

PONTIFICIA UNIVERSIDAD CATÓLICA DEL PERÚ

FACULTAD DE CIENCIAS E INGENIERÍA



**PONTIFICIA
UNIVERSIDAD
CATÓLICA
DEL PERÚ**

**DESARROLLO DE UNA HERRAMIENTA PARA LA ANOTACIÓN
SEMÁNTICA AUTOMÁTICA DE DOCUMENTOS PDF BASADO EN
ONTOLOGÍAS**

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

Gustavo Coronado Altamirano

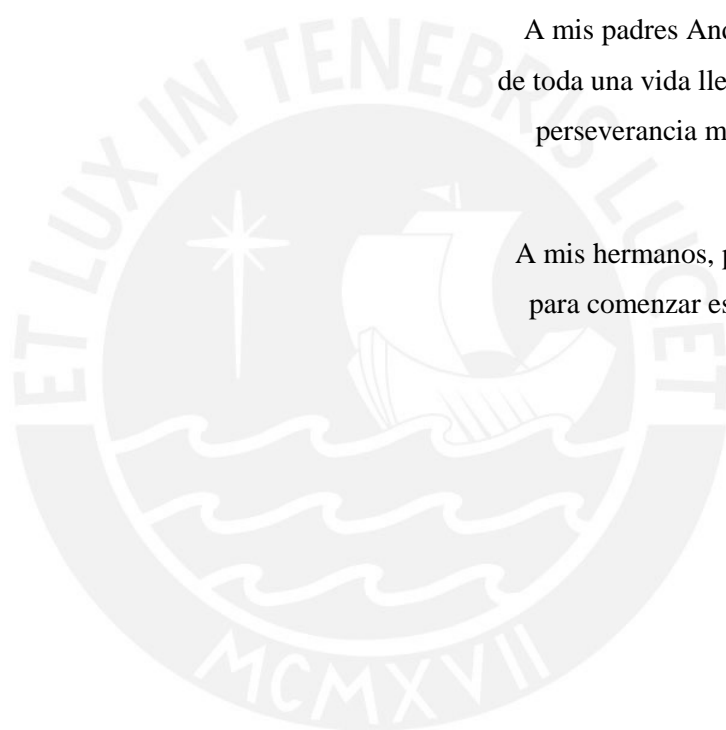
ASESOR: Dr. Héctor Andrés Melgar Sasieta

Lima, julio de 2017

Dedicatoria

A mis padres Andrés y Eufemia que a lo largo de toda una vida llena de ejemplos de esfuerzo y perseverancia me han impulsado a luchar por todos mis objetivos.

A mis hermanos, por el impulso que me dieron para comenzar esta carrera profesional en una prestigiosa universidad.

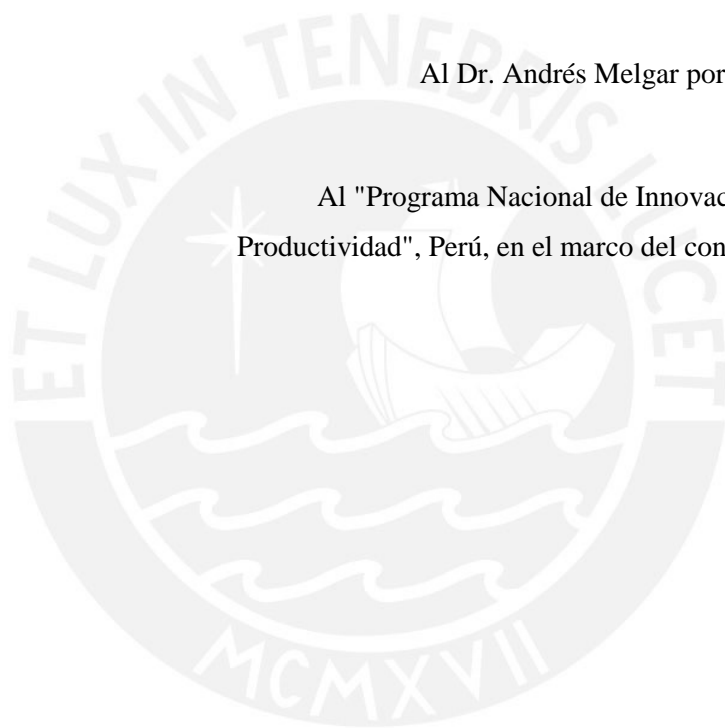


Agradecimientos

A todos mis amigos con los que compartí horas de estudios y trabajos grupales.

Al Dr. Andrés Melgar por su apoyo y asesoría en todas las etapas de este proyecto.

Al "Programa Nacional de Innovación para la Competitividad y Productividad", Perú, en el marco del contrato 124-PNICP-PIAP-2015 por el apoyo brindado.



RESUMEN

Actualmente, Internet es una de las fuentes más accesibles y utilizadas para buscar información sobre determinado tema, a través de la cual las personas pueden conectarse a una gran colección de recursos, servicios y contenidos. En ese sentido, el uso de motores de búsqueda es indispensable para poder encontrar contenido específico y relevante para el usuario, es decir, información precisa y alineada con el tema de su interés.

Sin embargo, los buscadores pueden presentar dificultades para brindar al usuario la información deseada. Estas dificultades se presentan por motivos tales como las características propias del lenguaje natural como la polisemia, sinonimia y ambigüedad; así, también, por el desconocimiento de los temas que son de interés para el usuario. Otra de las causas que dificultan la recuperación de información relevante es que la búsqueda de resultados se realiza de manera sintáctica, esto es, buscando en los documentos la coincidencia exacta de los términos ingresados en la cadena de búsqueda. Del mismo modo, otra razón importante es que los formatos e interfaces de contenido se presentan en formatos comprensibles solo por las personas y no por un computador.

Ante esto, el presente proyecto propone una alternativa de solución de forma tal que los documentos contengan información adicional que describa los conceptos y entidades principales del contenido. Esta información adicional se añadirá de manera automática a los documentos mediante anotaciones semánticas en base a un dominio de conocimiento que sea de interés para el usuario. De esta manera, se pretende apoyar el concepto de Web semántica cuya propuesta es clasificar, estructurar y anotar los recursos con semántica explícita para que puedan ser procesados por sistemas inteligentes.

Contenido

Contenido	iii
Ilustraciones	vi
Tablas	vii
1. DEFINICIÓN DEL PROBLEMA	1
1.1 PROBLEMÁTICA	1
1.2 OBJETIVO GENERAL.....	4
1.3 OBJETIVOS ESPECÍFICOS	4
1.4 RESULTADOS ESPERADOS	5
1.5 HERRAMIENTAS, MÉTODOS, METODOLOGÍAS Y PROCEDIMIENTOS	6
1.5.1. Herramientas.....	7
1.5.2. Métodos y Procedimientos	11
1.5.3. Metodologías	11
1.6 ALCANCE	12
1.6.1. Riesgos.....	12
1.7 JUSTIFICACIÓN	13
2. MARCO CONCEPTUAL.....	14
2.1 WEB SEMÁNTICA	14
2.2 ONTOLOGÍA.....	15
2.3 ANOTACIÓN.....	17
2.4 METADATO	17
2.5 ANOTACIÓN SEMÁNTICA	19
2.6 LINKED OPEN DATA.....	20
2.7 CONCLUSIÓN	21
3. ESTADO DEL ARTE.....	22
3.1 INTRODUCCIÓN.....	22
3.2 MÉTODO DE REVISIÓN DEL ESTADO DEL ARTE	22
3.3 ¿De qué manera se ha venido realizando anotaciones semánticas automáticas haciendo uso de las ontologías?	24
3.4 ¿Cuáles son los principales estándares para definir metadatos?.....	25
3.5 ¿Cuáles son las herramientas de anotación semántica automática que existen actualmente?.....	26
3.5.1. THEOPHRASTUS	26

3.5.2.	FLERSA	27
3.5.3.	UTOPIA DOCUMENTS.....	27
3.5.4.	OnTheFly.....	28
3.5.5.	DOMEO ANNOTATION TOOLKIT.....	29
3.5.6.	Semantic Annotation Tool for Annotating Arabic Web Documents 30	
3.6	CONCLUSIÓN	32
4.	CAPÍTULO 4: PROCESAMIENTO DE LA INFORMACIÓN TEXTUAL DE LOS DOCUMENTOS	33
4.1	Objetivo Específico N° 1: Implementar un módulo que permita procesar la información textual de los documentos.	33
4.2	Resultado Alcanzado N°1: Módulo para obtener los términos que se encuentran en un documento.	34
4.3	Resultado Alcanzado N°2: Módulo que permite reducir las palabras a su forma base o lema.....	36
4.4	Discusión de los resultados alcanzados N°1 y N°2	36
5.	CAPÍTULO 5: REPRESENTACIÓN DE LA INFORMACIÓN MEDIANTE LENGUAJE ESTRUCTURADO.....	39
5.1	Objetivo Específico N° 2: Representar la información de los documentos mediante un lenguaje estructurado para la generación de anotaciones.	39
5.2	Resultado Alcanzado N° 3: Módulo que permita realizar el etiquetado gramatical.	39
5.3	Resultado Alcanzado N° 4: Módulo que permita realizar el análisis semántico.	41
5.4	Discusión de los resultados alcanzados N° 3 y N° 4.	42
6.	CAPÍTULO 6: DESAMBIGUACIÓN DE TÉRMINOS EN UN DOMINIO ESPECÍFICO.....	43
6.1	Objetivo Específico N° 3: Implementar un módulo que permita resolver la ambigüedad de los términos contenidos en los documentos en un dominio específico.....	43
6.2	Resultado Alcanzado N° 5: Módulo de desambiguación de palabras dentro de un dominio.	43
6.3	Discusión del resultado alcanzado N° 5.	45
7.	CAPÍTULO 7: ALMACENAMIENTO DE ANOTACIONES SEMÁNTICAS EN REPOSITORIOS DIGITALES.....	47
7.1	Objetivo Específico N° 4: Implementar un módulo que permita almacenar anotaciones semánticas en repositorios digitales.	47
7.2	Resultado Alcanzado N° 6: Formato definido para el almacenamiento de las anotaciones semánticas en repositorios.....	47

7.3	Resultado Alcanzado N° 7: Módulo de almacenamiento de anotaciones semánticas.....	49
7.4	Discusión de los resultados alcanzados N° 6 y N°7.....	49
8.	CONCLUSIONES	51
9.	LIMITACIONES	53
10.	TRABAJOS FUTUROS	54
11.	BIBLIOGRAFÍA	55



Ilustraciones

Ilustración 1: Diferencia entre Web Semántica y Web actual. Imagen adaptada de [10].....	15
Ilustración 2: Ejemplo de Ontología. Imagen adaptada de [14].....	16
Ilustración 3: Ejemplo de anotación sobre un documento Web. Imagen de autoría propia.....	17
Ilustración 4: Ejemplo de metadato del documento Web en RDF. Imagen de autoría propia.	19
Ilustración 5: Ejemplo de un documento con anotaciones semánticas en una Web. Imagen de autoría propia.....	20
Ilustración 6: Proceso soportado por el sistema Theophrastus. Imagen adaptada de [41]	26
Ilustración 7: La arquitectura de Utopia Documents, mostrando la relación entre la GUI, plugins y ontologías. Imagen adaptada de [43].....	28
Ilustración 8: Aplicación Web OnTheFly. Imagen adaptada de [45].....	29
Ilustración 9: Arquitectura del sistema de Anotación Semántica. Imagen recuperada de [32].....	30
Ilustración 10: Funcionalidad de extracción de contenido textual. Imagen tomada de la herramienta desarrollada (Autoría propia).	34
Ilustración 11: Funcionalidad de extracción de términos. Imagen tomada de la herramienta desarrollada (Autoría propia).	35
Ilustración 12: Funcionalidad de reducción de palabras a su forma base. Imagen tomada de la herramienta desarrollada (Autoría propia).....	37
Ilustración 13: Resultados de prueba de Precisión y Recall.....	38
Ilustración 14: Funcionalidad de etiquetado gramatical de cada palabra. Imagen tomada de la herramienta desarrollada (Autoría propia).....	40
Ilustración 15: Resultados de prueba de Precisión y Recall (Autoría propia).....	45
Ilustración 16: Ejemplo de anotación semántica generada por la herramienta. Imagen tomada de la herramienta desarrollada (Autoría propia).48	
Ilustración 17: Tabla anotacion_semantica (Autoría propia).	49
Ilustración 18: Resultados de prueba de Precisión y Recall (Autoría propia).....	50

Tablas

Tabla 1: Resultados Esperados.....	5
Tabla 2: Herramientas Métodos y Procedimientos.	7
Tabla 3: Tabla de riesgos del proyecto.....	13
Tabla 4: Resumen de la cantidad de resultados devueltos por las bases de datos.	23
Tabla 5: Resumen de la respuesta a la pregunta 3.3.....	24
Tabla 6: Resumen de la respuesta a la pregunta 3.4.....	25
Tabla 7: Cuadro comparativo de aplicaciones de anotación semántica.	31



1. DEFINICIÓN DEL PROBLEMA

1.1 PROBLEMÁTICA

El incremento de la información y el conocimiento determina la complejidad y variabilidad de la sociedad. Por esto, los factores básicos de desarrollo socioeconómico del sistema social lo constituyen la información y su derivado: el conocimiento (Ziamba and Zelazny 2013).

Actualmente, una de las fuentes más accesibles y utilizada para buscar información sobre determinado tema es Internet. Según *Internet Live Stats*¹, al consultar la información el 15 de abril de 2017, la cantidad de páginas registradas bordea los 1,000 millones, mientras que el número de usuarios es aproximadamente 3,000 millones (Internet Live Stats, 2015); esto, comparado con los más de 60,000 sitios Web registrados a finales de 1995, evidencia un crecimiento a pasos agigantados de la información en la Web (NetCraft, 2015). Ciertamente, la finalidad de la Internet es que las personas puedan conectarse a una gran colección de recursos, servicios y contenidos (Caporuscio and Ghezzi 2015).

Debido a la sobrecarga de información existente, se hace indispensable el uso de motores de búsqueda que permitan encontrar información específica y apropiada (Fung and Thanadechtemapat 2010). Ante esta situación, los actuales motores de búsqueda en la Web colocan mayor énfasis en la velocidad de búsqueda y en cubrir la mayor cantidad posible de información que en la efectividad del resultado que se brinda al usuario (Costa, Printista, and Marin 2006). En ese sentido, es importante para el usuario que los documentos obtenidos al realizar la búsqueda contengan información relevante; es decir, que sea precisa y alineada con el tema de su interés (J. A. P. Sánchez 2011).

Aunque los resultados obtenidos por los buscadores son generalmente buenos, se pueden presentar los siguientes problemas para encontrar la información deseada. Uno de estos se genera al obtener como resultado páginas Web cuyo contenido no es relevante para el objetivo de la búsqueda, lo que representa una pérdida de tiempo al usuario porque tiene que revisar todos los resultados de búsqueda (García and Fernández 2005). Otro problema, debido a la sobrecarga de información, es que se obtiene como resultado una larga lista de páginas Web, lo cual ocasiona que el usuario tenga que revisar individualmente cada enlace y solo pueda procesar una parte muy pequeña de los resultados, ya que procesar toda la

¹ <http://www.internetlivestats.com/>

información obtenida sería una tarea inabarcable (Costa, Printista, and Marin 2006; Josephine and Sathiyadevi 2011).

Estos problemas pueden ser aún mayores tomando en cuenta los siguientes tres factores que influyen en los resultados de búsqueda y hacen que, en ocasiones, estos sean incorrectos: la polisemia, debido a que una palabra puede tener más de un significado, se pueden encontrar documentos que contienen las palabras ingresadas pero en un contexto distinto al deseado; la sinonimia, al realizar una búsqueda no figura entre las páginas encontradas aquellas que contienen sinónimos de los términos de la consulta, lo cual sería muy útil; por último, el multilingüismo que no permite que se muestren resultados en otros idiomas acerca de los mismos términos ingresados en la consulta (García and Fernández 2005).

Una de las principales razones por la cual los buscadores pueden perder eficacia y calidad en los resultados de la búsqueda es la forma en cómo recuperan la información (Tello 2001). En primer lugar, las búsquedas se realizan tomando las palabras claves de la consulta y se obtiene como resultado una lista de documentos que incluyen dichas palabras en su contenido; luego, los resultados se muestran ordenando esta lista en base a estadísticas y nivel de visitas de las páginas pero sin considerar necesariamente el contexto (Li, Xie, and Dong 2009). A esta forma de realizar las búsquedas se le denomina búsqueda sintáctica, en la cual los resultados son documentos que contienen las palabras clave (García and Fernández 2005).

Otra razón importante que contribuye a la dificultad de búsqueda, es que los formatos e interfaces de contenidos y servicios se presentan en formatos comprensibles solo por personas y no por las máquinas (J. A. P. Sánchez 2011; Tello 2001). Así también, otro motivo importante por el cual la Web puede presentar resultados deficientes es el espacio de búsqueda en donde se realizan las consultas (Josephine and Sathiyadevi 2011). Las búsquedas se realizan en lo que se denomina “Web superficial”, la cual es formada por documentos estáticos accesibles en la Web; sin embargo, existe también la denominada “Web profunda” formada por bases de datos no directamente accesibles que se hacen visibles mediante páginas generadas dinámicamente. La “Web profunda” contiene información varios cientos de veces mayor que la “Web superficial” y de mucha mejor calidad (Costa, Printista, and Marin 2006).

Considerando los problemas mencionados, una alternativa para superar las limitaciones de la Web actual es que las páginas incorporen metadatos estructurados que

describan los conceptos y entidades principales sobre su contenido en un determinado dominio de conocimiento (Castells 2003). Los metadatos son datos que brindan información adicional acerca del contenido del documento para permitir su procesamiento por agentes de software (Kiryakov et al. 2004).

Una forma de proporcionar esa información adicional a los documentos es realizando anotaciones semánticas con el objetivo de anotar metadatos sobre las páginas o documentos en base a un dominio de información seleccionado por el usuario (UREN et al. 2006). Las anotaciones semánticas permiten establecer relaciones entre las entidades identificadas en un documento y los conceptos del dominio (Kiryakov et al. 2004). De esta manera, las búsquedas ya no se realizarían solamente por palabras claves, sino que, además, se basarían en las anotaciones de los documentos para entender su contenido y discriminar su importancia para la búsqueda del usuario (García and Fernández 2005).

Esto conduce al concepto de una Web semántica, cuyo objetivo primordial es que la información que actualmente es comprensible solo por los humanos también esté disponible de manera formal para sistemas inteligentes (Berners-Lee, Hendler, and Lassila 2001).

Para explotar los beneficios de la Web semántica y enfocarse en lo que realmente le interesa al usuario se deben construir sistemas basados en conocimiento, los cuales presentan dos componentes básicos: una base de conocimiento de los hechos conocidos por el sistema y un motor de inferencia. La base de conocimiento se irá construyendo tomando como base las anotaciones recogidas por la aplicación mientras navega por las páginas Web semánticas (García and Fernández 2005).

Con respecto a las anotaciones, para evitar el uso inadecuado de lenguajes de representación de conocimiento, la *World Wide Web Consortium*² ha definido un Marco de Descripción de Recursos (RDF, por sus siglas en inglés) (W3C, Documento en línea, s.f.). Esto debido a que se requiere que cualquiera pueda realizar anotaciones sobre páginas Web o documentos y que dichas anotaciones no estén restringidas a un vocabulario fijo (W3C, RDF, s.f.). Para poder recuperar conocimiento a partir de las anotaciones realizadas por distintas personas o aplicaciones surge la necesidad de que estas sean almacenadas y accesibles desde repositorios digitales de información (Kiryakov et al. 2004).

Sin embargo, el manejar una estructura con anotaciones formales para poder “entender” el contenido de cada documento no sería suficiente para las aplicaciones que

² <http://www.w3.org/>

realicen búsquedas en la Web semántica, pues persiste el problema de la ambigüedad de términos, ya que los motores de búsqueda no se enfocan en un modelo de dominio específico sino en cualquier fuente disponible a la que se tenga acceso (Tello 2001). Es por ello que las aplicaciones necesitarán un modelo de dominio sobre el cual trabajar, y dicho modelo puede ser especificado mediante el uso de ontologías (García and Fernández 2005). En este contexto, las ontologías permiten representar el vocabulario de los conceptos relevantes a ese dominio, las propiedades y relaciones entre los diferentes conceptos, así como también las reglas de gobierno del dominio, de tal forma que permita expresar los aspectos semánticos de los recursos de información (Davies, Fensel, and Van Harmelen 2003; Gruber 1995a).

Entonces, la Web actual está evolucionando hacia un nuevo y prometedor concepto denominado Web Semántica que, según Berners-Lee, permitirá solucionar las necesidades de contextualización de la información a partir de su dominio (Berners-Lee, Hendler, and Lassila 2001; Villazón and Villa 2014). El fin esencial de la Web semántica es facilitar la localización, compartición e integración de información y servicios para lograr el mayor beneficio de los recursos de la Web (J. A. P. Sánchez 2011).

Frente a los problemas de la Web actual de presentar una semántica implícita, el crecimiento exponencial y desordenado de la información, y la ausencia de una organización clara, la Web semántica propone clasificar, estructurar y anotar los recursos con semántica explícita para que puedan ser procesados por máquinas (Castells 2003). Por esta razón, surge la interrogante ¿De qué manera se puede estructurar el contenido de los documentos en la Web como soporte a la recuperación de información relevante para el usuario haciendo uso de las ontologías?

1.2 OBJETIVO GENERAL

Desarrollar una herramienta que permita realizar anotaciones semánticas automáticas sobre documentos PDF basado en ontologías.

1.3 OBJETIVOS ESPECÍFICOS

- Objetivo 1. Implementar un módulo que permita procesar la información textual de los documentos.
- Objetivo 2. Representar la información de los documentos mediante un lenguaje estructurado para la generación de anotaciones.
- Objetivo 3. Implementar un módulo que permita resolver la ambigüedad de los términos contenidos en los documentos en un dominio específico.

- Objetivo 4. Implementar un módulo que permita almacenar anotaciones semánticas en repositorios digitales.

1.4 RESULTADOS ESPERADOS

En la tabla 1 se muestra los resultados esperados para cada objetivo y los medios o fuentes que servirán para verificar que los resultados esperados se han alcanzado.

Tabla 1: Resultados Esperados.

Resultado Esperado	Descripción	Medios / Fuentes de Verificación
Objetivo 1: Implementar un módulo que permita procesar la información textual de los documentos.		
1. Módulo para obtener los términos que se encuentran en un documento.	Este módulo permitirá procesar todo el contenido textual de un documento con el fin de identificar las palabras o frases que contiene y luego verificar si tienen relación con el dominio de conocimiento configurado.	Documento ICONIX.
2. Módulo que permita reducir las palabras a su forma base o lema.	Este módulo permitirá estructurar el texto en <i>tokens</i> que luego se agruparán en oraciones y se podrá identificar acrónimos y abreviaturas.	Documento ICONIX.
Objetivo 2: Representar la información de los documentos mediante un lenguaje estructurado para la generación de anotaciones.		
3. Módulo que permita realizar el etiquetado gramatical.	Este módulo permitirá asignar a cada <i>token</i> su categoría gramatical.	Documento ICONIX.
4. Módulo que permita realizar el análisis semántico.	Este módulo permitirá determinar la estructura sintáctica de cada oración para luego construir el mapeamiento semántico. Con ello, se codificarán las	Documento ICONIX.

	anotaciones semánticas mediante un lenguaje estructurado.	
Objetivo 3. Implementar un módulo que permita resolver la ambigüedad de los términos contenidos en los documentos en un dominio específico.		
5. Módulo de desambiguación de palabras dentro de un dominio.	Este módulo permitirá resolver el problema de la ambigüedad de término que tienen distintos significados según el contexto en que se encuentren. Se determinará el significado adecuado de un término en base a su relación con los conceptos del dominio configurado.	El software desarrollado. Documento ICONIX.
Objetivo 4. Implementar un módulo que permita almacenar anotaciones semánticas en repositorios digitales.		
6. Formato definido para el almacenamiento de las anotaciones semánticas en repositorios.	Estructura de metadatos que permita la realización de anotaciones digitales que contengan información semántica del dominio del conocimiento.	Documento ICONIX.
7. Módulo de almacenamiento de anotaciones semánticas.	Este módulo permitirá la persistencia de las anotaciones semánticas realizadas de manera que puedan ser utilizadas por otras aplicaciones de la Web semántica para realizar búsqueda y gestión de información.	Documento ICONIX.

1.5 HERRAMIENTAS, MÉTODOS, METODOLOGÍAS Y PROCEDIMIENTOS

El objetivo principal de este proyecto de fin de carrera es desarrollar una herramienta para realizar de forma automática anotaciones semánticas sobre documentos en base a una ontología determinada por el usuario con el fin de dar soporte a la recuperación de conocimiento mediante herramientas de búsqueda y a la gestión de la información. En la tabla 2 se muestra el mapeo de las herramientas, métodos y procedimientos que se usarán durante el desarrollo del proyecto para alcanzar los resultados esperados.

Tabla 2: Herramientas Métodos y Procedimientos.

Resultado Esperado	Herramientas, Métodos y Procedimientos.
RE1: Módulo para obtener los términos que se encuentran en un documento.	FreeLing, Apache PDFBox, Lenguaje de Programación Java, NetBeans.
RE2: Módulo que permita reducir las palabras a su forma base o lema.	FreeLing, Lenguaje de Programación Java, NetBeans.
RE3: Módulo que permita realizar el etiquetado gramatical.	FreeLing, Lenguaje de Programación Java, NetBeans.
RE4: Módulo que permita realizar el análisis sintáctico.	FreeLing, Lenguaje de Programación Java, NetBeans.
RE5: Módulo de desambiguación de palabras dentro de un dominio.	Lenguaje de programación Java, Apache Jena, OWL, SPARQL, NetBeans.
RE6: Formato definido para el almacenamiento de las anotaciones semánticas en repositorios.	CommonKADS, RDF
RE7: Módulo de almacenamiento de anotaciones semánticas.	Lenguaje de programación Java, MySQL, NetBeans.

1.5.1.Herramientas

1.5.1.1. FreeLing

*FreeLing*³ es una librería de código abierto para el procesamiento multilingüe automático que proporciona una amplia gama de servicios de análisis lingüístico para diversos idiomas. El proyecto *FreeLing* fue concebido en el centro de investigación TALP⁴ de la Universidad Politécnica de Cataluña con el objetivo de poner a disposición recursos y herramientas básicos de Procesamiento de Lenguaje Natural (NLP por sus siglas en inglés, *Natural Language Processing*) (Padró 2012).

³ <http://nlp.lsi.upc.edu/freeling/>

⁴ <http://www.talp.upc.edu/>

La librería, desde su lanzamiento en el año 2010, brindaba servicios de análisis en tres idiomas (inglés, español y catalán); sin embargo, debido a su naturaleza de código abierto, la comunidad de usuarios ha ampliado el número de idiomas a nueve, aunque los idiomas que tienen la mayor cantidad de servicios de análisis son el español e inglés. Entre los servicios que brinda destacan el análisis morfológico de texto, el etiquetado gramatical de palabras y la identificación de sus bases morfológicas (Padró and Stanilovsky 2012). *FreeLing* ha sido desarrollado en el lenguaje C++, pero se tienen extensiones para Java, Python y PHP. Para este proyecto de fin de carrera se trabajará con la API para Java que utiliza los recursos de *FreeLing* en C++.

La elección de esta herramienta se debe a que, para los idiomas español e inglés, están disponibles todos los servicios de análisis que ofrece, lo cual facilitará el procesamiento y análisis de los textos escritos en lenguaje natural que se encuentran dentro de los comentarios.

1.5.1.2. Apache PDFBox

Apache PDFBox es una herramienta de análisis de contenido de código abierto para Java concebida como un proyecto de *The Apache Software Foundation*⁵. La herramienta trabaja con documentos PDF, que es el formato de archivo con el que trabajará la herramienta de anotación semántica a desarrollar en este proyecto. PDFBox permite la creación de nuevos documentos PDF, manejar documentos existentes y extraer el contenido de los mismos. Los archivos pueden ser analizados a través de una única interfaz siendo útil para la indexación del motor de búsqueda, análisis de contenido y traducción (The Apache Software Foundation, s.f.).

Le elección de esta herramienta se basa en que posee los componentes necesarios para la extracción de texto y metadatos de un documento, lo que permitirá identificar la relación del contenido con la ontología configurada en la herramienta que se desarrollarán en este proyecto.

1.5.1.3. Apache Jena

Jena es un *framework* de código abierto hecho en Java para construir aplicaciones de Web Semántica y datos enlazados. Está compuesto de diferentes APIs que interactúan en conjunto para procesar datos en formato RDF. Jena es compatible con los modelos RDFS y Lenguaje de Ontologías Web (OWL, por sus siglas en inglés), lo que permite añadir semántica adicional a los datos RDF (Apache Jena, s.f.). También cuenta con un motor de consultas ARQ que soporta el lenguaje de consultas SPARQL. Para el presente proyecto el API Jena permitirá leer las ontologías y recuperar la estructura de la misma.

⁵ <http://www.apache.org/>

La elección de esta herramienta se debe a que una de las APIs, Jena Ontology API, provee una interfaz de programación que facilita la carga de ontologías en lenguaje OWL para su posterior tratamiento, lo cual permitirá implementar el módulo de desambiguación de términos.

1.5.1.4. RDF

El Marco de Descripción de Recursos (RDF, por sus siglas en inglés) es un modelo estándar para el intercambio de información en la Web. RDF es una recomendación de la *World Wide Web Consortium*⁶ (W3C) que utiliza el lenguaje XML para escribir los documentos en este formato y está diseñado para ser leído y entendido por las aplicaciones de computadoras (W3C, s.f.).

RDF amplía la estructura de enlaces de la Web mediante el uso de un Identificador Universal de Recursos (URI, por sus siglas en inglés) para nombrar la relación entre dos objetos, así como los dos extremos del enlace, este elemento es llamado Triple que consta de tres elementos: sujeto, predicado y objeto (Dublin Core, s.f.).

La elección de RDF se debe a que es un modelo estándar para la descripción de recursos que facilitará el intercambio de datos en la Web y es usado ampliamente en la *Linked Open Data*, por lo que su uso es indispensable si se desea delimitar los recursos ofrecidos en el repositorio semántico.

1.5.1.5. OWL

El Lenguaje de Ontologías Web (OWL, por sus siglas en inglés) es un lenguaje de marcado semántico que permite publicar y compartir ontologías en la Web. OWL está desarrollado como una extensión del vocabulario RDF y la información es recuperada en ontologías que pueden ser almacenadas en la Web como documentos codificados en lenguaje XML (Bechhofer 2009).

OWL es un lenguaje basado en la lógica computacional de modo que el conocimiento expresado en OWL puede ser aprovechado por los programas de ordenador, por ejemplo, para verificar la consistencia de ese conocimiento o para hacer explícito el conocimiento implícito. En ese sentido, OWL facilita un mecanismo para interpretar el contenido Web proporcionando un vocabulario adicional junto con una semántica formal (W3C, s.f.)

⁶ <http://www.w3.org/>

La elección de esta herramienta se basa en que es un estándar usado ampliamente para la representación de ontologías debido a que tiene mayor capacidad para expresar significados y semántica que XML, RDF y RDF-S. De este modo, OWL va más allá de estos lenguajes respecto a su capacidad para representar contenido interpretable por un computador en la Web. La representación de ontologías permitirá delimitar los contenidos con el fin de solucionar las ambigüedades.

1.5.1.6. SPARQL

SPARQL es un lenguaje estándar para consulta de grafos RDF recomendado por la *World Wide Web Consortium*. SPARQL se puede utilizar para expresar consultas a través de diversas fuentes de datos, si los datos se almacenan de forma nativa como RDF.

SPARQL contiene capacidades para consultar los patrones gráficos obligatorios y opcionales, además de sus conjunciones y disyunciones. Asimismo, es compatible con las pruebas de valor extensible y que limitan las consultas por fuentes de grafo RDF. Los resultados de consultas SPARQL pueden ser conjuntos de resultados o gráficos RDF (W3C, 2008).

La elección de esta herramienta se debe a que es una herramienta clave para realizar consultas sobre fuentes de información usando RDF, que es el estándar que se empleará en este proyecto para la estructuración de las anotaciones semánticas. SPARQL permitirá hacer consultas sobre las ontologías para solucionar las ambigüedades que se presenten en el contenido de un documento (Pérez, Arenas, and Gutierrez 2006).

1.5.1.7. MySQL

MySQL es un sistema de administración de base de datos relacional, multiplataforma, y multiusuario que cuenta con dos tipos de licenciamiento, una versión de código libre y otra de código privativo. La herramienta ha sido desarrollada en C y C++ y es adaptable a diferentes entornos de desarrollo, lo que permite su interacción con los lenguajes de programación más utilizados y su integración con distintos sistemas operativos (MySQL, 2015).

La elección de esta herramienta se basa en la necesidad de la persistencia de anotaciones semánticas de la aplicación que se va a desarrollar. MySQL, además de ser rápida y segura, es accesible por su licencia GNU GPL.

1.5.1.8. Lenguaje de Programación Java

Java es un lenguaje de programación orientado a objetos, cuya sintaxis está basada principalmente en C y C++.

Una de las principales características de Java es que permite desarrollar aplicaciones que se pueden ejecutar en una gran variedad de dispositivos; es decir, el código que es ejecutado en una plataforma no tiene que ser recompilado para correr en otra (Oracle, 2015).

1.5.1.9. NetBeans

NetBeans es un entorno de desarrollo integrado creado en Java que, además de ofrecernos un entorno de desarrollo en Java, soporta múltiples lenguajes de programación tales como C, C++, PHP, HTML, entre otros. Es de software libre y multiplataforma gracias al lenguaje sobre el que fue implementado. Entre sus principales características está la inclusión de herramientas que permiten generar el diseño gráfico de distintas aplicaciones en poco tiempo (NetBeans, 2015).

La elección de esta herramienta se debe a que es una IDE que permite la integración de SPARQL, Java y MySQL y, por ende, permitirá tener un mejor manejo del desarrollo del producto del proyecto.

1.5.2. Métodos y Procedimientos

1.5.2.1. Precisión y Recall

Precisión y *Recall* son métricas estadísticas para evaluar el rendimiento de un sistema de recuperación de información. La Precisión se encuentra dada por la relación entre el número de instancias recuperadas que son relevantes y el número de instancias recuperadas, mientras que el *Recall* está definido como el número de instancias recuperadas que son relevantes entre el número total de instancias relevantes en la colección que se esperan recuperar (Euzenat 2007).

La elección de este método se basa en que se desea contar con un indicador que permita medir la precisión de la herramienta para extraer información relevante de un documento, lo que es indispensable para verificar si se alcanzaron los resultados esperados.

1.5.3. Metodologías

1.5.3.1. CommonKADS

CommonKADS es una metodología que da soporte a la Ingeniería del Conocimiento estructurado, mediante una suite de modelos de ingeniería elaboradas a partir del conocimiento

humano. El modelo central en esta metodología es el modelo de la experiencia que organiza el comportamiento de la solución de problemas de un agente en términos de los conocimientos que aplica para realizar una determinada tarea. Los otros modelos capturan aspectos relevantes de la realidad como el contexto organizacional o la distribución de tareas a diferentes agentes (Schreiber 2000).

Una importante contribución de la metodología *CommonKADS* es su propuesta para estructurar el modelo de experiencia, que distingue tres tipos diferentes de los conocimientos necesarios para resolver una tarea en particular. Básicamente, los tres tipos diferentes corresponden a una visión estática, funcional y dinámica del sistema basado en conocimiento que se ha construido (Studer, Benjamins, and Fensel 1998).

La elección de esta metodología se basa en que se busca extraer la información requerida de los documentos en base a una ontología por lo que aportará de manera significativa a los objetivos del proyecto.

1.6 ALCANCE

El proyecto se encuentra dentro del área de las Ciencias de la Computación, y se realizará bajo el enfoque de la Ingeniería del Conocimiento. La realización de este proyecto de fin de carrera implica el desarrollo de una herramienta para la anotación semántica automática sobre documentos en la Web con formato PDF en base al procesamiento del contenido del documento; esto con el fin de dotarlos de información adicional al título o etiquetas, de manera que apoye la recuperación y gestión de información por otras aplicaciones de la Web semántica. Para realizar las anotaciones semánticas es necesario tener una ontología que va a ser configurada por el usuario, en base a la cual se identificarán los términos relevantes del documento anotado.

Para este proyecto se reusará y adaptará una ontología ya existente en el campo de la biomedicina, con el fin de realizar las pruebas y mostrar los resultados obtenidos por la herramienta que se desarrollará.

1.6.1. Riesgos

En la tabla 3 se muestra los riesgos identificados del presente proyecto, su impacto y las medidas correctivas que se tomarán para mitigar estos riesgos.

Tabla 3: Tabla de riesgos del proyecto.

N°	Riesgo Identificado	Impacto en el proyecto	Medidas correctivas para mitigar
1	Falta de ontologías del dominio elegido que impidan el desarrollo del proyecto de tesis	Alto	Verificar la disponibilidad de ontologías en el dominio elegido antes de comenzar la realización del proyecto
2	Actualizaciones imprevistas de las librerías o herramientas a usar que requieran modificar lo que se ha desarrollado del proyecto y genere retrasos en los entregables.	Bajo	Verificar constantemente y con anticipación las nuevas versiones de las herramientas para identificar si se requiere alguna modificación en el desarrollo del proyecto.

1.7 JUSTIFICACIÓN

Con el presente proyecto se espera aportar nuevo conocimiento en el área de la Web Semántica poniendo énfasis en una plataforma de dominio genérico. Se desarrollará una herramienta de anotación semántica automática, la cual permitirá realizar anotaciones automáticas que pueden ser usadas para investigaciones en el área de Recuperación de Información, Minería de Datos e Indexación de Recursos.

Esta herramienta permitirá a los estudiantes e investigadores que realicen búsquedas en la Web semántica, pues su propósito es recuperar información relevante para el tema de interés del usuario para sus estudios o proyectos de investigación. Asimismo, este modelo no está restringido a funcionar únicamente en búsquedas en un solo dominio, ya que la ontología utilizada como base de conocimiento será configurable por el usuario para realizar anotaciones en documentos relacionados a otras áreas, ampliando así la cobertura de beneficiarios a estudiantes de diferentes especialidades.

2. MARCO CONCEPTUAL

En el presente marco conceptual se mencionarán los términos más relevantes para el entendimiento del concepto de Web Semántica extraídos de la investigación bibliográfica realizada. La conceptualización de estos términos permitirá el entendimiento de cómo se puede representar el contenido de los documentos en la Web en una estructura formal en base a una ontología.

Se busca comprender los conceptos relacionados a Anotaciones Semánticas de Documentos, así como de las herramientas y tecnologías con las que se cuenta para realizar anotaciones, de manera que se pueda entender la problemática central de este proyecto y el entorno en el que se presenta.

2.1 WEB SEMÁNTICA

La Web semántica se concibe como una extensión de la Web que se conoce actualmente. Se puede decir que hoy en día la mayor parte de la Web es sintáctica, lo que indica que la búsqueda de información se realiza en base a coincidencias exactas de palabras o frases que se introduzcan en la consulta. Por otro lado, la Web semántica permite estructurar el contenido de cada página o documento en la Web con el objetivo de que los sistemas informáticos puedan procesar el contenido realizando búsquedas eficientes para el usuario (Davies, Fensel, and Van Harmelen 2003)

Para lograr que los sistemas inteligentes puedan entender el contenido de los documentos en la Web semántica se requiere que estos sean anotados con metadata. Esta metadata es información acerca de la información contenida en los documentos, que define sobre qué temas tratan dichos documentos en una forma procesable por las máquinas (Kim, Yoo, and Park 2012). La representación de la metadata bajo un dominio de conocimiento permitirá establecer una amplia red de conocimiento humano acompañado por el procesamiento de las máquinas (Davies, Fensel, and Van Harmelen 2003).

En ese sentido, los programas informáticos serán capaces de navegar, integrar y procesar información de una manera precisa y eficaz, pudiendo hacer inferencias para determinar en lo que el usuario realmente está interesado (García and Fernández 2005).

Una de las principales diferencias de la Web semántica con la Web actual es que esta última se asemeja a un grafo formado por nodos (documentos Web) del mismo tipo con arcos (enlaces) también del mismo tipo que no brindan información sobre el contenido o la

relación que hay entre los documentos. Por el contrario, en la Web semántica cada nodo tiene un tipo y los arcos presentan relaciones explícitamente diferenciadas (Castells 2003).

En la ilustración 1 se muestra la diferencia entre la Web actual y la Web semántica. Se observa que en la Web semántica cada nodo que representa un documento tiene un tipo definido (Familia de árboles oleáceas, árboles Olivo y Fresnos, frutos Aceituna y Samara) y los enlaces entre ellos contiene información sobre la relación que guardan (esTipoDe, esFamiliaDe, produce).

Bienvenido al sitio de información Oleaceae!

Oleaceae, las **Oleáceas**, son una familia de **plantas** perteneciente al **orden** **Lamiales**. Comprende 24 géneros de plantas leñosas, incluidos **arbustos**, **árboles** y **vides**. Se caracterizan por tener flores opuestas que pueden ser simples o pinnadas. Sus flores poseen **cáliz** y **corola** con cuatros lóbulos.

Descripción
 Son **árboles**, **arbustos** o **trepadoras** leñosas. Las plantas **hermafroditas** o **andromonoicas** o **dioicas**. Hojas opuestas (por ejemplo, en Nicaragua), trifoliadas o imparipinnadas. **Inflorescencias** terminales o axilares, cimosas o racimos, dicasiales, subumbeliformes, fasciculadas o panículas. Flores **actinomorfas**; sépalos 4 (–numerosos); pétalos 4 (–numerosos); anteras 2 (4), adnados al tubo de la corola; ovario súpero, bilocular, estilobdo bilobado o subcapitado, o estigma sésil. Frutona **drupa**, **baya**, cápsula o semilla.

Repositorio Semántico

Diagram illustrating the semantic relationships between concepts in the text:

- Lamiales** (order) is related to **Oleaceae** (family) via the relationship **order**.
- Oleaceae** (family) is related to **Plantae** (kingdom) via the relationship **Kingdom**.
- Plantae** (kingdom) is related to **Plantas** (type) via the relationship **label**.
- Plantas** (type) is related to **Plantae** (kingdom) via the relationship **label**.
- Plantae** (kingdom) is related to **Plant** (type) via the relationship **label**.

Ilustración 1: Diferencia entre Web Semántica y Web actual. Imagen adaptada de [10]

2.2 ONTOLOGÍA

Una de las definiciones más conocidas de ontología es la propuesta por Gruber (Gruber 1995b) : "una ontología es una especificación explícita de una conceptualización", de la cual se puede entender que una ontología es una abstracción esquemática de algún objeto o hecho del mundo real. La definición de Gruber fue extendida por Studer (Studer, Benjamins, and Fensel 1998): "Una ontología es una especificación formal y explícita de una conceptualización compartida". Esta definición logra captar los aspectos fundamentales del concepto de ontología. De ella se puede entender que una ontología representa los conceptos más relevantes y las relaciones existentes entre ellos dentro de un dominio, de una manera compartida y consensuada por un grupo; además, esta conceptualización debe

ser representada de manera formal, explícita y utilizable por los ordenadores (J. A. P. Sánchez 2011).

Las ontologías tienen los siguientes componentes que servirán para representar el conocimiento de algún dominio (Gruber 1995b):

- Conceptos: ideas básicas del dominio que se desea formalizar. Por ejemplo, pueden ser clases de objetos, métodos, planes, estrategias, procesos de razonamiento.
- Relaciones: representan la interacción y enlace entre los conceptos del dominio. Suelen formar la taxonomía del dominio. Por ejemplo: subclase-de, parte-de.
- Funciones: son un tipo concreto de relación donde se identifica un elemento mediante el cálculo de una función que considera varios elementos de la ontología. Por ejemplo, pueden aparecer funciones como categorizar-clase.
- Instancias: representan objetos determinados de un concepto.
- Axiomas: son teoremas que se declaran sobre relaciones que deben cumplir los elementos de la ontología. Por ejemplo, un axioma se puede definir de la siguiente manera: Si los árboles forestales y los árboles frutales son clases de árbol, entonces los árboles forestales no son una subclase de los árboles frutales.

En la ilustración 2 se muestra un ejemplo de ontología dentro del dominio de la botánica. Se pueden observar los conceptos de árbol, tipo de árbol (forestal y frutal), fruto y las relaciones y funciones que existen entre ellos.

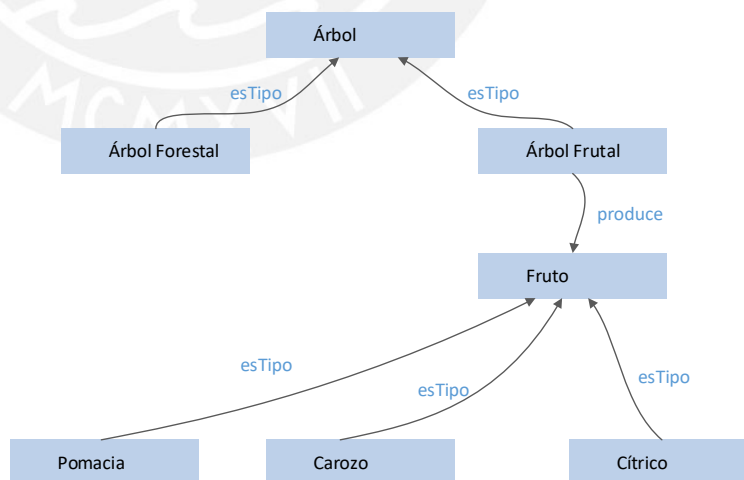


Ilustración 2: Ejemplo de Ontología. Imagen adaptada de [14]

El uso de ontologías en procesos de recuperación de información es muy importante, puesto que permite realizar consultas no limitando la búsqueda de elementos


solamente a partir del valor de sus propiedades, sino también amplía la búsqueda identificando los elementos involucrados en las relaciones y puede realizar inferencias a partir de las propiedades de herencia y las relaciones especificadas en los axiomas, lo que brinda mayor información (Tello 2001).

2.3 ANOTACIÓN

Una anotación es un objeto que es añadido o incrustado a un documento (Katayama et al. 2013). Las anotaciones pueden ser comentarios, notas o explicaciones. En una revisión de texto, las anotaciones incluyen texto subrayado, texto resaltado, notas y sellos que son añadidos a un documento. Cuando los estudiantes leen libros realizan anotaciones de subrayado o resaltado para enfatizar conceptos importantes y agregar notas cortas que resumen contenido (Niranatlamphong 2009). Por ejemplo, si se está revisando un documento de botánica se pueden hacer anotaciones en una ficha sobre información relevante de todo el libro o algún tema específico como las diferentes familias de árboles o las condiciones climáticas en las que pueden crecer.



Bienvenido al sitio de información Oleaceae!



Oleaceae, las Oleáceas, son una [familia](#) de [plantas](#) perteneciente al [orden Lamiales](#). Comprende 24 géneros que incluyen [arborescentes](#), [árboles](#) y [vides](#). Se caracterizan por tener flores opuestas que pueden ser simples o pinnadas. Sus flores poseen [cáliz](#) y [corola](#) con cuatro lóbulos.

Descripción
Son [árboles](#), [arborescentes](#) o [trepadoras](#) leñosas; [monocotiledóneas](#), [andromonoicas](#) o [dioicas](#). Hojas opuestas como en Nicaragua), trifoliadas o imparipinnadas, [infructosidad](#), [inflorescencias](#) terminales o axilares, cimosas o racimos, [diclasiales](#), subumbeliformes, fasciculadas o [racimosas](#). Flores [actinomorfas](#); sépalos 4 (–numerosos) valvados, a veces ausentes; corola generalmente simpétala, 4 (–numerosos) pétalos libres o ausentes; estambres 2 (4), adnados al tubo de la corola en las flores simples; anteras con dehiscencia longitudinal; ovario súpero, bilocular, estilo terminal con estigma bilobado o subcapitado, o estigma sésil. Frutona [drupa](#), [baya](#), cápsula o [sámara](#), a menudo con semilla.

Familia de :

- [Oleeae](#)
- [Picconia](#)
- [Priogymnanthus](#)
- [Schrebera](#)
- [Abeliophyllum](#)
- [Menodora](#)

Ilustración 3: Ejemplo de anotación sobre un documento Web. Imagen de autoría propia.

2.4 METADATO

Los metadatos se refieren a datos que brindan información acerca de otros datos. Son una porción de información secundaria que es independiente de la información primaria a la cual se refiere. Como ejemplos de metadatos se pueden mencionar a los esquemas,

restricciones de integridad, comentarios acerca de los datos, ontologías, los parámetros de calidad, anotaciones, de procedencia, y las políticas de seguridad (Srivastava and Velegrakis 2007).

En el campo de la Web semántica, la creación de metadatos es una de las técnicas más usadas para anotar documentos, ya que se pueden emplear con una amplia variedad de documentos y se pueden expresar en diferentes lenguajes y vocabularios (Corcho 2006). Estos proporcionan una estructura formal al contenido de los documentos en la Web con el propósito de que sean accedidos y gestionados de una manera más eficiente (Kiryakov et al. 2004).

Para la construcción de los metadatos es necesario definir una ontología que proporcione la semántica en la cual se basará la información estructurada (Corcho 2006). Las ontologías, así como los metadatos, han sido usadas para anotar documentos, puesto que las estructuras ontológicas incrementan el valor de las anotaciones semánticas permitiendo realizar inferencias y navegación conceptual, de esta forma los documentos son etiquetados con descripciones semánticas (Sasieta, Beppler, and do Santos Pacheco 2012).

Una ventaja de los metadatos es que pueden ser aplicados a los documentos en un variado grupo de formatos como HTML, PDF, Latex, entre otros. Esta aplicabilidad de metadatos a documentos se puede dar tanto en la Web, mediante servicios Web, así como en el disco duro de una computadora, a través de aplicaciones de escritorio. Los metadatos también pueden ser expresados en un variado rango de lenguajes (desde los naturales hasta los formales) y vocabularios; también, pueden ser expresados en diferentes formatos como electrónicos o físicos, y se les puede dar mantenimiento usando diferentes tipos herramientas (Corcho 2006).

Otra ventaja de los metadatos es que pueden ser usados como anotaciones digitales para aclarar las propiedades y semántica del contenido anotado. Esto conlleva un gran beneficio tanto para las personas como para las computadoras; por un lado, si se expresan en una forma entendible para las personas y se conocen su formato y conceptos, se puede obtener información útil y bien estructurada sobre determinado contenido; de otro lado, beneficia a las computadoras, ya que permite procesar automáticamente el contenido anotado (Agosti and Ferro 2007).

La ilustración 4 muestra un ejemplo del esquema conceptual de una anotación sobre un documento.

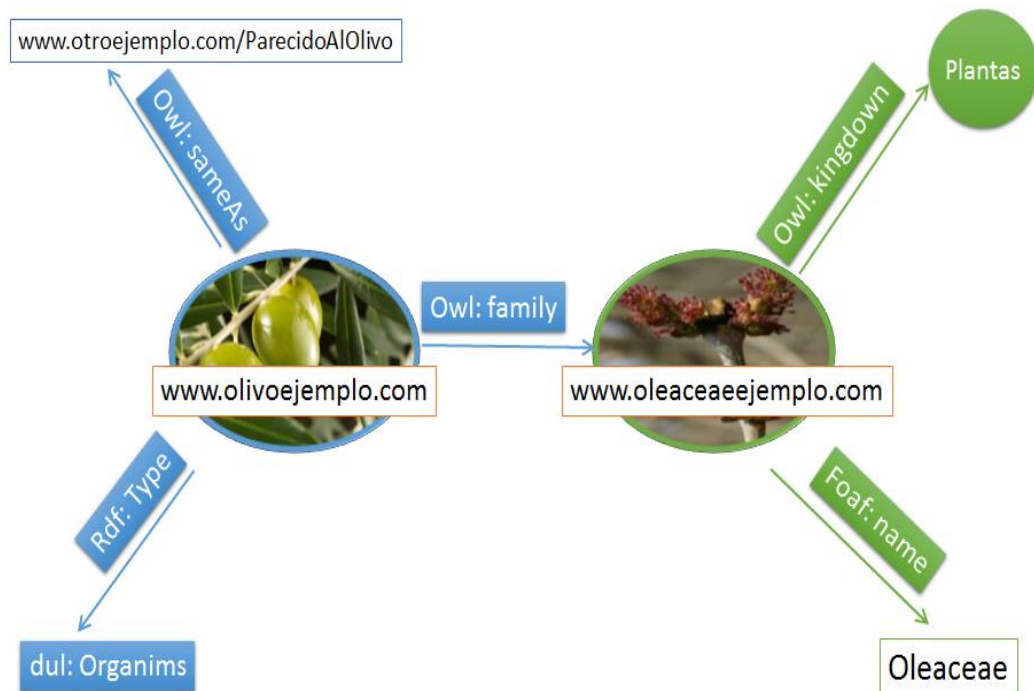


Ilustración 4: Ejemplo de metadato del documento Web en RDF. Imagen de autoría propia.

2.5 ANOTACIÓN SEMÁNTICA

Las anotaciones semánticas son un tipo específico de metadatos cuyo objetivo es permitir nuevos métodos de acceso a la información y extender los que ya existen (Kiryakov et al. 2004). En ese sentido, las anotaciones semánticas proporcionan referencias entre las entidades existentes en los recursos y conceptos de un dominio previamente modelado en una ontología (D. Sánchez, Isern, and Millan 2011).

En el ámbito de la Web semántica, las anotaciones brindan dos grandes mejoras relacionadas a la recuperación de información y a la interoperabilidad. Respecto a la primera mejora, las anotaciones semánticas permiten realizar búsquedas aprovechando las ontologías para hacer inferencias sobre datos provenientes de recursos heterogéneos. Sobre la interoperabilidad, la mejora se logra mediante el uso de una ontología común que proporciona un marco general para integrar la información de distintas fuentes de conocimiento (UREN et al. 2006).

Las anotaciones semánticas se almacenan en repositorios digitales que brindan mecanismos para optimizar el acceso y consulta sobre las anotaciones. Estas tienen asociados metadatos definidos mediante un esquema formal de descripción de recursos

como, por ejemplo, RDF. Los metadatos guardan información como fecha de creación de la anotación, nombre del autor, tipo de anotación, entre otros.

En la ilustración5 se puede observar un ejemplo de anotaciones semánticas hechas sobre un documento Web acerca de los conceptos relacionados a la ontología de la ilustración2.

The screenshot shows a webpage titled "Bienvenido al sitio de información Oleaceae!". On the left is a photograph of a plant with small, reddish flowers. To the right of the photo is text describing the Oleaceae family, including terms like "familia de plantas", "orden Lamiales", "árboles", "arborescentes", "hojas", "hermafroditas", "dioicas", and "estípulas". A red circle highlights the word "árboles" in the text. Below the text, there are three semantic annotations in blue boxes: "Owl: sameAs" pointing to "www.otroejemplo.com/ParecidoAlOlivo", "Rdf: Type" pointing to "dul: Organims", and "Owl: family" pointing to "dbpedia: Oleaceae".

Ilustración 5: Ejemplo de un documento con anotaciones semánticas en una Web. Imagen de autoría propia.

2.6 LINKED OPEN DATA

Linked Data fue propuesto por Tim Berners Lee en sus notas de arquitectura Web del 2009. Técnicamente hablando, Linked Data referencia a la data publicada en la Web de forma que pueda ser leída por las computadoras, su significado es explícitamente definido y está enlazado a otras bases de datos externas. Linked Open Data (LOD) viene a ser una consecuencia de Linked Data (J. A. P. Sánchez 2011).

LOD ha sido desarrollado por W3C, encargados de divulgar y explicar la Web semántica, y tiene por objetivo ampliar los alcances de la Web semántica mediante la indexación y publicación de bases de datos basados en RDF y su enlazamiento mediante el

mismo. LOD es Linked Data distribuida bajo una licencia abierta permitiendo su reutilización de manera gratuita (Berners-Lee, 2006).

Por ejemplo, al procesar un documento en el dominio de la botánica se revisa las anotaciones digitales del mismo y se buscaría en todos los documentos que están a otros documentos en la Linked Open Data mediante palabras claves todo lo relacionado al dominio especificado, como otros documentos de botánica, enlaces a páginas Web, entre otra información.

2.7 CONCLUSIÓN

Luego de identificar los conceptos más importantes relacionados a la Web Semántica se puede tener una idea del gran potencial que tiene para el beneficio de los usuarios en el sentido que permitirá recuperar información relevante de una manera eficiente y lograr que esta información pueda ser mantenida de forma ordenada evitando que exista redundancia en el contenido de los documentos en la Web. Esto será posible haciendo uso de las anotaciones semánticas que permitirán que la información sea entendible, tanto por los humanos como por las computadoras.

El entendimiento de estos conceptos permite comprender el objetivo de la herramienta que propone este proyecto de fin de carrera que trabajará en conjunto con las anotaciones semánticas, metadatos y ontologías.

3. ESTADO DEL ARTE

3.1 INTRODUCCIÓN

La anotación semántica es una herramienta esencial para alcanzar el objetivo de transformar la Web actual en una Web Semántica, ya que permite que el contenido de los documentos pueda ser entendido y procesado por las computadoras por medio del uso de ontologías para estructurar el conocimiento de un dominio (Al-Bukhitan, Helmy, and Al-Mulhem 2014). En ese sentido, debido al crecimiento exponencial de los recursos en la Web, se requiere de una herramienta que permita realizar anotaciones semánticas sobre los documentos de manera rápida y automática. El objetivo de esta revisión de estado del arte es dar a conocer los trabajos e investigaciones relacionados a las anotaciones semánticas automáticas que se han venido desarrollando en los últimos años.

3.2 MÉTODO DE REVISIÓN DEL ESTADO DEL ARTE

El método seleccionado para realizar la revisión del estado del arte es la revisión sistemática. Según Kitchenham, la revisión sistemática permite “identificar, evaluar e interpretar todas las investigaciones disponibles acerca de una pregunta de investigación particular, área temática o fenómeno de interés” (Kitchenham 2004).

La revisión sistemática consiste en tres fases: planificación, ejecución y presentación de informes sobre los resultados obtenidos. En la fase de planificación se debe identificar las necesidades de lo que se quiere investigar y definir un protocolo de revisión para realizar la búsqueda de material relacionado al tema que se está investigando (Kitchenham 2004; Riaz, Mendes, and Tempero n.d.).

El protocolo de revisión utilizado para esta sección incluye lo siguiente:

- Preguntas propuestas:
 1. ¿De qué manera se ha venido realizando anotaciones semánticas automáticas haciendo uso de las ontologías?
 2. ¿Cuáles son los principales estándares para definir metadatos?
 3. ¿Cuáles son las herramientas de anotación semántica automática que existen actualmente?
- Cadenas de búsqueda: para ejecutar la revisión se define la cadena de búsqueda para cada pregunta.

Pregunta 1: (*“automatic semantic annotation” or “automated semantic annotation” and “ontology”*)

Pregunta 2: (*“standards” and (“metadata definition” or “metadata editing”)*)

Pregunta 3: (*“(“automatic semantic annotation” or “automated semantic annotation”) and (“software” or “tool”)*)

- Fuentes de información: se seleccionarán los artículos en las siguientes bases de datos de fuentes primarias: ACM Digital Library⁷, IEEE Xplore Digital Library⁸, Science Direct⁹ y Scopus¹⁰.
- Criterio de inclusión: documentos de tipo primario en donde el título, resumen y palabras claves abarquen los siguientes temas: anotación semántica, ontologías, metadatos.
- Criterio de exclusión: documentos cuyos temas se alejen del objetivo de esta revisión o que sean de tipo secundario. También se excluyen documentos publicados antes de 2009, ya que se desea obtener información de estudios realizados en los últimos años.

Con el protocolo de revisión utilizado se procedió a la ejecución de la búsqueda de información. Esta fue efectuada el 15 de abril del 2017. En la siguiente tabla se muestra un resumen de la cantidad de estudios primarios identificados como relevantes en los resultados devueltos por cada cadena de búsqueda definida en la fase previa luego de aplicar los criterios de exclusión.

Tabla 4: Resumen de la cantidad de resultados devueltos por las bases de datos.

N°	Cadena	Número de documentos encontrados.
1	<i>“(“automatic semantic annotation” or “automated semantic annotation”) and “ontology”)</i>	23
2	<i>“(“standards” and (“metadata definition” or “metadata editing”))</i>	18
3	<i>“(“automatic semantic annotation” or “automated semantic annotation”) and (“software” or “tool”)</i>	15

⁷ <http://dl.acm.org/>

⁸ <http://ieeexplore.ieee.org/>

⁹ <http://www.sciencedirect.com/>

¹⁰ <http://www.scopus.com/>

3.3 ¿De qué manera se ha venido realizando anotaciones semánticas automáticas haciendo uso de las ontologías?

Tabla 5: Resumen de la respuesta a la pregunta 3.3.

N°	Método	Documentos de referencia
1	Obtener automáticamente la metadata y anotar los documentos con la metadata extraída.	(Corcho 2006; De Maio et al. 2014)
2	Emplear una ontología previamente creada.	(Handschuh and Staab 2003; Rajput 2014; D. Sánchez, Isern, and Millan 2011)

La falta de la semántica en los documentos Web actuales hace que estos sean comprensibles solo por humanos y no por las computadoras (Corcho 2006). El problema de algunos motores de búsqueda es que se basan en búsquedas por palabras clave y no distinguen entre el contenido de diferentes dominios de conocimiento (Rajput 2014).

La Web semántica impulsa la anotación de documentos o contenidos generales de recursos Web mediante la creación de metadatos, utilizando la información semántica de las ontologías de dominio (De Maio et al. 2014).

Las anotaciones semánticas van más allá de anotaciones textuales, que están dirigidas a la inserción de palabras clave para su uso por el creador del documento; en cambio, las anotaciones semánticas permiten individualizar los conceptos, y las relaciones que existen entre ellos, en el contenido de un documento con el objetivo de que sean entendibles por humanos, pero sobre todo por las máquinas, de manera que otorgue valor y sea consistente con los esquemas y ontologías que se adopten (De Maio et al. 2014; Rajput 2014). Diferentes ontologías representan diferentes conceptualizaciones de conocimiento; por lo tanto pueden ser utilizadas para recuperar información del mismo documento en función del conocimiento especificado en la ontología (Rajput 2014).

Para lograr que las anotaciones semánticas se realicen de forma automática se maneja básicamente dos opciones. Una forma es obtener automáticamente la metadata y anotar los documentos con la metadata extraída, sin embargo, este enfoque enfrenta el difícil problema que

implica la generación de la ontología por lo que es muy poco utilizado para realizar anotaciones (Handschuh and Staab 2003; D. Sánchez, Isern, and Millan 2011). El enfoque más utilizado para realizar anotaciones semánticas automáticas consiste en emplear una ontología previamente creada; de esta forma, la creación de la ontología se trabaja separadamente y la tarea principal se dirige hacia la anotación semántica solamente (Handschuh and Staab 2003).

3.4 ¿Cuáles son los principales estándares para definir metadatos?

Tabla 6: Resumen de la respuesta a la pregunta 3.4.

Nº	Estándares para definición de metadatos	Documentos de referencia
1	HTML	(W3C, Metadata, s.f.) , (Duval et al. 2002), (Kobayashi and Takeda 2000)(Corcho 2006; De Maio et al. 2014)(Corcho 2006; De Maio et al. 2014), (Corcho 2006)
2	XML	(W3C, Metadata, s.f.) , (Duval et al. 2002), (Kobayashi and Takeda 2000)
3	RDF	(W3C, Metadata, s.f.) , (Duval et al. 2002), (Kobayashi and Takeda 2000)

La *World Wide Web Consortium* (W3C) ha compilado una lista de recursos para las propuestas de información y estandarización de metadatos (W3C, Documento en línea, s.f.). Los estándares de metadatos más publicitados son *Dublin Core Metadata Standard*, *IEEE Learning Object Metadata* y *Warwick Framework* (Ur Rehman, Anwer, and Iftikhar 2005). De estos estándares, el más extendido es el *Dublin Core* que es un conjunto de 15 elementos de metadatos propuesto para facilitar la recuperación de información rápida y exacta (W3C, Metadata, s.f.)(Kobayashi and Takeda 2000).

Los elementos que propone *Dublin Core* para definir recursos generales son título, autor, tema, descripción, editor, contribuyentes, fecha, tipo de recurso, formato, identificador de recurso, fuente, lenguaje, relación, cobertura y derechos (W3C, Metadata, s.f.). Con estos elementos se puede describir diferentes tipos de materiales como videos, sonidos, imágenes, textos y páginas Web (Duval et al. 2002).

Para la descripción de los recursos se utiliza un esquema de codificación de metadatos que pueden ser HTML, XML o RDF. El esquema para etiquetado de documentos HTML es el

más simple, pero presenta el inconveniente de que sólo se puede utilizar para describir el documento al que está unido (Duval et al. 2002). Para resolver este inconveniente, la W3C propuso el Marco de Descripción de Recursos, RDF, que es usado como esquema de codificación de metadatos para documentos Web (Kobayashi and Takeda 2000).

3.5 ¿Cuáles son las herramientas de anotación semántica automática que existen actualmente?

En este punto se da a conocer las herramientas de anotación semántica que han sido desarrolladas hasta la fecha, abril de 2017.

3.5.1. THEOPHRASTUS

Theophrastus es un sistema que soporta anotaciones semánticas en la fuente de información original y provee servicios de exploración a través de la minería de datos y la explotación de Linked Open Data. Estos servicios son proporcionados en tiempo real y según demanda del usuario. El sistema está basado en los requisitos del dominio de la biodiversidad, en el cual el principal problema que afrontan los miembros de la comunidad es la identificación y recuperación de información relacionada acerca de una especie tornándose un proceso lento y dificultoso (Fafalios and Papadakos 2014).

En la ilustración 6 se muestra el proceso soportado por el sistema para su funcionamiento:

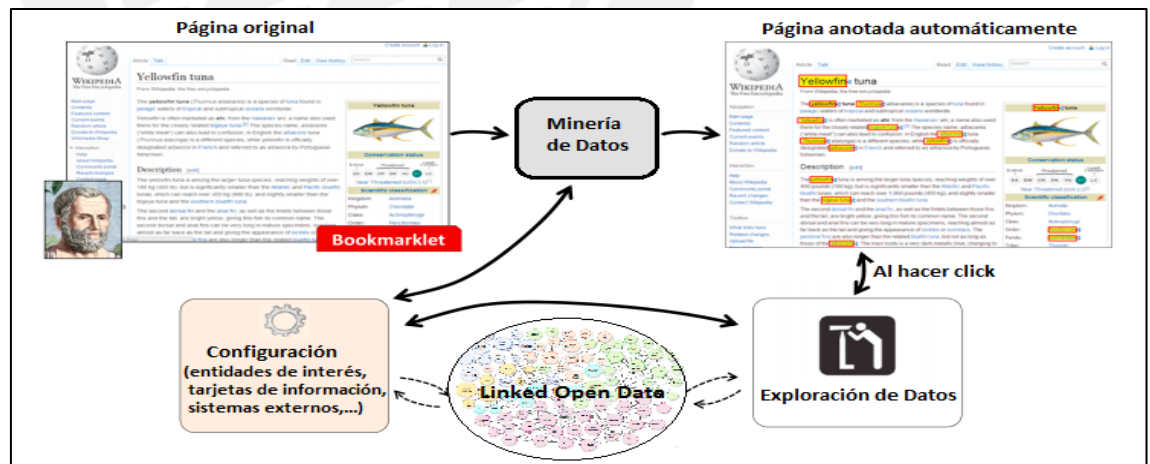


Ilustración 6: Proceso soportado por el sistema Theophrastus. Imagen adaptada de [41]

En la imagen se muestra que en primer lugar se realiza la minería de datos a partir del documento original sobre el cual se harán las anotaciones semánticas. Una vez realizadas las anotaciones, continúa la exploración de datos en Linked Open Data de las entidades reconocidas.

Finalmente, se observa que el sistema es configurable por el usuario; por ejemplo, se puede especificar cómo hacer coincidir semánticamente una entidad identificada y cómo encontrar recursos relacionados ingresando plantillas de consultas SPARQL e indicando la base de conocimiento donde se realizará la búsqueda; del mismo modo, se puede especificar sistemas externos de búsqueda y agregar nuevas categorías de entidades.

Profundizando en el proceso de anotación semántica, *Theophrastus* lee el contenido del documento seleccionado y utiliza una herramienta de reconocimiento de entidades, *Gate Annie* (Annie, s.f.), para identificar las mismas. Para ello no procesa la microdata o RDF que pueda estar incrustado en el documento, sino solo identifica las entidades de interés en base a su actual configuración. En el caso de documentos Web las anotaciones se realizan en el documento original, mientras que para archivos PDF las anotaciones se muestran en una barra lateral externa al texto.

3.5.2. FLERSA

FLERSA (*Flexible Range Semantic Annotation*) es una herramienta que permite convertir el contenido de documentos Web, expresados en lenguaje natural, en contenido semántico, el cual es enriquecido con metadatos en formato estructurado (Navarro-Galindo and Samos 2010).

La herramienta fue desarrollada para permitir tanto anotaciones como búsquedas semánticas. En lo referente a anotaciones semánticas, permite que estas se puedan hacer de forma manual o automática; para el primer caso emplea una técnica de marcado de rangos flexibles (partes arbitrarias del documento analizado delimitadas por un comienzo y un fin) basada en el estándar RDFa, mientras que, para las anotaciones automáticas emplea un enfoque híbrido basado en técnicas de aprendizaje tales como el Modelo de Espacio Vectorial, en el que se asignan pesos a cada elemento del documento, y N-gramas, que se basa en la frecuencia de los elementos dentro del documento.

FLERSA es fácil de usar desde la Web y está integrada con los navegadores para hacer sencilla su interacción con los usuarios. Hace uso exclusivo de estándares abiertos tales como RDF, OWL y RDFa para promover la interoperabilidad y extensibilidad. En cuanto a la arquitectura, se emplea el modelo Cliente-Servidor permitiendo que múltiples usuarios realicen anotaciones en múltiples páginas Web de forma simultánea.

3.5.3. UTOPIA DOCUMENTS

Utopia Documents es una aplicación de escritorio para la lectura y exploración en formato PDF que integra semánticamente herramientas de visualización y análisis de datos con

artículos con artículos de investigación publicados. La herramienta transforma artículos estáticos en un puente hacia un conocimiento adicional enlazando la información, tanto implícita como explícita, de los documentos con recursos en línea, así como proveyendo acceso transparente a los datos adicionales y a herramientas de visualización y análisis de datos interactivas. La implementación de las mejoras mencionadas las realiza sin comprometer la integridad de los del archivo PDF (Attwood et al. 2010).

La arquitectura de software mostrada en la ilustración 7 se compone de tres componentes principales: El Núcleo que provee los mecanismos generales para la visualización y manipulación de los artículos PDF, los *plugins* que analizan, anotan y visualizan las características del documento de forma automática o bajo la guía del usuario, y, por último, las ontologías que son usadas para integrar semánticamente los otros componentes.

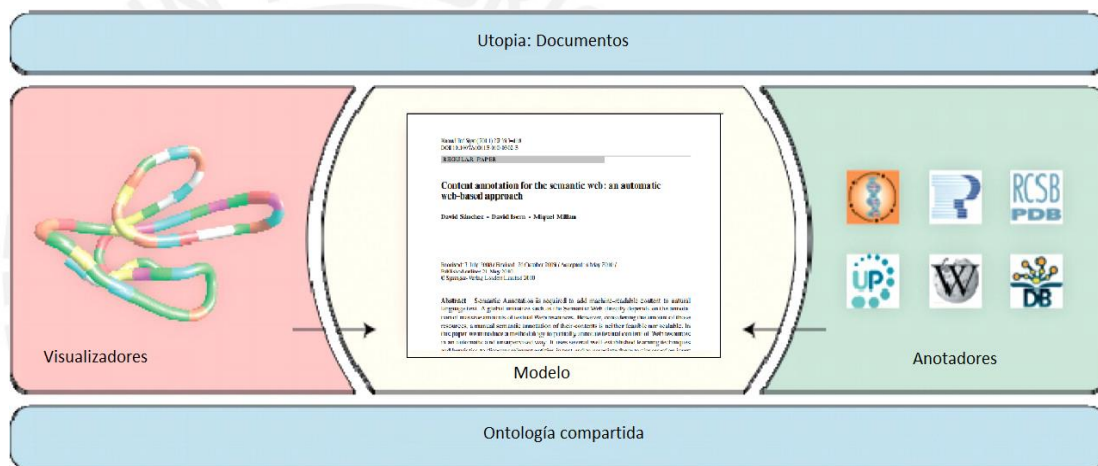


Ilustración 7: La arquitectura de Utopia Documents, mostrando la relación entre la GUI, plugins y ontologías. Imagen adaptada de [43]

3.5.4. OnTheFly

OnTheFly es una aplicación basada en Web que permite el reconocimiento de entidades en el dominio de la biología para enriquecer documentos como Microsoft Word, Power Point, Excel, PDF y de texto plano. Para ello convierte los archivos de entrada al formato HTML y los envía al servidor de etiquetado *Reflect* (Pafilis et al. 2009), en el cual se resaltan los nombre de entidades como genes, proteínas y productos químicos; luego agrega código JavaScript para invocar una ventana emergente de resumen que ofrece una visión general de la información relevante sobre la entidad y enlaces a otros recursos en línea relevantes. Esta aplicación también es capaz de extraer las bioentidades reconocidas en un conjunto de archivos y producir una

representación gráfica de la red de conocimiento y las asociaciones entre entidades (Pavlopoulos et al. 2009).

OnTheFly emplea una arquitectura Cliente-Servidor que se muestra en la ilustración 8. En ella se observa, además, el funcionamiento de la aplicación. Primero se muestra una tabla de las anotaciones hechas sobre un documento (A); luego, a partir de estas anotaciones, se genera una ventana emergente (B) con la información relevante de cada entidad anotada. Finalmente, se muestra la representación gráfica de la red de conocimiento (C) y la arquitectura de la aplicación (D) en la que todo el intercambio de datos se da mediante el protocolo HTTP. En la arquitectura de la aplicación se hace uso de dos servidores, el servidor de conversión transforma los documentos cargados en la interfaz al formato HTML, luego el documento en este formato es enviado al servidor reflector que retorna el documento con las anotaciones hechas.

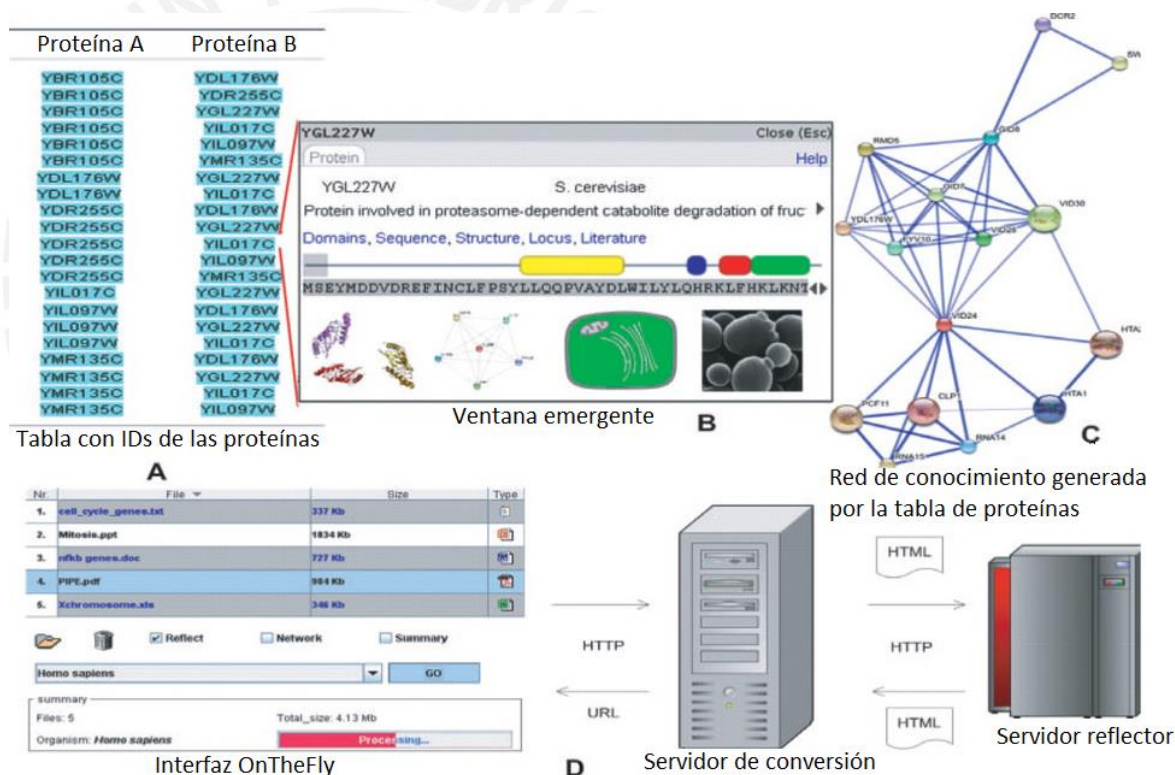


Ilustración 8: Aplicación Web OnTheFly. Imagen adaptada de [45]

3.5.5. DOME0 ANNOTATION TOOLKIT

DOME0 (*Document Metadata Exchange Organizer*) es una aplicación Web extensible que permite a los usuarios crear y compartir metadatos de anotaciones basadas en ontologías hechas sobre documentos HTML o XML de una manera eficiente, usando el modelo RDFa para Anotación mediante Ontologías (AO). Esta herramienta que está limitada al dominio de la

biomedicina soporta anotaciones manuales, semi-automáticas y totalmente automáticas con una total representación de la procedencia de las anotaciones, así como anotaciones personales o comunitarias con autorización y control de acceso (Ciccarese, Ocana, and Clark 2012).

La interfaz de usuario de DOMEEO es un componente Web extensible en el cual los usuarios pueden invocar directamente anotaciones de documentos en línea en formato HTML, XHTML y XML. La aplicación permite cargar documentos HTML como si se cargaran directamente desde el navegador; esto puede ser realizado copiando y pegando la dirección URL del documento en la barra de direcciones de DOMEEO o usando un *plugin* en el explorador Firefox que agrega un icono de la aplicación en su barra de estado.

El servidor de DOMEEO, en el cual se busca la información para realizar las anotaciones semánticas, puede ser desarrollado en cualquier lenguaje o plataforma que permita publicar página Web que aloje el JavaScript de la aplicación. La aplicación minería de entidades y accede a ontologías, así como a otras facilidades para el marcado automático a través de llamadas de servicios Web.

3.5.6. Semantic Annotation Tool for Annotating Arabic Web Documents

El idioma árabe ha recibido poca atención en las investigaciones sobre Web Semántica en comparación con lenguajes latinos, especialmente en el campo de la anotación semántica. Por ello, este estudio propone una herramienta para la anotación semántica automática de documentos Web en idioma árabe. La herramienta toma la URL del recurso Web y la ontología involucrada para producir anotaciones externas al documento usando el lenguaje RDF (Al-Bukhitan, Helmy, and Al-Mulhem 2014).

En la ilustración 9 se muestra la arquitectura del sistema de anotación semántica propuesto.

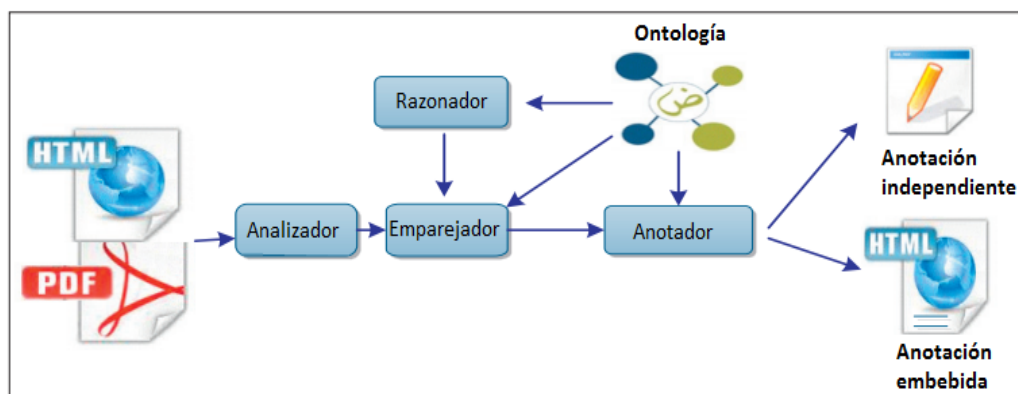


Ilustración 9: Arquitectura del sistema de Anotación Semántica. Imagen recuperada de [32].

El módulo Analizador se encarga del análisis y extracción de la data textual del documento Web. El módulo emparejador es responsable de identificar las similitudes entre las frases contenidas en el texto de los documentos y los términos de recursos de ontología. Este módulo interactúa con el componente razonador con el fin de identificar la correspondencia entre los recursos de la ontología. Finalmente, el módulo Anotador se encarga de generar las anotaciones sobre el documento original, estas pueden ser embebidas o externas al documento. Los documentos anotados pueden estar disponibles en la Web para el uso de los usuarios y sistemas de recuperación de información inteligentes.

En el siguiente cuadro, se muestra un resumen de las principales herramientas encontradas permitiendo comparar el dominio de conocimiento sobre el cual se aplican, los recursos utilizados para el desarrollo, el tipo de anotación semántica, los tipos de documentos que pueden ser procesados, y con qué otras aplicaciones son compatibles.

Tabla 7: Cuadro comparativo de aplicaciones de anotación semántica.

Propuesta	Dominio	Recursos Utilizados	Tipo de anotación	Tipo de documentos	Compatibilidad
THEOPHRASTUS	Biología	RDF, OWL	Automática	Documentos Web, documentos PDF.	Aplicación Web
FLERSA	Múltiples ontologías	XML, RDF, RDFa, W3C's OWL	Manual, Automática.	Documentos Web	Internet Explorer, Mozilla Firefox, Chrome y Opera
UTOPIA DOCUMENTS	Bioquímica	XML, RDF, OWL	Automática	Documentos PDF	Aplicación de Escritorio
ONTHEFLY	Biología	CMI, Red	Manual, Automática.	Documentos Office	Aplicación WEB
DOMEO	Biomedicina	XML, RDF, OWL	Semi-automática.	Aplicación WEB	Aplicación WEB

3.6 CONCLUSIÓN

En esta sección se ha revisado las propuestas tecnológicas desarrolladas en los últimos años que buscan una alternativa de solución a la problemática planteada de poder estructurar formalmente el contenido de los documentos en la Web usando ontologías con el objetivo de dar soporte a la recuperación de información para que esta sea relevante para el usuario. En base a estas propuestas y las investigaciones ya realizadas, se tiene una idea más clara de las tecnologías estándares con las que se cuenta para poder desarrollar una herramienta que satisfaga la necesidad de la Web Semántica en lo concerniente a la estructuración formal de la información mediante anotaciones semánticas automáticas.

Se puede concluir que el uso de OWL (*Ontology Web Language*) y RDF (*Resource Description Framework*) está estandarizado para el desarrollo de ontologías y la realización de anotaciones, respectivamente, ya que están presentes en prácticamente todos los proyectos revisados. Estos estándares son de mucha ayuda, puesto que se cuenta con repositorios semánticos y repositorios para anotaciones RDF que permiten ser reutilizadas y facilitan la búsqueda y extracción de contenido.



4. CAPÍTULO 4: PROCESAMIENTO DE LA INFORMACIÓN TEXTUAL DE LOS DOCUMENTOS

La herramienta que se va a desarrollar como proyecto de fin de carrera permitirá realizar anotaciones semánticas a partir del análisis del contenido de los documentos que ingresen como parámetros. Estos documentos pueden estar elaborados en diversos formatos; sin embargo, para probar los resultados alcanzados se optó por el formato PDF, debido a la gran cantidad de documentos que se encuentran en la Web en este formato. El formato PDF es ampliamente utilizado para el intercambio y presentación de documentación porque es estándar y multiplataforma, además de ser livianos en cuanto al tamaño del archivo y la característica de no ser modificables lo que mantiene la integridad de la información contenida en él.

Los documentos PDF no pueden ser procesados directamente por la computadora para el análisis de su contenido; por esta razón, se escogió la herramienta Apache PDFBox que mediante las funciones de su librería permite extraer el texto plano de los documentos que se utilizan como datos de entrada.

En ese sentido, en este capítulo se detalla la implementación de un módulo que permita extraer el contenido textual de los documentos. Los resultados alcanzados con este objetivo específico son importantes para el posterior análisis semántico de los términos contenidos en el documento en base a la ontología que defina el usuario de la herramienta.

4.1 Objetivo Específico N° 1: Implementar un módulo que permita procesar la información textual de los documentos.

Una razón importante que contribuye a la dificultad de búsqueda en la Web, es que los formatos e interfaces de contenidos y servicios están representados en formatos comprensibles solo por personas y no por las máquinas (J. A. P. Sánchez 2011; Tello 2001). Por ello, el primer objetivo del proyecto para poder realizar anotaciones semánticas es implementar un módulo que soporte el procesamiento de la información textual de los documentos que estén representados en formato PDF.

Las funciones de este módulo serán las de extraer, analizar y procesar el contenido textual del archivo de entrada. Para realizar estas funciones se aplicará técnicas y herramientas de NLP. NLP es un campo de las Ciencias de la Computación que ofrece un conjunto de técnicas computacionales para el análisis y representación de textos de origen natural en uno o más niveles de análisis lingüístico con el propósito de procesar el lenguaje humano para determinadas tareas (Liddy et al. 2003).

El procesamiento del contenido textual consistirá en el análisis de cada oración dentro del documento, esto será posible mediante el uso de las técnicas de NLP como la separación de oraciones, tokenización y lematización (reducción de palabras a su forma base).

4.2 Resultado Alcanzado N°1: Módulo para obtener los términos que se encuentran en un documento.

Para alcanzar este resultado se implementó un módulo que recibe como entrada un documento en formato PDF y generará como salida una lista con los términos contenidos en él.

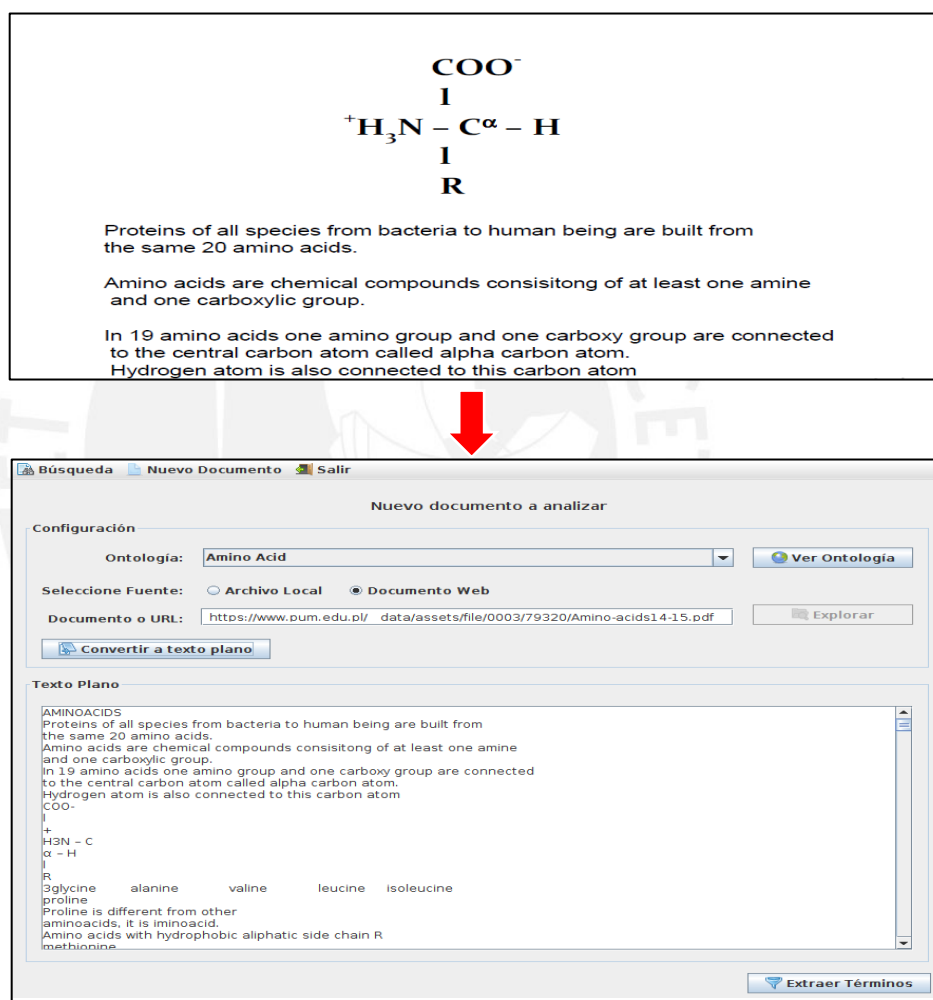


Ilustración 10: Funcionalidad de extracción de contenido textual. Imagen tomada de la herramienta desarrollada (Autoría propia).

Para realizar esta tarea, como primer paso, se debe extraer todo el contenido textual del documento y convertirlo en un conjunto de caracteres (texto plano). La ejecución de esta tarea se realizó haciendo uso de la herramienta Apache PDFBox, la cual permite ingresar como parámetro el nombre del archivo en formato PDF y llamando al método *extraerContenido*

implementado en la clase *FuncionesPDF* convierte el contenido en texto plano como se muestra en la ilustración 10.

Como segundo paso, a partir del contenido convertido en texto plano se genera una lista de todos los términos contenidos en el documento. Esto se realiza mediante el empleo de una herramienta de NLP que para este caso será la API en lenguaje Java de la librería FreeLing. La forma en la cual FreeLing realiza esta tarea es dividiendo el contenido en párrafos que a su vez son divididos en oraciones y estos, finalmente, se dividen en palabras.

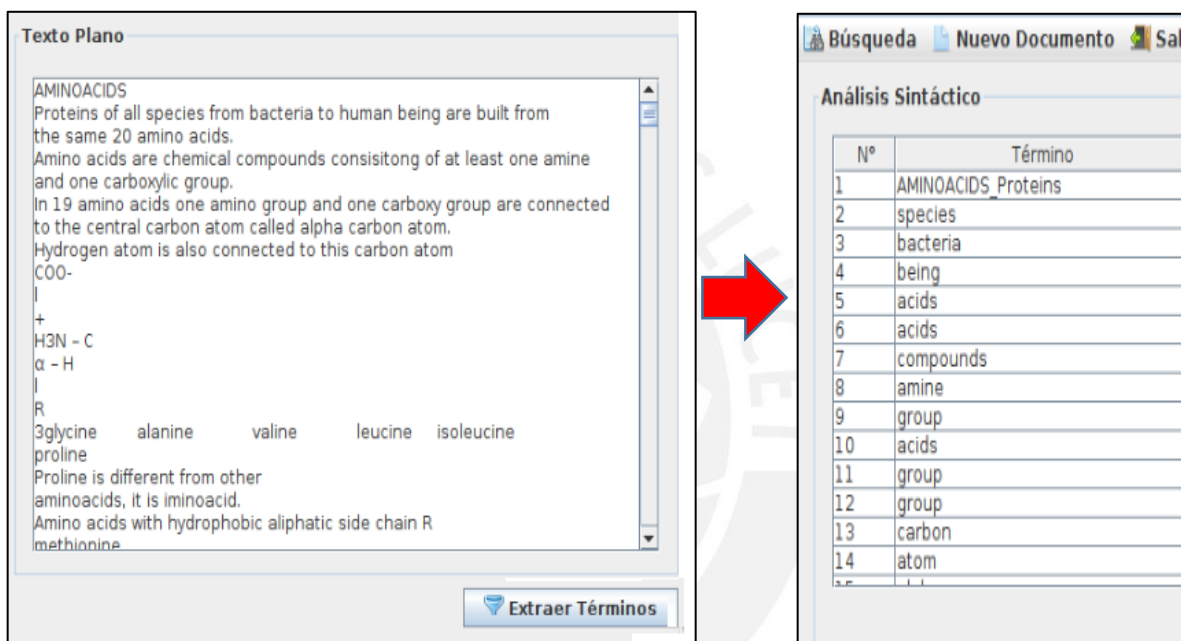


Ilustración 11: Funcionalidad de extracción de términos. Imagen tomada de la herramienta desarrollada (Autoría propia).

Finalmente, sobre la lista de palabras obtenidas la herramienta se centrará en extraer todos los sustantivos y adjetivos; esto, con el fin de obtener la mayor cantidad de conceptos posibles en el documento. El resultado de la implementación de este módulo se muestra en la ilustración 11, en la cual se observa una lista con los términos extraídos del texto plano que recibe como parámetro.

Posteriormente, se analizará semánticamente cada palabra obtenida para verificar si tiene relación con la ontología.

4.3 Resultado Alcanzado N°2: Módulo que permite reducir las palabras a su forma base o lema.

El resultado alcanzado para este objetivo es importante para el análisis semántico que se realizará de los conceptos del documento en base a la ontología configurada por el usuario porque busca recuperar el concepto de cada palabra al obtener su lema.

Como primer paso para alcanzar este resultado se realizó un pre-procesamiento de la lista de palabras obtenidas con el Resultado Alcanzado N° 1 con la finalidad de revisar que todas las palabras estén completas y no separadas por caracteres especiales; por ejemplo, una palabra puede estar separada por un guion debido a que se produce un cambio de línea en el documento. El pre-procesamiento consiste en analizar cada palabra extraída del texto plano para saber su posición en el documento y dependiendo del motivo por el cual pueda estar incompleta una palabra verificar si es necesario realizar una concatenación para extraer la palabra completa.

Para el análisis semántico de los conceptos la herramienta procesa, principalmente, la forma base de la palabra encontrada en el texto, debido a que en el idioma existen múltiples conjugaciones para sustantivos verbos y adjetivos. De esta manera, representar todas las conjugaciones de una palabra en una ontología sería un trabajo muy tedioso; es por ello que en el plano de la gestión del conocimiento se usa, generalmente, la forma base del concepto para representarlo.

Sin embargo, existen conceptos o sustantivos que no poseen lema, por lo tanto, para efectuar este análisis semántico se consideró tanto el concepto en la forma en que se encuentre en el documento, así como la forma base obtenida por la herramienta FreeLing, esto con el objetivo de que se pueda encontrar la mayor cantidad de conceptos posibles en la ontología.

En la ilustración 12 se muestra el resultado obtenido de la implementación de este módulo, en el cual se observan las palabras del texto y su correspondiente lema.

4.4 Discusión de los resultados alcanzados N°1 y N°2

Luego de la extracción del contenido de los documentos es necesario procesarlo para obtener los términos que serán analizados con la ontología. Para ello, se eligió la herramienta FreeLing luego de comparar las funcionalidades que ofrece contra otras herramientas de NLP con soporte para el idioma español. La elección consideró la variedad de funciones ofrecidas como análisis morfológico, etiquetado gramatical, entre otros. Del mismo modo, se consideró el constante soporte que ofrece su equipo de desarrollo para el idioma y los resultados de las pruebas realizadas.

N°	Término	Lema
1	AMINOACIDS_Proteins	aminoacids_proteins
2	species	species
3	bacteria	bacteria
4	being	being
5	acids	acid
6	acids	acid
7	compounds	compound
8	amine	amine
9	group	group
10	acids	acid
11	group	group
12	group	group
13	carbon	carbon
14	atom	atom

Ilustración 12: Funcionalidad de reducción de palabras a su forma base. Imagen tomada de la herramienta desarrollada (Autoría propia).

Del mismo modo, se observa en los resultados que el procesamiento del contenido textual puede perder eficiencia al encontrar palabras que son divididas por un guion al final de una línea por falta de espacio, ya que en esa situación no será posible identificar la palabra mediante la herramienta. Para contrarrestar esta situación se optó por realizar un procesamiento posterior de todos los términos extraídos para verificar que estén correctamente identificadas.

También, se pudo observar que el reconocimiento de nombres de entidades es poco efectivo, pues los identifica porque comienzan con mayúsculas y los concatena mediante un guión pudiendo unir entidades diferentes como en el caso del término obtenido “AMINOACIDS_Proteins”. En el texto se observa que la palabra “AMINOACIDS” hace referencia al título del documento y la palabra con la que comienza el primer párrafo es “Proteins”.

La deficiencia en la extracción de términos y nombres en general se puede mitigar empleando herramientas de reconocimiento nombres en textos en español como, por ejemplo, *Smorph* y *Nooj* que son sistemas que detectan automáticamente expresiones y estructuras propias del idioma español (Tramallino 2014). Esta sería una buena opción para trabajo futuro que contribuiría a mejorar el rendimiento de la herramienta.

Para validar el resultado obtenido por el procesamiento textual del documento y la generación de la lista de palabras en su forma base se comparó con el procesamiento manual que se hizo del contenido, en el cual se identificó las palabras y su forma base.

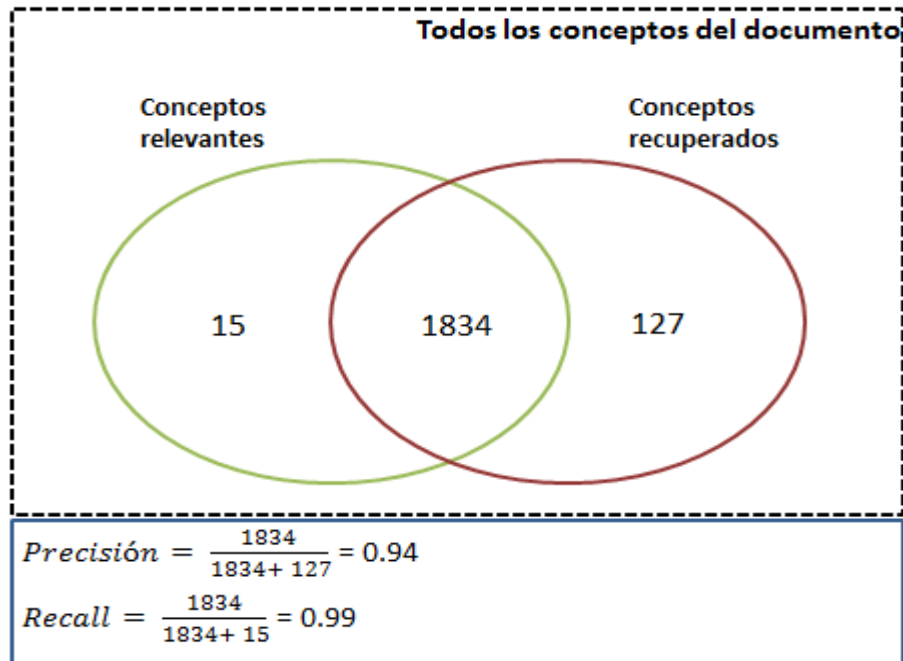


Ilustración 13: Resultados de prueba de Precisión y Recall.

Según los resultados obtenidos, los valores de Precisión de 0.94 y Recall de 0.99 permiten concluir que el método de extracción de términos utilizado puede perder eficacia al extraer términos que no son parte del contenido, como, por ejemplo, notas al pie de página, bibliografía, numeraciones, entre otros. Sin embargo, los resultados son buenos para poder realizar el análisis semántico, pues los conceptos relevantes que no son recuperados son pocos en comparación con los recuperados.

5. CAPÍTULO 5: REPRESENTACIÓN DE LA INFORMACIÓN MEDIANTE LENGUAJE ESTRUCTURADO.

Uno de los principales problemas que dificulta la recuperación de información relevante para el usuario es que los documentos que se pueden encontrar en la Web no presentan una estructura estándar en su elaboración (D. Sánchez, Isern, and Millan 2011). Esto puede darse por diferentes motivos, uno de ellos es que los usuarios que suben contenido a la Web son libres de elaborar sus documentos en cualquier formato. Así también, la escritura de estos documentos se realiza usando lenguaje natural y en diferentes idiomas, lo cual hace que la construcción de las consultas sea complicada.

Por tal motivo, en este capítulo se detalla la implementación de un módulo que permita estructurar de manera formal el contenido del documento analizado por la herramienta en base a la ontología seleccionada por el usuario.

5.1 Objetivo Específico N° 2: Representar la información de los documentos mediante un lenguaje estructurado para la generación de anotaciones.

Luego de haber obtenido los resultados del primero objetivo específico se tiene una lista de todos los términos contenidos en el documento analizado por la herramienta. Además, se tiene una lista de las formas base o lema de los términos extraídos. Esta información será utilizada como entrada para lograr los resultados del segundo objetivo específico del proyecto que busca representar la información del documento mediante un lenguaje estructurado para luego poder generar las anotaciones semánticas.

De esta manera, para representar la información de manera estructurada es necesario, como primer paso, clasificar cada término del documento con su categoría gramatical correspondiente para discriminar los términos que brindan mayor información acerca del contenido del archivo. Seguidamente, estos términos son procesados por la herramienta para determinar si tienen relación con la ontología definida por el usuario.

5.2 Resultado Alcanzado N° 3: Módulo que permita realizar el etiquetado gramatical.

El etiquetado gramatical de cada término de la lista obtenida con el primer objetivo específico es importante porque permite filtrar solo los términos que corresponden a las categorías de sustantivos y adjetivos. La elección de solo estas dos categorías se debe a que se busca obtener la mayor cantidad de conceptos posibles en el documento analizado y las demás categorías gramaticales no brindan mayor información acerca del contenido (Tramallino 2014).

La ejecución de esta tarea se realizó implementando la clase *AnalizadorSintactico*, en la cual se define la función *etiquetadogramatical* que recibe como parámetro los términos obtenido

N°	Término	Lema	Categoría Gramatical
1	AMINOACIDS_Proteins	aminoacids_proteins	NP00000
2	species	species	NN
3	bacteria	bacteria	NNS
4	being	being	NN
5	acids	acid	NNS
6	acids	acid	NNS
7	compounds	compound	NNS
8	amine	amine	NN
9	group	group	NN
10	acids	acid	NNS
11	group	group	NN
12	group	group	NN
13	carbon	carbon	NN
14	atom	atom	NN

Ilustración 14: Funcionalidad de etiquetado gramatical de cada palabra. Imagen tomada de la herramienta desarrollada (Autoría propia).

del documento y de forma iterativa se realiza la clasificación de cada palabra.

La clasificación de las palabras encontradas en el documento se hizo con ayuda de la herramienta FreeLing que asigna un código alfanumérico para cada categoría gramatical, tal como se muestra en la ilustración 14. El código de la categoría gramatical se obtiene del análisis morfológico que se hace de cada palabra para representar su información utilizando un conjunto de etiquetas propuestas por el grupo EAGLES¹¹, estas etiquetas gramaticales fueron propuestas para la anotación morfosintáctica de lexicones y corpus para todas las lenguas europeas (Padró 2012).

El código de cada categoría gramatical puede tener diferentes longitudes según la categoría. Por ejemplo, la categoría Sustantivo o Nombre se representa con siete caracteres, donde cada carácter indica un atributo de la palabra. El primer carácter indica la categoría Nombre (N), el segundo indica el tipo de la categoría Común (C) y Propio (P), el tercer carácter indica el género Masculino (M), Femenino (F) y Común (C), el cuarto indica el número Singular (S), Plural (P) e Invariable (N), el quinto y sexto carácter indican la clasificación semántica

¹¹ <http://www.ilc.cnr.it/EAGLES96/home.html>

(Persona, Lugar, Organización y Otros) y el último indica el grado Aumentativo (A) y Diminutivo (D). De esta manera, por ejemplo, la palabra “medicina” tendría la clasificación NCFS000.

Si una palabra, dependiendo del idioma, no presenta alguno de los atributos mencionados se omite el carácter correspondiente en el código de la categoría gramatical como se observa en la ilustración 14. Como siguiente paso, luego de la clasificación gramatical de cada término, se procedió a descartar todas las palabras cuya categoría gramatical sea diferente de sustantivo y adjetivo, esto con el fin de obtener la mayor información acerca del contenido del documento. En la ilustración 14 se muestra el resultado alcanzado con la lista final de los términos filtrados por categoría.

Esta lista final de términos se utiliza como parámetro de entrada para el siguiente módulo que permite realizar el análisis semántico en base a la ontología configurada, con lo cual se construyen las anotaciones considerando solo los términos del documento que tienen relación con el dominio.

5.3 Resultado Alcanzado N° 4: Módulo que permita realizar el análisis semántico.

El siguiente paso para el desarrollo de la herramienta es la identificación de los términos contenidos en el documento que pueden ser anotados según su relación con la ontología definida por el usuario. Para ello, se debe tener en cuenta que cada concepto de la ontología está relacionado a un set de predicados y objetos. En ese sentido, se recorre de forma iterativa los recursos representados en la ontología para compararlos con los términos extraídos del texto y determinar la relación existente.

La búsqueda de conceptos en la ontología se realizó haciendo uso de la herramienta Apache Jena y el motor de búsquedas del lenguaje SPARQL. Mediante el motor de consultas es posible recorrer la ontología de manera eficiente y rápida aprovechando que la ontología está estructurada como un grafo y no como un repositorio digital del conceptos (Pérez, Arenas, and Gutierrez 2006). De este modo, todas las clases de la ontología son guardadas en una lista y luego se realiza la iteración para encontrar una coincidencia con los términos contenidos en el documento.

En el caso de encontrar una coincidencia se llama a la función *generar_annotacion* que recibe como parámetro el término encontrado en el documento y la URI del concepto de la ontología. En la función *generar_annotacion* se implementa la funcionalidad para extraer los atributos del concepto de la ontología y se enlaza al término anotado.

Considerando que una palabra del documento puede carecer de lema en el idioma, como, por ejemplo, los nombres propios y acrónimos, se realizó la comparación del recurso de la ontología tanto con el término en su forma original, así como con el lema identificado por la herramienta FreeLing, esto con el fin de encontrar la mayor cantidad de conceptos posibles en la ontología.

En el caso de que un término del texto tenga coincidencia con un concepto de la ontología se guarda en una estructura de datos la información referente al recurso de la ontología para construir luego las anotaciones semánticas. Esta información es el Identificador Universal de Recurso (URI), la jerarquía dentro de la ontología y su relación con otros recursos de la misma.

Como resultado final se obtiene una lista de los términos encontrados en la ontología. Esta lista es el parámetro de entrada para el siguiente módulo de desambiguación con el cual se determinará el significado correcto de cada término según el contexto del documento.

5.4 Discusión de los resultados alcanzados N° 3 y N° 4.

La principal característica de la herramienta desarrollada es que el usuario tiene la flexibilidad para seleccionar la ontología que sea de su interés. Por esta razón, y para efectos de probar los resultados de cada objetivo se optó por obtener un conjunto de ontologías biomédicas del repositorio BioPortal¹². En este repositorio se puede encontrar ontologías tanto en idioma inglés como en español; sin embargo, las ontologías disponibles para la herramienta son del idioma inglés, debido a que presentan información más completa acerca de cada dominio pues son actualizadas periódicamente, siendo la actualización más reciente en junio del 2016.

¹² <http://biportal.bioontology.org/>

6. CAPÍTULO 6: DESAMBIGUACIÓN DE TÉRMINOS EN UN DOMINIO ESPECÍFICO.

Uno de los problemas principales que dificulta la recuperación de información es la ambigüedad de términos, pues cuando se realizan búsquedas en la Web los motores de búsqueda no se enfocan en un modelo de dominio específico sino en cualquier fuente disponible a la que se tenga acceso, lo cual puede generar que se obtenga como resultados documentos que contienen los términos de la consulta, pero en un contexto irrelevante para el usuario.

La identificación precisa del sentido de una palabra en un documento es esencial para la indexación de documentos, búsquedas en la Web e integración de datos (Hassell, Aleman-Meza, and Arpinar 2006). La desambiguación de términos es la capacidad de determinar el sentido correcto de una palabra dentro de un conjunto de varios candidatos.

En el presente capítulo se propone una estrategia de desambiguación basada en el contexto en el cual se emplea la palabra y la ontología configurada por la herramienta.

6.1 Objetivo Específico N° 3: Implementar un módulo que permita resolver la ambigüedad de los términos contenidos en los documentos en un dominio específico.

El problema de determinar el sentido de una palabra empleada en un texto se conoce como Desambiguación del Sentido de una Palabra (WSD, por sus siglas en inglés). Para intentar resolver este problema existen diferentes métodos, entre los cuales se encuentran los que se basan en el análisis de las oraciones o párrafos donde está contenido el término a desambiguar, y los basados en la contabilización de coincidencias de palabras clave del concepto referido por una palabra en el documento (Hotho, Staab, and Stumme 2003; Plaza, Stevenson, and Díaz 2010). Otros métodos de desambiguación emplean fuentes de conocimiento alojadas en la Web, como por ejemplo Wordnet, determinando el sentido de una palabra en base a la mayor cantidad de uso que tenga el concepto dentro de la fuente de conocimiento.

En este objetivo específico se propone una alternativa de solución al problema de WSD en base al contexto del documento en el cual se emplea una palabra y a la ontología que esté configurada por el usuario.

6.2 Resultado Alcanzado N° 5: Módulo de desambiguación de palabras dentro de un dominio.

La estrategia de desambiguación de términos empleada en este proyecto se realiza en función al contexto del documento considerando las palabras cercanas al término a desambiguar

dentro de una oración o párrafo. En ese aspecto, se puede determinar el sentido adecuado de un término t que está referido a un conjunto de conceptos diferentes $Ref_c = \{concepto1, concepto2, concepto3, \dots\}$ siguiendo los siguientes pasos:

1. Definir una vecindad semántica de un concepto c , la cual incluye todos los subconceptos y superconceptos de c .
2. Obtener todos los términos del documento que podrían referirse a un concepto de la vecindad semántica de c . Esto se identifica como la vecindad conceptual, $VecConc$, del término
3. Emplear la función N° 1 para desambiguar un término t en base al contexto del documento d .

$$desambiguar(d, t) = primer\{c \in Ref_c(t) \mid Max(Clasificación(d, VecConc(c)))\}$$

Ecuación 1: Función de desambiguación. Adaptada de [50]

Como se puede observar en la función 1, el sentido de la una palabra se puede desambiguar obteniendo una Clasificación de los términos que pueden ser expresados con algún concepto de la vecindad de c ($VecConc(c)$). Luego se obtiene el concepto con la mayor clasificación, el cual es asignado al término que ingresó como parámetro.

Como resultado de aplicar la función se obtiene el sentido del término que guarda mayor relación con el contexto del documento. Esto significa que los elementos del conjunto dentro de la función serán todos los conceptos que pueden ser representados por el término ingresado a la función como parámetro. Como siguiente paso de la función, se aplicó el criterio de relación que existe entre los conceptos del conjunto con el contexto del documento. Esta relación se evaluó calculando la frecuencia absoluta de la coincidencia de los conceptos pertenecientes a la vecindad del concepto evaluado. Esta vecindad es una colección de los conceptos más cercanos al concepto evaluado según la ontología configurada por la herramienta.

Con respecto a la frecuencia, no se puede evaluar de la misma manera la coincidencia de conceptos que se encuentren en niveles lejanos al concepto evaluado, debido que la ontología tiene la estructura de un grafo; en ese caso, el grado de cercanía de los conceptos se calculará tomando la distancia en aristas que exista entre ellos.

La estrategia de desambiguación se aplicó al conjunto de términos obtenidos por el procesamiento de lenguaje natural realizado al documento. El resultado es un conjunto de conceptos que se relaciona respectivamente con los términos procesados.

6.3 Discusión del resultado alcanzado N° 5.

El proceso de desambiguación de términos según el contexto del documento permite trabajar en conjunto con los resultados del procesamiento de lenguaje natural y la ontología definida por el usuario. La ontología sirve como base de conocimiento, pues permite obtener los conceptos relevantes para el documento analizado.

Los resultados alcanzados están limitados por el conjunto de términos extraídos del documento y también por la comparación sintáctica que se realiza con cada uno de los conceptos representados en la ontología. A pesar de que al tomar como base de conocimiento una ontología se obtiene un valor semántico en la identificación de conceptos, también existe un componente sintáctico limitado por la eficacia del procesamiento de lenguaje natural de la herramienta.

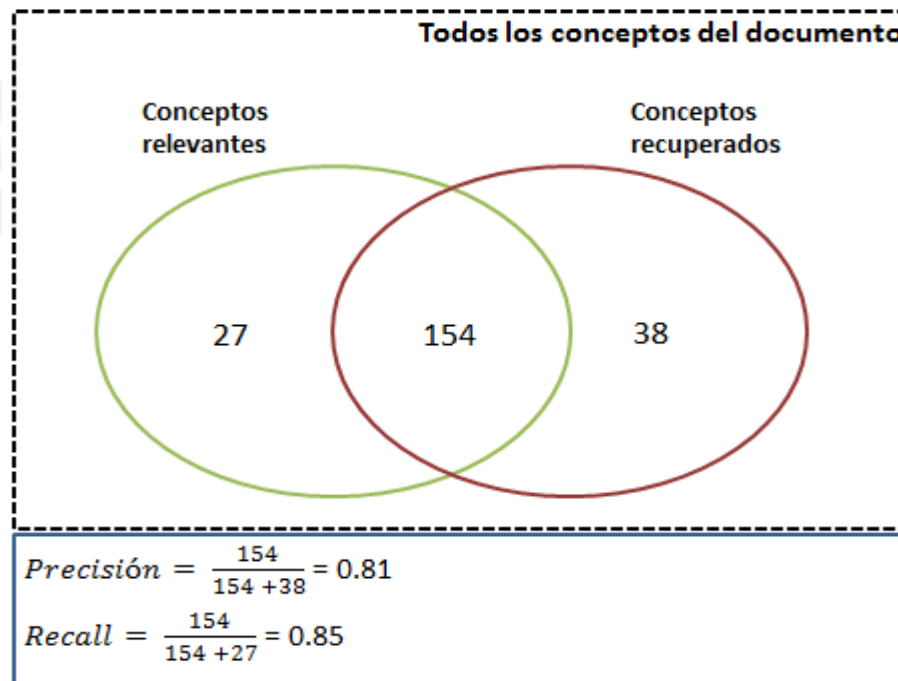


Ilustración 15: Resultados de prueba de Precisión y Recall (Autoría propia).

En ese sentido, puede ocurrir que no se encuentren conceptos de la ontología relacionados con los términos del documento en el caso de que en el documento no se trate un tema relacionado con la ontología. También puede darse el caso de que en la ontología no se tenga mapeado el término procesado, esto debido a que las ontologías son actualizadas constantemente y no necesariamente representa todos los conceptos referentes a un dominio.

Para evaluar la eficacia de la función y la merma en su rendimiento que es causada por las dificultades para extraer términos del documento se realizó pruebas con un corpus de 20 documentos cuyos conceptos más relevantes fueron mapeados con la ontología de Aminoácidos. Las pruebas realizadas fueron las de Precisión y *Recall*, las cuales son métricas estadísticas para evaluar el rendimiento de un sistema de recuperación de la información. La Precisión se encuentra dada por la relación entre el número de instancias recuperadas que son relevantes y el número de instancias recuperadas, mientras que el *Recall* está definido como el número de instancias recuperadas que son relevantes entre el número total de instancias relevantes en la colección que se esperan recuperar.

Una instancia relevante es aquella cuyo significado tiene una relación directa con el dominio de la ontología y una instancia recuperada es toda aquella que fue encontrada en el documento. En la siguiente figura se puede observar los resultados de las pruebas realizadas:

Según los resultados obtenidos, los valores de Precisión de 0.81 y *Recall* de 0.85 permiten concluir que la función de desambiguación puede perder eficacia por las comparaciones sintácticas que se realizan con los conceptos de la ontología, a pesar de contemplar una evaluación semántica de los términos del documento. Además, la falta de una función de reconocimiento de entidades disminuye eficiencia de la función de desambiguación. Sin embargo, los resultados son buenos para las especificaciones planteadas por la herramienta, cuyo valor mínimo de Precisión y *Recall* es 0.80.

7. CAPÍTULO 7: ALMACENAMIENTO DE ANOTACIONES SEMÁNTICAS EN REPOSITORIOS DIGITALES.

Uno de los problemas principales que dificulta la recuperación de información en la Web es que las búsquedas se realizan de manera sintáctica; esto es, en los resultados se busca la coincidencia exacta de las palabras ingresadas como cadena de búsqueda. De esta forma, en la búsqueda de información no se tiene en cuenta información adicional sobre el contenido de los documentos como el dominio de conocimiento y las relaciones que pueden existir entre los términos de la consulta con otros conceptos de los documentos.

Una alternativa de solución ante este problema es que los documentos sean anotados con información explícita acerca de su contenido y la relación que existe entre el documento y la ontología o dominio de conocimiento definida por el usuario. Es por ello que en este capítulo se desarrolla un módulo que permita almacenar las anotaciones semánticas generadas por la herramienta en repositorios digitales.

7.1 Objetivo Específico N° 4: Implementar un módulo que permita almacenar anotaciones semánticas en repositorios digitales.

Como objetivo final de este proyecto se busca que la herramienta desarrollada soporte la persistencia de las anotaciones semánticas generadas automáticamente luego de haber procesado el contenido del documento con relación a la ontología definida por el usuario. Esta persistencia de las anotaciones semánticas sirve de soporte a la recuperación de información puesto que permite que las búsquedas ya no se realicen solo de manera sintáctica, sino que se puede consultar las anotaciones para analizar información adicional acerca del contenido de un documento.

Por ello, se propone contar con una base de datos de todas las anotaciones semánticas correspondientes a cada documento que sea analizado por la herramienta. Las anotaciones semánticas permiten enriquecer el contenido de los documentos con información adicional en base al contexto del documento y su relación con la ontología, lo cual facilita la clasificación y recuperación del documento al momento de realizar búsquedas de documentos.

7.2 Resultado Alcanzado N° 6: Formato definido para el almacenamiento de las anotaciones semánticas en repositorios.

Como primer paso para alcanzar este objetivo, se plantea definir la estructura que tiene cada anotación semántica. Dado que las anotaciones semánticas son generadas en base al

dominio de conocimiento que es la ontología, los valores que se guardarán respecto a cada anotación son los siguientes:

- 1) Ontología, dado que la herramienta permite al usuario configurar el dominio de interés.
- 2) URL Documento, correspondiente al identificador URL del documento procesado.
- 3) Término contenido en el documento que es anotado por la herramienta y está relacionado con la ontología.
- 4) URI Concepto, correspondiente al identificador único que corresponde al concepto dentro de la ontología.
- 5) Clase Padre (SubClassOf), correspondiente a la clase padre a la que pertenece el concepto identificado dentro de la ontología. Según la ontología seleccionada un concepto puede tener una o más clases padre.
- 6) Fecha de Creación (Datetime), correspondiente a la fecha y hora en la que es generada una anotación semántica.

En la ilustración 16, se muestra un ejemplo de anotación semántica obtenida por la herramienta al procesar un documento con una ontología de Aminoácidos.

```
<rdf:RDF
  xmlns:rdf="http://www.w3.org/1999/02/22-rdf-syntax-ns#"
  xmlns:ontology="http://bioportal.bioontology.org/ontologies/AMINO-ACID" ①
  xmlns:document="http://www.ncbi.nlm.nih.gov/pmc/articles/PMC438520/pdf/jcinvest00616-0029.pdf/" ②
  xmlns:xsd="http://www.w3.org/2001/XMLSchema#" >
  <rdf:Description rdf:about="http://www.ncbi.nlm.nih.gov/pmc/articles/PMC438520/pdf/jcinvest00616-0029.pdf/SERINE">③
    <document:URI rdf:resource="http://www.co-ode.org/ontologies/amino-acid/2006/05/18/amino-acid.owl#5"/> ④
    <document:SubClassOf>TinyPolarAminoAcid</document:SubClassOf> ⑤
    ⑥<document:DateTime rdf:datatype="http://www.w3.org/2001/XMLSchema#date">2016-09-26T00:05:25.273</document:DateTime>
  </rdf:Description>
</rdf:RDF>
```

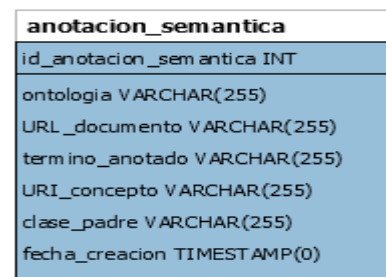
Ilustración 16: Ejemplo de anotación semántica generada por la herramienta. Imagen tomada de la herramienta desarrollada (Autoría propia).

Con el formato definido se procedió a procesar el contenido de todo el documento para generar las anotaciones semánticas, las cuales se almacenan también en un archivo de salida en formato RDF. RDF es un modelo estándar para el intercambio de información en la Web que utiliza el lenguaje XML para escribir los documentos en este formato y está diseñado para ser leído y entendido por las aplicaciones de computadoras (W3C, s.f.). De esta manera, se facilita que las anotaciones puedan ser consultadas por diversos sistemas de recuperación de información.

7.3 Resultado Alcanzado N° 7: Módulo de almacenamiento de anotaciones semánticas.

Finalmente, luego de haber generado las anotaciones semánticas, estas deben ser almacenadas en una base de datos para facilitar las tareas de organización de contenido. Para lograr este resultado se empleó un JavaBean cuyos objetos son los definidos previamente en el formato de las anotaciones semánticas. Un JavaBean permite el encapsulamiento de varios objetos, que en este caso son los atributos de cada anotación, y será guardado en una base de datos relacional MySQL por medio de una clase elaborada en Java con métodos que permiten la interacción entre la plataforma de Java y el manejador de base de datos MySQL.

La base de datos tiene una tabla que es la que permite guardar toda la información referente a cada anotación semántica. Los atributos de la tabla corresponden a los elementos de cada anotación como se puede apreciar en la ilustración 17.



anotacion_semantica	
id_anotacion_semantica	INT
ontologia	VARCHAR(255)
URL_documento	VARCHAR(255)
termino_annotado	VARCHAR(255)
URI_concepto	VARCHAR(255)
clase_padre	VARCHAR(255)
fecha_creacion	TIMESTAMP(0)

Ilustración 17: Tabla anotacion_semantica (Autoría propia).

Con la información de la ontología y el URI del término anotado es posible acceder a mayor información que esté disponible en la ontología, de acuerdo a la necesidad de información que tenga la herramienta.

7.4 Discusión de los resultados alcanzados N° 6 y N°7.

Para definir la estructura de una anotación semántica se consideró la información que es necesaria que esté enlazada al concepto a analizar. Para el caso, se consideró agregar a la URI del concepto el nombre de la ontología, además de la jerarquía del concepto dentro de la ontología, pues permite notar con mayor claridad el contexto al que pertenece el término anotado.

Para validar el resultado obtenido por el proceso de generación de anotaciones semánticas se utilizó la métrica de Precisión y Recall en base a los conceptos extraídos del

documento y los conceptos mapeados en la ontología. En la ilustración 18 se muestra los resultados obtenidos.

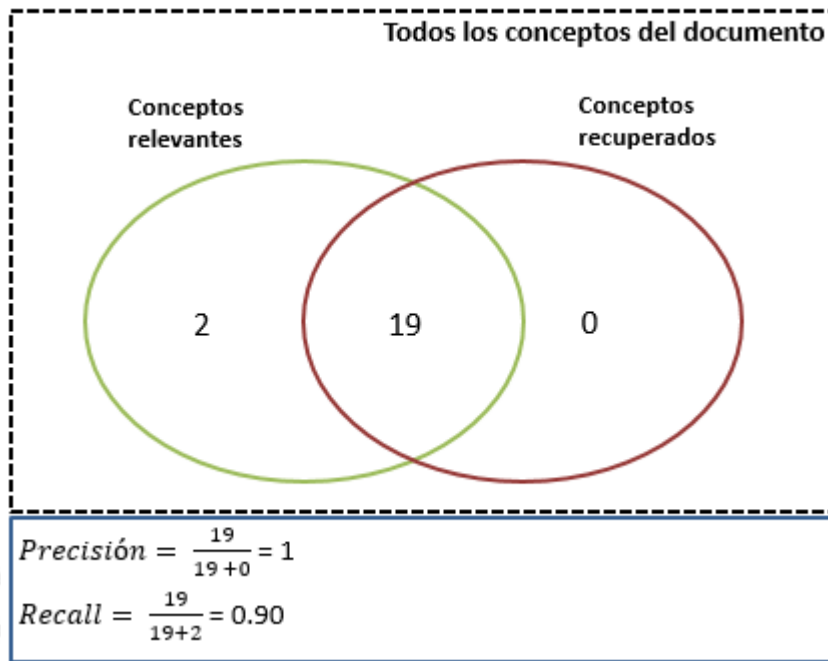


Ilustración 18: Resultados de prueba de Precisión y Recall (Autoría propia).

Según los resultados obtenidos, los valores de Precisión de 1 y *Recall* de 0.90 permiten concluir que la función que permite generar las anotaciones semánticas puede perder eficacia por las comparaciones sintácticas que se realizan con los conceptos de la ontología, a pesar de contemplar una evaluación semántica de los términos del documento. Esto se debe a que se encontró en el documento términos que están mapeados en la ontología, pero que en el documento se encontraban concatenados a símbolos u otros términos mediante guiones. Sin embargo, los resultados son buenos para las especificaciones planteadas por la herramienta, cuyo valor mínimo de Precisión y *Recall* es 0.80.

Finalmente, sobre la tabla de base de datos donde se almacenan las anotaciones, se toma como llaves primarias la ontología, el término anotado y la clase padre, puesto que un concepto dentro de la ontología puede tener más de una clase padre, según el dominio que esté configurado.

8. CONCLUSIONES

El objetivo general del presente proyecto de fin de carrera es poder realizar anotaciones semánticas de manera automática sobre un documento en formato PDF en base a la relación que su contenido tiene con la ontología configurada por la herramienta. Para lograr este objetivo se desarrolló un conjunto de módulos que permiten llevar a cabo las principales funcionalidades de la herramienta: procesamiento de texto del documento PDF, análisis sintáctico del contenido, mapeamiento semántico con la ontología seleccionada, desambiguación de términos y la persistencia de anotaciones.

Con respecto al procesamiento del contenido textual del documento, una de las principales dificultades para aplicar las técnicas de NLP (Procesamiento de Lenguaje Natural) fue que los documentos pueden presentar diversos tipos de contenido como, por ejemplo, imágenes, tablas, fórmulas, datos numéricos, acrónimos, nombres propios, entre otros. Esto dificulta la extracción y correcta identificación de términos del documento, lo cual repercute en la eficiencia del análisis realizado por la herramienta. Es por ello que se optó por realizar el análisis del contenido no solo con los términos encontrados en su forma original, sino también con su forma base o lema, lo cual supuso una mejora en los resultados obtenidos.

Otra dificultad en el proceso de extracción de términos fue que en algunos documentos del corpus se pueden encontrar palabras separadas por guiones cuando se encuentran al final de una línea, o concatenación de dos palabras, entre otras formas de escritura. Para ello se realizó un pre-procesamiento del contenido del documento antes de aplicar las técnicas de NLP, con el objetivo de mejorar la identificación de conceptos de la ontología.

En cuanto al análisis sintáctico del contenido, en un documento se puede encontrar una gran cantidad de términos, sin embargo, todos estos no brindan información relevante acerca del tema que se trata en el documento. Por ello, fue necesario clasificar cada término según su categoría gramatical para poder tomar en cuenta solo las categorías que brindan información sobre el contenido como son los sustantivos y adjetivos. Una dificultad que se presenta al identificar la categoría gramatical es el idioma en el cual está elaborado un documento, para ello se empleó una funcionalidad de la librería FreeLing que permite identificar el lenguaje automáticamente y así permitir que la herramienta sea flexible tanto para el idioma español e inglés.

Sobre el análisis semántico de los términos extraídos del documento, este es realizado en base a la ontología seleccionada por el usuario y su eficiencia está ligada a la cantidad de clases mapeadas en la ontología, ya que la ontología no fue desarrollada como parte del proyecto. Las

ontologías disponibles para el proyecto son extraídas del repositorio BioPortal, en el cual se almacenan ontologías del ámbito de la biomedicina tanto en idioma inglés como español. Para probar los resultados obtenidos por la herramienta se utilizó un corpus de documentos PDF en el idioma inglés relacionados a la ontología de Aminoácidos (*Amino Acid Ontology*¹³), debido a que las ontologías en inglés tienen mayor soporte por parte de la comunidad de BioPortal.

Respecto a la desambiguación de términos, la técnica empleada permite discriminar el sentido de una palabra en base al contexto en el que se encuentra, verificando si los términos cercanos también tienen relación con la ontología y así determinar el sentido correcto. Los resultados obtenidos por las pruebas de Precisión y Recall demuestran que la técnica de desambiguación es eficiente para el análisis realizado por la herramienta, para ello es necesario contar con una ontología debidamente estructurada y con resultados óptimos obtenidos por el procesamiento de NLP.

Por último, las anotaciones semánticas son generadas en formato RDF/XML y se almacenan de manera externa al documento PDF en un archivo de texto con extensión .rdf y, además, los datos de la anotación se guardan en una base de datos relacional con el objetivo de facilitar la búsqueda de información acerca del documento, pues se almacena la relación de la URL del documento, en caso sea un documento Web, con las anotaciones generadas. En caso que el documento analizado sea un archivo local se almacena de manera referencial la ruta y nombre del archivo.

La estructura de las anotaciones está diseñada para guardar información del documento analizado y del término anotado. Acerca del término anotado se optó por considerar la información del URI y la jerarquía del concepto dentro de la ontología. Esta estructura de anotación se puede rediseñar según la necesidad de información que se desee guardar.

¹³ <http://biportal.bioontology.org/ontologies/AMINO-ACID>

9. LIMITACIONES

En la aplicación de técnica de NLP se tiene la limitación del reconocimiento de nombres de entidades que se expresan con más de una palabra, así como del reconocimiento de palabras técnicas de determinado dominio. Para el caso de nombre de entidades compuestos por más de una palabra se opta por identificarlas porque comienzan con mayúsculas y se concatenan mediante subguiones, aunque no siempre se obtiene el nombre correcto.

Por otro lado, una de las principales limitaciones presentadas a lo largo del desarrollo del proyecto es la cantidad de repositorios activos para obtener ontologías alojadas en la Web, puesto que la herramienta debe dar al usuario la flexibilidad de configurar diversos temas de interés. En ese sentido, la herramienta está limitada al tiempo de respuesta y estado de actividad de los repositorios para realizar el mapeamiento semántico.

Otra limitación está ligada al desarrollo de la ontología seleccionada, debido a que son ontologías desarrolladas por terceros y puede haber ontologías muy completas con una gran cantidad de clases acerca de determinado dominio, pero, también hay ontologías que no tienen mucho soporte y sus clases abarcan pocos conceptos acerca del dominio.

Otra limitación importante es el tiempo de respuesta para las consultas SPARQL que se realizan sobre las ontologías para obtener las clases y propiedades. Dependiendo de la ontología configurada por el usuario, la consulta puede tomar mucho tiempo debido a la gran cantidad de clases mapeadas y la consulta puede exceder el tiempo de respuesta por lo que el servicio es interrumpido.

10. TRABAJOS FUTUROS

Respecto al procesamiento de NLP, la deficiencia en la extracción de términos y nombres en general se puede mitigar empleando herramientas de reconocimiento nombres en textos en español como, por ejemplo, *Smorph* y *Nooj* que son sistemas que detectan automáticamente expresiones y estructuras propias del idioma español (Tramallino 2014). Otra opción sería la realización de una herramienta que reconozca nombres de entidades y palabras técnicas de acuerdo a los temas que contengan los documentos a utilizar.

Respecto a las ontologías disponibles para el uso de la herramienta, un aspecto que se puede abordar es el desarrollo de un conjunto de ontologías especializadas en determinados temas para garantizar la eficiencia de las anotaciones semánticas generadas por la herramienta.

Respecto a la desambiguación de términos, se puede realizar como trabajo futuro el empleo de más de una técnica o algoritmo para mejorar la eficiencia en la identificación del sentido de una palabra.

Por último, respecto a las anotaciones semánticas, como trabajo futuro se puede realizar inferencias sobre los términos anotados y las propiedades y relaciones que tengan con otros conceptos de la ontología para enriquecer el contenido del documento analizado.

11. BIBLIOGRAFÍA

- AMecological. (s.f.). *BIOHEALTH BS WSG*. Recuperado el 08 de Mayo de 2015, de www.amecological.cl:
http://www.amecological.cl/archivos/biohealth/etiqueta_biohealth_bs.pdf
- Annie, G. (s.f.). *Gate's Annie System*. Obtenido de <https://gate.ac.uk/ie/annie.html>
- Apache Jena. (s.f.). *Apache Jena*. Recuperado el 05 de Junio de 2015, de <https://jena.apache.org/>
- Berners-Lee, T. (2006). *Documento en Línea*. Recuperado el 15 de Abril de 2015, de <http://www.w3.org/DesignIssues/LinkedData.html>
- Dublin Core. (s.f.). *Expressing Dublin Core metadata using the Resource Description Framework (RDF)*. Recuperado el 05 de Junio de 2015, de <http://dublincore.org/documents/dc-rdf/>
- GATE. (s.f.). *General Architecture for Text Engineering*. Recuperado el 05 de Junio de 2015, de <https://gate.ac.uk/>
- Internet Live Stats*. (2015). Recuperado el 19 de Abril de 2015, de <http://www.internetlivestats.com/>
- MySQL. (2015). *MySQL*. Recuperado el 05 de Junio de 2015, de <http://www.mysql.com/>
- NetBeans. (2015). *NetBeans IDE*. Recuperado el 05 de Junio de 2015, de <https://netbeans.org/>
- NetCraft*. (2015). Recuperado el 19 de Abril de 2015, de <http://www.netcraft.com/>
- Oracle. (2015). *Oracle Technology Network*. Recuperado el 05 de Junio de 2015, de Java: <http://www.oracle.com/technetwork/java/index.html>
- Stats, I. L. (2015). *Internet Live Stats*. Recuperado el 21 de Abril de 2015, de <http://www.internetlivestats.com/>
- The Apache Software Foundation. (s.f.). Recuperado el 05 de Junio de 2015, de Apache Tika: <https://tika.apache.org/>
- Vallez, M., Rovira, C., Codina, L., & Pedraza, R. (2010). *Procedimientos para la extracción de palabras clave de páginas web basados en criterios de posicionamiento en buscadores*. Recuperado el 17 de Abril de 2015, de Hipertext.net, 8,: http://www.upf.edu/hipertextnet/numero-8/extraccion_keywords.html
- W3C. (s.f.). Recuperado el 15 de Abril de 2015, de Documento en línea: <http://www.w3.org/standards/semanticweb/ontology>
- W3C. (15 de Enero de 2008). *SPARQL Query Language for RDF*. Recuperado el 05 de Junio de 2015, de <http://www.w3.org/TR/rdf-sparql-query/>
- W3C. (s.f.). *Metadata*. Recuperado el 07 de Mayo de 2015, de Metadata: <http://www.w3.org/Metadata>

- W3C. (s.f.). *RDF*. Recuperado el aBRIL de 21 de 2015, de RESOURCE DESCRIPTION SOFTWARE: <http://www.w3.org/RDF/>
- W3C. (s.f.). *RDF*. Recuperado el 05 de Junio de 2015, de <http://www.w3.org/RDF/>
- W3C. (s.f.). *Web Ontology Language*. Recuperado el 06 de Junio de 2015, de <http://www.w3.org/2001/sw/wiki/OWL>
- Agosti, Maristella, and Nicola Ferro. 2007. "A Formal Model of Annotations of Digital Content." *ACM Transactions on Information Systems* 26(1): 3–es. <http://www.scopus.com/inward/record.url?eid=2-s2.0-37049000565&partnerID=tZOtx3y1> (February 27, 2015).
- Al-Bukhitan, Saeed, Tarek Helmy, and Mohammed Al-Mulhem. 2014. "Semantic Annotation Tool for Annotating Arabic Web Documents." *Procedia Computer Science* 32: 429–36. <http://www.scopus.com/inward/record.url?eid=2-s2.0-84902661096&partnerID=tZOtx3y1> (March 27, 2015).
- Attwood, Teresa K et al. 2010. "Utopia Documents: Linking Scholarly Literature with Research Data." *Bioinformatics* 26(18): i568–i574.
- Bechhofer, Sean. 2009. "OWL: Web Ontology Language." In *Encyclopedia of Database Systems*, Springer. incollection, 2008–9.
- Berners-Lee, Tim, James Hendler, and Ora Lassila. 2001. "The Semantic Web." *Scientific American* 284(5): 34–43. <http://www.scopus.com/inward/record.url?eid=2-s2.0-79551602978&partnerID=tZOtx3y1> (January 9, 2015).
- Caporuscio, Mauro, and Carlo Ghezzi. 2015. "Engineering Future Internet Applications: The Prime Approach." *Journal of Systems and Software* (0). <http://www.sciencedirect.com/science/article/pii/S0164121215000783>.
- Castells, Pablo. 2003. "La Web Semántica." *Sistemas interactivos y colaborativos en la web*: 195–212.
- Ciccarese, Paolo, Marco Ocana, and Tim Clark. 2012. "Open Semantic Annotation of Scientific Publications Using DOME0." *J. Biomedical Semantics* 3(S-1): S1.
- Corcho, Oscar. 2006. "Ontology Based Document Annotation: Trends and Open Research Problems." *International Journal of Metadata, Semantics and Ontologies* 1(1): 47. <http://www.scopus.com/inward/record.url?eid=2-s2.0-33750709439&partnerID=tZOtx3y1> (April 16, 2015).
- Costa, V G, A M Printista, and M Marin. 2006. "Improving Web Searches with Distributed

- Buckets Structures.” In *Web Congress, 2006. LA-Web '06. Fourth Latin American*, . inproceedings, 119–26.
- Davies, John, Dieter Fensel, and Frank Van Harmelen. 2003. “Towards the Semantic Web.” *Ontology-driven Knowledge Management. Wiley, First Edition January 21*.
- Duval, Erik, Wayne Hodgins, Stuart Sutton, and Stuart L Weibel. 2002. “Metadata Principles and Practicalities.” *D-lib Magazine* 8(4): 16.
- Euzenat, Jérôme. 2007. “Semantic Precision and Recall for Ontology Alignment Evaluation.” In *IJCAI*, . inproceedings, 348–53.
- Fafalios, Pavlos, and Panagiotis Papadakos. 2014. “Theophrastus: On Demand and Real-Time Automatic Annotation and Exploration of (Web) Documents Using Open Linked Data.” *Web Semantics: Science, Services and Agents on the World Wide Web* 29: 31–38. <http://www.scopus.com/inward/record.url?eid=2-s2.0-84906071639&partnerID=tZOtx3y1> (December 17, 2014).
- Fung, Chun Che, and Wigrai Thanadechteemapat. 2010. “Discover Information and Knowledge from Websites Using an Integrated Summarization and Visualization Framework.” In *Knowledge Discovery and Data Mining, 2010. WKDD'10. Third International Conference on*, . inproceedings, 232–35.
- García, Norberto Fernández, and Luis Sánchez Fernández. 2005. “La Web Semántica: Fundamentos Y Breve estado Del Arte.” *Novática: Revista de la Asociación de Técnicos de Informática* (178): 6–11.
- Gruber, Thomas R. 1995a. “Toward Principles for the Design of Ontologies Used for Knowledge Sharing?” *International Journal of Human-Computer Studies* 43(5–6): 907–28. <http://www.sciencedirect.com/science/article/pii/S1071581985710816> (February 8, 2015).
- . 1995b. “Toward Principles for the Design of Ontologies Used for Knowledge Sharing?” *International Journal of Human-Computer Studies* 43(5–6): 907–28. <http://www.scopus.com/inward/record.url?eid=2-s2.0-58149365542&partnerID=tZOtx3y1> (February 8, 2015).
- Handsuh, Siegfried, and Steffen Staab. 2003. *96 Annotation for the Semantic Web*. IOS Press. book.
- Hassell, Joseph, Boanerges Aleman-Meza, and I Budak Arpinar. 2006. “Ontology-Driven Automatic Entity Disambiguation in Unstructured Text.” In *International Semantic Web*

- Conference*, . inproceedings, 44–57.
- Hotho, Andreas, Steffen Staab, and Gerd Stumme. 2003. “Ontologies Improve Text Document Clustering.” In *Data Mining, 2003. ICDM 2003. Third IEEE International Conference on*, . inproceedings, 541–44.
- Josephine, J Anitha, and S Sathiyadevi. 2011. “Ontology Based Relevance Criteria for Semantic Web Search Engine.” In *Electronics Computer Technology (ICECT), 2011 3rd International Conference on*, . inproceedings, 60–64.
- Kim, Younghwan, Sujin Yoo, and Seongbin Park. 2012. “A Semantic Web Browser for Novice Users.” In *2012 Sixth International Conference on Complex, Intelligent, and Software Intensive Systems*, IEEE, 806–9. <http://www.scopus.com/inward/record.url?eid=2-s2.0-84866618639&partnerID=tZOtx3y1> (April 16, 2015).
- Kiryakov, Atanas et al. 2004. “Semantic Annotation, Indexing, and Retrieval.” *Web Semantics: Science, Services and Agents on the World Wide Web* 2(1): 49–79. <http://www.scopus.com/inward/record.url?eid=2-s2.0-9944223230&partnerID=tZOtx3y1> (December 18, 2014).
- Kitchenham, Barbara. 2004. “Procedures for Performing Systematic Reviews.” *Keele, UK, Keele University* 33(2004): 1–26.
- Kobayashi, Mei, and Koichi Takeda. 2000. “Information Retrieval on the Web.” *ACM Comput. Surv.* 32(2): 144–73. <http://doi.acm.org/10.1145/358923.358934>.
- Li, Lizhen, Kemin Xie, and Zhifeng Dong. 2009. “Towards the Ontology Organization for Semantic Searching.” In *2009 Sixth International Conference on Fuzzy Systems and Knowledge Discovery*, IEEE, 603–6. <http://www.scopus.com/inward/record.url?eid=2-s2.0-76549136258&partnerID=tZOtx3y1> (April 16, 2015).
- Liddy, Elizabeth D et al. 2003. “Natural Language Processing.” *Encyclopedia of library and information science* 2.
- De Maio, Carmen et al. 2014. “Formal and Relational Concept Analysis for Fuzzy-Based Automatic Semantic Annotation.” *Applied intelligence* 40(1): 154–77.
- Navarro-Galindo, José L, and José Samos. 2010. “Manual and Automatic Semantic Annotation of Web Documents: The FLERSA Tool.” In *Proceedings of the 12th International Conference on Information Integration and Web-Based Applications & Services, iiWAS '10*, New York, NY, USA: ACM. inproceedings, 542–49.

<http://doi.acm.org/10.1145/1967486.1967570>.

- Niranatlamphong, Winyu. 2009. "A Conceptual Framework for Digital Annotation System on WWW." In *Computer Science and Information Technology, International Conference on*, eds. Worasit Choochaiwattana and Michael B Spring. . CONF, 27–31. <http://doi.ieeecomputersociety.org/10.1109/ICCSIT.2009.5234462>.
- Padró, Lluís. 2012. "Analizadores Multilingües En Freeling." *Linguamática* 3(2): 13–20.
- Padró, Lluís, and Evgeny Stanilovsky. 2012. "Freeling 3.0: Towards Wider Multilinguality."
- Pafilis, Evangelos et al. 2009. "Reflect: Augmented Browsing for the Life Scientist." *Nat Biotech* 27(6): 508–10. <http://dx.doi.org/10.1038/nbt0609-508>.
- Pavlopoulos, Georgios A et al. 2009. "OnTheFly: A Tool for Automated Document-Based Text Annotation, Data Linking and Network Generation." *Bioinformatics* 25(7): 977–78. <http://bioinformatics.oxfordjournals.org/content/25/7/977.abstract>.
- Pérez, Jorge, Marcelo Arenas, and Claudio Gutierrez. 2006. "Semantics and Complexity of SPARQL." In *The Semantic Web-ISWC 2006*, Springer. incollection, 30–43.
- Plaza, Laura, Mark Stevenson, and Alberto Díaz. 2010. "Improving Summarization of Biomedical Documents Using Word Sense Disambiguation." In *Proceedings of the 2010 Workshop on Biomedical Natural Language Processing*, . inproceedings, 55–63.
- Rajput, Quratulain. 2014. "Ontology Based Semantic Annotation of Urdu Language Web Documents." *Procedia Computer Science* 35(C): 662–70. <http://www.scopus.com/inward/record.url?eid=2-s2.0-84924203419&partnerID=tZOtx3y1> (March 27, 2015).
- Riaz, Mehwish, Emilia Mendes, and Ewan Tempero. "Protocol for Systematic Literature Review of Maintainability Prediction of Software Applications."
- Sánchez, David, David Isern, and Miquel Millan. 2011. "Content Annotation for the Semantic Web: An Automatic Web-Based Approach." *Knowledge and Information Systems* 27(3): 393–418.
- Sánchez, Juan Antonio Pastor. 2011. *Tecnologías de La Web Semántica*. Editorial UOC. book.
- Sasieta, Héctor Andrés Melgar, Fabiano Duarte Beppler, and Roberto Carlos do Santos Pacheco. 2012. "Um Modelo Para a Visualização Do Conhecimento Baseado Em Arquétipos Visuais-Doi: 10.4025/actascitechnol. v34i4. 10435." *Acta Scientiarum. Technology* 34(4): 381–89.

- Schreiber, Guus. 2000. *Knowledge Engineering and Management: The CommonKADS Methodology*. MIT press. book.
- Srivastava, Divesh, and Yannis Velegrakis. 2007. "Intensional Associations between Data and Metadata." In *Proceedings of the 2007 ACM SIGMOD International Conference on Management of Data - SIGMOD '07*, New York, New York, USA: ACM Press, 401. <http://www.scopus.com/inward/record.url?eid=2-s2.0-35148840335&partnerID=tZOtx3y1> (April 16, 2015).
- Studer, Rudi, V. Richard Benjamins, and Dieter Fensel. 1998. "Knowledge Engineering: Principles and Methods." *Data and Knowledge Engineering* 25(1-2): 161-97. <http://www.scopus.com/inward/record.url?eid=2-s2.0-0032026995&partnerID=tZOtx3y1>.
- Tello, Adolfo Lozano. 2001. "Ontologías En La Web Semántica." *Jornadas de Ingeniería Web* 1.
- Tramallino, Carolina Paola. 2014. "ANÁLISIS MORFOLÓGICO CON HERRAMIENTAS INFORMÁTICAS. RECONOCIMIENTO DE NOMBRES EN TEXTOS DE ESPAÑOL CON EL SISTEMA NOOJ." *Lingüística y Literatura* (63): 33-48.
- Ur Rehman, Mohib, Mohammad Haseeb Anwer, and Nadeem Iftikhar. 2005. "Universal Metadata Definition." In *Proceedings - WEC'05: 3rd World Enformatika Conference*, , 144-46. <http://www.scopus.com/inward/record.url?eid=2-s2.0-30844451582&partnerID=tZOtx3y1>.
- UREN, V et al. 2006. "Semantic Annotation for Knowledge Management: Requirements and a Survey of the State of the Art." *Web Semantics: Science, Services and Agents on the World Wide Web* 4(1): 14-28. <http://www.scopus.com/inward/record.url?eid=2-s2.0-30544450075&partnerID=tZOtx3y1> (November 21, 2014).
- Villazón, Boris, and Daniel Villa. 2014. "Tecnologías Semánticas: Datos Enlazados En Bibliotecas Y Repositorios Digitales."
- Ziamba, Ewa, and Rafal Zelazny. 2013. "Measuring the Information Society in Poland- Dilemmas and a Quantified Image." In *Computer Science and Information Systems (FedCSIS), 2013 Federated Conference on*, . inproceedings, 1185-92.