



Casa abierta al tiempo

UNIVERSIDAD AUTÓNOMA METROPOLITANA

Integración Semántica de Recursos en una Memoria Corporativa

Idónea Comunicación de Resultados para obtener el grado de

MAESTRO EN CIENCIAS
(CIENCIAS Y TECNOLOGÍAS DE LA INFORMACIÓN)

por

Erik Alarcón Zamora

Asesores:

Dra. Reyna Carolina Medina Ramírez

Dr. Héctor Pérez Urbina

23 de octubre de 2013

Resumen

El área de Redes y Telecomunicaciones (RyT) del departamento de Ingeniería Eléctrica (IE) de la Universidad Autónoma Metropolitana (UAM) tiene una amplia y rica variedad (heterogeneidad en formato, contenido y estructura) de recursos de información. Algunos ejemplos de estos recursos de información son: los profesores y alumnos del departamento IE, artículos científicos, notas de curso, bases de datos de los trabajadores del dpto. IE, libros, presentaciones, manuales, inventarios, especificaciones de circuitos eléctricos.

Cada recurso representa el conocimiento sobre investigaciones, colaboraciones, proyectos, cursos y temas de interés de los profesores y alumnos en el dominio RyT. Por ejemplo, los artículos científicos, presentaciones, notas de curso e inclusive el propio profesor autor de estos documentos y multimedia son fuentes de información. Todo el conocimiento de una organización representado a través de los recursos, se conoce como memoria corporativa [1].

Una adecuada gestión del conocimiento en una memoria corporativa (MC) se traduce en varias ventajas a nivel operacional, como: una organización bien informada y con mejores tomas de decisión, una herramienta de aprendizaje para las personas adscritas a la organización, una base de conocimiento persistente y accesible para estas personas, un instrumento para búsqueda, recuperación e intercambio de conocimiento entre personas, por mencionar algunas.

Para llevar a cabo esta gestión de los recursos en una MC, se necesitan dos operaciones: 1) la representación del conocimiento sobre los recursos y 2) la búsqueda sobre esta representación. En las tecnologías de la Información, hay varios enfoques tradicionales de representar/buscar el conocimiento de los recursos, como: motores de búsqueda sintácticos y bases de datos relacionales. Pero, el enfoque que nos llamó la atención, es el de las Tecnologías Semánticas.

Las Tecnologías Semánticas se basan en el uso de tecnologías, herramientas y estándares para: la representación de los recursos en un formato estándar, establecer un vocabulario conceptual, la explotación del conocimiento mediante reglas, la búsqueda y recuperación de la información a partir de la representación estándar, el uso de aplicaciones genéricas para la creación, manipulación y visualización de la información sobre los recursos, y para que los expertos en el dominio sean los encargados de suministrar y evaluar la información sobre los recursos.

En esta tesis de maestría, se propone una metodología para la representación, búsqueda, explotación e integración del conocimiento de los recursos de información en una memoria corporativa, mediante el uso de tecnologías semánticas. Esta metodología está guiada por

dos casos de uso base y la memoria corporativa es del área de RyT de la UAM.

- El primer caso de uso (Cartografía de competencias) consiste en la búsqueda de las personas (adscritas o relacionadas al depto. IE) a partir de sus características profesionales. En particular, se buscan a las personas por las competencias de profesionales, lingüísticas y sobre los temas que conocen de Redes y Telecomunicaciones. Por ejemplo, "todos los profesores de la UAM con conocimientos en radios cognitivos y que lean en inglés". Este primer caso también contempla la búsqueda de profesores que pueden impartir un curso, a partir de un conjunto de temas básicos que debe saber para dicho curso.
- El segundo caso de uso (Búsqueda de recursos digitales) consiste en la búsqueda de documentos y archivos multimedia, con base a uno o varios criterios de búsqueda (autor, título, año, temas de RyT, entre otros). Por ejemplo, "todos los artículos de Tim Berners Lee sobre Web Semántica y mayores al 2009".

La metodología para el desarrollo del modelo, la explotación y la integración del conocimiento sobre los recursos en una MC, se ha dividido en varias etapas que concuerdan con cada uno de los objetivos de la tesis. Los objetivos de la tesis son los siguientes:

- Un modelo (representación del conocimiento) de los recursos a partir de los dos casos de uso en un formato estándar.
- Un modelo coherente y del cual se explote el conocimiento sobre los recursos (ontología), a partir del uso de axiomas y un programa razonador.
- La búsqueda y recuperación (integración) de los recursos que satisfagan las necesidades informativas de los usuarios, a partir de un motor de consulta.
- Un prototipo (navegación y consultas específicas) para la interacción fácil y visual de los usuarios con el modelo .
- Evaluar los resultados devueltos y el tiempo de ejecución de las consultas a la ontología.

En las tecnologías de la web semántica, el marco de descripción de recursos (RDF) es la solución para la representación del conocimiento de manera formal sobre los recursos en la MC. La representación se basa en la descripción de las características significativas o relaciones semánticas de/entre los recursos. Por ejemplo, Jorge Aparicio Reyes tiene 29 años, vive en el Estado de México, lee en Inglés, conoce a Erik Alarcón, estudia en la UAM y tiene conocimientos en sistemas operativos, java y flash.

Si bien cada recurso de la MC tiene un nombre propio, en el marco RDF cada persona, documento, multimedia o concepto tiene un identificador único de recurso [2] (URI). Con la finalidad de no tener ambigüedades a la hora de referirse a un recurso. Por ejemplo, el URI de Jorge Aparicio es <http://www.mi-ejemplo.com/JorgeAparicio>. Para cada recurso

(identificado con URI) se describen las características/relaciones en forma de triples (sujeto-predicado-objeto) y cada elemento de un triple es un URI o en algunos casos el objeto es una Literal.

Esta representación de las características se encuentra en un formato estándar y para almacenar estos triples, se emplea un triplestore. En este trabajo de tesis se empleó el triplestore Apache Jena que proporciona almacenamiento, un motor de consulta y un razonador.

Las descripciones representan la información explícita de los recursos, pero, esta información explícita tiene conocimiento implícito. Por ejemplo, un alumno, niño, profesor, empleado, madre, hijo son personas, pero éstas como tal no tienen un triple que establezca que son personas. Entonces, para explotar este conocimiento implícito de los recursos, se proponen un conjunto de reglas o axiomas que permiten establecer estas relaciones. Aunque, para materializar estos triples a partir de los axiomas, es necesario un programa razonador que infiera estos triples. Este razonador también permite encontrar inconsistencias en el modelo. Algunos triplestores integran o permiten importar un razonador, en el caso de Jena permite las dos opciones.

El modelo que captura el conocimiento explícito (descripciones) de los recursos y los axiomas que completan el conocimiento sobre éstos, se denomina ontología. En esta tesis se hicieron dos ontologías; una para cada caso de uso, y también se modificó una ontología legada que tiene conceptos del área de RyT. Esta última ontología se emplea para vincular a personas, documentos y multimedia con los tópicos de RyT.

La consulta de los triples en el modelo, ya sea únicamente descripciones (triples explícitos) o una ontología con razonador (triples explícitos e inferidos), se hace con un motor de búsqueda (integrado en el triplestore) que compara los triples con un conjunto de patrones; aquellos triples que concuerden, se recuperará la información que se solicitó en la consulta.

Un motor de consulta y un razonador que materializa triples en una ontología, son una buena combinación, ya que permiten consultar el conocimiento inferido (triples inferidos) y reducir la complejidad de las consultas. Por ejemplo, se tienen seis individuos que afirman que son alumno, niño, profesor, empleado, madre, hijo respectivamente, también se tienen los axiomas que establecen que alumno, niño, profesor, empleado, madre, hijo son personas y se tiene la siguiente pregunta "¿Quiénes son personas?". Si se emplea solamente un motor de búsqueda, entonces no habrá ningún resultado, pero si se emplea la combinación motor y razonador, los seis individuos serán respuesta, porque estos seis individuos tienen el triple que afirma que son personas.

Los usuarios del área de RyT no están familiarizados con las tecnologías semánticas y en particular, al uso de la sintaxis de consulta. Entonces para facilitar a éstos la interacción y consulta del conocimiento de la ontología, se propone un prototipo que medie (interfaz) entre los usuarios y la ontología, específicamente este prototipo tiene los siguientes objetivos:

- Navegación a través de la información de los recursos; guiada por los casos de uso.
 - Estructurar la pregunta de un usuario.
 - Mapear las preguntas a consultas para el motor de consulta.
-

- Ejecutar la consulta con el motor de consulta, el razonador y la ontología.
- Publicar la información de los recursos respuesta en un formato visual agradable al usuario.

En esta tesis dos de los aspectos importantes a evaluar son: el desempeño de Apache Jena a la hora de consultar la ontología, así como el número y cuáles resultados responden estas consultas. Para llevar a cabo estas dos evaluaciones se obtuvieron un conjunto básico de preguntas para interrogar el modelo, para cada pregunta se sabe de ante manos el número y los recursos que la responden. En la primer evaluación, para cada consulta básica se calcula 20 veces el tiempo aproximado en milisegundos y se saca un tiempo promedio. Mientras, en la segunda evaluación, para cada consulta se compara el número/recursos que responde el motor con los recursos que previamente se sabe que la responden.

Las contribuciones de esta tesis son:

1. Una metodología para la Integración Semántica de Recursos en la MC de Redes y Telecomunicaciones.
 2. Identificación y descripción de los principales escenarios de búsqueda/recuperación de los recursos en la MC de RyT.
 3. Ontologías (Triples RDF + axiomas) que capturan el conocimiento de los recursos (apegados a los dos casos de uso) en la memoria corporativa RyT.
 4. Prototipo para la consulta interactiva de los usuarios con las ontologías de RyT.
 5. Evaluación del desempeño y calidad de resultados del triplestore Jena para la consulta de información.
-

Agradecimientos

Contenido

| | |
|---|-------------|
| Lista de Tablas | XI |
| Lista de Figuras | XIII |
| Acrónimos | XV |
| 1. Introducción | 1 |
| 2. Descripción del Problema | 7 |
| 2.1. Memoria Corporativa | 8 |
| 2.1.1. Administración de una Memoria Corporativa | 9 |
| 2.1.2. Naturaleza de una Memoria Corporativa | 11 |
| 2.1.3. Integración del Conocimiento | 12 |
| 2.2. Casos de uso | 13 |
| 2.2.1. Cartografía de Competencias | 14 |
| 2.2.2. Búsqueda de Recursos Digitales | 14 |
| 2.3. Hipótesis | 15 |
| 3. Tecnologías Semánticas | 17 |
| 3.1. Introducción y definiciones | 17 |
| 3.2. Marco de Descripción de Recursos | 17 |
| 3.3. Lenguaje de consulta sobre grafos RDF (SPARQL) | 21 |
| 3.4. Reglas de inferencia (RDF(S)/OWL) y razonadores | 21 |
| 3.5. Ventajas de las tecnologías Semánticas | 25 |
| 4. Estado del arte | 31 |
| 4.1. Integración semántica de recursos de información | 32 |
| 4.2. Herramientas para la integración semántica de recursos | 35 |
| 5. Integración semántica de recursos de información en una memoria corporativa | 39 |
| 5.1. Representación del conocimiento en los recursos | 40 |
| 5.2. Enriquecimiento del conocimiento en el modelo | 40 |

| | |
|--|-----------|
| 5.3. Búsqueda y recuperación de información en el modelo | 40 |
| 6. Prototipo | 41 |
| 7. Evaluación experimental | 43 |
| 7.1. Escenarios de experimentación | 43 |
| 7.2. Experimentación | 43 |
| 7.3. Resultados | 43 |
| 8. Conclusiones y Trabajo Futuro | 45 |
| Appendices | 49 |
| Apéndice A. Códigos interfaz de Usuario | 51 |
| Bibliografía | 53 |

Lista de Tablas

| | | |
|------|---|----|
| 3.1. | Ejemplos de identificadores (URI) asignados a distintos recursos. | 18 |
| 3.2. | Ejemplos de identificadores asociados a distintas propiedades. | 18 |
| 3.3. | Ejemplos de tripletas que emplean la propiedad <i>rdf:type</i> para asignar un recurso a una determinada clase. | 20 |
| 4.1. | Criterios considerados para el <i>estado del arte</i> de la integración semántica de recursos. | 31 |
| 4.2. | Comparativa entre los trabajos estudiados y nuestros criterios para la integración semántica de recursos. | 34 |

Lista de Figuras

| | | |
|------|--|----|
| 2.1. | Diagrama de casos de uso para la integración de los recursos de una memoria corporativa. | 13 |
| 3.1. | Ejemplos de tripletas asociadas a las declaraciones para los recursos Juan y libro de matemáticas discretas. | 19 |
| 3.2. | Ejemplo de un grafo RDF o grafo de conocimientos. | 20 |
| 3.3. | Estructura básica de una consulta SPARQL. | 21 |
| 3.4. | Regla para indicar que un Metal-Líquido pertenece a las clases Metal y Líquido. | 25 |
| 3.5. | ABox y TBox para ejemplificar el beneficio de utilizar un razonador y un motor de búsqueda. | 29 |
| 3.6. | Consulta SPARQL para recuperar todos los individuos que son personas. | 29 |
| 3.7. | Ontología con tripletas que han sido inferidas mediante un razonador. | 29 |

Acrónimos

| Acrónimo | Descripción | Definición |
|----------|--|------------|
| RyT | Redes y Telecomunicaciones | 7 |
| IE | Ingeniería Eléctrica | 7 |
| UAMI | Universidad Autónoma Metropolitana Unidad Iztapalapa | 7 |
| MC | Memoria Corporativa | 8 |
| MO | Memoria Organizacional | 8 |
| TI | Tecnologías de la Información | 10 |
| GBDR | Gestor de Bases de Datos Relacional | 10 |
| BD | Base de Datos | 10 |
| TS | Tecnologías Semánticas | 17 |
| ABox | Componente Asertivo | 23 |
| TBox | Componente Terminológico | 23 |
| RDF | Resource Description Framework | 17 |
| URI | Identificador Único de Recursos | 18 |
| W3C | World Wide Web Consortium | 21 |
| RDF(S) | Schema RDF | 23 |
| OWL | Web Ontology Languages | 23 |
| FOAF | Friend Of A Friend | 27 |
| XML | Lenguaje de Marcado eXtensible | 33 |
| GUI | Interfaz Gráfica de Usuario | 35 |
| IDE | Entorno de Desarrollo Integrado | 35 |
| API | Interfaz de Programación de Aplicaciones | 37 |
| ISR | Integración Semántica de los Recursos | 39 |

Capítulo 1

Introducción

Las personas todos los días están en contacto con diferentes organizaciones. Por ejemplo, el niño que asiste a la **escuela primaria**, el estudiante que asiste a la **universidad**, la ama de casa que compra productos en una **tienda departamental**, la persona que hace un depósito o cobrar en una **institución bancaria**, la personas que solicita un servicio en alguna **dependencia gubernamental**, el empleado trabaja en una **empresa**, inclusive una **familia** es una organización.

El concepto de organización tiene diferentes definiciones, nosotros elegimos la siguiente definición: “*una organización es una entidad a través de la cual las personas realizan actividades y de las cuales por lo menos algunas se dirigen a la consecución de fines comunes (metas) de las personas del grupo*” [3]. De esta definición, se tiene que una organización alcanza mayores logros, porque varias personas se coordinan y dirigen sus esfuerzos conjuntamente. Las organizaciones deben poner atención en las siguientes actividades para alcanzar sus metas y objetivos [4]:

1. Reunir recursos para alcanzar las metas y los resultados deseados.
2. Producir bienes y servicios de manera eficiente.
3. Buscar formas innovadoras de producir y distribuir con mayor eficiencia bienes y servicios.
4. Utilizar tecnologías de información y manufactura.
5. Adaptar, evolucionar e influir en un entorno que cambia con rapidez.
6. Crear valor para dueños, empleados y clientes.
7. Hacer frente y adaptarse a los cambios que plantea la diversidad del mundo laboral, problemas éticos, responsabilidad social y coordinación de los empleados.

La **administración** es un concepto importante para una organización y éste se define como: “un conjunto de actividades dirigido a aprovechar los recursos de manera eficiente y eficaz con el propósito de determinar y alcanzar los objetivos de la organización” [5]. A partir de esta definición, se tienen dos elementos importantes: actividades y recursos. Las

actividades en una organización pueden ser *búsqueda de información, almacenamiento de los recursos, intercambio de información, control de bienes y materiales, control de inventario, colaboración con otras personas, solo por mencionar algunas*. Mientras, los **recursos** son “el medio que posee una organización para realizar las actividades que le permitan lograr los objetivos” [6]. Una organización puede tener los siguientes recursos: materiales o físicos, humanos (personas), financieros (dinero) e informáticos. La finalidad de la administración en una organización es que ésta sea estable, crezca y prospere.

La *administración en una organización* tiene diferentes enfoques que dependen de los principales elementos de la misma, por ejemplo: las metas, el proceso interno y los recursos. En particular, nuestro foco de atención son los recursos de información. Porque éstos son los instrumentos que representan y encapsulan el conocimiento de una organización. Algunos ejemplos de estos recursos son: una persona, una base de datos, un libro, un archivo multimedia, informes anuales, un equipo de cómputo, un servidor, por mencionar algunos. De esta manera, el enfoque para esta tesis es *con base en recursos* y éste se define como: “la capacidad de la organización para adquirir recursos valiosos, integrarlos y administrarlos exitosamente” [5].

La administración de los recursos puede realizarse con alguna herramienta de las Tecnologías de la Información. La finalidad de estas herramientas es facilitar, efficientar y agilizar las actividades relacionadas a la administración de los recursos. Los dos enfoques mediante el uso de estas tecnologías son: el manual y automático. Por un lado, el enfoque manual consiste en almacenar y organizar los recursos digitales (documentos, archivos de audio, presentaciones, documentos escaneados, etc) en carpetas que tienen cierta estructura. Por otro lado, el enfoque automático permite delegar ciertas tareas de gestión a programas computacionales; las dos herramientas comunes de este enfoque son: *sistemas gestores de bases de datos relacionales* o *motores de búsqueda basados en keywords*. Un *motor de búsqueda* [7] es un sistema de recuperación de la información que a partir de las palabras clave, realiza una búsqueda documental. Este motor responde al usuario con aquellos documentos que tienen las palabras clave en su contenido. Por otro lado, un *gestor de bases de datos relacional* es un mecanismo para el almacenamiento y recuperación de la información sobre una Base de Datos. Estos gestores se basan en esta idea: *la base de datos es percibida como un conjunto de tablas (relaciones) bajo un mismo contexto, donde, una tabla es una matriz que guarda datos* [8]. Un gestor emplea un *esquema conceptual* para las tareas de almacenamiento de información. El esquema permite describir un conjunto de objetos, aspectos relevantes y las interrelaciones de/entre estos, así como restricciones de integridad. Mientras, para fines de recuperación de la información, se emplean lenguajes de consulta sobre las bases de datos.

El enfoque manual y las dos herramientas del enfoque automático tienen algunos detalles que dificultan la gestión en los recursos de una organización. En el caso de una solución manual, si hay un crecimiento explosivo de los archivos (recursos digitales), entonces la búsqueda de recursos se vuelve un proceso tardado, pesado y cansado para las personas. Mientras, las dificultades de las dos herramientas son: 1) *un motor de búsqueda en ocasiones recupera documentos innecesarios para los usuarios*, 2) *un motor proporciona resultados inadecuados por problemas de ambigüedad en las palabras*, 3) *una representación deficiente en una BD*

relacional, puede causar anomalías en los datos encontrados cuando el modelo crece, 4) un modelo relacional inadecuado propicia a tener datos inconsistentes, lo que provoca, problemas en la generación y validación de la información [8], y 5) pérdida de información en el modelo cuando se representan las atributos de los recursos [8].

Las tecnologías semánticas [9] son un conjunto de metodologías, lenguajes, aplicaciones, herramientas y estándares, para obtener y suministrar el significado de la información ¹. Estas tecnologías permiten la representación y la administración del conocimiento, por ello, son una solución interesante para la administración de los recursos en una organización. A continuación, se presentan los beneficios del uso de las mismas:

- **Formato estándar:** una persona, documento, objeto físico o digital, concepto, idea, en general, cualquier recurso posee información significativa y útil para las personas. Esta información puede estar embebida en el recurso o es referente a éste, por ejemplo, en un libro nos interesa saber sobre qué trata, el título, los autores, la fecha de edición, entre otros. Por otro lado, los datos de los recursos pueden ser de distintas formas: estructurados (bases de datos), semiestructurados (lenguajes de etiquetas, como XML y HTML) o sin estructura (orientados al texto). También, cada recurso puede estar almacenado en distintos tipos de archivo, por ejemplo, un documento digital puede ser un doc, pdf, odp, rtf, etc. Esta diversidad en los recursos hace difícil la administración de los mismos. Por ello, las tecnologías proponen representar los recursos a través de sus características significativas en un formato estándar, para que, los procesos automáticos puedan acceder, procesar, razonar, combinar, reutilizar y compartir esta información.
- **Enriquecer el conocimiento:** Las tecnologías semánticas permiten la introducción de reglas de inferencia para enriquecer el modelo de conocimiento implícito. La finalidad de estas reglas es que un programa especial realice inferencia sobre éstas para hacer explícito el conocimiento implícito. De esta manera, los procesos automáticos pueden aprovechar este conocimiento, para fines de búsqueda de la información. Por ejemplo, una persona, un perro y un gato pertenecen al campo semántico mamíferos, si se introduce la regla que establece que todo gato, perro o persona es un mamífero, entonces, un proceso automático podrá identificar quienes son mamíferos.
- **Flexibilidad e interoperabilidad:** una característica importante en las tecnologías semánticas es la flexibilidad. Esta característica se refiere a la facilidad para representar y mantener el conocimiento de un dominio. Esta representación se basa en la descripción de los recursos a partir de sus características significativas y relaciones en un formato estándar. Otra característica relacionada a la flexibilidad, es la interoperabilidad. Este concepto se refiere a que gracias a los estándares pueden emplearse una variedad de herramientas y aplicaciones.

¹L. Feigenbaum, "Semantic Web vs. Semantic Technologies", Disponible en: <http://www.cambridgesemantic.com/semantic-university/semantic-web-vs-semantic-technologies>

Existen distintos tipos de organizaciones que dependen del enfoque con el que se mira. Si es con respecto al alcance, se tienen corporaciones multinacionales, pequeños y medianos negocios, así como negocios familiares. Cuando el enfoque es el objeto final, se tienen organizaciones que fabrican productos o proveen servicios. Si es a partir de la naturaleza de la organización, se tienen instituciones económicas (empresas), fundaciones, organizaciones sin fines de lucro e instituciones públicas.

Esta tesis de maestría se enfoca en las organizaciones de investigación (institutos o universidades), porque tienen áreas o equipos de investigación. En concreto, la organización electa como caso de estudio es el grupo de investigación del **área de Redes y Telecomunicaciones** de la **Universidad Autónoma Metropolitana Unidad Iztapalapa**. Los recursos significativos en esta organización son: *personas (profesores, alumnos y colegas empleados), documentos (artículos científicos, libros, tesis), bases de datos, archivos multimedia (presentaciones, vídeos, imágenes), solo por mencionar algunos*. Porque representan el conocimiento de los profesores (miembros de esta organización) sobre sus investigaciones, colaboraciones, proyectos, actividades, cursos y temas de interés. Una adecuada administración de los recursos, se traduce en un grupo de investigación bien informado con mejores tomas de decisiones, así como una base de conocimiento persistente y accesible para los profesores y alumnos.

El principal objetivo de esta tesis es *ver la viabilidad del uso de las tecnologías semánticas para la construcción de una base de conocimiento, con la finalidad de integrar la información sobre los recursos del área de RyT de la UAM, para responder las necesidades informativas de los usuarios de RyT*. Mientras, los objetivos particulares son:

- Desarrollar una metodología para la integración semántica de recursos de información pertenecientes a un dominio particular.
- Determinar los casos de uso para esta integración semántica de los recursos.
- Representar y enriquecer el conocimiento de los recursos de información (identificados en los casos de uso) del dominio de redes y telecomunicaciones de la UAM en una o varias representaciones de conocimiento.
- Implementar un prototipo de interfaz de usuario que permita a éstos últimos una interacción amigable para la integración semántica de los recursos.
- Evaluar los resultados devueltos en la integración semántica para el dominio de redes y telecomunicaciones.

Mientras, las contribuciones de tesis son las siguientes:

1. Metodología para la integración semántica de recursos en un dominio particular.
 2. Identificar los principales casos de uso para la integración semántica (cartografía de competencias y búsqueda de recursos digitales)
-

3. Estados del arte para la integración semántica de recursos y las herramientas semánticas.
4. Tres modelos de conocimiento que representan el conocimiento e información en los recursos del dominio de redes y telecomunicaciones.
5. Prototipo de interfaz de usuario para la interacción amigable de los usuarios con los modelos de conocimientos del dominio de redes y telecomunicaciones.
6. Evaluación del desempeño en el proceso de consulta y evaluación de la precisión de los resultados del gestor de modelos semánticos con nuestros modelos de conocimiento.

Al organizar esta tesis, hemos querido establecer un camino coherente para alcanzar cada uno de los objetivos planteados. Los capítulos se organizan de la siguiente manera:

El capítulo 2 se describe la problemática principal de esta tesis, la cual es la integración semántica de recursos en una memoria corporativa, así como algunos conceptos básicos (memoria corporativa, integración, recurso). Los principales conceptos, definiciones, estándares de los elementos pertenecientes a las tecnologías semánticas, se presentan en el capítulo 3. En el capítulo 4 se presentan los dos estados del arte: el primero es sobre la integración semántica de los recursos en una memoria corporativa, mientras el segundo es sobre las herramientas para la generación de triples, editores de ontologías y triplestore. El capítulo 5 describe nuestra metodología para la integración semántica de recursos en una memoria corporativa. El capítulo 6 describe los objetivos y características del prototipo para la integración semántica de recursos. Las pruebas y resultados (desempeño y calidad de las respuestas) hechos/obtenidos al gestor del modelo semántico, así como al modelo para el área de redes y telecomunicaciones, se presentan en el capítulo 7. Finalmente, las conclusiones sobre la integración semántica de los recursos, el uso de las tecnologías semánticas y los resultados de nuestra experimentación, se presentan en el capítulo 8. En esta sección también se presentan algunos trabajos futuros que identificamos.

Capítulo 2

Descripción del Problema

El *área de Redes y Telecomunicaciones* (RyT) es una de las cinco áreas académicas en que se organiza el departamento de Ingeniería Eléctrica (IE) de la Universidad Autónoma Metropolitana Unidad Iztapalapa (UAM-I). En esta área se cultivan las siguientes líneas de investigación: *Redes y Servicios de Telecomunicaciones, Sistemas de Comunicación Digital, Sistemas Distribuidos y Web Semántica*.

El área de RyT es una organización que se constituye por un conjunto de personas. Ellas desempeñan las actividades de investigación, académicas, preservación y difusión de la cultura. Las personas de RyT pueden ser clasificadas en dos tipos: las que pertenecen al núcleo del área y las temporales. Las personas del núcleo del área son los **profesores-investigadores**. Ellos se encargan de realizar las siguientes actividades: *planear, definir, dirigir, coordinar y evaluar los cursos de las licenciaturas en Computación, Ingeniería Electrónica, Posgrado en Ciencias y Tecnologías de la Información, investigación, así como la investigación y desarrollo de proyectos asociados a sus líneas de investigación*. Ahora bien, las **personas temporales** trabajan con el personal del núcleo, ya sea en la investigación, colaboración, ayuda o servicios administrativos. Estas personas tienen un rol menos activo en el área, porque el tiempo en que trabajan es un periodo corto. Algunos ejemplos de este tipo de personas son: 1) *estudiantes que realizan algún proyecto o servicios social y cuyo responsable de ellos es un profesor del núcleo*, 2) *profesores temporales que imparten cursos relacionados con los temas de Redes y Telecomunicaciones*, 3) *empleados de la universidad que proporcionan servicios administrativos a los profesores del núcleo* y 4) *empleados de otras organizaciones que colaboran con los profesores del núcleo*.

En cuanto a la cantidad de personas involucradas en el área RyT. Para el núcleo se tienen trece profesores-investigadores. Mientras el número de personas temporales, que han participado o participan con las personas del núcleo, no hay un número exacto de éstas. Porque cada profesor-investigador tiene su lista de personas conocidas (estudiantes y colegas) y cada trimestre estas listas se van incrementando.

El conjunto de personas del área es el elemento más importante para ésta. Porque ellas realizan las actividades para lograr las metas y objetivos del área de RyT. Las personas al realizar sus actividades cotidianas y estructuradas, se convierten en las constructoras del conocimiento para la organización. Las etapas para la construcción del conocimiento son la *adquisición y representación*.

1. Las personas consiguen y hacen propio el conocimiento por distintas maneras, como: *la experiencia, al realizar sus actividades cotidianas; la observación, análisis, experimentación, evaluación y en general por distintas actividades de la investigación; la búsqueda, obtención, almacenamiento, recopilación, lectura, visualización y consulta de distintos soportes (documento, imagen, audio, vídeo); enseñanza y aprendizaje entre personas; por mencionar algunas*. Estas personas utilizan este conocimiento para ejecutar sus actividades y tareas en la organización.
2. Las personas realizan dos actividades con el conocimiento: *1) mantener el conocimiento en su mente o 2) hacer presente el conocimiento con palabras, imágenes, sonidos, símbolos en algún soporte como documento, imagen, audio, presentación, base datos, hoja de cálculo o vídeo*. En la primera actividad, la *representación del conocimiento es intangible*, como habilidades, destrezas profesionales, conocimiento privado o el conocimiento de la organización. La finalidad este conocimiento es que las personas sean instrumentos de conocimiento para realizar determinadas tareas o solucionar problemas específicos en la organización. Mientras, en la segunda actividad la *representación del conocimiento es tangible*. La finalidad de esta representación es que los objetos (recursos inanimados) conserven y transmitan la información a las personas de la organización.

Personas y recursos inanimados se agrupan bajo el concepto de **recurso de información o conocimiento**. En el área de RyT, los recursos de información son el conocimiento de *investigaciones, colaboraciones, proyectos, cursos, temas de interés, objetos, ideas o conceptos vinculados con los **tópicos de Redes y Telecomunicaciones***. Esta área tiene las siguientes clases de recursos: *artículos científicos, presentaciones, libros, equipos de cómputo, bases de datos, tesis, reportes técnicos, audios, vídeo tutoriales, notas de curso, tareas, imágenes, páginas web, profesores, estudiantes, empleados de otras organizaciones, servidores computacionales, programas y aplicaciones computacionales científicas-académicas*.

2.1. Memoria Corporativa

Los recursos de información expresan el conocimiento en la organización. A este conocimiento se denomina memoria corporativa (MC) o memoria organizacional (MO). Una definición formal de este concepto es la siguiente: *“una memoria corporativa es la representación explícita, tácita, consistente y persistente del conocimiento en una organización”* [1]. Por explícita, se refiere a que el conocimiento se expresa de manera clara y formal. Representación tácita significa que ciertas partes del conocimiento no se mencionan formalmente, sino que deben inferirse; por ejemplo, una mujer y un hombre son personas. Por consistente, se traduce en que el conocimiento es estable y no sufre grandes cambios. Persistente, es una cualidad temporal y se refiere a que el conocimiento debe durar por un tiempo prolongado.

Una memoria corporativa conserva y mantiene el conocimiento de una organización [10], con la finalidad de *facilitar el acceso, intercambio y difusión del mismo*. De esta manera, las personas adscritas o interesadas en la organización podrán *adquirir, reutilizar y razonar* con

este conocimiento y realizar nuevas actividades o mejorarlas. Por ejemplo, aportar nuevas ideas, modificar ciertos aspectos en su trabajo, colaborar e intercambiar puntos de vista con sus colegas, abarcar otros mercados, generar mayor conocimiento, actualizar la información, por mencionar algunas.

En una organización existen distintas razones para tener una memoria corporativa. Rose Dieng et al. proponen una lista básica de razones [10]:

- Prevenir la pérdida del conocimiento de los expertos, cuando éstos salgan de la organización.
- Aprovechar las experiencias buenas y malas de trabajos pasados, con la finalidad de mejorar el trabajo y no caer en los mismos errores.
- Aprovechar el conocimiento global para mejores tomas de decisión en la organización.
- Mejorar las capacidades de la organización para reaccionar y adaptarse a los cambios.
- Mejorar la circulación de la información y la comunicación entre las personas de la organización.
- Mejorar el aprendizaje de las personas en la organización.
- Integrar el conocimiento fundamental de una organización, como flujos de trabajo, productos, técnicas, información secreta.

2.1.1. Administración de una Memoria Corporativa

Una memoria corporativa (MC) es uno de los principales elementos para una organización y las personas adscritas o interesadas en ésta, por esta razón, es importante la **administración de la memoria corporativa**. La *administración* es un concepto interesante para las organizaciones. Este concepto se define como: “*un conjunto de actividades dirigidas a aprovechar los recursos de manera eficiente y eficaz, con el propósito de determinar y alcanzar los objetivos en la organización*” [5].

La administración del conocimiento es un problema complejo que puede ser abordado de distintos enfoques: financieros y económicos, técnicos, metas, proceso interno, entre otros. En particular, el conocimiento prioritario para esta tesis son los **recursos de información**: 1) *elementos tangibles* como datos, procedimientos, planes, documentos, audios, vídeos, presentaciones, tesis, libros, por mencionar algunos y 2) *elementos intangibles* como habilidades, destrezas profesionales, conocimiento privado y el conocimiento del contexto en la organización.

Los **objetivos** en la administración de una memoria corporativa son: *integrar el conocimiento disperso en la organización, preservar y difundir el conocimiento, facilitar el acceso y visibilidad del conocimiento, tener con un instrumento para el aprendizaje, facilitar la búsqueda y recuperación del conocimiento, promover la comunicación y cooperación entre personas,*

emplear un lenguaje técnico entendido por todas las personas, promover el crecimiento e intercambio del conocimiento, facilitar la compartición de nuevas ideas, mejorar las tomas de decisión, por mencionar algunos.

Un analogía de la administración de los *recursos de información* se presenta a continuación. Una biblioteca es una organización dedicada a la *adquisición, conservación, exposición y préstamo* de libros. Para llevar a cabo estas tareas, la biblioteca realiza distintas actividades de administración con los libros. Las actividades básicas en la administración de los libros son: *caracterizar los libros, generar las fichas bibliográficas, clasificar las fichas de acuerdo a ciertos parámetros, asignar un identificador a cada libro, acomodar el libro de acuerdo a la clasificación y al identificador, generar un catálogo de todos los libros; consultar el catálogo, retirar el libro del estante, dar de baja un libro en el catálogo, indicar a quién se le presta el libro, indicar una fecha de devolución; dar de alta el libro en el catálogo y regresar el libro a su ubicación.* Este flujo de actividades las podemos agrupar en seis actividades generales: **representar, almacenar, clasificar, consultar, recuperar y actualizar.**

Una memoria corporativa debe administrar los recursos de información, de manera semejante a como, una biblioteca administra los libros. En la actualidad, la administración de los recursos se hace mediante el uso de las **Tecnologías de la Información** (TI). Estas tecnologías proporcionan un conjunto de herramientas, enfoques y aplicaciones para facilitar, agilizar y automatizar distintas actividades o procesos.

En el área de Redes y Telecomunicaciones (RyT), la administración del conocimiento se hace de manera individual, es decir, cada profesor, estudiante o empleado administra sus recursos de información. Porque cada persona tiene intereses particulares (líneas de investigación) y emplea la herramienta que más le conviene. Estas personas administran sus recursos mediante dos enfoques:

- El **enfoque manual** consiste en almacenar los recursos de información (recolectados o generados) en carpetas organizadas. Estas carpetas están estructuradas de forma jerárquica y cada recurso tiene un nombre significativo. Las personas para recuperar los recursos, tienen que buscar en las carpetas e identificar el recurso con base al nombre o al contenido de éste.
- El **enfoque automático** consiste en emplear alguna aplicación para automatizar el almacenamiento, búsqueda y recuperación de los recursos. Los profesores emplean como aplicaciones a motores de búsqueda sintácticos basados en keywords y gestores de bases de datos relacionales. Los *motores de búsqueda sintácticos basados en keywords* (MBSK) hacen una búsqueda documental de acuerdo a las palabras (keywords) que un usuario escribe. Los resultados de esta búsqueda se presentan como un ranking de enlaces a los documentos fuente. Un motor de búsqueda no realiza actividades que se relacionan al almacenamiento de los documentos. Estos motores generan índices del contenido de los documentos, para facilitar el trabajo de búsquedas futuras. Mientras, un *gestor de bases de datos relacional* (GBDR) almacena, modifica y recupera la información en una base de datos (BD). La consulta de información se hace mediante un lenguaje de consulta

estructurado. Los resultados asociados a las consultas, se presentan en forma de tabla. Un GBDR necesita de esquema relacional para almacenar y actualizar la información en la base de datos.

Estos dos enfoques en la administración de recursos de información se aplican a fragmentos de la memoria corporativa. Sin embargo, todos los recursos de la memoria corporativa no se administran bajo un mismo enfoque. Ahora bien, *cuál es el enfoque o herramienta para aprovechar los recursos de manera eficiente y eficaz*. Para tomar esta decisión, deben ser analizadas: 1) las características de una memoria corporativa y 2) los beneficios de los distintos enfoques de las Tecnologías de información.

2.1.2. Naturaleza de una Memoria Corporativa

En una memoria corporativa, los recursos de información tienen distintas cualidades que deben considerarse para administrar el conocimiento de éstos. Porque estas cualidades pueden causar dificultades en etapas tempranas del proceso de administración. Esta tesis presenta las principales características a considerar en la gestión de una memoria corporativa. En particular, las características de la memoria del área RyT.

Diversidad en formato

Esta característica tiene que ver con los recursos digitales. En el área de RyT, los recursos digitales se clasifican de acuerdo al soporte (documento, audio, vídeo, presentación, imagen, base de datos y código). Los recursos pertenecientes a un determinado soporte, no tienen el mismo formato que otros recursos pertenecientes a otros soportes. Inclusive, recursos pertenecientes al mismo soporte, no necesariamente todos tienen el mismo formato. Esto se debe a la gran ***diversidad de formatos*** que emplean las aplicaciones como: *procesadores de texto, hojas de cálculo, editores de código, bases de datos, por mencionar algunas*. Por ejemplo, los recursos de información que son documentos tienen los siguientes formatos: *pdf, doc, txt, docx, odp, tex y html*. Idealmente, se podría pensar que todos estos recursos sean guardados con el mismo formato. Sin embargo, esto no sucede porque las personas emplean distintas aplicaciones computacionales. En la gestión del conocimiento se debe considerar esta *diversidad en formato* que cambien se denomina ***heterogeneidad en formato***.

Diversidad en Contenido

El conocimiento del área de Redes y Telecomunicaciones se clasifica en las cuatro líneas de investigación: *Redes y Servicios de Telecomunicaciones, Sistemas de Comunicación Digital, Sistemas Distribuidos y Web Semántica*. Cada línea tiene un conjunto de temas que se relacionan a ésta. Por ejemplo, la línea de Sistemas Distribuidos tienen los siguientes temas: *p2p, middleware, estado global, sistema operativo, replicación, concurrencia, sincronización, por mencionar algunos*.

En una memoria corporativa, un recurso en su contenido representa el conocimiento de uno o más temas de una línea de investigación. Por ejemplo, un conjunto de documentos

pueden tener el mismo formato, pueden pertenecer a la misma organización, pero éstos pueden representar distintos temas como: p2p, middleware o estado global. De esta manera, se puede afirmar que una memoria corporativa tiene una *variedad en el contenido de los recursos*. Esta diversidad también se conoce como *heterogeneidad en contenido*.

Diversidad en la Estructura

Los datos en los recurso digitales aparecen en distintas formas. Éstos se pueden clasificar en tres formas: 1) **datos estructurados**: *la información se apega a una estructura formal, como el modelo relacional en las bases de datos*, 2) **datos semi-estructurados**: *la información está contenida entre etiquetas para marcar el contenido de recurso*, y 3) **datos sin estructura**: *la información es orientada al texto*. Ejemplos de estos tres tipos son los siguientes: una base de datos con los datos de los profesores del área de RyT es ejemplo de datos estructurados, páginas web son ejemplos de datos semi-estructurados, notas de un curso son ejemplos de datos sin estructura.

Significado de la Información

Los recursos de información contienen palabras (escritas o habladas), símbolos lingüísticos, expresiones o situaciones, en general, información. Esta información usualmente puede ser entendida e interpretada sin ningún problema. Sin embargo, la naturaleza de nuestro lenguaje (escrito y oral) puede llevar a confusiones y malas interpretaciones. En particular, se puede tener dificultades con las siguientes cualidades de las palabras: *homonimia y la sinonimia*. La **homonimia** es la relación entre palabras que se escriben o pronuncian igual y tienen distinto significado. La **sinonimia** es la relación entre palabras que se escriben o pronuncian diferente y tienen el mismo significado. Un ejemplo de homonimia es la palabra radio, ya que esta palabra tiene distintos significados que se asocian a la Química, Comunicación, Anatomía o Geometría. Mientras, un ejemplo de sinonimia son las palabras resumen, sumario, síntesis y recapitulación.

2.1.3. Integración del Conocimiento

La *administración en una memoria corporativa (MC)* contempla varias actividades (representar, almacenar, clasificar, consultar, recuperar, actualizar, entre otras) que puede prolongar el tiempo y la complejidad de ésta. Además, en esta administración se debe contemplar las características de una memoria corporativa. Por estas razones se debe limitar el conjunto de actividades a una menor cantidad, es decir, ajustar el alcance de esta administración.

En la administración de una memoria corporativa existen distintos objetivos que son los elementos prioritarios, para alcanzar la finalidad de ésta (promover el acceso, intercambio y difusión de conocimiento). En particular, los siguientes objetivos prioritarios tienen una relación cercana: integrar el conocimiento disperso en la organización, facilitar el acceso y

visibilidad del conocimiento, tener con un instrumento para el aprendizaje y facilitar la búsqueda y recuperación del conocimiento.

El análisis de estos objetivos, nos lleva a un problema de integración de la información o del conocimiento. La **integración del conocimiento** es el proceso de representar y utilizar el conocimiento de un dominio dado (Memoria Corporativa), con el fin de llevar a cabo actividades de búsqueda, recuperación y combinación de la información de los recursos. Esta integración debe proporcionar información correcta a la consulta o pregunta del usuario.

2.2. Casos de uso

Esta tesis presenta la integración de la *memoria corporativa del área de RyT*. Los **principales usuarios** de la integración son: *los profesores-investigadores del área RyT, estudiantes de Computación y Electrónica, así como personas interesadas en el área (colegas de los profesores)*.

La memoria corporativa de RyT tiene una gran cantidad de recursos de información. Esto hace difícil las actividades de integración del conocimiento. Por ello, se propone descubrir y registrar los principales **casos de uso**. La finalidad de éstos, es *identificar las operaciones básicas o aspectos funcionales en la integración de los recursos, describir situaciones específicas, así como identificar los principales recursos de información y el contexto de éstos*.

En este trabajo, los casos de uso se identificaron a través del análisis de los principales recursos de información. Los principales recursos del área son **las personas y los recursos digitales**. De esta manera, los casos de uso identificados son: *Cartografía de competencias* para personas y *Búsqueda de recursos digitales*. La Figura 2.1 presenta el **diagrama de casos de uso**, en la cual, se ve la interacción entre los usuarios y los dos casos de uso.



Figura 2.1: Diagrama de casos de uso para la integración de los recursos de una memoria corporativa.

2.2.1. Cartografía de Competencias

El elemento dinámico en el área de RyT es el conjunto de personas que se clasifican en: *profesores, investigadores, estudiantes y empleados*. Estas personas tienen actitudes, valores, conocimientos técnicos, habilidades individuales y colectivas. Las caracterizadoras profesionales son importantes para la organización. Porque con base en éstas se pueden identificar las personas para: *realizar determinadas tareas, solucionar problemas específicos, hacer colaboraciones o tener un determinado cargo*.

La **cartografía de competencias** es la búsqueda y recuperación de las personas a partir de las características profesionales. Los principales parámetros en la búsqueda de estas personas son: las competencias profesionales (*trabajo en equipo, liderazgo, organizar, planificar*), conocimientos en temas de Redes y Telecomunicaciones (*sistemas operativos, capa enlace, filtros, ontologías, radios cognitivos*), capacidades lingüísticas (*lee en inglés, escribe en español, habla en francés*), relaciones profesionales (*colega, asesor o conocido*) y finalmente por la ocupación (*estudiante, empleado o profesor*).

Para cada *persona* identificada en la memoria corporativa, debe ser recuperada *información significativa* de ésta. La finalidad esto, es proporcionar al usuario mayor información, para que pueda localizar y contactar a la persona o filtrar los resultados de acuerdo a otros criterios (*sexo, edad, habilidades*).

2.2.2. Búsqueda de Recursos Digitales

En la memoria corporativa de RyT, los recursos digitales representan *ideas, objetos, teorías, procesos, flujos de trabajo y conocimiento estático de la organización* en un formato digital. Estos recursos se clasifican en: *artículos científicos, libros, reportes técnicos, páginas web, tesis, otros documentos, audios, vídeos, presentaciones, imágenes y otros archivos multimedia*. Las personas emplean a estos recursos como *objetos de aprendizaje*. Por esta razón, deben identificarse los recursos que solucionen las *necesidades informativas* de los usuarios.

La **búsqueda de recursos digitales** es la búsqueda y recuperación de los documentos y archivos multimedia a partir del contenido de éstos. Los principales parámetros de búsqueda de los recursos digitales son: el autor, la extensión (*ppt, wav, mp3, mpg, jpg*), relaciones con los temas de Redes y Telecomunicaciones (*sistemas operativos, capa enlace, filtros, ontologías, radios cognitivos*), el idioma fuente (*inglés, español, francés, ruso, chino*), tipo de recurso digital (*artículos, reportes técnicos, páginas web, tesis, libros, audios, vídeos, imágenes y presentaciones*) y la organización a la que pertenece (*uam, unam, ipn, iee, acm, oracle*).

Para cada *recurso digital* identificado en la memoria corporativa, debe recuperarse información significativa de éste, con la finalidad de proporcionar al usuario mayor información. De esta manera, el usuario verifica la importancia del recursos filtrar los resultados de acuerdo a otros criterios (*número de páginas, extensión, lenguaje fuente*).

2.3. Hipótesis

En este capítulo, se ha descrito de manera explícita el alcance, los principales elementos y características para la integración de los recursos en una Memoria Corporativa. Esta integración se ha planteado de manera genérica con respecto al uso de una determinada tecnología, con la finalidad de poder desarrollar la integración con algún enfoque, herramienta, metodología o aplicación de las Tecnologías de la Información.

Nosotros no elegimos alguna de las dos herramientas que ocupan las personas del área (MBSK y GBDR). En cambio, seleccionamos a las Tecnologías Semánticas como enfoque para solucionar esta integración. De esta manera, *nuestra hipótesis de investigación* para esta tesis es: *¿Acaso es posible usar a las Tecnologías Semánticas para solucionar la integración de los recursos en una memoria corporativa?*

Capítulo 3

Tecnologías Semánticas

3.1. Introducción y definiciones

La *semántica* [11] es un subcampo de la lingüística que determina la relación entre palabras y el significado de éstas; así como el estudio de cómo las palabras, frases y otros símbolos lingüísticos, se relacionan entre sí para formar un significado estructurado.

Las *tecnologías semánticas* (TS) [9] son *un conjunto de metodologías, lenguajes, aplicaciones, herramientas y estándares para suministrar u obtener el significado de las palabras, información y las relaciones entre éstos*¹. En estas tecnologías existen varios enfoques para la aplicación del concepto. Estos enfoques se agrupan en dos categorías: 1) *mejorar las capacidades de los procesos automáticos para analizar y comprender el lenguaje* y 2) *técnicas para describir formalmente las palabras, información y el conocimiento para un dominio especializado*.

La categoría para la integración de la información (búsqueda) es la segunda (*técnicas para describir formalmente el conocimiento*), porque al describir formalmente la información y el conocimiento en los recursos, se crea una *capa de conocimiento* en los recursos. La finalidad de esta capa es que los procesos automáticos puedan *acceder, procesar, razonar, combinar, reutilizar y compartir la información y su significado* [12]. De esta manera, se podrá mejorar la búsqueda de información, ya que se evitan problemas de ambigüedad y las personas obtendrán resultados más significativos de acuerdo al contexto del dominio dado.

3.2. Marco de Descripción de Recursos

Las *tecnologías semánticas* proponen al *marco de descripción de recursos* (RDF²) como *marco de trabajo para representar el conocimiento e información acerca de los recursos en un formato estándar* [13]. La finalidad de *expresar este conocimiento (modelar)* es *proveer a los recursos con un significado que sea comprensible por los procesos automáticos*. Mientras,

¹L. Feigenbaum, “Semantic Web vs. Semantic Technologies,” Disponible en: <http://www.cambridgesemantics.com/semantic-university/semantic-web-vs-semantic-technologies>

²W3C, “RDF 1.1 Concepts and Abstract Syntax,” Disponible en: <http://www.w3.org/TR/rdf11-concepts/>

la finalidad de un *formato estándar* es tener un formato compatible y universal para que los procesos automáticos interpreten, mezclen y compartan la información.

En el marco RDF se tienen tres conceptos claves [14]: **1) Recurso**, **2) Propiedad** y **3) Sentencia**.

El **recurso** es cualquier *persona, lugar, documento, página web, objeto abstracto o físico* que se quiera representar. Cualquier recurso en rdf debe tener un identificador único de recursos (URI), para distinguirlo de otros. Un URI es “*una cadena compacta de caracteres que proporciona un medio simple y extensible para la identificación de un recurso*” [15]. En la Tabla 3.1, se muestran algunos identificadores URI para cinco recursos.

| Recurso | Identificador (URI) |
|-------------|---|
| Juan López | http://www.mi-ejemplo.com/Juan_Lopez |
| UAM | http://www.mi-ejemplo.com/UAM |
| kitty | http://www.mi-ejemplo.com/kitty |
| celda solar | http://www.mi-ejemplo.com/celda_sol |
| Mamífero | http://www.mi-ejemplo.com/mamifero |

Tabla 3.1: Ejemplos de identificadores (URI) asignados a distintos recursos.

La **propiedad** es un *aspecto significativo, característica, metadato (datos de datos) o relación* que se describe del recurso. Por ejemplo, en una persona los metadatos y relaciones interesantes son: *nombre, edad, teléfono, correo electrónico, habilidad lingüística, nivel de estudios o relación amistad*; en un libro los metadatos interesantes son: *título, autor, isbn, resumen, edición, editorial, año de publicación, volumen o referencia*.

Estas propiedades indican acción entre dos recursos, por ello, es común que el *nombre de una propiedad* empiece por un verbo. Estas propiedades se identifican con URI y deben tener un significado bien definido, para expresar sin ambigüedad su funcionalidad. En la Tabla 3.2 se ejemplifican las propiedades asociadas a determinados metadatos.

| Metadato/Relación | Propiedad (URI) |
|-------------------|---|
| nombre | http://www.mi-ejemplo.com/tiene-nombre |
| conocido | http://www.mi-ejemplo.com/conoce-a |
| autor | http://www.mi-ejemplo.com/tiene-autor |
| referencia | http://www.mi-ejemplo.com/refiere-a |

Tabla 3.2: Ejemplos de identificadores asociados a distintas propiedades.

Los identificadores URI de los recursos y propiedades son cadenas con una longitud larga. Para abreviar estas cadenas se emplea un **prefijo**. Un prefijo sustituye la secuen-

cia de caracteres desde *http://* hasta el comienzo del **nombre del recurso o propiedad** por una **abreviación**. Por ejemplo, el prefijo “*exp*” es la abreviación de esta URI: *http://www.mi-ejemplo.com/*. De esta manera, los recursos y propiedades *Juan Lopez*, *Mamífero*, *nombre y conocido* de las Tablas 3.1 y 3.2 se escriben de la siguiente manera.

- *exp:Juan_Lopez*
- *exp:Mamifero*
- *exp:tiene-nombre*
- *exp:conoce-a*

La **declaración** [13] (sentencia o descripción) es una **afirmación de un hecho explícito** de un recurso, en términos de una **propiedad de objeto o dato** y el **valor** asignado a ella (otro recurso o literal). Estas declaraciones representan el conocimiento o información explícita de los recurso. La forma básica para escribir una declaración, es la **tripleta** [16]. La notación de una tripleta es: *sujeto-predicado-objeto*.

1. **Sujeto** es el recurso que se describe.
2. **Predicado** es la propiedad.
3. **Objeto** es otro recurso o una literal (cadena o entero) que describe el predicado.

En la Figura 3.1 se ejemplifican las tripletas que están asociadas a las siguientes declaraciones: 1) *Juan estudia en la UAM*, 2) *Juan tiene como mascota a kitty*, 3) *Juan es conocido de Jorge*, 4) *Jorge tiene 28 años y* 5) *El libro de matemáticas discretas fue escrito por Jorge*.

```
exp:Juan exp:estudia-en exp:UAM ..... (1)
exp:Juan exp:tiene-mascota exp:kitty ..... (2)
exp:Juan exp:conoce-a exp:Jorge ..... (3)
exp:Jorge exp:tiene-edad "28 años" ..... (4)
exp:mate_disc exp:tiene-autor exp:Jorge ..... (5)
```

Figura 3.1: Ejemplos de tripletas asociadas a las declaraciones para los recursos Juan y libro de matemáticas discretas.

El marco RDF proporciona la propiedad **tipo (type)** para indicar que *un recurso pertenece a una determinada clase*. Esta propiedad tiene el siguiente URI *http://www.w3.org/1999/02/22-rdf-syntax-ns#type* o en su forma compacta “*rdf:type*”. La propiedad *rdf:type* es una de las más importantes para describir y hacer declaraciones sobre los recursos, porque nos permite clasificar a los recursos. La tripleta asociada a esta descripción es: *prefijo:recurso*

| Clase persona | Clase mascota |
|--------------------------------|--------------------------------|
| exp:Jorge rdf:type exp:Persona | exp:fido rdf:type exp:Mascota |
| exp:Juan rdf:type exp:Persona | exp:kitty rdf:type exp:Mascota |
| exp:Pablo rdf:type exp:Persona | exp:orion rdf:type exp:Mascota |

Tabla 3.3: Ejemplos de tripletas que emplean la propiedad *rdf:type* para asignar un recurso a una determinada clase.

rdf:type *prefijo:Clase*. Para ejemplificar esto: los recursos Jorge, Juan, y Pablo son personas, mientras los recursos fido, kitty, orión son mascotas. Las respectivas tripletas de estos se muestran en la Tabla 3.3.

Un **grafo estructurado y dirigido** es la estructura para visualizar las tripletas. Este **grafo RDF** [14] está compuesto por nodos, aristas y etiquetas para representar las tripletas. El nodo origen es el sujeto, el nodo destino es objeto, mientras la etiqueta de la arista es la propiedad que vincula al nodo origen y al nodo destino.

La Figura 3.2 muestra un grafo RDF asociado a las tripletas de la Figura 3.1. En este grafo, los nodos circulares son recursos y los nodos rectangulares son literales, el nodo destino es aquel a quien apunta la *punta de flecha*, mientras el otro nodo es el origen.



Figura 3.2: Ejemplo de un grafo RDF o grafo de conocimientos.

En el marco RDF, existen distintas sintaxis para escribir y almacenar las tripletas. Estas

sintaxis son: N3³, turtle⁴, RDF/XML⁵, N-triples⁶. El Consorcio de la Web (W3C) establece como sintaxis estándar al RDF/XML. Aunque, la sintaxis Turtle es equivalente a las tripletas de la Figura 3.1.

3.3. Lenguaje de consulta sobre grafos RDF (SPARQL)

Las tecnologías semánticas proponen al lenguaje **SPARQL** como *lenguaje de consulta y protocolo de acceso a RDF* [17], para la búsqueda y recuperación de la información en un grafo RDF.

La idea básica de una **consulta SPARQL** es encontrar conjuntos de tripletas en el grafo RDF que coincidan con un **patrón triplete**. Un *patrón triplete* es parecido a una *tripleta RDF*, *excepto que el sujeto, predicado y objeto pueden ser una variable*. La estructura genérica de una consulta SPARQL se presenta en la Figura 3.3.

```

###Lista de prefijos
PREFIX rdf: <http://www.w3.org/1999/02/22-rdf-syntax-ns#>
PREFIX exp: <http://www.mi-ejemplo.com/>

### Variables a recuperar
SELECT ?x
WHERE {
    ### Lista de patrones tripletas
    ?x exp:propiedad1 exp:objeto1.
    ?x exp:propiedad2 ?y.
}

```

Figura 3.3: Estructura básica de una consulta SPARQL.

Un motor de consulta SPARQL a partir de estas consultas básicas, realiza las siguientes operaciones: 1) interpretar una consulta SPARQL, 2) comparar los *patrones triplete* con el *grafo RDF*, y 3) recuperar los valores asociados a las variables de la cláusula SELECT. Los resultados que proporciona este motor son *conjuntos de datos*.

3.4. Reglas de inferencia (RDF(S)/OWL) y razonadores

En las tecnologías semánticas, el concepto clave es la **ontología** para representar (modelar) y gestionar el conocimiento de un dominio particular. Varios investigadores en las TI, como: Newell, Genesereth y Nilsson, Neches y Gruber, han definido este concepto. Nosotros

³W3C, "Notation3 (N3)," Disponible en: <http://www.w3.org/TeamSubmission/n3/>

⁴W3C, "Turtle," Disponible en: <http://www.w3.org/TR/2013/CR-turtle-20130219/>

⁵W3C, "RDF/XML Syntax Specification," Disponible en: <http://www.w3.org/TR/rdf-syntax-grammar/>

⁶W3C, "N-Triples," Disponible en: <http://www.w3.org/2001/sw/RDFCore/ntriples/>

elegimos la siguiente definición: “*Una ontología es una especificación formal y explícita de una conceptualización compartida* [18]. En esta definición se tienen las siguiente características [12], [19].

- **Conceptualización** es una visión simplificada de algún fenómeno en el mundo que queremos representar a partir de los conceptos, funciones, relaciones, restricciones y otros objetos relevantes en dicho fenómeno.
- **Explícita** consiste en definir expresa y claramente los conceptos así como las restricciones sobre ellos.
- **Formal** significa que los elementos de una conceptualización deben ser representados en un lenguaje para que sea comprensible por los procesos automáticos.
- **Compartida** se refiere a que la conceptualización debe ser consensuada y aceptada por el grupo de personas.

La finalidad de una **ontología de un área investigación** es permitir encontrar información pertinente sobre temas especializados para los grupos de investigación. De esta manera, estas personas en vez de dedicar tiempo en la búsqueda, mejor pasen más tiempo en realizar sus actividades de investigación.

Los principales objetivos en el uso de una ontología son [18]: 1) *La construcción de un vocabulario conceptual formal y consensuado para un dominio dado.* 2) *Un conjunto de reglas para combinar los conceptos y relaciones, de esta manera, componer expresiones complejas en el vocabulario.* 3) *Un vocabulario para construir descripciones y comunicar hechos.* 4) *Personas y procesos automáticos interpreten sin ambigüedad el conocimiento y vocabulario de un dominio dado.* 5) *Personas y procesos intercambien y reutilicen el conocimiento para diferentes propósitos.* 6) *La inferencia de información a partir de un programa especializado (razonador) y los hechos en una ontología.* 6) *Personas y procesos consulten información mediante motores de búsqueda y razonadores*

Una ontología tiene tres elementos clave [20], [21]:

- Clase (Class) representa una colección de objetos que comparten características comunes.
- Individuo (Individual) es el nombre de un objeto específico que pertenece a alguna clase.
- Propiedad (Property) describe relaciones binarias entre los objetos.
 - Propiedad de Objeto (Object Property) son relaciones entre objetos.
 - Propiedad de Dato (Data Property) son relaciones entre un objeto y una literal (cadena, entero).

Una ontología tiene dos componentes [22]:

- Un componente asertivo (ABox) representa el conocimiento e información explícita en los recursos del dominio. Este componente está constituido por las declaraciones (descripciones o hechos verdaderos) de los recursos que afirman que los individuos son instancias de una clase o propiedad. Por ejemplo, puede afirmarse que: *el curso **Temas Selectos de Bases de Datos** pertenece al plan de estudios de la **Licenciatura en Computación**, el alumno **Jorge Aparicio** está cursando **Temas Selectos de Bases de Datos** o el **Laboratorio de Análisis y Rendimiento de Teleservicios** está en la **Universidad Autónoma Metropolitana Unidad Iztapalapa**.*
- Un componente terminológico (TBox) representa el conocimiento implícito en los recursos del dominio. Este componente describe las clases y propiedades relevantes, así como los axiomas que permiten aprovechar la manera en que las instancias se relacionan entre sí. Por ejemplo, se puede afirmar que: 1) *todo **alumno** está inscrito a un **curso** y pertenece a una **universidad**,* 2) *toda **universidad** es una **institución educativa** o 3) *todo **estudiante de universidad** pertenece a la **comunidad universitaria**.**

Los axiomas [18] son expresiones para enriquecer el conocimiento explícito en el grafo RDF. Estos axiomas tienen diferentes propósitos [22], como son: describir relaciones entre clases, definir propiedades en términos de otras, definir relaciones entre conceptos, definir restricciones de cómo las propiedades se relacionan, por mencionar algunos.

Los axiomas deben serializarse en varias tripletas y los vocabularios para escribirlas, son el *esquema RDF* (RDF(S)⁷) y al *Lenguaje de Ontologías Web* (OWL⁸). Estos dos vocabularios son los estándares propuestos por las tecnologías semánticas. En una ontología el prefijo “owl” abrevia el siguiente URI “<http://www.w3.org/2002/07/owl#>”, mientras el prefijo “rdfs” abrevia al URI “<http://www.w3.org/2000/01/rdf-schema#>”.

Las funcionalidades de los axiomas para *relacionar clases en términos de otras*, se listan a continuación. Estos axiomas son los que generalmente se emplean en la construcción y mantenimiento de ontologías [23], [21].

- **Subclase** (*rdfs:subClassOf*) afirma que una *clase A* se subsume por una *clase B*, es decir, la clase A es un caso particular de la *clase B*. En este caso, las instancias de la clase A son instancias de la clase B. Este axioma permite especificar la jerarquía entre clases. Por ejemplo, *todo animal o planta es un ser vivo. esto significa que, las clases **Animal** y **Planta** son subclases de la clase **Ser vivo**.*
- **Clases equivalente** (*owl:equivalentClass*) afirma que la *clase A* y *clase B* representan al mismo conjunto de individuos y el significado de ambas clases es el mismo, es decir, son sinónimas. Por ejemplo, *todas las personas son humanos y todos los humanos son personas, esto significa que todas las instancias de la clase **Persona** deben ser instancias de la clase **Humano** y viceversa.*

⁷W3C, “RDF Vocabulary Description Language 1.0: RDF Schema,” Disponible en: <http://www.w3.org/TR/2004/REC-rdf-schema-20040210/>

⁸W3C, “OWL 2 Web Ontology Language Structural Specification and Functional Style Syntax,” Disponible en: <http://www.w3.org/TR/owl2-syntax/>

- **Clases disjuntas** (*owl:disjointWith*) afirma que la *clase A* y *clase B* no tienen instancias en común. Por ejemplo, *ninguna mujer es hombre, esto significa que ninguna instancia de la clase **Mujer** debe pertenecer a la clase **Hombre** y viceversa.*

Las funcionalidades de los axiomas *para definir propiedades en términos de otras*, se listan a continuación.

- **Subpropiedad** (*rdfs:subPropertyOf*) afirma que todos los recursos que se relacionan por la *propiedad X*, también se relacionan por la *propiedad Y*. Este axioma permite especificar la jerarquía entre propiedades. Por ejemplo, *la propiedad es padre es un caso particular de la propiedad es familiar, de esta manera, si Juan es padre de Pedro, entonces Juan es familiar de Pedro.*
- **Propiedad equivalente** (*owl:equivalentProperty*) afirma que la *propiedad X* y la *propiedad Y* relacionan a los mismos recursos y éstas tienen el mismo significado. Por ejemplo, *las propiedad **tienen automóvil** es sinónimo de la propiedad **tienen carro**, por ello, si Juan tiene un automóvil tipo sedan, entonces Juan tiene un carro tipo sedan y viceversa.*
- **Propiedad inversa** (*owl:inverseOf*) afirma que si la *propiedad X* relaciona al *individuo A* con el *individuo B*, entonces hay una *propiedad Y* que relaciona al *individuo B* con el *individuo A*. Por ejemplo, *la propiedad inversa de **es abuelo**, es la propiedad **es nieto**, por ello, si Juan es abuelo de Antonio, entonces Antonio es nieto de Juan.*
- **Propiedad transitiva** (*owl:TransitiveProperty*) afirma que si la *propiedad X* relaciona al *individuo A* con el *individuo B* y también ésta relaciona al *individuo B* con el *individuo C*, entonces debe relacionar a los *individuos A* y *C*. Por ejemplo, *si Juan tiene parentesco de consanguinidad con Pedro y Pedro tiene parentesco de consanguinidad con Arturo, entonces Juan tiene parentesco de consanguinidad con Arturo.*
- **Propiedad simétrica** (*owl:SymmetricProperty*) afirma que la *propiedad X* es su propia propiedad inversa, es decir, si la *propiedad X* relaciona al *individuo A* con el *individuo B*, entonces, esta propiedad debe relacionar al *individuo B* con el *individuo A*. Por ejemplo, *si Juan es familiar de Pedro, entonces Pedro es familiar de Juan.*
- **Propiedad reflexiva** (*owl:ReflexiveProperty*) afirma la *propiedad X* relaciona al *individuo A* consigo mismo. Por ejemplo, *Juan se conoce a sí mismo.*

Las funcionalidades de los axiomas *para asociar restricciones a las propiedades*, se listan a continuación.

- **Dominio** (*rdfs:domain*) especifica qué clase se aplica a una propiedad. Por ejemplo, *todo individuo que emplea la propiedad **es madre**, debe ser una **Mujer**, por ello, si Rocío es madre de Arturo, entonces Rocío es una instancia de la clase **Mujer**.*

- **Rango** (*rdfs:range*) especifica los valores (clase o tipo de literal) que puede asumir una propiedad. Por ejemplo, *toda persona que **tiene abuelo** debe vincularse con un individuo de la clase **Hombre***, esto es, si María tiene por abuelo a Ramón, entonces Ramón es una instancia de la clase **Hombre**.

El vocabulario OWL ofrece otros axiomas que tienen otras funcionalidades⁹ (restricciones cardinalidad, valores de literales, existenciales o universales) para enriquecer el conocimiento en un dominio [24]. Los axiomas en los lenguajes OWL y RDF(S) pueden ser combinados para construir clases y propiedades complejas [22], [24]. Un ejemplo de esta combinación es el siguiente. *Todo **metal líquido** es aquel elemento que pertenece a la intersección de la clase **Metal** y la clase **Líquido***. Este ejemplo, se representa en la Figura 3.4.

exp:Metal-Liquido *rdfs:subClassOf* (**exp:Metal** **and** **exp:Liquido**)

Figura 3.4: Regla para indicar que un Metal-Liquido pertenece a las clases Metal y Líquido.

En las tecnologías semánticas, un *razonador* [23], [25] es un programa que deduce declaraciones a partir de los axiomas y declaraciones explícitas en la ontología. Este programa también se denominan razonador semántico o motor de inferencias. Un razonador permite realizar dos actividades importantes con una ontología:

- Un razonador como *instrumento de validación de consistencia de una ontología*. La validación consiste en deducir información con los axiomas y encontrar si hay contradicciones en el modelo. Si no existen contradicciones en el modelo, entonces, éste es consistente. Por el contrario si hay una contradicción, entonces el modelo no es consistente. Por ejemplo, si en la ontología se establece que la clase Hombre y Mujer son disjuntas, y el recurso Antonio es instancia de estas dos clases, entonces el modelo tiene una contradicción, por tanto, el modelo no es consistente.
- Un razonador para *mejorar la búsqueda de la información en una ontología*. Las declaraciones explícitas y un motor SPARQL solamente permiten recuperar información explícita de los recursos. Un motor de búsqueda SPARQL junto con un razonador, permiten recuperar mejor información en el grafo RDF. Esto es, el razonador expande el grafo RDF con las declaraciones inferidas, donde esta expansión puede ser o no ser explícita. De esta manera, el motor consulta y recupera la información en el grafo.

3.5. Ventajas de las tecnologías Semánticas

Las tecnologías semánticas proporcionan varias características y funcionalidades que benefician la representación y gestión del conocimiento. Algunos de estos beneficios son *facilitar*

⁹W3C, "OWL 2 Web Ontology Language Primer," Disponible en: <http://www.w3.org/TR/2012/REC-owl2-primer-20121211/>

la percepción y representación de dominios particulares, integrar el conocimiento de fuentes heterogéneas (formato, contenido, estructura) de información, compartir y reutilizar el conocimiento a partir de modelos dados e inferir conocimiento a partir de los axiomas. A continuación, se describen estos y otros beneficios que ofrecen las tecnologías semánticas.

Las tecnologías semánticas proporcionan una **manera fácil y sencilla** de representar el conocimiento de un dominio particular en una ontología. Esta facilidad para modelar, se debe a los siguientes hechos: 1) *todo recurso debe tener un URI*, 2) *las características y relaciones en los recursos se representan en forma de tripletas*, 3) *una tripleta se compone por un sujeto, verbo y un objeto*, 4) *las definiciones de clases, propiedades y axiomas se representan en forma de tripletas*, 5) *las tripletas del conocimiento explícito e implícito pueden ser visualizadas en un grafo (nodos, etiquetas y arcos)* y 6) *el grafo de conocimiento constituye la ontología de dominio*. De esta manera, dominios particulares con una gran cantidad de objetos pueden representarse a partir de elementos básicos y sencillos en un formato estándar (tripletas). Esta **facilidad** en la construcción y mantenimiento de una ontología, se denomina **flexibilidad de las tecnologías semánticas** [1], [26], [27].

En una ontología el conocimiento no se limita a unas cuantas características sobre un recurso, sino que toda información significativa puede describirse para un determinado recurso. Por ejemplo, se desea modelar las personas dependientes de un empleado en una BD relacional de una organización. En esta BD pueden realizarse dos cosas: 1) *para cada empleado se asigna un determinado número de dependientes en la tabla de información personal de los empleados*, o bien, 2) *se construye una nueva tabla con todos los dependientes de los empleados y se hace una relación entre esta tabla y la tabla de los empleados*. Ahora bien, en una ontología para cada recurso empleado se representan los dependientes en forma de tripleta. Esta propiedad de desarrollar ontologías de forma creciente, se denomina **extensibilidad en una ontología** [1], [28].

El marco RDF es una herramienta para *solucionar la heterogeneidad en formato, contenido y estructura* en los recursos. Porque este marco permite representar cualquier recurso a partir de sus características significativas y relaciones con otros recursos. En concreto, el marco RDF permite realizar las siguientes actividades para cada tipo de heterogeneidad.

- **Formato** El marco RDF permite modelar cualquier recurso, sin importar si es un documento con extensión doc, pdf, odp, html, xml, o un archivo multimedia con extensión ppt, mp3, mpeg, jpg, o incluso si es una persona, organización o cualquier otro recurso físico. De hecho, una característica importante en un documento y archivo multimedia es la extensión del archivo. Otra característica importante es indicar a cuál clase pertenece un recurso, por ejemplo, Documento, Multimedia, Persona, Organización, por mencionar algunas.
- **Contenido** Por definición el marco describe las características en torno o en los recursos. De esta manera, si los recursos hacen referencia a distintos temas, el marco RDF permite establecer las tripletas que vinculan a un recurso con uno o varios temas.

- **Estructura** La flexibilidad del marco, permite representar cualquier recurso, sin importar que este recurso sea estructurado, semi-estructurado o sin estructura. De hecho, el estándar R2RML¹⁰ es el lenguaje estándar para trasladar BD relacionales a modelos con tripletas RDF.

Las tecnologías semánticas solucionan problemas de *ambigüedad* en la representación de un dominio. Para empezar, una ontología soluciona el problema de homonimia. Porque todo recurso, clase y propiedad tiene un *identificador único*. Esto significa que si un recurso o propiedad tiene distintos significados, entonces para cada significado se le asigna un identificador único. Por ejemplo, el término **radio** tiene un significado distinto para cada uno de estos cuatro dominios: *Matemáticas*, *Anatomía*, *Geometría* o *Telecomunicaciones*. Este término por cada dominio se asigna un identificador único, es decir, *mat:radio*, *anat:radio*, *geo:radio* y *tel:radio*. De esta manera, si se emplea el recurso *mat:radio* en una tripleta, entonces, esta tripleta describe un objeto del dominio de Matemáticas.

En una ontología, puede definirse que un recurso, clase o propiedad es *sinónimo* de otro objeto del mismo tipo. Esta propiedad de sinonimia se hace con base en el uso de *axiomas*. Estos son los axiomas para definir objetos sinónimos: *clase equivalente* (*owl:equivalentClass*), *propiedad equivalente* (*owl:equivalentProperty*) e *individuos idénticos* (*owl:sameAs*). Esta propiedad de sinonimia es importante para fines de búsqueda. Porque, al hacer una consulta mediante un objeto que tiene un sinónimo, puede recuperarse mayor información del otro objeto, o bien, puede simplificarse la consulta. Por ejemplo, el recurso *computadora* tiene los siguientes sinónimos: *computador*, *ordenador*, *equipo de cómputo*, *por mencionar algunos*. Al hacer una consulta sobre alguna característica de una computadora, es importante, recuperar también la información de los recursos: *computador*, *ordenador* y *equipo de cómputo*. Si se emplea el axioma de equivalencia de clase entre estos recursos y un razonador, entonces, la información de estos recursos es recuperada por el motor de búsqueda.

Una utilidad importante en las tecnologías semánticas es la *interoperabilidad* [29], [27]. Este concepto se refiere a *la facilidad de reutilizar y compartir las ontologías entre personas o aplicaciones, gracias a que una ontología emplea y se elabora con varios estándares*. En concreto, el marco RDF propone la *estructura estándar* para describir el conocimiento, RDF(S) y OWL proponen los *lenguajes estándares* para escribir los axiomas, y SPARQL el *lenguaje de consulta estándar* sobre grafos de conocimiento.

Ejemplos de ontologías¹¹ (modelo de referencia) que proporcionan interoperabilidad son: 1) **Dublin Core** es un vocabulario genérico de metadatos que proporciona información descriptiva de cualquier documento en un sistema de información¹², y 2) **Friend Of A Friend** (FOAF) es un vocabulario para describir a las personas y las relaciones de éstas en la Web¹³.

La interoperabilidad promueve la realización de diversas actividades para mejorar de manera eficiente y eficaz la gestión del conocimiento. Estas actividades se listan a continuación:

- integrar el conocimiento desde distintas fuentes de información.

¹⁰W3C, "R2RML: RDB to RDF Mapping Language," Disponible en: <http://www.w3.org/TR/r2rml/>

¹¹W3C, "Good Ontologies," Disponible en: http://www.w3.org/wiki/Good_Ontologies

¹²Dublin Core Semantic Initiative, "Dublin Core," Disponible en: <http://dublincore.org/>

¹³FOAF project, "The Friend of a Friend (FOAF) project," Disponible en: <http://www.foaf-project.org/>

- independizar el uso de una única herramienta o sobre determinado sistema operativo.
- realizar tareas de inferencia a partir de los vocabularios OWL y RDF(S).
- recuperar información o construir subgrafos mediante consultas en una ontología.
- liberar a las organizaciones del uso de formatos propietarios que tienen un costo económico o de propiedad intelectual.
- producir ontologías genéricas para dominios particulares, como: Biomédica, Economía, Matemáticas, Ciencias de la Computación, Física, etc.
- construir rápidamente modelos de conocimiento a partir de ontologías básicas.
- mezclar ontologías y construir modelos de conocimiento complejos.

Una ventaja de tener un modelo flexible y estándar es desarrollar **aplicaciones genéricas** para aprovechar estos modelos. Los objetivos de estas herramientas son: *procesar datos, facilitar la visualización del grafo de conocimiento a los usuarios, incrementar el conocimiento en las ontologías mediante la introducción de descripciones o axiomas, facilitar el mantenimiento a una ontología, proporcionar tareas de inferencia en una ontología, mejorar la búsqueda de la información, integrar y mezclar ontologías, facilitar de uso y mejorar la integración de los usuarios.*

Estas herramientas genéricas posibilitan que personas expertas en el dominio sean las principales constructoras del grafo de conocimiento. De esta manera, la información en el grafo será confiable, ya que estas personas son las que tienen los conocimientos en el dominio. Mientras, los desarrolladores son los encargados de construir de estas aplicaciones genéricas.

En esta tesis, el beneficio más interesante es el **uso de un razonador y un motor de búsqueda** para *mejorar la búsqueda y recuperación de la información*. Un **razonador** a partir de los axiomas, expande el grafo RDF con las declaraciones inferidas, donde esta expansión puede ser o no ser explícita. A partir de este grafo de conocimiento, un motor de consulta puede compararlo para responder una consulta dada.

Para ejemplificar el beneficio de combinar un motor de búsqueda y razonador. Primero, se parte de utilizar solamente un motor de búsqueda y la ontología. La Figura 3.5 muestra el TBox y ABox para la ontología de ejemplo.

Ahora bien, un usuario desea recuperar *todos los individuos que son personas*. La consulta en lenguaje SPARQL se presenta en la Figura 3.6.

Un motor de consulta SPARQL procesa esta consulta, seguido de esto, el motor no arroja ningún resultado. Porque en la ontología (Figura 3.5) no hay una tripleta explícita que establezca que un recurso pertenece a la clase Persona.

Esta ontología implícitamente tiene las tripletas que indican que un recurso pertenece a la clase Persona. Porque los axiomas establecen que los individuo de las clases Mujer, Hombre y Estudiante son instancias de la clase Persona. Ahora bien, un razonador infiere las tripletas a partir de los axiomas y descripciones en la ontología de la Figura 3.5. La Figura 3.7 muestra la ontología con las tripletas materializadas.



Figura 3.5: ABox y TBox para ejemplificar el beneficio de utilizar un razonador y un motor de búsqueda.

```
PREFIX rdf: http://www.w3.org/1999/02/22-rdf-syntax-ns#>
PREFIX exp: <http://www.mi-ejemplo.com/>
SELECT ?x
WHERE
{
  ?x rdf:type exp:Persona.
}
```

Figura 3.6: Consulta SPARQL para recuperar todos los individuos que son personas.



Figura 3.7: Ontología con tripletas que han sido inferidas mediante un razonador.

Si de nuevo se hace la consulta de la Figura 3.6 con un motor de consulta SPARQL en la ontología de la Figura 3.7, se obtienen los siguientes resultados: *exp:Juan*, *exp:Laura* y *exp:Luis*. Por esta razón, la combinación de un razonador y motor de consulta es un mecanismo que permite recuperar más respuestas porque el conocimiento implícito se vuelve explícito.

Capítulo 4

Estado del arte

La *integración de los recursos de información en una memoria corporativa* ha sido poco explotada por las organizaciones o áreas de investigación. Existen algunos trabajos sobre la *integración de información* que incorporan a las tecnologías semánticas para representar y enriquecer el conocimiento de un dominio dado, así como, la búsqueda de información a partir de este conocimiento. En este estado del arte, se consideran los trabajos que cumplen con alguno de nuestros criterios de investigación. Estos criterios son descritos en la Tabla 4.1

| Criterio | Descripción | Definición formal |
|--|--|---------------------|
| Integración de información en los recursos | Ésta consiste en el proceso de búsqueda y recuperación de la información sobre los recursos de información. | Sección 2.1.3 |
| Memoria corporativa | Ésta es la representación consistente y formal del conocimiento en una organización. | Sección 2.1 |
| Modelo semántico | Éste es la representación del conocimiento a partir de las tecnologías semánticas. | Secciones 3.2 y 3.4 |
| Inferencia en el modelo | Ésta consiste en deducir información a partir de los axiomas en el modelo. | Sección 3.4 |
| Interfaz visual para la integración | Ésta es un aplicación con una enfoque visual para que las personas pregunten o naveguen a través de la información en el modelo semántico. | Sección 6 |

Tabla 4.1: Criterios considerados para el *estado del arte* de la integración semántica de recursos.

La Sección 4.1 describe los trabajos que se estudiaron para la integración de los recursos y al final de ésta sección se presenta una tabla comparativa de estos trabajos, así como los valores asociados a los criterios de investigación.

En este *estado del arte*, se contempla un estudio de las aplicaciones para realizar la *integración semántica de los recursos en una memoria corporativa*. Las aplicaciones estudiadas,

se agrupan de acuerdo a las siguientes funcionalidades.

1. Escribir las declaraciones en forma de triple y guardarlos en alguna sintaxis estándar.
2. Escribir los axiomas mediante los vocabularios estándar (OWL y RDF(S)).
3. Gestión del grafo RDF, es decir, carga del modelo, consulta de información e inferencia.

La Sección 4.2 muestra y describe las aplicaciones estudiadas con base en su funcionalidad; al final de cada agrupación, se da a conocer: *cuál herramienta se eligió para facilitar y efectuar la funcionalidad dada*.

4.1. Integración semántica de recursos de información

El principal objetivo de la *integración de los recursos* es buscar y recuperar información que está en los recursos, para responder las necesidades informativas de las personas. Una *integración semántica* de recursos emplea las tecnologías semánticas con la finalidad de recuperar información significativa en los recursos a partir de las características y relaciones de estos. El uso de una *memoria corporativa* para la integración semántica, se traduce en información y conocimiento de los recursos bajo un dominio particular.

Algunos trabajos exploran o emplean el enfoque de las *tecnologías semánticas* para fines de integración del conocimiento, representación de una memoria corporativa o búsqueda de información. A continuación, se describe el estado actual del conocimiento referente a estos trabajos.

- La ***arquitectura del modelo dual*** [30] es una propuesta para la representación consistente y comprensible de la información clínica de cualquier persona. La finalidad de esta arquitectura es facilitar el acceso de historial clínico de los pacientes a los profesionales de la salud. La información de estos historiales está distribuida en varios sistemas independientes y heterogéneos. Esta arquitectura se basa en un modelo que por un lado representa la información y por el otro el conocimiento. En la representación de la información se describen las estructuras de datos comunes. Mientras, en la representación del conocimiento se emplean arquetipos para representar el conocimiento formal de conceptos clínicos. Este trabajo presenta una herramienta para desarrollar los arquetipos de datos clínicos. Esta herramienta es llamada LinkEHR-Ed. La finalidad de esta es que los profesionales de la salud y expertos en tecnologías de la información sean los principales constructores del conocimiento.

Nosotros ***elegimos este trabajo*** por las siguientes razones: 1) *la arquitectura representa el conocimiento de los historiales clínicos (dominio de la salud)*, 2) *la arquitectura solventa la información distribuida y formatos propios de un sistema*, 3) *la arquitectura modela el conocimiento a manera de un componente terminológico y un componente asertivo*, 4) *los arquetipos agregan una capa semántica y proporcionar el conocimiento formal para el sector salud* y 5) *una herramienta para construir arquetipos asociados al sector salud*.

- El *marco de integración semántica* [31] es una propuesta para solucionar de manera eficaz y flexible la integración de la información en el dominio del alojamiento en-línea. La finalidad de esta integración es facilitar la reunión y compartición de información referente al alojamiento en-línea, donde, esta información está en constante cambio. Este marco de integración tiene un conjunto de características básicas: 1) emplear una ontología para facilitar el acceso a la información integrada y solucionar la heterogeneidad en la estructura de la información, 2) emplear un proceso que resolver la naturaleza dinámica de las fuentes de información, 3) permitir a los propietarios de la información participar en el proceso de integración, 4) emplear una serie de esquemas para el intercambio de información.

Este trabajo *se eligió* por estas razones: 1) *proporciona un marco de trabajo para la integración de la información*, 2) *emplea una ontología para modelar la información del dominio de alojamiento en-línea* y 3) *emplear la ontología como instrumento para resolver la heterogeneidad en la estructura de la información y el acceso a la información integrada*.

- Jun Zhai et al. [32] proponen una *integración semántica con base en ontologías para sistemas de información de energía eléctrica*, donde, estos sistemas son heterogéneos con funciones y organizaciones descentralizadas. Esta integración, por un lado, *emplea al lenguaje de marcado extensible (XML para el intercambio de información entre estos sistemas)*. Por otro lado, esta integración *utiliza una ontología para describir formalmente la información a nivel conceptual en el dominio de la electricidad*. Este trabajo propone una arquitectura de tres capas para esta integración semántica: 1) capa fuentes de datos heterogéneos distribuidos, 2) capa de integración de la información y 3) capa de sistemas de aplicación.

Este trabajo de Jun Zhai *es interesante*, en primera, *porque emplea una ontología como especificación para integrar la información a nivel semántico*, y en segunda, *el uso de XML para el intercambio de información y solucionar la heterogeneidad de datos en los sistemas de energía eléctrica*.

- Xin y Guangleng [33] emplean un *enfoque basado en las ontologías para capturar la información del design rationale*. Este design rationale es un conocimiento para explicar qué y cómo se diseña un producto, así como para apoyar la reutilización, comunicación y verificación de diseños en empresas manufactureras. En este trabajo, se emplea una memoria corporativa para las actividades de gestión del conocimiento, en particular, las actividades de captura y disponibilidad en el design rationale. Las ontologías permiten el acceso uniforme a las fuentes de información, y en este trabajo, éstas modelan el design rational para el background del diseño de autos de carga.

Este trabajo *es importante* por estas razones: 1) *utilizar una memoria corporativa para el conocimiento del design rational*, 2) *emplear ontologías para capturar la información del design rational* y 3) *acceder de manera uniforme a los recursos de información mediante las ontologías*.

- **PCOGEME** [34] es un *entorno de colaboración* para la *creación, gestión, difusión, mantenimiento de memorias corporativas*. En este trabajo, las memorias corporativas son mecanismos para la gestión del conocimiento y documentos. PCOGEME propone un modelo de interacción basado en las ontologías para la representación y gestión de estas memorias. El funcionamiento de PCOGEME se basa en la lluvia de ideas y un mecanismo de toma de decisiones consensuadas, para la construcción de memorias corporativas mediante el uso de ontologías.

Nosotros *elegimos* este trabajo, porque emplea los siguientes elementos: 1) *memoria corporativa como instrumentos de gestión del conocimiento*, 2) *representación del conocimiento mediante el uso de ontologías* y 3) *el uso de un entorno de gestión del conocimiento*.

La Tabla 4.2 es un sumario de los valores asociados a nuestros criterios investigación para cada trabajo estudiado. Las cabeceras en esta Tabla están en forma abreviada y estos son sus significados: **IIR** = integración de información en los recursos, **MC** = memoria corporativa, **MS** = modelo semántico, **IeM** = inferencia en el modelo y **IVpI** = interfaz visual para la integración.

| Trabajo | IIR | MC | MS | IeM | IVpI |
|---|-----|-------------------|------------------------------------|--------------|-----------|
| Arquitectura del modelo dual | Si | No | Modelo de referencia y arquetipos | No | No |
| Marco de integración semántica | Si | No | Información XML y ontología global | No | Prototipo |
| Arquitectura para la integración en SIEE ^a | Si | No | tripletas RDF y axiomas RDF(S) | Consistencia | No |
| Metodología para ontologías en el design rationale | Si | Design rationale | Ontología | No | No |
| Construcción de MC en forma colaborativa | No | SSII ^b | Ontología | No | No |

^a Sistemas de Información de Energía Eléctrica.

^b Sociedad de Servicios en Ingeniería Informática.

Tabla 4.2: Comparativa entre los trabajos estudiados y nuestros criterios para la integración semántica de recursos.

En esta Tabla 4.2, el valor ‘**Si**’ indica que el trabajo cumple con ese criterio, mientras, el

valor ‘No’ indica lo contrario.

4.2. Herramientas para la integración semántica de recursos

Un *descriptor semántico de recursos* [35] es una herramienta para crear y almacenar tripletas RDF a partir de la *información explícita en los recursos*. Las tripletas que son generadas por esta herramienta, están escritas en una de las siguientes sintaxis: *RDF/XML*, *Turtle*, *N-triple* y *N3*. El principal objetivo de un *descriptor* es construir instancias y relacionar éstas con determinados valores u otras instancias (*concepto de triple*). Algunas de estas herramientas requieren un TBox para saber cuáles clases y propiedades, pueden emplearse en los triples. Un descriptor proporciona una *interfaz gráfica de usuario* (GUI) para simplificar a los usuarios la creación y modificación de las declaraciones. Algunos descriptores sugieren información para las declaraciones a partir de un proceso de aprendizaje en un corpus documental o de imágenes.

En la siguiente lista, se presentan los descriptores semánticos que nosotros estudiamos.

- **OntoMat Annotizer** [36] es una herramienta para hacer anotaciones semánticas de páginas web, documentos basados en texto plano y lenguajes de marcado¹. El objetivo de esta herramienta es que el usuario cree de manera amigable instancias y declaraciones de éstas, mediante la funcionalidad de arrastrar y soltar (drag-and-drop).
- **MnM** [35] es una herramienta que proporciona apoyo automatizado y semiautomatizado para describir páginas Web con contenido semántico². MnM tiene GUI que integra un editor de ontología, navegador Web, un editor de instancias y de propiedades. El objetivo de esta herramienta es la descripción de documentos a partir de declaraciones derivadas de ontologías preexistentes.
- **GATE** [37] es un entorno de desarrollo integrado (IDE) para el desarrollo de componentes en el procesamiento del lenguaje humano y el procesamiento de texto³. Las tareas en el procesamiento de texto son: *minería web*, *extracción de información y descripciones semánticas*.
- **Aktive Media** [35] es una GUI para la descripción automática de una colección de imágenes o documentos (batch annotation) para un contexto específico. “*El objetivo de Aktive es automatizar el proceso de descripción, mediante la sugerencia interactiva de la información al usuario, mientras éste está describiendo*”⁴. Estas sugerencias se hacen con base en axiomas y descripciones previas.

¹M. Siroker, “OntoMat Annotizer,” Disponible en: <http://projects.semwebcentral.org/projects/ontomat/>

²The Open University, “MnM,” Disponible en: <http://projects.kmi.open.ac.uk/akt/MnM/>

³The University of Sheffield, “GATE,” Disponible en: <http://gate.ac.uk/>

⁴A. Chakravarthy, V. Lanfranchi, F. Ciravegna, “AKTive Media,” Disponible en: <http://www.aktors.org/technologies/aktivemedia/index.html>

La finalidad de un descriptor es facilitar la generación de descripciones en forma de triple. Sin embargo, hay varias razones, por las cuáles, no se elige una de estas herramientas para alcanzar este fin. Las razones son: 1) *todas estas aplicaciones permiten hacer declaraciones de documentos e imágenes, por tal razón, no proporcionan una solución a la heterogeneidad en formato*, 2) *OntoMat Annotizer y MnM no interpretan los axiomas que están escritos con los vocabularios OWL y RDF(S)*, 3) *Aktive Media y GATE cambian las URIs en las tripletas por sus propios URIs*, 4) *OntoMat Annotizer y MnM no tienen versión estable* y 5) *Aktive Media, GATE y MnM no tienen documentación disponible para solucionar problemas de configuración*.

Un **script** es un código que se escribe en un lenguaje de programación y se utiliza para la escritura y almacenamiento de descripciones en forma de tripletas. El propósito es facilitar y agilizar el proceso de generación de tripletas en alguna sintaxis estándar. Aunque un script no posee una interfaz gráfica para seleccionar la información de los recursos. Esto se puede solucionar mediante el uso de formularios web que capturen la información sobre los recursos. Posteriormente, la información es guardada en algún documento de texto plano, para que un script transforme esta información en triples RDF. Por tal razón, **un script es la opción electa** para representar el conocimiento explícito en forma de tripletas.

Un **editor de ontología** [38] es una herramienta que proporciona una serie de interfaces amigables para la construcción y mantenimiento de ontologías. Estos editores proporcionan las siguientes funcionalidades básicas a los usuarios: 1) *definir las clases, propiedades, instancias y axiomas*, 2) *cargar, almacenar, importar y exportar ontologías que son escritas con lenguajes estándar (RDF(S) y OWL)* y 3) *visualizar las clases, propiedades e individuos*.

- **Protégé** [39] es una plataforma con herramientas para la creación, visualización y manipulación de ontologías en diversos formatos de representación⁵. Esta plataforma proporciona al usuario una interfaz amigable para la definición de clase, propiedades y axiomas, así como la introducción de datos. La arquitectura de esta herramienta se puede extender a través de plug-ins y APIs. Esta herramienta tiene licencia open-source Mozilla Public License⁶.
- **pOWL** [40] es una herramienta para la visualización y edición de ontologías vía web⁷. Esta herramienta soporta la carga y edición de ontologías con vocabularios RDF(S) y OWL, generación de consultas y almacenamiento del modelo en una base de datos relacional.
- **TopBraid Composer** [41] es un IDE para "*desarrollar, gestionar y probar configuraciones de los modelos de conocimiento e instancias de las bases de conocimiento*"⁸. Esta

⁵Stanford Center for Biomedical Informatics Research, "Protégé," Disponible en: <http://protege.stanford.edu/>

⁶Mozilla, "Mozilla Public License," Disponible en: <http://www.mozilla.org/MPL/>

⁷Sören Auer, "pOWL," Disponible en: <http://aksw.org/Projects/Powl.html>

⁸TopQuadrant, Inc., "TopBraid Composer," Disponible en: http://www.topquadrant.com/products/TB_Composer.html

herramienta proporciona un conjunto de editores para visualizar grafos RDF y diagramas de clase. Existen tres versiones de esta herramienta: maestro, estándar y gratuita. La versión gratuita permite crear y editar archivos OWL/XML, así como consultar con el lenguaje SPARQL.

- **SWOOP** [42] es un editor para crear y editar ontologías, comprobar inconsistencias, navegar por las ontologías, compartir y reutilizar los datos existentes⁹. Este editor ofrece un entorno con aspecto de navegador web para facilitar la navegación y edición de ontologías OWL. Este editor provee una interfaz amigable y eficaz para los usuarios web promedios.

Protégé es el editor electo para representar los axiomas en una ontología. Porque este editor proporciona estos beneficios: 1) *una interfaz amigable e intuitiva para el usuario*, 2) *amplia documentación y tutoriales, así como una comunidad de desarrolladores*, 3) *facilidad de extender la funcionalidad de esta herramienta, gracias a su arquitectura de plug-ins*, 4) *variedad de sintaxis para las ontologías, como: Turtle, Manchester, OWL/XML o XML/RDF*, 5) *visualización del grafo (axiomas, clases y propiedades)*, 6) *incorporar razonadores, como: Pellet¹⁰, Fact++¹¹ y HermiT¹²* y 7) *incorporar un motor de consulta SPARQL*.

Un **triplestore** [43] es un programa para *el almacenamiento e indexación de tripletas RDF*, con el fin de permitir la consulta eficiente de información sobre estas tripletas. Estos triplestores emplean el estándar SPARQL como lenguaje de consulta para consultar el grafo RDF. Algunos triplestores soportan la capacidad de inferir en el grafo RDF a partir de axiomas, mediante la incorporación o importación de un razonador para ello. Los triplestores se idealizan como *sistema gestor de bases de datos para modelos basados en tripletas RDF*.

En el siguiente listado, se presentan y describen los triplestores que estudiamos.

- **Apache Jena** [44] es un *marco de trabajo* que proporciona un conjunto de interfaces de programación de aplicaciones (API) para Java. Estas APIs ofrecen las siguientes funcionalidades: *lectura, procesamiento y escritura de triples RDF, así como axiomas RDF(S) y OWL, un motor de inferencia y un motor de consulta SPARQL*. La finalidad de Jena es desarrollar aplicaciones que usan las tecnologías semánticas para la representación del conocimiento¹³.
- **Stardog** [45] es una base de datos para modelos semánticos. El propósito de esta herramienta es la ejecución de consultas sobre los datos RDF que están bajo su gestión directa¹⁴. Esta herramienta emplea los protocolos *HTTP* y *SNARL* para *acceder y*

⁹University of Maryland, “SWOOP,” Disponible en: <https://code.google.com/p/swoop/downloads/list>

¹⁰Clark & Parsia, LLC, “Pellet,” Disponible en: <http://clarkparsia.com/pellet/>

¹¹Clark & Parsia, LLC, “Pellet,” Disponible en: <http://clarkparsia.com/pellet/>

¹²Oxford University, “HermiT,” Disponible en: <http://hermit-reasoner.com/>

¹³The Apache Software Foundation, “Apache Jena,” Disponible en: <http://jena.apache.org/>

¹⁴Clark & Parsia, LLC, “Stardog,” Disponible en: <http://stardog.com/>

controlar de manera remota el modelo de datos RDF, inferencia a partir de axiomas en lenguaje OWL y consultas SPARQL.

- **4store** [46] *es un sistema para el almacenamiento RDF que incorpora un motor de consultas SPARQL*¹⁵. Las principales fortalezas de esta herramienta son el rendimiento, seguridad, escalabilidad y estabilidad.
- **Sesame** [47] *es un marco de trabajo estándar de facto para el análisis, almacenamiento, inferencia y consulta de datos RDF*¹⁶. Este marco proporciona una API que puede emplearse sobre los distintos *sistemas de almacenamiento RDF* para consultar y acceder a esta información de manera remota.

Cualquiera de estos triplestore es una opción viable para efectuar tareas de almacenamiento y búsqueda de información en grafos RDF. Aunque, el más interesante desde nuestra perspectiva es Apache Jena. Las razones del *porqué emplear esta herramienta*, son: 1) *amplia documentación y tutoriales para el desarrollo de modelos semánticos*, 2) *integración de Jena en IDEs para el lenguaje Java, como Eclipse*¹⁷, 3) *proyecto open-source bajo la licencia Apache*¹⁸ *versión 2*, 4) *un conjunto de librerías para crear, cargar, almacenar y consultar declaraciones, así como axiomas en OWL y RDF(S)*, 5) *un motor de inferencia para realizar razonamiento en ontologías que emplean axiomas OWL y RDF(S)* y 6) *una amplia comunidad de desarrolladores*.

¹⁵Garlik, “4store,” Disponible en: <http://4store.org/>

¹⁶Aduna, “Sesame,” Disponible en: <http://www.openrdf.org/index.jsp>

¹⁷The Eclipse Foundation, “Eclipse IDE,” Disponible en: <http://www.eclipse.org/>

¹⁸The Eclipse Foundation, “Licencia Apache v. 2.0 ,” Disponible en: <http://www.apache.org/licenses/LICENSE-2.0.html>

Integración semántica de recursos de información en una memoria corporativa

La *integración de los recursos* es el proceso de búsqueda y recuperación significativa de información existente en los recursos, para responder una consulta dada por un usuario. Si esta integración se hace mediante el uso de herramientas, estándares, metodologías y aplicaciones pertenecientes a las *tecnologías semánticas*, entonces, se dice que ésta es una *integración semántica de los recursos (ISR)*.

Esta *integración semántica de recursos* se emplea en una *memoria corporativa (MC)*, porque una memoria tiene un conjunto diverso de recursos de información que representan el conocimiento de una organización (dominio particular). Las principales razones para realizar la *integración semántica de recursos en una memoria corporativa*, son las siguientes: 1) *solucionar la heterogeneidad de los recursos y la ambigüedad de la información en una memoria corporativa*, 2) *adaptar el conocimiento cambiante o explosivo en los recursos*, 3) *extender y mantener un modelo (representación) del conocimiento*, 4) *permitir consultas específicas a partir de las características y relaciones de los recursos*, 5) *recuperar información significativa de los recursos para que respondan las preguntas de las personas adscritas en la organización* y 6) *emplear herramientas, aplicaciones, vocabularios y formatos estándar*.

El desarrollo de la *integración semántica de recurso* se hace con base en una *secuencia ordenada de métodos (metodología)*. Esta tesis describe una *propuesta de metodología* para la *integración semántica de recursos en una memoria corporativa*, la cual es guiada por *casos de uso*. La finalidad de esta metodología es: facilitar y guiar a los desarrolladores en las tareas de construcción y búsqueda de información de/en un modelo semántico.

La finalidad de esta propuesta es *facilitar y guiar a los desarrolladores en las tareas de construcción de un modelo semántico, así como en la búsqueda de información sobre éste*. Mientras, los principales objetivos de ésta son:

- realizar la ISR en cualquier memoria corporativa, por ejemplo *Biomédica, Química, Biología, Computación, Economía, Zoología, por mencionar algunas*
- emplear distintos *casos de uso* para la ISR y no limitar el número de éstos.
- representar una MC en un formato estándar con un vocabulario consensuado y asociado al contexto de la MC.

- utilizar vocabularios estándares para los axiomas, así como el uso del lenguaje SPARQL para las consultas.

Esta metodología está organizada en tres etapas principales:

1. ***Representación del conocimiento en los recursos*** consiste en identificar los recursos de la memoria corporativa y representar los metadatos (conocimiento explícito) de estos recursos mediante el marco RDF.
2. ***Enriquecimiento del conocimiento en el modelo*** consiste en introducir axiomas en OWL y RDF(S) para extender, completar y adaptar el conocimiento explícito de los recursos.
3. ***Búsqueda y recuperación de la información en el modelo*** consisten en identificar las principales consultas de los usuarios en el dominio y ejecutar éstas mediante el uso de un *motor de búsqueda SPARQL* junto con un *razonador*, para recuperar información de los recursos.

En esta metodología, uno de los elementos clave es el ***caso de uso***. Porque este elemento guía el desarrollo de la integración semántica. En concreto, los casos de uso permiten lo siguiente para las tres etapas: 1) *qué características y relaciones son significativas*, 2) *qué reglas de inferencia son necesarias* y 3) *cuáles consultas son importantes*.

Si bien, nuestra *metodología* es abierta en estos dos aspectos (memoria corporativa y casos de uso). Esta propuesta se pone en práctica para la *memoria corporativa* del grupo de investigación en el área de Redes y Telecomunicaciones del departamento de Ingeniería Eléctrica de la Universidad Autónoma Metropolitana Unidad Iztapalapa. Mientras, los *casos de uso* usados en esta propuesta, son *la cartografía de competencias* y *la búsqueda de recursos digitales*. Los principales usuarios en la integración son los profesores-investigadores del núcleo del área de Redes y Telecomunicaciones, así como los estudiantes que realizan algún proyecto o servicios social y que son asesorados por un profesor del núcleo.

La arquitectura de la integración semántica se diseñó con base en el modelo de tres capas (nivel usuario, nivel negocio y nivel de datos).

- En el nivel de usuario: se tiene un conjunto de páginas Web dinámicas y estáticas que proporcionan la interfaz visual. Esta interfaz proporciona una manera fácil y sencilla de estructurar las preguntas de los usuarios, así como la visualización de los resultados vinculados a estas preguntas. Las páginas estáticas proporcionan los formularios para que los usuarios *estructuren las preguntas y capturen la información* a buscar en la MC. Mientras, las páginas dinámicas proporcionan la información que responde las preguntas en un formato visual agradable al usuario.
 - En el nivel de negocios: una aplicación transforma la información recopilada de las páginas estáticas en patrones tripletas y construir una consulta SPARQL. Posteriormente, esta aplicación invoca al triplestore para efectuar estas actividades: 1) solicitar y cargar
-

la ontología, 2) hacer inferencia en una ontología mediante el uso de un razonador y 3) buscar y recuperar la información en el modelo inferido mediante el uso de motor de búsqueda SPARQL y la consulta SPARQL.

- En el nivel de datos (conocimiento): la ontología modela el conocimiento de los recursos de una memoria corporativa en un formato estándar y con un vocabulario consensual. El componente asertivo contiene las descripciones de las características y relaciones explícitas de los recursos. Mientras, el componente terminológico contiene los axiomas que definen y restringen la manera en que se relacionan los recursos.

La propuesta de metodología permite construir, utilizar e implementar los componentes de esta arquitectura.

Esta arquitectura es genérica y puede implementarse en cualquier memoria corporativa. La Figura ?? muestra esta arquitectura, en la que cada componente se desarrolla en una de las tres etapas de nuestra metodología.

Esta arquitectura es genérica y se puede implementar en cualquier memoria corporativa asociada a un dominio. La Figura 1 muestra la arquitectura, en la cual varios de los componentes son desarrollados en alguna de las tres etapas de esta ISR.

La arquitectura para la integración semántica de recursos se muestra en la Figura ??.

Esta arquitectura se desarrollo

5.1. Representación del conocimiento en los recursos

5.2. Enriquecimiento del conocimiento en el modelo

Hay que mostrar las jerarquías de clases y de propiedades, así como los axiomas de nuestras ontologías.

Dado que los casos de uso son independientes, decidimos utilizar una ontología para cada uno y otra que es de uso común (ODARyT).

5.3. Búsqueda y recuperación de información en el modelo

Aquí hay que retomar nuestros casos de uso y expandirlos con consultas específicas que se hacen sobre las ontologías definidas en la sección anterior.

Es importante hacer notar que ciertas consultas explotan a los axiomas, por lo que se requiere de un razonador.

Capítulo 6

Prototipo

La manera de probar el enfoque semántico y la integración semántica es desarrollar un prototipo de sistema.

Este prototipo basa su funcionamiento en los elementos, tecnologías y estándares actuales de la Web Semántica.

Nosotros para construir el sistema seguimos una serie de actividades. En donde estas actividades se apegan a la metodología que propusimos.

Evaluación experimental

7.1. Escenarios de experimentación

Algún texto...

7.2. Experimentación

Más texto...

7.3. Resultados

Más texto...

Conclusiones y Trabajo Futuro

Aunque los buscadores actuales entregan un conjunto de resultados en poco tiempo. Muchos de éstos no satisfacen la pregunta dada por un usuario. En cambio, al hacer una búsqueda basada en la semántica de los recursos. Los resultados entregados serán más significativos para un usuario. Nuestra propuesta se basa en ésta idea. Así como el uso de conceptos y estándares de la Web Semántica. Un razonador para una interrogación más inteligente. En donde el dominio para nuestra propuesta es el de Redes y Telecomunicaciones. Los recursos que sean devueltos por nuestra propuesta. Los vamos a evaluar con base en la opinión de los usuarios del dominio. Así como los valores proporcionados por las métricas de la Recuperación de la Información. Aunque nuestra propuesta es para el dominio RyT. El objetivo a largo plazo de nuestra propuesta se implemente en otros dominios.

Los recursos son los elementos clave para la adquisición del conocimiento en una organización. Si existe una buena gestión de este conocimiento, la organización tiene ventajas competitivas. En las tecnologías de la información existen muchos instrumentos para lograr esta gestión. El enfoque semántico que estamos investigando requiere varias actividades y mucho tiempo para adaptarlo a una aplicación. Sin embargo, hay varias ventajas de usar este enfoque. Por ejemplo, tener bien definido el significado de los recursos, aprovechar el potencial de los metadatos en los recursos, aprovechar la información explícita e implícita, hacer el conocimiento comprensible por las personas y procesos automáticos, colaborar con otras personas y con procesos automáticos, establecer estándares en la Web, etc. Este enfoque es curioso y nosotros lo asociamos con una analogía de la vida. ¿Para entender y comprender a una persona, debemos comunicarnos con esta persona. Podemos empezar por conocer su nombre, edad, interés y otras cualidades. De esta manera, podemos inferir si esta persona es afín a nuestros intereses o si es una persona de la que podemos aprender?. En este sentido, un recurso no sabemos qué utilidad tiene. Solamente hasta que empezamos a entender su significado. Por esta razón un proceso automático debe ¿comprender que representa un recurso? para que los resultados que devuelva sea los apropiados para las personas. Nuestro prototipo de sistema aún está en fase de construcción. Sin embargo, nosotros hemos obtenido experiencia y también experimentado con todos los elementos de la Web Semántica. Específicamente, las actividades hechas hasta la redacción de este artículo son: 1) investigar los elementos de la web semántica, 2) identificar los casos de uso, 3) construir nuestros modelos de ontologías en OWL RDF y posteriormente transformarlo a un archivo OWL con la herramienta protégé, 4) verificar que el modelo es consistente y hacer inferencias con el razonador, 6) consultar las

descripciones semánticas empleando SPARQL, 7) construir un GoogleForm para construir las descripciones semánticas y finalmente 8) proponer una metodología para la construcción de un sistema Semántico Integrador de los Recursos (SIR) de una memoria corporativa. En nuestra experiencia hemos identificado varios nichos de oportunidades. Específicamente, nosotros hemos identificado la necesidad de automatizar varias actividades. También, aunque nuestro sistema solo emplea un subconjunto de recursos del área de Redes y Telecomunicaciones. Nosotros estamos seguros que se puede extender el alcance (en futuros proyectos) para gestionar todos los recursos del área. En donde, todos los recursos deben guiarse por los casos de uso. De esta manera, habrá orden y se construirá un adecuado modelo de ontología. Finalmente, nosotros aún tenemos actividades por investigar, desarrollar y terminar. Como, terminar las descripciones semánticas de los recursos de la memoria corporativa, agregar los axiomas necesarios a las ontologías del sistema SIR, instalar y probar el Framework Jena, así como implementar el sistema SIR y evaluarlo. Este trabajo es laborioso, pero nuestra motivación es obtener los beneficios del enfoque semántico.

Los recursos son los elementos clave para la adquisición del conocimiento en una organización. Si existe una buena gestión de este conocimiento, la organización tiene ventajas competitivas. En las tecnologías de la información existen muchos instrumentos para lograr esta gestión. El enfoque semántico que estamos investigando requiere varias actividades y mucho tiempo para adaptarlo a una aplicación. Sin embargo, hay varias ventajas de usar este enfoque. Por ejemplo, tener bien definido el significado de los recursos, aprovechar el potencial de los metadatos en los recursos, aprovechar la información explícita e implícita, hacer el conocimiento comprensible por las personas y procesos automáticos, colaborar con otras personas y con procesos automáticos, establecer estándares en la Web, etc. Este enfoque es curioso y nosotros lo asociamos con una analogía de la vida. ¿Para entender y comprender a una persona, debemos comunicarnos con esta persona. Podemos empezar por conocer su nombre, edad, interés y otras cualidades. De esta manera, podemos inferir si esta persona es afín a nuestros intereses o si es una persona de la que podemos aprender?. En este sentido, un recurso no sabemos qué utilidad tiene. Solamente hasta que empezamos a entender su significado. Por esta razón un proceso automático debe ¿comprender que representa un recurso? para que los resultados que devuelva sea los apropiados para las personas. Nuestro prototipo de sistema aún está en fase de construcción. Sin embargo, nosotros hemos obtenido experiencia y también experimentado con todos los elementos de la Web Semántica. Específicamente, las actividades hechas hasta la redacción de este artículo son: 1) investigar los elementos de la web semántica, 2) identificar los casos de uso, 3) construir nuestros modelos de ontologías en OWL RDF y posteriormente transformarlo a un archivo OWL con la herramienta protégé, 4) verificar que el modelo es consistente y hacer inferencias con el razonador, 6) consultar las descripciones semánticas empleando SPARQL, 7) construir un GoogleForm para construir las descripciones semánticas y finalmente 8) proponer una metodología para la construcción de un sistema Semántico Integrador de los Recursos (SIR) de una memoria corporativa. En nuestra experiencia hemos identificado varios nichos de oportunidades. Específicamente, nosotros hemos identificado la necesidad de automatizar varias actividades. También, aunque nuestro sistema solo emplea un subconjunto de recursos del área de Redes y Telecomuni-

caciones. Nosotros estamos seguros que se puede extender el alcance (en futuros proyectos) para gestionar todos los recursos del área. En donde, todos los recursos deben guiarse por los casos de uso. De esta manera, habrá orden y se construirá un adecuado modelo de ontología. Finalmente, nosotros aún tenemos actividades por investigar, desarrollar y terminar. Como, terminar las descripciones semánticas de los recursos de la memoria corporativa, agregar los axiomas necesarios a las ontologías del sistema SIR, instalar y probar el Framework Jena, así como implementar el sistema SIR y evaluarlo. Este trabajo es laborioso, pero nuestra motivación es obtener los beneficios del enfoque semántico.

Appendices

Apéndice A

Códigos interfaz de Usuario

Bibliografía

- [1] L. Gandon, Fabien. Ontology Engineering: a Survey and a Return on Experience. Technical Report RR-4396, INRIA, March 2002.
- [2] T. Berners-Lee, R. Fielding, and L. Masinter. Uniform resource identifiers (uri): Generic syntax. 1998.
- [3] James G. March, Herbert A. Simon, and Harold S. Guetzkow. *Teoría de la Organización*. Ariel, 1987.
- [4] Richard L. Daft. *Teoría Y Diseño Organizacional*. Cengage Learning, 09 edition, 2007.
- [5] Reinaldo O. Silva. *Teorías de la administración*. Thomson, 01 edition, 2002.
- [6] Juan J. Gilli. *Diseño organizativo: estructura y procesos*. Granica, 2007.
- [7] Sergey Brin and Lawrence Page. The anatomy of a large-scale hypertextual web search engine. In *Proceedings of the seventh international conference on World Wide Web 7*, WWW7, pages 107–117, Amsterdam, The Netherlands, The Netherlands, 1998. Elsevier Science Publishers B. V.
- [8] Peter Rob and Carlos Coronel. *Sistemas de bases de datos: diseño, implementación y administración*. Thomson, 05 edition, 2004.
- [9] S. Alfred, A. Arpah, L. H S Lim, and K. K S Sarinder. Semantic technology: An efficient approach to monogenean information retrieval. In *Computer and Network Technology (ICCNT), 2010 Second International Conference on*, pages 591–594, 2010.
- [10] Rose Dieng, Olivier Corby, Alain Giboin, and Myriam Ribi re. Methods and Tools for Corporate Knowledge Management. Technical Report RR-3485, INRIA, September 1998.
- [11] John Lyons. *Sem ntica Ling  stica: Una Introducci n*. Paid s Iberica, 1997.
- [12] Torcoroma Vel squez P rez, Andr s Puentes Vel squez, and Jaime Guzm n Luna. Ontologias: una tecnica de representacion de conocimiento. *Avances en Sistemas e Inform tica*, 8(2), 2011.
- [13] S. Bouzid, C. Cauvet, and J. Pinaton. A survey of semantic web standards to representing knowledge in problem solving situations. In *Information Retrieval Knowledge Management (CAMP), 2012 International Conference on*, pages 121–125, 2012.

-
- [14] C. Gueret, S. Schlobach, K. Dentler, M. Schut, and G. Eiben. Evolutionary and swarm computing for the semantic web. *Computational Intelligence Magazine, IEEE*, 7(2):16–31, 2012.
 - [15] Tim Berners-Lee, Roy T. Fielding, and Larry Masinter. Uniform resource identifier (URI): Generic syntax. RFC 3986, RFC Editor, January 2005.
 - [16] Nigel Shadbolt, Tim Berners-Lee, and Wendy Hall. The semantic web revisited. *IEEE Intelligent Systems*, 21(3):96–101, May 2006.
 - [17] T. Fujino and N. Fukuta. A sparql query rewriting approach on heterogeneous ontologies with mapping reliability. In *Advanced Applied Informatics (IIAIAI), 2012 IIAI International Conference on*, pages 230–235, 2012.
 - [18] Thomas R. Gruber. A translation approach to portable ontology specifications. *Knowl. Acquis.*, 5(2):199–220, June 1993.
 - [19] Nicola Guarino, Daniel Oberle, and Steffen Staab. What is an ontology? In Steffen Staab and Rudi Studer, editors, *Handbook on Ontologies*, International Handbooks on Information Systems, pages 1–17. Springer Berlin Heidelberg, 2009.
 - [20] Markus Krötzsch, František Simančík, and Ian Horrocks. A description logic primer. *Computing Research Repository (CoRR)*, abs/1201.4089, 2012.
 - [21] Matthew Horridge, Holger Knublauch, Alan Rector, Robert Stevens, and Chris Wroe. A Practical Guide To Building OWL Ontologies Using The Protege-OWL Plugin and CO-ODE Tools Edition 1.0. August 2004.
 - [22] Magdalena Ortiz. Introducción a las Lógicas Descriptivas. Technical report, Vienna University of Technology, 2009.
 - [23] Yun Lin and John Krogstie. Semantic annotation of process models for facilitating process knowledge management. *Int. J. Inf. Syst. Model. Des.*, 1(3):45–67, July 2010.
 - [24] Bernardo Cuenca Grau, Ian Horrocks, Boris Motik, Bijan Parsia, Peter Patel-Schneider, and Ulrike Sattler. Owl 2: The next step for owl. *Web Semant.*, 6(4):309–322, November 2008.
 - [25] R. B. Mishra and Sandeep Kumar. Semantic web reasoners and languages. *Artif. Intell. Rev.*, 35(4):339–368, April 2011.
 - [26] A.Q. Al-Namiy and F.S. Majeed. Towards automatic extracted semantic annotation (esa) for web documents. In *Information Processing, 2009. APCIP 2009. Asia-Pacific Conference on*, volume 2, pages 614–617, 2009.
 - [27] A. Norta, R. Yangarber, and L. Carlson. Utility evaluation of tools for collaborative development and maintenance of ontologies. In *Enterprise Distributed Object Computing Conference Workshops (EDOCW), 2010 14th IEEE International*, pages 207–214, 2010.
 - [28] Li Ding, Pranam Kolari, Zhongli Ding, and Sasikanth Avancha. Using ontologies in the semantic web: A survey. In Raj Sharman, Rajiv Kishore, and Ram Ramesh, editors, *Ontologies*, volume 14 of *Integrated Series in Information Systems*, pages 79–113. Springer US, 2007.
-

-
- [29] T. Aruna, K. Saranya, and C. Bhandari. A survey on ontology evaluation tools. In *Process Automation, Control and Computing (PACC), 2011 International Conference on*, pages 1–5, 2011.
- [30] *Archetype-Based Semantic Integration and Standardization of Clinical Data*, 2006.
- [31] Kai Yang and R. Steele. A semantic integration solution for online accommodation information integration. In *Industrial Electronics and Applications (ICIEA), 2011 6th IEEE Conference on*, pages 1105–1110, 2011.
- [32] Jun Zhai, Jianfeng Li, and Qinglian Wang. Using ontology and xml for semantic integration of electricity information systems. In *Electric Utility Deregulation and Restructuring and Power Technologies, 2008. DRPT 2008. Third International Conference on*, pages 2197–2201, 2008.
- [33] Wang Xin and Xiong Guangleng. Design rationale as part of corporate technical memory. In *Systems, Man, and Cybernetics, 2001 IEEE International Conference on*, volume 3, pages 1904–1908 vol.3, 2001.
- [34] R. Chakhmoune, H. Behja, and A. Marzak. Building corporate memories in collaborative way using ontologies: Case study of a ssii. In *Next Generation Networks and Services (NGNS), 2011 3rd International Conference on*, pages 23–28, 2011.
- [35] Victoria Uren, Philipp Cimiano, José Iria, Siegfried Handschuh, Maria Vargas-Vera, Enrico Motta, and Fabio Ciravegna. Semantic annotation for knowledge management: Requirements and a survey of the state of the art. *Web Semantics: Science, Services and Agents on the World Wide Web*, 4(1):14–28, January 2006.
- [36] Oscar Corcho. Ontology based document annotation: trends and open research problems. *Int. J. Metadata Semant. Ontologies*, 1(1):47–57, 2006.
- [37] Hamish Cunningham, Diana Maynard, Kalina Bontcheva, Valentin Tablan, Niraj Aswani, Ian Roberts, Genevieve Gorrell, Adam Funk, Angus Roberts, Danica Damjanovic, Thomas Heitz, Mark A. Greenwood, Horacio Saggion, Johann Petrak, Yaoyong Li, and Wim Peters. *Text Processing with GATE (Version 6)*. 2011.
- [38] N. Islam, M.S. Siddiqui, and Z.A. Shaikh. Tode : A dot net based tool for ontology development and editing. In *Computer Engineering and Technology (ICCET), 2010 2nd International Conference on*, volume 6, pages V6–229–V6–233, 2010.
- [39] Holger Knublauch, Ray W. Ferguson, Natalya F. Noy, and Mark A. Musen. The Protégé OWL Plugin: An Open Development Environment for Semantic Web Applications. In Sheila .. McIlraith, Dimitris Plexousakis, and r. a. n. k. van, Harmelen, editors, *The Semantic Web - ISWC 2004*, volume 3298 of *Lecture Notes in Computer Science*, chapter 17, pages 229–243. Springer Berlin / Heidelberg, Berlin, Heidelberg, 2004.
- [40] Sören Auer. Powl - a web based platform for collaborative semantic web development. In *Proceeding of 1st Workshop Scripting for the Semantic Web (SFSW'05), Hersonissos, Greece, May 30*. CEUR Workshop Proceedings, May 2005.
-

- [41] Walter Waterfeld, Moritz Weiten, and Peter Haase. Ontology management infrastructures. In Martin Hepp, Pieter Leenheer, Aldo Moor, and York Sure, editors, *Ontology Management*, volume 7 of *Computing for Human Experience*, pages 59–87. Springer US, 2008.
 - [42] Aditya Kalyanpur, Bijan Parsia, Evren Sirin, Bernardo Cuenca Grau, and James Hendler. Swoop: A web ontology editing browser. *Web Semant.*, 4(2):144–153, June 2006.
 - [43] Mohamed Morsey, Jens Lehmann, Sören Auer, Claus Stadler, and Sebastian Hellmann. DBpedia and the Live Extraction of Structured Data from Wikipedia. *Program: electronic library and information systems*, 46:27, 2012.
 - [44] B. McBride. Jena: a semantic web toolkit. *Internet Computing, IEEE*, 6(6):55–59, 2002.
 - [45] Karlis Cerans, Guntis Barzdins, Renars Liepins, Julija Ovcinnikova, Sergejs Rikacovs, and Ar-turs Sprogis. Graphical schema editing for stardog owl/rdf databases using owlged/s. In Pavel Klinov and Matthew Horridge, editors, *OWLED*, volume 849 of *CEUR Workshop Proceedings*. CEUR-WS.org, 2012.
 - [46] M. Salvadores, G. Correndo, T. Omitola, N. Gibbins, S. Harris, and N. Shadbolt. 4s-reasoner: Rdfs backward chained reasoning support in 4store. In *Web Intelligence and Intelligent Agent Technology (WI-IAT), 2010 IEEE/WIC/ACM International Conference on*, volume 3, pages 261–264, 2010.
 - [47] Jeen Broekstra, Arjohn Kampman, and Frank van Harmelen. Sesame: A generic architecture for storing and querying rdf and rdf schema. In *Proceedings of the First International Semantic Web Conference on The Semantic Web*, ISWC '02, pages 54–68, London, UK, UK, 2002. Springer-Verlag.
-