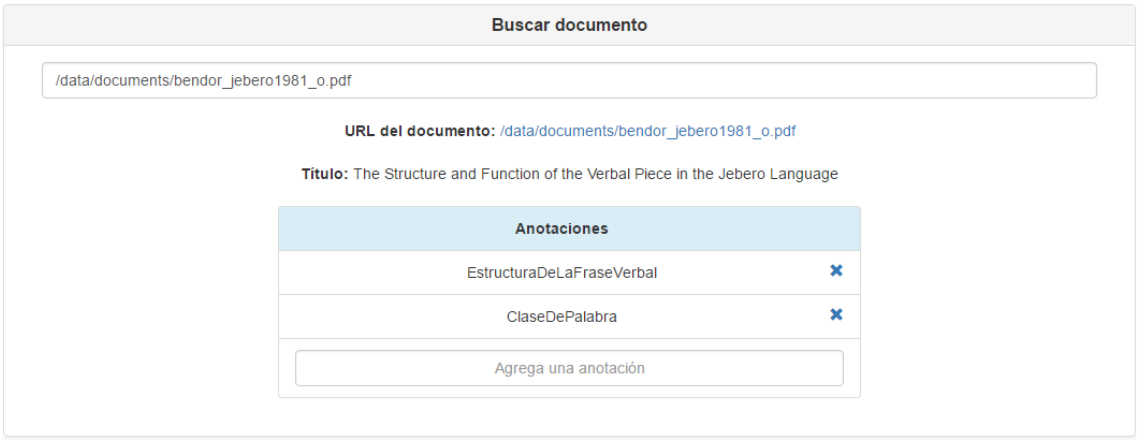

Figura 16: Ingreso manual de las anotaciones semánticas para un determinado documento. Autoría propia.

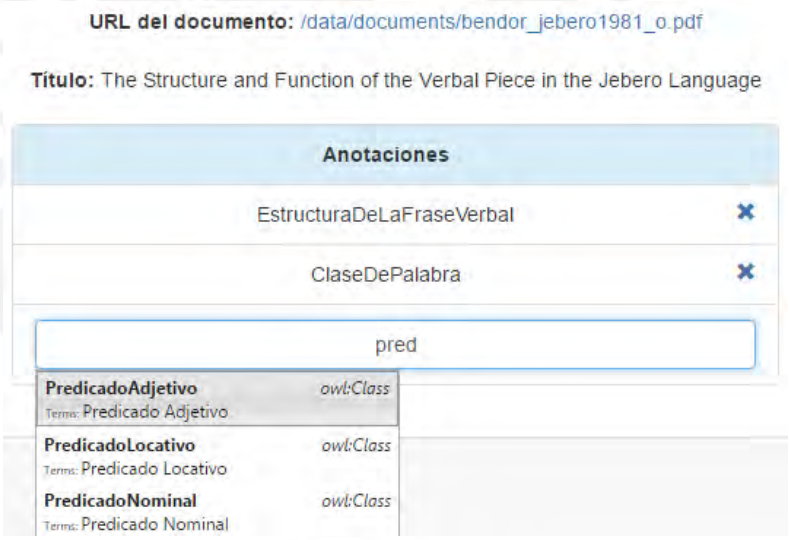

Figura 17: Ingreso de una anotación. Autoría propia.

Luego de ingresar las anotaciones por cualquiera de las dos maneras, se verifica que estas anotaciones sean entidades de la ontología. En caso se quiera anotar con palabras que no representan ninguna entidad de la ontología, la aplicación simplemente no considera esta anotación. Por último, las anotaciones son almacenadas en una base de datos MySQL con la estructura que se definió en la sección anterior.

Finalmente, se tiene el módulo de anotación semántica implementado, el cual se muestra en la Figura 18.

Figura 18: Módulo de anotación semántica. Autoría propia.

6 BUSCADOR SEMÁNTICO EN EL DOMINIO DE LA LINGÜÍSTICA

En esta sección, se presentan los resultados esperados que permitieron la implementación de las tareas de búsqueda. A continuación, se explican unas consideraciones importantes que permiten tener un mejor entendimiento del funcionamiento del buscador semántico.

El buscador semántico implementado requiere que el usuario ingrese en las consultas un concepto de la ontología desarrollada, para proceder con la búsqueda semántica a partir de este concepto. Sin embargo, si el usuario no ingresa en la consulta un concepto de la ontología, no se puede realizar directamente una búsqueda semántica correspondiente. Cuando sucede esto, se realiza una búsqueda textual previa con el objetivo de encontrar el o los conceptos de la ontología que se asemejen textualmente a la consulta ingresada. Luego, se realiza la búsqueda semántica a partir de estos conceptos encontrados. Para poder implementar el buscador textual es necesaria la construcción de un indexador de conceptos de la ontología sobre el cual se realicen estas búsquedas basadas en texto.

De esta manera; la implementación del indexador, del buscador basado en texto y del buscador semántico mencionados en el párrafo anterior corresponden a los tres siguientes resultados esperados. El último resultado esperado consiste en la evaluación de las búsquedas semánticas empleando los mecanismos de precisión y *recall*.

6.1 Indexación de los documentos mediante elementos de la ontología

Como se mencionó anteriormente, además de una búsqueda semántica, la aplicación implementada también realiza una búsqueda textual en caso el usuario no ingrese directamente un concepto de la ontología en su consulta. Esta búsqueda textual requiere de la implementación de un mecanismo que indexe los conceptos de la ontología.

La indexación se realizó, entonces, mediante las entidades o conceptos de la ontología desarrollada. De esta manera, luego el buscador textual puede pasar los

conceptos encontrados en el índice al buscador semántico, y éste recuperar los documentos relacionados.

Para la implementación de este resultado esperado se utilizaron el *framework* Jena y la librería Lucene. Jena permitió recorrer la jerarquía de clases en la ontología con el objetivo de extraer los conceptos que serán indexados. Luego de esto, se implementa el indexador, utilizando la librería Lucene, de la siguiente manera: por cada concepto extraído se crea un documento (el cual representa a un índice) y se le agrega dos campos. Estos campos son: el URI del concepto en la ontología y un campo denominado "contents", el cual contiene datos sobre el nombre del concepto para hacer el *match* con la consulta del usuario. Luego de ser creados todos los documentos, éstos son colocados en un *dataset*, el cual estará corriendo en RAM mientras la aplicación esté en funcionamiento.

En la Figura 19 se muestra cómo se implementó el indexador, extrayendo primero las entidades de la ontología para luego construir los índices con los datos de éstas.

```
private void indexDocs() throws IOException{

    analyzer = new SpanishAnalyzer();

    // Almacenar el índice en memoria
    directory = new RAMDirectory();

    IndexWriterConfig config = new IndexWriterConfig(analyzer);
    IndexWriter iwriter = new IndexWriter(directory, config);

    // obtener las clases a indexar
    OntClass root_class = model.getOntClass( namespace + "LinguisticaDescriptiva" );
    ExtendedIterator classes = root_class.listSubClasses();

    // indexar las clases
    while (classes.hasNext()){
        OntClass thisClass = (OntClass) classes.next();

        org.apache.lucene.document.Document doc = new org.apache.lucene.document.Document();
        doc.add(new Field("URI", thisClass.getURI(), TextField.TYPE_STORED));
        String contents = thisClass.getLocalName() + " " + Helper.splitCamelCase(thisClass.ge
        doc.add(new Field("contents", contents, TextField.TYPE_STORED));

        iwriter.addDocument(doc);
    }

    iwriter.close();
}
```

Figura 19: Código para la indexación. Autoría propia.

6.2 Herramienta de software que permite realizar búsquedas basadas en texto previas a las búsquedas semánticas

En el caso que el usuario no ingrese directamente un concepto de la ontología en la consulta, no se puede determinar inequívocamente el concepto que éste desea buscar. Cuando sucede esto, se realiza una búsqueda que se encarga de encontrar los conceptos de la ontología que se asemejan textualmente a la consulta ingresada, para luego pasar estos conceptos encontrados al buscador semántico. La herramienta que permite esta búsqueda basada en texto se implementa en el presente resultado esperado utilizando la librería Lucene. El objetivo de esta herramienta es poder brindarle al usuario resultados a pesar de no ingresar exactamente un concepto de la ontología.

La herramienta funciona de la siguiente manera:

1. La herramienta recibe la consulta ingresada por el usuario.

2. La herramienta busca en el índice (construido en el resultado esperado anterior) los conceptos de la ontología que coincidan con la consulta del usuario.

3. La herramienta pasa los conceptos encontrados al buscador semántico.

La Figura 20 muestra el método que realiza la búsqueda en el índice. Este método recibe como parámetro el texto de la consulta del usuario y retorna una lista con las URIs de los conceptos encontrados.

```
private List<String> search_index(String search_string) throws IOException, ParseException{
    List<String> result = new ArrayList<String>();

    // Buscar en el índice
    DirectoryReader ireader = DirectoryReader.open(directory);
    IndexSearcher isearcher = new IndexSearcher(ireader);
    // Parsear la consulta con el texto de búsqueda
    QueryParser parser = new QueryParser("contents", analyzer);
    Query query = parser.parse(search_string);
    ScoreDoc[] hits = isearcher.search(query, null, 1000).scoreDocs;

    for (int i = 0; i < hits.length; i++) {
        org.apache.lucene.document.Document hitDoc = isearcher.doc(hits[i].doc);
        System.out.println(hitDoc.get("URI") + " - " + hits[i].score);
        result.add(hitDoc.get("URI"));
    }
    ireader.close();

    return result;
}
```

Figura 20: Código para la búsqueda textual en el índice. Autoría propia.

La Figura 21 muestra la prueba que se hizo de la herramienta ingresando en la consulta el texto "clausula". Esta prueba devuelve como resultado las URIs de los conceptos encontrados con un respectivo *score*. Este *score* determina cuánto tiene que ver el texto de la consulta en relación con el nombre del concepto. Cuanto más alto es el *score*, la relevancia del resultado es mayor.

```
#ClausulaComplementaria - 2.0398068
#ClausulaDeCopula - 2.0398068
#ClausulaPosesiva - 2.0398068
#ClausulaSinVerbo - 2.0398068
#ClausulaAdverbial - 2.0398068
#ConcatenacionDeClausulas - 2.0398068
#ClausulaRelativa - 2.0398068
```

Figura 21: Resultados de una búsqueda textual. Autoría propia.

Una ventaja importante que brinda esta herramienta es que permite que el usuario pueda ingresar más de un concepto de la ontología en su consulta y obtener resultados relevantes. Por ejemplo, el usuario puede ingresar la siguiente consulta: "morfología sintaxis y léxico". Esta consulta no puede ser recibida directamente por el buscador semántico, ya que no encontrará en la ontología un concepto que tenga de nombre todo ese texto ("morfología sintaxis y léxico"). La herramienta de búsqueda textual permitirá encontrar los conceptos que el usuario quiso ingresar.

Esta herramienta encontrará los conceptos “Morfología”, “Sintaxis” y “Léxico (entre otros, posiblemente) y los pasará al buscador semántico. De esta manera, el usuario puede realizar búsquedas con más de un solo concepto.

6.3 Aplicación que permite realizar búsquedas semánticas de documentos en el dominio de la lingüística

Esta aplicación o herramienta permite realizar las búsquedas semánticas en el dominio de la lingüística, mediante la interacción con la ontología desarrollada utilizando el *framework* Jena. El buscador semántico puede recibir los conceptos de la ontología de dos formas distintas:

- El buscador semántico recibe un concepto de la ontología directamente de la consulta. Con el objetivo de ayudar al usuario a elegir un concepto existente se le muestra una lista con los conceptos de la ontología que se asemejan a su consulta. De esta manera puede elegir un concepto de la lista (ver Figura 22). Cuando el usuario selecciona un concepto de la lista, se realiza directamente una búsqueda semántica a partir de este concepto seleccionado. Mediante esta forma, el buscador semántico solo puede recibir como máximo un concepto.
- El buscador semántico recibe el o los conceptos encontrados por la herramienta de búsqueda textual. Esto se da cuando el usuario realiza una búsqueda sin seleccionar ningún concepto de la lista que se le muestra. Cuando sucede esto, se considera que el usuario no quiere limitar la búsqueda a un solo concepto (de lo contrario, hubiera seleccionado el concepto en la lista). Por lo tanto, se realiza primero la búsqueda textual descrita en el resultado esperado anterior. Mediante esta forma, el buscador semántico puede recibir uno o más conceptos.

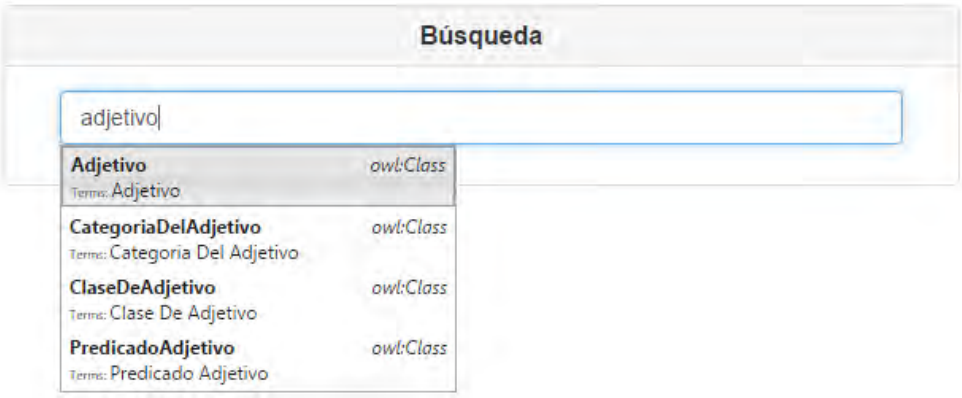

Figura 22: Ingreso de una consulta. Autoría propia.

Luego de recibir el o los conceptos, el buscador semántico procede a recorrer la ontología a partir de éstos para encontrar todos los demás conceptos relacionados semánticamente y obtener sus URIs (ver Figura 23). Teniendo las URIs de estos conceptos, la aplicación consulta a la base de datos en la que se encuentran almacenadas las anotaciones semánticas. Esta consulta consiste en obtener los documentos que se hayan anotado con los conceptos representados por las URIs obtenidas por el buscador y, por último, devolver las URLs de estos documentos.

```
// lista con los conceptos encontrados por la busqueda semantica
List<String> concepts_list = new ArrayList<>();
// lista con los documentos encontrados por la busqueda semantica
List<Document> results_list = new ArrayList<>();

OntClass ontConcept = model.getOntClass(namespace + search_string);

if(ontConcept != null){
    // el usuario ingresó un concepto de la ontologia

    concepts_list.add(ontConcept.getURI());
    // se recorre la ontologia para obtener los conceptos relacionados
    ExtendedIterator subconcepts = ontConcept.listSubClasses();
    while (subconcepts.hasNext()){
        OntClass thisConcept = (OntClass) subconcepts.next();
        concepts_list.add(thisConcept.getURI());
    }

    modelMap.put("root_class", ontConcept.getLocalName());
}
```

Figura 23: Obtención de conceptos relacionados a partir de uno. Autoría propia.

Dentro de la aplicación de búsqueda semántica, con el objetivo de brindarle al usuario información importante que pueda ayudarle al momento de realizar sus consultas, se desarrolló una interfaz que muestra la jerarquía de clases de la ontología a modo de un árbol de conceptos (ver Figura 24). Al hacerle *click* a un nodo del árbol (concepto), se mostrarán los nodos hijos de éste. Esto permite al usuario conocer qué conceptos existen en la ontología y cómo están relacionados de manera que pueda guiar bastante mejor sus consultas. Además, cuando el usuario realiza una búsqueda seleccionando un concepto de la lista que se le muestra (ver Figura 22), se selecciona automáticamente este concepto en el árbol de conceptos. De esta manera, el usuario puede ver la información del concepto que buscó (conceptos padre y conceptos hijo) y tomar una mejor decisión en caso lo desee.

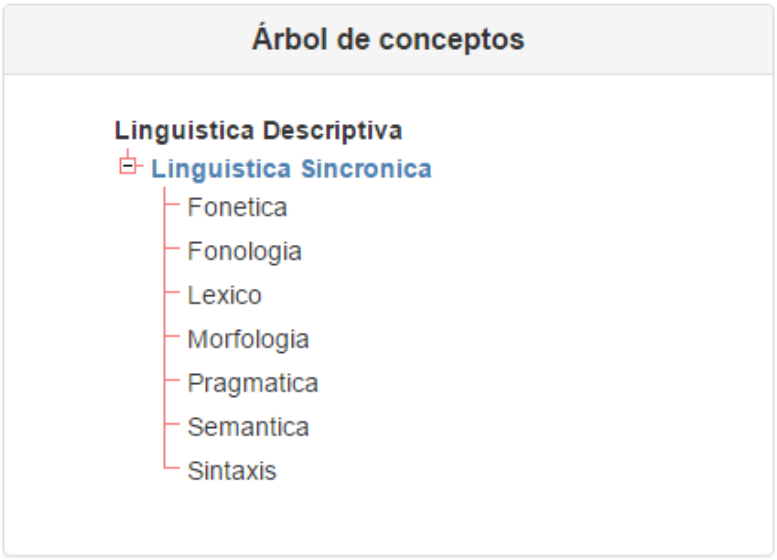

Figura 24: Árbol de conceptos de la ontología con el nodo "Lingüística Sincrónica" seleccionado. Autoría propia.

Finalmente, la Figura 25 muestra la pantalla principal de la aplicación, la cual consiste en la herramienta que permite obtener los documentos mediante búsquedas semánticas junto con el árbol de conceptos mencionado anteriormente. Se puede observar en la figura que se realizó una consulta con el concepto "Morfología", y el árbol de conceptos mostró automáticamente este concepto.

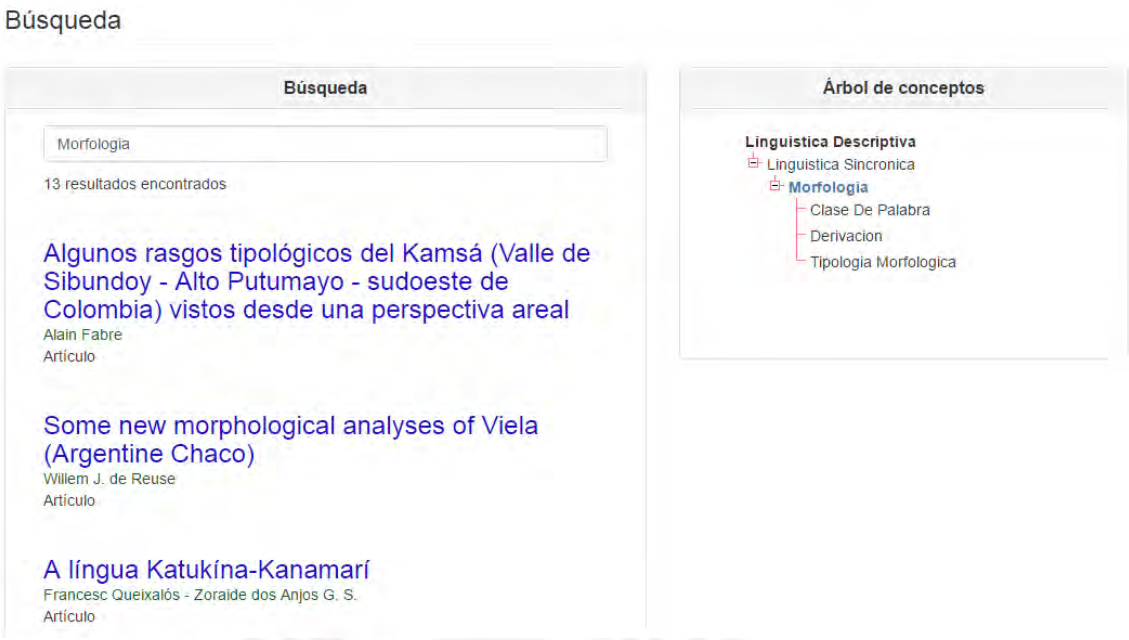

Figura 25: Aplicación de recuperación semántica de documentos. Autoría propia.

6.4 Evaluación de los resultados obtenidos por las búsquedas semánticas usando precisión y *recall*

Los resultados obtenidos por la aplicación de búsqueda semántica implementada se evaluaron utilizando las medidas de precisión y *recall*, con el objetivo de comprobar la buena performance de las búsquedas semánticas en estos aspectos. Las fórmulas para calcular la precisión y el *recall* son las siguientes:

- Precisión = Número de documentos recuperados relevantes / Número total de documentos recuperados
- *Recall* = Número de documentos recuperados relevantes / Número total de documentos relevantes (recuperados y no recuperados)

Para poder realizar esta evaluación, se utilizaron un conjunto de 34 documentos los cuales fueron anotados semánticamente utilizando la aplicación implementada para este fin. Luego, se ejecutó la aplicación de búsqueda semántica con diferentes consultas y se midieron los resultados para cada una de éstas. La Tabla 8 muestra la precisión y el *recall* de los resultados de las diferentes consultas realizadas.

Tabla 8: Precisión y recall de diferentes consultas

#	Consulta	Número de documentos recuperados relevantes	Número de documentos no recuperados relevantes	Número total de documentos recuperados	Precisión	*Recall*
1	Sintaxis	24	0	24	100%	100%
2	Morfologia	13	0	13	100%	100%
3	Relacion Gramatical	7	0	7	100%	100%
4	clase de palabra	10	0	10	100%	100%
5	clase abierta	5	0	10	50%	100%
6	valencia	2	0	2	100%	100%
7	morfología y sintaxis	34	0	34	100%	100%
8	frase nominal y verbal	4	0	9	44,44%	100%
9	relacion gramatical, alineamiento y distincion de transitividad	7	0	7	100%	100%
10	ergatividad, jerarquia y principio semantico	4	0	24	17%	100%

En las tres primeras consultas se seleccionó un concepto de la ontología (de la lista que se le muestra al usuario), realizándose directamente las búsquedas semánticas respectivas. Estas búsquedas obtuvieron siempre una precisión y *recall* de 100%.

En las consultas 4, 5 y 6 se ejecutó primero la búsqueda textual (no se seleccionó el concepto de la lista). La consulta 5 no alcanzó un 100% de precisión debido a que el buscador textual obtuvo conceptos que no estaban relacionados semánticamente al concepto ingresado ("clase abierta").

Para las consultas 7 y 8, en cada una se quiso buscar dos conceptos a la vez, realizándose primero la búsqueda textual. La consulta 8 no alcanzó el 100% de precisión por el mismo motivo que la consulta 5.

Para las últimas dos consultas, en cada una se quiso buscar tres conceptos a la vez. En ambos casos, el buscador textual encontró conceptos que no se encontraban relacionados semánticamente a los ingresados. Sin embargo, la consulta 9 no se vio afectada por esto, ya que los conceptos encontrados que pudieron disminuir su precisión no estaban anotados a ningún documento (de los 34 que se utilizaron para esta evaluación). Esto quiere decir que un futuro, cuando se cuenten con más documentos, esta precisión sí puede disminuir. Y finalmente, la consulta 10 sí se vio afectada disminuyendo considerablemente su precisión.

Luego de analizar los resultados de esta evaluación, se tienen las siguientes conclusiones:

- Siempre se obtendrá un *recall* de 100%, debido al mecanismo empleado por las búsquedas.
- Las búsquedas semánticas directas (sin una búsqueda textual previa) obtendrán siempre el 100% de precisión.
- Cuando se realiza primero la búsqueda textual, existe la probabilidad de que la precisión disminuya. Esto se da debido a que el buscador textual puede encontrar conceptos que se asemejen textualmente pero que no estén semánticamente relacionados a los conceptos ingresados en la consulta.

- Como se puede apreciar en las consultas 5, 8 y 10; mientras más larga sea la consulta (incluya más palabras o conceptos), la probabilidad de que la precisión disminuya aumentará y su valor será cada vez menor. Esto se da ya que el buscador textual tendrá más palabras que analizar y obtendrá una mayor cantidad de conceptos.

7 CONCLUSIONES Y TRABAJOS FUTUROS

7.1 Conclusiones

El presente proyecto de fin de carrera ha implementado una aplicación web que permite realizar búsquedas semánticas en el dominio de la lingüística. A continuación, se presentan las conclusiones que se han obtenido luego de haber logrado los objetivos específicos que se plantearon.

En el modelamiento del dominio de la lingüística, se desarrolló una ontología que permitió representar el conocimiento de este dominio y codificarlo. Para esto, se usó la herramienta Protégé y la metodología *Ontology Development 101*. Lo primero que se tuvo en cuenta antes de iniciar la construcción de la ontología fue que el dominio de la lingüística era demasiado amplio, por lo que necesitaba ser limitado. La metodología usada permitió realizar un análisis del dominio y determinar el verdadero alcance, el cual fue la parte de la lingüística denominada lingüística descriptiva.

En la construcción de la ontología, se observó que debido a los requerimientos de búsqueda y a la propia naturaleza del dominio, no se necesitaron incluir propiedades a las clases ni crear instancias de las mismas. La dificultad de esta construcción estuvo en la gran cantidad de conceptos (175 clases en la ontología) y su modelamiento a través de una clasificación jerárquica. Toda la información que se requirió para la construcción se recabó mediante reuniones con un profesor de lingüística del Departamento de Humanidades de la universidad.

Luego de tener la ontología desarrollada, se realizaron pruebas que permitieron verificar su consistencia. Se probó que a partir de un concepto se puedan obtener todos los subconceptos dentro de éste, lo cual era requerido por la posterior búsqueda semántica a implementar.

Lo último que se puede concluir del modelamiento del conocimiento del dominio de la lingüística es que no existe un único correcto modelamiento. El modelamiento depende del autor, su conocimiento del dominio y los fines para los que será usado

este conocimiento. Por lo tanto, dos personas o dos aplicaciones pueden tener modelamientos diferentes para un mismo dominio.

Un segundo objetivo específico que se planteó el proyecto fue el desarrollo de una herramienta de software para la anotación semántica de los documentos. Se implementó el mecanismo y se realizó la anotación de 34 documentos que se usaron para las pruebas de la aplicación. Debido a que se trata de una anotación manual, ésta implicó un esfuerzo el cual aumentará y hará más ineficiente el proceso a medida que aumenten los documentos a anotar.

La herramienta de anotación se implementó con el objetivo de que las anotaciones de los documentos puedan ser mantenidas en el tiempo. Esto se logró con la segunda forma de anotación que se definió, la cual permite buscar un documento registrado en la aplicación y modificar sus anotaciones.

El proceso de anotación desarrollado es un aspecto muy importante y que debe ser realizado por personas con conocimiento del dominio de la lingüística, ya que la eficacia de las búsquedas semánticas depende directamente de estas anotaciones. La precisión de las búsquedas se verá afectada si no se realiza una correcta anotación de los documentos.

Por último, la ontología y el mecanismo de anotación semántica desarrollados permitieron implementar el buscador semántico. La implementación del buscador semántico, relacionado al tercer objetivo específico, es el principal foco de este proyecto; sin embargo, un buscador en el ámbito de la Web Semántica no es posible sin una ontología y un proceso de anotación semántica previo.

La ontología lingüística es la que permite potenciar las búsquedas semánticas; es decir, brinda el conocimiento necesario para realizar las búsquedas. De esta manera, las búsquedas van a depender de cómo el conocimiento haya sido modelado en la ontología. Además, el buscador debe conocer este modelo para poder recorrer la ontología de manera adecuada. El presente proyecto empleó el lenguaje OWL para el modelado de la ontología y el *framework* Jena que permitió recorrer la ontología en este lenguaje. Una ventaja de tener el conocimiento separado a la propia herramienta

de búsqueda es que las búsquedas pueden ser potenciadas con el tiempo; es decir, brindar mayor conocimiento añadiendo más conceptos a la ontología, sin la necesidad de modificar la herramienta.

La herramienta de búsqueda semántica implementada depende también de un proceso de anotación semántica previo. Las anotaciones de los documentos le permiten al buscador reconocer el contenido de estos documentos; es decir, la semántica de la información contenida. Para el caso de este proyecto, las anotaciones definen los conceptos lingüísticos, que pueden considerarse como temas, que tratan los documentos. De esta manera, el buscador puede determinar si un documento es relevante o no para el usuario de acuerdo a la consulta.

Además de la herramienta de búsqueda semántica, se implementó una herramienta de búsqueda textual. Esta herramienta permite brindar resultados relevantes al usuario a pesar de que no haya ingresado un concepto de la ontología en la consulta, proporcionando mayor flexibilidad a las búsquedas. La ventaja de esta búsqueda textual es que permite que las consultas no se limiten a solo un concepto de la ontología, en caso el usuario lo desee así. Sin embargo, se tiene la desventaja de que puede disminuir la precisión de las búsquedas semánticas, lo cual se pudo observar y analizar en la evaluación realizada.

7.2 Trabajos futuros

Luego de haber implementado la aplicación de búsqueda semántica en el dominio de la lingüística con el alcance establecido, se identificaron los siguientes trabajos futuros que podrían mejorar las capacidades del buscador semántico:

- *Añadir un módulo de ontología:* Con el objetivo de brindar mayor conocimiento al buscador semántico, la ontología lingüística puede ser extendida añadiéndole más conceptos, mejorando así las búsquedas. Se podría implementar un módulo de ontología en la aplicación de manera que esta tarea pueda ser realizada por determinados usuarios con conocimientos lingüísticos (profesores, por ejemplo). Este módulo solo podría ser accedido por usuarios designados y no por cualquier usuario de la aplicación. Esto

debido a que, al igual que las anotaciones semánticas, se trata de tareas que deben ser realizadas de manera consiente y cuidadosa ya que afectan directamente la precisión de las búsqueda semánticas. Mediante una interfaz amigable que oculte la complejidad de la construcción de una ontología, se le brindaría al usuario los permisos suficientes solo para añadir y quitar clases de la taxonomía.

.

- *Incluir búsqueda por lenguas o idiomas:* Los documentos pertenecientes al Departamento de Humanidades de la universidad tratan sobre temas lingüísticos de diferentes lenguas o idiomas. El presente proyecto, como se estableció en su alcance, considera los temas o conceptos lingüísticos; sin embargo, no considera las lenguas. De este modo; se podría permitir en las búsquedas, además de conceptos, consultar sobre determinadas lenguas. Por ejemplo, se podría tener la siguiente consulta: "ideófono surui"; siendo la primera palabra el concepto a buscar y la segunda palabra la lengua. De esta manera, se recuperarían los documentos que traten sobre ideófonos en la lengua Surui.

- *Recibir consultas en lenguaje natural:* El buscador se implementó de manera que reciba en la consulta conceptos de la ontología. En caso se añadiera la búsqueda por lenguas o idiomas descrita en el punto anterior, se podría implementar un método de parseo que permita recibir consultas en lenguaje natural. Siguiendo el ejemplo anterior, se podría tener la siguiente consulta: "ideófonos en la lengua surui"; de manera que se pueda identificar los términos que se refieren a conceptos de la ontología y los que se refieren a lenguas.

- *Anotación automática de los documentos:* Como se mencionó anteriormente, el proceso de anotación implementado en el proyecto es ineficiente debido a que se trata de una anotación manual. Esta ineficiencia podría solucionarse mediante un mecanismo de anotación automática. Este mecanismo se puede implementar con una herramienta que procese el texto de los documentos y los anote con los conceptos de la ontología identificados.

.

- *Mejorar las búsquedas textuales para reducir el impacto en la precisión:* Como se pudo analizar anteriormente, las búsquedas textuales son un factor que pueden disminuir la precisión de las búsquedas semánticas posteriores. Para evitar o reducir esta disminución en la precisión, se podría investigar más a fondo sobre la librería Lucene utilizada para mejorar los resultados de la herramienta de búsqueda textual. Por ejemplo, esta librería devuelve los resultados con un *score* que determina cuánto se asemeja la consulta al resultado, de manera que se podrían considerar solo los resultados con un *score* alto.

Referencias bibliográficas

[1] Manning, C. D., & Raghavan, P. (2009). An Introduction to Information Retrieval. *Online*. http://doi.org/10.1109/LPT.2009.2020494

[2] Wahlster, W., Dengel, A., Telekom, D., Dengel, W., Dengler, C. D., Heckmann, D., ... & Sintek, M. (2006). Web 3.0: Convergence of Web 2.0 and the semantic web. In *In Technology Radar, Feature Paper, 2nd ed.; Deutsche Telekom Laboratories*.

[3] Berners-Lee, T., Hendler, J., & Lassila, O. (2001). The Semantic Web. *Scientific American*. http://doi.org/10.1038/scientificamerican0501-34

[4] Giunchiglia, F., Kharkevich, U., & Zaihrayeu, I. (2009). Concept search. In *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)* (Vol. 5554 LNCS, pp. 429–444). http://doi.org/10.1007/978-3-642-02121-3_33

[5] Yu, L. (2007). Introduction to the Semantic Web and Semantic Web Services. *CRC Press*.

[6] Cardoso, J. (2007). Semantic Web Services: Theory, Tools and Applications: Theory, Tools and Applications. *Information Science Reference*.

[7] Kiryakov, A., Popov, B., Terziev, I., Manov, D., & Ognyanoff, D. (2004). Semantic annotation, indexing, and retrieval. *Web Semantics: Science, Services and Agents on the World Wide Web*. http://doi.org/10.1016/j.websem.2004.07.005

[8] Uren, V., Cimiano, P., Iria, J., Handschuh, S., Vargas-Vera, M., Motta, E., & Ciravegna, F. (2006). Semantic annotation for knowledge management: Requirements and a survey of the state of the art. *Journal of Web Semantics*, *4*(1), 14–28. http://doi.org/10.1016/j.websem.2005.10.002

[9] Reeve, L., & Han, H. (2005). Survey of semantic annotation platforms. In *Proceedings of the 2005 ACM symposium on Applied computing - SAC '05* (pp. 1634–1638). http://doi.org/10.1145/1066677.1067049

[10] Calero, C., Ruiz, F., & Piattini, M. (2006). Ontologies for Software Engineering and Software Technology. *Springer*.

[11] Lacy, L. W. (2005). Owl: Representing Information Using the Web Ontology Language. *Trafford.*

[12] Lim, E. H. Y., Liu, J. N. K., & Lee, R. S. T. (2011). Knowledge Seeker - Ontology Modeling for Information Search and Management: A Compendium. *Springer Berlin Heidelberg.*

[13] Sicilia, M. A., Kop, C., & Sartori, F. (2010). Ontology, Conceptualization and Epistemology for Information Systems, Software Engineering and Service Science: 4th International Workshop, ONTOSE 2010, Held at CAiSE 2010, Hammamet, Tunisia, June 7-8, 2010, Revised Selected Papers. *Springer.*

[14] Baca, M. (2008). Introduction to Metadata. *Getty Research Institute.*

[15] Dunsire, G., & Willer, M. (2011). Standard library metadata models and structures for the Semantic Web. *Library Hi Tech News*. http://doi.org/10.1108/07419051111145118

[16] Wittenburg, P., & Broeder, D. (2015). Metadata Overview and the Semantic Web.

[17] Castells, P., Perdrix, F., Pulido, E., Rico, M., Benjamins, R., Contreras, J., & Lorés, J. (2004). Neptuno: Semantic web technologies for a digital newspaper archive. *Lecture Notes in Computer Science (including Subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*. http://doi.org/10.1007/978-3-540-25956-5_31

[18] Bikakis, N., Giannopoulos, G., Dalamagas, T., & Sellis, T. (2010). Integrating keywords and semantics on document annotation and search. In *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)* (Vol. 6427 LNCS, pp. 921–938). http://doi.org/10.1007/978-3-642-16949-6_19

[19] Berrueta, D., Labra, J. E., & Polo, L. (2006). Searching over Public Administration Legal Documents Using Ontologies. In *Knowledge-based Software Engineering: Proceedings of the Seventh Joint Conference of Knowledge-Based Software Engineering* (p. 167).

[20] Lemnitzer, L., Simov, K., Osenova, P., Mossel, E., & Monachesi, P. (2008). Using a domain-ontology and semantic search in an e-learning environment. In *Innovative Techniques in Instruction Technology, E-Learning, E-Assessment, and Education* (pp. 279–284). http://doi.org/10.1007/978-1-4020-8739-4-4

[21] Kitchenham, B. (2004). Procedures for performing systematic reviews. *Keele, UK, Keele University*, *33*(TR/SE-0401), 28. http://doi.org/10.1.1.122.3308

[22] Levene, M. (2011). An introduction to search engines and web navigation. *John Wiley & Sons.*

[23] Schreiber, G. (2000). Knowledge engineering and management: the CommonKADS methodology. *MIT press.*

[24] Noy, N., & McGuinness, D. L. (2001). Ontology development 101. *Knowledge Systems Laboratory, Stanford University*.

[25] Farrar, S., & Langendoen, D. T. (2003). A linguistic ontology for the semantic web. *Glot International*, *7*(3), 97-100.

[26] Noruzi, A. (2007). Folksonomies-Why do we need controlled vocabulary?.*Webology*, *4*(2).

[27] descriptive linguistics. (n.d.). Collins English Dictionary - Complete & Unabridged 10th Edition. Retrieved April 04, 2016 from Dictionary.com website http://www.dictionary.com/browse/descriptive-linguistics

[28] Russell, J. & Cohn, R. (2012). *Semantic Reasoner*. Book on Demand.