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

24 de noviembre 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 busqueda 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 busqueda 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 tambien 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 está tesis se hicieron dos ontologías; una para cada caso de uso, y tambien 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 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, tambien 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

Lista de Figuras

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	34
IDE	Entorno de Desarrollo Integrado	36
API	Interfaz de Programación de Aplicaciones	37
ISR	Integración Semántica de los Recursos	39
CSV	Valores Separados por Coma	48

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, eficientar 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) *uun modelo relacional inadecuado propicia a tener datos inconsistentes, lo que provoca, problemas en la generación y validación de la información* [8], y 5) *perdida 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 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.

[1] L. Feigenbaum, "Semantic Web vs. Semantic Technologies", Disponible en: `http://www.cambridgesemantics.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 telecomunicaicones.

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). Esta 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.

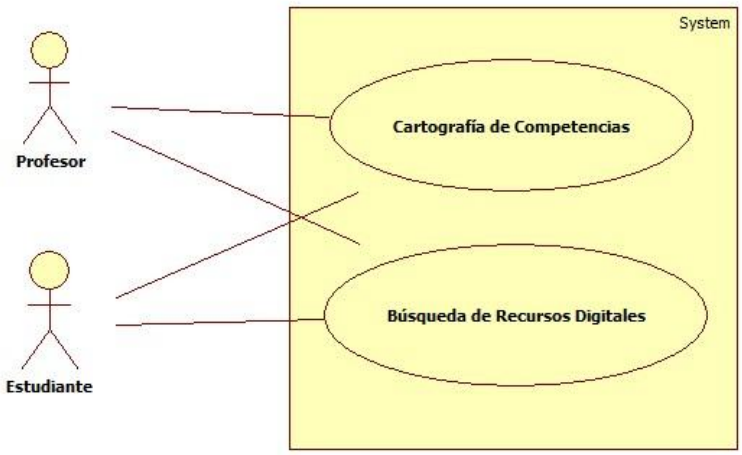

Figura 2.1: Diagrama de casos de uso para la integración de los recursos de una memoria corporativa.

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 los 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*** (RDF[2]) 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,

[1] L. Feigenbaum, "Semantic Web vs. Semantic Technologies," Disponible en: `http://www.cambridgesemantics.com/semantic-university/semantic-web-vs-semantic-technologies`

[2] W3C,"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.

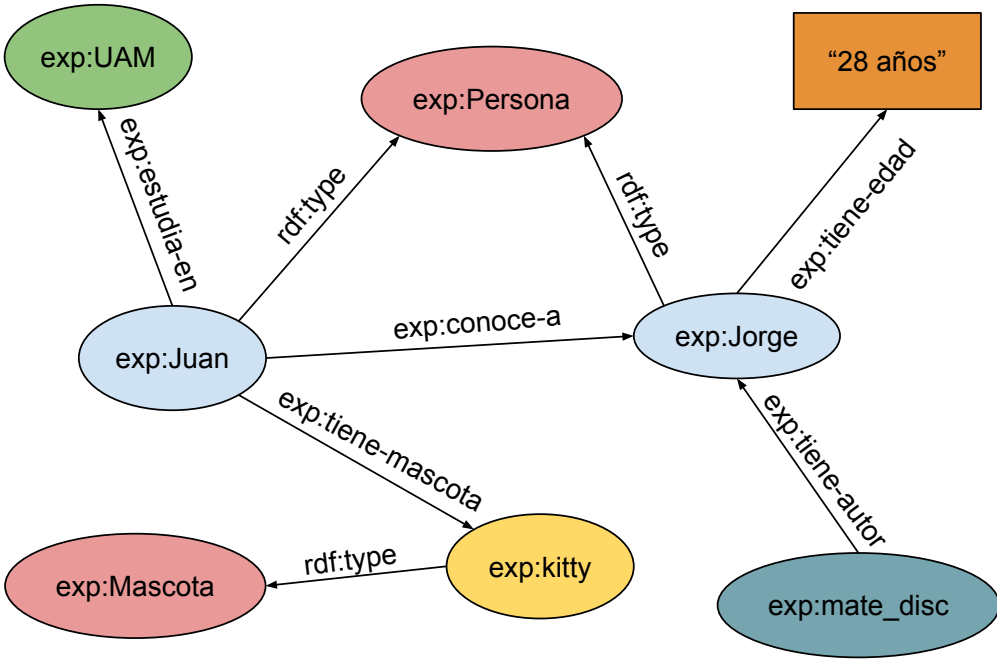

Figura 3.2: Ejemplo de un grafo RDF o grafo de conocimientos.

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

sintaxis son: N3[3], turtle[4], RDF/XML[5], N-triples[6]. 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 tripleta***. Un *patrón tripleta* 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 tripleta* 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

[3]W3C,"Notation3 (N3)," Disponible en: `http://www.w3.org/TeamSubmission/n3/`
[4]W3C,"Turtle," Disponible en: `http://www.w3.org/TR/2013/CR-turtle-20130219/`
[5]W3C,"RDF/XML Syntax Specification," Disponible en: `http://www.w3.org/TR/rdf-syntax-grammar/`
[6]W3C,"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. 7) 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 Ontologias Web* (OWL[8]). 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.*

[7]W3C, "RDF Vocabulary Description Language 1.0: RDF Schema," Dispoible en: `http://www.w3.org/TR/2004/REC-rdf-schema-20040210/`

[8]W3C,"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 Rocio 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 funcionalidades[9] (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*

[9]W3C,"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 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 lo 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 R2RML[10] 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 ontologías[11] (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 información[12], y 2) ***Friend Of A Friend*** (FOAF) es un vocabulario para describir a las personas y las relaciones de éstas en la Web[13].

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.

[10]W3C, "R2RML: RDB to RDF Mapping Language," Disponible en: `http://www.w3.org/TR/r2rml/`
[11]W3C,"Good Ontologies," Disponible en: `http://www.w3.org/wiki/Good_Ontologies`
[12]Dublin Core Semantic Initiative,"Dublin Core," Disponible en: `http://dublincore.org/`
[13]FOAF 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.

Ontología

```
ABox =
    {
        exp:Juan rdf:type exp:Hombre,
        exp:Laura rdf:type exp:Mujer,
        exp:Luis rdf:type exp:Estudiante.
    }

TBox =
    {
        exp:Hombre rdfs:subClassOf exp:Persona,
        exp:Mujer rdfs:subClassOf exp:Persona,
        exp:Estudiante rdfs:subClassOf exp:Persona.
    }
```

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.

Ontología+

```
ABox+ =
    {
        exp:Juan rdf:type exp:Hombre,
        exp:Laura rdf:type exp:Mujer,
        exp:Luis rdf:type exp:Estudiante,
        exp:Juan rdf:type exp:Persona,
        exp:Laura rdf:type exp:Persona,
        exp:Luis rdf:type exp:Persona.
    }

TBox =
    {
        exp:Hombre rdfs:subClassOf exp:Persona,
        exp:Mujer rdfs:subClassOf exp:Persona,
        exp:Estudiante rdfs:subClassOf exp:Persona.
    }
```

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 guardarlas 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 éstos. 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 esta 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 está es que los profesionales de la salud y expertos en tecnologías de la información sean los principales constructores del conocimiento.

Éstas son las características asociadas a nuestros criterios de investigación (Ver Tabla 4.1): 1) ***Integración de la información****, el objetivo de la arquitectura dual es facilitar la información clínica de los pacientes a los profesionales de la salud* y 2) ***Modelo semántico****, la arquitectura dual por un lado representa las estructuras de datos comunes y por otro lado, representar el conocimiento formal de conceptos clínicos a partir del uso de arquetipos.*

- 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 resuelva 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.

Éstas son características del *marco de integración semántica* que están vinculadas a nuestros criterios de investigación (Ver Tabla 4.1): 1) ***Integración de la información****, este trabajo es una propuesta para reunir y compartir la información del alojamiento en-línea* y 2) ***Modelo semántico****, este marco propone el uso de una ontología para modelar la información del dominio de alojamiento en-línea.*

- 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.

En el trabajo de Jun Zhai, éstas son las características asociadas a nuestros criterios de investigación: 1) ***Integración de información****, este trabajo propone una integración semántica para especificar la información a nivel semántico, proveniente de los sistemas de información de energía eléctrica*, 2) ***Modelo semántico****, Jun Zhai et al. emplean a XML para el intercambio de información entre estos sistemas, así como una ontología para describir formalmente los conceptos en el dominio de la electricidad* y 3) ***Inferencia en el modelo****, este trabajo emplea al proceso de inferencia como herramienta para la construcción de modelos (conceptos y datos XML) consistentes.*

- Xin y Guangleng [33] emplean un *enfoque basado en las* ***ontologías*** *para capturar la* ***información de la "justificación del diseño"***. Esta *justificación del diseño (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 la *justificación del diseño.* Las ontologías permiten el acceso uniforme a las fuentes de información, y en este trabajo, éstas modelan la *justificación del diseño* para el background del diseño de autos de carga.

En este trabajo de Xin y Guangleng, éstas son las características asociadas a nuestros criterios de investigación: ***Integración de la información****, la finalidad de este trabajo es capturar, reutilizar y comunicar la información de las distintas tareas de la "justificación del diseño"*, 2) ***Memoria Corporativa****, este trabajo emplea una memoria corporativa para las actividades de captura y disponibilidad en la "justificación del diseño"* y 3) ***Modelo semántico****, Xin y Guangleng emplean a las ontologías para capturar la información de la "justificación del diseño" y tener el acceso de manera uniforme a los recursos de información*

- ***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.

Éstas son características del *PCOGEME* que están vinculadas a nuestros criterios de investigación: 1) ***Memoria corporativa****, en este trabajo se emplea a la memoria corporativa como mecanismo para la gestión del conocimiento y los documentos* y 2) ***Modelo semántico****, este trabajo propone a las ontologías como medio para la representación y gestión de/en una memoria corporativa*

Para concluir esta Sección 4.1, se presenta la Tabla 4.2, la cual resume todos 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.

En esta Tabla 4.2, el valor '**Sí**' 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.

Trabajo	IIR	MC	MS	IeM	IVpI
Arquitectura del modelo dual	Sí	No	Modelo de referencia y arquetipos	No	No
Marco de integración semántica	Sí	No	Información XML y ontología global	No	No
Arquitectura para la integración en SIEE[a]	Sí	No	tripletas RDF y axiomas RDF(S)	Consistencia	No
Metodología para ontologías en la "justificación del diseño"	Sí	Justificación del diseño	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 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 marcado[1]. 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 semántico[2]. 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.
- ***Aktive Media*** [35] es una GUI para la descripción automática de una colección de

[1] M. Siroker, "OntoMat Annotizer," Disponible en: `http://projects.semwebcentral.org/projects/ontomat/`
[2] The Open University, "MnM," Disponible en: `http://projects.kmi.open.ac.uk/akt/MnM/`

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*[3]". Estas sugerencias se hacen con base en axiomas y descripciones previas.

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 cambia las URIs de los recursos y propiedades por sus propios URIs*, 4) *OntoMat Annotizer y MnM no tienen versión estable* y 5) *Aktive Media y MnM no tienen documentación disponible para solucionar problemas de configuración.*

Un ***editor de ontología*** [37] 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*** [38] es una plataforma con herramientas para la creación, visualización y manipulación de ontologías en diversos formatos de representación[4]. 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 License[5].

- ***pOWL*** [39] es una herramienta para la visualización y edición de ontologías vía web[6]. 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*** [40] es un *entorno de desarrollo integrado*(IDE) para "*desarrollar, gestionar y probar configuraciones de los modelos de conocimiento e instancias de las bases de conocimiento*"[7]. Esta herramienta proporciona un conjunto de editores para visualizar grafos RDF y diagramas de clase. Existen tres versiones de esta

[3]A. Chakravarthy, V. Lanfranchi , F. Ciravegna, "AKTive Media," Disponible en: `http://www.aktors.org/technologies/aktivemedia/index.html`

[4]Stanford Center for Biomedical Informatics Research, "Protégé," Disponible en: `http://protege.stanford.edu/`

[5]Mozilla, "Mozilla Public License," Disponible en: `http://www.mozilla.org/MPL/`

[6]Sören Auer, "pOWL," Disponible en: `http://aksw.org/Projects/Powl.html`

[7]TopQuadrant, Inc., "TopBraid Composer," Disponible en: `http://www.topquadrant.com/products/TB_Composer.html`

herramienta: maestro, estándar y gratuita. La versión gratuita permite crear y editar archivos OWL/XML, así como consultar con el lenguaje SPARQL.

- ***SWOOP*** [41] es un editor para crear y editar ontologías, comprobar inconsistencias, navegar por las ontologías, compartir y reutilizar los datos existentes[8]. 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: Pellet*[9]*, Fact++*[10] *y HermiT*[11] y 7) *incorpora un plugin para escribir y ejecutar consultas SPARQL desde la interfaz de usuario.*

Un ***triplestore*** [42] 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 describen cuatro triplestores que estudiamos de muchos que hay disponibles[12].

- ***Apache Jena*** [43] 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 conocimiento[13].

- ***Stardog*** [44] 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 directa[14]. Esta herramienta emplea los protocolos *HTTP y SNARL* para *acceder y*

[8]University of Maryland, "SWOOP," Disponible en: `https://code.google.com/p/swoop/downloads/list`

[9]Clark & Parsia, LLC, "Pellet," Disponible en: `http://clarkparsia.com/pellet/`

[10]Clark & Parsia, LLC, "Pellet," Disponible en: `http://clarkparsia.com/pellet/`

[11]Oxford University, "HermiT," Disponible en: `http://hermit-reasoner.com/`

[12]W3C, "Triple Store," Disponible en: `http://www.w3.org/2001/sw/wiki/Category:Triple_Store`

[13]The Apache Software Foundation, "Apache Jena," Disponible en: `http://jena.apache.org/`

[14]Clark & 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*** [45] *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*** [46] *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 cuatro triplestores son opciones viables para efectuar las 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 que soporta varios axiomas*[19] *OWL y RDF(S)* y 6) *una amplia comunidad de desarrolladores.*

4.3. Resumen

Necesitamos una sección que resuma este capítulo. Termina muy abruptamente. Hay que decir claramente (1) cuáles son las características distintivas de nuestro enfoque, (2) por qué no utilizamos una herramienta como MnM para generar RDF y qué hicimos en su lugar (utilizamos un script para generar RDF artificial), y (3) qué editor de ontologías y triplestore utilizamos y para qué.

Un ***script*** es un programa para generar tripletas RDF. 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.

[15]Garlik, "4store," Disponible en: `http://4store.org/`

[16]Aduna, "Sesame," Disponible en: `http://www.openrdf.org/index.jsp`

[17]The Eclipse Foundation, "Eclipse IDE," Disponible en: `http://www.eclipse.org/`

[18]The Eclipse Foundation, "Licencia Apache v. 2.0 ," Disponible en: `http://www.apache.org/licenses/LICENSE-2.0.html`

[19]The Apache Software Foundation, "OWL coverage," Disponible en: `http://jena.apache.org/documentation/inference/index.html#OWLcoverage`

Capítulo 5

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)***.

La *integración semántica de recursos* puede implementarse en una *memoria corporativa (MC)*. Porque una memoria tiene un conjunto diverso de recursos de información, los cuales representan el conocimiento en una organización (dominio particular). Estas son las principales razones de esta *integración en una memoria corporativa*: 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 está guiada por dos *casos de uso*.

La ***finalidad*** de esta propuesta es *facilitar y guiar a los desarrolladores en estas dos tareas: 1) construir un modelo semántico (otologías) y 2) consultar información en este modelo*. Mientras, los principales objetivos de esta integración semántica 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, la primera y segunda etapa sirven para construir el modelo semántico a partir del conocimiento explícito e implícito en los recursos de información. La tercera etapa consiste en identificar y efectuar las principales consultas sobre el modelo semántico.

En esta metodología, el elemento clave es el ***caso de uso***. Porque un caso de uso encamina en el desarrollo de la integración semántica. En concreto, los *casos de uso* permiten encontrar: 1) *qué características y relaciones son significativas*, 2) *qué reglas de inferencia son necesarias* y 3) *cuáles consultas son importantes*.

La Figura 5.1 muestra la arquitectura para la *integración semántica de recursos*. Esta arquitectura es genérica para ser desarrollada e implementada en cualquier memoria corporativa. Los componentes de esta arquitectura se construyen, utilizan e implementan mediante el uso de nuestra propuesta de metodología.

Esta arquitectura se diseño con base en el modelo de tres capas: ***usuario, negocio*** y ***datos***.

- En la capa 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 la capa 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.

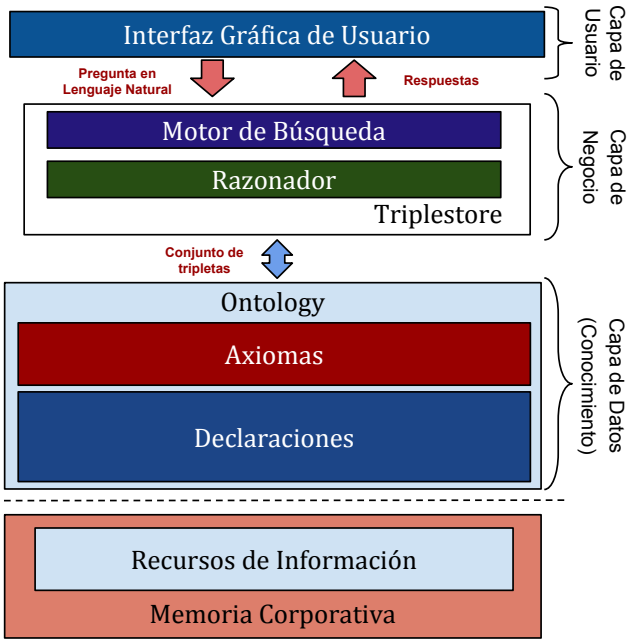

Figura 5.1: Arquitectura general para la Integración Semántica de Recurso en una Memoria Corporativa.

- En la capa 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.

Esta propuesta de metodología se pone en práctica para la *memoria corporativa* del grupo de investigación perteneciente al área de Redes y Telecomunicaciones del departamento de Ingeniería Eléctrica de la Universidad Autónoma Metropolitana Unidad Iztapalapa. 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 están a cargo de profesor del núcleo.

Los *casos de uso* básicos en esta metodología son *la cartografía de competencias* y *la búsqueda de recursos digitales*.

1. La cartografía de competencias es la búsqueda y recuperación de las personas a partir de las características personales y profesionales (competencias, capacidades, conocimientos en los temas del dominio).

2. La búsqueda de recursos digitales es la búsqueda y recuperación de los documentos y archivos multimedia a partir del contenido de éstos (temas del dominio, autor, año).

Estos dos *casos de uso* son independientes entre ellos, por tal razón, cada uno tiene una respectiva ontología. La ontología de la cartografía de competencias modela el conocimiento explícito e implícito de los recursos persona, con base en las características personales y profesionales de éstos. Mientras, la ontología de los recursos digitales modela el conocimiento explícito e implícito del contenido y acerca de éstos.

En ambos casos de uso, un aspecto importante es que tanto personas como recursos digitales se vinculan con los temas del área de Redes y Telecomunicaciones(RyT). Específicamente, los conocimientos de las personas son relaciones entre personas y temas de RyT. Mientras, los tópicos pertenecientes en los recursos digitales son las relaciones entre recursos digitales y temas de RyT. Por tal razón, se construye una tercer ontología para modelar el vocabulario consensual del área de Redes y Telecomunicaciones.

La Figura 5.2 muestra el modelo semántico en forma de un diagrama de venn. En este diagrama, las circunferencias representan las tres ontologías: cartografía de competencias, recursos digitales y vocabulario de RyT. Estas circunferencias no tienen intersección, porque cada ontología representa un determinado recurso de información (personas, documentos y multimedia, así como conceptos de RyT). En esta figura, las flechas representan el vinculo entre: *personas personas - conceptos de RyT* y *recursos digitales - conceptos de RyT*.

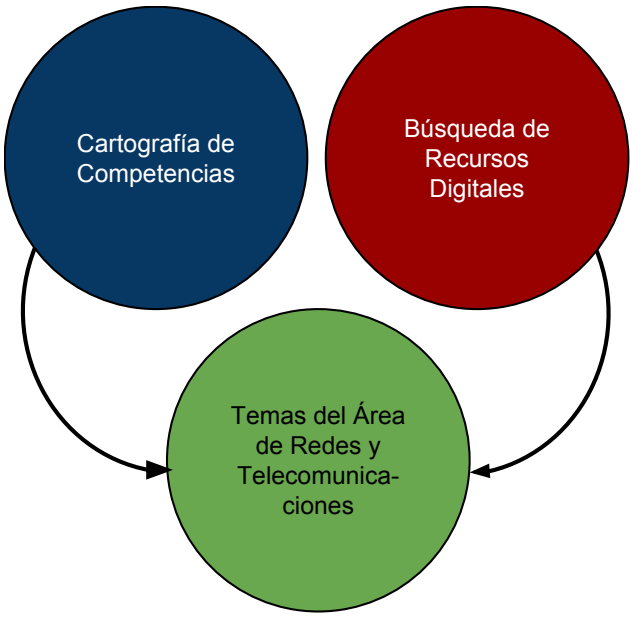

Figura 5.2: Diagrama de Venn para visualizar las tres ontologías que conforman el modelo semántico

5.1. Representación del conocimiento en los recursos

La primera actividad es ***identificar los principales recursos de información*** para construir la memoria corporativa. Esta identificación se hace a partir del análisis de los *casos*

de uso. Los recursos asociados al primer caso de uso son: profesores adscritos al área de RyT, estudiantes asociados a uno de éstos profesores, empleados de otras organizaciones que colaboran con los profesores. Mientras, los recursos digitales son: artículos científicos relacionados a los temas de investigación, libros y páginas Web de referencia, tesis de maestría y doctorado de los alumnos, reportes técnicos de los profesores y sus estudiantes, presentaciones de cursos o congresos, audios de reuniones o clases, vídeo tutoriales e imágenes de referencia.

La Figura 5.3 muestra el esquema de la memoria corporativa de área de Redes y Telecomunicaciones, donde los *recursos de información* están clasificados por caso de uso.

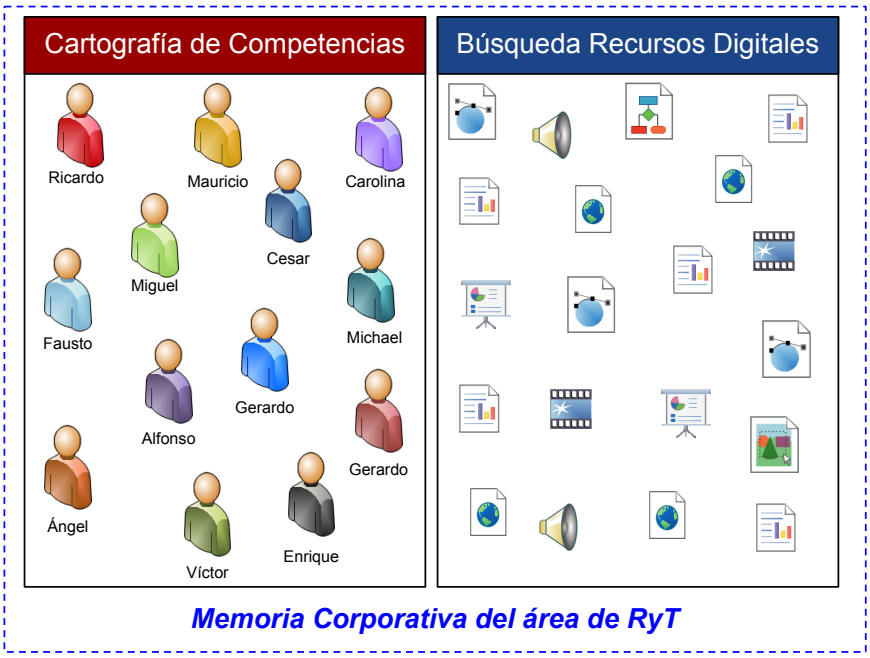

Figura 5.3: Recursos de información agrupados por casos de uso para nuestra memoria corporativa.

La siguiente actividad es ***adquirir el conocimiento o información*** en los recursos de información, mediante la utilización de los dos *casos de uso*. En esta adquisición debe considerarse un hecho importante del ***marco de trabajo RDF***, el cual es "*cualquier persona, lugar, documento, objeto abstracto o físico se representa a partir de una serie de características y relaciones significativas de éste*". Por esta razón, la *adquisición del conocimiento* se hace con base en las características y relaciones de los *recursos de información*.

Un ***diagrama de clases*** es una manera visual para mostrar un conjunto de clases, colaboraciones y sus relaciones [47]. Los diagramas de clases se utilizan para especificar y formalizar (modelar) las abstracciones y sus relaciones en un momento dado.

En la Figura 5.4 se presentan las clases, atributos(características) y relaciones entre éstas para visualizar y especificar el comportamiento de los *recursos de información* que pertenecen a la cartografía de competencias. Mientras, la Figura 5.5 muestra las clases y sus relaciones para los documentos y archivos multimedia pertenecientes a la búsqueda de recursos digitales.

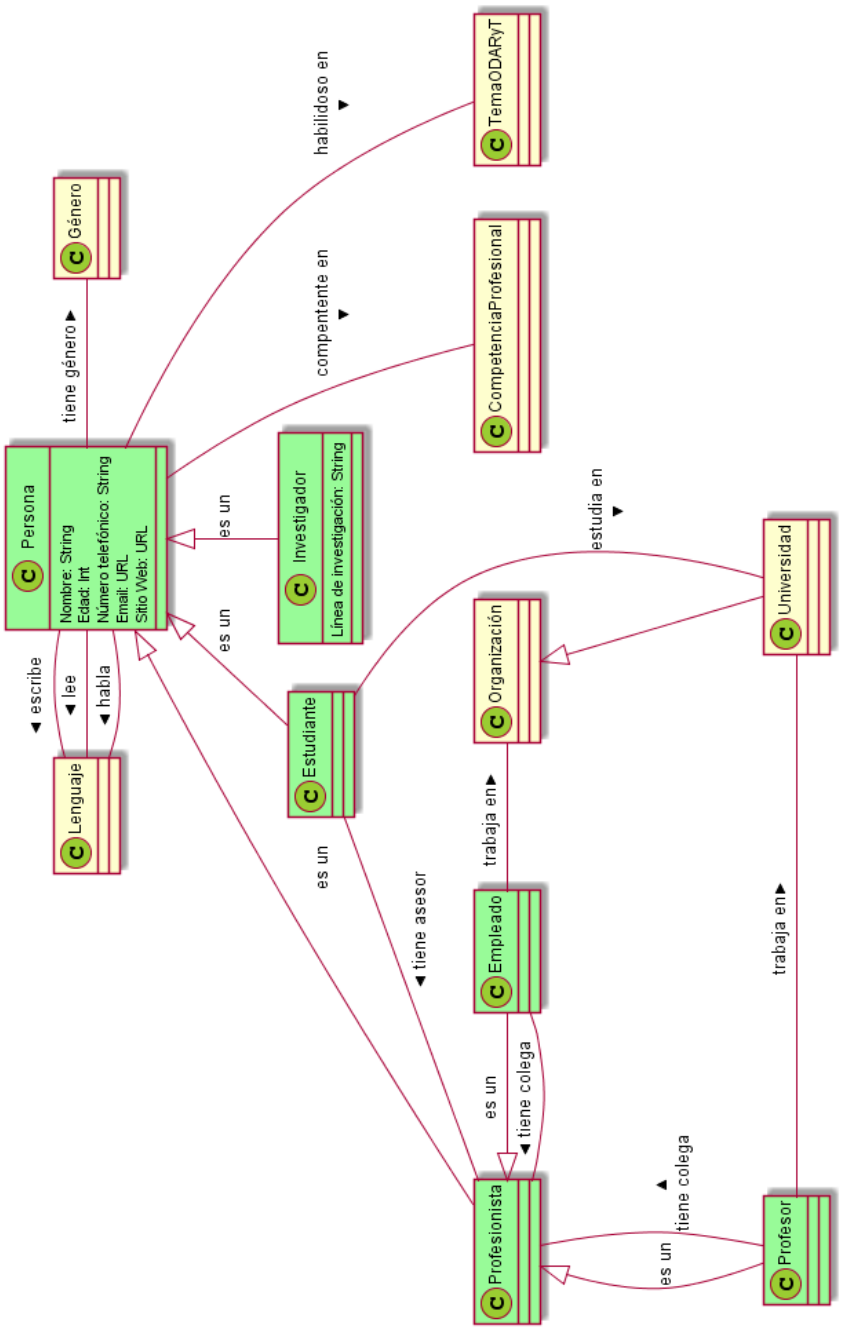

Figura 5.4: Diagrama de clases para la cartografía competencias.

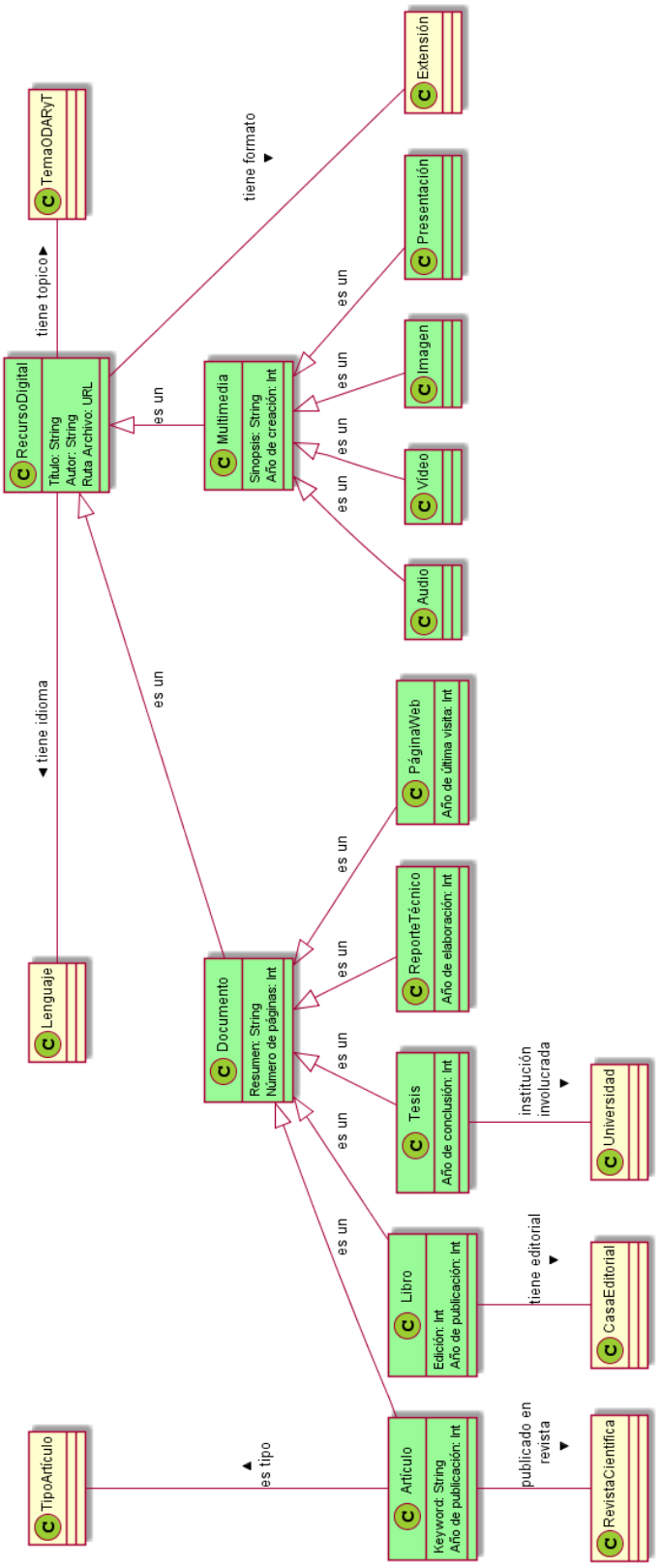

Figura 5.5: Diagrama de clases para la búsqueda de recursos digitales.

La siguiente actividad es la ***representación del conocimiento*** e información mediante el ***marco de trabajo RDF***. Esta representación tiene cuatro pasos: 1) *asignar identificadores a los recursos*, 2) *asignar identificadores a las propiedades*, 3) *reconocer si los valores de las propiedades son otros recursos o literales* y 4) *construir tripletas.*

El primer paso, es asignar un URI para cada recurso de información. Esta asignación de identificadores URI se hace con base en los casos de uso:

- Los identificadores de los recursos en la cartografía de competencia emplean el prefijo "`http://arte.izt.uam.mx/ontologies/personRyT.owl#`", el cual se abrevia "*sirp*".
- Los identificadores de los recursos pertenecientes a la búsqueda de recursos digitales, utilizan el prefijo "`http://arte.izt.uam.mx/ontologies/digiResourceRyT.owl#`", el cual se abrevia "*sird*".

La Tabla 5.1 enuncia los identificadores URI de algunos profesores del área de RyT. En esta Tabla, la primera columna tiene los nombres de los profesores y la segunda columna enuncia los identificadores URI de éstos.

Nombre	**Identificador**
Alfonso Prieto	***sirp:****AlfonsoPrieto*
Michael Pascoe	***sirp:****MichaelPascoe*
Reyna Carolina Medina	***sirp:****CarolinaMedinaRamirez*
Ricardo Marcelin	***sirp:****RicardoMarcelinJimenez*
Miguel Lopez	***sirp:****MiguelLopez*
Victor Manuel Ramos	***sirp:****VictorRamosVictorRamos*
Fausto Marcos Casco	***sirp:****FaustoCasco*
Cesar Jalpa	***sirp:****CesarJalpa*
Enrique Rodriguez	***sirp:****EnriqueRodriguez*

Tabla 5.1: Ejemplos de identificadores URI asociados a los recursos persona para la cartografía de competencias.

La Tabla 5.2 enuncia los identificadores URI de algunos recursos digitales pertenecientes a los profesores de RyT. En la primer columna de esta tabla se enuncia el nombre completo del recursos digital y la segunda columna presenta los identificadores (URI) de estos recursos.

El siguiente paso es asignar un identificador URI para cada propiedad. Estos identificadores so construidos con base en las características y relaciones en los diagramas de clases (Figuras 5.4 y 5.5). Los identificadores URI de estas propiedades dependen del caso de uso. Por un lado, si éstas pertenecen a la cartografía de competencias, entonces emplean el prefijo "*sirp*". Por otro lado, si pertenecen a la búsqueda de recursos digitales, entonces emplean el prefijo "*sird*".

Nombre	Identificador
Ontology engineering	***sird:****RR-4396-2002-pdf*
A Description Logic Primer	***sird:****A-DescriptionLogicPrimer-2012-pdf*
Introduction to Ontologies and OWL	***sird:****Introduction2Ontol-2005-pdf*
The Semantic Web - An Overview	***sird:****TheSemanticWeb-AnOverview-2011-flv*
What is Linked Data?	***sird:****What-isLinkedData-2012-flv*

Tabla 5.2: Ejemplos de identificadores URI asociados a los documentos y archivos multimedia para la búsqueda de recursos digitales.

Característica o relación	Identificador
Nombre	***sirp:****has-name*
Sitio Web	***sirp:****has-webSite*
Lugar de trabajo	***sirp:****worksIn*
Línea de investigación	***sirp:****researchesOn*
Colega	***sirp:****has-colleague*
Competencias	***sirp:****competentIn*
Habilidades en de Redes y Telecom.	***sirp:****expertiseIn*

Tabla 5.3: Ejemplos de identificadores URI de las propiedades pertenecientes a la cartografía de competencias.

La Tabla 5.3 presenta algunos identificadores de las propiedades que pertenecen a la cartografía de competencias. La primera columna enuncia las características o relaciones, mientras la segunda columna presenta los identificadores URI de estas característica.

La Tabla 5.4 presenta algunos identificadores de las propiedades de la búsqueda de recursos digitales. La primera columna enuncia las características o relaciones, mientras la segunda columna presenta los identificadores URI de estas característica.

Característica o relación	Identificador
Título	***sird:****has-title*
Autor	***sird:****has-author*
Ruta Archivo	***sird:****has-filePath*
Año de creación	***sird:****has-yearOfCreation*
Lenguaje Fuente	***sird:****has-languageSource*
Temas de Redes y Telecom.	***sird:****has-topic*

Tabla 5.4: Ejemplos de identificadores URI de las propiedades pertenecientes a la búsqueda de recursos digitales.

El siguiente paso es ***identificar el tipo de valor*** de las características en el *diagrama*

de clases. Por un lado, si el *objeto* es una cadena o un número, entonces es una *literal*. Por otro lado, si el *objeto* es otro recurso, entonces es un *identificador URI*. En el caso de las ***relaciones***, los valores son otros recursos, por ello, éstos deben ser *identificadores URI*.

En la representación del conocimiento explícito, el último paso es la ***generación*** de *tripletas RDF* asociadas a las declaraciones de los recursos.

En esta tesis, la generación de tripletas se hace mediante la combinación de formularios y scripts. Por un lado, los formularios son las herramientas para la recuperación y almacenamiento de la información acerca de los recursos en *hojas de cálculo*. Por el otro lado, los scripts mapean la información de las *hojas de cálculo* en forma de tripletas RDF para que posteriormente éstas sean almacenadas en archivos "*.rdf*".

Este es el procedimiento de generación de tripletas RDF mediante el uso de formularios y scripts:

1. Identificar la información que debe ser adquirida en los recursos con base en los diagramas de clases.

2. Construir los formularios para los recursos de información (persona, documento y multimedia) mediante Google Form[1], con el propósito de agilizar el proceso de recopilación de la información en los recursos.

3. Enviar los formularios vía email a los profesores o alumnos, para que ellos escriban la información sobre las características y relaciones de los recursos.

4. Recuperar y almacenar las respuestas de cada formulario en una de tres *hojas de cálculo* (persona, documento y multimedia).

5. Descargar la información de cada *hoja de cálculo* en un archivo CSV (persona, documento y multimedia).

6. Transformar cada fila de un archivo csv a un conjunto de tripletas RDF, mediante los scripts que están escritos en Java y con la librería Jena.

7. Almacenar las tripletas RDF asociadas a una fila (descripción semántica de un recurso) en un archivo "*rdf*" con la sintaxis de serialización Turtle.

En la Figura 5.6 se muestra la representación del Dr. Ricardo Marcelin Jiménez en forma de tripletas RDF. En la parte izquierda de esta figura, se presenta una foto representativa del Dr. Ricardo. Mientras, la parte derecha se dan a conocer las tripletas RDF en formato de serialización Turtle que describen las características y relaciones básicas de esta persona.

El identificador asociado al recurso *Ricardo Marcelin Jiménez* es "***sirp:****RicardoMarcelin Jimenez*". En la Figura 5.6, el conjunto de tripletas RDF indican las siguientes declaraciones:

[1]Google, "Formularios," Disponible en: `https://support.google.com/drive/topic/1360904?hl=es&ref_topic=2811744`

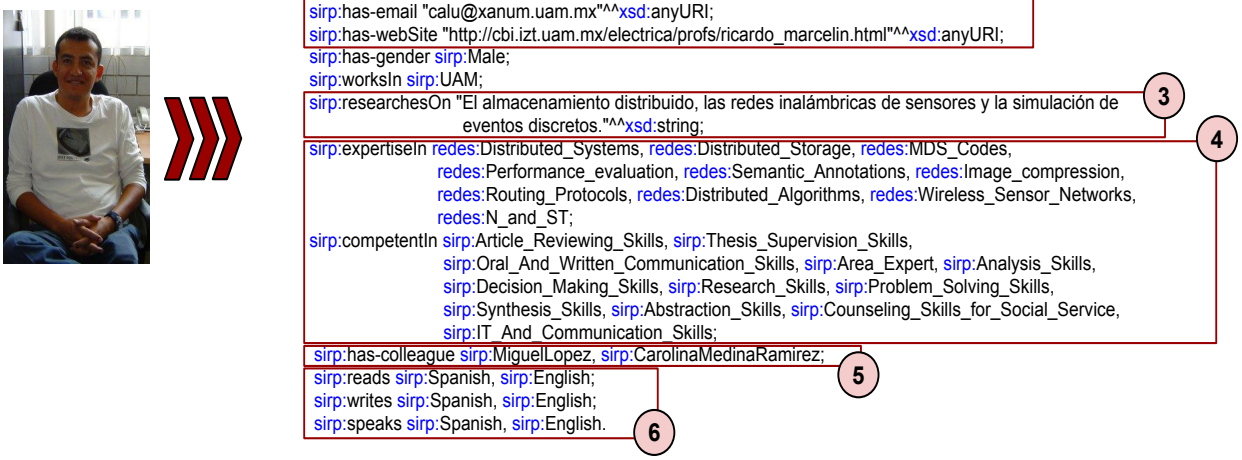

Figura 5.6: Declaraciones del Dr. Ricardo Marcelin Jiménez en forma de tripletas RDF

- En el primer recuadro, la tripleta RDF establece que este recurso pertenece a la clase Profesor.
- En el segundo recuadro, las tripletas establecen los valores para los atributos *nombre*, *email* y *sitio Web*.
- En el tercer recuadro, la tripleta RDF indica el tema o línea de investigación del Dr. Ricardo.
- En el cuarto recuadro, se establecen las habilidades en los temas de Redes y Telecomunicaciones, así como las competencias profesionales que se obtienen a partir de las competencias propuestas en el proyecto Tuning en latinoamérica [**?**].
- El quito recuadro indica las tripletas RDF que vinculan al Dr. Ricardo con sus colegas: *Miguel López* y *Carolina Medina*.
- El sexto recuadro establece las habilidades lingüísticas (*lee, habla y escribe*) para distintos idiomas.

La Figura 5.7 presenta la representación del vídeo "*What is Linked Data?*" en forma de tripletas RDF. La parte izquierda de esta figura, muestra una captura de pantalla de éste vídeo, mientras la parte derecha muestra las declaraciones de este vídeo en forma de tripletas RDF.

Las tripletas RDF de la Figura 5.7 establecen las siguientes características y relaciones para el vídeo *What is Linked Data?*.

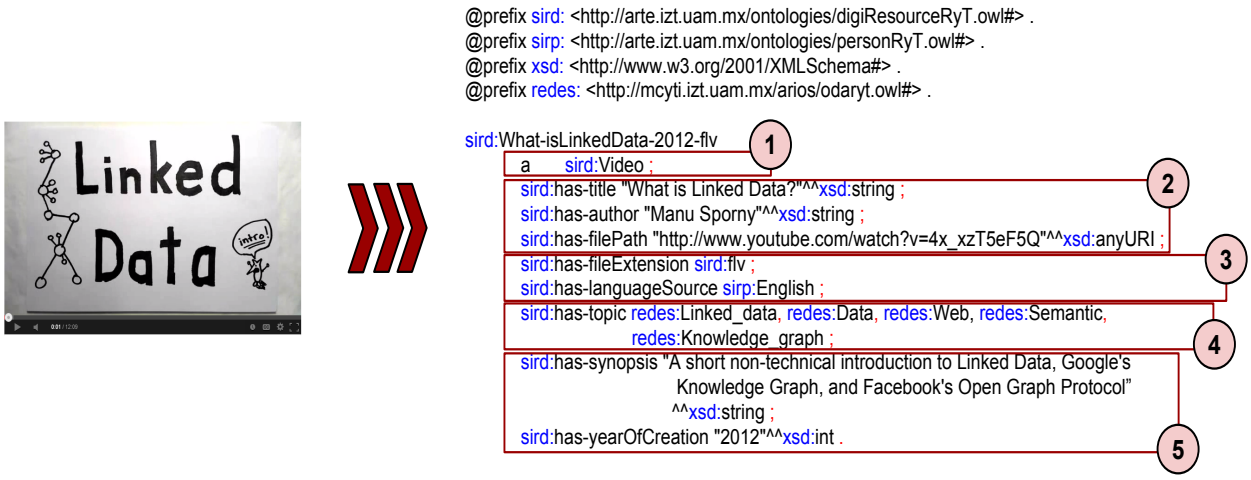

Figura 5.7: Declaraciones de vídeo “What is Linked Data?” en forma de tripletas RDF

- En el primer recuadro, la tripleta RDF indica que el recurso *What is Linked Data?* pertenece a la clase Video.
- En el segundo recuadro se indican los valores asociados a las características *título*, *autor* y *ruta del archivo*.
- En el tercer recuadro, la tripleta RDF establece la extensión o formato del recurso digital, así como el idioma fuente en que está escrito o que se habla en este recurso.
- En el cuarto recuadro, se establecen los temas de Redes y Telecomunicaciones que están en el contenido de este recurso digital.
- En el quinto recuadro, se establecen las tripletas para indicar la sinopsis y el año de creación de este vídeo.

Todas las *tripletas RDF* que están asociadas a las declaraciones de los recursos persona, conforman el *componente asertivo* de la ontología *cartografía de competencias*. De igual manera, todas las *tripletas RDF* asociadas a las declaraciones de los documentos y archivos multimedia, constituyen el *componente asertivo* de la ontología *búsqueda de recursos digitales*.

5.2. Enriquecimiento del conocimiento en el modelo

La etapa de representación del conocimiento, nos permite describir el conocimiento explícito en los *recursos de información*. Ahora bien, este conocimiento puede ser enriquecido mediante la introducción de axiomas o reglas de inferencia. Los axiomas permiten representar el conocimiento implícito sobre: los recursos y las relaciones de éstos. Por ejemplo, los profesores, empleados y estudiantes son personas, por ello, deben tenerse tres axiomas que

establezcan que un profesor es una persona, un empleado es una persona y un estudiante es una persona.

Para cada *caso de uso* debe encontrarse el respectivo conjunto de axiomas (TBox). Este proceso de búsqueda de axiomas debe guiarse a partir de los siguientes elementos: 1) *diagramas de clase*, 2) *cualidades en las relaciones* y 3) *operaciones de la teoría de grupos*. A continuación, se describen y argumentan todos los ***axiomas*** que se identificaron para las ontologías de *cartografía de competencias y recursos digitales*.

Protégé es el editor que se emplea, para construir, visualizar y manipular los axiomas (TBox) en nuestra ontologías. Esta herramienta permite a los usuarios manipular las clases, propiedades y axiomas desde una interfaz gráfica de usuario. También, esta herramienta permite escribir a los axiomas en distintas sintaxis de serialización (XML/RDF, Sintaxis Manchester, OWL/XML, Sintaxis Funcional o Turtle[2]).

5.2.1. Herencia de clases

El primer *conocimiento implícito* a representar, es la jerarquía de clases para cada *caso de uso*. El objetivo de ésto es construir modelos de organización jerárquicos del conocimiento para la cartografía de competencias y los recursos digitales. La búsqueda y construcción de las jerarquías se hace mediante el análisis de los *diagramas de clases* de las Figuras 5.4 y 5.3.

A continuación, se describen las dos *jerarquías de clases* para la cartografía de competencias y recursos digitales.

- En el área RyT, las personas pueden agruparse en cuatro clases básicas: ***Estudiante (Student)***, ***Profesor (Teacher)***, ***Investigador (Researcher)*** y ***Empleado (Employee)***. Las personas pueden pertenecer a más de una clase, por ejemplo, un profesor puede ser un investigador o un empleado, así como un estudiante puede ser un profesor o un empleado. Los profesores y empleados son profesionistas, por ello, la clase ***Profesionista (Professional)*** es super-clase de ***Profesor*** y ***Empleado***. Finalmente, cualquier individuo perteneciente a una de estas cinco clases, pertenece es una persona. De esta manera, las cinco clases tienen como super-clase la clase ***Persona (Person)***. La Figura 5.8 muestra la jerarquía de clases para los recursos persona.

En la Figura 5.9 se presenta un ejemplo de inferencia a partir del uso de axiomas de *jerarquía de clases*. En la parte izquierda de la Figura 5.9, se muestra una rama de la ontología *cartografía de competencias*. Esta rama tiene dos axiomas que establecen que un *Profesor es un Profesionista* y un *Profesionista es una Persona*. Esta rama también indica que el recurso Ricardo Marcelin Jiménez pertenece a la clase Profesor. Mientras, en la parte de la derecha de la Figura 5.9, se muestra la misma rama con inferencia, la cual indica que el recurso Ricardo Marcelin Jiménez pertenece las clases: Profesor, Profesionista y Persona.

[2]W3C, "OWL 2 Web Ontology Language Document Overview (Second Edition)," Disponible en: `http://www.w3.org/TR/owl2-overview/`

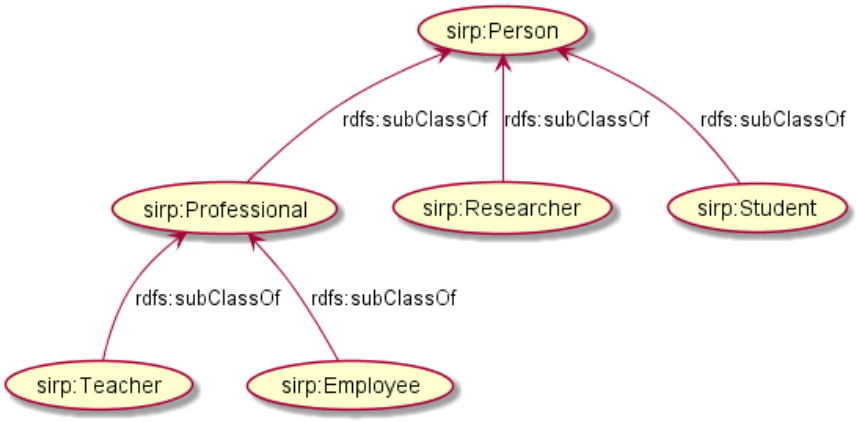

Figura 5.8: Jerarquía de clases para los recursos persona

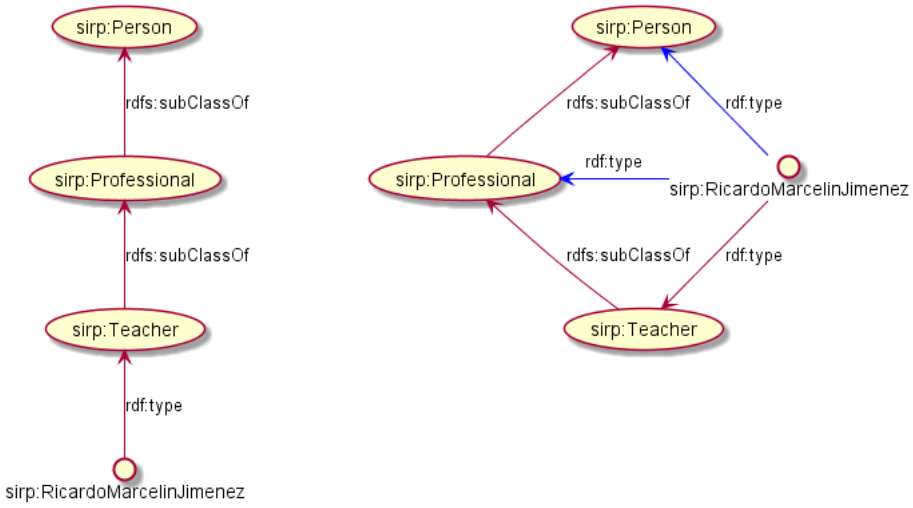

Figura 5.9: Ejemplo de inferencia para los axiomas de jerarquía de clases y el uso del recurso Ricardo Marcelin.

- En el área RyT, los principales recursos digitales se pueden agrupar en ocho clases básicas: ***Artículo (Paper)***, ***Libro (Book)***, ***Tesis (Thesis)***, ***Página Web (Webpage)***, ***Reporte Técnico (TechnicalReport)***, ***Audio (Audio)***, ***Vídeo (Video)***, ***Presentación (Presentation)*** e ***Imagen (Image)***. Estas nueve clases son disjuntos, por ello, no tienen individuos en común. Los recursos digitales se agrupan en dos clases generales: ***Documento (Document)*** y ***Mutimedia (Multimedia)***. Las primeras cinco clases Artículo, Libro, Tesis, Página Web, Reporte Técnico tienen como super-clase a la clase ***Documento***. Mientras las otras cuatro clases básicas tienen como super-clase a la clase ***Multimedia***. Finalmente, cualquier individuo de estas once once clases es un recurso digital, por ello, las once clases son subclases de la clase ***Recurso Digital (DigitalResource)***. La Figura 5.10 muestra la jerarquía de clases para los recursos

digitales.

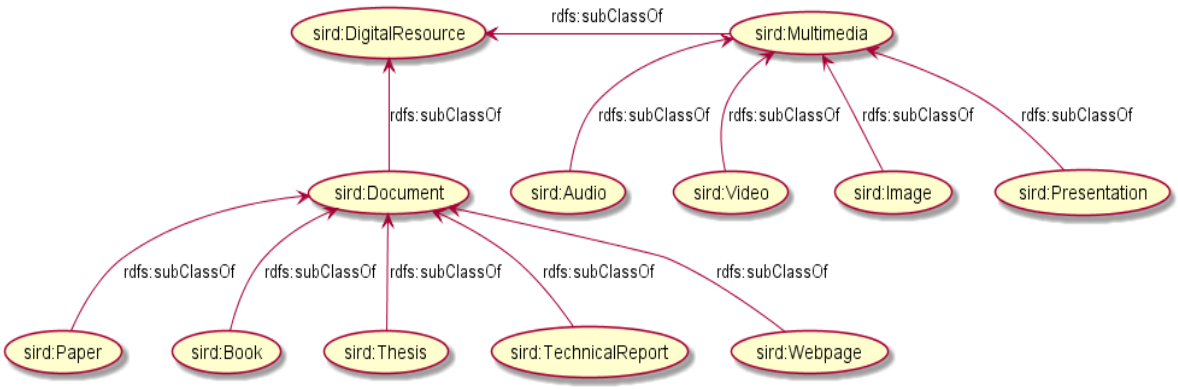

Figura 5.10: Jerarquía de clases para los recursos digitales

En la Figura 5.11 se presenta un ejemplo para la inferencia a partir de axiomas de *jerarquía de clases*. En esta figura, los axiomas son tomados de la ontología *búsqueda de recursos digitales*. La parte izquierda de esta Figura 5.11 ilustra una rama de la *ontología recursos digitales*. En esta rama, se exalta la asignación del recurso *What is Linked Data?* a la clase Video. Mientras, la parte derecha de la Figura 5.11, se presenta esta rama después de realizar inferencia. En esta rama, el recurso *What is Linked Data?* está asignado a estas clases: *Video, Multimedia* y *Recurso Digital*.

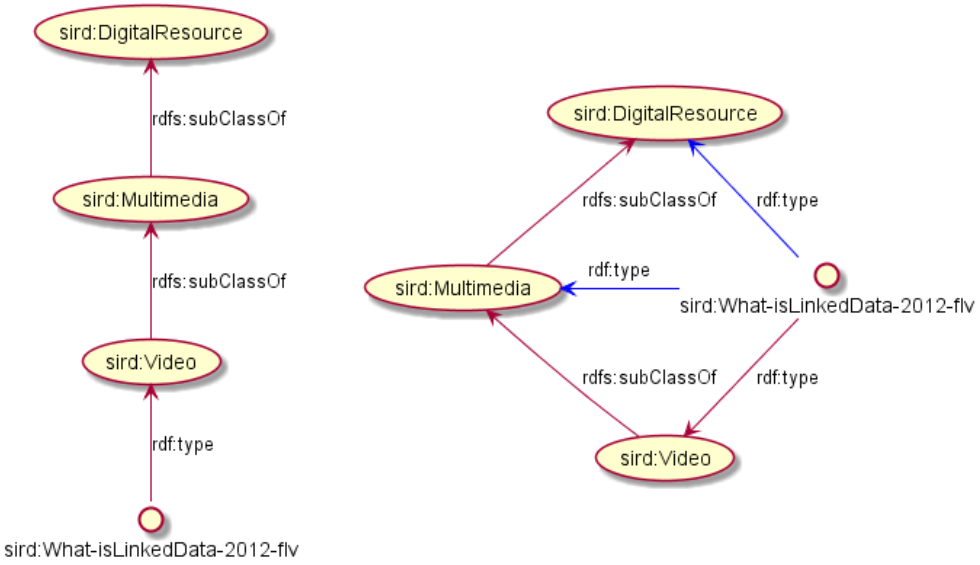

Figura 5.11: Ejemplo de inferencia para los axiomas de jerarquía de clases y el uso del recurso What is Linked Data?

5.2.2. Herencia de propiedades

Las propiedades al igual que las clases, pueden generalizarse mediante propiedades comunes. Esta jerarquización se hace mediante el análisis de los atributos en los caso de uso; averiguando que propiedades tienen un significado común.

La siguiente lista describe las *jerarquías de propiedades* que se identificaron para la *cartografía de competencias*. Para cada item de esta lista, se presenta un diagrama que muestra esta jerarquía en forma de un grafo.

- ***Lee***, ***habla*** y ***escribe*** son propiedades para indicar las habilidades lingüísticas de una persona. Estas propiedades pueden generalizarse a partir de la propiedad ***tiene lenguaje***, con el fin de indicar que una persona tiene algún conocimiento lingüísticos en un idioma. Por ello, las propiedades *lee, habla y escribe* son subpropiedades de la propiedad *tiene lenguaje*. En la Figura 5.12 se muestra la jerarquía de propiedades para describir las habilidades lingüísticas de las personas.

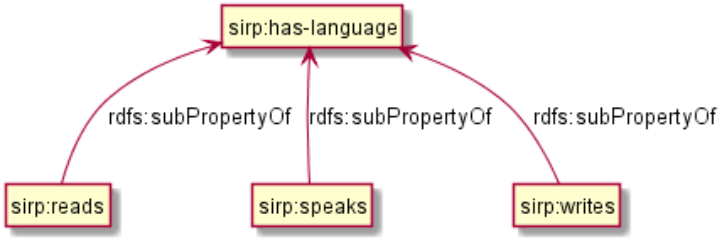

Figura 5.12: Jerarquía de propiedades para las habilidades lingüísticas.

- Las propiedades ***trabaja en*** y ***estudia en*** son relaciones que permiten vincular a una persona con el lugar donde labora (trabajo o estudio), por ello, estas dos propiedades son generalizadas a partir de la propiedad *tiene lugar de trabajo*. En la Figura 5.13 se muestra la jerarquía de propiedades para el lugar de trabajo.

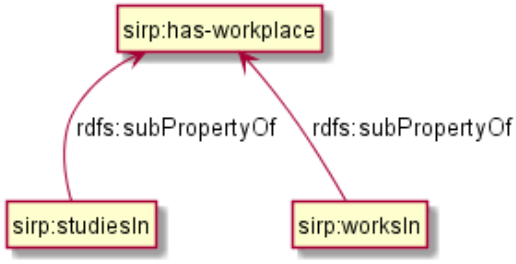

Figura 5.13: Jerarquía de propiedades para el lugar de trabajo.

- Las propiedades ***tiene asesor*** y ***tiene colega*** son relaciones que vinculan a dos personas. La propiedad *tiene asesor* vincula a un estudiante con un profesor. Mientras, la propiedad *tiene colega* vincula a dos profesionistas. Estas dos propiedades pueden generalizarse mediante la propiedad ***conoce a***, es decir, *tiene colega* y *tiene asesor* son subpropiedades de la propiedad *conoce a*. En la Figura 5.14 se muestra la jerarquía de propiedades que representan las relaciones profesionales entre las personas del área RyT.

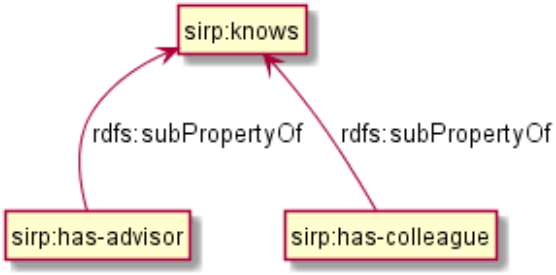

Figura 5.14: Jerarquía de propiedades para las relaciones profesionales entre personas.

De la misma manera que en la cartografía de competencias, el siguiente listado presenta las tres *jerarquías de propiedades* que se identificaron para el TBox de la *búsqueda de recursos digitales*. Al final de cada item, se presenta una imagen que muestra la respectiva jerarquía en forma de un grafo.

- La propiedad ***tiene resumen*** se emplea para describir el contenido básico de un documento. Mientras, la propiedad ***tiene sinopsis*** describe el contenido básico de un recurso multimedia. Estas dos propiedades pueden generalizarse mediante la propiedad ***tiene compendio***, la cual vincula a cualquier *recurso digital* con la descripción básica del contenido de éste. Por esta razón, *tiene resumen* y *tiene sinopsis* son subpropiedades de *tiene compendio*. Esta jerarquía de propiedades se presenta en la Figura 5.15.

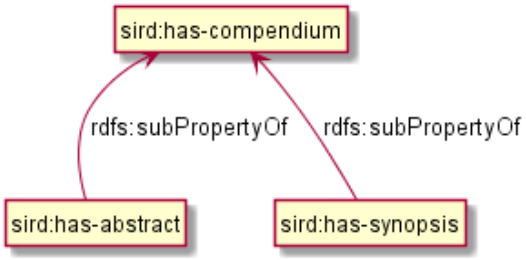

Figura 5.15: Jerarquía de propiedades para describir el contenido de un recurso digital.

- La propiedad ***tiene año de conclusión*** indica el año de conclusión de una tesis. La propiedad ***tiene año de última visita*** indica el año de la última visita para una página Web. La propiedad ***tiene año de publicación*** indica el año en que fue publicado un artículo científico o un libro. Mientras, las propiedades ***tiene año de creación*** y ***tiene año de elaboración*** se emplean para indicar el año de creación o elaboración de un recurso multimedia. Estas cinco propiedades pueden ser generalizadas a partir de la propiedad ***tiene año***. De esta manera, la propiedad *tiene año* es la superpropiedad de las otras cinco propiedades. La Figura 5.16 muestra gráficamente esta jerarquía de propiedades para el año de construcción de un recurso.

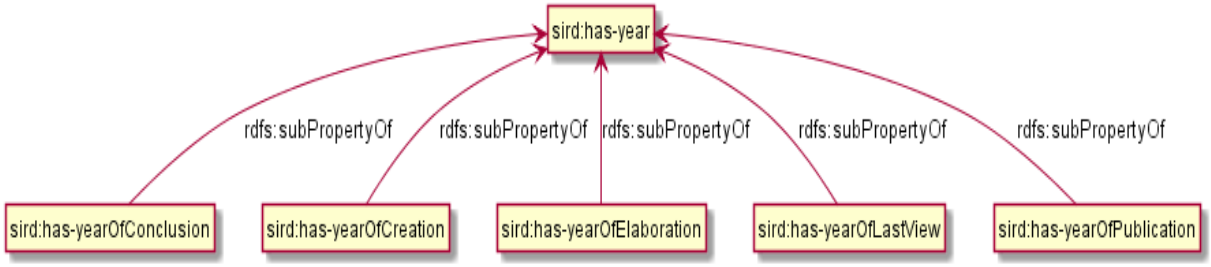

Figura 5.16: Jerarquía de propiedades para indicar el año de un recurso digital.

- La propiedad ***publicado en revista*** indica que un artículo se publica en una revista científica. La propiedad ***tiene editorial*** establece que un libro se publica por una determinada casa editorial. La propiedad ***tiene institución involucrada*** indica que una tesis pertenece a una determinada universidad o institución educativa. Estas tres propiedades se generalizan a partir de la propiedad ***publicado en***, por tal razón *publicado en revista*, *tiene editorial* y *tiene institución involucrada* son subpropiedades de la propiedad *publicado en*. Esta jerarquía de propiedades gráficamente se muestra en la Figura 5.17.

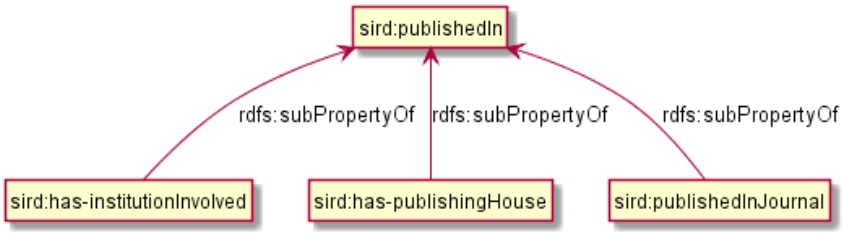

Figura 5.17: Jerarquía de propiedades para vincular a una organización con un recurso digital.

5.2.3. Domingo y rango de propiedades

Los axiomas de ***Dominio*** y ***Rango*** permiten definir qué recursos o recurso-literal pueden relacionarse en una propiedad. Esta idea es parecida a una función matemática, donde el dominio es el conjunto formado por *los valores que puede tomar una función*. Mientras, el rango es el conjunto de *los valores que resultan de evaluar la función*.

La Tabla 5.5 presenta el ***dominio*** y ***rango*** de las propiedades que pertenecen a la cartografía de competencias. En esta Tabla, la primer columna enuncia las propiedades de la cartografía, la segunda columna muestra las clases para el dominio y la tercera columna enuncia las clases o tipo de literales del rango. De la misma manera, la Tabla 5.6 presenta el ***dominio*** y ***rango*** para las propiedades en el TBox de la *búsqueda de recursos digitales*. La primera columna enuncia las propiedades, la segunda las clases del dominio y la tercera las literales o clases del rango.

La Figura 5.18, 5.19 y 5.20 ejemplifican el proceso de inferencia a partir del uso de axiomas de Dominio y Rango. Estos axiomas pertenecen a la cartografía de competencias y son utilizados para restringir las relaciones profesionales entre personas.

En la Figura 5.18 se presentan los axiomas de dominio y rango para las propiedades "*tiene asesor*" y "*tiene colega*". Por un lado, la clase *Profesionista* es tanto el ***dominio*** como el ***rango*** de la propiedad *tiene colega*. Por otro lado, las clases *Estudiante* y *Profesionista* son respectivamente el ***dominio*** y ***rango*** de la propiedad *tiene asesor*.

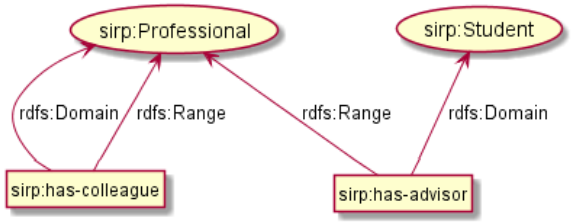

Figura 5.18: Axiomas de dominio y rango para modelar las relaciones profesionales entre personas del área de Redes y Telecomunicaciones.

En la Figura 5.19 se muestra un grafo RDF, en el cual las tripletas indican lo siguiente: *la Dra. Carolina Medina y el Dr. Héctor Pérez son asesores de Erik Alarcón*, *Laura Mendez tiene por asesores a la Dra. Carolina Medina y al Dr. Enrique Rodriguez*, *el Dr. Ricardo Marcelin y el Dr. Enrique Rodriguez son colegas de la Dra. Carolina Medina* y *la Dra. Carolina Medina es colega del Dr. Héctor Pérez*.

En la Figura 5.20, se presentan las tripletas de vinculación profesional entre individuos, así como las tripletas de asignación a una de estas dos clases: *Estudiante (sirp:Student)* y *Profesionista (sirp:Professional)*. Las tripletas de asignación son obtenidas después de realizar el proceso de inferencia. Los individuos *Ricardo Marcelin, Enrique Rodriguez, Carolina Medina y Héctor Pérez son Profesionistas*, porque el *dominio* y *rango* de la función *tiene colega (sirp:has-colleague)* es la clase *Profesionista*, también porque el *rango* de la función *tiene asesor (sirp:has-advisor)* es la clase *Profesionista*. Mientras, los individuos *Erik Alarcón*

Propiedad	Dominio	Rango
sirp:*has-name*	***sirp:****Person*	***xsd:****string*
sirp:*has-age*	***sirp:****Person*	***xsd:****integer*
sirp:*has-email*	***sirp:****Person*	***xsd:****anyURI*
sirp:*has-webSite*	***sirp:****Person*	***xsd:****anyURI*
sirp:*competentIn*	***sirp:****Person*	***xsd:****Competence*
sirp:*expertiseIn*	***sirp:****Person*	***redes:****TopicRyT*
sirp:*has-gender*	***sirp:****Person*	***sirp:****Gender*
sirp:*reads*	***sirp:****Person*	***sirp:****Language*
sirp:*speaks*	***sirp:****Person*	***sirp:****Language*
sirp:*writes*	***sirp:****Person*	***sirp:****Language*
sirp:*researchesOn*	***sirp:****Researcher*	***xsd:****string*
sirp:*has-advisor*	***sirp:****Student*	***sirp:****Professional*
sirp:*has-colleague*	***sirp:****Professional*	***sirp:****Professional*
sirp:*studiesIn*	***sirp:****Student*	***sirp:****University*
sirp:*worksIn*	***sirp:****Professional*	***sirp:****Organization*

Tabla 5.5: Dominio y Rango para las propiedades asociadas a la cartografía de competencias.

Propiedad	Dominio	Rango
sird:*has-title*	***sird:****DigitalResource*	***xsd:****string*
sird:*has-author*	***sird:****DigitalResource*	***xsd:****string*
sird:*has-filePath*	***sird:****DigitalResource*	***xsd:****anyURI*
sird:*has-fileExtension*	***sird:****DigitalResource*	***sird:****Extension*
sird:*has-languageSource*	***sird:****DigitalResource*	***sird:****Language*
sird:*has-topic*	***sird:****DigitalResource*	***redes:****TopicRyT*
sird:*has-abstract*	***sird:****Document*	***xsd:****string*
sird:*has-numberOfPages*	***sird:****Document*	***xsd:****integer*
sird:*has-description*	***sird:****Multimedia*	***xsd:****string*
sird:*has-edition*	***sird:****Book*	***xsd:****integer*
sird:*has-publishingHouse*	***sird:****Book*	***sirp:****PublishingHouse*
sird:*publishedInJournal*	***sird:****Paper*	***sirp:****Journal*
sird:*has-institutionInvolved*	***sird:****Thesis*	***sirp:****University*
sird:*isTypeOfArticle*	***sird:****Paper*	***sirp:****TypeArticle*
sird:*has-yearOfConclusion*	***sird:****Thesis*	***xsd:****integer*
sird:*has-yearOfElaboration*		***xsd:****integer*
sird:*has-yearOfLastView*	***sird:****Webpage*	***xsd:****integer*
sird:*has-yearOfPublication*	***sird:****Book*	***xsd:****integer*
sird:*has-yearOfPublication*	***sird:****Paper*	***xsd:****integer*

Tabla 5.6: Dominio y Rango para las propiedades asociadas a la búsqueda de recursos digitales.

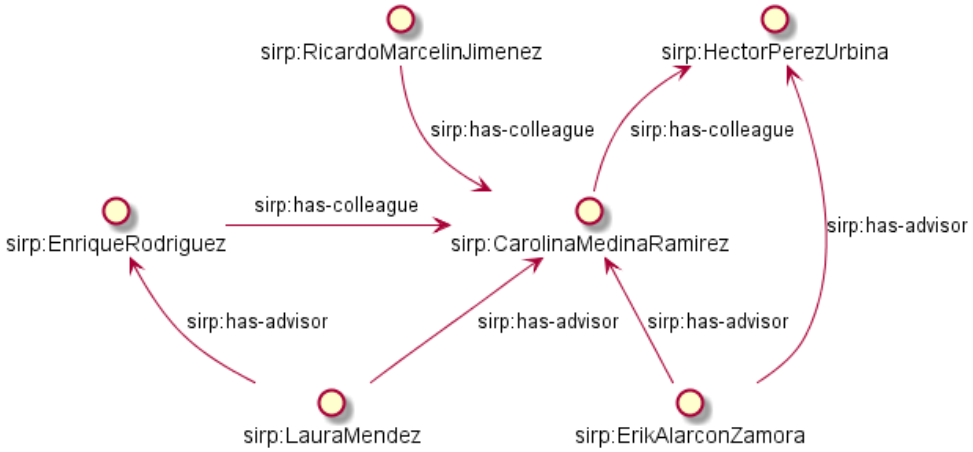

Figura 5.19: Tripletas RDF que describen las relaciones profesionales entre personas del área de Redes y Telecomunicaciones.

y Laura Mendez son Estudiantes, porque el *dominio* de la propiedad *tiene asesor* es la clase *Estudiante*.

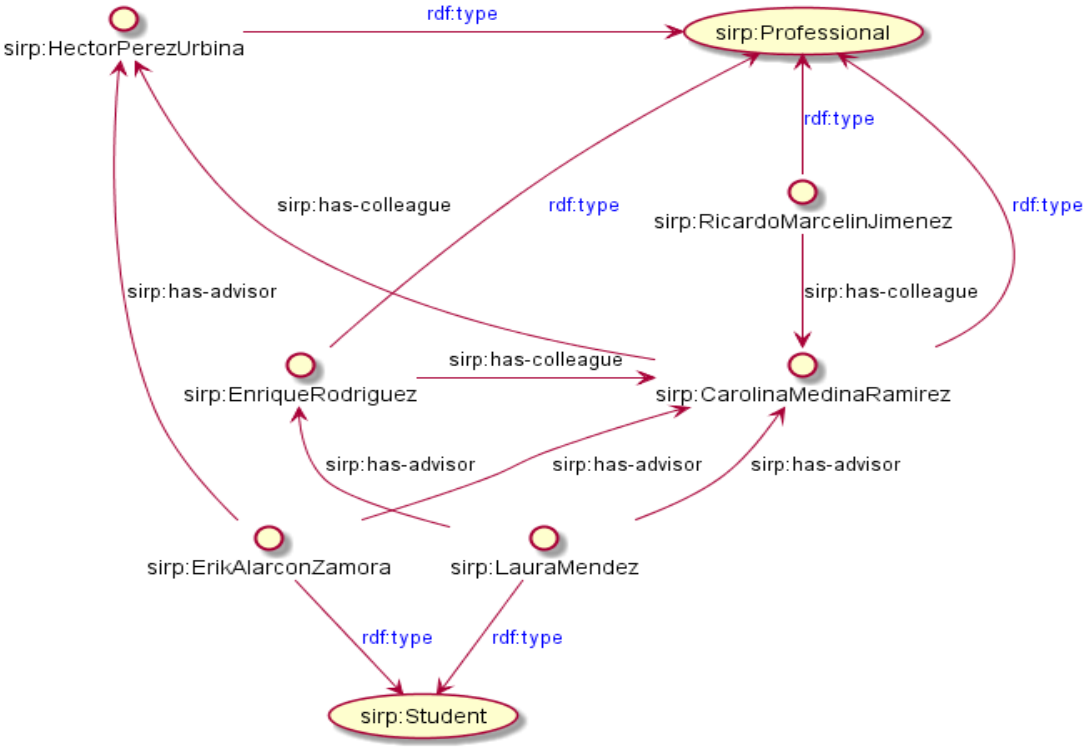

Figura 5.20: Ejemplo de inferencia para los axiomas de Dominio y Rango que pertenecen a la cartografía de competencias.

5.2.4. Simetría en las propiedades

Una tripleta RDF es una relación unidireccional, es decir, parte de un *individuo p* y termina en *individuo q*. En la Figura5.21 se ejemplifica este comportamiento unidireccional, donde la relación *p tiene género q*, indica que todos los *individuos p* se relacionan con un *individuo q*.

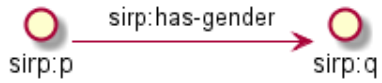

Figura 5.21: Ejemplo del comportamiento unidireccional de una propiedad.

Sin embargo, existen algunas propiedades que requieren tener un comportamiento bidireccional, con la finalidad de tener una mejor representación del dominio. En la Figura 5.22 se presenta un ejemplo genérico del uso de relaciones simétricas ("relación bidireccional"). En este ejemplo, los individuos p y q son el nodo origen y el nodo destino, es decir, en una tripleta RDF pueden ser el sujeto o el objeto.

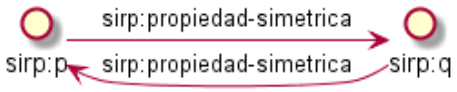

Figura 5.22: Ejemplo de simetría en una propiedad genérica.

A continuación, se describen las propiedades simétricas que se identificaron para la *cartografía de competencias*. Con respecto al caso de la búsqueda de recursos digitales, no se identificó alguna propiedad que tenga la característica simétrica.

- La propiedad ***tiene colega (sir:has-colleague)*** es una relación del tipo laboral que vincula a dos Profesionistas (Profesor o Empleado). La idea básica de esta relación es indicar quiénes son los colegas de trabajo o con quiénes ha hecho colaboración un *Profesionista*.

  La Figura 5.23 muestra un subgrafo RDF con relaciones *tiene colega* entre cuatro profesores del área de Redes y Telecomunicaciones (RyT). En este subgrafo las tripletas indican las siguientes declaraciones: *el Dr. Ricardo Marcelin tiene por colegas a la Dra. Carolina Medina, Dr. Enrique Rodrígez y Dr. Miguel López*, *el Dr. Enrique Rodrígez tiene por colegas a la Dra. Carolina Medina y al Dr. Miguel López* y *la Dr. Carolina Medina tiene por colega al Dr. Miguel López*.

  En la Figura 5.23, el recurso "*sirp:MiguelLopez*" es el ***objeto*** en todas las tripletas del subgrafo RDF, es decir, el Dr. Miguel López no tiene alguna declaración propia que indique quienes son sus colegas. El comportamiento de la propiedad *tiene colega*

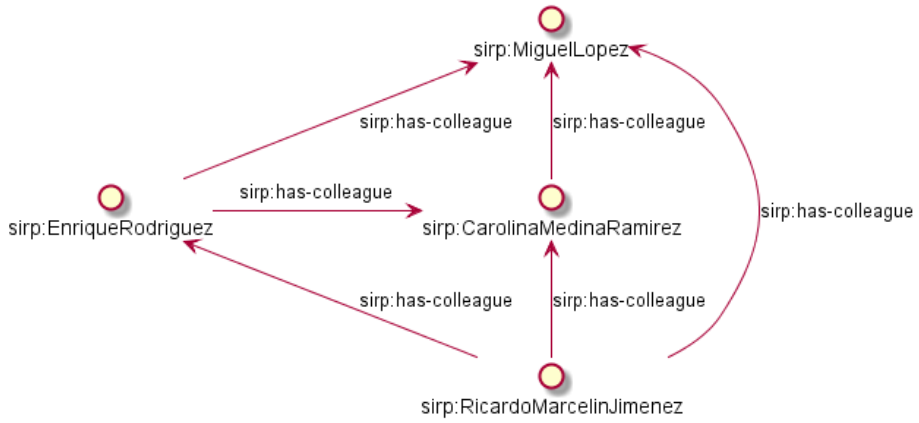

Figura 5.23: Subgrafo RDF con tripletas que indican las relaciones profesionales entre profesores del área.

indica una relación unidireccional. Sin embargo, esta propiedad debe tener un comportamiento bidireccional. Por ello, la propiedad *tiene colega (sirp:has-colleague)* tiene la *característica simétrica*.

La Figura 5.24 muestra el subgrafo RDF inferido. Este subgrafo se obtiene a partir de las tripletas en la Figura 5.23 y del axioma que establece que la propiedad *tiene colega* es simétrica. En esta Figura, el recurso "*sirp:MiguelLopez*" tiene el rol de *objeto*, así como el rol de ***sujeto*** para otras tres tripletas RDF. Ésto significa que el *Dr. Miguel López tiene por colegas a la Dra. Carolina, Dr. Enrique y al Dr. Ricardo*. De la misma manera, los otros recursos (Doctores) carentes de su propia afirmación, ahora tienen esta relación.

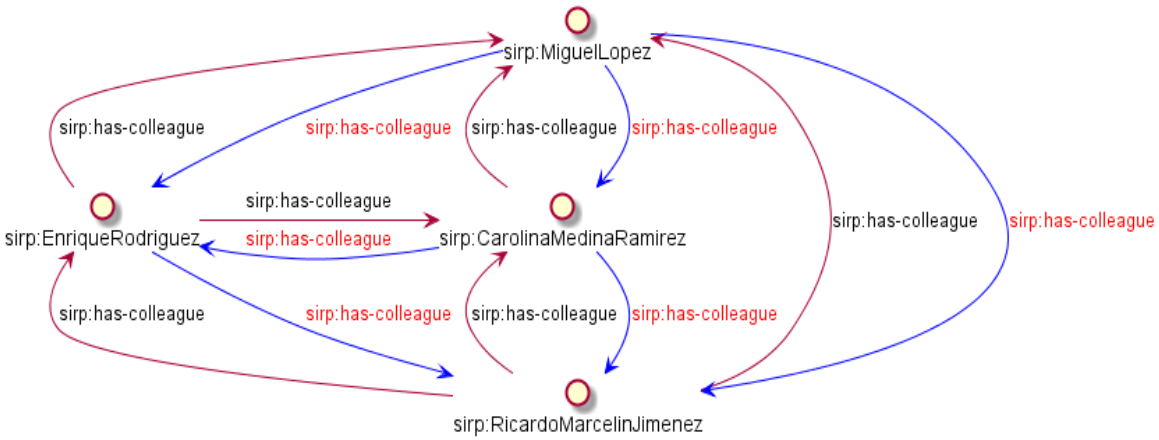

Figura 5.24: Ejemplo de inferencia a partir de la propiedad tiene-colega como propiedad simétrica.

- La propiedad ***conoce a (sirp:knows)*** es una relación profesional para la interacción entre las personas (profesor, empleado o estudiante) del área de Redes y Telecomunicaciones. Esta propiedad, al igual que la propiedad *tiene colega*, tiene un comportamiento bidireccional. Por ello, la propiedad *conoce a* es una propiedad simétrica.

  La Figura 5.25 ejemplifica un subgrafo RDF, en el cual las relaciones *conoce a* tienen solo una dirección. En esta Figura, las tripletas indican lo siguiente: *Erik Alarcón conoce a la Dra. Carolina Medina, Dr. Héctor Pérez y Laura Méndez*, *Laura Méndez conoce a la Dra. Carolina Medina y Dr. Enrique Rodríguez*, la *Dra. Carolina conoce al Dr. Héctor Pérez* y el *Dr. Enrique Rodríguez conoce a la Dra. Carolina Medina.*

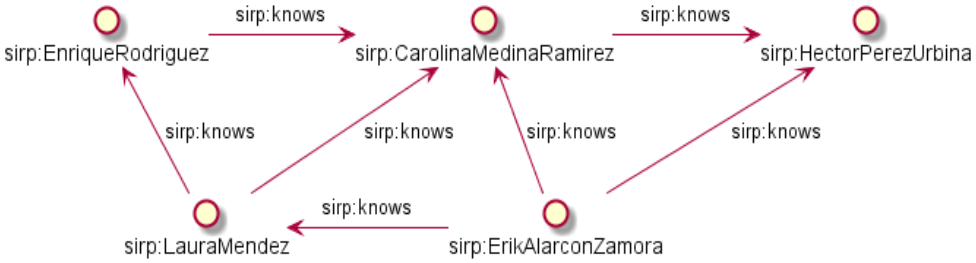

Figura 5.25: Subgrafo RDF de las relaciones conoce-a entre personas del área de Redes y Telecomunicaciones.

La Figura 5.26 presenta el subgrafo de la Figura 5.25 después de realizar el proceso de inferencia, donde la propiedad *conoce a* tiene la característica de ser simétrica. En esta Figura 5.26, todos los recursos (personas) son sujeto y objeto en las tripletas RDF. Por ejemplo, el Dr. Héctor Pérez en la Figura 5.25 no tiene sus propias declaraciones que indican a quienes conoce. Mientras, el Dr. Héctor Pérez en la Figura 5.26 sí tiene sus propias declaraciones que indican que conoce a la Dra. Carolina Medina y Erik Alarcón.

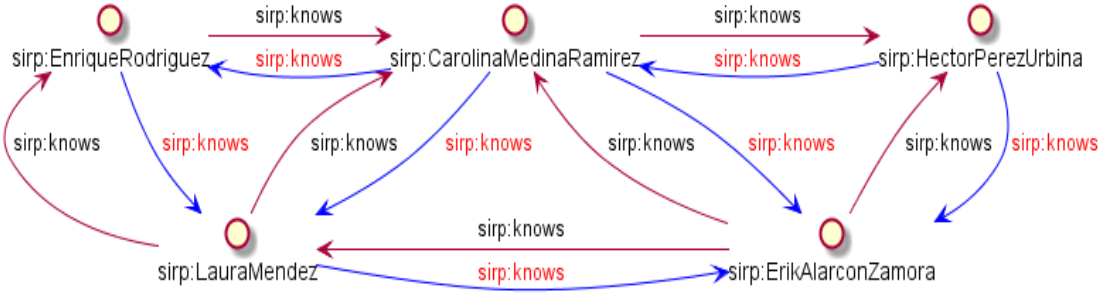

Figura 5.26: Ejemplo de inferencia a partir de la propiedad conoce-a como propiedad simétrica.

5.3. Búsqueda y recuperación de información en el modelo

Las etapas *representación y enriquecimiento del conocimiento* permiten construir nuestras bases de conocimiento, es decir, las ontologías para la cartografía de competencias, búsqueda de recursos digitales, así como el vocabulario de Redes y Telecomunicaciones. Estas ontologías son contenedores de conocimiento con un significado bien definido.

En esta tesis, las ontologías tienen como finalidad la búsqueda y recuperación de la información para responder las preguntas o necesidades informativas de los usuarios del área de Redes y Telecomunicaciones (RyT). Esta búsqueda y recuperación se hace a partir de tres puntos clave: 1) *identificar las preguntas en lenguaje natural*, 2) *transformar las preguntas a una consultas SPARQL* y 3) *ejecutar las consultas mediante un motor de búsqueda SPARQL*.

La ***identificación de las preguntas en lenguaje natural*** es el proceso de análisis y hallazgo de las principales preguntas a partir de los casos de uso.

Nuestros *casos de uso* son: 1) la *cartografía de competencias* que es la búsqueda de personas a partir de sus habilidades profesionales y conocimientos en RyT, así como 2) ***la búsqueda de recursos digitales*** que es el hallazgo de documentos y archivos multimedia a partir de los metadatos de éstos y los temas que están vinculados con el área de RyT.

Los parámetros de búsqueda en la cartografía competencias son:

- Habilidades profesionales como trabajo en equipo, creatividad, liderazgo, administración de proyectos, auto-aprendizaje, toma de decisiones, por mencionar algunas.
- Habilidades lingüísticas como hablar en español, inglés y francés; escribir en español, inglés y japones; leer en español y francés, entre otros idiomas.
- Conocimientos de Redes y Telecomunicaciones como sabiduría en sistemas operativos, capa enlace, filtros, ontologías, hilos de procesamiento, radios cognitivos, datos ligados, por mencionar algunos.
- Relaciones profesionales como conoce a la Dra. Carolina, Dr. Héctor, Dr. Ricardo, Dr. Miguel, Erik, Laura, entre otros.
- Clase de persona en RyT como profesionista, profesor, estudiante o empleado.
- Intervalo de edad como como 18 a 40, <40, >18, 18 a 18, por mencionar.
- Lugar de trabajo como la Universidad Autónoma Metropolitana, Clark & Parsia LLC, Infotec, Canonical Ltd, entre otras.
- Género, masculino, femenino, hombre y mujer.

A partir de estos parámetros de búsqueda, se hallaron un conjunto de preguntas básicas. El siguiente listado presenta las diecinueve preguntas básicas para la *cartografía de competencias*, donde cada pregunta tiene un respectivo *identificador de consulta*.

- ***Q1.1***.- ¿Cuáles son los nombres, correos, sitios web, géneros y edades de las personas del área de RyT?
- ***Q1.2***.- ¿Cuáles son los nombres, sitios web y los lugares donde laboran las personas del RyT?
- ***Q1.3***.- ¿Quiénes son mayores de 20 años y menores de 45 años?
- ***Q1.4***.- ¿Cuáles son los nombres y sitios web de los profesionistas del área de RyT?
- ***Q1.5***.- ¿Quiénes trabajan en la Clark & Parsia y son del sexo Masculino?
- ***Q1.6***.- ¿Quiénes son estudiantes y leen en inglés?
- ***Q1.7***.- ¿Quienes hablan, leen y escriben en inglés?
- ***Q1.8***.- ¿Qué estudiantes saben algo de inglés?
- ***Q1.9***.- ¿Qué profesores tienen la capacidad de síntesis?
- ***Q1.10***.- ¿Qué profesionistas tienen conocimiento en los temas de Web Semántica?
- ***Q1.11***.- ¿Qué profesores tienen conocimientos en Sistemas Distribuidos?
- ***Q1.12***.- ¿Quiénes tienen conocimiento en Java, OWL, RDF, Threads, C, OpenMP?
- ***Q1.13***.- ¿Qué estudiantes tienen algún conocimiento en los subtemas de Sistemas Operativos?
- ***Q1.14***.- ¿Quiénes trabajan en una Universidad?
- ***Q1.15***.- ¿Quienes laboran en la UAM y tienen algún conocimiento en Web Semántica?
- ***Q1.16***.- ¿Qué personas tienen como asesor a Carolina Medina?
- ***Q1.17***.- ¿Quiénes son los colegas de Ricardo Marcelin?
- ***Q1.18***.- ¿Cuáles son los nombres y correos de las personas que conocen a Carolina Medina Ramírez?
- ***Q1.19***.- ¿Qué personas son profesores-investigadores?

De la misma manera que la cartografía, los principales criterios de búsqueda para los recursos digitales son:

- Nombre del autor como Carolina Medina, Héctor Pérez, Ricardo Baeza, Tim Berners Lee, por mencionar algunos.
- Lenguaje como inglés, español, francés, chino, ruso, entre otros.

- Extensión del archivo como pdf, doc, odp, txt, html, xml, ppt, flv, mpg, mp3, wav, jpg, entre otras.
- Institución involucrada como editorial (Mc Graw Hill, Grijalbo, Pearson), universidad (UNAM, UAM, IPN), revista científica (IEEE, ACM, Springer).
- Año de cota inferior, por ejemplo >1980, >2000, >2012.
- Clase de recurso digital en RyT como libro, artículo, reporte técnico, tesis, página Web, audio, vídeo, imagen, presentación, documento y multimedia.
- Temas vinculados al área de RyT como sistemas operativos, capa enlace, filtros, ontologías, hilos de procesamiento, radios cognitivos, datos ligados, por mencionar algunos.

Estos parámetros de búsqueda dieron pauta a la búsqueda de las principales preguntas para este *caso de uso*. A continuación, se listan las diez preguntas principales que se hallaron para la búsqueda de recursos digitales.

- ***Q2.1***.- ¿Cuáles son los títulos, rutas, extensión, idioma de todos los recursos digitales de RyT?
- ***Q2.2***.- ¿Cuáles libros tratan sobre algunos temas de Sistemas Distribuidos?
- ***Q2.3***.- ¿Qué recursos fueron publicados por la UAM?
- ***Q2.4***.- ¿Qué documentos sirven para dar un curso de Sistemas P2P?
- ***Q2.5***.- ¿Qué recursos multimedia son mayores al año 2009?
- ***Q2.6***.- ¿Cuáles documentos tratan sobre Ontologías?
- ***Q2.7***.- ¿Qué recursos fueron publicados en una Revista científica?
- ***Q2.8***.- ¿Qué recursos tienen en su descripción las palabras "linked data¿
- ***Q2.9***.- ¿Cuáles documentos en Inglés y mayores al año 2000 son de autoría de Erik Alarcon Zamora?
- ***Q2.10***.- ¿Cuál son las tesis de Samuel Hernandez Maza?

Las preguntas para los dos casos de uso fueron electas, porque algunas utilizan el conocimiento explícito y otras aprovechan el conocimiento implícito.

Con base en las preguntas de ambos listados, el siguiente paso es ***transformar estas preguntas a consultas SPARQL***. Esta transformación para cada *pregunta-consulta* se hace de esta manera.

1. Analizar e identificar estos componentes: a) los valores que se desean recuperar, a los cuales se les asocia una *variable resultado*, b) las *propiedades* y c) los criterios de búsqueda que son *variables auxiliares* o *valores específicos* (recursos o literales).

2. Escribir las *variables resultado* en la cláusula SELECT.

3. Construir y escribir los *patrones tripletas* a partir de las variable resultado, propiedades, variables auxiliares y valores específicos.

4. Introducir los *patrones tripletas* en la cláusula WHERE.

A continuación, se presentan y describen algunos ejemplos para transformar las preguntas en *lenguaje natural* a sus respectivas consultas SPARQL. Estas preguntas son electas del listado de preguntas para la cartografía de competencias y la búsqueda de recursos digitales.

La Figura 5.27 presenta la consulta SPARQL que está asociada a la pregunta (Q1.1) *¿Cuáles son los nombres, correos, sitios web, géneros y edades de las personas del área de RyT?*. En esta pregunta, los valores de recuperación son: *nombre*, *correo*, *sitio Web*, *género* y *edad* para cualquier recurso persona. Cada valor tiene asociado una respectiva variable: *?name*, *?mail*, *?ws*, *?gender* y *?age*. Las propiedades asociadas a estos valores son: *tiene nombre (sirp:has-name)*, *tiene email (sirp:has-email)*, *tiene sitio Web (sirp:has-webSite)* y *tiene género (sirp:has-gender)*. Finalmente, la variable auxiliar "*?x*" establece que sin importar el identificador URI del recurso, deben recuperarse los valores de las variables respuesta.

¿Cuáles son los nombres, correos, sitios web, géneros y edades de las personas del área de RyT?

```
PREFIX sirp: <http://arte.izt.uam.mx/ontologies/personRyT.owl#>

SELECT  ?name ?mail ?ws ?gender ?age
    WHERE
    {
            ?x  sirp:has-name ?name;
                sirp:has-email ?mail;
                sirp:has-webSite ?ws;
                sirp:has-gender ?gender;
                sirp:has-age ?age.
    }
```

Figura 5.27: Consulta SPARQL asociada a la pregunta Q1.1 de la cartografía de competencias.

La consulta asociada a la pregunta Q1.1, es un ejemplo sencillo de construcción de consultas SPARQL. Ahora bien, existen preguntas que al transformarse a consultas SPARQL, emplean varios *patrones tripleta*. Pero, estas consultas pueden reducirse a una menor cantidad de *patrones tripleta*, si se da por hecho que hay razonamiento en la ontología. Un ejemplo de este tipo de consultas se da para la pregunta (Q1.18) *¿Cuáles son los nombres y sitios Web de las personas que conocen a Carolina Medina Ramírez?*. La Figura 5.28 muestra la consulta asociada a la pregunta Q1.18, en la cual no se da por hecho el uso de razonamiento. Mientras, la Figura 5.29 presenta la consulta para la misma pregunta, donde se da por hecho el uso del razonamiento.

A continuación, se listan los elementos de la consulta que están asociados a la pregunta Q1.18 de la Figura 5.28.

- Variable resultado: *?name* - nombre y *?ws* - sitio Web.
- Propiedades: *tiene nombre (sirp:has-name)*, *tiene colega (sirp:has-colleague) tiene asesor(sirp:has-advisor)*, *conoce a (sirp:knows)* y *tiene sitio Web (sirp:has-webSite)*.
- Variables auxiliares: *?x* - identificador del recurso.
- Valores específicos: *sirp:CarolinaMedinaRamirez* - identificador de la Dra. Carolina Medina Ramírez.

En esta consulta SPARQL, la parte importante es el conjunto de *patrones tripleta* que indican las relaciones profesionales de cualquier individuo con la Dra. Carolina. Porque, para indicar *a quiénes conoce* o *quiénes conocen a* la Dra. Carolina se deben emplear cinco *patrones tripleta*, los cuales aparecen en el recuadro con el símbolo del asterisco.

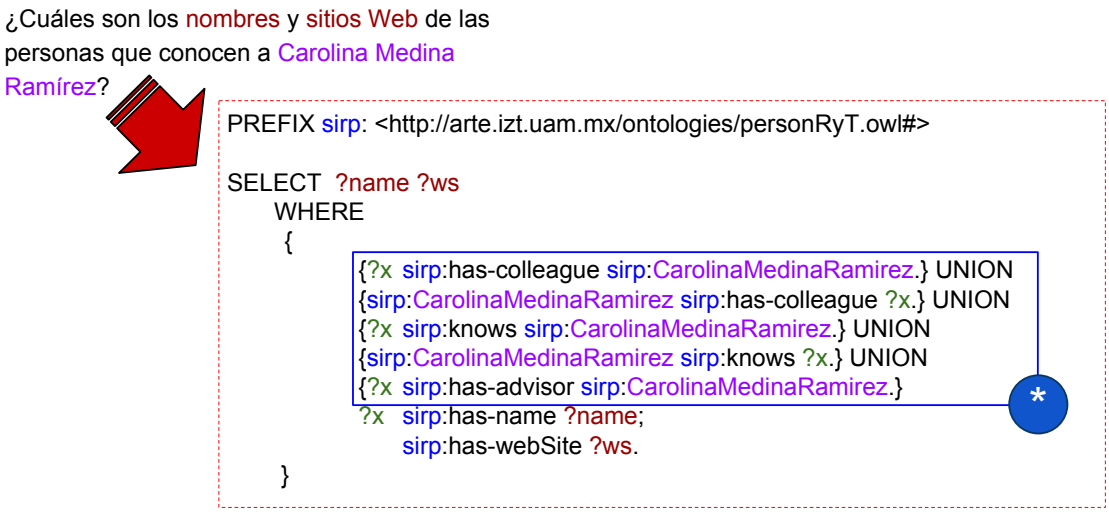

Figura 5.28: Consulta SPARQL asociada a la pregunta Q1.18, en la cual no se da por hecho el uso del razonamiento.

La consulta de la pregunta Q1.18, puede simplificarse mediante la asunción de emplear razonamiento en la ontología. La Figura 5.29 presenta esta consulta SPARQL simplificada. En la cual, el número de patrones tripletas del recuadro con asterisco de la Figura 5.28 se reduce a un solo patrón marcado con el símbolo más. Porque, se da por hecho que las propiedades *tiene colega (has-colleague)* y *tiene asesor* son subpropiedades de la propiedad *conoce a(knows)*, además esta superpropiedad tiene la característica simétrica.

En el caso de uso *búsqueda de recursos digitales*, existen consultas que mediante la inferencia en una ontologia, pueden reducirse el número de patrones en sus respectivas clausulas WHERE. A continuación, se presenta un ejemplo de este tipo de consultas. Este ejemplo esta

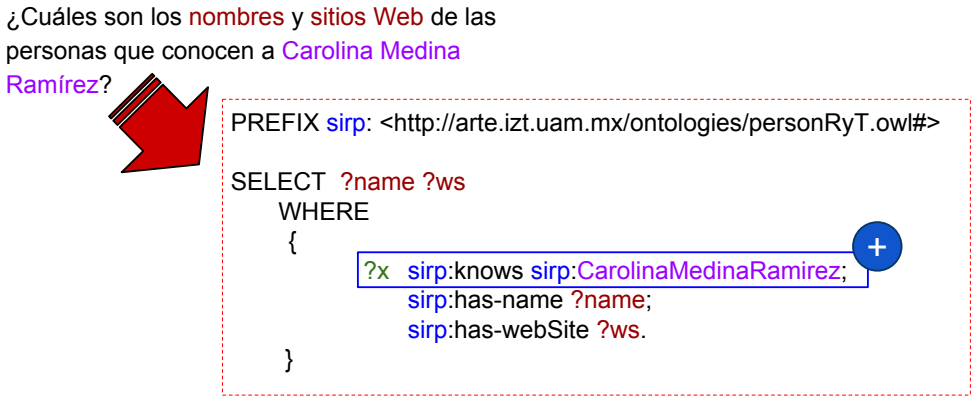

Figura 5.29: Simplificación de la Consulta SPARQL asociada a la pregunta Q1.18 mediante la asunción de emplear razonamiento.

dividido en dos partes: 1) *la consulta se construye sin dar por hecho el uso de un razonador* y 2) la consulta se construye a partir de la aseveración de emplear razonamiento en una ontología.

La Figura 5.30 presenta la consulta asociada a la pregunta (Q2.6) *¿Cuáles documentos tratan sobre Ontologías?*. Esta consulta se construye a partir de las tripletas explícitas en la ontología de *recursos digitales*.

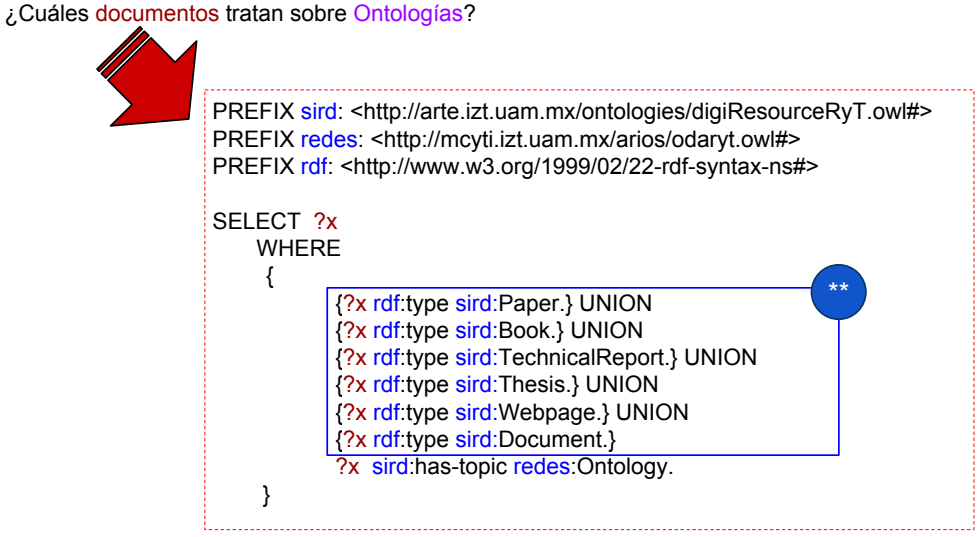

Figura 5.30: Consulta SPARQL asociada a la pregunta Q2.6, en la cual no se da por hecho el uso del razonamiento.

Los elementos importantes en la consulta (Q2.6) de la Figura 5.30, se listan a continuación:

- Variable resultado: *?x* - identificador del recurso.

- Propiedades: *tipo (rdf:type)* y *tiene tópico (sird:has-topic)*.
- Variables auxiliares: *?x* - identificador del recurso.
- Valores específicos: *sird:Book* - identificador de la clase Libro, *sird:Paper* - identificador de la clase Artículo, *sird:Thesis* - identificador de la clase Tesis, *Webpage* - identificador de la clase Página Web, *sird: TechnicalReport* - identificador de la clase Reporte Técnico, *sird:Document* - identificador de la clase Documento y *redes:Ontology* identificador del tema Ontología.

En esta consulta SPARQL, se resaltan los patrones del recuadro marcado con dos asteriscos. Estos seis patrones indican los recursos que son documentos. Porque, si solamente se emplea un patrón (*?x rdf:type sird:Document*), entonces se omiten artículos, libros, reportes técnicos, tesis y páginas Web que no tienen la asignación a la clase *Documento (sird:Document)*.

La consulta SPARQL de la Figura 5.30 puede simplificarse mediante la asunción de emplear el proceso de inferencia en la ontología de recursos digitales. La Figura 5.31 presenta la consulta simplificada para la pregunta Q2.6. En esta consulta los seis patrones son reducidos a un patrón que está marcado con el doble signo de más. Porque, se da por hecho que los *individuos* de las clases *Artículo (sird:Paper)*, *Libro (sird:Book)*, *Reporte Técnico (sird:TechnicalReport)*, *Tesis (sird:Thesis)* y Página Web (sird:Webpage) son individuos de la clase *Documento (sird:Document)*.

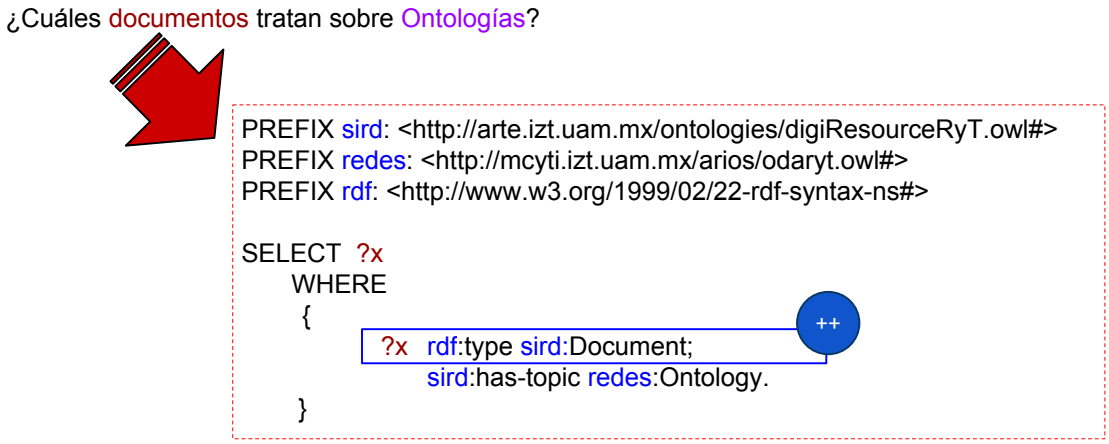

Figura 5.31: Simplificación de la Consulta SPARQL asociada a la pregunta Q2.6 mediante la asunción de emplear razonamiento.

Estos son cinco ejemplos de transformación de preguntas a consultas SPARQL. Algunas consultas se construyen mediante tripletas del ABox, porque la información a recuperar pertenece al conocimiento explícito; por ejemplo, la consulta de la Figura 5.27. Otras consultas se construyen mediante la agrupación de varios patrones, porque éstos permiten recuperar mejor información en el grafo RDF; por ejemplo las consultas de las Figuras 5.28 y 5.30.

Finalmente, existen consultas que aprovechan el uso de inferencia, para reducir el número de patrones y obtener mejores resultados; por ejemplo, las consultas SPARQL de las Figuras 5.29 y 5.31.

En la recuperación de la información, el *tercer punto clave* es la ***ejecución de las consultas***. Esta ejecución se basa en el uso de un *motor de búsqueda* SPARQL, para buscar y recuperar la información en los grafos RDF de nuestras ontologías. El funcionamiento general de un motor SPARQL es el siguiente. Un motor de consulta SPARQL a partir de una consulta SPARQL y un grafo RDF, interpreta esta consulta SPARQL. Posteriormente, este motor compara los patrones de la cláusula WHERE con todos los triples en un *grafo RDF* (modelo). Aquellas tripletas que concuerdan con los patrones, el motor recupera la información de las *variables resultado*. Finalmente, el motor regresa la información de estas variables al usuario. Esta información comúnmente se presenta forma de tabla. En la Figura 5.32 se presenta este funcionamiento de un motor SPARQL.

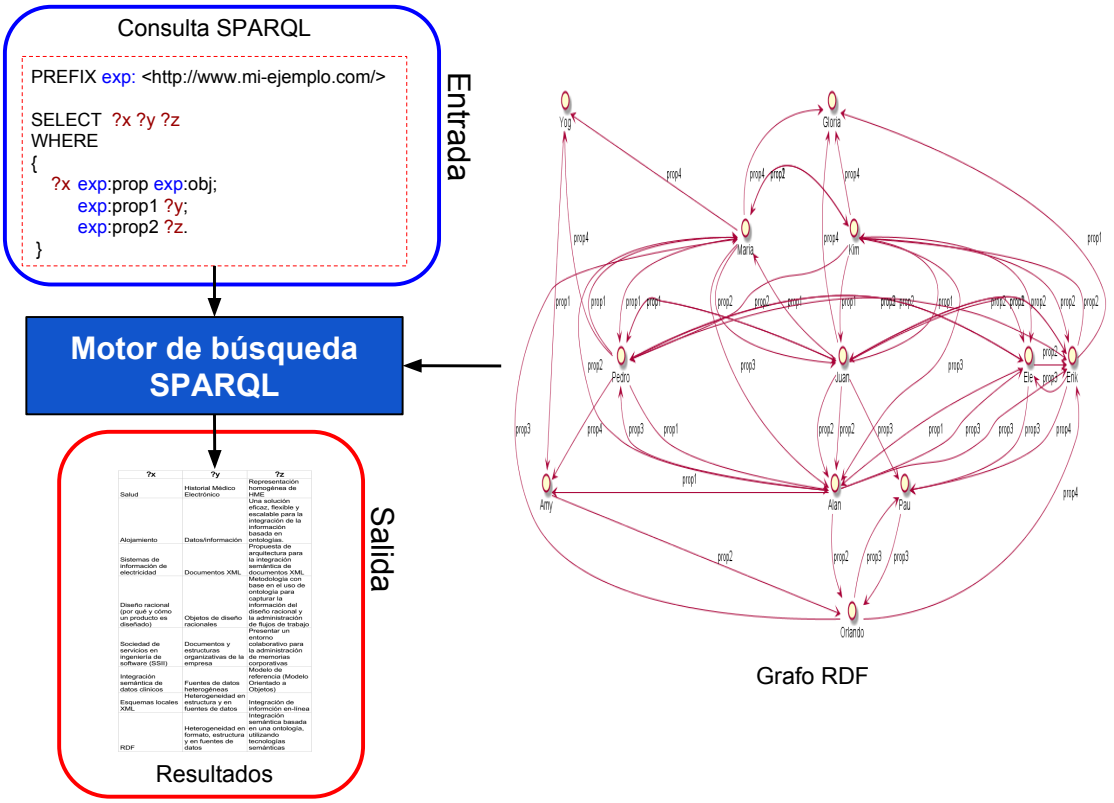

Figura 5.32: Proceso básico de consulta de información para un motor de búsqueda SPARQL.

Un motor SPARQL puede encontrarse en un triplestore. El triplestore Jena[3] posee el motor de consulta (ARQ), el cual soporta el lenguaje SPARQL. Este motor ARQ permite recuperar los resultados de una consulta para mostrarlos en pantalla en forma de una tabla u obtenerlos de forma iterativa mediante JAVA.

[3]The Apache Software Foundation, "Apache Jena," Disponible en: `http://jena.apache.org/`

Los resultados proporcionados por un motor SPARQL, dependen del uso o no del proceso de inferencia en una ontología. Si el motor emplea una ontología sin inferencia, entonces los resultados pueden ser menores a los esperados. Para ejemplificar ésto, supongamos que el grafo en la Figura 5.33 es el modelo de entrada para el motor SPARQL. Este modelo es un subgrafo que se tomo de la ontología *búsqueda de recursos digitales*.

A continuación, se listan las declaraciones que están asociadas a las tripletas de la Figura 5.33.

- El recurso *tesis 01 (sird:tesis01-pdf)* tiene dos declaraciones; la primera establece que este recurso pertenece a la clase *tesis (sird:Thesis)* y la segunda indica que la *UAM (sirp:UAM)*es la institución involucrada en este recurso.

- Los recursos *tesis 02 (sird:tesis02-doc)*, *tesis 03 (sird:tesis03-odp)* y *tesis 05 (sird:tesis05-pdf)* tienen estas declaraciones. La *UAM (sirp:UAM)* es la institución involucrada en la *tesis 02* y *tesis 05*. Mientras, el *IPN (sird:IPN)* es la institución en la *tesis 03*.

- El recurso *tesis 04 (sird:tesis04-pdf)* tiene estas declaraciones. Este recurso pertenece a la clase *Tesis (sird:Thesis)* y el *IPN (sirp:IPN)* es la institución involucrada en esta *tesis 04*.

- Los recursos *libro 01 (sird:book01-pdf)*, *libro 03 (sird:book03-odp)* y *libro 05 (sird:book05-pdf)* tienen las siguientes declaraciones. *Mc Graw Hill (sirp:McgrawHill)* es la editorial del recurso *libro 01*. Mientras, los recursos *libro 03* y *libro 05* tienen a la editorial *Oceano (sirp:Oceano)*.

- El recurso *libro 02 (sird:book02-doc)* es un recurso que pertenece a la clase *libro (sird:Book)* y la editorial de éste es *Mc Graw Hill (sirp:McgrawHill)*.

- El recurso *libro 04 (sird:book04-docx)* tiene estas declaraciones. Este recurso pertenece a la clase *libro (sird:Book)* y *Oceano (sirp:Oceano)* es la editorial de éste.

- Los recursos *artículo 01 (sird:paper01-pdf)* y *artículo 02 (sird:paper02-odp)* son publicados en la *IEEE (sirp:IEEE)* y la *ACM (sirp:ACM)* respectivamente.

- El recurso *artículo 03 (sird:paper03-doc)* tiene dos declaraciones. La primera establece que este recurso se publico en la *IEEE (sirp:IEEE)*. Mientras, la segunda indica que éste pertenece a la clase *Artículo (sird:Paper)*.

- El recurso *artículo 04 (sird:paper04-docx)* pertenece a la clase *artículo (sird:Paper)* y se publico en la *ACM (sirp:ACM)*.

- La clase *Documento (sird:Document)* es superclse de las clases *artículo (sird:Paper)*, *libro (sird:Book)* y *Tesis (sird:Thesis)*.

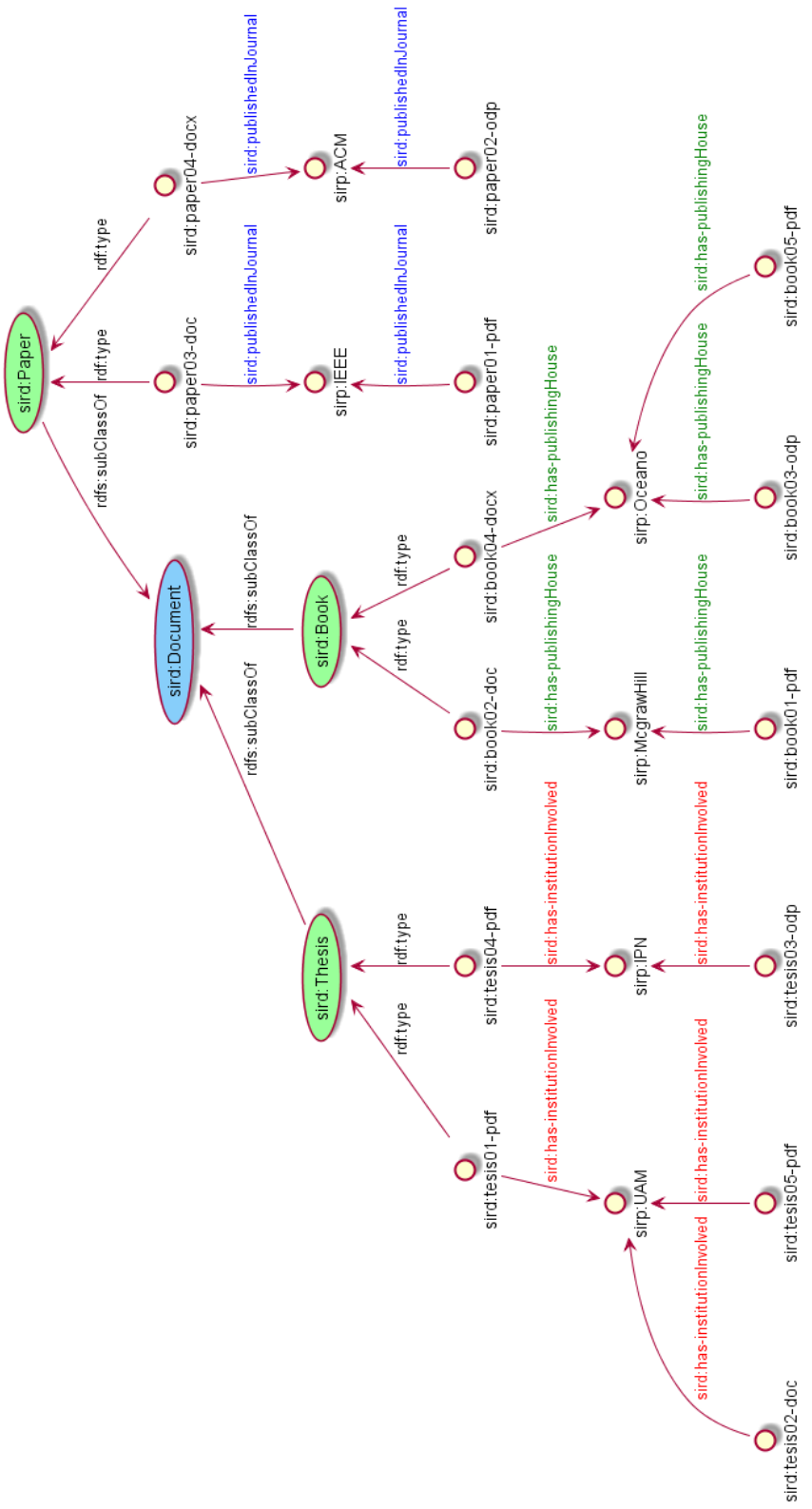

Figura 5.33: Ejemplo de modelo sin inferencia para el proceso de consulta de información.

```
PREFIX sird: <http://arte.izt.uam.mx/ontologies/digiResourceRyT.owl#>
PREFIX rdf: <http://www.w3.org/1999/02/22-rdf-syntax-ns#>

SELECT  ?x
    WHERE
    {
            {?x rdf:type sird:Paper.} UNION
            {?x rdf:type sird:Book.} UNION
            {?x rdf:type sird:TechnicalReport.} UNION
            {?x rdf:type sird:Thesis.} UNION
            {?x rdf:type sird:Webpage.} UNION
            {?x rdf:type sird:Document.}
    }
```

Figura 5.34: Consulta de ejemplo para el proceso de consulta de información.

La consulta a interpretar por el motor SPARQL, es la asociada a la pregunta *¿Cuáles son los recursos que son documentos?*. Esta consulta se presenta en la Figura 5.34 y emplea seis patrones para establecer cuales recursos son documentos (artículos, libros, tesis, reportes técnicos, páginas web).

Un motor SPARQL interpreta esta consulta, compara estos patrones con el grafo de la Figura 5.33 y arroja una serie de resultados. La Tabla 5.7 muestra los identificadores de los recursos que son documentos.

?x
sird:tesis01-pdf
sird:tesis04-pdf
sird:book02-doc
sird:book04-docx
sird:paper01-pdf
sird:paper02-odp

Tabla 5.7: Identificadores de los recursos resultantes sin inferencia para la consulta de información.

Esta *ejecución de la consulta* por un motor SPARQL arrojo seis resultados. Sin embargo, existen otros recursos que son documentos y no tienen la declaración para indicar que pertenece a la clase *Documento* o a las subclases de ésta. De tal manera, es necesario el uso de un modelo con inferencia. La Figura 5.35 presenta el modelo que se obtiene por inferencia. En este modelo, todos los recursos están asignado a la clase *Documento (sird:Document)*, por estas razones:

- Todos los recursos que pertenecen a las clases *artículo (sird:Paper)*, *libro (sird:Book)* y *tesis (Thesis)*, también pertenecen a la clase *documento (Documento)* , por los axiomas de jerarquía (rdfs:subClassOf).

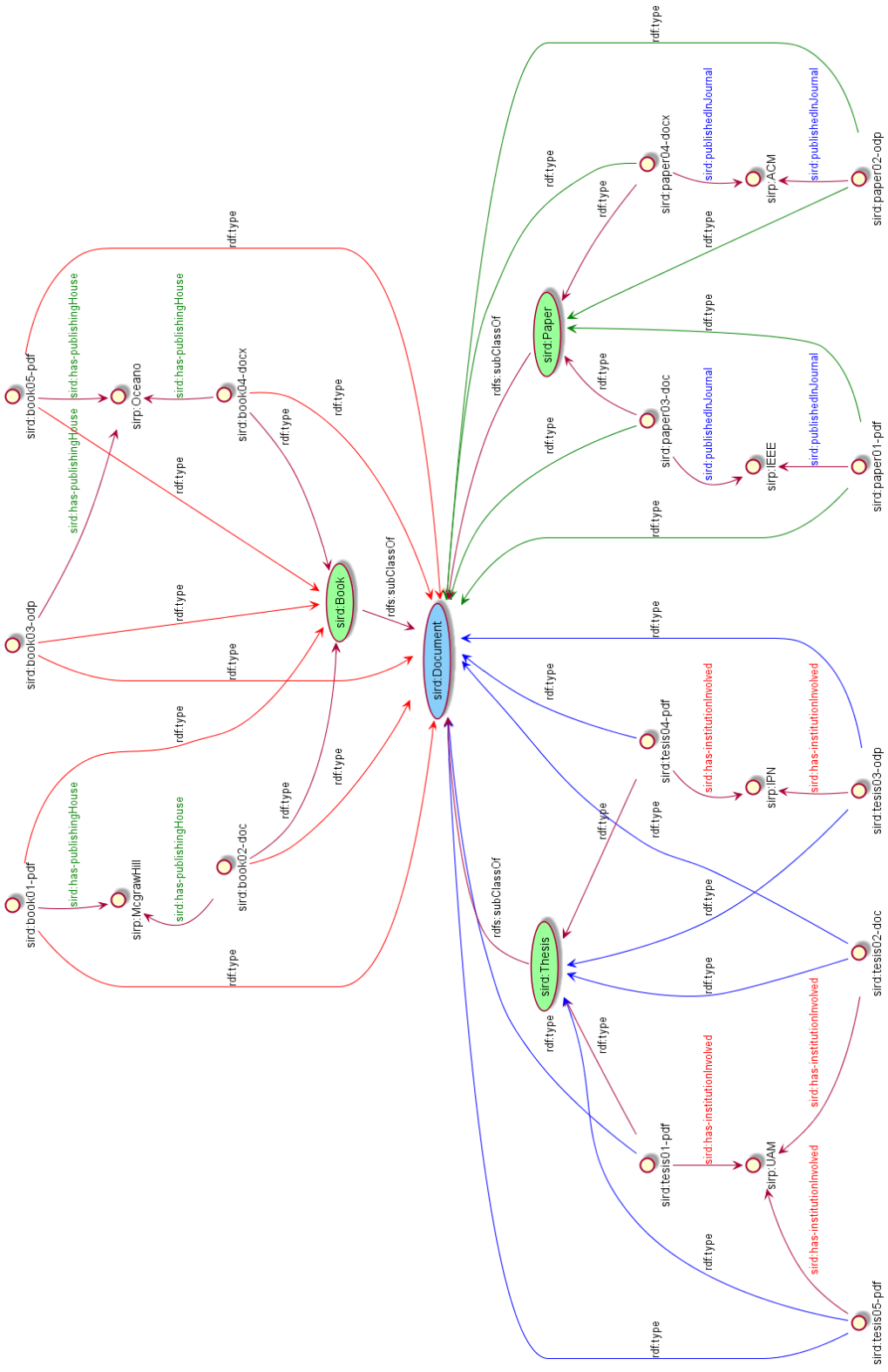

Figura 5.35: Ejemplo de modelo con inferencia para el proceso de consulta de información.

- La propiedad *publicado en la revista (publishedInJournal)* tiene por *dominio* la clase *artículo (sird:Paper)*. Por ello, los recursos que utilizan esta propiedad, pertenecen a la clase Artículo, además la clase *artículo* es subclase de *documento*.
- De igual manera, las propiedades *tiene institución involucrada (sird:has-institutionInvolved)* y *tiene editorial (sird:has-publishingHouse)* tienen por dominio las clases *Tesis (sird:Thesis)* y *libro (sird:Book)* respectivamente. Estas dos clases son subclases de *documento*.

Al darse por hecho el uso de inferencia, la consulta 5.34 puede simplificarse a un solo *patrón tripleta*. Este patrón indica que se busca cualquier recurso que sea del tipo *documento (sird:Document)*. En la Figura 5.36 se presenta la consulta simplifica para la pregunta *¿Cuáles son los recursos que son documentos?*.

```
PREFIX sird: <http://arte.izt.uam.mx/ontologies/digiResourceRyT.owl#>
PREFIX rdf: <http://www.w3.org/1999/02/22-rdf-syntax-ns#>

SELECT  ?x
    WHERE
     {   ?x rdf:type sird:Document. }
```

Figura 5.36: Consulta simplificada para el proceso de consulta de información.

Finalmente, la Tabla 5.8 presenta los resultados arrojados por un motor SPARQL para la consulta de la Figura 5.36 y empleando el grafo de la Figura 5.35. En esta Tabla, todos los identificadores de los recursos son los recursos que se esperan para responder esta pregunta.

?x
sird:tesis01-pdf
sird:tesis02-doc
sird:tesis03-odp
sird:tesis04-pdf
sird:tesis05-pdf
sird:book01-pdf
sird:book02-doc
sird:book03-odp
sird:book04-docx
sird:book05-pdf
sird:paper01-pdf
sird:paper02-odp
sird:paper03-doc
sird:paper04-docx

Tabla 5.8: Identificadores de los recursos resultantes a partir del uso de inferencia para la consulta de información.

Capítulo 6
Prototipo

Cualquier usuario no puede realizar el proceso de consulta (interacción) de información (integración ISR) en las ontologías de la *cartografía de competencias* y *búsqueda de recursos digitales*. Porque los usuarios deben tener un conocimiento básico en: las tecnologías semánticas (tripletas y consultas SPARQL), el vocabulario utilizado en las tripletas (recursos y propiedades), manejo del triplestore (carga, inferencia).

La integración semántica de recursos (ISR) a partir del uso de un triplestore, no es una tarea que cualquier usuario (profesor o estudiante) puede hacer, ya que éste debe estar familiarizado con el triplestore, el lenguaje de consulta SPARQL y las ternas RDF.

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.

La integración semántica es un proceso de búsqueda y recuperación de información sobre los recursos de una Memoria Corporativa, bajo el enfoque de las Tecnologías Semánticas. Esta integración se basa en la representación del conocimiento de los recursos en forma de un grafo RDF y en la consulta del mismo, para satisfacer la pregunta de un usuario. En esta integración semántica, el grafo se constituye de triples y las consultas están están compuestas de patrones parecidos a triples. Ahora bien, cualquier usuario del área de Redes y Telecomunicaciones (RyT) que quiera consultar el modelo RDF, debe aprender a construir consultas SPARQL y tener un conocimiento general sobre los triples del modelo RDF. Sin embargo, esta situación puede ser molesta para los usuarios. Así que, nosotros proponemos un prototipo para la interacción amigable de los usuarios con el modelo RDF, de esta forma, un usuario podrá consultar al modelo RDF y visualizar los resultados asociados a la misma, sin que éste tenga conocimientos en las Tecnologías Semánticas. Los objetivos particulares del prototipo son: 1) permitir a los usuarios estructurar su pregunta, para que se mapee a una consulta SPARQL. 2) cargar el modelo RDF y los axiomas, e invocar el razonador para inferir nuevas relaciones al grafo RDF. 3) invocar el motor de consulta SPARQL, para que ejecute la consulta SPARQL al grafo RDF inferido. 4) mostrar para cada resultado un conjunto de datos significativos (nombre, ruta, lenguaje, etc.). A partir de estos objetivos se plantea la siguiente arquitectura.

Capítulo 7
Evaluación experimental

La *integración semántica de los recursos* es el proceso de búsqueda y recuperación de información en un *grafo de conocimiento (ontología)*. Un *motor de búsqueda de tripletas* es el mecanismo encargado de realizar la consulta de información en los grafos de conocimiento, para responder una consulta dada. Este *motor de tripletas* generalmente pertenece a un triplestore. En esta tesis, se emplea el *triplestore Jena*. Éste proporciona dos componentes importantes: 1) *un motor de búsqueda* para tripletas RDF, denominado ***ARQ*** y 2) *un motor de inferencia* que soporta los axiomas de nuestras ontologías, los cuales son descritos en la Sección 5.2.

La *calidad de los resultados* depende del uso o no de inferencia en nuestros modelo semántico. Si Jena emplea un *modelo sin inferencia* como el ejemplo de la Figura 5.33, entonces el motor ARQ puede proporcionar todos, varios o ningún resultado. Esta variedad en la entrega de resultados, depende de las consultas SPARQL: 1) *consultas sobre las declaraciones de los recursos*, como la consulta de la Figura 5.27, 2) *consultas que agrupan varios patrones para un criterio de búsqueda*, como las consultas de las Figuras 5.28 y 5.30, y 3) *consultas simplificadas*, como en las Figuras 5.29 y 5.31. En contraste, Jena puede entregar mejores resultados cuando emplea un modelo que se obtiene de la inferencia en una ontología. Un ejemplo de este modelo se presenta en la Figura 5.35.

Una característica asociada al uso de inferencia, es el impacto en el *tiempo de procesamiento* para responder las consultas. Por un lado, se ha observado que este *tiempo* es pequeño (menor a medio segundo) cuando se usa un *modelo sin inferencia*. Mientras tanto, el *tiempo para un modelo con inferencia* es mayor en comparación con con el tiempo del modelo sin inferencia.

Con base en estas observaciones, nuestras dos hipótesis de experimentación son éstas:

1. *El* ***triplestore Jena*** *obtiene* ***mejores resultados*** *cuando utiliza nuestros modelos con inferencia.*

2. *El* ***tiempo de consulta*** *es* ***mayor*** *para nuestros modelos con inferencia en comparación con nuestros* ***modelos sin inferencia****.*

Esta experimentación consiste en la realización de dos actividades para probar nuestras hipótesis de experimentación. *La* ***primera actividad*** *es* ***evaluar la calidad de los resultados*** *para los modelos con y sin inferencia.* Esta evaluación consiste en estas etapas: 1)

establecer una serie de consultas para interrogar nuestros modelos, 2) encontrar manualmente cuántos y cuáles recursos responden las consultas, 3) ejecutar las consultas con el motor ARQ de Jena y 4) comparar los recursos dados por Jena con las respuestas manuales.

*La **segunda actividad** consiste en medir los **tiempos promedio de procesamiento** para las consultas de la primera actividad.* La finalidad de esta segunda actividad es comparar los tiempos de procesamiento para un modelo con inferencia y otro que no emplea ésta. La determinación del tiempo para modelos sin inferencia, consiste en medir los tiempos de: 1) *ejecución de la consulta en en el modelo* y 2) *recuperación de la información.* De la misma manera, la medición de tiempos para un modelo con inferencia es parecida a la medición en un modelos sin inferencia. La excepción es que en un modelo con inferencia, se toman en cuenta los tiempos para: *el proceso de inferencia en el modelo* y *la ejecución de la consulta al modelo inferido.*

En esta tesis, el proceso de *integración semántica* está asociado a dos *casos de uso* (cartografía de competencias y búsqueda de recursos digitales). Ahora bien, nuestra experimentación consiste en probar la *calidad de los resultados* y el *tiempo de procesamiento* para la *integración semántica de recursos* en la *memoria corporativa* del área de Redes y Telecomunicaciones. Por esta razón, las dos actividades de nuestra experimentación deben ser aplicadas a nuestros dos *casos de uso.*

Los sujetos de nuestra experimentación son un conjunto de personas, documentos y archivos multimedia que son generados artificialmente. Esta *generación artificial* consiste en el uso de scripts para: 1) *asignar un identificador URI para un conjunto de **recursos de información ficticios*** y 2) *generar tripletas RDF para estos recursos con base en las **propiedades** y **clases** de nuestras ontologías, así como **datos aleatorios**.*

Un script genera un conjunto de declaraciones para los recursos persona. Mientras, otro script genera las declaraciones para los documentos y archivo multimedia. El algoritmo 1 presenta el funcionamiento general de ambos scripts para la generación y almacenamiento de tripletas RDF.

1 $N \leftarrow$ número de *recursos ficticios de información* a describir;
2 $modelo_{rdf} \leftarrow$Crear un modelo rdf;
3 **for** $i \leftarrow 1$ **to** N **do**
4 | $\sigma_i \leftarrow$Crear el recurso i y establecer un identificador URI para éste;
5 | Elaborar los valores para cada característica significativa de este recurso (σ_i);
6 | Escribir las aserciones, concatenando el URI del recurso (σ_i), las propiedades de la ontología y los valores del paso 5;
7 **end**
8 Guardar el $modelo_{rdf}$ en un archivo con extensión "*rdf*" y sintaxis de serialización *Turtle*;

Algorithm 1: Funcionamiento básico de scripts para la generación de tripletas artificialmente

El apéndice A presenta los dos algoritmos con el funcionamiento detallado de los scripts.

Un algoritmo para los datos simulados de los recursos persona y el otro para los recursos digitales.

La finalidad del uso de *información simulada* es tener rápidamente un volumen grande de datos en nuestros ABox. La *cantidad de información* en estos ABox debe ser realista con respecto al área de Redes y Telecomunicaciones (RyT). Ya que al tener información realista, nuestra experimentación se ajusta a la cantidad de datos que esperamos manejar en la integración semántica. Otra razón del uso de información simulada es ver si Jena soporta esta escala realista de datos (según los profesores del área RyT).

El número de

La cantidad de recursos persona y digitales

se construyo a partir de un cuestionario que se hizo a varios profesores del área de redes y telecomunicaciones.

—->Aquí voy

Este proyecto contempla dos casos de uso: la Cartografía de Competencias (C1) y la Búsqueda de Recursos Digitales (C2).

Los recursos persona agrupan a los miembros del Área de Redes y Telecomunicaciones. Los recursos digitales conjuntan todos los documentos digitales y archivos multimedia que emplean los miembros del área.

Los recursos persona (C1) se clasifican en cuatro tipos básicos: Profesor, Empleado, Investigador y Estudiante. Un profesor es aquella persona que imparte docencia y está adscrito a una Universidad. Un empleado es aquella persona que trabaja en una organización y realiza otras actividades distintas a la docencia. Un investigador es aquella persona que realiza actividades de investigación científica. Finalmente un estudiante es la persona que estudia en una Universidad y realiza algún proyecto (terminal, investigación o doctoral).

Estas cuatro agrupaciones básicas se pensaron para ser conjuntos no disjuntos. Por este motivo, un persona puede ser estudiante y también empleado, inclusive una persona puede pertenecer a los cuatro grupos básicos. En general, un recurso persona puede estar en varias agrupaciones, siempre y cuando, este recurso tenga explícitamente las aserciones que lo vincule con más de un tipo básico.

Por otro lado existen personas que no pertenecen a los conjuntos básicos, por tal razón, estos individuos no poseen ninguna aserción (rdf:type) a un tipo básico. Este hecho es importante, porque en la búsqueda de información hay personas que carecer de la propiedad tipo, de esta manera, estas personas pueden ser excluidas de los resultados de la búsqueda.

Nuestra experimentación con los recursos persona se basa en la asignación explícita de solo uno de los cuatro tipos básicos o la carencia de esta asignación. En concreto, se tienen 73 recursos persona de los cuales: 51 son son profesores o empleados (profesionista) y 23 estudiantes. De estas 51 personas profesionistas: 19 son profesores, 9 empleados y el resto (33) no están catalogados pero sabemos que trabajan. Ahora bien de los 23 estudiantes se afirma explícitamente que 9 son estudiantes y el resto (14) solo se sabe que estudian, porque en descripción RDF se tiene la propiedad ?estudiaEn?. Con respecto a los investigadores, explícitamente ninguna persona se identifica como investigador pero sabemos que 13 hacen investigación (en su descripción RDF tienen la propiedad investiga-sobre).

Nosotros para representar mejor los valores de la tabla 1, decidimos hacer un diagrama de Venn. La figura 1 muestra los 4 tipos básicos, así como otras agrupaciones a partir de estos 4 tipos. También este diagrama contempla aquellos recursos que se deducen a partir del razonador. Específicamente, aquellas personas que estudian y cuya propiedad ?rdf:type? no es explícita, a través del razonador se infiere que estos son estudiantes. De la misma manera, aquellos recursos que no estamos seguros si pertenecen a uno de los 4 tipos básicos, empleando el razonador se puede inferir si son profesionistas o personas.

Los Recursos Digitales (C2) se clasifican en ocho grupos: Artículos, Reportes Técnicos, Páginas Web, Tesis, Libros, Presentaciones, Imágenes, Audios y Videos. La clasificación de los recursos digitales se hace con base en las características de los mismos. Un artículo es una documento de carácter científico que es publicado en revistas o ponencias. Un reporte técnico es un documento que informa el estatus técnico de un problema. Una Página Web es un documento que emplea un lenguaje de etiquetado, típicamente html. Un tesis es un documento sobre un tema particular que se emplea para obtener algún grado académico. Un libro es un documento literario o científico que debe poseer más de 49 páginas [1]. Una presentación es un recurso multimedia empleado para mostrar visualmente información sobre un determinado tema, puede contener imágenes, videos, y oraciones cortas. Una Imagen es un recurso multimedia de contenido visual. Un Audio es un recurso multimedia de contenido auditivo y típicamente se emplea para: grabar una conversación importante o como archivos de prueba. Un Video es un recursos multimedia de contenido audiovisual.

Estas ocho agrupaciones son conjuntos disjuntos, por tanto, ningún recurso digital debe estar en más de una agrupación de las ocho básicas. Ésto significa que los recursos digitales tienen explícitamente sólo una asignación de las ocho básicas. Por otra parte, estos recursos digitales también pueden carecer de esta asignación. De esta manera, los recursos digitales pueden tener una o ninguna asignación de los ocho agrupaciones básicas. Estos ocho tipos básicos se agrupan en dos tipos generales. Por un lado, los documentos conjuntan los artículos, libros, reportes técnicos, páginas web, tesis y también todo aquel recurso escrito. Mientras los recursos multimedia conjuntan audios, videos, imágenes y presentaciones.

Nuestra experimentación cuenta con 1330 recursos digitales. Estos recursos se dividen en: 156 artículos, 366 libros, 34 reportes técnicos, 146 páginas web, 73 tesis, 42 videos, 42 audios, 77 imágenes y 112 presentaciones. Para los artículos hay 89 recursos que explícitamente tienen esta asignación. Las páginas web, libros y tesis respectivamente tienen 79, 185 y 31 recursos que explícitamente tienen la asignación del tipo. El resto de recursos digitales no poseen explícitamente la asignación del tipo. Pero, nosotros sabemos que hay 815 documentos y 515 recursos multimedia. La tabla 2 lista los tipos de recursos digitales y el número de recursos que tiene cada tipo.

A partir de estos números y las aserciones de todos los recursos, se hace un listado de consultas estáticas. La finalidad de estas consultas es interrogar la información de los recursos a partir de sus aserciones. Estas consultas se escriben en lenguaje natural y para cada una se enumeran los recursos que las responden. En particular, se tiene una lista de preguntas para los recursos persona (C1) y una lista para los recursos digitales (C2). La tabla 3 contiene las consultas sobre los recursos persona y para cada consulta se muestra el número de recursos

que la responden. Mientras la tabla 4 lista las consultas para los recursos digitales, así como el número recursos que responden a estas consultas.

Para evaluar este costo en tiempo, calculamos el tiempo promedio que tarda cada consulta en responderse. Específicamente, nosotros tomamos el tiempo desde que se consulta la información del modelo hasta que se presentan los resultados en pantalla. Esta operación la repetimos veinte veces por consulta, de esta manera, sacamos el tiempo promedio por consulta de nuestro modelo.

Esta medición del tiempo promedio se hace a partir de un script en Java. Este script se ejecuta en una computadora que tiene un procesador Intel core I7 a 2.3GHz con 8Gb en ram y 8 núcleos de procesamiento. Las siguientes dos tablas presentan los tiempos de respuesta para nuestro conjunto de consultas. En la tabla 7 se muestran los valores de las consultas para el caso de uso Cartografía de Competencias, en tanto, la tabla 8 muestra los tiempos para el caso de uso Recursos Digitales. En ambas tablas, se contemplan los tiempos de respuesta y el número de resultados para las consultas que emplean sólo con ABox, así como para las consultas que utilizan ABox, TBox y un Razonador (ATR).

——— En esta experimentación hay 73 personas y 1330 recursos digitales. Nosotros escribimos manualmente los triples de 11 personas que están adscritas al Departamento de Ingeniería Eléctrica de la Universidad Autónoma Metropolitana y los triples de las otras 60 fueron generados artificialmente. Por otro lado, nosotros escribimos manualmente los triples de 16 recursos digitales, mientras que los triples de los otros 1314 fueron generados artificialmente.

Las 73 personas se clasifican en: Profesor, Estudiante, Empleado, Profesionista, Investigador y Persona. Mientras los 1330 recursos digitales se clasifican en: Artículos, Reportes Técnicos, Páginas Web, Tesis, Libros, Presentaciones, Imágenes, Audios, Videos, Documento, Multimedia y Recurso Digital. En particular, para cada clase de Persona y Recursos Digital se tienen las siguientes cantidades:

Para los recursos persona en esta experimentación se tienen 1750 triples. Mientras, para los recursos digitales se tienen 20429 triples. De esta manera, el número total de triples en esta experimentación son 22179.

Por otro lado, basándonos en las consultas básicas para nuestros modelos, el análisis de los triples escritos manualmente y el uso de variables contador en los dos scripts, se identificaron para cada consulta el número de recursos que responden a la misma. En la tabla 3 se listan solamente 10 de nuestras 28 preguntas y el número de resultados de las mismas. ———

Ahora bien, la finalidad de este listado es tomar los tiempos promedios y el número de respuestas, cuándo las consultas SPARQL se hacen con el motor sin razonador y con razonador. Con la finalidad de averiguar la precisión [5], así como el desempeño del motor de SPARQL y el razonador de Jena.

Las dos variables de experimentación son tiempo y número de resultados. La variable tiempo nos permite sacar el tiempo promedio que toma una consulta en ser ejecutada K veces. Mientras la variable número de resultados almacena la cantidad de recursos que fueron recuperados para una consulta dada.

Para encontrar los valores de estas variables, nosotros empleamos un programa en Java.

Este programa se ha diseñado para ejecutar únicamente una consulta e imprimir los valores de las variables en pantalla. El programa permite elegir al usuario el modelo RDF a consultar. Si es modelo RDF con triples explícitos, entonces solo cargan los triples (ABox) de los recursos. Por el contrario, si el modelo es con triples inferidos, entonces se cargan los triples (ABox), los axiomas (TBox) y se hace inferencia con el razonador de Jena. Este programa se ejecutó en una computadora con las siguientes capacidades: Procesador Intel Core I7 a 2.3GHz con 8Gb en RAM y 8 núcleos de procesamiento, y los valores resultantes de las variables se muestran en la Tabla 2.

——————— En esta tabla se tienen dos columnas compuestas, la columna (titulada conocimiento explícito) muestra los valores de las consultas que emplearon únicamente el modelo con triples explícitos. Mientras, la segunda columna (titulada conocimiento inferido) muestra los valores de las consultas que emplearon el modelo con triples explícitos, axiomas y un razonador. Para las columnas sencillas, la columna ?Número de resultados? muestra el número de recursos recuperados del total esperado para la consulta dada, mientras la columna ?Tiempo promedio? muestra el tiempo promedio de consulta en milisegundos.

En algunos casos, la consulta al conocimiento explícito recupera todos los recursos esperados y los tiempos de respuesta son pequeños (no pasan del segundo). Sin embargo, en otras consultas se descartaron varios recursos que si responden la consulta. Esto se debe a que algunos recursos carecen un determinado triple. Por otro lado, las consultas al grafo con triples inferidos permitieron recuperar todos los recursos esperados, porque mediante los axiomas y el razonador se deducen triples (materializaron) que serán considerados por el motor de búsqueda . Sin embargo, el tiempo de procesamiento es mucho mayor porque se invierte tiempo en procesar e inferir relaciones en el grafo RDF.

Todo tiene un costo, cuando el razonador materializa los triples en el modelo, éste consume tiempo en procesamiento y al hacer una consulta, el motor debe comparar más aserciones. El desarrollador no debe abusar de la axiomatización, en algunos casos cuando la consulta es sobre hechos explícitos, no es necesario el uso del razonador, basta con escribir y hacer la consulta sobre el conocimiento explícito. ———————

La primer conclusión afirma que el performance mejora cuando se usan axiomas. Esta afirmación resulta cierta, porque un razonador deduce una relación que vincula directamente dos objetos. Análogamente, resulta más rápido ir por el camino directo que por una serie de rutas hasta el mismo objeto.

La segunda conclusión tiene que ver con el número de resultados. Si bien, las aserciones establecen un conjunto directo y estático de enlaces entre los distintos recursos de nuestro modelo. En muchas ocasiones, al momento de construir una consulta SPARQL no se contemplan algunos de estos enlaces, inclusive en otros casos, estos enlaces no están escritos explícitamente. Por consiguiente, mucho recursos no se contemplan como respuesta para una consulta. En contraste, los axiomas, aserciones y un razonador, establecen estos enlaces entre recursos de forma explícita en memoria, de esta manera, las consultas respondan más resultados que no se habían contemplado.

Funcionamiento de los scripts

7.1. Escenarios de experimentación

Algún texto...

Id. Consulta	Pregunta	No. de Recursos
Q1	¿Cuáles son los títulos, rutas, extensión, idioma de todos los recursos digitales de RyT?	1330
Q2	¿Cuáles libros tratan sobre algunos temas de Sistemas Distribuidos?	103
Q3	¿Qué recursos fueron publicados por la UAM?	18
Q4	¿Qué documentos son para dar un curso de Sistemas P2P?	31
Q5	¿Qué recursos multimedia son mayores al año 2009?	119
Q6	¿Cuáles documentos tratan sobre Ontologías?	30
Q7	¿Qué recursos fueron publicados en una Revista científica?	156
Q8	¿Qué recursos tienen en su contenido las palabras "linked data¿	159
Q9	¿Cuáles documentos en inglés y mayores al año 2000 son de autoría de Erik Alarcón Zamora?	2
Q10	¿Cuáles la tesis de Samuel Hernández Maza?	4

7.2. Experimentación

Más texto...

7.3. Resultados

Más texto...

La integración semántica consiste en la búsqueda y recuperación de la información de los recursos. En este punto, es importante evaluar esa recuperación y medir los tiempos de procesamiento para un modelo con razonamiento en Jena. Para evaluar la recuperación de los recursos consiste en enumerar los recursos que responden a una consulta. Mientras, el tiempo promedio de procesamiento es la media en tiempo (milisegundos) que tarda una consulta en ejecutarse. En esta experimentación hay 73 personas y 1330 recursos digitales. Nosotros escribimos manualmente los triples de 11 personas que están adscritas al Departamento de Ingeniería Eléctrica de la Universidad Autónoma Metropolitana y los triples de

Id. Consulta	Modelo (ABox)		Modelo (Razonador+Ontología)	
	Tiempo promedio (ms)	No. Recursos	Tiempo promedio (ms)	No. Recursos
Q1	12	1330/1330	138	1330/1330
Q2	10	0/103	194	103/103
Q3	8	18/18	406	18/18
Q4	28	15/31	129	31/31
Q5	7	66/119	157	119/119
Q6	9	15/30	4016	30/30
Q7	12	156/156	3520	156/156
Q8	16	159/159	3472	159/159
Q9	42	0/2	3451	2/2
Q10	13	3/4	3312	4/4

las otras 60 fueron generados artificialmente. Por otro lado, nosotros escribimos manualmente los triples de 16 recursos digitales, mientras que los triples de los otros 1314 fueron generados artificialmente. Las 73 personas se clasifican en: Profesor, Estudiante, Empleado, Profesionista, Investigador y Persona. Mientras los 1330 recursos digitales se clasifican en: Artículos, Reportes Técnicos, Páginas Web, Tesis, Libros, Presentaciones, Imágenes, Audios, Videos, Documento, Multimedia y Recurso Digital. En particular, para cada clase de Persona y Recursos Digital se tienen las siguientes cantidades: Para los recursos persona en esta experimentación se tienen 1750 triples. Mientras, para los recursos digitales se tienen 20429 triples. De esta manera, el número total de triples en esta experimentación son 22179. Por otro lado, basándonos en las consultas básicas para nuestros modelos, el análisis de los triples escritos manualmente y el uso de variables contador en los dos scripts, se identificaron para cada consulta el número de recursos que responden a la misma. En la tabla 3 se listan solamente 10 de nuestras 28 preguntas y el número de resultados de las mismas. Ahora bien, la finalidad de este listado es tomar los tiempos promedios y el número de respuestas, cuándo las consultas SPARQL se hacen con el motor sin razonador y con razonador. Con la finalidad de averiguar la precisión [5], así como el desempeño del motor de SPARQL y el razonador de Jena. Las dos variables de experimentación son tiempo y número de resultados. La variable tiempo nos permite sacar el tiempo promedio que toma una consulta en ser ejecutada K veces. Mientras la variable número de resultados almacena la cantidad de recursos que fueron recuperados para una consulta dada. Para encontrar los valores de estas variables, nosotros empleamos un programa en Java. Este programa se ha diseñado para ejecutar únicamente una consulta e imprimir los valores de las variables en pantalla. El programa permite elegir al usuario el modelo RDF a consultar. Si es modelo RDF con triples explícitos, entonces solo cargan los triples (ABox) de los recursos. Por el contrario, si el modelo es con triples

inferidos, entonces se cargan los triples (ABox), los axiomas (TBox) y se hace inferencia con el razonador de Jena. Este programa se ejecutó en una computadora con las siguientes capacidades: Procesador Intel Core I7 a 2.3GHz con 8Gb en RAM y 8 núcleos de procesamiento, y los valores resultantes de las variables se muestran en la Tabla 2.

Interpretación de resultados En esta tabla se tienen dos columnas compuestas, la columna (titulada conocimiento explícito) muestra los valores de las consultas que emplearon únicamente el modelo con triples explícitos. Mientras, la segunda columna (titulada conocimiento inferido) muestra los valores de las consultas que emplearon el modelo con triples explícitos, axiomas y un razonador. Para las columnas sencillas, la columna ?Número de resultados? muestra el número de recursos recuperados del total esperado para la consulta dada, mientras la columna ?Tiempo promedio? muestra el tiempo promedio de consulta en milisegundos. En algunos casos, la consulta al conocimiento explícito recupera todos los recursos esperados y los tiempos de respuesta son pequeños (no pasan del segundo). Sin embargo, en otras consultas se descartaron varios recursos que si responden la consulta. Esto se debe a que algunos recursos carecen un determinado triple. Por otro lado, las consultas al grafo con triples inferidos permitieron recuperar todos los recursos esperados, porque mediante los axiomas y el razonador se deducen triples (materializaron) que serán considerados por el motor de búsqueda . Sin embargo, el tiempo de procesamiento es mucho mayor porque se invierte tiempo en procesar e inferir relaciones en el grafo RDF. Todo tiene un costo, cuando el razonador materializa los triples en el modelo, éste consume tiempo en procesamiento y al hacer una consulta, el motor debe comparar más aserciones. El desarrollador no debe abusar de la axiomatización, en algunos casos cuando la consulta es sobre hechos explícitos, no es necesario el uso del razonador, basta con escribir y hacer la consulta sobre el conocimiento explícito.

Documento importante

Capítulo 8
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 explicita 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 explicita 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
Algoritmos para le generación de datos simulados

Algoritmo para generar datos artificialmente de los recursos persona

1. $N \leftarrow$ número de personas a describir;
2. Tópicos de Redes y Telecomunicaciones $TopRyT \leftarrow \{\theta_1 \dots \theta_k\}$;
3. **for** $i \leftarrow 1$ **to** N **do**
   - *a*) Elegir el nombre, apellido paterno y apellido materno de unas listas predefinidas;
   - *b*) Concatenar el nombre y apellidos electos, donde cada palabra sea separada por un espacio en blanco;
   - *c*) Guardar este nombre en una lista tipo cola ($Names$);
4. Lista de Nombres $Names \leftarrow \{nombre_1 \dots nombre_N\}$;
5. $modelo_{rdf} \leftarrow$ Crear un modelo rdf para los recursos persona;
6. **for** $i \leftarrow 1$ **to** N **do**
   - *a*) $\sigma_i \leftarrow$ crear el recurso i y establecer un identificador URI para éste;
     1) Seleccionar el i-ésimo nombre de la lista de nombres ($nombre_i$);
     2) Quitar los espacios en blanco de esta cadena;
     3) Establecer esta cadena como identificador del recurso;
   - *b*) Para cada característica significativa de una persona elaborar las literales y elegir los objetos de ésta;
     1) Literal *nombre* $\leftarrow$ seleccionar el i-ésimo nombre de la lista de nombres ($nombre_i$);
     2) Objeto *Género* $\leftarrow$ elegir el sexo de la persona a partir del nombre;
     3) Literal *Año* $\leftarrow$ establecer el año aleatoriamente en un intervalo del 2000 al 2013;

4) Objeto *Ocupación* ← elegir una ocupación para la persona de las tres posibles (Estudiante, Profesor o Empleado);
   /** *Establecer la edad a partir de la ocupación de la persona* **/
5) **if** (Ocupación ≡ 'Empleado' ‖ Ocupación ≡ 'Profesor') **then**
   Literal *Edad* ← un número aleatorio en el intervalo de 27 a 57;
6) **else if** (Ocupación ≡ 'Estudiante') **then**
   Edad ← un número aleatorio en el intervalo de 18 a 28;
   /** *Elegir una organización de un listado a partir de la ocupación de la persona* **/
7) **if** (Ocupación ≡ 'Estudiante' ‖ Ocupación ≡ 'Profesor') **then**
   Literal *Organización* ← se elige una universidad de un listado preestablecido de universidades;
8) **else if** (Ocupación ≡ 'Empleado') **then**
   Organización ← se elige una organización de una lista preestablecida de organizaciones;
9) Literal *Email* ← construir el email a partir del nombre de la persona (cambiando los espacios en blanco por guiones bajos), la organización donde labora y otras palabras especiales;
10) Literal *Sitio Web* ← construir el URL concatenando el nombre (cambiando espacios en blanco por guiones), la organización donde labora, la ocupación, una extensión de páginas web y otras palabras especiales;
    /** Elegir otros individuos que se relacionan con esta persona a partir de la ocupación de la persona **/
11) **if** (Ocupación ≡ 'Estudiante') **then**
    Conocidos ← se eligen dos nombres aleatorios de la lista nombres, cuya ocupación de éstos sea Empleado o Profesor;
    Quitar los espacios en blanco de los nombres;
    Guardar estas cadenas en una lista de conocidos;
12) **else if** (Ocupación ≡ 'Estudiante' ‖ Ocupación ≡ 'Profesor') **then**
    Conocidos ← se eligen dos nombres a cinco nombre aleatorios de la lista nombres, cuya ocupación de éstos sea Empleado o Profesor;
    Quitar los espacios en blanco de los nombres;
    Guardar estas cadenas en una lista de conocidos;
13) Objeto *Habilidades lingüísticas* ← se eligen de forma aleatoria de 1 a 3 idiomas de una lista preestablecida de idiomas;
14) Objeto *Conocimientos de RyT* ← se eligen de manera aleatoria de 5 a 7 Tópicos de Redes y Telecomunicaciones ($TopRyT$);
15) Objeto *Competencias profesionales* ← se eligen de forma aleatoria de 3 a 4 competencias de un listado preestablecido de competencias;

c) Para cada característica significativa de una persona construir sus aserciones respectivas.

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 Ribière. Methods and Tools for Corporate Knowledge Management. Technical Report RR-3485, INRIA, September 1998.

[11] John Lyons. *Semántica Lingüística: Una Introducción*. Paidos Iberica, 1997.

[12] Torcoroma Velásquez Pérez, Andrés Puentes Velásquez, and Jaime Guzmán Luna. Ontologias: una tecnica de representacion de conocimiento. *Avances en Sistemas e Informática*, 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 (IIAIAAI), 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] 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.

[38] Holger Knublauch, Ray W. Fergerson, 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.

[39] 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.

[40] 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.

[41] 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.

[42] 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.

[43] B. McBride. Jena: a semantic web toolkit. *Internet Computing, IEEE*, 6(6):55–59, 2002.

[44] Karlis Cerans, Guntis Barzdins, Renars Liepins, Julija Ovcinnikova, Sergejs Rikacovs, and Arturs Sprogis. Graphical schema editing for stardog owl/rdf databases using owlgred/s. In Pavel Klinov and Matthew Horridge, editors, *OWLED*, volume 849 of *CEUR Workshop Proceedings*. CEUR-WS.org, 2012.

[45] 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.

[46] 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.

[47] Ivar Jacobson, James Rumbaugh, and Grady Booch. *El lenguaje unificado de modelado*. ADDISON-WESLEY, 2001.