PONTIFICIA UNIVERSIDAD CATÓLICA DEL PERÚ

FACULTAD DE CIENCIAS E INGENIERÍA

DISEÑO DE UN MODELO PARA LA RECUPERACIÓN DE DOCUMENTOS BASADO EN ONTOLOGÍAS EN EL DOMINIO DE LA INGENIERÍA INFORMÁTICA

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

Héctor Erasmo Gómez Montoya

ASESOR: Héctor Andrés Melgar Sasieta

Lima, junio del 2014

RESUMEN

La selección de información relevante de documentos digitales es uno de los principales problemas para los estudiantes de pregrado de la especialidad de Ingeniería Informática. Para facilitar dicha tarea, es necesario un modelo que represente la relación entre las entidades en las que se define toda la información disponible. Por ello, se decidió llevar a cabo una revisión sistemática acerca de las posibles soluciones que representen dicho domino.

Como resultado de la revisión realizada, se propone el uso de ontologías como estructura básica para la representación del conocimiento por su eficacia a la hora de realizar la recuperación. Además, se plantea utilizar un proceso de etiquetación semántica de documentos para relacionar cada documento digital con - al menos - una entidad de la ontología con la finalidad de poder realizar búsquedas mediante el uso de etiquetas y lenguaje natural.

Se concluye que las ontologías son una estructura flexible y que soportan la recuperación de conocimiento en un dominio específico y que el modelo planteado cumple con las necesidades de búsqueda y etiquetación para los usuarios.

FACULTAD DE CIENCIAS E INGENIERÍA
ESPECIALIDAD DE INGENIERÍA INFORMÁTICA

PONTIFICIA UNIVERSIDAD CATÓLICA DEL PERÚ

TEMA DE TESIS PARA OPTAR EL TÍTULO DE INGENIERO INFORMÁTICO

TÍTULO: Diseño de un modelo para la recuperación de documentos basado en Ontologías en el dominio de la Ingeniería Informática.

ÁREA: Ciencias de la Computación
PROPONENTE: Héctor Andrés Melgar Sasieta
ASESOR: Héctor Andrés Melgar Sasieta
ALUMNO: Héctor Erasmo Gómez Montoya
CÓDIGO: 20082060
TEMA N°: S 37
FECHA: 09 de septiembre de 2014

PONTIFICIA UNIVERSIDAD CATÓLICA DEL PERÚ
Facultad de Ciencias e Ingeniería
07 OCT. 2014
MIGUEL MEJIA PUENTE
DECANO

DESCRIPCIÓN

La selección de información relevante de documentos digitales es uno de los principales problemas para los estudiantes de pregrado de la especialidad de Ingeniería Informática. Para facilitar dicha tarea, es necesario un modelo que represente la relación entre las entidades en las que se define toda la información disponible. Por ello, se decidió llevar a cabo una revisión sistemática acerca de las posibles soluciones que representen dicho domino. Como resultado de la revisión realizada, se propone el uso de ontologías como estructura básica para la representación del conocimiento por su eficacia a la hora de realizar la recuperación. Además, se plantea utilizar un proceso de etiquetación semántica de documentos para relacionar cada documento digital con - al menos - una entidad de la ontología con la finalidad de poder realizar búsquedas mediante el uso de etiquetas y lenguaje natural. Se concluye que las ontologías son una estructura flexible y que soportan la recuperación de conocimiento en un dominio específico y que el modelo planteado cumple con las necesidades de búsqueda y etiquetación para los usuarios.

OBJETIVO GENERAL

Diseñar un modelo basado en ontologías en el dominio de la Ingeniería Informática, Que facilite la recuperación de documentos almacenados en repositorios de contenido digital.

OBJETIVOS ESPECÍFICOS

O1: Analizar el escenario para establecer la cobertura, vacíos existentes y las tecnologías que se emplean para la representación del conocimiento.

O2: Proponer una estructura capaz de soportar la recuperación de información así como los métodos y técnicas que nos facilitaran la extracción.

Av. Universitaria 1801
San Miguel, Lima – Perú
Apartado Postal 1761
Lima 100 – Perú
Teléfono:
(511) 626 2000 Anexo 4801

O3: Establecer un proceso para la etiquetación semántica de documentos digitales.

O4: Diseñar un prototipo que valide la viabilidad del modelo planteado.

ALCANCE

El presente proyecto se enfoca en proponer un modelo para la recuperación de documentos a partir de repositorios digitales utilizando herramientas de la Ingeniería del Conocimiento como son las ontologías, anotaciones semánticas y recuperación de información. Para esto se analizará de manera sistemática los problemas, oportunidades, contexto y posibles soluciones a la problemática presentada. Vale recalcar que se utilizará un proceso de anotación semántica manual en 36 documentos alojados en un banco de documentos digitales con la información de prácticas y exámenes pasados de los cursos de Ciencias de la Computación de la carrera de Ingeniería Informática de la PUCP. Para, así, obtener data de prueba que permita validar el modelo planteado.

Máximo: 100 páginas

Av. Universitaria 1801
San Miguel, Lima – Perú

Apartado Postal 1761
Lima 100 – Perú

Teléfono:
(511) 626 2000 Anexo 4801

A mi hermano Julio por preocuparse siempre por mí y ser una fuente de inspiración, a mi hermana Lida porque de donde esté siempre me está protegiendo, a mi tía Luz por demostrarme que el mejor regalo que se le puede dar a una persona es la educación y a mis dos ángeles Fidencio y Lida, que sin tener una gota de mi sangre me quisieron y cuidaron como a uno de los suyos, gracias por motivarme a cumplir siempre mis metas.

AGRADECIMIENTOS

Al Dr. Andrés Melgar, por compartir el conocimiento obtenido mediante su trayectoria profesional, generando nuevas oportunidades de investigación y especialización a futuros ingenieros.

A Bonnie Carranza y Rodrigo Espinoza, un excelente grupo de investigación con quienes tuve el honor de trabajar

ÍNDICE GENERAL

Índice de Figuras

Índice de Tablas

CAPÍTULO I: PLANTEAMIENTO

1.1 Problemática

La información es de gran importancia para la sociedad debido a los beneficios que esta ofrece. Incluso se dice que es fuente de riqueza de la misma (Lytras & Sicilia, 2005) , ya que permite mejorar el aprendizaje y como consecuencia, genera mejores perfiles profesionales. Actualmente la información tienen como objetivo brindar acceso al conocimiento y aprendizaje a los individuos, grupos u organizaciones dentro de la sociedad (Stewart, 2001).

Actualmente, la información digital ha aumentado drásticamente. Para darnos cuenta de la magnitud de este crecimiento, consideremos que en el año 1994 solo existían 100,000 páginas que se podían acceder por web. Por el año de 1998, esta suma se incrementó a 30 millones de páginas accesibles vía web con un total de 225,000 web servers.(Haverkamp & Gauch, 1998).

Para la época del desarrollo de la tesis, estos valores se han incrementado considerablemente, tal es así que existen 155 millones de websites corriendo es aproximadamente 75 millones de servers y siguen incrementándose a razón de 50% por año. (International Data Corporation, 2014)

El principal medio en que los usuarios accedan a toda esta gran cantidad de información es mediante los meta-buscadores. La probabilidad de que los usuarios lleguen a comprender correctamente la información brindada para su consulta dependerá de dos factores muy importantes. El primero es qué tan confiable sea la fuente de información, lo que involucra su orden y estructura. El segundo es el tipo y forma de consulta con la que los usuarios solicitan información. (Gilchrist, 2003)

La fuente de información se ve representada de diferentes formas, desde una libreta de anotaciones hasta la biblioteca virtual de Google Scholar[1]. Pero, ¿qué pasaría si estas fuentes no tuvieran un orden que facilite las búsquedas? Entonces los usuarios tendrían un proceso extra que hacer antes de pensar en solicitar información, lo que generaría que estos no tengan la motivación para seguir investigando. Lamentablemente esto pasa muy a menudo debido a la aglomeración de conocimiento desordenado que no satisface las necesidades del usuario (Hou & Pai, 2009), esto

[1] http://scholar.google.com/

trae como consecuencia que los usuarios utilicen información totalmente diferente a lo que ellos buscaban, burlando de nuevo la validez de la información.

La comunicación del usuario con la fuente de conocimiento es mayormente por medio de consultas, estas son realizadas con la finalidad de conseguir y aprovechar al máximo la fuente, para esto se utiliza el proceso de recuperación del conocimiento. El gestor de información debe ser capaz de entender la consulta y brindar la información más adecuada que sirva para responder la consulta en mención (Gilchrist, 2003). Estas afirmaciones nos dejan con una pregunta ¿cómo se hace para que el computador entienda lo que el usuario le solicita?

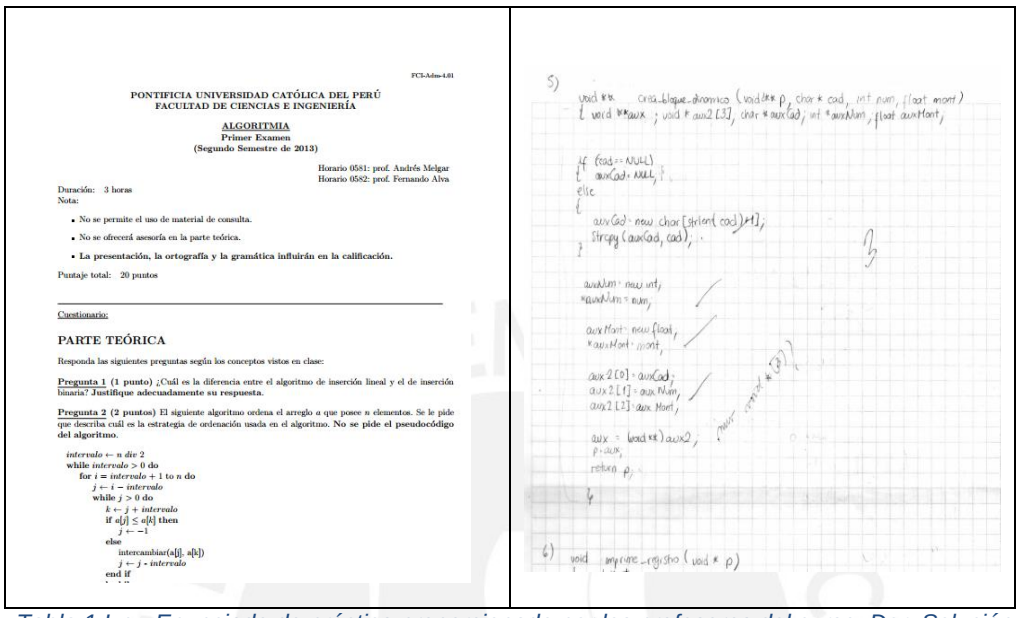
Una tentativa de respuesta a la pregunta planteada son las ontologías, que nos ayudan a estructurar el conocimiento en un dominio en específico y construir relaciones, reglas y formalismos capaces de ser reconocidos por ambas partes (Melgar & Pacheco, 2010).

En este contexto, en la especialidad de Ingeniería Informática de la PUCP existe la necesidad en los estudiantes de buscar documentos que contengan información sobre los cursos de la carrera, ya sean estos exámenes, prácticas o incluso información de tesis pasadas, con la finalidad de estudiar y prepararse para rendir evaluaciones programadas en el plan curricular de la carrera.

Como primer problema que se enfrenta en esta situación es que el ámbito o dominio del que se habla es bastante amplio, pues la carrera de ingeniería informática se divide en 5 áreas principales cada una de ellas con mención en las sumillas de cursos de la carrera[1]. Además este ámbito se incrementa a gran cantidad debido a la sinergia entre especialidades y así abarcar ramas con toda clase de información específica, como por ejemplo: bioinformática, electrónica computacional, marketing digital, etc.

Actualmente, se cuenta con repositorios digitales de documentos, que almacenan imágenes escaneadas o documentos en formato PDF, a la que los alumnos y profesores tienen acceso. Estos documentos contienen información acerca de evaluaciones anteriores de los cursos dictados en la especialidad. Ya sean enunciados proporcionados por los profesores o solucionarios compartidos por los alumnos. Debido a que no se tiene un estándar al subir estos documentos, en este repositorio se pueden encontrar ya sean fotos de ejercicios solucionados o pdfs con teoría obtenidos de la web. En la Tabla 1 se observa los tipos de documentos que se pueden

encontrar en el repositorio, cabe destacar que el documento de la izquierda se encuentra en el repositorio en formato PDF y el de la derecha en formato JPG ya que es una imagen escaneada.



PONTIFICIA UNIVERSIDAD CATÓLICA DEL PERÚ
FACULTAD DE CIENCIAS E INGENIERÍA

ALGORITMIA
Primer Examen
(Segundo Semestre de 2013)

Horario 0581: prof. Andrés Melgar
Horario 0582: prof. Fernando Alva

Duración: 3 horas
Nota:

- No se permite el uso de material de consulta.
- No se ofrecerá asesoría en la parte teórica.
- La presentación, la ortografía y la gramática influirán en la calificación.

Puntaje total: 20 puntos

Cuestionario:

PARTE TEÓRICA

Responda las siguientes preguntas según los conceptos vistos en clase:

Pregunta 1 (1 punto) ¿Cuál es la diferencia entre el algoritmo de inserción lineal y el de inserción binaria? Justifique adecuadamente su respuesta.

Pregunta 2 (2 puntos) El siguiente algoritmo ordena el arreglo *a* que posee *n* elementos. Se le pide que describa cuál es la estrategia de ordenación usada en el algoritmo. **No se pide el pseudocódigo del algoritmo.**

```
intervalo ← n div 2
while intervalo > 0 do
    for i = intervalo + 1 to n do
        j ← i − intervalo
        while j > 0 do
            k ← j + intervalo
            if a[j] ≤ a[k] then
                j ← −1
            else
                intercambiar(a[j], a[k])
                j ← j - intervalo
            end if
```

Tabla 1 Izq.: Enunciado de práctica proporcionado por los profesores del curso, Der: Solución proporcionada por un estudiante del curso

Sin embargo, actualmente estos documentos no proporcionan alguna manera de reconocer su contenido y, así, facilitar su recuperación. Además se sabe que este repositorio sigue en aumento ya que cada ciclo los alumnos suben solucionarios de prácticas y exámenes así como información que crean pertinente. Esto nos lleva a preguntarnos sobre cómo poder identificar dichos documentos en una base de datos muy grande.

Para poder anexar información a documentos, la ingeniería del conocimiento nos ofrece métodos de anotaciones semánticas. Esto es la clave para diferenciar las búsquedas sintácticas (bases de datos convencionales) con las búsquedas semánticas (ontologías). Estas anotaciones deberán pertenecer a un documento en específico, para poder relacionarlo con las diferentes propiedades de la ontología y así poder producir inferencias que nos permitan recuperar el documento (Kiryakov, Popov, Terziev, & Ognyanoff, 2004) .

[1] http://facultad.pucp.edu.pe/ingenieria/informatica/plancurricular

Las anotaciones semánticas ofrecen la ventaja de afinar bastante las fuentes de información y, así, ampliar la diferencia con las bases de datos convencionales pues con la ayuda de las ontologías y la anotación semántica se hace posible la búsqueda por temas de interés, en comparación a una base de datos convencional estructurada que entregará respuestas directas sin posibilidad de profundizar más en la investigación (Greengrass, 2000).

La recuperación de información no es un tema trivial. Para poder brindar información al usuario información correcta y útil, se debe cumplir una serie de validaciones. Desde asegurarse que la fuente de información tenga la estructura correcta, hasta analizar, con detenimiento, el tipo y forma de la consulta, para entender que es lo que realmente el usuario necesita. Se encuentran a disposición un conjunto de herramientas de la ingeniería del conocimiento, que nos proporcionan lo necesario para cumplir con estas validaciones (Rujiang & Junhua, 2009).

Considerando todo lo expuesto, este proyecto de fin de carrera propone un modelo para la recuperación de conocimiento, representado en documentos en repositorios digitales, basado en ontologías en el contexto de la Ingeniería Informática, buscando facilitar a los usuarios con herramientas que les permitan acceder a la información de un tema específico.

1.2 Objetivo general

Diseñar un modelo basado en ontologías en el dominio de la Ingeniería Informática, que facilite la recuperación de documentos almacenados en repositorios de contenido digital.

1.3 Objetivos específicos

Con la finalidad de alcanzar el objetivo general algunos de los objetivos específicos son los siguientes:

- **O1:** Analizar el escenario para establecer la cobertura, vacíos existentes y las tecnologías que se emplean para la representación del conocimiento.
- **O2:** Proponer una estructura capaz de soportar la recuperación de información así como los métodos y técnicas que nos facilitaran la extracción.
- **O3:** Establecer un proceso para la etiquetación semántica de documentos digitales.
- **O4:** Diseñar un prototipo que valide la viabilidad del modelo planteado.

1.4 Resultados esperados

- **Resultado esperado del O1:** Documento de revisión sistemática que contiene el protocolo de revisión y el análisis de los estudios y reportes.
- **Resultado esperado del O2:** Documento de la arquitectura planteada, que contiene un formulario OM-1 con el resumen del problema, contexto y solución, arquitectura lógica y un prototipo de la arquitectura final.
- **Resultado esperado del O3:** Documento conteniendo el proceso manual de anotación semántica, además de las ontologías a partir de las cuales se realiza y el planteamiento de la estructura donde se realizará la persistencia de la anotación.

- **Resultado esperado del O4:** Un piloto de prueba del prototipo que comprueba el adecuado uso del conocimiento estructurado.

A continuación se muestra un mapeo de los resultados esperados las herramientas a usarse.

Resultados esperado	Herramientas a usarse
RE1: Documento de revisión sistemática conteniendo el protocolo de revisión y el análisis de los estudios y reportes	Revisión Sistemática
RE2: Documento de arquitectura planteada, conteniendo una plantilla OM-1, arquitectura lógica y un prototipo de la arquitectura final	Metodología CommonKads, para la identificación de problemas y oportunidades, el contexto y la solución usando Plantilla OM-1 OM-2.
RE3: Documento conteniendo el proceso manual de anotación semántica, además de las ontologías a partir de las cuales se realiza y el planteamiento de la estructura donde se realizará la persistencia de la anotación.	RDF/XML, como lenguaje de etiquetado OWL (Ontology Web Language) lenguaje aprobado por la W3C para el manejo de ontologías. PROTÉGÉ, plataforma de diseño para el modelado de conocimiento usando ontologies
RE4: Un piloto de prueba del prototipo que compruebe el adecuado uso del contenido representado	JavaScript,HTML5 y PHP, como tecnologías web base Slim, micro framework para php 5.0 JOWL, librería en javascript para recorrer una ontlogía MySQL, para la persistencia de datos.

Tabla 2 Mapeo de herramientas y resultados esperados (Realizado por el autor)

1.5 Herramientas, métodos y procedimientos

En esta sección se muestra un mapeo de herramientas, métodos y procedimientos que se utilizaron, y una breve descripción de los mismos.

1.5.1 Metodologías

- **Revisión sistemática**

La revisión sistemática es una forma de identificar, evaluar e interpretar las investigaciones logradas hacia un tema en específico (Melgar Sasieta, 2011). Esta metodología muy conocida actualmente por los centros de investigación en el campo de la medicina, pero se ha extendido su uso en la Ingeniería del Conocimiento, pues a partir de la inferencia se logran resultados más adecuados para las interpretaciones. Una revisión sistemática se diferencia de una revisión narrativa pues cuenta con un proceso científico de investigación que asegura el entendimiento de una problemática a partir de su contexto, de manera que los resultados de la misma puedan ser utilizados por otros profesionales (The Cochrane Collaboration, 2011).

The Cochrane Collaboration[1] es una organización que promueve y difunde el uso de esta metodología para la investigación, la cual nos brinda una secuencia de fases para la revisión sistemática, entre los cuales tenemos:

1. **Planeamiento de la revisión**: incluye la identificación de las necesidades de la revisión, la elaboración de una propuesta para la revisión y finalmente un protocolo de revisión.

2. **Ejecución de la revisión**: ejecución del protocolo de revisión y la extracción y síntesis de datos.

3. **Elaboración de la documentación**: de manera que los estudios puedan ser re-utilizados por otros investigadores en el tema.

- **CommonKADS**

La metodología CommonKADS (Schreiber, 2000) nace de la necesidad de estructurar investigaciones, proyectos y sistemas basados en conocimiento. Se basa en la realización de 3 preguntas básicas: ¿Por qué? ¿Qué? y ¿Cómo?, ya que con ellas se puede llegar a definir un problema, verificar el contexto del mismo y plantear posibles soluciones.

[1] http://www.cochrane.org/

Para la obtención de un resultado final CommonKADS nos propone escalar entre ciertos niveles y modelos representados en la Figura 1, en ella observamos que para la obtención del modelo de conocimiento o de comunicación no es necesario realizar todo los modelos del nivel anterior, si no que esto dependerá del tipo de proyecto que se esté realizando.

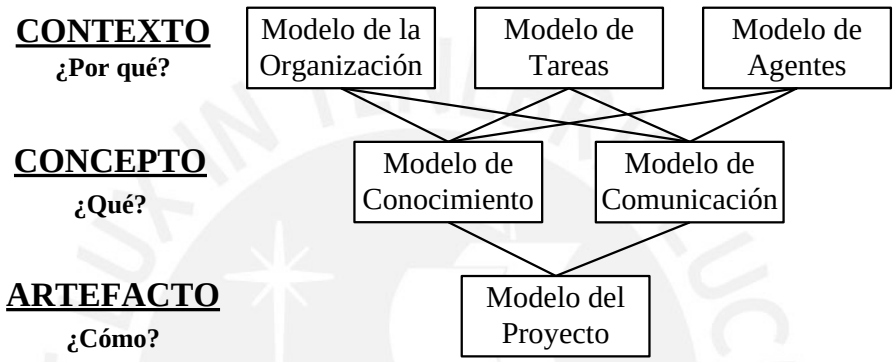


Figura 1 Modelos de la Metodología CommonKADS (Schreiber, 2000)

En nuestro caso se utilizará el Modelo de la organización para analizar las características principales de los estudiantes y así identificar problemas y oportunidades, contexto de la organización y posibles soluciones a los mismos, para ello CommonKADS nos provee de plantillas para la identificación de Problemas y Oportunidades dentro del Modelo de lo estudiado, como por ejemplo la plantilla OM-1 que podemos observarla en la Figura 2, nos recomienda hacer un listado de los problemas y oportunidades de la organización, analizar el contexto de la misma y proponer soluciones a esos problemas.

Modelo de Organización	Plantilla OM-1. Problemas y Oportunidades
PROBLEMAS Y OPORTUNIDADES	[Listado]
CONTEXTO DE LA ORGANIZACIÓN	Misión, visión, objetivos de la organización: [Texto]
	Los factores externos que la organización tiene que tratar: [Texto]
	Estrategia de la organización:[Texto]
	Su cadena de valor y los conductores de valor principales: [Mapa de los procesos]
SOLUCIONES	[Listado]

Figura 2 Plantilla OM-1 de la Metodología CommonKADS (Schreiber, 2000)

1.5.2 Herramientas de Desarrollo

- **RDF/XML**

RDF es el modelo de datos para metadatos, se basa en tripletes modelo entidad-atributo-valor. Un ejemplo de triplete puede ser: Objeto (curso), atributo (alumnos matriculados) y el valor (64). Este mecanismo para describir recursos es utilizado como recomendación de la W3C.

La colección de varios tripletes relacionados, se puede describir en un grafo dirigido y etiquetado, como se puede observar en Figura 3 Gráfico de un ejemplo de RDF/XML Figura 3. RDF/XML es la sintaxis normativa propuesta por la W3C, para expresar/serializar un grafo RDF como un documento XML.

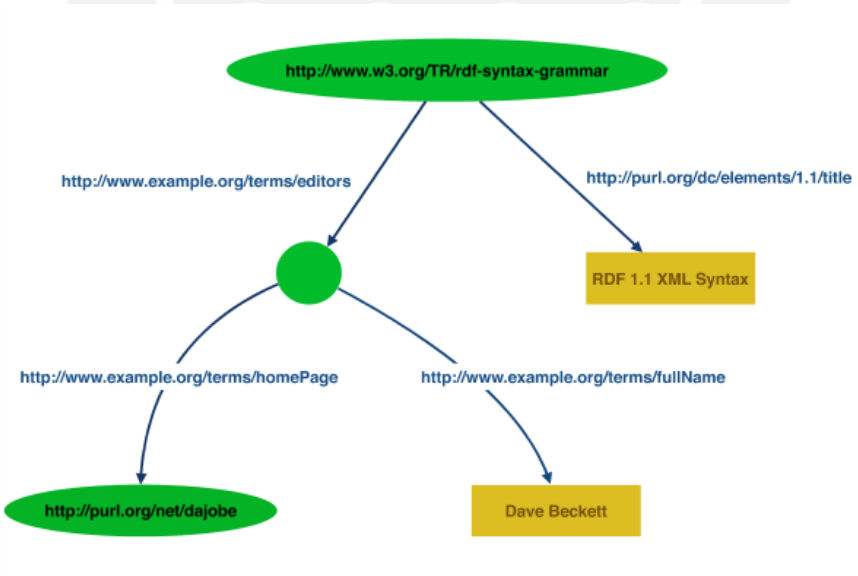


Figura 3 Gráfico de un ejemplo de RDF/XML (W3C, 2014)

- **OWL**

Web Ontology Language por sus siglas en inglés, es un lenguaje de marcación semántica para la publicación de ontologías en la Web. Se

desarrolla como una extensión del lenguaje RDF propuesto por la W3C para los procesos que se refieren a web semántica (W3C, 2014).

Dependiendo del nivel de exactitud que se desee emplear el lenguaje OWL se subdivide en, OWL LITE, OWL FULL y OWL DL, ésta última que se usará en este proyecto pues nos garantiza integridad de información, al generar clases y subclases en la ontología . (W3C, 2014).

- **PROTÉGÉ**

Es un editor de ontologías de código abierto desarrollado por la Universidad de Stanford[1], que ofrece una representación gráfica de las ontologías para un mejor entendimiento utilizando librerías gráficas en JAVA.

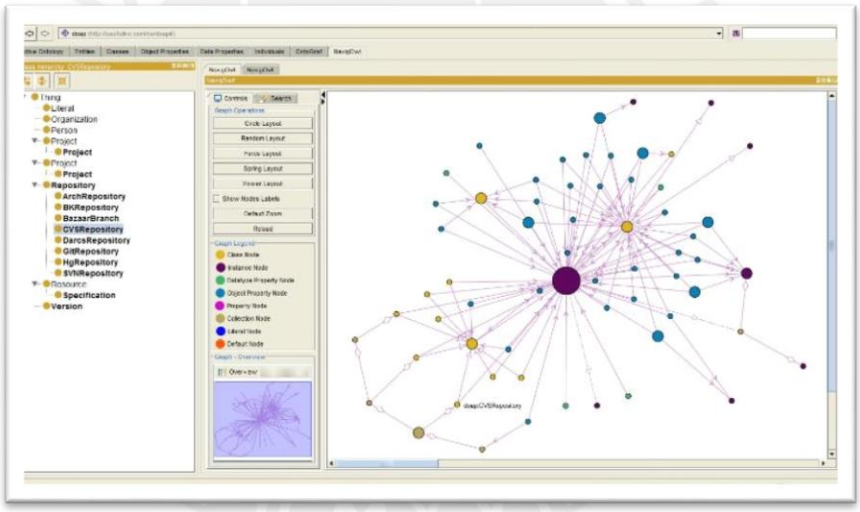


Figura 4 Ejemplo de ontología en Protegé http://protege.stanford.edu/

- **JOWL**

Es un plugin Jquery para navegar y visualizar documentos en formato XML/RDF. Tiene diferentes componentes, entre los que destacan un navegador de ontologías en vista de árbol, esto permite al usuario tener un control de lo que se busque en la estructura rdf.

Además un visualizador de contenidos de las entidades de la estructura ontológica, es decir un detalle de todas las relaciones que una entidad pueda

[1] http://protege.stanford.edu/

tener, desde dataRelationships (Entity - DataType) hasta propertyRelationsships (Entity - Entity)

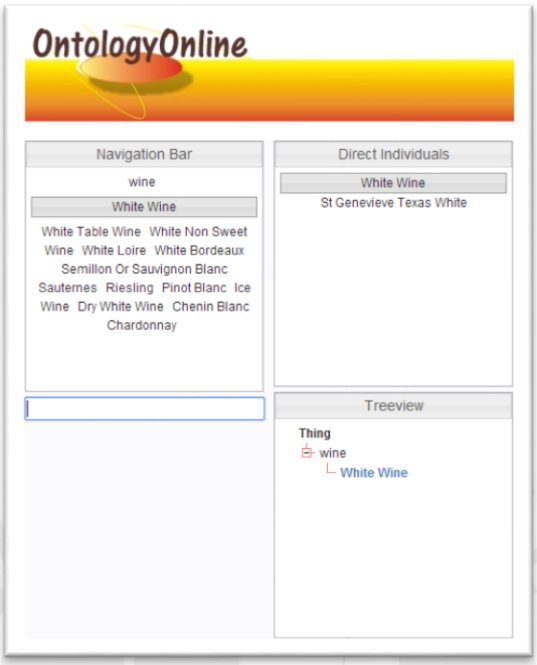


Figura 5 jOWL ontology online

1.6 Alcance

El presente proyecto se enfoca en proponer un modelo para la recuperación de documentos a partir de repositorios digitales utilizando herramientas de la Ingeniería del Conocimiento como son las ontologías, anotaciones semánticas y recuperación de información. Para esto se analizará de manera sistemática los problemas, oportunidades, contexto y posibles soluciones a la problemática presentada. Vale recalcar que se utilizará un proceso de anotación semántica manual en 36 documentos alojados en un banco de documentos digitales con la información de prácticas y exámenes pasados de los cursos de Ciencias de la Computación de la carrera de Ingeniería Informática de la PUCP. Para, así, obtener data de prueba que permita validar el modelo planteado.

1.7 Justificativa del proyecto de tesis

Como ya se ha mencionado, este proyecto tiene como finalidad proponer un modelo ontológico que sirva como base para la recuperación de documentos digitales de interés. Obtener resultados precisos para una consulta es lo que el usuario

persigue siempre que la realiza; sin embargo brindarle aún más información pertinente que servirá en su búsqueda genera un valor agregado para el usuario.

En cuanto al valor teórico e implicaciones prácticas, el desarrollo del modelo en el contexto de la Ingeniería Informática es conveniente ya que permitirá brindar al estudiante información sobre preguntas y soluciones de prácticas y exámenes que corresponden a ciclos anteriores y así reducir tiempos en el proceso de estudio para las evaluaciones. Además, este modelo permite la vinculación de conocimiento con conocimiento. Esto quiere decir, poder anexar información adicional, de importancia, en las respuestas, como por ejemplo: inferir estadísticas, preguntas tipo y resoluciones de las mismas de manera que entusiasmen al alumno a seguir practicando.

Un modelo ontológico se caracteriza por tener ciertos aspectos que hacen del mismo reutilizable para diferentes contextos. Si partimos del uso de una ontología base y un mapeo semántico, lo que resta para cambiar el contexto es la utilización de otra ontología de dominio, donde se representa el área que se desea estudiar. Es por ello que siguiendo este modelo clásico de ontologías se puede representar el conocimiento de diferentes problemáticas con sus respectivas características.

CAPITULO II: MARCO DE REFERENCIA

Para la redacción e investigación de este capítulo se utilizó la metodología de Revisión Sistemática a manera de validar el proceso de búsqueda. Por lo tanto, en el anexo 1 se explicará cómo se adaptó esta metodología para la realización de este proyecto de fin de carrera.

2.1 Marco conceptual

En esta sección se presentará algunos conceptos de importancia para la comprensión de este proyecto.

2.1.1 Objetivo del marco conceptual

El objetivo del marco conceptual es entender los conceptos relacionados a la problemática central del proyecto y el entorno en el que se desarrolla, para luego comprender los conceptos relacionados a las posibles soluciones presentadas por este proyecto.

2.1.2 Conceptos

A continuación se presentan una lista de conceptos que son importantes para el entendimiento del proyecto.

2.1.2.1 Metadato

Los metadatos se definen como datos sobre otros datos (Woodley, 2005), Por ejemplo si un bibliotecario quisiera registrar libros que se ofrecen en un estante, él tendría que hacer una lista sobre los datos trascendentales[1] de los mismos para su mejor administración. Es decir, que se obtendrá una lista de atributos distintivos de cada libro del estante y los almacenará en algún documento con una estructura definida.

En informática, metadato[2] se utiliza en diferentes aspectos. Por ejemplo, en las bases de datos una metadata podría ser un diccionario de datos representado en una tabla que contenga información sobre lo que se está almacenando en la BD. También es usado en la web, dado que se puede almacenar información en una página web especificando, por ejemplo, en qué lenguaje se escribió, qué herramientas se usaron y dónde debo ir para obtener más información relacionada al tema.

[1] El prefijo META se usa para indicar algo transcendental o poco natural
[2] Otro nombre para Metadatos

La DCMI[1] toma como referencia a Weibel y Lagoze (Weibel & Lagoze, 1997) para decirnos que la asociación de metadatos descriptivos tiene el potencial para mejorar la localización y recuperación de información, ya que permite la indización de objetos y, así, facilitar el acceso al contenido referenciado. Actualmente esta organización ofrece un marco de desarrollo (Coyle, 2009) y estándares para la creación de metadata en la web expresándolos en formato XML, RDF/XML, etc. Dentro de este marco, ayudará a crear un conjunto de elementos de datos que describan los documentos digitales con el fin de facilitar su búsqueda y recuperación. En la Tabla 3, podemos observar los 15 elementos de datos que la DCMI considera como estándar.

Contenido del Recurso	*Recurso como propiedad Intelectual*	*Instancia de los Recursos*
Título	Tipo	Origen
Editorial	Fecha	Idioma
Creador	Formato	Cobertura
Descripción	Identificador	Relación
Asunto	Contribuyente	Derecho

Tabla 3. Estándar DCMI para metadatas (DCMI, 2007)

Las aplicaciones de la metadata son muy amplias, pero principalmente se usan para recuperar información de un conjunto de datos, estructurados o no (DCMI, 2007), pues su función es darle cierto orden, de manera que el proceso de búsqueda sea más sencillo. Esto se aplica, por ejemplo, en la creación de la web semántica.

2.1.2.2 Ontología

El término ontología, en su sentido filosófico, trata del ser y de sus propiedades trascendentales de manera sistemática y la relación con su entorno (Bunge M. , Treatise on Basic Philosophy: Ontology I", 1977) (Bunge M. , Treatise on Basic Philosophy: Ontology II, 1979). Por otro lado, desde el punto de vista tecnológico, Gruber (Gruber, 1993) lo define como "una especificación explícita de una conceptualización". Esto va muy relacionado al concepto de metadatos. Para completar esta definición, Pease (The Knowledge Engineering Review, 2002), define

[1] Dublin Core Metadata Initiative: Organismo internacional que se encarga de fomentar el

ontología como “Un conjunto de conceptos, axiomas y relaciones que describen un dominio de interés”. Estas definiciones nos dan una idea más clara para poder responder la pregunta ¿Qué es una ontología?, y lo definimos como un cuerpo estructurado de metadatos que nos sirve para la comprensión de un dominio o ámbito del conocimiento.

La IC[1] conoce la importancia del conocimiento en la resolución de problemas, y utiliza diferentes herramientas para la representación del mismo. A lo largo de los años, se ha ido evolucionando en este aspecto, desde la representación más simple como terminologías y tesauros[2] hasta llegar a la utilización de redes semánticas, éstas últimas divididas en Redes IS-AS, cuya característica se basó en que la relación entre metadatas es etiquetada[1].

Las ontologías son herramientas útiles para la comunicación entre expertos en un área determinada. Se puede formalizar conceptos en un dominio para que luego sean validadas por computadores y obtener respuestas más favorables a la hora de consultarlas (Greengrass, 2000).

Formalmente las ontologías son usadas para representar el conocimiento, la razón por la cual se usan estos esquemas, es debido a que nos permite realizar búsquedas de información con carácter semántico, con la finalidad de obtener resultados mucho más óptimos (Melgar Sasieta, 2011).

Una búsqueda semántica se puede definir como un proceso de optimización de los resultados obtenidos por un motor de búsqueda, utilizando las propiedades de la semántica. En este contexto, las ontologías nos ofrecen mecanismos de relación conceptual que nos permiten transformar la consulta en una más enriquecida y representada de manera estructurada, para luego compararla con la ontología que maneja el motor utilizado y así devolver mejores resultados (Rujiang & Junhua, 2009).

Para la representación de ontologías se utilizan diferentes lenguajes que han ido evolucionando en el tiempo, la finalidad de estos lenguajes es de representar a las mismas de manera que sean entendibles por el computador y por el usuario, por ejemplo: RDF/XML y OWL.

[1] IC: Ingeniería del Conocimiento
[2] Tesauro: Se define como una jerarquía arbórea de conceptos

2.1.2.3 Web Semántica

La W3C realiza diversas actividades como parte del proyecto de lograr una web donde armonicen sistemáticamente los usuarios y la tecnología. Para ello se basa en la idea de añadir información estructurada en metadatos a la web, con la ayuda de diferentes marcos y metodologías aprobadas por ellos. (W3C, 2014)

Según Tim Bernners-Lee (Berners-Lee, The Semantic Web, 2001), "La web semántica brindará una estructura al contenido significativo de las páginas web". Además de ello, asegura que la web semántica no está desligada de la web actual, si no que se convierte en una extensión de la misma. Esto se refiere a que la web semántica es un espacio donde la información tendrá un significado bien definido de manera que pueda ser entendido por el computador y por las personas.

Para el propósito de lograr una web semántica, se necesita representar el conocimiento de manera que ambas partes tengan una comunicación fluida (persona-computador). Como ya sabemos, el esquema utilizado será la ontología, la cual representa una capa superior dentro de la estructura de la Web Semántica, como se aprecia en la Figura 6 (Berners-Lee, World Wide Web Consortium: XML and The Web, 2000). La capa del esquema de RDF se encuentra por debajo de la de Ontología, esto debido a que la primera define todas las relaciones entre clases para que luego la RDF se encargue de la definición de la semántica de los datos.

Si bien es cierto que en la web se pueden manejar diferentes lenguajes de programación para la realización de páginas web, se maneja un lenguaje de etiquetación propuesto por la W3C para el desarrollo de la web semántica y para su representación: el lenguaje RFD. Un ejemplo de esta representación lo vemos en la Figura 7. En ella se aprecia la estructura básica de un triplete, como primer nodo tenemos el encabezado, que representa la base de la identificación de la data. Además, de la descripción del objeto a desarrollar, identificado por la etiqueta "rdf:about".

Para finalizar, las características y valores, relacionando, así, la data con una respectiva URI[2] que se construye a partir de la base y características del objeto, de manera que los recursos sean de más fácil acceso.

[1] Se le pone un nombre a la relación
[2] Uniform Resource Identifier http://www.w3.org/Addressing/

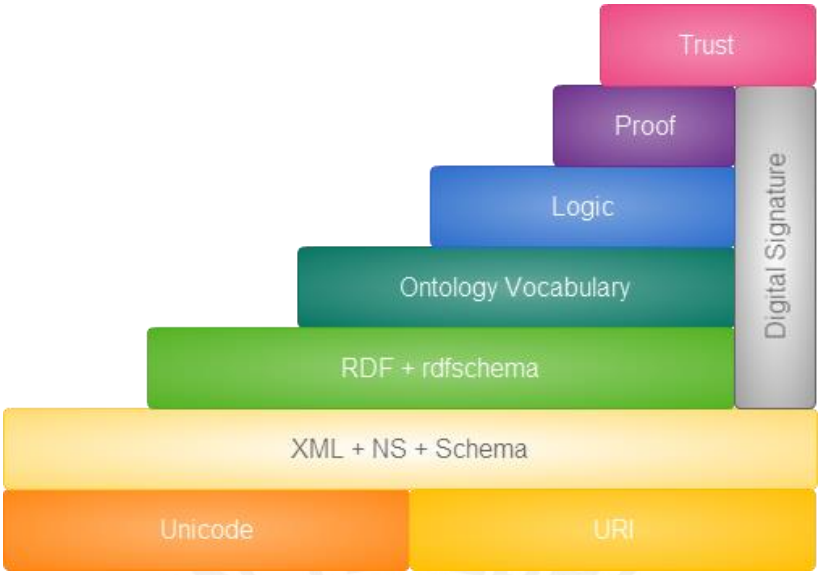


Figura 6 Capas de la Web Semántica (Berners-Lee, The Semantic Web, 2001)

```
<rdf:RDF
    xmlns:rdf="http://www.w3.org/1999/02/22-rdf-syntax-ns#"
    xmlns:lsldREL="http://cbgp.upm.es/lsld.rdf#"
    xml:base="http://cbgp.upm.es/lsld">

    <rdf:Description rdf:about="#prot_b"/>
    <rdf:Description rdf:about="#cell_cycle"/>
    <rdf:Description rdf:about="#prot_a">
      <lsldREL:participates_in rdf:resource="#cell_cycle"/>
      <lsldREL:interacts_with rdf:resource="#prot_b"/>
    </rdf:Description>
</rdf:RDF>
```

Figura 7 Ejemplo de representación RDF (Berners-Lee, World Wide Web Consortium: XML and The Web, 2000)

En la Figura 8, podemos apreciar claramente lo diferente que es una Web semántica frente a la web actual, mientras una se encuentra relacionada solo por términos la otra se relaciona por conceptos. Esto ayuda claramente a tener una fuente de información más ordenada y entendible para los humanos.

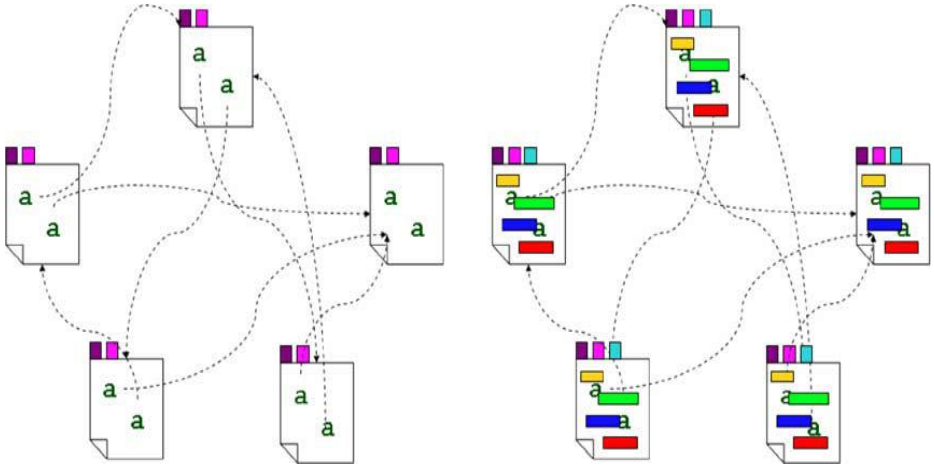


Figura 8 Actual WWW (izq.) y Web Semántica (der) (Kiryakov, Popov, Terziev, & Ognyanoff, 2004)

2.1.2.4 Anotación Semántica

Por anotaciones entendemos comentarios, notas u otro tipo de reseñas que se pueden asociar a un documento. En lingüística computacional se puede entender como el proceso mediante el cual se introduce información relevante dentro de archivos que correspondan a páginas web para así enriquecerlos, y así se pueda relacionar de manera conceptual con otros metadatos, y se representado por cierto identificador conceptual, una URI tomado de una ontología o de otra fuente de conocimiento (Kiryakov, Popov, Terziev, & Ognyanoff, 2004).

Existen 2 alternativas para realizar anotaciones: anotaciones empotradas y anotaciones externas. La primera se refiere a que la metadata se incluye dentro de los archivos, ya sea páginas web o documentos digitales. La segunda se refiere a que las anotaciones se almacenan en un servidor externo, cada usuario puede agregar anotaciones y puede compartirlas con otros usuarios, volviéndose así más dinámicas.

Las “Named Entities” (Kiryakov, Popov, Terziev, & Ognyanoff, 2004) son definidas en el campo del Procesamiento del Lenguaje Natural, como el nombre que se le da a entidades ya sean nombres de personas, nombres de lugares, etc. Además, incluyen valores numéricos como fechas, números y cantidades. La anotación semántica consiste en asociar las entidades a sus respectivas definiciones. Como podemos ver en la Figura 9, a gran escala se refiere a la identificación de entidades por medio de híper-links a sus definiciones a fin de enriquecer los documentos y así permitir el acceso a ellos de diferentes maneras.

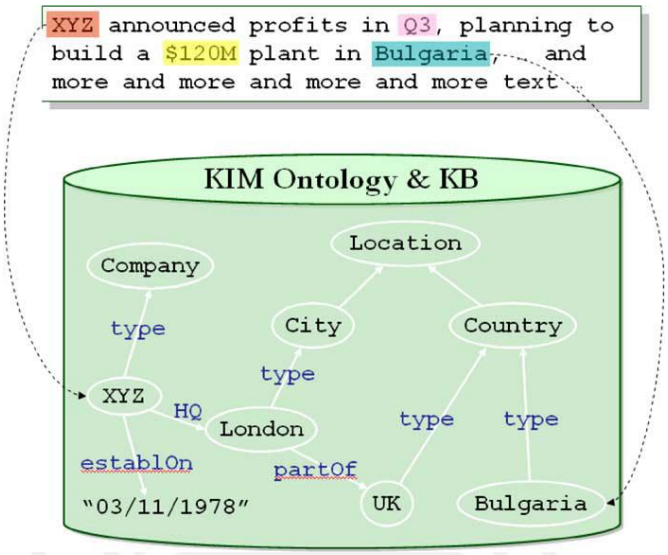


Figura 9 Anotación semántica (Kiryakov, Popov, Terziev, & Ognyanoff, 2004)

En (Oren, Moller, Scerri, Handschuh, & Sintek, 2006), nos explican la siguiente analogía acerca de las URI's y la anotación semántica:

> "Metaphorically, we can see URIs as the "atoms" of the Semantic Web and semantic annotations as the "molecules". The Semantic Web is about shared terminology, achieved through consistent use of URIs. Annotations create a relationship between URIs and build up a network of data"

Es decir que estas moléculas (anotaciones) permitirán que los átomos (URIs) se relaciones con otros átomos y así crear redes consistentes de información. Se puede afirmar, entonces, que con la ayuda de anotaciones contribuimos a la realización del proyecto de Web Semántica, ya que con su uso realizamos las conexiones entre páginas web y entidades de archivo, que ya se mencionaron.

Para la realización de una anotación semántica, se necesita un lenguaje de marcado, basado en RDF. El lenguaje que propone la W3C es OWL, Web Ontology Language. Este lenguaje nos proporciona más vocabulario de manera que nos permita describir clases y la relación entre ellas, también el uso y representación de la cardinalidad. Con ello se tiene cubierto el primer nivel de representación. Luego de ello, se tiene que buscar tecnologías que nos permitan crear ontologías y anotaciones

de manera práctica. Entre ellos tenemos JOWL y entornos de desarrollo como Protegé[1], los cuales se explicaran más adelante como herramientas de desarrollo.

2.1.2.5 Recuperación de Información

La recuperación de Información (Mitra, 2000) es el proceso mediante el cual se busca información sobre una determinada base de datos, mediante una consulta realizada por el usuario. La tarea de la IR, se puede ver como parte del proceso de la recuperación del conocimiento, que inicia con la consulta del usuario, para luego seguir y realizar las demás tareas del proceso, como por ejemplo la extracción de información, Question Answering[1] y Sumarización de Textos (Melgar & Pacheco, 2010).

La finalidad de la recuperación de información es devolver al usuario la mayor cantidad de información relevante sobre el tema buscado, y a su vez reducir la cantidad de información irrelevante para así optimizar la recuperación.

En este contexto, para la problemática que se desarrollará en este proyecto de tesis, se responderá a la siguiente pregunta: ¿cómo la recuperación de información puede ser aplicada a recuperar documentos digitales? Es aquí donde los demás conceptos y herramientas de la recuperación de información se utilizan para dar respuesta a la pregunta.

La IR se basa en las ontologías y búsqueda semántica para poder recuperar información de cualquier tipo de estructura, ya sean archivos estructurados (web pages) o no estructurados como son los archivos PDF. Usando la anotación semántica en los documentos PDF, la IR soporta la recuperación de documentos, a diferencia de si es que se usara una base de datos convencional, ya que la posibilidad de realizar inferencias es menos común (Melgar & Pacheco, 2010).

En (Greengrass, 2000), el autor describe modelos clásicos de recuperación de información. En estos modelos se indica que los documentos se pueden representar por medio de palabras representativas que les denominan términos índice. Estos se

[1] Página Oficial de Protege - http://protege.stanford.edu/

utilizarán para resumir los documentos e indizarlos y así hacer más rápido el acceso a ellos. Estos índices, para su mejor manejo, se representaran también por medio de números que indicarán la importancia que tienen dichos términos en el documento. Estos 2 modelos son el Modelo Booleano y el Modelo Vectorial.

El modelo Booleano, utiliza la lista de términos antes mencionada pero su representación es binaria; es decir, que solo se representará si es que se encuentra el término o no en el documento. Esto nos indica que para la elección de documentos relevantes solo bastará con la aparición del término en el documento, es decir que aparezca en la lista de términos.

Por las carencias de la lógica, este modelo no soporta la relevancia de documentos. Es decir, un documento se declara como relevante o no relevante, no existe el "parcialmente" relevante. Por ejemplo, si tuviéramos la siguiente cadena de búsqueda: "Algoritmia AND Punteros AND Void" y obtenemos que en los términos 1 y 2 arrojó positivo es decir que en el documento "INF263-PB4-20141.pdf" si se encontró dichos términos, pero el tercer término arrojo negativo (no se encuentra en el documento), automáticamente descarta el documento como relevante. Este modelo se asemeja en demacia a la teoría de conjuntos (ver Figura 10).

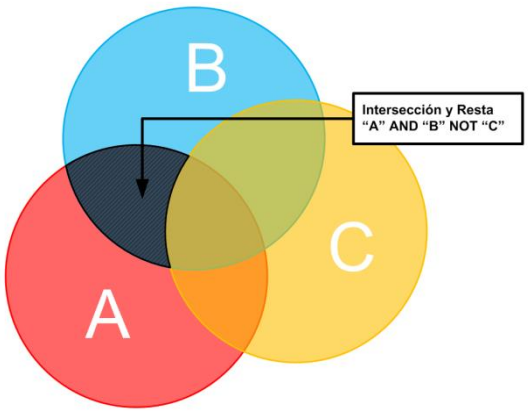


Figura 10 Modelo booleano como teoría de conjuntos (Blazquéz Ochando, 2011)

El modelo Vectorial, tiene como principal característica usar un vector de valores no binarios, los que permiten diferentes valores de acuerdo a la frecuencia de las apariciones de los términos en el documento y de la importancia que representa en él. Además, resuelve la molestia de que un conjunto de palabras (*Stop Words*) que

[1] Búsqueda de Respuestas: es un tipo de IR donde se recupera información a través de preguntas en lenguaje natural.

aparecen en muchos documentos, no sean consideradas dentro de la lista principal de términos del documento, pues puede que no sea útil considerarlas ahorrando mucho tiempo de ejecución y afinando la respuesta de documentos relevantes.

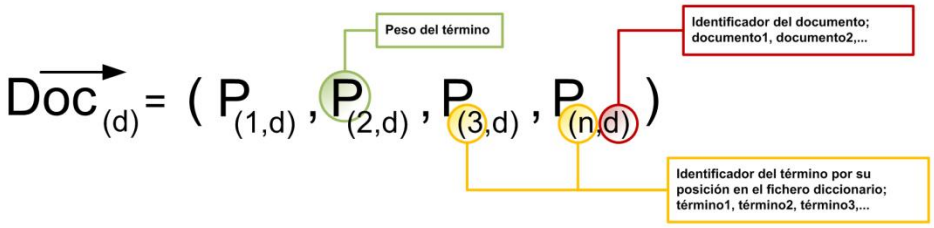


Figura 11 Modelo Vectorial - Arreglo de frecuencias y Pesos (Blazquéz Ochando, 2011)

Existen estructuras que nos ayudan en la recuperación de información: Vector Semántico y Vector Invertido, el primero de ellos utiliza la ISL (Indexación semántica latente) o LSI, como nos explica Greengrass en (Greengrass, 2000), su característica fundamental es la relación de conceptos; es decir, un documento puede ser recuperado si comparte el mismo concepto que otro que sea relevante a la consulta.

Por otro lado, el vector invertido, nos provee de una lista de palabras relevantes y la frecuencia de aparición en todos los archivos de la base de datos en la que se realiza la búsqueda. La creación de un vector invertido se puede describir de la siguiente manera:

- Realizar una lista parseando los documentos y extrayendo “tokens” que serán guardados junto con su respectivo identificador de documento.
- Ordenar esa lista alfabéticamente
- Se agrega una frecuencia de aparición para así eliminar términos repetidos con el mismo identificador de documento.
- Con esa lista se crea un diccionario de datos completo y un “postings” que nos servirá para comparar frecuencias en archivos.

Este tipo de vector permite una búsqueda de información más rápida, dónde por cada peso de un término se tendrá un triplete conteniendo Id del documento, la frecuencia del termino en el documento y posición del termino en el documento (ver Figura 11).

2.1.3 Conclusiones

Como se puede observar, estos conceptos nos permiten entender el "core" del proyecto que tiene que ver con recuperar conocimiento a partir de documentos digitales y permitir que los usuarios obtengan información de interés.

2.2 Estado del Arte

A continuación se presentan proyectos actuales de diferentes tipos, que hayan hecho uso de las ontologías para la recuperación de Información. Además se hará énfasis en las técnicas utilizadas en los proyectos, a manera de verificar como se encuentra el estado del arte del tema a tratar.

2.2.1 Objetivos de la revisión del estado del arte

El objetivo de la revisión del estado del arte es dar a conocer las diferentes propuestas que se han planteado a la fecha acerca del uso de las ontologías en el proceso de modelamiento de conocimiento y recuperación de información como respuesta al problema identificado.

2.2.2 Método usado en la revisión del estado del arte

El método utilizado para la revisión del estado del arte es la Revisión Sistemática, este se encuentra detallado en el Anexo 1.

2.2.3 Alternativas de solución

Actualmente, existen diversos proyectos relacionados al área de recuperación de información. Estos utilizan diferentes técnicas y herramientas para lograr su objetivo. A continuación se detallaran algunos proyectos representativos.

2.2.3.1 Proyecto Mesh

MeSH (U.S. National Library of Medicine) es un vocabulario controlado utilizado para la indización de artículos en MedLine[1]. Está formado por conjuntos de términos denominados descriptores organizados en una estructura jerárquica que permite la búsqueda a diferentes niveles de especificidad. Actualmente, existen varios proyectos que usa la Ontología Mesh como base para la recuperación de información utilizando

[1] Base de datos de la Biblioteca Nacional de Medicina de los Estados Unidos http://www.nlm.nih.gov/

expansión de consulta en el campo de la medicina (Mariano Crespo & Maña, 2011) (Perea Ortega, Montejo Ráaez, Díaaz Galiano, & García Cumbreras, 2011).

2.2.3.2 Proyecto Wolfram Alpha

Wolfram Research es una compañía de software cuyas investigaciones en el área de la matemática e informática son muy reconocidas a nivel mundial. Uno de sus proyectos actuales es Wolfram Alpha, que consiste en un meta buscador basado en ontologías (Wolfram Research) para recuperar información en base a consultas, lo que provee la habilidad de resolver preguntas de manera directa y no se limita a proporcionar los links de las páginas web que podrían solucionar su consulta, sino que además resume la información para que pueda computada nuevamente por el usuario. (Wolfram Research, 2012)

Figura 12 Una consulta en wolframalpha.com

Actualmente, se siguen realizando comparaciones entre este buscador y los que ya conocemos por ejemplo en (Talbot, 2014), el autor hace una comparación usando consultas idénticas y compara sus resultados para ver qué tan acertada pueden ser sus resultados.

2.2.3.3 Proyecto Xonto

XONTO es un proyecto para la realización de un sistema basado en ontologías para la extracción de información semántica de documentos PDF (Oro & Ruffolo,

2008). Aplica la arquitectura GATE[1] (General Architecture for Text Engineering) la cual es un conjunto de herramientas desarrolladas en Java. Actualmente es desarrollada por la Universidad de Calabria, y es usada por una gran cantidad de científicos y estudiantes interesados en PLN[2], incluyendo extracción de información.

2.2.3.4 Proyecto Ontoghobi

Es un modelo de un meta-buscador, que incorporando la ontología WORDNET[3] , el sistema se desarrolla en la Universidad de Cauca – Colombia y consta de los siguientes módulos: módulo de *consulta remota bilingüe,* un módulo de *indexación y filtrado bilingüe,* módulo de *filtrado y ordenamiento,* y módulo de *perfil de usuario* (Ordoñez & Cobos, 2010).

2.2.3.5 Proyecto Aondë

Servicio web basado en ontologías, orientado hacia el dominio de la biología, el cual trabaja con ontologías biológicas que proporcionan descripciones acerca del estado de las especies en la cadena alimenticia, con la finalidad de responder a los biólogos cuando ellos requieran información acerca de plantas, animales, etc. Su arquitectura se caracteriza por el uso de dos repositorios: un repositorio semántico, organizado en ontología y espacios de metadata, y un repositorio de ontologías en donde publican datos de ontologías vía servicios web (Bauzer Medeiros & Daltio, 2008).

2.2.4 Resumen Comparativo

Para la comparación de los proyectos investigados, se utilizó un conjunto de atributos que nos ayudaron a filtrar otros proyectos. Además estos atributos resaltan las características del modelo que se plantea. En la Tabla 4 Cuadro comparativo de Proyectos que usan Ontologías para IR.se encuentran mapeados los proyectos y las propiedades escogidas.

[1] http://gate.ac.uk/
[2] Procesamiento de Lenguaje Natural
[3] http://wordnet.princeton.edu/

	Proyecto MeSH	Wolfram Alpha	XONTO	ONTOGHOBI	AONDË
Tipo Proyecto	Ontología	Meta-Buscador	Sistema	Meta-Buscador	Web Service
Lenguaje y herramientas utilizadas	-	-	OWL/DLP	Word- Net	RDF, Jena, OWL, RDQL, SPARQL, Protégé, PostgreSQL
Arquitectura utilizada	-	-	GATE	Cliente - Servidor	WeBIOS
Técnica de IR utilizada	Expansión de Consulta	Expansión de Consulta	SDO	Stemming	Vector Invertido, Repositorio Semántico
Usan Ontologías	SI	SI	SI	SI	SI
Procesamiento	Semi-Automático	Automático	Semi - Automático	Automático	Semi - Automático
Responsable de desarrollo	National Library of Medicine - EEUU	Wolfram Research	Universidad de Calabria - Italia	Universidad del Cauca - Colombia	Universidad Estatal de Campiña - Brasil

Tabla 4 Cuadro comparativo de Proyectos que usan Ontologías para IR.

Actualmente, en la web semántica se utilizan muchos meta-buscadores que permiten encontrar soluciones a consultas de carácter científico. Tal es el caso de Ontoghobi y Wolfram Alpha. Algunas de las ontologías utilizadas por esos meta-buscadores no son conocidas por motivos empresariales lo que reduce en ciertos aspectos la investigación global de los proyectos así como las diferentes aplicaciones de los mismo. Esto contrasta por ejemplo con la ontología MeSH, la cual ha generado diferentes trabajos de investigación en diferentes partes del mundo y permite generar avances en el campo.

Para este proyecto, se propone realizar una ontología en el dominio de Ingeniería Informática, con un alcance en la rama de Ciencias de la Computación.

La ontología propuesta será de libre acceso para la comunidad de la universidad para futuras investigaciones y mejoras, promoviendo así el área de investigación de la especialidad de Ingeniería Informática de la PUCP en temas de Ingeniería del Conocimiento.

En resumen, lo que se desarrollará en este proyecto tratará de contener las mejores prácticas definidas por la W3C, además de las características más importantes de cada proyecto descrito, al mismo tiempo que incluir nuevas características que no se encontraron en los proyectos mencionados.

CAPITULO III: PROCESO DE RECUPERACIÓN DE INFORMACIÓN, ANÁLISIS DEL DOMINIO

3.1 Presentación del Proceso

En los siguientes 3 capítulos se presentará el diseño de la estructura que soportará recuperar conocimiento basado en: el análisis del dominio realizado, las ontologías ya existentes y en la metodología CommonKADS. Para el diseño de la estructura, se dividió el proceso de recuperación de conocimiento en 4 secciones: análisis del dominio, anotación y persistencia de documentos, recuperación de información y aplicación.

Con el análisis del dominio se plantea la construcción de una ontología que tiene como objetivo consolidar la información del dominio y así hacerla computable. Anotación y persistencia de conocimiento, es donde se definen ciertas estructuras que brindarán soporte a la recuperación de conocimiento. Los mecanismos utilizados se explicarán en la parte de recuperación de información. Finalmente, para lograr la aplicación se implementarán y utilizarán componentes que nos permitan visualizar los documentos en un entorno web, utilizando los métodos anteriores para su recuperación. En este capítulo se presentará la primera fase del proceso de recuperación de conocimiento.

Recuperación de Conocimiento			
CommonKADS	Especificación de la plataforma hw/sw	Mecanismos de desambiguación y Navegación en la ontología	Prototipo WEB
Analisis del dominio	Anotación y Persistencia	Recuperacion de Información	Detalle del diseño de la aplicación
PASO 1	PASO 2	PASO 3	PASO 4

Tabla 5 Fases de la Recuperación de Conocimiento (Realizado por el autor)

La Tabla 5, es una guía para la realización de cada uno de los resultados esperados de este proyecto de fin de carrera, estos mapeados en los capítulos 3, 4 y 5 además del anexo 1. En este capítulo se desarrollara la primera fase del proceso de

recuperación de información. Esta fase se mapea directamente al Objetivo Especifico 2, que nos habla del planteamiento de estructuras capaces de soportar la RI.

3.2 CommonKADS

Se utilizó el modelo de organización de CommonKADS buscando caracterizar y estructurar el modelo del conocimiento que se estudia en el este proyecto. Para lograr esto, se utilizan un conjunto de plantillas propuesta por la metodología (OM1, OM2). Estas plantillas logran identificar los problemas y oportunidades, así como las relaciones entre los componentes en el proceso de recuperación de conocimiento.

La plantilla OM-1, que se muestra en la Tabla 6, nos muestra una descripción general de la organización. En este caso, nos preocupamos por describir a los alumnos de la facultad de ingeniería informática y sus necesidades de información. En este contexto, se aprecia que dada la necesidad de obtener información relevante que sirva para que puedan repasar lo visto en clase, los alumnos de la facultad requieren de mecanismos que les permitan sacar el máximo provecho de los materiales a disposición.

Problemas	Dificultad de los usuarios para encontrar información de importancia para estudiar en los cursos de informática.
Oportunidades	Apoyo de los profesores brindando material de estudio
Contexto	Crecimiento desmesurado de información en formato digital. Grupos de estudio en los cursos “Fuertes” de la especialidad.
Soluciones	Integrar y estructurar la información digital en un repositorio que permita realizar procesos de recuperación de conocimiento.

Tabla 6 Plantilla OM-1 (Realizado por el autor)

La plantilla OM-2 se enfoca en resaltar los procesos que se llevan a cabo en la organización, las personas o tipos de usuarios que se desenvuelven en el contexto, los recursos que ellos utilizan para realizar los mencionados procesos y el conocimiento utilizado y adquirido por los usuarios.

En la Tabla 7, se presenta el mapeo realizado para este contexto y la identificación de procesos.

Procesos	P1. Recuperación de conocimiento P2. Recuperación de documentos
Personas	Alumnos de la facultad de informática que necesitan información estructurada
Recursos	R1. Repositorios digitales de documentos
Conocimientos	C1. Conocimiento relacionado del dominio C2. Entidades etiquetadas en los documentos digitales

Tabla 7 Plantilla OM-2 (Realizado por el autor)

En este sentido, luego de realizar el análisis de la organización, se propuso una serie de entidades clave para la representación del dominio de Ingeniería Informática. En la Tabla 8 se aprecia un resumen de las entidades y sus principales propiedades de relación, ya sea con otras entidades o individuos de la ontología.

Entidad	Propiedad	Valor
Facultad	tieneEspecialidad	Especialidad
Especialidad	tieneCurso	Curso
Curso	tieneProgramaAnalitico	Programa Analítico
Programa Analítico	tieneUnidadAprendizaje	Unidad Aprendizaje
Unidad Aprendizaje	tieneConcepto	Concepto
Concepto	-	Individuos

Tabla 8 Entidades y relaciones (Realizado por el autor)

Las entidades fueron sacadas en su mayoría de las páginas web de la universidad, donde presentan la distribución de especialidades en diferentes facultades. Además de la distribución de cursos, sus respectivas unidades de aprendizaje y conceptos a tratar, que se encuentran en los programas analíticos disponibles para el alumnado.

Los individuos de la ontología, para este caso en particular, nos proporcionaran la última instancia de relación entre los conceptos finales y los documentos digitales relacionados, en el capítulo 4 se explicará a detalle esta relación.

En la Figura 13, observamos que existen instancias para cada una de las entidades, esto nos ayudará a esquematizar los hechos para poder realizar las inferencias. Un ejemplo de hechos sacado de la ontología podría ser:

- Facultad: Ciencias e Ingeniería
    - tieneEspecialidad: IngenieriaInformatica
    - Especialidad: IngenieriaInformatica
        - tieneCurso: Algoritmia
        - Curso: Algoritmia
            - tienePrograma Analitico: INF263
            - ProgramaAnalitico: INF263
                - tieneUnidadAprendizaje: ProgramacionDinamica
                - UnidadAprendizaje: ProgramacionDinamica
                    - tieneConcepto: BinPacking

Figura 13 Ontología planteada en el dominio de la Ingeniería Informática

3.3 Consideraciones Finales

En este capítulo, se han presentado el análisis del dominio de Ingeniería Informática, para proponer una estructura ontológica capaz de soportar y almacenar todas las relaciones e inferencias producidas en el dominio.

Se recalca que, todas las versiones de la ontología propuesta se encuentran en formato owl, que en su interior se representan por la sintaxis en RDF/XML. Se utilizó esta sintaxis debido a que las diferentes librerías utilizadas en el proyecto usan este formato por defecto, de esta manera evitaremos pérdida de información e incongruencias.

Se utilizaron una serie estándares de codificación en la ontología que se definieron al inicio del proceso de construcción y se tomó como guía los estándares de desarrollo que proponen la W3C en su documentación de PROTEGÉ.

CAPITULO IV: PROCESO DE RECUPERACIÓN DE INFORMACIÓN, ANOTACIÓN Y PERSISTENCIA

Luego de modelar la ontología en el dominio de Ingeniería Informática de la PUCP, se vio la necesidad de utilizar un conjunto de documentos que permitan probar el modelo planteado y sus aplicaciones.

En este capítulo, se presenta la segunda fase del proceso de recuperación de conocimiento, que implica la organización del corpus[1] de documentos, haciendo énfasis en el proceso de etiquetado y en su almacenamiento en un repositorio digital.

4.1 Modelo de persistencia de datos

Para poder recuperar conocimiento a partir de un conjunto de documentos, estos deben estar representados de alguna manera en nuestro aplicativo. Para ello, se definieron características y propiedades a manera de emular clases en nuestro código. Estas clases deberían ser fáciles de recuperar y guardar, por lo que se estableció un modelo básico de persistencia de datos.

Este modelo se caracteriza por una estructura en MySql donde se realice el "match" entre los documentos electrónicos y las entidades de la ontología. Para lograr esto, se utilizó como identificador general las URI's, debido a que nos provee de un identificador universal que permite reducir a 0 el riesgo de inconsistencia de información.

Para la representación de un documento, se utilizó una URL, esto indicaba su ubicación lógica en el repositorio digital. Las propiedades que se almacenaron fueron: nombre del curso, ciclo de dictado, tipo de evaluación y url. Todo el repositorio digital fue almacenado en un servidor proporcionado por la universidad y que es compartido por los tesistas que utilizan este mismo corpus de documentos.

La relación indispensable para el proyecto es la que se genera a partir de una entidad de la ontología y un documento. Es por ello que se decidió utilizar una tabla en la que se permita almacenar el etiquetado de los documentos y relacionados a manera de uno a muchos, para así permitir que un documento pueda ser recuperado por más

[1] Corpus: Conjunto lo más extenso y ordenado posible de datos o textos científicos, literarios, etc. , que pueden servir de base a una investigación
http://buscon.rae.es/drae/srv/search?val=corpus

de una inferencia en la ontología. En la Figura 14, se muestra como un documento puede estar mapeado a diferentes entidades de la ontología.

En este ejemplo en particular, se escogió una práctica del curso de Algoritmia, que se dicta en 5to ciclo de la especialidad de Informática, en el silabo del curso se puede apreciar la lista de las unidades de aprendizaje y los conceptos que se verán durante el ciclo. Tal es así, que en una práctica pueden estar mapeados diferentes conceptos de una o más unidades de aprendizaje.

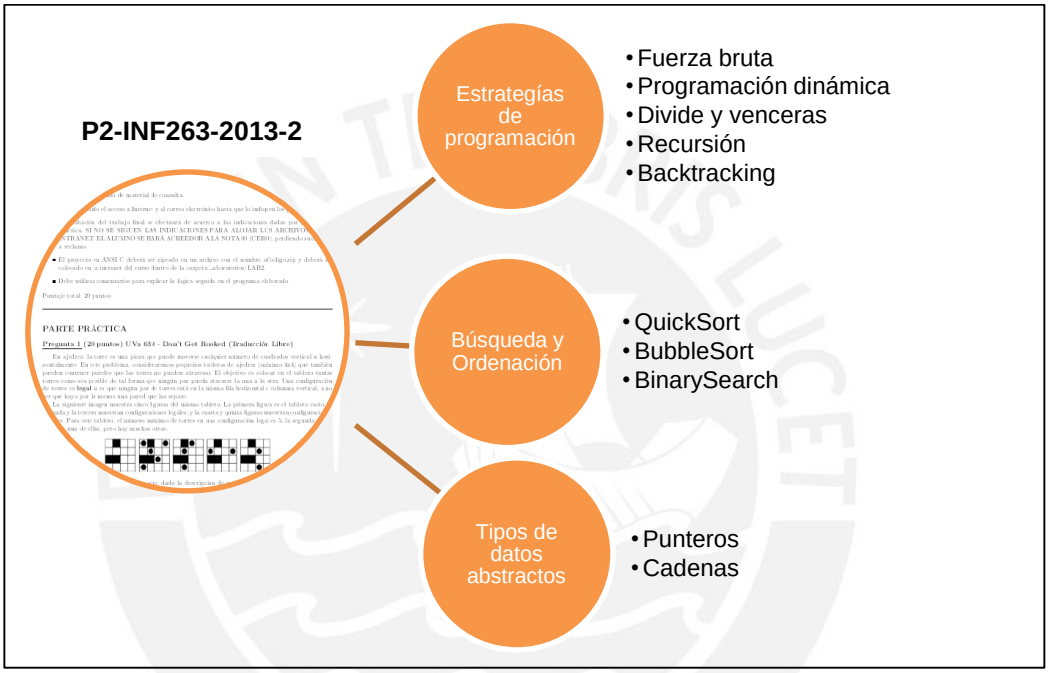


Figura 14 Mapeo de documntos a Clases e Individuos

Figura 15 se observa la estructura propuesta, en la que se almacenan los documentos, donde la utilización de las URI's para realizar las búsquedas mejoran considerablemente esta relación. Otra ventaja de utilizar las URI's es que al realizar la etiquetación se podría agregar un atributo adicional "TAG" que nos permita utilizar y contribuir con el proceso de desambiguación.

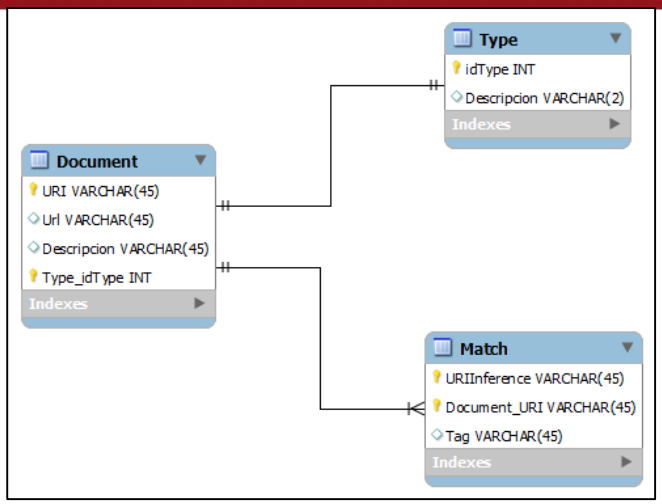


Figura 15 Modelo de persistencia de objetos

4.2 Etiquetación del corpus de documentos

Debido al dominio en el que se desarrolló el proyecto de fin de carrera, se encontró la necesidad de tener un conjunto de documentos que tengan relación al dominio. Se utilizaron enunciados de prácticas y exámenes de cursos de la carrera de ingeniería informática. En la Tabla 9, se especifican la distribución de documentos utilizados.

Curso	Ciclo	Tipo	Cantidad
Algoritmia	2013-1	PB	5
	2013-1	PE	2
	2013-1	PS	2
	2013-2	PB	5
	2013-2	PE	2
	2013-2	PS	1
Fundamentos de Programación	2013-2	PA	4
	2013-2	PE	2
Lenguajes de Programación 1	2012-1	PA	4
	2012-1	PB	5
	2012-1	PE	2
	2013-2	PA	4
	2013-2	PB	5
	2013-2	PE	2
Sistemas	2013-1	PA	4

Operativos	2013-1	PB	5
	2013-1	PE	2
	2013-2	PA	4
	2013-2	PB	5
	2013-2	PE	2

Tabla 9 Distribucion de los documentos digiales obtenidos

De manera que la data pueda seguir incrementándose, se definió un atributo tipo que se asigna a cada documento dependiendo de su contenido, estos tipos pueden ser:

- PA: indicador de que el documento es una práctica calificada.
- PB: hace referencia a una práctica de laboratorio.
- PE: se refiere a las evaluaciones tipo exámenes ya sean parciales o finales.
- PS: se refiere a solucionarios de evaluaciones proporcionado por los alumnos o por los profesores.

Dado que el proceso de etiquetación que desarrolla este proyecto es de manera manual, se podría utilizar cualquier tipo de documento ya sean PDF o alguna extensión de Microsoft Office, imágenes, etc. Sin embargo para este prototipo se ha utilizado solo documentos en formato PDF.

El proceso de etiquetado manual se realizó siguiendo las características de nuestro modelo, es decir se mantuvieron los estándares mencionados con respecto a las nomenclaturas de los documentos así también se estableció un formato de etiquetado que se representa en la Tabla 10. Este formato sirve como ayuda para luego realizar la inserción de los mismos en la base de datos. Para este proyecto se generó un procedimiento en MySQL que reciba como input el archivo en formato csv y genere los scripts en la base de datos del proyecto.

Ubicación	Curso	Tipo	Ciclo	Etiqueta
URL **ejem.** **www.coruja.com/corpus/algoritmia/2013-2/pc2**	Algoritmia	PA2	2013-2	{Etiqueta1, Etiqueta2, Etiqueta3}

Tabla 10 Esquema para el etiquetado

Como se puede observar en la Figura 16, la persona etiquetadora deberá reconocer un conjunto de palabras significativas que se encuentren en los documentos a etiquetar.

La primera restricción del modelo es que las palabras escogidas por el etiquetador deberán pertenecer al conjunto de entidades, individuos o sinónimos que se maneja en la ontología. Esto con la finalidad de que exista un match concreto entre cada documento y por lo menos 1 miembro de la ontología.

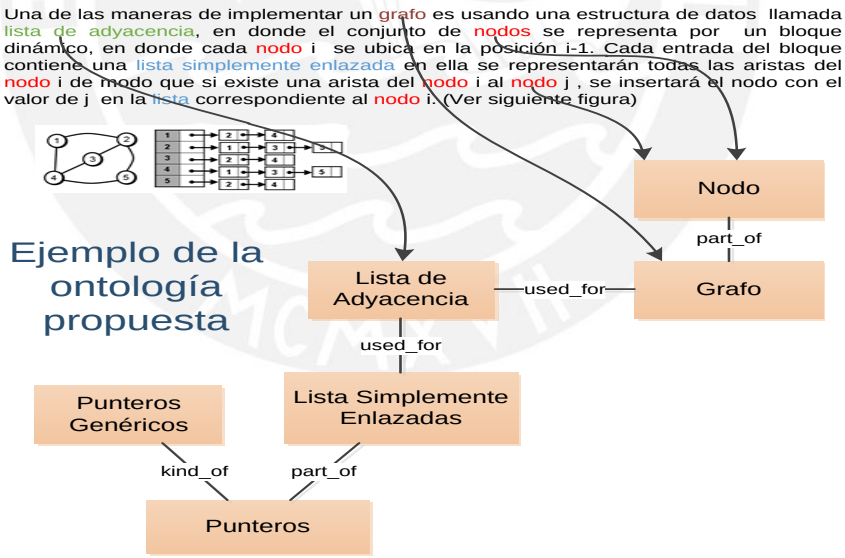


Figura 16 Ejemplo de selección de entidades (Elaborado por el autor)

4.3 Consideraciones finales

En este capítulo, se han presentado un conjunto de especificaciones en relación a la etiquetación semántica de documentos.

Cabe resaltar, que el proceso que se está utilizando en este proyecto de fin de carrera, es un proceso de etiquetación manual. Esto debido a que la data que se encuentra en el repositorio contiene todo tipo de documentos (PDF, Imágenes Escaneadas, WebSites con información) y por ello no se puede aplicar algún procesamiento semi-automático o automático.

CAPITULO V: PROCESO DE RECUPERACION DE INFORMACIÓN, RECUPERACIÓN DE CONOCIMIENTO Y DISEÑO DEL PROTOTIPO

Buscando demostrar la viabilidad del modelo planteado, en este capítulo se desarrollaran las técnicas aplicadas para recuperar información y el desarrollo de un prototipo aplicado al dominio de ingeniería informática. Esto como 3era y 4ta fase del proceso de recuperación de conocimiento planteado.

Hasta el momento, el conjunto de herramientas utilizadas en cada una de las fases nos proporciona todo lo necesario para la creación de este aplicativo. Para esta fase, se explicará la interacción entre una librería grafica para la exploración de ontologías en web como JOWL y frameworks para el manejo de documentos digitales, con la finalidad de construir un buscador de documentos para un dominio específico.

5.1 Técnicas de recuperación de documentos

La ontología se encuentra en formato RDF/XML. Este formato nos ofrece diferentes funcionalidades a la hora de realizar lecturas del mismo documento. Incluso sin la ayuda de algún framework es muy sencillo poder recuperar lo que contiene el rdf como se aprecia en la Figura 17. Son ejemplos de relaciones de entidades a manera de tripletes, si quisiéramos recuperar esto en cualquier lenguaje, podríamos utilizar un lector XML para lograrlo, y así formar una estructura que permita ser consultada, haciendo uso de las etiquetas del lenguaje de marcado.

```
<!-- http://www.semanticweb.org/hector/ontologies/2014/2/pucp#tieneConcepto -->

<owl:ObjectProperty rdf:about="http://www.semanticweb.org/hector/ontologies/2014/2/pucp#tieneConcepto"
>
    <rdfs:range rdf:resource="http://www.semanticweb.org/hector/ontologies/2014/2/pucp#Concepto"/>
    <rdfs:domain rdf:resource="http://www.semanticweb.
    org/hector/ontologies/2014/2/pucp#UnidadAprendizaje"/>
</owl:ObjectProperty>


<!-- http://www.semanticweb.org/hector/ontologies/2014/2/pucp#tieneCurso -->

<owl:ObjectProperty rdf:about="http://www.semanticweb.org/hector/ontologies/2014/2/pucp#tieneCurso">
    <rdf:type rdf:resource="&owl;TransitiveProperty"/>
    <rdfs:range rdf:resource="http://www.semanticweb.org/hector/ontologies/2014/2/pucp#Curso"/>
    <rdfs:domain rdf:resource="http://www.semanticweb.org/hector/ontologies/2014/2/pucp#Especialidad"/
    >
</owl:ObjectProperty>


<!-- http://www.semanticweb.org/hector/ontologies/2014/2/pucp#tieneEspecialidad -->

<owl:ObjectProperty rdf:about="http://www.semanticweb.
org/hector/ontologies/2014/2/pucp#tieneEspecialidad">
    <rdf:type rdf:resource="&owl;TransitiveProperty"/>
    <rdfs:range rdf:resource="http://www.semanticweb.org/hector/ontologies/2014/2/pucp#Especialidad"/>
    <rdfs:domain rdf:resource="http://www.semanticweb.org/hector/ontologies/2014/2/pucp#Facultad"/>
</owl:ObjectProperty>
```

Figura 17 Ejemplo de RDF de la ontología propuesta

Para poder potenciar el uso de la ontología se utilizó el framework JOWL que nos permitió estructurar la data en un conjunto de clases, miembros y demás propiedades. Con él se desarrollaron técnicas para mejorar la recuperación, y así permitir diferentes resultados en cuanto a precisión y escalabilidad.

Con este framework se realizaron las funciones necesarias para esquematizar las consultas que se utilizan como punto de partida en el proceso, como se puede observar en la Figura 18. Una indización de entidades, miembros y propiedades, que son representados por su URI se recuperan fácilmente con un recorrido que dependa del índice y la relación que la entidad tenga con otra.

```
jOWL.query = function(match, options){
    //console.log(options);
    options = $.extend({exclude : false}, options);
    //console.log(options);
    if(options.filter == 'Class'){ options.filter = __.owl("Class");}
    var that = this;
    //filter : [], exclude : false
    var items = new jOWL.Ontology.Array();
    var jsonobj = {};
    var test = jOWL.index("dictionary");

    function store(item){
            var include = true, i = 0;
            if(options.filter){
                if(typeof options.filter == 'string'){ include = (options.filter == item[3])
                else { for(i = 0;i<options.filter.length;i++){ if(options.filter[i] == item[
                }
            else if(options.exclude){
                include = true;
                if(typeof options.exclude == 'string'){ include = (options.exclude !== item[
                else { for(i = 0;i<options.exclude.length;i++){ if(options.exclude[i] == ite
            }
            else { include = true;}
            if(!include){ return;}
            if(!jsonobj[item[1]]){ jsonobj[item[1]] = [];}
            jsonobj[item[1]].push( { term : item[0], locale: item[2], type: item[3] });
    }
```

Figura 18 Codigo para la realizacion de consultas en la ontología

En este proceso de construcción de mecanismo, se vio la necesidad de utilizar diccionarios que nos permitan realizar filtros de entidades y palabras que guarden relación.

El primer diccionario utilizado, fue el de stopwords, que nos provee de una serie de palabras que no se deben considerar en lo queries realizado. Debido a como estaba diseñada la ontología, es decir, sin el uso de stopwords solo fue necesario filtrar algunos nombres que si contenían estas palabras.

Se usaron diccionarios que contengan todos los “Names” (propiedad de la ontología), e internamente para hacer la comparación con las cadenas de consulta, las entidades se sometieron a un proceso de tokenización. Es así que, una consulta dada “Programación”, el aplicativo puede devolver ya sea datos correspondientes al curso

de Lenguaje de Programación o Fundamentos de Programación. Sin embargo esto acarrea problemas de ambigüedades.

Para solucionar el problema de la ambigüedad, se optó por indicar al usuario el CONTEXTO en el que se está realizando la consulta. Es decir, si la consulta fuera “Arboles”, se realizaron las modificaciones adecuadas para que el aplicativo devuelva lo siguiente:

{“Algoritmia: Arboles”, “Fundamentos de Programación: Arboles”}

5.2 Presentación del prototipo

El proceso que se desea presentar con este prototipo es el de recuperación de conocimiento, el cual se presenta en la Figura 19.

- Realizar un Query: Ingresar la entidad o concepto buscado.
- Realizar el Matching: Parseo del query a una entidad de la estructura ontológica.
- Mostrar Inferencias: Visualización de las inferencias de la ontología, representadas por un visor de documentos.

Adicionalmente, se vio la necesidad de agregar un agente visual al prototipo; es decir que en el resultado final las inferencias serán representadas por documentos digitales y así finalizar el proceso de recuperación de documentos.

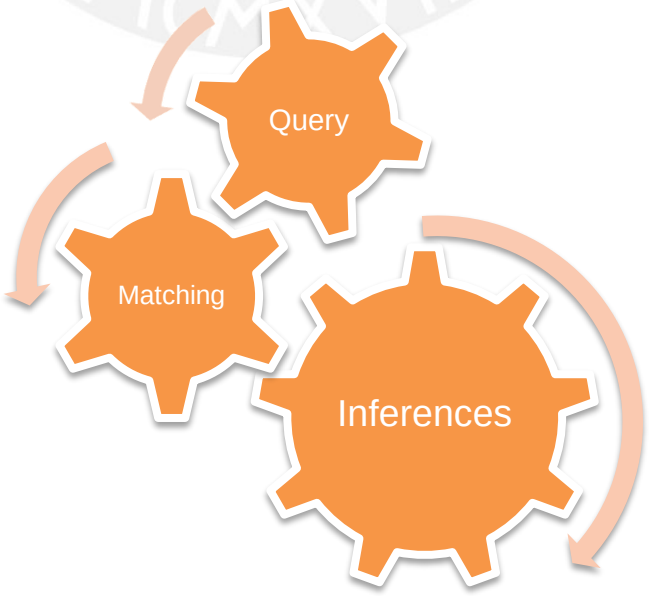


Figura 19 Proceso de recuperación de conocimiento

Una pantalla inicial del prototipo se muestra en la Figura 20.

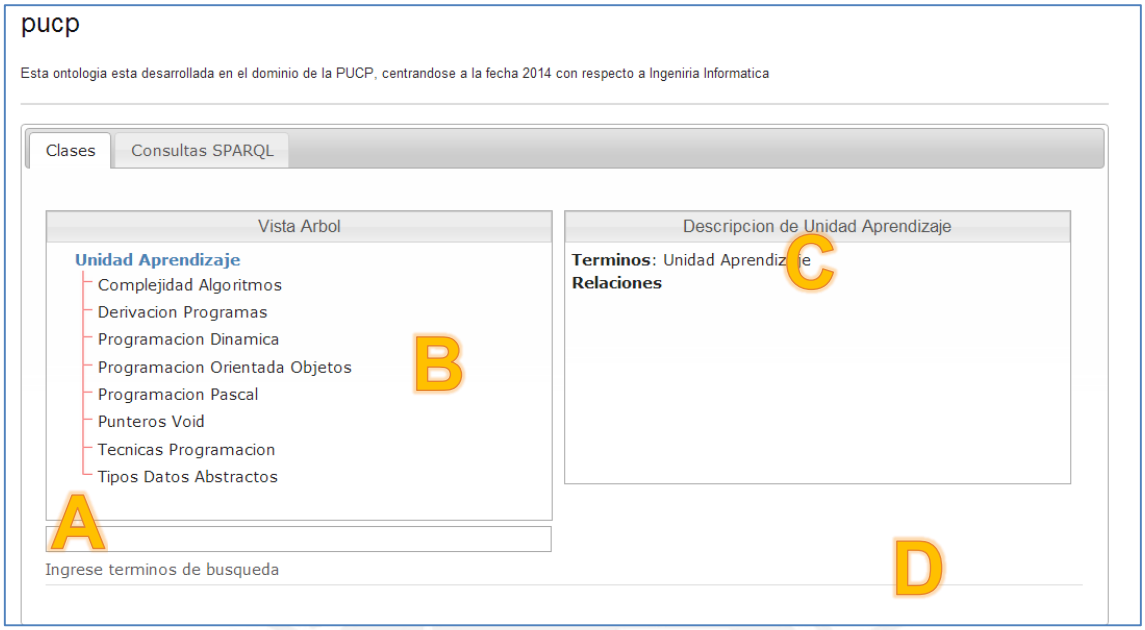


Figura 20 Pantalla inicial del prototipo

Aquí se pueden identificar 4 zonas que hacen referencia a cada una de las funcionalidades del aplicativo:

- Zona A: Es una zona de búsqueda en la que el usuario puede insertar una consulta. Para este en, particular, se mostrará una ayuda para que el usuario pueda saber que términos están disponibles en la ontología como lo muestra la Figura 21

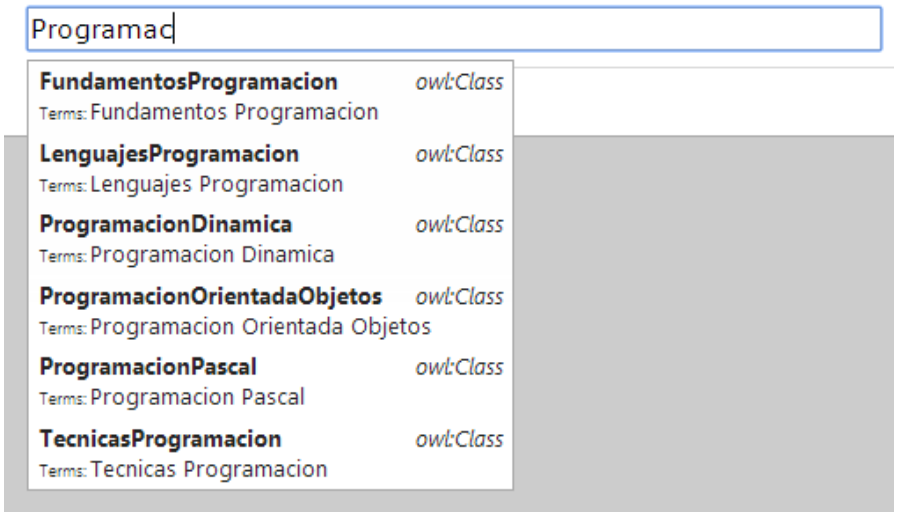


Figura 21 Ejemplo de búsqueda

- Zona B: Se refiere a la zona donde se muestra el árbol de la ontología para que el usuario, si es que así lo desee, también pueda realizar consultas desde ese medio.

- Zona C: En la zona de resumen se muestra un listado de lo concerniente a la entidad buscada o seleccionada. Incluye las propiedades que hacen referencia a la entidad y principalmente las inferencias realizadas a partir de la consulta. Es en esta zona donde se interactúa con las inferencias para poder obtener el documento requerido. Un ejemplo se aprecia en la Figura 22, para una consulta como “Tipos de datos abstractos”, que representa a la entidad UnidadAprendizaje::TiposDatosAbstractos y se listan el conjunto de conceptos relacionados a la entidad.

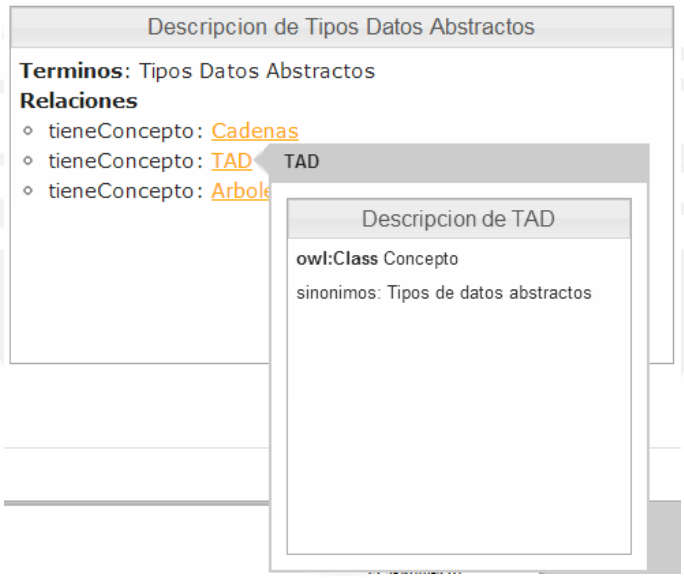


Figura 22 Ejemplo de descripción

- Zona D: Visualizador de los documentos que se obtienen a partir de las consultas

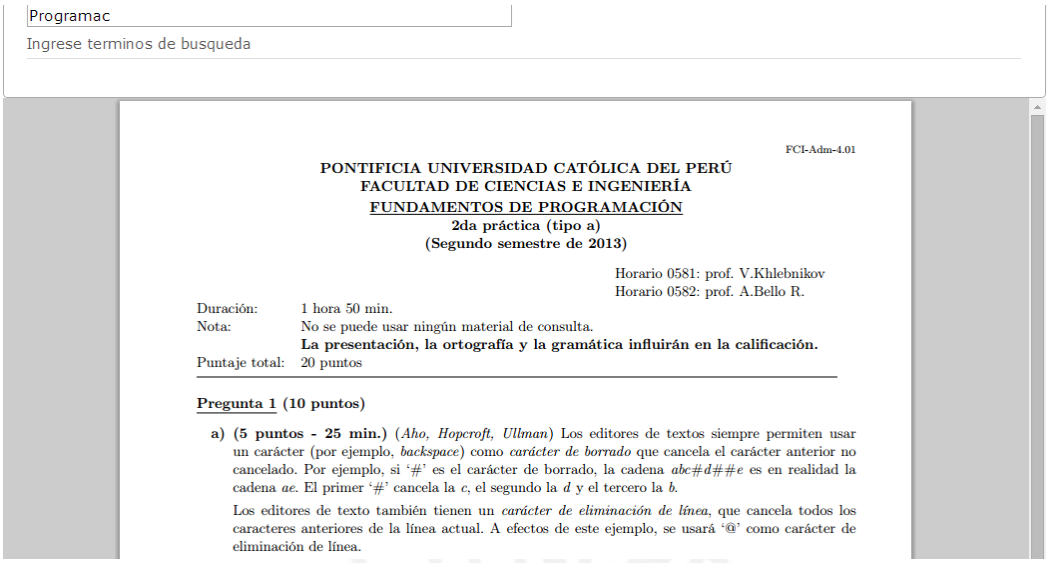


Figura 23 Ejemplo de visualización de documentos

5.3 Consideraciones finales

En este capítulo, se ha presentado las últimas fases del proceso planteado. Habiendo culminado con la implementación de las técnicas de procesamiento de datos y algoritmos de recuperación, se procedió a elaborar el prototipo que utilice todo lo antes mencionado, con la finalidad de comprobar que pueda funcionar.

Cabe mencionar algunos aspectos técnicos que el aplicativo web utilizó:

- PHP 5.0
- RDF/XML
- JOWL
- SPARQL
- Slim framework

CAPITULO VI: CONCLUSIONES

6.1 Presentación de conclusiones

Dada la investigación que se desarrolló a lo largo del proyecto de fin de carrera, se ve que los proyectos y aplicaciones para la Recuperación de Información aplicando ontologías, es bastante amplio. Sin embargo, estas herramientas tienen muchas limitaciones, entre ellas el idioma, el costo y el procesamiento.

En el análisis del dominio, se establecieron una serie de restricciones a medida que se fue desarrollando la ontología, esto con la finalidad de reducir incongruencias en la estructura, tal y como se observa en el Capítulo 3. La utilización de un corpus de documentos originales (36 documentos) fue pieza clave para realizar el proceso entero, desde buscar entidades que posiblemente se encuentren en estos documentos a plantear una estructura de persistencia que almacenen las referencias a las entidades de la ontología.

.También, se destacó la importancia de realizar un proceso de etiquetación semántico de documentos como se observa en el Capítulo 4. Se sabe que el proceso manual de etiquetación, a medida que la data incremente, se vuelve más ineficiente. Es por ello que se debe establecer una manera de distinguir los documentos de tipo PDF de las imágenes, con la finalidad de realizar con los primeros un proceso de etiquetación automático.

Debido a la gran ayuda que nos brindó la librería JOWL, con respecto al manejo de una ontología, se decidió realizar un aplicativo web que cumpla con los parámetros antes mencionados y que utilice la estructura de información que se plantea. Sin embargo, se observó que debido a que la librería fue construida hace más de 3 años, se mantiene en un formato grafico simple, por lo que no permitió agregar más funcionalidades al aplicativo.

Luego de probar el aplicativo, se observa que el modelo planteado cubre en gran medida las necesidades planteadas. Aun no se puede cumplir con realizar consultas en lenguaje natural. Debido a que el procedimiento de analisis se propuso, a manera que, limitemos al usuario a utilizar consultas que directamente sean entidades de la ontología, como se explica en el Capítulo 4.

6.2 Trabajos Futuros

Debido al alcance del proyecto se pueden realizar las siguientes variaciones:

- *Ampliar la cantidad de documentos en el corpus,* Abarcar una gran cantidad de facultades y demás especialidades, ayudarían a probar diferentes maneras de desambiguar entidades. Por ejemplo, si se añadiera las especialidades de la facultad de Ciencias y Artes de la comunicación, tendríamos muchos ejemplos sobre ambigüedades para una búsqueda como “Información”.

- *Mejorar el método de parseo para poder recibir lenguaje natural,* dado la restricción que se estableció, sobre ingresar palabras que parseen a entidades de la ontología. Esto no permite realizar consultas que sean de mucho más ayuda para el usuario. Como por ejemplo, “arboles, con nota mayor a 15”, esto último debería parsear a una entidad y a una de sus propiedades. Es decir, parsear a algo parecido a esto: ”Algoritmia::Arboles::Nota::>15”. Así poder devolver todas la inferencias a prácticas o exámenes donde el tema a tratar sea árboles y que su nota sea mayor a 15, todo con la ayuda de PLN.

- *Mejorar la librería JOWL, s*i bien nos permitió un sin número de aplicativos, existen diferentes mejoras que, para un desarrollo futuro servirían de mucho, como por ejemplo: Mejorar el entorno grafico utilizando una versión más actualizada de jquery y añadir funcionalidades que permitan al usuario desenvolverse mejor en el aplicativo.

- *Realizar los ajustes para poder aplicar etiquetación automática,* Como ya se sabe, la cantidad de documentos en el repositorio se va incrementando cada vez más. Esto hará que el método propuesto solo sea útil para aquellos documentos que NO tienen alguna manera de soportar una etiquetación semántica. Ya que estas no se pueden convertir en texto puro, con una precisión muy alta como Imágenes de solucionarios, links a teoría de clases, audios, etc.

  Los documentos en formato PDF puro, si podrían soportar etiquetación automática, realizar con más rapidez la actualización de la base de datos.

BIBLIOGRAFÍA

Bauzer Medeiros, C., & Daltio, J. (2008). *Aond^e: An Ontology Web Service for Interoperability across.* Obtenido de http://www.lis.ic.unicamp.br/projects/webios/aonde-an-ontology-web-service

Berners-Lee, T. (2000). *World Wide Web Consortium: XML and The Web.* Obtenido de http://www.w3.org/2000/Talks/0906-xmlweb-tbl/

Berners-Lee, T. (2001). The Semantic Web. *Scientific American, 1*(1).

Blaxter, L., Hughes, C., & Tight, M. (2006). *How to research.* England: Open University Press.

Bunge, M. (1977). *Treatise on Basic Philosophy: Ontology I".* Reidel.

Bunge, M. (1979). *Treatise on Basic Philosophy: Ontology II.* Reidel.

Bunge, M. A. (1997). Mechanism and explanation. *Philosophy of the Social Sciences, 27*(4), 410.

Bunge, M. A. (2003). *Emergence and convergence: Qualitative novelty and the unity of knowledge.* Canada: Universidad de Toronto Pr.

Coyle, K. (18 de 05 de 2009). *Dublin Core Metadata Initiative.* Obtenido de http://dublincore.org/

DCMI. (2007). *DCMI Metadata Basics*. (DCMI) Recuperado el 08 de 05 de 2013, de http://dublincore.org/metadata-basics/

Gilchrist, A. (2003). Thesauri, taxonomies and ontologies-an etymological note. *Journal of Documentation, 59*(1), 7-18.

Greengrass, E. (30 November 2000). *Information Retrieval: A Survey.*

Gruber, T. (1993). Toward Principles for the Design of Ontologies Used for Knowledge Sharing. Technical Report. *Knowledge Systems Laboratory, Stanford University*.

Haverkamp, D., & Gauch, S. (1998). Intelligent information agents: Review and challenges for distributed information sources. *Journal of the American Society for Information Science, 49*(4), 304-311.

Hou, J., & Pai, S. (2009). A spatial knowledge sharing platform using the visualization approach. .*International Journal of Production Research, 47*(1), 25-50.

Kiryakov, A., Popov, B., Terziev, I., & Ognyanoff, D. (2004). Semantic annotation, indexing, and retrieval. *Journal of Web Semantics, 2*(1), 49-79.

Kitchenham, B. (2004). *Procedures for Performing Systematic Reviews.* Australia: Software Engineering Group, Department of Computer Science, Keele University and Empirical Software Engineering National ICT.

Lytras, M., & Sicilia, M. (2005). The knowledge society: a manifesto for knowledge and learning. *International Journal of Knowledge and Learning, 1*(1), 1-11.

Mariano Crespo, J. M., & Maña, M. J. (2011). Using Ontologies for Query Expansion in Image Retrieval in the Biomedical. *Procesamiento del Lenguaje Natural*(47), 39-46.

Melgar Sasieta, H. A. (2011). *Un modelo para la visualización de conocimiento basado en imagenes semánticas.* Brasil: Tesis para optar el grado de Doctor en Ingeniería y Gestion del Conocimiento: Universidad Federal de Santa Catarina.

Melgar, A., & Pacheco, R. (2010). La Ingeniería del Conocimiento en la Sociedad del Conocimiento. *The Knowledge Engineering in the Knowledge Society, 1*(1).

Mitra, M. (2000). Information Retrieval from Documents: A Survey. *Information Retrieval, 2*(1), 141-163.

Ordoñez , H., & Cobos, C. (2010). OntoGhobi - Meta Buscador Semántico Que Incorpora. *Grupo de I+D en Tecnologías de la Información, 1*(1), 1-10.

Oren, E., Moller, K., Scerri, S., Handschuh, S., & Sintek, M. (2006). What are Semantic Annotations? *Digital Enterprise Research Institute, National University of Ireland, Galway, 1*(1).

Oro, E., & Ruffolo, M. (2008). XONTO: An Ontology-based System for Semantic Information Extraction from. *20th IEEE International Conference on Tools with Artificial Intelligence, 1*(1), 1-8.

Pease, A. (2002). IEEE Standar Upper Ontology: A progress Report. *17*(1), 65-70.

Perea Ortega, J. M., Montejo Ráaez, A., Díaaz Galiano, M. C., & García Cumbreras, M. A. (2011). Analysis of the query expansion for medical collections using mutual. *Procesamiento del Lenguaje Natural*(47), 13-20.

Petticrew, M., & Roberts, H. (2005). *Systematic reviews in the social sciences: A practical guide.* Blackwell Publishing.

Rujiang, B., & Junhua, L. (2009). Improving Documents Classification with Semantic Features. *2009 Second International Symposium on Electronic Commerce and Security, 1*(1), 640-643.

Schreiber, A. T. (2000). *Knowledge Engineering and Managment - The CommonKADS Methodology.* Massachusets: Library of Congress Cataloging-in-Publication Data.

Stewart, T. (2001). *The Wealth of Knowledge: Intellectual Capital and the Twenty-1rst Century Organization.* [S.l.]:Doubleday.

Talbot, D. (s.f.). *MIT Tecnology Review*. Recuperado el 06 de 06 de 2013, de http://www.technologyreview.com/news/413358/wolfram-alpha-and-google-face-off/?nlid=2001

The Cochrane Collaboration. (2011). *Cochrane Handbook for Systematic Reviews of Interventions.* Recuperado el 05 de 06 de 2013, de http://handbook.cochrane.org/

U.S. National Library of Medicine. (s.f.). *Medical Subject Heading*. Recuperado el 06 de 06 de 2013, de U.S. National Library of Medicine

University of Maryland at College Park. (s.f.). *SHOE*. Obtenido de http://www.cs.umd.edu/projects/plus/SHOE/#team

W3C. (s.f.). *W3C Recomendations*. Recuperado el 05 de Junio de 2013, de http://www.w3.org/2001/sw/

W3C. (s.f.). *W3C-Owl Reference.* Recuperado el 05 de Junio de 2013, de http://www.w3.org/TR/owl-ref/#Intro

Weibel, & Lagoze. (1997). *Dublin Core Metadata for simple resource description.*

Wolfram Research. (s.f.). *Wolfram Alpha*. Recuperado el 06 de 06 de 2013, de http://www.wolframalpha.com/faqs9.html

Woodley, D. -M. (2005). *DCMI Glosary*. (DCMI) Recuperado el 8 de 05 de 2013, de http://dublincore.org/documents/usageguide/glossary.shtml#M

Zachman, J. A. (2003). *The Zachman Framework for enterprise architecture : Primer for enterprise engineering and manufacturing.*

Zachman, J. (s.f.). *Zachman Internationals*. Recuperado el 05 de Junio de 2013, de http://www.zachman.com/about-the-zachman-framework