Metadatos En Documentos Digitalizados A Través Del Uso De Ontologías (Metadoc)

OSWALDO COTES SOLANO

Informe final del Trabajo de Investigación presentado como requisito para obtener el título Magíster en Ingeniería: **Área Informática Y Ciencias de la Computación**.

Director: Sergio Castillo Castelblanco, Ph. D.

Universidad Industrial de Santander

Escuela de Ingeniería de Sistemas e Informática

Maestría en Ingeniería

Bucaramanga

2006

Nota Aceptación

AGRADECIMIENTOS

- A todas las personas que de manera directa o indirecta participaron en la elaboración de este proyecto.
- A mi mama y hermanos porque sin ellos hubiera sido imposible la culminación de este trabajo
- A mi novia, compañera en las incansables jornadas.

TABLA DE CONTENIDO

LISTA DE FIGURAS

RESUMEN

TITULO: METADATOS EN DOCUMENTOS DIGITALIZADOS A TRAVÉS DEL USO DE ONTOLOGÍAS (METADOC) *

Autor: Oswaldo Alexander Cotes Solano.**

Palabras claves: Ontologías, Metadatos, anotaciones semánticas

Descripción

En el contexto de este trabajo el uso de ontologías se define como la extensión de los actuales sistemas de gestión del conocimiento cuyo objetivo es que no sólo los humanos, sino también las máquinas, sean capaces de "comprender" el contenido de los documentos. Para que este objetivo sea posible es necesario proveer de mecanismos y tecnologías con las que poder definir la semántica de los documentos, meta información que posteriormente podrá ser utilizada por agentes y buscadores inteligentes con el fin de ofrecer resultados precisos y contextualizados. Las ontologías hacen posible una semántica para construir infraestructuras de meta datos, permitiendo no solo almacenar la información, sino también poder buscarla y recuperarla. Define los términos y las relaciones básicas para la compresión de un área del conocimiento, así como las reglas para poder combinar los términos para definir las extensiones de este tipo de vocabulario controlado. Este trabajo describe un caso de estudio de cómo las tecnologías de Web semántica pueden ser usadas para anotar y recuperar información en documentos estructurados. Se presenta una ontología de documentos y una herramienta automática para la construcción anotaciones y recuperación de información utilizando las API de OWL y JENA usando como escenario la Universidad Industrial de Santander.

**Facultad de Ingenierías Físico-Mecánicas. Escuela de Ingeniería de Sistemas e Informática. Maestría en Ingenierías, área informática y ciencias de la computación. Director: Sergio Castillo Castelbalnco Ph.D.

SUMMARY

TITLE: METADATA IN DIGITALIZED DOCUMENTS USING ONTOLOGIES (METADOC) *

Author: Oswaldo Alexander Cotes Solano.**

Key words: Ontologiesn, Metadata, Semantics Annotation

Description

In the context of this work the use of ontologies defines like the extension of the present-day systems of steps of knowledge whose objective itself he is than no only humans, but also hardware, be capable to understand the contents of documents. In order that this objective be possible it is necessary to provide with mechanisms and technologies with them that to be able to define the semantics of documents, goal information that at a later time will be able to be used by agents and intelligent seekers with the aim of offering precise results. The ontologies make semantics to forge infrastructures of goal possible data, permitting not only to store the information, but also to be able to look for her and to recover her. Define the terms and basic relations for the compression of an area of knowledge, as well as rules to be able to combine the terms to define the suchlike extensions of controlled vocabulary. This work describes a case study of how the technologies of semantic Web can be used to note down and to recover information in structured documents. Ontology of documents and an automatic tool for the construction encounters notes and information retrieval utilizing them API of OWL and JENA, using like scenario the Universidad Industrial De Santander

** Facultad de Ingenierías Físico-Mecánicas. Escuela de Ingeniería de Sistemas e Informática. Maestría en Ingenierías, área informática y ciencias de la computación. Director: Sergio Castillo Castelblanco Ph.D.

INTRODUCCIÓN

Toda organización conserva su memoria en los documentos producidos en su diaria operación. Cada fuente escrita que se genera es fundamental en el proceso de toma de decisiones. Como consecuencia de esta actividad, el área de la informática ha venido desarrollando muchas técnicas para la recuperación de la información que está contenida en los documentos.

El modo de operación de los sistemas de consulta y recuperación de información se basa en guardar solamente los datos clave del documento, y el grueso del texto se deja para que los motores de búsqueda filtren el texto usando comparación binaria de caracteres y operadores lógicos de concatenación. Esta característica hace que la búsqueda de información que no es clave dentro del documento arroje demasiadas coincidencias y que se pierda tiempo por parte del usuario al buscar el documento requerido.

Para garantizar que la información presente en un documento pueda ser buscada utilizando el significado del texto que la compone , Se propone convertir la información en conocimiento mediante unas estructuras formalizadas (las ontologías) que referencien los datos, por medio de metadatos, bajo un esquema

común normalizado sobre algún dominio del conocimiento. Los metadatos no sólo especificarán el esquema de datos que debe aparecer en cada instancia, sino que también podrán contener información adicional de cómo hacer deducciones sobre ellos.

1. PANORAMA GENERAL DEL PROYECTO

Este libro trata del proyecto de maestría titulado “Metadatos En Documentos Digitalizados A Través Del Uso De Ontologías (Metadoc)”, desarrollado dentro del marco investigativo del grupo GITSI, en el área de Web Semántica. En este capítulo se hace un bosquejo del trabajo realizado, partiendo de la descripción del plan de proyecto, los objetivos, la justificación, Estado del arte, las etapas de desarrollo y los logros obtenidos. Al final se presenta una descripción general de la organización y temática de los demás capítulos de este texto.

1.1. DESCRIPCIÓN DEL PROBLEMA DE INVESTIGACIÓN

Respaldar los documentos utilizando medios digitales sigue siendo la mejor de las opciones en el mercado. Recuperar de manera ágil y sobre todo correcta la información respaldada asegura el éxito de las instituciones, en la manera en que aseguran su acervo documental.

Recuperar esa información utilizando solo los datos claves al “bautizar” el documento y el texto que se ha reconocido implica que los requerimientos de búsqueda deben ir de acuerdo a lo que el sistema ha declarado como sus

parámetros de consulta. Que pasa si cierta persona quiere buscar un literal que no se encuentra dentro de la información clave sino en cierta zona del documento, las ocurrencias de este literal serán muchas debido a que sin importar donde se encuentre el motor de búsqueda del software solo se limitara a una comparación plana de caracteres.

Permitir al buscador “leer” y extraer la información utilizando la anotación semántica como guía en el proceso de consulta permite no solo asegurar que la búsqueda arroje pocos y precisos resultados sino que el respaldo sea también inteligente.

1.1.1. Planteamiento Del Problema

El auge de la informática y las grandes prestaciones que los computadores han alcanzado en el transcurso de finales del siglo XX y principios del siglo actual, ha modernizado la forma como las instituciones manejan el acervo documental que representa su historia y tradición en el mundo empresarial. Desde la automatización de los procesos de manejo de archivos hasta el uso de scanners de alta gama para la digitalización de documentos impresos, las ciencias de la computación han estado presentes aprovechando la analogía de un sistema de archivos con el concepto de sistema en informática.

El advenimiento de gran cantidad de paquetes informáticos en donde la finalidad principal es **Digitalizar** y **Guardar** documentos dentro de una base de datos robusta que logre mantener fiel copia de los documentos impresos originales proporciona una solución efectiva como respaldo pero no tan efectiva a la hora de ejecutar una búsqueda exitosa, es como decir ¡Señor Usuario encontré lo que buscaba, está en el planeta tierra! ¿Pero que es una búsqueda exitosa? Una búsqueda exitosa se traduce en encontrar, sino exactamente lo que se esta buscando, un cúmulo reducido de respuestas en las cuales este la respuesta exacta, se debe tener en cuenta que no son 10 papeles digitalizados sino que se puede hablar fácilmente de un millón o dos millones de documentos dentro de una institución media. De ahí la necesidad de otorgar a la maquina la capacidad de no solamente comparar literales sino que pueda “comprender” la búsqueda.

Se podría decir que los sistemas actuales incorporan buscadores eficientes en donde el uso de una alta gama de operadores proporciona al usuario una capacidad de filtrado de documentos excelente. Desafortunadamente el problema se mantiene, ya que el inconveniente no es filtrar la información sino buscar en la parte exacta donde muy factiblemente se encuentre la intención de conocimiento de la cadena de búsqueda. En este caso no solo la información resultante es equivoca sino que el tiempo empleado en buscar en otras zonas del documento donde es muy seguro que la

equivalencia de caracteres se encuentra pero el conocimiento que se desea extractar está lejos de esa ubicación.

Empezar a mirar la búsqueda dentro de un documento como una operación inteligente de conocimiento mas no de simple comparación de literales, favorece la interoperabilidad de las aplicaciones basadas en la digitalización de documentos. Estandarizar los modelos de conocimientos usados en la elaboración de los documentos facilitaría la comunicación entre agentes documentales.

Para la Universidad Industrial De Santander, institución pública dedicada a generar conocimiento es de una utilidad inmensa proteger y salvaguardar la memoria institucional presente en sus documentos, por consiguiente, se hace indispensable no solo respaldar la documentación en papel en medios digitales, sino, encontrar la forma de otorgar una organización estructural de contenido a cada documento incorporando dentro de cada uno, los metadatos o anotaciones que expresen que contienen y permita que se pueda encontrar en el momento justo de su requerimiento.

1.2. JUSTIFICACIÓN

Encontrar exactamente lo que se busca es siempre la mejor justificación de cualquier proyecto, literalmente la aplicación del modelo de datos basado en ontologías en la búsqueda de documentos que han sido digitalizados asegura que mediante la estructuración de la información contenida en los documentos se logre éxito en las consultas, es decir, no solamente el lector humano la entiende sino que la maquina puede “entender” de lo que trata el texto mas allá de tratarlo como simples caracteres. Lo dicho anteriormente aceleraría y haría efectiva una búsqueda por muy vaga que la información inicial sea, porque la maquina estaría en capacidad de “leer” y extractar la información coincidente, proceso más efectivo que encontrar coincidencias numéricas o lógicas sin conocer la estructura del texto.

1.3. OBJETIVOS

1.3.1. Objetivo General

Aplicar el modelo de datos basado en ontologías para la incorporación de metadatos dentro de los documentos digitalizados en la Universidad Industrial De Santander.

1.3.2. Objetivos Específicos

- Analizar los sistemas de clasificación, organización de archivos y perfiles de acceso a documentos actualmente utilizados en la Universidad Industrial De Santander.
- Estructurar el conocimiento contenido en la clasificación actual de documentos a través del uso del modelo de datos basado en ontologías.
- Desarrollar un prototipo software que permita el uso de anotaciones y búsqueda inteligente dentro del modelo ontológico de metadatos de cada documento.

1.4. METODOLOGÍA DE DESARROLLO

Para el cumplimiento de los objetivos planteados en este proyecto se diseñaron las siguientes fases.

Fase 1. Estudio de la literatura

En esta etapa se investigó acerca de las nuevas investigaciones y desarrollos que en la actualidad la comunidad científica ha presentado o implementado. Se analizaron las últimas herramientas disponibles para la

creación de ontologías y anotaciones en documentos digitales, así como, la actualidad en cuanto a los lenguajes de Web Semántica. En este punto del proyecto toda la información y herramientas estudiadas servirán para conocer de cerca la filosofía de la llamada “Web Del Conocimiento”, de igual manera, la ampliación del estudio de estas técnicas proporcionaran un acercamiento real y no ideal sobre la verdadera aplicación de la Web Semántica.

Fase 2. Análisis del sistema de digitalización de documentos actual en la institución

Se analizó la situación local actual en cuanto los sistemas utilizados para digitalizar documentos, en esta fase se estudiaron los procesos de digitalización, clasificación, y organización que el sistema otorga a los documentos. Se seleccionó una oficina de la institución donde se utilizaba algún sistema de estas características, dicho análisis se centró en que clase de documentos se manejan dentro de la oficina para poder empezar a generar el modelo de conocimiento que tiene cada tipo de documento.

En esta etapa se diseño la ontología que conceptualiza los documentos estructurados que la oficina produce. Se observó las funciones de búsqueda de documentos que el sistema actual este empleando, para determinar que se busca dentro de una oficina, y poder generar una ontología orientada hacia la búsqueda.

Fase 3. Desarrollo de prototipo software

Se desarrolló un prototipo software, que utiliza la ontología desarrollada como modelo de datos para la generación automática de anotaciones dentro de los documentos. El prototipo incorporo un modulo de recuperación de información después del marcado, que explora la estructura de cada anotación seleccionando la información requerida como criterio de búsqueda. Para probar la evolución del prototipo se utilizo información original de la oficina seleccionada, tanto en el proceso de anotación, como en el proceso de búsqueda.

Fase Documentación

Se elaboró el informe final y del artículo de investigación, donde se consignan las experiencias desarrolladas, los conocimientos adquiridos y demás información obtenida a lo largo del trabajo de investigación.

1.5. LOGROS

El proyecto también realizó un aporte a la academia, al aplicar los conceptos de Web Semántica en una problemática real planteada dentro de la institución. Los resultados del proyecto ofrecen bases teóricas confiables del tema que permitirán

plantear trabajos posteriores que mantengan la línea de investigación activa aportándole nuevos conocimientos.

A la culminación de este proyecto, se pueden destacar los siguientes aspectos:

- ✓ Se fortaleció la línea de investigación de Web semántica y recuperación de información, perteneciente al grupo de Telemática y sistemas inteligentes.

- ✓ Se escribió un artículo para la revista de la facultad de Ingenierías Físico mecánicas de la UIS

- ✓ Se dejó el camino abierto para la realización de proyectos que involucren Web semántica y recuperación de información.

1.6. BOSQUEJO DEL PRESENTE TEXTO

En el capítulo 2 se expone el estado del arte en relación a la utilización de ontologías como herramientas para la estructuración y recuperación de información en fuentes digitales. También se explican las bases legales en las que se soporta el trabajo de investigación.

En el capítulo 3 se describe el esquema de metadatos y el modelo ontológico utilizado para la herramienta de anotación, se realiza una descripción de las clases

y las propiedades que conforman el dominio documental de la oficina seleccionada.

El capítulo 4 describe la forma como se desarrolló el prototipo, partiendo del análisis del sistema actual, diseño de datos y procesos, plataformas y metodología de programación.

Finalmente se presentan las conclusiones y recomendaciones.

2. ESTADO DEL ARTE

En este capítulo se presentan los fundamentos técnicos derivados del estado del arte, sobre los que se basa este trabajo.

2.1. TECNOLOGIAS DE WEB SEMANTICA

2.1.1. Xstensible Markup Language *XML*

El Lenguaje Extensible de Marcas, abreviado XML, describe una clase de objetos de datos llamados documentos XML y describe parcialmente el comportamiento de los programas de computadora que los procesan.

Los documentos XML están compuestos por unidades de almacenamiento llamadas entidades, que contienen tanto datos analizados como no analizados. Los datos analizados están compuestos de caracteres, algunos de los cuales, de la forma datos carácter, y otros de la forma marca. Las marcas codifican una descripción de la estructura de almacenamiento del documento y su estructura lógica. XML proporciona un mecanismo para imponer restricciones al almacenamiento y a la estructura lógica.

Se utiliza un módulo software llamado **procesador XML** para leer documentos XML y proporcionar acceso a su contenido y estructura. Se

asume que un procesador XML hace su trabajo dentro de otro módulo, llamado **aplicación**. Esta especificación describe el comportamiento requerido de un procesador XML en términos de cómo leer datos XML y la información que debe proporcionar a la aplicación.

2.1.2. Resource Description Framework RDF

RDF es uno de los lenguajes desarrollados por el *World Wide Web Consortium* (W3C), que permite la representación de metadatos en la Internet. Facilita la descripción de recursos en la Web a través de un modelo de datos de grafos etiquetados y dirigidos donde el orden no es relevante.

El objetivo general de RDF es definir un mecanismo para describir recursos que no cree ninguna asunción sobre un dominio de aplicación particular, ni defina (a priori) la semántica de algún dominio de aplicación. La definición del mecanismo debe ser neutral con respecto al dominio, sin embargo el mecanismo debe ser adecuado para describir información sobre cualquier dominio.

El modelo de datos RDF utiliza los siguientes componentes para la creación de infraestructuras de descripción de recursos:

- **Recursos:** Todas las cosas descritas por expresiones RDF se denominan *recursos*. Un recurso puede ser una página Web completa; tal como el documento HTML "http://www.w3.org/Overview.html" por ejemplo. Un recurso puede ser una parte de una página Web; p. ej. un elemento HTML o XML específico dentro del documento fuente. Un recurso puede ser también una colección completa de páginas; p. ej. un sitio Web completo. Un recurso puede ser también un objeto que no sea directamente accesible vía Web, p. ej. un libro impreso. Los recursos se designan siempre por URIs más identificadores de anclas opcionales. Cualquier cosa puede tener un URI; la extensibilidad de URIs permite la introducción de identificadores para cualquier entidad imaginable.

- **Propiedades:** Una *propiedad* es un aspecto específico, característica, atributo, o relación utilizado para describir un recurso. Cada propiedad tiene un significado específico, define sus valores permitidos, los tipos de recursos que puede describir, y sus relaciones con otras propiedades.

- **Sentencia:** Un recurso específico junto con una propiedad denominada, más el valor de dicha propiedad para ese recurso es una *sentencia RDF* [RDF statement]. Estas tres partes individuales de una sentencia se denominan, respectivamente, sujeto, predicado

y objeto. El objeto de una sentencia (es decir, el valor de la propiedad) puede ser otro recurso o pude ser un literal; es decir, un recurso (especificado por un URI) o una cadena simple de caracteres [string]

Considérese la siguiente frase, *Secretaria General UIS es la creadora del recurso http://www.uis.edu.co/sec/docs/acu001.html,* Gráficamente el modelo de datos RDF expresa esta sentencia de la siguiente manera:

Fig 1. Representación Grafica De Una Sentencia RDF

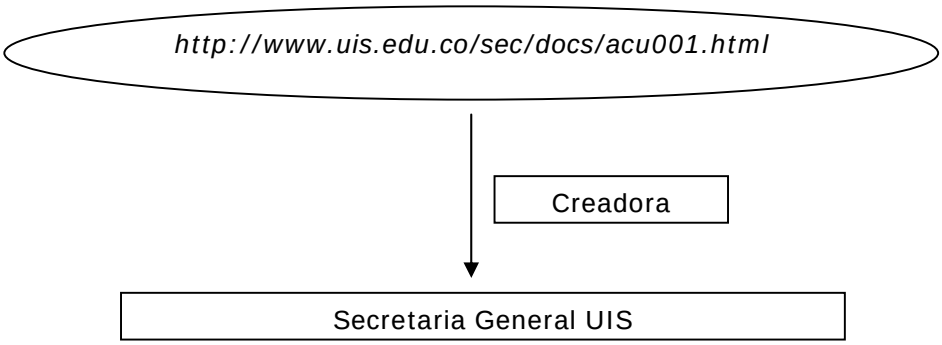


Tabla 1. Descripción De Una Sentencia RDF

Sujeto(Recurso)	*http://www.uis.edu.co/sec/docs/acu001.html*
Predicado(Propiedad)	*creadora*

Objeto(Literal)	*Secretaria General UIS*

La Figura 1 describe la representación gráfica de una sentencia RDF. Usa grafos etiquetados (también denominados "diagramas de nodos y arcos"). En estos grafos, los nodos (dibujados como óvalos) representan recursos y los arcos representan propiedades denominadas. Los nodos que representan cadenas de literales pueden dibujarse como rectángulos.

La sintaxis abreviada RDF/XML proporciona la descripción adecuada del recurso, en la Fig. 2 se observa el codigo que formaliza el grafo dirigido expuesto en la Fig. 1.

Fig 2. Sintaxis RDF/XML Abbrev

```
<rdf:RDF>
  <rdf:Description about="http://www.uis.edu.co/sec/docs/acu001.html">
    <s:Creator>Secretaria General UIS</s:Creator>
  </rdf:Description>
</rdf:RDF>
```

2.1.3. Resource Description Framework Schema RDF-S

El esquema de RDF denominado RDF-S, describe como usar RDF para describir vocabularios. Además, la especificación define un vocabulario

para este propósito y menciona otros vocabularios construidos inicialmente en RDF.

En el caso de RDF es fundamental utilizar palabras que transmitan un significado inequívoco con el fin de que las aplicaciones entiendan el enunciado para un procesamiento correcto. En RDF, este significado se expresa a través de un esquema. Podemos pensar en un esquema como una especie de diccionario que define los términos que se utilizarán en una declaración o sentencia RDF para otorgarle significados específicos. Con RDF se pueden utilizar una gran variedad de formas de esquema, incluyendo la definida en RDFSchema que posee unas características especiales para automatizar tareas utilizando RDF, pero también otras muchas formas.

Para la declaración de clases y propiedades del dominio en estudio los esquemas RDF se valen del siguiente vocabulario:

- **rdfs::Resource**: Todas las cosas que se describan por expresiones RDF se denominan recursos *(resources)*, y se consideran como *instances* (objetos específicos de la categoría) de la clase **rdfs:Resource**.

- **rdfs::Property**: representa el subconjunto de recursos RDF que son propiedades, es decir, todos los elementos del conjunto presentados como propiedades.

- **rdfs::Class**: corresponde con el concepto genérico de un tipo *(Type)* o categoría *(Category)*, semejante a la noción de Clase en los lenguajes de programación orientados a objetos tales como Java. Cuando un esquema define una nueva clase, el recurso que representa esa clase debe tener una propiedad **rdf:type** cuyo valor es el recurso **rdfs:Class**. Las clases RDF pueden definirse para representar cualquier cosa, como páginas *web*, personas, tipos de documentos, bases de datos o conceptos abstractos.

2.1.4. Web Ontology Language OWL

Uno de los desarrollos mas recientes del W3C World Wide Web Consortium) es el OWL acrónimo de Web Ontology Language, OWL esta basado en un modelo lógico diferente permite definir los conceptos complejos en definiciones simples.

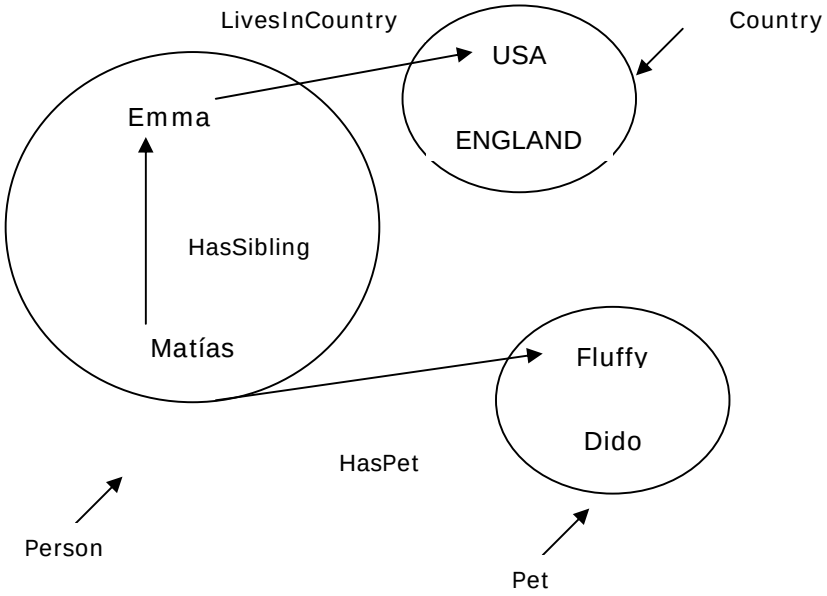
OWL esta diseñado para el uso de aplicaciones que necesiten procesar e interpretar información. Garantiza alta interpretabilidad de contenidos por parte de las máquinas que soportan XML, RDF, RDF-S. Lenguajes que aportan semántica al contenido de recursos Web.

Una ontología OWL se compone de:

- Individuos: representan los objetos de nuestro dominio de interés.
- Propiedades: Son las relaciones que se dan entre los individuos.
- Clases: Grupos que contienen individuos, representan los conceptos de nuestro dominio.

La representación de un dominio en OWL se describe en la Fig. 3, describe las clases, propiedades e individuos que actúan en un entorno personal.

Fig 3. Representacion de domínio en OWL



2.1.5. Ontologías en la descripción de contenidos documentales

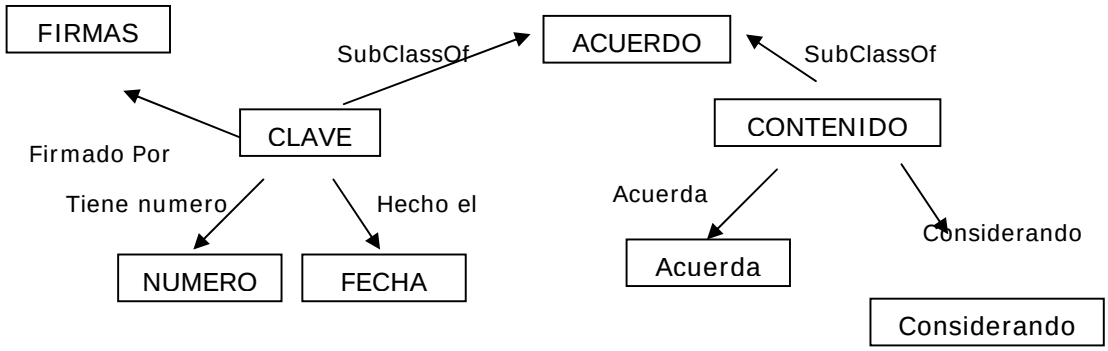
Las ontologías son un elemento importante en la Web Semántica, y es un término filosófico que trata sobre la naturaleza de la existencia y los tipos de objetos que existen [1]. También, la ontología se puede definir como una especificación explícita de una conceptualización compartida [2].

Actualmente, se les está dando más uso a las ontologías en el campo de la Inteligencia Artificial. En este campo se define la ontología como una descripción formal y explícita de conceptos de un dominio (*classes,* a las que nos referiremos como conceptos), propiedades de cada concepto, las cuales describen las características y atributos de los conceptos (*slots* llamados también como propiedades o roles) y restricciones de las propiedades (*facets* conocidas como restricciones) [3].

Con el uso de modelos de bases de datos relacionales, los documentos son organizados y clasificados utilizando relaciones de pertenencia, entre el contenido del documento y los campos claves que el diseñador de datos ha escogido para identificarlo. El uso de ontologías describe el contenido del texto teniendo en cuenta la semántica del texto y no la relación que haya con algún vínculo externo; dándole al buscador la oportunidad de “leer” el texto y filtrar no por tipos sino por contenidos.

Para lograr su objetivo la ontología debe describir eficazmente de que trata el documento, la descripción se logra usando anotaciones que seccionen el documento en partes que se relacionan entre si de forma explicita informando que contienen y que tan importante son. En la Fig. 4 se propone una ontología para describir un acuerdo generado en la Secretaria General de la Universidad Industrial de Santander.

Fig 4. Prototipo de ontología para describir acuerdos.



La ontología propuesta en la Fig. 4 describe el conocimiento del acuerdo, los rectángulos son los conceptos que están presentes en su contenido, las flechas indican las propiedades o roles de los conceptos, nótese que el sentido de la flecha indica la prelación de la relación.

2.2. BASES LEGALES

El presente trabajo de investigación se fundamenta en el siguiente cuerpo normativo:

Constitución Política de Colombia: Son pilares de normatividad aplicada a los documentos y el fundamento de la política archivística en el país, los artículos de la constitución de 1991 detallados a continuación:

Art. 8: Obligación del estado y de las personas de proteger las riquezas culturales de la nación.

Art. 15: Derecho a la intimidad personal, familiar y al buen nombre, actualizar y rectificar las informaciones que se tengan en los archivos de entidades públicas y privadas.

Art. 20: Derecho a recibir información veraz e imparcial.

Art. 23: Derecho de petición.

Art. 70: Obligación del estado de promover y fomentar el acceso a la información.

Art. 71: Deber a incluir en los planes de desarrollo económico y fomento a la cultura social.

Art. 72: El patrimonio cultural de la nación está bajo la protección del estado y son inalienables, inembargables e imprescriptibles.

Art. 74: Derecho de las personas de acceder a los documentos públicos.

Art. 95: Deberes y obligaciones de los ciudadanos de proteger los recursos culturales del país y velar por su conservación en un ambiente sano.

Art. 112: Garantía de acceso a la información y a la documentación oficial a los partidos y movimientos políticos que no participen en el gobierno.

Además de los mandatos constitucionales existen muchas disposiciones legales vigentes con relación al manejo de los documentos y a las responsabilidades de los servidores públicos frente a ellos:

Ley 80 de diciembre 22 de 1989: Por la cual se crea el archivo general de la nación.

Acuerdo 012 de octubre 1o de 1991: Por el cual se fijan plazos para la presentación de las tablas de retención documental.

Acuerdo 07 de junio 29 de 1994: Por el cual se adopta y se expide el reglamento general de archivos.

Acuerdo 09 de octubre 18 de 1995: Por el cual se reglamenta la presentación de las tablas de retención documental al archivo general de la nación.

Ley 190 de junio de 1995: Estatuto anticorrupción.

Art. 27: Utilización indebida de la información.

Art. 79: Es causal de mala conducta obstaculizar, retardar y negar inmotivadamente el acceso a los documentos. Los casos de reservas deben estar amparados en la constitución y las leyes.

Ley 594 de julio 14 de 2000: Ley general de archivos, "por medio de la cual de dicta la ley general de archivos y se dictan otras disposiciones".

Art. 8: Clasificación de los archivos desde el punto de vista territorial.

Art. 10: La creación de los archivos en todos los niveles de la administración pública es de carácter obligatorio.

Art. 14: La documentación de la administración pública es producto y propiedad del estado y este ejercerá el pleno control de sus recursos informativos.

Parágrafo 1: La administración pública podrá contratar con personas naturales o jurídicas los servidos de custodia, organización y conservación de documentos de archivos.

Art. 24: Obligatoriedad de la elaboración, adopción y aplicación de las tablas de retención documental.
Circular externa 001 de enero de 2001: En referencia con la elaboración y adopción de tablas de retención documental y establecimiento del siguiente calendario de presentación:

a) Municipios de categorías 5 y 6, fecha límite julio de 2002.
b) Municipios de categorías 3 y 4, fecha límite septiembre de 2002.
c) Municipios de categorías 1 y 2, fecha límite noviembre de 2002.

Ley 734 de febrero 5 de 2002: Nuevo código disciplinario único.

Art. 34: Deberes de los servidores públicos.

Numeral 1: Cumplir y hacer cumplir la constitución, los tratados internacionales, las leyes, los decretos, las ordenanzas, los acuerdos, los estatutos de la entidad, los reglamentos y Manuales de función, etc.

Numeral 5: Custodiar y cuidar la documentación e información que conserve bajo su cuidado e impedir o evitar la sustracción, destrucción, ocultamiento o utilización indebida.

Numeral 22: Responder por la conservación de los documentos y rendir cuenta oportuna de su utilización.

Art. 35: Prohibiciones de los servidores públicos.

Numeral 7: Omitir, negar, retardar o entrabar el despacho de los asuntos a su cargo.

Numeral 8: Omitir, retardar o no suministrar debida y oportuna respuesta a las peticiones respetuosas.
Numeral 12: Proporcionar dato inexacto o presentar documentos ideológicamente falsos.

Numeral 13: Ocasionar daños o dar lugar a la pérdida de documentos o expedientes.

Numeral 21: Dar lugar al acceso o exhibir expedientes, documentos o archivos a personas no autorizadas

3. ESQUEMA DE METADATOS

Los metadatos son datos altamente estructurados que describen información, describen el contenido, la calidad, la condición y otras características de los datos[4].

A los metadatos se les define como información sobre información; brindando la capacidad para la descripción de documentos [5]. Con los metadatos se puede:

- Dar un significado global a los datos.
- Facilitar la búsqueda de datos.
- Relacionar recursos en linea.

3.1. DESCRIPCION DE LOS METADATOS

Para la descripción e identificación de los metadatos se utilizo como caso de estudio la oficina de Secretaria General De La Universidad Industrial De Santander, donde el flujo de creación, organización y búsqueda de documentos ofrece un escenario ideal de trabajo. Los documentos que genera esta sección son altamente estructurados en su contenido.

3.1.1. Identificación De Contenido Documental

Se realizó una charla con los funcionarios de la oficina para conocer los documentos que se manejan, en la charla se resaltaron aspectos como:

- Tipos de documentos.
- Frecuencia de creación.
- Estructuración del contenido de cada documento.
- Frecuencia de búsqueda.
- Criterios importantes en la búsqueda.
- Criterio de organización.

Cada elemento explorado aportó información sobre el estado actual del sistema informático actual de organización de documentos incorporado en la oficina. Se encontró que la oficina produce y recibe documentos. Los documentos producidos son el resultado de la toma de decisiones por parte de las autoridades de la universidad. Los documentos recibidos, en su gran mayoría se relacionaban con correspondencia externa.

3.1.2. Documentos Seleccionados

Debido a la naturaleza estructurada del modelo de datos basado en ontologías se seleccionaron los documentos que aseguraban un alto volumen de interpretabilidad a la hora de conceptualizar su dominio. En esta medida se escogieron las resoluciones y acuerdos.

Este tipo de documentos se componen de:

- Tipo De Documento
- Número.
- Fecha de expedición.
- Resumen de contenido.
- Autoridad que expide.
- Ítems de consideración.
- Ítems de resolución o acuerdo.
- Firmas.

Las resoluciones y acuerdos son documentos numerados que expide una autoridad, considerando cierta problemática; para resolver o acordar su solución o atenuación. Este concepto se represento a través de los metadatos usando ontologías para la abstracción de dicho concepto.

3.1.3. Construcción De La Ontología.

La construcción de una ontología para representar el conocimiento de los documentos, permitió definir la naturaleza de los tipos escogidos. Esta naturaleza esta representada por las clases, propiedades y atributos descritos en la ontología. Se utilizo una ontología para:

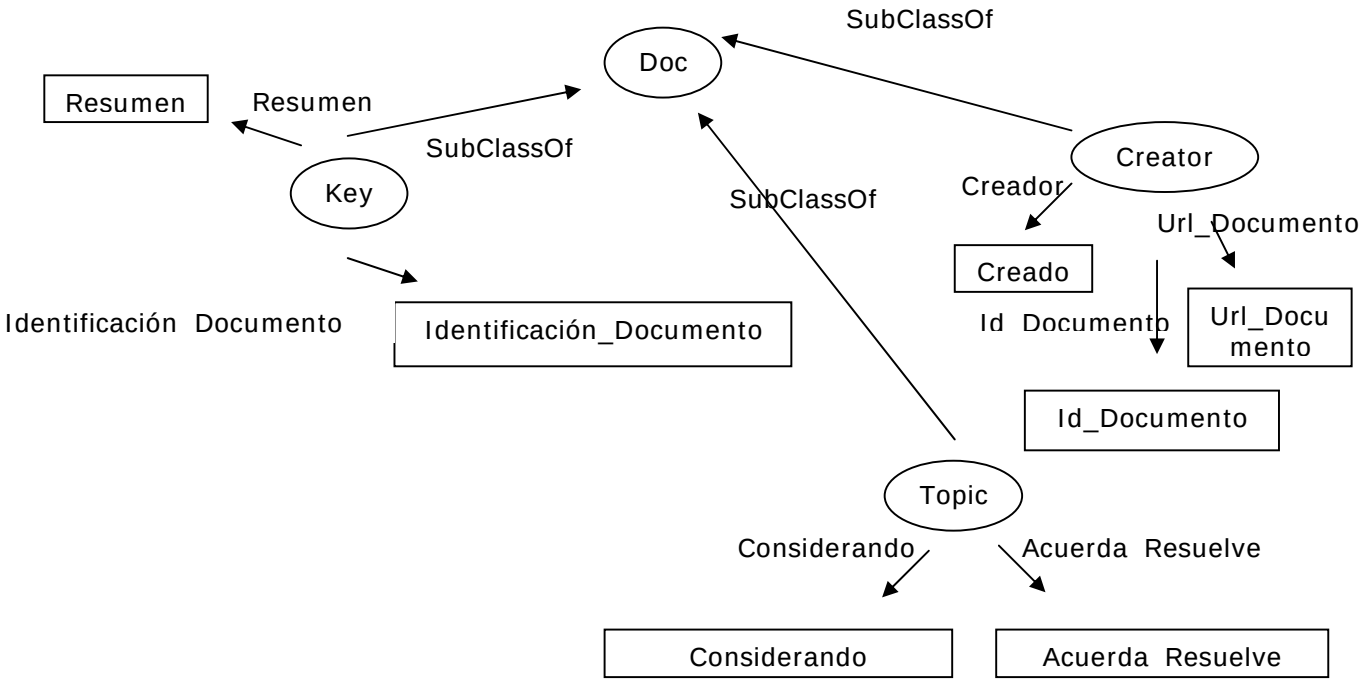
- Representar una estructura común de información para los tipos de documentos escogidos.
- Representar el concepto de forma clara.
- Definir un esquema para la inserción de los metadatos.
- Formalizar una infraestructura de metadatos para compartir y reutilizar información.
- Favorecer la implementación de agentes documentales.

3.1.4. Construcción de clases

Una vez identificadas las clases, se empezó a construir la jerarquía de clases. Se usó la metodología **TOP-DOWN** para su construcción. Los documentos seleccionados difieren solamente en el tipo y en el concepto que define la solución, la definición de clases utiliza una sola ontología para describir este tipo de documentos utilizando la funcionalidad de sus

atributos como identificadores de tipo. En la Fig. 5 se precia la ontología que se diseñó.

Fig. 5 Ontología de documentos seleccionados



La clase Doc oficia como raíz del diseño de la ontología, las subclases Creator, Key, Sing y Topic identifican al documento en información clave, firmas y contenido respectivamente.

3.1.5. Descripción De Propiedades

Cada una de las clases describe un concepto dentro del dominio de cada documento, en cada clase existen propiedades que aportan alcances y limitaciones de la ontología. Las propiedades que describen cada clase son las siguientes:

- ✓ **Clase Key**

Tabla 2. Descripción de las propiedades de la clase Key		
	Resumen	**Identificación_Documento**
Cardinalidad	1	1
Tipo	string	String
Rango	Key	Key
Dominio	Doc	Doc

- **Resumen:** información acerca del tema expuesto en el documento.
- **Identificación_Documento:** Tipo de documento y fecha de elaboración.

✓ **Clase Topic**

Tabla 3. Descripción de las propiedades de la clase Topic		
	Considerando	**Acuerda_Resuelve**
Cardinalidad	1	1
Tipo	string	string
Rango	Topic	Topic
Dominio	Doc	Doc

- **Considerando:** Ítems de consideración.
- **Acuerda_Resuelve:** Ítems de resolucion o acuerdo

✓ **Clase Creador**

Tabla 4. Descripción de las propiedades de la clase Creator			
	Creador	**Url_Documento**	**Id_Documento**
Cardinalidad	1	1	1
Tipo	string	string	int
Rango	Creator	Creator	Creator
Dominio	Doc	Doc	Doc

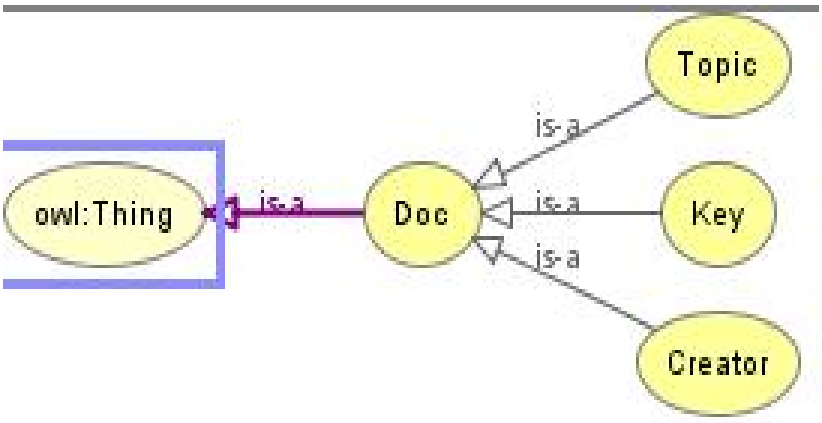
- **Creador:** información acerca del creador del documento
- **Url_Documento: d**ireccion http del recurso.

- **Id_Documento:** numero de identificación interno del documento en la base de datos original.

3.1.6. Edición de la ontología con Protege - OWL 3.1

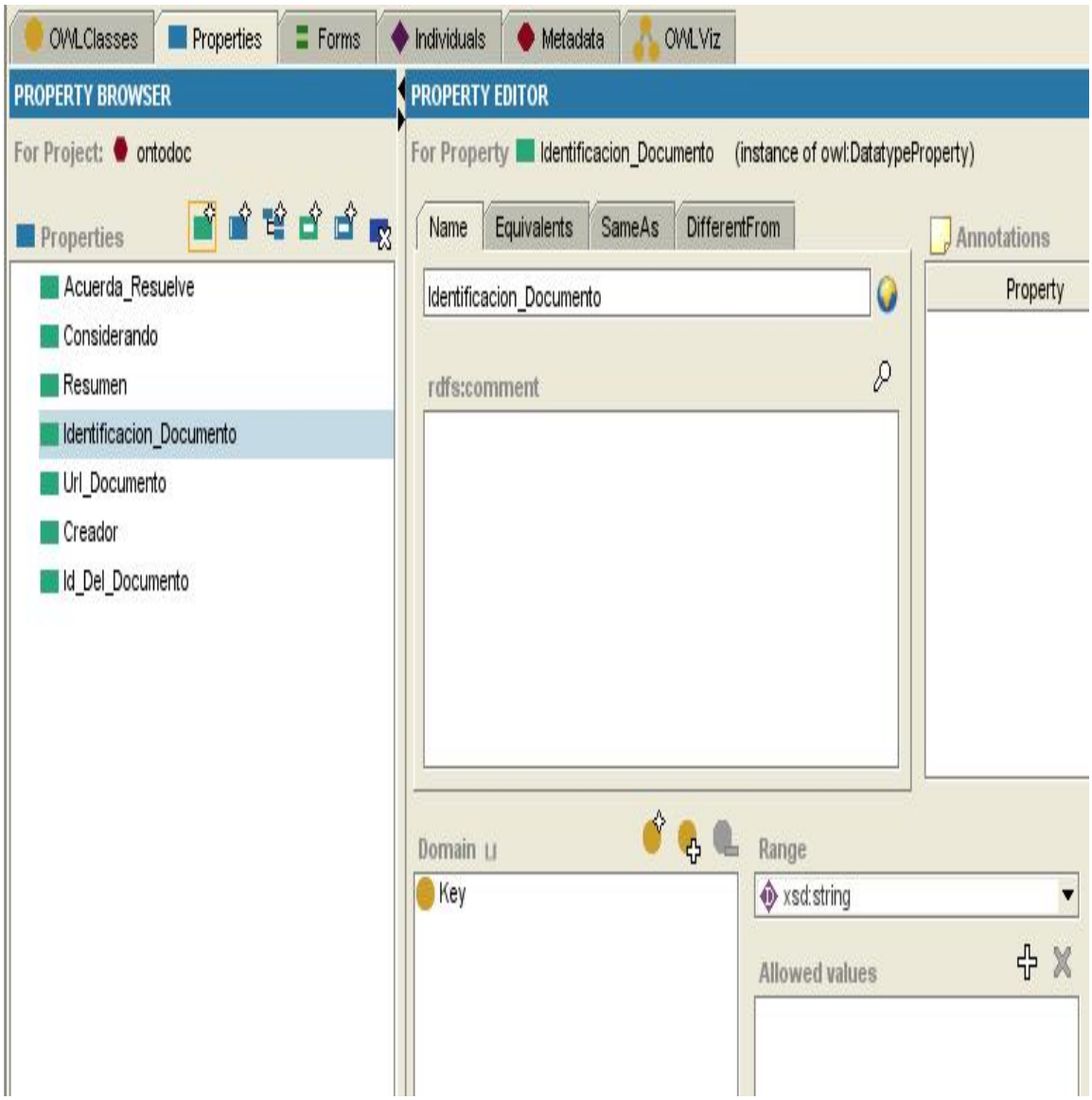
Protégé [7] es un entorno para la edición de ontologías y la adquisición de conocimiento. Se uso el plugin de protégé OWL para describir la ontologia. En la Fig. 6 se visualizan las clases descritas.

Fig 6. Grafo dirigido de la ontología de documentos



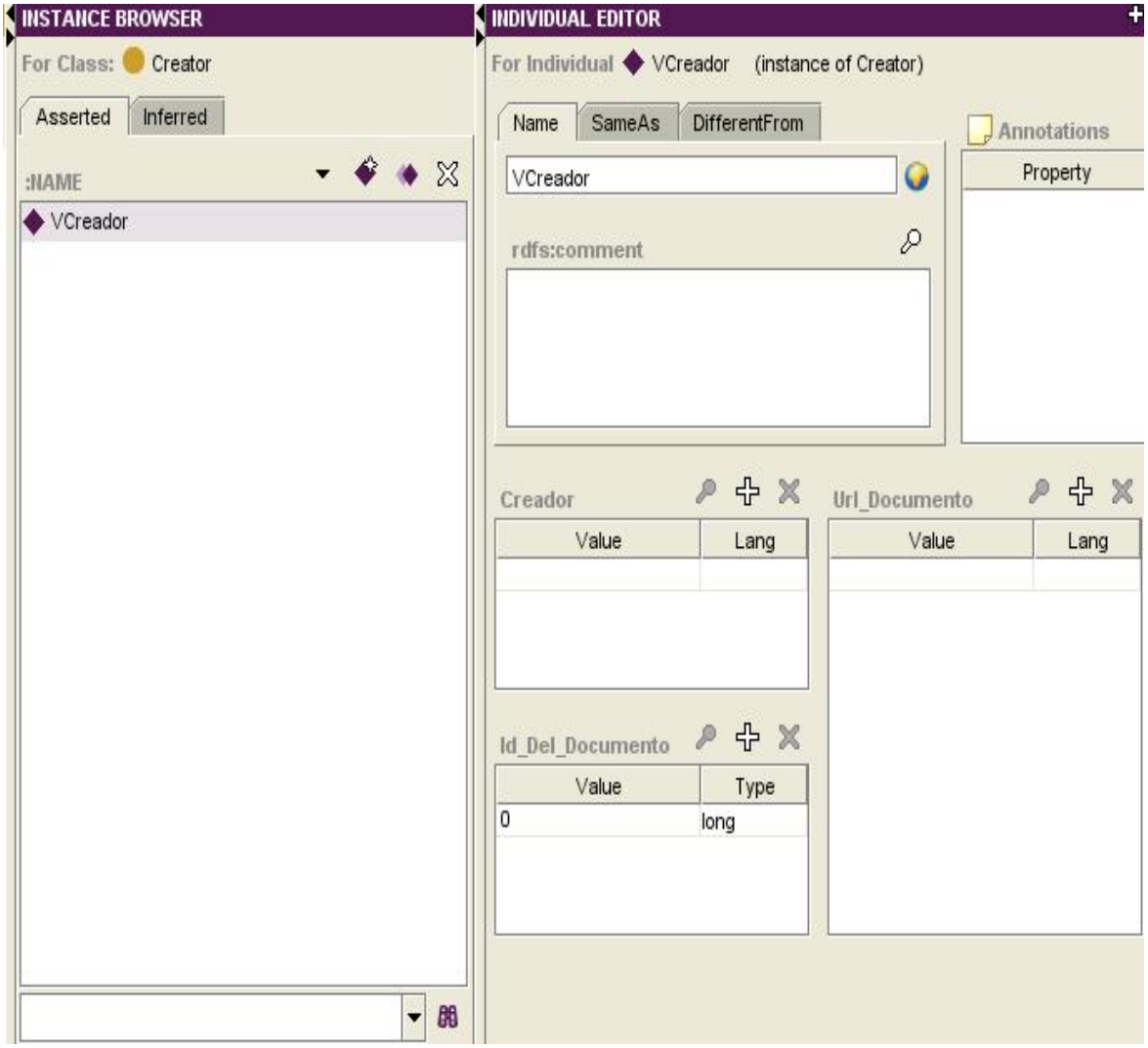
Se diseñaron las propiedades aplicando los valores correspondientes al diseño de la ontología. En la Fig 7. se observan los tipos de datos, así como el dominio de cada propiedad.

Fig 7. Propiedades descritas en Protégé – OWL 3.1



Se crearon las instancias o individuos Vcreator, Vkey, Vtopic para la representación final de la ontología y sus propiedades, la Fig. 8 muestra la edición de los individuos.

Fig 8. Edición de instancias



4. CONSTRUCCION DEL PROTOTIPO DE ANOTACION Y BUSQUEDA SEMANTICA

La inserción de metadatos posee características que pueden ser utilizadas en aplicaciones para la recuperación de recursos que se encuentran guardados en bases de datos en campos de tipo texto. La información acerca de los recursos que se desean recuperar es analizada y buscada de acuerdo a la necesidad del usuario.

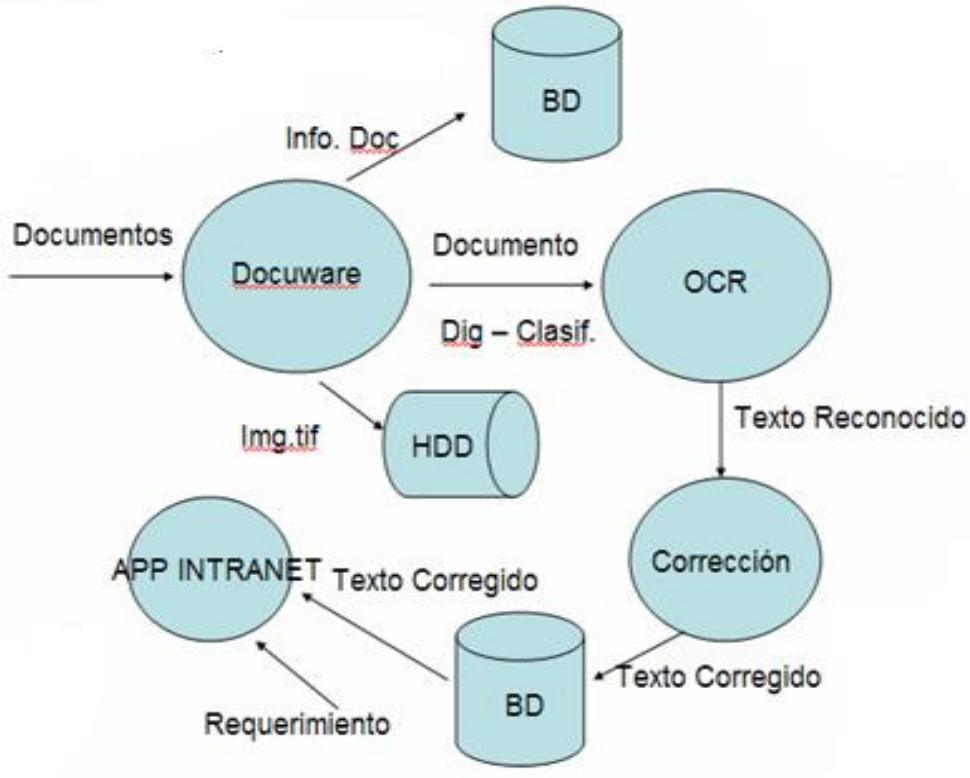
En este capítulo se presentan los pasos dados en la construcción del prototipo que se desarrollo para insertar y recuperar metadatos de documentos digitalizados. Se presentará una descripción de las herramientas utilizadas en su desarrollo.

4.1. Análisis Del Actual Sistema De Digitalización y Consulta De Documentos.

Para el desarrollo del prototipo software se realizo un análisis del sistema actual de digitalización y consulta de documentos, se seleccionó la oficina de secretaria general, oficina que maneja una gran cantidad de documentos emitidos por diferentes secciones de la universidad.

Actualmente la oficina de secretaria general utiliza el software comercial Docuware para digitalizar y posteriormente reconocer el texto en la imagen, operación que permite mediante una previa corrección hecha por el personal de la oficina, guardar texto exacto de cada acuerdo o resolución en un campo tipo full texto. La imagen es guardada en disco magnético con formato tiff y las referencias se codifican dentro de la base de datos. En la Fig. 9 se muestra el flujo de información del sistema actual.

Fig 9. Flujo de información sistema actual



Para la consulta de documentos, la oficina de Sistemas de información desarrollo una pequeña aplicación Web que toma como fuente el texto

corregido guardado en la base de datos. La búsqueda se realiza a través de sentencias SQL como like, y comparaciones exactas usando operaciones de relación.

El sistema actual funciona bien, respalda la información documental de manera eficiente y permite buscar con facilidad el texto requerido en una consulta de información exacta, por ejemplo: rangos de fechas, números de documentos, nombre exactos. Sin embargo estos documentos tratan situaciones que en ciertas ocasiones también son consultadas por los usuarios por ejemplo: buscar un nombre en el resumen, buscar un literal dentro lo que considera el acuerdo 01 de 2006, etc. En estas circunstancias donde los datos de la búsqueda no son tan exactos y hasta se pueden decir que son vagos, es donde el sistema arroja una gran cantidad de documentos donde posiblemente esta el que se requiere. Esta problemática pone al usuario navegar por muchos documentos, los cuales no están ordenados por ningún criterio,

4.1.1. Tecnología Utilizada En El Sistema Actual

Utiliza como plataforma en el servidor, el sistema operativo Microsoft Windows 2000 Profesional:

- ✓ Licencia Docuware para Windows por cada Equipo, software de digitalización y reconocimiento de documentos.

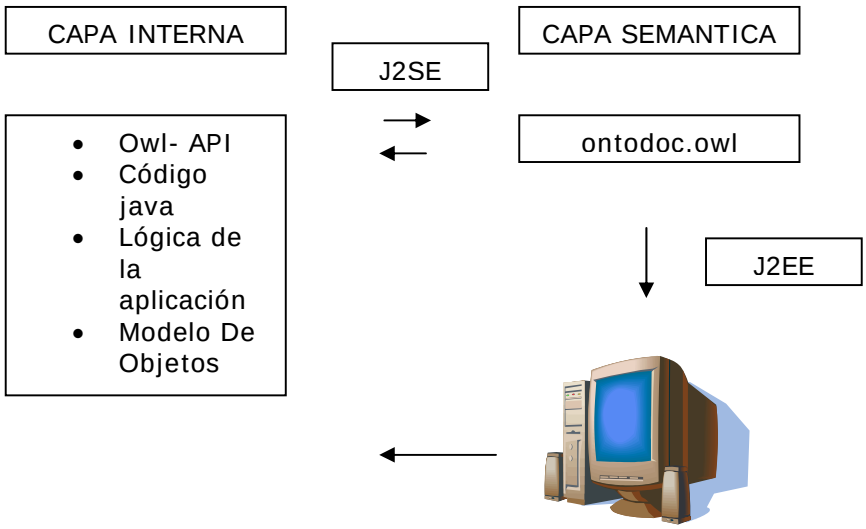
- ✓ Microsoft SQL SERVER 2000, como motor de base de datos.
- ✓ ASP(Active Server Pages), para el desarrollo de la aplicación Web de búsqueda.
- ✓ Auto Feed Scanner Hewllet Packard, Hardware de digitalización.

4.2. Arquitectura Del Prototipo

En el capitulo 3 se describió el esquema de metadatos y la construcción de la ontología base del prototipo. Para el desarrollo del prototipo se utilizó Protégé – OWL - API 3.1, plataforma que permite la edición de ontologías y la ejecución e implementación de herramientas de anotación y búsqueda de metadatos. Esta desarrollado en Java permitiendo su trabajo en entornos J2SE y J2EE.

En la Fig 10 se puede observar la arquitectura del prototipo de aplicación de Web semántica.

Fig. 10 Arquitectura del prototipo de aplicación de Web semántica



4.2.1. Representación De La Ontología En Lenguaje OWL – RDF/XML ABBREV

El lenguaje semántica Onto Web Language OWL provee a través de un nuevo modelo lógico donde se identifican clases, propiedades e individuos, una alta interpretabilidad y abstracción de conocimientos en dominios documentales. Se representa la información interna de cada acuerdo o resolución, su identificación, creador, tópicos de contenido como las propiedades de las clases que componen el documento.

Fig 11. Ontologia Propuesta

```
<?xml version="1.0" ? encoding = “ISO-8859-1”>
- <rdf:RDF xmlns:rdf="http://www.w3.org/1999/02/22-rdf-syntax-ns#"
xmlns:xsd="http://www.w3.org/2001/XMLSchema#"
xmlns:rdfs="http://www.w3.org/2000/01/rdf-schema#"
xmlns:owl="http://www.w3.org/2002/07/owl#" xmlns="http://www.owl-
ontologies.com/unnamed.owl#" xml:base="http://www.owl-
ontologies.com/unnamed.owl">
  <owl:Ontology rdf:about="" />
  <owl:Class rdf:ID="Doc" />
+ <owl:Class rdf:ID="Creator">
+ <owl:Class rdf:ID="Topic">
+ <owl:Class rdf:ID="Key">
+ <owl:DatatypeProperty rdf:ID="Creador">
+ <owl:DatatypeProperty rdf:ID="Identificacion_Documento">
+ <owl:DatatypeProperty rdf:ID="Acuerda_Resuelve">
+ <owl:DatatypeProperty rdf:ID="Considerando">
+ <owl:DatatypeProperty rdf:ID="Resumen">
+ <owl:DatatypeProperty rdf:ID="Id_Del_Documento">
+ <owl:DatatypeProperty rdf:ID="Url_Documento">
- <Key rdf:ID="VKey">
  <Resumen rdf:datatype="http://www.w3.org/2001/XMLSchema#string" />
  <Identificacion_Documento rdf:datatype="http://www.w3.org/2001/XMLSchema#string"
/>
  </Key>
- <Creator rdf:ID="VCreador">
  <Id_Del_Documento
rdf:datatype="http://www.w3.org/2001/XMLSchema#long">0</Id_Del_Documento>
```

```
<Creador rdf:datatype="http://www.w3.org/2001/XMLSchema#string" />
<Url_Documento rdf:datatype="http://www.w3.org/2001/XMLSchema#string" />
</Creator>
- <Topic rdf:ID="VTopic">
<Acuerda_Resuelve rdf:datatype="http://www.w3.org/2001/XMLSchema#string" />
<Considerando rdf:datatype="http://www.w3.org/2001/XMLSchema#string" />
</Topic>
</rdf:RDF>
```

En la Fig. 11 se representa la ontología propuesta donde se observan objetos de tipo:

- <Owl:class>: Son los conceptos del dominio de los documentos. Las clases Creador, Key y Topic son subclases de la Superclase Doc
- <Owl:DataTypeProperty>: son las propiedades de cada clase, permiten identificar y declarar los datos que describen a los conceptos. Relacionan las clases con la información del documento
- <clase rdf:ID=”Indiviudo”>: declaración de los individuos o instancias de cada clase, agrupan las clases y las propiedades en un solo ente. Los individuos o instancias son los objetos que permiten las anotaciones.

4.2.2. Generación Automática De Anotaciones Semánticas

Una anotación se puede considerar como una información sobre las entidades o conceptos de una ontología, que aparecen en un texto y su situación en el mismo, se guardan clasificadas como objetos o propiedades correspondientes a un concepto de una ontología.

Fig 12 Anotaciones a un acuerdo

ACUERDO No 01 DE 2006 (Enero 24) Por el cual se aprueba la convocatoria para ampliar la base de profesores de cátedra elegibles y se establece el cronograma de actividades EL CONSEJO ACADÉMICO DE LA UNIVERSIDAD INDUSTRIAL DE SANTANDER en uso de sus atribuciones legales y, CONSIDERANDO: a. Que mediante Acuerdo Superior No. 004 del 7 de febrero de 2005 se aprobó el Reglamento del Profesor de Cátedra de la Universidad Industrial de Santander. b. Que mediante Acuerdo Académico No. 142 del 4 de octubre 4 de 2005 se desarrolló el artículo 54 del Acuerdo Superior No. 004 de 2005 c. Que después de surtida la convocatoria del ano 2005 para conformar la base de profesores de cátedra elegibles, todavía existen unidades académicas que requieren profesores de cátedra para atender las asignaturas que se ofrecerán durante el primer semestre académico de 2006. d. Que según el Reglamento del Profesor de Cátedra, articulo 18, en caso de no existir o haberse agotado los candidatos elegibles en la base de datos para proveer docentes en determinadas asignaturas, la Vicerrectoría Académica podrá realizar una convocatoria pública extraordinaria. e. Que según el Reglamento del Profesor de Cátedra, articulo 6, el Consejo Académico debe aprobar la convocatoria pública para el concurso de inclusión en la base de profesores de cátedra elegibles. ACUERDA: ARTICULO l°: Aprobar la convocatoria para ampliar la base de profesores de cátedra elegibles y establecer el cronograma de actividades. Enero 25 Cierre de inscripciones. Se recibirá la documentación hasta las 8:00 p.m. Enero 26 Verificación del cumplimiento de requisitos de participación y valoración de hojas de vida Enero 26 Publicación de la lista de los candidatos que reúnen los requisitos de la convocatoria y publicación de la lista de aspirantes ordenada por el puntaje obtenido por hoja de vida y conformación de la base de profesores de cátedra elegibles organizada por áreas de desempeño, a partir de las 3 p.m. COMUNÍQUESE Y CÚMPLASE, Expedido en Bucaramanga, a los veinticuatro (24) días del mes de enero de 2006. LA PRESIDENTA DEL CONSEJO ACADÉMICO, LUCILA NIÑO BAUTISTA, Vicerrectora Académica EL SECRETARIO GENERAL, CRISÓSTOMO BARAJAS FERREIRA

En la Fig. 12 se resaltan los elementos del texto que serán agregados a la ontología. Las anotaciones son de tipo externo, es decir, no son hechas directamente sobre el recurso sino que son guardadas en un repositorio de datos RDF.

El proceso de anotación es automático debido a la estructuración del documento. El texto es simultáneamente guardado en la base de datos SQL SERVER 2000 y en el repositorio de datos RDF, En la Fig 13 se observa un documento anotado con la herramienta.

Fig 13 Documento resultante del proceso de anotación

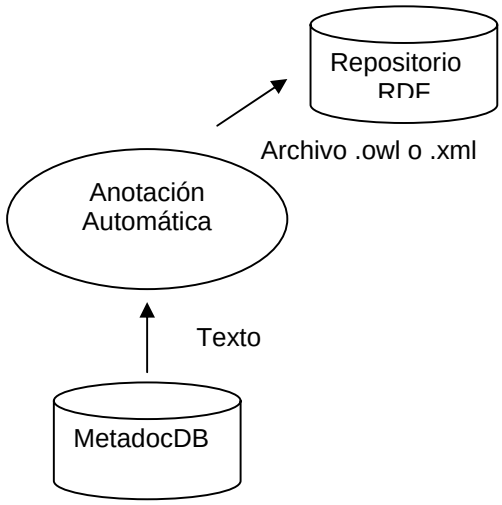
- <Creator1 rdf:ID="**VKey**">
<Identificacion_Documento rdf:datatype="**http://www.w3.org/2001/XMLSchema#string**">**ACUERDO No 01 DE 2006 (Enero 24)**</Identificacion_Documento>
<Resumen rdf:datatype="**http://www.w3.org/2001/XMLSchema#string**">**"Por el cual se aprueba la convocatoria para ampliar la base de profesores de catedra elegibles y se establece el cronograma de actividades"**</Resumen>
</Creator1>
- <Creator1 rdf:ID="**VCreator**">
<Creador rdf:datatype="**http://www.w3.org/2001/XMLSchema#string**">**EL CONSEJO ACADEMICO DE LA UNIVERSIDAD INDUSTRIAL DE SANTANDER**</Creador>
<Id_Del_Documento rdf:datatype="**http://www.w3.org/2001/XMLSchema#int**">**01**</Id_Del_Documento>
<Url_Documento rdf:datatype="**http://www.w3.org/2001/XMLSchema#string**">**C:\oacs\Proyecto\onto\012006AcuConsAca.htm**</Url_Documento>
</Creator1>
- <Creator1 rdf:ID="**VTopic**">
<Considerando rdf:datatype="**http://www.w3.org/2001/XMLSchema#string**">**a. Que mediante Acuerdo Superior No. 004 del 7 de febrero de 2005 se aprobo el Reglamento del Profesor de Catedra de la Universidad Industrial de Santander. b. Que mediante Acuerdo Academico No. 142 del 4 de octubre 4 de 2005 se desarrollo el artículo 54 del Acuerdo Superior No. 004 de 2005 c. Que despues de surtida la convocatoria del ano 2005 para conformar la base de profesores de catedra elegibles, todavía existen unidades academicas que requieren profesores de catedra para atender las asignaturas que se ofreceran durante el primer semestre academico de 2006. d. Que segun el Reglamento del Profesor de Cátedra, articulo 18, en caso de no existir o haberse agotado los candidatos elegibles en la base de datos para**

proveer docentes en determinadas asignaturas, la Vicerrectoría Académica podrá realizar una convocatoria pública extraordinaria. e. Que según el Reglamento del Profesor de Cátedra, articulo 6, el Consejo Académico debe aprobar la convocatoria pública para el concurso de inclusión en la base de profesores de cátedra elegibles.</Considerando>
<Acuerda_Resuelve
rdf:datatype="**http://www.w3.org/2001/XMLSchema#string**">**ARTICULO l°: Aprobar la convocatoria para ampliar la base de profesores de cátedra elegibles y establecer el cronograma de actividades. Enero 25 Cierre de inscripciones. Se recibirá la documentación hasta las 8:00 p.m. Enero 26 Verificación del cumplimiento de requisitos de participación y valoración de hojas de vida Enero 26 Publicación de la lista de los candidatos que reúnen los requisitos de la convocatoria y publicación de la lista de aspirantes ordenada por el puntaje obtenido por hoja de vida y conformación de la base de profesores de cátedra elegibles organizada por áreas de desempeño, a partir de las 3 p.m.**</Acuerda_Resuelve>
</Creator1>

4.2.2.1. Implementación De Anotaciones Automáticas

La herramienta extrae de la bases de datos original el texto del documento y exporta el documento anotado a un repositorio de datos RDF, Ver Fig. 14

Fig 14 proceso de anotación externa



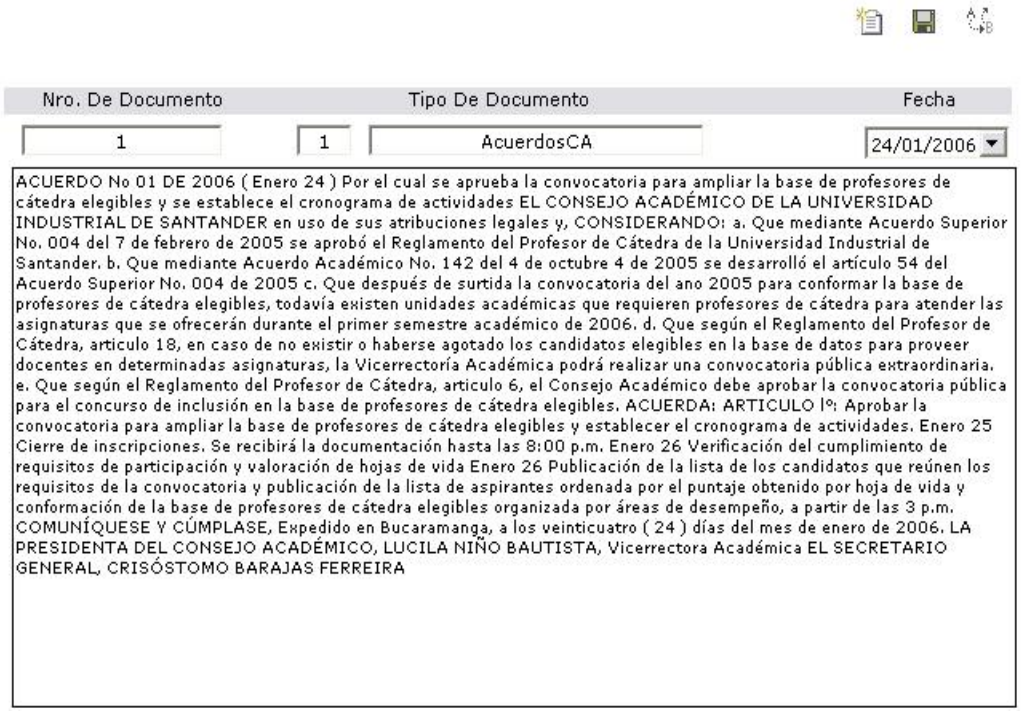
Para generar el nuevo documento OWL se usa la combinación de las API's OWL y JENA, se crean a partir de las siguientes clases definidas en el javadoc de las API´s utilizadas:

- ✓ **createJenaOWLModel():** Crea un nuevo modelo ontologico OWL y permite exportarlo a lenguaje RDF/XML Abrev.

- ✓ **createOWLNamedClass(),createOWLNamedSubClass() :** Crea una nueva clase o subclase
- ✓ **createOWLDatatypeProperty():** Crea una nueva propiedad para la clase designada, especifica tipo de datos, cardinalidad y anotaciones extras a cada propiedad

- ✓ **createOWLIndividual():** Crea un instancia de la clase y sus propiedades, permitiendo realizar la inserción de los metadatos en la definición de cada uno de ellos.

Fig 15 Interfaz de Anotación



4.2.3. Recuperación y Búsqueda De Metadatos

Cada documento anotado posee una replica en el repositorio de datos RDF, este lenguaje provee primitivas para representar expresiones que en su estructura tiene tres elementos importantes *sujeto, predicado* y *objeto*. Los metadatos de los documentos están representados en expresiones RDF, de esta forma se representarán todas las propiedades concernientes al texto. En la Fig. 16 se puede observar un fragmento de la descripción de un acuerdo en lenguaje RDF/XML.

Fig 16 Ejemplo de acuerdo almacenado en el repositorio de datos

- <Topic1 rdf:ID="**VTopic**">
<Considerando rdf:datatype="**http://www.w3.org/2001/XMLSchema#string**">**a. Que mediante Acuerdo Superior No. 004 del 7 de febrero de 2005 se aprobo el Reglamento del Profesor de Catedra de la Universidad Industrial de Santander. b. Que mediante Acuerdo Academico No. 142 del 4 de octubre 4 de 2005 se desarrollo el artículo 54 del Acuerdo Superior No. 004 de 2005 c. Que despues de surtida la convocatoria del ano 2005 para conformar la base de profesores de catedra elegibles, todavía existen unidades academicas que requieren profesores de catedra para atender las asignaturas que se ofreceran durante el primer semestre academico de 2006. d. Que segun el Reglamento del Profesor de Cátedra, articulo 18, en caso de no existir o haberse agotado los candidatos elegibles en la base de datos para proveer docentes en determinadas asignaturas, la Vicerrectoría Académica podrá realizar una convocatoria pública extraordinaria. e. Que según el Reglamento del Profesor de Cátedra, articulo 6, el Consejo Académico debe aprobar la convocatoria pública para el concurso de inclusión en la base de profesores de cátedra elegibles.**</Considerando>
<Acuerda_Resuelve rdf:datatype="**http://www.w3.org/2001/XMLSchema#string**">**ARTICULO l°: Aprobar la convocatoria para ampliar la base de profesores de cátedra elegibles y establecer el cronograma de actividades. Enero 25 Cierre de inscripciones. Se recibirá la documentación hasta las 8:00 p.m. Enero 26 Verificación del cumplimiento de requisitos de participación y valoración de hojas de vida Enero 26 Publicación de la lista de los candidatos que reúnen los requisitos de la convocatoria y publicación de la lista de aspirantes ordenada por el puntaje obtenido por hoja de vida y conformación de la base de profesores de cátedra elegibles organizada por áreas de desempeño, a partir de las 3 p.m.**</Acuerda_Resuelve>
</Topic1>

Para extraer los metadatos de la base de conocimiento utilizamos JENA que provee un conjunto de librerías, protocolos y herramientas que permiten manipular el modelo RDF de manera jerárquica, navegando desde la definición de las clases, propiedades, instancias.

En este caso en particular las consultas al modelo de datos RDF/XML de cada documento se realizan por búsqueda de subcadenas en el texto de cada propiedad RDF/XML, lo cual exige extractar todo el contenido para realizar las operaciones. La API de JENA permite navegar dentro del documento a través de las siguientes instrucciones:

- ✓ **ModelFactory.createOntologyModel(“Ruta De Modelo De Datos”):** Crea un modelo de datos a partir del repositorio RDF existente y se almacena en la memoria temporal del servidor.

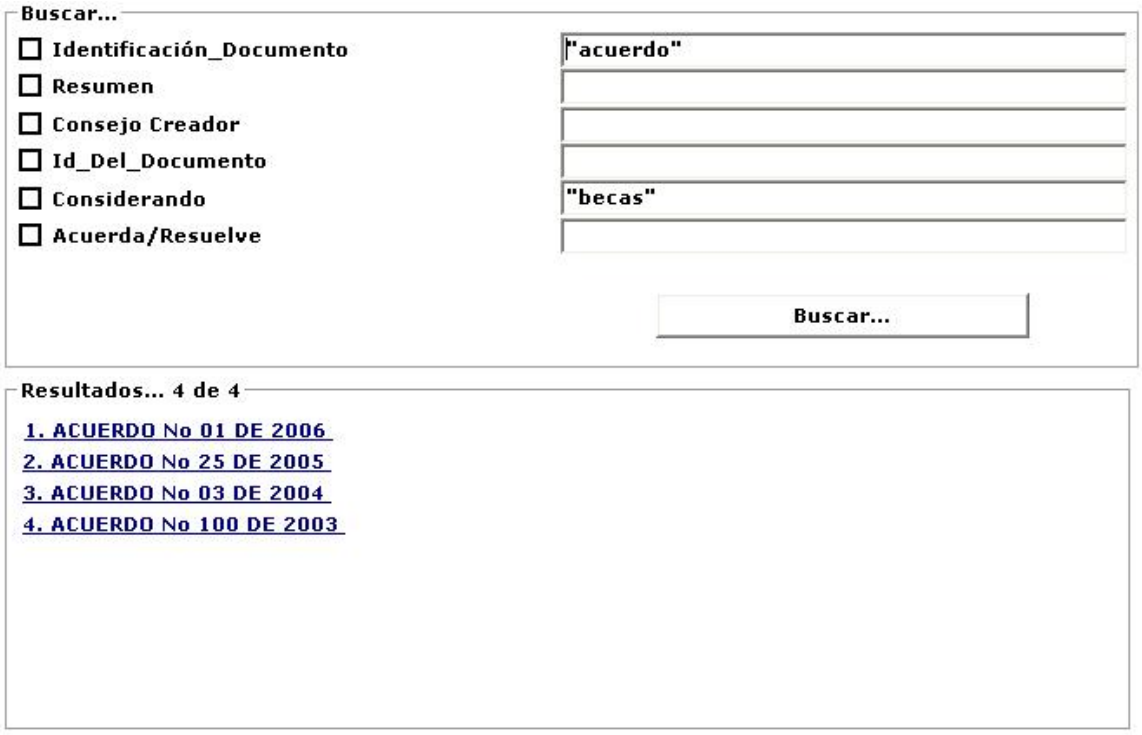
- ✓ **m.listClasses():** se extrae la lista de clases del modelo de datos

- ✓ **c.listInstances():** se extrae la lista de instancias del modelo de datos

- ✓ **c.listDeclaredProperties():** se extrae el valor de cada propiedad y se realizan las comparaciones tipo like, búsqueda de subcadenas, búsquedas exactas, etc.

Se realiza una búsqueda de metadatos en todos los modelos existentes en el repositorio RDF, consultando solamente las propiedades que el usuario haya seleccionado, permitiendo una mayor eficacia en la muestra de resultados. En la Fig 17 se muestra la pantalla de busqueda de metadatos

Fig 17 Busqueda De Metadatos



La Fig 17 muestra la estructura de la interfaz de búsqueda semántica, donde se buscan las documentos usando los conceptos de la ontología diseñada. Nótese la posibilidad de búsquedas por herencia usando el literal, por ejemplo: al buscar la cadena “becas”, el sistema arroja además de las coincidencias en cada concepto, aquellas palabras que tienen un significado parecido al dado inicialmente.

4.2.4. Tecnologías utilizadas

La tecnología utilizada en el desarrollo del prototipo son:

- ✓ Sql Server 2000, Motor de base de datos utilizado para guardar el texto del documento identificando su procedencia y tipo
- ✓ JENA API, middleware desarrollado por HP LABS que permite manejar los modelos de datos RDF y OWL para editar y buscar información dentro de los repositorios de datos
- ✓ OWL API, librería que permite navegar por los modelos de datos OWL
- ✓ JAVA, lenguaje de programación utilizado por las librerías JENA y OWL API
- ✓ JDBC SQL DRIVER, conector usado en java para manipular bases de datos SQL Server.
- ✓ JSP, interfaz final del usuario
- ✓ Tomcat 5.X, JSP/Servlet Container.

5. CONCLUSIONES Y RECOMENDACIONES

En este trabajo de investigación se construyó una ontología de descripción de documentos y su representación en el lenguaje semántico RDF/XML como una alternativa para compartir y reutilizar el conocimiento de un dominio. También, se investigó y se diseñó un prototipo funcional para la anotación automática búsqueda y la recuperación de documentos para el cual se utilizó como caso de estudio los documentos estructurados (acuerdos y resoluciones) generados en la oficina de Secretaria General de la Universidad Industrial De Santander.

5.1. CONCLUSIONES

La aplicación de un modelo de datos basado en ontologías para la clasificación y recuperación de información de documentos estructurados permitió aclarar los conceptos sobre la capacidad y el futuro de las aplicaciones semánticas. También, otorga la oportunidad de acercar a la comunidad universitaria hacia el desarrollo de aplicaciones Web que enfoquen su contenido hacia el conocimiento y no hacia el diseño.

El análisis de los sistemas actuales de digitalización y búsqueda de documentos utilizados por la institución dejo ver que el aumento progresivo de los documentos dificulta su posterior recuperación.

Dar paso a sistemas que extraigan el conocimiento y aseguren el patrimonio documental no solo de forma digital sino también de forma inteligente, facilita la búsqueda por parte del usuario final que, a fin de cuentas, somos todos, debido al carácter publico de los documentos procesados.

5.2. RECOMENDACIONES

- ✓ Permitir la implementación final de aplicaciones semánticas para la gestión de documentos estructurados
- ✓ Introducir en el ámbito local el uso de metadatos como solución al problema de integración de información.
- ✓ Con las herramientas existentes para el desarrollo de sistemas basados en RDF/XML, es posible crear aplicaciones como la producida durante el trabajo descrito, pero no se nota una amplia movilidad en las especificaciones de las tecnologías utilizadas, lo que produce que las herramientas (APIs y lenguajes) no están implementadas con un nivel de optimización para llevarlas a producción, sino que más bien se encuentran resolviendo los problemas de ceñirse a las especificaciones que las guían.

BIBLIOGRAFÍA

1. Berners-Lee, T., J. Hendler y O. Lassila. 2001 The Semantic Web: A new Form of Web Content that is Meaningful to Computers will Unleash a Revolution of New posibilities. http://www.sciam.com/article.cfm?articleID=00048144-10D2-1C70-4A9809EC588EF21.

2. Duineveld, A. J., R. Stoter, M.R. Weiden, B. Kenepa y V.R. Benjamins. 2000. WonderTools?: a comparative study of ontological engineering tools. Internacional Journal of Human-Computer Studies, Volume 52, Issue 6, p: 1111-1133.

3. Fridman, N. y D. McGuinness. 2001. Ontology Development 101: A Guide to Creating Your First Ontology. Stanford Knowledge Systems Laboratory Technical Report KSL-01-05 and Stanford Medical Informatics Technical Report SMI-2001-0880.

4. Dinos R. Juan L. 2004. Arquitectura De Un Sistema Basado En Agentes Para La Recuperación De Metadatos RDF En Base A Una Ontología De Documentos

5. Steinacker, A., A. Ghavam y R. Steinmetz. 2001. Metadata Standards for Web-Based Resources. Multimedia, IEEE, vol: 8, issue: 1, pp: 70-17.

6. Peis Redondo, Eduardo, Hassan Montero, Yusef, Herrera Viedma, Enrique, Herrera, Juan Carlos. Ontologías, metadatos y agentes: recuperación “semántica” de la información. 2003

7. Sergio F. Castillo**,** Juan R. Velasco. Agentes moviles para la composición de servicios web. IV Jornadas de Ingeniería Telemática. 2003

8. World Wide Web Consortium. Resource Description Framework (RDF). http: //www.w3.org/RDF/.

9. World Wide Web Consortium. Resource Description Framework Schema Specification 1.0http://www.w3.org/TR/2000/CR-rdf-schema-20000327/

10. P. Patel-Schneider, P. Hayes, and I. Horrocks. OWL Web Ontology Language (OWL) Abstract Syntax and Semantics. http://www.w3.org/TR/owl-semantics/, 2003.

11. M. Smith, C. Welty, and D. McGuinness. OWL Web Ontology Language Guide.http://www.w3.org/TR/owl-guide/, 2003

12. [Atanas,03] Atanas, K. et al.. Semantic Annotation, Indexing, and Retrieval. Human Language Technologies Workshop at the 2nd International Semantic Web

Conference (ISWC2003), 20 October 2003, Florida, USA. Disponible en: http://www.ontotext.com/publications/SemAIR_ISWC169.pdf (3/07/04)

13. S. Bechhofer, I. Horrocks, C. Goble, and R. Stevens. OilEd: a Reason-able Ontology Editor for the Semantic Web. In Proc. of KI2001, Joint German/Austrian

14. Sean Bechhofer, Raphael Volz, and Phillip Lord. Cooking the Semantic Web with the OWL API

15. Hewlett Packard. Jena Semantic Web Toolkit. http://www.hpl.hp.com/semweb/jena.htm.